University of São Paulo Luiz de Queiroz College of Agriculture

Process-based crop models calibration and structure affect data assimilation for estimating sugarcane yield

Izael Martins Fattori Junior

Dissertation presented to obtain the degree of Master in Science. Area: Agricultural Systems Engineering

Piracicaba 2022 Izael Martins Fattori Junior Agronomic Engineer

Process-based crop models calibration and structure affect data assimilation for estimating sugarcane yield

Advisor: Prof. Dr. **FÁBIO RICARDO MARIN**

Dissertation presented to obtain the degree of Master in Science. Area: Agricultural Systems Engineering

Piracicaba 2022

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

Fattori Junior, Izael Martins

Process-based crop models calibration and structure affect data assimilation for estimating sugarcane yield / Izael Martins Fattori Junior. - - Piracicaba, 2022.

70 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Saccharum officinarum L. 2. DSSAT 3. SAMUCA 4. WOFOST 5. Calibração 6. Índice de área foliar I. Título

ACKNOWLEDGEMENTS

To my family, especially my wife, Valentina Fazzolari for her love, support, comprehension, and belief in me during all these years of study. To all my family for their support especially my Mother, Father and Sisters Ana Maria Bussaneli Martins, Izael Martins Fattori, Anaisa Bussaneli Martins and Mariana Bussaneli Martins;

To my professor advisor, Dr. Fábio Ricardo Marin, for his mentoring and assistance during the development of this research and since under graduation (2016), as well as, the friendship, teaching and motivational support to keep on crop modelling research;

To CNpq, by the financial support (24 months) through the Master scholarship (Process n° 132833/2020-4);

To my colleague Murilo Vianna, for his mentoring and assistance during the development of this research;

To professors and employees of "Luiz de Queiroz" College of Agriculture for teaching and support, especially to Davilmar;

To "Luiz de Queiroz" College of Agriculture, University of Sao Paulo, for the infrastructure for experimental field and many possibilities to improve my skills in research and teaching in Biosystem Engineering through the Graduate Program of Agricultural Systems Engineering;

To ESALQ/CLIMA group for tremendous support on field experiment, and especially by collecting and providing dataset of the experiment;

To everyone that contributed to my personal e professional development, helping to finish my thesis and obtain the degree of Master in Agricultural Systems Engineering. Thank you very much

CONTENTS

RESUMO 6
ABSTRACT
1. INTRODUCTION
1.1. Objectives
1.1.1. Specific objectives
References
2. ASSIMILATING LEAF AREA INDEX DATA INTO A SUGARCANE PROCESS-
BASED CROP MODEL FOR IMPROVING YIELD ESTIMATION 17
Abstract
2.1. Introduction
2.2. Material and Methods
2.2.1. Field experiments description
2.2.2. Brief description of DSSAT/SAMUCA and calibration process
2.2.3. Description of data assimilation procedure
2.2.4. Ensemble Kalman filter method 22
2.2.5. Ensemble smoother method
2.2.6. Weighted mean method
2.2.7. Data analysis
2.2.8. Effect of genotype-specific calibration on DA methods
2.2.9. Effect of the number and timing of observations on DA methods
2.3. Results
2.3.1. Influence of using fixed crop parameters in the performance of DA methods across
different cultivars
2.3.2. Comparing data assimilation methods
2.3.3. Analyzing DA methods for different experiments
2.3.4. Impact of leaf area index observations in sugarcane growth stage
2.4. Discussion
2.5. Conclusion
References
3. ASSESSING THE INFLUENCE OF CROP MODEL STRUCTURE ON THE
PERFORMANCE OF DATA ASSIMILATION FOR SUGARCANE 41

Abstract	
3.1. Introduction	41
3.2. Material and Methods	
3.2.1. Field experiments data for calibration	
3.2.2. Field experiments for data assimilation evaluation	
3.2.3. Satellite image pre-processing	
3.2.4. Brief description of DSSAT/SAMUCA	
3.2.5. Brief description of WOFOST	
3.2.6. Calibration process	
3.2.7. Description of data assimilation procedure	47
3.2.8. Ensemble Kalman filter method	
3.2.9. Data analysis	
3.3. Results	
3.3.1. Calibration results	
3.3.2. Data Assimilation Evaluation	
3.4. Discussion	
3.5. Conclusions	
References	
APPENDICES	

RESUMO

Calibração e estrutura de modelos baseados em processos para cana-de-açúcar influenciam na assimilação de dados para estimativa de produtividade

A cana-de-açúcar é uma importante matéria-prima para produção de açúcar e etanol, portanto, estratégias para acompanhar a quantidade e a disponibilidade de cana-de-açúcar são essenciais. Diferentes métodos têm sido desenvolvidos para estimativa de produtividade, destacando-se o uso de modelos de cultura baseados em processos (MBP) junto à assimilação de dados (AD). Pois esses métodos juntos usam duas fontes diferentes de informação, e suas respectivas incertezas para a estimativa de produtividade. Entretanto, inconsistências entre o MBP e os dados assimilados foram relatadas na literatura, o que leva a erros sistemáticos e baixo desempenho das simulações. Essa limitação está estritamente ligada à ausência ou má calibração do MBP e às simplificações dos processos do MBP em representar as características de desenvolvimento da cultura. Portanto, este estudo teve como objetivo avaliar o impacto do uso de uma calibração especifica para um cultivar de cana-de-açúcar em outras cultivares, utilizando três métodos de AD e comparar essa fonte de incertezas com outras. Ademais, avaliar como a diferença entre dois MBP, em relação às estruturas e descrições das características específicas da cana-de-acúcar, afetam o desempenho dos métodos de AD. Para isso, primeiramente, com o DSSAT/SAMUCA (DS) foram simulados 22 experimentos de campo, para quantificar o impacto do uso dos parâmetros calibrados para cultivar (cv. RB867515) nas simulações de experimentos com cultivares não calibradas (cv. NCo376, SP832847, R570, RB72454), para estimar peso de colmo fresco (PCF). Isso foi realizado para três métodos diferentes de AD, Ensemble Kalman filter (EnKF), Ensemble smoother (ES), e Weighted mean (WM) para assimilar o índice de área foliar (IAF) obtido diretamente do campo e comparado com a simulação sem AD (Open-loop, OP). Além disso, analisamos a influência do momento e da quantidade de dados de IAF, para comparar com o impacto da calibração. Em segundo lugar, dois MBP diferentes, em termos de estrutura, um mais detalhado para descrever características da cana-de-açúcar (DS) e outro mais generalista (WOFOST, WO), foram comparados para acessar o desempenho de simular PCF com o uso de EnKF assimilando IAF. O IAF foi obtido dos sensores Landsat 7 ETM+ e 8 OLI, de talhões de um banco de dados de uma usina de açúcar. Portanto, as simulações com o EnKF desses talhões foram comparadas com as simulações OP. Os resultados mostraram que o uso de uma calibração específica de genótipo teve acurácia substancialmente maior em comparação com as não calibradas. A simulação de experimentos com as cultivares não calibradas apresentou um aumento de acurácia maior, para EnkF e ES, porém, o WM teve resultados opostos. Portanto, a acurácia das simulações com AD apresentou uma alta correlação com a acurácia das simulações OP, sendo essa correlação superior a influência do número de observações de IAF assimilados. Nesse sentido, nossos resultados indicaram que o desempenho das calibrações e a estrutura dos MBPs influenciaram as simulações de OP, com DS apresentando desempenho superior ao WO. No entanto, com o AD o desempenho foi limitado pela inconsistência entre o IAF do Landsat e o IAF simulado pelos MBPs, apesar do aumento de acurácia e precisão. Com isso, assimilar IAF obtido pelo Landsat apresentou potencial devido a melhora de estimativa de PCF, entretanto, a melhor descrição dos processos do DS não conseguiu inibir a inconsistência entre o MBP e o IAF assimilado. Portanto, este estudo ressalta que para a utilização da AD, os MBP devem ser previamente calibrados, seguindo as características das cultivares, para garantir um melhor desempenho. Ademais, os MBPs mais específicos podem se beneficiar de sua descrição detalhada para melhorar o desempenho das simulações de OP e consequentemente com AD.

Palavras-chave: Saccharum officinarum L., Sensoriamento remoto, DSSAT, WOFOST, Calibração, Índice de área foliar

ABSTRACT

Process-based crop models calibration and structure affect data assimilation for estimating sugarcane yield

Sugarcane is an important feedstock of sugar and ethanol. Thus, strategies to follow the quantity and the availability of sugarcane are essential. Different methods have been developed for yield estimation and the use of process-based crop models (PBM) with data assimilation (DA) stands out. Due to the capability of using two different sources of information, and their respective uncertainty for crop yield estimation. However, inconsistencies between PBM and the assimilated data were reported in the literature, which led to systematic errors and low performance of the simulations. Such limitation was connected to the absence or poor calibration of PBMs and to the simplification of the PBMs structure to represent crop development traits. Thus, this study aimed to evaluate the impact of using one sugarcane cultivar-based calibration on other cultivars with DA methods and compare this source of uncertainties with others. Moreover, evaluate how the difference between two PBM, regarding their structure of specific sugarcane traits, affects the performance of DA methods. For that, firstly, the DSSAT/SAMUCA (DS) was used to simulate 22 field experiments and quantify the impact of using one cultivarspecific calibration (cv. RB867515) compared to four non-calibrated cultivars (cv. NCo376, SP832847, R570, RB72454), on stalk fresh yield (SFY) predictions. This was performed for three different DA methods, Ensemble Kalman filter (EnKF), Ensemble smoother (ES), and Weighted mean (WM) to assimilate leaf area index (LAI) retrieved from field observation and compared to the PBM simulation without DA (Open-Loop, OP). Moreover, we analyzed the influence of the timing and amount of LAI data, to compare with the impact of calibration. Second, two different PBM, in terms of structure, one more detailed in terms of structure (DS) and other more general (WOFOST, WO), were compared to the performance of simulate SFY with the use of EnKF and LAI. The LAI was retrieved from Landsat 7 ETM+ and 8 OLI, from fields of a sugarmill database. Thus, the simulations with the EnKF of these fields were compared to the OP simulations. The results showed that the use of a genotype-specific calibration had substantially higher accuracy compared to non-calibrated, for the three DA methods. The simulation of noncalibrated cultivar experiments had a higher accuracy increase, for EnkF and ES, however, WM had opposite results. In this regard, the accuracy of the simulations with DA had a high correlation OP simulations accuracy, which was higher than the correlation with the number of LAI observations assimilated. Furthermore, our results indicated that the calibration performance and the structure of the PBMs influenced the OP simulations, with DS showing higher performance, compared to WO. However, with DA the performance was limited by the inconsistency between Landsat LAI and the LAI simulated by the PBMs, despite the improvements. Thus, assimilated Landsat LAI had the potential to improve yield estimation, but the better descriptions of DS did not inhibit the error inconsistency. Therefore, this study emphasized that the use of DA required previously calibrated PBMs regarding cultivar traits to ensure a higher performance. In addition, more detailed PBMs in terms of process description can benefit from their detailed description to improve the performance of OP simulations and consequently with DA.

Keywords: *Saccharum officinarum* L., Remote sensing, DSSAT, WOFOST, Calibration, Leaf area index

1. INTRODUCTION

Brazil is the largest world sugarcane producer, accounting for ca. 38% of the world's production (FAOSTATS, 2022). Brazil allocates a unique position in the world due to the pioneering policies to utilize renewable energy, which results in safety and flexibility in the production of clean energy. Thus, this process of substituting fossil fuels with biofuels resulted in an increase in sugarcane's growing area between 2002 and 2009. After that, the economic crisis and the decrease in the international market sugar price decelerated Brazil's sugar-energy sector (Marin, 2016). Nevertheless, the national policy that aimed to establish the yearly task of decarbonization to the fuel sector, such as RENOVABIO, established 40% of the national energetic matrix as renewable, and 18% of biofuels in the transport sector. Consequently, the sugarcane sector was consolidated as a source of biofuels to the energetic matrix (Marin et al., 2019a)

The supply of sugarcane and ethanol involves different sector levels: local, national, and global (Goldemberg et al., 2014). Therefore, strategies to follow the quantity and the availability of sugarcane for sugar and energy production are essential. Those strategies involve the knowledge of the growing area, yield, and quality of sugarcane. Among these, the yield estimation has more considerable complexity, because of a result of the various factors, such as weather and crop management (Everingham et al., 2016).

Along the time, different methodologies were developed for yield forecast, and these can be classified into two groups: the direct methods, which consist of the acquisition of field data to measure the yield, and the indirect ones, which use empirical relationship between yield data and weather, remote sensing, or both (Bégué et al., 2010; Luciano et al., 2021; Morel et al., 2014). Further, other indirect methods are process-based crop models (PBM), a framework that consists of organized algorithms that describe the physical and physiological process that occurs in crop development (Jones et al., 2017). The PBM has been largely used for studies involving yield estimation and forecast (Basso and Liu, 2019). For sugarcane, different PBMs were developed and presented satisfactory results, such as: AUSCANE (Jones et al., 1989); CANEGRO (Inman-bamber, 1989; Singels et al., 2008); QCANE (Liu and Kingston, 1995); APSIM (Holzworth et al., 2014; Keating et al., 2003); MOSICAS (Martiné, 2003); SAMUCA (Marin and Jones, 2014; Vianna et al., 2020) and CASUPRO (Villegas et al., 2005).

Nevertheless, these PBMs were developed for deterministic simulations, which ignore the uncertainty caused by the dynamic of environmental variabilities, such as physics and chemistry soil properties, genotype response to the environment, weather, and management practices (Marin et al., 2017). Moreover, the uncertainty is intensified for simulations extrapolated for large areas (Manivasagam and Rozenstein, 2020). These uncertainties influence the crop canopy simulations, which mainly affect the process of light interception and the potential photosynthesis rates. It also influences the soil water content simulations, which control part of the evapotranspiration and determine the crop water stress (de Wit and van Diepen, 2007).

Recently Marin et al. (2015) and Dias and Sentelhas (2017) recommended that the use of an ensemble of sugarcane PBMs improves the performance of crop simulations. However, applying multiple PBM increased the number of inputs and parameters used to describe the crop genotype, soil, and weather conditions, which are difficult to quantify due to spatial and temporal variations (Marin et al., 2017). Furthermore, the sugarcane crop modeling has some particularity compared to the mainly simulated crops (wheat, maize, and soybean), because sugarcane growing seasons are larger (8 to 24 months with a mean of 12 months), increasing the time and the sensitivity to adverse weather conditions (Bégué et al., 2010; Hu et al., 2019). In the management of Brazilian sugarcane production, the crop regrows after consecutive harvests, which leads to a decrease in sugarcane yield that

can change by environment and management (Marin et al., 2019b). These factors increase the complexity of sugarcane yield estimation and cannot be simulated by models, without changes in the model parameterization or the use of empirical relationships (Dias and Sentelhas, 2017).

One solution to reduce the model uncertain is the data assimilation (DA), a methodology that incorporates external data of the environment into the deterministic dynamics of the PBM, to increase the simulations accuracy of one or more state variables and, therefore, the variable of interest, generally crop yield (Huang et al., 2019). Furthermore, the use of external observations represents a mechanism to account for processes that are not simulated by the PBM, such as crop management, adverse weather conditions (storms), lodging, pest and disease damages (Hu et al., 2019).

The DA algorithms can be classified into three groups: first, the forcing methods consist of directly replacing the simulated state variable with the observed data, which ignores the uncertainty in the PBM and observation data, and the performance relies only on the observation quality (Morel et al., 2014). Second, the calibration methods consist of reparametrize the PBM inputs, generally, the soil and crop genotype parameters, to match the PBM outputs with the observed data. Thus, this method attributes all the uncertainty of PBM's simulations to the parameters and initial conditions and neglects the observation error (Hu et al., 2019). Third, the update methods consist of algorithms that consider both the PBMs and observations uncertain to sequentially update the PBM state variable, during the model simulations when observations are available (Hu et al., 2019; Huang et al., 2019).

Among the three methods, the update is the most used (Kang and Özdoğan, 2019) and has been widely applied for different crops: maize (de Wit and van Diepen, 2007; Ines et al., 2013; Kang and Özdoğan, 2019; Lu et al., 2021; Zhao et al., 2013), wheat (Huang et al., 2016, 2015; Nearing et al., 2012; Pan et al., 2019; Tewes et al., 2020a, 2020b, 2020c; Xie et al., 2017; Zhuo et al., 2019); sugarcane (Abebe et al., 2022; Hu et al., 2019; Yu et al., 2020, 2022), grass (Huang et al., 2021), potato and cotton (Linker and Ioslovich, 2017).

Moreover, the majority of the studies on DA methods showed some improvement in PBM simulations, thus, the performance of PBM coupled with DA methods was related to the variable assimilated (Ines et al., 2013; Pan et al., 2019; Yu et al., 2022), the DA method used (Hu et al., 2019; Huang et al., 2021) and the methodology or sensor used to retrieve the assimilated variable (Abebe et al., 2022; Yu et al., 2020). Different variables were used as observations for DA methods, such as leaf area index (LAI) (Ines et al., 2013; Pan et al., 2019), soil moisture (Pan et al., 2019; Zhuo et al., 2019), evapotranspiration (Huang et al., 2015), the fraction of absorbed photosynthetically active radiation (Lu et al., 2019; Morel et al., 2014), biomass (Lu et al., 2022), and plant height (Yu et al., 2020). Usually, the most common variables were LAI and soil moisture, because of their relevance in plant growth and development and had been widely studied and correlated to different sensors (Hu et al., 2019). Furthermore, the variable chosen for DA was linked to the amount and timing of observation acquisition (Tewes et al., 2020c).

In this regard, different methodologies and sensors were used to retrieve information about the environments, such as direct field measurement (Hu et al., 2019; Tewes et al., 2020a); the use of remote sensing satellite sensors such as MODIS (Huang et al., 2021; Ines et al., 2013), Landsat (Kang and Özdoğan, 2019), Sentinel-2 (Pan et al., 2019; Tewes et al., 2020b; Tewes et al., 2020c; Zhuo et al., 2019); and synthetic aperture radars (SARs) such as Sentinel-1 (Abebe et al., 2022; Pan et al., 2019) and UAV-based observations (Yu et al., 2020). Despite the methodology or sensor used, it is important to quantify the uncertainties of the retrieved data. This can be done by testing different values of uncertainty or comparing them to field observations (Abebe et al., 2022; Nearing et al., 2012; Yu et al., 2020).

Furthermore, the uncertainties of the assimilated variable need to be compared with the PBM uncertainties to estimate and update the PBM state variable sequentially. For that, different DA algorithms were evaluated, such as Ensemble Kalman Filter (EnKF), Four-Dimensional Variational Strategy, Particle Filter, and Weighted Mean (Jin et al., 2018; Tewes et al., 2020a). Among them, the EnKF was the most tested and used, the EnKF use the Bayesian approach and the Monte Carlo method (Evensen, 2003), which sequentially accounts for uncertainties from both simulations and observed data. Kang and Özdoğan (2019) emphasized that prior model calibration improved the PBMs performance assimilating LAI for maize yield estimation, using the SAFY model. Thus, applying the same genetic coefficients, ignoring the variability of the environment, led to errors and inconsistency between PBM and the LAI retrieved from remote sensing. However, after the calibration, the error inconsistency still occurred and the author attributed part of it to the simplifications in the PBM used.

Therefore, calibrating the model parameters that are specific to cultivar or region should be targeted before DA methods (Huang et al., 2021), and this should be more stressed in crop-specific PBM, due to a large number of parameters to describe the crops cultivar (Silvestro et al., 2017). Moreover, error inconsistency is strictly connected to model structure, and a hypothesis that arises from literature is that the use of a more complex PBM in terms of structure would reduce the inconsistency in the use of DA. Because these models benefit from their detailed description of crop development, traits and plant phenology (Curnel et al., 2011; Vianna et al., 2020).

However, some points can be better explored to understand how the DA methods are affected by the PBMs calibrations, and structures, regarding how detailed their processes were described. First, no other studies evaluated the impact of using one genotype-based calibration on other cultivars on DA methods performance, with the hypothesis that DA methods could neglect the use of a priori model calibration. Second, most of the studies focus on understanding the impact of observations quality, quantity and timing, however, no other studies compare if these points had a higher influence on the PBM parametrization. Third, assess the performance of the PBM with DA, regarding how detailed their processes were described, with the hypothesis that a more detailed description of crop development could lower the inconsistency or BIAS in DA.

1.1. Objectives

The general objective of this study was to evaluate the performance of sugarcane PBMs with LAI assimilation on the stalk fresh yield (SFY) estimation, regarding the influence of prior PBM calibration, based on genotype cultivar parameters, and the PBMs structure related to the description of crop biophysical process.

1.1.1. Specific objectives

- i. Develop a framework to assimilate LAI retrieved from field measurement or remote sensing into the DSSAT/SAMUCA and WOFOST model;
- ii. Assess the performance of using crop-specific PBM in conditions of no or indirect calibrations based on sugarcane cultivars;
- Evaluate other factors that could impact the PBM performance, e.g. the amount and timing of LAI observations;

- iv. Evaluate two types of PBMs regarding their model structure for estimating SFY, with LAI assimilation;
- v. Assess how the processes involved in LAI calculation affect the error inconsistency between PBM and DA.

References

- Abebe, G., Tadesse, T., Gessesse, B., 2022. Assimilation of leaf Area Index from multisource earth observation data into the WOFOST model for sugarcane yield estimation. Int. J. Remote Sens. 43, 698–720. https://doi.org/10.1080/01431161.2022.2027547
- Basso, B., Liu, L., 2019. Seasonal crop yield forecast: Methods, applications, and accuracies, 1st ed, Advances in Agronomy. Elsevier Inc. https://doi.org/10.1016/bs.agron.2018.11.002
- Bégué, A., Lebourgeois, V., Bappel, E., Todoroff, P., Pellegrino, A., Baillarin, F., Siegmund, B., 2010. Spatiotemporal variability of sugarcane fields and recommendations for yield forecast using NDVI. Int. J. Remote Sens. 31, 5391–5407. https://doi.org/10.1080/01431160903349057
- Chen, Y., Tao, F., 2022. Potential of remote sensing data-crop model assimilation and seasonal weather forecasts for early-season crop yield forecasting over a large area. F. Crop. Res., 276, 108398. https://doi.org/10.1016/j.fcr.2021.108398
- Curnel, Y., de Wit, A.J.W., Duveiller, G., Defourny, P., 2011. Potential performances of remotely sensed LAI assimilation in WOFOST model based on an OSS Experiment. Agric. For. Meteorol. 151, 1843–1855. https://doi.org/10.1016/j.agrformet.2011.08.002
- de Wit, A.J.W., van Diepen, C.A., 2007. Crop model data assimilation with the Ensemble Kalman filter for improving regional crop yield forecasts. Agric. For. Meteorol. 146, 38–56. https://doi.org/10.1016/j.agrformet.2007.05.004
- Dias, H.B., Sentelhas, P.C., 2017. Evaluation of three sugarcane simulation models and their ensemble for yield estimation in commercially managed fields. F. Crop. Res. 213, 174–185. https://doi.org/10.1016/j.fcr.2017.07.022
- Evensen, G., 2003. The Ensemble Kalman Filter: Theoretical formulation and practical implementation. Ocean Dyn. 53, 343–367. https://doi.org/10.1007/s10236-003-0036-9
- Everingham, Y., Sexton, J., Skocaj, D., Inman-Bamber, G., 2016. Accurate prediction of sugarcane yield using a random forest algorithm. Agron. Sustain. Dev. 36. https://doi.org/10.1007/s13593-016-0364-z
- FAO, 2022. Food and Agriculture Organization of the United Nations. FAOSTAT. URL http://www.fao.org/faostat/en/#home
- Goldemberg, J., Mello, F.F.C., Cerri, C.E.P., Davies, C.A., Cerri, C.C., 2014. Meeting the global demand for biofuels in 2021 through sustainable land use change policy. Energy Policy 69, 14–18. https://doi.org/10.1016/j.enpol.2014.02.008
- Holzworth, D.P., Huth, N.I., deVoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Chenu, K., van Oosterom,
 E.J., Snow, V., Murphy, C., Moore, A.D., Brown, H., Whish, J.P.M., Verrall, S., Fainges, J., Bell, L.W., Peake,
 A.S., Poulton, P.L., Hochman, Z., Thorburn, P.J., Gaydon, D.S., Dalgliesh, N.P., Rodriguez, D., Cox, H.,
 Chapman, S., Doherty, A., Teixeira, E., Sharp, J., Cichota, R., Vogeler, I., Li, F.Y., Wang, E., Hammer, G.L.,
 Robertson, M.J., Dimes, J.P., Whitbread, A.M., Hunt, J., van Rees, H., McClelland, T., Carberry, P.S.,
 Hargreaves, J.N.G., MacLeod, N., McDonald, C., Harsdorf, J., Wedgwood, S., Keating, B.A., 2014. APSIM -

Evolution towards a new generation of agricultural systems simulation. Environ. Model. Softw. 62, 327–350. https://doi.org/10.1016/j.envsoft.2014.07.009

- Hu, S., Shi, L., Huang, K., Zha, Y., Hu, X., Ye, H., Yang, Q., 2019. Improvement of sugarcane crop simulation by SWAP-WOFOST model via data assimilation. F. Crop. Res. 232, 49–61. https://doi.org/10.1016/j.fcr.2018.12.009
- Huang, J., Gómez-Dans, J.L., Huang, H., Ma, H., Wu, Q., Lewis, P.E., Liang, S., Chen, Z., Xue, J.H., Wu, Y., Zhao, F., Wang, J., Xie, X., 2019. Assimilation of remote sensing into crop growth models: Current status and perspectives. Agric. For. Meteorol. 276–277, 107609. https://doi.org/10.1016/j.agrformet.2019.06.008
- Huang, J., Ma, H., Su, W., Zhang, X., Huang, Y., Fan, J., Wu, W., 2015. Jointly Assimilating MODIS LAI and et Products into the SWAP Model for Winter Wheat Yield Estimation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 4060–4071. https://doi.org/10.1109/JSTARS.2015.2403135
- Huang, J., Sedano, F., Huang, Y., Ma, H., Li, X., Liang, S., Tian, L., Zhang, X., Fan, J., Wu, W., 2016. Assimilating a synthetic Kalman filter leaf area index series into the WOFOST model to improve regional winter wheat yield estimation. Agric. For. Meteorol. 216, 188–202. https://doi.org/10.1016/j.agrformet.2015.10.013
- Huang, X., Zhao, G., Zorn, C., Tao, F., Ni, S., Zhang, W., Tu, T., Höglind, M., 2021. Grass modelling in data-limited areas by incorporating MODIS data products. F. Crop. Res. 271. https://doi.org/10.1016/j.fcr.2021.108250
- Ines, A.V.M., Das, N.N., Hansen, J.W., Njoku, E.G., 2013. Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. Remote Sens. Environ. 138, 149–164. https://doi.org/10.1016/j.rse.2013.07.018
- Inman-Bamber, N. G., Thompson, G.D., 1989. Models of dry matter accumulation by sugarcane. Proc. South African Sugarcane Technol. 63, 212–216.
- Jin, X., Kumar, L., Li, Z., Feng, H., Xu, X., Yang, G., Wang, J., 2018. A review of data assimilation of remote sensing and crop models. Eur. J. Agron. 92, 141–152. https://doi.org/10.1016/j.eja.2017.11.002
- Jones, C.A., Wegener, M.K., Russell, J.S., McLeod, I.M., Williams, J., 1989. AUSCANE-Simulation of Australian sugarcane with EPIC. CSIRO.
- Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., Keating, B.A., Munoz-Carpena, R., Porter, C.H., Rosenzweig, C., Wheeler, T.R., 2017. Brief history of agricultural systems modeling. Agric. Syst. 155, 240–254. https://doi.org/10.1016/j.agsy.2016.05.014
- Kang, Y., Özdoğan, M., 2019. Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. Remote Sens. Environ. 228, 144–163. https://doi.org/10.1016/j.rse.2019.04.005
- Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N.G., Meinke, H., Hochman, Z., McLean, G., Verburg, K., Snow, V., Dimes, J.P., Silburn, M., Wang, E., Brown, S., Bristow, K.L., Asseng, S., Chapman, S., McCown, R.L., Freebairn, D.M., Smith, C.J., 2003. An overview of APSIM, a model designed for farming systems simulation. Eur. J. Agron. 18, 267–288. https://doi.org/10.1016/S1161-0301(02)00108-9
- Linker, R., Ioslovich, I., 2017. Assimilation of canopy cover and biomass measurements in the crop model AquaCrop. Biosyst. Eng. 162, 57–66. https://doi.org/10.1016/j.biosystemseng.2017.08.003
- Liu, D.L., Kingston, G., 1995. QCANE: a simulation model of sugarcane growth and sugar accumulation. In: Research and Modelling Approaches to Assess Sugercane Production Opportunities and Constraints. University of Queensland, St. Lucia, 29–25.

- Lobell, Lu, Y., Chibarabada, T.P., Ziliani, M.G., Onema, J.M.K., McCabe, M.F., Sheffield, J., 2021. Assimilation of soil moisture and canopy cover data improves maize simulation using an under-calibrated crop model. Agric. Water Manag. 252, 106884. https://doi.org/10.1016/j.agwat.2021.106884
- Lu, Y., Wei, C., McCabe, M. F., Sheffield, J., 2022. Multi-variable assimilation into a modified AquaCrop model for improved maize simulation without management or crop phenology information. Agric. Water Manag, 266, 107576. https://doi.org/10.1016/j.agwat.2022.107576
- Luciano, A.C. dos S., Picoli, M.C.A., Duft, D.G., Rocha, J.V., Leal, M.R.L.V., le Maire, G., 2021. Empirical model for forecasting sugarcane yield on a local scale in Brazil using Landsat imagery and random forest algorithm. Comput. Electron. Agric. 184. https://doi.org/10.1016/j.compag.2021.106063
- Manivasagam, V.S., Rozenstein, O., 2020. Practices for upscaling crop simulation models from field scale to large regions. Comput. Electron. Agric. 175. https://doi.org/10.1016/j.compag.2020.105554
- Marin, F. R., Vianna, M. S., Nassif, D. S., 2019a. Challenges, Constraints, and Limitations of Cane Biofuels. In Sugarcane Biofuels (pp. 389-407). Springer, Cham. https://doi.org/10.1007/978-3-030-18597-8_17
- Marin, F., Jones, J.W., Boote, K.J., 2017. A stochastic method for crop models: Including uncertainty in a sugarcane model. Agron. J. 109, 483–495. https://doi.org/10.2134/agronj2016.02.0103
- Marin, F.R., 2016. Understanding sugarcane production, Biofuels, and market volatility in Brazil—A research perspective. Outlook Agric. 45, 75–77. https://doi.org/10.1177/0030727016649802
- Marin, F.R., Jones, J.W., 2014. Process-based simple model for simulating sugarcane growth and production. Sci. Agric. 71, 1–16. https://doi.org/https://doi.org/10.1590/S0103-90162014000100001
- Marin, F.R., Rattalino Edreira, J.I., Andrade, J., Grassini, P., 2019b. On-farm sugarcane yield and yield components as influenced by number of harvests. F. Crop. Res. 240, 134–142. https://doi.org/10.1016/j.fcr.2019.06.011
- Marin, F.R., Thorburn, P.J., Nassif, D.S.P., Costa, L.G., 2015. Sugarcane model intercomparison: Structural differences and uncertainties under current and potential future climates. Environ. Model. Softw. 72, 372–386. https://doi.org/10.1016/j.envsoft.2015.02.019
- Martiné, J.-F., 2003. Modélisation de la production potentielle de la canne à sucre en zone tropicale, sous conditions thermiques et hydriques contrastées. Applications du modèle. Institut national agronomique Paris-Grignon.
- Morel, J., Todoroff, P., Bégué, A., Bury, A., Martiné, J.F., Petit, M., 2014. Toward a satellite-based system of sugarcane yield estimation and forecasting in smallholder farming conditions: A case study on reunion island. Remote Sens. 6, 6620–6635. https://doi.org/10.3390/rs6076620
- Nearing, G.S., Crow, W.T., Thorp, K.R., Moran, M.S., Reichle, R.H., Gupta, H. V., 2012. Assimilating remote sensing observations of leaf area index and soil moisture for wheat yield estimates: An observing system simulation experiment. Water Resour. Res. 48, 1–13. https://doi.org/10.1029/2011WR011420
- Pan, H., Chen, Z., de Allard, W., Ren, J., 2019. Joint assimilation of leaf area index and soil moisture from sentinel-1 and sentinel-2 data into the WOFOST model for winter wheat yield estimation. Sensors (Switzerland) 19. https://doi.org/10.3390/s19143161
- Silvestro, P.C., Pignatti, S., Pascucci, S., Yang, H., Li, Z., Yang, G., Huang, W., Casa, R., 2017. Estimating wheat yield in China at the field and district scale from the assimilation of satellite data into the Aquacrop and simple algorithm for yield (SAFY) models. Remote Sens. 9, 1–24. https://doi.org/10.3390/rs9050509
- Singels, A., Jones, M., Van Der Berg, M., 2008. DSSAT v4.5-Canegro Sugarcane Plant Module: Scientific Documentation.

- Tewes, A., Hoffmann, H., Krauss, G., Schäfer, F., Kerkhoff, C., Gaiser, T., 2020a. New approaches for the assimilation of LAI measurements into a crop model ensemble to improve wheat biomass estimations. Agronomy 10, 1–21. https://doi.org/10.3390/agronomy10030446
- Tewes, A., Hoffmann, H., Nolte, M., Krauss, G., Schäfer, F., Kerkhoff, C., Gaiser, T., 2020b. How do methods assimilating sentinel-2-derived LAI combined with two different sources of soil input data affect the crop model-based estimation of wheat biomass at sub-field level? Remote Sens. 12. https://doi.org/10.3390/rs12060925
- Tewes, A., Montzka, C., Nolte, M., Krauss, G., Hoffmann, H., Gaiser, T., 2020c. Assimilation of Sentinel-2 Estimated LAI into a Crop Model: Influence of Timing and Frequency of Acquisitions on Simulation of Water Stress and Biomass Production of Winter Wheat. Agronomy 10, 1813. https://doi.org/10.3390/agronomy10111813
- Vianna, M. dos S., Nassif, D.S.P., dos Santos Carvalho, K., Marin, F.R., 2020. Modelling the trash blanket effect on sugarcane growth and water use. Comput. Electron. Agric. 172, 105361. https://doi.org/10.1016/j.compag.2020.105361
- Villegas, F.D., Daza, O.H., Jones, J.W., Royce, F.S., 2005. CASUPRO: An industry-driven sugarcane model. 2005 ASAE Annu. Int. Meet. 0300. https://doi.org/10.13031/2013.19062
- Xie, Y., Wang, P., Bai, X., Khan, J., Zhang, S., Li, L., Wang, L., 2017. Assimilation of the leaf area index and vegetation temperature condition index for winter wheat yield estimation using Landsat imagery and the CERES-Wheat model. Agric. For. Meteorol. 246, 194–206. https://doi.org/10.1016/j.agrformet.2017.06.015
- Yu, D., Zha, Y., Shi, L., Jin, X., Hu, S., Yang, Q., Huang, K., Zeng, W., 2020. Improvement of sugarcane yield estimation by assimilating UAV-derived plant height observations. Eur. J. Agron. 121, 126159. https://doi.org/10.1016/j.eja.2020.126159
- Yu, D., Zha, Y., Shi, L., Ye, H., Zhang, Y., 2022. Improving sugarcane growth simulations by integrating multisource observations into a crop model. Eur. J. Agron. 132, 126410. https://doi.org/10.1016/j.eja.2021.126410
- Zhao, Y., Chen, S., Shen, S., 2013. Assimilating remote sensing information with crop model using Ensemble Kalman Filter for improving LAI monitoring and yield estimation. Ecol. Modell. 270, 30–42. https://doi.org/10.1016/j.ecolmodel.2013.08.016
- Zhuo, W., Huang, J., Li, L., Zhang, X., Ma, H., Gao, X., Huang, H., Xu, B., Xiao, X., 2019. Assimilating Soil Moisture Retrieved from Sentinel-1 and Sentinel-2 Data into WOFOST Model to Improve Winter Wheat Yield Estimation. Remote Sens. 11, 1618. https://doi.org/10.3390/rs11131618

2. ASSIMILATING LEAF AREA INDEX DATA INTO A SUGARCANE PROCESS-BASED CROP MODEL FOR IMPROVING YIELD ESTIMATION

Abstract

The ability to estimate sugarcane yield is an important factor to improving the planning capacity of public and private sectors, and so food and energy security. One way of achieving this is by employing process-based crop models (PBM), which can be coupled to data assimilation (DA) algorithms to correct predictions along the crop season. While the application of PBMs often need careful parameterization or genotype-specific parameters, few studies focus on understanding the impacts of crop parametrization with different crop genotypes with DA. Moreover, dimensioning the number and timing of observations is key to effectively improve predictions with DA. This study assesses the performance of a new sugarcane PBM (DSSAT/SAMUCA) coupled to three DA methods, and when the genotype-specific parameters are available or not. Data from 22 field experiments is utilized to compare the performance of using the Ensemble Kalman filter (EnKF), Ensemble smoother (ES) and Weighted mean (WM) for assimilating leaf area index (LAI) to improve yields estimates. We also quantify the impact of using one genotype-specific calibration (cv. RB867515) on yield predictions of four non-calibrated genotypes (cv. NCo376, SP832847, R570, RB72454). Simulations of DA methods had better performance than employing the PBM without DA, so called open-loop (OP). The ES method resulted in the best performance ($R^2 = 0.498$ and RMSE = 20.268 Mg ha⁻¹) followed by EnKF and WM. Utilizing a genotype-specific calibration showed substantially smaller RMSE for the three DA methods (EnKF = 16.76, ES = 16.70 and WM $= 15.36 \text{ Mg ha}^{-1}$ compared to non-calibrated (EnKF = 21.44 - 26.23, ES = 21.50 - 26.27 and WM = 23.38 - 28.37 Mg ha⁻¹). Nevertheless, we also verified a higher improvement of model performance when applying EnKF and ES method to experiments where the cultivar does not match the genotypespecific calibration employed. While the WM had the opposite results, with the calibrated cultivar showing a higher improvement of model performance. As the number of LAI data assimilation increases, the DA methods tend to outperform the OP, but observations at late crop phenological stage of development showed a higher positive influence on SFY predictions.

Key-words: Ensemble Kalman filter, Ensemble smoother, Weighted mean, DSSAT/SAMUCA

2.1. Introduction

Sugarcane yield estimations are key for agricultural development and planning in public and private sectors, and so to food and energy security as the crop is a major feedstock of sugar and ethanol. Brazil is the largest sugarcane producer in the world, accounting for ca. 40% of the global production, and ca. 50% of global sugar exports (FAO, 2019). It is also the second largest ethanol producing country after the USA, thus sugarcane production chain is crucial for food security and for the country's energy matrix production (Goldemberg et al., 2014).

Sugarcane grown in Brazil shows a great seasonal and spatial yield variability (Marin et al., 2008; Marin and de Carvalho, 2012), which can be attributed to interannual climate variability, pests and diseases, soil fertility, as well as to the intensive sequential harvesting practice (e.g. ratooning) (Lisboa et al., 2018; Marin et al., 2019, 2021). Thus, the significance of Brazil's production and the yield variability emphasizes the importance of accurate crop yield predictions for assisting decision-makers and ensuring better market planning and regulation (Marin, 2016).

Process-based crop models (PBM) are widely used tools to predict crop yield as a product of Genotype \times Environment \times Management interactions (Wang et al., 2019). These models consist of organized algorithms that describe physical and biological processes that occur in crop growth and development, and so mimic the interactions among soil, plant, and atmosphere components (Jones et al., 2017). For sugarcane, several PBMs were developed and evaluated against observed data across the world, such as AUSCANE (Jones et al., 1989); CANEGRO (Inman-Bamber and Thompson, 1989; Singels et al., 2008); QCANE (Liu and Kingston, 1995); APSIM (Holzworth et al., 2014; Keating et al., 2003); MOSICAS (Martiné, 2003); SAMUCA (Marin and Jones, 2014; Marin et al., 2017; Vianna et al., 2020); and CASUPRO (Villegas et al., 2005). The DSSAT/SAMUCA (DS) model considers physiological processes such as biomass partitioning at phytomer level and canopy carbon assimilation using leaf assimilation rates and carboxylation efficiency. Furthermore, DS was recently updated and evaluated considering the effect of the green cane trash blanket effect, an arising management practice in Brazil (Vianna et al., 2020).

The deterministic approach employed in most PBMs generally relies on a substantial amount of input data and coefficients to characterize the soil-plant-atmosphere and management system. This information is sometimes difficult to be determined and may vary within the spatial and temporal domain, transferring this uncertainty to simulations of crop yields (Marin et al., 2017; Zhao et al., 2013). In addition, PBMs are not yet ready to account for the reduction factors in yield simulations because of the high complexity in measuring the various biotic and abiotic interactions, which is generally accounted as an empirical yield-gap factor (van Ittersum et al., 2003; Lobel et al., 2009). For instance, biotic factors such as pests and diseases, lodging, and storms can damage the sugarcane canopy by reducing the leaf area index (LAI), which is not well simulated by most of available PBMs (Hu et al., 2019). Such factors considerably affect two key crop physiological processes: (i) plant canopy light interception and potential photosynthesis; and (ii) soil water content and evapotranspiration, which in turn determine the actual photosynthesis rate (de Wit and van Diepen, 2007).

An option to reduce uncertainty in yield prediction by PBMs is the use of data assimilations methods (DA) that incorporate observations into the PBM simulations, to produce more accurate estimates of state variables (Huang et al., 2019). In this regard, updating methods are among the most used tools because they are computationally less expensive and capable of accounting for both model and observation uncertainty (Ines et al., 2013). The most used update method is the Ensemble Kalman filter (EnKF) (Kang and Özdoğan, 2019; Pan et al., 2019), a Bayesian approach of the Monte Carlo method developed by Evensen (2003), which sequentially accounts for uncertainties from crop simulation and measured data (Xie et al., 2017).

The EnKF is flexible and fits well with PBMs frameworks but updating only a few state variables may affect the model integrity and cause undesired model states in some circumstances (Nearing et al., 2012; Tewes et al., 2020a). To overcome such limitations, other DA methods were proposed such as the Weighted mean (WM), which can account for observational errors without changing the PBM state variables (Tewes et al., 2020a). A third DA is the ensemble smoother (ES), which uses similar assumptions of EnKF, but assimilates all available observations simultaneously, ensuring consistency between the PBM state variables and DA state vector (Lee et al., 2016; Van Leeuwen and Evensen, 1996; Yu et al., 2020).

Recently, DA methods were tested for sugarcane crop models through LAI, soil moisture and plant height data assimilation, for all showed improved crop growth simulations, mainly in water-limited environments (Hu et al., 2019; Yu et al., 2020). However, these two previous studies used the PBM SWAP-WOFOST adapted for sugarcane (Supit et al., 1994; van Dam et al., 1997), which lacks the representation of some key physiological and morphological process, such as tillering and sucrose accumulation. Thus, a more generalist PBM is more practical for applying DA due to the lower number of parameters and simplicity of the plant process included (Kang and Özdoğan, 2019). Crop-specific PBMs, like those available in the DSSAT platform, have more detailed descriptions

of plant physiological processes, which may lead to less uncertain simulations (Jones and Singels, 2018; Vianna et al., 2020). Nevertheless, the use of DA in such sophisticated models may also cause model inconsistency due to the model's complexity and the number of parameters involved (Nearing et al., 2012).

Moreover, due to the difficulty of quantifying the variability of genotypes in large areas, this information is commonly resumed to just one calibration in DA simulations (Huang et al., 2016; Tewes et al., 2020a). Lu et al. (2021) proposed a framework to use the EnKF to simulate Maize, without using a specific genotype calibration and presented satisfactory improvement compared to open-loop (OP) simulations. However, recent studies emphasized that calibrating the model parameters that are specific for cultivar or region should be targeted before DA methods (Huang et al., 2021), and this should be more stressed in crop-specific PBM, due to a large number of parameters to describe the crops cultivar (Kang and Özdoğan 2019). To our knowledge, no other studies explored how the use of one genotype-based calibration for sugarcane, affected other cultivars on DA methods performance. This study aimed to assess the performance of using crop specific PBM in conditions of no or indirect calibrations based on sugarcane cultivars, and also analyses other factors that could impact the PBM performance, e.g. the amount and timing of LAI observations. To fill this knowledge gap, we developed a framework to assimilate sugarcane LAI measurements in the DS model using three different DA methods: EnKF, ES, and WM. A dataset of field experiments conducted with five cultivars in a diversity of soils and climates across Brazil was utilized, part of them to calibrate the DS for the cultivar RB867515, and other to evaluate and compare the three techniques in estimating stalk fresh yield (SFY).

2.2. Material and Methods

2.2.1. Field experiments description

In total, we used a dataset of 22 experiments described in Table 1. Experiments 1 to 3 were conducted at the College of Agriculture "Luiz de Queiroz" (Esalq) of University of São Paulo (USP) in Piracicaba, Brazil (Lat. 22°41'55" S, long. 47°38'34" W, alt. 540 m a.m.s.l.). Chopped stalks of the RB867515 cultivar (Exps. 1 and 2) were used for planting 13-15 buds m^{-1} at 1.4 m row spacing down to a depth of 0.2 m. Experiment 3 was conducted in the first ration of experiment 1 (sugarcane re-growth). Agricultural practices were adopted to represent high-yield farming systems and to ensure the crop was free from pests, diseases, and nutritional stress. The climate is characterized by a hot and humid summer with dry winter (Cwa - Köppen classification), and the soil of the three experiments is classified as Typic Hapludox. For these experiments, daily solar radiation, maximum and minimum air temperature, rainfall, wind speed, and relative air humidity were collected adjacent to the experiment site using an automatic weather station. The experimental area was irrigated by a center-pivot. The water balance was monitored to manage the water applications and ensure crops were not exposed to water stress throughout the growing cycle; for these, irrigation was triggered every time the soil moisture reached 80% of the total available soil water. Five locations were randomly selected at the beginning of each season, where LAI samples were taken with an LAI-2200 instrument (LI-COR Bioscience), following user manual recommendations for row crops (Gonçalves et al. 2020). LAI measurements were taken in a frequency of 5 - 20 days in experiments 1-3. Stalk fresh yield (SFY, Mg ha⁻¹) was measured only at the end of the crop season by mechanical harvesting for the three experiments. Experiments 1 and 2 during the crop season suffered from lodging, after heavy rain events and after the crop reaches a high SFY and stalks height. This caused changes in LAI during the season and accelerate the decrease in LAI at the late stage.

We also used experimental data of previous studies (Experiments 4-22, Tab. 1) conducted in a diversity of environments and using different cultivars (Marin et al., 2015, 2011; Suguitani, 2006; Vianna et al., 2020), and following a similar protocol of the experiments 1-3 above described. Experiments 4-15 had the LAI sampled 1-7 times during the crop season, and SFY sampled only at the end of the crop season. The experiments 16 - 22 were only used for DS calibrations and had tiller population, stalk diameter, stalk height, LAI, SFY, stalk, and leaf dry mass, and sucrose content on fresh cane basis (POL) obtained by regular sampling (Vianna et al., 2020).

Soil characteristics and management practices such as planting and harvesting dates, row spacing, mulch cover and irrigation applications (mm d-1) on each site were prescribed to the model as input information. Also, for experiments 3, 17, 19 and 21 a total of 12 Mg ha⁻¹ of green cane straw were considered for simulations. All other experiments were conducted under bare soil conditions.

The experiments 1 to 3 and 16 to 22 had the LAI sampled with the LAI-2200 and LAI-2000 respectively, which had an accuracy of 4%, according to the user manual (Gonçalves et al., 2020). The other experiments had their LAI sampled with different sensors and methodology (Marin et al., 2015; Marin et al., 2011; Suguitani, 2006); to maintain the response of the DA methods with the DS, we assumed the same accuracy for all observations.

Experim ent Number	Location	Planting date	Harvest date	Planting Type	Weath er [†]	Soil‡	Water treatme nt	Variety	Referen ce		
1		10/05/20 18	11/30/20 19	Plant			Irrigate d	RB8675 15			
2		11/06/20 18	11/30/20 19	Plant			Irrigate d	RB8675 15	¥		
3		11/31/20 19	11/04/20 20	1 st Ratoon*			Irrigate d	RB8675 15			
4	-	10/29/20 04	07/15/20 05	Plant	-		Irrigate d	R570			
5	Piracicaba/	10/29/20 04	07/15/20 05	Plant	21.6	Typic	Rainfed	R570			
6	SP 22°41 S,	10/29/20 04	07/15/20 05	Plant	°C, 1230	30 Haplud m, Typic	1230 Haplud ox	1230 Haplud	Irrigate d	RB7245 4	
7	47 38 W, 560m	10/29/20 04	07/15/20 05	Plant	mm, CWa		Rainfed	RB7245 4	Suguita		
8		10/29/20 04	07/15/20 05	Plant			Irrigate d	SP8328 47	ni (2006)		
9		10/29/20 04	07/15/20 05	Plant			Rainfed	SP8328 47			
10		10/29/20 04	07/15/20 05	Plant			Irrigate d	NCo37 6			
11		10/29/20 04	07/15/20 05	Plant			Rainfed	NCo37 6			
12	Aparecida do Taboado/ MS 20°05S, 51°18'W, 335 m	07/01/20 06	09/08/20 07	Plant	23.5 °C, 1560 mm, Aw	Typic Haplud ox Typic	Rainfed	RB8675 15	Marin et al. (2015)		

Table 1. Description of experimental datasets used in the simulations. Only experiments 16 to 22 were used for model calibration.

13	Colina/SP 20°25'S, 48°19'W, 590 m	02/10/20 04	06/15/20 05	Plant	22.8 °C, 1363 mm, Cwa	Typic Haplud ox Typic	Rainfed	RB8675 15	
14	Olimpia/S P 20°26'S, 48°32'W, 500 m	02/10/20 04	06/15/20 05	Plant	23.3 °C, 1349 mm, Cwa	Typic Haplud ox Typic	Rainfed	RB8675 15	
15	Coruripe/ AL 10°07'S, 36°10'W, 16 m	08/16/20 05	09/15/20 06	Plant	21.6 °C, 1401 mm, As	Fragiud ult Typic	Rainfed	RB8675 15	
16		12/06/20 12	10/15/20 13	Plant			Irrigate d	RB8675 15	
17		10/15/20 13	07/15/20 14	1 st Ratoon*			Irrigate d	RB8675 15	
18	Piracicaba/	10/15/20 13	07/15/20 14	1 st Ratoon	21.6	Typic	Irrigate d	RB8675 15	
19	SP 22°41' S,	07/15/20	06/08/20	2 nd Ratoon*	°C, 1230	Haplud	Irrigate d	RB8675	Vianna et al.
20	47°38'W, 560m	07/15/20 14	06/08/20 15	2 nd Ratoon	mm, CWa	Typic	Irrigate d	RB8675 15	(2020)
21		06/08/20	06/08/20	3rd			Irrigate	RB8675	
22		15 06/08/20 15	$16 \\ 06/08/20 \\ 16$	Katoon* 3 rd Ratoon			d Irrigate d	15 RB8675 15	

† Respectively: mean annual temperature, annual total rainfall, Koeppen Classification.

‡ U.S. Soil Taxonomy

¥ Experimental data collected for this study and not previously published.

* With mulch cover

2.2.2. Brief description of DSSAT/SAMUCA and calibration process

The SAMUCA model is a PBM firstly developed by Marin and Jones (2014), which is capable to simulate the growth and development of sugarcane crop, implementing an algorithm to describe processes related to phenology, canopy development, tillering, biomass accumulation, root growth, and water stress (Marin et al., 2017). Vianna et al. (2020) improved the SAMUCA model by including recent scientific findings on sugarcane growth at phytomer level, canopy assimilation, and tillering. In this new version, the model was adapted to operate the onedimensional "tipping bucket" soil water balance and to incorporate the soil temperature to account for the trash blanket effect on sugarcane growth and water use. This presented a superior performance compared with the previous version and was comparable to other widely used PBMs for sugarcane. For this study, we used the SAMUCA model incorporated into the Decision Support System for Agrotechnology Transfer (DSSAT) platform version 4.8 (Jones et al., 2003; Hoogenboom et al., 2019; Vianna et al., 2020), namely DSSAT/SAMUCA (DS).

The model was calibrated for cultivar RB867515 by Vianna at al. (2020), using experiments 16 to 22. A routine was designed to find crop parameters that minimize simulation errors by means of the RMSE (Wallach et al., 2018). To avoid unrealistic parameters estimation, the constrained BFGS ("Broyden–Fletcher–Goldfarb–Shanno") optimization method (Byrd et al., 1996) was employed assigning plausible range of parameters based on field

observation and literature. To adapt the plant-module of SAMUCA within the soil-plant-atmosphere framework of DSSAT, the root growth parameters SRLMAX (Specific Root Length at Root Front), SRLMIN (Specific Root Length at Inner Roots Profile), and DSHOOT_EXT (Below ground shoots expansion rate) were re-calibrated (Appendix A - Table 1). The procedure was performed by eye-fitting using the same set of observations employed by Vianna et al. (2020) (experiments 16 to 22) for which we obtained similar performance (Appendix A - Table 2).

2.2.3. Description of data assimilation procedure

Three different updating methods of DA were investigated in this study: EnKF, ES, and WM, which are described below. Also, standard DS simulations without DA were performed, which are thereafter called open-loop (OP) simulations. For each DA, the ability to deal with LAI assimilation and their performance with DS simulations were assessed, by comparing the simulation results of SFY at the end of the crop cycle. Therefore, the DS was adapted to read an input file with a new estimated vector of state variables at any time, regardless of the method. When new LAI values were assimilated by DS, the leaf area and dry weight were also updated at phytomer and field level to ensure the consistency of canopy representation.

2.2.4. Ensemble Kalman filter method

The EnKF employs an analytic solution based on two related sources of information, in this case: PBMs outputs and field observations. These are synthesized to provide a better estimation, with lower variance. For that, the EnKF assumes that the observed data can be related to the state variable x_t (LAI in the case of this study) at time t as shown in Eq. 1:

$$\mathbf{y} = \mathbf{H}\mathbf{x}_{\mathsf{t}} + \boldsymbol{\varepsilon} \tag{1}$$

where y is the observations vector; H is the observation operator that relates to y; ε is a Gaussian random error vector with a mean of zero and observation error covariance R. Also, the forecast of x_t at t = k is Gaussian with mean $x_{t=k}^{f}$ and error covariance $P_{t=k}^{f}$. Under these assumptions, the estimated state and error covariance (P) are updated as:

$$x_{t=k}^{a} = x_{t-k}^{f} + K(y_{t=k} - Hx_{t=k}^{f})$$

$$P_{t=k}^{a} = (I - KH)P_{t=k}^{f}$$
(2)
(3)

where t is the time index; k is the time of the observed data; f represents the prior state (called forecast) and a is the posterior state (called analysis); I is the identity matrix; and K represents the Kalman gain calculated by Eq. 4:

$$K = P_{t=k}^{f} H^{T} \left(H P_{t=k}^{f} H^{T} + R_{t=k} \right)^{-1}$$
(4)

The EnKF forecast and analysis error covariance P^{f} come directly from an ensemble of the model simulations:

$$P^{f}H^{T} = (N_{e} - 1)^{-1} \sum_{n=1}^{N_{e}} (x_{n}^{f} - \bar{x}^{f}) (Hx_{n}^{f} - H\bar{x}^{f})^{T}$$
(5)

where *Ne* is the number of ensemble members, *n* is a running index for an ensemble member, and \bar{x}^f are the ensemble mean calculated as:

$$\bar{x}^{f} = N_{e}^{-1} \sum_{n=1}^{N_{e}} x_{n}^{J}$$
(6)

In our study, we only used the LAI retrieved from ground measurements as a state variable for DA methods. Thus, H can be taken as an identity matrix (H= 1), with that we can rewrite the Eq. 2, 4, and 5 as Eq. 7, 8 and 9.

$$x_{t=k}^{a} = x_{t=k}^{f} + K(y_{t=k} - x_{t=k}^{f})$$
(7)

$$K = P_{t=k}^{f} \left(P_{t=k}^{f} + R_{t=k} \right)^{-1}$$
(8)

$$P^{f} = (N_{e} - 1)^{-1} \sum_{n=1}^{N_{e}} (x_{n}^{f} - \bar{x}^{f}) (x_{n}^{f} - \bar{x}^{f})^{T}$$
(9)

In EnKF, the observed data are perturbed with the Monte Carlo approach to generate an ensemble, based on the data uncertainty represented by the variance. When used together with PBMs, there are two methods to generate the ensemble members (Zhuo et al., 2019): the first method adds a Gaussian perturbation to the PBM state variables output. The second, add a Gaussian perturbation to the model input parameters.

In this study, we used the second method to generate the ensemble members. Thus, we selected the most sensitive parameters to LAI based on a sensitivity analysis, using the Fourier amplitude sensitivity test (FAST) (Cukier et al., 1973; Saltelli et al., 1999) from the SALlib library (https://salib.readthedocs.io/en/latest/index.html). These were MAXGL (maximum number of green leaves a tiller can hold), MLA (maximum leaf area), PLASTOCHRON (thermal time required for the appearance of one new phytomer), INIT_LF_AREA (initial leaf area of first appeared leaf), and MID_TT_LF_GRO (thermal time where leaves can achieve half of its maximum biomass) (Appendix A - Table 1). These parameters were then perturbed to generate an ensemble (40 members), with a gaussian distribution and an uncertainty level of 10% before the simulation starts, as recommended for Ines et al. (2013) and Curnel et al. (2011) to optimize the time of the simulation and model accuracy.

After generating the set of parameters, DS runs until the first observed LAI is available. At this point, we calculated in sequence K and the vector $x_{t=k}^{a}$ (Eq. 7 and 8), that was considered the optimal estimation of LAI. This step also included small inflation of 1.5 for LAI in ensemble members, in the case of their variability becoming too low (Ines et al., 2013). This step ensured that the observations were not systematically rejected during assimilation. After that, the estimated LAI estimation is stored in an input file for the next simulations, and runs were re-initialized until the next observations became available.

2.2.5. Ensemble smoother method

The ES has the same assumptions and equations as the EnKF. The difference between them is the number of assimilations. The ES assimilates all observed data at once, regardless of the acquisition time. Thus, the DS predicted all the state variables until the end of the simulations, using 40 ensemble members and considering the parameter perturbation procedure. At the prediction step, the DS output is then compared with all observations. For each observation, the Eq. 7, 8 and 9 were applied, and the term $x_{t=k}^{a}$ was estimated. Thereafter, the DS was reinitialized and the $x_{t=k}^{a}$ was assimilated at each time, which reduces the number of model re-initialization and made it easier to couple with any program that was not created to DA adaptation, like the majority of PBM (Lee et al., 2016; Yu et al., 2020).

2.2.6. Weighted mean method

The WM methodology follows the approach proposed by Tewes et al. (2020a), which assumes that the PBMs OP ensemble simulation runs from the beginning to the end of the crop growing cycle. One or a few members of the PBMs simulations that were close to the observed variable receive a greater weight for the state vector estimation.

The same crop parameters used for EnKF and ES were used to create the ensemble members in this method. However, different from the other DA, a uniform distribution is assumed to create the 40 sets of parameters, and the maximum and the minimum values of the distribution range were assumed to be $\pm 10\%$ of the selected parameters mean value (Appendix A - Table 1). After the ensemble simulations run, a Python script reads the output file of the DS and performs the WM calculation as follows.

To predict the state $\hat{X}(t)$ of the system, we used the weighted mean of the ensemble $X_i(t)$:

$$\hat{X}(t) = \frac{\sum_{i=1}^{N} w_i(t) X_i(t)}{\sum_{i=1}^{N} w_i(t)}$$
(10)

where each weight w of ensemble member i at day t is calculated as follow:

$$w_i(t) = P(O(t_k)|X_i(t_k)) \text{ for } t_k \le t \le t_{k+1}$$
(11)

where P is the likelihood from that the observational O at time t_k approximates the simulated value. The observational error was assumed as a normal distribution, where $O(t_k)$ is mean and σ_k the standard deviation of the distribution. Thus, we applied the following equation for the calculation of the likelihood P:

$$P(O(t_k)|X_i(t_k)) = \frac{1}{\sqrt{2\pi\sigma_k^2}} exp\left(-\frac{(h((t_k)) - O(t_k))^2}{\sigma_k^2}\right)$$
(12)

where h mapped the states to the observational variables.

The calculated weights for the first observation were then propagated until the next observation, and they were also used to calculate the weighted mean of other state variables. For example, the SFY retrieved from the simulation members that have LAI closer to the observations will receive more weight. When a new observation is available, the weighted mean is recalculated, and when observations were outside the ensemble members, the entire weights were given to the closest member. Contrary to other existing DA updating methods, no state variables are updated during the simulation runs. Therefore, observations are used to select the output simulations' members of the PBM, not change the state variable directly in the PBM, during the run, which is common for the other DA methods.

2.2.7. Data analysis

The performance of each DA was evaluated by comparing simulated and observed SFY by using root mean squared error (RMSE) and coefficient of determination (R²). We also compared each DA with OP simulation to select the best approach for each experiment. It should be noted that we used the absolute error (AE) to compare the SFY at the end of the crop cycles for each experiment. To quantify if and how much each DA method reduced the SFY simulation error, the difference between the AE of three DA methods against the AE of OP simulations is calculated. Negative differences between DA and OP error, indicate better performance in DA method than the OP simulation.

2.2.8. Effect of genotype-specific calibration on DA methods

The DS model was previously calibrated and evaluated for the RB867515 cultivar, with experiments 16 to 22 (Tab. 1). This cultivar is one of the main Brazilian cultivars present in more than 20% of sugarcane plantations in Brazil (Vianna et al., 2020); and which is also present in seven of the fifteen field experiments not utilized in the calibration procedure (Tab. 1). The remaining experiments (4 to 11) used four different cultivars, where three were also commonly planted in Brazil (R570, RB72454, and SP832847). The NCo376 is one of the main South African cultivars for which the DSSAT/Canegro model was extensively tested (Marin et al., 2011; Singels et al., 2008).

Therefore, this analysis aimed to investigate the influence of assimilating data from different genotypes but considering fixed crop parameters previously calibrated for the RB867515 cultivar. For that, we performed simulations for OP and three DA methods (EnKF, ES, and WM) using the same calibration for all experiments (Appendix A - Table 1). The effect of employing genotype-specific calibration was then evaluated by grouping the experiments by cultivar type and comparing the RMSE of the cultivar group with the calibrated cultivar (RB867515). This resulted in two groups of experiments: (i) the experiments with cultivar RB867515, considering in this group only those not directly used for calibration (1 to 3 and 12 to 15); (ii) the experiments with different cultivars, with no genotype-specific calibration. These results provide practical information to whether PBMs must be calibrated prior to being used with DA methods across cultivars, and what is the level of uncertainty of this procedure.

2.2.9. Effect of the number and timing of observations on DA methods

To assess the impact of number and timing of LAI observations on the performance of SFY simulations, we ran a simulation experiment omitting and prescribing LAI observations to DA methods at different combinations of crop developmental stages. We divided the crop cycle in three main stages: stage 1 – early stage of development from planting to the maximum number of tillers; stage 2 – medium stage of development from a maximum number of tillers to the maximum LAI; and stage 3 – late stage of development from maximum LAI to harvest. Seven study cases scenarios were tested (Tab. 2) considering these three stages. Cases 1 to 3 assimilated LAI observations from only one specific stage. For verifying the relationship between crop stages, we performed study cases 4 to 6, which grouped LAI observations two at a time. Finally, we simulated DAs considering all observations (Case 7). For this analysis, we only used experiments 1 and 2 (Tab. 1), because they presented enough observations in each stage, different from the other experiments. For each stage of development, we used only four observations, to maintain the same number of observations for all stages (Appendix A - Table 3).

Casa	Description	Number of LAI observations		
Case	Description	Exp. 1	Exp. 2	
1	Observations only at stage 1	4	4	
2	Observations only at stage 2	4	4	
3	Observations only at stage 3	4	4	
4	Observations only at stages 1 and 2	8	8	
5	Observations only at stages 1 and 3	8	8	
6	Observations only at stages 2 and 3	8	8	
7	All observations	12	12	

Table 2. Study cases based on the sugarcane crop development and the number of leaf area index (LAI) observations presented in each stage for experiments 1 and 2. The description of crop phenological stages is shown in section 2.4.6.

2.3. Results

2.3.1. Influence of using fixed crop parameters in the performance of DA methods across different cultivars

The RMSE simulated by the DS model with OP and three DA methods was grouped by cultivar (Tab. 3). Amongst the DA methods, the WM had the best accuracy for the calibrated cultivar (RB867515) (RMSE = 15.36 Mg ha⁻¹), followed by ES (RMSE = 16.70 Mg ha⁻¹) and EnKF (RMSE = 16.76 Mg ha⁻¹) (Tab. 3). For the other cultivars, the WM presented a lower accuracy compared to EnKF and ES. Between EnKF and ES, the RMSE per cultivars were very similar (Tab. 3).

For the OP simulations, our results showed that the calibrated cultivar (RB867515), had the best accuracy for OP simulations (RMSE = 19.94 Mg ha⁻¹), followed by R570 (RMSE = 21.26 Mg ha⁻¹). However, the other cultivars presented a larger RMSE, with cultivar RB72454 showing the worse accuracy (RMSE = 32.79 Mg ha⁻¹, Tab. 3).

Nevertheless, the cultivars NCo376, R72454 and SP832847 presented a higher RMSE reduction compared to the experiments with cultivar RB867515, for ES and EnKF, with an average reduction of -22.4% and -22.5%, whereas the RB867515 showed a reduction of -16.2% and -15.9%, respectively (Tab. 3). For the WM method, the error reduction was lower than other methods, for the non-calibrated cultivars, and obtained the higher reduction in RMSE (-23.0%) for the RB867515 (Tab. 3). For the cultivar R570, the three DA methods resulted in an RMSE increase of 2.5%, 2.6% and 10%, respectively for ES, EnKF and WM. This may be caused to the distinct traits of R570, compared to the other cultivars and RB867515. We describe possible reasons for these differences in section 2.5.3.

The correlation between the RMSE of the three DA with the OP was higher ($R^2 = 0.650$, p<0.01, Fig. 1), which demonstrates that OP simulations had a direct relation with results of DA methods. As result, DA methods had their simulation performance linked to the DS parametrization capacity of capturing the dynamics of key model states variables and influencing the ensemble spread.

C1ti]	RMSE (Mg ha ⁻¹)	
Cultivar	OP	WM	EnKF	ES
NCo376	29.36	24.90 (-15.2%)	21.83 (-25.7%)	21.85 (-25.6%)
R 570	21.26	23.38 (10.0%)	21.82 (2.6%)	21.79 (2.5%)
RB72454	32.79	28.37 (-13.5%)	26.23 (-20.0%)	26.27 (-19.9%)
SP832847	27.45	23.87 (-13.0%)	21.44 (-21.9%)	21.50 (-21.7%)
RB867515	19.94	15.36 (-23.0%)	16.76 (-15.9%)	16.70 (-16.2%)

Table 3. Root mean squared error (RMSE Mg ha⁻¹) for stalk fresh yield grouped by cultivar, simulated with open-loop (OP) and three data assimilation methods: Weighted mean (WM), Ensemble Kalman filter (EnKF), and Ensemble smoother (ES). The values in brackets are the RMSE reduction compared to the OP simulations.



Figure 1. Relation between the root mean squared error (RMSE Mg ha⁻¹) for stalk fresh yield, grouped by cultivar, simulated with open-loop (OP) and three data assimilation methods: Weighted mean (WM), Ensemble Kalman filter (EnKF), and Ensemble smoother (ES). Dashed black line represent the 1:1 adjustment, whereas the red dashed line is the regression between RMSE OP and RMSE DA.

2.3.2. Comparing data assimilation methods

The regression analyses between DS simulations and observations of all experiments showed that OP simulations of SFY presented an agreement of $R^2 = 0.323$ and an RMSE = 24.588 Mg ha⁻¹. The DA methods improved the accuracy for SFY simulation, being the best results observed with ES ($R^2 = 0.498$, RMSE = 20.268 Mg ha⁻¹), followed by EnKF ($R^2 = 0.497$, RMSE = 20.276 Mg ha⁻¹) and WM ($R^2 = 0.458$, RMSE = 21.588 Mg ha⁻¹) (Fig. 2) Also, all the simulations presented a high significance (p<0.01) (Fig. 2). The EnKF and ES had similar values of RMSE and slope coefficient (b) (Fig 2), as well as closer R^2 values corresponding to an increase of 54% as compared to OP (Fig. 2). This suggested that, for our data, the ES method, which performs just one global assimilation, resulted in similar accuracy and precision to the multiples model re-initialization used by EnKF method. The WM had a lower RMSE and R^2 compared to the other DA methods but presented an increase in simulations precision of 41.54 % and a decrease on RMSE of 13.82%.



Figure 2. Comparison of observed and simulated stalk fresh yield (SFY) (a) with open-loop simulations (OP) and three data assimilation methods: (b) Ensemble Kalman filter (EnKF), (c) Ensemble smoother (ES), and (d) Weighted mean (WM). Dashed black line represents the 1:1 adjustment, whereas the red dashed line is the regression between RMSE OP and RMSE DA methods. Squared symbols represent the experiments with cultivar RB867515 and triangles represent the experiments with no genotype-specific set of parameters (NCo376, R570, RB72454, and SP832847).

2.3.3. Analyzing DA methods for different experiments

We compared the AE of SFY simulations individually for each experiment (Tab. 4) and found DA methods having better results than OP simulations in most cases. The WM presented the best accuracy for six experiments (1, 2, 3, 6, 10 and 12), but in average it showed the highest AE average (17.17 Mg ha⁻¹, Tab. 4). The EnKF approach was the best method for four experiments (7, 9, 11, and 15), and showed an average AE of 16.58 Mg ha⁻¹ (Tab. 4); the ES performed best for only one experiment (5), however presented the lower average AE (16.55 Mg ha⁻¹) (Tab. 4) and for four experiments (4, 8, 13 and 14) the OP simulation showed the best accuracy. Across experiments, EnKF and ES outperformed the OP in the 11 experiments and the WM for 10 experiments.

Between those experiments for which OP showed the best performance, experiments 8 and 13 presented low AE values (1.25 and 0.77 Mg ha⁻¹) (Tab. 4). This can be attributed to the fact that DA methods, when generating their ensembles, made the simulations more spread than OP, which resulted in higher errors compared to the OP simulations that had very high accuracy. As a result, in the assimilation step of the DA methods, the dispersion of the ensemble resulted in a high analysis error variance compared to the LAI observations variance. This led the DA methods to estimate LAI closer to the observed values, which in this case, resulted in less accurate estimation of SFY.

We also verified an increase of AE for Experiment 4, where we employed the DS parameters calibrated for the RB867515 to simulate R570 cultivar under irrigated conditions. Experiment 5 was also conducted in parallel to Experiment 4 to evaluate the growing performance of cultivar R570 under rainfed and irrigated conditions, yet both trials showed very similar SFY values (\approx 110 Mg ha⁻¹). In contrast, the DS simulations resulted in substantially different SFY values (Tab. 4), where the irrigated experiment 4 showed higher yields (SFY = 139-143 Mg ha⁻¹) than the rainfed experiment 5 (SFY = 101-110 Mg ha⁻¹). This could indicate that cultivars RB867515 and R570 have distinct traits which may require a specific calibration for cultivar R570.

Experiments 13 and 14 had only one LAI measurement in the database that was not enough to improve model accuracy by any DA method (Tab. 4). For these experiments, one LAI observation resulted in increased AE values as compared to the OP simulations. The number of LAI observations had a significant correlation with the difference of AE between OP and the three DA methods, 1% for WM (p<0.01) and 5% for ES and EnKF (p<0.05) (Fig. 3). The WM presented the highest correlation ($R^2 = 0.508$), followed by ES ($R^2 = 0.372$) and EnKF ($R^2 = 0.368$) (Fig. 3).

Moreover, Experiments 1 and 2 presented a higher number of LAI observations compared to others, 22 and 23 respectively, in these conditions the DA methods promoted the best absolute improvements in SFY simulations (Fig. 4 and 5). The WM had the best accuracy (AE = 1.68 and 2.63 Mg ha⁻¹) followed by ES (AE = 2.88 and 4.61 Mg ha⁻¹) and EnKF (AE = 4.86 and 2.99 Mg ha⁻¹) (Fig. 4 and 5). For experiments with 4 LAI observations (e.g. Experiments 4 to 11), the AE differences were positive for two (Experiments 4 and 8) and negative in the others, for all DA methods. For the experiments with 5 and 7 LAI observations, the AE differences were negative for all experiments. In Experiment 15, where 6 LAI observations were available, the AE difference was negative for EnKF and ES and positive for WM.

However, excluding experiments 1 and 2, and considering the EnKF and ES methods, the higher reduction on SFY's AE was for experiment 11, with 4 LAI observations (AE difference = -10.50 and -10.47 Mg ha⁻¹), followed by experiments 3, with 7 LAI observations (AE difference = -9.15 and -9.60 Mg ha⁻¹) and experiment 7 with 4 LAI observations (AE difference = -8.77 and -8.71 Mg ha⁻¹). Moreover, the higher relative reduction was for experiment 5, with 4 observations (98% and 99%) followed by experiment 1 (84% and 75%) and 2 (72% and 83%). Despite the increase in the number of LAI observations tended to reduce AE for DA simulations, the magnitude of these reductions was not directly affected by the number of LAI observations, mainly for experiments with 4 to 7 observations. Furthermore, considering experiments 6 to 12, and 15, there was a high correlation between the AE of OP simulations to the AE reduction. Thus, the initial errors of the model influenced the AE reduction, because it should be easier to reduce higher values of AE than lower values, due to the model structure.

		Number of I AI		AE (N	<u> </u>	Best simulation	
Experiment	Location	observations	OP	WM	EnKF	ES	method
1	PIRA	22	17.27	1.68	4.86	2.88	WM
2	PIRA	23	18.35	2.63	2.99	4.61	WM
3	PIRA	7	28.41	13.51	19.26	18.81	WM
4	PIRA	4	28.76	32.88	30.85	30.81	OP
5	PIRA	4	8.79	3.58	0.13	0.07	ES
6	PIRA	4	15.25	10.96	12.20	12.21	WM
7	PIRA	4	43.79	38.60	35.03	35.08	EnKF
8	PIRA	4	1.25	3.15	1.36	1.35	OP
9	PIRA	4	38.79	33.61	30.29	30.37	EnKF
10	PIRA	4	19.25	15.73	16.17	16.19	WM
11	PIRA	4	36.79	31.51	26.29	26.32	EnKF
12	APTA	5	29.61	21.84	25.07	24.84	WM
13	COSP	1	0.77	3.89	2.56	2.56	OP
14	OLSP	1	9.25	22.04	26.44	26.44	OP
15	COAL	6	19.46	21.94	15.15	15.72	EnKF
Mean			21.05	17.17	16.58	16.55	ES
SD			12.70	12.42	11.68	11.70	

Table 4. Absolute error (AE Mg ha⁻¹) for different experiments used for open-loop (OP) and data assimilations methods: Weighted mean (WM), Ensemble Kalman filter (EnKF) and Ensemble smoother (ES) for final stalk fresh yield simulations.

PIRA: Piracicaba/SP, APTA: Aparecida do Taboado/MS, COSP: Colina/SP, OLSP: Olímpia/SP, COAL: Coruripe/AL, LAI: Leaf area index.



Figure 3. Relation between the number of leaf area index observations assimilated in each experiment and the absolute error (AE) difference between stalk fresh yield (SFY) simulated with one of the three data assimilation methods: (a) Weighted mean (WM), (b) Ensemble Kalman filter (EnKF) or (c) Ensemble smoother (ES) methods and the open-loop simulation (OP).



Figure 4. Comparison between simulated (lines) and observed (circles) leaf area index (LAI) and stalk fresh yield (SFY) for experiment 1, (a, e) open-loop simulations (OP) and three methods of data assimilations: (b, f) Weighted mean (WM), (c, g) Ensemble smoother (ES) and (d, h) Ensemble Kalman filter (EnKF).



Figure 5. Comparison between simulated (lines) and observed (circles) leaf area index (LAI) and stalk fresh yield (SFY) for experiment 2, (a, e) open-loop simulations (OP) and three methods of data assimilations: (b, f) Weighted mean (WM), (c, g) Ensemble smoother (ES) and (d, h) Ensemble Kalman filter (EnKF).

2.3.4. Impact of leaf area index observations in sugarcane growth stage

Seven cases were studied to investigate the impact of timing of LAI observations in DA simulations (Fig. 6). Differences in DA performances among development stages (cases 1 to 3) were observed. For stages 1 and 2 (case 1 and 2) there was an accuracy improvement only for experiment 1, for the three methods (Fig. 6), with AE values lower than OP simulations. However, at stage 3 (case 3), for both experiments and three methods, there was a substantial accuracy improvement (Fig. 6) compared to OP simulations.

When the LAI observations of stages 1 and 2 were assimilated together (case 4), it increased the AE compared to other cases and OP simulations. Also, case 4 presented the highest AE for experiments 1 and 2 for the ES and EnKF methods (Fig. 6). Nevertheless, the assimilation of LAI data at stages 1 and 3 in case 5 improved the accuracy of DS simulations for all methods, and presented the lower AE, as compared to all cases. In Case 6, assimilation of LAI observations at stages 2 and 3, also presented an accuracy improvement compared to OP to cases 1 to 4, for all methods. Across crop stages, the observations assimilated at stage 3 showed a high influence for the accuracy improvement of the DS simulations, for all DA methods. However, LAI observations assimilated at stages 1 and 2 did not always result in better accuracy than OP simulations. Cases 1, 2 and 4 presented higher AE than OP simulations in experiment 2 and were not optimal in experiment 1 (Fig 6). This may have occurred because the sugarcane growing season is long, and assimilating LAI observations only at early stages may not be able to correct factors occurring in the remaining part of the cycle. Conversely, assimilating LAI measurements only at stage 2, when the crop is experiencing full canopy development still may not correct simulations after the onset of tillering senescence process, leading to overestimated LAI simulations. On the other hand, data assimilation during stage 3, covers most of the stalk formation stage of sugarcane which largely characterizes SFY. Moreover, at stage 3, is common for sugarcane, with high SFY and height, to suffer from lodging, which can reduce the LAI and the final SFY.



Figure 6. Absolute error (AE, Mg ha⁻¹) for stalk fresh yield for seven study cases (Tab. 2), for (a) experiment 1 and (b) experiment 2, and for three data assimilation methods: Weighted mean (WM), Ensemble smoother (ES) and Ensemble Kalman filter (EnKF). Dashed lines represent the open-loop AE.

2.4. Discussion

In this study, three DA methods were used (WM, EnKF, and ES) to assimilate LAI observations along the growing cycle to improve SFY simulations as compared to the OP method. The ES and EnKF had very similar results and presented the best beneficial impact on the accuracy of SFY simulations across all field experiments and cultivars (Fig 2). Nevertheless, for experimental conditions with high LAI sampling such as those found in field experiments 1 and 2, the WM method presented the best improvements in SFY simulations, followed by ES and EnKF (Fig. 4 and 5). Thus, our results do not show a predominance of a unique DA method for all experiments analyzed (Tab. 4).

The WM performs the assimilation without updating model's internal variables, only utilizing the assimilated observation to select the ensemble member that is closer to crop state. This method works well with DS,

and this agreed with the results found by Tewes et al. (2020a), for wheat simulations with LINTUL5 model. In addition, WM can be easily applied with DS, because it did not require adaptations in the code for DA. However, our data showed that this method resulted in a low accuracy gain, compared to others DA methods when the DS was run for not calibrated cultivars.

The EnKF and ES presented very similar results. Part of this can be attributed to the lower error of LAI observations (4%), which narrows the EnKF and ES estimates around the observations (Fig. 4 and 5). A derivation method of ES, the iterative ensemble smoother, was applied by Yu et al. (2020) and resulted in an interesting improvement of stalk dry mass accuracy, for sugarcane simulated with SWAP-WOFOST with plant-high observations. However, more studies are required to understand the ES dynamic with remote sensing data for large areas.

The assimilation of LAI within the DS model improved the accuracy of SFY simulations for the majority of experiments, highlighting the potential for using DA methods with crop models. However, despite the improvements some experiments and DA methods did not show superior performance than the OP simulations or presented a lower accuracy gain. Therefore, this can be linked to three key factors for DA methods performance: number of observations, model parametrization and the state variables assimilated.

Because sugarcane crop remains in the field for longer periods, this may increase the need for observations throughout the crop cycle to be assimilated in the model. Morel at al. (2014) recommended more than five LAI observations for forcing methods. For this study, almost all the experiments with four observations resulted in improved SFY simulations, for the three DA methods tested. However, the absolute decrease in AE was only partly explained by the number of LAI observations. Moreover, the crop phenological stage has high importance for the DA methods (Pan et al., 2019; Tewes et al., 2020b; Xie et al., 2018; Yu et al., 2020). Our results demonstrated that the late stage was the more important, which corroborate with Yu et al. (2020).

Our results showed that the performance of the DA methods, in situations where there were no previous cultivar-specific calibration, resulted in lower accuracy likely due to the uncertainty of OP simulations. This agrees with Kang and Özdoğan (2019), which showed that the EnKF had better yield estimation for maize when calibration was applied prior to LAI data assimilation. In contrast, Lu et al. (2021) improved the accuracy of maize yield estimation without prior genotype-specific calibration. In their study, the EnKF method was employed to assimilate canopy cover data for adjusting the phenological parameters of the AquaCrop model for each crop season. Huang et al. (2016) also improved wheat yield estimations of the WOFOST model using the EnKF with LAI observations and fixed crop parameters for an entire region in China. Although an overall improvement of performance was found across cultivars using a fixed set of parameters for DS, the simulation performance was superior in experiments involving the calibrated cultivar (RB867515) (Fig. 1 and Tab. 3). The DS model is a PBM which simulates various processes specific to the sugarcane crop system, with good performance (Vianna et al., 2020). Thus, it is recommended for this type of model a previous calibrated set of parameters, differently from that observed for more generic crop models, which obtained more accurate results using DA even without any previous calibration (Hu et al., 2019).

The observation's ability to correct the PBM imperfections is the success of DA methods (Huang et al, 2019). However, updating one state variable only (e.g. LAI) does not always fully correct model behavior and may not properly translate the effects of limiting and reducing factors to other physiological processes simulated. Hence, mechanistically consistent simulations of SFY are unguaranteed, despite the use of LAI in DA having already been tested and demonstrated a more considerable relevance than others for yield estimates (Ines et al., 2013; Pan et al.,

2019). Still, the use of LAI allied with other variables might improve the performance of DA methods. Hu et al. (2019) demonstrated that when the water stress is high, assimilating LAI with soil moisture presents better results compared to LAI alone.

Moreover, comparing our results to Hu et al. (2019), the EnKF methods with SWAP-WOFOST and LAI assimilations had an $R^2 = 0.45$ and RMSE = 8.73 Mg ha⁻¹, which was more accurate than those found in our study (Fig. 2). However, these results were conducted with an LAI frequency of 8 – 16 days and part of the experiment was conducted at water stress and with leaf strip (Hu et al., 2019), which differs from most of all our experiments. Compared to experiments 1 and 2, which have a similar number of LAI observations, the AE showed similar results (Fig. 3 and 4). Furthermore, the OP simulation of the SWAP-WOFOST showed a lower agreement with R2 = 0.05, compared with DS, despite the great accuracy (RMSE = 10.16 Mg ha⁻¹). However, more studies are needed to intercompare sophisticated and generalist crop models at the same conditions and calibration. Understanding in which cases the robustness of the PBM affects the DA simulations could be useful for choosing the best DA methods and configurations in different conditions.

Although the DA techniques employed in this study showed improved performance of SFY simulations, we found some limitations in our study. As follows, we emphasize some points that can be attributed to improving the use of DA with DS for subsequent studies. Firstly, the calibration of the model should be better explored, apparently requiring one set of calibrated parameters for each cultivar because of the complexity of the PBM considered in our study. Secondly, the experiments used in our study followed a strict protocol to ensure there was no interference of reducing factors, and these are required conditions for PBM to simulate more efficiently. Therefore, studies with this methodology in areas where the reduction factors are present, or there would be data flaws in describing the environment, such as soil characterization and climate data, may result in a more pronounced accuracy gain due to the use of DA methods (Gilardelli et al., 2019). Thirdly, other state variables simulated by DS, such as above ground biomass (Lu et al., 2021) plant height (Yu et al., 2020), soil moisture (Hu et al., 2019; Zhuo et al., 2019), canopy nitrogen accumulation (Li et al., 2015), and canopy cover (Lu et al., 2021) could be also used to enhance the model accuracy. Yet, the used PBM was developed to run without assimilations from start to end, for this study we just adjusted the state variables directly related to the LAI, plant weight, and leaf area. However, further studies could explore the allometric relations between LAI with the number of stalks, stalk height, and other related crop variables to simultaneously update these variables without direct measurements.

2.5. Conclusion

This study aimed to investigate the performance of DA methods when a genotype-specific calibration for a PBM is not available, as well as the combination of data amount and timing. This was done by means of a robust experimental field database for sugarcane, comprising 5 cultivars across different edaphoclimatic conditions in Brazil. The three DA methods obtained a better accuracy (RMSE) and precision (R²) if compared to the OP simulations. In general, considering all experiments the best results can be attributed to ES followed by EnKF and WM. The high accuracy occurred, for the three DA methods, when the cultivar-specific calibration was used to corresponding cultivar (e.g. RB867515). Moreover, the three DA methods outperformed the OP for three of four non-calibrated cultivars. However, the response for relative reduction on SFY's RMSE, changed between the methods, with EnKF and ES presenting higher reduction for the non-calibrated cultivars and WM for the calibrated cultivar. As the number of LAI observations increase, the DA methods tend to outperform the OP, nevertheless, this just explained part of the accuracy improvement by applying DA methods. In addition, we observed that LAI measurements at late stages more strongly influence the SFY accuracy. This study provides a potential methodology that can be applied for regional applications with LAI retrieved from remote sensing techniques.

References

- Byrd, R.H., Peihuang, L., Nocedal, J., 1996. A limited-memory algorithm for bound-constrained optimization. https://doi.org/10.2172/204262
- Cukier, R.I., Fortuin, C.M., Shuler, K.E., Petschek, A.G., Schaibly, J.H., 1973. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory. J. Chem. Phys. 59, 3873–3878. https://doi.org/10.1063/1.1680571
- Curnel, Y., de Wit, A.J.W., Duveiller, G., Defourny, P., 2011. Potential performances of remotely sensed LAI assimilation in WOFOST model based on an OSS Experiment. Agric. For. Meteorol. 151, 1843–1855. https://doi.org/10.1016/j.agrformet.2011.08.002
- de Wit, A.J.W., van Diepen, C.A., 2007. Crop model data assimilation with the Ensemble Kalman filter for improving regional crop yield forecasts. Agric. For. Meteorol. 146, 38–56. https://doi.org/10.1016/j.agrformet.2007.05.004
- Evensen, G., 2003. The Ensemble Kalman Filter: Theoretical formulation and practical implementation. Ocean Dyn. 53, 343–367. https://doi.org/10.1007/s10236-003-0036-9
- FAO, 2019. Food and Agriculture Organization of the United Nations. FAOSTAT. URL http://www.fao.org/faostat/en/#home
- Gilardelli, C., Stella, T., Confalonieri, R., Ranghetti, L., Campos-Taberner, M., García-Haro, F.J., Boschetti, M., 2019. Downscaling rice yield simulation at sub-field scale using remotely sensed LAI data. Eur. J. Agron. 103, 108– 116. https://doi.org/10.1016/j.eja.2018.12.003
- Goldemberg, J., Mello, F.F.C., Cerri, C.E.P., Davies, C.A., Cerri, C.C., 2014. Meeting the global demand for biofuels in 2021 through sustainable land use change policy. Energy Policy 69, 14–18. https://doi.org/10.1016/j.enpol.2014.02.008
- Gonçalves, A.O., Silva, E.H.F.M.D., Gasparotto, L.G., Rosa, J.M., Carmo, S.D., Fattori Junior, I.M., Marin, F.R., 2020. Improving indirect measurements of the leaf area index using canopy height. Pesqui. Agropecu. Bras. 55, 1–9. https://doi.org/10.1590/S1678-3921.PAB2020.V55.01894
- Holzworth, D.P., Huth, N.I., deVoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Chenu, K., van Oosterom, E.J., Snow, V., Murphy, C., Moore, A.D., Brown, H., Whish, J.P.M., Verrall, S., Fainges, J., Bell, L.W., Peake, A.S., Poulton, P.L., Hochman, Z., Thorburn, P.J., Gaydon, D.S., Dalgliesh, N.P., Rodriguez, D., Cox, H., Chapman, S., Doherty, A., Teixeira, E., Sharp, J., Cichota, R., Vogeler, I., Li, F.Y., Wang, E., Hammer, G.L., Robertson, M.J., Dimes, J.P., Whitbread, A.M., Hunt, J., van Rees, H., McClelland, T., Carberry, P.S., Hargreaves, J.N.G., MacLeod, N., McDonald, C., Harsdorf, J., Wedgwood, S., Keating, B.A., 2014. APSIM Evolution towards a new generation of agricultural systems simulation. Environ. Model. Softw. 62, 327–350. https://doi.org/10.1016/j.envsoft.2014.07.009
- Hoogenboom, G., Porter, C.H., Boote, K.J., Shelia, V., Wilkens, P.W., Singh, U., White, J. W., Asseng, S., Lizaso, J.I., Moreno, L.P., Pavan, W., Ogoshi, R., Hunt, L.A., Tsuji, G. Y., Jones, J.W., 2019. The DSSAT crop modeling ecosystem. In: Boote, K.J. (Ed.), Advances in Crop Modeling for a Sustainable Agriculture. Burleigh Dodds Science Publishing, Cambridge, United Kingdom, pp. 173–216.

- Hu, S., Shi, L., Huang, K., Zha, Y., Hu, X., Ye, H., Yang, Q., 2019. Improvement of sugarcane crop simulation by SWAP-WOFOST model via data assimilation. F. Crop. Res. 232, 49–61. https://doi.org/10.1016/j.fcr.2018.12.009
- Huang, J., Gómez-Dans, J.L., Huang, H., Ma, H., Wu, Q., Lewis, P.E., Liang, S., Chen, Z., Xue, J.H., Wu, Y., Zhao, F., Wang, J., Xie, X., 2019. Assimilation of remote sensing into crop growth models: Current status and perspectives. Agric. For. Meteorol. 276–277, 107609. https://doi.org/10.1016/j.agrformet.2019.06.008
- Huang, J., Sedano, F., Huang, Y., Ma, H., Li, X., Liang, S., Tian, L., Zhang, X., Fan, J., Wu, W., 2016. Assimilating a synthetic Kalman filter leaf area index series into the WOFOST model to improve regional winter wheat yield estimation. Agric. For. Meteorol. 216, 188–202. https://doi.org/10.1016/j.agrformet.2015.10.013
- Huang, X., Zhao, G., Zorn, C., Tao, F., Ni, S., Zhang, W., Tu, T., Höglind, M., 2021. Grass modelling in datalimited areas by incorporating MODIS data products. F. Crop. Res. 271, 108250. https://doi.org/10.1016/j.fcr.2021.108250
- Ines, A.V.M., Das, N.N., Hansen, J.W., Njoku, E.G., 2013. Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. Remote Sens. Environ. 138, 149–164. https://doi.org/10.1016/j.rse.2013.07.018
- Inman-Bamber, N. G., Thompson, G.D., 1989. Models of dry matter accumulation by sugarcane. Proc. South African Sugarcane Technol. 63, 212–216.
- Jones, C.A., Wegener, M.K., Russell, J.S., McLeod, I.M., Williams, J., 1989. AUSCANE-Simulation of Australian sugarcane with EPIC. CSIRO.
- Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., Keating, B.A., Munoz-Carpena, R., Porter, C.H., Rosenzweig, C., Wheeler, T.R., 2017. Brief history of agricultural systems modeling. Agric. Syst. 155, 240–254. https://doi.org/10.1016/j.agsy.2016.05.014
- Jones, J.W., Hoogenboom, G., Porter, C., Boote, K., Batchelor, W., Hunt, L., Wilkens, P., Singh, U., Gijsman, A., Ritchie, J., 2003. The DSSAT cropping system model. Eur. J. Agron. 18, 235–265. https://doi.org/10.1016/S1161-0301(02)00107-7
- Jones, M.R., Singels, A., 2018. Refining the Canegro model for improved simulation of climate change impacts on sugarcane. Eur. J. Agron. 100, 76–86. https://doi.org/10.1016/j.eja.2017.12.009
- Kang, Y., Özdoğan, M., 2019. Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. Remote Sens. Environ. 228, 144–163. https://doi.org/10.1016/j.rse.2019.04.005
- Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N.G., Meinke, H., Hochman, Z., McLean, G., Verburg, K., Snow, V., Dimes, J.P., Silburn, M., Wang, E., Brown, S., Bristow, K.L., Asseng, S., Chapman, S., McCown, R.L., Freebairn, D.M., Smith, C.J., 2003. An overview of APSIM, a model designed for farming systems simulation. Eur. J. Agron. 18, 267–288. https://doi.org/10.1016/S1161-0301(02)00108-9
- Lee, K., Jung, S., Choe, J., 2016. Ensemble smoother with clustered covariance for 3D channelized reservoirs with geological uncertainty. J. Pet. Sci. Eng. 145, 423–435. https://doi.org/10.1016/j.petrol.2016.05.029
- Li, Z., Wang, J., Xu, X., Zhao, C., Jin, X., Yang, G., Feng, H., 2015. Assimilation of two variables derived from hyperspectral data into the DSSAT-CERES model for grain yield and quality estimation. Remote Sens. 7, 12400–12418. https://doi.org/10.3390/rs70912400

- Lisboa, I.P., Cherubin, M.R., Lima, R.P., Cerri, C.C., Satiro, L.S., Wienhold, B.J., Schmer, M.R., Jin, V.L., Cerri, C.E.P., 2018. Sugarcane straw removal effects on plant growth and stalk yield. Ind. Crops Prod. 111, 794–806. https://doi.org/10.1016/j.indcrop.2017.11.049
- Liu, D.L., Kingston, G., 1995. QCANE: a simulation model of sugarcane growth and sugar accumulation. In: Research and Modelling Approaches to Assess Sugercane Production Opportunities and Constraints. University of Queensland, St. Lucia, 29–25.
- Lobell, Lobell, D.B., Cassman, K.G., Field, C.B., 2009. Crop Yield Gaps: Their Importance, Magnitudes, and Causes. Annu. Rev. Environ. Resour. 34, 179–204. https://doi.org/10.1146/annurev.environ.041008.093740
- Lu, Y., Chibarabada, T.P., Ziliani, M.G., Onema, J.M.K., McCabe, M.F., Sheffield, J., 2021. Assimilation of soil moisture and canopy cover data improves maize simulation using an under-calibrated crop model. Agric. Water Manag. 252, 106884. https://doi.org/10.1016/j.agwat.2021.106884
- Marin, F., Jones, J.W., Boote, K.J., 2017. A stochastic method for crop models: Including uncertainty in a sugarcane model. Agron. J. 109, 483–495. https://doi.org/10.2134/agronj2016.02.0103
- Marin, F.R., 2016. Understanding sugarcane production, Biofuels, and market volatility in Brazil—A research perspective. Outlook Agric. 45, 75–77. https://doi.org/10.1177/0030727016649802
- Marin, F.R., de Carvalho, G.L., 2012. Spatio-temporal variability of sugarcane yield efficiency in the state of São Paulo, Brazil. Pesqui. Agropecu. Bras. 47, 149–156. https://doi.org/10.1590/S0100-204X2012000200001
- Marin, F.R., Jones, J.W., 2014. Process-based simple model for simulating sugarcane growth and production. Sci. Agric. 71, 1–16. https://doi.org/https://doi.org/10.1590/S0103-90162014000100001
- Marin, F.R., Jones, J.W., Royce, F., Suguitani, C., Donzeli, J.L., Filho, W.J.P., Nassif, D.S.P., 2011. Parameterization and evaluation of predictions of DSSAT/CANEGRO for Brazilian sugarcane. Agron. J. 103, 304–315. https://doi.org/10.2134/agronj2010.0302
- Marin, F.R., Lopes-Assad, M.L., Assad, E.D., Vian, C.E., Santos, M.C., 2008. Sugarcane crop efficiency in two growing seasons in São Paulo State, Brazil. Pesqui. Agropecuária Bras. 43, 1449–1455. https://doi.org/10.1590/s0100-204x2008001100002
- Marin, F.R., Rattalino Edreira, J.I., Andrade, J., Grassini, P., 2019. On-farm sugarcane yield and yield components as influenced by number of harvests. F. Crop. Res. 240, 134–142. https://doi.org/10.1016/j.fcr.2019.06.011
- Marin, F.R., Rattalino Edreira, J.I., Andrade, J.F., Grassini, P., 2021. Sugarcane Yield and Yield Components as Affected by Harvest Time. Sugar Tech. https://doi.org/10.1007/s12355-020-00945-5
- Marin, F.R., Thorburn, P.J., Nassif, D.S.P., Costa, L.G., 2015. Sugarcane model intercomparison: Structural differences and uncertainties under current and potential future climates. Environ. Model. Softw. 72, 372–386. https://doi.org/10.1016/j.envsoft.2015.02.019
- Martiné, J.-F., 2003. Modélisation de la production potentielle de la canne à sucre en zone tropicale, sous conditions thermiques et hydriques contrastées. Applications du modèle. Institut national agronomique Paris-Grignon.
- Morel, J., Todoroff, P., Bégué, A., Bury, A., Martiné, J.F., Petit, M., 2014. Toward a satellite-based system of sugarcane yield estimation and forecasting in smallholder farming conditions: A case study on reunion island. Remote Sens. 6, 6620–6635. https://doi.org/10.3390/rs6076620
- Nearing, G.S., Crow, W.T., Thorp, K.R., Moran, M.S., Reichle, R.H., Gupta, H. V., 2012. Assimilating remote sensing observations of leaf area index and soil moisture for wheat yield estimates: An observing system simulation experiment. Water Resour. Res. 48, 1–13. https://doi.org/10.1029/2011WR011420

- Pan, H., Chen, Z., de Allard, W., Ren, J., 2019. Joint assimilation of leaf area index and soil moisture from sentinel-1 and sentinel-2 data into the WOFOST model for winter wheat yield estimation. Sensors (Switzerland) 19. https://doi.org/10.3390/s19143161
- Saltelli, A., Tarantola, S., Chan, K.P.S., 1999. A quantitative model-independent method for global sensitivity analysis of model output. Technometrics 41, 39–56.
- Singels, A., Jones, M., Van Der Berg, M., 2008. DSSAT v4.5-Canegro Sugarcane Plant Module: Scientific Documentation.
- Suguitani, C., 2006. Entendendo o crescimento e produção da cana de açúcar: avaliação do modelo MOSICAS. Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz."
- Supit, I., Hooijer, A., van Diepen, C., 1994. Improvement of sugarcane crop simulation by SWAP-WOFOST model via data assimilation Volume 1: Theory and Algorithms.
- Tewes, A., Hoffmann, H., Krauss, G., Schäfer, F., Kerkhoff, C., Gaiser, T., 2020a. New approaches for the assimilation of LAI measurements into a crop model ensemble to improve wheat biomass estimations. Agronomy 10, 1–21. https://doi.org/10.3390/agronomy10030446
- Tewes, A., Montzka, C., Nolte, M., Krauss, G., Hoffmann, H., Gaiser, T., 2020b. Assimilation of Sentinel-2 Estimated LAI into a Crop Model: Influence of Timing and Frequency of Acquisitions on Simulation of Water Stress and Biomass Production of Winter Wheat. Agronomy 10, 1813. https://doi.org/10.3390/agronomy10111813
- van Dam, J.C., Huygen, J., Wesseling, J.G., Feddes, R.A., Kabat, P., van Walsum, P.E. V., Groenendijk, P., van Diepen, C.A., 1997. Theory of SWAP version 2.0; simulation of water flow, solute transport and plant growth in the soil-water-atmosphere-plant environment.
- Van Ittersum, M.K., Leffelaar, P.A., Van Keulen, H., Kropff, M.J., Bastiaans, L., Goudriaan, J., 2003. On approaches and applications of the Wageningen crop models. Eur. J. Agron. 18, 201–234. https://doi.org/10.1016/S1161-0301(02)00106-5
- Van Leeuwen, P.J., Evensen, G., 1996. Data Assimilation and Inverse Methods in Terms of a Probabilistic Formulation. Mon. Weather Rev. 124, 2898–2913. https://doi.org/doi.org/10.1175/1520-0493(1996)124<2898:DAAIMI>2.0.CO;2
- Vianna, M. dos S., Nassif, D.S.P., dos Santos Carvalho, K., Marin, F.R., 2020. Modelling the trash blanket effect on sugarcane growth and water use. Comput. Electron. Agric. 172, 105361. https://doi.org/10.1016/j.compag.2020.105361
- Villegas, F.D., Daza, O.H., Jones, J.W., Royce, F.S., 2005. CASUPRO: An industry-driven sugarcane model. 2005 ASAE Annu. Int. Meet. 0300. https://doi.org/10.13031/2013.19062
- Wallach, D., Makowski, D., Jones, J.W., Brun, F., 2018. Working with Dynamic Crop Models: Methods, Tools and Examples for Agriculture and Environment. Academic Press.
- Wang, E., Brown, H. E., Rebetzke, G. J., Zhao, Z., Zheng, B., Chapman, S. C., 2019. Improving process-based crop models to better capture genotype× environment× management interactions. Journal of Experimental Botany, 70, 2389-2401. https://doi.org/10.1093/jxb/erz092
- Xie, Y., Wang, P., Bai, X., Khan, J., Zhang, S., Li, L., Wang, L., 2017. Assimilation of the leaf area index and vegetation temperature condition index for winter wheat yield estimation using Landsat imagery and the CERES-Wheat model. Agric. For. Meteorol. 246, 194–206. https://doi.org/10.1016/j.agrformet.2017.06.015

- Yu, D., Zha, Y., Shi, L., Jin, X., Hu, S., Yang, Q., Huang, K., Zeng, W., 2020. Improvement of sugarcane yield estimation by assimilating UAV-derived plant height observations. Eur. J. Agron. 121, 126159. https://doi.org/10.1016/j.eja.2020.126159
- Zhao, Y., Chen, S., Shen, S., 2013. Assimilating remote sensing information with crop model using Ensemble Kalman Filter for improving LAI monitoring and yield estimation. Ecol. Modell. 270, 30-42. https://doi.org/10.1016/j.ecolmodel.2013.08.016
- Zhuo, W., Huang, J., Li, L., Zhang, X., Ma, H., Gao, X., Huang, H., Xu, B., Xiao, X., 2019. Assimilating Soil Moisture Retrieved from Sentinel-1 and Sentinel-2 Data into WOFOST Model to Improve Winter Wheat Yield Estimation. Remote Sens. 11, 1618. https://doi.org/10.3390/rs11131618

3. ASSESSING THE INFLUENCE OF CROP MODEL STRUCTURE ON THE PERFORMANCE OF DATA ASSIMILATION FOR SUGARCANE

Abstract

Process-base crop models (PBM) are important tools to describe how the agricultural system responds to environmental conditions. Sugarcane represents a major world source of sugar and ethanol and its PBMs had different levels of complexity in terms of structure, i.e. how detailed their processes were described. Yet, literature has widely demonstrated that data assimilation techniques (DA) represent a valuable option for reducing model uncertainty, but the inconsistency between PBM and the assimilated variable can significantly affect the performance of DA. Such limitation is strictly connected to model structure, and a hypothesis that arises from literature is that the use of more complex models would reduce model uncertainty after DA. We accessed the performance of using two different PBMs, one more detailed (DSSAT/SAMUCA, DS) and the other more general (WOFOST, WO), by assimilating leaf area index (LAI) retrieved from Landsat 7 ETM+ and 8/OLI, using the Ensemble Kalman Filter (EnKF). Both PBMs were calibrated and evaluated with a robust database of 13 experiments and evaluated against a sugarmill database to evaluate the EnKF performance, compared with simulations without DA (Open-loop, OP). Moreover, the processes involved in LAI simulations were analyzed to access the EnKF performance. The DS had superior performance in the calibration and evaluation step with EF = 0.907, 0.878, 0.458 for stalk dry mass, stalk fresh vield (SFY), and LAI, while WO showed EF = 0.622, 0.610, 0.417 for the same variables, respectively. The calibration step affected the OP simulation of commercial fields with DS having higher accuracy (RMSE = 31.678 Mg ha⁻¹) and precision (R² = 0.509), compared with WO (RMSE = 39.593 Mg ha⁻¹; $R^2 = 0.458$). However, after the assimilation, both PBM presented error inconsistency with EnKF, despite the decrease in RMSE (-44.73% and -29.58%) and increase in R² (22.15% and 36.50%) of DS and WO, respectively. The error inconsistency diverged from each PBM: the OP simulation of DS overestimated the Landsat LAI; after DA, simulated LAI decreased resulting in SFY underestimation (BIAS = -11.469 Mg ha⁻¹). The WO showed OP simulations for LAI closer to Landsat's LAI values, despite the positive BIAS in SFY estimation, and so EnKF slightly reduced the SFY overestimation (BIAS = 22.944 Mg ha⁻¹). Thus, the better descriptions of DS in terms of structure did not inhibit the error inconsistency. We suggested that new studies are required to understand how the assimilated variables impact on the other state variables of the PBM.

Key-words: Ensemble Kalman filter, DSSAT/SAMUCA, WOFOST, Landsat, Remote sensing, Leaf area index

3.1. Introduction

Process-based crop models (PBM) have been largely used for decision making and planning in agriculture, because of the capability of describing how the agricultural system responds to environmental conditions (Morell et al., 2016). The PBMs are capable of a variety of tasks, including yield forecasts in response to weather variables and the impacts of management changes. For that, the PBMs works mechanistically and attempts to explain the genotype \times environment \times management interactions (Wang et al., 2019). These models consist of organized algorithms that describe physical and biological processes that occur in crop growth and development, and so mimic the interactions among soil, plant, and atmosphere components (Jones et al., 2017). Studies with PBMs are important for advancing the scientific knowledge on crop ecophysiology and management and, as an applied tool, to support government and private agencies, food security policies, and planning (Curnel et al., 2011). Sugarcane is a key crop because it represents a major source of sugar and ethanol. Brazil has ca. 8.6 million ha planted with sugarcane, producing ca. 654 million metric tons (Mt) of harvested stalk fresh yield (CONAB 2022). Brazil is the world largest producing country, accounting for ca. 38% of global sugarcane production and 50% of global sugar exports (Marin et al. 2019a).

Different PBMs have been developed and evaluated for sugarcane across the world, and these models present different levels of complexity in terms of biophysical process representation in their code (Marin et al., 2015). For example, SAMUCA model considers physiological processes such as biomass partitioning at phytomer level, canopy carbon assimilation using leaf assimilation rates, and carboxylation efficiency, and can simulate specific sugarcane traits like plant height, tiller population, sucrose accumulation, and stalk moisture (Marin and Jones, 2014; Marin et al., 2017). Furthermore, SAMUCA was recently updated and evaluated considering the effect of the green cane trash blanket effect, arising from management practice in Brazil (Vianna et al., 2020). On the other hand, the WOFOST (WO) (World Food Studies, de Wit et al., 2019) is a more generalist PBM, that can simulate different crops, with a reasonable description of photosynthesis, respiration, biomass portioning, and soil water balance. The WO was adapted for sugarcane simulation (Hu et al., 2019; Scarpare, 2011; van Heemst, 1988), but without simulating specific traits of sugarcane growth and development.

Nevertheless, the mechanist principles of PBMs are only approximations of reality and require a substantial amount of input data and parameters to characterize the soil-plant-atmosphere and the management system, which are sometimes difficult to provide and may vary within the spatial and temporal domains (Manivasagam and Rozenstein, 2020; Marin et al., 2017). One way to reduce the uncertainty of PBM simulations is to insert real-time information by using a data assimilation algorithm (DA) (Huang et al., 2019). These DA methods can be classified into two groups: 1) the variational algorithms, which use all observations of a timestamp to update the model trajectory; 2) the sequential algorithms, which use only the corresponding time observation to estimate the new model state (Huang et al., 2019; Kang and Özdoğan, 2019). One sequential algorithm that has been widely used is the Ensemble Kalman Filter (EnKF) for crop yield estimation (Evensen, 2003; Ines et al., 2013; Pan et al., 2019). The EnKF is capable of counting with both model and observation error, due to the Bayesian approach and a Monte-Carlo realization of the PBM to estimate model error covariance (Wu et al., 2021). For that, EnKF method assumes the model and observation errors had random white noise, with zero means (Ines et al., 2013).

However, due to the difficulties in model mechanists and the PBM parameterization, the application of PBM frequently includes BIAS (Kang and Özdoğan, 2019), which neglects the assumptions of EnkF that can lead to systematic errors. One example is the error inconsistency between LAI and yield, which happens when PBM simulates LAI values higher than the assimilated, and the open-loop (OP) simulated yield has negative BIAS. In these cases, after DA, it is likely that simulations would result in even higher yield underestimation due to the further LAI reduction (Kang and Özdoğan, 2019). That was partly demonstrated by Nearing et al. (2012), who showed that EnKF did not improve wheat yield estimation because of the low correlation between assimilated LAI and crop yield. The absence of prior PBM calibration can also result in error inconsistency, because the use of a set of cultivar-specific parameters cannot describe specific traits of other cultivars, such as higher potential yield and canopy properties (Fattori Junior et al., 2022; Kang and Özdoğan, 2019; Huang et al., 2021). Moreover, model errors related to crop phenology generally resulted in lower accuracy for yield estimation with EnKF, because PBM with more general structures usually account for phenology phase to calculate crop development (e.g. biomass partitioning). In this regard, considering the LAI assimilation, the error in phenology might lead to inconsistency in the magnitude and observed LAI (Curnel et al., 2011; Kang and Özdoğan, 2019).

In contrast, several studies showed improvements in yield estimation after DA. For instance, Yu et al. (2020) used a variant of the EnKF, the Ensemble Smoother, to assimilate sugarcane plant height with an adapted version of WO, finding an expressive improvement for both LAI and yield estimations. Moreover, Fattori Junior et al. (2022) used the EnKF to assimilate LAI with DSSAT/SAMUCA (DS), finding a reasonable improvement in yield estimation compared to the OP simulation. Part of the improvement was due to a detailed description of plant physiology and soil water process, which can benefit the model simulation with the EnKF (Ines et al., 2013). Yet, such benefit was also related to the accuracy of the model without DA, demonstrating that models with high accuracy before DA usually improves the performance with EnKF (Fattori Junior et al., 2022; Kang and Özdoğan, 2019). Moreover, due to a more detailed description of the plant development, it is possible to update variables related to the assimilated variable (e.g. LAI, leaf weight and specific leaf area), and this might reduce the sensitivity of EnKF to model BIAS (Ines et al., 2013; Kang and Özdoğan, 2019).

In this regard, the study of Silvestro et al. (2017) compared two different PBMs [Aquacrop and Simple Algorithm for Yield (SAFY)] by assimilating LAI and canopy cover for wheat yield estimation at a district scale. They found the Aquacrop showing lower accuracy after DA due to a more detailed description of the crop physiology related to water stress. This, in turn, increased the difficulty of the model calibration, due to a large number of parameters, compared to SAFY. However, for each PBM, a different DA method was used, thus the results were not exclusively influenced by the different PBM structures. Therefore, it is still not clear how a detailed description of the crop development could improve PBM simulation with DA methods and reduce the error inconsistency with EnKF.

To our knowledge, this is the first study to investigate how the PBM structure affected the performance of EnKF. We then aimed to assimilate sugarcane LAI data in two types of PBMs regarding their structures for estimating stalk fresh yield (SFY) and to analyze how the processes related to LAI calculation and the prior calibration affected the DA methods performance. To fill this knowledge gap, we calibrated both models with a robust experimental database collected across several producing regions of Brazil and developed a framework to assimilate LAI retrieved from Landsat 7 ETM+ and 8/OLI into DS and WO models. We then used the EnKF and tested this framework using a large on-farm sugarmill dataset collected in the most important sugarcane producing region of Brazil. Moreover, we compared the PBMs in terms of the structure involved in the LAI calculation to understand the impact on DA performance.

3.2. Material and Methods

3.2.1. Field experiments data for calibration

In total, we used a dataset of 13 experiments conducted in a diversity of environments and used the cultivar RB867515, for calibrating (experiments 1-7) and evaluating (experiments 8-13) the PBMs (Tab. 5). All of the experiments received adequate N, P, and K fertilization and regular weed and pest control and were planted using healthy cuttings with 13-15 buds m⁻². Row spacing varied from 1.4 m to 1.5 m. Experiments 1-7 had tiller population (TILL), stalk height (SH), LAI, SFY, stalk dry mass (SDM), and sucrose content on a fresh cane basis (POL) obtained by regular sampling. Experiments 8-13, had at last two of these variables sampling during the crop season. A full description of these experiments can be found in (Marin et al., 2015; Vianna et al., 2020).

Soil characteristics and management practices such as planting and harvesting dates, row spacing, mulch cover, and irrigation applications (mm d^{-1}) on each site were prescribed to the model as input information. Also, for

experiments 2, 4, and 6 (Tab. 5) a total of 12 Mg ha⁻¹ of green cane straw was considered for simulations. All other experiments were conducted under bare soil conditions. Other details of the experimental data can be found in Table 5.

Experime nt Number	Location	Planting date	Harvest date	Planting Type	Weather †	Soil‡	Water treatmen t	Referenc e
1		12/06/2012	10/15/2013	Plant			Irrigated	
2		10/15/2013	07/15/2014	1 st Ratoon*			Irrigated	
3	Piracicaba/S	10/15/2013	07/15/2014	1 st Ratoon	21.6 °C,	Typic	Irrigated	Vianna
4	P 22°41' S,	07/15/2014	06/08/2015	2 nd Ratoon*	1230	Hapludo	Irrigated	et al.
5	560m	07/15/2014	06/08/2015	2 nd Ratoon	CWa	x Typic	Irrigated	(2020)
6		06/08/2015	06/08/2016	3 rd Ratoon*			Irrigated	
7		06/08/2015	06/08/2016	3 rd Ratoon			Irrigated	
8	Aparecida do Taboado/M S 20°05S, 51°18'W, 335 m	07/01/2006	09/08/2007	Plant	23.5 °C, 1560 mm, Aw	Typic Hapludo x Typic	Rainfed	
9	Colina/SP 20°25'S, 48°19'W, 590 m	02/10/2004	06/15/2005	Plant	22.8 °C, 1363 mm, Cwa	Typic Hapludo x Typic	Rainfed	
10	Olimpia/SP 20°26'S, 48°32'W, 500 m	02/10/2004	06/15/2005	Plant	23.3 °C, 1349 mm, Cwa	Typic Hapludo x Typic	Rainfed	Marin et al. (2015)
11	Coruripe/A L 10°07'S, 36°10'W, 16 m	08/16/2005	09/15/2006	Plant	21.6 °C, 1401 mm, As	Fragiudul t Typic	Rainfed	
12	União/PI, 4°41'S, 42°52°W, 68 m	03/29/2007	06/16/2008	Plant	27 °C, 1500 mm, Aw	Oxisol	Irrigated	
13	União/PI, 4°41'S, 42°52°W, 68 m	03/29/2007	06/16/2008	Plant	27 °C, 1500 mm, Aw	Oxisol	Rainfed	

Table 5. Description of experimental datasets used for model calibration and evaluation.

† Respectively: mean annual temperature, annual total rainfall, Koeppen Classification.

‡ U.S. Soil Taxonomy

* With mulch cover

3.2.2. Field experiments for data assimilation evaluation

To evaluate the DA with EnKF in our study area, data from 32 blocks were collected from a sugarcane mill database located in São Paulo state in Brazil, between the years 2012 to 2015. All the blocks were managed following the standard for the region and had only the same cultivar RB867515. Each block was mechanically planted and harvested as a unit and received uniform management and inputs (e.g., fertilizer, pesticides). Only blocks with sugarcane plant-cane were used in this study to reduce the influence of management practices on the sugarcane yield (Marin et al., 2019b, 2021). In sugarcane mill blocks the harvest and planting are usually performed in many adjacent areas at once, thus these blocks were grouped into the 7 fields (Tab. 6). Therefore, all blocks of a field had the same planting and harvest date and observed SFY, however, the soil classification was not the same. In this regard, the SFY was simulated for each block, and to estimate the SFY of a field the simulated SFY was weighed by the planted area of each block. The following data were collected from each block: localization, size, planting date, harvest date, soil classification, and final SFY. The weather data [maximum and minimum air temperature (°C), rainfall (mm), solar radiation (MJ m⁻²d⁻¹) and relative humidity (%)] were collected daily from a weather station installed close to the fields.

Table 6. Description of field datasets used for data assimilation evaluation.

Fields	Planting date	Harvest date	Soils descriptions	Number of blocks
1	07/15/2012	12/05/2013	PVe ar/md and LVPal ar/md-ar	5
2	08/15/2013	09/13/2014	LVPd ar/md-ar	4
3	08/15/2013	09/07/2014	LVPal md-ar/md and LVPd ar/md-ar	6
4	08/15/2013	09/13/2014	LVPd ar/md-ar	2
5	08/15/2013	09/18/2014	LVPe ar/md-ar and Aqd	5
6	09/15/2014	07/24/2015	PVe ar/md	1
7	09/15/2014	12/05/2015	LVPal md/md-arg and PVal ar/md-ar	9

PVe ar/md – Argissolo Vermelho Eutrófico textura muito argilosa; LVPal ar/md-ar - Latossolo Vermelho Argissóilico Álico textura arenosa/média-arenosa; LVPd ar/md-ar - Latossolo Vermelho Argissóilico Álico textura arenosa/média-arenosa; PVal ar/md-ar - Argissolo Vermelho Álico textura arenosa/média-arenosa; Aqd - Areias Quartzosas distróficas

3.2.3. Satellite image pre-processing

For this study, the LAI retrieved from Landsat 7 ETM+ and 8/OLI were used as observed data to assimilate into WO and DS models. Both images had a spatial resolution of 30m and a revisit frequency of 16-day. The surface reflectance data were obtained by the Google Earth Engine (GEE) (Gorelick et al., 2017). where the images were atmospheric corrected. We used cloud/shadow masks available in eemont, a python package developed by Montero (2021).

The LAI was obtained based on the normalized difference vegetation index (NDVI), calculated with the surface reflectance of Landsat 7 ETM+ and 8/OLI, following the relationship proposed by Xavier and Vettoriazzi (2004). The relationship between LAI and NDVI was based on different surface vegetations, but 68% of the data were from sugarcane vegetated surfaces. Thus, for different months of the year, a different equation was obtained

(January, March, August, November), being all significant at 1% level (p < 0.01), R² ranging between 0.54 and 0.74, and the standard error between 0.38 and 0.67. Therefore, for each field, during the growing season (plant to harvest), an NDVI time series was obtained. Also, the NDVI time series were filtered by the Saviky-Golay method, to reduce the noise caused by the sensor error and atmospheric perturbation (Kang and Özdoğan, 2019; Zhao et al., 2013), after that, the LAI time series were calculated for each block. The LAI time series was used to represent the observation for the DA with EnKF. Based on the results of Xavier and Vettoriazzi (2004) and Abebe et al., (2022a) for the LAI retrieved from remote sensing we considered an error of 30%, for DA with EnKF.

3.2.4. Brief description of DSSAT/SAMUCA

The SAMUCA model is a PBM firstly developed by Marin and Jones (2014), which is capable to simulate the growth and development of sugarcane crop, implementing an algorithm to describe processes related to phenology, canopy development, tillering, biomass accumulation, root growth, and water stress (Marin et al., 2017). Vianna et al. (2020) improved the SAMUCA model by including recent scientific findings on sugarcane growth at phytomer level, canopy assimilation, and tillering. In this new version, the model was adapted to operate the onedimensional "tipping bucket" soil water balance and to incorporate the soil temperature to account for the trash blanket effect on sugarcane growth and water use. This presented a superior performance compared with the previous version and was comparable to other widely used PBMs for sugarcane. For this study, we used the SAMUCA model incorporated into the Decision Support System for Agrotechnology Transfer (DSSAT) platform version 4.8 (Jones et al., 2003; Hoogenboom et al., 2019; Vianna et al., 2020), namely DSSAT/SAMUCA (DS).

3.2.5. Brief description of WOFOST

The WOFOST (WO) model was developed by Wageningen University, the Netherlands, and it was used in this study due to its extensive application and evaluation (Abebe et al., 2022a). The WO model is a mechanistic PBM, which simulates crop growth as a function of solar radiation, temperature and crop properties (Wang et al., 2013). The basis of the simulation is the physiological and ecological progress of crops which includes light interception, CO₂ assimilation, respiration, transpiration, phenological development, dry matter accumulation, and portioning to various organs (Ma et al., 2013). Recently, the WO model was incorporated within the Python Crop Simulation Environment (PCSE) (de Wit et al., 2019). In this study, we implemented DA with the WO model using PCSE, parametrized for sugarcane. The WO is not a simplistic crop model such as SAFY, but rather a generalist PBM that can be adapted for different crops. Thus, WO version used here was not able to simulate specific traits of sugarcane, such as tiller population, plant height, sucrose content, and stem moisture, which are important to define the sugarcane yield. We used the parameters collected and calibrated by Scarpare (2011), for a few Brazilian sugarcane cultivars, as standard for our study.

3.2.6. Calibration process

The models were calibrated for cultivar RB867515 using the experiments (1-7) present in Tab. 5. The method used for calibration was the Generalized Likelihood Uncertainty Estimation (GLUE). For that we used the calibration process described by Marin et al. (2011), Li et al. (2018) and Pereira et al. (2021), following the steps: (i) Develop prior parameter distributions, in this case, we assume a uniform distribution from a predefined range of variation of genotype parameters, as recommended by Marin et al. (2017), selecting the most sensitive parameters based on Pereira et al., (2021) for DS and Scarpare, (2011) for WO (Appendix B - Table 1); (ii) Generate a random set of parameters values from prior parameter distributions based on the Monte Carlo method, for this study we used a sample of 3000 set of parameters; (iii) Run the model with different parameters sets; (iv) The calculus of each likelihood values for each observation (O) was used along with the corresponding simulated outputs to compute the likelihood values, $L(\theta_i|O)$, for each of the N generated parameter vector θ_i following Eq. 13:

$$L(\theta_i|O) = \prod_{J=1}^{M} \frac{1}{\sqrt{2\pi\sigma_o^2}} \exp\left(-\frac{(\theta_2 - f(\theta_i))^2}{2\sigma_0^2}\right), (i = 1, 2, 3, \dots N)$$
(13)

$$L_{comb}(\theta_i) = \prod_{k=1}^{K} L_k(\theta_i | O_k)$$
(14)

where *M* is the number of observations replicates; $f(\theta_i)$ is the model output referring to θ_i ; σ_o^2 is the variance model errors; *K* is the number of observations type; $L_{comb}(\theta_i)$ is the combined likelihood value of ith parameter set θ_i . Then, the probability $p(\theta_i)$ of each parameter set was computed with the following Eq. 15:

$$p(\theta_i) = \frac{L(\theta_i|Y)}{\sum_{i=1}^N L(\theta_i|Y)}$$
(15)

(v) Construct posterior distribution and statistics. The pairs of parameter set and probabilities, (θ_i, p_i) , i = 1, ..., N, were used to construct empirical posterior distributions and to compute the means and variance of selected parameters using the following equations:

$$\hat{\mu}(\theta) = \sum_{i=1}^{N} p(\theta_i) \ \theta_i \tag{16}$$

$$\hat{\sigma}(\theta) = \sum_{i=1}^{N} p(\theta_i) \left(\theta_i - \hat{\mu}\right)^2 \tag{17}$$

where $\hat{\mu}(\theta)$ and $\hat{\sigma}(\theta)$ are the mean and the variance of the posterior distribution, thus, $\hat{\mu}(\theta)$ is the optimum set of parameters and was considered the calibrated genotype set of parameters and was used in the evaluation step.

The measured data collected between the experiments in Table 5 were SDM, SFY, LAI, SH, TILL, and POL. For DS all variables were used for performing the GLUE, for the WO model, we only used the SDM, SFY and LAI to calibrate the model, because the WO cannot simulate the other variables. The parameters results in the calibration step were present in Appendix B - Table 1.

3.2.7. Description of data assimilation procedure

The sequential DA method, EnKF, was used in this study to assimilate LAI derived from Landsat 7 ETM+ and 8/OLI. To evaluate the improvements of the EnKF in the model's simulations, the results were compared to the simulation without DA, called open-loop (OP). The EnKF algorithm is described below. To implement the EnKF with DS, the model was adapted to read an input file with a new estimated vector of state variables at any time. Yet, when new LAI values were assimilated by DS, the leaf area and dry weight were also updated at phytomer and field level to ensure the consistency of canopy representation. The version of WOFSOT

used in this study inserted in PCSE, is more flexible to be adapted with EnKF. As follows, the simulations can be paused at any time and update the new state variable. For both models, a Python script was developed to read a control file, with each field description, create the input file for model simulation, run the model and pause any time to perform the EnKF and estimate the new LAI values, reinitialize the model and read the outputs. The WO model cannot simulate the SFY, thus we used the relation between SDM and SFY in the experiments in Table 5 to transform the simulated SDM in SFY during the crop cycle (Fig. 7).



Figure 7. The relationship between sugarcane stalk fresh yield and corresponding stalk dry mass of experiments in Table 5.

3.2.8. Ensemble Kalman filter method

The EnKF employs an analytic solution based on two related sources of information, in this case: PBMs outputs and field observations. These are synthesized to provide a better estimation, with lower variance. For that, the EnKF assumes that the observed data can be related to the state variable x_t (LAI in the case of this study) at time t as shown in Eq. 18:

 $y = Hx_t + \epsilon$

(18)

where y is the observations vector; H is the observation operator that relates to y; ε is a Gaussian random error vector with a mean of zero and observation error covariance R. Also, the forecast of x_t at t = k is Gaussian with mean $x_{t=k}^{f}$ and error covariance $P_{t=k}^{f}$. Under these assumptions, the estimated state and error covariance (P) are updated as:

$$x_{t=k}^{a} = x_{t-k}^{f} + K(y_{t=k} - Hx_{t=k}^{f})$$
⁽¹⁹⁾

$$P_{t=k}^{a} = (I - KH)P_{t=k}^{j}$$
(20)

where t is the time index; k is the time of the observed data; f represents the prior state (called forecast) and a is the posterior state (called analysis); I is the identity matrix and K represents the Kalman gain calculated by Eq. 21:

$$K = P_{t=k}^{f} H^{T} \left(H P_{t=k}^{f} H^{T} + R_{t=k} \right)^{-1}$$
(21)

The EnKF forecast and analysis error covariance P^{f} come directly from an ensemble of the model simulations:

$$P^{f}H^{T} = (N_{e} - 1)^{-1} \sum_{n=1}^{N_{e}} (x_{n}^{f} - \bar{x}^{f}) (Hx_{n}^{f} - H\bar{x}^{f})^{T}$$
(22)

where Ne is the number of ensemble members, n is a running index for an ensemble member, and \bar{x}^{f} are the ensemble mean calculated as:

$$\bar{x}^{f} = N_{e}^{-1} \sum_{n=1}^{N_{e}} x_{n}^{f}$$
(23)

In our study, we only used the LAI retrieved remote sensing as a state variable for DA methods. Thus, H can be taken as an identity matrix (H= 1), with that we can rewrite Eq. 19, 21, and 22 as Eq. 24, 25 and 26.

$$x_{t=k}^{a} = x_{t=k}^{f} + K(y_{t=k} - x_{t=k}^{f})$$
(24)

$$K = P_{t=k}^{f} \left(P_{t=k}^{f} + R_{t=k} \right)^{-1}$$
(25)

$$P^{f} = (N_{e} - 1)^{-1} \sum_{n=1}^{N_{e}} (x_{n}^{f} - \bar{x}^{f}) (x_{n}^{f} - \bar{x}^{f})^{T}$$
(26)

In EnKF, the observed data were perturbed with the Monte Carlo approach to generate an ensemble, based on the data uncertainty represented by the variance. Because, the observations needed to be treated as random variables and it is commonly assumed that observation errors have a Gaussian distribution (Zhuo et al., 2019). The PBM uncertainties are accounted for by the model ensemble. There are two methods to generate the ensemble members (Zhuo et al., 2019): the first method adds a Gaussian perturbation to the PBM state variables output. The second, add a Gaussian perturbation to the model input parameters. In this study, we used the second method to generate the ensemble members. Thus, to select the most sensitive parameters to LAI, for the DS we used the parameters selected by Fattori Junior et al., (2022), which were MAXGL (maximum number of green leaves a tiller can hold), MLA (maximum leaf area), PLASTOCHRON (thermal time required for the appearance of one new phytomer), INIT_LF_AREA (initial leaf area of first appeared leaf), and MID_TT_LF_GRO (thermal time where leaves can achieve half of its maximum biomass). For the WO we used the parameters that show higher influence in LAI, as reported by Hu et al., (2019) and Scarpare (2011). The parameters selected were TSUMEA (temperature from emergence to anthesis), RGRLAI (maximum relative increase in LAI), TBASE (lower threshold temperature for aging of leaves), EFF (light-use efficiency for real leaf) and CVL (efficiency of conversion into leaves). These parameters were then perturbed to generate an ensemble (40 members), with a gaussian distribution and an uncertainty level of 10% before the simulation started, as recommended by Ines et al. (2013) and Curnel et al. (2011) to optimize the time of the simulation and model accuracy.

After generating the set of parameters, DS runs until the first observed LAI is available. At this point, we calculated in sequence K and the vector $x_{t=k}^{a}$ (Eq. 24 and 25), that was considered the optimal estimation of LAI. This step also included small inflation of 1.5 for LAI in ensemble members, in the case of their variability becoming too low (Ines et al., 2013). This step ensured that the observations were not systematically rejected during assimilation. After that, the estimated LAI is stored in an input file for the next simulations, and runs were reinitialized until the next observations became available.

3.2.9. Data analysis

The performance of the calibration step was evaluated using the following statistical indices: root mean squared error (RMSE), determination index (R^2), Nash-Sutcliff model efficiency (EF), model bias (BIAS) and Willmot accuracy index (d) (Willmot et al., 2012). Secondly, for the evaluation step with sugarmill blocks, we only compared the SFY at the end of the cycle, simulated and observed. Thus, we used the RMSE, R^2 and BIAS to compare all the field simulations and observations, with the two models and two methods, OP and EnKF.

3.3. Results

3.3.1. Calibration results

Considering the experiments used for calibration (Tab. 7), the DS overperformed WO, simulating SDM, SFY, and LAI, with an EF = 0.907, 0.878, and 0.458 for DS and 0.622, 0.610 and 0.417 for WO, respectively (Tab. 7). The accuracy of DS was higher than WO for SDM, SFY, and LAI; Regarding the LAI, WO had RMSE = 0.981 m² m⁻² and DS RMSE = 0.946 m² m⁻². For SDM and SFY, WO had RMSE = 6.442 and 24.845 Mg ha⁻¹ and for DS RMSE = 3.198 and 16.964 Mg ha⁻¹ (Tab. 7). Furthermore, WO had good precision and accuracy for simulating SDM and SFY, with R² > 0.87 and d > 0.90; LAI presented a lower precision and accuracy with R² = 0.535 and d = 0.850 (Tab. 7). DS had similar performance results, with SDM and SFY showing R² > 0.92 and d > 0.96, as well as for LAI, with R² = 0.645 and d = 0.867 (Tab. 7). The DS underestimated SDM, SFY, and LAI, with BIAS = -1.313 Mg ha⁻¹, -9.837 Mg ha⁻¹, and -0.456 m² m⁻², respectively, while WO had positive BIAS for SDM and SFY (5.358 Mg ha⁻¹, 24.845 Mg ha⁻¹), and negative for LAI (-0.157 m² m⁻²) (Tab. 7).

The sucrose content in the stalk (POL) is a variable only simulated by DS, for which the model had an acceptable precision ($R^2 = 0.767$; d = 0.920; EF = 0.638) and accuracy ($RMSE = 1.262 \ \%_{[fresh]}$) (Tab. 7). The SH had a higher precision compared to LAI and TILL ($R^2 = 0.925$; d = 0.853; EF = 0.618) and had RMSE = 0.530 m (Tab. 7). The TILL had lower performance compared to other variables, except for LAI ($R^2 = 0.667$; d = 0.829; EF = 0.545; $RMSE = 3.392 \ \# \text{ m}^{-2}$) (Tab. 7). The SH and TILL presented a negative BIAS (-0.331 m; -1.551 \ \# m^{-2}) and POL was the only variable in the calibration step that had a positive BIAS (0.544 $\%_{[fresh]}$) (Tab. 7).

The WO presented a lower accuracy for the experiments used for evaluation with RMSE = 4.399 Mg ha⁻¹, 24.416 Mg ha⁻¹, 1.243 m² m⁻², respectively for SDM, SFY and LAI when compared to DS (RMSE = 4.304, 22.284 Mg ha⁻¹, 0.607 m² m⁻²). Different from the results on calibration step, WO had a negative BIAS for SDM (BIAS = -0.034 Mg ha⁻¹), and DS positive for SDM and SFY (BIAS = 0.657, 2.675 Mg ha⁻¹). Yet, both models had negative BIAS for LAI, being -0.101 m² m⁻² and -0.818 m² m⁻² respectively for DS and WO (Tab. 7). The WO showed lower

performance for simulating LAI (EF = -0.749, $R^2 = 0.252$, d = 0.641), but it was satisfactory for SDM and SFY (EF > 0.81, $R^2 > 0.87$, d > 0.95) (Tab. 3). The DS had good performance for simulating SDM and SFY (EF > 0.84, $R^2 > 0.91$, d > 0.96), although lower for LAI (EF = 0.458, $R^2 = 0.645$, d = 0.867) (Tab. 7).

Model	Туре	Variables	BIAS	RMSE	EF	\mathbb{R}^2	d	Х	Y
		SDM	-1.313	3.198	0.907	0.926	0.974	12.587	11.274
		SFY	-9.837	16.964	0.878	0.932	0.964	69.267	59.43
	Calibration	LAI	-0.456	0.946	0.458	0.645	0.867	3.214	2.759
	Cambration	POL	0.544	1.262	0.683	0.761	0.92	8.942	9.485
		SH	-0.331	0.53	0.618	0.925	0.853	1.204	0.873
		TILL	-1.551	3.392	0.545	0.667	0.829	14.544	12.993
DSSA1/SAMUCA		SDM	0.657	4.304	0.903	0.951	0.98	19.443	20.1
	Evaluate	SFY	2.675	22.284	0.843	0.915	0.968	84.907	87.582
		LAI	-0.101	0.607	0.583	0.635	0.887	2.595	2.494
		POL	-0.533	1.144	0.547	0.733	0.901	13.232	12.699
		SH	-1.608	3.585	0.517	0.666	0.812	14.03	12.422
		TILL	0.43	0.495	0.673	0.924	0.925	0.963	1.393
		SDM	5.358	6.442	0.622	0.887	0.911	12.587	17.944
	Calibration	SFY	24.845	30.349	0.61	0.872	0.903	69.267	94.112
WOFOST		LAI	-0.157	0.981	0.417	0.535	0.85	3.214	3.058
		SDM	-0.034	4.399	0.899	0.899	0.972	19.443	19.409
	Evaluate	SFY	12.736	24.416	0.811	0.871	0.953	84.907	97.642
		LAI	-0.818	1.243	-0.749	0.252	0.641	2.595	1.777

 Table 7. Statistical indexes of performance of the DSSAT/SAMUCA and WOFOST models in simulating sugarcane crop components across experiments.

RMSE: Root mean squared error; EF: Modeling efficiency; R²: Determination index; d: accuracy index of Wilmot; X: Mean observations; Y: Mean simulations; Bias = Y – X; SDM: Stalk dry mass (Mg ha⁻¹); SFY: Stalk fresh yield (Mg ha⁻¹); LAI: Leaf area index (m² m⁻²); POL: sucrose content on a fresh cane basis (% [fresh]); SH: Stalk height (m); TILL: Tiller population (# m⁻²).

3.3.2. Data Assimilation Evaluation

The DS OP simulations had a higher accuracy (RMSE = 31.678 Mg ha⁻¹) compared to WO (RMSE = 39.593 Mg ha⁻¹) and higher precision (R² = 0.509 and 0.458 for DS and WO respectively) (Tab. 8). Both models had a positive BIAS for OP simulations, 25.406 Mg ha⁻¹ for DS and 31.282 Mg ha⁻¹ for WO (Tab. 8).

When DA based on EnKF using LAI was performed, the DS had RMSE = 17.508 Mg ha⁻¹ which represented a decrease of -44.73 % compared to the OP method (Tab. 8). The WO had lower accuracy compared to DS (RMSE = 27.880 Mg ha⁻¹), but it represented a decrease of -29.58% compared with OP methods (Tab. 8). The precision also improved after DA, as DS showed $R^2 = 0.622$, an increase of 22.15% compared to OP. The WO had an even higher increase in precision (36.50%) with $R^2 = 0.625$, which was higher than DS (Tab. 8). The DS with EnKF had BIAS = -11.469 Mg ha⁻¹, while WO showed a positive BIAS = 22.944 Mg ha⁻¹ after DA (Tab. 8).

Model	Method	BIAS	RMSE	R ²	Х	Y	
DSSAT/SAMUCA	OP	25.406	31.678	0.509*	94.701	120.107	
	EnKF	-11.469	17.508	0.622*	94.701	83.232	
WOFOST	OP	31.282	39.593	0.458*	94.701	125.983	
	EnKF	22.944	27.880	0.625*	94.701	117.645	
DMCT D	1 D2	D	1 37 3.6	1 . 37	NC 1.1	$\mathbf{D}\mathbf{I}\mathbf{A}\mathbf{C} = \mathbf{V}$	

Table 8. Statistical indexes of performance of DSSAT/SAMUCA and WOFOST without data assimilation, open-loop method (OP), and with data assimilation using the Ensemble Kalman Filter (EnKF).

RMSE: Root mean squared error; R²: Determination index; X: Mean observations; Y: Mean simulations; BIAS = Y - X

After DA, DS simulated lower values of SFY, with a mean = 83.232 Mg ha⁻¹, compared to SFY simulated by WO, for which mean = 117.645 Mg ha⁻¹ (Tab. 8). Further, after DA, DS had some fields with SFY simulated lower than the values observed, different from WO, that had a lower accuracy reduction and any SFY simulated lower than observed ones (Fig. 8). In this regard, the WO improved the simulations of 5 fields out of 8, while DS improved 4 fields out of 8. For WO, however, the fields without improvements in accuracy were those that already had high accuracy before DA. (Fig. 8).

During the sugarcane crop development, the DS showed higher LAI values for OP simulations compared to WOFOST and Landsat LAI, in general for all blocks, as shown for two selected blocks in Figure 9. Higher differences between DS and Landsat LAI were observed in the early phases of crop development when the LAI increases following a linear relationship (Fig. 9 a and i). The LAI simulated by DS had a peak period in the early stage of development and, after that, LAI decreased and stabilized at a lower level. This was different from what Landsat LAI showed, for which the peak period occurred at later stages of development (Fig. 9 a and i). For WO, LAI profiles also had a peak period in an earlier stage compared to Landsat LAI, but the values were closer to the observed ones along the crop season (Fig. 9 c and k).

Therefore, two types of error inconsistency occurred in the simulations, for each PBM. For DS, first, the blocks with lower SFY, simulated by OP, (Fig. 9 m) had LAI values higher than the Landsat LAI (Fig. 9 j). However, the OP simulations resulted in SFY lower than the observed (Fig. 9 m). Thus, after DA, the negative BIAS for SFY was even increased (Fig. 9 n). Second, a mismatch between the LAI's peak period was observed for all subfields (Fig. 9 a and i), which mainly reduced the performance of LAI assimilated in fields with low SFY (Fig 9 j and n). For WO, the OP simulated LAI was closer to the Landsat LAI (Fig 9 c and k), however, the SFY simulations had positive BIAS mainly for blocks with high observed SFY (Fig 9 g). Thus, after DA, the high BIAS in SFY was slightly lowered (Fig. 9 h). Moreover, the WO simulations had a mismatch between OP LAI and Landsat LAI, with higher values in the early development phases and lower at the end of the crop cycle, mainly for sub-fields with higher SFY simulations (Fig. 9 c and k).

Nevertheless, after DA, for WO, it was observed a higher accuracy for the blocks with lower observed SFY, because the simulated LAI had better agreement with the Landsat LAI, and closer SFY simulated values with observed ones (Fig. 9 l and p). For DS, the simulations after DA had higher accuracy for fields with high SFY (Fig. 9 e and f).

The higher decrease in DS SFY estimation followed a higher decrease in LAI values, after DA, along crop cycle, compared to WO (Fig 9). In summary, the OP simulations of LAI presented a mean value $(1.347 \text{ m}^2 \text{ m}^{-2})$ close to the Landsat LAI (1.350 m² m⁻²), while DS showed mean LAI = 2.485 m² m⁻², which was 84% higher than mean

^{*}p < 0.01



Figure 8. Comparison of observed and simulated stalk fresh yield (SFY) (a, b) with open-loop simulations (OP) and (c, d) with Ensemble Kalman filter (EnKF), using (a, c) DSSAT/SAMUCA and (b, d) WOFOST. The dashed black line represents the 1:1 adjustment, whereas the red dashed line is the regression between observed SFY and simulated SFY.



Figure 9. Comparison between simulated (dark blue lines) and observed (green circles) leaf area index (LAI) and stalk fresh yield (SFY) for one of the blocks of field 1 (a-h) and 3 (i-p), with (a, c, e, g, i, k, m, and o) open-loop and Ensemble Kalman filter (b, d, f, h, j, l, n, and p) method, using the DSSAT/SAMUCA (a, b, e, f, I, j, m and h) and WOFOST (c, d, g, h, k, l, o and p). Greys lines are the ensemble simulations, blue lines are the ensemble mean.



3.4. Discussion

In this study, the models DS and WO were calibrated for cultivar RB867515 and had the SFY simulations, using the EnKF, compared with the same block database by assimilating LAI retrieved from Landsat 7 ETM+ and 8 OLI. In the calibration step, the DS had better performance for calibration and evaluation, despite the higher number of observed variables used for GLUE. Thus, it seems that simulating different plant variables resulted in a lower uncertainty, despite the higher complexity and interactions. For example, SDM is partly derived from the TILL and SH, and adding these relationships to the model and using this observed variable for calibration, we found better simulations for SDM and SFY. Furthermore, these results may also be due to a more detailed description of the soil-water balance and soil layers of DS, which is different from the WO that considers only one soil layer.

One important part of the lower performance of WO came from the ration experiments (2-7, Tab. 5), as the model overestimated the SDM, SFY, and LAI for them. This can be related to some parameters of WO, such as initial total crop dry weight (TDWI); degree-days from emergence to anthesis (TSUM1); and initial rooting depth (RDI), that were not retrieved from these experiments. Therefore, the parameters retrieved from Scarpare (2011) might not represent the conditions of these experiments, because they were derived from different environments and genotypes. Moreover, despite both PBMs having the same database for calibration and the same weather database, the uncertainty in PBM structure and the difference in soil and genotype parameters affected the performance of the simulations (Marin et al., 2015).

The calibration performance affected the sugarmill blocks simulations, with DS showing higher performance for OP simulations than WO (Tab. 7 and 8). However, OP simulations with DS had lower accuracy, compared to the results of Fattori Junior et al. (2022), which use the same PBM for simulating different experiments with EnKF and LAI retrieved from field observations. Moreover, comparing the results of WO with other studies, the OP simulations presented lower accuracy, with higher RMSE than studies with the same model (Abebe et al., 2022a; Hu et al., 2019). The lower accuracy of both PBMs can be a result of management adopted in sugarmill blocks that were more susceptible to reduction factors, which decreased the SFY and were not simulated by the PBMs (Dias and Sentelhas et al., 2017).

Nevertheless, the use of EnKF reduced the RMSE and increased R² for both models (Tab. 8), showing the potential of using LAI retrieved from Landsat 7 ETM+ and 8 OLI for reducing model uncertainty (Abebe et al., 2022a; Huang et al., 2015; Kang and Özdoğan, 2019). It showed, also, the potential for using DA methods to correct the model simulation when reduction factors were present (Hu et al., 2019). Yet, matters to highlight that in spite of both models having the same assimilated variable and used the same database for calibration, the results after DA were considerably different, which indicates that the PBM structure highly affected the performance of SFY estimation after DA (Silvestro et al., 2017).

In this regard, the light interception mechanics and the distribution over the canopy are similar between both PBMs. Both account for solar radiation being fractionated between direct and diffuse proportions (Vianna et al., 2020; de Wit et al., 2019), and the leaf area is a result of total living leaves and the specific leaf area. Therefore, the difference between models was the mechanism used to calculate the living leaf biomass and the values used for specific leaf area. For DS, the leaf biomass was calculated by phytomer and the total leaf biomass per area was a result of the number of phytomers with living leaves, leaf biomass of each phytomer, and the number of stalks (Vianna et al., 2020). Moreover, the senescence rate of leaves was related to the process of shading leaves, the maximum number of leaves, leaves age, and tiller age, in conditions without water stress. Further, the specific leaf area was considered a fixed parameter during the crop simulation. The WO, in turn, does not simulate the sugarcane number of tiller and phytomer, and the leaf biomass was calculated by leaves age (de Wit et al., 2019), which also affects the calculus of LAI, leaf area, and leaf senescence. The relation between leaf area and stalk mass was only related to partitioning factors, over the crop development stage. Different from DS, the specific leaf area changes during the crop development stages, following the input values.

Therefore, the DS had a more complex structure for LAI simulations, which was affected by the number of stalks and phytomers, which represent a better approximation of the sugarcane development (Lou et al., 2013; Zhou and Shoko, 2011). This mechanism might explain part of the better accuracy and precision of DS for calibration and evaluation steps (Tab. 7), because enough information (LAI and TIL) was provided for calibration (Vianna et al., 2020). This also may explain part of the better performance of DS in OP simulations, for the sugarmill fields simulations (Tab. 8).

The DA performance had a close relationship with the OP performance (Fattori Junior et al., 2022), and after DA, the DS had also higher accuracy compared to WO. However, the DS underestimated the SFY (Tab. 8), because the changes in LAI values after DA had a direct relationship with SFY simulation. In this regard, when OP simulations had LAI and yield (SFY) both with positive BIAS, but the BIAS in SFY was slightly lower, after DA the SFY significantly reduced, resulting in large errors (Kang and Özdoğan, 2019). However, when both LAI and SFY had high BIAS, after DA, the accuracy was notably improved. For WO, the resulted LAI from OP and the Landsat LAI had close values, thus when the model had high positive BIAS for SFY estimation, the SFY after DA was only slightly improved (Kang and Özdoğan, 2019); when however the BIAS for LAI was close to zero, the DA was not able to improve the SFY estimations.

Therefore, our results emphasize that the structure of DS and WO could not reduce the sensitivity of EnKF of both PBM and assimilated variable BIAS. This disagrees with the hypothesis that updating state variables related to LAI reduces the sensitivity of EnKF to model BIAS (Ines et al., 2013; Kang and Özdoğan, 2019). The target variable to be improved was the SFY, but the variable assimilated was LAI, and between LAI and SFY there are several complex relationships affected by many other factors (Nearing et al., 2012). Also, the LAI of sugarcane has a considerable variation among genotype and environment (Lou et al., 2013; Marin et al., 2011), which reflects in the relationship between LAI and other variables important to define the SFY, such as SH and TILL (Yu et al., 2020; Zhou and Shoko, 2011). Thus, assimilating only LAI and updating close related variables was not sustainable to inhibit the error inconsistency. So, the correlation between LAI and variables such as TILL and SH should be better explored and improved when the LAI is assimilated into the PBM. Further, using one or more assimilated variables would reduce the impact of biased variables (Yu et al., 2022, Pan et al., 2019), and might lower the EnKF sensitivity. Further studies should then explore this considering different model structures.

In terms of practicalities, the WO model was easily coupled with EnKF due to object-oriented construction in python. Thus, multiple simulations can be initialized and paused at any time during the simulation runs, to perform the DA. Different from the DS, the structure of DSSAT platforms did not allow for a pause in the simulations during the model run. Therefore, to perform the DA, the models need to be reinitialized at each observation, increasing the time of simulations, despite the FORTRAN language being faster. Ines et al. (2013) overcome such limitation using a modified version of CSM-Maize model, outside the DSSAT, thus allowing EnKF to control the simulation ensemble with independent crop model runs to improve the speed and applicability with EnKF.

Although the EnKF techniques employed in this study showed improved performance of SFY simulations for both models, we emphasized some limitations of our study. The LAI time series retrieved from Landsat 7 ETM+ and 8/OLI, used the relationship of Xavier and Vettoriazzi (2004), which in turn used data from surface vegetation cover that was partly retrieved from other crops and may not well represent the sugarcane LAI. In this regard, during the DA process, we assumed a relatively high LAI error for generating the ensembles and this may interfere with DA performance. Therefore, further studies should explore the differences between PBMs with LAI observations retrieved from more sophisticated models such as biophysics models (Pan et al., 2019) and the Gaussian process (Abebe et al., 2022b). Finally, in this study we used the standard version of EnKF, but other studies developed different variations of the EnKF to overcome some limitations of the method (Jamal and Linker, 2022; Wu et al., 2021), and so these can be tested in future researchers to compare the interference in more PBMs.

3.5. Conclusions

This study aimed to investigate how two different PBMs in terms of structure affected the performance of EnKF to assimilate LAI retrieved from Landsat 7 ETM+ and 8 OLI. Both PBM had satisfactory performance in the calibration and evaluation step, but the DS had better performance. This was reflected in the OP simulations of sugarmill blocks, with DS presenting a superior accuracy (RMSE) and precision (R²) than WO. The DS had a more detailed description of the relationship between LAI and the tiller population, in PBM structure, which benefited OP simulations. However, after DA both PBMs showed error inconsistency, despite the improvement in accuracy and precision. This was caused by the different simulated LAI values of each PBM, with DS having higher LAI values than the observed, and so, after DA, the SFY was underestimated. Differently, WO had LAI values closer to the assimilated, and this resulted in minor changes in the large positive BIAS of SFY after DA. Thus, the structure of both BPM did not lower the sensitivity of EnKF to PBM BIAS, despite the more robust structure of DS for LAI simulations. In this regard, this study emphasizes that the relationship between LAI and other related variables, during the DA, should be better understood and added to the PBM routine.

References

- Abebe, G., Tadesse, T., Gessesse, B., 2022a. Assimilation of leaf Area Index from multisource earth observation data into the WOFOST model for sugarcane yield estimation. Int. J. Remote Sens. 43, 698–720. https://doi.org/10.1080/01431161.2022.2027547
- Abebe, G., Tadesse, T., Gessesse, B., 2022b. Estimating Leaf Area Index and biomass of sugarcane based on Gaussian process regression using Landsat 8 and Sentinel 1A observations. Int. J. Image Data Fusion 00, 1–31. https://doi.org/10.1080/19479832.2022.2055157
- CONAB National Supply Company, 2022 Products and services Available at: https://www.conab.gov.br/infoagro/safras/cana (Accessed 07 February 2022)
- Curnel, Y., de Wit, A.J.W., Duveiller, G., Defourny, P., 2011. Potential performances of remotely sensed LAI assimilation in WOFOST model based on an OSS Experiment. Agric. For. Meteorol. 151, 1843–1855. https://doi.org/10.1016/j.agrformet.2011.08.002

- de Wit, A., Boogaard, H., Fumagalli, D., Janssen, S., Knapen, R., van Kraalingen, D., Supit, I., van der Wijngaart, R., van Diepen, K., 2019. 25 years of the WOFOST cropping systems model. Agric. Syst. 168, 154–167. https://doi.org/10.1016/j.agsy.2018.06.018
- Dias, H.B., Sentelhas, P.C., 2017. Evaluation of three sugarcane simulation models and their ensemble for yield estimation in commercially managed fields. F. Crop. Res. 213, 174–185. https://doi.org/10.1016/j.fcr.2017.07.022
- Evensen, G., 2003. The Ensemble Kalman Filter: Theoretical formulation and practical implementation. Ocean Dyn. 53, 343–367. https://doi.org/10.1007/s10236-003-0036-9
- Fattori Junior, I. M., Vianna, M.S., Marin, F. R., 2022. Assimilating leaf area index data into a sugarcane processbased crop model for improving yield estimation. Eur. J. Agron, 136, 126501. https://doi.org/10.1016/j.eja.2022.126501
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetaryscale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031
- Hoogenboom, G., Porter, C.H., Boote, K.J., Shelia, V., Wilkens, P.W., Singh, U., White, J. W., Asseng, S., Lizaso, J.I., Moreno, L.P., Pavan, W., Ogoshi, R., Hunt, L.A., Tsuji, G. Y., Jones, J.W., 2019. The DSSAT crop modeling ecosystem. In: Boote, K.J. (Ed.), Advances in Crop Modeling for a Sustainable Agriculture. Burleigh Dodds Science Publishing, Cambridge, United Kingdom, pp. 173–216.
- Hu, S., Shi, L., Huang, K., Zha, Y., Hu, X., Ye, H., Yang, Q., 2019. Improvement of sugarcane crop simulation by SWAP-WOFOST model via data assimilation. F. Crop. Res. 232, 49–61. https://doi.org/10.1016/j.fcr.2018.12.009
- Huang, J., Gómez-Dans, J.L., Huang, H., Ma, H., Wu, Q., Lewis, P.E., Liang, S., Chen, Z., Xue, J.H., Wu, Y., Zhao, F., Wang, J., Xie, X., 2019. Assimilation of remote sensing into crop growth models: Current status and perspectives. Agric. For. Meteorol. 276–277, 107609. https://doi.org/10.1016/j.agrformet.2019.06.008
- Huang, J., Tian, L., Liang, S., Ma, H., Becker-Reshef, I., Huang, Y., Su, W., Zhang, X., Zhu, D., Wu, W., 2015. Improving winter wheat yield estimation by assimilation of the leaf area index from Landsat TM and MODIS data into the WOFOST model. Agric. For. Meteorol. 204, 106–121. https://doi.org/10.1016/j.agrformet.2015.02.001
- Huang, X., Zhao, G., Zorn, C., Tao, F., Ni, S., Zhang, W., Tu, T., Höglind, M., 2021. Grass modelling in datalimited areas by incorporating MODIS data products. F. Crop. Res. 271, 108250. https://doi.org/10.1016/j.fcr.2021.108250
- Ines, A.V.M., Das, N.N., Hansen, J.W., Njoku, E.G., 2013. Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. Remote Sens. Environ. 138, 149–164. https://doi.org/10.1016/j.rse.2013.07.018
- Jamal, A., Linker, R., 2020. Genetic Operator-Based Particle Filter Combined with Markov Chain Monte Carlo for Data Assimilation in a Crop Growth Model. Agri., 10(12), 606. https://doi.org/10.3390/agriculture10120606
- Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., Keating, B.A., Munoz-Carpena, R., Porter, C.H., Rosenzweig, C., Wheeler, T.R., 2017. Brief history of agricultural systems modeling. Agric. Syst. 155, 240–254. https://doi.org/10.1016/j.agsy.2016.05.014

- Jones, J.W., Hoogenboom, G., Porter, C., Boote, K., Batchelor, W., Hunt, L., Wilkens, P., Singh, U., Gijsman, A., Ritchie, J., 2003. The DSSAT cropping system model. Eur. J. Agron. 18, 235–265. https://doi.org/10.1016/S1161-0301(02)00107-7
- Kang, Y., Özdoğan, M., 2019. Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. Remote Sens. Environ. 228, 144–163. https://doi.org/10.1016/j.rse.2019.04.005
- Li, Zhenhai, He, J., Xu, X., Jin, X., Huang, W., Clark, B., Yang, G., Li, Zhenhong., 2018. Estimating genetic parameters of DSSAT-CERES model with the GLUE method for winter wheat (Triticum aestivum L.) production. Comput. Electron. Agric. 154, 213–221. https://doi.org/10.1016/j.compag.2018.09.009
- Luo, J., Que, Y., Zhang, H., Xu, L., 2013. Seasonal variation of the canopy structure parameters and its correlation with yield-related traits in sugarcane. Sci. World J. 2013. https://doi.org/10.1155/2013/801486
- Ma, H., Huang, J., Zhu, D., Liu, J., Su, W., Zhang, C., Fan, J., 2013. Estimating regional winter wheat yield by assimilation of time series of HJ-1 CCD NDVI into WOFOST-ACRM model with Ensemble Kalman Filter. Math. Comput. Model. 58, 759–770. https://doi.org/10.1016/j.mcm.2012.12.028
- Manivasagam, V.S., Rozenstein, O., 2020. Practices for upscaling crop simulation models from field scale to large regions. Comput. Electron. Agric. 175. https://doi.org/10.1016/j.compag.2020.105554
- Marin, F. R., Jones, J. W., Royce, F., Suguitani, C., Donzeli, J. L., Filho, W. J. P., Nassif, D. S., 2011. Parameterization and evaluation of predictions of DSSAT/CANEGRO for Brazilian sugarcane. Agronomy Journal, 103(2), 304-315. https://doi.org/10.2134/agronj2010.0302
- Marin, F. R., Vianna, M. S., Nassif, D. S., 2019a. Challenges, Constraints, and Limitations of Cane Biofuels. In Sugarcane Biofuels (pp. 389-407). Springer, Cham. https://doi.org/10.1007/978-3-030-18597-8_17
- Marin, F., Jones, J.W., Boote, K.J., 2017. A stochastic method for crop models: Including uncertainty in a sugarcane model. Agron. J. 109, 483–495. https://doi.org/10.2134/agronj2016.02.0103
- Marin, F.R., Jones, J.W., 2014. Process-based simple model for simulating sugarcane growth and production. Sci. Agric. 71, 1–16. https://doi.org/https://doi.org/10.1590/S0103-90162014000100001
- Marin, F.R., Rattalino Edreira, J.I., Andrade, J., Grassini, P., 2019b. On-farm sugarcane yield and yield components as influenced by number of harvests. F. Crop. Res. 240, 134–142. https://doi.org/10.1016/j.fcr.2019.06.011
- Marin, F.R., Rattalino Edreira, J.I., Andrade, J.F., Grassini, P., 2021. Sugarcane Yield and Yield Components as Affected by Harvest Time. Sugar Tech. https://doi.org/10.1007/s12355-020-00945-5
- Marin, F.R., Thorburn, P.J., Nassif, D.S.P., Costa, L.G., 2015. Sugarcane model intercomparison: Structural differences and uncertainties under current and potential future climates. Environ. Model. Softw. 72, 372–386. https://doi.org/10.1016/j.envsoft.2015.02.019
- Montero, D., 2021. eemont: A Python package that extends Google Earth Engine. J. Open Source Softw. 6, 3168. https://doi.org/10.21105/joss.03168
- Morell, F.J., Yang, H.S., Cassman, K.G., Wart, J. Van, Elmore, R.W., Licht, M., Coulter, J.A., Ciampitti, I.A., Pittelkow, C.M., Brouder, S.M., Thomison, P., Lauer, J., Graham, C., Massey, R., Grassini, P., 2016. Can crop simulation models be used to predict local to regional maize yields and total production in the U.S. Corn Belt? F. Crop. Res. 192, 1–12. https://doi.org/10.1016/j.fcr.2016.04.004
- Nearing, G.S., Crow, W.T., Thorp, K.R., Moran, M.S., Reichle, R.H., Gupta, H. V., 2012. Assimilating remote sensing observations of leaf area index and soil moisture for wheat yield estimates: An observing system simulation experiment. Water Resour. Res. 48, 1–13. https://doi.org/10.1029/2011WR011420

- Pan, H., Chen, Z., de Allard, W., Ren, J., 2019. Joint assimilation of leaf area index and soil moisture from sentinel-1 and sentinel-2 data into the WOFOST model for winter wheat yield estimation. Sensors (Switzerland) 19. https://doi.org/10.3390/s19143161
- Pereira, R.A. de A., Vianna, M. dos S., Nassif, D.S.P., Carvalho, K. dos S., Marin, F.R., 2021. Global sensitivity and uncertainty analysis of a sugarcane model considering the trash blanket effect. Eur. J. Agron. 130, 126371. https://doi.org/10.1016/j.eja.2021.126371
- Scarpare, F. V., 2011. Simulação do crescimento da cana-de-açúcar pelo modelo agrohidrológico SWAP/WOFOST. Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz". Piracicaba, Brasil [In Portuguese].
- Silvestro, P.C., Pignatti, S., Pascucci, S., Yang, H., Li, Z., Yang, G., Huang, W., Casa, R., 2017. Estimating wheat yield in China at the field and district scale from the assimilation of satellite data into the Aquacrop and simple algorithm for yield (SAFY) models. Remote Sens. 9, 1–24. https://doi.org/10.3390/rs9050509
- Van Heemst, H. D. J., 1988. Plant data values required for simple crop growth simulation models: review and bibliography.
- Vianna, M. dos S., Nassif, D.S.P., dos Santos Carvalho, K., Marin, F.R., 2020. Modelling the trash blanket effect on sugarcane growth and water use. Comput. Electron. Agric. 172, 105361. https://doi.org/10.1016/j.compag.2020.105361
- Wang, E., Brown, H.E., Rebetzke, G.J., Zhao, Z., Zheng, B., Chapman, S.C., 2019. Improving process-based crop models to better capture genotype×environment×management interactions. J. Exp. Bot. 70, 2389–2401. https://doi.org/10.1093/jxb/erz092
- Wang, J., Li, X., Lu, L., Fang, F., 2013. Parameter sensitivity analysis of crop growth models based on the extended Fourier Amplitude Sensitivity Test method. Environ. Model. Softw. 48, 171–182. https://doi.org/10.1016/j.envsoft.2013.06.007
- Willmott, C. J., Robeson, S. M., Matsuura, K., 2012., A refined index of model performance. International Journal of climatology, 32(13), 2088-2094. https://doi.org/10.1002/joc.2419
- Wu, S., Yang, P., Ren, J., Chen, Z., Li, H., 2021. Regional winter wheat yield estimation based on the WOFOST model and a novel VW-4DEnSRF assimilation algorithm. Remote Sens. Environ. 255. https://doi.org/10.1016/j.rse.2020.112276
- Xavier, A.C., Vettorazzi, C.A., 2004. Mapping leaf area index through spectral vegetation indices in a subtropical watershed. Int. J. Remote Sens. 25, 1661–1672. https://doi.org/10.1080/01431160310001620803
- Yu, D., Zha, Y., Shi, L., Jin, X., Hu, S., Yang, Q., Huang, K., Zeng, W., 2020. Improvement of sugarcane yield estimation by assimilating UAV-derived plant height observations. Eur. J. Agron. 121, 126159. https://doi.org/10.1016/j.eja.2020.126159
- Yu, D., Zha, Y., Shi, L., Ye, H., Zhang, Y., 2022. Improving sugarcane growth simulations by integrating multisource observations into a crop model. Eur. J. Agron. 132, 126410. https://doi.org/10.1016/j.eja.2021.126410
- Zhao, Y., Chen, S., Shen, S., 2013. Assimilating remote sensing information with crop model using Ensemble Kalman Filter for improving LAI monitoring and yield estimation. Ecol. Modell. 270, 30–42. https://doi.org/10.1016/j.ecolmodel.2013.08.016
- Zhou, M. M., Shoko, M. D., 2011. Seasonal and varietal effects on tiller population development of sugarcane (Saccharum officinarum L.). South African Journal of Plant and Soil, 28(1), 11-16. https://hdl.handle.net/10520/EJC88005

Zhuo, W., Huang, J., Li, L., Zhang, X., Ma, H., Gao, X., Huang, H., Xu, B., Xiao, X., 2019. Assimilating Soil Moisture Retrieved from Sentinel-1 and Sentinel-2 Data into WOFOST Model to Improve Winter Wheat Yield Estimation. Remote Sens. 11, 1618. https://doi.org/10.3390/rs11131618

APPENDICES

APPENDIX A. Understanding the impact of data assimilation in process-based crop models for sugarcane yield estimation

Table 1. List of DSSAT/SAMUCA crop parameters for the RB867515 cultivar (Vianna et al., 2020), with their respective definitions, units and values used in the simulations. Values market with "*" were calibrated in this study. The genotype file (CUL: Cultivar, ECO: Ecotype, SPE: Species) for each parameter is indicated in brackets followed its description.

Parameter	Description	Values	Units
MAXGL	Maximum number of green leaves a tiller can hold (CUL)	12.00	#/tiller
N_LF_STK _EM	Number of leaves appeared before stalks emerges at soil surface(CUL)	4.00	#/tiller
N_LF_IT_ FORM	Number of leaves appeared before internode formation(CUL)	3.00	#/tiller
MAXDGL	Maximum number of developed (dewlap) green leaf a tiller can hold (CUL)	6.00	#/tiller
AMAX	Assimilation rate at light saturation point (CUL)	44.90	$\mu mol/m^2/s$
DEE		0.069	$\mu mol_{[CO2]}/m^2/s$
EFF	Carboxylation efficiency (CUL)		$(\mu mol_{PPFD}/m^2/s)^{-1}$
CHUSTK	Thermal time for stalk emergence (CUL)	650.00	°Cdays
CHUPEAK	Thermal time for tillering peak (CUL)	1400.0 0	°Cdays
CHUDEC	Thermal time for tillering senescence (CUL)	1600.0 0	°Cdays
CHUMAT	Thermal time for population stabilization (CUL)	3200.0 0	°Cdays
POPMAT	Tiller population at tillering stabilization (CUL)	9.50	tillers/m ²
POPPEAK	Number of tillers at peak of population (CUL)	22.00	tillers/m ²
TILLOCH RON	Thermal time required for the emergence of one new tiller (CUL)	69.00	°Cdays/tiller
PHYLLOC HRON	Thermal time required for the appearance of one new leaf (CUL)	132.00	°Cdays/leaf
SLA	Specific Leaf Area (CUL)	120.00	cm ² /g
MLA	Maximum leaf area (CUL)	600.00	cm^2
PLASTOC HRON	Thermal time required for the appearance of one new phytomere (CUL)	132.00	°Cdays/phytomer
INIT_LF_ AREA	Initial leaf area of first appeared leaf (CUL)	15.00	cm ²
MAX_INI_ LA	Initial leaf area of leaves appeared after top parts formation (n_lf_max_ini_la) (CUL)	120.00	cm ²
MAX_IT_ DW	Maximum Dry Biomass of Internodes (CUL)	28.00	G
MID_TT_I T_GRO	Thermal time where internodes can achieve half of its maximum biomass (CUL)	400.00	°Cdays
END_TT_I T_GRO	Thermal time for completion of internode growth (CUL)	1200.0 0	°Cdays
MID_TT_L	Thermal time where leaves can achieve half of its maximum biomass (CUL)	700.00	°Cdays

F_GRO			
END_TT_		1300.0	
LF_GRO	Thermal time for completion of leaf growth (CUL)	0	°Cdays
NS_LF_TI L	Number of dry leaves to consider on canopy light transmission (ECO)	3.00	#/tiller
N_LF_MA X_ILA	Number of leaves when leaves appear at it maximum initial area (ECO)	15.00	#/tiller
TB0PHO	Minimum base temperature for photosynthesis (ECO)	12.00	°C
TB1PHO	Optimum temperature for photosynthesis (ECO)	16.00	°C
ТВ2РНО	Optimum temperature for photosynthesis (ECO)	32.00	°C
TBFPHO	Maximum base temperature for photosynthesis (ECO)	40.00	°C
TBPER	Minimum base temperature for plant expansion (ECO)	18.00	°C
TBMAX_P ER	Maximum base temperature for plant expansion (ECO)	35.00	°C
LTTHRES HOLD	Threshold of light transmitted through canopy to start tiller senescence (ECO)	0.40	0-1
FDEADLF	Fraction of dry leaves blade area considered on canopy light transmission (ECO)	0.75	0-1
RDM	Maximum root depth (ECO)	120.00	Cm
DPERCOE FF	Maximum plant expansion rate (ECO)	2.75	mm/h
RWUEP1	Water Stress Sensitivity of Photosynthesis (ECO)	1.00	Dml
RWUEP2	Water Stress Sensitivity of Plant Expansion (ECO)	1.50	Dml
T_MAX_W S_PHO	Supply/Demand ratio where water stress effect is maximum for photosynthesis (ECO)	0.00	0-1
T_MID_W S_PHO	Supply/Demand ratio where half of maximum water stress effect (0.5) for photosynthesis occurs (ECO)	0.20	0-1
T_MIN_W S_PHO	Supply/Demand ratio where water stress effect on photosynthesis onsets (ECO)	0.60	0-1
T_MAX_W S_EXP	Supply/Demand ratio where water stress effect is maximum for expansion (ECO)	0.00	0-1
T_MID_W	Supply/Demand ratio where half of maximum water stress effect (0.5) for expansion	0.40	0-1
S_EXP T_MIN_W S_EXP	Supply/Demand ratio where water stress effect on expansion onsets (ECO)	1.00	0-1
FRAC_SU C_BG	Sucrose fraction of total sugars of below ground internode (ECO)	0.50	0-1
FRAC_HE X_BG	Hexose fraction of total sugars of below ground internode (ECO)	0.50	0-1

INIT_PD_	Initial plant depth of ratoon cane (ECO)		Cm
RAT			
IT_STR_T B_INI	Minimum base temperature for internode structural partitioning (ECO)	18.00	°C
IT_STR_T O1	Optimum temperature for internode structural partitioning (ECO)	31.00	°C
IT_STR_T O2	Optimum temperature for internode structural partitioning (ECO)	55.00	°C
IT_STR_T B_END	Maximum base temperature for internode structural partitioning (ECO)	60.00	°C
IT_STR_P F_MAX	Maximum biomass partitioning fraction to structural parts of internodes (ECO)	0.95	0-1
IT_STR_P F_MIN	Minimum biomass partitioning fraction to structural parts of internodes (ECO)	0.00	0-1
IT_STR_P F_TB	Thermal time for onset of biomass partitioning to internodes structural parts (ECO)	0.00	°Cdays
IT_STR_P F_TM	Thermal time where biomass partitioning to internodes structural parts is at maximum (ECO)	50.00	°Cdays
IT_STR_P F_TE	Thermal time for biomass partitioning to internodes structural parts cessation (ECO)	800.00	°Cdays
IT_STR_P F_D	Shape coefficient for biomass partitioning to internodes structural parts (ECO)	1.00	Dml
IT_STR_T _RED	Maximum reduction of internode structural partitioning due to temperature (ECO)	0.15	0-1
IT_STR_W _RED	Maximum reduction of internode structural partitioning due to water stress (ECO)	0.15	0-1
MAX_PER _IT	Maximum expansion rate of internodes (ECO)	4.00	mm/day
DSWAT_D DWS	Water weight increment per unit of structural biomass gain (ECO)	9.88	$d_{[H2O]}/d_{[STR]}$
DSWAT_D SUC	Water weight decrease per unit of sugars biomass gain (ECO)	11.25	d[H2O]/d[SUG]
HEX_MIN	Minimum hexoses fraction of total sugars in a growing internode (ECO)	0.00	g/g
SUC_ACC_ INI	Internode total sugars concentration where sucrose accumulation onsets (ECO)	0.35	0-1
DSUC_FR	Sucrose weight increment per unit of total sugars increment in internodes (TSUG >	2.50	
AC_TS	suc_acc_ini) (ECO)		d[suc]/d[rsug]
TT_CHUM	Thermal time required after peak of population for tillering stabilization (ECO)	1600.0	°Cdays

AT_LT		0	
SRLMAX*	Specific Root Length at Root Front (SPE)	14.02	m/g
SRLMIN*	Specific Root Length at Inner Roots Profile (SPE)	11.30	m/g
ROOTDR ATE	Root depth rate (SPE)	0.048	cm/°Cdays
MAX_RT_ DW	Maximum Dry Biomass of Roots (SPE)	120.00	g/m²
END_TT_ RT_GRO	Thermal time for completion of root growth (SPE)	3000.0 0	°Cdays
ROOTLEF TFRAC	Fraction of roots left alive after harvesting (SPE)	0.03	0-1
KDIF	Diffuse light extinction coefficient (SPE)	1.00	Dml
KC_MIN	Minimum Crop Coefficient (SPE)	0.79	Dml
EORATIO	Direct relation of LAI with Crop Coefficient (SPE)	1.15	dKc/dLAI
SO2SI_US ERES	Source-sink ratio threshold crop can use its reserves (SPE)	3.00	Dml
MAXLAI_ EO	Maximum Leaf Area Index for Crop Coefficient Increment (SPE)	6.00	m^2/m^2
GRESP	Growth Respiration (SPE)	0.24	g[CH2O]/g[DW]
KMR_LEA F	Michaelis-Menten kinetics of the O2 on Leaves (SPE)	0.01	Km
KMR_STE M	Michaelis–Menten kinetics of the O2 on Stems (SPE)	0.0004	Km
KMR_RO OT	Michaelis-Menten kinetics of the O2 on Roots (SPE)	0.01	Km
KMR_STO R	Michaelis–Menten kinetics of the O2 on Storage (SPE)	0.0002	Km
Q10_LEAF	Leaf proportional change in respiration with a 10 $^{\circ}\mathrm{C}$ increase in temperature (SPE)	1.58	dml
Q10_STEM	Stem proportional change in respiration with a 10 °C increase in temperature (SPE)	1.80	dml
Q10_ROO T	Root proportional change in respiration with a 10 °C increase in temperature (SPE)	1.80	dml
Q10_STOR	Storage proportional change in respiration with a 10 °C increase in temperature (SPE)	1.80	dml
TREF_MR	Reference temperature of adjustment of Q10 respiration function (SPE)	25.00	°C
ТВ	Base temperature for crop development (SPE)	12.00	°C
TBM	Maximum base temperature for crop development (SPE)	50.00	°C
THRESHE WS	Supply/Demand ratio threshold for extreme drought condition counter (SPE)	0.05	0-1
DSHOOT_	Below ground shoots expansion rate (SPE)	0.046	cm/°Cdays

EXT*			
MID_TT_	Thermal time where roots can achieve half of its maximum biomass (SPE)		961
RT_GRO			Cdays
MAX_IT_		5.00	
DW_BG	Maximum Structural Dry Biomass of Below Ground Internodes (SPE)		g
SUC_MIN	Minimum sucrose content of internodes (SPE)	0.07	g[suc]g[dw]
TILLAGEF		1.00	
AC	Kelative age difference among tillers shape coefficient (SPE)		dml
ROOTSHA		2.00	
PE	Root profile shape factor (geotropism function) (SPE)		dml
RWUMAX	Maximum Root Water Uptake Rate (SPE)	0.07	$cm^{3}_{[H2O]}/cm_{[RLD]}$
FRES_USE	Number of times total reserves pool should exceed the total reserves used for emergence	2.00	1 1
D_EM	to allow reserves use for growth (SPE)		ami
AGEFAC_		-2.00	1 1
AMAX	Age factor parameter for CO2 assimilation rate at light saturation point (SPE)		dml
AGEFAC_		-5.00	
PER	Age factor parameter for CO2 assimilation rate for plant expansion (SPE)		dml
C_SCATT		0.15	
ERING	Canopy scattering coefficient (SPE)		0-1
К	Canopy light extinction coefficient (SPE)	0.75	dml
RT_FRON		10.00	
T_SIZE	Vertical size of root front (SPE)		cm

Table 2. Statistical index of performance of the SAMUCA adapted for the DSSAT v4.8 in simulating sugarcane crop components calibrated for cultivar RB867515. Calibration was performed using experiments 16 to 22, whereas model validation (e.g. results presented in this table) utilized experiments 12 to 15 (Tab. 1). Full description of model calibration can be found in Vianna et al. (2020).

Experiments	Variables	Bias	RMSE	EF	R ² d		X	Y
	Dry cane (Mg ha ⁻¹)	-0.993	3.229	0.905	0.924	0.972	12.587	11.594
	Sucrose content (% [Fresh])	1.092	1.543	0.525	0.767	0.882	8.942	10.033
16 += 22	Tiller population (# m ⁻²)	-0.055	3.013	0.641	0.645	0.889	14.544	14.49
10 to 22	Stalk fresh yield (Mg ha-1)	-9.165	18.002	0.863	0.927	0.957	69.267	60.101
	Leaf area index (m ² m ⁻²)	-0.614	1.026	0.361	0.626	0.837	3.214	2.600
	Stalk height (m)	-0.369	0.567	0.562	0.925	0.83	1.204	0.835
	Dry cane (Mg ha ⁻¹)	2.455	3.583	0.942	0.982	0.987	19.290	21.745
	Sucrose content (% [Fresh])	-0.403	1.020	0.640	0.748	0.916	13.232	12.829
10 . 15	Tiller population (# m ⁻²)	0.221	2.601	0.669	0.672	0.897	12.362	12.583
12 to 15	Stalk fresh yield (Mg ha-1)	8.369	13.409	0.952	0.977	0.989	83.242	91.611
	Leaf area index (m ² m ⁻²)	-0.038	0.614	0.573	0.612	0.879	2.595	2.556
	Stalk height (m)	0.437	0.502	0.663	0.925	0.924	0.963	1.401

RMSE: Root mean squared error; EF: Modeling efficiency; R²: Determination index; d: accuracy index of Wilmot; X: Mean observation; Y: Mean simulation; BIAS = Y - X.

Table 3. Description of the sugarcane's stage of development and the leaf area index (LAI) observation date for experiment 1 and 2 (Tab. 1) used for analyses the effect of the number and timing of observations on DA methods (section 2.4.6).

Stage	Description -	LAI obser	vation date
	Description	Exp. 1	Exp. 2
1	Planting to the maximum number of tillers	12/19/2018, 01/02/2019, 01/16/2019 and 02/08/2019	12/19/2018, 01/02/2019, 01/31/2019 and 02/21/2019
2	Maximum number of tillers to the maximum LAI	03/21/2019, 04/23/2019, 05/23/2019 and 07/09/2019	03/09/2019, 03/21/2019, 04/15/2019 and 05/06/2019
3	Maximum LAI to harvest	07/24/2019, 08/07/2019, 08/21/2019 and 09/03/2019	06/05/2019, 07/09/2019, 08/07/2019 and 09/03/2019

APPENDIX B. Assessing the influence of crop model structure on the performance of data assimilation for sugarcane

				Values Calibrated	
Model	Parameters	Description	Range*		
	SPAN	Life span of leaves growing at 35 Celsius (day)	40-90	54.945	
WOFOST	TBASE	Lower threshold temperature for ageing of leaves (°C)	10-20	13.759	
	EFFTB	Initial light-use efficiency single leaf as function of daily mean temperature (kg ha ⁻¹ hr ¹ J ⁻¹ m ² s ¹)	0.44-0.65	0.495(0)/0.456(40)**	
	SLATB	Specific leaf area as a function of DVS (ha kg ⁻¹)	0.0005-0.0008	0.000517(0)/0.000558(0.3)/0.000 614(0.44)***	
DSSAT/S AMUCA	N_LF_STK_EM	Number of leaves appeared before stalks emerges at soil surface (#/tiller)	3-8	4.833	
	N_LF_IT_FORM	Number of leaves appeared before internode formation (#/Tiller)	3-8	4.549	
	TILLOCHRON	Thermal time required for the emergence of one new tiller (°C days/leaf)	48.1-134.8	84.94	
	PHYLLOCHRON	Thermal time required for the appearance of one new leaf (°C days/leaf)	107-169	138.88	
	MLA	Maximum leaf area (cm2)	450-800	610.45	
	PLASTOCHRON	Thermal time required for the appearance of one new phytomere (oC days/phytomer)	107-169	126.84	

Table 1. List of DSSAT/SAMUCA and WOFSOT crop parameters for the RB867515 cultivar, with their respective definitions, units ranged used for calibration and values used in the simulations.

*Minimum and maximum values used for calibration step.

** Values in brackets are the temperature

*** Values in brackets are the development stage