Data assimilation in crop models: old experiences in new contexts

Monique Pires Gravina de Oliveira[1], Luiz Henrique Antunes Rodrigues[1]

[1]Faculdade de Engenharia Agrícola – Universidade Estadual de Campinas (Unicamp)

moniquepgoliveira@gmail.com, lique@unicamp.br

Abstract. *Data assimilation has been widely used for improvement of crop models' estimates, for example to incorporate the effects of external events or compensate calibration errors in large areas. There are then many well-established approaches for those who want to take advantage of satellite imagery and reduce uncertainty or model error. However, its use in different contexts requires exploring aspects of the pipeline that are not as well established, such as which variables to assimilate or how to ascribe uncertainty to observations or model estimates. In this study, we assess the impacts of different noise levels for performing data assimilation in a tomato growth model, with artificial observations of fruit and mature fruit biomass.*

Resumo. *Assimilação de dados é uma técnica que tem sido amplamente utilizada para melhorar as estimativas de modelos de crescimento de plantas, por exemplo, para incorporar os efeitos de eventos externos. Existem muitas abordagens bem estabelecidas para realizar assimilação com imagens de satélite, mas seu uso método em novos contextos, como cultivo protegido, requer a exploração de aspectos da metodologia que não estão tão bem estabelecidos para estes casos. Neste trabalho, avaliamos os impactos de diferentes níveis de incerteza associados às observações, realizando assimilação de dados em um modelo de crescimento de tomateiros, com observações artificiais de biomassa de frutos e frutos maduros.*

Introduction

Data assimilation on crop models has mostly been performed by the integration of remote sensing Earth observations into mechanistic models, often with the goal of improving agricultural systems' models' predictive capability. Technical aspects of the discipline have frequently been revisited given the evolution in computational capacity and available state estimation techniques [Dorigo et al. 2007; Fischer et al. 1997; Huang et al. 2019; Jin et al. 2018; Luo et al. 2023]. These reviews, which detail how the approach has been used in crop modeling, have looked into the subject from different perspectives, such as the methods used to derive biophysical and biochemical canopy

state variables from optical remote sensing data in the VNIR-SWIR regions [Dorigo et al. 2007], the sources of errors in each element of the data assimilation process [Jin et al. 2018], the theoretical basis for methods as well as a walkthrough of the steps required to apply them [Huang et al. 2019], and the models and quantities being assimilated [Luo et al. 2023]. These studies, however, emphasize limitations of satellite-derived observations, e.g. the spatial and temporal scale of satellite images [Huang et al. 2019; Jin et al. 2018; Luo et al. 2023].

In this sense, many of the lessons that have been learned by the crop modeling and remote sensing community could still be discussed and extended into other domains. For instance, in soil water monitoring and irrigation, some studies assess the effect of local measurements on the quality of assimilation [Valdes-Abellan et al. 2019], or how parameter importance shifts as a consequence of irrigation regimes [Orlova and Linker 2023]. For crop growth, greenhouse environments also allow for more intense monitoring, e.g. with daily pictures [Liu et al. 2022; Moon et al. 2022], without the adverse effects of large scales. It would be useful to explore assimilation techniques to enhance accuracy and reduce uncertainty in model estimates obtained for these environments. [Luo et al. 2023] quantified which variables and models are the most used in data assimilation studies and, for variables, the leaf area index was unquestionably the most used. They give multiple reasons, but one relevant aspect not mentioned is the existence of products that allow for coupling model estimates and outputs of satellite images. In new contexts, these relationships must be established. Additionally, they should represent the variables that could in fact be useful for assimilation since not always updating one variable would lead to improvement in another [Nearing et al. 2012]. These new relationships and observations also include uncertainty aspects that need to be quantified for some of the methods more frequently used. The models [Luo et al. 2023] mentioned as most used would not likely be used in protected environments, so uncertainty quantification would also be required for the models explored. Additionally, while the remote sensing realm is dependent on revisit frequency and is vulnerable to unfavorable atmospheric conditions, leading to fewer observations available, high-frequency noisy observations could become a hindrance.

To better understand the subject, it is often useful to explore artificial data, so that it is possible to investigate the behavior of the system with more methodological control. These types of studies have been called Observation System Synthetic Experiments (OSSE) [Nearing et al. 2012; Pellenq and Boulet 2004] and synthetic twin [Lei et al. 2020] and have been used for answering questions such as if the assimilation of an observation improves all components of the model's simulations, if calibration errors can be compensated by assimilation [Pellenq and Boulet 2004], which are limitations imposed by the model, the assimilation method, and uncertainty in model inputs and observations [Nearing et al. 2012], and appropriate ensemble size [Lei et al. 2020].

In this study, we assess how aspects of uncertainty from the decision-making process of performing data assimilation relate to performance and use as an example a greenhouse tomato growth model — the Reduced-State Tomgro model [Jones et al. 1999] —, aiming at improving yield estimates through assimilating artificial observations of tomatoes in a greenhouse environment.

Materials and Methods

Data sources

Environmental data collection was performed in research greenhouses cultivated with tomatoes. The dataset includes photosynthetically active radiation and air temperature from three growth cycles. The first cycle took place from Jul/2019 to Oct/2019 (Exp 1), the second, from Nov/2020 to Feb/2021 (Exp 2), and the third, from Mar/2021 to Jun/2021 (Exp 3). Dry mass from aboveground plants' organs and leaf area index from destructive analyses of one tomato growth cycle (Exp 3) were also collected for model calibration.

Crop model

With the environmental data from greenhouses, we simulated growth using the Reduced State Tomgro model [Jones et al. 1999]. We performed assimilation in the Reduced Tomgro model with parameters obtained for the original experiment in Gainesville, using artificial observations, as explained in the Data assimilation section. As a source for the ground truth and a reference of performance, calibration was performed by minimizing the relative squared error of data obtained in the experiment and models' estimates through growth. Code was implemented in python language and difference equations were integrated by the Euler method. Model code, as all code used in this study and reference to related materials (i.e. data, other studies), is available at [Oliveira 2023].

Data assimilation

For assimilation, we used the Unscented Kalman Filter (UKF) and evaluated the impacts of different approaches for performing data assimilation for yield estimates. Ground truth values corresponded to the simulation performed in each of the three environments with the calibrated Reduced Tomgro. An overview of the elements assessed is detailed below.

- Two assimilated state variables: Fruit dry weight (Wf) and mature fruit dry weight (Wm).
- Observations: Three different noise levels (10%, 30% and 50%) were ascribed to observations of the calibrated Reduced Tomgro Model. The level multiplied by the observation was treated as the standard deviation of a normal distribution from which the perturbation was sampled.
- Uncertainty in crop model: UKF requires determining a value for uncertainty in model estimates. As assimilation was performed in the non-calibrated Reduced Tomgro model, these were ascribed as the relative absolute error of the non-calibrated model to the samples used for calibration (Table 1).
- Uncertainty in observations: The noise level multiplied by the observation was treated as the uncertainty ascribed to that observation.

- We subsampled the observations to study the effect of frequency. Subsampling used 50% and 10% of the data available in the cycle. In one of the repetitions, sampling was regularly spaced through the cycle while in the others, it was randomly sampled.

We repeated the process 20 times to avoid biasing the results due to sampling of the artificial observations.

Table 1. Values ascribed to the filters as uncertainty estimates [%].

Simulation day	State variable	
	Wf	Wm
1-10	0.01*	0.01*
11-27	0.01*	0.01*
28-38	100	0.01*
39-52	94	0.01*
53-66	82	51
67-90	66	68
91-end	66	75

*Placeholder to avoid 0 variance.

We evaluated our approaches by calculating the daily absolute relative error through growth. Our focus on evaluating daily results is related to the indeterminate growth. Differently from other crops in which one value is ascribed to yield, harvest for indeterminate crops is continuous, and therefore, model errors through the growth cycle affect estimates along harvests. As the excess of zeros from the vegetative phase could skew these results, they were not included in the calculation.

Results and Discussion

In this case study, the assimilation of fruits' observations (Wf), except for the highest noise level, only slightly reduced errors in yield (Wm) estimates, when compared to the model without calibration, while the counterpart assimilation of mature fruits' observations (Wm) led to improvements for most experiments (Figure 1). The first result happened because improving Wf estimates did not ensure Wm being correctly estimated. The simulations from the non-calibrated model (OL) showed very close estimates for Wf and Wm, representing a maturity rate much larger than the one from simulated truth (Figure 2). This is a consequence of the Wm estimate depending on a parameter that differed by more than 100% from the non-calibrated to the truth scenarios. The simulated truth also pointed to much larger overall biomass values than the estimates obtained by the non-calibrated model. So, while the model previously underestimated Wm, assimilation led to an overestimation (Figure 2), since Wm was then obtained by the non-calibrated model with the larger parameter, and based on a much larger value of Wf. Interestingly, the largest noise in measurements allowed for model estimates to be more explored by the filter, so the process led to improved estimates by averaging the higher and lower estimates of Wf.

On the other hand, since the assimilation of Wm itself did not depend on the step of the model processing the updated value, improvements, in particular for lower noise levels, are more noticeable. Kalman filter methods are an optimized approach for performing a weighted average, in which the weights are related to the covariance of each estimate. Since we used a non-calibrated model to provide models' estimates, one

could already expect, from Table 1, that the models' large covariances would lead assimilation to take more advantage of observations when their covariances were low. And we see that in general, for the assimilation of Wm, the lower the noise level, the closer the estimates were brought to the truth. It is the case then that not always assimilation is going to improve results, and this outcome depends on how the model will use the updated value and on its sensitivity to this input, i.e., how much the estimate relies solely on the input.

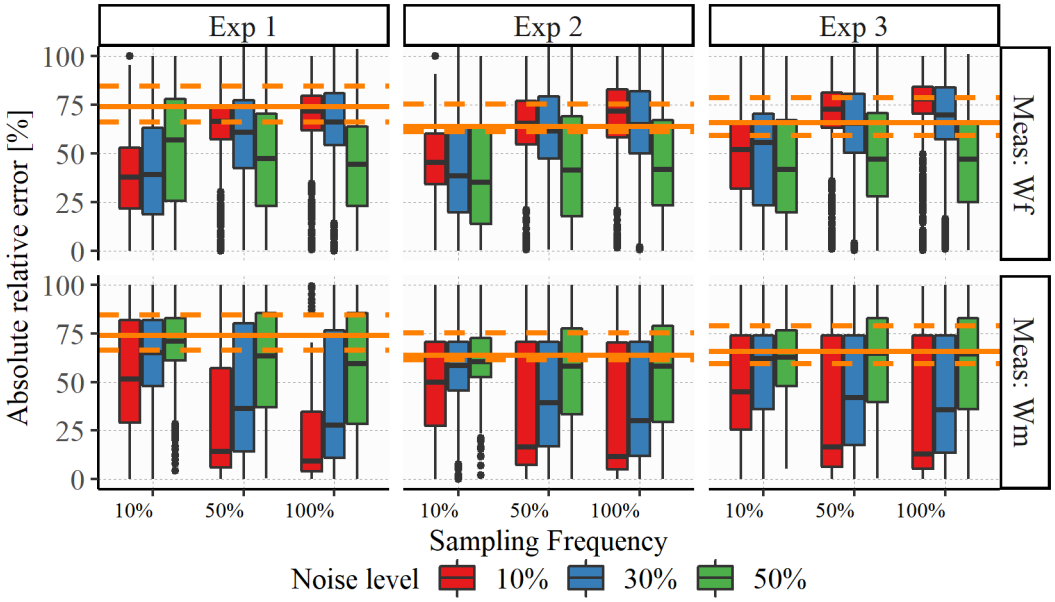

Figure 1. Relative errors for daily estimates of mature fruit dry mass after assimilation of artificial observations of fruit dry biomass (Wf) or mature fruit dry biomass (Wm), obtained by different degrees of perturbations in the outputs of the calibrated Reduced Tomgro model (10%, 30% and 50%), for three weather conditions, with different fractions of the full observation dataset. Horizontal orange lines refer to the relative errors of the Reduced Tomgro model in estimating mature fruits without assimilation: full line corresponds to the median and dashed lines to the 25th and 75th percentiles. Y-axis is truncated at 100%.

Assimilation frequency may be considered complementary to the acceptable noise level in observations, since assimilating an observation with large errors very frequently may not allow for the model to correct the estimates. In our example, this effect depends on which variable is being assimilated, with worse outcomes for the more frequent assimilation of fruits' observations and the opposite being true for the assimilation of mature fruits, with a more pronounced effect for the lowest noise level in both cases. As pointed out previously, the highest noise level in the assimilation of fruits led to the model estimate being more explored, causing a slightly different behavior.

Overall, assimilation with half the observations led to very similar results when compared to the assimilation of the full dataset. This result is interesting when connected to processing capacity. For instance, if the observations are obtained by pictures of plants growing, there would be no need for obtaining, storing, and extracting the related biomass from them every day.

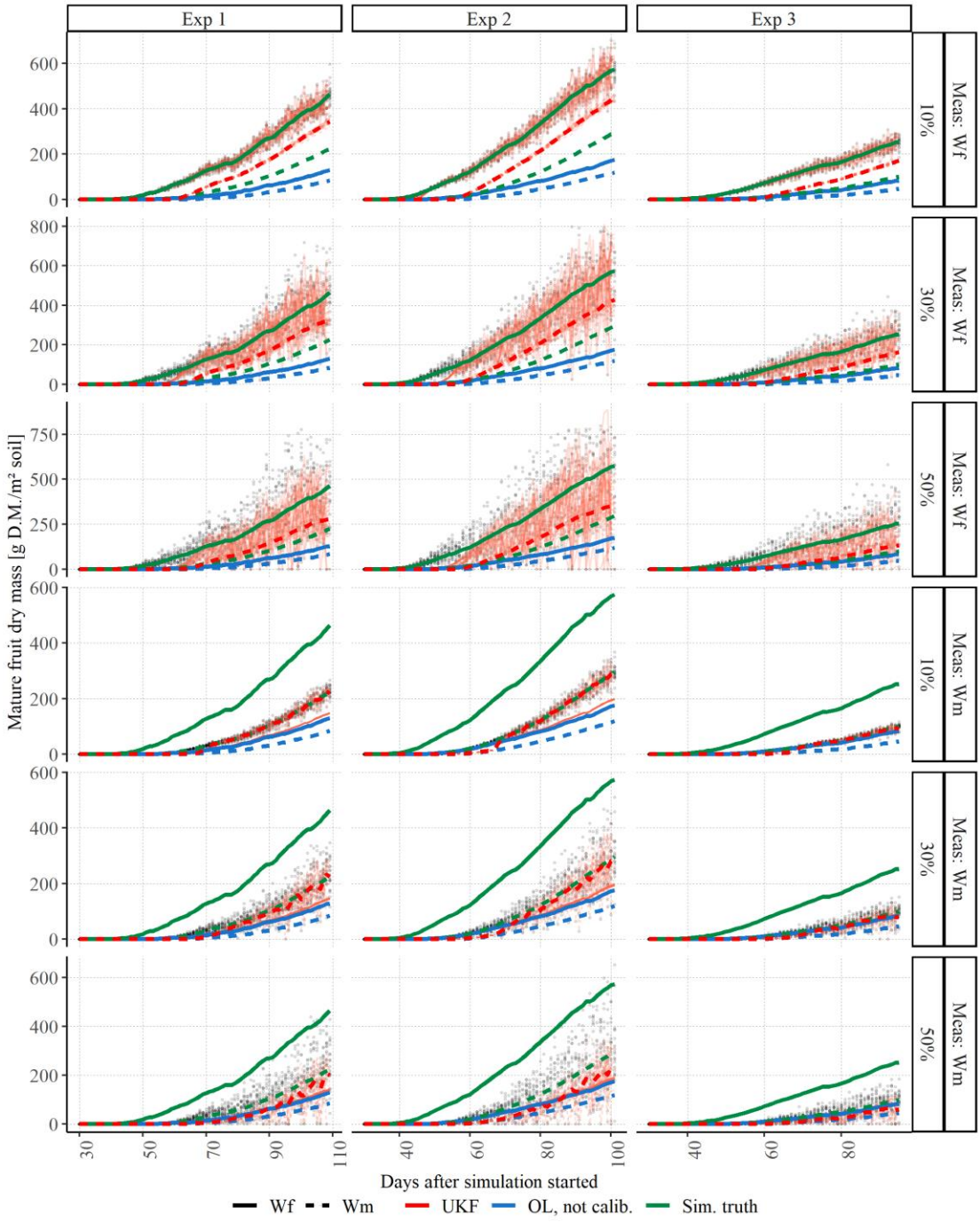

Figure 2. Growth curves [g m^{-2}] for fruit (continuous line) and mature fruit dry biomass (dashed line) with and without assimilation. Assimilation was performed with observations of fruit dry biomass (Meas: Wf) or mature fruit dry biomass (Meas: Wm) with different noise levels (10%, 30%, 50%) in the Reduced Tomgro Model, with the complete observation dataset for the three weather experiments. Assimilation curves in light red refer to all 20 repetitions of the experiment. The darker red curve for mature fruit biomass refers to the average result of the assimilation runs. Dots refer to all observations used in the multiple runs. The x-axis was truncated at 30 days since the previous period corresponds to the vegetative stage.

Conclusions

The discussions on the use of data assimilation often focus on field crops, and for the use of assimilation in protected environments, little has been explored. The main characteristics of this new context for application are the new data sources, as well as the frequency for obtaining them. They could, for instance, rely on daily digital images to estimate fruit mass. In this new context, in which the availability of observations may not be a restriction, some aspects of the process must be reevaluated. In our preliminary assessment of possible noise levels and frequency of assimilation, we aimed at observing how imperfect measurements can allow for improvements in estimates of tomato yield of a non-calibrated tomato growth model. Overall, we observed improvements, but in some cases, when observations and model estimates were equally poor, filtering impaired them even more. As an overview of the method, however, we showed how even with imperfect measurements there may be improvements that lead model performance toward the performance of the calibrated model. While useful in the context of a lack of available data for calibration, these results could be expanded by assessing the same filter parameters in the context of the calibrated model.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and by grant #2018/12050-6, São Paulo Research Foundation (FAPESP).

References

Dorigo, W. A., Zurita-Milla, R., De Wit, A. J. W., et al. (may 2007). A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. *International Journal of Applied Earth Observation and Geoinformation*, v. 9, n. 2, p. 165–193. doi: 10.1016/j.jag.2006.05.003

Fischer, A., Kergoat, L. and Dedieu, G. (19 feb 1997). Coupling Satellite Data with Vegetation Functional Models: Review of Different Approaches and Perspectives Suggested by the Assimilation Strategy. *Remote Sensing Reviews*, v. 15, n. 1–4, p. 283–303. doi: 10.1080/02757259709532343

Huang, J., Gómez-Dans, J. L., Huang, H., et al. (15 oct 2019). Assimilation of remote sensing into crop growth models: Current status and perspectives. *Agricultural and Forest Meteorology*, v. 276–277, p. 107609. doi: 10.1016/j.agrformet.2019.06.008

Jin, X., Kumar, L., Li, Z., et al. (jan 2018). A review of data assimilation of remote sensing and crop models. *European Journal of Agronomy*, v. 92, n. November 2017, p. 141–152. doi: 10.1016/j.eja.2017.11.002

Jones, J. W., Kenig, A. and Vallejos, C. E. (1999). Reduced state-variable tomato growth model. *Transactions of the ASAE*, v. 42, n. 1, p. 255–265. doi: 10.13031/2013.13203

Lei, F., Crow, W. T., Kustas, W. P., et al. (15 mar 2020). Data assimilation of high-resolution thermal and radar remote sensing retrievals for soil moisture monitoring in a drip-irrigated vineyard. *Remote Sensing of Environment*, v. 239, p. 111622. doi: 10.1016/j.rse.2019.111622

Liu, L., Yuan, J., Gong, L., Wang, X. and Liu, X. (20 nov 2022). Dynamic Fresh Weight Prediction of Substrate-Cultivated Lettuce Grown in a Solar Greenhouse Based on Phenotypic and Environmental Data. *Agriculture*, v. 12, n. 11, p. 1959. doi: 10.3390/agriculture12111959

Luo, L., Sun, S., Xue, J., et al. (aug 2023). Crop yield estimation based on assimilation of crop models and remote sensing data: A systematic evaluation. *Agricultural Systems*, v. 210, n. June, p. 103711. doi: 10.1016/j.agsy.2023.103711

Moon, T., Kim, D., Kwon, S., Ahn, T. I. and Son, J. E. (12 oct 2022). Non-Destructive Monitoring of Crop Fresh Weight and Leaf Area with a Simple Formula and a Convolutional Neural Network. *Sensors*, v. 22, n. 20, p. 7728. doi: 10.3390/s22207728

Nearing, G. S., Crow, W. T., Thorp, K. R., et al. (1 may 2012). Assimilating remote sensing observations of leaf area index and soil moisture for wheat yield estimates: An observing system simulation experiment. *Water Resources Research*, v. 48, n. 5. doi: 10.1029/2011WR011420

Oliveira, M. (2023). Leveraging high frequency data for improving crop growth estimates. doi: 10.5281/zenodo.7632419

Orlova, Y. and Linker, R. (2023). Data assimilation with sensitivity-based particle filter: A simulation study with AquaCrop. *Computers and Electronics in Agriculture*, v. 204, n. July 2022, p. 107538. doi: 10.1016/j.compag.2022.107538

Pellenq, J. and Boulet, G. (may 2004). A methodology to test the pertinence of remote-sensing data assimilation into vegetation models for water and energy exchange at the land surface. *Agronomie*, v. 24, n. 4, p. 197–204. doi: 10.1051/agro:2004017

Torres-Monsivais, J. C., López-Cruz, I. L., Ruíz-García, A., Ramírez-Arias, J. A. and Peña-Moreno, R. D. (2017). Data assimilation to improve states estimation of a dynamic greenhouse tomatoes crop growth model. *Acta Horticulturae*, n. 1170, p. 433–440. doi: 10.17660/ActaHortic.2017.1170.53

Valdes-Abellan, J., Pachepsky, Y., Martinez, G. and Pla, C. (2019). How Critical Is the Assimilation Frequency of Water Content Measurements for Obtaining Soil Hydraulic Parameters with Data Assimilation? *Vadose Zone Journal*, v. 18, n. 1, p. 1–30. doi: 10.2136/vzj2018.07.0142