Publication status: Not informed by the submitting author

Kalman Filters in crop models: old experiences in new contexts

Monique Pires Gravina de Oliveira, Thais Queiroz Zorzeto-Cesar, Romis Ribeiro de Faissol Attux, Luiz Henrique Antunes Rodrigues

https://doi.org/10.1590/SciELOPreprints.8033

Submitted on: 2024-02-04
Posted on: 2024-02-06 (version 1)
(YYYY-MM-DD)

Preprint of manuscript *Kalman Filters in crop models: old experiences in new contexts* This version of the manuscript has not been peer-reviewed.

2024-02-03

Kalman Filters in crop models: old experiences in new contexts

Monique Pires Gravina de Oliveira[a1*], Thais Queiroz Zorzeto-Cesar[a2], Romis Ribeiro de Faissol Attux[b3], Luiz Henrique Antunes Rodrigues[a4]

[a] Universidade Estadual de Campinas (UNICAMP), Faculdade de Engenharia Agrícola
Av. Cândido Rondon, 501 - Cidade Universitária - 13083-875 - Campinas, SP, Brazil

[b] Universidade Estadual de Campinas (UNICAMP), Faculdade de Engenharia Elétrica e de Computação, Departamento de Engenharia de Computação e Automação Industrial.
Av. Albert Einstein, 400 - Cidade Universitária -13083-970 - Campinas, SP, Brazil

[1] moniquepgoliveira@gmail.com , https://orcid.org/0000-0001-7167-4473

[2] thaisqzc@unicamp.br , https://orcid.org/0000-0001-6959-7990

[3] attux@unicamp.br , https://orcid.org/0000-0002-2961-4044

[4] lique@unicamp.br , https://orcid.org/0000-0002-1756-7367

[*] Corresponding author

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. The term describes multiple approaches for those who want to take advantage of satellite imagery to reduce uncertainty or improve accuracy of model estimates. Kalman Filters are among the most used methods for achieving these goals. But their use in new contexts, i.e., from open field to protected environments, requires untangling aspects of the pipeline that are often performed in many different ways without guidelines, such as which variables to assimilate or how to ascribe uncertainty to observations or model estimates. This study is then divided in two parts. In the first, we review details on how uncertainty is ascribed on crop model estimates and in observations for applications of the Kalman Filter and three variations of the

method, i.e., the Extended, Unscented and Ensemble, as well as which state variables are often updated and the frequency with which assimilation may occur, as well as how these aspects are connected to each other. In the second part, we apply different approaches from the reviewed literature in a greenhouse tomato crop model. We use artificial data with controlled noise levels as well as artificial data generated by simulation using other tomato crop model. We assess the impacts of using different methods and different approaches for ascribing uncertainty in model estimates and in observations, by assimilating artificial observations of fruit and of mature fruit biomass. We note that covariances should not be fixed values, that there are trade-offs between ascribing model uncertainty to the state itself and to other elements of the process, that observation covariance may have been considered disproportionality higher when using some ensemble generation approaches in the EnKF, and that bias in model estimates may lead to worse outcomes even when observations are high-quality ones. While we discussed aspects that should be considered in a new environment, many of them are also important for field crops, and we concluded assimilation should follow an assessment of which variables could be useful for assimilation.

Keywords: crop model; data assimilation; protected environments; uncertainty; state estimation

1 INTRODUCTION

Data assimilation on crop models has mostly been performed by the integration of remote sensing Earth observations into process-based crop 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 [1–5]. These reviews, which detail how the approach has been used in crop modelling, 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 [2], the sources of errors in each element of the data assimilation process [3], the theoretical basis for methods as well as a walkthrough of the steps required to apply them [4], and the models and quantities being assimilated [5]. These studies, however, emphasize limitations and aspects related to satellite-derived observations, e.g. the spatial and temporal scale of satellite images [3–5].

In this sense, many of the lessons that have been learned by the crop modelling and remote sensing community could still be discussed and extended into other agricultural domains. Protected environments, for instance, allow for more intense monitoring, e.g. with daily pictures [6,7], without the adverse effects of large scale. It could be useful to explore assimilation techniques to enhance accuracy and reduce uncertainty in model estimates obtained in the context of these environments. In the steps suggested by Huang et al. (2019) [4], assimilating observations requires some pre-data assimilation considerations concerning the model and the observations, characterizing uncertainties, and solving a scale mismatch problem between observations and models. In the proposed new context, the first two would be the more relevant steps.

Luo et al (2023) [5] quantified which state variables and models are the more frequently explored in data assimilation studies with remote sensing data and, for variables, leaf area index was unquestionably the most used. They give multiple reasons, but one relevant aspect not mentioned is the previous existence of products that allow for coupling model estimates and outputs of satellite images. For different crops and data sources, these are relationships that must be established. And they should be defined to represent the variables that could in fact be useful for assimilation, since not always updating one variable would lead to improvement in another [8,9]. Furthermore, the models they mentioned as most used would likely not be used in protected environments, for horticultural or ornamental crops. Therefore, the uncertainty

quantification step would be required for them, as well as for these new relationships and observations. Additionally, while the remote sensing realm is dependent on revisit frequency and is vulnerable to unfavorable atmospheric conditions, leading to fewer observations available, the high-frequency potentially noisy observations of a greenhouse environment [10,11] could become a hindrance.

To better understand how these differences could impact new studies, it is 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 sometimes been called Observation System Synthetic Experiments (OSSE) [9,12] or synthetic twin [13] 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 [12], which are limitations imposed by the model, the assimilation method, and uncertainty in model inputs and observations [9], appropriate ensemble size [13], and to test new methods [14].

This study is then divided in two parts. In the first part, we revisit applications of data assimilation in crop models, describing the general use of Kalman Filters and focusing on the steps of uncertainty characterization. In the following sections, we assess how they relate to performance and use as example a greenhouse tomato growth model — the Reduced-State Tomgro model [15] —aiming at improving yield estimates through assimilating artificial observations of tomatoes in a greenhouse environment.

2 BACKGROUND

The reviews of data assimilation in crop models previously mentioned [1–5] categorize three types of data assimilation: forcing, calibration and update. Given the broad scope of three approaches and the several methods that exist to explore them, this study will only comment on the update approach using filters from the Kalman Family, which are very frequently used [5], focusing on the elements more connected to the uncertainty characterization. The reviews also

explain the methods in detail, so we refer to them, especially Huang et al. (2019) [4], for the mathematical equations of the filters.

2.1 Crop models

Briefly, crop models are mathematical representations of plants throughout their growth as affected by the factors that influence said growth, e. g. crop's genetics and the environment. De Vries (1982) proposed three levels of complexity for categorizing crop models: preliminary, comprehensive, or summary models. While the first type contains basic features of the system and aims at a first understanding of the subject, once more knowledge into the processes is gained and imbued into the model, it becomes closer to the second type which, to be made more accessible to others, may then be simplified into the third type, depending on the intended use of the model. Passioura (1996) discussed two other categories, which are related to the intent of their development and use: they could be aimed at farmers, to aid in their decision-making and, for that, understanding of the underlying mechanisms would not be required, or they could be aimed at scientific purposes, which means that the description of mechanisms should be related to theories and validated by hypothesis testing. The first case is generally associated with the terms functional, empirical, statistical and phenomenological and the second, to the terms mechanistic and process-based [18]. In this study, we explore process-based crop models, given their ability to represent interactions between state variables.

2.2 Kalman Filter methods and their requirements

2.2.1 The Kalman Filter (KF)

The main assumptions of the Kalman Filter are that the model estimates and the observations follow a normal distribution, and that the process model and the observation function are linear, but if the non-linear crop model can be assumed locally linear between adjacent time steps, the standard Kalman Filter could be a viable choice [4].

Given its restrictions, there are fewer examples of the application of this technique. In some of them, the premise of the filter is used, but with modifications. Instead of calculating the gain, Vazifedoust et al. (2009) [19] tested different values, using the best one as fixed, circumventing the need for identifying the source error values. This approach was repeated by Chen, Zhang and Tao (2018) [20], who also normalized simulated and observation values according to the maximum value obtained so that they would be in the same range. Operating in this normalized space allowed them to focus on spatial variability and, in part, trends, instead of absolute values. Later, Chen and Tao (2020) [21] explored more approaches for defining an appropriate value for the fixed gain, by a grid search of an optimal value, as well as exploring historical values.

2.2.2 Extended Kalman Filter (EKF)

The Extended Kalman Filter is an adaptation of the Kalman Filter to deal with non-linear cases. To do so, it takes advantage of local linearization by replacing the model and the measurement function by their partial derivatives. The use of this technique is also limited, as it requires access to the Jacobian of the model or, in some cases, to an approximation by finite differences that often will not scale to higher dimensions [4], so there are also few examples on crop modeling that apply this technique.

One of these few examples in which there is an explanation of how the filter was used is the study of Linker and Ioslovich (2017) [22]. The authors used data from growth experiments of cotton and potatoes aiming at improving estimates of canopy cover and biomass through state assimilation and through the recalibration of three parameters from the Aquacrop model. Given there were two different approaches for improving estimates, they estimated the covariance matrix of the errors in the state variables in two ways. For the assimilation process, by calculating the difference between the square of the model residuals and the dispersion of the measurements. They chose not to propagate the matrix along the process, given its strong

nonlinearity, and recalculated it at each new time of measurement. They justified this choice by claiming the propagation without assimilation of new measurements would only increase the uncertainty related to the linearization and to the unknown initial data of the model errors. For the recalibration process, the matrix was calculated using an assumption that the corresponding standard deviation of each of the chosen parameters is equal to 20% of the current value of corresponding parameter. In their assimilation approach, the H matrix, i.e., that used for the linearization, corresponded to the unit matrix, as the measurements directly corresponded to the states and, in the recalibration one, the components of the partial derivatives matrix H were calculated numerically at each instance of canopy cover measurement.

2.2.3 Ensemble Kalman Filter (EnKF)

Overall, in the Ensemble Kalman Filter, an ensemble of initial states is generated and each individual ensemble member is propagated through the model until an observation is available. Then the update step is performed individually in each member. This allows for recalculation of the ensemble mean for the states and generation of a new ensemble. The ensemble approach operates on the underlying assumption that at least some of the particles will represent the true state [12]. There are, however, different ways of approaching this problem and the elements of uncertainty are intimately connected to other decisions.

Composition of ensemble elements

The uncertainty associated with the model derives from the variability in the ensemble. Therefore, the choice of how to generate these elements is reflected on how much the filter will rely on the models' estimates. There have been many approaches for doing so, dealing with every element involved in the simulations, i.e., inputs, parameters, or states.

Input perturbation examples come from Lei et al. (2020) [13], who perturbed precipitation and irrigation inputs via multiplicative rescaling with mean-unity lognormally distributed random errors that have a standard deviation equal to 20% of the corresponding input, and from

De Wit and Van Diepen (2007) [23], who generated precipitation ensembles based on a highly accurate precipitation dataset that was perturbed with an additive error component and a multiplicative component that generated binary rain or no-rain events on locations in which the records pointed to the absence of precipitation. In Oliveira et al. (2023) [24], the authors explored the uncertainty existent in measurements of low-cost luxmeters to perturb the photosynthetically active radiation measurements used as inputs in the model.

In cases in which states are perturbed, Xie et al. (2017) [25] input the initial states and parameters into the CERES-Wheat and, at the beginning of the green-up stage, LAI and soil moisture were perturbed according to the errors between the field measurements and the simulated LAI and soil moisture. Ines et al. (2013) [26] randomly sampled, at the start of the simulation, values of leaf weight at emergence and plant leaf area at emergence, to increase the variability of the ensemble. Beyond inputs, Lei et al. (2020) [13] also applied direct perturbations to soil moisture states at all depths independently with random errors sampled from a mean-zero, normal distribution with temporally varying standard deviation equal to 10% of the state value, followed by the introduction of a vertical auto-correlation at the different depths. Kang and Özdoğan (2019) [8] not only perturbed the initial aboveground biomass or LAI for each ensemble member to simulate model noise, but also initial soil water content, to simulate variations in soil properties and water balance.

Researchers have used multiple ways of ascribing uncertainty to parameters. Kang and Özdoğan (2019) [8] added random noises drawn from a uniform distribution ranges from 0 to 10% of the original values of soil hydraulic parameter. They also chose a parameter that should vary over the course of the growing season and that accounts for stress not simulated by the model to perturb with noise sampled from a uniform distribution centered in the original value. Huang et al. (2016) [27] chose the parameters based on the results of a sensitivity analysis and set the values of the standard deviations of two parameters, following the results of a previous

study. Ines et al. (2013) [26] identified which parameters had major influence in the model and, with an uncertainty level of 10%, perturbed each model parameter using a Gaussian distribution, generating ensemble members by randomly sampling model parameter combinations from the perturbed arrays. Zhao, Chen and Shen (2013) [28] even tried to evaluate the impact of using parameter uncertainty to generate the ensembles. They chose one parameter that was mostly correlated to crop yield and ascribed a distribution to it, multiplying its standard deviation by different fixed values. Lu et al. (2021) [29] took advantage of the existing uncertainty in parameters and used this as an artifact to generate ensembles without calibrating the model. They sampled parameters that they called variant as well as a fixed factor to scale phenological parameters for the canopy in a given year.

One issue in perturbing parameters or inputs for generating the ensembles is what Curnel et al. (2011) [30] denominated phenological shift. This effect refers to ensemble elements that are in different phenological stages, which leads, at the same point in the simulation, to different modules in the model to be active and, therefore, the assimilation of an observation having different a meaning for each ensemble member.

As for the perturbation used in observations — which are also treated as random variables to add to the variance of the updated ensemble [31] — Ines et al. (2013) [26] state that the variance used in its generation is based on the uncertainty of the data. But more precisely, Huang et al. (2016) [27] mentions that the standard deviation of the Gaussian white noise error needs to be a realistic value for it to represent the uncertainty of the remotely sensed observation. In section 2.3, uncertainties in observations are more thoroughly described, but as an example, Xie et al. (2017) [25] used the differences between field measurements and remote sensing observations to determine the standard deviations of the observed LAI and soil moisture.

Ensemble size

The choice of ensemble size is often performed in three different ways: testing, referencing a theoretical result or referencing other assimilation study on the literature. Pellenq and Boulet (2004) [12] affirmed a preliminary study must be performed to find the minimum ensemble size that ensures particles may follow the same trajectory as the true state. They say the number usually corresponds to value above which assimilation results are identical. With this approach, Nearing et al. (2012) [9] showed an example in which the number depended on the goal of the assimilation. The authors tested different values when assimilating LAI and soil moisture aiming at improving estimates of wheat yield, LAI and soil moisture. In the cases of assimilation of the state variable, RMSE became stable with number of elements of 25. In the other cases, the stability came with an ensemble of 100 elements. Lu et al. (2021) [29] evaluated ensemble sizes for simultaneous assimilation of canopy cover and soil moisture from 10 to 400 and overall observed little improvement for more than 200, even though in some years 10 elements were enough for stable results.

Several studies, however, refer to the experiences of other authors. Frequently, authors refer to De Wit and Van Diepen (2007) [23] when commenting on their choice for the ensemble size [28,32,33] The study, however, applies to assimilating soil moisture with an ensemble obtained by perturbing precipitation and with an initial state ascribed by sampling a calculated Gaussian acceptable value and it is possible that they do not generalize for other approaches. Additionally, the authors mention that although they observed reduced RMSE in soil moisture estimates, this was not applied to the variance. Despite that, their results were compatible with other results for soil moisture, and Mishra, Cruise and Mecikalski (2021) [34] followed the suggestion from the study of Yin et al. (2015) [35], who theoretically and through an example showed that the ideal ensemble size for assimilating soil moisture is 12, which suggests 50 would be a reasonable estimate in similar situations.

2.2.4 Unscented Kalman Filter (UKF)

Similarly to the EnKF, the Unscented Kalman Filter uses the average of an ensemble as the state estimate, instead of the direct estimates provided by the model. However, the ensemble is not just sampled from a distribution. It uses what is called the unscented transform to generate particles — the sigma points — and weights for those particles that, when combined, are more representative of the expected state value. These sigma points are propagated through the non-linear model, which provides more accurate approximations of the mean and covariance matrix of the state vector, and thus more accurate state estimation. [36].

Torres-Monsivais et al. (2017) [37] evaluated the UKF along with data simulated with the Reduced State Tomgro model, perturbed by several noise levels. Ruíz-García et al. (2014) [38] used data from destructive analyses of lettuce in a greenhouse to assess uncertainty of the NICOLET model. In the study with tomato, the authors ascribed low covariances to the model and higher to the measurements, and assessed the UKF performance given noise added to model estimates, while in the study with lettuce, the values were tuned until reasonable results were obtained. Recently, Belozerova (2023) [39] emphasized the UKF advantages over the EnKF, namely, its non-stochastic behavior and its being computationally cheaper, with comparable performance.

2.3 Errors and uncertainty

How to identify errors in the elements involved in assimilation and quantify their uncertainties is widely discussed by Jin et al. (2018) [3] and Huang et al. (2019a) [4], as they are central in filtering approaches. For crop models, the sources of uncertainty they list include not only issues referring to how processes are represented and the quality of input data, but also the difference between simulations and actual growth, which is impacted by pests and diseases. For observations, they mention errors in the measurement themselves and in retrieval methods. In both cases, often their studies emphasize aspects of satellite-derived observations, such as

errors in spatial data and scale mismatches. This section aims to revisit this topic, with more details and examples on how these uncertainties have been quantified and applied in data assimilation studies. Although the discussion in this section focus on trying to ascribe meaning and understanding the uncertainties, these are filter hyperparameters that may be estimated from data [40].

2.3.1 Observations

Overall, data assimilation in crop models rely on observations retrieved from satellite monitoring of Earth's surface. Dorigo et al. (2007) [2] covered methods used to derive canopy state variables from optical remote sensing data in the visible to near-infrared and shortwave infrared regions. These methods either rely on statistical relationships between the spectral signature and the measured biophysical or biochemical properties of the canopy or they derive the states from the known behaviors of leaf reflectance and radiation propagation through the canopy. Both are used to obtain remote sensing products, which directly estimate the state for the final user. And both remote sensing products and reflectance itself, are used in assimilation. For those products, Huang et al. (2019a) [4] mention how guidelines for uncertainty quantification are still being established by the community and that many EO-derived products have poor or no uncertainty information available. Particularly for satellite-derived leaf area index (LAI) products, Fang et al. (2019) [41] also comment on how given the complexity associated to the retrieval process, a comprehensive quantitative assessment of the quality of LAI products is still missing. In the case of assimilating reflectance or albedo, the crop model is coupled with a radiative-transfer model (RTM), which allows for quantifying uncertainty in the measurements directly [42].

By assimilating products, several studies [26–28] are able to consider the observation as the same quantity as the state, which means the relationship between states and observations may be obtained by the unit matrix, simplifying the approach. The adverse effects of this choice

303 are not often discussed and the main example we found that mentions them comes from the
304 study of De Wit and Van Diepen (2007) [23], which makes it explicit that the variance they
305 ascribed to observations did not account for deficiencies in the conversion model itself, later
306 concluding that the value they ascribed to the variance was indeed underestimated.

307 When measurements do not correspond directly to the state, filters require observation
308 operators for the conversion. One observation operator often used is the RTM PROSAIL [42–
309 45]. It combines the PROSPECT leaf optical properties model and the SAIL canopy
310 bidirectional reflectance model, and is used for retrieval of vegetation biophysical properties
311 [46]. Huang et al. (2019b) [42] used the RTM PROSAIL, arguing this is a good way to avoid
312 the process of regional LAI retrieval, and Li et al. (2017) [43] used the PROSAIL model and
313 characterized errors in the observations, pointing to values from 0.09 to 0.51 $m^2 m^{-2}$ of error in
314 LAI in the different development stages of wheat. For those who develop their own
315 measurement functions, they often establish them with empirical relationships and characterize
316 their uncertainty based on field data [32,47]. Either in their case, in the case of using the
317 PROSAIL or other RTMs or, for example, in the case of Kang and Özdogan (2019) [8], who
318 used a relationship between the Enhanced Vegetation Index extracted from Landsat images and
319 field LAI observations derived from a global dataset with RMSE of 1.01$m^2 m^{-2}$ and MAE of
320 0.81$m^2 m^{-2}$, these metrics refer to the operator, which, as mentioned, may affect both estimates
321 of state and covariance. For instance, depending on the quality of the observation operator, the
322 conversion may lead to bias in the residual, and consequently, the updated estimate to the wrong
323 value and, or by affecting the gain, to the wrong weight of the residuals in the new estimate.
324 While observation uncertainties should be accounted for separately, these should not be
325 disregarded when understanding the results.

326 As for observation uncertainties themselves, since these studies often refer to large areas,
327 their estimates may be considered as the variability across fields. For instance, Zhao, Chen and

Shen (2013) [28] understood that neighboring pixels had similar uncertainties for the same period and used the variance among fields as uncertainty of remote sensing LAI. Huang et al. (2016) [27] stated that by random sampling their synthetic 30-m LAI series, the observational errors well for each 1-km grid cell were well-represented. For a time-series derived from Landsat, they obtained linear regressions between the observations and LAI values from the field-measured subplots and used the RMSE as observational errors. Kang and Özdogan (2019) [8], on the other hand, assumed constant LAI observation variance at 0.5 m^2 m^{-2}, which they claimed corresponds to the average variance of LAI within a county across the growing season.

Other than satellite retrieved data, there are other sources for observations to which error is ascribed in other ways. For instance, Linker and Ioslovich (2017) [22] and Ruíz-García et al. (2014) [38] used destructive measurements of the assimilated state. In the first case, the authors used direct measurements of aboveground biomass of potatoes and cotton and in the second case, of lettuces. As for non-destructive measurements, Linker and Ioslovich (2017) [22] also used pictures taken from 1.5 and 2 m above the crop to determine canopy cover. Since these were processed to obtain the fraction of the soil surface covered by the canopy, for the Aquacrop model this corresponded to a direct measurement of the state. On these cases, errors corresponded to variance from measurements. Data retrieved by unmanned aerial vehicles (UAVs) often have similar limitations as satellites regarding scale, but brings into discussions other aspects, particularly, as cameras are able to capture other types of data. For instance, Yu et al. (2020) [48] used field measured plant height as well as detected by UAVs and discussed the effects of multiple values ascribed to errors, arguing the trial-and-error procedure could provide a guideline when the true field observation error is unknown. Recently, Han et al. (2022) [49] commented on how, for smallholders, pictures could be a source of observations for assimilation and emphasized the need to proper characterize observation errors. They evaluated an approach in which the measurement and its error were estimated using the

probabilistic output of a convolutional neural network trained with pictures collected during growth using smartphones and labels obtained by destructive and non-destructive measurements.

Finally, one relevant aspect refers to how soil-crop systems may not have a constant value for the error. Nearing et al. (2012) [9] explain how the soil moisture observation uncertainty is variable throughout time, since measurement accuracy degrades as vegetation water content increases throughout the season. They ascribed to error measurement a value derived from the relationship between variance in the soil moisture retrieval and this fraction of plant population and plant biomass that corresponds to water. Lei et al. (2020) [13] evaluated a time-varying error for soil moisture observations as a function of LAI. They observed an overall improvement in soil moisture estimates, but also a somewhat less stable DA performance. Also for soil moisture, Mishra; Cruise and Mecikalski (2021) [34], chose a constant error for the observation, but they were aware that the errors in the sensors used behaved in contrasting ways over crop growth stages, and that this choice may have led to errors that were too low in the early growth season and larger later in the season. Lu et al. (2021) [29] used the multi-year average value of the daily standard deviation of the observations from the 4 soil moisture profiles. But for canopy cover, they noted the error varies dramatically during the growing season, with significant variability in the exponential growth stage and the decay period canopy cover, and only marginal when the canopy was near maximum. So, they assumed canopy cover observation error as dynamic, and the standard deviation of the samplings from the different zones on each sampling day was used separately. Li et al. (2017) [43] considered the standard deviation of the LAI observations as 10% of the measured value, based on their observations of LAI, and Curnel et al. (2011) [30] used a coefficient of variation to characterize uncertainty, thus ascribing to this hyperparameter of the filter a value that corresponded to a fraction of the observation. Belozerova (2023) [39] built a trend curve of LAI observations and used the

deviation from the trend curve as a proxy of measurement uncertainty, ascribing a different error estimate for each observation.

2.3.2 Model estimates

For Wallach and Thorburn (2017) [50], uncertainty means the distribution of the errors of prediction. By defining the model error (e) as in Equation 1, these distributions may be ascertained in two different ways, depending on the predictor ($f(X\ ;\ \theta)$) being treated as fixed or random [51]. If the predictor is treated as fixed, the model error may be ascertained by hindcasts, determining the discrepancy between the prediction and an observed value (Y). If it is not treated as fixed, each of its elements may then be treated as random variables, with several possible values and, therefore, uncertainty in the predictor may have as sources uncertainty in inputs (X), model structure itself ($f(X\ ;\ \theta)$) and parameters (θ).

$$e\ =\ Y\ -\ f(X\ ;\ \theta) \qquad \text{Equation 1}$$

In the case of Pellenq and Boulet (2004) [12], they had two situations, and the differences in model behavior, regarding soil moisture and biomass, required different approaches for determining sources of model uncertainty. When analyzing the effects of initial input values, they observed that for biomass, as the state value is propagated throughout growth, there is no compensation for previous errors, and errors in the estimates of initial conditions could impact the following behavior. And while for soil water, the reliance on previous values is lower, with shorter “memory” of the system, in the coupled case, the initial water content could strongly impact biomass evolution. As for crop model noise, they assumed there would be at least one parameter set in the ensemble that could satisfactorily reproduce natural conditions. So, they decided by generating ensembles ascribing uncertainty to parameters and to inputs. On the other hand, in the case of soil moisture, since it tends towards low variance and equilibrium, they suggested including model noise as well, which should be nonetheless calibrated to avoid the loss of model integrity. Nearing et al. (2012) [9] evaluated uncertainty in weather inputs,

through correlated perturbations in weather time-series. Their results were not conclusive as in one of their systems, the assimilation of LAI improved yield estimates, but not the exclusive assimilation of soil moisture.

Uncertain inputs also manifest through unusual events, which are often not included in models. Therefore, for some authors, an advantage of filter assimilation methods is that they can incorporate these dynamic changes [33]. For example, Hu et al. (2019) [52] improved sugarcane yield estimates by assimilating leaf area index into the SWAP-Wofost model, after the interference in LAI caused by artificial leaf stripping and natural storms, and in Zhao, Chen and Shen (2013) [28], the authors observed high errors when simulating yield for four regions in which meteorological disasters had occurred, which were then reduced to some extent by assimilating observations.

Calibration is an issue that is often mentioned regarding model errors, as it makes the model more consistent with the spatially limited field measurements and calculated uncertainty in parameters could be propagated through the model [4]. Kang and Özdoğan (2019) [8] commented on how, because over large areas, calibration is no longer specific for cultivar, sowing dates or management, model estimates become biased, which violates the assumptions of assimilation techniques that require model errors to have zero means and results in inefficiency of the method. It is then the case that before performing assimilation, models are frequently calibrated [4]. Lu et al. (2021) [29] believed the calibration standard was lower, aiming at having an ensemble of non-calibrated simulations that could capture the dynamics of key model states and that its spread reflected the model state variability. Their assimilation of canopy cover and soil moisture was able to improve yield when compared to the no-assimilation case. Following a similar premise, Fattori Junior et al (2022) [53] also assessed the effect of a non-genotype-specific calibration for multiple cultivars, and observed an overall improvement

in model estimates after assimilation in the non-calibrated model, using three assimilation techniques.

2.4 Variables

In a way, the largest restrictions to performing data assimilation in crop models are which additional data is available and if the knowledge or ability of how to relate them to models' state variables exists. This is one reason why LAI, canopy cover and soil moisture are frequently explored as observations, as there are several satellite products available for them as well as they are often simulated by crop models. To counter availability restrictions, Luo et al. (2023) [5] point to how coupling multi-source remote sensing observations has become more explored and how some studies have shifted from univariate to multivariate assimilation to improve the robustness of crop yield estimates. But being able to perform data assimilation does not ensure that assimilation will be effective. As summarized by Lei et al. (2020) [13], the performance of any data assimilation algorithm is fundamentally related to the strength of the relationship between observations and model states.

For Mishra; Cruise and Mecikalski (2021) [34], assimilation of soil moisture, especially in irrigated areas, led to improvements in yield estimates, which is a very direct relationship, but for Ines et al. (2013) [26], they expected assimilation of soil moisture in the DSSAT-CSM-Maize model to update the rootzone soil moisture, affecting soil nitrogen and, therefore, yield. There is then no guarantee that the included observations will improve estimates. For instance, Linker and Ioslovich (2017) [22] discuss how since the Aquacrop model is water-driven, and as such, solar radiation is not considered explicitly, the effect of canopy cover on crop development may be underestimated. And if assimilation not improving the outcomes is undesirable, it should be noted that it could even have an adverse effect on the estimates, depending on how variables interact with each other. Tewes et al. (2020a) [54] argue that as model complexity rises, sequential update of only one or few state variables could threaten the

model's integrity and cause an undefined state of the model, such as when the simulation triggers a new module by reaching a threshold value, but the filter updates the estimate to a value lower than the threshold.

Time-averaged correlation has been suggested as not very helpful when determining best assimilating state variables by Nearing et al. (2012) [9]. In their experiments, they point to several cases, using different realistic uncertainty scenarios, in which high correlation is not connected to improvement in yield estimates. Nearing et al. (2018) [55] framed this discussion by relying on concepts of information theory, proposing a method to quantify how efficient data assimilation may be, through the quantification of information content on simulated model states and of the retrieval data relative to the imperfect evaluation data, and then measuring the fraction of this information that is extracted by a given DA implementation or algorithm.

2.5 Timing and frequency

An issue that interacts with which variable is going to be assimilated to improve an estimate is at what time of growth and how often should the estimate be updated. Frequently, the discussion is connected to at which moment of the cycle the observation available will be most informative. Dente et al. (2008) [47] evaluated the exclusion of one more precise image and observed that for wheat, within the conditions they observed, the data should include images from either the end of stem elongation stage or the beginning of heading, when the LAI reaches the maximum value. Timing of assimilation in wheat has been widely discussed [8,25,30,43,47] with some authors reaching the conclusion that images from the whole cycle presented the best results [8,43]. For sugarcane, on the other hand, Yu et al. (2020) [48] concluded that assimilation of height in the late period of the elongation stage, involving the maximum plant height, can be the most useful, without the need for its sampling over the whole development stage.

As remote sensing observations are often only available with large intervals between them, their assimilation allows for the model to adjust to the updates, but local assimilation of, for example, soil moisture, would present a different situation. Lu et al. (2021) [29] commented on how their use of local probes for monitoring soil moisture allowed for daily assimilation of this state, which likely improved their results. As crop systems models often present daily steps, it is not the case that assimilation would be performed in more frequent intervals, but in other contexts, such as weather forecasts, it has been argued that very frequent updates could insert noise in models, degrading forecasts [56].

3 MATERIALS AND METHODS

3.1 Data sources

Environmental data collection was performed in research greenhouses cultivated with tomatoes and refer to 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. Data is publicly available [57,58].

3.2 Crop models

With the environmental data from greenhouses, we simulated growth using the Reduced State Tomgro model (RT) [15] and the Vanthoor model (VM) [59]. We performed assimilation in the RT using observations obtained by simulation with both models, as detailed in section 3.3. Because we focus on the RT model, a brief explanation on how it calculates fruit biomass is warranted. In the RT model, fruit biomass (Equation 1) is obtained directly from the balance between photosynthesis and respiration (GR_{net}, i.e., net aboveground biomass growth rate),

which is affected by suboptimal higher ($g(T_{daytime})$) and lower temperatures ($f_F(T_d)$), as well as the maximum partitioning of new growth to fruit (α_F). From fruit biomass, another function of temperature $D_F(T_d)$ impacts the conversion of fruit biomass (W_F) into mature fruit biomass (W_M) (Equation 2). In Equation 1, ϑ and N_{FF} are parameters connected to the transition between vegetative and reproductive stages and N is the number of nodes.

$$\frac{dW_F}{dt} = GR_{net} \cdot \alpha_F \cdot f_F(T_d) \cdot \left[1 - e^{\vartheta-(N-N_{FF})}\right] \cdot g\left(T_{daytime}\right) \quad \text{Equation 2}$$

$$\frac{dW_M}{dt} = D_F(T_d) \cdot (W_F - W_M) \quad \text{Equation 3}$$

For assimilation, RT model was used without local calibration, i.e., with parameters obtained in the original experiment in Gainesville. For obtaining the artificial observations, calibration was performed in the RT and in the VM through minimizing the relative squared error of data obtained in the field and models' estimates through growth. Parameters values for the Reduced State Tomgro model are presented in Supplementary Table A1.

3.3 Data assimilation

From the elements presented in Section 2.2, we chose two techniques, the Ensemble Kalman Filter (EnKF), and the Unscented Kalman Filter (UKF). We evaluated the impacts of different approaches for structuring model uncertainty when performing data assimilation for improving yield estimates. Ground truth values corresponded to the simulations performed in each of the three environments with the calibrated Reduced Tomgro.

- Two assimilated state variables: Fruit dry weight (Wf) and Mature fruit dry weight (Wm).
- Two sources of observations: one (Case 1) in which three noise levels (10%, 30% and 50%) were ascribed to the simulations of the calibrated Reduced Tomgro (X_{truth}) and one (Case 2) in which calibrated Vanthoor model was used for simulations. While a 50% noise level could seem excessive, it should be remarked that it could comprise the error from the

observation as well as the observation operator. In Case 1, the level multiplied by the observation was treated as the standard deviation of a normal distribution from which the perturbation was sampled (Equation 4), leading to the artificial observation (Y_{artif}). Case 2 was proposed as another artificial data source, but one in which the noise level was not chosen, similar to what would happen with observations retrieved from cameras, for example. Case 2 was not explored through all combinations below, but only for the scenarios with best performance for the more similar noise level of Case 1.

$$Y_{artif_i} = X_{truth_i} + N\left(0, X_{truth_i} * Noise\right) \quad \text{Equation 4}$$

- Uncertainty in crop model: UKF requires determining a value for uncertainty in model estimates. Since assimilation was performed in the non-calibrated Reduced Tomgro model, we decided to ascribed these uncertainties as two variants of a relative error metric of the non-calibrated model (Table 1). The first is obtained by multiplying the estimate by the mean absolute percentage error of the model (MAPE, Equation 5, in which X_i is an estimated value from the non-calibrated model and Y_i is an observation measured through the cycle, also used for calibration) from the non-calibrated Reduced Tomgro model for the assimilated variable. This is denominated the fixed approach. The second approach, denominated variable, was obtained by multiplying the estimate by different relative errors through the cycle. Instead of the average value obtained in Equation 5, each relative absolute error (RAE) used in the calculation of MAPE is considered the as uncertainty associated with the estimates for that period (Equation 6). For the EnKF, as this uncertainty could be ascribed in four different ways, we made an assessment of the different approaches: 1. ascribing to the initial states the standard deviation of the observed initial values, 2. ascribing to a parameter, conditioned to which state variable would be assimilated (DFMax for mature fruit and alpha_F for fruits, Supplementary Table A1), a perturbation of 10% of its value, 3.

ascribing to the hourly photosynthetically active radiation a perturbation of 10% of its value and 4. ascribing the same values used for the UKF directly to the model state.

$$MAPE = \frac{1}{n}\sum_i \frac{|Y_i - X_i|}{|X_i|}, Xi > 0 \quad \text{Equation 5}$$

$$RAE = \frac{|Y_i - X_i|}{|X_i|} \quad \text{Equation 6}$$

- Uncertainty in observations: In Case 1, the noise level multiplied by the observation was treated as treated as the uncertainty (R_i) ascribed to that observation (Equation 7). In Case 2, we ascribed the normalized root mean squared error from the calibrated Vanthoor model for the assimilated variable, either using the RAE of the different observations in one day or MAPE of the whole cycle (Table 1). In modes 2 and 3 of the EnKF uncertainty assessment, as these perturbations affect the model difference equation output, instead of the state variable itself, the uncertainty ascribed to observations should take this into account, or variances in the filter will not be connected to the same quantity. To do so, in Case 1, we defined the uncertainty as the difference between the current observation and its previous value, multiplied by the respective noise level (Equation 8). In Case 2, for simplicity, we applied the RAE to the difference between the current artificial observation and the previous one (Equation 9) for variable observation error and, in the fixed case, the value used was the MAPE.

$$R_i = \left(Y_{artif_i} \times Noise\right)^2 \quad \text{Equation 7}$$

$$R_i = \left[\left(Y_{artif_i} - Y_{artif_{i-1}}\right) \times Noise\right]^2 = (R_i - R_{i-1})^2 \quad \text{Equation 8}$$

$$R_i = \left[\left(Y_{artif_i} - Y_{artif_{i-1}}\right) \times RAE\right]^2 \quad \text{Equation 9}$$

- We subsampled the observations to determine the effect of frequency. Subsampling used 50% and 10% of the data available in the cycle. We repeated the process 20 times to

avoid biasing the results due to sampling, and in one of the repetitions, sampling was regularly spaced through the cycle while in the others, it was random.

- The number of elements in the Ensemble Kalman Filter was defined as 250, after evaluation (results not shown).

Table 1. Errors ascribed to the filters as uncertainty estimates.

	Simulation day	Model [%]		Observations (Case 2) [%]	
State variable		Wf	Wm	Wf	Wm
Fixed – Non-calibrated model	All	81	65	43	37
Variable	1-10	0.01*	0.01*	0.01*	0.01*
	11-27	0.01*	0.01*	0.01*	0.01*
	28-38	100	0.01*	164	26
	39-52	94	0.01*	15	2010
	53-66	82	51	20	20
	67-90	66	68	3.8	38
	91-end	66	75	7.9	52

*Placeholder to avoid 0 variance.

There are few examples of how to assess model uncertainty, so we chose a simple approach of using a relative error metric that could be comparable to the noise applied. We evaluated our approaches by calculating the daily absolute relative error through growth (Equation 6). 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. Since the excess of zeros from the vegetative phase could skew these results, they were not included in the calculation. We also show mean error (Equation 10) in order to discuss bias.

$$ME = Y_i - X_i \qquad \text{Equation 10}$$

The code for the crop models was written in python and difference equations were integrated by the Euler method. Filtering used the python filterpy library[1], in which the

[1] https://github.com/rlabbe/filterpy

ensemble Kalman filter function was slightly modified to accommodate the multiple approaches for generating the ensemble. Data preparation for running simulations and simulation and assimilation results were processed using R. All code used in this study, is available in an online repository [60].

4 RESULTS AND DISCUSSION

We have identified in the Background section that even when the problem is well established, ascribing the covariances of filters' parameters is performed in multiple ways. In our evaluation of strategies to determine how to ascribe uncertainty to model and observations, we aimed at observing how imperfect measurements can allow for improvements on estimates of greenhouse tomato yield, when compared to using the model without calibration (Open Loop).

4.1 Assimilation variable

The first aspect we can comment on is the effect of assimilating an observation that is different from the target, i.e., the variable for which the improvement is desired. In this case study, Open Loop simulations severely underestimated the truth (Figure 1), but followed its same trend of largest estimates for Exp 2, which was conducted during spring-summer. Since the artificial observations of Case 1 followed the trends in the truth, relying on them to update the state would lead to increasing fruit biomass, which was observed. However, assimilation of fruits' observations (Wf) in general led to an overestimation of yield, which for the lowest noise level, only slightly reduced the absolute errors in yield (Wm) estimates — whenever it did reduce the error —, when compared to the model without calibration. Table 2 shows mean errors, to make explicit the differences in over and underestimation caused by assimilation, but improvement in absolute errors is highlighted in bold. On the other hand, assimilation of mature fruits' observations (Wm) led to improvements for all experiments.

Table 2. Mean error (after fruit appearance) of mature fruit biomass with and without assimilation, using the variable approach for observation (only Case 2) and for model uncertainties (both cases), with simulated observations of fruit biomass (Wf) and mature fruit biomass (Wm). Bold numbers correspond to the cases in which the absolute value is lower than the case without assimilation.

Case	Noise level	Assimilated variable	Filter	Exp 1	Exp 2	Exp 3
No assimilation				-29.0	-36.6	-11.5
1	10 %	Wf	EnKF	**21.1**	**29.2**	**11.1**
			UKF	**27.3**	37.2	13.7
		Wm	EnKF	**-13.1**	**-16.3**	**-4.02**
			UKF	**-1.42**	**-2.94**	**-1.59**
	30 %	Wf	EnKF	**5.75**	**13.0**	**4.41**
			UKF	**25.7**	**34.0**	12.3
		Wm	EnKF	**-20.8**	**-26.0**	**-6.90**
			UKF	**-4.89**	**-7.60**	**-3.49**
	50 %	Wf	EnKF	**-5.35**	**-5.77**	**-1.12**
			UKF	**14.3**	**20.1**	**5.71**
		Wm	EnKF	**-23.2**	**-30.6**	**-8.89**
			UKF	**-14.1**	**-21.2**	**-8.04**
2	-	Wf	EnKF	**-14.7**	**-29.6**	**1.33**
		Wm	EnKF	**-26.8**	**-38.1**	**-8.9**

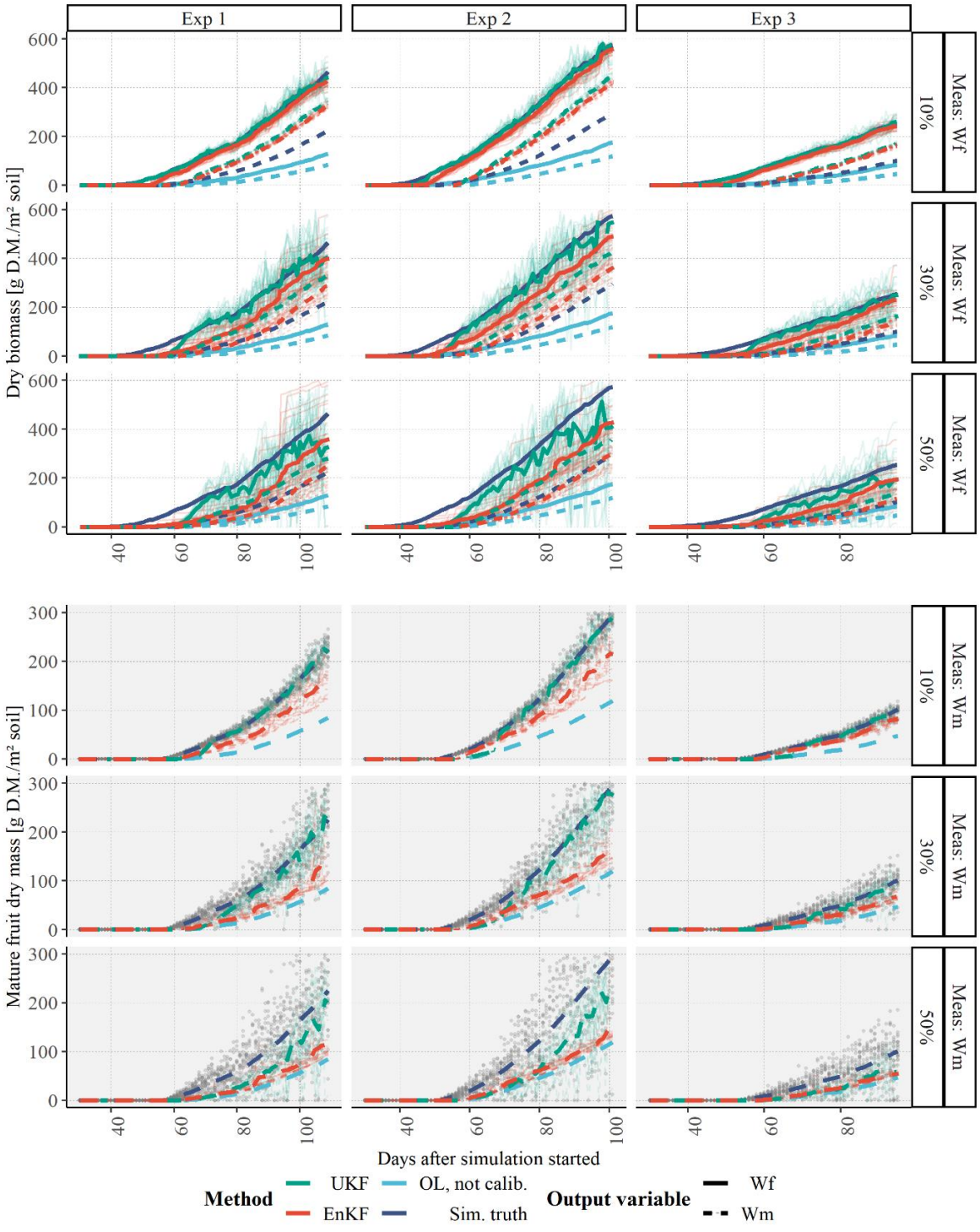


Figure 1. Overview of data assimilation results. Growth curves [g m^{-2}] for fruit (continuous line) and mature fruit dry mass (dashed line) obtained by assimilation of observations of fruit dry biomass (Meas: Wf) or mature fruit dry biomass (Meas: Wm), with different noise levels, in the Reduced Tomgro Model, with the complete observation dataset for the three weather experiments and ascribing variable uncertainty for both model and observations. The Ensemble Kalman Filter used Approach 2 (parameter perturbation) for ensemble generation. Assimilation curves in lighter colors refer to each of the 20 repetitions of the experiment. The darker curve refers to the average result of the assimilation runs. Dots refer to all observations used in the multiple runs. For better visibility, fruit mass observations were truncated at 800 g m^{-2}. The x-axis was truncated at 30 days since the previous period corresponds to the vegetative stage.

The first result happened because improving Wf estimates could 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. This is a consequence of the Wm estimate depending on a parameter that differed in more than 100% from the non-calibrated to the truth scenarios (Supplementary Table A1). 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, since Wm was then obtained by the non-calibrated model with the larger parameter, and based on a much larger value of Wf, as updated by the filter. This is similar to the problem described by Kang and Özdogan (2019) [8] as error inconsistency. 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 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. This effect is largely connected to the relationships between uncertainties as used by the Kalman Filters. 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.

An additional issue, that is obfuscated by assessing metrics only, is the behavior associated with the estimates. Directly updating the state, as is the case with the UKF, may lead to frequent non-monotonic behavior that is not compatible with plant growth. In Figure 1, although the average curve represents a smoothed result, individual curves, represented in

lighter colors, are very erratic, in particular for higher noise levels. There are two possibilities that could attenuate this issue. The first is the frequency of assimilation. When Torres-Monsivais et al. (2017) [37] assessed the performance of the UKF in LAI assimilation with the RT, it seems that because they used fewer observations, the results they show do not lead to unreasonable curves, similar to the cases of updates performed with satellite imagery. The effect of frequent observation assimilation is further discussed on section 4.5. The second is the source of error in observations. In the example presented, the effect was amplified by sampling observations from a normal distribution with zero mean but increasingly larger variances. Despite the variability observed in the assimilation of fruits, the erratic behavior is attenuated when the updated fruit biomass estimate is processed by the model to estimate mature fruit biomass, since the update occurs in the level of the daily increment.

Finally, there are trade-offs in the decision of which variable to assimilate. We chose in this study to simulate observations more directly connected to yield. This choice would still allow for using the crop model for forecasting, ensuring the starting point of these chosen variables is as correct as possible, but it nevertheless limits the advantage of exploring all the model's potential. As stated in the section 2.4, other variables are more often used either given data availability or how much they affect the desired state variable. Therefore, while updating mature fruit biomass directly gives improved updates, it does not take advantage of the models' interaction components and, for example, impacts on leaf area that could lead to a reduction in fruit biomass will only be detected whenever fruit biomass is itself affected.

4.2 Noise level

It could be said that it is pointless to assimilate observations that are known to be incorrect, but since every measurement include some sort of error, the discussion is on the threshold accepted. For instance, the overall relative root mean squared error of high resolution remote sensing data used as reference values for the leaf area index of crops used in the

validation of moderate resolution LAI products ranged from 2.1% to 37% with a median of 22% [41]. In our case, visually, noise levels corresponding to 30% still reasonably represent the truth, but noise levels of 50% may lead to very unreasonable, non-monotonic, estimates. In our case, sampling was centered on the truth, so the fluctuations represent noise in measurements, but observation errors could also be caused by systematic bias, which we did not represent.

In this Case 1, observations were expected to improve estimates, since the model was not calibrated and observations followed the trend in the truth. We also expected that increasing noise would lead to deterioration of filter performance. Both happened for the assimilation of mature fruit observations. But as the noise in observation increases and this information is passed to the filter, along with the observation covariance, the model estimate is further explored by the filter. And when the errors are in opposite directions, which is the case of assimilation of the biomass of fruits, they compensate and the result is an overall improvement in mature fruit biomass estimate.

4.3 Ensemble generation approach

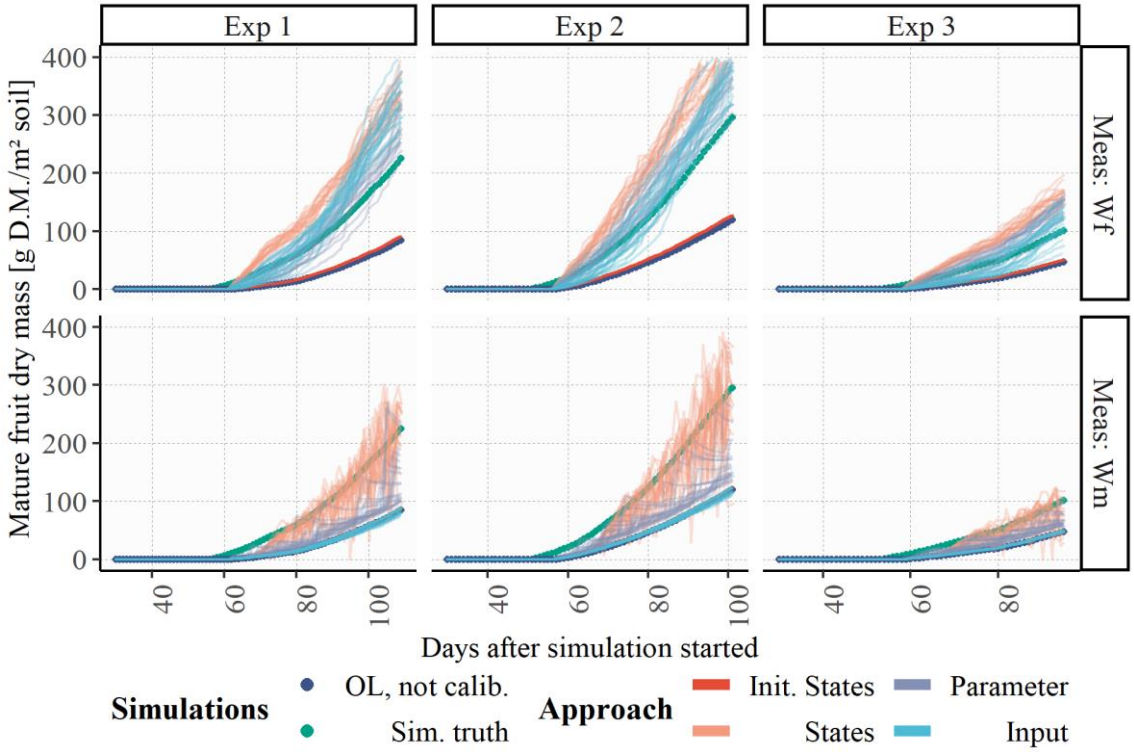
In Figure 1, we show only the results of the Ensemble Kalman Filter for the second approach of obtaining the ensemble described in section 3.3. We decided on further using this approach for the assessments after we observed the effects of the different approaches on yield (Figure 2) using the intermediate noise level. We noticed how, differently from directly updating the states either with the EnKF or the UKF, this approach led to reasonable results while not generating implausible curves even for the noisier observations, as discussed on section 4.1.

Furthermore, the Unscented Kalman Filter, when assimilating mature fruits, led to higher mature fruit biomass than fruit biomass (result not shown), which would not be realistic. This effect was not present in the EnKF as it relies more on model structure. Leaving the ensemble to be obtained by perturbing inputs and parameters likely lead to smoother results since they

affect the increment of the state variable and not the state itself. This means that, for observation uncertainty to be compatible with the variability generated within the ensemble, we needed to calculate uncertainties also based on the how these uncertainties affected the increments. This specific approach of associating the covariance of the observations with the magnitude of the increment is not explored in other studies.

On the other hand, changing only initial states or perturbing inputs did not improve the results as much. Changing initial states, while sometimes used, is less appropriate for a variable that will only change later in the cycle, differently from soil moisture, aboveground biomass, or leaf area index. To better assess this approach, perturbations could be performed at fruit appearance. Moreover, the choice of noise level ascribed to inputs likely led to less variability in the state variable being assimilated than the noise level that affected the parameters. This, in turn, led to lower covariance in the ensemble and more reliance in model estimates than in the observations. Similarly, even though we chose the parameters with highest values in a sensitivity analysis [61], the same 10% perturbation led to very different curve dispersion for fruits or for mature fruits.

It should also be noted that by leaving the uncertainty to perturbation, we imply that the estimate should be somewhere around the variability caused by this parameter or input, while ascribing it to the state considering model error, also carries an information with regard to the bias of the model. Albeit larger, it allows for the extra compensation, if needed.



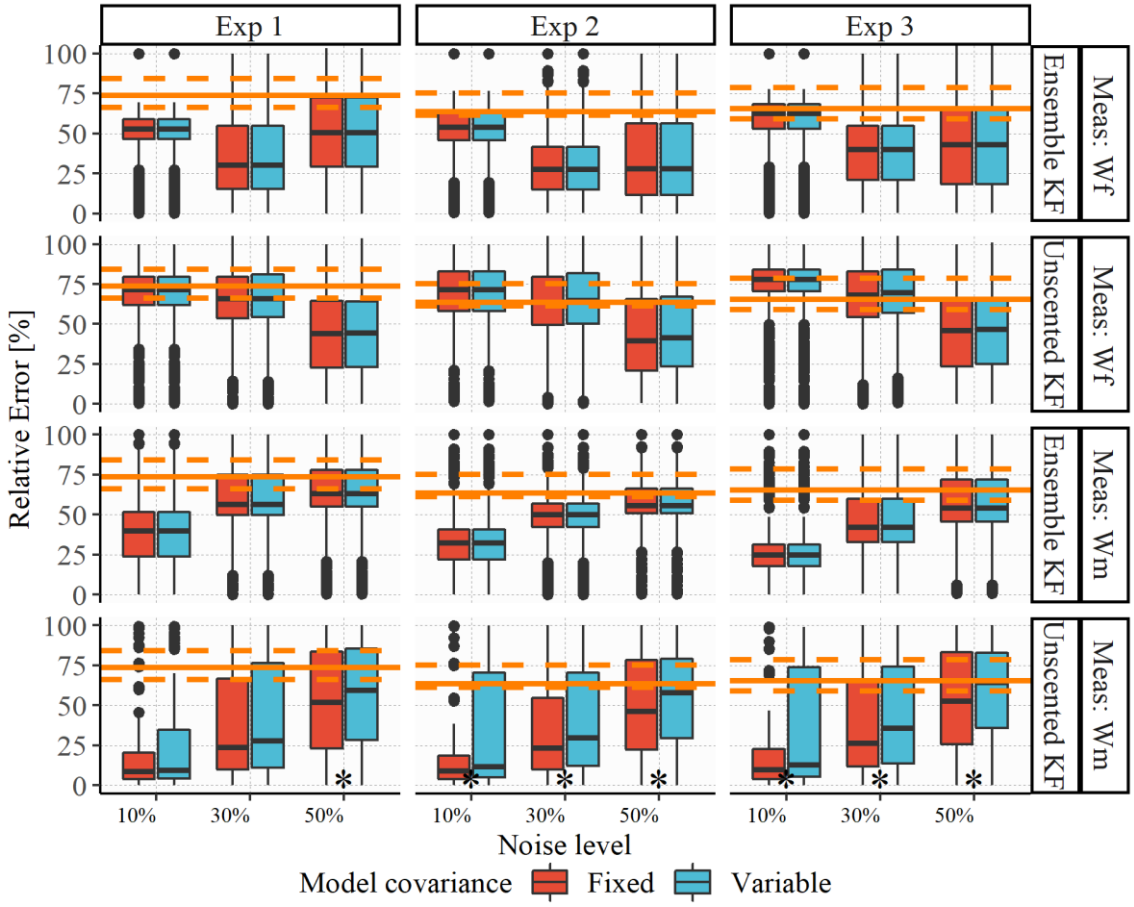
724

725 Figure 2. Outcomes for different ensemble approaches. Growth curves for the estimates of mature fruit
726 biomass [g m^{-2}], by assimilation of fruit dry biomass (Meas: Wf) or mature fruit dry biomass (Meas:
727 Wm), using the Ensemble Kalman Filter with different approaches for ascribing uncertainty to the
728 model: to the initial state (Init. States), directly to the state variables (States), perturbing a parameter
729 (Parameter) or perturbing an input (Input). All curves for the 20 repetitions are shown, in each of the
730 three weather experiments, for the case of variable model uncertainty in simulations using 30% of noise
731 in observations.
732

4.4 Effect of variable model error

After determining strategies for generating the ensemble, we assessed the effect of treating the noise level as fixed or variable. In this assessment, we meant to deal with an aspect of crop models that is that they often do not have constant variance [62]. The few studies that do not leave uncertainty to be determined in the ensemble define both a fixed value [37] or variable ones [22]. We did not observe a large difference between approaches (Figure 3), even though some are statistically different, likely because this effect was already mitigated by the use of relative values. Even though we present the results, there were, as expected, no differences between approaches in the assimilation using the EnKF, since for the ensemble generation method used, the covariance is derived from the perturbation in the parameters and is not an input. In the EnKF case, the model covariance naturally varies throughout the simulation.

For the UKF, the assimilation of fruits led to similar results between the approaches, within the same method and noise level, likely because the overall magnitudes of the uncertainties used either fixed or through growth (Table 1), are similar and the updated estimates are further modified by the model process. As previously discussed in section 4.1, the cases in which there was improvement in the estimates occur for the highest noise levels in observations. For the assimilation of mature fruit biomass, overall, the UKF was sensitive to the differences in covariance and even though the median error is similar, the variable approach frequently led to higher errors. These errors may be connected to small differences during mature fruit appearance, but are nonetheless quantified as high relative errors.



754

755 Figure 3. Assessment of the error in using different model covariance approaches. Relative errors for
756 daily estimates of mature fruit dry mass obtained by assimilation of artificial observations of fruit dry
757 biomass (Meas: Wf) or mature fruit dry biomass (Meas: Wm), obtained by different degrees of
758 perturbations in the outputs of the calibrated Reduced Tomgro model (noise levels 10%, 30% and 50%),
759 through the whole cycle. The results refer to three weather conditions, with two different approaches of
760 assigning uncertainty to the model, using both the Unscented and the Ensemble Kalman Filters.
761 Horizontal orange lines refer to the relative errors of the Reduced Tomgro model in estimating mature
762 fruits without assimilation: full line corresponds to the median and dashed lines to the 25th and 75th
763 percentiles. Y-axis is truncated at 100%. Asterisk corresponds to different distributions using the
764 Kolmogorov-Smirnov test with a 0.05 significance level.

4.5 Frequency of assimilation

Assimilation frequency may be considered complementary to the acceptable noise level in observations, since assimilating an observation very frequently may not allow for the model to have an input in the estimates. In our example, this happens, counterintuitively, to the low noise level in the assimilation of fruits. We commented on section 4.1 how assimilating fruit observations overestimates mature fruit estimates. Therefore, the more frequent assimilation of them lead to increasing the error in yield estimates. When the noise increases, the opposite is true, and because the filter relies more on the model, the more frequent assimilation of these observations lead to a better result.

For the assimilation of mature fruits, we had already seen (section 4.2) that the assimilation of observations with the lowest noise levels led estimates to be closer to the truth, but the magnitude of the improvement depends on the covariance associated with model estimates (Figure 4). For the EnKF, the covariance generated on the ensemble seems lower than the one ascribed to the UKF so, as the noise in the observations increases, observations are less explored by the EnKF than by the UKF. Therefore, for the EnKF, reducing the frequency has a more pronounced effect when the noise level is the lowest, but less so when it increases.

The performance of the UKF follows overall the same pattern, but the covariance ascribed to the model allows the filter to explore more of the observations even with higher noise levels. So the effects of more frequent assimilation of fruit biomass leading to an increase in the error of yield estimates and of increasing median errors with fewer observations of mature fruit biomass occur up to the noise level of 30%.

By sampling twenty times randomly we do not assess the effect of timing of observations, but we can discuss the usefulness of continuous monitoring. In this example, for the highest noise level, assimilation with half the observations led to very similar results, in particular for the assimilation of mature fruit observations. 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.

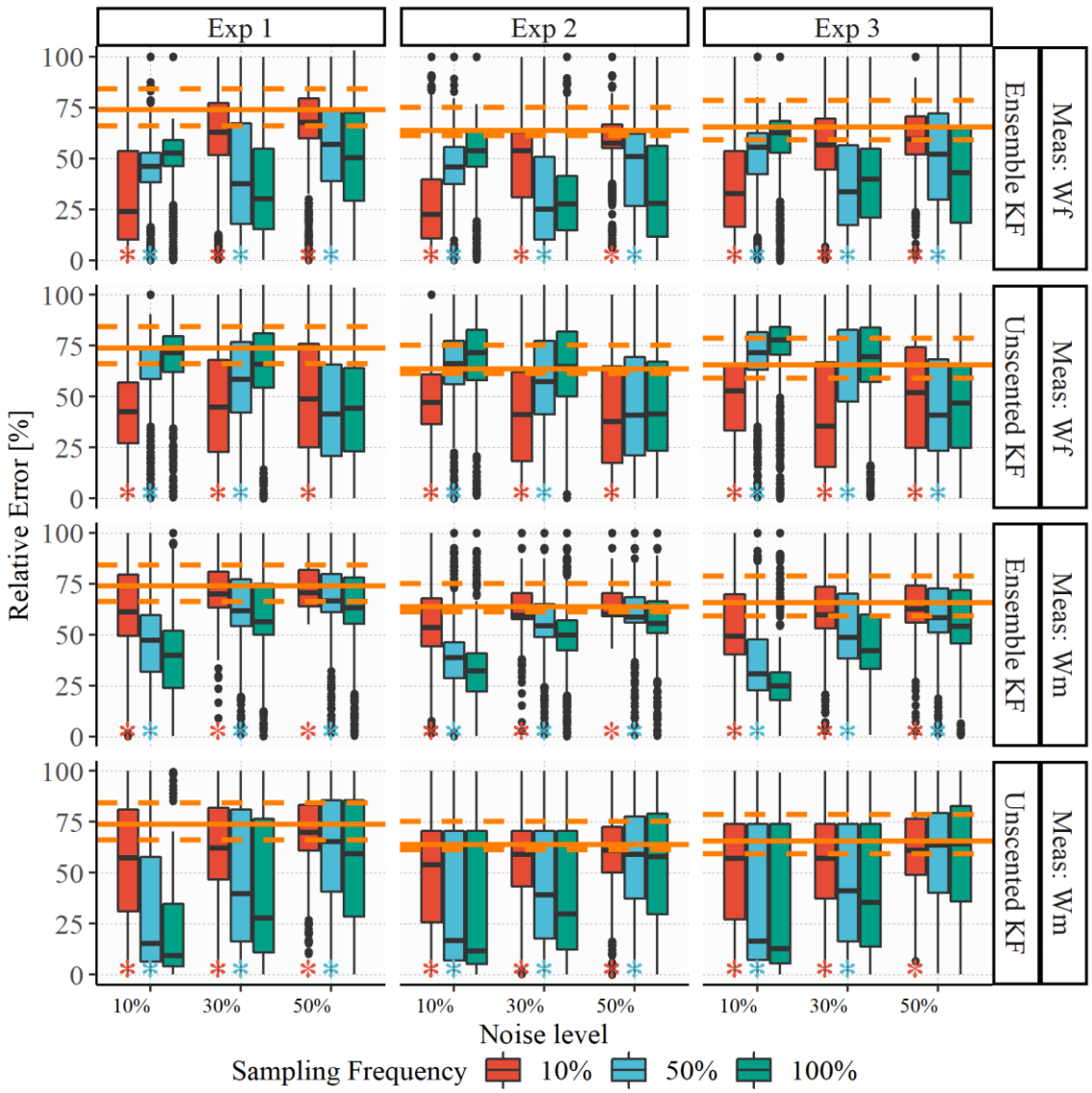


Figure 4. Sampling frequency effect on assimilation. Relative errors for daily estimates of mature fruit dry mass obtained by assimilation of artificial observations of fruit dry biomass (Meas: Wf) or mature fruit dry biomass (Meas: Wm), obtained by different degrees of perturbations in the outputs of the calibrated Reduced Tomgro model (noise levels 10%, 30% and 50%), through the whole cycle. The results refer to three weather conditions, with different fractions of the complete observation dataset, using both the Unscented and the Ensemble Kalman Filters. 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%. Asterisk corresponds to different distributions using the Kolmogorov-Smirnov test with a 0.05 significance level.

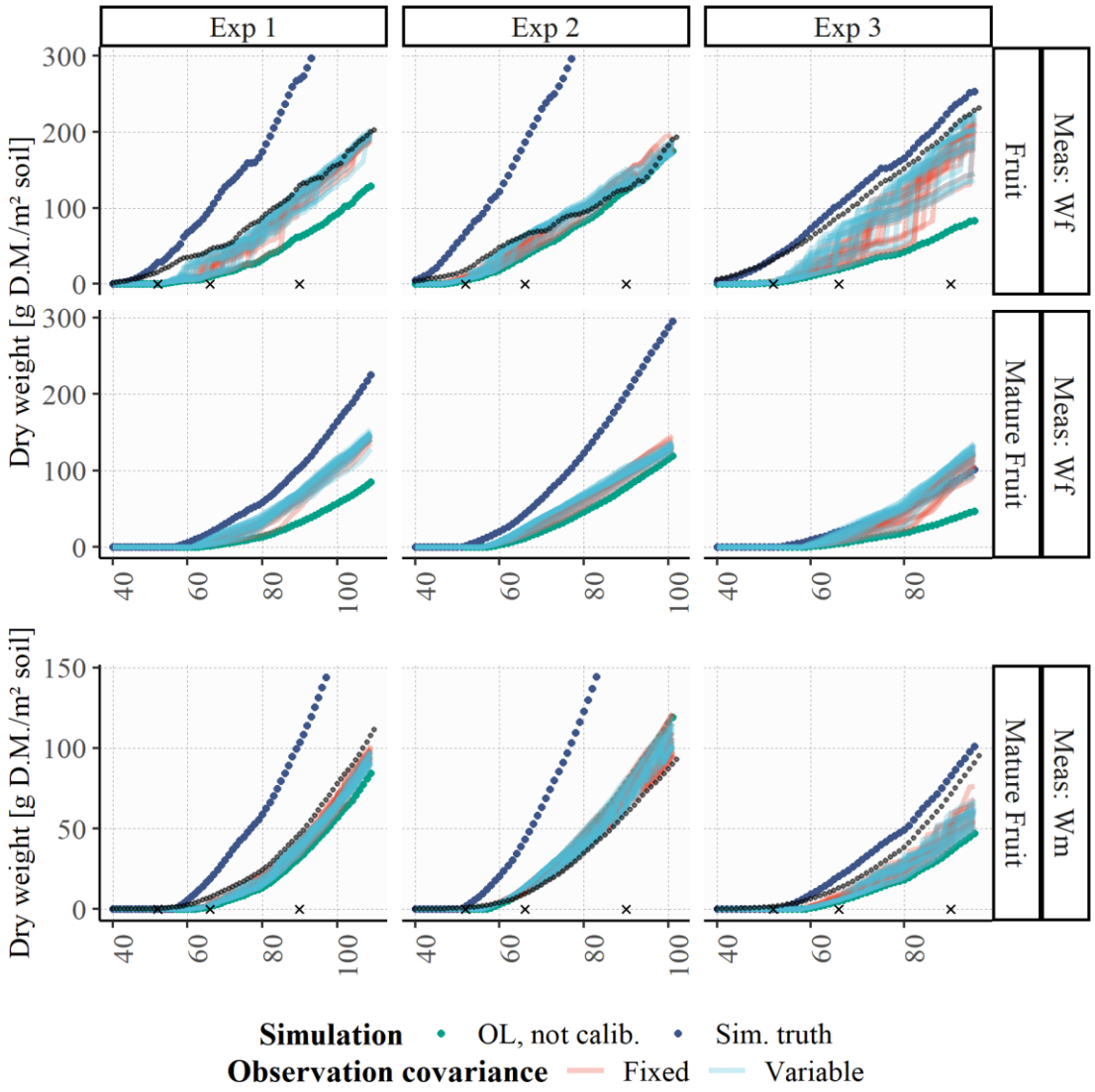
4.6 Different source for observations

In this example, observations were obtained by simulation using the calibrated Vanthoor model. From the previous results, given its more stable performance, we only use the EnKF, which also allows for not having to choose the covariance behavior through the cycle. We again chose the approach of ensemble generation by using perturbation of parameters. Since in this case the uncertainty on observations is not by design, the covariance ascribed to them was also evaluated as variable through the cycle or fixed. While we could assess this case without repetitions associated with the observations, since they are simulated instead of obtained by adding noise, we decided to use 10% of observations to make more visible the differences in the two observation covariance approaches, thus leading to repeating the procedure to account for variability in sampling.

This is a scenario in which both model and observations underestimate the truth (Figure 5), leading to the high errors seen in Table 2. But in all examples, the observation covariance ascribed was such that observations are quite explored by the filter, which also resulted in the improvement of the estimates. The exception occurs during fruit appearance, when the updated estimates are much closer to the ones obtained by the model. For the assimilation of fruits, with both approaches of ascribing the covariance, this is quite visible before the first date in which the covariance changes, marked by an ‘x’. For the assimilation of mature fruits, this shift happens at the second date, more visibly in Exp 3.

We explored the approach of scaling the uncertainty in the observations by the period to which they refer, observing that without it, the gain would always be too low. By scaling, in all three experiments, for both approaches in the assimilation of fruits, we observed gain values ranging from 0 to almost 1, varying through the cycle. We mostly observed no difference between varying or fixing the covariance in the observations. Even the seemingly high fixed values ascribed (43% and 37%, Table 1) were apparently often of a compatible magnitude to the covariance generated by the ensemble.

By using fewer observations, we more clearly see the shifts in estimates caused by assimilation. After a shift in the magnitude, the Open Loop curve no longer represents model estimates, since the following estimates are based on the shifted updated value. This explains why, in Exp 2, some estimates are larger than both observation and open loop estimates.



836

837 Figure 5. Assimilation results considering a different source for observations. Growth curves for fruit
838 and mature fruit dry mass obtained by assimilation of fruit dry biomass (Meas: Wf) and for mature fruit
839 dry biomass by assimilation of mature fruit dry biomass (Meas: Wm). Observations correspond to
840 estimates of the Vanthoor model, using the Ensemble Kalman Filter (EnKF) with subsamples of 10%
841 of the observation dataset for the three weather experiments and ascribing variable uncertainty for both
842 model and observations and only to model. The x-axis was truncated at 40 days, since the previous
843 period corresponds to the vegetative stage. Curves refer to all 20 repetitions of the experiment. For better
844 visibility, fruit mass observations were truncated at 300 g m^{-2} in the upper panels and at 150 g m^{-2} in the
845 lower panels.
846

5 FINAL REMARKS AND CONCLUSION

The recently published reviews help to quantify and better understand the prevalent approaches of remote sensing assimilation into crop models, but when the specific application is different in all aspects including which models, observations and state variables are to be used, there are practical decisions that go beyond what has been discussed. Our goal in this study was to contextualize and apply data assimilation in the context of protected environments. These environments are used to grow different crops, e.g. horticultural or ornamental, which entails new models and new data sources. These data sources could be connected to multiple variables represented by crop models, free from the previous restrictions of what satellite imagery could provide. They could, for instance, rely on digital images to estimate either fruit mass or leaf area. This would also have implications in the frequency, which would no longer be restricted to satellites temporal resolution. But not only it is unclear which new variables could be useful, but also how to ascribe uncertainty to them would have to be thought through, since filter uncertainty parameter choices really impacts outcomes. And as these elements interact, there were many aspects to be discussed.

Giving meaning to filters' hyperparameters is a choice, as they could be tuned to obtain the best outcome [20,38]. But recent discussions in the crop modeling community have been moving from strictly data driven approaches towards hybrid ones [63] and the need for machine learning models to explore the knowledge constraints present in crop models has been emphasized [64,65]. For parameter calibration, instead of blind sensitivity analysis followed by optimization, the community has been proposing a mixture of expert knowledge in parameter and bound selection, followed by optimization using mathematical criteria (AIC and loss function) [66,67]. Therefore, while no meaning could lead to better outcomes, they would also hinder understandability in the process. We focused on approaching explainable ways to ascribing uncertainty to filters. One takeaway from our review and analysis is that observation covariances should not be fixed values. Kalman filter methods are an optimized approach for

873 performing a weighted average, in which the weights are related to the covariance of each
874 estimate. Since errors grow in magnitude through the cycle, the process is more informed if
875 covariances can vary with the magnitude of the observations and estimates, which naturally
876 occurs in the Ensemble Kalman Filter. Ascribing covariance values as a proportion of the
877 magnitude of estimates or observations can mitigate this issue.

878 We also noted how since these models work at the difference step, perturbation of
879 elements of the equation to generate the ensemble leads to the need of scaling observation
880 covariance. It is possible that this outcome is being achieved by using model uncertainty
881 inflation, as suggested by Ines (2013) [26]. We noted in our experiments that, without scaling,
882 the observation uncertainty was too large in comparison to the one generated in the ensemble,
883 leading to gains almost always equal to zero. This was possible, however, because we directly
884 assimilated observations. It's unclear how scaling could work when an observation operator is
885 involved. This empirical result could be further investigated.

886 For model error, much of the discussion refers to bias, since many studies use the EnKF,
887 which generates the model uncertainty. Using estimates from a non-calibrated model means
888 more bias and uncertainty in the model estimate, so it is well-established that both should be
889 reduced by parameter calibration to obtain a better estimate. Our example highlighted the aspect
890 of updating the state itself being a weighted average which depended on the covariances
891 ascribed both to model and observation, but that update of a different variable could be heavily
892 affected by the relationship of the updated and the target variable.

893 Finally, we noted that which variable to assimilate is an open question for field data as
894 well. Previous assessments with artificial data could be useful for avoiding trial-and-error ad-
895 hoc studies with laborious data gathering steps. In a new context, in which there is little previous
896 research pointing to which variables could be useful, choosing the variable means defining
897 which data will be collected. In that sense, good understanding of the models is helpful. Our

study allowed for directly assessing the relationships between the assimilated variable and a parameter or another state, but this is often not possible, since the models are usually more complex and the relationships between variables are not so direct.

In part, extended state estimation is solving the issue by leveraging sensitivity analysis, so that at least one component of the model, i.e., parameters, are changed with assurance of impact on the target variable. Orlova and Linker (2023) [14] used sensitivity analysis immediately before the assimilation step, to account for changes in parameter importance through the growth cycle. This approach, however, still relies on how the model represents the process, which prevents the assimilation advantage of capturing processes not included in the model. It could be said that parameter would behave as dynamic, accommodating these processes, but in this case, they may lose their original meaning. Furthermore, the value would be identified within the process. To be able to choose which variable and through it, which data to collect, the analysis would have to be performed prior to assimilation. A sequential sensitivity analysis for multiple state variables [68] could allow for identifying which parameter impacts both the various state variables that could be assimilated as well as the target state variable, suggesting an interesting variable to assimilate.

Sensitivity analyses methods are well-established to characterize which elements impact an outcome the most. It focuses, however, on parameters, and effects from bias on other state variables and weather conditions are not included. Ines et al. (2013) [26] concluded that weather conditions expected during the growing season could provide information as to when a variable is best to be assimilated, and this would include a beneficial constraint in a sensitivity analyses across multiple years [61].

To address the two main points of how to characterize uncertainties and how to choose useful variables, it could be the case that protocols, such as those being proposed for calibration [67,69,70] could also be explored. At least a framework could be designed on how uncertainties

should be understood before moving into data gathering. The would be helpful both for practitioners that are now approaching the subject, but also for those that have already been using remote sensing observations, but with new models and crops.

6 Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, by grant #2018/12050-6, São Paulo Research Foundation (FAPESP), and by CNPq (grant #308811/2019-4). The funding sources had no role in the study design; the collection, analysis, and interpretation of data; the writing of the report; or the decision to submit the article for publication.

7 Authors' contributions

Monique Oliveira: Conceptualization, Methodology, Software, Investigation, Data Curation, Writing - Original Draft; Thais Zorzeto-Cesar: Resources, Writing - Review & Editing; Romis Attux: Writing - Review & Editing; Luiz Rodrigues: Writing - Review & Editing, Supervision, Funding acquisition

8 Conflicts of interest

The authors declare that there is no conflict of interest.

9 References

[1] A. Fischer, L. Kergoat, G. Dedieu, Coupling Satellite Data with Vegetation Functional Models: Review of Different Approaches and Perspectives Suggested by the Assimilation Strategy, Remote Sens. Rev. 15 (1997) 283–303. https://doi.org/10.1080/02757259709532343.

[2] W.A. Dorigo, R. Zurita-Milla, A.J.W. de Wit, J. Brazile, R. Singh, M.E. Schaepman, A review on reflective remote sensing and data assimilation techniques for enhanced

946 agroecosystem modeling, Int. J. Appl. Earth Obs. Geoinf. 9 (2007) 165–193.
947 https://doi.org/10.1016/j.jag.2006.05.003.
948 [3] X. Jin, L. Kumar, Z. Li, H. Feng, X. Xu, G. Yang, J. Wang, A review of data assimilation
949 of remote sensing and crop models, Eur. J. Agron. 92 (2018) 141–152.
950 https://doi.org/10.1016/j.eja.2017.11.002.
951 [4] J. Huang, J.L. Gómez-Dans, H. Huang, H. Ma, Q. Wu, P.E. Lewis, S. Liang, Z. Chen,
952 J.-H. Xue, Y. Wu, F. Zhao, J. Wang, X. Xie, Assimilation of remote sensing into crop
953 growth models: Current status and perspectives, Agric. For. Meteorol. 276–277 (2019)
954 107609. https://doi.org/10.1016/j.agrformet.2019.06.008.
955 [5] L. Luo, S. Sun, J. Xue, Z. Gao, J. Zhao, Y. Yin, F. Gao, X. Luan, Crop yield estimation
956 based on assimilation of crop models and remote sensing data: A systematic evaluation,
957 Agric. Syst. 210 (2023) 103711. https://doi.org/10.1016/j.agsy.2023.103711.
958 [6] L. Liu, J. Yuan, L. Gong, X. Wang, X. Liu, Dynamic Fresh Weight Prediction of
959 Substrate-Cultivated Lettuce Grown in a Solar Greenhouse Based on Phenotypic and
960 Environmental Data, Agriculture. 12 (2022) 1959.
961 https://doi.org/10.3390/agriculture12111959.
962 [7] T. Moon, D. Kim, S. Kwon, T.I. Ahn, J.E. Son, Non-Destructive Monitoring of Crop
963 Fresh Weight and Leaf Area with a Simple Formula and a Convolutional Neural
964 Network, Sensors. 22 (2022) 7728. https://doi.org/10.3390/s22207728.
965 [8] Y. Kang, M. Özdoğan, Field-level crop yield mapping with Landsat using a hierarchical
966 data assimilation approach, Remote Sens. Environ. 228 (2019) 144–163.
967 https://doi.org/10.1016/J.RSE.2019.04.005.
968 [9] G.S. Nearing, W.T. Crow, K.R. Thorp, M.S. Moran, R.H. Reichle, H. V. Gupta,
969 Assimilating remote sensing observations of leaf area index and soil moisture for wheat
970 yield estimates: An observing system simulation experiment, Water Resour. Res. 48

971 (2012). https://doi.org/10.1029/2011WR011420.
972 [10] J. Rong, H. Zhou, F. Zhang, T. Yuan, P. Wang, Tomato cluster detection and counting
973 using improved YOLOv5 based on RGB-D fusion, Comput. Electron. Agric. 207 (2023)
974 107741. https://doi.org/10.1016/j.compag.2023.107741.
975 [11] F. Zhang, Z. Lv, H. Zhang, J. Guo, J. Wang, T. Lu, L. Zhangzhong, Verification of
976 improved YOLOX model in detection of greenhouse crop organs: Considering tomato
977 as example, Comput. Electron. Agric. 205 (2023) 107582.
978 https://doi.org/10.1016/j.compag.2022.107582.
979 [12] J. Pellenq, G. Boulet, A methodology to test the pertinence of remote-sensing data
980 assimilation into vegetation models for water and energy exchange at the land surface,
981 Agronomie. 24 (2004) 197–204. https://doi.org/10.1051/agro:2004017.
982 [13] F. Lei, W.T. Crow, W.P. Kustas, J. Dong, Y. Yang, K.R. Knipper, M.C. Anderson, F.
983 Gao, C. Notarnicola, F. Greifeneder, L.M. McKee, J.G. Alfieri, C. Hain, N. Dokoozlian,
984 Data assimilation of high-resolution thermal and radar remote sensing retrievals for soil
985 moisture monitoring in a drip-irrigated vineyard, Remote Sens. Environ. 239 (2020)
986 111622. https://doi.org/10.1016/j.rse.2019.111622.
987 [14] Y. Orlova, R. Linker, Data assimilation with sensitivity-based particle filter: A
988 simulation study with AquaCrop, Comput. Electron. Agric. 204 (2023) 107538.
989 https://doi.org/10.1016/j.compag.2022.107538.
990 [15] J.W. Jones, A. Kenig, C.E. Vallejos, Reduced state-variable tomato growth model,
991 Trans. ASAE. 42 (1999) 255–265. https://doi.org/10.13031/2013.13203.
992 [16] F.W.T.P. De Vries, Phases of development of models, in: Simul. Plant Growth Crop
993 Prod., Pudoc, 1982: pp. 20–25.
994 [17] J.B. Passioura, Simulation models: Science, snake oil, education, or engineering?,
995 Agron. J. 88 (1996) 690–694.

https://doi.org/10.2134/agronj1996.00021962008800050002x.

[18] J.W. Jones, J.M. Antle, B. Basso, K.J. Boote, R.T. Conant, I. Foster, H.C.J. Godfray, M. Herrero, R.E. Howitt, S. Janssen, B.A. Keating, R. Munoz-Carpena, C.H. Porter, C. Rosenzweig, T.R. Wheeler, Brief history of agricultural systems modeling, Agric. Syst. 155 (2017) 240–254. https://doi.org/10.1016/j.agsy.2016.05.014.

[19] M. Vazifedoust, J.C. van Dam, W.G.M. Bastiaanssen, R.A. Feddes, Assimilation of satellite data into agrohydrological models to improve crop yield forecasts, Int. J. Remote Sens. 30 (2009) 2523–2545. https://doi.org/10.1080/01431160802552769.

[20] Y. Chen, Z. Zhang, F. Tao, Improving regional winter wheat yield estimation through assimilation of phenology and leaf area index from remote sensing data, Eur. J. Agron. 101 (2018) 163–173. https://doi.org/10.1016/J.EJA.2018.09.006.

[21] Y. Chen, F. Tao, Improving the practicability of remote sensing data-assimilation-based crop yield estimations over a large area using a spatial assimilation algorithm and ensemble assimilation strategies, Agric. For. Meteorol. 291 (2020) 108082. https://doi.org/10.1016/j.agrformet.2020.108082.

[22] R. Linker, I. Ioslovich, Assimilation of canopy cover and biomass measurements in the crop model AquaCrop, Biosyst. Eng. 162 (2017) 57–66. https://doi.org/10.1016/j.biosystemseng.2017.08.003.

[23] A.J.W. de Wit, C.A. van Diepen, Crop model data assimilation with the Ensemble Kalman filter for improving regional crop yield forecasts, Agric. For. Meteorol. 146 (2007) 38–56. https://doi.org/10.1016/j.agrformet.2007.05.004.

[24] M.P.G. De Oliveira, T.Q. Zorzeto-Cesar, R.R.D.F. Attux, L.H.A. Rodrigues, Can accuracy issues of low-cost sensor measurements be overcome with data assimilation?, Eng. Agrícola. 43 (2023) 2–8. https://doi.org/10.1590/1809-4430-eng.agric.v43n2e20220170/2023.

1021 [25] Y. Xie, P. Wang, X. Bai, J. Khan, S. Zhang, L. Li, L. Wang, Assimilation of the leaf area
1022 index and vegetation temperature condition index for winter wheat yield estimation using
1023 Landsat imagery and the CERES-Wheat model, Agric. For. Meteorol. 246 (2017) 194–
1024 206. https://doi.org/10.1016/J.AGRFORMET.2017.06.015.
1025 [26] A.V.M. Ines, N.N. Das, J.W. Hansen, E.G. Njoku, Assimilation of remotely sensed soil
1026 moisture and vegetation with a crop simulation model for maize yield prediction, Remote
1027 Sens. Environ. 138 (2013) 149–164. https://doi.org/10.1016/J.RSE.2013.07.018.
1028 [27] J. Huang, F. Sedano, Y. Huang, H. Ma, X. Li, S. Liang, L. Tian, X. Zhang, J. Fan, W.
1029 Wu, Assimilating a synthetic Kalman filter leaf area index series into the WOFOST
1030 model to improve regional winter wheat yield estimation, Agric. For. Meteorol. 216
1031 (2016) 188–202. https://doi.org/10.1016/j.agrformet.2015.10.013.
1032 [28] Y. Zhao, S. Chen, S. Shen, Assimilating remote sensing information with crop model
1033 using Ensemble Kalman Filter for improving LAI monitoring and yield estimation, Ecol.
1034 Modell. 270 (2013) 30–42. https://doi.org/10.1016/j.ecolmodel.2013.08.016.
1035 [29] Y. Lu, T.P. Chibarabada, M.G. Ziliani, J.K. Onema, M.F. Mccabe, J. Sheffield,
1036 Assimilation of soil moisture and canopy cover data improves maize simulation using an
1037 under-calibrated crop model, Agric. Water Manag. 252 (2021) 106884.
1038 https://doi.org/10.1016/j.agwat.2021.106884.
1039 [30] Y. Curnel, A.J.W. de Wit, G. Duveiller, P. Defourny, Potential performances of remotely
1040 sensed LAI assimilation in WOFOST model based on an OSS Experiment, Agric. For.
1041 Meteorol. 151 (2011) 1843–1855. https://doi.org/10.1016/j.agrformet.2011.08.002.
1042 [31] G. Burgers, P. Jan van Leeuwen, G. Evensen, Analysis Scheme in the Ensemble Kalman
1043 Filter, Mon. Weather Rev. 126 (1998) 1719–1724. https://doi.org/10.1175/1520-
1044 0493(1998)126<1719:ASITEK>2.0.CO;2.
1045 [32] T. Bai, S. Wang, W. Meng, N. Zhang, T. Wang, Y. Chen, B. Mercatoris, Assimilation

of Remotely-Sensed LAI into WOFOST Model with the SUBPLEX Algorithm for Improving the Field-Scale Jujube Yield Forecasts, Remote Sens. 11 (2019) 1945. https://doi.org/10.3390/rs11161945.

[33] Y. Li, Q. Zhou, J. Zhou, G. Zhang, C. Chen, J. Wang, Assimilating remote sensing information into a coupled hydrology-crop growth model to estimate regional maize yield in arid regions, Ecol. Modell. 291 (2014). https://doi.org/10.1016/j.ecolmodel.2014.07.013.

[34] V. Mishra, J.F. Cruise, J.R. Mecikalski, Assimilation of coupled microwave / thermal infrared soil moisture profiles into a crop model for robust maize yield estimates over Southeast United States, Eur. J. Agron. 123 (2021) 126208. https://doi.org/10.1016/j.eja.2020.126208.

[35] J. Yin, X. Zhan, Y. Zheng, C.R. Hain, J. Liu, L. Fang, Optimal ensemble size of ensemble Kalman filter in sequential soil moisture data assimilation, Geophys. Res. Lett. 42 (2015) 6710–6715. https://doi.org/10.1002/2015GL063366.

[36] M. Mansouri, B. Dumont, M.-F. Destain, Modeling and prediction of nonlinear environmental system using Bayesian methods, Comput. Electron. Agric. 92 (2013) 16–31. https://doi.org/10.1016/j.compag.2012.12.013.

[37] J.C. Torres-Monsivais, I.L. López-Cruz, A. Ruíz-García, J.A. Ramírez-Arias, R.D. Peña-Moreno, Data assimilation to improve states estimation of a dynamic greenhouse tomatoes crop growth model, Acta Hortic. (2017) 433–440. https://doi.org/10.17660/ActaHortic.2017.1170.53.

[38] A. Ruíz-García, I.L. López-Cruz, A. Ramírez-Arias, E. Rico-Garcia, Modeling uncertainty of greenhouse crop lettuce growth model using Kalman Filtering, Acta Hortic. 1037 (2014) 361–368. https://doi.org/10.17660/ActaHortic.2014.1037.44.

[39] O.D. Belozerova, Enhancing WOFOST crop model with unscented Kalman filter

1071 assimilation of leaf area index, Int. J. Image Data Fusion. 00 (2023) 1–16.
1072 https://doi.org/10.1080/19479832.2023.2287037.
1073 [40] D. Wallach, D. Makowski, J.W. Jones, F. Brun, Data Assimilation for Dynamic Models,
1074 in: D. Wallach, D. Makowski, J.W. Jones, F. Brun (Eds.), Work. with Dyn. Crop Model.,
1075 Third Edit, Elsevier, 2019: pp. 487–518. https://doi.org/10.1016/B978-0-12-811756-
1076 9.00014-9.
1077 [41] H. Fang, F. Baret, S. Plummer, G. Schaepman-Strub, An Overview of Global Leaf Area
1078 Index (LAI): Methods, Products, Validation, and Applications, Rev. Geophys. 57 (2019)
1079 739–799. https://doi.org/10.1029/2018RG000608.
1080 [42] J. Huang, H. Ma, F. Sedano, P. Lewis, S. Liang, Q. Wu, W. Su, X. Zhang, D. Zhu,
1081 Evaluation of regional estimates of winter wheat yield by assimilating three remotely
1082 sensed reflectance datasets into the coupled WOFOST–PROSAIL model, Eur. J. Agron.
1083 102 (2019) 1–13. https://doi.org/10.1016/J.EJA.2018.10.008.
1084 [43] H. Li, Z. Chen, G. Liu, Z. Jiang, C. Huang, Improving Winter Wheat Yield Estimation
1085 from the CERES-Wheat Model to Assimilate Leaf Area Index with Different
1086 Assimilation Methods and Spatio-Temporal Scales, Remote Sens. 9 (2017) 190.
1087 https://doi.org/10.3390/rs9030190.
1088 [44] L. Zhang, C.L. Guo, L.Y. Zhao, Y. Zhu, W.X. Cao, Y.C. Tian, T. Cheng, X. Wang,
1089 Estimating wheat yield by integrating the WheatGrow and PROSAIL models, F. Crop.
1090 Res. 192 (2016) 55–66. https://doi.org/10.1016/j.fcr.2016.04.014.
1091 [45] C. Guo, Y. Tang, J. Lu, Y. Zhu, W. Cao, T. Cheng, L. Zhang, Y. Tian, Predicting wheat
1092 productivity: Integrating time series of vegetation indices into crop modeling via
1093 sequential assimilation, Agric. For. Meteorol. 272–273 (2019) 69–80.
1094 https://doi.org/10.1016/J.AGRFORMET.2019.01.023.
1095 [46] S. Jacquemoud, W. Verhoef, F. Baret, C. Bacour, P.J. Zarco-Tejada, G.P. Asner, C.

François, S.L. Ustin, PROSPECT + SAIL models: A review of use for vegetation characterization, Remote Sens. Environ. 113 (2009) S56–S66. https://doi.org/10.1016/J.RSE.2008.01.026.

[47] L. Dente, G. Satalino, F. Mattia, M. Rinaldi, Assimilation of leaf area index derived from ASAR and MERIS data into CERES-Wheat model to map wheat yield, Remote Sens. Environ. 112 (2008) 1395–1407. https://doi.org/10.1016/j.rse.2007.05.023.

[48] D. Yu, Y. Zha, L. Shi, X. Jin, S. Hu, Q. Yang, Improvement of sugarcane yield estimation by assimilating UAV-derived plant height observations, Eur. J. Agron. 121 (2020) 126159. https://doi.org/10.1016/j.eja.2020.126159.

[49] J. Han, L. Shi, Q. Yang, Z. Chen, J. Yu, Y. Zha, Rice yield estimation using a CNN-based image-driven data assimilation framework, F. Crop. Res. 288 (2022) 108693. https://doi.org/10.1016/j.fcr.2022.108693.

[50] D. Wallach, P.J. Thorburn, Estimating uncertainty in crop model predictions: Current situation and future prospects, Eur. J. Agron. 88 (2017) A1–A7. https://doi.org/10.1016/j.eja.2017.06.001.

[51] D. Wallach, P. Thorburn, S. Asseng, A.J. Challinor, F. Ewert, J.W. Jones, R. Rötter, A. Ruane, Estimating model prediction error: Should you treat predictions as fixed or random?, Environ. Model. Softw. 84 (2016) 529–539. https://doi.org/10.1016/j.envsoft.2016.07.010.

[52] S. Hu, L. Shi, K. Huang, Y. Zha, X. Hu, H. Ye, Q. Yang, Improvement of sugarcane crop simulation by SWAP-WOFOST model via data assimilation, F. Crop. Res. 232 (2019) 49–61. https://doi.org/10.1016/j.fcr.2018.12.009.

[53] I.M. Fattori Junior, M. dos Santos Vianna, F.R. Marin, Assimilating leaf area index data into a sugarcane process-based crop model for improving yield estimation, Eur. J. Agron. 136 (2022) 126501. https://doi.org/10.1016/j.eja.2022.126501.

1121 [54] A. Tewes, H. Hoffmann, G. Krauss, F. Schäfer, C. Kerkhoff, T. Gaiser, New Approaches
1122 for the Assimilation of LAI Measurements into a Crop Model Ensemble to Improve
1123 Wheat Biomass Estimations, Agronomy. 10 (2020) 446.
1124 https://doi.org/10.3390/agronomy10030446.
1125 [55] G. Nearing, S. Yatheendradas, W. Crow, X. Zhan, J. Liu, F. Chen, The Efficiency of
1126 Data Assimilation, Water Resour. Res. (2018). https://doi.org/10.1029/2017WR020991.
1127 [56] H. He, L. Lei, J.S. Whitaker, Z. Tan, Impacts of Assimilation Frequency on Ensemble
1128 Kalman Filter Data Assimilation and Imbalances, J. Adv. Model. Earth Syst. 12 (2020).
1129 https://doi.org/10.1029/2020MS002187.
1130 [57] M.P.G. de Oliveira, R.P. Amaro, H.B. Pescarini, L.H.A. Rodrigues, Tomato growth in
1131 production-like setting, (2021). https://doi.org/10.25824/redu/EP4NGO.
1132 [58] M. Oliveira, R. Amaro, H. Pescarini, L.H. Rodrigues, Dataset of tomato plants growth
1133 observations obtained from multiple sources in a production-like setting, Scielo Prepr.
1134 (2023) 2023–2035. https://doi.org/10.1590/SCIELOPREPRINTS.7667.
1135 [59] B.H.E. Vanthoor, P.H.B. de Visser, C. Stanghellini, E.J. van Henten, A methodology for
1136 model-based greenhouse design: Part 2, description and validation of a tomato yield
1137 model, Biosyst. Eng. 110 (2011) 378–395.
1138 https://doi.org/10.1016/j.biosystemseng.2011.08.005.
1139 [60] M. Oliveira, Leveraging high frequency data for improving crop growth estimates,
1140 (2023). https://doi.org/10.5281/zenodo.7632419.
1141 [61] M.P.G. de Oliveira, Leveraging data assimilation and monitoring data for improvement
1142 of crop growth estimates in protected environments, Universidade Estadual de
1143 Campinas, 2022. https://hdl.handle.net/20.500.12733/4792.
1144 [62] D. Wallach, D. Makowski, J.W. Jones, F. Brun, Regression Analysis, Frequentist, in: D.
1145 Wallach, D. Makowski, J.W. Jones, F. Brun (Eds.), Work. with Dyn. Crop Model., Third

1146 Edit, Elsevier, 2019: pp. 161–205. https://doi.org/10.1016/B978-0-12-811756-9.00005-
1147 8.
1148 [63] N. Zhang, X. Zhou, M. Kang, B.-G. Hu, E. Heuvelink, L.F.M. Marcelis, Machine
1149 learning versus crop growth models: an ally, not a rival, AoB Plants. 15 (2022) 1–7.
1150 https://doi.org/10.1093/aobpla/plac061.
1151 [64] S.H. van Delden, M. SharathKumar, M. Butturini, L.J.A. Graamans, E. Heuvelink, M.
1152 Kacira, E. Kaiser, R.S. Klamer, L. Klerkx, G. Kootstra, A. Loeber, R.E. Schouten, C.
1153 Stanghellini, W. van Ieperen, J.C. Verdonk, S. Vialet-Chabrand, E.J. Woltering, R. van
1154 de Zedde, Y. Zhang, L.F.M. Marcelis, Current status and future challenges in
1155 implementing and upscaling vertical farming systems, Nat. Food. 2 (2021) 944–956.
1156 https://doi.org/10.1038/s43016-021-00402-w.
1157 [65] M.G.J. Kallenberg, B. Maestrini, R. van Bree, P. Ravensbergen, C. Pylianidis, F. van
1158 Evert, I.N. Athanasiadis, Integrating processed-based models and machine learning for
1159 crop yield prediction, ArXiv. (2023). http://arxiv.org/abs/2307.13466.
1160 [66] D. Wallach, S. Buis, D.-M. Seserman, T. Palosuo, P. Thorburn, H. Mielenz, E. Justes,
1161 K.-C. Kersebaum, B. Dumont, M. Launay, S.J. Seidel, A calibration protocol for soil-
1162 crop models, BioRxiv. (2023). https://doi.org/10.1101/2023.10.26.564162.
1163 [67] D. Wallach, T. Palosuo, P. Thorburn, H. Mielenz, S. Buis, Z. Hochman, E. Gourdain, F.
1164 Andrianasolo, B. Dumont, R. Ferrise, T. Gaiser, C. Garcia, S. Gayler, M. Harrison, S.
1165 Hiremath, H. Horan, G. Hoogenboom, P.E. Jansson, Q. Jing, E. Justes, K.C. Kersebaum,
1166 M. Launay, E. Lewan, K. Liu, F. Mequanint, M. Moriondo, C. Nendel, G. Padovan, B.
1167 Qian, N. Schütze, D.M. Seserman, V. Shelia, A. Souissi, X. Specka, A.K. Srivastava, G.
1168 Trombi, T.K.D. Weber, L. Weihermüller, T. Wöhling, S.J. Seidel, Proposal and
1169 extensive test of a calibration protocol for crop phenology models, Agron. Sustain. Dev.
1170 43 (2023) 1–14. https://doi.org/10.1007/s13593-023-00900-0.

1171 [68] M. Lamboni, D. Makowski, S. Lehuger, B. Gabrielle, H. Monod, Multivariate global
1172 sensitivity analysis for dynamic crop models, F. Crop. Res. 113 (2009) 312–320.
1173 https://doi.org/10.1016/j.fcr.2009.06.007.
1174 [69] S.J.J. Seidel, T. Palosuo, P. Thorburn, D. Wallach, Towards improved calibration of crop
1175 models – Where are we now and where should we go?, Eur. J. Agron. 94 (2018) 25–35.
1176 https://doi.org/10.1016/j.eja.2018.01.006.
1177 [70] D. Wallach, T. Palosuo, P. Thorburn, Z. Hochman, E. Gourdain, F. Andrianasolo, S.
1178 Asseng, B. Basso, S. Buis, N. Crout, C. Dibari, B. Dumont, R. Ferrise, T. Gaiser, C.
1179 Garcia, S. Gayler, A. Ghahramani, S. Hiremath, S. Hoek, H. Horan, G. Hoogenboom,
1180 M. Huang, M. Jabloun, P.-E. Jansson, Q. Jing, E. Justes, K.C. Kersebaum, A.
1181 Klosterhalfen, M. Launay, E. Lewan, Q. Luo, B. Maestrini, H. Mielenz, M. Moriondo,
1182 H. Nariman Zadeh, G. Padovan, J.E. Olesen, A. Poyda, E. Priesack, J.W.M. Pullens, B.
1183 Qian, N. Schütze, V. Shelia, A. Souissi, X. Specka, A.K. Srivastava, T. Stella, T. Streck,
1184 G. Trombi, E. Wallor, J. Wang, T.K.D. Weber, L. Weihermüller, A. de Wit, T. Wöhling,
1185 L. Xiao, C. Zhao, Y. Zhu, S.J. Seidel, The chaos in calibrating crop models: Lessons
1186 learned from a multi-model calibration exercise, Environ. Model. Softw. 145 (2021)
1187 105206. https://doi.org/10.1016/j.envsoft.2021.105206.
1188
1189

Supplementary Material A

Table A1. Parameters values before and after calibration.

Parameter	Calibrated value	Non-calibrated value (Original set)
alpha_F	0.92	0.8
beta	0.5	0.169
delta	0.07	0.038
DFmax	0.033	0.08
E	0.75	0.75
K	0.58	0.58
K_F	8	5
m	0.1	0.1
N_b	15.5	16
N_FF	19.6	22
N_max	0.57	0.5
p_1	2	2
Q10	1.4	1.4
Qe	0.09	0.08
rm	0.016	0.016
sl_N1	0.025	0.025
sl_N2	0.05	0.05
sl_R	0.0032	0.0032
T_crit	28.4	24.4
tau1	0.07428	0.07428
tau2	0.05848	0.05848
tmaxPg	35	35
tmin_fr_gr	8.5	8.5
TQe	23	23
TSlop	20.6	20.6
V	0.135	0.135
V_max	8	8

This preprint was submitted under the following conditions:

- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.