Publication status: Preprint has been published in a journal as an article
DOI of the published article: https://doi.org/10.1016/j.inpa.2025.02.003

Leveraging data from plant monitoring into crop models

Monique Oliveira, Thais Zorzeto-Cesar, Romis Attux, Luiz Henrique Rodrigues

https://doi.org/10.1590/SciELOPreprints.7663

Submitted on: 2023-12-08
Posted on: 2023-12-08 (version 1)
(YYYY-MM-DD)

Leveraging data from plant monitoring into crop models

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

Researchers using crop models have been devising new roles for data and crop modeling based on the former's increased availability and the new techniques developed for the latter. From the various available techniques, modeling may be tackled by data-driven methods or through a process-based approach. Process-based or mechanistic models may nonetheless take advantage of real-time observations through data assimilation. And while this approach has been widely used for field crops, this is not the case for crops grown in protected environments. We present a case study of data assimilation in a protected environment, capturing tomato growth data from different sources. We updated growth estimates of the Reduced State TOMGRO model, by assimilating

observational data obtained through the continuous monitoring of plant mass and images captured by low-cost cameras, using the Unscented Kalman Filter and the Ensemble Kalman Filter. Since these techniques had not been used yet in the protected cultivation of tomatoes, it was necessary to develop the observation models as well, establishing the relationship between the observed variables and the ones estimated by the process-based model. The employed measurements, i.e., area of organs observed in pictures and plant-water mass, seemed suitable for tracking plant growth and for obtaining good approximations of the state variables estimated by the model. However, the quality of observations and of observation models was crucial for good performance of the assimilation techniques. As with other crops, it was not the case that assimilating one observation was useful for improving the value of others, including yield. We also observed that the assimilation performed better than calibrated models when there was a need to adjust the estimates to growth disturbances and that when filters lead to better yield estimates, continuous observations may not be required. There are then several steps and decisions that should be considered when bringing the idea from its application in field crops to protected environments and more studies are required to better determine the best approach.

Keywords

Crop model, data assimilation, greenhouse, proximal sensing

1. Introduction

Food production using as few resources as possible is the goal of researchers and growers all over the world. For protected environments, this would mean determining the exact requirements and trade-offs of energy demands on heating, cooling, carbon dioxide enrichment, and lighting, in

addition to irrigation and fertilization. To quantify those requirements and trade-offs, it is necessary to understand how the plant would respond to changes in these elements without sacrificing yield and quality, which is an application of simulation models.

Until now, crop simulation models have relied on crop growth data mainly for their development and calibration, steps prior to their use, as well as for performance assessment. However, the recent developments in sensors and computational power allow for more data regarding growth and development to be collected and processed, providing information in real-time. Keating and Thorburn (2018) posed that these developments further new roles for data in crop modeling, such as model-data fusion and inverse modeling, and new roles for remote and proximal sensing in model initialization and calibration.

Currently, real-time data obtained in protected environments has been explored by data-driven modeling approaches, such as in Gong et al. (2021). However, in their recent review of the status of vertical farming systems, van Delden et al. (2021) mentioned the downsides of relying solely on machine learning techniques and suggested sensor-informed artificial intelligence (AI) could be used to update self-learning dynamic growth prediction models that are partially process-based and partially data-driven, as a strategy to increase radiation use efficiency. Given the overall goal of reducing the use of resources in food production, the several paths available for its achievement should be explored. A different approach that could be explored to combine the outcomes of process-based and real-time data is called data assimilation.

Data assimilation is a widely used method in hydrology and meteorology which consists of combining observed values to the states estimated by models, taking into account the uncertainty that exists in model estimates and observations (Pellenq and Boulet, 2004). In the case of crop modeling, data assimilation has mostly been used with remote sensing images of field crops to update state variables related to crop canopy or soil properties, obtaining better estimates of yield,

leaf area index (LAI), and soil moisture (Dorigo et al., 2007; Jin et al., 2018; Luo et al., 2023). Recently, Luo et al. (2023) quantified applications of data assimilation in research using satellite images, pointing to the Ensemble Kalman Filter (EnKF) as the most explored method and to WOFOST, DSSAT, AquaCrop, and SAFY as popular crop models used in data assimilation for yield estimation studies.

For vegetables and other crops usually grown in greenhouses and other protected environments, there have been few studies and little is known about the use of monitoring data to update their simulations of growth, even though the idea has been adopted for environmental monitoring of greenhouses (van Mourik et al., 2019). In an application of data assimilation in a protected environment, Ruíz-García et al. (2014) worked with lettuces and the Nicolet model, but used data from destructive analyses. Differently from field crops, for which monitoring often depends on the frequency of satellite images, in protected environments, it is possible to obtain observations for data assimilation at greater temporal and spatial resolutions. Automatic plant-related measurements in protected environments have been applied to obtain different kinds of information. For example, growers monitoring plant growth through load cells had been registered in the Netherlands by de Koning and Bakker (1992) and has been suggested by studies such as Helmer et al. (2005) and Lee and Son (2019), and Chen et al. (2016) for plant factories. It has also been suggested for determining water demand (De Graaf et al., 2004). But mostly, images are easily explored to provide indirect estimates of growth (Liu et al., 2022; Zhang et al., 2023).

Therefore, while for field crops reviews are being published discussing which crops, variables, models, data sources and techniques are being used and their limitations (Huang et al., 2019; Jin et al., 2018; Luo et al., 2023), there have been few studies exploring this approach for coupling proximal monitoring data and horticultural crops. This means little has been discussed with regard to which type of data could be useful, which variables could be explored or how

models' estimates respond to assimilation. Our goal was to revisit the steps required to perform data assimilation, applying them in a simple greenhouse tomato growth example, and to discuss these aspects of the approach. We monitored growth of tomato plants, obtained the measurement functions required for assimilation, and, using the Reduced State Tomgro model (Jones et al., 1999), obtained estimates of tomato growth, which were updated by data assimilation techniques with data extracted from pictures of the plants and from a weighting system.

2. Material and methods

This study is based on the steps of crop monitoring, observation model building, and data assimilation. Figure 1 provides an overview of what a daily step in the assimilation process entails. Monitoring the environment provides the inputs for the crop model (Reduced State Tomgro), which is used, without being locally calibrated, to obtain an estimate of the state variables of the tomato plant. This estimate is converted by an observation model, which is a simple relationship between destructive and non-destructive observations of plants grown in the same environment, into a measurement analogous to that being obtained by the monitoring systems. Both the converted estimates and the continuously measured observations were used in the filters, along with the pre-determined uncertainties estimates, to obtain the updated model estimates. The updated estimates are then combined with those obtained by the calibrated model. Since the model is executed in a daily step, the updated state is used as an input for the following day.

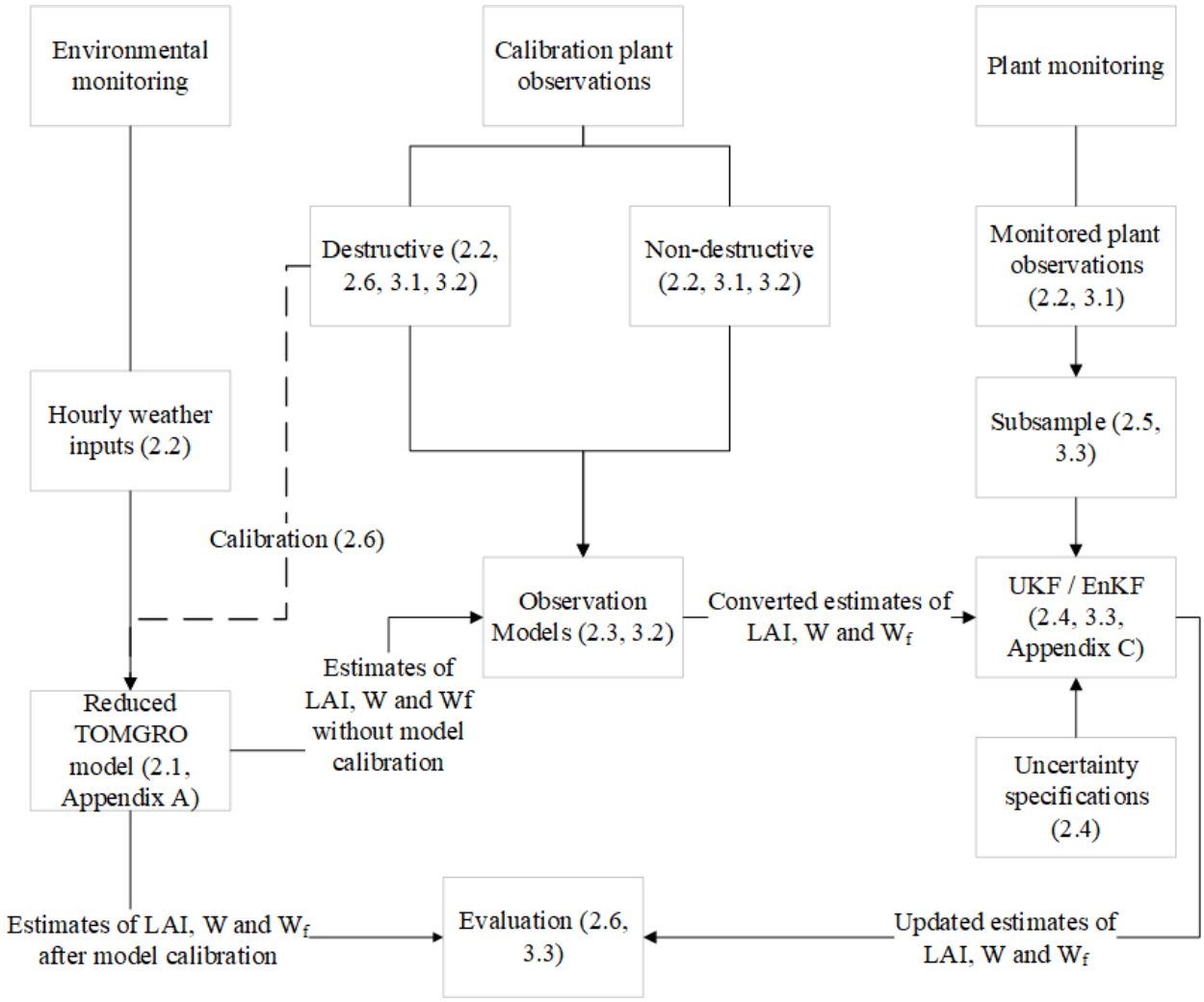


Figure 1. Flowchart of a daily step of data assimilation in this study. References in parentheses indicate the section of the manuscript in which the item is further described. The state variables assimilated are the leaf area index (LAI), aboveground dry biomass (W), and fruit dry biomass (W_f). The filter assimilation techniques used are the Unscented Kalman Filter (UKF) and the Ensemble Kalman Filter (EnKF).

2.1. Crop model

This study uses the Reduced State Tomgro (RT) model, which is a summary model derived from the TOMGRO model, aimed at being used in greenhouse control systems (Jones et al., 1999). The model quantifies the growth and development of the tomato plant when water and nutrients do not limit growth, based on hourly air temperature (T) and photosynthetically active solar radiation

data (PPFD). Its equations are presented in Appendix A. Briefly, the model has only five state variables: number of nodes (N), leaf area index (LAI), aboveground dry biomass (W), fruit dry biomass (W_f), and mature fruit dry biomass (W_m). Development is indicated by the number of nodes, and growth, by the other state variables. The leaf area index influences photosynthesis, which, along with respiration, determines the total carbohydrates available for growth of aboveground biomass and fruit biomass. The RT model, which has its variables and parameters further detailed in Jones et al. (1999), does not include root growth or irrigation, and this simplification may help in this first assimilation approach.

The RT model used in this study was implemented in Python, using Jones et al. (1991), Jones et al. (1999), and the spreadsheet provided by Dr. James W. Jones as sources. The difference equations were integrated by the Euler method. All code used in this study, including the model code, is available in Oliveira (2023).

2.2. Data sources

The experiments were conducted in research greenhouses at the School of Agricultural Engineering of the University of Campinas (22º 49' 06" S, 47º 03' 40" W, 635 m altitude). Three cycles of tomato growth were performed from 2019 to 2021. In each cycle, environmental and plant growth data were obtained. Plant growth data refers both to the plants that were continuously monitored, which provided inputs for assimilation, and to calibration plants, which provided inputs for model calibration and for the development of observation models. All data is available in Oliveira et al. (2021). Full experiment description, with extensive details regarding plant management and infrastructure, will be submitted separately as a data paper and can currently be seen in Oliveira (2022).

For environmental monitoring, measurements of air temperature (T) were obtained by SHT75 transducers protected by porous capsules which, by their turn, were protected by tubes of polyvinylchloride tubes coated with aluminum foil and which included downstream fans. The sensors were installed in a hardware platform for wireless sensor networks (Radiuino BE900), with daily backup. For measurements of photosynthetic photon flux density (PPFD), we used quantum sensors Licor LI-190SA with a datalogger Licor LI-1400. Each sensor node was positioned adjacent to the monitored plants.

Plant growth was characterized by destructive analyses (calibration plants) and by non-destructive observations (calibration and monitored plants). Destructive analyses were performed to determine the biomass of plant organs and the leaf area. In intervals of one to three weeks, three plants were removed from the greenhouse, and the leaves, stem, and green and mature fruits were separated. Leaves were digitized with a scanner and the digitization included a reference of known dimensions, which was used for calculation of the area occupied by leaves in the image. Plant material was then weighed before and after drying, in a forced-air circulation oven (60 °C) for four days or until constant mass was reached, so both fresh and dry mass data were determined.

For both calibration and monitored plants, non-destructive data includes pictures taken from side and top-down views. For monitored plants, they were obtained by fixed Raspberry Pi Camera Modules v2, connected to Raspberry Pi Zero, one positioned above the plant and one in an adjacent row (Appendix B, Figure B1 and Figure B2). For calibration plants, non-destructive observations were obtained prior to removal for destructive analyses. Before removal, those plants were photographed laterally and from a top-down view with a smartphone camera (Figure 2 and Appendix B, Figure B3). Pictures included references of known size.

Labeling of the plant organs in all images was done manually, using the software GIMP. Annotations were made only in areas in which there was certainty the organ corresponded to the

correct plant, i.e., if there was uncertainty or occlusion, the area was not marked. The organs were colored differently (Figure 2). The size references were used to calculate the area corresponding to plant organs in the images. All image processing — of plant photos and of digitized leaves — was performed by scripts in python, also included in Oliveira (2023).

Monitored plants were also subject to continuous — every minute — weight monitoring. To do so, they were placed in a support isolated from contact with other surfaces, which was hanged in force transducers HBM S2M with nominal force of 10 N (0.02 % accuracy). Data was stored in a data logger PMX WGX002. As fluctuations in weight caused by changes in water content in the substrate were not determined, we refer to this measurement as plant-water system mass.

Figure 2. Example of non-destructive observation of lateral leaf area and fruit area of a calibration plant. Leaves are marked in green, mature fruits are marked in red, and non-mature fruits are marked in yellow. The A4 sheet used as a reference is marked in cyan.

Therefore, plant monitoring led to three types of observations used in assimilation: top-down images, lateral images and plant-water system mass. We used images captured every other day as the full set of image observations. Data from the weighting system was averaged in the hour and averages from between 4 am and 5 am, before sunrise in all cycles, were used as the observation

for each day, providing the plant-water system mass dataset. The choice allowed for the lowest interference of water content in the measurement.

The three cycles presented different developments, which are explored in this study: the first may be characterized by low irrigation, subjecting plants to water deficit (Cycle 1), the second, by an excess of nitrogen fertilization and an attack of tomato rust mite by the end of the cycle (Cycle 2), and the third cycle was conducted closer to full water and fertilization (Cycle 3).

2.3. Observation models

In order to establish a relationship that would allow for comparing the outcomes of the model and the observed values, observation models or measurement functions are required. In this study, observation models were created from data obtained from the calibration plants, from which destructive and non-destructive data is available. Our modeling strategy focused on obtaining simple empirical relationships, and the generalized least squares method was used to account for the heteroscedasticity and correlation between residues. To avoid data leakage, we made the choice of obtaining different models for each cycle. Therefore, despite the different growth conditions, data from the cycle was not used to obtain the relationship that would be used for the conversion of measurements in that cycle e.g. the observation model of Cycle 1 was developed only with observations from Cycles 2 and 3. All observations obtained per plant were multiplied by plant density to make units compatible with the state variables in the model.

We used areas identified from lateral and top images as proxies of the leaf area index, fruit area extracted from lateral images to infer fruit dry mass, and weight as determined by the weighting system to infer aboveground biomass. In this last case, we had to make an assumption

as to which percentage of the system corresponded to the roots, and we did so based on the developmental stage (20% before fruit appearance and 10% after).

2.4. Data assimilation

Two assimilation techniques were used: Ensemble Kalman Filter (EnKF) and Unscented Kalman Filter (UKF). The first was pointed by Luo et al. (2023) among the most used with satellite images while the latter, was used by Torres-Monsivais et al. (2017), in a problem similar to ours. Both of these methods belong to the class of state estimators based on Kalman filtering. The problem of state estimation, in simple terms, corresponds to the task of determining the values of a set of "hidden" state variables based on the available observations and on knowledge about the system dynamic model and certain signal properties. Ideally, this must be done using information provided by several samples for the sake of improving the attainable precision (Haykin, 2013).

The properties of the system of interest and of the relevant signals are essential to determine the complexity of the problem. The classical Kalman Filter (KF) is optimal under assumptions that are typically summarized as linearity and Gaussianity. However, for more general system and signal/noise models, the KF is not optimal and may not even lead to satisfactory performance (Arulampalam et al., 2002).

The Ensemble Kalman Filter, proposed in 1994 (Evensen, 1994), adopts an ensemble-based estimation process that is very well-suited for high dimensional systems. It was initially applied in practical applications related to geophysics but became a widespread filtering approach in other large-scale scenarios (e.g. meteorological analysis, marine ecosystem modeling, etc.) (Evensen, 2003). Although the method relies on the Gaussian assumption, it is employed also in nonlinear contexts (Mandel, 2006).

The Unscented Kalman Filter is an approach aimed at dealing with the problem of recursive nonlinear estimation. Instead of resorting to a linearization process, it represents the state distribution by a carefully chosen set of points, which are able to capture its true mean, as well as the posterior mean and covariance, when propagated through the true non-linear system (Julier and Uhlmann, 1997; Wan and Van Der Merwe, 2000), leading to a more precise estimation when the state equations are pronouncedly nonlinear.

In this study, three of the state variables simulated by the model were assimilated: aboveground dry biomass, leaf area index, and fruit dry biomass. Uncertainty in the initial states was determined as the variance of the samples analyzed on the transplanting day. For the EnKF, uncertainty in the process was ascribed to a model parameter, depending on the state variable being assimilated. An added perturbation to the parameter was sampled from a normal distribution with zero mean and a standard deviation of 10% of its value. Parameters were chosen following a sensitivity analysis (Oliveira, 2022): for the leaf area index, the maximum leaf area expansion per node was perturbed, for aboveground biomass, the leaf quantum efficiency, and for fruit biomass, the maximum partitioning of new growth to fruit. The number of elements in the ensemble was tested and defined as 100. Uncertainty in the model for the UKF was determined as the squared absolute error, calculated by the difference between observations of the states in the calibration samples in the cycle and the simulated value of the uncalibrated model. We used the implementation of the library filterpy[1] with modifications to account for the perturbation of model parameters in the EnKF. More details on the theoretical aspects of the filters implemented in the library are presented in Appendix C.

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

Variance in the measurements was determined as the variance of the non-destructive observations of the calibration samples in the growth cycle. While we are aware that this corresponds to partial data leakage, we believe this was the best way to provide an estimate for these filter parameters, as the different characteristics of each cycle would lead to unreasonable estimates of uncertainty.

2.5. Frequency of assimilation

We subsampled the full datasets of monitoring observations to assess the effect of the frequency of assimilation. Subsampling corresponded to using 50% and 10% of the data available in the cycle. To avoid results being biased by sampling, the procedure was performed 20 times. One of the repetitions was sampled in regular intervals through the cycle, and the others were randomly sampled, so that the days from which observations were retrieved varied in each sampling.

2.6. Evaluation

To conclude the common steps of data assimilation in a crop model, we compared the results from data assimilation in the model without calibration to the results of the calibrated model (open loop, OL, calibrated). We also present the non-calibrated simulations (open loop not calibrated) as a reference. To avoid errors caused by internal variability of the greenhouse environment, simulations for each monitored plant were performed individually with the data provided by each sensor node associated with them.

Calibrated runs used data from the destructive analyses of the respective cycle, in order to fully represent the growth conditions and non-calibrated runs used parameters' values from the

calibration performed by the developers of the model for their Gainesville data. Calibration followed the approach of minimizing the relative squared error using a global optimization algorithm.

In addition to the comparison of the Open Loop Calibrated simulation and Updated Non-Calibrated simulation, we compared the Open Loop Calibrated of Cycles 1 and 2 to the assimilation using model calibration of Cycle 3, i.e., under full fertilization and irrigation, as a manner of incorporating particularities of growth in tropical conditions. This would be equivalent to updating simulations by a model calibrated under appropriate conditions with observations from real-world problematic observations. We calculated the root mean squared error through the cycle, using samples from destructive analysis, comparing non-calibrated, calibrated, and filtered series.

3. Results and discussion

3.1. Plant monitoring

While several studies have estimated biomass from images, we had to ensure that data provided by continuous monitoring was representative of growth, and that these observations were equivalent to the observations obtained by the other methods, destructive or not, used in the calibration plants. Both aspects were suggested by the results (Figure 3). Since the measurements obtained by photographs from the same angle but different sources were comparable, they allowed for the approach of using smartphone pictures of calibration plants for obtaining observation models that represent data from the fixed cameras. Similarly, measurements of growth from monitored plants obtained through the weighting system were similar to the estimated fresh weight of the whole plant in calibration plants.

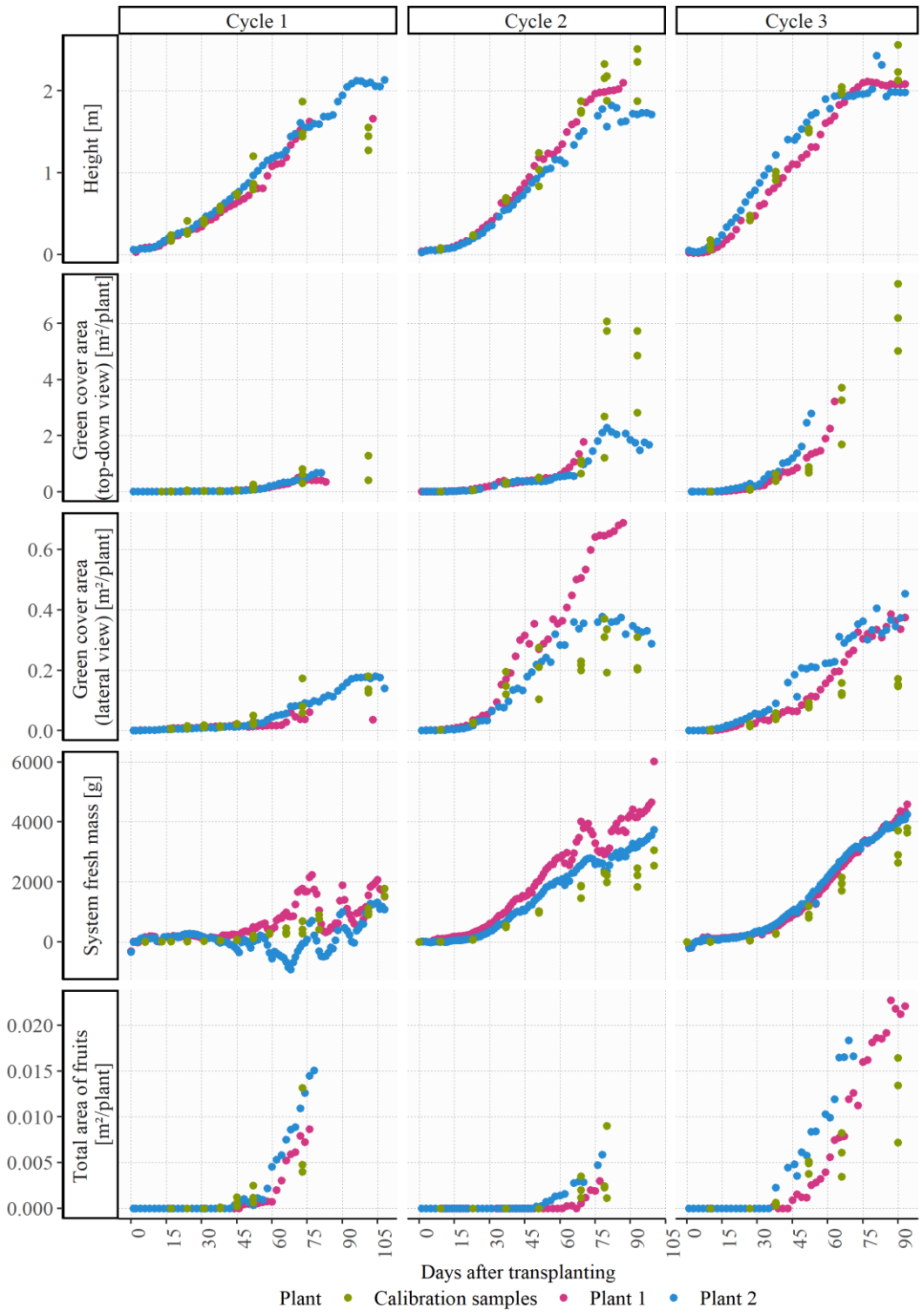


Figure 3. Time series of observations for Cycles 1 to 3. from monitored plants, i.e. Plants 1 and 2, and values for the same variables from the calibration plants' samples. For the calibration samples associated with the system fresh mass, the observations refer to the aboveground fresh mass of calibration plants, except for the last observations, which correspond to destructive analyses of the monitored plants.

Areas corresponding to leaves, i.e. green cover area on lateral and top-down views, and total area of fruits extracted from the images obtained with smartphones, for calibration plants, usually followed the same trends and magnitudes of the ones extracted from images obtained by the fixed cameras, for monitored plants (Figure 3). Discontinuities on these curves are often explained by changes in lighting and by occlusion, which is a downside of this type of measurement. When visible areas, i.e. the areas annotated in pictures, of the lateral view of monitored plants were larger than visible areas in pictures from calibration plants, this effect likely can be attributed to occlusion in calibration plants. Furthermore, because of the weighting system, monitored plants were slightly dislocated from the planting line, allowing for larger visible area in lateral images, while visibility of calibration plants was more affected by adjacent plants, an effect that is more noticeable in Cycle 3. Conversely, visible areas of calibration plants in Cycles 1 and 3 are equivalent, despite maximum leaf area per plant having reached an average of 0.44 m^2 leaves/plant in Cycle 1 and in Cycle 3, 1.91 m^2 leaves/plant, i.e., similar visible areas representing quite different values of leaf area index. This was likely a consequence of fewer leaves reducing the complexity of annotations in the environment of Cycle 1, which led to larger visible area, affecting observation models, as will be further discussed.

Curves of total area of fruits in Cycles 1 and 2 are interrupted before the end of the cycle because as plants were harvested, observations did not correspond to total fruit mass any longer and were not compatible with the principle of accumulated biomass used in the growth model. To fully account for this effect, the area of each truss would have to be identified and summed to measurements obtained after harvest.

In Figure 3, height was also included with two purposes: as a reference of information extracted from pictures in a comprehensible unit, but also to show how monitored plants having their growth interrupted earlier than plants used in calibration — when they reached the height of

the fixed cameras — is particularly noticeable in Cycle 3, and the measurement remains constant by the end of the cycle. In the green cover area identified from the top-down view, interruption often refers to this moment when plants reached the camera and occupied all visible area, so that these observations were no longer informative. The very low values observed in Cycle 1 for the top-down view may be connected not only to lower leaf area, but also to a slight dislocation of the camera, so that it did not fully capture the plant. As for the system fresh mass, one can observe how Cycle 1 did not present a regular mass growth trend, and variability mostly corresponded to water fluctuation in the system. Negative values correspond to measurements lower than the initial state of the weighting system, which only included the support and moist substrate, indicating loss of water. In the second and third cycles, these fluctuations are less common and there is a clear trend of biomass increase, suggesting the validity of the approach of using hourly average before sunrise to represent the weight of the system. In those two cycles, we can observe more clearly that values from monitored plants are larger than for calibrated samples, since roots are also included in the system.

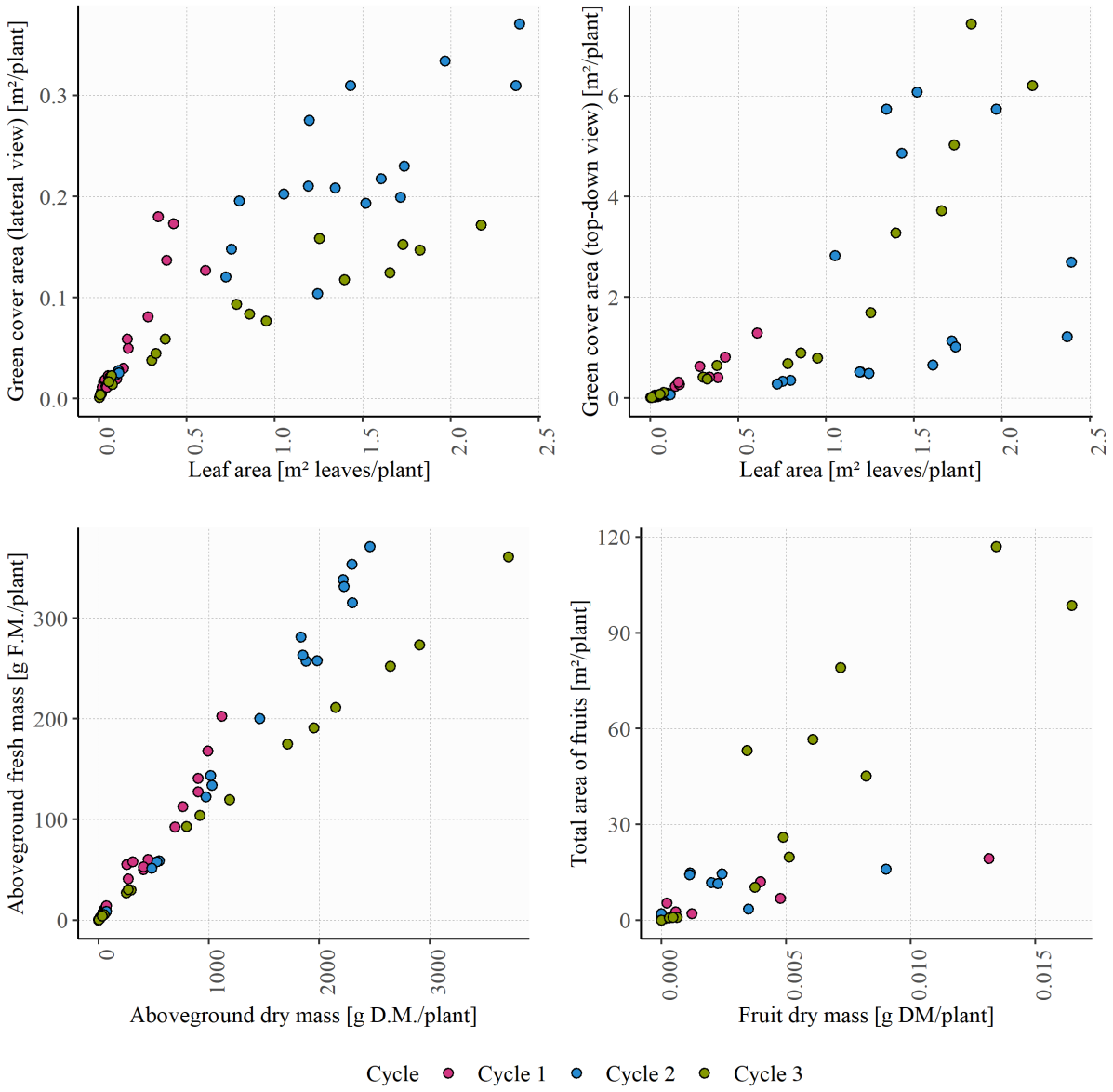
3.2. Observation models

In the case of assimilation for large areas using remote sensing images as the source of observations, several products are already available. This step, however, was still needed in this study, as there is no established relationship between the non-destructive observations used and the state variables (LAI, W and W_f) to be updated.

Although linear relationships may be observed in the scatterplots of the relationships that gave basis to the observation models (Figure 4), particularities of each cycle are noticeable, such as larger fruit biomass in Cycle 3 and lower leaf area in Cycle 1. For Cycle 2, excessive nitrogen

355 fertilization at the beginning of the cycle, followed by mite attack after the beginning of the fruit
356 stage, led to large leaf areas but low fruit mass.

357 As mentioned for the indirect observations, as plants reach the camera, one can no longer
358 distinguish leaf area from images taken from the top-down view. In this sense, the relationship
359 becomes non-linear, and images from the lateral view represent the leaf area better than from top-
360 down. One can also see that in all cases of observations extracted from images, particularly for
361 fruits, observed values are concentrated in the lower range, making it harder to obtain good
362 relationships for larger values. Besides the length of the cycles, lower values may also have been
363 caused by our choice of not compensating for occlusion and only including areas in which there
364 was certainty the area corresponded to the plant. These choices did not impair our attainment of
365 reasonably good linear relationships between said areas and their corresponding observed state
366 within a cycle. For aboveground biomass, although the linear relationship is very visible,
367 uncertainty relates to the other aspect of conversion i.e. the fraction of system biomass ascribed to
368 aboveground biomass, which could not be measured and had to be inferred.



369

370 Figure 4. Scatterplots of relationships between observations of plants used in the destructive
371 analyses, used for the development of observation models in each growth cycle. In the Y-axis,
372 observations that refer to an area were extracted from images and aboveground fresh mass was
373 obtained by weighting plants before drying.
374

375 We noted in the previous section that different leaf area per plant in Cycle 1 and Cycle 3 led
376 to similar visible areas in pictures. This is noticeable in Figure 4, through the similar ranges in the
377 y-axis despite the different ranges in the x-axis, which also shows that the model obtained with

data from Cycles 2 and 3 will overestimate leaf area index for Cycle 1, by ascribing larger leaf area values to the range of green cover observations of this cycle.

However, even as the differences between growth cycles existed, the individual models obtained presented similar slope coefficients (Table 1) and, except for the conversion of fruit biomass (W_f), since these coefficients are at least one order of magnitude larger than the intercept, they dominate the results. Visibly, the most discrepant model corresponds to fruit biomass (W_f), with the largest coefficient of variation for the slope coefficient (62% for slopes 1.19×10^{-4}, 1.31×10^{-4} and 3.34×10^{-4}), particularly in Cycle 3, which has a much lower intercept (1.36×10^{-4}). The difference affected assimilation, as will be discussed, but the other results suggest it should be possible to generate simple models that could be generalized for several conditions.

388 Table 1. Equations derived from calibration data for observation models, i.e., conversion of state
389 variables (W_f, W and LAI) into the equivalent observation.

Growth Cycle	Equation	State variable	Observed variable
Cycle 1	$wf\_lat = 1.19 \times 10^{-4} \times W_f + 1.97 \times 10^{-3}$	W_f	Area Wf
Cycle 2	$wf\_lat = 1.31 \times 10^{-4} \times W_f + 1.77 \times 10^{-3}$		
Cycle 3	$wf\_lat = 3.34 \times 10^{-4} \times W_f + 1.36 \times 10^{-4}$		
Cycle 1	$w\_fm = 8.32 \times W + 4.74 \times 10^{-1}$	W	W_fm
Cycle 2	$w\_fm = 7.81 \times W - 5.26 \times 10^{-1}$		
Cycle 3	$w\_fm = 6.90 \times W + 6.79 \times 10^{-1}$		
Cycle 1	$lai\_lat = 1.42 \times 10^{-1} \times LAI + 3.44 \times 10^{-3}$	LAI	GC Lat
Cycle 2	$lai\_lat = 2.54 \times 10^{-1} \times LAI + 2.97 \times 10^{-3}$		
Cycle 3	$lai\_lat = 2.26 \times 10^{-1} \times LAI + 4.75 \times 10^{-3}$		
Cycle 1	$lai\_td = 9.87 \times 10^{-1} \times LAI + 5.74 \times 10^{-4}$		GC TD
Cycle 2	$lai\_td = 1.30 \times LAI - 1.28 \times 10^{-3}$		
Cycle 3	$lai\_td = 1.16 \times LAI - 4.81 \times 10^{-4}$		

390 W_f: dry mass of fruits [g m^{-2}], W: aboveground dry biomass [g m^{-2}], LAI: leaf area index [m^2 m^{-2}],
391 Area Wf: area of fruits on images [m^2 m^{-2}], W_fm: aboveground fresh mass [m^2 m^{-2}], GC Lat: area
392 of leaves identified on images from lateral view [m^2 m^{-2}], GC TD: area of leaves identified on
393 images from the top-down view [m^2 m^{-2}].

Correlations between state variables and indirect observations are compatible with what is visible from the scatterplots i.e. reasonably linear relationships within cycles in most cases, but error metrics point to large uncertainties for some variables, even within the training set (Table 2).

While an error of 0.40 g FM m^{-2} may be considered very small when compared to the mass of the weighting system, 0.61 m^2 m^{-2} could be considered very large for the area visible on images and is likely caused by the different behaviors from the three growth cycles. It is not uncommon for remote sensing LAI products used in assimilation to reach errors larger than 0.5 m^2 m^{-2}. Fang et al. (2019) report validation RMSE errors from moderate and high-resolution leaf area index products for crops ranging from 0.2 to 0.8 m^2 m^{-2}. However, Mean Absolute Percentage Errors suggest that given the error in the unseen cycle, using these models to convert state variables in the assimilation may lead to lower efficiency of the process. In the case of mature fruit biomass, which was not explored here as there were not enough observations to develop the models, obtaining good relationships should be a lot easier, as occlusion by leaves is minimized by pruning practices.

Table 2. Standard error (SE), mean absolute percentage error (MAPE) and coefficient of determination (R^2) from each observation model for data from each cycle. SE from a cycle is reported from training in the other cycles and MAPE is reported from validation in the same cycle.

Assimilated variable		LAI	LAI	W	Wf
Observed Variable		GC TD [m^2 m^{-2}]	GC Lat [m^2 m^{-2}]	W_fm [g FM m^{-2}]	Area Wf [m^2 m^{-2}]
Cycle 1	SE (training)	0.988	0.311	0.395	0.062
Cycle 2		0.597	0.613	0.339	0.084
Cycle 3		0.921	0.436	0.257	0.252
Cycle 1	R^2 (training)	0.64	0.73	0.94	0.81
Cycle 2		0.86	0.82	0.98	0.77
Cycle 3		0.44	0.87	0.99	0.36
Cycle 1	MAPE (validation)	30 %	41 %	34 %	373 %
Cycle 2		122 %	55 %	29 %	82 %
Cycle 3		26 %	113 %	58 %	122 %

*Observations as: GC Lat: green cover (lateral view), GC TD: green cover (top-down view), W_fm fresh mass from destructive analyses, Area Wf: total area of fruits.

3.3. Data assimilation

Overall, assimilation of aboveground biomass led to the largest improvements in Cycles 2 and 3 (Table 3). Without the disturbances that happened in Cycle 1, this was the most precise measurement, as suggested by sensor accuracy and by observed trends of monitored and calibration plants (Figure 3). Additionally, the observation model for fresh and dry aboveground biomass was not affected by the fluctuation in water content of the substrate, which did not compromise the conversion (Table 2). The combination of high-quality observations and observation model allowed for the UKF to reach the same performance as the calibrated model.

Table 3. Average root mean squared error (RMSE) for state variables assimilated in the evaluations.

State variable	Filter	Assim. State	Obs. Variable*	Cycle 1	Cycle 2	Cycle 3
LAI [m^2 m^{-2}]	None – Calib.	-	-	0.08	0.53	0.31
	None – Not Calib.	-	-	0.17	1.17	1.76
	EnKF	LAI	GC TD	**0.07**	**1.04**	**0.78**
	UKF			**0.09**	**0.98**	**0.67**
	EnKF		GC Lat	**0.08**	**0.70**	1.83
	UKF			**0.10**	**0.70**	1.83
W [g m^{-2}]	None – Calib.		-	42.4	29.8	124.8
	None – Not Calib.	-	-	30.8	148.8	275.4
	EnKF	W	W_fm_full	61.6	**59.0**	**142.8**
	UKF			68.1	**66.5**	**124.4**
Wf [g m^{-2}]	None – Calib.	-	-	15.9	82.7	12.4
	None – Not Calib.	-	-	25.8	34.8	90.1
	EnKF	Wf	Area Wf	**25.2**	**30.9**	**88.7**
	UKF			**23.4**	**16.1**	**86.2**

*Observations as: GC TD: green cover (top-down view), GC Lat: green cover (lateral view), W_fm_full: weighting system, Area Wf: total area of fruits. Bold numbers refer to root mean squared errors lower than the larger RMSE between the non-calibrated and calibrated error.

Another example in which the observation model had lower validation error (Table 2) and in which observations tracked quite well the calibration plants (Figure 3) comes from using the pictures from the top-down angle to estimate the leaf area index in Cycle 3 (Figure 5). Both techniques were able to increase the estimates, with the UKF presenting slightly better results. The absence of images led the final half of the growth cycle to be simulated by the non-calibrated model, without the increase in magnitude that the observations provided.

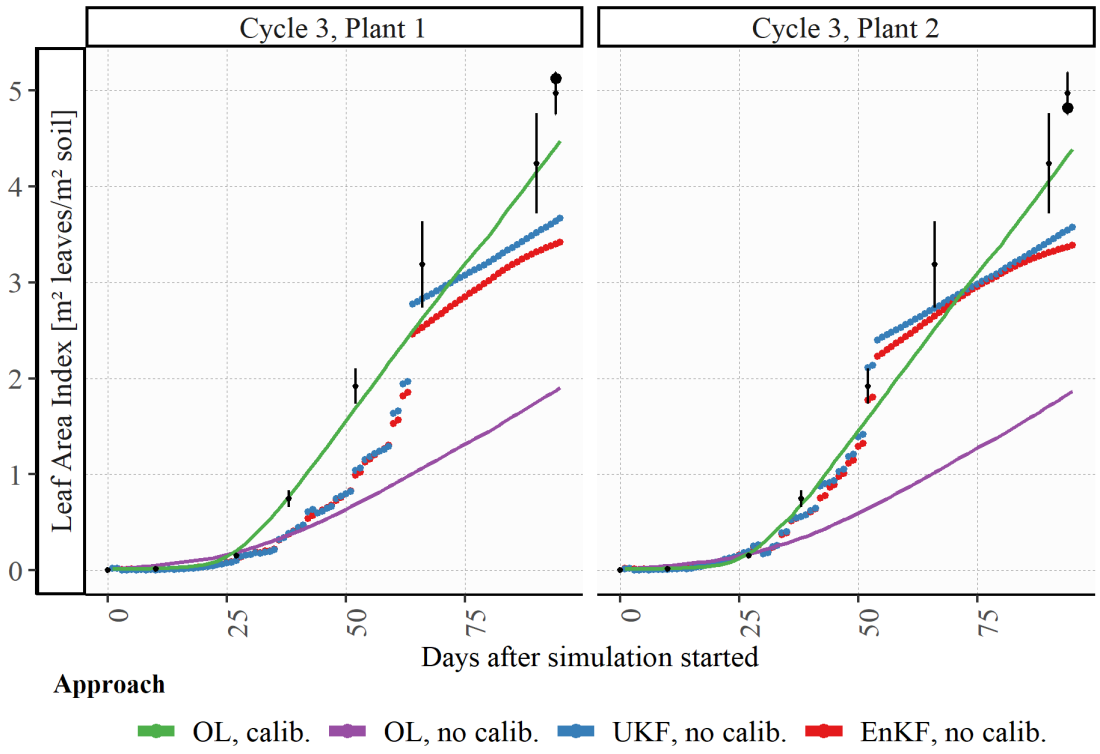


Figure 5. Assimilation of leaf area observations. Growth curves for each monitored plant without assimilation (OL, calib. And OL, no calib.) and with the different methods used for assimilation (UKF and EnKF) of the indirect measurement of leaf area index corresponding to the area of leaves in images obtained from a top-down angle. Dots refer to the average equivalent observation determined by destructive measurements, and the bar represents the associated standard deviation. The final equivalent value for the monitored plant is represented by a larger dot. The equivalent observations were converted from the original value to account for plant density.

To comment on the other results, we have to take into account the other elements of our study. By proposing the use of new observations and their respective observation models, with a previously not used model, in three different conditions, we have to dissect each aspect.

3.3.1. Effect of growth disturbances

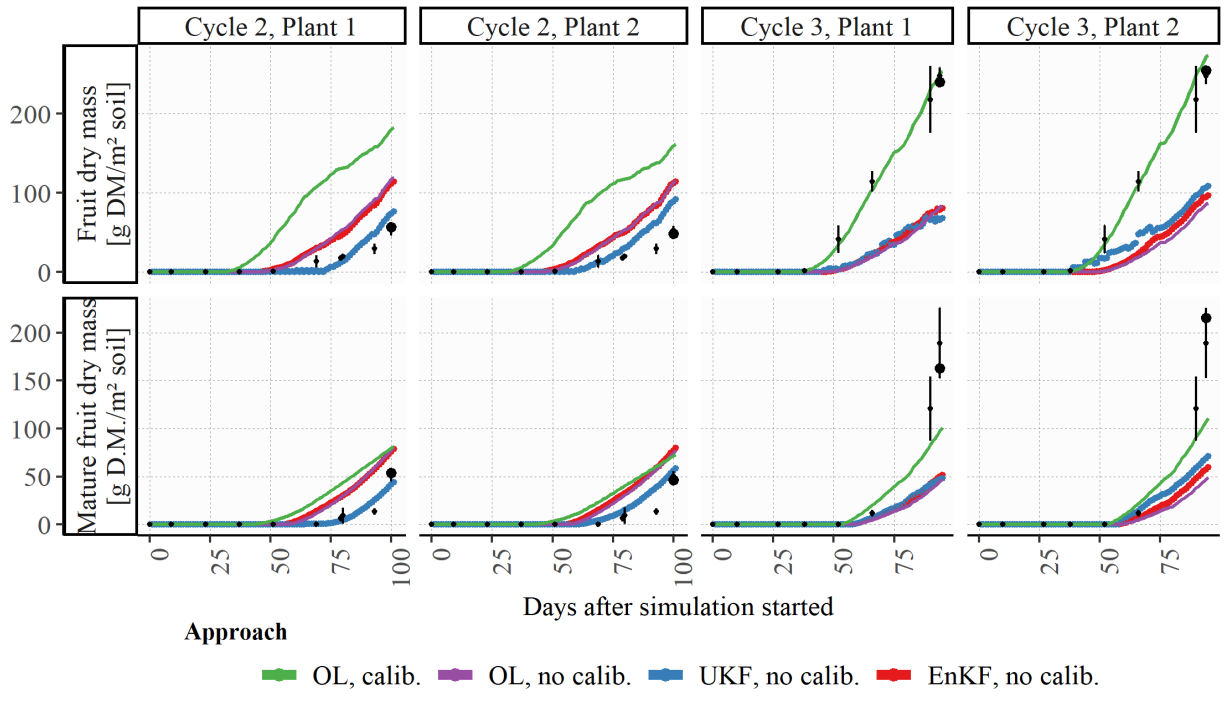
Assimilation studies in crop modeling often focus on using additional information to account for spatial variability, planting dates, or other uncertainties that impair model best estimates (Hu et al., 2019; Jin et al., 2018). And as crop models often simulate potential — sometimes limited — growth, actual growth requires this additional information to be included in the estimates. From the three growth cycles, one represented potential growth while the others were at least limited. As such, we dealt with multiple scenarios that serve as examples of issues in assimilation.

For instance, while locally calibrating the model usually led to the best results (Table 3), this was not the case when the state variable was impaired by growth conditions. Therefore, there were two references to which the results of assimilation could be compared: the calibrated model, which should give the best performance in potential conditions, and the non-calibrated model, which provided the estimates for assimilation. In our study, using data assimilation with the non-calibrated model was often able to improve the outcomes in comparison with the poorest performance from the standalone model (Table 3). No assimilation technique from both assessed was consistently better either across variables or across growth cycles.

As previously mentioned, although calibration is expected to improve model performance, this was not observed in some cases, since growth not always corresponded to the situation for which the model was developed, i.e. potential growth. In Cycle 1, in which total biomass was impaired by faulty irrigation, the optimization used in calibration could not determine parameters that would generate compatible estimates between all variables. Similarly, in Cycle 2, excessive

462 nitrogen led to much lower fruit production, and this effect was not properly captured by the
463 parameters selected.

464 Assimilation results in both cycles in which growth was disturbed also depended on the
465 quality of observations. In Cycle 1, in which the system biomass, i.e., the observation assimilated,
466 was affected by water content in the substrate, and measurements didn't correspond strictly to the
467 expected quantity — plant biomass — poor estimates of aboveground biomass were obtained,
468 leading to the best results coming from the non-calibrated simulation (Table 3). In Cycle 2, on the
469 other hand, assimilation of fruit images led to the adjustment of the estimates to the lower values
470 that actually happened (Figure 6, leftmost panels).



471

472 Figure 6. Assimilation of fruit area observations. Growth curves for each monitored plant without
473 assimilation (OL, calib. And OL, no calib.) and with the different methods used for assimilation
474 (UKF and EnKF) of the indirect measurement of fruit dry biomass corresponding to the area of
475 fruits in images. Dots refer to the average equivalent observation determined by destructive
476 measurements, and the bar represents the associated standard deviation. The final equivalent value
477 for the monitored plant is represented by a larger dot. The equivalent observations were converted
478 from the original value to account for plant density.

3.3.2. Observation models

In Cycle 3, fruit area observations were apparently barely used, and the assimilated curve closely resembles the simulation without calibration (Figure 6, rightmost panels). This is likely caused by bias in the observation model. While Figure 3 showed that monitored plants in Cycle 3 had the largest areas of fruits by the end of the growth cycle, the scatterplot from Figure 4 shows how a model obtained with data from the other two cycles would likely underestimate the fruit mass observation. Therefore, when the model-estimated fruit biomass value was converted to the equivalent observation, the difference in magnitude between what was estimated by the model and the observation could not be captured in the residual calculation.

It is likely that the high errors of some observation models affected the ability of the techniques to extract information from the observations. These errors lead to poorer estimates of the residuals and the gain, shifting how much the filter should rely on observations. Our observation models were limited, particularly for fruit area, as there were fewer observations, but obtaining good relationships seemed feasible, as suggested by the overall high correlations (Table 2). Additionally, we chose simple approaches both for monitoring and for the observation models, i.e. only manual annotation of images and linear relationships for the observation models, but the use of deep learning techniques could enhance the results obtained. Han et al. (2022) recently used these techniques to extract observations and their uncertainties using smartphone pictures in an assimilation experiment. In greenhouses, for monitoring, deep learning could be used for tracking a fruit truss (Y. Ge et al., 2022), to account for their influence in simulated biomass after harvesting. They are also useful to separate foreground and background leaves (Baar et al., 2022). Fruit classification and automatic segmentation (Fonteijn et al., 2021) would allow for monitoring more plants and making the assimilation results more robust and based on more observations, which also impacts the observation models. In the case of observation models, deep learning methods could

503 extract more features from the observations and better characterize them. Moon et al. (2022)
504 explored deep learning techniques to model fresh weight and leaf area of sweet peppers. While the
505 error in fresh weight error measurements was similar to the ones obtained in this study, leaf area
506 ones were much lower. As the few previous assimilation studies in protected environments did not
507 go through this step of requiring observation models, their outcomes are not comparable (Ruíz-
508 García et al., 2014; Torres-Monsivais et al., 2017).
509
510 **3.3.3. Yield and assimilated variables**
511 Since the update of a variable should impact their own estimates as well as the ones from
512 other variables, we expected the assimilation of several different state variables to improve yield
513 outcomes in different ways. For instance, changes in leaf area index should affect aboveground
514 biomass and fruit biomass, since it affects photosynthesis. There are, however, several mechanisms
515 that govern other state variables, such as respiration and fruit abortion, so that it is not necessarily
516 the case that, even if there are improvements in the assimilated state, they are going to be
517 propagated to other variables.
518 When analyzing the results focusing on yield, often assimilation of either leaf area or
519 aboveground biomass did not improve the estimates (Table 4). While counterintuitive, this has
520 been observed in other studies, for other crops, such as wheat (Nearing et al., 2012), cotton, and
521 potato (Linker and Ioslovich, 2017). The absence of improvement in yield does not mean the
522 approach is not useful. Even though in this study the model used did not include irrigation or other
523 aspects that are relevant to the management of tomato growth, such as fertilization or supplemental
524 lighting, it could be the case that improvement in estimates of the state variables themselves could
525 be informative for other model states and possibly, management decisions.

There is no established method for determining previously which variables should be updated so that model performance is improved when estimating yield. Recently, Orlova and Linker (2023) used an interesting approach, in which sensitivity analysis is performed before an update step of assimilation, and a particle filter estimates relevant models' parameters. In this manner, they address the problem of parameter importance being variable through growth, which is analogous to the issue of how one state variable impacts the other. A version of this approach could be explored to better understand up to what point yield estimates could benefit from an update given its representation encoded in a model.

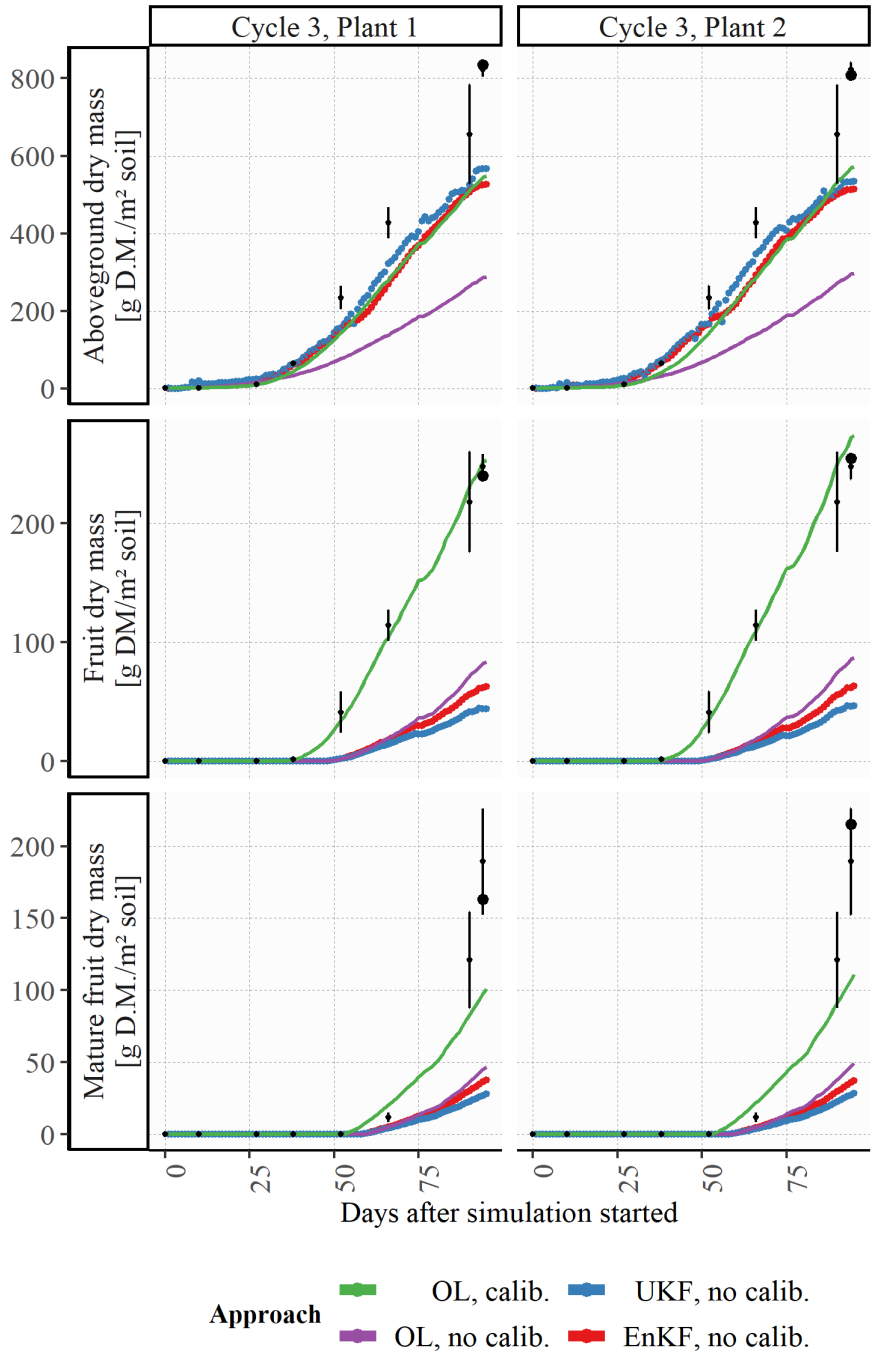
534 Table 4. Root mean squared error [g m^{-2}] from estimates of yield updated with data from different
535 sources with the Unscented Kalman Filter and the Ensemble Kalman Filter, and from the model
536 without assimilation with and without calibration.

Filter	Assim. state	Obs. variable*	Cycle 1		Cycle 2		Cycle 3	
			Plant 1	Plant 2	Plant 1	Plant 2	Plant 1	Plant 2
None – Calib.	-	-	5.05	10.1	29.1	25.3	36.0	31.5
None – Not Calib.	-	-	20.5	18.0	22.4	21.7	62.2	61.1
EnKF	LAI	GC TD	22.5	19.8	**27.5**	**19.6**	**47.1**	**40.1**
UKF			22.9	18.8	31.4	**21.2**	**45.7**	**39.2**
EnKF		GC Lat	25.5	19.2	64.9	41.5	63.7	**58.5**
UKF			24.4	18.1	65.7	41.7	63.6	**58.8**
EnKF	W	W_fm_full	**20.2**	**14.9**	**26.89**	28.1	67.3	67.6
UKF		-	20.8	**9.93**	**7.19**	**11.4**	71.7	71.7
EnKF	Wf	Area Wf	**19.7**	**16.5**	**20.3**	**21.5**	**60.6**	**57.0**
UKF			**14.4**	**9.30**	**5.87**	**9.00**	**61.6**	**51.2**

537 *Observations as: GC TD: green cover (top-down view), GC Lat: green cover (lateral view),
538 W_fm_full: weighting system, Area Wf: total area of fruits. Bold numbers refer to root mean
539 squared errors lower than the larger RMSE between the non-calibrated and calibrated error.

One example of how assimilation of one variable may unexpectedly affect the outcome of another refers to the assimilation of aboveground biomass using the weighting system (Figure 7). Even though it thoroughly improved aboveground biomass, it had an adverse effect on fruit dry biomass. Because the Reduced Tomgro model calculates yield based on photosynthesis and respiration (Equation A9, Appendix A), instead of previous values of aboveground biomass, the increase in biomass may lead to an increase in respiration that is not compensated by an increase in photosynthesis through LAI, thus decreasing assimilates available for fruits.

547



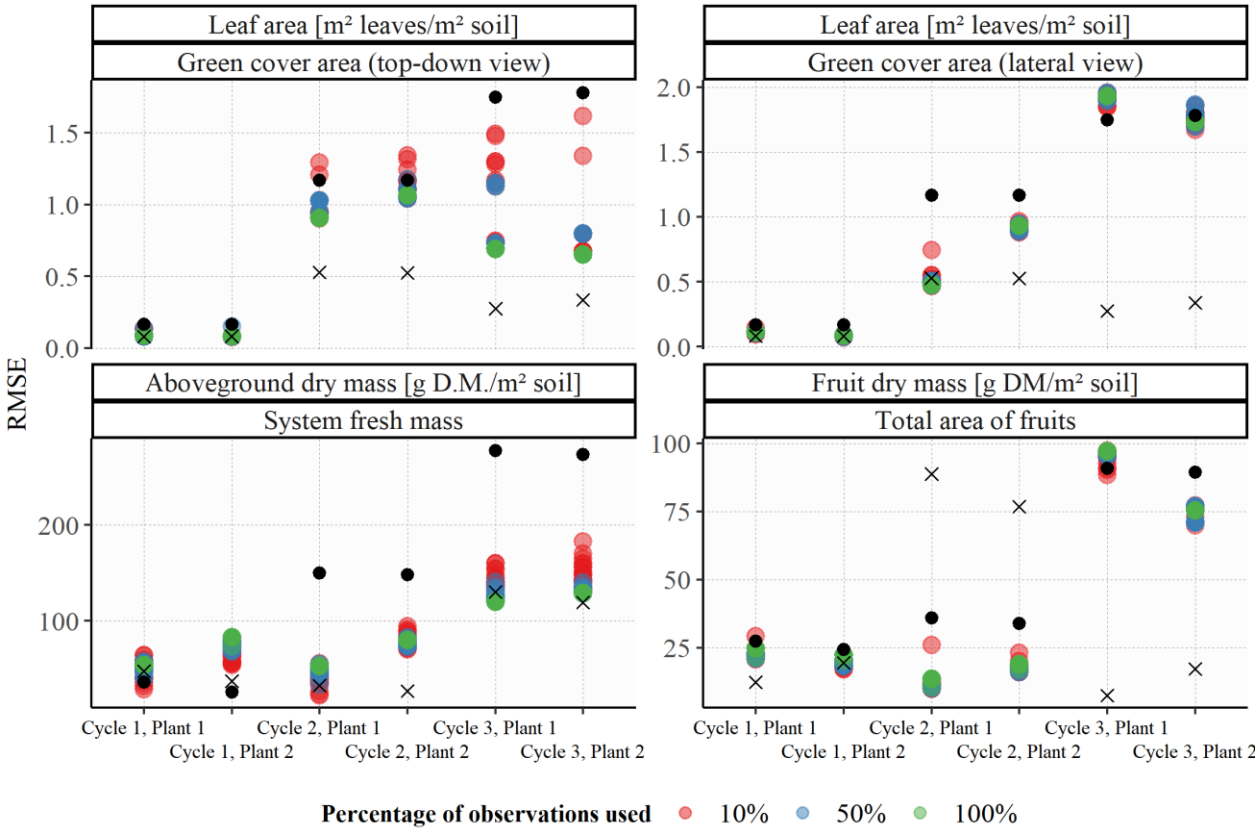
548

549 Figure 7. Assimilation of aboveground biomass observations. Growth curves for each monitored
550 plant without assimilation (OL, calib. And OL, no calib.) and with the different methods used for
551 assimilation (UKF and EnKF) of the indirect measurement of aboveground biomass corresponding
552 to fresh plant weighting system. Dots refer to the average equivalent observation determined by
553 destructive measurements, and the bar represents the associated standard deviation. The final
554 equivalent value for the monitored plant is represented by a larger dot. The equivalent observations
555 were converted from the original value to account for plant density.

3.3.4. Frequency

For field crops, observation availability has been mostly reliant on satellite temporal resolution, and only recently unmanned aerial vehicles (H. Ge et al., 2022; Yu et al., 2020) or pictures (Han et al., 2022) have become an additional source of data. In the high-frequency observation scenario of a protected environment, data availability allows for obtaining more prompt responses to plants' changes not captured by the model. However, since computational capacity may limit exploring an excessive number of images, we assessed the difference in performance caused by using fewer observations. While we observed similar outcomes of using all available data (Figure 8), by reducing the frequency, in particular for the assimilation of top-down images and the system fresh mass, the ranges of the errors obtained in the multiple simulations increased.

This increase in range may have just been caused by the need of frequent assimilation given the poor estimates of the non-calibrated model. But the variability among samples suggests it could also have been caused by the usefulness of observations not being equal across time. For images, it is also the case that later observations are connected to poorer data quality, since the environment becomes more complex. This adds to the problem of determining which variable could improve the outcome after assimilation, as variable importance, similar to what has been commented for parameters, also depends on timing.



574

575 Figure 8. Differences in errors caused by sampling of observations. Root mean squared error
576 (RMSE) for each state variable assimilated, from different sources, with the Unscented Kalman
577 Filter for all the twenty samplings of the dataset. The black dot corresponds to the error of the
578 simulation without calibration and the X mark, to the simulation with calibration.
579

580 ### 3.3.5. Different model calibration

581 For the most part of this study, we made a choice of not calibrating the model before
582 assimilation, as in Lu et al. (2021), which assimilated canopy cover and soil moisture to improve
583 maize yield. We did so because, even though it could influence filters' performances (Kang and
584 Özdoğan, 2019), calibration may be an excessively laborious step, as the parameters obtained from
585 calibration with data from one population do not necessarily correspond to the true parameter
586 values (Wallach, 2011), leading to the necessity of calibration for every different location and
587 cultivar. Within our exploration of data assimilation for a protected environment, we also tested

how well the assimilation of indirect measurements would bypass the need for calibration of the available models.

However, as the results observed are also a consequence of which were the parameters in the open loop simulation to which assimilation of observations was added, we also evaluated the results using another set of parameters, i.e., the ones obtained in calibration in Cycle 3, for assimilation in Cycles 1 and 2. The best example from the results is the case of the fruit estimates in Cycle 2 (Figure 9). One can see the model now overestimates fruit and mature biomass (OL, calib external). Assimilation lowers the estimates while there are observations available, but as the rate of fruit biomass is calculated independently from the current value of the state — relying exclusively on net biomass, development stage, some parameters and air temperature (Equation A8, Appendix A) —, the steep increase observed in the non-calibrated estimates is then also observed in the curves of the assimilated cases as soon as observations are no longer available. Likely, for the EnKF, the covariance generated by perturbation of the parameter was much lower than the ones in the observations, leading to higher reliance on the model's estimates.

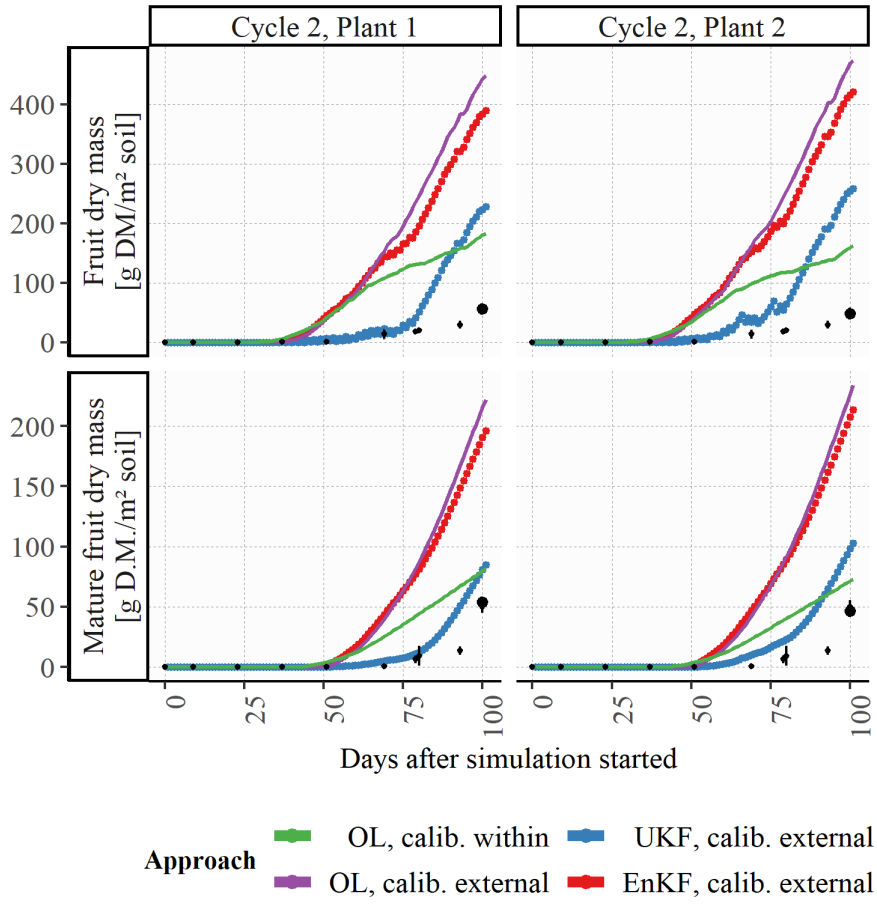


Figure 9. Assimilation of fruit area observations in a different model calibration. Growth curves for each monitored plant without assimilation (OL, within and OL, external) and with the different methods used for assimilation (UKF and EnKF) of fruit dry biomass corresponding to the area of fruits in images. Calibration with data from the cycle is referred to as "within" and with data from a different cycle, as "external". Dots refer to the average equivalent observation determined by destructive measurements, and the bar represents the associated standard deviation. The final equivalent value for the monitored plant is represented by a larger dot. The equivalent observations were converted from the original value to account for plant density.

4. Conclusions

This study covered aspects of the data assimilation framework of research, redirecting previous knowledge of assimilation in large areas with satellite images to protected environments. As an area not yet largely explored, several intermediate steps were required before achieving the

goal of assimilation, including determining viable monitoring approaches and observation models. Our case study brings attention to the impacts these steps and decisions may have on the outcome.

One decision was connected to which variables would lead to effective assimilation. We used observations that could lead to reasonable estimates of most state variables represented by the Reduced-State Tomgro model, i.e., aboveground biomass, leaf area index and fruit biomass. Overall, assimilation of observations only slightly improved estimates obtained by models, but the approach seemed especially valuable to adjust estimates in growth cycles in which potential growth was not observed. Simple measurements extracted from photos obtained during growth were able to represent state variables, providing a good estimate to be used along with the one provided by the model. The ability to further explore the approach also relies on the availability of good and generalizable observation models.

As for the improvement of yield estimates, in our case, assimilation led to two different outcomes: both improvement and deterioration of them. It is often unclear how assimilation may affect the results, since how much one state affects the other is not fixed, even in potential growth, because changes in the weather inputs or cultivars could modify crop behavior. A previous overall assessment of how variables relate to each other is recommended for a more efficient application of the technique.

As a first proposal of using real-time data for data assimilation in a protected environment, several aspects could not be investigated, such as exploring the uncertainty reduction in final estimates, and the impacts of assimilation of multiple observations simultaneously. Additionally, deep learning identification of observations could increase data availability and when more observations are available, one can delve deeper into the question of which is the more appropriate timing for obtaining them, as it seems possible, with fewer observations, to achieve results as good as with observations every other day.

5. 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). We thank Dr James W Jones for providing access to the spreadsheet that was originally used to perform model calculations.

6. 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

7. Conflicts of interest

The authors declare that there is no conflict of interest.

8. References

Arulampalam, M.S.; Maskell, S.; Gordon, N.; Clapp, T. 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on Signal Processing 50: 174–188. https://doi.org/10.1109/78.978374

Baar, S.; Kobayashi, Y.; Horie, T.; Sato, K.; Suto, H.; Watanabe, S. 2022. Non-destructive Leaf Area Index estimation via guided optical imaging for large scale greenhouse environments. Computers and Electronics in Agriculture 197: 106911.

https://doi.org/10.1016/j.compag.2022.106911

Chen, W.-T.; Yeh, Y.-H.F.; Liu, T.-Y.; Lin, T.-T. 2016. An Automated and Continuous Plant Weight Measurement System for Plant Factory. Frontiers in Plant Science 7: 392. https://doi.org/10.3389/fpls.2016.00392

De Graaf, R.; De Gelder, A.; Blok, C. 2004. Advanced weighing equipment for water, crop growth and climate control management. Acta Horticulturae 664: 163–167. https://doi.org/10.17660/ActaHortic.2004.664.17

de Koning, A.N.M.; Bakker, J.C. 1992. In situ plant weight measurement of tomato with an electronic force gauge. Acta Horticulturae 183–186. https://doi.org/10.17660/ActaHortic.1992.304.20

Dorigo, W.A.; Zurita-Milla, R.; de Wit, A.J.W.; Brazile, J.; Singh, R.; Schaepman, M.E. 2007. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. International Journal of Applied Earth Observation and Geoinformation 9: 165–193. https://doi.org/10.1016/j.jag.2006.05.003

Evensen, G. 2003. The Ensemble Kalman Filter: Theoretical formulation and practical implementation. Ocean Dynamics 53: 343–367. https://doi.org/10.1007/s10236-003-0036-9

Evensen, G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research 99: 10143–10162. https://doi.org/10.1029/94JC00572

Fang, H.; Baret, F.; Plummer, S.; Schaepman-Strub, G. 2019. An Overview of Global Leaf Area Index (LAI): Methods, Products, Validation, and Applications. Reviews of Geophysics 57: 739–799. https://doi.org/10.1029/2018RG000608

Fonteijn, H.; Afonso, M.; Lensink, D.; Mooij, M.; Faber, N.; Vroegop, A.; Polder, G.; Wehrens, R. 2021. Automatic phenotyping of tomatoes in production greenhouses using robotics and

686 computer vision: From theory to practice. Agronomy 11: .
687 https://doi.org/10.3390/agronomy11081599
688 Ge, H.; Ma, F.; Li, Z.; Du, C. 2022. Estimating rice yield by assimilating UAV-derived plant
689 nitrogen concentration into the DSSAT model: Evaluation at different assimilation time
690 windows. Field Crops Research 288: 108705. https://doi.org/10.1016/j.fcr.2022.108705
691 Ge, Y.; Lin, S.; Zhang, Y.; Li, Z.; Cheng, H.; Dong, J.; Shao, S.; Zhang, J.; Qi, X.; Wu, Z. 2022.
692 Tracking and Counting of Tomato at Different Growth Period Using an Improving YOLO-
693 Deepsort Network for Inspection Robot. Machines 10: 489.
694 https://doi.org/10.3390/machines10060489
695 Gong, L.; Yu, M.; Jiang, S.; Cutsuridis, V.; Pearson, S. 2021. Deep Learning Based Prediction on
696 Greenhouse Crop Yield Combined TCN and RNN. Sensors 21: 4537.
697 https://doi.org/10.3390/s21134537
698 Han, J.; Shi, L.; Yang, Q.; Chen, Z.; Yu, J.; Zha, Y. 2022. Rice yield estimation using a CNN-
699 based image-driven data assimilation framework. Field Crops Research 288: 108693.
700 https://doi.org/10.1016/j.fcr.2022.108693
701 Haykin, S. 2013. Adaptive Filter Theory, Fifth edit. ed. Pearson.
702 Helmer, T.; Ehret, D.L.; Bittman, S. 2005. CropAssist, an automated system for direct
703 measurement of greenhouse tomato growth and water use. Computers and Electronics in
704 Agriculture 48: 198–215. https://doi.org/10.1016/j.compag.2005.04.005
705 Hu, S.; Shi, L.; Huang, K.; Zha, Y.; Hu, X.; Ye, H.; Yang, Q. 2019. Improvement of sugarcane
706 crop simulation by SWAP-WOFOST model via data assimilation. Field Crops Research 232:
707 49–61. https://doi.org/10.1016/j.fcr.2018.12.009
708 Huang, J.; Gómez-Dans, J.L.; Huang, H.; Ma, H.; Wu, Q.; Lewis, P.E.; Liang, S.; Chen, Z.; Xue,
709 J.-H.; Wu, Y.; Zhao, F.; Wang, J.; Xie, X. 2019. Assimilation of remote sensing into crop

710 growth models: Current status and perspectives. Agricultural and Forest Meteorology 276–
711 277: 107609. https://doi.org/10.1016/j.agrformet.2019.06.008
712 Jin, X.; Kumar, L.; Li, Z.; Feng, H.; Xu, X.; Yang, G.; Wang, J. 2018. A review of data assimilation
713 of remote sensing and crop models. European Journal of Agronomy 92: 141–152.
714 https://doi.org/10.1016/j.eja.2017.11.002
715 Jones, J.W.; Dayan, E.; Allen, L.H.; Van Keulen, H.; Challa, H. 1991. A dynamic tomato growth
716 and yield model (TOMGRO). Transactions of the ASAE 34: 0663–0672.
717 https://doi.org/10.13031/2013.31715
718 Jones, J.W.; Kenig, A.; Vallejos, C.E. 1999. Reduced state-variable tomato growth model.
719 Transactions of the ASAE 42: 255–265. https://doi.org/10.13031/2013.13203
720 Julier, S.J.; Uhlmann, J.K. 1997. New extension of the Kalman filter to nonlinear systems, in:
721 Kadar, I. (Ed.), Signal Processing, Sensor Fusion, and Target Recognition VI. SPIE, p. 182.
722 https://doi.org/10.1117/12.280797
723 Kang, Y.; Özdoğan, M. 2019. Field-level crop yield mapping with Landsat using a hierarchical
724 data assimilation approach. Remote Sensing of Environment 228: 144–163.
725 https://doi.org/10.1016/J.RSE.2019.04.005
726 Keating, B.A.; Thorburn, P.J. 2018. Modelling crops and cropping systems—Evolving purpose,
727 practice and prospects. European Journal of Agronomy 100: 163–176.
728 https://doi.org/10.1016/j.eja.2018.04.007
729 Lee, J.W.; Son, J.E. 2019. Nondestructive and Continuous Fresh Weight Measurements of Bell
730 Peppers Grown in Soilless Culture Systems. Agronomy 9: 652.
731 https://doi.org/10.3390/agronomy9100652
732 Linker, R.; Ioslovich, I. 2017. Assimilation of canopy cover and biomass measurements in the crop
733 model AquaCrop. Biosystems Engineering 162: 57–66.

https://doi.org/10.1016/j.biosystemseng.2017.08.003

Liu, L.; Yuan, J.; Gong, L.; Wang, X.; Liu, X. 2022. Dynamic Fresh Weight Prediction of Substrate-Cultivated Lettuce Grown in a Solar Greenhouse Based on Phenotypic and Environmental Data. Agriculture 12: 1959. https://doi.org/10.3390/agriculture12111959

Lu, Y.; Chibarabada, T.P.; Ziliani, M.G.; Onema, J.K.; Mccabe, M.F.; Sheffield, J. 2021. Assimilation of soil moisture and canopy cover data improves maize simulation using an under-calibrated crop model. Agricultural Water Management 252: 106884. https://doi.org/10.1016/j.agwat.2021.106884

Luo, L.; Sun, S.; Xue, J.; Gao, Z.; Zhao, J.; Yin, Y.; Gao, F.; Luan, X. 2023. Crop yield estimation based on assimilation of crop models and remote sensing data: A systematic evaluation. Agricultural Systems 210: 103711. https://doi.org/10.1016/j.agsy.2023.103711

Mandel, J. 2006. Efficient implementation of the ensemble Kalman filter. University of Colorado at Denver and Health Sciences Center, Center for Computational Mathematics Reports.

Moon, T.; Kim, D.; Kwon, S.; Ahn, T.I.; Son, J.E. 2022. Non-Destructive Monitoring of Crop Fresh Weight and Leaf Area with a Simple Formula and a Convolutional Neural Network. Sensors 22: 7728. https://doi.org/10.3390/s22207728

Nearing, G.S.; Crow, W.T.; Thorp, K.R.; Moran, M.S.; Reichle, R.H.; Gupta, H. V. 2012. Assimilating remote sensing observations of leaf area index and soil moisture for wheat yield estimates: An observing system simulation experiment. Water Resources Research 48: . https://doi.org/10.1029/2011WR011420

Oliveira, M.P.G. de. 2022. Leveraging data assimilation and monitoring data for improvement of crop growth estimates in protected environments. Universidade Estadual de Campinas. URL: https://hdl.handle.net/20.500.12733/4792

Oliveira, M.P.G. de; Amaro, R.P.; Pescarini, H.B.; Rodrigues, L.H.A. 2021. Tomato growth in

production-like setting. https://doi.org/10.25824/redu/EP4NGO

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

Orlova, Y.; Linker, R. 2023. Data assimilation with sensitivity-based particle filter: A simulation study with AquaCrop. Computers and Electronics in Agriculture 204: 107538. https://doi.org/10.1016/j.compag.2022.107538

Pellenq, J.; Boulet, G. 2004. A methodology to test the pertinence of remote-sensing data assimilation into vegetation models for water and energy exchange at the land surface. Agronomie 24: 197–204. https://doi.org/10.1051/agro:2004017

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

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

van Delden, S.H.; SharathKumar, M.; Butturini, M.; Graamans, L.J.A.; Heuvelink, E.; Kacira, M.; Kaiser, E.; Klamer, R.S.; Klerkx, L.; Kootstra, G.; Loeber, A.; Schouten, R.E.; Stanghellini, C.; van Ieperen, W.; Verdonk, J.C.; Vialet-Chabrand, S.; Woltering, E.J.; van de Zedde, R.; Zhang, Y.; Marcelis, L.F.M. 2021. Current status and future challenges in implementing and upscaling vertical farming systems. Nature Food 2: 944–956. https://doi.org/10.1038/s43016-021-00402-w

van Mourik, S.; van Beveren, P.J.M.; López-Cruz, I.L.; van Henten, E.J. 2019. Improving climate monitoring in greenhouse cultivation via model based filtering. Biosystems Engineering 181:

40–51. https://doi.org/10.1016/J.BIOSYSTEMSENG.2019.03.001

Wallach, D. 2011. Crop Model Calibration: A Statistical Perspective. Agronomy Journal 103: 1144–1151. https://doi.org/10.2134/agronj2010.0432

Wan, E.A.; Van Der Merwe, R. 2000. The unscented Kalman filter for nonlinear estimation. IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium, AS-SPCC 2000 153–158. https://doi.org/10.1109/ASSPCC.2000.882463

Yu, D.; Zha, Y.; Shi, L.; Jin, X.; Hu, S.; Yang, Q. 2020. Improvement of sugarcane yield estimation by assimilating UAV-derived plant height observations. European Journal of Agronomy 121: 126159. https://doi.org/10.1016/j.eja.2020.126159

Zhang, F.; Lv, Z.; Zhang, H.; Guo, J.; Wang, J.; Lu, T.; Zhangzhong, L. 2023. Verification of improved YOLOX model in detection of greenhouse crop organs: Considering tomato as example. Computers and Electronics in Agriculture 205: 107582. https://doi.org/10.1016/j.compag.2022.107582

Appendix A Reduced State Tomgro (RT) model equations

The model calculates growth and development of tomato plants and its main calculations are based on Equations A1 to A9, extracted from (Jones et al., 1999) and (Jones et al., 1991). The model is represented by five difference equations of the states number of nodes (N), leaf area index (LAI), aboveground dry biomass (W), fruit dry biomass (W_f) and mature fruit dry biomass (W_m). It uses photosynthetic photon flux density (PPFD) and hourly air temperatures (T) as inputs. Parameters names are presented in Table A1.

$$\frac{dN}{dt} = N_m \cdot f_N(T) \quad \text{A1}$$

$$\frac{dLAI}{dt} = \rho \cdot \delta \cdot \lambda(T_d) \cdot \frac{e^{[\beta(N-N_b)]}}{1+e^{[\beta(N-N_b)]}} \cdot \frac{dN}{dt} \quad \text{A2}$$

$$GR_{net} = \mathrm{E} \cdot \left(P_g - R_m\right) \cdot [1 - f_R(N)] \quad \text{A3}$$

$$R_m = \sum Q_{10}^{(T-20)} \cdot r_m \cdot (W - W_m)\,\mathrm{dt} \quad \text{A4}$$

$$P_g = \sum \left\{ \frac{\mathrm{D} \cdot LF_{max} \cdot \mathrm{PGRED(T)}}{K} \cdot \ln\left[\frac{(1-m) \cdot LF_{max} + Q_e \cdot \mathrm{K} \cdot \mathrm{PPFD}}{(1-m) \cdot LF_{max} + Q_e \cdot \mathrm{K} \cdot \mathrm{PPFD} \cdot \mathrm{e}^{-\mathrm{K} \cdot \mathrm{LAI}}} \right] \right\} \quad \text{A5}$$

$$\frac{dW}{dt} = GR_{net} - p_1 \cdot \rho \cdot \frac{dN}{dt} \quad \text{A6}$$

$$\frac{dW}{dt}_{max} = \frac{dW_F}{dt} + (V_{max} - p_1) \cdot \rho \cdot \frac{dN}{dt} \quad \text{A7}$$

$$\frac{dW_F}{dt} = GR_{net} \cdot \alpha_F \cdot f_F(T_d) \cdot \left[1 - e^{\vartheta - (N - N_{FF})}\right] \cdot g\left(T_{daytime}\right) \quad \text{A8}$$

$$\frac{dW_M}{dt} = D_F(T_d) \cdot (W_F - W_M) \quad \text{A9}$$

814 Table A1. Parameters from the Reduced Tomgro model.

815

Parameter	Meaning
δ	Maximum leaf area expansion per node [m^2 leaf $node^{-1}$]
β	Coefficient in expolinear equation [$node^{-1}$]
N_b	Project of linear segment of LAI vs N to horizontal axis [node]
α_F	Maximum partitioning of new growth to fruit [fraction d^{-1}]
ϑ	Transition coefficient between vegetative and full fruit growth [$node^{1}$]
N_{FF}	Nodes per plant when first fruit appears [node]
T_{CRIT}	Mean daytime air temperature above which fruit abortion starts [˚C]
DFmax	Average development rate used to move fruits from green to mature stage [d^{-1}]
V_{max}	Maximum increase in vegetative tissue dry weight growth per node [g dry weight $node^{-1}$]
N_m	Maximum rate of node appearance (at optimal air temperatures) [node d^{-1}]
Q_e	Leaf quantum efficiency [µmol (CO_2 fixed) µmol $(photon)^{-1}$]
E	Growth efficiency, ratio of biomass to photosynthate available for growth [g d.w. g^{1}CH2O]
K	Light extinction coefficient
ρ	Plant density [plant m^{-2}]
p_1	Loss of leaf d.w. per node after LAI_max is reached [g leaf $node^{-1}$]

816

Appendix B Examples of data obtained through non-destructive methods

Figure B1. Example of photo of a monitored plant, captured by the fixed Raspberry Pi Camera Module v2 positioned above the tomato plant weighting system. Image includes reference of known dimensions.

Figure B2. Example of photo of a monitored plant, captured by the fixed Raspberry Pi Camera Module v2 positioned in a row adjacent to the tomato plant weighting system. Image includes reference of known dimensions.

Figure B3. Example of image of a calibration plant, obtained before removal from the greenhouse, by a smartphone camera, from a top-down view. Image includes reference of known dimensions.

Appendix C Equations for the EnKF and the UKF

Equations C1 to C24 summarize the steps and relationships between states, observations and their uncertainties, expressed by their covariances. Overall, they follow a pattern of using the process or model equation to predict the next estimate and updating the estimate by including the information brought by the observation, accounting for the uncertainty present both in the model estimate and the observation. The first step is often called forecast, prior or predict and the second step, update, posterior or analysis. Relevant moments from the update step include the calculation of the residual, which consists in the difference between the estimate and the observation, and of the gain, which corresponds to the weight of the residual when modifying the value estimated by the model, as well as the uncertainty of the outcome.

EnKF

$$\boldsymbol{\chi} \sim \boldsymbol{\mathcal{N}}(\mathbf{x}_0, \mathbf{P}_0) \quad \text{C1}$$

$$\boldsymbol{\mathcal{Y}} = \mathrm{f}(\boldsymbol{\chi}) + \boldsymbol{\upsilon}_Q \quad \text{C2}$$

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_1^N \boldsymbol{\mathcal{Y}} \quad \text{C3}$$

$$\bar{\mathbf{P}} = \frac{1}{N-1}\sum_1^N (\boldsymbol{\mathcal{Y}} - \bar{\mathbf{x}})(\boldsymbol{\mathcal{Y}} - \bar{\mathbf{x}})^{\boldsymbol{T}} \quad \text{C4}$$

$$\boldsymbol{\mathcal{Z}} = \mathrm{h}(\boldsymbol{\mathcal{Y}}) \quad \text{C5}$$

$$\boldsymbol{\mu}_{\boldsymbol{z}} = \frac{1}{N}\sum_1^N \boldsymbol{\mathcal{Z}} \quad \text{C6}$$

$$\mathbf{y} = \mathbf{z} - \boldsymbol{\mathcal{Z}} + \boldsymbol{\upsilon}_R \quad \text{C7}$$

$$\mathbf{P}_{\mathbf{zz}} = \frac{1}{N-1}\sum_1^N (\boldsymbol{\mathcal{Z}} - \boldsymbol{\mu}_{\boldsymbol{z}})(\boldsymbol{\mathcal{Z}} - \boldsymbol{\mu}_{\boldsymbol{z}})^{\boldsymbol{T}} + \mathbf{R} \quad \text{C8}$$

$$\mathbf{P}_{\mathbf{xz}} = \frac{1}{N-1}\sum_1^N (\boldsymbol{\mathcal{Y}} - \bar{\mathbf{x}})(\boldsymbol{\mathcal{Z}} - \boldsymbol{\mu}_{\boldsymbol{z}})^{\boldsymbol{T}} \quad \text{C9}$$

$$\mathbf{K} = \mathbf{P}_{\mathbf{xz}}\mathbf{P}_{\boldsymbol{zz}}{}^{-\mathbf{1}} \quad \text{C10}$$

$$\boldsymbol{\chi} = \bar{\mathbf{x}} + \mathbf{K}\mathbf{y} \quad \mathbf{C}11$$

$$\mathbf{x} = \frac{1}{N}\sum_1^N \boldsymbol{\chi} \quad \text{C12}$$

$$\mathbf{P} = \bar{\mathbf{P}} - \mathbf{K}\mathbf{P}_{\mathbf{zz}}\mathbf{K}^{\mathrm{T}} \quad \text{C13}$$

UKF

$$\boldsymbol{\mathcal{Y}} = \mathrm{f}(\boldsymbol{\chi}) \quad \text{C14}$$

$$\bar{\mathbf{x}} = \sum w^m \boldsymbol{\mathcal{Y}} \quad \text{C15}$$

$$\bar{\mathbf{P}} = \sum w^c (\boldsymbol{\mathcal{Y}} - \bar{\mathbf{x}})(\boldsymbol{\mathcal{Y}} - \bar{\mathbf{x}})^{\boldsymbol{T}} + \mathbf{Q} \quad \mathbf{C}16$$

$$\boldsymbol{\mathcal{Z}} = \mathrm{h}(\boldsymbol{\mathcal{Y}}) \quad \text{C17}$$

$$\boldsymbol{\mu}_{\boldsymbol{z}} = \sum w^m \boldsymbol{\mathcal{Z}} \quad \text{C18}$$

$$\mathbf{y} = \mathbf{z} - \boldsymbol{\mu}_{\boldsymbol{z}} \quad \mathbf{C}19$$

$$\mathbf{P}_{\mathbf{zz}} = \sum w^c (\boldsymbol{\mathcal{Z}} - \boldsymbol{\mu}_{\boldsymbol{z}})(\boldsymbol{\mathcal{Z}} - \boldsymbol{\mu}_{\boldsymbol{z}})^{\boldsymbol{T}} + \mathbf{R} \quad \mathbf{C}20$$

$$\mathbf{P}_{\mathbf{xz}} = \sum w^c (\boldsymbol{\mathcal{Y}} - \bar{\mathbf{x}})(\boldsymbol{\mathcal{Z}} - \boldsymbol{\mu}_{\boldsymbol{z}})^{\boldsymbol{T}} \quad \text{C22}$$

$$\mathbf{K} = \mathbf{P}_{\mathbf{xz}}\mathbf{P}_{\mathbf{zz}}{}^{\mathbf{-1}} \quad \mathbf{C}22$$

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{K}\mathbf{y} \quad \mathbf{C}23$$

$$\mathbf{P} = \bar{\mathbf{P}} - \mathbf{K}\mathbf{P}_{\mathbf{zz}}\mathbf{K}^{\mathrm{T}} \quad \text{C24}$$

In the predict equations (C1 to C4 and C14 to C16), **x** and **P** are the state mean and covariance, **F** is the process function in matrix form, while f is the equivalent nonlinear process model and **Q** is the process covariance. In the update equations (C5 to C13 and C17 to C24), **z** and **R** are the measurement mean and noise covariance, **H** is the measurement function while h is the equivalent nonlinear observation model, **y** and **K** are the residual and Kalman gain. In the UKF, $\boldsymbol{\chi}$

878 are the sigma points and $\boldsymbol{\mathcal{Y}}$ are the transformed sigma points while $\boldsymbol{w^m}$ and $\boldsymbol{w^c}$ are weights. In the

879 EnKF, $\boldsymbol{\upsilon_Q}$ and $\boldsymbol{\upsilon_R}$ correspond to added noise.

880

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.