Agricultural Systems 201 (2022) 103456

Contents lists available at ScienceDirect

Agricultural Systems

journal homepage: www.elsevier.com/locate/agsy

Assimilation of wheat and soil states for improved yield prediction: The APSIM-EnKF framework

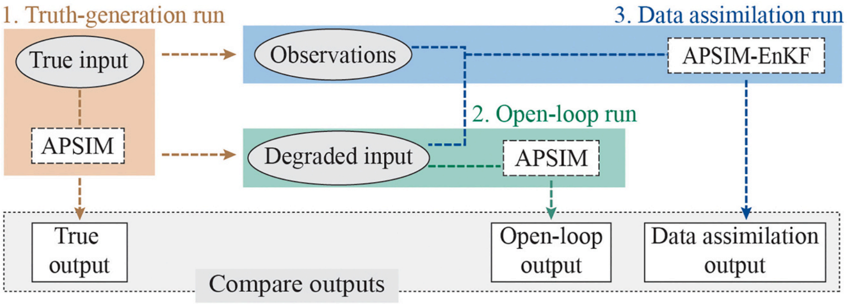
Yuxi Zhang, Jeffrey P. Walker [*], Valentijn R.N. Pauwels

Department of Civil Engineering, Monash University, Clayton, Victoria, Australia

HIGHLIGHTS

- Data assimilation integrates remote sensing with crop models for increased temporal resolution and improved model accuracy
- A data assimilation framework was developed to assimilate wheat and soil observations into APSIM and is available on GitHub
- Yield was improved by some wheat and soil states, with the improvement varying according to phenological stage
- This comprehensive synthetic study provides a guide for data assimilation practices and field experiment design

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Mark van Wijk

Keywords:
APSIM-wheat
Data assimilation
EnKF
Yield forecasting

ABSTRACT

Context.

Accurate prediction of within-field crop yield in response to spatial and temporal variability provides essential information for farm managers to improve productivity and ensure optimal use of inputs. Understanding yield spatial and temporal variability cannot be solely addressed by crop modelling or remote sensing but by integrating the instantaneous spatial information from remote sensing and the temporal information from crop modelling. Sequential data assimilation techniques allow wheat and soil observations to be assimilated into the crop model while it evolves and evaluate model and observational uncertainties to improve the accuracy of crop monitoring and yield prediction.

Objectives: The objective of this study was to comprehensively explore the potential yield estimation improvement by assimilating observations of all prognostic wheat and soil states, including various repeat intervals and accuracy, allowing recommendations on implementation to be made.

Methods: This study develops an Ensemble Kalman filter (EnKF) data assimilation framework for the APSIM-Wheat model and illustrates potential improvements in wheat yield estimation through a synthetic study. Through several scenarios, assimilation of wheat and soil observations into APSIM was explored, by assimilating these variables solely or collectively, and in various phenological stages.

Results and conclusions.

The results showed, under the specific weather and soil conditions assumed in this study, that while open-loop (no data assimilation) provided a yield estimation with a bias of 10.1%, assimilation in the flowering to end of

Abbreviations: EnKF, Ensemble Kalman filter.

[*] Corresponding author.

E-mail address: jeff.walker@monash.edu (J.P. Walker).

https://doi.org/10.1016/j.agsy.2022.103456
Received 26 February 2022; Received in revised form 24 May 2022; Accepted 22 June 2022
Available online 29 June 2022

grain filling stage reduced the bias to 1.4%, 2.9%, 4.4%, and 1.0% when constraining with leaf area index, leaf weight, stem weight, surface soil moisture observations, respectively. When assimilating in the floral initiation to the flowering stage, the yield estimation bias was reduced to 7.1%, 9.8%, 1.1%, and 1.2% when constraining with leaf nitrogen, stem nitrogen, top-layer soil ammonium-nitrogen and nitrate-nitrogen, respectively. Leaf area index, biomass and surface soil moisture are recommended for data assimilation especially with observations from remote sensing.

Significance: This study developed a data assimilation framework for the APSIM-Wheat model and can be extended to over 20 crop modules integrated with APSIM. This synthetic study provided a exhaustive data assimilation experiment for wheat and soil states that are measurable by current techniques with a rigorous justification on uncertainties. It thus provides a guide for future agricultral data assimilation practices in choosing crop and soil states for assimilation, and for planning the timing and frequency of data collection. It should also inspire researchers to develop new techniques for measuring wheat states.

1. Introduction

With increased human population comes the demand for more food from the same amount of land and water. Accordingly, farming practices need to be made more efficient to achieve this outcome, and this can only be done with a combination of numerical prediction models and observations. With better insight into the predicted spatial and temporal variability of yield, site-specific management can be performed so as to achieve increased farm productivity, reduced cost and a reduced impact on the environment (Noori and Panda, 2016; Panda et al., 2010; Paustian and Theuvsen, 2017; Shaw et al., 2016).

While physically-based crop simulation models can provide a real-time estimation of the crop growth on a daily basis, the data required for accurate simulation across spatial scales are not usually available (Batchelor et al., 2002; Mosleh et al., 2015). In contrast, remotely sensed data provide broad spatial coverage and fine resolution, but is only available every couple of days at best and does not provide insights into the interaction of the crop with the environment for management purposes (Mosleh et al., 2015). Nor does it allow a direct propagation of estimated yield. Therefore, a combination of crop simulation models and remote sensing data provides a potential pathway for providing spatially variable information on the current crop status and expected evolution of yield.

The approach of combining crop simulation models and remotely sensed data has been discussed by many researchers (Bouman, 1995; Jin et al., 2018; Maas, 1988; Wiegand et al., 1986). Methods include: 1) using remotely sensed weather data as input to drive the model (Maas, 1988); 2) direct insertion that substitutes remotely sensed observations for model states in the simulation (Maas, 1988); 3) re-calibration whereby model initial conditions and/or parameters are optimised using a cost function, usually being the sum of squared difference between the model estimates and the remotely sensed observations (Jin et al., 2016; Launay and Guerif, 2005; Novelli et al., 2019; Thorp et al., 2012); and 4) state-updating using data assimilation techniques to sequentially update model states with the introduction of external observations (Curnel et al., 2011; Ines et al., 2013; Li et al., 2017b; Nearing et al., 2012). However, using remotely sensed as model input only makes it a surrogate of input from other sources. Direct insertion is a simplified state-updating that only trusts the observed rather than the modelled state values. The re-calibration method uses an iteration process that requires observations over the entire simulation window to be collected prior to the re-calibration, making it unsuitable for in-season forecasting. Although the term “data assimilation” generally means the fusing of a model with data, the discussion of data assimilation is commonly (and in this paper) constrained to a narrow sense, namely, state-updating.

Data assimilation frameworks have been developed for several popular crop prediction models: WOFOST (Curnel et al., 2011), DSSAT-CERES (Nearing et al., 2012), and AquaCrop (Silvestro et al., 2017). While these models have been widely applied and evaluated in regions over Europe (Eitzinger et al., 2004; Langensiepen et al., 2008; Mavromatis, 2016), America (Chipanshi et al., 1997; Mearns et al., 1992; Rosenzweig and Tubiello, 1996), Asia (Ahmed et al., 2016; Patel et al., 2010; Timsina and Humphreys, 2006; Xiong et al., 2008; Zhang et al., 2013), and Africa (Sadras et al., 2015), validation experiments for their application in Australia are rarely found. Rather, the Agricultural Production Systems sIMulator (APSIM), a highly advanced crop simulation system model developed over the last 20 years and well-validated (Ahmed et al., 2016; Asseng et al., 1998; Asseng et al., 2003; Asseng et al., 2000; Zhang et al., 2012; Zhao et al., 2014) is commonly used in Australia, but until recently there has been no data assimilation framework for this model (Kivi et al., 2022; Zhang et al., 2022).

Current practices in the domain of crop model data assimilation have primarily used limited types of remote sensing observations, including leaf area index (LAI) from MODIS (Chen et al., 2018; Ines et al., 2013; Vazifedoust et al., 2009; Zhao et al., 2013) and Landsat (Huang et al., 2016; Kang and Özdoğan, 2019), surface soil moisture from SMOS (Liu et al., 2019) and AMSR-E (Ines et al., 2013; Liu et al., 2019), and spectral reflectance from HJ-1 (Guo et al., 2019; Li et al., 2009; Ma et al., 2013). These practices successfully improved crop yield estimation. However, important wheat and soil states such as biomass, phenology and soil nitrogen have not been tested. Accordingly, a comprehensive and robust exploration of assimilating all potential state variables into crop models is lacking and so has been included here.

The estimation of model and observation uncertainties is crucial to any data assimilation implementation. Previous studies were generally based on simple assumptions for background uncertainties, assuming that they are only sourced from parameters and/or initial conditions (e.g., Huang et al., 2016; Ines et al., 2013; Li et al., 2017a; Zhao et al., 2013) or considered as Gaussian noise added to the state by a fixed value or proportional to the estimation (e.g., Kang and Özdoğan, 2019; Ma et al., 2013; Silvestro et al., 2017). A more sophisticated and common practice in meteorological, land surface and hydrological studies is to generate background uncertainties based on the individual uncertainty sources (primarily weather, model parameters, and initial conditions) with good estimation. However, this was only considered in a few crop modelling studies (Wit and Van Diepen, 2007; Nearing et al., 2012).

Against the above background, this paper determined which state observations can improve yield estimation through assimilation into the APSIM-Wheat model, and at which phenological stage(s) and repeat interval had the best impact on results. This paper presents an Ensemble Kalman filter (EnKF) algorithm-based state-updating data assimilation framework for the APSIM-Wheat model, and applies it in the context of a synthetic study such that a wide range of observation types, accuracies and repeat intervals etc. could be tested. Assimilated states were selected according to an extensive sensitivity analysis (Zhang, 2020) and included LAI, soil moisture, soil nitrogen, total biomass, individual organ weight, and phenology. Additionally, the observations at different phenological stages and repeat intervals were assimilated.

2. The crop model-data assimilation framework

2.1. APSIM-wheat

The physically-based APSIM-Wheat model is used to simulate wheat growth with a daily time step, accounting for interactions of the plant with the environment. The description of the model hereafter is based on the APSIM documentation for the wheat (Zheng et al., 2014) and soil (Probert et al., 1998) modules. The model considers winter wheat phenology according to ten phases (1-sowing, 2-germination, 3-emergence, 4-end of juvenile, 5-floral initiation, 6-flowering, 7-start of grain filling, 8-end of grain filling, 9-maturity and 10-harvest, with the phenological stage being the period between two phases), controlled by air temperature, day length, vernalisation and stress factors (e.g., water and nitrogen). The biomass accumulation is based on a simple radiation use efficiency light utilisation approach. Total daily incoming biomass is allocated to the above-ground wheat organs (leaf, stem, spike, and grain) with a proportion that varies with the phenological stage. The model considers plant extractable water and soil nitrogen as two stress factors to the growth of wheat, with a water balance model coupled with the wheat module to simulate soil water movement and estimate plant extractable water. The model considers rainfall-runoff, evapotranspiration, infiltration, unsaturated flow, saturated flow, and lateral flow processes. The soil nitrogen module uses three organic matter pools (fresh, hum and biom) to simulate the soil carbon and nitrogen transformation through decomposition, nitrification, denitrification, mineralisation, and immobilisation processes.

The simulation window was taken as a full year that had a moderate weather condition from January 1, 1996 to December 30, 1996, using a set of APSIM example weather data and parameters. The example weather input was provided in the APSIM software package as a dataset instance, including rainfall, temperature, solar radiation, and vapour pressure. The sowing date was set to Day of Year (DoY) 131 (May 10) after a 5-day rainfall event. The time series of perturbed weather inputs, including radiation, temperature and rainfall, are shown in Supplementary Fig. S2. The cultivar parameters, soil parameters and initial conditions used as model input are shown in Tables 1 & 2.

Table 2
Cultivar parameters selected within the range of typical parameter values in APSIM and used as model input in this study.

Cultivar parameter	Unit	Values or standard deviation used in		
		Truth	Open-loop	Assimilation (std.)[a]
VernSens (vernalization sensitivity)	–	2	2.1	0.1
PhotopSens (photoperiod sensitivity)	–	3.5	3.6	0.1
TT4 (target thermal time in stage 4)	°C day	400	410	20
TT5 (target thermal time in stage 5)	°C day	580	600	30
TT6 (target thermal time in stage 6)	°C day	120	125	6
TT7 (target thermal time in stage 7)	°C day	590	610	30
Potential Grain Filling Rate	10^{-3} grain g^{-1} d^{-1}	2	2.1	0.2
Potential Grain Growth Rate	10^{-3} grain g^{-1} d^{-1}	1	1.1	0.1
Potential Grain N (nitrogen) Filling Rate	10^{-5} grain g^{-1} d^{-1}	5.5	5	0.5

[a] The value presented is the standard deviation (std.) of the random term added to the respective variable of the open-loop run in producing the ensemble for the data assimilation run.

2.2. Ensemble Kalman filter

The Ensemble Kalman filter data assimilation algorithm is based on a Monte Carlo approach, where an ensemble of stochastic model simulations is used to approximate the probability distribution of the state. The description of the EnKF algorithm in this section is according to Marc (2014) and Stuart and Zygalakis (2015). Implementation of the EnKF consists of forecast and analysis steps. The forecast step utilises a space-time model that maps the analysis state from the previous timestep k-1 to the current step k. This 'forecast' (sometimes known as 'background') state is used as the input to the analysis step (otherwise known as data assimilation step), where external observations are assimilated into the system. The terms' analysis' and 'forecast'/'background' are sometimes denoted by 'posterior' and 'prior', meaning that the state or error covariance is obtained posterior and prior to the analysis step,

Table 1
Soil parameters and initial conditions selected within the range of typical parameter values in APSIM and as model input in this study.

Soil parameter	Unit	Values or standard deviation used in …	Layer						
			1	2	3	4	5	6	7
Soil depth (depth of each soil layer)	cm		0–15	15–30	30–60	60–90	90–120	120–150	150–200
InitialSM (initial soil moisture) [a]	m^3	Truth	0.25						
	m^{-3}	Open-loop	0.24						
		Assimilation (std.) [b]	0.1	0.09	0.08	0.07	0.06	0.06	0.06
WheatLL (lower limit of soil moisture that is extractable by the plant)	m^3	Truth	0.10	0.11	0.12	0.13	0.14	0.14	0.14
	m^{-3}	Open-loop	0.09	0.10	0.11	0.12	0.13	0.13	0.13
		Assimilation (std.) [b]	0.02						
DUL (drained upper limit)	m^3	Truth	0.36	0.35	0.34	0.33	0.32	0.32	0.32
	m^{-3}	Open-loop	0.38	0.37	0.36	0.35	0.34	0.34	0.34
		Assimilation (std.) [b]	0.02						
AirDry (soil moisture of air-dry soil)	m^3 m^{-3}	All	WheatLL - 0.02 [c]						
SAT (soil moisture of saturated soil)	m^3 m^{-3}	All	DUL + 0.1 [d]						
BD (bulk density)	g cm^{-3}	All	1.3	1.35	1.4	1.45	1.5	1.5	1.5

[a] No perturbation was applied to the initial soil moisture because a warm-up period of four months before sowing was included in the simulation window to allow the spread of ensemble soil moisture to be sufficiently large.

[b] The value presented is the standard deviation (std.) of the random term added to the respective variable of the open-loop run in producing the ensemble for the data assimilation run.

[c] AirDry was assumed 0.02 lower than the perturbed WheatLL value.

[d] SAT was assumed 0.1 higher than the perturbed DUL value.

respectively.

The forecast step estimates the forecast states as a direct model estimation based on the analysis states from the previous step by

$$\mathbf{x}_k^{i,f} = \mathscr{M}\left(\mathbf{x}_{k-1}^{i,a}, \mathbf{f}_k, \theta\right) \tag{1}$$

where the state vector $\mathbf{x}$ consists of 34 wheat and soil states, $\mathbf{f}_k$ is the time-dependent driving force (weather input for the case of APSIM), and θ is a set of parameters uniform throughout the simulation window. The ensemble of model state variables with an ensemble size of N is represented as

$$\mathbf{X} = \left[\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^N\right] \tag{2}$$

Observations of the system are mapped from the state vector through an observational matrix as

$$\mathbf{y}_k = \mathbf{H}\mathbf{x}_k + \mathbf{v}_k \tag{3}$$

where $\mathbf{H}$ is the observation matrix, and $\mathbf{v}_k$ is the observation error randomly drawn from a known Gaussian distribution $N(0, \mathbf{R}_k)$. In the analysis step, the forecast error covariance is calculated according to

$$\mathbf{P}_k^f = \frac{1}{N-1}\mathbf{D}_k^f\,\mathbf{D}_k^{f\,T} \tag{4}$$

where $\mathbf{D}_k$ is calculated by all $\mathbf{x}$ forecast states at timestep k as

$$\mathbf{D}_k^f = \left[\mathbf{x}^{1,f} - \mu_k^f, \mathbf{x}^{2,f} - \mu_k^f, \dots, \mathbf{x}^{N,f}\mu_k^f\right]. \tag{5}$$

The analysis state of the ith ensemble is calculated by

$$\mathbf{x}_k^{i,a} = \mathbf{x}_k^{i,f} + \mathbf{K}_k\left(\mathbf{y}_k + \mathbf{v}_k^i - \mathbf{H}\mathbf{x}_k^{i,f}\right) \tag{6}$$

where the Kalman gain obtained as

$$\mathbf{K}_k = \mathbf{P}_k^f\mathbf{H}^T\left(\mathbf{P}_k^f\mathbf{H}^T + \mathbf{R}_k\right)^{-1} \tag{7}$$

The forecast and analysis state of the ensemble are taken as the forecast and analysis ensemble means μ_k^a and μ_k^f at timestep k, expressed by

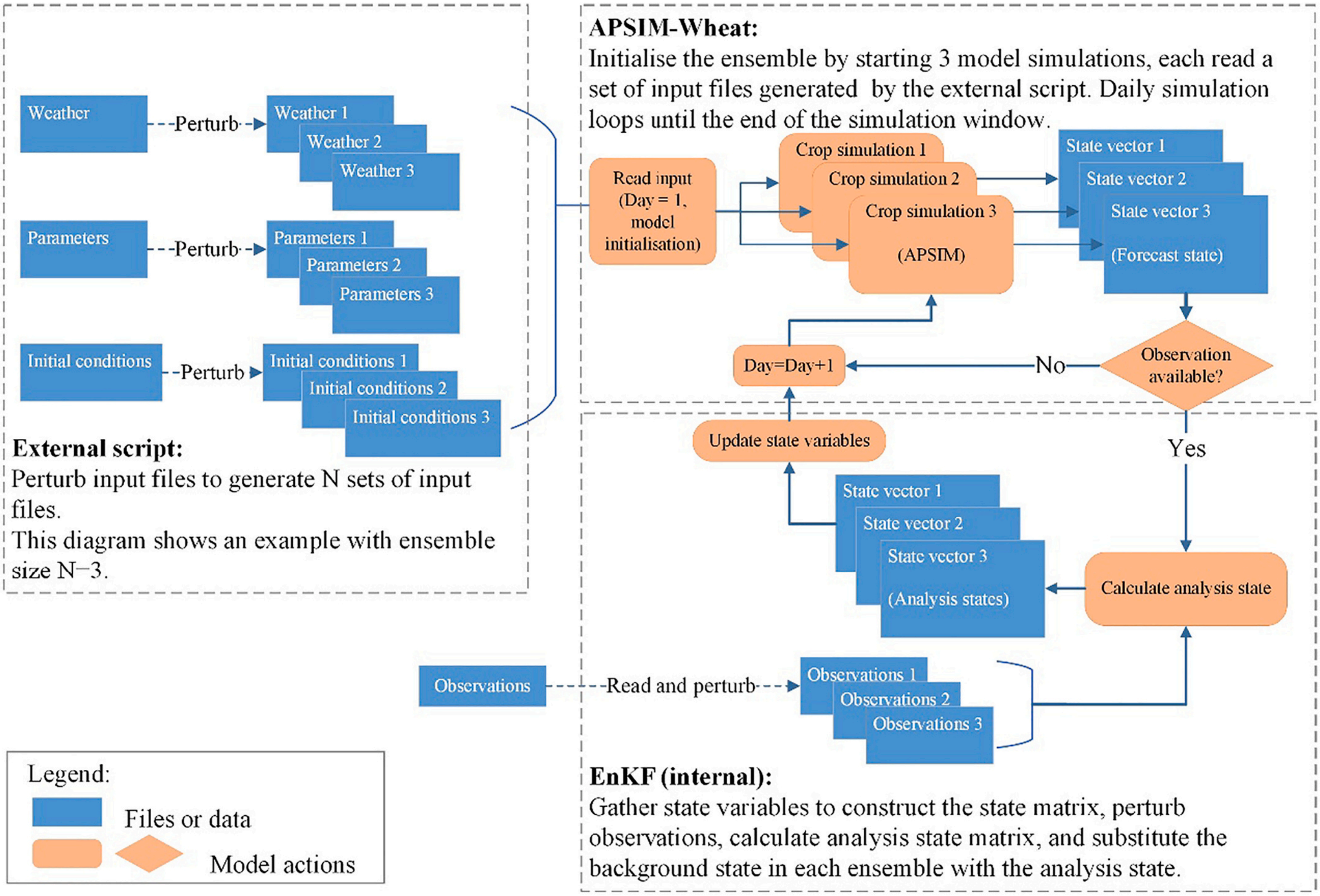
$$\mu_k^f = E\left[\mathbf{x}_k^{i,f}\right] = \frac{1}{N}\sum_{i=0}^{N}\mathbf{x}_k^{i,f} \tag{8}$$

$$\mu_k^a = E\left[\mathbf{x}_k^{i,a}\right] = \frac{1}{N}\sum_{i=0}^{N}\mathbf{x}_k^{i,a} \tag{9}$$

for forecast and analysis states of the ith ensemble $\mathbf{x}_k^i$, respectively.

2.3. APSIM-EnKF framework

Development of the APSIM-EnKF data assimilation framework (source code provided on: https://github.com/yuxi-research/APSIM-EnKF) was based on the version APSIM Next Generation (APSIMX, Holzworth et al., 2018) that was under development on GitHub in 2016, when the version APSIM 7.5 (Holzworth et al., 2014) was migrated to the new version with the core code for the model physics remaining unchanged. The framework was developed as a built-in module of the model, working cooperatively with an external program to generate and perturb input files for the EnKF. This data assimilation framework is extendable to all future plant modules and can be switched to other state updating data assimilation algorithms with minor modification to the source code.



APSIM-Wheat:
Initialise the ensemble by starting 3 model simulations, each read a set of input files generated by the external script. Daily simulation loops until the end of the simulation window.

External script:
Perturb input files to generate N sets of input files.
This diagram shows an example with ensemble size N=3.

EnKF (internal):
Gather state variables to construct the state matrix, perturb observations, calculate analysis state matrix, and substitute the background state in each ensemble with the analysis state.

Fig. 1. Schematic diagram of the data assimilation framework, coupling an EnKF with APSIM-Wheat for an example of 3 ensemble members.

Fig. 1 shows the schematic diagram of the data assimilation framework. Ensembles were generated by adding Gaussian errors to weather, model parameters and initial conditions, using a script running external to the model and read as input data at the beginning of the model simulation. After initialisation, the ensemble of models is run in parallel and looped at a daily timestep. Whilst APSIM allows multiple simulations to be run, this framework uses a multi-threaded environment to ensure the simulations to be synchronised at the end of each loop, and that state values from the ensemble can communicate. Therefore, at the end of each timestep, each simulation in the ensemble pauses and waits until all simulations finish their calculation of the day. Then an observation availability check is performed: if an observation of the model state is available the EnKF module is invoked. With the forecast step finished by the daily model simulation, the EnKF module gathers the model states from all ensemble members to construct the forecast state matrix $\mathbf{x}_k^b$, calculates the forecast error covariance $\mathbf{P}_k^b$ and the Kalman gain $\mathbf{K}_k$, perturbs the observations, calculates the analysis state matrix $\mathbf{x}_k^a$, and updates the state values in the model.

An important step of the EnKF is the ensemble generation, including uncertainties in the weather forcing, model parameters, initial conditions, and model physics. Accordingly, the weather forcing was perturbed using the method described by Turner et al. (2008). This method avoids biases caused by selecting physically unrealistic perturbations from an unconstrained Gaussian distribution. According to Turner et al. (2008), weather data can be classified into three types: 'un-restricted', 'restricted', and 'semi-restricted', depending on whether the range of data is unrestricted, restricted at both bounds, or only restricted by an upper or lower bound. The forcing vector at time step k for the ith ensemble was therefore generated by:

$$\mathbf{f}_k^i = \mathbf{f}_k + \zeta_k^i + \beta^i \tag{10}$$

where $\mathbf{f}_k$ is the observed point forcing data vector at time step k, ζ_k^i is a Gaussian error with zero mean and standard deviation of σ_1, and β^i represents an offset taken as a single sample from the Gaussian distribution with zero mean and standard deviation of σ_2. Unrestricted means that the data is not physically constrained and thus will not exceed its normal range when the perturbation is applied. For the unrestricted data type (e.g., air temperature), the standard deviation σ_1 and σ_2 are calculated by:

$$\sigma_1 = \xi \tag{11}$$

$$\sigma_2 = \chi \tag{12}$$

The semi-restricted data type has a lower boundary $\mathbf{f}_{\min}$ to constrain the minimum value, including rainfall and radiation that must be larger than or equal to zero. The standard deviations are calculated by:

$$\sigma_1 = (\mathbf{f}_k - \mathbf{f}_{\min})\xi \tag{13}$$

$$\sigma_2 = (\mathbf{f}_k - \mathbf{f}_{\min})\chi \tag{14}$$

where ξ and χ are constants. The values of ξ and χ used in this study were taken from Turner et al. (2008) as presented in Table 3.

Parameter and initial condition ensembles were generated from the open-loop by adding Gaussian noise with a mean of zero and the standard deviation (std.) values in Table 1 & Table 2 for each quantity, albeit as a single vector of values. Although sensitivity analysis of APSIM cultivar parameters has been discussed (Zhao et al., 2014), an estimation of uncertainties that can happen in real observations or from calibration cannot be found from literature. Thus, these parameters were assumed to have an uncertainty of 5% of the parameter values (Table 2), so that the ensemble spread produced by the perturbation of parameters was large enough to represent a non-perfect model, but not excessively large. Thus, when generating the ensemble, a Gaussian noise with a mean of zero and a standard deviation equal to 5% of the parameter value was added to each cultivar parameter. Moreover, initial state values such as soil moisture were found to have a strong impact on model state and yield estimation in a sensitivity analysis, and so required careful perturbation. Therefore, the initialised states were determined by perturbation with a 4-month warm-up period to allow equilibrium conditions to be reached with suitable ensemble spread before the sowing date. Accordingly, further perturbation of initial conditions at sowing was not required.

Table 3
Parameters for uncertainty estimation of weather forcing, according to Turner et al. (2008).

Data	Unit	Restriction type	ξ	χ
Rainfall	mm	Semi-restricted with lower bound	0.25	0.25
Radiation	MJ m^{-2}	Semi-restricted with lower bound	0.864	0.864
Temperature	°C	Unrestricted	1.4	0.6

Observational uncertainties are also required for the assimilation, and are the result of instrument inaccuracy and imperfect retrieval algorithms. The techniques for measuring LAI and surface soil moisture from remote sensing are mature, and so their observational uncertainties (Table 4) were aligned with those for remote sensing products and based on several validation experiments in the literature. For example, the MODIS LAI product was reported to have an uncertainty of 0.38 m^2/m^2 in grass and cereal crop areas (Tan et al., 2005), and SMOS near-surface soil moisture products were reported to have an accuracy of 0.04 m^3/m^3 (Kerr et al., 2012). Soil moisture in the sub-surface is less dynamic than surface, and thus was assumed to have a lower uncertainty (0.03 m^3/m^3). The remainder states, nitrogen amount of leaf, stem and spike, dry weight of leaf and stem, and soil nitrogen, currently require destructive sampling and/or laboratory analysis to measure. As an initial test, a small uncertainty of 5% for each state was used, by assuming the measurement was collected in a relative homogeneous field. A wide range of uncertainty levels (0 to 50%) were further tested in the study of observational accuracy impact detailed in Supplementary Material.

3. Data assimilation experiments

The synthetic study is an Observing System Simulation Experiment

Table 4
Wheat and soil state variables of APSIM included in the synthetic data assimilation.

State variable (s)	Description	Unit	Interval	Uncertainty
Wheat states				
LAI	Leaf area index	m^2 m^{-2}	8 days	0.4
LeafWt	Leaf weight	g cm^{-2}	7 days	5% of the state values
LeafN	Leaf nitrogen	g cm^{-2}	7 days	5% of the state values
StemWt	Stem weight	g cm^{-2}	7 days	5% of the state values
StemN	Stem nitrogen	g cm^{-2}	7 days	5% of the state values
PodN	Spike nitrogen	g cm^{-2}	7 days	5% of the state values
Soil states				
SM1	Volumetric soil moisture in layer 1	m^3 m^{-3}	3 days	0.04
SM2, …, SM7	Volumetric soil moisture in layer 2, …, 7	m^3 m^{-3}	3 days	0.03
NO3N1, NO3N2, …, NO3N7	Soil nitrogen in the form of nitrate in layer 1, 2, …, 7	kg ha^{-1}	7 days	5% of the state values
NH4N1, NH4N2, …, NH4N7	Soil nitrogen in the form of ammonium in layer 1, 2, …, 7	kg ha^{-1}	7 days	5% of the state values

(OSSE) procedure designed to test the performance of data assimilation and its sensitivity similar to that outlined in Moradkhani (2008). Given the fact that many variables are difficult to measure, a synthetic study makes it possible to undertake an exhaustive analysis of all state variables, although not all possibilities of uncertainty encountered when using real data are captured, including observations not being homogeneously distributed through time, not balanced, with outliners, and with a wider range. A typical synthetic experiment consists of three components (Fig. 2): a truth generation run, an open-loop run, and a data assimilation run (Curnel et al., 2011). Here the synthetic true states and outputs from the truth generation run simulation were used as a reference, and the degraded simulation without assimilation used as the open-loop run (or control run) to benchmark the assimilation (closed-loop) run improvement in performance. The procedure is shown in Fig. 2. The “true” dataset is a collection of weather, parameter and initial condition data which is assumed to be known accurately in the synthetic study. The truth generation run is achieved by running a single ‘perfect’ APSIM model to generate the model output as the ‘true’ crop status. The synthetic observations were generated by applying observational uncertainties to the true model states. The single degraded weather dataset is generated by a single random draw from a Gaussian distribution characterised by weather values as the mean and their uncertainties as the standard deviation. Degraded parameters were selected to have a small discrepancy with the truth so that the truth is enclosed by the perturbed ensemble (Tables 1 & 2). Starting from the single degraded input dataset, the open-loop and the assimilation scenarios were performed as an ensemble of stochastic simulations with different perturbed input datasets generated by adding random noise to the degraded input. The ensemble of perturbed simulations runs in parallel, with the resultant model states and output taken as the mean of all ensemble members. The EnKF algorithm plays the role of updating state variables at each observation time step by merging the model forecast states with the external observations to deliver a set of posterior states.

In a typical synthetic study, both the open- and closed-loop scenarios describe a situation where the uncertainties of weather data, parameters, initial conditions, and imperfection of the model physics are considered to contribute to the uncertainties of the model states. The simulation of both scenarios was based on the same single “degraded” input dataset generated from the truth to mimic a realistic situation, where all model input and parameters are not known accurately.

The open-loop run imitates a realistic situation where input data suffers from uncertainties. It also gives the errors in the estimated model state and output compared to the truth. The data assimilation scenario used the same ensemble of simulations with the same input data as the open-loop, but with the synthetic observations of model state variables assimilated during the model evolution. The EnKF updating process accounts for the ensemble error covariance given by the probability distribution of both simulated model states (forecast uncertainties) and their observations (observational uncertainties). Successful implementation of data assimilation can be inferred if the estimation error from the open-loop is reduced with the assimilation of external observations.

The performance of the data assimilation can depend on the assimilation set-up. For example, the EnKF outcome is affected by the assimilated state variables, ensemble size, assimilation frequency, and the timing of phenological stages when observations are available. Therefore, the data assimilation experiments presented in this work include ensemble size determination, the assimilation of single and multiple state variable types, the assimilation of state variables in different observation availability scenarios, and assimilation when the phenology is constrained.

3.1. Ensemble size determination

The EnKF uses a finite-size ensemble to evaluate the error covariance and so often causes sampling errors (Evensen, 2009). The ensemble size is ideally to be as large as possible to approximate the probability distribution of the states, but it results in a high computational cost, particularly in multi-dimensional assimilation with fine resolution pixels. The determination of ensemble size is, therefore, a trade-off

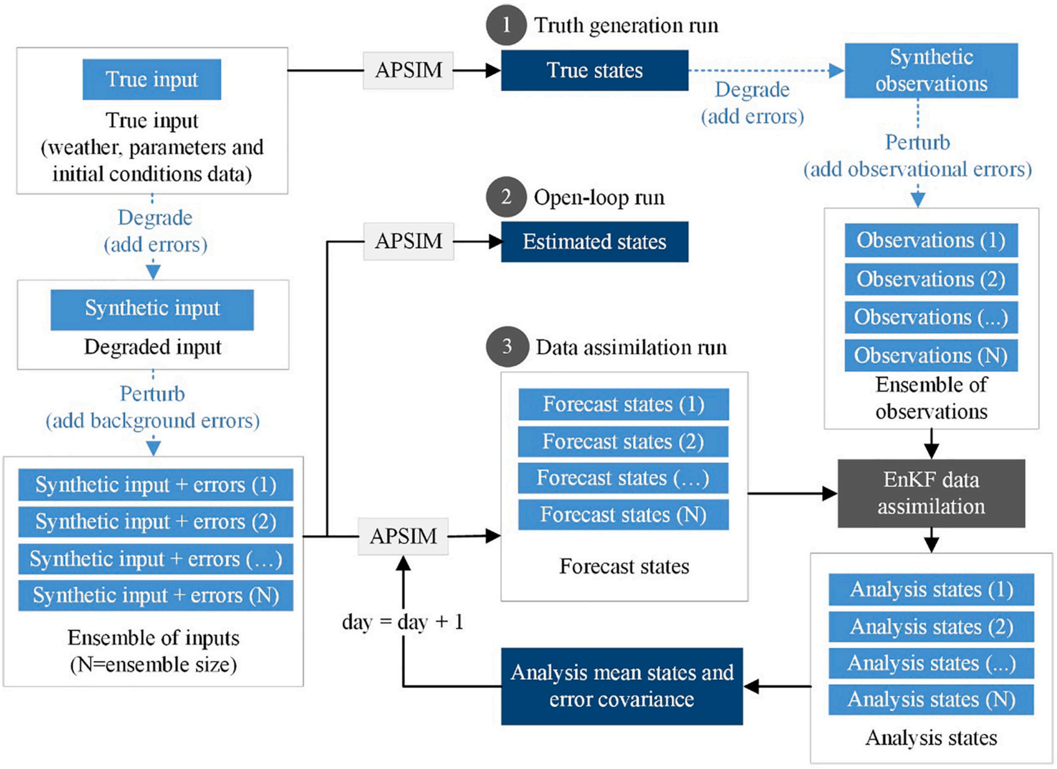


Fig. 2. Schematic of the synthetic study.

between more statistical samples and computational efficiency. This experiment aimed to determine the minimum ensemble size required for assimilation experiments with APSIM-Wheat. Six open-loop runs were tested, having ensemble sizes of 10, 20, 50, 100, 200 and 400, respectively.

3.2. Baseline assimilation scenario

This experiment aimed to explore the potential of improving yield estimation by constraining all prognostic wheat and soil state variables in APSIM. Based on the previous sensitivity analysis in Zhang (2020), the wheat and soil states that were found to have an impact on the APSIM output are presented in Table 4.

Wheat and soil are two separate modules in APSIM, and the impact of assimilating state variables is therefore discussed according to those two separate groups. The assimilation of single wheat states included LAI, leaf weight (LeafWt), leaf nitrogen (LeafN), stem weight (StemWt), stem nitrogen (StemN), and spike nitrogen (PodN), while soil states included soil moisture, ammonium, and nitrites in the top, medium, and bottom layers (layers 1, 4, and 7, respectively) and all seven soil layers. The combined data assimilation was conducted by assimilating several wheat and soil states together. The frequency and uncertainties of state variables presented in Table 4 were applied to all assimilation experiments if not specially clarified, and is referred to as the "reference" configuration for observation availability hereafter. The reference availability of soil states is the whole simulation period, while wheat states were only taken to be available during phenology phase 4 (end of juvenile) to phase 7 (end of grain filling), because the assimilation of wheat states in early stages can cause failure of the model when the wheat states are updated to a value close to zero.

In realistic situations, the measurement of observations is usually not always available at the same time. For example, LAI observations from satellite missions are available every 8 or 16 days (e.g., MODIS and Landsat) and depend on cloud cover, while near-surface soil moisture is usually available every 2 to 4 days (e.g., SMOS and SMAP), depending on the latitude. Thus, the synthetic observation data used here for the baseline experiment were assumed to be available at a specified time interval by sampling the observations from a continuous daily time series with specified acquisition intervals (Table 4).

3.3. Observation-limited scenario and phenology-constrained scenario

In addition to the baseline scenarios, two scenarios were explored: an observation-limited scenario and a phenology-constrained scenario.

The availability of remote sensing observations varies in frequency, accuracy and weather conditions. Therefore, the observation-limited scenario aimed to explore the minimum requirement of observation availability (sapling interval, phenological stages and observational accuracy) where assimilation was capable of providing a more accurate yield estimation. With phenology being a key feature controlling crop development in APSIM, it was also investigated whether the assimilation of phenology stage itself reduced model uncertainties in a phenology-constrained scenario. The description, results and discussion of the observation-limited and phenology-constrained scenarios are elaborated in the Supplementary Material Sections 1 and 2, respectively.

3.4. Evaluation of data assimilation results

The outcomes of the data assimilation experiments were evaluated with the root mean square error (RMSE) of the state variables, as an indicator of state variable estimation accuracy, and the relative difference of yield ($\mathrm{RD_{yield}}$, note that the yield refers to the grain weight at harvest), as an indicator of yield estimation accuracy, expressed as:

$$\mathrm{RMSE} = \frac{1}{\mathrm{L}}\sum\nolimits_{\mathrm{k}=1}^{\mathrm{L}}\left(\mathbf{X}_{\mathrm{k}}^{\mathrm{est}} - \mathbf{X}_{\mathrm{k}}^{\mathrm{true}}\right) \tag{15}$$

$$\mathrm{RD_{yield}} = \frac{\mathrm{yield_{est}} - \mathrm{yield_{true}}}{\mathrm{yield_{true}}} \tag{16}$$

where L is the total time step. The estimated states $\mathbf{X}_{\mathrm{k}}^{\mathrm{est}}$ is the analysis ensemble mean for the assimilation or open-loop run while the $\mathbf{X}_{\mathrm{k}}^{\mathrm{true}}$ is the true states at time step k from the truth generation run. The $\mathrm{yield_{est}}$and $\mathrm{yield_{true}}$ are the estimated and true grain weight in kg $\mathrm{ha^{-1}}$ at the date of harvest, respectively. Therefore, a value of $\mathrm{RD_{yield}}$ close to zero means that the yield was close to the truth. An absolute value of $\mathrm{RD_{yield}}$ from a data assimilation experiment being less than that from the open-loop indicates that the assimilation of external observations contributed to a better yield estimation, compared to no observations assimilated. A negative value indicates that the yield estimation error was over-corrected.

4. Results and discussion

4.1. Ensemble size

An ensemble size of 50 was found to adequately represent the probability distribution of the stochastic implementation of APSIM-Wheat model (Fig. 3), being a reasonable approximation of ensemble spread obtained from a larger ensemble. Although the difference of standard deviation estimated for SM1 from different ensemble sizes was generally small, there was an underestimation at smaller standard deviations that was consistent with even an ensemble size of 100 members, when compared with 200 and 400 ensembles. In other crop model ensemble size experiments in literature, ensemble sizes of 50 (Wit and Van Diepen, 2007) and 70 (Ma et al., 2013) were recommended for the WOFOST model and 100 (Nearing et al., 2012) were recommended for the CERES-Wheat model. Therefore, the ensemble size of 50 was selected for the data assimilation experiments in the remainder of this paper.

4.2. Assimilation of wheat and soil states

4.2.1. Correlation between wheat and soil states

A strong correlation was found to exist within each group of state variables, meaning that the assimilation of any of the states in the wheat or soil group was able to correct the estimation errors of other states in the same group, while the impact on states in the other group was negligible. With the assimilation of a single wheat state type, the RMSE of all wheat state variables were reduced compared to the open-loop simulation, but with no distinct reduction found in the RMSE of soil states (Table 5). In the LAI assimilation experiment (Fig. 4-b), the time series of LAI were updated to approach the truth with the introduction of external LAI observations into the model, but the values of soil moisture were only slightly changed at the grain-filling stage (phenological stage 7). Similarly, assimilation of either the soil moisture, ammonium or nitrate state variables provided a better estimation of the remaining soil state types in all seven layers, but with little impact on the wheat state variables. In the SM1 assimilation experiment (Fig. 5), the posterior LAI was slightly improved from the prior values at each time step the SM1 observations were assimilated. As a result, LAI had a slightly better estimate during the leaf growth stage but was over-estimated afterwards when the leaves senesced.

The weak correlation between wheat and soil states can be explained by the weak link between the wheat and soil modules. The two modules were primarily linked by extractable water and nutrients, and the mechanism of water and nutrients impact on wheat growth is cumulative and slow. Thus, abrupt change of states in one module did not immediately affect the value of states in the other module. The sensitivity analysis (Zhang, 2020) also showed that changing the states in either the wheat or the soil group only affected the other group in stage 6–7, giving a weak correlation of errors between the two groups.

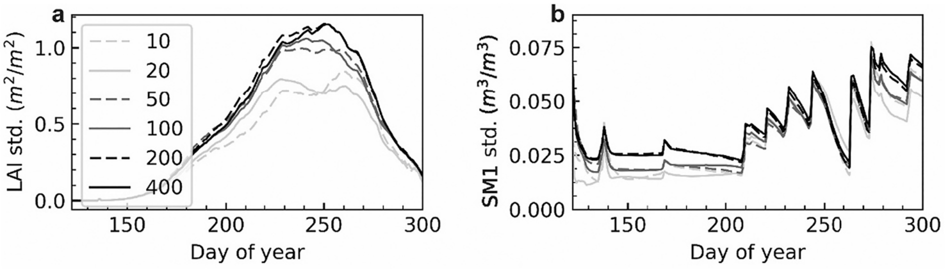


Fig. 3. Ensemble standard deviation (std.) of LAI (a) and SM1 (b) estimation from open-loop simulations with different ensemble size.

Therefore, the states in one group were not considerably affected when assimilating states in the other group.

4.2.2. *Impact on yield estimation*

By assimilating wheat and soil states over the whole growing season, LAI, LeafWt, StemWt and LeafN, among the wheat states, were found to improve yield estimation. In Table 5, the error of the estimated yield caused by uncertainties was partially corrected by the assimilation of these states, as found in the relative difference (RD) of yield reduced from 10.1% in the open-loop to 0.8%, −2.6%, 1.7% and 3.8%, respectively.

The assimilation of soil states contributed to a better yield estimation, and the assimilation of soil states in the top layers was more effective than the bottom layers in improving yield estimation. Nearly all relative difference of yield given by the assimilation of soil states in the different soil layers was found to be less than the open-loop result, indicating a better yield estimation with data assimilation. It is intuitive to understand that errors of yield estimation caused by the rainfall uncertainty could be corrected by updating soil moisture. However, the better yield estimation caused by assimilating soil nitrogen states was likely to be due to the associated improvement in soil moisture, which is the key state variable affecting APSIM yield estimation according to the previous sensitivity analysis (Zhang, 2020), with the soil moisture being better estimated due to the strong association among state variables in the soil group.

A better time series of GrainWt was found to not necessarily be linked to a better yield at harvest, and vice versa. An example was in the assimilation of some wheat state types (e.g., GrainWt, StemN, PodN) which were found to have a lower RMSE for the GrainWt estimation. However, the yield at harvest remained nearly unchanged compared to the open-loop. Similarly, in the soil moisture assimilation: although the yield at harvest was well-fitted to the truth, no direct update of GrainWt from assimilation of soil moisture was found during the grain filling stages. This was likely due to the physics applied in APSIM for modelling the grain filling: the grain demand is determined by StemWt at flowering, cultivar parameters and stress factors affected by temperature and nitrogen, and that once the grain demand is met, the daily incoming biomass is not allocated to grain but to other organs. Therefore, even if GrainWt is accumulated daily, the yield at harvest is constrained by a maximum GrainWt under the impact of other factors.

4.2.3. *Impact of assimilating combined states*

The combined assimilation of multiple wheat state types improved the estimation of all wheat states and gave a more correct yield at harvest than the open-loop. This was found in assimilating the combination of two or three of the wheat states, where the RMSE of all wheat states and the RD of yield were lower than the open-loop (Table 5), indicating a better-estimated wheat state combination and a partially corrected yield. When all three states in the LAI, LeafWt and StemWt were assimilated, the wheat states and yield estimation was found to be substantially better than only assimilating two or one of them (Table 5). Like the combination of wheat states, the assimilation of multiple soil state types improved the estimation of all soil states and gave a more correct yield at harvest, as the RMSE of all soil states were reduced, and the yield was closer to the truth (Table 5).

The combined assimilation of mixed wheat and soil state types improved the estimation of almost all the wheat and soil states than the open-loop, but the yield was sometimes over-corrected. For instance, combined assimilation of LAI and SM1 (Fig. 6) resulted in the RMSE of all state variables being smaller than the open-loop, but the RD of crop yield resulting from the combined assimilation was lower than zero (Table 5). This over-correction indicates a conflict between the state variables in the wheat and soil groups: when the assimilation of one group of states impacts the estimated yield, but the states in the other group remained almost unchanged. Therefore, solely assimilating either wheat or soil states led to a lower yield estimation that approached the truth, while the combined assimilation seemed to amplify this reduction and gave an underestimate even greater than the overestimate in the open-loop. The cause of this conflict is considered to be two-fold. First, when LAI was assimilated while not constraining other wheat states, the other wheat states may be pushed in a wrong direction if the correlation in the errors is wrongly estimated and thus leads to worse yield estimation. Second, although the data assimilation was able to constrain and better estimate some states, the uncertainties caused by some cultivar parameters could not be cancelled because their parameters directly control the grain filling process that is not affected by state variables.

5. Discussion on remotely sensed data availability

This paper demonstrated the potential of assimilating LAI, wheat organ weight, and soil moisture in improving yield estimation of APSIM-Wheat. The key wheat and soil states and their observation availability (phenological stage, repeat interval and accuracy) was summarised in Table 6, according to the observation-limited scenario (details presented in Supplementary Material Section 1). However, not all the wheat and soil states in the APSIM-Wheat model can be measured by mature spaceborne remote sensing techniques; some require in-situ measurements. While destructive sampling is a common approach to in-situ measurement, in-field sensors and machine-mounted GPS-enabled equipment are also increasingly used by farmers to collect crop and soil information that could be useful for data assimilation purposes. This section discusses the availability of satellite remotely sensed data of the key wheat and soil states that benefited the yield estimation in assimilation.

The existing remote sensing LAI products from satellites satisfy the accuracy requirement of LAI assimilation for APSIM-Wheat. Several studies validating LAI products with ground measurements reported the uncertainties of LAI products from a range of satellites. In moderate resolution satellite products, for instance, MODIS LAI was reported to have an uncertainty (RMSE) of 0.38 m^2/m^2 in grass and cereal crop areas (Tan et al., 2005), and 0.66 m^2/m^2 in biomes including grass, crop, shrubs, savannas and forest (Yang et al., 2006), respectively. LAI data retrieved from Sentinel-2 high resolution (up to 10 m) spectral reflectance was reported to have RMSE values of 0.69 m^2/m^2 in the multi-crop area (Pasqualotto et al., 2019). These uncertainties are in the range of 0 to 1 m^2/m^2 where a clear improvement from data assimilation was

Table 5

Relative difference (RD) of yield and root mean square error (RMSE) of wheat and soil states with the assimilation of single or combined types of state variables. RMSE values smaller than 80% of the open-loop is shaded in light grey, with that small than 50% of the open-loop shaded in dark grey.

State(s) assimilated	RD of yield	RMSE (wheat states) of							RMSE (soil states) of					
		LAI	Grain Wt	Leaf Wt	Stem Wt	Leaf N	Ste m N	Pod N	SM1	SM4	NH 4N1	NH4N 4	NO3N 1	NO3N 4
	%	m^2 m^{-2}	g cm^{-2}	g cm^{-2}	g cm^{-2}	g cm^{-2}	g cm^{-2}	g cm^{-2}	10^{-2} m^3 m^{-3}	10^{-2} m^3 m^{-3}	kg ha^{-1}	10^{-4} kg ha^{-1}	kg ha^{-1}	kg ha^{-1}
Open-loop	10.1	0.5	44	22.8	100.7	0.8	0.59	0.16	4.9	1.9	0.12	7	3.9	1.1
Assimilation of wheat states														
LAI	0.8	0.32	41.1	13.8	62.1	0.55	0.6	0.15	5	1.8	0.12	4	2.4	1.1
GrainWt	9.9	0.36	27.8	15.6	60.5	0.48	0.3	0.11	4.9	2	0.12	5.9	3.9	1.1
LeafWt	-2.6	0.19	35.7	8.4	55.1	0.35	0.58	0.14	5	1.6	0.12	3.3	2.8	1.1
StemWt	1.7	0.34	22.5	14.3	62.7	0.39	0.3	0.12	4.5	1.3	0.11	5.5	3.7	1.1
LeafN	3.8	0.3	21	13	74.8	0.27	0.27	0.14	4.6	1.4	0.11	3.9	2.4	1.1
StemN	11.8	0.41	36.5	17.6	57.2	0.58	0.27	0.12	4.8	1.9	0.12	4	2.7	1.1
PodN	11.4	0.41	35.6	17.5	67.2	0.61	0.36	0.08	4.9	2	0.12	5.3	3.9	1.1
LAI, LeafWt	4.9	0.21	33.9	9.1	51.4	0.39	0.54	0.15	4.7	1.5	0.11	3.4	2.6	1.1
LAI, StemWt	7.2	0.29	26.4	12.5	49.6	0.35	0.24	0.11	5.4	1.9	0.13	3.5	3.9	1.1
LeafWt, StemWt	6.1	0.19	25.7	8.1	42.9	0.27	0.4	0.12	4.9	1.6	0.11	2.1	2.7	1
LAI, LeafWt, StemWt	-1.0	0.2	27.2	8.4	42.4	0.32	0.36	0.12	5.1	2.1	0.12	3.1	3.1	1.1
Assimilation of soil states														
SM1	2	0.5	43	21.7	85.2	0.77	0.61	0.16	1.8	1	0.06	6.1	7.8	1
SM4	7.4	0.64	41.5	27.7	93	0.82	0.57	0.17	3.8	0.6	0.1	5.3	3.9	0.8
SM7	7.5	0.5	43	21.5	96.6	0.78	0.61	0.16	4.8	1.5	0.12	5.7	3.2	0.4
NH4N1	-2	0.58	32.4	25.6	83.4	0.59	0.5	0.19	3.3	0.7	0.05	3.8	3.5	1.2
NH4N4	12.7	0.82	43.6	35.7	102.4	1.03	0.58	0.19	4.8	1.5	0.11	2.4	3.3	1.1
NH4N7	11.3	0.56	43.5	24.2	100.6	0.86	0.62	0.16	4.7	1.7	0.12	7.2	4.9	0.8
NO3N1	-5.5	0.42	43.9	18.8	78.7	0.54	0.59	0.18	3.3	1	0.08	4.4	1.5	0.3
NO3N4	-2.2	0.54	39	23.8	54.4	0.8	0.52	0.15	4.5	1.6	0.1	5.1	3.2	0.2
NO3N7	7.8	0.49	40.9	21.3	90.6	0.78	0.61	0.15	3.5	1.5	0.09	5.6	4	0.6
SM1, NH4N1	1.9	0.5	34.1	21.9	82.3	0.64	0.51	0.17	1.3	0.7	0.05	4.3	6.6	1.1
SM1, NO3N1	-3.0	0.24	32.1	10.7	47.4	0.36	0.44	0.13	1.3	0.8	0.05	3.2	1.2	0.9
NO3N1, NH4N1	0.2	0.58	32.8	25.6	76.6	0.58	0.45	0.19	3.2	0.8	0.05	3.4	1.3	0.5
Assimilation of mixed wheat and soil states														
LAI, SM1	-5	0.3	41.9	13.3	51	0.48	0.58	0.17	1.3	0.7	0.05	4.5	2.2	1
LAI, SM1, NH4N1	-3.9	0.24	34.1	10.4	47.2	0.37	0.47	0.13	1.4	0.7	0.05	3.2	4.4	1.6
LAI, SM1, NO3N1	-3.6	0.25	36.4	11	45.1	0.42	0.49	0.14	1.6	0.8	0.06	3.4	1.2	0.6
LAI, NO3N1, NH4N1	-9.6	0.22	39.2	10	43.3	0.31	0.47	0.15	3.4	0.9	0.05	3	1.4	0.5

found. It is important to note that a maximum uncertainty of 1 m^2/m^2 is only acceptable for assimilation when wheat LAI is sufficiently large (e.g., in stages 5 or 6–7). In in early states, this uncertainty is too high and may cause the model to fail. Overall, the existing remote sensing techniques can provide LAI remote sensing data with sufficient spatial resolution and accuracy for successful LAI data assimilation in the APSIM-Wheat model.

The existing remote sensing near-surface soil moisture products from medium resolution satellite products satisfy the accuracy requirement for surface soil moisture assimilation in APSIM-Wheat, having an accuracy of around 0.04 m^3/m^3 (SMAP, Colliander et al., 2017; SMOS, Sanchez et al., 2012). Downscaling techniques can be applied to these medium resolution products to map soil moisture at finer resolutions (less than 1 km, Sabaghy et al., 2018) to match typical agricultural field sizes. A high-resolution product from Sentinel-1 retrieved surface soil moisture with RMSE of 0.08–0.12 m^3/m^3 in a vegetated area (Pulvirenti et al., 2018). As soil moisture in the top layer with an uncertainty of less than 0.15 m^3/m^3 improved yield estimation (Table 6), the existing remote sensing technology can provide surface soil moisture data with sufficient spatial resolution and accuracy SM1 data assimilation in the APSIM-Wheat model.

Plant organ weight is difficult to measure as it requires cutting the

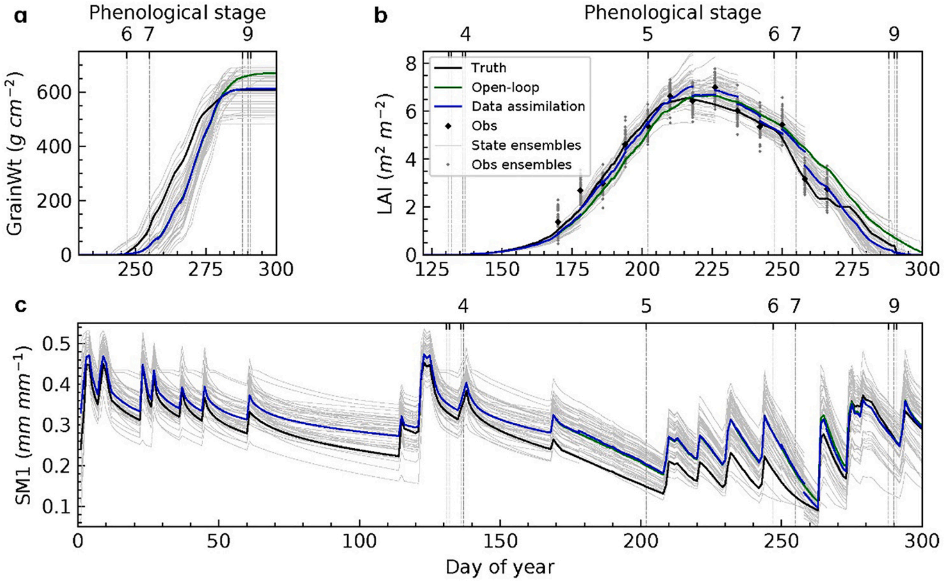


Fig. 4. Evolution of GrainWt (a), LAI (b), SM1 (c) in the LAI assimilation experiment. The legend applies to all subsequent figures.

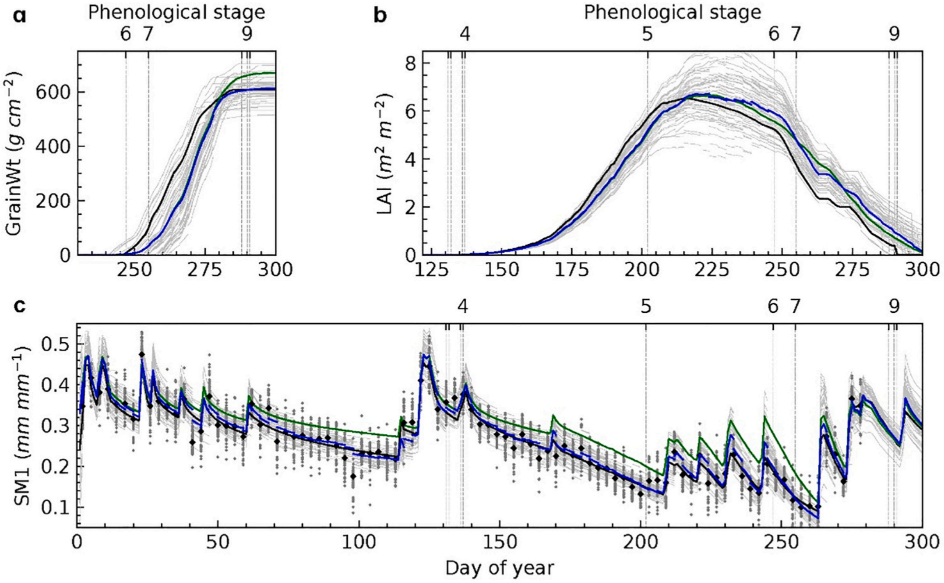


Fig. 5. As for Fig. 4 but for the SM1 assimilation experiment. See Fig. 4 for the legend.

whole plant, but the above-ground dry biomass (biomass) is measurable from remote sensing, and can be assimilated through an observational matrix linking it to the total weight of above-ground organs (leaf, stem, spike and grain). A common approach to estimate biomass from remote sensing is to use statistical regression between ground measurements of biomass and vegetation indices (VIs). For instance, Bao et al. (2009) estimated biomass for winter wheat in the vegetative stage with an RMSE of 664 kg ha^{-1}, using regression between the biomass and normalised difference vegetation index (NDVI) of Landsat TM and MODIS images. Satellite-borne remotely sensed (optical, radar) data are widely applied in the forest, grassland and rangeland biomass estimation (Kumar et al., 2015). However, cereal biomass estimation requires local training of regression models and is often limited to using Unmanned Aerial Vehicle data (e.g., barley, Bendig et al., 2014; maize, Han et al., 2019; winter wheat, Yue et al., 2017).

Techniques of detecting crop phenology from remote sensing have focused on leaf phenology for green-up, peak LAI, and senescence (e.g., Boschetti et al., 2009; Reed et al., 1994; Sakamoto et al., 2005; Vina et al., 2004; You et al., 2013; Zhang et al., 2003). Essential wheat phenological stages (anthesis and grain-filling) are currently unavailable from remote sensing. Thus, wheat phenology needs to be obtained by ground observation at the present time.

6. Conclusion

This paper presented a framework for assimilating all wheat and soil state variables into APSIM-wheat for improved wheat growth and yield estimation. Through an extensive synthetic case study, this paper demonstrated the potential of assimilating LAI, wheat organ weight, and soil moisture in improving yield estimation of APSIM-Wheat.

Under the specific weather and soil conditions assumed in this study, synthetic observations of wheat and soil states, synthetic observations of wheat and soil states generally improved the estimation of other states in APSIM-Wheat, leading to a better yield estimation. Key states that

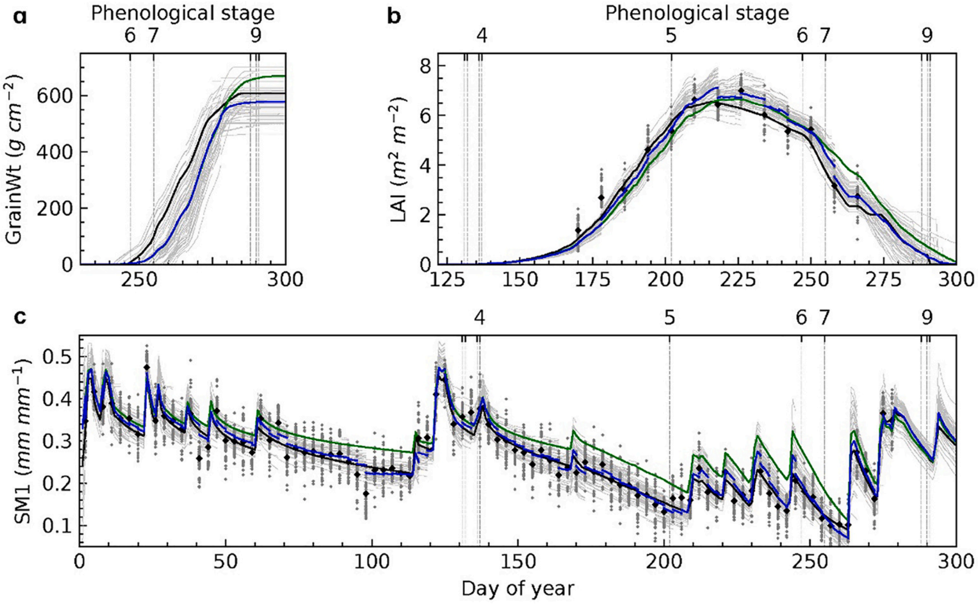


Fig. 6. As for Fig. 4 but for the combined LAI/SM1 assimilation experiment. See Fig. 4 for the legend.

Table 6
Summary of the key assimilated states that improved yield estimation, with a recommended phenological period(s), minimum observation interval, and observation accuracy for assimilation. A dash means no clear trend of improved yield was found when increasing the accuracy of this state.

State	Recommended stage, minimum assimilation interval and observation accuracy		
	Phenological state(s)	Interval	Accuracy
LAI	Stage 6–7	16 days	1 m^2/m^2
GrainWt	Stage 6–7	2 days	–
LeafWt	Stage 6–7	16 days	–
LeafN	Stage 5, 6–7	16 days	–
StemWt	Stage 6–7	16 days	10% (relative)
StemN	Stage 5, 6–7	16 days	–
SM1	Stage 6–7	16 days	0.15 m^3/m^3
NH4N1	Stage 5	16 days	–
NO3N1	Stage 5	16 days	2% (relative)
Phenology	NA	NA	NA

NA: not applicable.

improved yield estimation included leaf area index, grain weight, leaf weight, stem weight, leaf nitrogen, stem nitrogen, soil moisture, nitrate and ammonium nitrogen. A summarising table (Table 6) was presented for the recommended phenological period(s), minimum observation interval and accuracy of the key states.

Among the state variables that improved yield estimation in this study, LAI and surface soil moisture are already routinely obtained from current remote sensing techniques (e.g., LAI from Landsat/MODIS/Sentinel-2 and surface soil moisture from SMOS/SMAP/Sentinel-1) with sufficient temporal repeat and accuracy. Thus, LAI and surface soil moisture are the most promising states for assimilation. However, wheat organ weight states (leaf and stem weight) provided good yield estimation, and so biomass is expected to be a promising state for APSIM-Wheat assimilation if the biomass remote sensing techniques become mature in the future.

The combined assimilation of mixed wheat and soil state types generally improved the estimation of all the wheat and soil states relative to the open-loop, but was not better in estimating a specific state when compared to the individual assimilation of this state. Moreover, the yield from the combined assimilation was sometimes over-corrected, while the assimilation of a single state type provided good yield estimation. Therefore, for the purpose of yield estimation, it is recommended to assimilate only one state among leaf area index, biomass (as a sum of above-ground plant organ weight), and surface soil moisture to avoid over-correction. When focusing on developing insights of wheat evolution, collective assimilation of multiple states is recommended.

The constraint of phenology reduced uncertainties caused by temperature and cultivar parameters, leading to a better yield estimation (elaborated in the phenology-constrained scenario, Supplementary Material Section 2). However, the detection of wheat phenological stages is rarely found in literature, especially those specially defined by the APSIM-Wheat model that is slightly different from popular phenology scales such as Zadoks et al. (1974). The forcing of phenology presented in this study was a preliminary direct insertion assimilation attempt based on a simple assumption that the observation is accurate. In a realistic study, advanced data assimilation methods that take observation uncertainties into account can be a superior approach for phenology assimilation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We thank Monash University for the Faculty of Engineering International Postgraduate Research Scholarship to cover the doctoral tuition fees, and for the solid and persistent support in completing the research. Acknowledgment is made to the APSIM Initiative which takes responsibility for quality assurance and a structured innovation programme for APSIM's modelling software, which is provided free for research and development use (see www.apsim.info for details).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.agsy.2022.103456.

References

Ahmed, M., Akram, M.N., Asim, M., Aslam, M., Hassan, F.-U., Higgins, S., Stöckle, C.O., Hoogenboom, G., 2016. Calibration and validation of APSIM-wheat and CERES-Wheat for spring wheat under rainfed conditions: models evaluation and application. Comput. Electron. Agric. 123, 384–401.

Asseng, S., Keating, B., Fillery, I., Gregory, P., Bowden, J., Turner, N., Palta, J., Abrecht, D., 1998. Performance of the APSIM-wheat model in Western Australia. Field Crop Res. 57, 163–179.

Asseng, S., Van Keulen, H., Stol, W., 2000. Performance and application of the APSIM Nwheat model in the Netherlands. Eur. J. Agron. 12, 37–54.

Asseng, S., Turner, N.C., Botwright, T., Condon, A.G., 2003. Evaluating the impact of a trait for increased specific leaf area on wheat yields using a crop simulation model. Agron. J. 95, 10–19.

Bao, Y., Gao, W., Gao, Z., 2009. Estimation of winter wheat biomass based on remote sensing data at various spatial and spectral resolutions. Frontiers of earth science in China 3, 118.

Batchelor, W.D., Basso, B., Paz, J.O., 2002. Examples of strategies to analyze spatial and temporal yield variability using crop models. Eur. J. Agron. 18, 141–158.

Bendig, J., Bolten, A., Bennertz, S., Broscheit, J., Eichfuss, S., Bareth, G., 2014. Estimating biomass of barley using crop surface models (CSMs) derived from UAV-based RGB imaging. Remote Sens. 6, 10395–10412.

Boschetti, M., Stroppiana, D., Brivio, P.A., Bocchi, S., 2009. Multi-year monitoring of rice crop phenology through time series analysis of MODIS images. Int. J. Remote Sens. 30, 4643–4662.

Bouman, B., 1995. Crop modelling and remote sensing for yield prediction. NJAS wageningen journal of life sciences 43, 143–161.

Chen, Y., Zhang, Z., Tao, F., 2018. Improving regional winter wheat yield estimation through assimilation of phenology and leaf area index from remote sensing data. Eur. J. Agron. 101, 163–173.

Chipanshi, A.C., Ripley, E.A., Lawford, R.G., 1997. Early prediction of spring wheat yields in Saskatchewan from current and historical weather data using the CERES-wheat model. Agric. For. Meteorol. 84, 223–232.

Colliander, A., Jackson, T.J., Bindlish, R., Chan, S., Das, N., Kim, S., Cosh, M., Dunbar, R., Dang, L., Pashaian, L., 2017. Validation of SMAP surface soil moisture products with core validation sites. Remote Sens. Environ. 191, 215–231.

Curnel, Y., de Wit, A.J., Duveiller, G., Defourny, P., 2011. Potential performances of remotely sensed LAI assimilation in WOFOST model based on an OSS experiment. Agric. For. Meteorol. 151, 1843–1855.

Eitzinger, J., Trnka, M., Hösch, J., Žalud, Z., Dubrovský, M., 2004. Comparison of CERES, WOFOST and SWAP models in simulating soil water content during growing season under different soil conditions. Ecol. Model. 171, 223–246.

Evensen, G., 2009. Chapter 11: Sampling strategies for the EnKF. In: Data Assimilation: The Ensemble Kalman Filter. Springer Science & Business Media, pp. 157–174.

Guo, C., Tang, Y., Lu, J., Zhu, Y., Cao, W., Cheng, T., Zhang, L., Tian, Y., 2019. Predicting wheat productivity: integrating time series of vegetation indices into crop modeling via sequential assimilation. Agric. For. Meteorol. 272-273, 69–80.

Han, L., Yang, G., Dai, H., Xu, B., Yang, H., Feng, H., Li, Z., Yang, X., 2019. Modeling maize above-ground biomass based on machine learning approaches using UAV remote-sensing data. Plant Methods 15, 1–19.

Holzworth, D.P., Huth, N.I., deVoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Chenu, K., van Oosterom, E.J., Snow, V., Murphy, C., Moore, A.D., Brown, H., Whish, J.P.M., Verrall, S., Fainges, J., Bell, L.W., Peake, A.S., Poulton, P.L., Hochman, Z., Thorburn, P.J., Gaydon, D.S., Dalgliesh, N.P., Rodriguez, D., Cox, H., Chapman, S., Doherty, A., Teixeira, E., Sharp, J., Cichota, R., Vogeler, I., Li, F.Y., Wang, E., Hammer, G.L., Robertson, M.J., Dimes, J.P., Whitbread, A.M., Hunt, J., van Rees, H., McClelland, T., Carberry, P.S., Hargreaves, J.N.G., MacLeod, N., McDonald, C., Harsdorf, J., Wedgwood, S., Keating, B.A., 2014. APSIM – evolution towards a new generation of agricultural systems simulation. Environ. Model Softw. 62, 327–350.

Holzworth, D., Huth, N.I., Fainges, J., Brown, H., Zurcher, E., Cichota, R., Verrall, S., Herrmann, N.I., Zheng, B., Snow, V., 2018. APSIM next generation: overcoming challenges in modernising a farming systems model. Environ. Model Softw. 103, 43–51.

Huang, J., Sedano, F., Huang, Y., Ma, H., Li, X., Liang, S., Tian, L., Zhang, X., Fan, J., Wu, W., 2016. Assimilating a synthetic Kalman filter leaf area index series into the WOFOST model to improve regional winter wheat yield estimation. Agric. For. Meteorol. 216, 188–202.

Ines, A.V., Das, N.N., Hansen, J.W., Njoku, E.G., 2013. Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. Remote Sens. Environ. 138, 149–164.

Jin, X., Kumar, L., Li, Z., Xu, X., Yang, G., Wang, J., 2016. Estimation of winter wheat biomass and yield by combining the AquaCrop model and field hyperspectral data. Remote Sens. 8.

Jin, X., Kumar, L., Li, Z., Feng, H., Xu, X., Yang, G., Wang, J., 2018. A review of data assimilation of remote sensing and crop models. Eur. J. Agron. 92, 141–152.

Kang, Y., Özdoğan, M., 2019. Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. Remote Sens. Environ. 228, 144–163.

Kerr, Y.H., Waldteufel, P., Richaume, P., Wigneron, J.P., Ferrazzoli, P., Mahmoodi, A., Al Bitar, A., Cabot, F., Gruhier, C., Juglea, S.E., 2012. The SMOS soil moisture retrieval algorithm. IEEE Trans. Geosci. Remote Sens. 50, 1384–1403.

Kivi, M.S., Blakely, B., Masters, M., Bernacchi, C.J., Miguez, F.E., Dokoohaki, H., 2022. Development of a data-assimilation system to forecast agricultural systems: a case study of constraining soil water and soil nitrogen dynamics in the APSIM model. Sci. Total Environ. 820, 153192.

Kumar, L., Sinha, P., Taylor, S., Alqurashi, A.F., 2015. Review of the use of remote sensing for biomass estimation to support renewable energy generation. J. Appl. Remote. Sens. 9, 097696.

Langensiepen, M., Hanus, H., Schoop, P., Gräsle, W., 2008. Validating CERES-wheat under north-German environmental conditions. Agric. Syst. 97, 34–47.

Launay, M., Guerif, M., 2005. Assimilating remote sensing data into a crop model to improve predictive performance for spatial applications. Agric. Ecosyst. Environ. 111, 321–339.

Li, H., Kalnay, E., Miyoshi, T., 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography 135, 523–533.

Li, H., Chen, Z., Liu, G., Jiang, Z., Huang, C., 2017a. Improving winter wheat yield estimation from the CERES-wheat model to assimilate leaf area index with different assimilation methods and Spatio-temporal scales. Remote Sens. 9, 190.

Li, H., Jiang, Z.-W., Chen, Z.-X., Ren, J.-Q., Liu, B., Hasituya, 2017b. Assimilation of temporal-spatial leaf area index into the CERES-wheat model with ensemble Kalman filter and uncertainty assessment for improving winter wheat yield estimation. J Integr Agr 16, 2283–2299.

Liu, D., Mishra, A.K., Yu, Z., 2019. Evaluation of hydroclimatic variables for maize yield estimation using crop model and remotely sensed data assimilation. Stoch. Env. Res. Risk A. 33, 1283–1295.

Ma, H., Huang, J., Zhu, D., Liu, J., Su, W., Zhang, C., Fan, J., 2013. Estimating regional winter wheat yield by assimilation of time series of HJ-1 CCD NDVI into WOFOST–ACRM model with ensemble Kalman filter. Math. Comput. Model. 58, 759–770.

Maas, S.J., 1988. Use of remotely-sensed information in agricultural crop growth models. Ecol. Model. 41, 247–268.

Marc, B., 2014. Introduction to the Principles and Methods of Data Assimilation in the Geosciences.

Mavromatis, T., 2016. Spatial resolution effects on crop yield forecasts: an application to rainfed wheat yield in North Greece with CERES-wheat. Agric. Syst. 143, 38–48.

Mearns, L.O., Rosenzweig, C., Goldberg, R., 1992. Effect of changes in interannual climatic variability on CERES-wheat yields: sensitivity and 2× CO2 general circulation model studies. Agric. For. Meteorol. 62, 159–189.

Moradkhani, H., 2008. Hydrologic remote sensing and land surface data assimilation. Sensors (Basel) 8, 2986–3004.

Mosleh, M.K., Hassan, Q.K., Chowdhury, E.H., 2015. Application of remote sensors in mapping rice area and forecasting its production: a review. Sensors (Basel) 15, 769–791.

Nearing, G.S., Crow, W.T., Thorp, K.R., Moran, M.S., Reichle, R.H., Gupta, H.V., 2012. Assimilating remote sensing observations of leaf area index and soil moisture for wheat yield estimates: an observing system simulation experiment. Water Resour. Res. 48.

Noori, O., Panda, S.S., 2016. Site-specific management of common olive: remote sensing, geospatial, and advanced image processing applications. Comput. Electron. Agric. 127, 680–689.

Novelli, F., Spiegel, H., Sandén, T., Vuolo, F., 2019. Assimilation of Sentinel-2 leaf area index data into a physically-based crop growth model for yield estimation. Agronomy 9.

Panda, S.S., Hoogenboom, G., Paz, J.O., 2010. Remote sensing and geospatial technological applications for site-specific management of fruit and nut crops: a review. Remote Sens. 2, 1973–1997.

Pasqualotto, N., Delegido, J., Van Wittenberghe, S., Rinaldi, M., Moreno, J., 2019. Multi-crop green LAI estimation with a new simple Sentinel-2 LAI index (SeLI). Sensors (Basel) 19, 904.

Patel, H., Shroff, J.P., Shekh, A.V., Bhatt, B., 2010. Calibration and validation of CERES-wheat model for wheat in middle Gujarat region. Journal of Agrometeorology 12, 114–117.

Paustian, M., Theuvsen, L., 2017. Adoption of precision agriculture technologies by German crop farmers. Precis. Agric. 18, 701–716.

Probert, M., Dimes, J., Keating, B., Dalal, R., Strong, W., 1998. APSIM's water and nitrogen modules and simulation of the dynamics of water and nitrogen in fallow systems. Agric. Syst. 56, 1–28.

Pulvirenti, L., Squicciarino, G., Cenci, L., Boni, G., Pierdicca, N., Chini, M., Versace, C., Campanella, P., 2018. A surface soil moisture mapping service at national (Italian) scale based on Sentinel-1 data. Environ. Model Softw. 102, 13–28.

Reed, B.C., Brown, J.F., VanderZee, D., Loveland, T.R., Merchant, J.W., Ohlen, D.O., 1994. Measuring phenological variability from satellite imagery. J. Veg. Sci. 5, 703–714.

Rosenzweig, C., Tubiello, F.N., 1996. Effects of changes in minimum and maximum temperature on wheat yields in the central US a simulation study. Agric. For. Meteorol. 80, 215–230.

Sabaghy, S., Walker, J.P., Renzullo, L.J., Jackson, T.J., 2018. Spatially enhanced passive microwave derived soil moisture: capabilities and opportunities. Remote Sens. Environ. 209, 551–580.

Sadras, V., Cassman, K., Grassini, P., Bastiaanssen, W., Laborte, A., Milne, A., Sileshi, G., Steduto, P., 2015. Yield Gap Analysis of Field Crops: Methods and Case Studies.

Sakamoto, T., Yokozawa, M., Toritani, H., Shibayama, M., Ishitsuka, N., Ohno, H., 2005. A crop phenology detection method using time-series MODIS data. Remote Sens. Environ. 96, 366–374.

Sanchez, N., Martínez-Fernández, J., Scaini, A., Perez-Gutierrez, C., 2012. Validation of the SMOS L2 soil moisture data in the REMEDHUS network (Spain). IEEE Trans. Geosci. Remote Sens. 50, 1602–1611.

Shaw, R., Lark, R., Williams, A., Chadwick, D., Jones, D., 2016. Characterising the within-field scale spatial variation of nitrogen in a grassland soil to inform the efficient design of in-situ nitrogen sensor networks for precision agriculture. Agric. Ecosyst. Environ. 230, 294–306.

Silvestro, P., Pignatti, S., Pascucci, S., Yang, H., Li, Z., Yang, G., Huang, W., Casa, R., 2017. Estimating wheat yield in China at the field and district scale from the assimilation of satellite data into the Aquacrop and simple algorithm for yield (SAFY) models. Remote Sens. 9.

Stuart, A., Zygalakis, K., 2015. Data Assimilation: A Mathematical Introduction. Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States).

Tan, B., Hu, J., Zhang, P., Huang, D., Shabanov, N., Weiss, M., Knyazikhin, Y., Myneni, R. B., 2005. Validation of moderate resolution imaging Spectroradiometer leaf area index product in croplands of Alpilles, France. Journal of Geophysical Research: Atmospheres 110.

Thorp, K., Wang, G., West, A., Moran, M., Bronson, K., White, J., Mon, J., 2012. Estimating crop biophysical properties from remote sensing data by inverting linked radiative transfer and ecophysiological models. Remote Sens. Environ. 124, 224–233.

Timsina, J., Humphreys, E., 2006. Performance of CERES-Rice and CERES-wheat models in rice–wheat systems: a review. Agric. Syst. 90, 5–31.

Turner, M., Walker, J., Oke, P., 2008. Ensemble member generation for sequential data assimilation. Remote Sens. Environ. 112, 1421–1433.

Vazifedoust, M., Van Dam, J., Bastiaanssen, W., Feddes, R., 2009. Assimilation of satellite data into agrohydrological models to improve crop yield forecasts. Int. J. Remote Sens. 30, 2523–2545.

Vina, A., Gitelson, A.A., Rundquist, D.C., Keydan, G., Leavitt, B., Schepers, J., 2004. Monitoring maize (Zea mays L.) phenology with remote sensing. Agron. J. 96, 1139–1147.

Wiegand, C.L., Richardson, A.J., Jackson, R.D., Pinter, P.J., Aase, J.K., Smika, D.E., Lautenschlager, L.F., McMurtrey, J., 1986. Development of agrometeorological crop model inputs from remotely sensed information. IEEE Trans. Geosci. Remote Sens. 90–98.

Wit, De, Van Diepen, C., 2007. Crop model data assimilation with the ensemble Kalman filter for improving regional crop yield forecasts. Agric. For. Meteorol. 146, 38–56.

Xiong, W., Conway, D., Holman, I., Lin, E., 2008. Evaluation of CERES-wheat simulation of wheat production in China. Agron. J. 100, 1720–1728.

Yang, W., Tan, B., Huang, D., Rautiainen, M., Shabanov, N.V., Wang, Y., Privette, J.L., Huemmrich, K.F., Fensholt, R., Sandholt, I., 2006. MODIS leaf area index products: from validation to algorithm improvement. IEEE Trans. Geosci. Remote Sens. 44, 1885–1898.

You, X., Meng, J., Zhang, M., Dong, T., 2013. Remote sensing based detection of crop phenology for agricultural zones in China using a new threshold method. Remote Sens. 5, 3190–3211.

Yue, J., Yang, G., Li, C., Li, Z., Wang, Y., Feng, H., Xu, B., 2017. Estimation of winter wheat above-ground biomass using unmanned aerial vehicle-based snapshot hyperspectral sensor and crop height improved models. Remote Sens. 9, 708.

Zadoks, J.C., Chang, T.T., Konzak, C.F., 1974. A decimal code for the growth stages of cereals. Weed Res. 14, 415–421.

Zhang, Y., 2020. Towards Improved Crop Growth and Yield Estimation: Observation Constrained Wheat Modelling. Department of Civil Engineering. Monash University.

Zhang, X., Friedl, M.A., Schaaf, C.B., Strahler, A.H., Hodges, J.C.F., Gao, F., Reed, B.C., Huete, A., 2003. Monitoring vegetation phenology using MODIS. Remote Sens. Environ. 84, 471–475.

Zhang, Y., Feng, L., Wang, E., Wang, J., Li, B., 2012. Evaluation of the APSIM-wheat model in terms of different cultivars, management regimes and environmental conditions. Can. J. Plant Sci. 92, 937–949.

Zhang, W., Liu, W., Xue, Q., Chen, J., Han, X., 2013. Evaluation of the AquaCrop model for simulating yield response of winter wheat to water on the southern loess plateau of China. Water Sci. Technol. 68, 821–828.

Zhang, Y., Walker, J.P., Pauwels, V.R.N., Sadeh, Y., 2022. Assimilation of wheat and soil states into the APSIM-wheat crop model: a case study. Remote Sens. 14, 65.

Zhao, Y., Chen, S., Shen, S., 2013. Assimilating remote sensing information with crop model using ensemble Kalman filter for improving LAI monitoring and yield estimation. Ecol. Model. 270, 30–42.

Zhao, G., Bryan, B.A., Song, X., 2014. Sensitivity and uncertainty analysis of the APSIM-wheat model: interactions between cultivar, environmental, and management parameters. Ecol. Model. 279, 1–11.

Zheng, B., Chenu, K., Doherty, A., Chapman, S., 2014. The APSIM-wheat module (7.5 R3008).