5-10-2022

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

Marissa S. Kivi
University of Illinois Urbana-Champaign

Bethany Blakely
University of Illinois Urbana-Champaign

Michael Masters
University of Illinois Urbana-Champaign

Carl J. Bernacchi
USDA ARS

Fernando E. Miguez
Iowa State University

See next page for additional authors

Science of the Total Environment 820 (2022) 153192

Contents lists available at ScienceDirect

Science of the Total Environment

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

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

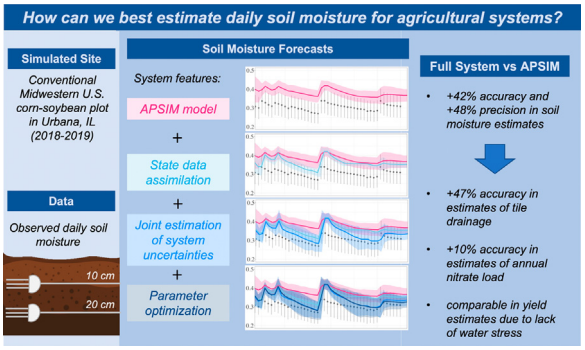
Marissa S. Kivi [a,*], Bethany Blakely [b,c], Michael Masters [b,d,e], Carl J. Bernacchi [a,b,f], Fernando E. Miguez [g], Hamze Dokoohaki [a]

[a] Department of Crop Sciences, University of Illinois at Urbana-Champaign, Turner Hall AW-101, 1102 S Goodwin Ave, Urbana, IL 61801, USA
[b] Department of Plant Biology, University of Illinois at Urbana-Champaign, Morrill Hall, 505 S. Goodwin Ave, Urbana, IL 61801, USA
[c] Center for Advanced Bioenergy and Bioproducts Innovation, University of Illinois at Urbana-Champaign, 1206 W. Gregory Drive, Urbana, IL 61801, USA
[d] Institute for Sustainability, Energy and Environment, University of Illinois at Urbana-Champaign, 1101 W. Peabody, Suite 350, Urbana, IL 61801, USA
[e] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, 1206 W. Gregory Drive, Urbana, IL 61801, USA
[f] Global Change and Photosynthesis Research, USDA-ARS, Urbana, IL 61801, USA
[g] Department of Agronomy, Iowa State University, Agronomy Hall 1206, 716 Farm House Ln, Ames, IA 50011, USA

HIGHLIGHTS

- Current tools in crop forecasting are limited in accuracy and/or precision.
- We develop a robust data assimilation system to improve APSIM model forecasts.
- Estimating uncertainties and model parameters enhances assimilation performance.
- Soil moisture data assimilation improves modeling of nitrate leaching.
- Without water stress, yield estimates are largely unaffected by assimilation.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 16 November 2021
Received in revised form 10 January 2022
Accepted 12 January 2022
Available online 19 January 2022

Editor: Ouyang Wei

Keywords:
Agricultural forecasting
State-parameter data assimilation
Filter divergence
APSIM
Soil moisture
Nitrate leaching

ABSTRACT

As we face today's large-scale agricultural issues, the need for robust methods of agricultural forecasting has never been clearer. Yet, the accuracy and precision of our forecasts remains limited by current tools and methods. To overcome the limitations of process-based models and observed data, we iteratively designed and tested a generalizable and robust data-assimilation system that systematically constrains state variables in the APSIM model to improve forecast accuracy and precision. Our final novel system utilizes the Ensemble Kalman Filter to constrain model states and update model parameters at observed time steps and incorporates an algorithm that improves system performance through the joint estimation of system error matrices. We tested this system at the Energy Farm, a well-monitored research site in central Illinois, where we assimilated observed in situ soil moisture at daily time steps for two years and evaluated how assimilation impacted model forecasts of soil moisture, yield, leaf area index, tile flow, and nitrate leaching by comparing estimates with in situ observations. The system improved the accuracy and precision of soil moisture estimates for the assimilation layers by an average of 42% and 48%, respectively, when compared to the free model. Such improvements led to changes in the model's soil water and nitrogen processes and, on average, increased accuracy in forecasts of annual tile flow by 43% and annual nitrate loads by 10%. Forecasts of aboveground measures

* Corresponding author.
E-mail addresses: mkivi2@illinois.edu (M.S. Kivi), blakely6@illinois.edu (B. Blakely), mmasters@illinois.edu (M. Masters), bernacch@illinois.edu (C.J. Bernacchi), femiguez@iastate.edu (F.E. Miguez), hamzed@illinois.edu (H. Dokoohaki).

http://dx.doi.org/10.1016/j.scitotenv.2022.153192

did not dramatically change with assimilation, a fact which highlights the limited potential of soil moisture as a constraint for a site with no water stress. Extending the scope of previous work, our results demonstrate the power of data assimilation to constrain important model estimates beyond the assimilated state variable, such as nitrate leaching. Replication of this study is necessary to further define the limitations and opportunities of the developed system.

1. Introduction

The ability to accurately and precisely forecast the behavior of agricultural systems could provide impactful insights for an array of agricultural issues, including yield gaps, drought, climate change, and nutrient losses (Reading et al., 2019; Silva and Giller, 2021). However, the predictive capacity of most tools developed for these real-world applications remains limited in accuracy, precision, or both (Dokoohaki et al., 2021). Consider process-based crop models, which have grown to be a powerful and well-recognized tool in agricultural forecasting (Jin et al., 2018; Silva and Giller, 2021). These models combine state-of-the-art knowledge on agricultural processes to more comprehensively monitor and simulate cropping systems than field experiments alone due to greater system complexity (Boote et al., 1996; Pasley et al., 2021). Nonetheless, although crop models can simulate crop growth and development in an internally consistent manner by conserving mass and energy, their weakness lies in the unaccounted uncertainties associated with their parameters and inputs (Dokoohaki et al., 2021). First, in their development, most process-based crop models are developed on top of deterministic schemes in which the uncertainties associated with model parameters and drivers are ignored. Then, later, when models are used, they are frequently unconstrained and/or hand-tuned. In the case that constraints are applied, the employed methodology is typically unable to utilize all available information (Dietze et al., 2013). Such modeling activities often focus on constraining a model to a single site with a single data product, an approach in direct contrast with the diverse range of available data products and the dimensionality of the true system (Dietze et al., 2013; Fer et al., 2021; Seidel et al., 2018).

Yet, crop models are not the only tool used to monitor agricultural systems. New technology has enabled the efficient and high-precision monitoring of agricultural field experiments, providing data on a variety of state variables like soil moisture, tile drainage flow, and leaf area index. Yet, while observational data from field experiments have been essential to improving our understanding of many processes in soil and cropping systems, analyses of field experiment data alone are often limited in dimensionality and, thus, fail to capture the complexity of high-level applied research questions ("Systems Thinking", 2020). In addition, combining measurements and data products from different instruments and experiments and across different temporal and spatial resolutions is rarely straightforward and often impossible (Dietze et al., 2013).

To overcome the limitations of these two powerful resources (i.e., process-based crop models and observational data), sequential data assimilation (SDA) has emerged as a viable solution in the crop modeling world (Jin et al., 2018). SDA fuses process-based crop models and observed agricultural data together, allowing them to speak to and build on one another despite differing temporal and spatial scales (Dietze et al., 2013). With this method, a variety of observations can be incorporated into crop models to reduce uncertainty around spatially-heterogenous and dynamic properties in agricultural systems. This increases precision and accuracy in simulations while decreasing dependence on extensive site-level model calibration (Mishra et al., 2021). The model provides a temporally continuous, high-dimensional scaffold in which observations can be smoothly integrated (Dietze et al., 2013; Liu et al., 2021). Several SDA techniques have been applied in crop simulation studies in the past (Huang et al., 2019). However, the ensemble Kalman filter (EnKF; Evensen, 2003) is one of the most popular SDA techniques for use with non-linear dynamic crop models due to its ease of implementation, its computational efficiency, and its ability to intuitively propagate uncertainty within model forecasts (Dietze, 2017; Mishra et al., 2021). At each observed time step, the filter combines information from available observed data and the model forecast distribution through the computation of an analysis distribution, which has lower uncertainty than either of the input distributions alone. One limitation of the EnKF is that its performance is highly dependent on the accurate estimation of the forecast and observation uncertainties in the system, which is a difficult task in practice due to computational limitations, time, and data availability (Huang et al., 2019). Several algorithms have been developed and tested to systematically and jointly estimate both uncertainty matrices within the EnKF system to overcome this issue (Tandeo et al., 2020). Other recent studies have advanced and generalized the EnKF by numerically solving the analysis step (in contrast to the original analytical approach) such that process error and state variables are estimated as latent variables in a fully Bayesian framework (Raiho et al., 2020). This approach adds extra flexibility by relaxing assumptions of the EnKF. All these filter improvement methods have been applied successfully with geophysical and ecosystem models (e.g., Hoffman et al., 2013; Dokoohaki et al., 2021b). However, they have yet to be employed with crop models.

Using SDA, a variety of data products have been successfully assimilated into crop models with the intention of improving model forecasts, including leaf area index (e.g., Nearing et al., 2012; Ines et al., 2013; Ma et al., 2013; Chen et al., 2018; Lu et al., 2021), biomass (e.g., Linker and Ioslovich, 2017) and evapotranspiration (e.g., Huang et al., 2015). Of these data products, soil moisture (in situ or remotely sensed) has emerged as a popular and effective choice for assimilation in crop models due to the high sensitivity of agricultural system function to soil moisture levels, as well as the natural heterogeneity of soil moisture in space (de Wit and van Diepen, 2007; Monsivais-Huertero et al., 2010; Chakrabarti et al., 2014; Mishra et al., 2021). Initially, studies that assimilated soil moisture into crop models focused on how the process impacted estimates of the assimilation state variable itself (i.e., soil moisture), as well as model estimates of crop yields (e.g., de Wit and van Diepen, 2007; Chakrabarti et al., 2014; Liu et al., 2019). Soil moisture assimilation was found to be especially beneficial for estimates of yield in water-stressed or irrigated study areas (Chakrabarti et al., 2014; Liu et al., 2021; Lu et al., 2021; Mishra et al., 2021).

Beyond crop yields, the impact of soil moisture assimilation on root-zone soil moisture estimates has also been evaluated within crop models (Monsivais-Huertero et al., 2010; Mishra et al., 2021), as well as within hydrological (Bolten et al., 2010) and land surface models (Lü et al., 2011; Wu et al., 2016; Liu et al., 2017). Lü et al. (2011) and Liu et al. (2017) determined that model estimates of root-zone soil moisture were more accurate when soil moisture states were assimilated, but optimal estimates of root-zone soil moisture were achieved when the assimilation system estimated soil hydraulic parameters in addition to the soil moisture states. Assuming uncertain dynamic model parameters to be constant in time and/or space can impose large biases in model state estimates (Hu et al., 2017). For example, soil bulk density or hydraulic conductivity are kept constant in crop models, but, in a field condition, these parameters are often dynamic due to freeze-thaw cycles or disturbances related to field operations (Quine and Zhang, 2002). To allow for variation in parameters in the EnKF, parameters can be included in the model forecast distribution and updated in the analysis time step according to their covariance with the assimilated states via the state augmentation technique (Evensen, 2009; Liu et al., 2017). Though this method has not yet been applied in soil moisture assimilation studies with crop models, its performance in hydrological models shows promise (Lü et al., 2011; Liu et al., 2017; Liu et al., 2021).

Past studies have been successful in using soil moisture assimilation as a method of constraining yield, canopy cover, and root-zone soil moisture. However, there are other important state variables related to agricultural issues that must also be considered in assimilation studies. One such variable is nitrate leaching. Over the past few decades, nitrate (NO_3) leaching from

agricultural soils has become an issue of increasing concern for the United States Midwest (Christianson et al., 2018). A shift in the region's typical agricultural practices to monoculture production systems, artificial subsurface tile drainage, excessive N fertilization, as well as an overall intensification of regional crop production, has been linked to increased NO_3 concentrations in local and downstream water sources, which is both an environmental and human health concern (Dinnes et al., 2002; Bijay-Singh and Craswell, 2021). However, current strategies to quantify agricultural NO_3 losses in the U.S. Midwest remain limited by the high costs associated with data collection and the resulting lack of direct NO_3 leaching observations (Liang et al., 2014; Gurevich et al., 2021). Limited observed data restricts not only our understanding of temporal and spatial trends but also our ability to accurately calibrate process-based models for broader areas (Liang et al., 2017; Reading et al., 2019). As a result, models are also insufficient for estimating NO_3 leaching at the regional scale.

In this study, we explore the potential of soil moisture SDA as a method to systematically and consistently improve the accuracy and precision of estimates of NO_3 leaching in a process-based crop model without the need for direct monitoring of NO_3 losses. The Agricultural Production Systems Simulator (APSIM) is a popular, well-validated, and comprehensive crop model that has been widely trusted to simulate agricultural systems in the U.S. Midwest (Keating et al., 2003; Archontoulis et al., 2014; Dokoohaki et al., 2018; Archontoulis et al., 2020) and has been used in past studies to estimate site-level (Puntel et al., 2016; Ojeda et al., 2018) and regional NO_3 leaching (Reading et al., 2019). Within APSIM, estimates of NO_3 leaching losses directly depend on estimates of tile drainage flow and soil NO_3 concentration in the lowest layer of the soil profile. APSIM's soil nitrogen (N) and soil water cycle are closely linked, such that rate factors controlling soil N transformations (i.e., denitrification, mineralization, etc.) are estimated as a function of soil moisture. Hence, both tile drainage flow and soil NO_3 concentration depend on previous model estimates of soil moisture. Based on this fact, we hypothesize that the successful assimilation of soil moisture observations into the APSIM model will constrain and improve estimates of NO_3 leaching (as well as crop yield, canopy cover, and tile drainage flow) later in the model process without the need for observing the states directly. To our knowledge, this work is the first to assimilate data into the APSIM model, the first to apply state-parameter assimilation and uncertainty estimation techniques to a crop model, and the first to explore the impact of soil moisture assimilation on crop model forecasts of several downstream processes including nitrate leaching.

This study has two main objectives:

1. To determine the optimal data assimilation scheme for constraining estimates of soil moisture in the APSIM model using in situ soil moisture observations. We hypothesize that the optimal system will increase accuracy and precision in model forecasts of the assimilated state variables (i.e., soil moisture) and will be scalable, flexible, and robust.
2. To evaluate the impact of soil moisture assimilation on the accuracy and precision of downstream model estimates including crop yield, leaf area index, tile drainage flow, and NO_3 leaching. We hypothesize that constraining uncertainty in soil moisture estimation, an upstream process, will result in lower uncertainty in downstream processes in APSIM.

2. Methods

In the following sections, we provide details on the features and data products employed in this study. In Sections 2.1 and 2.2, we introduce our study site and the data used to (1) set up the APSIM model for the study site (2.1, 2.2.1), (2) constrain APSIM estimates (2.2.2), and (3) evaluate system impacts (2.2.3, 2.2.4). Section 2.3 provides information on different features that are included and tested in our data-assimilation system and the platform on which the system rests (2.3.5). Section 2.4 outlines the configuration of the system and how the system is tested in this study, and Section 2.5 presents the metrics by which we evaluate the system.

2.1. Study site

To test our data assimilation system, we focused on the University of Illinois's Energy Farm in Urbana, IL, USA. Although this research farm has numerous experimental plots that are 4 ha. in size, all data used in this study are from the plot located at 40.06°N, −88.20°W from 2018 to 2019 (Fig. 1). This plot was selected due to the wealth of data available on soil conditions, yield, drainage, and management. It follows agricultural practices typical for maize production in the U.S. Midwest (Moore et al., 2021).

Since accurately specified management information is crucial to ensure accurate model predictions (Archontoulis et al., 2020), all known

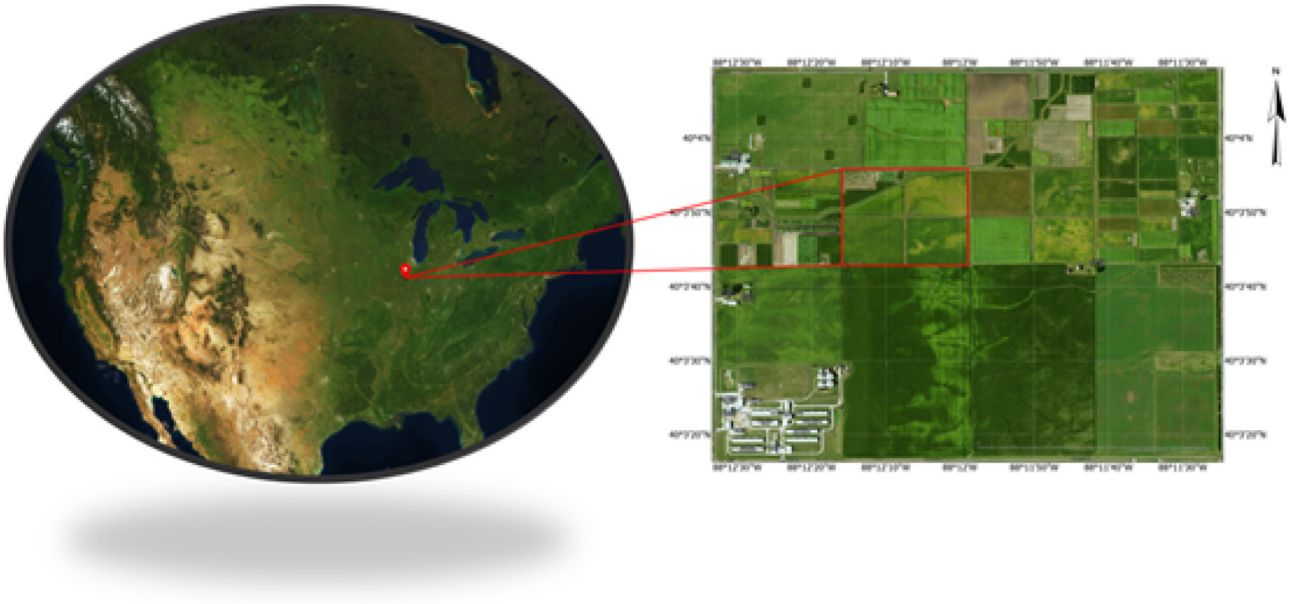


Fig. 1. Aerial image of the Energy Farm research plots outside of Urbana, IL. The “Maize Control” plot is outlined by the red square on the zoomed right panel.

management details were included as constants across our simulations. We collected management information through personal correspondence with Energy Farm personnel (Mies, personal communication, 2020). For the 2018 growing season, fertilizer was applied to the plot on the day of planting (8 May 2018) in the form of 32% liquid UAN (urea ammonium nitrate) at a rate of 202 kg/ha. Maize was planted at a rate of 8.4 plants/m^2. For the 2019 growing season, soybean was planted on 17 May at a rate of 34.6 plants/m^2, and no fertilizer was applied. Both crops were sown at a depth of 1.5 in/3.8 cm and in 76.8 cm/30 in. rows. Any residue on the plot at the beginning of each growing season was considered to be from the previous year's crop, which was maize in both cases. We did not include information on tillage, herbicide, nor pesticide practices in our simulations.

2.2. Observed data

2.2.1. Model drivers

There are two important model drivers used in our analysis—climate and soil drivers—which function to best recreate growing conditions at the Energy Farm for the years and location under study. To account for the uncertainty in these model drivers, we included 11 climate ensembles and 25 soil ensembles, which were randomly assigned to model ensembles within each simulation series. 10 of the 11 climate ensembles are products of the ERA5 dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 is a global gridded reanalysis data product that characterizes climate variables at hourly timesteps with associated uncertainties (Hersbach et al., 2020). Our derived ensembles include data on solar radiation, maximum air temperature, minimum air temperature, precipitation, and wind speed aggregated to daily resolution. The final climate ensemble was aggregated from observed weather data collected on site ("Water", 2021). Soil drivers for this analysis were derived from the SoilGrids global gridded soil database (Hengl et al., 2014) and characterize 30 soil properties, including those which define water holding capacity, soil pH, conductivity, albedo, and initial 2018 soil nutrient pools. 25 soil ensembles were generated based on the given mean and uncertainties in the SoilGrids dataset, and the depth of each soil profile was reduced to approximately the depth of the drainage tiles at the study site (i.e., roughly 1.4 m.).

2.2.2. Soil moisture

Soil moisture observations were collected at the Energy Farm plot for 2018–2019 at 30-min intervals. Measurements were taken at 5 different soil depths (i.e., 10, 20, 50, 75, and 100 cm.) using Hydra Probe II soil sensors and are measured as the volumetric water fraction at each depth (Moore et al., 2021). For the purposes of data assimilation, we focus on the measurements of soil moisture available at the 10 and 20 cm depths, which will be referred to as SM3 and SM4, respectively, hereinafter. Observations for the 75 and 100 cm depths were used for evaluation of lower-layer soil moisture estimates. Since assimilation occurs at the end of each model day, we computed the end-of-day soil moisture as the average value between the hours of 22:00 of the current day and 02:00 of the next day for each depth. These estimates constituted the state variables in our observed mean vector (Y_t) for every day where data was available. To account for instrument failures, days with fewer than 5 measurements for this 4-h period were excluded. We also removed data points from the winter months (i.e., January, February, and December) to avoid possible sensor inaccuracies related to freezing soils. Due to a low observation sample size, a 10% observation error is assumed around the mean for both depths.

2.2.3. Crop yield and LAI

Data on harvested yield for both growing seasons were measured at the time of harvest at the Energy Farm. Maize was harvested on 9 October 2018 with a yield of 13 Mg/ha, and soybean was harvested on 9 October 2019 with a yield of 4.15 Mg/ha. Maize and soybean harvests were recorded as dry grain-only biomass. Measurements of leaf area index (LAI) for the plot were collected using a LAI-2200 optical instrument at 3 different locations, approximately weekly. After removing observations without replication (i.e., $n = 1$), there are 10 and 14 LAI observations available for the 2018 and 2019 growing seasons, respectively (Bernacchi, 2020).

2.2.4. Tile flow and nitrate loads

For both growing seasons, the Energy Farm collected information on tile flow for the study plot at 15-min intervals using an area velocity sensor (pressure transducer, Hach Company, Loveland CO) to measure water height and flow speed above the weir within the drainage system. Flow was summed to give daily observed tile flow, as well as cumulative tile flow for each growing season. Tile flow data were unavailable for the study plot from 18 August 2019 until January 2020 due to sensor malfunction. However, based on data collected from nearby plots, the Energy Farm team believes the missing flow from this time period to be small relative to the year's total, so we assume there was no drainage at this time.

Measurements of NO_3 concentration in drainage waters were collected using an autosampler (American Sigma 900MAX portable sampler) that systematically collected samples at flow proportional intervals (i.e., every X number of liters). This value of X was adjusted based on historical measurements of drainage for the plot such that approximately 30 grab samples were collected each season. In practice, 29 and 42 NO_3 concentration samples were taken for the 2018 and 2019 growing seasons, respectively. These samples were filtered through a 0.45 μm membrane and analyzed by project collaborators at the University of Southampton. To calculate NO_3 loads for each 15-min interval, NO_3 concentrations were linearly interpolated between samples, multiplied by the instantaneous flow rate at each 15-min time point, averaged between the two values at the ends of each interval, and then multiplied by t. Loads were then summed to daily resolution for use within our analysis.

2.3. Data-assimilation system

2.3.1. Crop model

The Agricultural Production Systems Simulator (APSIM Classic Version 7.10) is a robust modular modeling framework which allows flexibility in management, cultivar parameterization, and model climate drivers (Holzworth et al., 2014). The model has been widely trusted as an aid for management decision making, production system design, supply chain analysis, and U.S. agricultural policy making, among other tasks (Keating et al., 2003). It has been calibrated and applied in numerous studies to simulate agricultural settings within the U.S. Corn Belt (e.g., Archontoulis et al., 2020; Pasley et al., 2021). For our APSIM simulations, we include the following available modules: *Fertiliser*, *SoilWat*, *SurfaceOM*, *SoilN*, *Soybean*, and *Maize*. Apart from those model parameters related to management (Table A.1), we made minimal changes to the model's parameterization. For the source code, we recompiled a new version of the model to allow for online communication with R statistical software (R Core Team, 2021) through RDotNet and the .NET framework. For more information on the APSIM modules and source code, see https://www.apsim.info/.

For the purposes of our analysis, the two APSIM modules controlling soil water and soil N are of particular importance. The APSIM *SoilWat* module operates as a cascading water balance model to estimate the movement of water and solutes between and across soil layers, on the soil surface (i.e., runoff and evaporation), and out of the system (i.e., drainage). It, therefore, estimates the soil moisture content of each soil layer as a balance of water input and output to the soil profile. Soil water can move between layers via three different types of flow: saturated flow, unsaturated flow, and above saturation flow. The water flow simulated by the module for each layer and each day depends on that layer's soil moisture content and how it relates to the soil moisture at saturation and the drained upper limit within that layer. Each flow type then has specified model equations and parameters which are used to calculate how much water (and relevant solutes, such as NO_3 or urea) moves between each layer. Estimates of daily drainage of soil water and dissolved NO_3 are calculated as the amount of water and solute that flow from the lowest layer in the soil profile each day. Complete mixing of the solute within the layer is assumed.

The *SoilN* module in APSIM controls N availability to plants, NO_3 concentrations in leachate, and N losses via denitrification. More specifically, the module tracks movement of nutrients through the N cycle through five processes: mineralization, immobilization, nitrification, denitrification, and urea hydrolysis. These five processes move nutrients between four pools of soil organic matter—fresh organic matter, a fast- and intermediate-decomposing pool, and an inert pool—and move N in and out of these pools into plant-available forms or, as in the case of denitrification, into system N losses. The rate at which these processes occur in a given day depends on rate factors related to daily soil moisture estimates, which are calculated within the *SoilWat* module. Fig. 2 demonstrates specifically how soil moisture affects the rate factors associated with each of these processes. Immobilization of mineral N occurs in tandem with mineralization, such that there is a balance between the N released during decomposition and microbial synthesis and humification.

2.3.2. Sequential data assimilation with the ensemble Kalman filter

Within our analysis, observed soil moisture data were assimilated into the APSIM model using an ensemble Kalman filter (EnKF). The EnKF is an extension of the Kalman filter that has been successfully employed to assimilate soil moisture data into crop models (e.g., de Wit and van Diepen, 2007; Chakrabarti et al., 2014; Liu et al., 2019; Lu et al., 2021; Mishra et al., 2021). It estimates the optimal state of the system at time t by combining the two pieces of available information—model forecasts and observed data—into an analysis distribution using Bayes' theorem.

$$P(X_t|Y_t)\tilde{P}(Y_t|X_t)\,P(X_t) \tag{1}$$

This system relies on two fundamental assumptions. First, it assumes observations (y) are related to the true state of the system (X) such that

$$y_{t=}HX_t + \varepsilon \tag{2}$$

$$\varepsilon\tilde{N}(0, R_t)$$

where H is the observation operator, connecting the model variable space to observation space. Second, the system assumes the distribution of forecasted states is Normal with mean vector X_f and covariance matrix P_f.

Founded on these assumptions, the system computes the analysis distribution (i.e., the posterior) at each time step using the Kalman Gain (K), which consequently gives the weighted mean of the forecast and observation distributions based on their respective precisions. The resulting posterior distribution is Normal with mean vector X_a and P_a.

$$K_t = P_{f,t}H^T\left(R_t + HP_{f,t}H^T\right)^{-1} \tag{3}$$

$$X_{a,t} = X_{f,t} + K_t\left(Y_t - HX_{f,t}\right) \tag{4}$$

$$P_{a,t} = (I - K_tH)P_{f,t} \tag{5}$$

Upon calculating the analysis distribution, the model forecast ensembles are updated from the analysis distribution based on their respective likelihood within the forecast distribution. The analysis distribution is thus used as the initial conditions for the model forecast into the next time step, thereby potentially constraining any model process in the next time step which depends on those adjusted initial conditions. Where data is available, the analysis step is repeated.

2.3.3. Filter tuning

Filter divergence is an issue commonly seen in data assimilation systems that rely on the ensemble Kalman filter. It occurs when observations are repeatedly rejected by the filter due to poorly estimated observation (R) and/or forecast uncertainty (P_f), which can result from low observation sample size, low ensemble size, and/or an overly confident model (Huang et al., 2019). In this case, the filter places too much weight on the forecast distribution, and, thus, neglects the observations when estimating the posterior distribution. Consequently, the correct specification of both error covariance matrices is imperative for proper filter performance (Park and Xu, 2009). Since the observation sample size for soil moisture at Energy Farm was limited at each time step (i.e., $n = 2$), we had insufficient information to accurately quantify R for our analysis at the outset of our study. Due to high computational cost, our ensemble size ($n = 50$) was also relatively

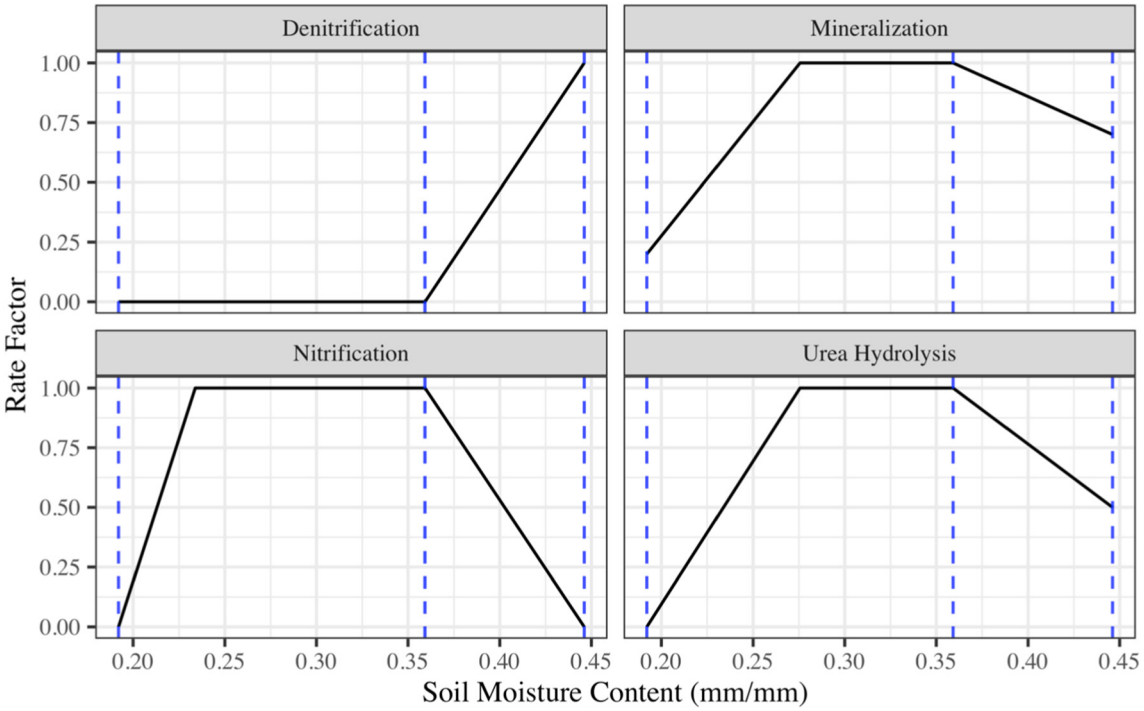


Fig. 2. Sample functions for determining soil moisture rate factors for soil N processes in APSIM based on soil moisture content. This example was generated based on the lower limit (SMC = 0.19), drained upper limit (SMC = 0.36), and saturated limit (SMC = 0.45) of Layer 3 at the study site. These limits are shown in the figure as blue, dashed vertical lines.

small, which limited our representation of P_f. To overcome these issues, we adapted a method presented by Miyoshi et al. (2013) that systematically and jointly estimates R and P_f at each analysis time step to better quantify uncertainties within the filter and avoid filter divergence.

The Miyoshi algorithm is based on innovation statistics derived as diagnostic checks for assimilation performance. At each analysis time step, it adaptively estimates a forecast inflation factor Δ and a diagonal R using known relationships between system innovations, P_f, and R. One important caveat of the algorithm rests in the circular nature of its assumptions, such that the estimation of forecast inflation depends on an accurate specification of R and vice versa. Therefore, in this study, the algorithm does not function to exactly estimate both values for a given analysis time step. Rather, the algorithm uses the estimates of all previous time steps to inform each successive analysis, allowing for the system to naturally adapt to new information and converge to optimal value ranges over the course of the simulation. We made three notable changes to the algorithm presented in Miyoshi et al. (2013) to better suit the needs of our analysis. First, we applied a constraint on estimates of P_f such that variance values never drop below 1. This ensures that we are only inflating and never shrinking forecast uncertainty. Second, we estimate a inflation matrix rather than an inflation scalar to account for possible scale differences across state variables. Only the diagonal terms of the computed inflation matrix are considered so that only forecast variance (not covariance) is inflated. Lastly, we assumed that observation errors between state variables at each time step were independent of one another, and, therefore, we only estimate the diagonal elements of R for our analysis.

In the case of our analysis, we appended the Miyoshi algorithm to our assimilation workflow as an offline estimator at time t for Δ_{t+1} and R_{t+1}. Prior to the start of assimilation, we initialize our estimates of Δ and R,

$$\Delta_1 = I$$

$$R_1 = \Sigma$$

where I is the identity matrix (only the diagonal values are relevant) and Σ is a diagonal matrix where the standard deviation of each observed variable is assumed to be 10% of the measured mean value at analysis time step $t = 1$. At each analysis time step t, Δ_t and R_t are used to compute the analysis distribution as follows:

$$K_t = \Delta_t P_{f,t} H^T \left(R_t + H\Delta_t P_{f,t} H^T\right)^{-1} \tag{6}$$

$$X_{a,t} = X_{f,t} + K_t\left(Y_t - HX_{f,t}\right) \tag{7}$$

$$P_{a,t} = (I - K_t H)\Delta_t P_{f,t} \tag{8}$$

Upon the completion of the analysis distribution at analysis timestep t, we then estimate a diagonal R using a relationship demonstrated by Desroziers et al. (2005),

$$\mathrm{E}\left(d_{o-a} d_{o-f}^T\right) = R_{est} \tag{9}$$

where $d_{o\text{-}a}$ and $d_{o\text{-}f}$ represent the observation-analysis and observation-forecast innovations for the current time step, respectively, E denotes the expectation operator. Only the diagonal values are maintained in the estimate of R as previously discussed.

Next, the algorithm employs the estimated R to estimate Δ using an equation first proposed by Wang and Bishop (2003),

$$\Delta_{est} = \frac{d_{o-f}^T\, d_{o-f} - R_{est}}{H\Delta P_f H^T} \tag{10}$$

where $d_{o\text{-}f}$ represents the observation-forecast innovations for the current time step, and P_f is the forecast covariance matrix and Δ is the inflation factor from the current time step. To preserve the forecast variance propagated by the model, we impose a lower bound of 1 on the estimated values of Δ_{est}.

Finally, the algorithm proposes values of Δ_{t+1} and R_{t+1} (i.e., values to be used in the next analysis time step) using a temporal smoother that combines the current values with the new estimates in a weighted average,

$$R_{t+1} = (\rho)R_{est} + (1-\rho)R_t \tag{11}$$

$$\Delta_{t+1} = (\rho)\Delta_{est} + (1-\rho)\Delta_t \tag{12}$$

where ρ is a user-defined weight given to the new estimate. We use $\rho = 0.05$ in our analysis to smooth noisy estimates. This ensures that a single estimate of observation error at time t will not heavily influence the error estimates informed by all previous time steps.

2.3.4. Sequential state-parameter data assimilation

In addition to the state variables, EnKF also allows for constraining model parameters such that they can be included in the state vector X_f and, thus, updated at each analysis step based on their covariance with the updated state variables (estimated in P_f; Evensen, 2009). This is a powerful function of the EnKF, as it adjusts both the initial conditions of the next model forecast and the underlying model processes generating the forecast, while state data assimilation only updates the former. Furthermore, PDA is useful because (1) it can adjust parameters that, by nature, are dynamic throughout the growing season, but are treated as constants in the model (e.g., bulk density, hydraulic conductivity parameters) and (2) it's online optimization of parameters has lower computational costs compared to classic Bayesian parameter optimization methods (e.g., Markov Chain Monte Carlo), which then also employ optimized parameters in a fixed manner (e.g., Dokoohaki et al., 2018). However, the extent to which model processes can be improved by EnKF is dependent on the parameter and the magnitude of its impact on the assimilated model states (Liu et al., 2017). To determine the parameters updated within our analysis, we considered innovations from a preliminary SDA simulation and determined which soil water flow type was associated most with large prediction error to maximize the potential for error reduction (see Supplementary Materials).

2.3.5. Data fusion framework

Our modeling framework consists of several diverse and important pieces that, together, enable us to perform comprehensive, flexible, and robust analyses using the high-performance computers on the campus cluster at the University of Illinois at Urbana-Champaign. At the base of our modeling workflow on the cluster, we use Docker containers to generate and execute each of our crop model ensemble simulations using the "parallel System for Integrating Impact Models and Sectors" (pSIMS). pSIMS is an open-source framework developed to enable large-scale ensemble simulations by integrating and translating data inputs at varying spatial scales for use with different site-based models and reformatting model output into useful and approachable datatypes (Elliott et al., 2014). The platform generates model ensembles for a given pixel location, formats site-specific drivers into model-appropriate inputs, and incorporates uncertainty through ensembles of model drivers and parameters as part of the system's "Campaign" feature. Though we established the capacity to perform regional model-data fusion exercises across broad tiled spaces by leveraging pSIMS, we utilize pSIMS functionalities for this analysis at a single pixel that best represents our study area. Additionally, for the purpose of this study, we developed new features within the pSIMS platform to perform ensemble-based simulations and include uncertainty in soil, climate, model parameters, and initial conditions for a single site. Given a fixed number of ensemble members and a series of priors on cultivar parameters, our uncertainty propagation workflow within the pSIMS platform uses a Monte Carlo sampling approach to generate random samples of soil, weather, and cultivar parameters for each ensemble member. Dokoohaki et al. (2021) describes these changes in more detail.

2.4. System set-up

The features presented in Section 2.3 comprise the fundamental pieces of our full data-assimilation system. On top of the "Dockerized" pSIMS platform, we perform crop model forecasts using APSIM, which operates in series with the ensemble Kalman filter and a filter tuning algorithm, which have been built into the model using the APSIM's C# manager functionality and are called at the end of each day's forecast. The full overall workflow is demonstrated in Fig. A.2 in the Supplementary Materials.

To evaluate our framework, we compared four series of simulations for our study site, incrementally building our data assimilation system. We refer to each of these series as a different scheme hereinafter. Table 1 outlines the different schemes and the features they include, as well as their naming protocol within this study. All schemes were completed with 50 ensembles, and each was performed separately for the 2018 and 2019 growing seasons. As demonstrated by Lu et al. (2021), this is an adequate ensemble size for achieving stability in crop model assimilation studies.

Within the model ensembles, initial soil N pools and water balance were randomized on 1 January 2018, and distributions of soil water and nutrients on 31 December 2018 were used to initialize the beginning of the 2019 model ensemble for each scheme. Like the simulation study performed by Archontoulis et al. (2020), model ensembles were begun on 1 January 2018 to initialize the soil water and nutrient pools in the profile and allow the model to reach an equilibrium prior to planting 4 months later. For those plot management details that were unavailable, we randomized associated model parameters across the model ensemble to account for uncertainty, drawing parameter values randomly from informed prior distributions to incorporate the full range of management possibilities within each scheme (see Table A.2 for prior distributions). Model parameters that were randomized included initial 2018 soil water, cultivar parameters, and initial residue amount for both years. Prior distributions for maize cultivar parameters were adopted from those presented by Archontoulis et al. (2020) who used experimental data from 56 site-years to calibrate APSIM maize parameters for Iowa, an important agricultural state in the U.S. Midwest. For summarized information on parameter priors and fixed management parameters, see the Supplementary Materials.

2.5. Evaluation of model performance

2.5.1. Post-hoc ensemble weight estimation

To more accurately interpret and evaluate results from each tested scheme, we applied an ensemble weighting strategy. Our use of ensemble weights rests on the assumption that ensembles which most accurately estimated the assimilation state variables were also more likely to have accurately estimated other components of the system. Therefore, to systematically emphasize our best available forecasts, we incorporate ensembles weights in our analysis of model output in the following way.

After the simulations were completed, we assigned a weight for each ensemble at each analysis time step by estimating the posterior probability of the ensemble's forecast as given below:

$$P(\, X \mid \mu_a, P_a \,) \tag{13}$$

where X is the forecast matrix of the assimilated state variables. This equation estimates a relative weight representing the likelihood of producing the model simulations given the posterior (analysis) state of the system, which follows a Normal distribution. We normalized the weights for each time step across all ensembles so that their sum was equal to 1, and then, we calculated the average weight of each ensemble for each year. The free model ensembles were given equal weights as no posterior distribution was computed.

Table 1
Overview of simulation components and naming conventions.

Simulation Name	Workflow Components	Variables included in X_f
Free	APSIM	N/A
SDA	APSIM + EnKF	Soil moisture (10 and 20 cm)
Miyoshi	APSIM + EnKF + adapted Miyoshi algorithm	Soil moisture (10 and 20 cm)
PDA	APSIM + EnKF + adapted Miyoshi algorithm	Soil moisture (10 and 20 cm), SWCON (10 and 20 cm)

2.5.2. Evaluation statistics

To compare differences in forecast precision of assimilated state variables across schemes, we use the spectral norm ($||.||_2$), which represents the maximum singular value of a matrix. The spectral norm allows us to compare the magnitudes of P_f for each of the different schemes to identify how forecast precision within each simulation changes with time. The spectral norm of P_f is calculated as

$$\|P_f\|_2 = \sqrt{Maximum\ Eigenvalue\ of\ P_f^H P_f} \tag{14}$$

where $P_f{}^H$ represents the conjugate transpose of P_f. We represent the precision of each simulation scheme in estimating all other relevant model variables using a weighted variance. This value was calculated for annual output values as

$$Variance = \frac{\sum_{i=1}^{N}\left(w_i * (x_i - \bar{x}_W)^2 \right)}{\frac{(N-1)}{N}} \tag{15}$$

where N is the number of ensembles, w_i is the average weight of the i^{th} ensemble, $\bar{x}_W$ is the weighted mean across ensembles, and x_i is the forecasted value of the i^{th} ensemble. For daily output values, variance was calculated as

$$Variance = \frac{1}{M}\sum_{m=1}^{M}\frac{\sum_{i=1}^{N}\left(w_i * (x_{i,m} - \bar{x}_{W,m})^2 \right)}{\frac{(N-1)}{N}} \tag{16}$$

where M is the number of simulation days, $x_{i,m}$ is the forecasted value of ensemble i on day m, and $\bar{x}_{W,m}$ is the weighted mean on day m across all ensembles.

Following the same notation, the accuracy of different simulation schemes was compared for annual output values with the following metrics

$$RMSE = \sqrt{\sum_{i=1}^{N}\left(w_i * (y_{annual} - x_i)^2 \right)} \tag{17}$$

where y_{annual} is the annual observed value. The following accuracy metrics were used for daily output values,

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} w_i * (y_t - x_{i,t})^2} \tag{18}$$

where T is the number of simulation days with observed data, y_t is the t^{th} observed daily value, and $x_{i,t}$ is the forecasted value of ensemble i on day t with observed data.

3. Results

We begin this section by first comparing the four different data assimilation schemes in our analysis and identifying the most robust scheme for soil moisture estimation in both accuracy and precision. Then, we evaluate and compare the performance of the optimal scheme with the free model in estimating daily soil moisture, soil N, LAI, annual yield, tile flow, and annual NO_3 loads.

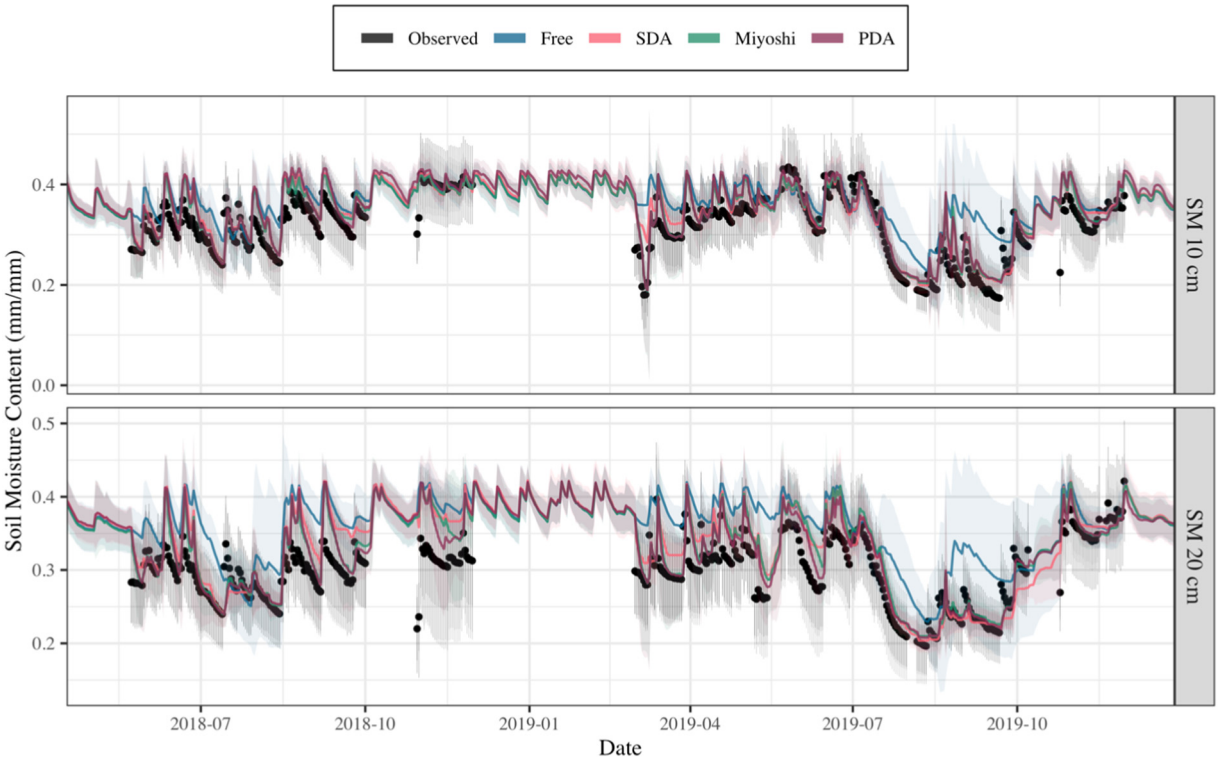


Fig. 3. Time series of simulated and observed soil moisture estimates for the 2 soil layers where assimilation is performed within the soil profile. SM3 refers to 9.1–16.6 cm (observed at 10 cm), and SM4 refers to 16.6–28.9 cm (observed at 20 cm). Weighted 95% confidence intervals are shown surrounding the mean line for the simulated estimates, and vertical bars around the mean observed value demonstrate the 95% confidence interval for those data as estimated by the Miyoshi algorithm in PDA.

3.1. Evaluation of different data assimilation schemes

The APSIM model performed sufficiently well without data assimilation and without intensive site-specific model calibration. As seen in the free model, the model was able to generally capture trends in LAI for both crop types (Fig. A.4) and trends in soil moisture throughout the soil profile (Fig. 3). Such performance points to the validity of both the underlying model processes and the model drivers. However, throughout the simulation period and, especially, during critical growth periods in the growing season (i.e., planting, vegetative phase), the free model overpredicts available soil moisture, which impacts downstream model estimates of crop water uptake, crop development, and tile flow, among others.

Compared to the free model, the assimilation of observed data via the EnKF helped to improve accuracy and precision of forecasts of soil moisture for the two soil layers (Table 2). Fig. 4a shows a smoothed time series of the spectral norms of P_f for each simulation. The forecast uncertainty for all simulation series is high at the initiation of 2018 following a wide prior on initial soil water balance but then drops by the spring of 2018. However, the SDA forecast uncertainty (as well as that of the other two assimilation schemes) remains low for the duration of the simulation period. The free model, on the other hand, experiences large jumps in uncertainty during both growing seasons, which we suspect to reflect uncertainties in crop parameters and/or model drivers. Since high precision and accuracy are most crucial within the growing season for the purpose of agricultural modeling, SDA clearly outperforms the free model by constraining soil water dynamics across the full parameter-input space.

Table 2
Comparison of soil moisture forecast accuracy and precision metrics.

Layer	Depth Range[a] *cm*	RMSE *proportion*				Average Variance 1×10^{-4}			
		Free	SDA	Miyoshi	PDA	Free	SDA	Miyoshi	PDA
SM3	9.1–16.6	0.061	0.038	0.034	0.036	6.1	3.5	3.7	3.6
SM4	16.6–28.9	0.064	0.042	0.037	0.035	7.2	3.3	3.9	3.3
SM6	49.3–82.9	0.084	0.073	0.072	0.074	5.8	4.4	5.1	5.0
SM7	82.9–138.3	0.142	0.076	0.074	0.076	6.0	4.1	4.5	3.2

[a] Layers SM1 (0–4.5 cm), SM2 (4.5–9.1 cm), and SM5 (28.9–49.3 cm) are excluded here as observed data was not available for these layers during our study period.

Yet, despite major improvements in forecast accuracy and precision, SDA shows filter divergence. An overestimated R and an underestimated P_f provide inaccurate weighting of the observed data and the model. As a result, the filter mostly ignores the observed data and overemphasizes the forecast distribution in the computation of the analysis distribution. By including the Miyoshi algorithm as an offline estimator of forecast and observed variances, we see this type of filter behavior mostly disappear in Miyoshi (See Fig. A.4 for a visualization of this phenomenon). Assuming divergence to be where the observed mean does not fall within the 95% confidence interval of the analysis distribution for at least one state variable, SDA diverged at 63.8% of analysis time steps, while Miyoshi diverged at 37.4%. This is a consequence of improved estimates of the two error matrices (i.e., R and P_f) when using the Miyoshi algorithm.

The final data assimilation scheme within our framework is parameter data assimilation. In our preliminary analyses of SDA innovations, we found that the module's prediction error for both soil layers was often the greatest on days with high precipitation and where end-of-day soil moisture was higher than or near the layer's drained upper limit. Since these conditions point to the use of the saturated flow model processes, we chose to update the *SWCON* model parameter for both soil layers (10 and 20 cm) within the EnKF. For each layer (T), the *SWCON* parameter controls the proportion of soil water (SW) above the drained upper limit (DUL) that flows into the next deepest layer for each day by the following equation:

$$Saturated\ Flow\ from\ Layer\ T = SWCON_T\ x\ (SW_T - DUL_T) \quad (19)$$

In PDA, we adjust the *SWCON* model parameter for both layers at each analysis time step according to the covariance between the parameter and the observed state variables. Though we see shifts in estimates of the two

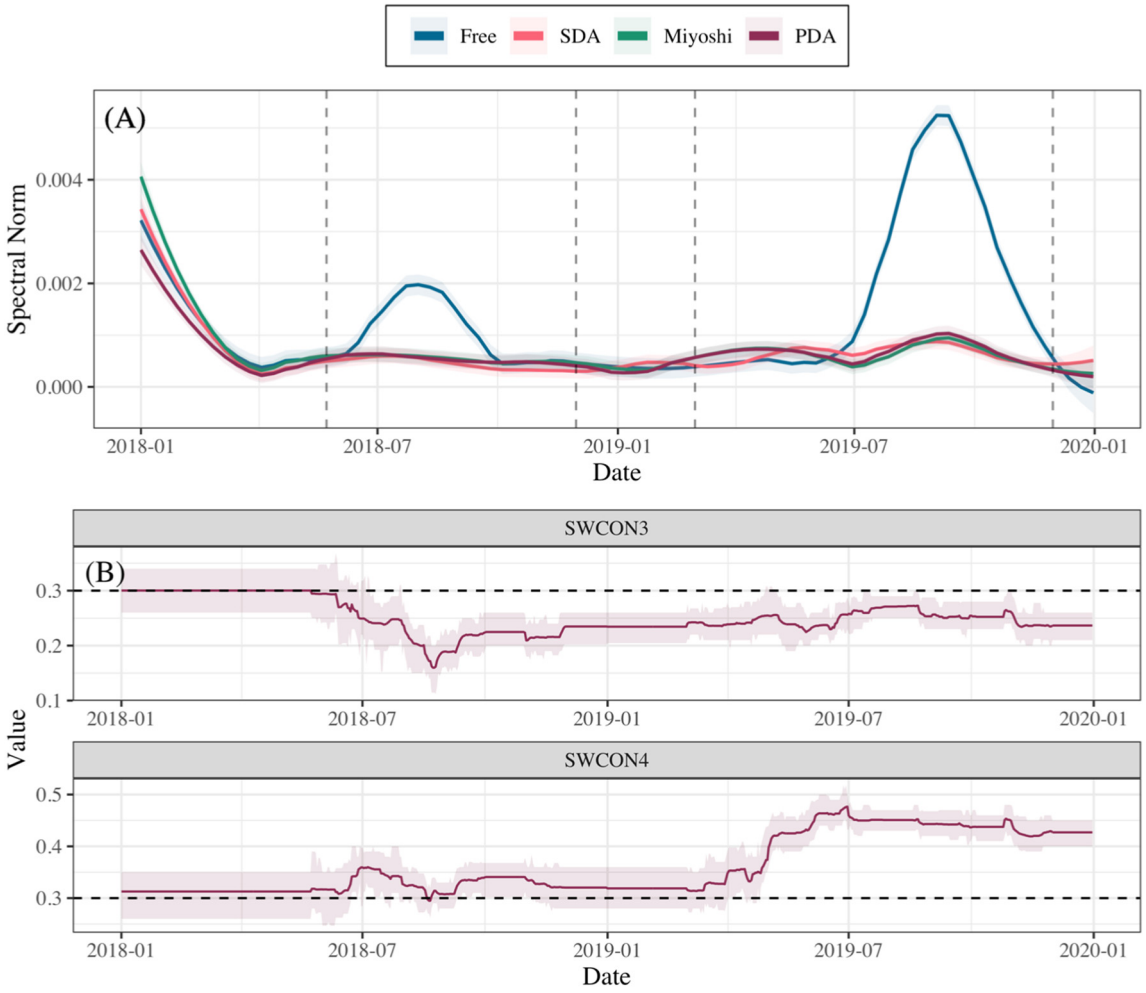


Fig. 4. (a) Time series of the spectral norm of the forecast covariance matrix (i.e., $||Pf||_2$) for SM3 and SM4 for all simulations for both years with local regression (LOESS) smoothing applied ($\alpha = 0.25$). Assimilation periods for both simulation years are indicated with the two sets of dashed lines. (b) Time series of SWCON parameter values under PDA optimization, where SWCON3 and SWCON4 correspond to the third and fourth soil layer, respectively. Dashed horizontal black lines denote the default model value for these two parameters.

SWCON parameters with PDA (Fig. 4b), these parameter adjustments did not lead to overall improved model performance in soil moisture estimation. PDA and Miyoshi performed similarly in terms of model accuracy and precision in estimates of soil moisture for the two assimilation layers. Yet, even though performance was not improved further in PDA, the final scheme allows for more flexibility in the model, which serves as an added benefit compared to Miyoshi. For this reason, we continue in our summary of our results by focusing on a comparison between the free model and our best-performing and most comprehensive data assimilation scheme: PDA.

3.2. Soil moisture

Estimates of soil moisture from the different simulation schemes are shown in Fig. 3, and Table 2 compares the accuracy and precision of daily soil moisture forecasts. Though the free model was able to capture the general trends of soil moisture in the soil profile, data assimilation helped to greatly improve soil moisture forecasts for the two layers with data assimilation. PDA was 40.2% and 44.3% more accurate, and 41.0% and 54.2% more precise for the two assimilation layers. However, assimilation also improved estimation of deeper soil layers; SM6 and SM7 estimate accuracy improved by 12.2% and 46.2%, and precision improved by 13.8% and 46.7%, respectively. Since the lower layers were not directly adjusted in the assimilation workflow, their improvement under assimilation is indicative of the "top-down" benefit that near-surface soil moisture assimilation can have for a model with a cascading water balance.

3.3. Soil nitrogen

With improved estimates of soil moisture, estimates of soil N dynamics throughout the soil profile were also impacted by data assimilation. On one hand, differences in estimates of total soil profile ammonium (NH_4) were minor for the duration of the simulation period with an average difference of 0.22 kg NH_4-N/ha and a maximum difference of 2.33 kg NH_4-N/ha. However, we do see great differences in estimates of total soil profile NO_3. Overall, the free model estimated lower NO_3 levels in the soil profile than PDA with an average difference of 3.92 kg NO_3-N/ha and a maximum difference of 9.35 kg NO_3-N/ha over the course of the study period. Large differences in total soil NO_3 are noticeable beginning in the middle of the 2018 growing season (Fig. 5a-b). We suspect these differences are the consequence of differences in soil moisture estimates. At that time, the free model often estimated soil moisture values above the drained upper limit for the assimilation layers, while PDA estimated soil moisture values below it. As shown in Fig. 2, this difference has the potential to alter the process rates within APSIM's soil N cycle, serving to increase the rate at which NO_3 was added to these layers (i.e., mineralization, urea hydrolysis, and/or nitrification) or decrease the rate at which NO_3 is lost (i.e., denitrification). The lower soil moisture estimates in the two assimilation layers also may have reduced the amount of soil water moving vertically through the soil profile and, thereby, limited the amount of NO_3 that leached into the lowest soil layer and lost from the system via leaching (Fig. 5c).

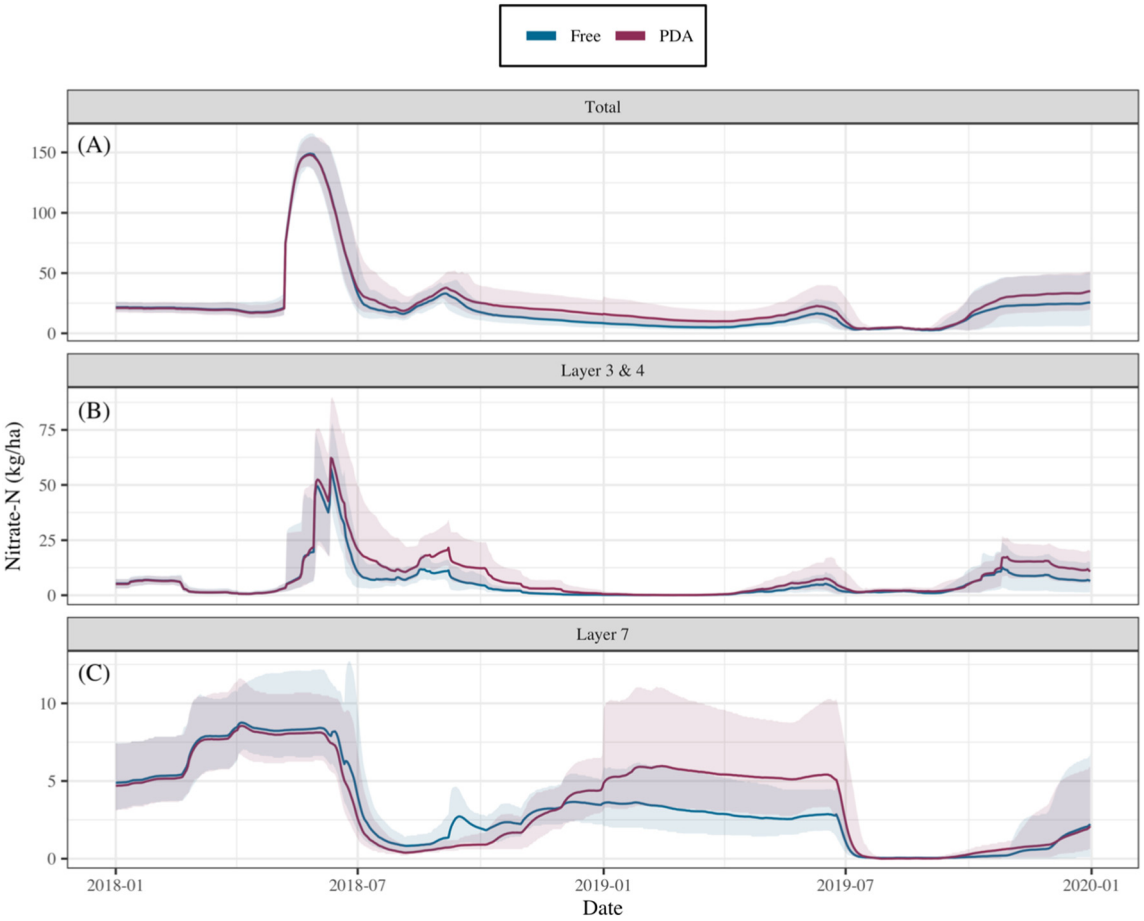


Fig. 5. (a-c) Time series of simulated soil NO3-N in (a) the total soil profile (b) the assimilation layers (i.e., Layers 3 and 4), and (c) the lowest soil layer (i.e., Layer 7). 95% confidence intervals are indicated by the shaded ribbon surrounding the mean lines for each scheme.

3.4. Leaf area index and annual yield

Aboveground measures of crop production, including LAI and annual yield, were less affected by assimilation, and changes in forecasting precision and accuracy were mixed. Table A.3 provides a more explicit comparison of accuracy and precision between simulation schemes and years for these variables. For maize in 2018, PDA was 10.2% and 0.1% less accurate than the free model when estimating yield and LAI. For soybean in 2019, PDA was 0.6% less accurate than the free model in estimating yield, but 14.1% more accurate when estimating LAI. Overall, though, the difference in accuracy was relatively minute between schemes when estimating aboveground variables. For precision, on the other hand, we see improvement with PDA in estimates of LAI for both crops and yield for maize. On average, variance was reduced by 12.9%, 9.8%, and 57.5% for estimates of maize LAI, soybean LAI, and maize yield, respectively. The average variance for soybean yield estimates increased by 36.7% with PDA.

3.5. Tile drainage and NO_3 loads

Following the improved soil moisture predictions with assimilation, similar improvements can be seen in estimates of daily and cumulative tile drainage. Although both the free model and PDA consistently overestimated daily tile drainage, PDA was more accurate and precise. PDA reduced RMSE by 23.0% and variance by 42.7% for daily tile flow estimates across both years. This improvement in PDA at the daily time scale led to similar improvements for cumulative annual estimates. On average, PDA reduced the RMSE by 43.1% and variance by 34.3% in annual estimates of tile drainage (Fig. 6a-b). As the free model often overestimated soil moisture in the two assimilation layers, we suspect that constraining soil moisture in PDA decreased the amount of soil water in the assimilation layers and, thus, decreased the amount of soil water drained from the system. Constraint of annual tile drainage with PDA was especially strong in 2018, where we saw great improvement in both accuracy and precision. This constraint was weaker in 2019, where we see exceptional improvement in accuracy but only slight improvement in precision.

For estimates of annual NO_3 loads, PDA was more accurate and precise than the free model for 2018. It predicted lower NO_3 loads and reduced RMSE by 19.3% and variance by 42.0%. However, PDA did not achieve the same constraint for annual NO_3 loads in 2019, where PDA's higher estimates reduced RMSE by just 1.82% and increased variance by about 120% compared to the free model (Fig. 6b). Such a large increase in uncertainty in 2019 likely stems from the large uncertainty associated with PDA estimates of NO_3 in the lowest soil layer (Fig. 5c). On the other hand, considering daily estimates of NO_3 load over the course of the simulation period, we see only a small 5.8% increase in accuracy and an 18.2% decrease in precision with PDA.

4. Discussion

Most crop modeling studies using data assimilation approaches focus on how SDA improves estimates of crop yield or biomass (e.g., de Wit and van Diepen, 2007; Fang et al., 2008; Ines et al., 2013; Mishra et al., 2021). However, in our analysis, crop yield estimates did not tell the full story. There are 4 key points to highlight within our results:

1. PDA effectively constrained soil moisture estimates for the two assimilation layers. One of the downstream impacts of this constraint was better

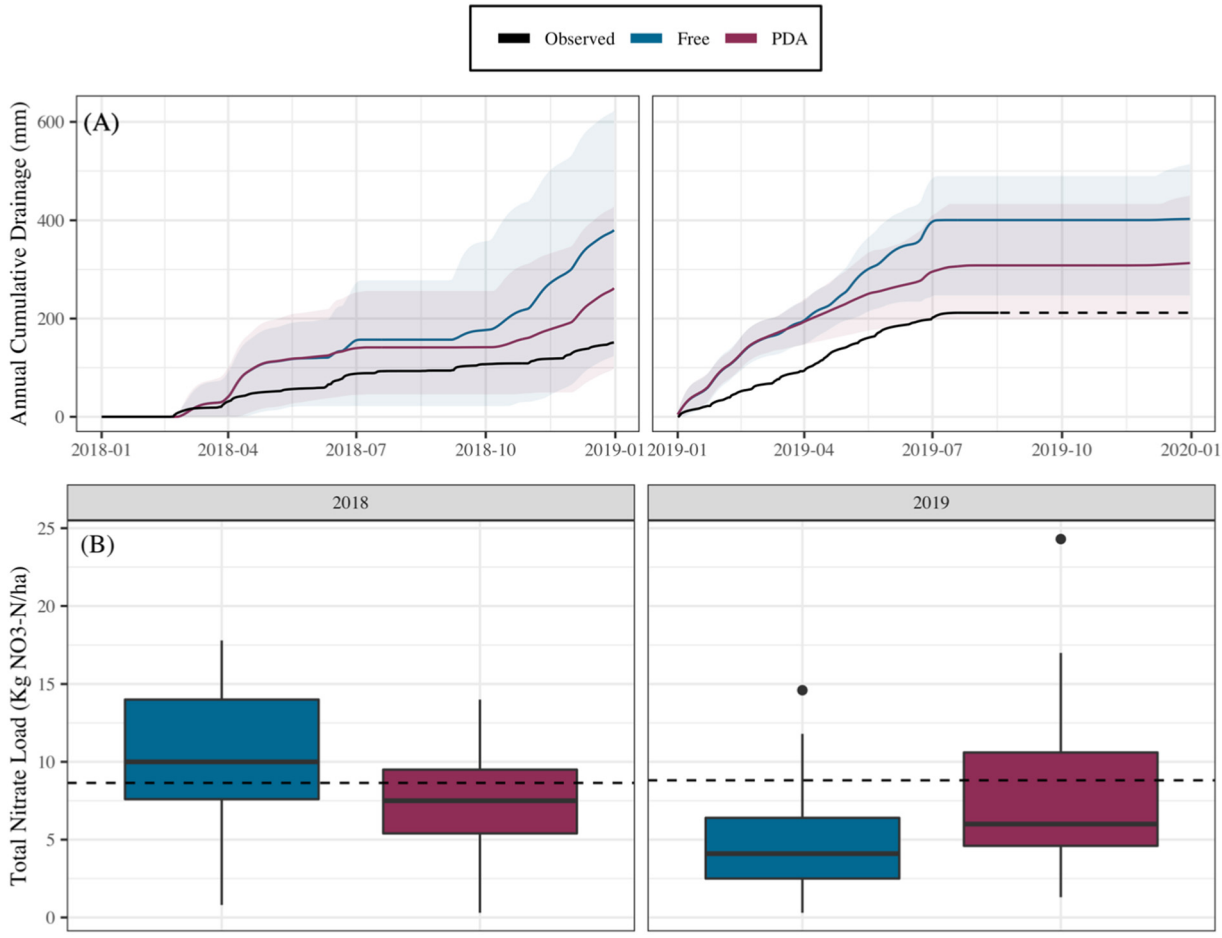


Fig. 6. (a) Time series of simulated and observed cumulative annual tile drainage (mm) for 2018 and 2019 from the study plot. 95% confidence intervals are indicated by the shaded ribbon surrounding the mean lines for each simulation. Black lines demonstrate the observed trends. Due to missing data from the end of 2019 as discussed in Section 2.2.4, we extrapolate the observed trend with a dashed line for 2019 following information from plot managers. (b) Boxplot summarizing the estimated distribution of total annual NO_3 load for each scheme in 2018 and 2019. Dashed horizontal lines mark the observed values for each growing season, with a load of 8.81 Kg NO_3-N/ha observed in 2018 and 8.65 Kg NO_3-N/ha observed in 2019.

soil moisture estimates for the two deeper layers (75 and 100 cm), where we see improvements in both forecast accuracy and precision with PDA (Table 2). In a similar study, Liu et al. (2017) attempted to use soil moisture assimilation to constrain root-zone soil moisture within the SWAT model by appending lower layers to the state vector at each analysis time step. However, due to the weak vertical coupling of SWAT, the improvement in soil moisture prediction in their analysis decreased with soil depth. The APSIM *SoilWat* module, on the other hand, operates as a cascading water balance model (Section 2.3.1), which exhibits strong downward vertical coupling between soil layers and, thus, increases the potential for constraint of those soil layers falling below the assimilation layers. Such potential is demonstrated by the strong constraint of soil moisture in Layer 7 in our results.

2. Our data assimilation workflow did not dramatically impact maize yield or LAI forecasts compared to the free model. However, considering the high levels of precipitation during the 2018 growing season (Moore et al., 2021) and the nature of the research site, which was managed to not be N-limited, there was little potential for data assimilation to impact aboveground estimates for maize at this study site. Assimilation typically lowered model soil moisture estimates, reducing the amount of soil water available to the crop, but the adjusted soil moisture value was often still greater than the maize crop's water demand. As a result, water uptake by maize was largely unaffected by the assimilation step (Fig. A.5). This result mirrors that of Lu et al. (2021) who found soil moisture assimilation to more effectively improve aboveground measures of maize in the presence of water stress.
3. Conversely, assimilation did play an impactful role in soybean LAI and yield estimates. In 2019, we see higher estimates of root-zone soil NO_3 with PDA compared to the free model, which we hypothesize to be the result of lower estimates of soil moisture in the two assimilation layers leading to changes in N cycle process rates. We suspect that the increased availability of soil NO_3 more aptly fulfilled the crop's N demand, which then increased N uptake, water demand, water uptake, and, consequently, crop growth. This can be shown in PDA's higher estimates of LAI in 2019. The soil N changes in PDA led to more accurate and precise estimates of soybean LAI in 2019 as compared with the free model. This improvement did not translate into improved estimates of soybean yield, however. Observed data on other portions of the water cycle, like plant water uptake, runoff, and evapotranspiration, could help to better understand these limitations of our assimilation system and identify missing or incorrectly defined model processes to improve them. For example, if estimates of LAI and water uptake but not yield are improved with data assimilation, parameters or processes that connect LAI to grain development may need to be closely investigated and possibly adjusted.
4. Compared to the free model, PDA was more accurate in its estimation of cumulative tile drainage than the free model, predicting lower cumulative tile drainage for both growing seasons. Since leaching is a function of both available soil NO_3 and tile drainage in APSIM, we would expect to see lower estimates of NO_3 load with reduced drainage if soil moisture was the only variable affected by assimilation. This partially explains the PDA results in 2018, where lower estimates of tile drainage aligned with lower and more accurate estimates of annual NO_3 load. However, in

2019, PDA estimated higher annual NO_3 load than the free model despite lower overall drainage from the system. Such a result highlights the downstream impact of assimilation on the soil N processes in the model and its interaction with the growing crop. Soil moisture could not have been the only variable affected by assimilation. Though a more comprehensive study of the *SoilN* module is necessary to draw conclusions on how assimilation specifically led to these improvements, these results demonstrate the potential for improving estimates related to NO_3 leaching via soil moisture data assimilation.

Upon highlighting the findings of our analysis, it is also imperative to highlight areas for improvement. Overall, the assimilation of soil moisture observations into the APSIM model was effective in improving model forecasts of soil moisture and downstream processes such as NO_3 leaching, which was a primary goal of the study. However, data assimilation algorithms—especially the EnKF—do not currently check for a water mass balance in the overall cropping system. This means that, at each analysis time step, assimilation is either erasing water or creating water within the modeled system rather than redistributing to other parts of the model (e.g., evaporation, crop water uptake, other soil layers, etc.). For our study, assimilation typically lessened soil moisture in the two assimilation layers and, thus, removed water from the forecasted soil profile when performing adjustment. With less water flowing through the soil profile, PDA estimated lower and more accurate tile drainage when compared with the observed data. Yet, by removing water with assimilation, PDA also disregards the system's water mass balance. The overestimation of soil moisture and tile drainage in the free model is indicative of inaccurate or missing processes within the APSIM model itself. Though PDA was able to improve tile drainage simulation, it does not account for these missing processes nor explain the ecological significance of the overestimation. Adding a water balance constraint (such as that presented in Wu et al., 2016) to this data-assimilation system, in conjunction with observed data on other water cycle components (e.g., evaporation, crop water uptake, runoff) would be useful to better understand where and why the model is making errors.

Further improvement to the assimilation workflow will also require reconsideration of the adjusted model parameter within the PDA workflow. As shown in our results, adjusting the SWCON model parameter for the two assimilation layers, though marginally helpful, did not dramatically improve soil moisture estimates as compared with Miyoshi. One possible explanation for the limited improvement with PDA could be the frequency with which SWCON is used for estimating water movement between soil layers. The SWCON parameter is associated with the saturated water flow process in the *SoilWat* module, which is only applied to those days and soil layers where soil moisture is above the drained upper limit but below saturation. In other words, soil moisture estimates (and, thus, innovations) in the two assimilation layers are dependent on the parameter value only when saturated flow happens in those layers. However, the modeling workflow assumes the estimates are correlated with the SWCON model parameters for the two layers and adjusts them accordingly at *all* analysis time steps. For more consistent and improved PDA performance, model parameters that are associated with soil moisture at all analysis time steps should be considered.

Another important consideration for future assimilation studies with the APSIM model concerns evaluating the model's soil N processes, an imperative component of cropping systems that remains poorly understood. At times within this study, assimilation of soil moisture had a dramatic impact on the soil N process rates and, thus, estimation of soil N pools. Since the model forecasts of soil moisture were improved in PDA, it would logically follow that the estimates of the soil moisture rate factors would also be improved, thereby improving soil N estimates. If data were available to evaluate how APSIM's *SoilN* changes with assimilation, one could feasibly distinguish weak points in the model process by identifying estimates that were not improved. Such a process could help to systematically improve the underlying processes in APSIM given adequate observed data for N cycle components. One process to investigate in the APSIM model that we highlighted in this study is crop uptake of mineral N forms. Currently, APSIM's *SoilN* module assumes that crops can only take up NO_3 and not NH_4, even though NH_4 fertilizer was also applied in the simulations and NH_4 uptake has lower energy requirements than NO_3 uptake in crops (Hachiya and Sakakibara, 2016). With adequate observed data, one could use soil moisture assimilation to understand the implications of this assumption more accurately.

The model-data fusion system introduced in this study provides a unique opportunity for the most complete account of uncertainty in modeling agricultural systems while allowing the dynamic constraint of uncertainties in both model parameters and state variables. Though the use of model-data fusion techniques in crop modeling is not new, the infrastructure developed, tested, and presented in this study is unique in that it (1) can be easily accommodated to assimilate other state variables or other types of observations (e.g., data collected from field experiments, flux towers, remote sensing, etc.), (2) jointly estimates the two error matrices in parallel with the simulation to dynamically improve filter performance, (3) can be expanded in space (allowing for performing regional data assimilation studies), (4) works well with all types of crops within the APSIM model, and (5) can leverage multi-data stream observations allowing for constraining different modules simultaneously. The authors are unaware of any other such system that shares all these advantages.

In expanding this analysis to the regional scale, the demonstrated results show that there is great potential for improved regional modeling of field-level NO_3 losses and tile drainage flow by using the presented system. Past regional studies were able to estimate NO_3 leaching with crop models by informing model inputs with coarse spatial data on soil type, land use, climate, water quality, and/or management information from the literature, public databases, and surveys (e.g., de Paz and Ramos, 2004; David et al., 2013; Roelsma and Hendriks, 2014; Reading et al., 2019; Li et al., 2020; Spijker et al., 2021). However, such applications fail to account for the fine-scale spatial variation in soil moisture and soil properties, which has been shown to be important for high accuracy and precision in estimates of NO_3 losses and tile drainage flow (Ojeda et al., 2018; Reading et al., 2019; Gurevich et al., 2021; Spijker et al., 2021). Given the appropriate observed data on soil moisture, the presented workflow has the capacity to improve on past regional studies by dynamically constraining soil moisture and soil hydraulic parameters at the field scale. By constraining the spatial and temporal variability of these model parameters and states across different fields, we can increase the accuracy and precision of NO_3 leaching estimates each field across a given region. This approach could potentially be applied to other regions given adequate data.

As a future direction, we will first focus on increasing the sample size of this study to enable a more thorough evaluation of the performance of our final assimilation system. We will investigate a variety of model parameters suitable to be adjusted within the PDA workflow to increase the potential for improvement in soil moisture estimates. To be able to expand this type of study to the regional scale, we will require soil moisture data that characterizes a broad range of cropping systems at high spatial and temporal resolutions. Soil moisture data products based on remote sensing (RS) imagery could be an invaluable tool for such a task. However, as this study was conducted with soil sensor data, the efficacy of the final assimilation system will need to be evaluated with RS data to ensure its performance does not suffer significantly with lower resolution soil moisture observations. Additionally, RS data typically characterizes surface soil moisture (0–5 cm), which has been shown to be less effective for constraining soil moisture in deeper layers of the soil profile (de Lannoy et al., 2007). Therefore, testing the utility of this workflow with RS-based soil moisture data products will be imperative prior to broader applications of the system. Finally, there is still room for improvement in model estimates of NO_3 leaching, such that estimates of tile drainage are still largely overpredicted and estimates of annual NO_3 load still lack high precision. To bridge these gaps, we will consider the addition of new data constraints within our data assimilation system, as well as the continued investigation and evaluation of the APSIM model processes.

5. Conclusion

In this study, we present a scalable and flexible data-assimilation system that improves forecasts through the systematic and robust combination of observed data and a crop model via the Ensemble Kalman Filter. In a case study, we demonstrate that the final system can effectively constrain soil moisture, maintain high filter performance by jointly estimating system uncertainties, and dynamically estimate soil properties in time by including parameters in the assimilation workflow. With such functionalities, our system stands apart from previous assimilation efforts in crop modeling.

Another novelty of this study is the focus and range of its system evaluation. Unlike other crop model assimilation studies which only consider changes in yield and soil moisture estimates, we assess system performance in estimating 5 different model states: yield, LAI, soil moisture, tile drainage, and annual NO_3 load. The system did not demonstrate strong constraint of aboveground measures, such as LAI or yield. However, assimilation did lead to changes in both the soil water and soil N cycle in APSIM, resulting in improved estimates of tile drainage flow and annual NO_3 load. These results point to soil moisture data assimilation as an improved method for estimating NO_3 leaching at the site level which can be expanded to broader regions. To verify these results, replication of this study for a range of site locations and observed data is necessary.

As we look to understand and resolve the large-scale agricultural issues of our time, the need for more accurate and precise agricultural forecasting methods only becomes more clear. The presented system provides a comprehensive solution for filling that gap. We have brought well-established forecasting methods from other disciplines to the realm of crop forecasting, advancing the predictive capacity in the field. In practice, our data-model fusion framework system can be applied to improve crop model performance through module-specific constraint and subsequent re-parameterization. In addition, for any site with available soil moisture observations, this system can quantify annual nitrate leaching losses for a production system with greater accuracy and precision than process-based models alone and with fewer resources and time than measuring the losses directly. Thus, our framework could serve as a practical and effective method for assessing the environmental impacts and informing future management design of agricultural production systems in the U.S. Midwest and beyond.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Marissa S. Kivi: Methodology, Software, Investigation, Writing – original draft. **Bethany Blakely:** Data curation, Writing – review & editing. **Michael Masters:** Data curation, Writing – review & editing. **Carl J. Bernacchi:** Writing – review & editing. **Fernando E. Miguez:** Writing – review & editing. **Hamze Dokoohaki:** Supervision, Conceptualization, Software, Writing – review & editing.

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.

Acknowledgements

The authors would like to thank all those on the Energy Farm team who made the presented case study possible. In particular, we would like to thank Grace Andrews and Heather Goring-Harford, who performed the chemical analyses for the nitrate leaching data, and Konrad Taube and Haley Ware, who helped with water collection and water filtering in 2018 and 2019. We also want to thank Caitlin Moore and Evan Dracup, who helped to collect and process much of the other data from the plot.

Additionally, we wanted to acknowledge those funding sources that supported the work of the Energy Farm team. First, the data used in this study was funded in part by (1) the Leverhulme Centre for Climate Change Mitigation, funded by the Leverhulme Trust through a Research Centre award (RC-2015-029), (2) the Center for Advanced Bioenergy and Bioproducts Innovation (CABBI) at the University of Illinois, and (3) the Global Change and Photosynthesis Research Unit of the USDA Agricultural Research Service. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Agriculture (USDA). Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. USDA is an equal opportunity provider and employer.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2022.153192.

References

Archontoulis, S.V., Miguez, F.E., Moore, K.J., 2014. Evaluating APSIM maize, soil water, soil nitrogen, manure, and soil temperature modules in the midwestern United States. Agron. J. 106 (3), 1025–1040. https://doi.org/10.2134/agronj2013.0421.

Archontoulis, S.V., Castellano, M.J., Licht, M.A., Nichols, V., Baum, M., Huber, I., Martinez-Feria, R., Puntel, L., Ordóñez, R.A., Iqbal, J., Wright, E.E., Dietzel, R.N., Helmers, M., Vanloocke, A., Liebman, M., Hatfield, J.L., Herzmann, D., Córdova, S.C., Edmonds, P., Lamkey, K.R., 2020. Predicting crop yields and soil-plant nitrogen dynamics in the US Corn Belt. Crop Science 60 (2), 721–738. https://doi.org/10.1002/csc2.20039.

Bernacchi, C., 2020. Data on Leaf Area Index for Energy Farm in 2018-2019. University of Illinois Unpublished raw data.

Bijay-Singh, Craswell, E., 2021. Fertilizers and nitrate pollution of surface and ground water: an increasingly pervasive global problem. SN Applied Sciences 3 (518). https://doi.org/10.1007/s42452-021-04521-8.

Bolten, J.D., Crow, W.T., Zhan, X., Jackson, T.J., Reynolds, C.A., 2010. Evaluating the utility of remotely sensed soil moisture retrievals for operational agricultural drought monitoring. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 3 (1), 57–66. https://doi.org/10.1109/JSTARS.2009.2037163.

Boote, K.J., Jones, J.W., Pickering, N.B., 1996. Potential uses and limitations of crop models. Agron. J. 88 (5), 704–716. https://doi.org/10.2134/agronj1996.00021962008800050005x.

Chakrabarti, S., Bongiovanni, T., Judge, J., Zotarelli, L., Bayer, C., 2014. Assimilation of SMOS soil moisture for quantifying drought impacts on crop yield in agricultural regions. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7 (9), 3867–3879. https://doi.org/10.1109/JSTARS.2014.2315999.

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. https://doi.org/10.1016/j.eja.2018.09.006.

Christianson, R., Christianson, L., Wong, C., Helmers, M., McIsaac, G., Mulla, D., McDonald, M., 2018. Beyond the nutrient strategies: common ground to accelerate agricultural water quality improvement in the upper Midwest. J. Environ. Manag. 206, 1072–1080. https://doi.org/10.1016/j.jenvman.2017.11.051.

David, M.B., McIsaac, G.F., Schnitkey, G.D., Czapar, G.F., Mitchell, C.A., 2013. Science Assessment to Support an Illinois Nutrient Loss Reduction Strategy.

Desroziers, G., Berre, L., Chapnik, B., Poli, P., 2005. Diagnosis of observation, background and analysis-error statistics in observation space. Q. J. R. Meteorol. Soc. 131 (613), 3385–3396. https://doi.org/10.1256/qj.05.108.

Dietze, M., 2017. Ecological Forecasting. Princeton University Press, Princeton https://doi.org/10.1515/9781400885459.

Dietze, M.C., Lebauer, D.S., Kooper, R., 2013. On improving the communication between models and data. Plant Cell Environ. 36 (9), 1575–1585. https://doi.org/10.1111/pce.12043.

Dinnes, D.L., Karlen, D.L., Jaynes, D.B., Kaspar, T.C., Hatfield, J.L., Colvin, T.S., Cambardella, C.A., 2002. Nitrogen management strategies to reduce nitrate leaching in tile-drained midwestern soils. Agron. J. 94, 153–171.

Dokoohaki, H., Miguez, F.E., Archontoulis, S., Laird, D., 2018. Use of inverse modelling and bayesian optimization for investigating the effect of biochar on soil hydrological properties. Agric. Water Manag. 208, 268–274. https://doi.org/10.1016/j.agwat.2018.06.034.

Dokoohaki, H., Kivi, M.S., Martinez-Feria, R., Miguez, F.E., Hoogenboom, G., 2021. A comprehensive uncertainty quantification of large-scale process-based crop modeling frameworks. Environ. Res. Lett. 16 (8), 084010. https://doi.org/10.1088/1748-9326/ac0f26.

Dokoohaki, H., Morrison, B.D., Raiho, A., Serbin, S.P., Dietze, M., 2021. A novel model–data fusion approach to terrestrial carbon cycle reanalysis across the contiguous U.S using SIPNET and PEcAn state data assimilation system v. 1.7.2 [Preprint]. Biogeosciences https://doi.org/10.5194/gmd-2021-236.

Elliott, J., Kelly, D., Chryssanthacopoulos, J., Glotter, M., Jhunjhnuwala, K., Best, N., Wilde, M., Foster, I., 2014. The parallel system for integrating impact models and sectors (pSIMS). Environ. Model. Softw. 62, 509–551.
Evensen, G., 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dyn. 53, 343–367. https://doi.org/10.1007/s10236-003-0036-9.
Evensen, G., 2009. The ensemble Kalman filter for combined state and parameter estimation. IEEE Control. Syst. 29 (3), 83–104. https://doi.org/10.1109/MCS.2009.932223.
Fang, H., Liang, S., Hoogenboom, G., Teasdale, J., Cavigelli, M., 2008. Corn-yield estimation through assimilation of remotely sensed data into the CSM-CERES-maize model. Int. J. Remote Sens. 29 (10), 3011–3032. https://doi.org/10.1080/01431160701408386.
Fer, I., Gardella, A.K., Shiklomanov, A.N., Campbell, E.E., Cowdery, E.M., De Kauwe, M.G., Desai, A., Duveneck, M.J., Fisher, J.B., Haynes, K.D., Hoffman, F.M., Johnston, M.R., Kooper, R., LeBauer, D.S., Mantooth, J., Parton, W.J., Poulter, B., Quaife, T., Raiho, A., Dietze, M.C., 2021. Beyond ecosystem modeling: a roadmap to community cyberinfrastructure for ecological data-model integration. Glob. Chang. Biol. 27 (1), 13–26. https://doi.org/10.1111/gcb.15409.
Gurevich, H., Baram, S., Harter, T., 2021. Measuring nitrate leaching across the critical zone at the field to farm scale. Vadose Zone J. 20 (2). https://doi.org/10.1002/vzj2.20094.
Hachiya, T., Sakakibara, H., 2016. Interactions between nitrate and ammonium in their uptake, allocation, assimilation, and signaling in plants. J. Exp. Bot., erw449 https://doi.org/10.1093/jxb/erw449.
Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km—global soil information based on automated mapping. PLoS ONE 9 (8), e105992. https://doi.org/10.1371/journal.pone.0105992.
Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Thépaut, J., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146 (730), 1999–2049. https://doi.org/10.1002/qj.3803.
Hoffman, R.N., Ardizzone, J.V., Leidner, S.M., Smith, D.K., 2013. Error estimates for ocean surface winds: applying desroziers diagnostics to the cross-calibrated, multiplatform analysis of wind speed. J. Atmos. Ocean. Technol. 30, 8.
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., Keating, B.A., 2014. APSIM – evolution towards a new generation of agricultural systems simulation. Environ. Model Softw. 62, 327–350. https://doi.org/10.1016/j.envsoft.2014.07.009.
Hu, S., Shi, L., Zha, Y., Williams, M., Lin, L., 2017. Simultaneous state-parameter estimation supports the evaluation of data assimilation performance and measurement design for soil-water-atmosphere-plant system. J. Hydrol. 555, 812–831. https://doi.org/10.1016/j.jhydrol.2017.10.061.
Huang, J., Ma, H., Su, W., Zhang, X., Huang, Y., Fan, J., Wu, W., 2015. Jointly assimilating MODIS LAI and ET products into the SWAP model for winter wheat yield estimation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8 (8), 4060–4071. https://doi.org/10.1109/JSTARS.2015.2403135.
Huang, J., Gómez-Dans, J.L., Huang, H., Ma, H., Wu, Q., Lewis, P.E., Liang, S., Chen, Z., Xue, J.-H., Wu, Y., Zhao, F., Wang, J., Xie, X., 2019. Assimilation of remote sensing into crop growth models: current status and perspectives. Agric. For. Meteorol. 276–277, 107609. https://doi.org/10.1016/j.agrformet.2019.06.008.
Ines, A.V.M., Das, N.N., Hansen, J.W., Njoku, E.G., 2013. Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. Remote Sens. Environ. 138, 149–164. https://doi.org/10.1016/j.rse.2013.07.018.
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. European Journal of Agronomy. 92. Elsevier B.V., pp. 141–152. https://doi.org/10.1016/j.eja.2017.11.002.
Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N.G., Meinke, H., Hochman, Z., McLean, G., Verburg, K., Snow, V., Dimes, J.P., Silburn, M., Wang, E., Brown, S., Bristow, K.L., Asseng, S., Smith, C.J., 2003. An overview of APSIM, a model designed for farming systems simulation. Eur. J. Agron. 18 (3–4), 267–288. https://doi.org/10.1016/S1161-0301(02)00108-9.
de Lannoy, G.J.M., Houser, P.R., Pauwels, V.R.N., Verhoest, N.E.C., 2007. State and bias estimation for soil moisture profiles by an ensemble Kalman filter: effect of assimilation depth and frequency. Water Resour. Res. 43 (6). https://doi.org/10.1029/2006WR005100.
Li, Z., Wen, X., Hu, C., Li, X., Li, S., Zhang, X., Hu, B., 2020. Regional simulation of nitrate leaching potential from winter wheat-summer maize rotation croplands on the North China plain using the NLEAP-GIS model. Agric. Ecosyst. Environ. 294, 106861. https://doi.org/10.1016/j.agee.2020.106861.
Liang, X., Harter, T., Porta, L., van Kessel, C., Linquist, B., 2014. Nitrate leaching in californian rice fields: a field- and regional-scale assessment. J. Environ. Qual. 43, 881–894. https://doi.org/10.2134/jeq2013.10.0402.
Liang, H., Qi, Z., DeJonge, K.C., Hu, K., Li, B., 2017. Global sensitivity and uncertainty analysis of nitrate leaching and crop yield simulation under different water and nitrogen management practices. Comput. Electron. Agric. 142, 201–210. https://doi.org/10.1016/j.compag.2017.09.010.
Linker, R., Ioslovich, I., 2017. Assimilation of canopy cover and biomass measurements in the crop model AquaCrop. Biosyst. Eng. 162, 57–66. https://doi.org/10.1016/j.biosystemseng.2017.08.003.
Liu, Y., Wang, W., Hu, Y., 2017. Investigating the impact of surface soil moisture assimilation on state and parameter estimation in SWAT model based on the ensemble Kalman filter in upper Huai River basin. J. Hydrol. Hydromech. 65 (2), 123–133. https://doi.org/10.1515/johh-2017-0011.
Liu, D., Mishra, A.K., Yu, Z., 2019. Evaluation of hydroclimatic variables for maize yield estimation using crop model and remotely sensed data assimilation. Stochastic Environ. Res. Risk Assess. 33 (7), 1283–1295. https://doi.org/10.1007/s00477-019-01700-3.
Liu, Z., Xu, Z., Bi, R., Wang, C., He, P., Jing, Y., Yang, W., 2021. Estimation of winter wheat yield in arid and semiarid regions based on assimilated multi-source sentinel data and the CERES-wheat model. Sensors 21 (4), 1247. https://doi.org/10.3390/s21041247.
Lü, H., Yu, Z., Zhu, Y., Drake, S., Hao, Z., Sudicky, E.A., 2011. Dual state-parameter estimation of root zone soil moisture by optimal parameter estimation and extended Kalman filter data assimilation. Adv. Water Resour. 34 (3), 395–406. https://doi.org/10.1016/j.advwatres.2010.12.005.
Lu, Y., Chibarabada, T.P., Ziliani, M.G., Onema, J.M.K., McCabe, M.F., Sheffield, J., 2021. Assimilation of soil moisture and canopy cover data improves maize simulation using an under-calibrated crop model. Agric. Water Manag. 252. https://doi.org/10.1016/j.agwat.2021.106884.
Ma, G., Huang, J., Wu, W., Fan, J., Zou, J., Wu, S., 2013. Assimilation of MODIS-LAI into the WOFOST model for forecasting regional winter wheat yield. Math. Comput. Model. 58 (3–4), 634–643. https://doi.org/10.1016/j.mcm.2011.10.038.
Mishra, V., Cruise, J.F., Mecikalski, J.R., 2021. Assimilation of coupled microwave/thermal infrared soil moisture profiles into a crop model for robust maize yield estimates over Southeast United States. Eur. J. Agron. 123. https://doi.org/10.1016/j.eja.2020.126208.
Miyoshi, T., Kalnay, E., Li, H., 2013. Estimating and including observation-error correlations in data assimilation. Inverse Prob. Sci. Eng. 21 (3), 387–398. https://doi.org/10.1080/17415977.2012.712527.
Monsivais-Huertero, A., Graham, W.D., Judge, J., Agrawal, D., 2010. Effect of simultaneous state–parameter estimation and forcing uncertainties on root-zone soil moisture for dynamic vegetation using EnKF. Adv. Water Resour. 33 (4), 468–484. https://doi.org/10.1016/j.advwatres.2010.01.011.
Moore, C.E., Haden, A.C., Burnham, M.B., Kantola, I.B., Gibson, C.D., Blakely, B.J., Dracup, E.C., Masters, M.D., Yang, W.H., DeLucia, E.H., Bernacchi, C.J., 2021. Ecosystem-scale biogeochemical fluxes from three bioenergy crop candidates: how energy sorghum compares to maize and miscanthus. GCB Bioenergy 13 (3), 445–458. https://doi.org/10.1111/gcbb.12788.
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 (5). https://doi.org/10.1029/2011WR011420.
Ojeda, J.J., Volenec, J.J., Brouder, S.M., Caviglia, O.P., Agnusdei, M.G., 2018. Modelling Stover and grain yields, and subsurface artificial drainage from long-term corn rotations using APSIM. Agric. Water Manag. 195, 154–171. https://doi.org/10.1016/j.agwat.2017.10.010.
Park, S.K., Xu, L., 2009. Data Assimilation for Atmospheric, Oceanic, and Hydrologic Applications. Springer.
Pasley, H., Nichols, V., Castellano, M., Baum, M., Kladivko, E., Helmers, M., Archontoulis, S., 2021. Rotating maize reduces the risk and rate of nitrate leaching. Environ. Res. Lett. 16 (6), 064063. https://doi.org/10.1088/1748-9326/abef8f.
de Paz, J.M., Ramos, C., 2004. Simulation of nitrate leaching for different nitrogen fertilization rates in a region of Valencia (Spain) using a GIS–GLEAMS system. Agric. Ecosyst. Environ. 103 (1), 59–73. https://doi.org/10.1016/j.agee.2003.10.006.
Puntel, L.A., Sawyer, J.E., Barker, D.W., Dietzel, R., Poffenbarger, H., Castellano, M.J., Moore, K.J., Thorburn, P., Archontoulis, S.V., 2016. Modeling long-term corn yield response to nitrogen rate and crop rotation. Front. Plant Sci. 7. https://doi.org/10.3389/fpls.2016.01630.
Quine, T.A., Zhang, Y., 2002. An investigation of spatial variation in soil erosion, soil properties, and crop production within an agricultural field in Devon, United Kingdom. J. Soil Water Conserv. 11.
R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Raiho, A., Dietze, M., Dawson, A., Rollinson, C.R., Tipton, J., McLachlan, J., 2020. Towards understanding predictability in ecology: a forest gap model case study [Preprint]. Ecology https://doi.org/10.1101/2020.05.05.079871.
Reading, L.P., Bajracharya, K., Wang, J., 2019. Simulating deep drainage and nitrate leaching on a regional scale: implications for groundwater management in an intensively irrigated area. Irrig. Sci. 37 (5), 561–581. https://doi.org/10.1007/s00271-019-00636-4.
Roelsma, J., Hendriks, R.F.A., 2014. Comparative study of nitrate leaching models on a regional scale. Sci. Total Environ. 499, 481–496. https://doi.org/10.1016/j.scitotenv.2014.07.030.
Seidel, S.J., Palosuo, T., Thorburn, P., Wallach, D., 2018. Towards improved calibration of crop models – where are we now and where should we go? Eur. J. Agron. 94, 25–35. https://doi.org/10.1016/j.eja.2018.01.006.
Silva, J.V., Giller, K.E., 2021. Grand challenges for the 21st century: what crop models can and can't (yet) do. Journal of Agricultural Science https://doi.org/10.1017/S0021859621000150 Cambridge University Press.
Spijker, J., Fraters, D., Vrijhoef, A., 2021. A machine learning based modelling framework to predict nitrate leaching from agricultural soils across the Netherlands. Environ. Res. Commun. 3 (4), 045002. https://doi.org/10.1088/2515-7620/abf15f.
Systems thinking, systems doing. Nat Food 1 (659). https://doi.org/10.1038/s43016-020-00190-9.
Tandeo, P., Ailliot, P., Bocquet, M., Carrassi, A., Miyoshi, T., Pulido, M., Zhen, Y., 2020. A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation. Mon. Weather Rev. 148 (10), 3973–3994. https://doi.org/10.1175/MWR-D-19-0240.1.
Wang, X., Bishop, C.H., 2003. A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. Am. Meteorol. Soc. 60, 1140–1158.
Water, Atmospheric Resources Monitoring Program, 2021. Illinois Climate Network. Illinois State Water Survey, 2204 Griffith Drive, Champaign, IL 61820-7495 https://doi.org/10.13012/J8MW2F2Q.
de Wit, A.J.W., van Diepen, C.A., 2007. Crop model data assimilation with the ensemble Kalman filter for improving regional crop yield forecasts. Agric. For. Meteorol. 146 (1), 38–56.
Wu, G., Dan, B., Zheng, X., 2016. Soil moisture assimilation using a modified ensemble transform Kalman filter based on station observations in the Hai River basin. Adv. Meteorol. 2016. https://doi.org/10.1155/2016/4569218.