A COMPREHENSIVE ASSESSMENT OF SOIL MOISTURE DATA ASSIMILATION AND ITS POTENTIAL TO INCREASE AGRICULTURAL FORECASTING CAPACITY

BY

MARISSA SUZANNE KIVI

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Crop Sciences
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Master's Committee:

Assistant Professor Hamze Dokoohaki, Chair
Associate Professor Rabin Bhattarai
Professor Bo Li
Assistant Professor Nicolas Martin

ABSTRACT

In the face of today's large-scale agricultural issues, the need for robust methods of agricultural forecasting has never been clearer. Yet, the accuracy and precision of recent forecasts remains limited by current tools and methods. Past studies have proposed and tested soil moisture data assimilation as a method to merge soil moisture observations with process-based crop models, thereby accounting for spatial heterogeneity in soil water dynamics and improving crop model estimates. Building on these previous studies, the following work systematically and comprehensively explored the potential for soil moisture data assimilation to serve as a powerful and generalizable tool for improving agricultural predictions in the U.S. Midwest. First, a scalable, flexible, and robust data-assimilation system was developed. The system (1) utilizes ensemble-based filtering approaches to constrain model states and update model parameters at observed time steps, (2) propagates uncertainties, and (3) incorporates an algorithm that improves system performance through the joint estimation of system error matrices. After assimilating *in situ* soil moisture observations into the APSIM crop model for an experimental site in central Illinois over two growing seasons, the system demonstrated strong constraint of soil moisture forecasts, improving soil moisture estimates in the two assimilation layers by 42% and 48%. Such constraint propagated into improved accuracy in estimates of lower layer soil moisture, annual tile flow, and annual nitrate loads, but did not have strong impacts on aboveground measures of crop productivity due to a lack of water stress at the site.

To further evaluate the developed system, the constraint of *in situ* soil moisture data assimilation was evaluated for 5 experimental sites across the U.S. Midwest using observations spanning 19 site-years. The system's impact on estimates of soil moisture, yield, NDVI, tile drainage, and nitrate leaching was assessed across all simulated growing seasons. For all site-years, the accuracy of soil moisture forecasts in the assimilation layers was improved. These changes also led to improved simulation of soil moisture in deeper parts of the soil profile in most cases. Although crop yield was improved for most site-years, the greatest improvement in yield accuracy was demonstrated in site-years with higher water stress, where assimilation served to increase available soil water for crop uptake. Alternatively, estimates of annual tile drainage and nitrate leaching were not well constrained across the study sites. Trends in drainage constraint suggest the importance of evapotranspiration observations as a next point for constraint.

Finally, to test the full generalizability of the developed system, the application of remote sensing surface soil moisture observations was investigated. Four different data products were assimilated within the developed data-assimilation system for the same 5 study sites. The assimilation of surface soil moisture showed weaker constraint of downstream model state variables when compared to the assimilation of root-zone soil moisture values from the previous analysis. The median reduction in soil moisture RMSE for observed soil layers was lower, on average, by a factor of 4. However, crop yield estimates were still improved overall with a median RMSE reduction of 17.2%, and there is strong evidence that yield improvement was higher when under water-stressed conditions. Comparisons of system performance across different combinations of remote sensing data products indicated the importance of high temporal resolution and accurate observation uncertainty estimates when assimilating surface soil moisture observations. This study highlighted many opportunities and challenges of soil moisture data assimilation as an agricultural forecasting tool and laid a strong foundation for future innovation and application of the approach.

To my grandmas, Deanna and Sue, for everything

ACKNOWLEDGEMENTS

Over the last two years, I can say with confidence that I have learned two things: (1) the immense potential for data-assimilation methods to change how we forecast agricultural systems and (2) the importance of family, friends, and colleagues. Over the course of my program at the University of Illinois at Urbana-Champaign, I have had the incredible support of old friends, new friends, and family members to motivate and guide me through the past few years of unexpected twists and turns. These people filled my daily walks with laughter and joy, connected me to home, and gave me new life when hours of investigating APSIM biases killed my spirit. Thank you to my mom and dad for loving and supporting me through my winding path and for being a sense of home when things feel uncertain. Thank you to my brothers for keeping me humble and for being the best friends that I do not particularly like but that I am grateful to have. Thank you to my extended family and old friends who continue to love me despite my aversion to digital communication and thank you to new friends who have made Urbana a bit more like home. Also, though many will not read this, I would like to say thank you to the Urbana Boulders community for a sense of belonging, hours of great climbing, new friendships, and short-person beta when I needed it most.

I want to extend my sincerest thanks to my thesis committee for their commitment to my research and to my growth. As someone new to the University of Illinois and the field of crop science, your expertise and dedication helped me navigate this program and research project with optimism and confidence. I want to also express my gratitude for Dr. Teerath Singh Rai and Nakian Kim for their support in my research work and, most importantly, for laughing at my jokes this past year. It has been nice to have friendly faces in the office. Also, a huge thank you to the Energy Farm team and Dr. Noemi Vergopolan of Princeton University who generously contributed data to this work and made this project possible.

Finally, I want to take a moment to recognize my advisor, Dr. Hamze Dokoohaki, who led me to this research position which drastically changed the course of my life. Thank you for this opportunity and for your unrelenting guidance, patience, and kindness over the past few years despite my inefficient writing strategies and my tendency to talk too much during meetings. Your innovative and enthusiastic approach to science is contagious and inspiring. This position and research project initiated my passion for the wonderful yet complex world of agriculture and helped me jumpstart my career in agricultural technology and forecasting. I cannot easily express how grateful I am for your mentorship over the past few years and for the skills and knowledge I have gained along the way.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The global community faces an array of complex, large-scale agricultural challenges. Population growth and climate change threaten to disrupt crop productivity and broaden yield gaps (Pradhan et al., 2015). Monoculture cropping systems, chemical inputs, and other traditional agricultural practices support rural livelihoods but degrade soil health, water quality, biodiversity, and natural habitats (Brummer et al., 1998; Christianson et al., 2018; Reading et al., 2019; Silva and Giller, 2021). To anticipate, measure, and manage these issues, there is great need for reliable, comprehensive agricultural forecasts. Such predictions could help to inform policy, reduce waste, and drive innovation, and there has been great effort to push forward new tools and compile expansive, comprehensive databases to support progress in the field (e.g., Marj and Meijerink, 2011; Abendroth et al., 2017; Dietze et al., 2017; Chighladze et al., 2021). However, despite huge strides in both data availability and method development, current methods in agricultural forecasting remain relatively poor and inefficient.

Today, the most popular approaches in agricultural forecasting leverage process-based crop models, crop monitoring data, and/or remote sensing imagery. Individually, each of these tools has its own unique advantages but is, nonetheless, limited in prediction accuracy, precision, or both. For example, process-based crop models are valuable as they combine state-of-the-art knowledge on agricultural processes to more comprehensively monitor and simulate cropping systems than field experiments due to greater system complexity (Boote et al., 1996; Jin et al., 2018; Pasley et al., 2021). This exhaustive representation of the system can simulate observed and unobserved system state variables and estimate their covariance while maintaining system mass balances (Archontoulis et al., 2020). Yet, the application of these models remains controversial despite their extensive development and demonstrated value. The strongest controversy focuses on the unaccounted uncertainties associated with crop model parameters, inputs, and structure (Huang et al., 2019; 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). Subsequent applications of the final calibrated model can be justified only within a narrow inference space and are, otherwise, unreliable.

Aside from crop models, agricultural forecasts have also been produced by fitting machine learning models with field observations. This approach focuses on the prediction of an observed response variable using a set of observed features. Depending on the complexity of the applied model, its interpretation can lead to a better understanding of agroecosystem processes at the field scale. In addition, new technology has broadened the application of this method by enabling the efficient and high-precision monitoring of a broader range of agricultural variables in field experiments, including soil moisture, tile drainage, and leaf area index. These advancements have increased the range of state variables that can be predicted through this method. However, despite recent improvements, the inference space for this method still falls short. While measured 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 inherently limited in dimensionality. They can only increase the predictive capacity for agricultural variables which can be directly measured, and the application of the fitted models is limited in time and space. Therefore, this approach fails to capture the complexity of the high-level applied research questions at the core of many agricultural issues ("Systems Thinking", 2020). This method also cannot readily leverage all available information, as combining measurements and data products from different instruments and experiments and across different temporal and spatial resolutions is rarely straightforward and often impossible with typical machine learning approaches (Dietze et al, 2013). Consequently, only a fraction of available information can be effectively used.

Agricultural forecasts can also be generated by combining remote sensing (RS) observations with machine learning models. RS data products are particularly valuable compared to *in situ* observations as they are broadly and consistently collected in space and time. This is an advantage as it eliminates between-site variability introduced by collection methods and can easily and thoroughly account for spatiotemporal variability agricultural state variables at broad scales. However, like crop models and *in situ* observations, the application of RS observations for agricultural predictions is limited. First, as discussed with *in situ* observations, RS data products can only be used to characterize a certain subset of agricultural system variables and, thus, can only build predictive models that leverage and estimate variables and relationships which can be approximated with spectral information. RS observations also introduce two new problems when making agricultural predictions (Huang et al., 2019). The first issue centers on the fact that RS data products are often, themselves, model estimates and, therefore, their application can impose additional biases and uncertainties, which are typically not well known. Next, RS data products provide estimates at spatial and temporal resolutions which are far coarser than what is needed to best characterize agroecosystems at the field-scale. Consequently, the information contained in RS observations may not be representative of the true in-field state variable since their estimation is based on spectral reflectance from a much broader region.

It is evident that, individually, current agricultural forecasting methods cannot effectively provide the necessary predictions for investigating large-scale agricultural issues. Yet, each tool boasts an important advantage for the task. Process-based models simulate higher system complexity, field observations accurately and locally represent an array of important state variables, and remote sensing data products account for important spatiotemporal variability in agricultural processes. To leverage each of these critical benefits, state data assimilation (SDA) has emerged as a viable method to bring these 3 methods together (Jin et al., 2018). SDA fuses process-based crop models and observed data together, allowing them to speak to and build on one another despite differing temporal and spatial scales (Dietze et al., 2013). As a foundation, the model provides a temporally continuous, high-dimensional scaffold in which a variety of observations can be smoothly integrated (Dietze et al., 2013; Liu et al., 2021). Then, observations (*in situ* or RS) are merged with model predictions using one of many robust, systematic algorithms, including the Ensemble Kalman Filter (EnKF), variational Bayes, and particle filters (Huang et al., 2019). Through this process, uncertainty around spatially-heterogenous and dynamic properties in agricultural systems can be reduced. This increases precision and accuracy in simulations while decreasing dependence on extensive site-level model calibration, a process that has limited the application of process-based models in the past (Mishra et al., 2021). The SDA process can also be a tool for investigating and reducing biases in model structure and parameters (Launay and Guerif, 2005; Lü et al., 2011; Hu et al., 2017).

Numerous studies have explored the potential for SDA to constrain crop model estimates. These studies have focused on a variety of state variables, 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), which have been made available through field experiments and remote sensing imagery. Additionally, these studies have employed a host of different process-based crop models, including WOFOST (e.g. de Wit and van Diepen, 2007), APSIM (e.g., Machwitz et al., 2014), and DSSAT (e.g., Ines et al., 2013), and have been executed at a range of spatial scales. Dorigo et al. (2008), Jin et al. (2018), and Huang et al. (2019) have all compiled and published reviews of the work completed in this field.

However, despite the investigation of assimilation applications in crop modeling, there is still an evident need for a benchmark systems approach in the field. Many past studies have been able to successfully assimilate observations into crop models, but, typically, their approaches have been designed to complete one-off objectives. To effectively innovate and utilize SDA for the purpose of agricultural forecasting, a scalable, flexible, and robust data-assimilation system will need to be developed. Scalability will allow for both small- and large-scale forecasts such that estimates can be generated at meaningful resolutions to answer a range of different research questions. Flexibility will ensure that a baseline system can be applied consistently and effectively across crop models, assimilation observation types, agronomic management practices, site-level uncertainties, and spatial scales. Lastly, a robust system will ensure accurate estimates of state variables, as well as their associated uncertainties through the application of state-of-the-art algorithms and comprehensive uncertainty propagation (Keenan et al., 2011; Tandeo et al., 2020; Dokoohaki et al., 2021). The development of a data-assimilation system with these 3 characteristics will best drive progress in agricultural forecasting methods.

In addition to benchmark infrastructure, another important gap in current SDA approaches in crop modeling concerns the evaluation of system performance. Although many different state variables have been directly constrained with SDA, studies typically do not evaluate the impact of assimilation beyond estimates of the assimilated state variable and annual crop yield. This makes sense in the case of regional studies where observations are difficult to collect at scale (e.g., de Wit and van Diepen, 2007; Dente et al., 2008; Wu et al., 2021). However, for studies based on field experiments (e.g., Launay and Guerif, 2005; Zhao et al., 2013; Jiang et al., 2014), this approach fails to leverage the full dimensionality of process-based crop models and the full suite of data-assimilation services, leaving many possible downstream model constraints uninspected and, therefore, undiscovered. When observations are available, future assimilation studies in agricultural forecasting should be rigorous in evaluating assimilation's impact on all observed model processes. Such an approach will help to highlight biases and new opportunities in the established system and methods, pushing further system innovation and application. A few published studies have performed more comprehensive evaluations of downstream assimilation impacts, including Thorp et al. (2010) who evaluated model estimates of canopy weight and evapotranspiration in a LAI-assimilation study and Lu et al. (2021) who evaluated model estimates of canopy cover, evapotranspiration, biomass, and yield. Such studies provide a more comprehensive understanding of how the assimilation system is functioning.

In the following work, a data-assimilation system will be developed, tested, and evaluated to help fill these demonstrated gaps in current agricultural forecasting methods. All 3 predictive methods (i.e., remote sensing imagery, field experiment observations, process-based crop models) will be explored and tested in the system through a multidimensional evaluation process that exhausts available observations on the true system. Chapter 2 focuses on the development of a scalable, flexible, and robust data-assimilation system which leverages well-established tools and algorithms to optimize system performance. In the chapter, the system is developed and evaluated using a range of field experiment observations from a single study site. Chapter 3 broadens the application and evaluation of the developed system to a larger number of study sites to test the generalizability of the single-site results and identify universal strengths and weaknesses of the system. Moving beyond *in situ* observations, Chapter 4 considers the application of RS observations in the developed system and explores how the selection of an RS data product can impact system performance. It includes an evaluation of system performance under RS assimilation and discusses the opportunities and challenges in assimilating RS information moving forward. Finally, in Chapter 5, key findings of this work, as well as their implications for the future of agricultural forecasting, are summarized and discussed.

CHAPTER 2

DEVELOPMENT AND EVALUATION OF AN OPTIMAL SOIL MOISTURE DATA ASSIMILATION SYSTEM AND ITS CONSTRAINT OF NITRATE LEACHING FORECASTS IN THE APSIM MODEL

INTRODUCTION

To support efficient and comprehensive innovation in agricultural forecasting, there is great need for a scalable, flexible, and robust data-assimilation system. Such a system must be composed of state-of-the-art tools and methods and be able to accommodate a diverse array of observations, operate at a range of spatiotemporal scales, and accurately estimate and propagate system uncertainties. To allow for these capabilities, the selection of a data-assimilation approach is of critical importance. In the past, many different state data assimilation (SDA) techniques have been applied to improve crop model predictions (Huang et al., 2019). However, the ensemble Kalman filter (EnKF; Evensen, 2003) stands out as one of the most popular SDA techniques for use with non-linear dynamic crop models due to its ease of implementation, computational efficiency, and 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 (De Lannoy et al., 2007; Zhao et al., 2013; 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. Of the many data products that have been assimilated into crop models, soil moisture (*in situ* or remotely

This chapter includes work that was previously published in Science of the Total Environment and is reprinted here with appropriate permission from the copyright owner: Kivi, M.S., Blakely, B., Masters, M., Bernacchi, C.J., Miguez, F.E., Dokoohaki, H., 2022. Development of a data-assimilation system to forecast agricultural systems: A case study of constraining soil water and soil nitrogen dynamics in the APSIM model. Science of The Total Environment 820, 153192. https://doi.org/10.1016/J.SCITOTENV.2022.153192.

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 Dipen, 2007; Monsivais-Huertero et al., 2010; Chakrabati 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; Chakrabati 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 (Chakrabati 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, soil moisture plays a much larger role within an agroecosystem, impacting an array of atmospheric, soil, crop, and water processes (Engman, 1991). Consequently, the assimilation of soil moisture into a crop model that includes these processes has the potential to constrain them, leading to improved forecasts of related agricultural state variables even if they are not directly observed. One critical candidate for such constraint 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 (Hansen et al., 2006; 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 (van der Laan et al., 2014; Liang et al., 2017; Reading et al., 2019). As a result, models are also insufficient for estimating NO_3 leaching at the regional scale. Soil moisture SDA has the potential to overcome these weaknesses and improve model estimates of NO_3 leaching through the constraint of its modeled relationship with soil moisture.

The following chapter focuses on the development of a robust and generalizable data-assimilation system for the purpose of agricultural forecasting and explores the potential of soil moisture SDA as a method to systematically improve the accuracy and precision of various agricultural state variables, including NO_3 leaching, in a process-based crop model. 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 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 and soil NO_3 concentration depend on previous model estimates of soil moisture. Based on this fact, it is hypothesized 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) later in the model process without the need for observing the states directly. To the author's 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 NO_3 leaching.

This chapter 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.
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, and NO_3 leaching.

MATERIALS AND METHODS

Study site

To test and evaluate an optimal data assimilation system, this study focuses 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. 2.1). 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). Both characteristics justify the use of this study site as a reasonable baseline for testing an agricultural forecasting system for the U.S. Midwest.

Since accurately specified management information is crucial to ensure accurate model predictions (Archontoulis et al., 2020), all known management details were included as constants across simulations. Management information was collected 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 assumed to be from the previous year's crop, which was maize in both cases. Information on tillage, herbicide, nor pesticide practices was not included in simulations at this point of the project.

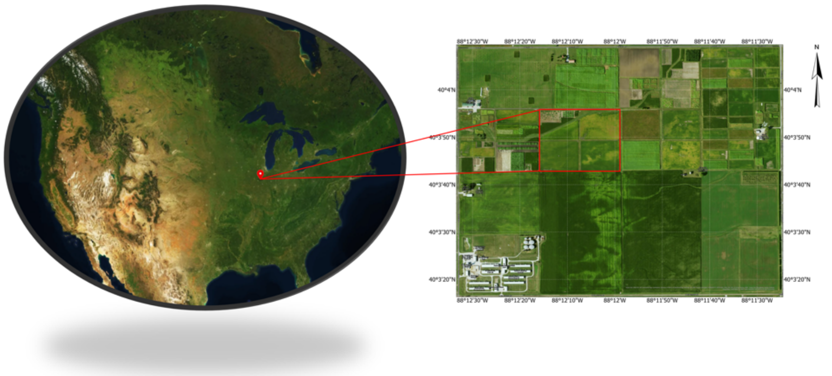

Figure 2.1. Aerial image of the Energy Farm research plots outside of Urbana, IL. The "Maize Control" plot is outlined by the red square on the zoomed right panel.

Observed data

Model drivers

There are two important model drivers used in this study—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, 11 weather ensembles and 25 soil ensembles were independently randomized across model ensembles for each simulation series. 10 of the 11 weather ensembles were 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 weather variables at hourly timesteps with associated uncertainties (Hersbach et al., 2020). The derived ensembles included data on solar radiation, maximum air temperature, minimum air temperature, precipitation, and wind speed aggregated to daily resolution. The remaining weather 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 provided in the SoilGrids dataset. 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.).

Soil moisture

Soil moisture observations were collected at the study site for 2018-2019 at 30-minute 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 measured as the volumetric water fraction at each depth (Moore et al., 2021). Only measurements of soil moisture available at the 10 and 20 cm depths were employed for the purposes of assimilation. These two state variables will be referred to as SM3 and SM4, respectively, hereafter. 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, end-of-day soil moisture was computed as the

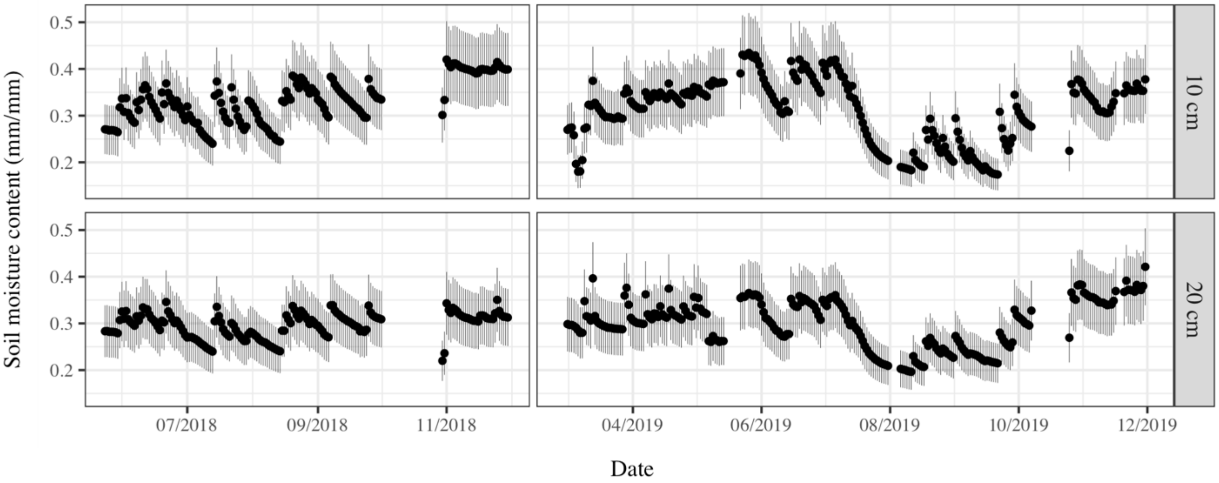

Figure 2.2. Time series of observed soil moisture from the study site for 2018-2019. Vertical black bars mark the breadth of a 95% confidence interval around the mean daily estimate with an assumed 10% observation standard error. Though some observations were available in the winter months of 2018 and 2019, those values were not included in our analysis and, thus, were excluded from this figure.

average value between the hours of 22:00 of the current day and 02:00 of the next day for each depth. These computed values constituted the state variables within the 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-hour period were excluded. Data points from the winter months (i.e., January, February, and December) were also removed to avoid possible sensor inaccuracies related to freezing soils. Due to a low observation sample size, a 10% observation error was assumed around the mean for both soil depths (Fig. 2.2).

Crop yield and leaf area index

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 were 10 and 14 LAI observations available for the 2018 and 2019 growing seasons, respectively (Bernacchi, 2020).

Tile flow and nitrate loads

For both growing seasons, the Energy Farm collected information on tile flow for the study plot at 15-minute 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 estimated the missing flow to be small relative to the year's total, so it was assumed 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-minute interval, NO_3 concentrations were linearly interpolated between samples, multiplied by the instantaneous flow rate at each 15-minute 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 this study.

Data-assimilation system

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 crop 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 the APSIM simulations in this study, the following available modules were included: *Fertiliser*, *SoilWat*, *SurfaceOM*, *SoilN*, *Soybean*, and *Maize*. Apart from those model parameters related to management (Table A.1), minimal changes were made to the model's parameterization. For the source code, a new version of the model was compiled to allow for online communication with R statistical software (R Core Team, 2021) through RDotNet and the .NET framework. APSIM module documentation and source code is available at https://www.apsim.info/.

For the purposes of this analysis and the larger project, the two APSIM modules controlling soil water and soil N were 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 soil water flow type 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 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 each day depends on rate factors related to daily soil moisture estimates, which are

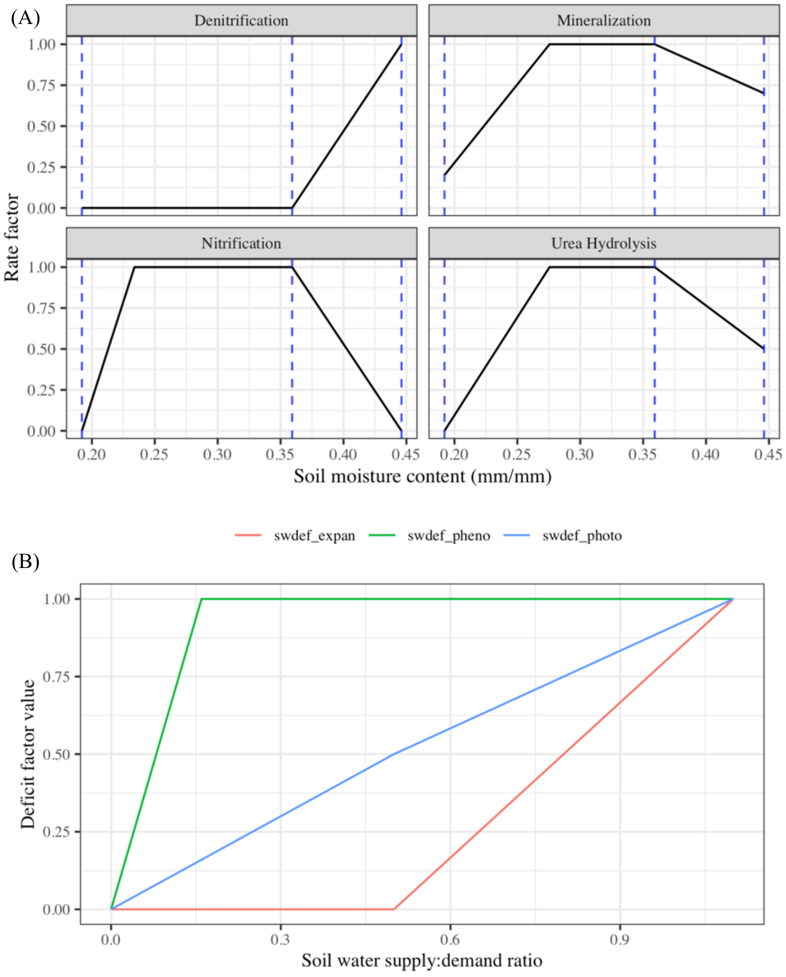

Figure 2.3. (a) Sample functions for determining soil moisture rate factors for soil N processes in APSIM based on soil moisture content. This example was generated based on the lower limit (SMC = 0.19), drained upper limit (SMC = 0.36), and saturated limit (SMC = 0.45) of Layer 3 at the study site. These limits are shown in the figure as blue, dashed vertical lines. (b) Piecewise functions that determine the 3 soil water deficit factors for maize leaf expansion, phenology, and photosynthesis based on the total soil supply to crop water demand ratio.

calculated within the *SoilWat* module. Figure 2.3a 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.

The APSIM modules responsible for growing maize and soybean at the study site were the *Maize* and the *Plant* module, respectively. The former is based on the CERES-Maize model and simulates maize growth on an area basis at each daily time step. The *Plant* module is a more general crop module which has been parameterized to simulate the development of several crops, including soybean. In both modules, crop phenology and productivity are impacted by weather, soil water, and soil N. Water stress is introduced to the crop via three soil water deficit factors (i.e., *swdef_pheno*, *swdef_photo*, and *swdef_expan*), which respectively impact phenology, photosynthesis, and leaf expansion. The *Plant* module also includes a soil water deficit factor for N fixation in soybean. These deficit factors are calculated as functions of a water availability ratio (i.e., actual to potential soil water supply) and can take on values between 0 and 1, where a value of 1 indicates no stress. The functions which determine the maize soil water deficit factors are shown in Figure 2.3b. Like water stress, N stress impacts photosynthesis, expansion, phenology, and grain filling through the computation of N deficit factors in the *Maize* module. The *Plant* module only includes N deficit factors for photosynthesis, phenology, and grain filling. These factors are calculated the same way in the two modules where the computed N concentration ratio for the plant stover is scaled by a different constant factor for each process. A scalar of 1, 1.25, 0.8, and 5.75 are used to slow grain N concentration, radiation use efficiency, leaf area expansion, and phenological development, respectively. Again, a deficit value of 1 indicates no stress, such that crop phenology is least affected by N stress.

Ensemble Kalman filter

Within this analysis, observed soil moisture data were assimilated into the APSIM model using the 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; Chakrabati 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—an ensemble of model forecasts and observed data—into an analysis distribution using Bayes' theorem.

$$P(X_t|Y_t) \sim P(Y_t|X_t)\, P(X_t)$$

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

$$y_{t\,=}\, HX_t + \varepsilon$$

$$\varepsilon \sim N(0, R_t)$$

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 EnKF computes the analysis distribution (i.e., the posterior) at each time step using the Kalman Gain (K). The result is the weighted mean of the forecast and observation distributions based on their respective precision values. The resulting posterior distribution is Normal with mean vector X_a and P_a.

$$K_t = P_{f,t}H^T(R_t + HP_{f,t}H^T)^{-1}$$

$$X_{a,t} = X_{f,t} + K_t(Y_t - HX_{f,t})$$

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

In the presented system, the EnKF was used to compute the analysis distribution at the end of each day where observations were available. Each model ensemble forecast was updated from the analysis distribution based on its respective likelihood within the forecast distribution. Thus, the analysis distribution was used as the initial condition for the model forecast into the next time step and could have potentially constrained any model process in the next time step which depended on the assimilation state variables.

Miyoshi algorithm

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). 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), there was insufficient information to accurately quantify R in this study. In addition, due to high computational cost, the potential ensemble size for this analysis was relatively small (n = 50), which limited the accurate representation of P_f. To overcome these issues, a method presented by Miyoshi et al. (2013) was adapted 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 the presented system, it did not function to exactly estimate both values for a given analysis time step. Rather, the algorithm used 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. Three notable changes to the algorithm presented in Miyoshi et al. (2013) were made to better suit the needs of this analysis. First, a constraint was applied to estimates of P_f such that variance values never dropped below 1. This ensured that the algorithm was inflating and never shrinking forecast uncertainty. Second, an inflation matrix was estimated rather than an inflation scalar to account for possible scale differences across state variables. Only the diagonal terms of the computed inflation matrix were considered so that only forecast variance (not covariance) was inflated. Lastly, observation errors were assumed to be independent between state variables at each time step, and, therefore, only the diagonal elements of R were estimated.

The Miyoshi algorithm was appended to the assimilation workflow as an offline estimator at time t for Δ_{t+1} and R_{t+1}. Prior to the start of assimilation, estimates of Δ and R were initialized as

$$\Delta_1 = I$$

$$R_1 = \Sigma$$

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 were used to compute the analysis distribution as follows:

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

$$X_{a,t} = X_{f,t} + K_t (Y_t - H X_{f,t})$$

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

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

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

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

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

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

where d_{o-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, a lower bound of 1 was imposed on the estimated values of Δ_{est}. Finally, the algorithm proposed 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 combined the current values with the new estimates in a weighted average,

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

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

where ρ is a user-defined weight given to the new estimate. This analysis used ρ = 0.05 to smooth noisy estimates and ensure that a single estimate of observation error at time t could not heavily influence the error estimates informed by all previous time steps.

State-parameter data assimilation

In addition to 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 this analysis, the innovations from a preliminary SDA simulation were used to determine which soil water flow type was associated most with large prediction error. This approach was taken to maximize the potential for error reduction.

Modeling platform

The complete modeling framework in this study consists of several diverse and important pieces that, together, allow for 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 the modeling workflow on the cluster, Docker containers generated and executed each of the 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 the presented system has the capacity to perform regional model-data fusion exercises across broad tiled spaces by leveraging pSIMS, the pSIMS functionalities for this analysis were utilized at a single pixel that best represented the study area. Additionally, for the purpose of this study, new features within the pSIMS platform were developed to perform ensemble-based simulations and include uncertainty in soil, weather, model parameters, and initial conditions for a single site. Given a fixed number of ensemble members and a series of priors on cultivar parameters, the uncertainty propagation workflow within the pSIMS platform used a Monte Carlo sampling approach to generate random samples of soil, weather, and cultivar parameters for each ensemble member. Dokoohaki et al. (2021) described these changes in more detail.

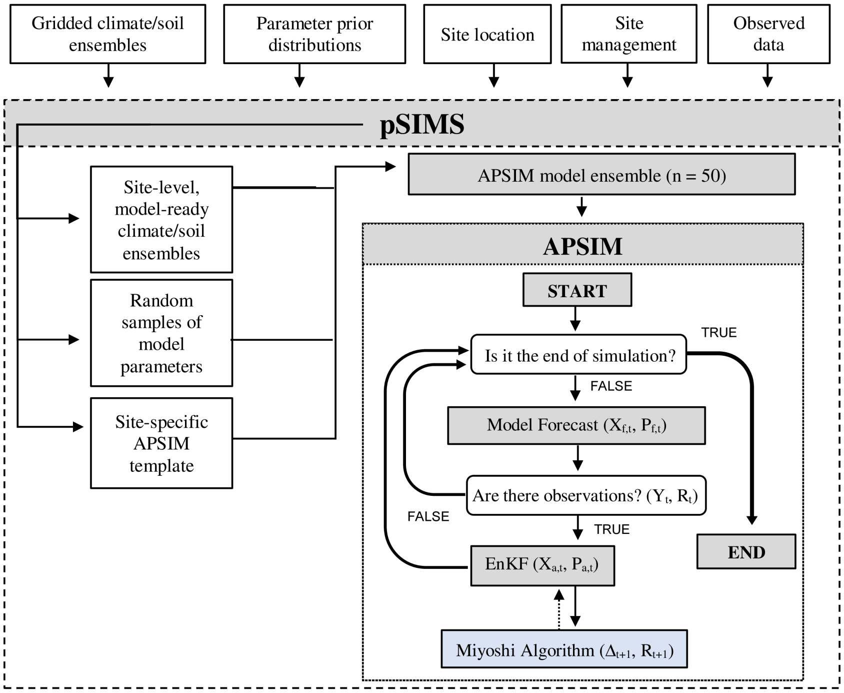
Gridded climate/soil ensembles
Parameter prior distributions
Site location
Site management
Observed data
pSIMS
Site-level, model-ready climate/soil ensembles
Random samples of model parameters
Site-specific APSIM template
APSIM model ensemble (n = 50)
APSIM
START
Is it the end of simulation?
TRUE
FALSE
Model Forecast ($X_{f,t}$, $P_{f,t}$)
Are there observations? (Y_t, R_t)
FALSE
TRUE
END
EnKF ($X_{a,t}$, $P_{a,t}$)
Miyoshi Algorithm (Δ_{t+1}, R_{t+1})

Figure 2.4. Flowchart of full data assimilation workflow. The Miyoshi algorithm steps, which are shown in blue, are not included in the SDA simulation workflow. Dashed arrows signify that the movement occurs in the next iteration (i.e., t + 1).

System set-up

The features presented above comprise the fundamental pieces of the full data-assimilation system. On top of the "Dockerized" pSIMS platform, crop model forecasts were performed using APSIM, which operated in series with the ensemble Kalman filter and the Miyoshi filter tuning algorithm, which were built into the model using the APSIM's C# manager functionality and were called at the end of each day's forecast. The full overall workflow is demonstrated in Figure 2.4.

The data-assimilation system was incrementally developed and tested with four series of simulations for the study site. These series will be referred to as schemes hereinafter. Table 2.1 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

Table 2.1. Overview of simulation components and naming conventions

Simulation name	Workflow components	Variables included in X_f
Free	APSIM	None
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)

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, associated model parameters were randomized across the model ensemble to account for uncertainty, where parameter values were drawn 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 weight 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 soybean cultivar, a set of APSIM-defined cultivars was selected based on preliminary performance in free model simulations and, then, cultivars were randomly assigned within the model ensemble. For summarized information on parameter priors and fixed management parameters, see Table A.1 and Table A.2.

Ensemble weights

An ensemble weighting strategy was applied to interpret and evaluate results from each tested scheme more accurately. The 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 the best available forecasts, the following ensembles weight strategy was applied.

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

$$P(\, X \mid \mu_a, P_a \,)$$

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. The weights were normalized for each time step across all ensembles (i.e., summed to 1), and later, the average weight of each ensemble was computed for each year. The free model ensembles were given equal weights as no posterior distribution was computed.

Evaluation statistics

The spectral norm (||.||$_2$), which represents the maximum singular value of a matrix, was computed to compare differences in forecast precision of assimilated state variables across schemes. The spectral norm represents the magnitude of P_f for a given scheme and can be compared to identify how forecast precision within each simulation changes with time. The spectral norm of P_f was calculated as

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

where P_f^H represents the conjugate transpose of P_f. A weighted variance was used to quantify the precision of each simulation scheme in estimating all other relevant model variables. This value was calculated for annual output values as

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

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

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

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 using the root mean squared error (RMSE), calculated as

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

where y_{annual} is the annual observed value. RMSE was also used for comparing daily output values,

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

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.

RESULTS

In this section, the four different data assimilation schemes in this study are compared and the most robust scheme for soil moisture estimation in both accuracy and precision is identified. Then, the performance of the optimal scheme is evaluated and compared with the free model in estimating daily soil moisture, soil N, LAI, annual yield, tile flow, and annual NO_3 loads.

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.1) and trends in soil moisture throughout the soil profile (Fig. 2.5). 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 overpredicted available soil moisture, which impacted downstream model estimates of crop water uptake, crop development, and tile flow, among others.

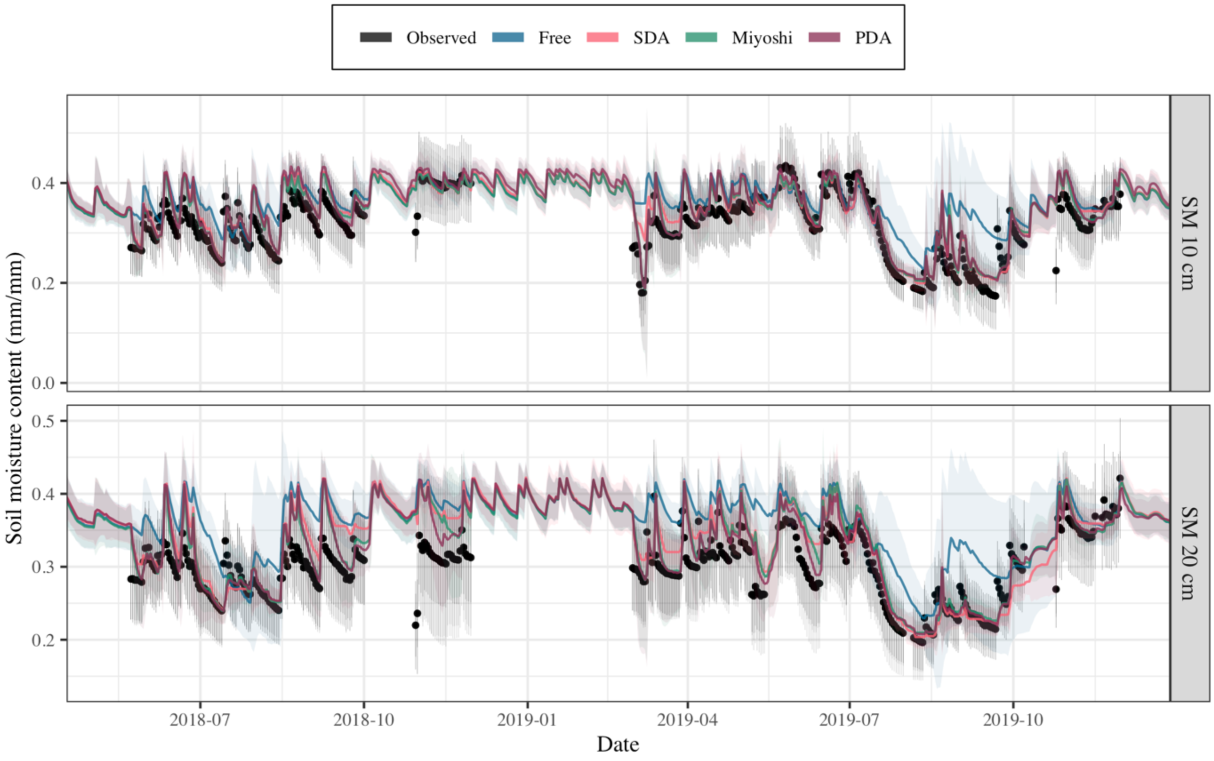

Figure 2.5. Time series of simulated and observed soil moisture estimates for the 2 soil layers where assimilation is performed within the soil profile. SM3 refers to 9.1-16.6 cm (observed at 10 cm), and SM4 refers to 16.6-28.9 cm (observed at 20 cm). 95% confidence intervals are shown surrounding the 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.

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 2.2). Figure 2.6a shows a smoothed time series of the spectral norms of P_f for each simulation. The forecast uncertainty for all simulation series was high at the initiation of 2018 following a wide prior on initial soil water balance but then dropped by the spring of 2018. However, the SDA forecast uncertainty (as well as that of the other two assimilation schemes) remained low for the duration of the simulation period. The free model, on the other hand, experienced large jumps in uncertainty during both growing seasons, which may 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 outperformed the free model by constraining soil water dynamics across the full parameter-input space.

Yet, despite major improvements in forecast accuracy and precision, SDA showed filter divergence. An overestimated R and an underestimated P_f provided inaccurate weighting of the observed data and the model. As a result, the filter mostly ignored the observed data and overemphasized the forecast distribution in the computation of the analysis distribution. By including the Miyoshi algorithm as an offline estimator of forecast and observed variances, this type of filter behavior mostly disappeared in Miyoshi (Fig. A.2). Assuming divergence to be where the observed mean did 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 was 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 tested in this study was parameter data assimilation. In preliminary analyses of SDA innovations, the module's prediction error for both soil layers was often found to be 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 pointed to the use of the saturated flow model processes, the *SWCON* model parameter for both soil layers (10 and 20 cm) was selected for update 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:

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

In PDA, the *SWCON* model parameter was adjusted for both layers at each analysis time step according to the covariance between the parameter and the observed state variables. Though there were shifts in estimates of the two *SWCON* parameters with PDA (Fig. 2.6b), 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 allowed for more flexibility in the model, which served as an added benefit compared to Miyoshi. For this reason, the rest of

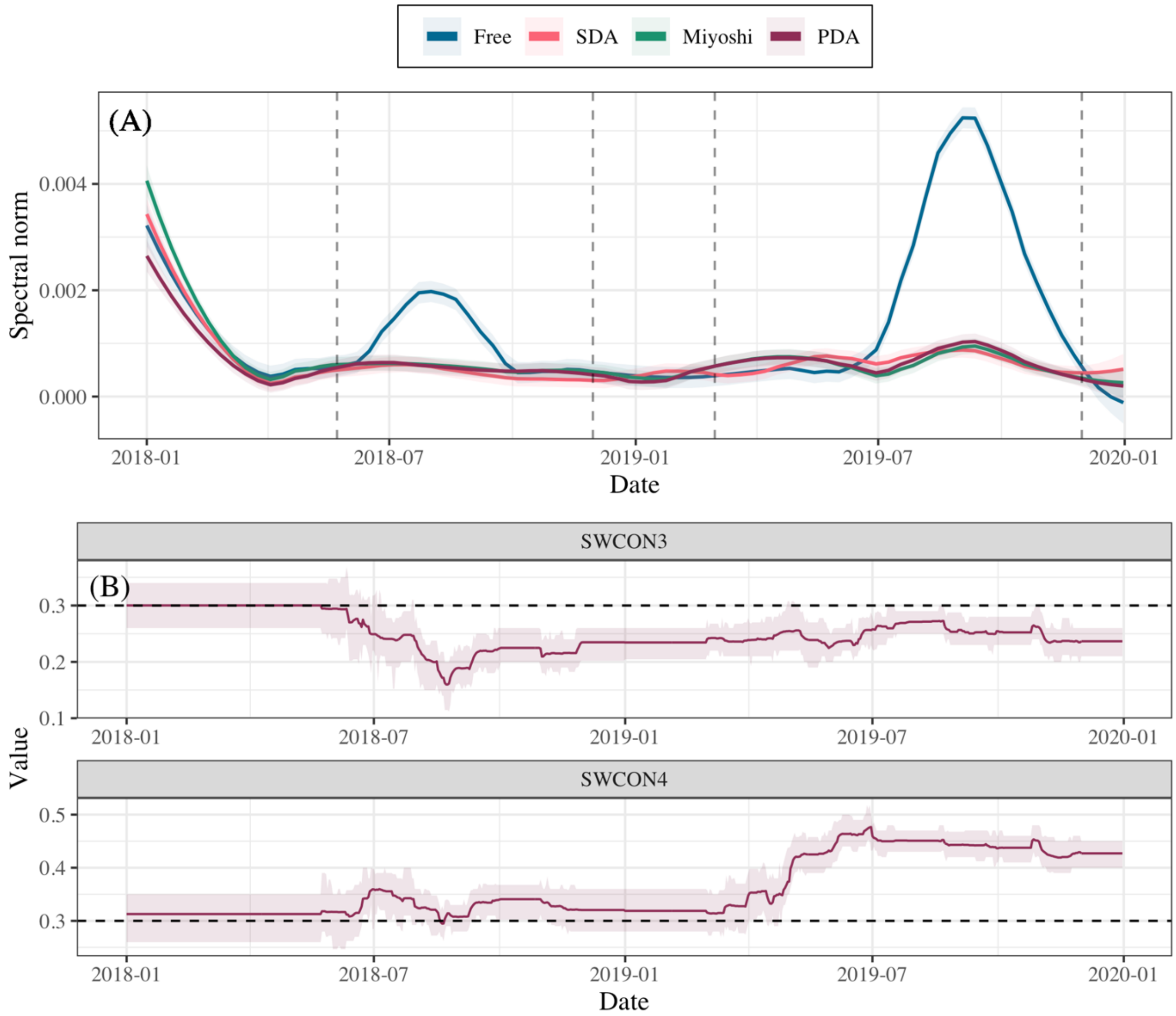

Figure 2.6. (a) Time series of the spectral norm of the forecast covariance matrix (i.e., $||Pf||_2$) for SM3 and SM4 for all simulations for both years with local regression (LOESS) smoothing applied ($\alpha = 0.25$). Assimilation periods for both simulation years are indicated with the two sets of 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.

this section focuses on a comparison between the free model and the best-performing and most comprehensive data assimilation scheme: PDA.

Soil moisture

Estimates of soil moisture from the different simulation schemes are shown in Figure 2.5, and Table 2.2 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

Table 2.2. Comparison of soil moisture forecast accuracy and precision metrics

Layer	**Depth range**[a] *cm*	**RMSE** *proportion*				**Average variance** *1 x 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.

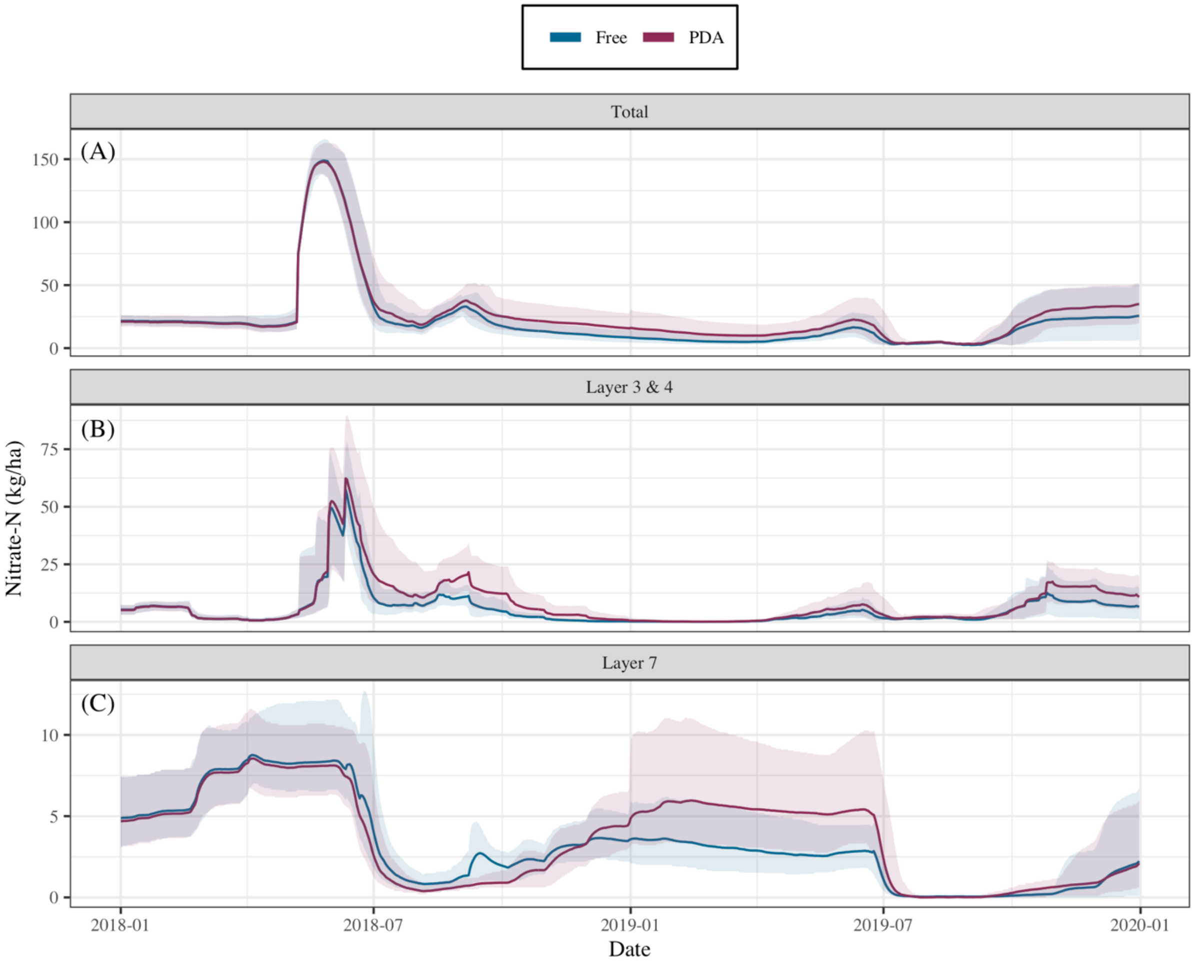

Figure 2.7. (a-c) Time series of simulated soil NO3-N in (a) the total soil profile (b) the assimilation layers (i.e., Layers 3 and 4), and (c) the lowest soil layer (i.e., Layer 7). 95% credibility intervals are indicated by the shaded ribbon surrounding the mean lines for each scheme.

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.

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, there were 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. 2.7a-b). These differences are suspected to be 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 Figure 2.3a, this difference had 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 was 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. 2.8c).

Leaf area index and annual yield

Aboveground measures of crop production, including LAI and annual yield, were less affected by assimilation, and changes in forecast 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, PDA improved 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.

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. 2.8a-b). As the free model often overestimated soil moisture in the two assimilation layers, it is suspected that constraining soil moisture in PDA decreased the amount of soil water in the assimilation layers and, thus, decreased the

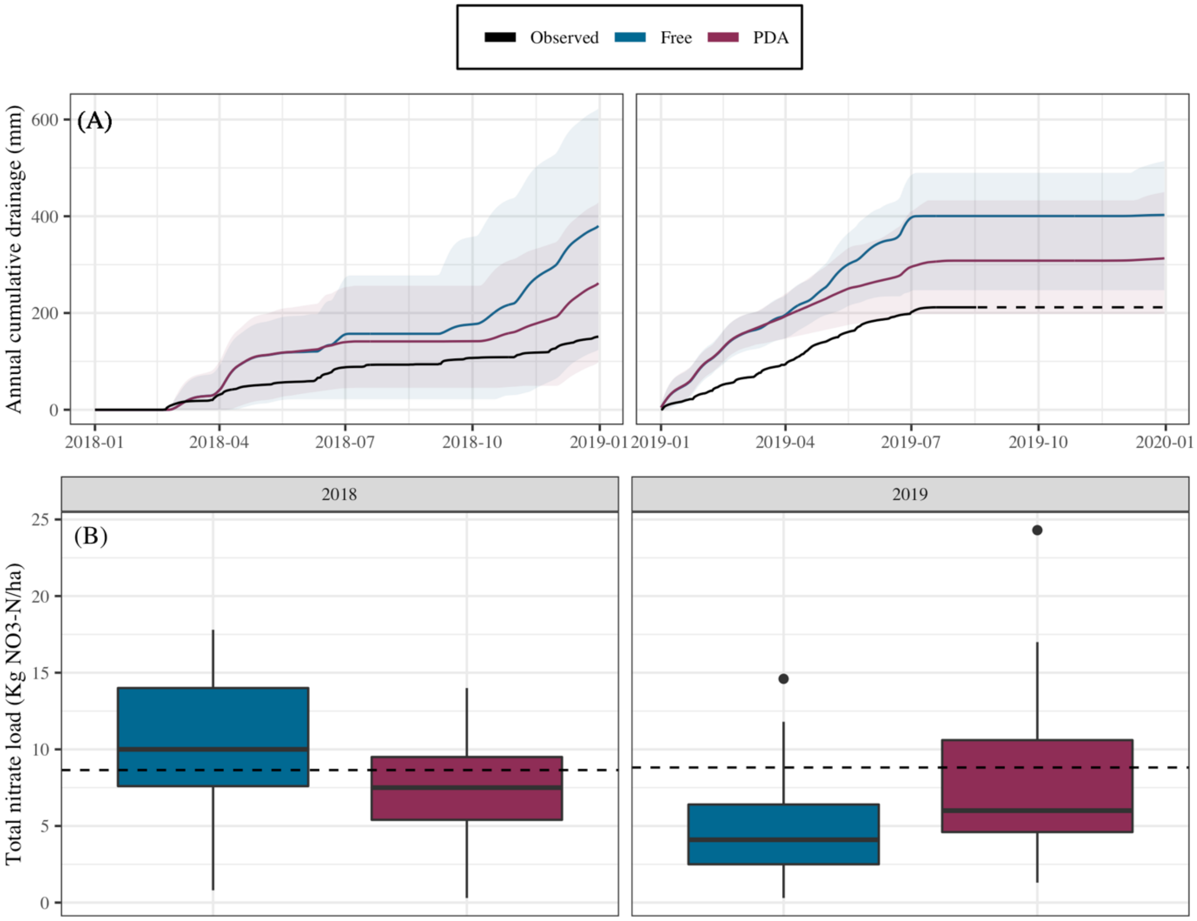

Figure 2.8. (a) Time series of simulated and observed cumulative annual tile drainage (mm) for 2018 and 2019 from the study plot. 95% credibility intervals are indicated by the shaded ribbon surrounding the mean lines for each simulation. Black lines demonstrate the observed trends. Due to missing data from the end of 2019 as discussed in Sec. 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.

amount of soil water drained from the system. Constraint of annual tile drainage with PDA was especially strong in 2018, where there was great improvement in both accuracy and precision. This constraint was weaker in 2019, where there was 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 2.8c). Such a large increase in uncertainty in 2019 likely stemmed from the large uncertainty associated with PDA estimates of NO_3 in the lowest soil layer (Fig. 2.7c). On the other hand, considering daily estimates of NO_3 load over the course of the simulation period, there was only a small 5.8% increase in accuracy and an 18.2% decrease in precision with PDA.

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 this study, crop yield estimates did not tell the full story. There are 4 key points to highlight within the results:

1. PDA effectively constrained soil moisture estimates for the two assimilation layers. One of the downstream impacts of this constraint was better soil moisture estimates for the two deeper layers (75 and 100 cm), where there were improvements in both forecast accuracy and precision with PDA (Table 2.2). 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 (Sec. 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 the presented results.

2. The presented 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.3). 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, PDA estimated greater root-zone soil NO_3 compared to the free model, which is believed to be the result of lower estimates of soil moisture in the two assimilation layers leading to changes in N cycle process rates. The increased availability of soil NO_3 may have 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 the presented data-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 were 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, lower estimates of NO_3 load would be expected 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 this study, 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 it to other parts of the model (e.g., evaporation, crop water uptake, other soil layers, etc.). For this 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 did 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 the presented 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 assumed the estimates were correlated with the SWCON model parameters for the two layers and adjusted 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 was 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 presented simulations and NH_4 uptake has lower energy requirements than NO_3 uptake in crops and is, therefore, preferred (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. No other known system 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 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, weather, 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 (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, one could 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.

Yet, further investigation is necessary to validate the performance of the presented assimilation system prior to broader application. In particular, the system should be applied for a range of sites across the U.S. Midwest where sufficient observations are available for the application and evaluation of the presented system. By increasing the sample size of this study, more robust conclusions can be drawn on which downstream model states can be constrained by the presented soil moisture data-assimilation system and where, when, and why the system's constraint is effective. This is the central purpose of Chapter 3.

CHAPTER 3

EXPANDED APPLICATION AND EVALUATION OF THE DEVELOPED SYSTEM ACROSS THE UNITED STATES MIDWEST

INTRODUCTION

To effectively address important large-scale agricultural issues, agricultural forecasting tools must exhibit high system performance at broad spatial and temporal scales. However, typical forecasting methods are often limited in their inference space due to limitations in observations and/or calibration and validation methods (Dietze et al., 2013). These limitations are especially strict when forecasting agricultural state variables due to the complexity of underlying biophysical processes (Silva and Giller, 2020). When using low dimensional calibration approaches to optimize complex, nonlinear models (i.e., process-based crop models), the final parameterized model could give the right answer for the wrong reason (e.g., fortuitous cancellation of errors). Such models would not be reliable for further application (van der Laan et al., 2014). For example, Pasley et al. (2021) calibrated the APSIM model to estimate yield, drainage, and NO_3 leaching using 56 site-years of data from sites across the U.S. Midwest and reported satisfactory performance in estimating NO_3 leaching. However, though performance was acceptable overall, the final calibrated model did not perform consistently across sites, with prediction errors in cumulative NO_3 leaching estimates as large as 100 kg NO_3-N/ha for some locations (see Supplementary Figure 8 in the work). These site-level inconsistencies were not explored in the published work. Li et al. (2014) experienced greater success in their calibration of the DNDC model for estimating NO_3 leaching in northern China. After calibration, RMSE dropped from 16 to 4 kg N/ha for NO_3 leaching estimates at their calibration site; however, their calibrated model was still not able to estimate the high leaching rate at one of their validation sites. They also note two other limitations—namely, regional data scarcity, soil heterogeneity, and low representation of management schemes—that restrict the application of their calibrated model to sites that fall outside of their study region and/or follow different tillage or crop rotation practices. Thus, it is clear that calibration or validation data must account for the true variability in soil properties, climate, management, etc. for the entire study region when training a process-based model for broader application. However, considering the resources and time that would be required to calibrate a process-based model across all of these dimensions, it is unlikely that current methods of process-based modeling will ever be able to answer complex, large-scale agricultural research questions with accuracy and precision.

On the other hand, machine learning (ML) approaches can achieve higher accuracy in predicting agricultural state variables across broader areas by fitting simpler models. These practices can for spatial variability by including representative predictors in the model. However, due to the heterogeneity of agricultural landscapes, these fitted models are limited in their inference space, as well as in the resolution in which they can be interpreted (Flathers and Gessler, 2018). For example, a machine learning approach by Spijker et al. (2021) employed nitrate data from a national monitoring program and an array of auxiliary spatial datasets to predict nitrate concentration in leachate from agricultural soils in the Netherlands using a Random Forest model. The final model performed satisfactorily, explaining about 58% of the variability in nitrate concentration estimates at the farm scale. However, even within the bounds of the study region, the fitted model generated "unexpected" results when applied to areas that were not included in the model training dataset and differed in farm management, land use, etc. Moreover, the application of the fitted model to other regions would require a similar intensive, large-scale auxiliary dataset as the one used in the study, which would necessitate immense time and resources. Thus, the presented approach in Spijker et al. is not conducive to broader forecasting applications. ML models are limited as they cannot be easily applied to areas that fall outside of the training parameter space and can still require intensive data collection. A study by Hoffman et al. (2020) revealed another shortcoming of ML approaches. In the work, they fit a Random Forest model to investigate the relationship between sorghum, maize, and soybean yields, technological advancements, and climate variables for counties in the U.S. Midwest. Although their fitted model was highly accurate in predicting yield for their compiled dataset and provided insights on crop sensitivity to climatic and technological conditions, their fitted model cannot provide detailed insights on the specific agronomic practices or agricultural processes that contributed to the high spatiotemporal variability in yields. ML approaches are also limited in the new information they provide on underlying agroecological processes as, inherently, they can only partially explain variability for a single state variable along explicitly measured dimensions.

To overcome the inference limitations of common forecasting methods, studies have shown that SDA can help reduce the need for site-level model calibration and the possibility of overfitting a model by systematically constraining model processes that are highly variable in space and time. Guerif and Duke (2000) were among the first to propose data assimilation as a substitute for site-level model calibration. They assimilated spectral reflectance information into a coupled SUCROS-SAIL model and were able to improve regional estimates of sugar beet yield in northern France through the constraint of LAI, as well as cultivar and management parameters. Similarly, Lu et al. (2021) assimilated *in situ* soil moisture and canopy cover information into the AquaCrop model for 6 years at an experimental station in Nebraska, U.S. to test SDA's ability to account for heterogeneity in maize cultivar. Their EnKF-based assimilation framework was able to better capture variability in maize phenology compared to the free model, reducing yield nRMSE from 18.61% to 11.48%. Lastly, a synthetic study by Zhu et al. (2017) found that the assimilation of coarse resolution surface soil moisture data into a water flow model could estimates of soil moisture in the first 50 cm of the soil profile despite explicitly unaccounted spatial heterogeneity in soil properties.

These studies provide clear evidence that SDA can efficiently overcome the need for explicit model parameter calibration when performing model simulations.

However, despite overall improvement in model forecasts, many past SDA studies have reported inconsistent model performance in downstream model constraint from site to site. For example, de Wit and van Diepen (2007) observed inconsistencies in yield constraint when assimilating remotely sensed soil wetness index (SWI) observations into the WOFOST model across agricultural regions of Spain, Germany, France, and Italy. They partially attributed poor predictions in certain regions to irrigation processes that were not captured by the model nor in the coarse resolution SWI observations. The abovementioned study by Lu et al. (2021) also saw year-to-year variability in the performance of their assimilation framework. When assimilating soil moisture independently, canopy cover estimates were better constrained in drier years. They suspected this to be the result of canopy cover's lower sensitivity to soil moisture when water is in surplus. Such trends in system behavior are important to highlight to identify the conditions where system performance is more reliable as well as to drive further improvement and advance performance more broadly. However, to investigate these trends, a system must be applied and evaluated across a variety of management and environmental conditions.

In Chapter 2, an optimal data-assimilation system was developed that can integrate a variety of observations, propagate system uncertainties, and generate robust agricultural forecasts of several important state variables without the need for site-level model calibration. The system was applied to simulate a corn-soybean rotation at an experimental site in Illinois, and, by assimilating *in situ* soil moisture observations, the system constrained estimates of soil moisture, tile drainage, and, to a lesser extent, NO_3 leaching at the study site. Aboveground measures of crop productivity (i.e.., crop yield and LAI) were not well-constrained. However, this was hypothesized to be the result of simulating a non-water limited system, where changes in soil moisture would not greatly affect crop growth. Overall, the results from Chapter 2 demonstrate great potential for the developed data-assimilation system to be a powerful tool in agricultural forecasting. However, as demonstrated by past studies, the performance of an SDA system for one site-year does not necessarily reflect how it will perform in another site-year. Since it is imperative that agricultural forecasting tools put forward for large-scale application exhibit high performance across broad environmental and management conditions, further evaluation of the system across different climate, management, and soil spaces is necessary to validate the results in Chapter 2. This need for further evaluation was the main driver of this chapter. In this study, the developed data-assimilation system was applied to assimilate *in situ* soil moisture observations for 19 site-years across the U.S. Midwest, a feat made possible by the remarkable database of field experimental observations compiled by the Transforming Drainage project. After application, the system's downstream constraint of APSIM estimates was evaluated and trends in system performance, as well as areas for future system improvement, were identified and discussed.

This chapter has two main objectives:

1. To test and evaluate soil moisture SDA in the presented system as a method for constraining downstream APSIM estimates of soil moisture, crop yield, leaf area index, tile drainage, and NO_3 leaching across several study sites in the U.S. Midwest.
2. To highlight the strengths and weaknesses of the presented system to drive future application and development.

MATERIALS AND METHODS

Study sites

In this chapter, the developed data-assimilation system is tested across 5 different sites in the U.S. Midwest and 19 site-years spanning 2011-2019. The 5 study sites are in 5 different states and include the Energy Farm (i.e., the study site from the previous chapter) and 4 experimental sites available through the Transforming Drainage (TD) project (Chighladze et al., 2021). The TD project database contains high-quality data from 42 research sites on an array of agricultural variables, including tile drainage, yield, water table, water quality, and soil characteristics, among many others. Of the numerous sites available as part of the project, the sites and years selected for this work included plots with the following:

- A free tile drainage system
- Available NO_3 load and tile flow data at the plot level
- Available soil moisture sensor data
- Maize and/or soybean cropping systems

To set up the APSIM model for each of the 5 sites, we included all available site information on year, cropping system, residue type, planting details, harvest date, tillage practices, and fertilizer applications as constants in the simulations. All sites were rain-fed, and tillage practices were included for all sites due to increased knowledge of Energy Farm management practices as defined by Moore et al. (2021). Site locations are shown in Figure 3.1, and plot and management information for the 5 sites is given in Table 3.1. Study sites will be referred to by their given study ID in the table hereinafter.

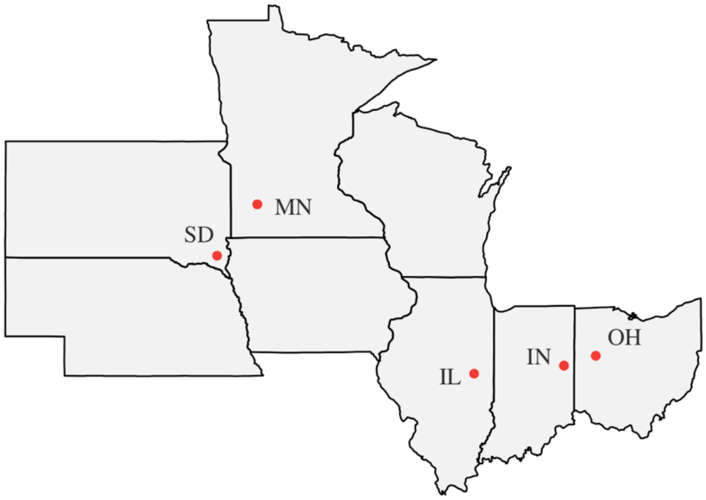

Figure 3.1. Geographic location of the study sites in this study.

Observed data

The following sections highlight the additional data products that were considered in this chapter relative to those described in Chapter 2. They also indicate any changes that were made to the data presented in Chapter 2 for the sake of this study.

Model drivers

The two model drivers applied in Chapter 2 were also employed in this analysis (i.e., SoilGrids and ERA5). 25 soil ensembles were generated for each site location, and the depth of each soil profile was reduced to the depth of the drainage tile placement at each study site to appropriately simulate vertical water movement (Table 3.1). The observed weather ensemble used in Chapter 2 for the Energy Farm was excluded from this analysis, and only 10 weather ensembles were used at each of the 5 sites to ensure consistency in inputs. Observed weather data were available through the TD database for all but one site-year among the TD sites in this analysis. These daily observations of precipitation, air temperature, and solar radiation were leveraged to identify climatic trends in system performance and soil moisture innovations through correlation analyses and feature selection.

Soil moisture

Soil moisture information for the TD sites is available as daily averages for an array of soil depths and is measured as the volumetric water fraction at each depth. This analysis focuses on the soil moisture data available for the 10- and 20-cm depths, which will be referred to as SM3 and SM4, respectively, hereinafter. To ensure data consistency across the study sites, the daily soil moisture values for the Energy Farm were re-computed as daily averages instead of end-of-day averages (as used in Chapter 2). Days

Table 3.1. Site management information as defined across all APSIM simulations in this study.

Study ID **(*Original ID)***	**Site information**	**Year**	**Crop**	**Planting date**	**Harvest date**	**Tillage**	**Fertilizer**[c]
IL (*Energy Farm*)	Plot ID: *Maize Control* Soy cultivar: *3.5 – 4.5* Tile depth: *1.383 m*	2018	maize	08 May	09 Oct	08 May: *chisel 50 mm*[a] 24 Oct: *chisel 150 mm*	08 May: *urea_N (101)* *NH_4NO_3 (288.57)*
		2019	soy	17 May	09 Oct	*N/A*	*N/A*
IN (*IN_Randolph*)	Plot ID: *SW* Soy cultivar: *3.0 – 4.0* Tile depth: *0.975 m*[b]	2011	soy	07 Jun	24 Oct	*N/A*	*N/A*
		2012	maize	23 Apr	10 Oct	12 Apr: *disc 50 mm*[a] 28 Nov: *disc 200 mm*	13 Feb: *NH_4_N (22.42)* *broadcast_P (46.25)* 23 Apr: *urea_N (17.88)* *NH_4NO_3 (51.09)* *banded_P (13.96)* 25 May: *urea_N (100.35)* *NH_4NO_3 (286.7)*
		2013	soy	20 May	14 Oct	09 Apr: *disc 50 mm*[a] 21 Oct: *chisel 150 mm*	21 Mar: *NH_4_N (18.48)* *broadcast_P (38.12)*
		2014	maize	27 Apr	21 Oct	13 Nov: *chisel 150 mm*	24 Apr: *NH_4_N (17.93)* *broadcast_P (36.99)* 23 Apr: *urea_N (17.88)* *NH_4NO_3 (51.09)* *banded_P (13.96)* 25 May: *urea_N (108.82)* *NH_4NO_3 (310.9)*
		2015	soy	06 Jun	12 Oct	*N/A*	*N/A*

Table 3.1. (cont.)

Study ID (*Original ID)*	**Site information**	**Year**	**Crop**	**Planting date**	**Harvest date**	**Tillage**	**Fertilizer**[c]
IN (cont).		2016	maize	26 Apr	07 Oct	*N/A*	26 Apr: *urea_N (16.71)* *NH_4NO_3 (47.73)* *banded_P (13.05)* 02 Jun: *urea_N (104.91)* *NH_4NO_3 (299.7)*
MN (*MN_Redwood1*)	Plot ID: *BE* Soy cultivar: *1.5 – 2.5* Tile depth: *1.22 m*	2012	maize	10 May	06 Oct	06 May: *disc 76.2 mm*[a] 01 Nov: *rip 228.6 mm*	06 May: *urea_N (177.1)* *NH_4_N (13.4)* *broadcast_P (34.2)* 10 May: *NH_4_N (7.84)* *banded_P (76.2)*
		2013	maize	24 May	31 Oct	23 May: *disc 76.2 mm*[a] 03 Nov: *rip 228.6 mm*	22 May: *urea_N (182.75)* *NH_4_N (13.4)* *broadcast_P (34.2)* 24 May: *NH_4_N (7.84)* *banded_P (11.63)*
		2014	maize	17 May	29 Oct	16 May: *disc 76.2 mm*[a] 01 Nov: *rip 228.6 mm*	16 May: *urea_N (150.47)* *NH_4_N (12.05)* *broadcast_P (30.8)* 17 May: *NH_4_N (7.84)* *banded_P (11.63)*
		2015	maize	30 Apr	13 Oct	29 Apr: *disc 50.8 mm*[a] 27 Oct: *rip 228.6 mm*	28 Apr: *urea_N (148.37)* *NH_4_N (14.15)* *broadcast_P (18.6)* 01 May: *NH_4NO_3 (47.8)* *urea_N (16.49)*

Table 3.1. (cont.)

Study ID (*Original ID*)	**Site information**	**Year**	**Crop**	**Planting date**	**Harvest date**	**Tillage**	**Fertilizer**[c]
MN (cont.)		2016	soy	13 May	18 Oct	11 May: *disc 50.8 mm*[a] 01 Nov: *rip 228.6 mm*	*N/A*
		2017	maize	06 May	03 Nov	06 May: *disc 50.8 mm*[a]	06 May: *NH_4_N (7.84)* *broadcast_P (11.63)*
OH (*OH_Auglaize2*)	Plot ID: *WS* Soy cultivar: *3.0 – 4.0* Tile depth: 0.975 *m*[b]	2013	maize	09 May	22 Oct	*N/A*	09 May: *broadcast_P (20.53)* *NH_4_N (11.76)* *urea_N (201.43)*
		2014	soy	15 May	20 Oct	05 Nov: *disc 200 mm*	*N/A*
		2015	maize	30 Apr	16 Oct	*N/A*	30 Apr: *NH_4_N (18.38)* *broadcast_P (38.12)* *urea_N (178.76)*
SD (*SD_Clay*)	Plot ID: *Plot7* Soy cultivar: *2.0 – 3.0* Tile depth: *1.22 m*	2016	maize	18 May	21 Oct	15 May: *disc 101.6 mm*[a]	14 Apr: *urea_N (180.32)*
		2017	soy	02 June	13 Oct	*N/A*	*N/A*

[a] Documentation on site-level management indicated the use of a field cultivator during spring tillage for several site-years. However, since the APSIM tillage module does not include parameterization for a field cultivator, the *disc* implement was applied at the documented depth instead due to similarities between the two implements in relation to incorporation.

[b] OH and IN both fall within the same tile of the gridded soil driver. Therefore, since the drainage tiles were placed at similar depths at the two sites (i.e., 0.91 and 1.04 m), the soil profile depth was adapted to the average depth of the two for simplicity.

[c] This column includes information on fertilizer application date, type, and amount as defined for each site-year. The notation for fertilizer type reflects fertilizer names in APSIM. If the fertilizer name contains *N* or *P*, amount is in Kg N/ha or Kg P/ha, respectively. Otherwise, amount is in Kg fertilizer/ha.

with fewer than 40 observations for a given day were excluded. A comparison of end-of-day soil moisture values from the Energy Farm (Chapter 2) with the newly computed daily average values show their differences to be negligible with a mean absolute difference of 0.006 mm/mm, a finding similar to that of Dietzel et al. (2015). Across all sites, data points from the winter months (December-March) were excluded from assimilation to avoid possible sensor inaccuracies related to freezing soils. Due to limited replication or limited information on replication, a 10% observation error is assumed around the mean for all soil depths. This observation error is estimated by the Miyoshi algorithm for the two assimilation layers. All other available soil moisture observations for lower soil layers were used in the evaluation of the downstream impacts of assimilation on the soil water profile. Observations were paired with an APSIM soil layer based on the recorded sensor depth and the site soil profile. The average soil moisture was computed for each day and layer with the assumption of uniform soil moisture content in each APSIM layer at a given time.

Crop yield and NDVI

Data on harvested yield are available for the TD sites was converted from grain at standard moisture content (i.e., 15.5% for maize and 13% for soybean) to dry-grain weight for best comparison with APSIM output. In Chapter 2, LAI observations from the Energy Farm were leveraged to evaluate APSIM's simulation of crop phenological development. However, since reliable and consistent measurements of LAI were not available for the TD sites, another type of observation was needed to serve this same purpose. The normalized difference vegetation index (NDVI) is a remote-sensing data product that can be used to quantify vegetation cover and reasonably track the phenological development of crops (Gao and Zhang, 2021). In this study, NDVI is used instead of LAI to assess APSIM's understanding of crop phenology at each of the tested site-years. NDVI time series were extracted for each of the sites from Landsat 7 remote sensing imagery via Google Earth Engine.

Tile flow and nitrate loads

Data tracking daily tile flow (mm) and daily NO_3 load (kg NO_3-N/ha) in the drainage water were available for all the TD sites considered in this analysis. Values were normalized by the total drainage area for each plot. Any missing daily values of drainage were imputed using an approach described by Helmers et al. (2022) and used to approximate missing daily values of NO_3. Helmers et al. (2022) provides further information on the specific methods and instrumentation used to collect and process these data at each of the TD sites. In this study, daily values for both tile flow and NO_3 load were summed to annual values for comparison. Days with NA values for tile flow were assumed to have no drainage and no NO_3 leaching.

Drought indices

Drought monitoring information from the United States Drought Monitor (USDM) was collected at the point level for each site-year using the Google Climate Engine. The drought monitor classifies drought intensity each week across the United States using five climatological indicators and local condition reports. Its classification system has the following categories: none (i.e., no drought conditions), abnormally dry, moderate drought, severe drought, extreme drought, and exceptional drought. In this study, the total number of observations between April and September with any drought classification above "none" was determined for each site-year. This value characterizes the proportion of weeks during the growing season where a site was drier than normal, and it is used in this analysis to better understand how drought impacts assimilation system performance.

Data-assimilation system

The workflow tested in this chapter closely resembles the Miyoshi iteration of the workflow presented in Chapter 2 in overall structure of the workflow and the major components (i.e., pSIMS, APSIM, EnKF, Miyoshi algorithm). However, the following sections note the important changes made to the data-assimilation system as presented in Chapter 2 that defined the system used in this chapter.

APSIM Operations module

To help reduce the resources required for site-level APSIM simulations, the *Operations* module was introduced into the modeling framework to enable multi-year simulations in APSIM and eliminate the time-consuming process of running multiple single-year simulations. The *Operations* module allows for the specification of day and year (rather than only day) when defining management events for an APSIM simulation and, thus, increases the flexibility of the model, allowing for crop rotations and changing management practices over time (e.g., reduced fertilizer applications, no tillage, etc.). In this analysis, this flexibility allowed for the most representative specification of management practices in APSIM for each of the sites. In addition, this change allowed for flawless continuity in all system pools across different growing seasons at each location. In Chapter 2, such continuity was made possible for the soil nitrogen and soil water pools by the manual adjustment of model initial conditions; however, some pools, such as the surface organic matter pool, were not carried over between years. Any differences between the Energy Farm simulations presented in Chapter 2 and those presented in this chapter can be attributed to this change in model structure. See the APSIM documentation for more information on the *Operations* module and its functionalities.

Table 3.2. Updated prior distributions and descriptions of ensemblized maize cultivar parameters. "Norm(mu,sd)" indicates a Normal prior distribution with mean *mu* and standard deviation *sd*, and "Unif(lower,upper)" indicates a uniform prior distribution with lower bound *lower* and upper bound *upper*.

APSIM Parameter	**Description**	**Distribution**
largestLeafParams1	Intercept in a fitted exponential regression which predicts the area of the largest leaf from total leaf number in maize (Eq. 13, Birch et al., 1998)	Norm(-1.09497, 0.02947)
leaf_init_rate	Thermal time (degree days) to initiate each leaf primordium until floral initiation	Norm(25.12144, 0.18726)
leaf_app_rate1	Thermal time (degree days) required to develop a leaf ligule for first leaves	Norm(64.34875, 0.92697)
tt_emerg_to_endjuv	Thermal time (degree days) between emergence and end of the juvenile phase	Norm(420.08108, 9.27756)
tt_flower_to_maturity	Thermal time (degree days) between flowering and maturity	Unif(780, 860)
tt_flower_to_start_grain	Thermal time (degree days) between flowering and start of grain fill	Unif(150, 200)
tt_maturity_to_ripe	Thermal time (degree days) between maturity and harvest ripe	Unif(150, 250)
head_grain_no_max	Maximum potential number of kernels per ear	Unif(750, 900)
grain_gth_max	Potential growth rate of grain (mg grain/day)	Unif(7.1, 8.57)

Model parameter priors

Within model ensembles, initial soil water, cultivar, initial residue weight, and, if unavailable in management data, planting depth were randomized across model ensembles for each site. In the case of planting depth, separate prior distributions were set for each crop, maize and soybean, in order to reflect reasonable ranges for the two crops in the Midwest as described in extension websites produced by the University of Missouri (Luce, 2016) and Michigan State University (Staton, 2012). Using a uniform prior distribution, planting depth ranged from 1.5 to 2.5 inches for maize, and from 1 to 2 inches for soybean.

Prior distributions for cultivar parameters changed notably in this chapter. For maize, the number of "ensemblized" cultivar parameters was increased from 6 to 9 to include three additional leaf-related parameters. The decision to add new maize parameters was based on a global sensitivity analysis of the APSIM *Maize* module in the U.S. Midwest by Dokoohaki et al. (in prep) which identified the maize cultivar parameters to which LAI estimates were most sensitive. The fourth parameter identified in this analysis, *tt_emerg_to_endjuv,* had already been included from the previous analysis. The global optimized value distributions for these four parameters, as computed in Dokoohaki et al. (in prep) through a hierarchical

Bayesian optimization approach, were used as the prior distributions in this analysis. Table 3.2 gives more detailed information on all randomized parameters and their prior distributions. A preliminary assessment of the *Maize* module at each of the sites demonstrated that, under these parameter value ranges, APSIM was capable of appropriately simulating the phenological development and grain yield for maize at each site.

In contrast to the approach described in Chapter 2, the selection of soybean cultivars included at each site was determined using a semi-systematic approach. First, a range of maturity groups was determined for each site based on a study by Mourtzinis and Conley (2017) which delineated soybean maturity groups across the U.S. The range for each site was bounded using the contour lines shown in the Figure 4 of the published study. Initial APSIM simulations were performed for each site using all APSIM-defined soybean cultivars falling within the prescribed maturity group range. The model results were compared to the observed soybean yields at each site, and the best-performing maturity group *MG* for each site was determined. The final range for each site was *MG* ± 0.5. In each ensemble, cultivar for each crop at each site was assumed to be constant across site-years.

PROSAIL model

Since APSIM does not currently estimate NDVI, the PROSAIL model was coupled with APSIM within the larger modeling framework to estimate daily NDVI values and enable the appropriate evaluation of the model's simulation of crop phenology at the study sites. The PROSAIL model is a radiative transfer tool that combines PROSPECT, a leaf optical properties model, and SAIL, a canopy bidirectional reflectance model, to estimate spectral reflectance for a given vegetative area based on soil and plant/canopy properties. In this study, APSIM's daily forecasts of soil and plant variables were transformed and used as inputs into the PROSAIL model to compute the spectral reflectance for each ensemble following the method presented by Dokoohaki et al. (in prep). Then, for each day and ensemble, the estimated spectral information was used to estimate NDVI using the *vegindex* function within the *hsdar* R library.

System set-up

In this analysis, two distinct simulation schemes were used to simulate each site. Following the same notation as given in Chapter 2, a Free run and a Miyoshi run were completed for each site, with the former serving as a basis for comparison for the latter. For simplicity, the Miyoshi scheme will be referred to as the SDA scheme hereinafter. Each scheme was tested with 100 ensembles, an increase from the ensemble number employed in Chapter 2 (i.e., n = 50). The number of ensembles was increased to more fully account for system uncertainties. All simulations were started on January 1 of the first simulated site-year for each site to allow for model pools to reach an equilibrium prior to the first growing season.

Estimates of Δ and R were not re-initialized within the Miyoshi algorithm each year but were carried over from one site-year to the next. Following the key discussion points from Chapter 2 and poor performance in preliminary results, the PDA scheme was not explored in this chapter.

System evaluation

Ensemble weighting scheme

This study follows the same ensemble weighting strategy as presented in Chapter 2, such that daily ensemble weights were summarized to annual ensemble weights for each site. Though this approach does not leverage all available information within the model ensemble, effectively smoothing over information on model accuracy for daily forecasts, the application of annual weights was found to be the most robust for evaluating yearly estimates in this study (e.g., yield, cumulative NO_3 load, cumulative tile drainage). Furthermore, a more detailed analysis of the use of daily vs. annual weights found only negligible differences in daily weighted estimates for all state variables except soil moisture. These two points justified the use of the simpler annual ensemble weights for evaluating SDA performance. See the Appendix for more detailed information on how ensemble weighting strategies were evaluated.

Evaluation statistics

In addition to the evaluation metrics applied in Chapter 2 (i.e., RMSE, spectral norm, and weighted variance), a few new metrics will be employed in this chapter to evaluate and compare system performance. First, to help standardize accuracy measures across site-years, a normalized v RMSE will be calculated as

$$nRMSE\ (\%) = 100 * \frac{RMSE}{\bar{Y}}$$

where $\bar{Y}$ is the average observed value.

The coefficient of determination (R^2) will be used to more effectively compare model performance for each state variable across all observed time points. It was calculated as

$$R^2 = 1 - \frac{\sum_{t=1}^{T}(Y_t - \bar{X}_t)^2}{\sum_{t=1}^{T}(Y_t - \bar{X}_t)^2 + \sum_{t=1}^{T}(\bar{X}_t - \bar{Y})^2}$$

where Y_t is the observed value at the tth observed time step and $\bar{X_t}$ is the simulated weighted mean at the tth observed time step. All observations (n = T) from all site-years were included in this calculation. Separate R^2 values were computed for Free and SDA results, and weighted mean estimates for each observed time point for each scheme were computed using annual ensemble weights.

To characterize both accuracy and direction of bias for system forecasts, the relative error for annual model forecasts for the kth site-year was calculated as

$$RE_k\ (\%) = 100 * \sum_{i\,=\,1}^{N} w_i\ * \frac{(Y_k - X_{i,k})}{Y_k}$$

where w_i and $X_{i,k}$ are, respectively, the average weight and simulated value for the i^{th} ensemble of the k^{th} site-year, and Y_k is the observed annual value of the k_{th} site-year.

To identify and quantify relationships between variables, one of two correlation statistics were employed depending on the sample size of the data. When comparing data with a sufficiently large sample size (i.e., n > 30), the Pearson correlation coefficient (r) was calculated to determine the direction and strength of the linear relationship between two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\ (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^{\ 2}} * \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^{\ 2}}}$$

where x and y denote the two variables being compared, n is the number of complete samples, and a bar represents the mean value. When comparing data at the site-level (i.e., n ≤ 19), the Spearman rank-order correlation coefficient (r_s) was applied, which is a nonparametric measure of the strength and direction of the monotonic relationship between two variables. Though the sample size in this case is still too small for appropriate application, the Spearman coefficient was applied as its assumptions are less strict than the Pearson coefficient and some metric for comparison was needed. It is calculated as

$$r_s = 1 - \frac{6\ \sum_{i=1}^{n} d_i^2}{n\ (n^2 - 1)}$$

where the d_i is the distance between the two ranks of the i^{th} complete pair (i.e., x_i and y_i). For both coefficients, a test for association between paired samples was used to determine significance.

Finally, to allow for comparison between soil moisture time series across site-years, a time series decomposition approach was used to estimate the magnitude of the noise in each site-year time series. An additive model was assumed

$$Y_t = T_t + \varepsilon$$

$$\varepsilon \sim Norm(0, \Sigma)$$

where the trend (T) was estimated using cubic splines for each soil depth. Then, a diagonal variance matrix Σ was computed for the residuals. Sensor errors were assumed to be independent across layers following the assumption declared in Chapter 2. The magnitude of the noise was then calculated as the spectral norm of Σ (see Chapter 2), which will be denoted as $||\Sigma||_2$. Higher values of $||\Sigma||_2$ indicate "noisier" time series.

Analysis of innovations

An analysis of soil moisture innovations was completed to better understand how different system conditions impacted accuracy of daily SDA soil moisture estimates. First, the total soil water adjustment was calculated for each ensemble at each analysis time step by, first, calculating the change in soil water in each assimilation layer due to assimilation as the change in soil moisture content after the analysis (i.e., given as mm/mm) multiplied by the layer's depth (mm). Then, the change in soil water was summed across the two assimilation layers for each ensemble and, finally, the average total value was computed across all ensembles. To determine the most important predictors of daily soil moisture innovation, daily information was compiled to characterize soil-crop interactions, soil water dynamics, input weather, observed weather, and assimilation conditions for each analysis time step for all site-years, and a Random Forest model was fit to identify the most important covariates for predicting daily soil moisture adjustment.

Method of comparison

When quantifying system performance across site-years, a classification system was defined to distinguish between situations where SDA led to improved, degraded, or similar performance. The relative change in accuracy and precision metrics were computed as

$$\Delta\,(\%) = 100 * \frac{(C_S - C_F)}{C_F}$$

where C_F and C_S were computed evaluation metrics for the Free and SDA schemes, respectively, at a given site-year. The RMSE was used for classifying changes related to model accuracy, and weighted variance was used for classifying changes related to model precision. Performance was classified as follows

- If Δ < -5%, SDA improved performance.
- If Δ > 5%, SDA degraded performance.
- Otherwise, SDA performed similarly to the free model.

RESULTS

Soil moisture

In soil moisture forecasts for the two assimilation layers (i.e., SM3 and SM4), SDA performed as well or better than the free model in accuracy across all site-years. The median change in RMSE due to SDA was -17% and -28% for SM3 and SM4, respectively. Average forecast precision was also increased with SDA in 84% of cases and by 23% on average. The three site-years where precision was not increased in SDA include OH in 2013 and 2014 and MN in 2013. Interestingly, these site-years where precision was not improved were among those with the greatest improvement in forecast accuracy for the same state variables. This relationship is intuitive considering the nature of the Miyoshi algorithm, which systematically inflates model forecast uncertainty at time steps when observed and forecasted soil moisture distributions differ greatly. At the cost of reduced forecast precision, such inflation allows for the filter to pull the model forecast toward the observed distribution and improve accuracy in future predictions. Looking generally, SDA constrained precision well for the assimilation state variables, reducing variance for SM3 and SM4 estimates, on average, by 3% and 49%, respectively.

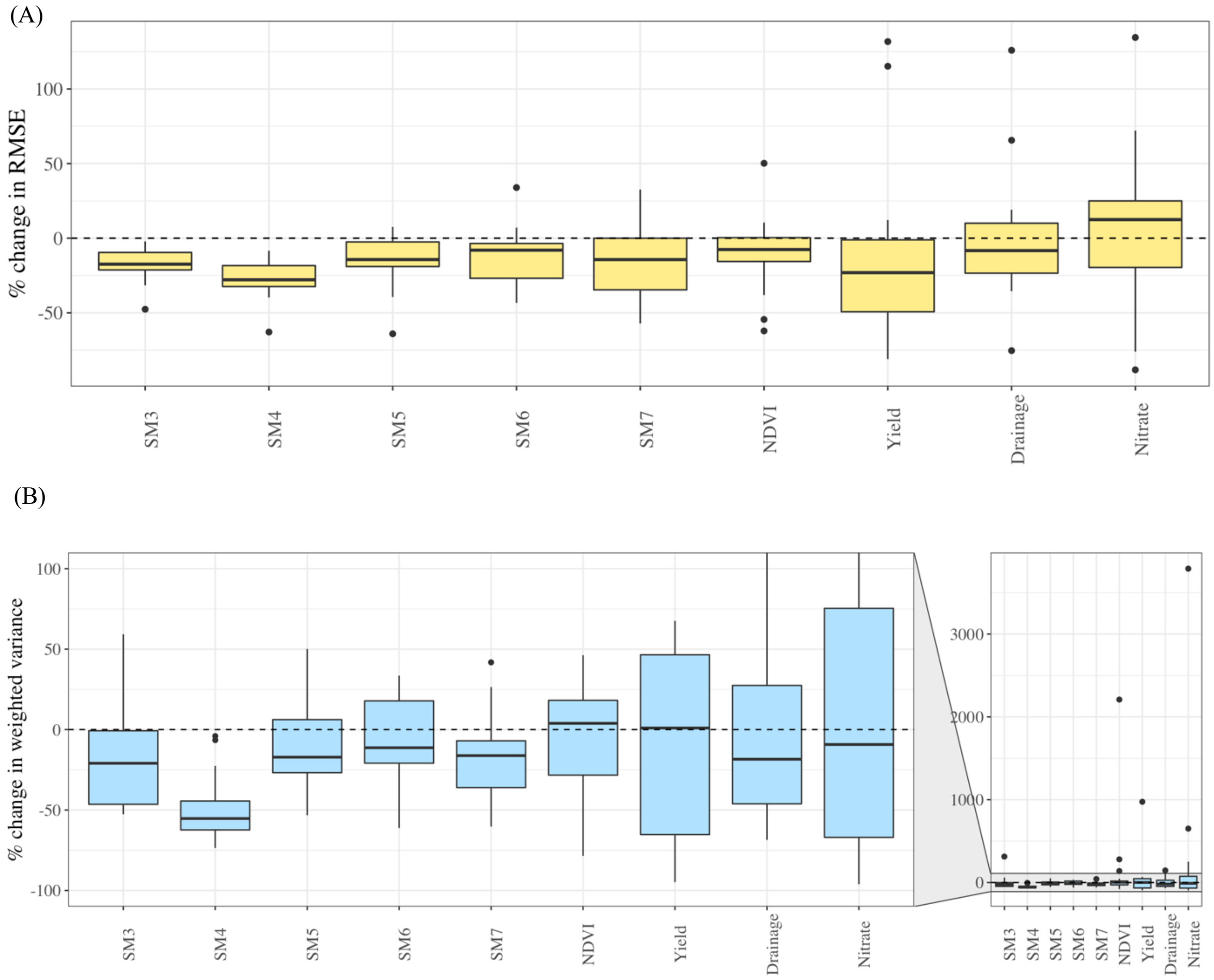

Figure 3.2. Boxplots demonstrating the distribution of relative change in (a) accuracy (RMSE) and (b) precision (weighted variance) due to SDA for each state variable across all site-years. Change is computed relative to the free model results. Negative values indicate improvement (e.g., $(RMSE_S - RMSE_F) / RMSE_F$).

Across all assimilation time steps, the model forecast tended to overpredict soil moisture within the two assimilation layers (Fig. 3.3), and, therefore, the adjustment in the analysis step typically reduced the total amount of water in the soil profile. An analysis of daily soil moisture innovations across all site-years highlighted strong positive autocorrelation in time, such that the strongest predictor of daily soil moisture adjustment (i.e., the increase or decrease in total soil water due to assimilation) was the adjustment value from the previous time step. Such a result points to consistent model biases in model soil water processes. Despite correcting these state variables at each time step in SDA, APSIM often reproduced a similar error in the next forecast.

Two other important predictors for daily soil water adjustment were daily precipitation and daily precipitation error, where daily precipitation is the average daily precipitation (mm) across weather ensembles for a given site-day and daily precipitation error (mm) is the difference in observed daily

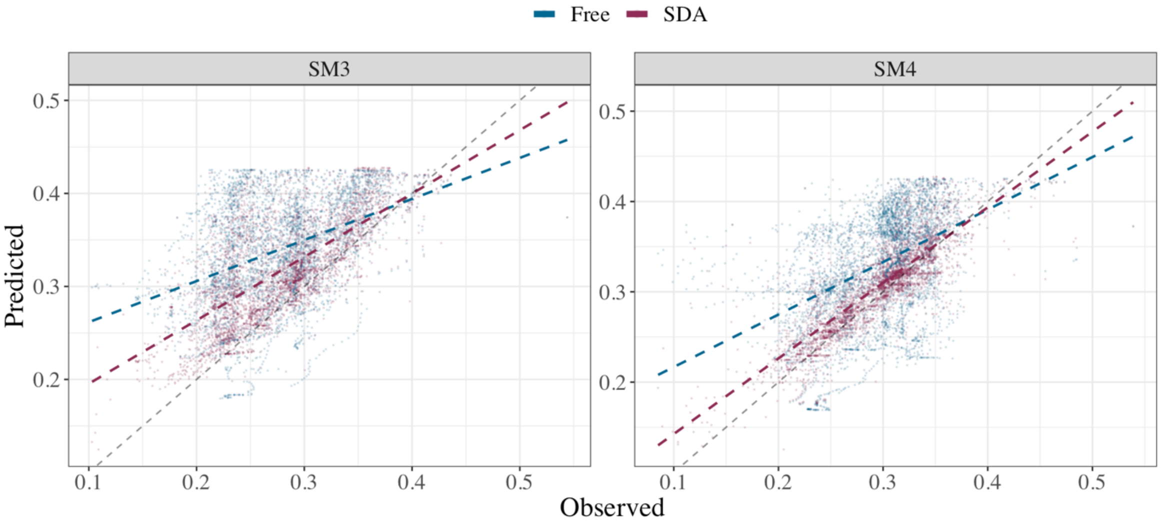

Figure 3.3. One-to-one plots for soil moisture estimates (mm/mm) in the two assimilation layers for the free model and SDA across all analysis time-steps and site-years. The least-squares regression line is shown for both schemes next to the black dashed line which demonstrates a perfect fit.

precipitation and the average daily precipitation in the model inputs. As demonstrated in Figure 3.4, with increasing input precipitation, the assimilation adjustment tended to remove more water from the two assimilation layers. Similarly, on days where input precipitation exceeded what was observed, the assimilation step typically reduced the amount of water in the soil profile to account for this error.

As seen in Chapter 2, SDA's constraint of SM3 and SM4 also led to the indirect constraint of soil moisture in deeper layers of the soil profile. Across all site-years with available data, the median change in RMSE for SDA estimates of SM5, SM6, and SM7 was -14%, -8%, and -14%, respectively. There were 1-2 site-years for each of these state variables where SDA functioned to increase RMSE, but most site-years saw improvement or, at the very least, similar performance compared to the free run (Fig. 3.2). In terms of precision, SDA had an overall positive impact on lower layer soil moisture estimates. The average change in weighted variance was -16%, -6%, and -20% for estimates of SM5, SM6, and SM7, respectively.

Yield and NDVI

Overall, soil moisture data assimilation improved yield estimates in this study. Compared to the free model, SDA predictions of yield explained 17.7% more variation in observed yield values (Table 3.3). It improved yield accuracy in 63% of site-years and performed comparably to the free model in 16% of site-years. Based on an analysis of site conditions, SDA was most effective at improving yield accuracy in site-years with a higher proportion of drought days during the growing season. This is an unintuitive result since SDA, overall, removed soil water from the profile over the course of each site-year. However, in those

Table 3.3. Summary statistics to quantify the impact of SDA on forecast accuracy of APSIM state variables. The "N" column indicates the number of site-years with available data for each state variable, and the "n" column indicates the total number of observations across site-years. (F) denotes a value computed for the free model estimates, and (S) denotes a value for the SDA estimates. The median change (Δ) in RMSE was computed as RMSE$_S$ - RMSE$_F$ / RMSE$_F$. Relative error (RE) is also given as a median value.

Variable	**N**	**n**	**RMSE (F)**	**RMSE (S)**	**Δ RMSE**	**R^2 (F)**	**R^2 (S)**	**RE (S)**
			median value					
SM3 *mm/mm*	19	12252	0.085	0.073	-17.4%	0.488	0.566	-19.9
SM4 *mm/mm*	19	12735	0.074	0.053	-27.9%	0.520	0.727	-12.4
SM5 *mm/mm*	17	11325	0.075	0.054	-14.3%	0.453	0.379	0.3
SM6 *mm/mm*	19	12846	0.065	0.068	-8.0%	0.424	0.341	9.2
SM7 *mm/mm*	9	5715	0.081	0.057	-14.3%	0.428	0.336	15.6
NDVI *unitless*	19	244	0.246	0.189	-7.6%	0.615	0.663	4.6
Yield *Mg/ha*	19	19	1.80	1.27	-23.1%	0.554	0.731	8.3
Annual drainage *mm*	19	19	151	145	-8.3%	0.472	0.457	6.2
Annual NO_3 load *Kg NO_3-N/ha*	19	19	36.2	21.6	+12.5%	0.416	0.449	50.8

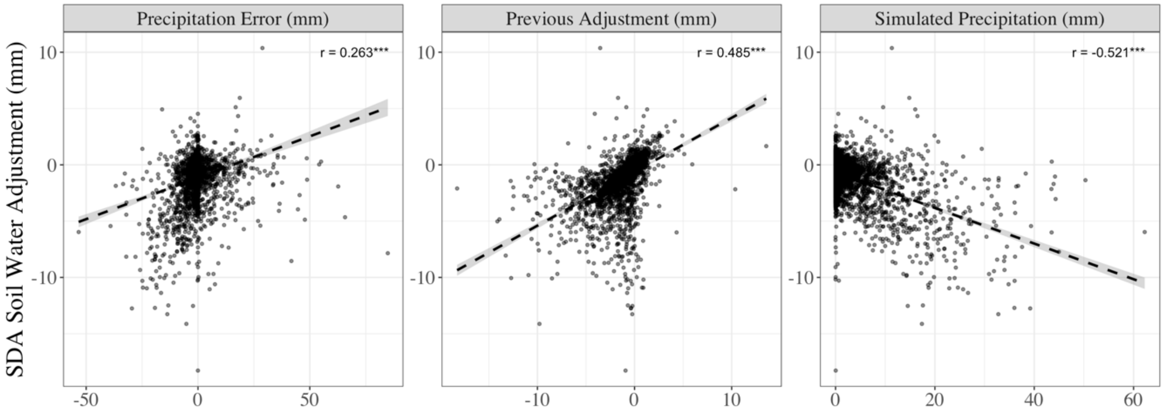

Figure 3.4. Scatterplots demonstrating the relationship between daily soil water adjustment and its three most important predictors. Dashed black lines demonstrate the least squares regression line for each relationship, and 3 asterisks (***) indicate a significant linear relationship at all significance levels (α ≅ 0). Together, these three predictors explain 47.6% of the variation in soil moisture innovations.

cases where yield estimates were improved, SDA often increased available soil water at critical points in crop development, leading to reduced crop soil water deficit factors and increased yield compared to the free model (Fig. 3.6, Fig A.4). The most evident example of SDA yield improvement is IN in 2012. Here, the free model estimated complete maize crop failure (i.e., no grain yield) due to leaf senescence in mid-July, but SDA estimated a harvestable crop. By constraining soil moisture, SDA increased available soil water in the soil profile in the weeks prior to the estimated crop failure. This increased soil moisture can

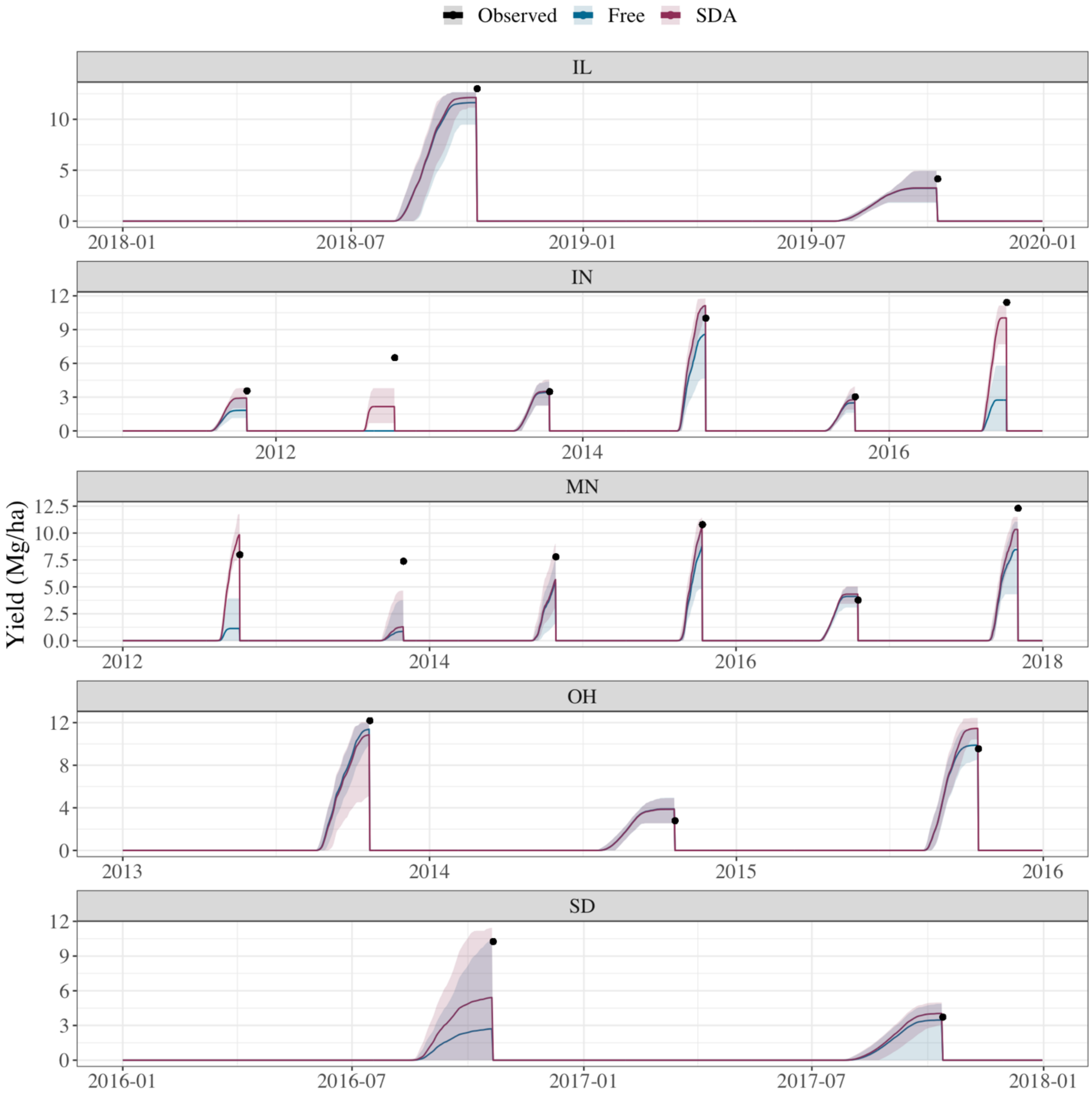

Figure 3.5. Time series of yield estimates from the two schemes with the mean daily estimates demonstrated with line graphs and the 95% credibility interval demonstrated by the shaded regions. Black points represent the observed harvest date and yield for each site-year.

explain the reduced soil water deficit factors, increased annual crop water uptake by 85 mm, and the survival of the maize crop into grain production demonstrated in SDA for that site-year. Looking instead to poor performance, the 4 site-years where SDA reduced yield accuracy include OH in 2013 and 2015, where RMSE increased by 131% and 115%, respectively. These two site-years, which comprise both maize growing seasons at OH, contradict the pattern noted above such that reduced overall crop water stress with SDA did not improve yield forecasts.

Generally, the free model was able to capture the phenological development of the cropping systems simulated in this study, as demonstrated by the good agreement between observed and simulated NDVI (Fig. A.5). SDA's impact on NDVI accuracy was similar to its impact on yield accuracy, such that it typically either increased accuracy due to lessened water stress or did not greatly affect model performance. A comparison of R^2 values demonstrates that SDA helped to explain 4.8% more variation in observed NDVI values compared to the free model. Intuitively, the site-years with the greatest jumps in NDVI accuracy also usually saw great improvement in yield accuracy, highlighting a generally well-defined physiological relationship between vegetation and grain yield in APSIM's simulation of maize and soybean development. The two site-years where SDA reduced NDVI accuracy were MN in 2015 (+50% RMSE) and IN in 2014 (+10% RMSE).

The impact of SDA on precision in both yield and NDVI estimates was largely mixed. Roughly 47% and 53% of site-years saw reduced precision (i.e., higher weighted variance) in estimates of NDVI and yield, respectively. On average, precision was reduced by 126% for NDVI estimates and by 40% for yield estimates. However, these average effects were highly skewed by two large outliers where precision was decreased by more than 900%. After removing the outlying site-year for both state variables SDA's average impact was more moderate with a 10% reduction in NDVI precision and a 15% *increase* in yield precision. The two site-years with the greatest reduction in NDVI precision (i.e., reduction by 2209% and 280%) correspond with the site-years where yield accuracy was also dramatically reduced with SDA (i.e., OH 2013 and 2015). Intuitively, reduced constraint in NDVI seems to have led to reduced yield performance.

Tile drainage

Across the 19 site-years simulated in this study, the free model and SDA showed overall poor performance in estimating annual drainage with nRMSE values ranging from 18-215% with a median value of 54.3% with SDA and from 20-250% in the free model with a median value of 52.4%. In the site-years with the lowest accuracy, APSIM often overpredicted drainage in both the free model and SDA. However, these cases of large overestimation in drainage were also among those site-years that were most improved with SDA. Since SDA resulted in a net loss of soil water for all site-years in this study, it is unsurprising

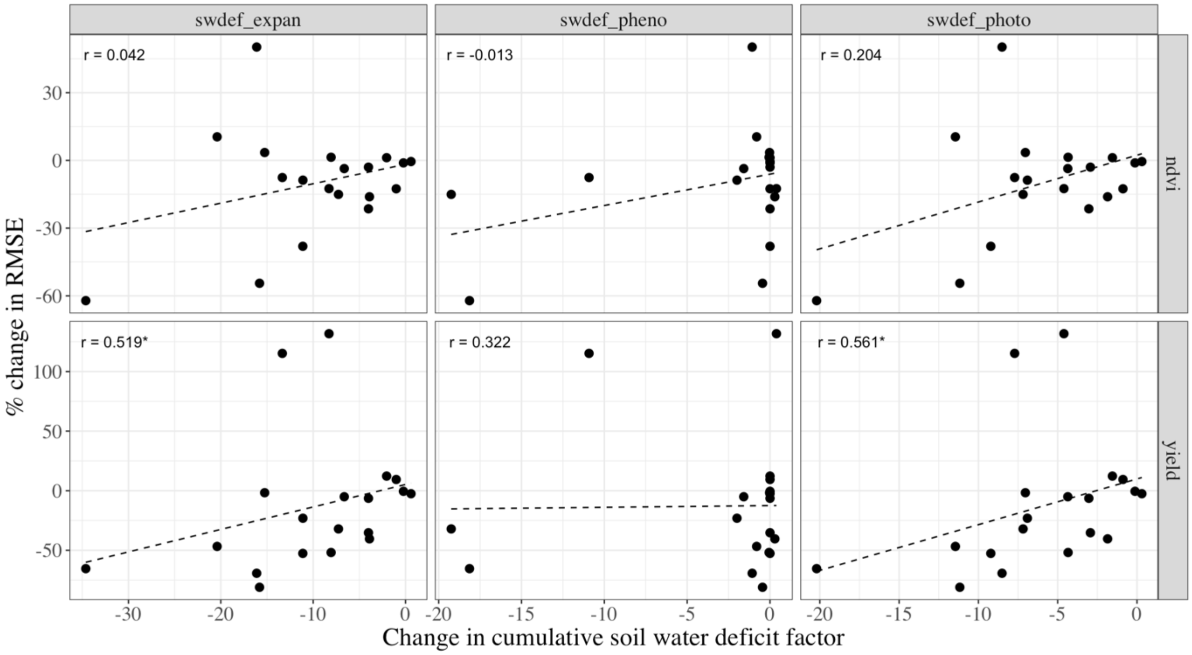

Figure 3.6. Scatterplots comparing changes in cumulative soil water deficits due to SDA with relative changes in yield and NDVI RMSE values across all site-years. Least squares regression lines are given by a black dashed line and the Spearman rank-order correlation coefficient is included for each scatterplot. Single asterisks (*) denote significant monotonic relationships with p-values < 0.05. Change on both axes is defined as the SDA estimate less the free model estimate. Change in cumulative soil water deficit is given as the difference of the mean annual sum of the 3 deficit factors across ensembles, such that a negative value on the x-axis indicates reduced crop water stress.

that 72% of the site-years where SDA improved estimates of annual drainage were cases where the free model overestimated tile flow (Fig. 3.7). In these cases, SDA functioned to remove available water from the soil profile and correctly lower the amount of water lost from the system.

SDA improved accuracy in 58% of drainage estimates compared to the free model; these improved site-years were either low-yield maize years or soybean years. The other 42% of site-years where SDA reduced accuracy in drainage estimates were typically maize years, aside from the soybean year at MN. Beyond crop type, two other variables help explain SDA performance in tile drainage constraint: (1) increased crop water uptake and (2) drought conditions. Cases where SDA reduced accuracy typically had higher uptake increases and lower drought intensity (Fig. 3.8). Conversely, site-years where SDA improved drainage accuracy were often drier, especially when crop water uptake was increased by the system.

SDA's impact on precision for annual drainage estimates was also highly variable. Although 63% of site-years saw improvements in precision, SDA's effect on precision across site-years was highly skewed to the left. Thus, on average, SDA decreased precision by 3.4%, but the median effect was an 18% increase. At the site-year level, the change in weighted variance for annual drainage estimates was strongly and positively correlated with the change in weighted variance for SM7 estimates, the state variable that describes the soil moisture content in the lowest soil layer (Fig. 3.9). This finding indicates that SM7 constraint propagated as a top-down constraint for tile drainage in the presented system.

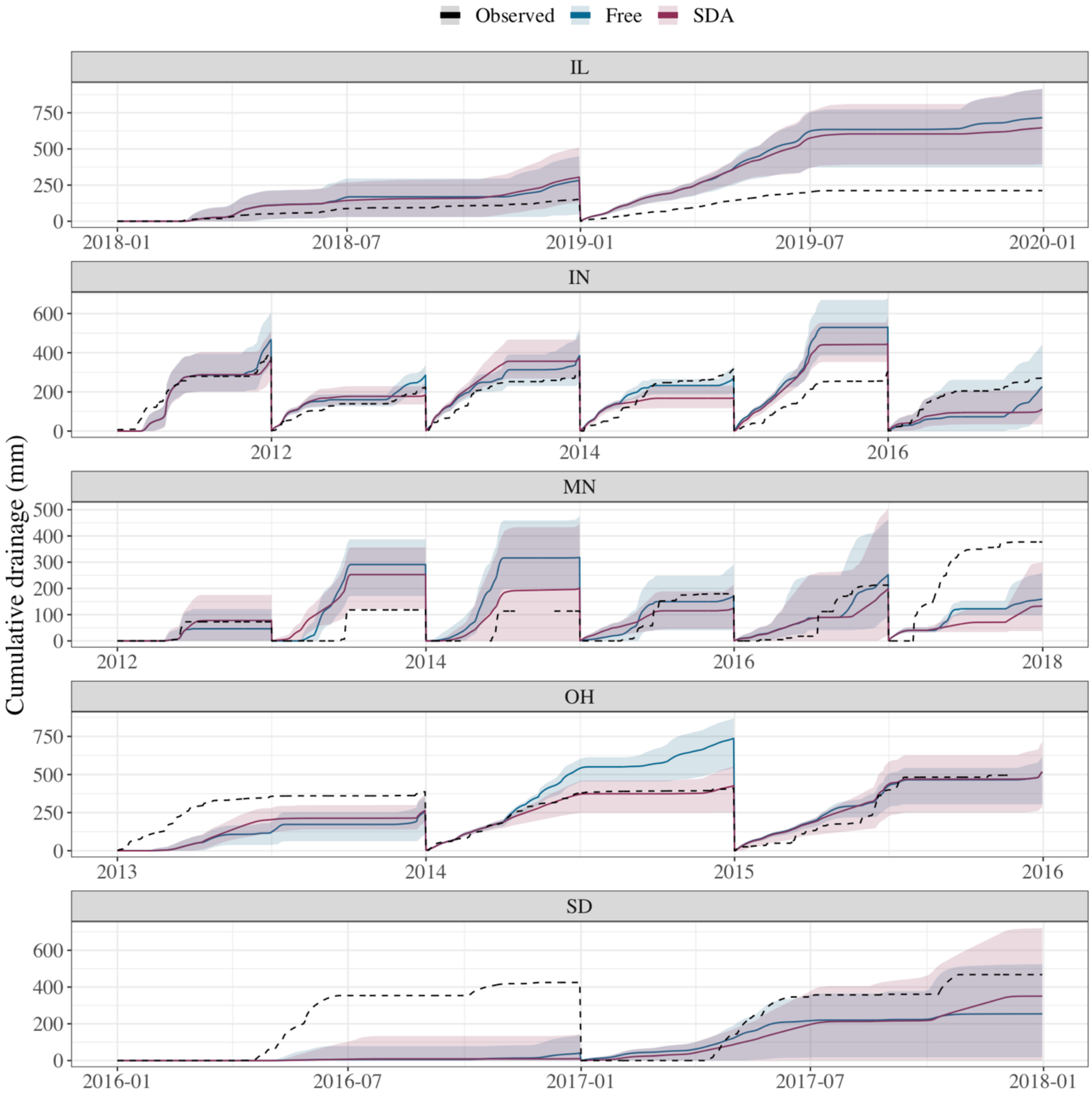

Figure 3.7. Time series of cumulative drainage estimates from the two schemes for each site-year with the mean daily estimates demonstrated with line graphs and the 95% credibility interval demonstrated by the shaded regions. Black dashed lines represent the observed cumulative drainage for each site-year.

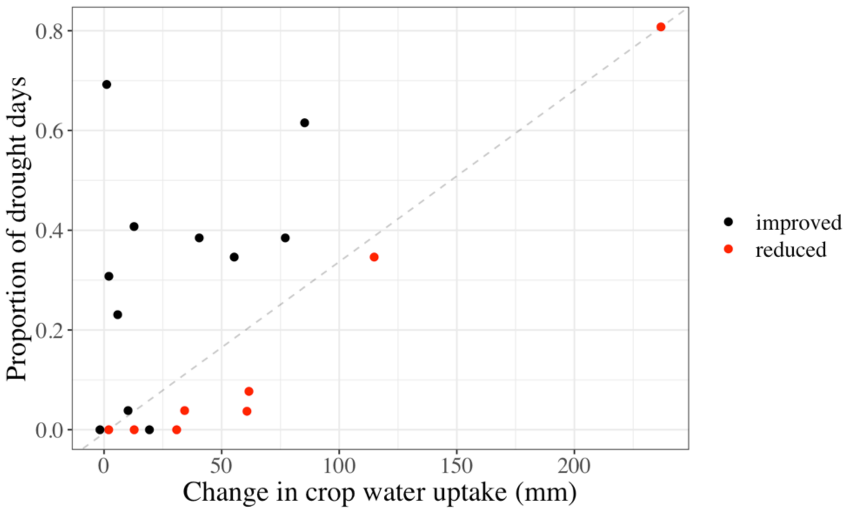

Figure 3.8. Scatterplot comparing the change in estimated mean crop water uptake between SDA and the free model with the proportion of drought days for each site-year. Color indicates the impact of SDA on annual drainage accuracy. A dashed black line is included to demonstrate an effective separation of the two classes across these two dimensions.

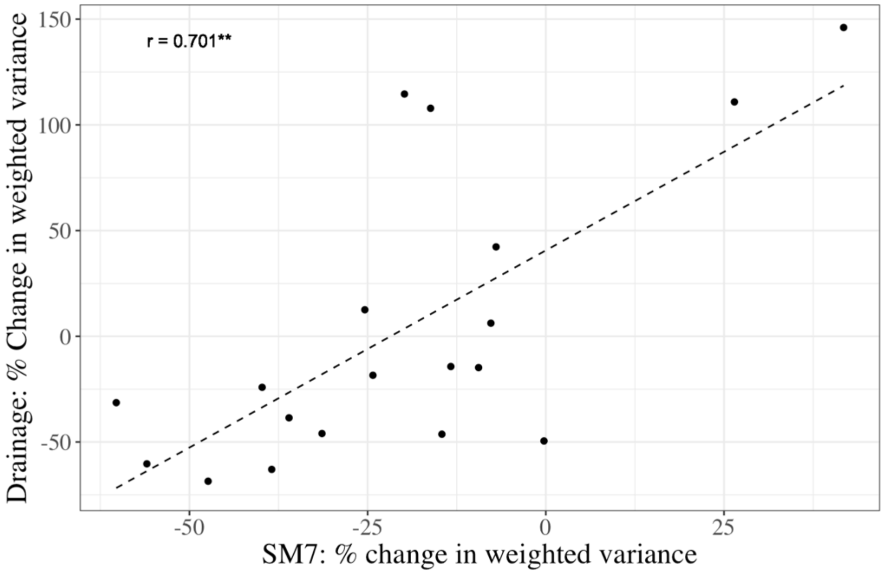

Figure 3.9. Scatterplot comparing the relative change in weighted variance for SM7 with that of annual drainage. The least squares regression line is shown with a blue dashed line, and the Spearman rank-order correlation coefficient is shown in text. The correlation coefficient is significant with a p-value < 0.01.

Nitrate load

Next to annual drainage, APSIM also struggled to accurately estimate annual NO_3 load for the tested site-years in this study (Fig. A.6). For the free model, nRMSE values ranged from 23-681% with a median value of 83.7% and, for SDA, nRMSE values ranged from 17-833% with a median value of 86.9%. At the site-year level, the free model commonly struggled when site-years had a low annual NO_3 load to annual drainage ratio. These were typically site-years at IL or MN (Fig. 3.10).

Among state variables considered in this analysis, estimates of annual NO_3 load were the most poorly constrained by SDA in terms of accuracy and precision. SDA's impact on precision was split, increasing precision in 53% of site-years and reducing precision in 42% of site-years. Moreover, accuracy was improved in 32% and reduced in 58% of SDA estimates. Among those 6 site-years where SDA

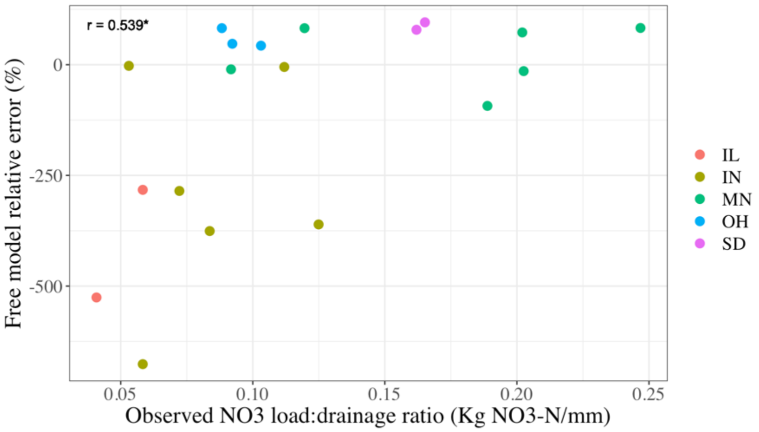

Figure 3.10. Scatterplot demonstrating the relationship between free model relative error in annual NO_3 load estimates and the observed ratio of annual NO_3 load to drainage for each site-year. Colors indicate the site location for each point. The Spearman correlation coefficient is shown in text and is significant with a p-value < 0.05.

improved NO_3 load accuracy, SDA typically reduced estimates compared to the free model. These sites were often maize years characterized by high amounts of input winter precipitation (Jan-Apr). On the other hand, no clear environmental nor agronomic trend was identified among those 11 site-years where SDA reduced accuracy.

Soil moisture time series quality

Across the 19 site-years, the distribution of $||\Sigma||_2$ was highly right-skewed. The greatest outlying points included MN in 2014 and 2017, which both had $||\Sigma||_2$ values that are at least 2x as great as the next noisiest site-year (Fig 3.11). The 3 OH soil moisture time series were the next noisiest. In comparing values of $||\Sigma||_2$ with system performance, the expectation was that noisier sensor observations would demonstrate poorer performance. At the least, performance in SM3 and SM4 estimates was expected to be strongly related to time series quality. However, a simple correlation analysis of $||\Sigma||_2$ against system performance for all state variables did not reflect these expectations. The analysis only indicated significant monotonic relationships with $||\Sigma||_2$ for NDVI accuracy ($r_s = 0.628$, $p = 0.005$) and SM7 accuracy ($r_s = -0.8$, $p = 0.014$). With increasing noise in the soil moisture time series, SDA reduced NDVI accuracy and increased SM7 accuracy. There were no significant relationships between noise and state variable precision.

DISCUSSION

This study highlights how SDA's impact on downstream model estimates depends on each state variable's sensitivity to the assimilated state variable (i.e., soil moisture). Lower layer soil moisture estimates—the most sensitive state variables evaluated—were the most strongly constrained. Figure A.8 demonstrates the significant linear relationship between daily changes in forecasted SM3 and SM4 due to SDA and daily changes in soil moisture estimates for all deeper soil layers. However, as expected with a cascading water balance model, the strength of the linear relationship weakens as the vertical distance

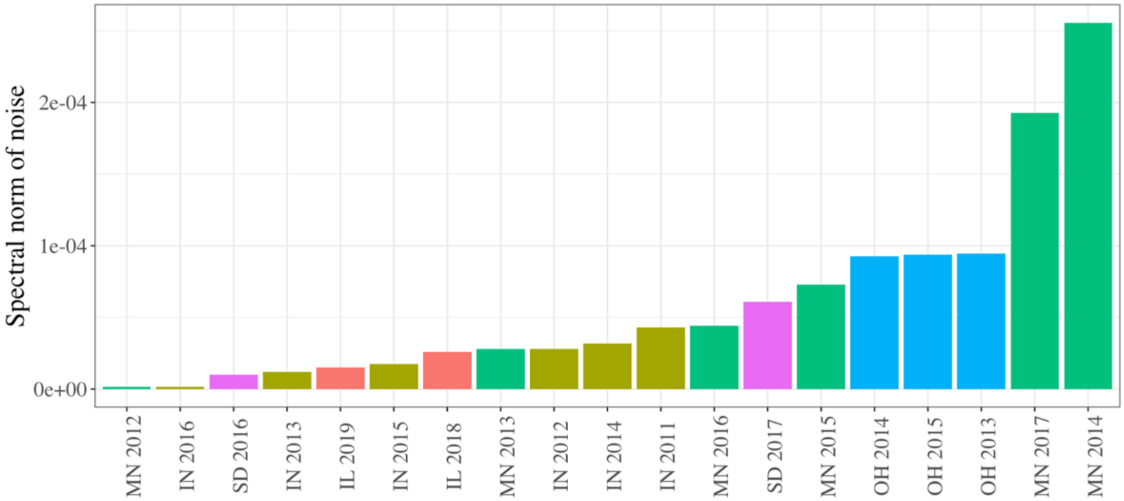

Figure 3.11. Column chart demonstrating the computed spectral norm of the noise variance matrix for each site-year soil moisture time series. The x-axis labels, as well as the fill color, indicate site location.

between soil layers increases. In the model, soil moisture in each layer can influence soil moisture estimates of deeper soil layers, but only indirectly through its influence on the soil moisture in the layer immediately below it. Therefore, the influence of the assimilation layers is reduced by each subsequent soil moisture process down through the soil profile and is weakest in the final soil layer (SM7). Yet, the constraint of SM7 was still quite strong in this study. By assimilating soil moisture for two upper soil layers into the APSIM model, the accuracy of soil moisture estimates improved immensely by simply leveraging the pre-existing model structure (compare to Liu et al., 2017).

Crop yield showed the next strongest constraint in this study. However, as noted in Chapter 2, its sensitivity to soil moisture and its constraint with soil moisture SDA were conditional (Lu et al., 2021). While changes in soil moisture influenced lower layer soil moisture at all analysis time steps, crop yield is only influenced when the changes affect crop water stress. Daily crop water uptake is determined in APSIM as the minimum of crop water demand and soil water supply. Therefore, SDA could only influence crop yield when the soil water adjustment pushes the soil water supply to be above or below the demand threshold. Any other type of change in soil moisture had no effect on daily crop development. There do exist other pathways that soil moisture can impact crop yield in APSIM, like soil N cycling, but these processes did not play a strong role in this study.

The impact of soil moisture SDA on APSIM drainage estimates can also be beneficial given certain conditions. As shown in the results, drainage was affected by SM3 and SM4 through 2 pathways: (1) changes in total soil water with assimilation adjustment and (2) changes in crop water uptake due to changes in crop water stress. The role of each of these pathways varied over the course of the year, such that the presence of a growing crop and root system weakened the sensitivity of drainage estimates to changes in the assimilation layers. To quantify this change in sensitivity, daily model forecasts were divided into two categories—with crop water uptake (June-Sept) and without crop water uptake—and the relationship between changes in SM3 and SM4 and changes in drainage was analyzed separately for each group. There was not a significant linear relationship when looking at SM3 changes in either case. However, the linear relationship between changes in SM4 and changes in daily drainage was stronger when no crop was present ($r = 0.23$, $p = 0.00$) than when a crop was present ($r = 0.14$, $p = 0.00$). This is similar to the findings of Hu et al. (2008), who identified notable changes in drainage dynamics during rapid crop growth compared to out-of-season dynamics in SPWS model simulations. In this study, annual drainage estimates tended to improve in dry years where crop yield was still underpredicted.

Among the state variables considered in this analysis, NO_3 leaching has the weakest and most complex relationship with SM3 and SM4 in APSIM. Therefore, it follows logically that the presented system performed most poorly in its constraint of annual NO_3 leaching estimates. In APSIM, daily NO_3 leaching estimates are computed as the product of two different daily values: estimated NO_3 concentration in the lowest soil layer and estimated tile drainage. Therefore, in addition to its impact on drainage, SDA can affect NO_3 load estimates through (1) changes in N cycle processes via soil moisture rate factors (Fig. 2.3) and (2) changes in the vertical movement of soil water (and N solutes) through the soil profile. The convoluted nature of this system, as well as limited observations tracking soil N and soil water cycle components, made identifying a trend in NO_3 leaching constraint difficult across the 19 site-years in this study. The prediction error for each site-year seemed to originate from its own unique pathway. In a validation study of APSIM N processes, Sharp et al. (2011) also observed inconsistent model behavior in annual leaching estimates for their experimental site in New Zealand, when simulating 3 years of a potato-rye rotation. Their final calibration of the model process only improved one of the annual estimates but did not constrain estimates in the other two years.

Several past studies have identified nitrate leaching estimates as a great forecasting challenge (Stewart et al., 2006; Sharp et al., 2011; van der Laan et al., 2014; Brilli et al., 2017). This study highlighted a few ways in which leaching estimates could be improved in APSIM by accounting for misrepresented or missing processes. First, the APSIM model currently does not account for snow, freezing soils, nor spring snowmelt in its soil water processes. Within numerous site-years, there were large overpredictions of drainage in the early winter months (i.e., Jan-Feb) and large underpredictions in the spring (i.e., Mar-May). Such behavior could point to missing snow-related processes in APSIM, an issue also highlighted by Ojeda et al. (2018) who used the model to simulate continuous-corn and corn-soybean rotations at 3 Indiana locations and saw great underpredictions of early spring drainage. Moreover, APSIM also does not currently account for the effect of tillage nor residue cover on water infiltration and movement in the soil profile (Malone et al., 2007; Brilli et al., 2017; Ojeda et al., 2018). Therefore, the model cannot distinguish between certain management practices across different locations when predicting infiltration. This can help explain the strong linear relationship between soil moisture innovations and simulated precipitation, such that the model overpredicted soil water inputs on high-precipitation days (Fig. 3.4). Accounting for management differences could help clarify the muddled results in this study's nitrate leaching errors. Though these model changes would be nontrivial, incorporating them could greatly benefit APSIM's understanding of the soil water and soil N cycles and improve future estimates of soil water and soil N fluxes.

In this study, the inconsistent impact of soil moisture SDA on estimates of annual drainage further emphasized the important caveats in the presented system first discussed in Chapter 2. Across all site-years in this study, SDA led to a net reduction in total soil water. However, instead of redistributing this water to other model components, the workflow simply eliminates this water from the system. Though this elimination can correct precipitation error in the inputs as previously discussed, it can also serve to violate the system's water balance, creating water deficits in other model estimates. Nonetheless, there are cases where this approach still successfully constrained drainage despite a possible violation. As shown in Figure 3.8, soil moisture SDA was more effective in improving drainage estimates in site-years with more frequent drought conditions and with smaller increases in crop water uptake. In these cases, it could be speculated that the water eliminated from the system via assimilation was "supposed" to be lost from the system via crop water uptake or evaporation. Therefore, its removal was beneficial for the sake of improving drainage estimates but costly for estimates of yield (which were largely underestimated in this study) or other soil water cycle components. Such a hypothesis reflects the findings of past studies where yield inaccuracies were partially attributed to inaccurate drainage predictions (Malone et al., 2007; van der Laan et al., 2014). However, this possible explanation is purely speculation in the context of this study and essentially useless for informing future system applications. The only way to improve constraint of annual drainage with soil moisture SDA is to confront model predictions with observations spanning more (or all) components of the soil water cycle. Considering the hypothesis above, observations on transpiration and evaporation could be critical for identifying biases and advancing system performance (Stewart et al., 2006; Lu et al., 2021). A potential future direction could be to leverage remote sensing data products, such as the ECOSTRESS evapotranspiration product, to fill this gap (Fisher et al., 2020). It would help constrain two important water loss pathways (i.e., evaporation and transpiration) in agricultural systems and, thus, improve the model's estimate of how much water is moving through the soil profile.

Due to the weak relationships between SM3 and SM4 and the other APSIM state variables evaluated in this study, the potential for perfect constraint of all or any state variables with soil moisture SDA was low. To improve the constraint of other state variables, other types of observations with more direct influence should be considered for assimilation. For example, observations of plant and soil N could help to constrain and correct N cycle processes, leading to better estimates of nitrate leaching (van der Laan et al., 2014). Such an approach could also lead to more accurate parameterization of crop N uptake, an APSIM process which has been recognized as inaccurate in the past (Malone et al., 2007). Additionally, the results of this study suggest the assimilation of daily drainage could improve estimates of nitrate leaching. Daily innovations for drainage explained about 21.5% of the variation in daily innovations of nitrate leaching in a linear regression analysis. Malone et al. (2007) identified a similar relationship between errors, attributing a large portion of error in predicting nitrate leaching to poorly estimated drainage. Furthermore, the assimilation of drainage observations could also serve to inform and nudge soil moisture estimates by leveraging their modeled covariance through the PDA configuration.

The analysis of soil moisture innovations revealed two important takeaways of this study. First, it highlighted an important benefit of soil moisture SDA: a systematic way to offset input error (Ines et al., 2013). Gridded weather datasets are necessary for making agricultural forecasts at broad scales but are not able to accurately represent local weather patterns, which are critical to understanding crop development. In particular, daily precipitation can be highly variable at small spatial scales yet is an important factor for estimating daily crop growth (de Wit and van Diepen, 2007; Thaler et al., 2018). By using a coarse resolution gridded weather dataset to estimate daily precipitation for the sites in this study, additional biases might be introduced into each day's forecast and propagated through the model (Fig. A.9). However, by assimilating observed soil moisture into the model at daily time steps, the adjustment step partially accounted for this input error, increasing total soil water when precipitation input was too low and decreasing total soil water when precipitation input was too high (Fig. 3.4). This is a strong advantage of SDA when observed weather information is not available for a location of interest.

The second important point of the soil moisture innovation analysis concerns the strong autocorrelation in soil moisture innovations. By assimilating observed soil moisture into the model, the system integrated all available information to optimize and set up the initial conditions for the next forecast. Yet, each day's soil water adjustment was highly correlated with the previous day's adjustment. This result points to biases in the model process that persist despite improved initial conditions. A study by de Wit and van Diepen (2007) highlighted similar temporal correlation in soil moisture innovations when assimilating maize fields and attributed the persistent errors to poor cultivar parameterization and inaccurate initial conditions. However, in this study, it is difficult to pinpoint the origin of these errors without sufficient observations. Daily forecasts of soil moisture depend on initial soil water, daily precipitation inputs, daily estimates of soil water outputs (e.g., evaporation, transpiration, drainage), and the previous day's soil moisture estimate, so any combination of those processes could have caused error in soil moisture predictions. Nonetheless, as other studies have observed similar soil moisture overpredictions in APSIM

(Sharp et al., 2011; Archontoulis et al., 2014), it seems likely that APSIM's soil water processes could be improved.

Considering the inherent variability of soil water dynamics, a relatively simply approach for decomposing the soil moisture time series was employed in this study (Basak et al., 2017). However, it provided a necessary measure of assimilation data quality and, thus, allowed for more appropriate comparison of the system applications presented in this study. Estimates of observation uncertainty (R) from the Miyoshi algorithm also helped validate this approach. Across all site-years, average spectral norm of observation uncertainty (R) was significantly and positively associated ($r_s = 0.707$, $p = 0.001$) with $||\Sigma||_2$ indicating that the applied decomposition method and Miyoshi understood observation uncertainty similarly for each site-year. Yet, contrary to expectations, differences in soil moisture data quality did not have a strong impact on SDA performance at the study sites. This, however, could point to the strength of the data-assimilation system itself. In the case of accuracy, the presented system was designed to adaptively estimate observation uncertainty based on filter performance using the Miyoshi algorithm. These estimates were likely able to account for sensor variability, reducing the influence of the observed distribution in the EnKF and suppressing the weight of noisy estimates when appropriate. In the case of precision, the impact of sensor noise may have been subdued on account of the EnKF. The EnKF computes an analysis distribution with lower uncertainty than either the observed or forecast distribution alone. If estimated observation uncertainty (or sensor noise) was high, a more confident forecast distribution would have helped to reduce the uncertainty propagated forward into the adjusted system state, diminishing the impact of sensor noise. Considering these 2 mechanisms, the presented EnKF-Miyoshi workflow could be an invaluable approach for assimilating datasets with low replication and poorly estimated uncertainties.

Considering past crop modeling works, this study is the first to apply data-assimilation methods to improve estimates of soil moisture, nitrate leaching, tile drainage, and yield at multiple locations across several years. However, other studies have employed calibrated crop models to achieve similar goals with, often, greater reported success than reported here. For example, Malone et al. (2007) used APSIM to simulate 12 site-years at an Iowa site and achieved RMSE values of 0.2525 Mg/ha and 0.372 Mg/ha for maize and soybean yields, respectively. This study reported a median RMSE value of 1.27 Mg/ha for both crops with SDA. Malone et al. also achieved lower RMSE values for drainage (15.5 vs. 145 mm) and nitrate leaching (6.0 vs. 21.6 Kg NO_3-N/ha) when compared to this study's median values. Martinez-Feria et al. (2019) used the APSIM model to simulate 7 long-term experimental sites across the U.S. Midwest, including 2 TD sites presented in this study (i.e., IN and MN). They reported nRMSE values for annual nitrate load and crop yield as 40.2% and ~13%, respectively, which also outperformed the median nRMSE values of 52.4% and 20.7% achieved here.

Yet, there is an important distinction to highlight between these past works and the results presented in this chapter. Though past modeling techniques were successful in estimating specific state variables for a few site-years, the model applied in those works could not be reasonably applied more broadly in space without great losses in accuracy due to the way in which they were carefully calibrated (Wallach, 2011; van der Laan et al., 2014; Seidel et al., 2018). For example, Puntel et al. (2016), who used a multistep calibration approach for their study, state that the application of their calibrated APSIM model can be justified for years outside of their time range but not for other locations. Their parameterization is limited to a certain genetic, environmental, and management inference space. Archontoulis et al. (2014), who also calibrated the APSIM model using 5 independent datasets for sites near Ames, Iowa, also discuss limitations in the application of their intensively calibrated model. They note that the model may require additional calibration for cropping systems with rotations or that fall outside of the evaluated climate-soil space. Li et al. (2014) predicted drainage satisfactorily for 5 sites in northern China with a calibrated DNDC model but expressed concern about broader application of the model due to soil heterogeneity.

Intensive calibration can be practical if only a single location is to be simulated. However, as demonstrated in this study, the immense variability in agricultural systems across time, space, and management factors cannot be well-represented by one single model parameterization. Consider, for example, the error in nitrate leaching estimates across the 19 tested site-years. Despite an intensive investigation, the error patterns were unpredictable in terms of timing, magnitude, and direction. No clear trend was found. This demonstrates the need for site-level calibration for accurate nitrate estimates since errors were so variable across site-years. However, the resources and time needed to apply the necessary calibration measures for 19 site-years, let alone an entire region, would be immense (Seidel et al., 2018). Intensive site-level calibration would also provide little to no benefit in terms of systematic model improvement and would not lessen calibration efforts for future applications. Instead, model calibration can lead to an overfit model and/or a fortuitous cancellation of errors within calibrated model processes. Such a model will not be reliable for extrapolated applications (van der Laan et al., 2014).

On the other hand, this work, which aims to develop and test a forecasting system that can be confidently applied at regional scales, holds generalizability as a central goal. All presented simulations were performed by the same version of the APSIM model, and little to no changes were made to the original model parameterization prior to application across the study sites. At the site-level, there were only two major changes applied: (1) the adjustment of the soil profile to account for artificial tile drainage and (2) the selection of soybean maturity group range. Apart from these changes, all other differences in model configuration between sites and site-years were applied systematically and with propagated uncertainty within the system. Thus, like the work of Guerif and Duke (2000), this study investigates how uncertainty propagation and data assimilation can help account for the site-level nuisances which traditional model calibration techniques target. As this system continues to be iteratively enhanced and generalized, the end goal is to be able to apply it uniformly and effectively across a gridded region. The reported differences in accuracy measures can be attributed to this difference in philosophy and serve as a baseline for further improvement.

Yet, for the sake of generalizability, this study has one great shortcoming: the availability of soil sensor soil moisture data. As they are expensive and difficult to collect, sensor-based soil moisture data would be an unrealistic data source for regional forecasting purposes. Alternative types of soil moisture data would need to be tested in the presented system if it were to be applied at regional scales. One possible solution could be soil moisture estimates derived from remote sensing (RS) imagery. RS soil moisture data products are readily available and have the strong advantage of broad and consistent spatiotemporal coverage. However, these data will likely introduce new caveats into the system, as they characterize soil moisture with greater (and poorly characterized) uncertainty and at different scales than soil sensors (Huang et al., 2019; Peng et al., 2021). Therefore, testing the utility of this workflow with RS-based soil moisture data products will be imperative prior to broader applications of the system. The current system with soil sensor soil moisture shows strong constraint of soil moisture and crop yields. Can the system still be effective in these constraints with RS soil moisture? The analysis presented in Chapter 4 investigates this possibility.

CHAPTER 4

SYSTEM APPLICATIONS AT SCALE: EXPLORING THE OPPORTUNITIES AND CHALLENGES OF ASSIMILATING REMOTELY SENSED SURFACE SOIL MOISTURE DATA ASSIMILATION

INTRODUCTION

Many of today's most pressing agricultural issues operate at scales that cannot be easily characterized by field experiment data alone. For example, there is great concern about the increasing incidence and intensity of droughts due to climate change and the impact that such natural disasters could have on agricultural productivity and stability in the U.S. Midwest (Bolten et al., 2010). However, to reliably investigate such questions at the regional scale, tools are needed that can consistently simulate agroecosystems across broad heterogenous landscapes with accuracy and precision. Such a tool must account for the spatiotemporal uncertainties in important agricultural variables—such as soil moisture, weather, soil properties, and management—at spatial scales that are relevant to the original research question. This will be especially important in the coming decades as social and environmental pressures continue to drive drastic changes in the spatial configuration and management of agroecosystems (Weiss et al, 2020). As demonstrated in Chapters 2 and 3, the soil moisture data-assimilation system presented in this work is scalable and capable of efficiently and systematically incorporating uncertainty across relevant dimensions when data are available to improve simulations. However, the data products previously employed in the system to constrain spatiotemporal variability in soil moisture (i.e., *in situ* soil moisture) are not available at the required spatial scales. To overcome this limitation, remote sensing (RS) soil moisture data products, which can effectively capture spatiotemporal variability in soil moisture dynamics across regions, could be invaluable for agricultural forecasting efforts, especially when employed in data-assimilation systems (Dorigo et al., 2007; Huang et al., 2019; Weiss et al., 2020). Over the past few decades, numerous studies have successfully assimilated RS soil moisture into process-based models to incorporate spatial heterogeneity within soil water processes and, thereby, improve predictions. However, due to the nuances and required resources for many such studies, there is still work to be done to make the assimilation of RS soil moisture more practical, effective, and generalizable for the purposes of broader predictions.

For example, an assimilation study by Liu et al. (2021) assimilated RS-based LAI and soil moisture estimates across cropland in China's Loess Plateau and experienced great success in constraining estimates of yield in the CERES-Wheat model across irrigated and non-irrigated regions. However, the high-resolution estimates of soil moisture assimilated in this study were derived based on a modeled relationship between Sentinel-1 radiance information, a water cloud model, and a high-quality, expansive field experiment dataset covering 45 sites across their specific study region. This method helped to localize the derivation of soil moisture estimates for their study. However, to recreate the performance of this study for a different region, a similarly comprehensive dataset with observed *in situ* soil moisture would be required. Similarly, a study by Chakrabati et al. (2014) downscaled the SMOS RS soil moisture data product from 25 km to 1 km and, then, assimilated the downscaled data product into the DSSAT crop model. Their system was able to improve estimates of soybean yield for 2 growing seasons in the lower La-Plata Basin in Brazil when compared to the free model. However, their downscaling approach required *in situ* observations of soil moisture and soil texture which were available at high temporal resolutions (i.e., every 3 hours) for roughly 26 site-years. Thus, this assimilation approach serves as another example of a systems approach that would be difficult to replicate for new locations and at broader scales.

Yet, the development of a RS soil moisture data-assimilation system that is both high-performing and generalizable is no easy task. The application of RS soil moisture (and other RS data products) in SDA poses several challenges for agricultural forecasting (Huang et al., 2019). First, uncertainty and biases in RS data products are typically poorly defined (Huang et al., 2019). RS soil moisture estimates are, themselves, based on modeled relationships, and, as they are predicted as a function of surface reflectance information, the largely unknown uncertainties in the raw radiance information and in the employed model propagate unsupervised into soil moisture estimates (Weiss et al., 2020). Additionally, although estimates of RS soil moisture are developed to represent surface soil moisture, this representation is typically imperfect in its characterization of the true, in-field values. This stems from a modeled relationship that has been generalized across diverse landscapes (not just agricultural landscapes) and a spatial resolution that is larger than agricultural fields (Huang et al., 2019). The downscaling approach by Chakrabati et al. (2014) and the localized derivation of Liu et al. (2021) helped their data-assimilation systems overcome this challenge. Compared to other state variables, surface soil moisture, which characterizes the first 5 cm of the soil profile, has also demonstrated limited constraint of soil-plant-water dynamics in past studies. Among others, De Lannoy et al. (2007) and Monsivais-Huertero et al. (2010) both found the assimilation of near-surface soil moisture to be far less effective than the assimilation of other soil layers when constraining soil moisture profiles due to largely de-coupled moisture pools (Mishra et al., 2021). Yet, since surface soil moisture is typically the layer where fertilizers are added, its accurate estimation is nonetheless important for today's agroecosystems (Verburg and CSIRO, 1996).

Considering these limitations, there are many clear opportunities for improvement when it comes to the generalizability and performance of RS soil moisture SDA in crop models. First, even though it is well known that uncertainty in RS data products is often not well characterized (Huang et al., 2019), there have been few attempts to estimate uncertainty in RS soil moisture estimates in the context of SDA. Instead, many studies utilize reported accuracy metrics (i.e., standard errors) from the literature (e.g., Dente et al., 2008; Ines et al., 2013; Lu and Steele-Dunne, 2019) or compare the RS estimates to *in situ* values available for their study region (e.g., Liu et al., 2021). However, the generalizability of such estimates is unknown. Since the performance of data-assimilation systems can be highly sensitive to prescribed values of observation uncertainty (Ouaadi et al., 2021), there is great potential for an algorithm, such as the Miyoshi algorithm, to better represent system uncertainties and to improve assimilation performance. Second, past SDA studies have typically assimilated soil moisture estimates from a single RS data product and, therefore, have leveraged only a small fraction of the available information. Combining information from several data products could help to reduce uncertainties related to the derivation of the estimates (Beck et al., 2020) and reduce the assimilation interval (Huang et al., 2019). In fact, Lu and Steele-Dunne (2019) found that assimilating SMOS and SMAP soil moisture estimates together increased assimilation frequency by 41% and outperformed the performance of the free model and the assimilation of each data product individually. Other studies have echoed the importance of high assimilation frequency when assimilating soil moisture to improve model constraint (e.g., De Lannoy et al., 2007; Pauwels et al., 2007). Finally, the data, derivation, and availability of RS soil moisture data products continues to improve. Soil moisture can be inferred from optical, thermal, and microwave RS information. However, the majority of recent RS data products have focused on the use of microwave radiance—from both passive and active sensors—to estimate soil moisture as it is less impacted by cloud cover (Peng et al., 2021). By combining information from passive and active sources, new data products aim to achieve higher spatial and temporal resolution (Lievens et al., 2017). Further evaluation and comparison of such RS soil moisture data products in SDA contexts can help drive innovation in agricultural forecasting methods and help guide the generation of future remote sensing data products.

In the following study, 4 critical issues regarding RS-based soil moisture SDA are investigated in the context of agricultural forecasting. These issues concern (1) the trade-off between temporal resolution and spatial resolution in assimilated RS soil moisture observations, (2) the impact of combining observations from different RS data products on system constraint, (3) the value of systematic estimates of observation uncertainty on EnKF performance, and, finally, (4) the potential for surface soil moisture to constrain model estimates related to root-zone soil moisture and crop productivity. To investigate these issues, the developed data-assimilation system was adapted and applied to assimilate surface soil moisture observations from 4 different RS soil moisture data products for 10 site-years across the U.S. Midwest. The included data products varied broadly in derivation, revisit time, and spatial resolution and were evaluated both individually and in combination with other data products with regard to their constraint of downstream APSIM state variables. This study presents a generalizable and robust approach for assimilating RS soil moisture into a crop model and highlights important insights for the future innovation and expansion of these methods.

This chapter has three main objectives:

1. To assimilate 4 RS soil moisture data products individually using the presented data-assimilation system and to evaluate and compare their downstream constraint of soil moisture, NDVI, and crop yield with respect to the spatial and temporal resolution of the observations.
2. To assess the added impact on system performance when combining soil moisture observations from different RS data product.
3. To investigate the potential downstream constraint of RS soil moisture data assimilation on soil water dynamics and crop development for experimental sites across the U.S. Midwest.

MATERIALS AND METHODS

Study sites

The 5 study sites presented in Chapter 3 were also considered in this study. However, due to the limited temporal coverage of some RS data products, only site-years in 2015 and later were evaluated. Aside from the RS soil moisture data products, no additional site-level observations were introduced for evaluating system performance. Evaluation focused on the soil sensor soil moisture, yield, and NDVI observations for the site-years as described in previous chapters.

Remote sensing soil moisture

In this chapter, 4 RS soil moisture data products, spanning different temporal and spatial resolutions, were extracted at the point-level for the study sites and utilized in the data-assimilation workflow (Table 4.1). Unlike the sensor-based soil moisture observations employed in previous chapters, the RS-based soil moisture observations represent surface soil moisture, which characterizes the first 5 cm of the soil profile. Only estimates from April through November were considered to avoid issues with snow cover and freezing soils in the winter months. Upon introducing each data product below, the study IDs provided in Table 4.1 will be used to identify them for the duration of this work.

ESA-CCI

The soil moisture dataset with the coarsest spatial resolution in this study is the ESA-CCI soil moisture product. Each year, the European Space Agency Climate Change Initiative (ESA CCI) algorithmically merges active and passive Level 2 microwave sensor data products to estimate daily surface soil moisture across the globe for over 40 years. Three soil moisture products are produced annually: ACTIVE, PASSIVE, and COMBINED. Dorigo et al. (2017) provides the full documentation on how these

Table 4.1. Overview of remote sensing soil moisture data products used in this study.

Product	Study ID	Temporal coverage	Temporal frequency	Spatial resolution	Average availability[a]	Average variance	Reference

ESA-CCI	ESA	1978-2019	1-2 days	0.25°	219 days	0.0003	Dorigo et al. (2017);
							Gruber et al. (2019);
							Gruber et al. (2017)

SMAP-Hydroblocks	SMAPHB	2015-2019	3 hours	30 m	127 days	0.0050	Vergopolan et al. (2021)
SMAP-Sentinel1	1KM/3KM	2015-now	12 days	1 km/3 km	7 days	0.0025	Das et al. (2019)

[a]Availability calculated after removing observations in the winter months (i.e., Dec-Mar) and are given on a per-year basis.

data products are produced. The COMBINED product (version v06.1), which includes daily estimates of uncertainty, was extracted for use in this study through the ESA-CCI website in January 2022. Several past studies have assimilated this data product into process-based models with varying levels of success (e.g., Zhou et al., 2016; Liu et al., 2017; Liu et al., 2018; Naz et al. 2019).

SMAP-HydroBlocks

The SMAP-HydroBlocks soil moisture dataset, the data product with the highest spatial resolution in this study, was first introduced in 2021 by Vergopolan et al. (2021). The data product leverages information from the HydroBlocks land surface model, a Tau-Omega radiative transfer model, machine learning, satellite-based data products, and *in situ* observations to estimate surface soil moisture with 30-meter resolution across the contiguous United States. The Hydroblocks model was coupled with a Tau-Omega radiative transfer model (HydroBlocks-RTM) and used to simulate soil moisture, soil temperature, and brightness temperature at a 3-hour, 30-meter resolution. Brightness temperature estimates from NASA's Soil Moisture Active Passive (SMAP) mission were then merged with the HydroBlocks-RTM estimates using a spatial cluster-based Bayesian merging scheme and, using the inverse HydroBlocks-RTM, soil moisture was estimated at the same 3-hour, 30-meter resolution. Vergopolan et al. (2021) report an RMSE of 0.07 mm^3/mm^3 after comparing SMAP-Hydroblocks estimates to *in situ* observations from 233 independent experimental sites. The published article provides more detail on the estimation materials and methods.

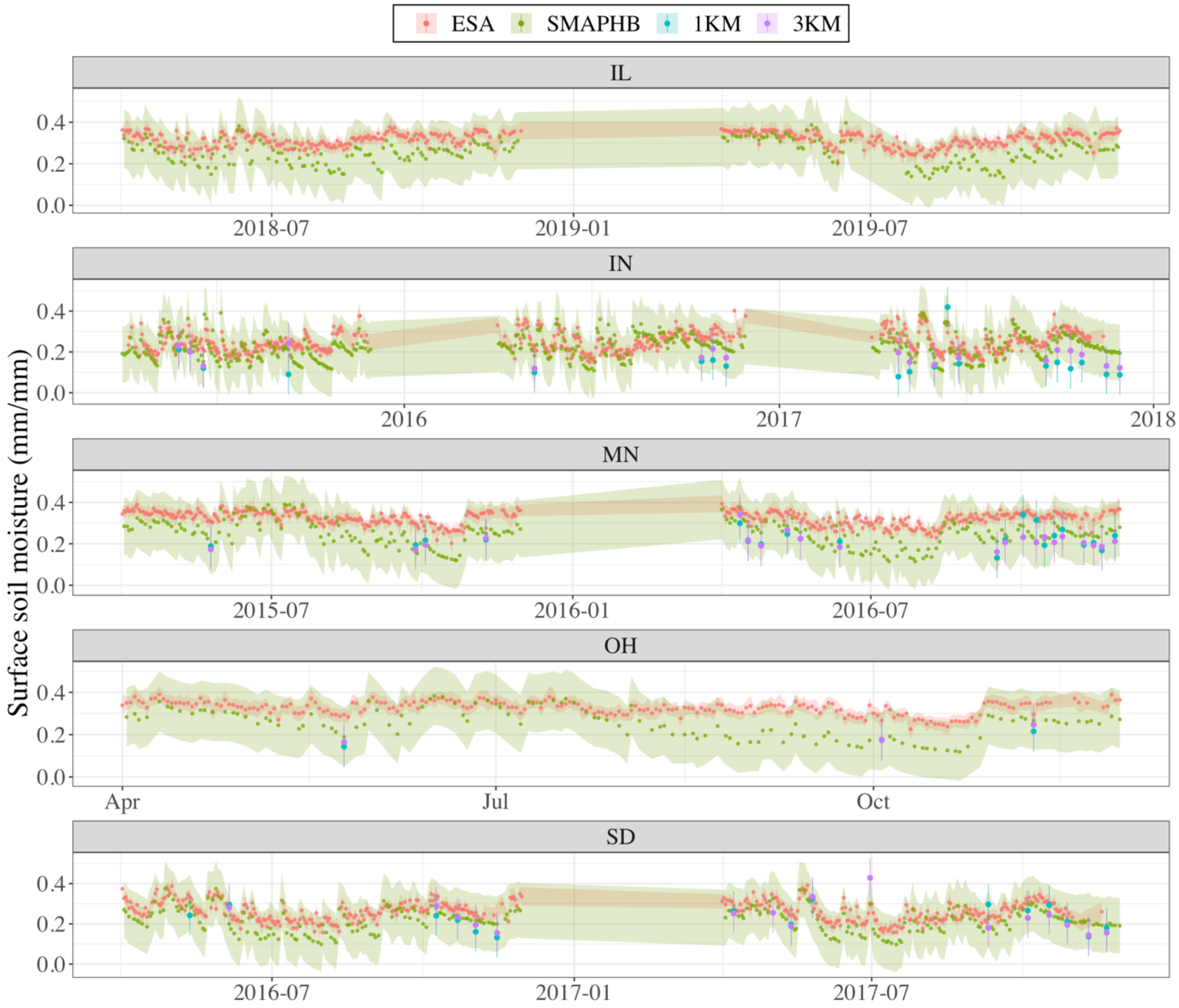

Figure 4.1. Time series for surface soil moisture estimates from the 4 RS data products included in this study. Points indicate the mean values for all data products. Ribbon plots are used to demonstrate the 95% confidence interval around the mean estimates for ESA and SMAPHB, as they are more complete time series. 95% confidence intervals for the sparser data products—1KM and 3KM—are represented by point ranges.

This study is the first to assimilate SMAP-HydroBlocks soil moisture estimates into a crop model. Soil moisture estimates were provided at the daily resolution, and site-level estimates were computed as the mean value of any data point within 0.0005° of the given site coordinates. The variance of each estimate was calculated based on the spatial variability of selected data points and the reported standard error (SE = 0.07 mm^3/mm^3) as

$$Var(Y_{s,t}) = Var(y_t) + SE^2$$

where, for site s at the tth available time step, Y represents the site-level soil moisture estimate and y presents soil moisture estimates within 0.0005° of the site location.

SMAP-Sentinel1

The final dataset considered in this study is the SMAP-Sentinel1 soil moisture product, which was produced by merging information collected by the SMAP L-band radiometer and the Copernicus Project Sentinel-1 C-band radar. The SMAP mission was intended to acquire high spatiotemporal resolution soil moisture estimates globally, but, in July 2015, the SMAP radar became inoperable. Therefore, Sentinel-1 active microwave data were used to supplement passive microwave sensor information from the still-operating SMAP radiometer to help account for the system malfunction and to allow for the continued use of the active-passive algorithm to estimate surface soil moisture content. The merge increased the revisit interval from 3 to 12 days, and data are available at two different spatial resolutions (i.e., 1 km and 3 km; Lievens et al., 2017). Upon comparison of the estimates with *in situ* soil moisture measurements, the reported RMSE for SMAP-Sentinel1 soil moisture estimates was roughly 0.05 m^3/m^3. In this study, this value was applied as the standard error for soil moisture estimates at both spatial resolutions and at all available time steps. Data was obtained from the NASA Distributed Active Archive Center (DAAC) at the National Snow and Ice Data Center (NSIDC) using Python notebooks provided on the NSIDC website. Estimates were available for all TD site-years but were not available for IL.

Data-assimilation system

In general, the data-assimilation system applied in this chapter is identical to the system presented in Chapter 3. However, two notable changes were made and are explained in the following sections. The first change concerns the method by which the posterior distribution and the inflation factor are estimated, while the second change concerns the formatting and application of the observed data and the simulation set-up.

Generalized ensemble filter

To increase flexibility in model definition and to allow for the relaxation of the normality assumptions of the EnKF, an alternative sequential data assimilation approach was tested in this analysis to replace the EnKF-Miyoshi method. The new method, which is known hereinafter as the Generalized Ensemble Filter (GEF), comprises a fully numerical Bayesian approach to estimating the analysis distribution, as well as a variance inflation scalar. The model resembles the approach presented by Raiho et al. (2020) and has the following form:

$$Q \sim U(0.001, 5)$$

$$X \sim N(\,\mu_f,\, P_f + (Q-1) * \left(I * P_f\right))$$

$$Y \sim N(X, R)$$

where Q is the estimated forecast inflation scalar and I is the identity matrix. The estimation of the analysis distribution (X) and Q was completed using a Markov Chain Monte Carlo (MCMC) approach by leveraging the NIMBLE R library (de Valpine et al., 2017).

For the purposes of this study, the GEF was preferred over the previously applied system when (1) assimilating more than 1 observation for a single state variable at a given time step and (2) the observation operator (H) was changing over the course of the simulation (i.e., not all data products are available on a given day). The Miyoshi algorithm, as implemented in the previous workflow, could not be easily adjusted to perform well under these conditions. The GEF also allows for the definition and estimation of more complex relationships between observations and model forecasts (e.g., nonlinear observation operators). Based on these two advantages, the GEF was developed, tested, and applied over the original workflow for all multi-observation assimilation schemes in this study. Preliminary applications of the GEF with the sensor-based soil moisture observations from Chapter 3 showed satisfactory performance with the new scheme (results not shown). See Figure A.10 for a schematic that highlights the differences in the GEF scheme compared to the Miyoshi scheme (Fig. 2.4).

Simulation set-up

The next major difference in the system concerns the observation operator and simulation set-up. In Chapter 2 and 3, each SDA run had the same observation format and utilized the same data-assimilation workflow: two observations of the same type (i.e., soil sensor) that characterize two different state variables (i.e., SM3 and SM4) were assimilated into the APSIM model using the EnKF-Miyoshi workflow at time steps where both observations were available. However, this chapter introduces new complexities to the data-assimilation configuration as all RS observations estimate the same quantity (i.e., surface soil moisture).

First, the Miyoshi scheme was applied to independently assimilate each individual RS data product into APSIM. The GEF was not applied when assimilating one observation as the MCMC did not converge in these conditions. These individual runs were performed to directly compare the value of different RS data products in the context of site-level soil moisture data-assimilation. Next, the GEF scheme was applied to jointly assimilate observations from multiple RS soil moisture data products into APSIM following an additive approach, such that each subsequent SDA run introduced another RS data product to the observation list. The first iteration included observations from one data product and the fourth iteration included all available observations. Since the RS data products varied in terms of temporal availability, the observation operator (H) was dynamically adjusted within the GEF scheme to reflect observation availability each day. A minimum of 2 observations per day were required for data assimilation due to limitations of the GEF (i.e., lack of MCMC convergence). Data products were added in succession based

Table 4.2. Overview of system structure, observation format, and naming protocol for data-assimilation runs presented in this chapter.

Group	Scheme	Study ID	Included observations
Individual observations	Miyoshi	SMAPHB	SMAPHB
		1KM[a]	1KM
		3KM[a]	3KM
		ESA	ESA
Additive observations	GEF	+SMAPHB	ESA + SMAPHB
		+1KM[a]	ESA + SMAHB + 1KM
		ALL[a]	ESA + SMAHB + 1KM + 3KM

[a] Soil moisture estimates from 1KM and 3KM were not available for IL so some runs were not completed for those site-years.

on availability, such that the first data product tested had the highest average number of observations per year. By sequentially adding new data products, the information contribution of each RS data product could be effectively quantified.

In this study, all RS soil moisture observations were merged with APSIM model forecasts of surface soil moisture (SM1). See Table 4.2 for more details on the configuration and naming protocol of the runs evaluated in the following sections. The same free model runs evaluated in Chapter 3 will serve again here as a baseline for comparison. The runs in this chapter are also consistent with Chapter 3 in ensemble number (i.e., n = 100), continuity of Miyoshi parameter values, and first-year initialization period length (i.e., Jan 1).

Evaluation metrics

For quantifying changes in forecast accuracy and precision, this study utilized only a subset of the metrics described in the previous 2 chapters. RMSE, nRMSE, weighted variance, spectral norm, and the Pearson correlation coefficient (R) were all employed in this work to help characterize changes in system performance across different configurations of RS data products. An innovation analysis of soil moisture adjustments was also completed as demonstrated in Chapter 3. The classification of site-years as “improved” or “reduced” in terms of forecast accuracy and precision was completed following the same approach as outlined in Chapter 3.

Evaluated state variables

Unlike Chapters 2 and 3, the evaluation of forecast accuracy and precision was not completed for the assimilation state variable (SM1) as no observations were available to serve as the "ground truth." Previous chapters employed sensor soil moisture data to evaluate model forecasts of soil moisture. However, surface soil moisture was not directly observed with soil sensors at any of the study sites. Furthermore, since biases within RS soil moisture data products are not well known and can vary greatly, the application of RS estimates for evaluation purposes would not be justifiable (Huang et al., 2019). Consequently, evaluation of each run in this work focused on the top-down effects of state data assimilation, quantifying how changes in SM1 due to assimilation indirectly impacted the accuracy and precision of observed downstream state variables. Evaluated state variables include soil moisture in all observed layers (SM3-7), NDVI, and crop yield. Following the inconsistent constraint by the system as presented in Chapter 3, annual drainage and nitrate load were not evaluated here.

Information contribution

A new method was introduced in this chapter to quantify the information contribution of each data product to system constraint. With each new data product added to the system within the additive runs, the average change in μ_a and P_a across all analysis time steps was calculated relative to the previous system estimates. This average change represents the *new* information contributed by each data product relative to the information that had already been added to the system. Higher values indicate more information. Since there were no observations to evaluate SM1 constraint, it was not possible to evaluate the quality of each data product's information contribution.

RESULTS

The results of this chapter are presented in three distinct sections. The first section focuses on the comparison of individual RS data products and their constraint of APSIM state variables in forecast accuracy and precision. The second section moves to the comparison of the additive SDA runs and the quantification of the information contribution of each added data product and its impact on system constraint in APSIM. Finally, the last section presents a closer investigation of APSIM constraint when assimilating all available RS information (i.e., ALL) compared to the free model.

Individual runs

As expected, the individual influence of each RS data product was heavily dependent on temporal availability. ESA, the most widely available data product, had the greatest impact on both assimilation and downstream state variables, while assimilation with 1KM and 3KM imposed only slight changes in

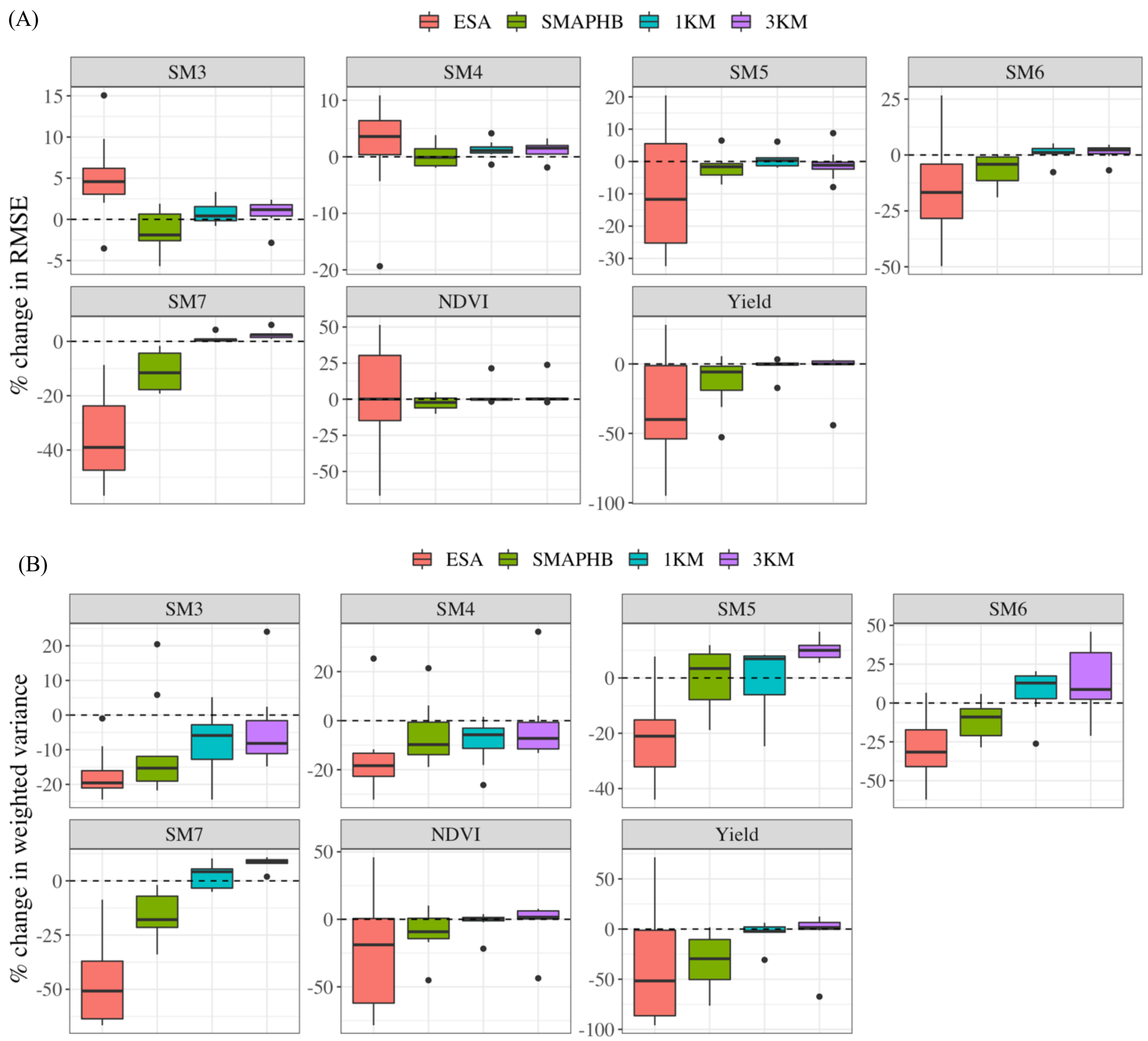

Figure 4.2. Boxplots demonstrating the distribution of relative change (%) in state variable (a) accuracy (RMSE) and (b) precision (weighted variance) due to the assimilaton of each of the RS data products individually across all site-years. Change is computed relative to the free model results. Negative values indicate improvement (e.g., (RMSE$_S$ – RMSE$_F$) / RMSE$_F$).

estimates when compared to the free model. However, ESA did not always lead to improvements in model performance. As demonstrated in Figure 4.2a, ESA results were more variable across site-years in terms of the accuracy of state variable estimates, in some cases leading to great improvement and, in other cases, leading to reduced performance. ESA led to reduced accuracy in predicting SM3 and SM4 in 40% of site-years but was the most effective in improving accuracy in estimates of annual yield, SM6, and SM7 compared to other data products. It had mixed impacts on NDVI and SM5. ESA also outperformed the other 3 RS data products in constraining forecast precision for all state variables (Fig. 4.2b), improving precision in 60-100% of site-years. Importantly, it showed the greatest reduction in the spectral norm of the

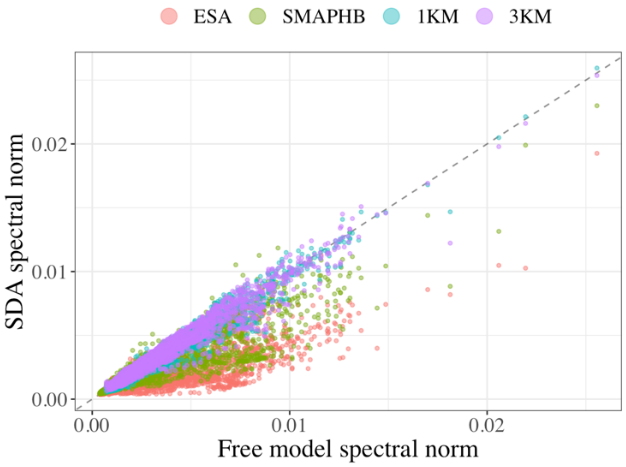

Figure 4.3. The spectral norm was computed for the weighted covariance matrix of daily soil moisture estimates across all soil layers (i.e., SM1-SM7) for each run as a measure of total soil moisture precision. The scatterplot compares the free model results with the results of each individual RS-SDA run. The 1:1 line is represented by the black dashed line.

soil moisture covariance matrix when compared to the free model, indicating the best constraint of soil moisture precision across the entire profile (Fig. 4.3).

Alternatively, the assimilation of SMAPHB, another temporally frequent RS data product, demonstrated more conservative performance than ESA across state variables. For almost all state variables, it typically performed similarly or better than the free model. Any improvements (or reductions) in forecast accuracy were more moderate than observed with ESA. For example, accuracy in NDVI estimates was never reduced with SMAPHB, but the greatest improvement observed in the tested site-years was a 10.1% accuracy increase. On the other hand, NDVI accuracy was reduced for 40% of site-years with ESA, but the maximum improvement was a 67% increase. This trend in the results highlights one important trade-off when assimilating more certain observations (i.e., ESA) over less certain observations (i.e., SMAPHB) when both data products have unknown biases. In terms of forecast precision, SMAPHB was overall quite effective in constraining state variable predictions, especially when compared to 1KM and 3KM. However, SMAPHB largely underperformed compared to ESA in this regard. 1KM and 3KM both underperformed in accuracy constraint when compared to ESA and SMAPHB, showing little to no change in RMSE when compared to the free model.

Considering the 4 individual runs, more frequent assimilation time steps also led to more robust performance of the EnKF-Miyoshi workflow. Filter divergence (i.e., when the observed mean falls outside of the 95% credibility interval of the analysis distribution) occurred at 52% and 59% of analysis time steps for 1KM and 3KM, respectively, but occurred at only 44% and 30% of analysis time steps for SMAPHB and ESA, respectively. For 1KM and 3KM, the Miyoshi algorithm also tended to estimate greater forecast inflation at analysis time steps, which could be a consequence of having great discrepancies between the observed and forecasted means at analysis time points (e.g., see Oct. and Nov. in Fig. 4.4). For estimates of observation uncertainty, the Miyoshi algorithm predicted lower uncertainty for all observations than

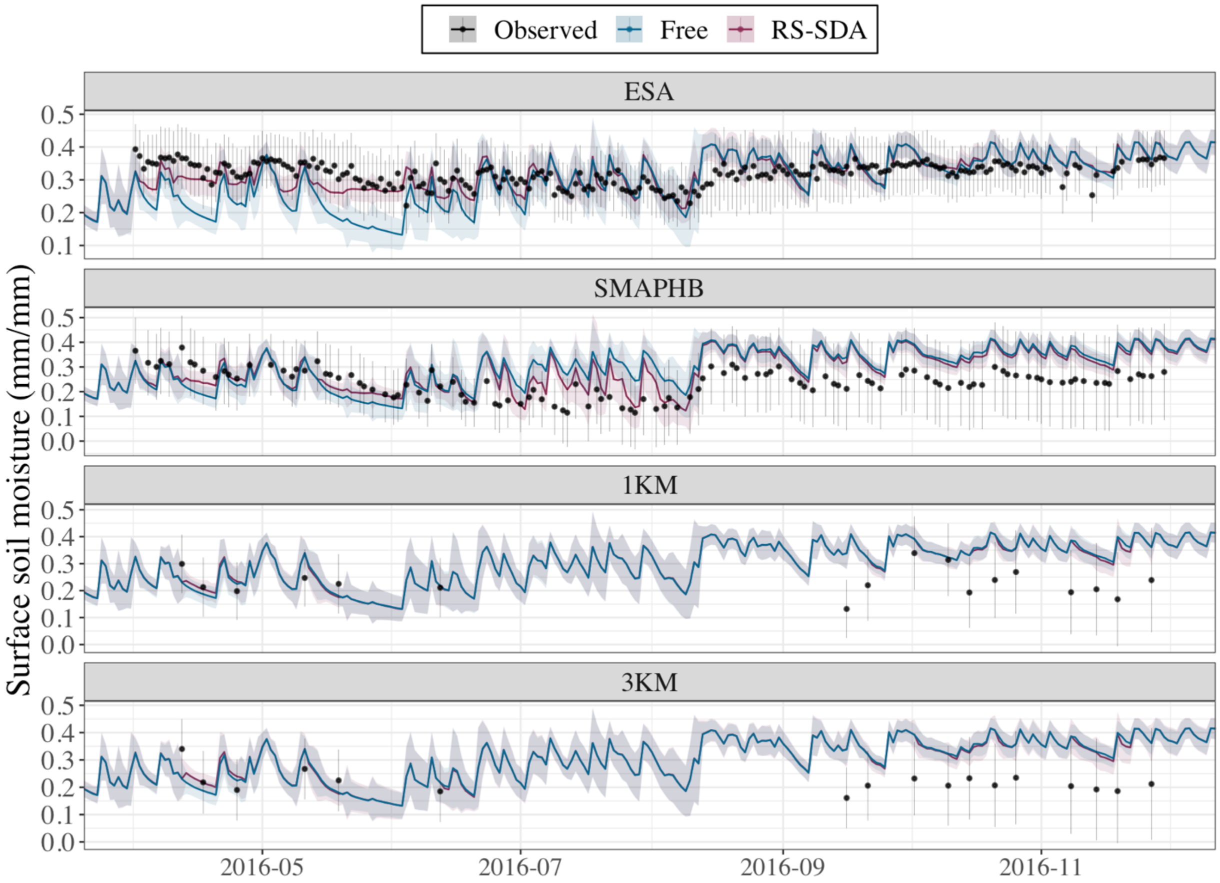

Figure 4.4. Time series of surface soil moisture estimate (SM1) distributions from the free model, SDA forecast, and RS observations for each individual RS data product. The lines represent the mean estimates, and the shaded regions indicate the 95% confidence interval at each time point. The black points represent observed means with vertical lines demonstrating 95% confidence intervals based on the reported standard error in the literature. This time series shows results from the IN 2016 site-year, which had the most observations for the 1KM and 3KM data products of all site-years.

reported in the literature. Standard error in SMAPHB estimates was reported as 0.07 mm^3/mm^3 but estimated to be 0.005 ± 0.003 mm^3/mm^3. Standard errors in 1KM and 3KM estimates were reported as 0.05 m^3/m^3 but estimated by the system to be 0.005 ± 0.002 mm^3/mm^3. SMAPHB had the greatest estimated uncertainty and ESA had the lowest estimated uncertainty according to Miyoshi estimates, a result that mimics reported values.

Additive runs

The baseline run for the additive RS-SDA runs was ESA, which demonstrated inconsistent constraint of forecast accuracy and strong constraint of forecast precision. As the second most available data product, SMAPHB was the next RS data product added and assimilated into the system. New SMAPHB observations, on average, imposed a -0.012 mm/mm change in μ_a and a -0.0003 change in P_a for

SM1 estimates. These changes translated into lower overall surface soil moisture estimates and increased precision. For downstream forecast accuracy, the addition of SMAPHB led to improved and/or more consistent constraint for all state variables except SM7 (Fig. 4.5a). At times, the added information from SMAPHB dampened the benefit of SDA, decreasing the degree to which accuracy measures are improved and producing estimates that are closer to the free model. However, the incidence and magnitude of reduced accuracy was decreased. This is clear when considering yield in Figure 4.5a, where a larger part of the

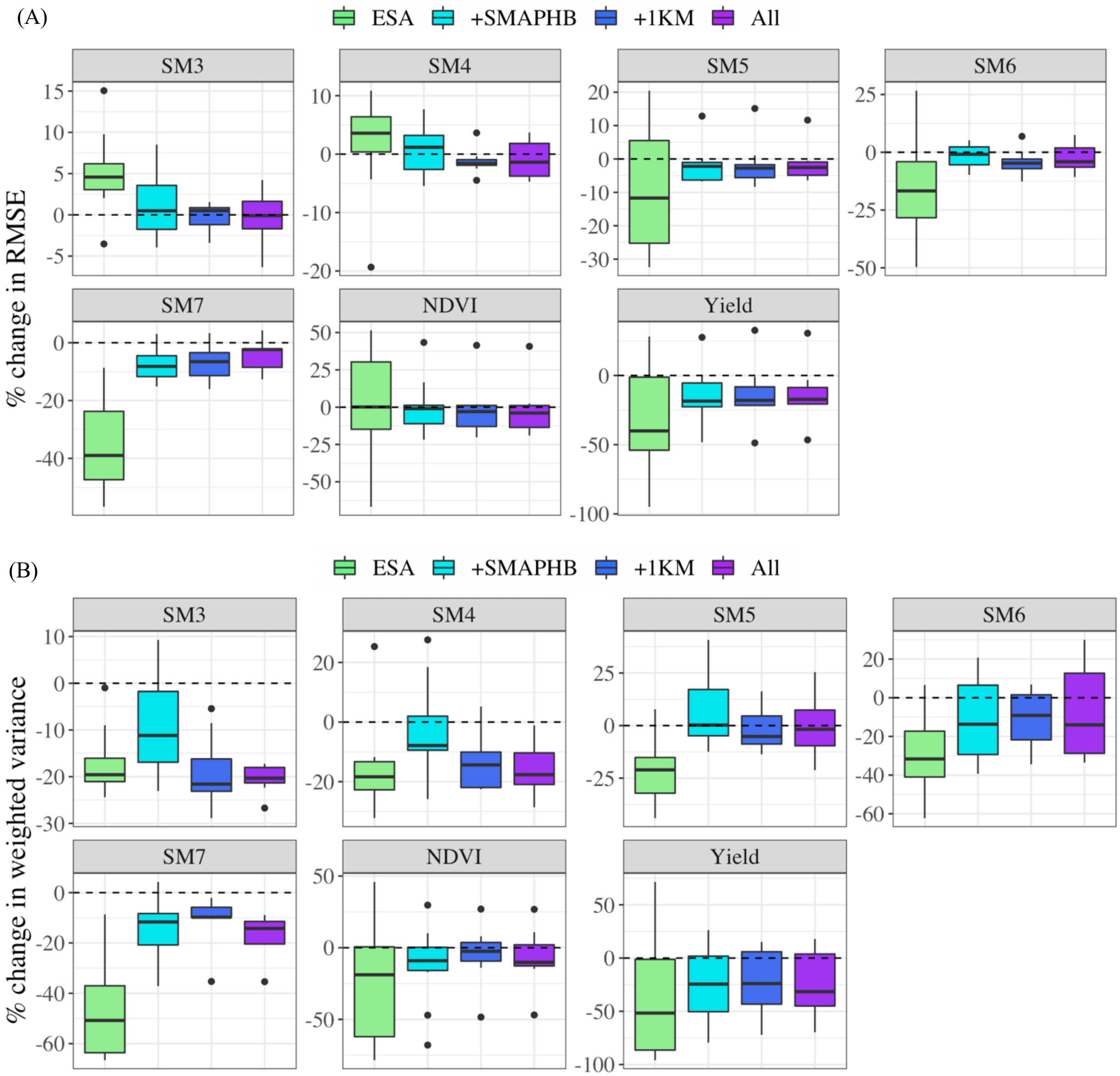

Figure 4.5. Boxplots demonstrating the distribution of relative change (%) in state variable (a) accuracy (RMSE) and (b) precision (weighted variance) due to the assimilaton of each iteration of the additive RS data products across all site-years. Change is computed relative to the free model results. Negative values indicate improvement (e.g., ($RMSE_S$ – $RMSE_F$) / $RMSE_F$).

distribution falls below zero in +SMAPHB compared to ESA (i.e., more site-years saw net reduction in RMSE) but the lower bound is much higher. For forecast precision, the addition of SMAPHB observations reduced performance for all state variables compared to ESA (Fig. 4.5b). However, in most cases, +SMAPHB performance in precision was still better than or similar to that of the free model. SM4, SM5, and SM6 were exceptions to this rule, such that forecast precision was notably reduced for those state variables.

The subsequent additions of the sparser 1KM and 3KM RS data products were less impactful than the addition of SMAPHB. New 1KM observations imposed an average -0.0004 mm/mm change in μ_a, and, later, new 3KM observations imposed an average -0.0003 change in μ_a. These changes are less than 4% of the change imposed by the initial addition of SMAPHB. Neither additional data product effected a notable average change in P_a. Following these minimal changes in the analysis adjustment of SM1, there was also little change in forecast accuracy and precision for downstream state variables in +1KM and ALL as compared to +SMAPHB (Fig. 4.5). Nevertheless, adding 1KM observations to +SMAPHB did hold some benefit for estimates of SM3 and SM4 in terms of accuracy and precision. Forecast precision for the two state variables, which had decreased with the addition of SMAPHB, was again constrained to a similar extent as observed with ESA. Accuracy for SM3 and SM4 was better in +1KM than both ESA and +SMAPHB. Therefore, the 1KM observations were generally useful additions to the system despite their limited information contribution. The effect of the 3KM observations was almost negligible or, even at times, harmful to system performance.

Analysis of downstream impact

The following sections look closer at the downstream impacts of RS-SDA on the soil water cycle and aboveground crop estimates. Although the addition of 3KM observations did not have a dramatic impact on system performance compared to +1KM, the following sections focus on the results from ALL to include all RS observations for the sake of completeness. Hereafter, ALL will be referred to as RS-SDA. As IL did not have available 1KM or 3KM observations, the +SMAPHB will replace ALL for those site-years as RS-SDA.

Soil water cycle

The assimilation of RS soil moisture had minor impacts on the soil water cycle. Figure 4.6 demonstrates differences between the free model and RS-SDA in SM1 estimates, the state variable directly constrained by assimilation. For several site-years, RS-SDA estimated significantly higher SM1 values in the early growing season (i.e., May-Jun) compared to the free model. Then, in the late season and fall, RS- SDA often estimated lower SM1 values. As seen previously, the impact of these SM1

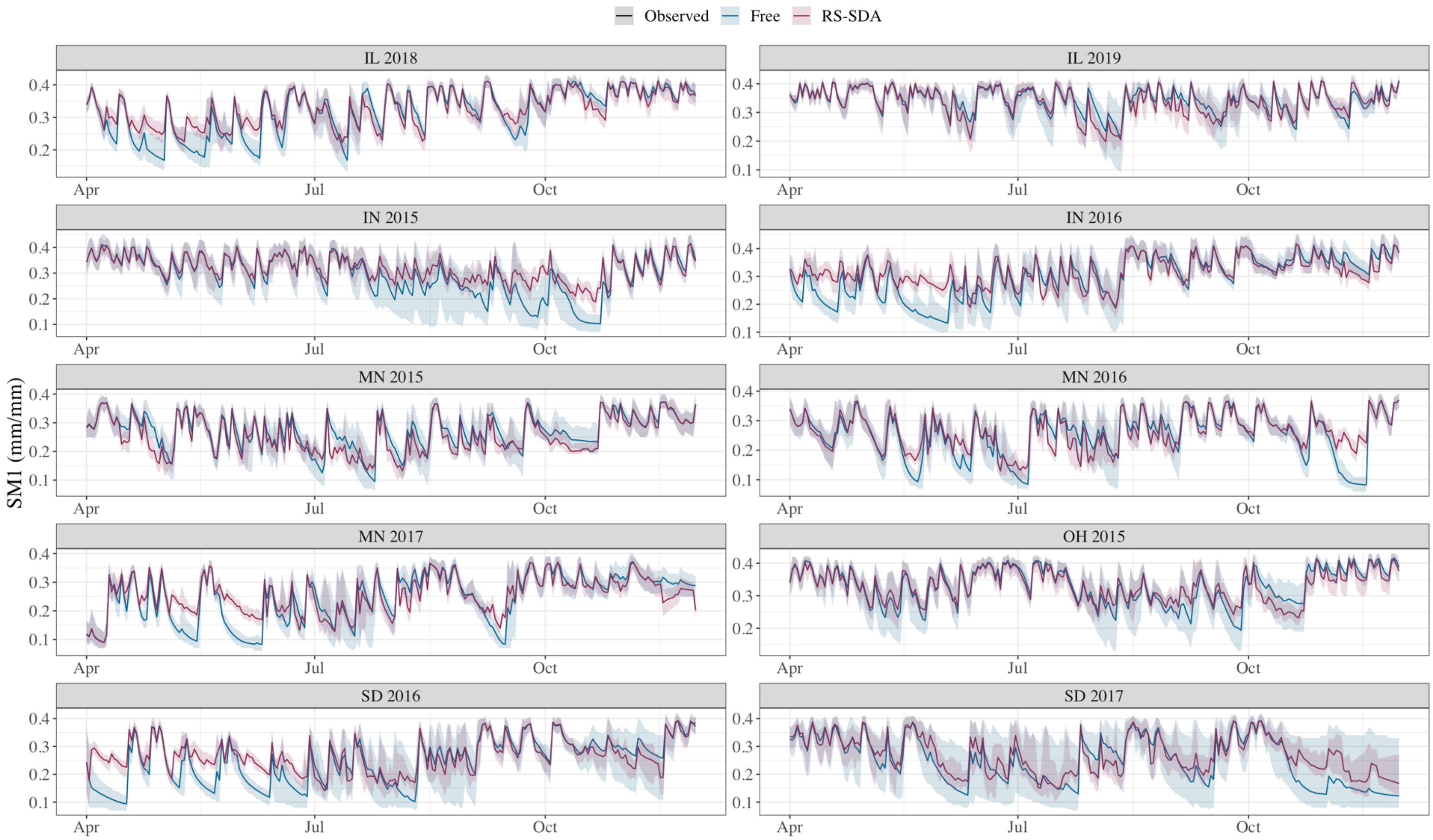

Figure 4.6. Time series of SM1 estimates from the free model and RS-SDA with the mean daily estimates demonstrated with line graphs and 95% credibility intervals are indicated by the shaded regions.

Table 4.3. Summary statistics to quantify the impact of RS-SDA on forecast accuracy of APSIM state variables. The "N" column indicates the number of site-years with available data for each state variable, and the "n" column indicates the total number of observations across site-years. (F) denotes a value computed for the free model estimates, and (S) denotes a value for the RS-SDA estimates. Relative error (RE) is also given as a median value. The median change (Δ) in RMSE was computed as RMSE_S - RMSE_F / RMSE_F.

Variable	N	n	RMSE (F)	RMSE (S)	Δ RMSE	R^2 (F)	R^2 (S)	RE (S)
			median value					
SM3 *mm/mm*	10	5592	0.082	0.084	-0.9%	0.48	0.48	-29.2
SM4 *mm/mm*	10	6141	0.068	0.069	-2.8%	0.43	0.43	-18.2
SM5 *mm/mm*	8	5101	0.061	0.059	-2.6%	0.45	0.45	5.83
SM6 *mm/mm*	10	6169	0.075	0.075	-1.0%	0.43	0.42	12.8
SM7 *mm/mm*	6	3265	0.088	0.077	-5.4%	0.44	0.43	21.5
NDVI *unitless*	10	134	0.206	0.201	-1.8%	0.69	0.71	9.11
Yield *Mg/ha*	10	10	1.45	1.24	-17.2%	0.53	0.69	13.4

changes on lower layer soil moisture values seemed to decrease with depth, such that differences between the free model and RS-SDA mean estimates were more subtle in deeper layers (Fig. A.13). This reduced impact on lower layers is also, in part, a reflection of the increasing total soil water volume represented by soil layers down through the profile (see Table 2.2 for soil layer depths). Nonetheless, any differences in soil moisture estimates did not lead to notable improvement in accuracy for any soil moisture layer (Table 4.3). Moreover, the increased available soil water with assimilation did not propagate into large differences in soil evaporation, drainage, nor runoff estimates (Fig. 4.7). Notable changes were visible, however, in the soil water deficit factors for several growing seasons, such that RS-SDA led to reduced water stress for the growing crop. This could be the result of increased available soil water in the root zone during initial periods of crop water uptake (i.e., June).

An analysis of soil moisture innovations identified 3 important predictors for soil water adjustment in this work. Two of these predictors were also highlighted as important innovation predictors in Chapter 3, including daily error in precipitation input and daily simulated precipitation. The first of these points to the correction of input error when assimilating soil moisture, such that the workflow tended to add water when precipitation inputs missed observed rainfall, and the second indicates that assimilation was more likely to remove water on days with high precipitation inputs. Unlike the innovation analysis results presented in Chapter 3, assimilation's previous soil water adjustment was not identified as a strong predictor

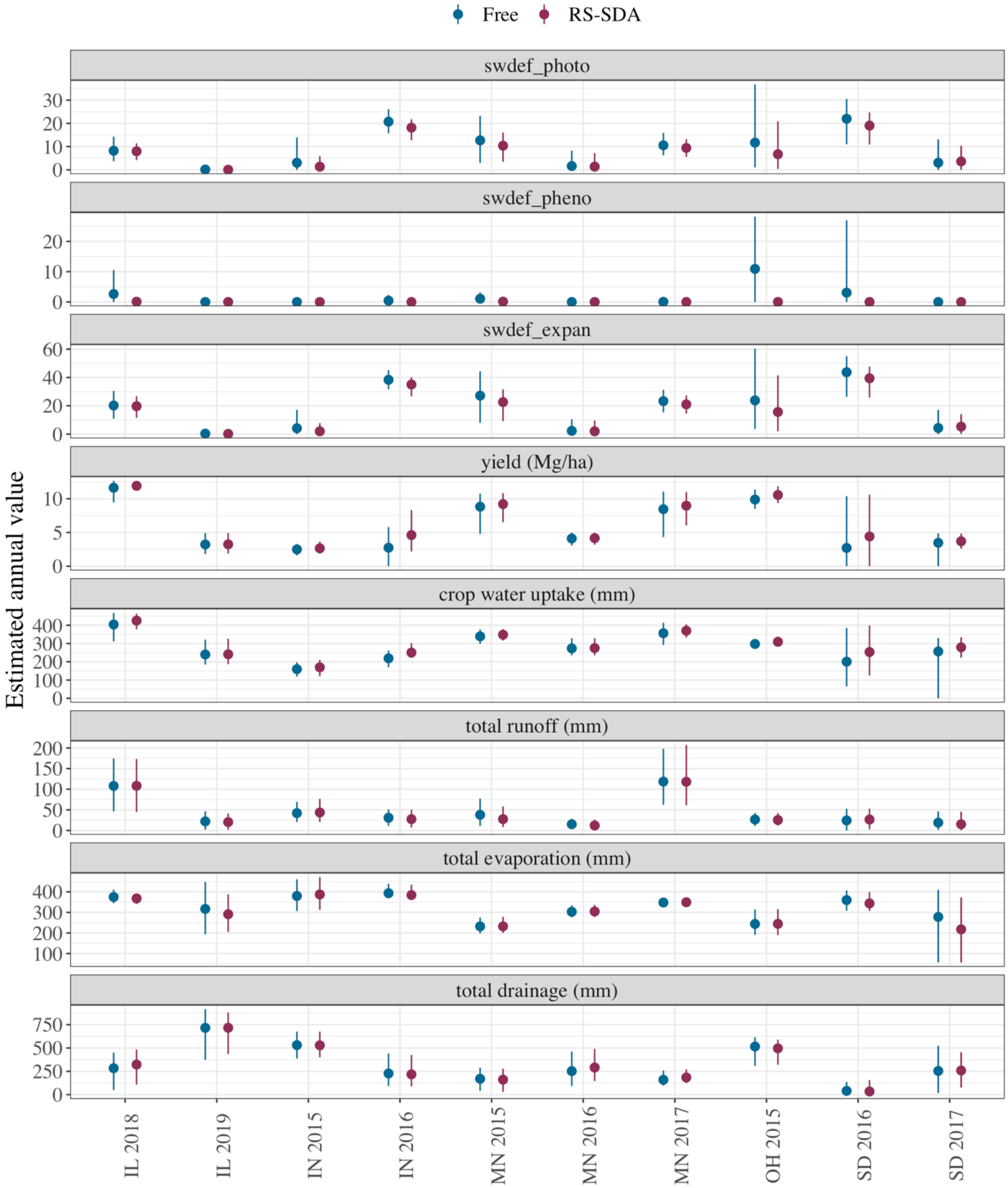

Figure 4.7. Comparison of annual estimates for major components of the soil water cycle for the free model and RS-SDA. Mean values are indicated by the points and 95% credibility intervals are demonstrated by vertical lines.

for soil moisture innovations in this work. Instead, simulated daily radiation was found to be strongly correlated with soil water adjustment, where days with higher simulated radiation were more likely to see added soil water with assimilation.

Forecast precision for soil water-related estimates also did not change substantially with assimilation, but, in most cases, the small changes were beneficial. For SM1 estimates, assimilation greatly reduced variability across site-years (Fig. 4.6). This constraint of soil moisture in the surface soil layer, in many cases, did not propagate into large changes for precision in lower layer estimates (Fig. A.14). However, on average, precision was improved rather than reduced with assimilation, with the greatest downstream constraint in the soil layers closest to the surface. Beyond soil moisture state variables, forecast precision was also noticeably improved for other components of the soil water cycle. Figure 4.7 demonstrates reduced variability in estimates of cumulative soil water deficit factors, drainage, crop water uptake, and evaporation for most site-years in this study.

Aboveground measures

With limited constraint of the soil water cycle, RS-SDA did demonstrate some constraint of NDVI and annual crop yield. Considering the R^2 values reported in Table 4.3, RS-SDA explained roughly 2% and 16% more of the variation in NDVI and yield observations than the free model, respectively. Based on these results, there is evidence that surface soil moisture data assimilation can constrain, to some extent, estimates of annual yield. All site-years except OH 2015 demonstrated increased yield accuracy and 60% of sites demonstrated increased yield precision with RS-SDA. However, unlike previous chapters, there

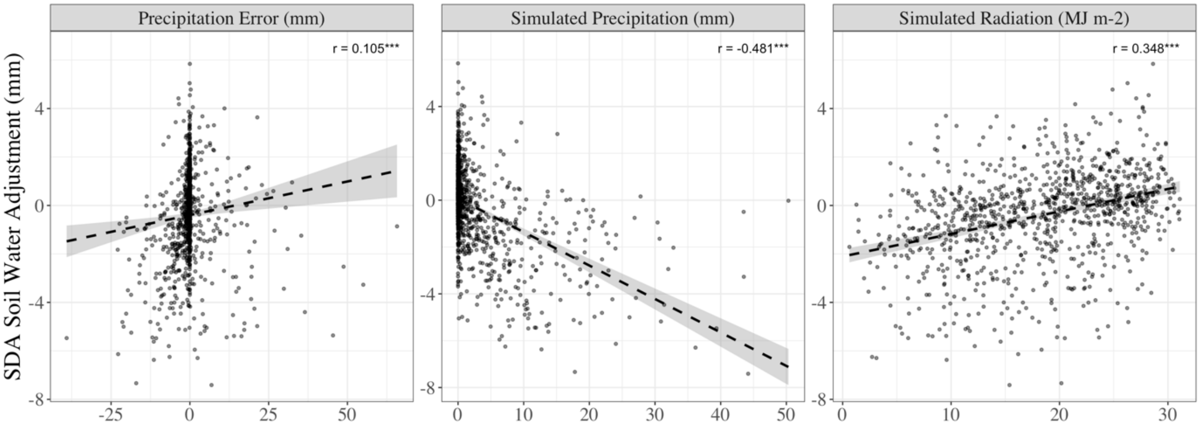

Figure 4.8. Scatterplots demonstrating the relationship between daily soil water adjustment and its three most important predictors. Dashed black lines demonstrate the least squares regression line for each relationship, and 3 asterisks (***) and 2 asterisks (**) indicate a significant linear relationship at all significance levels (α ≅ 0). Together, these three predictors explain 26.6% of the variation in soil moisture innovations.

was no significant trend among site-years between yield improvement and dry conditions, though this could be a consequence of sample size. Compared to estimates of yield, RS-SDA was less impactful in its constraint of NDVI. However, since the free model was able to reasonably predict NDVI ($R^2 = 0.69$), there was less potential for improvement with assimilation. Of the site-years in this study, 60% saw reduced RMSE and 70% saw reduced weighted variance for NDVI estimates.

DISCUSSION

Compared to results seen in Chapter 3, the assimilation of RS surface soil moisture observations imposed far weaker constraint on APSIM state variables compared to the assimilation of the soil sensor observations at the same study sites. For example, the median reduction in soil moisture RMSE ranged from 7-27% across different layers of the soil profile with soil sensor observations (Table 3.3), but, with RS observations, it ranged from roughly 1-5% (Table 4.3). The weakened constraint with RS-SDA was likely not an issue of inaccuracy in the observations, though. Instead, there is greater evidence to show that changes in SM1 simply had less influence on downstream state variables than changes in SM3 and SM4. This is due, in part, to the increased vertical distance between the surface soil moisture layer (SM1) and other observed soil layers in this study (i.e., SM3-7). Increased distance between 2 layers weakens the relationship between their soil moisture estimates as discussed in Chapter 3. Thus, the assimilation adjustment of SM1 estimate would not be as strongly tied to lower layer estimates by a top-down approach as the adjustment of SM3 and SM4. Surface soil moisture data assimilation did notably change SM2 estimates, the soil moisture estimates for the layer just below it. This result reflects the findings of Lu and Steele-Dunne (2019), who assimilated RS surface soil moisture observations into a surface energy balance model. They found SDA improved SM estimates in the second layer to a greater extent than in lower layers when comparing estimates to observations. Since observations were not available for SM2 at the study sites, this hypothesis cannot be tested within this system.

Furthermore, the two different assimilation protocols (i.e., assimilation of SM1 vs. assimilation of SM3 and SM4) were also markedly different in the quantity of soil water associated with their assimilation adjustments. Where soil layers 3 and 4 corresponded to almost 14% of the soil profile (20 cm depth), the near-surface soil layer only corresponded to about 3.6% of the soil profile (5 cm depth). Thus, when considering the top-down effect of soil moisture assimilation on lower layers, each adjustment with RS assimilation had just 25% of the impact of the previous system given the same adjustment in volumetric soil water content. This 5-fold reduction in potential impact closely mirrors the change in RMSE reduction for soil moisture layers as highlighted in the above paragraph (i.e., 7-27% to 1-5%). One way to overcome this limitation of surface soil moisture is to leverage the strong covariance between SM1 and soil moisture in nearby layers (i.e., SM2 and SM3) to directly nudge their values within the analysis time step through the PDA framework (see Chapter 2), as demonstrated in Ines et al. (2013). This would increase the soil profile depth directly adjusted with assimilation, thereby increasing the potential impact of assimilation for downstream estimates. This approach would also reduce or eliminate the need for RS-based root-zone soil moisture (RZSM) data products generated through the assimilation of RS surface soil moisture observations into land surface models (Pablos et al., 2018; Peng et al., 2021).

Despite its limited potential for constraint when compared to previous systems, surface soil moisture data assimilation still demonstrated strong potential for improving APSIM forecasts within this study. First, the assimilation of surface soil moisture still functioned to improve estimates of crop yield overall when compared to the free model, with a median RMSE reduction of 17.2%. Past RS soil moisture data assimilation studies had similar success in improving estimates of crop yield, and several attributed the improvement to increased surface soil moisture and reduced crop water stress with soil moisture assimilation (e.g., Ines et al., 2013; Chakrabati et al., 2014). The results of the current study were inconclusive regarding the impact of water stress on RS-SDA performance (Fig A.15), but, considering the major changes in crop water stress due to SDA (Fig. 4.7), the results do indicate that water stress had an important role in this study. Although observations are not available for crop water uptake to test this hypothesis, one possible explanation could be that RS-SDA increased available soil water at critical growth stages and, thus, increased crop water uptake.

Second, as seen in previous chapters, RS-SDA demonstrated the capacity to account for errors associated with coarse resolution precipitation inputs. This is clear in the results of the innovation analysis where precipitation error was identified as a critical variable for understanding variability in soil water adjustments (Fig. 4.8). At times where precipitation inputs exceeded what was measured at the site, the assimilation workflow was able to compensate for the error by removing water from the SM1 soil layer. The workflow helped to account for the scale mismatch between point-level simulations and gridded weather estimates, “localizing” weather inputs for more accurate simulations. This is an important capacity of RS soil moisture data products that has been demonstrated and discussed in several previous works (e.g., Ines et al., 2013; Lu and Steele-Dunne, 2019; Nair and Indu, 2019). Studies have also shown surface soil moisture assimilation can go beyond weather inputs and identify missing water that was applied through irrigation (Peng et al., 2021). For instance, Nair and Indu (2019) found that assimilating ESA soil moisture estimates (COMBINED v4.2) into the Noah land surface model correctly identified winter irrigation periods and accounted for missing rainfall in precipitation inputs in an assimilation study covering agricultural areas in India. Although the study sites in this chapter were all rainfed, this capability could be critical for scaling up the presented system for regional forecasting since irrigation information would be almost impossible to obtain at the needed resolution.

The results of the innovation analysis in this chapter also provided further evidence for structural and/or parameter errors in APSIM soil water processes. First, as discussed in Chapter 3, the strong linear relationship between soil moisture innovations and precipitation inputs demonstrates that APSIM tends to overpredict surface soil moisture on days with heavy rainfall. This could point to an array of missing processes in APSIM related to tillage, soil properties, preferential flow, runoff, etc. (Malone et al., 2007; van der Laan et al., 2014; Brilli et al., 2017; Ojeda et al., 2018). More observations would be needed to properly diagnose and address the origin[s] of these errors. Second, an analysis of the innovations identified input radiation as a significant predictor of assimilation adjustment. On simulation days with greater solar radiation, assimilation was more likely to add soil water to the profile. This finding suggests potential errors in the model processes related to evaporation, which is the only process where solar radiation can impact surface soil moisture estimates in APSIM (Verburg and CSIRO, 1996). Observations on evaporation would be necessary to be able to further investigate this hypothesis. In fact, the joint assimilation of RS soil moisture and evapotranspiration could function as a systematic and robust way to evaluate and re-parameterize the soil water dynamics in APSIM, leading to better forecasts in the future.

In this study, four different RS soil moisture data products were independently assimilated into the APSIM model and evaluated in their constraint of downstream model variables. These data products varied quite broadly in terms of spatial resolution. SMAPHB observations were available at the finest spatial resolution (30-meter) and ESA observations were available at the coarsest spatial resolution (0.25°). However, despite these grand differences in spatial scale, the individual performance of RS data products seemed to be most closely tied to the temporal availability of observations, such that the order of best to worst data product in terms of performance exactly reflected the ordering from the highest to lowest number of available observations. ESA, which had, on average, 219 observations per growing season, showed the best overall constraint of forecast precision and good constraint of forecast accuracy in downstream state variables among the 4 individual data products. Alternatively, the 1KM and 3KM data products, which each had an average of 8 observations per growing season, had almost no impact on forecast accuracy and only a slight impact on forecast precision. This study was not designed to independently test the impact of temporal and spatial resolution. However, these results echo the findings of Lu et al. (2019), who found high temporal resolution to be far more important to assimilation performance than high spatial resolution. They suspected that increased time between assimilation adjustments allowed errors in model structure, inputs, and/or parameters to go unchecked for longer periods of time, thereby allowing the magnitude of simulation errors to become large and unreasonable. This is especially important in the case of crop modeling since the timing of crop phenology—which can be highly sensitive to soil moisture—is critical to accurately estimating yield. More specifically, in the APSIM *Maize* module, the time between the sowing and germination stages and the time between the emergence and flowering stages both depend on soil moisture. In the latter example, water stress delays phenology stages through the *swdef_pheno* deficit factor (see Fig 2.3). If model error goes unconstrained by assimilation for just a few weeks and imposes inaccurate phenological delays, such delays will be carried forward throughout the model, delaying all the stages that follow, including the start of grain fill. More frequent assimilation helps to mitigate the impact of such model errors and improve overall crop model predictions by correcting errors more often (De Lannoy et al., 2007; Pauwels et al., 2007; Lu et al., 2021). Alternatively, in the case of low temporal resolution, a recalibration-based assimilation approach or the inclusion of a bias correction method might be more appropriate (De Lannoy et al., 2007; Curnel et al., 2011).

When comparing RS data products in this study, it is important to recognize that all data products considered in this work are based, in part, on SMAP radiometer data. SMAPHB merged SMAP brightness temperature data with the HydroBlocks-RTM model, ESA includes SMAP as one of its 10 passive microwave sensors, and 1KM and 3KM rely exclusively on SMAP for passive microwave information within their derivation. Since all 4 data products include SMAP data, it can be assumed that they will overlap to some degree in terms of the information they contribute to the SDA system. In fact, this redundancy is evident when comparing the additive runs. In the first iteration, ESA contributed most of the information provided by the SMAP radiometer to the model and, therefore, imposed large changes in SM1 estimates. Then, with each additional data product, the overall impact on the analysis distribution weakened as much of the new information had already been provided to the system by the data products before it. Though the analysis of somewhat redundant data products may see ineffectual, this overlap in information across data products served to further validate the importance of temporal resolution in this study. Each data product contained similar information, but, when assimilated individually, system performance still varied considerably due to temporal availability.

Estimates of observation uncertainty (R) also had important implications for system performance. In the individual runs, the Miyoshi algorithm estimated notably different levels of uncertainty in the data products. ESA, which was associated with the lowest estimated R, led to the strongest overall constraint of forecast precision and the largest (good and bad) shifts in downstream model estimates. Alternatively, SMAPHB, which was associated with a higher estimated R, had a more moderate impact on model predictions, generating predictions that were closer to the free model. These differences in estimated R can be partially explained by the number of sources employed in the derivation of each RS data product. ESA, which included information from 10 passive and 3 active microwave sensors, had the lowest uncertainty, while SMAPHB, which included just 1 passive microwave sensor and a process-based model, had the greatest uncertainty. 1KM and 3KM were estimated to have slightly lower uncertainty than SMAPHB estimates and included 1 passive and 1 active microwave sensor. Though it is difficult to disentangle the impact of temporal resolution from those of observation uncertainty in the individual runs, differences in forecast precision between ESA and SMAPHB, which had similar temporal resolution, suggest that the added information in the derivation of ESA was beneficial to improvements in forecast precision.

Alternatively, within the additive runs, estimates of observation uncertainty (R) for SMAPHB, 1KM, and 3KM were based on reported values in the literature. However, these estimates of uncertainty were likely inaccurate for the purposes of this study. It is well known in the literature that RS soil moisture data products, like most RS data products, have poorly characterized uncertainties (Peng et al., 2021). For each data product, uncertainty was typically reported as a standard error value after comparing the data product to a limited set of observations. These values do not capture all possible sources of uncertainty and cannot be generalized across space and time (Huang et al., 2019); yet, in the additive analysis, they were applied uniformly across time and space and without consideration of their native spatiotemporal resolutions. Consequently, it is likely these estimates of R did not well represent the true variability in the observations. This is one of the main caveats of this study, as inaccurate estimates of uncertainty directly impact the computed analysis distribution with the EnKF and the GEF and, therefore, impact all downstream forecast distributions. For example, in the case of SMAPHB, the reported RMSE of 0.07 mm^3/mm^3 for the data product served to generate an observed distribution that spanned most of the soil moisture value range and was, thereby, almost uninformative for downstream model estimates (Fig 4.1., Fig. 4.4). The RS data product likely contributed more information than this distribution would suggest, and, thus, poor estimates of observation uncertainty for SMAPHB, 1KM, and 3KM likely contributed to the limited impact of adding these observations in the additive runs. Their added information could have been more effectively leveraged by providing better approximations of uncertainty for each observation, a task that could be performed by the Miyoshi algorithm. In the individual runs, the Miyoshi algorithm improved, varied, and, often, lowered estimates of observation uncertainty over time for each data product, mitigating the impact of misrepresented uncertainties. Future applications of the GEF scheme could benefit from additional terms in the model that could capture R or the use of the Miyoshi algorithm.

When selecting a RS soil moisture data product for data assimilation applications, the results of this study indicate that temporal resolution and accurately estimated observation uncertainty are two critical components to consider for optimal system performance. They also provide evidence that increasing the number of sensors in the derivation step or combining several data products can help to reduce uncertainty in soil moisture estimates. However, further investigation is needed to independently test the impact of observation sample size (i.e., number of data products), temporal resolution, spatial resolution, and uncertainty on system performance. Furthermore, the data products considered in this study do not represent the full range of RS soil moisture data products that are available publicly. This work should be expanded to evaluate data products derived from other satellites/derivations both individually and in combination with other sources to exhaust all available options.

CHAPTER 5

MAJOR CONCLUSIONS

In the face of pressing, large-scale agricultural issues, there is great need for accurate and precise agricultural forecasting methods that are scalable, flexible, robust, consistent, and comprehensive. However, most current forecasting methods fall short. To help fill this gap, this project completed three important objectives. First, a crop model data-assimilation system was developed that estimates uncertainty matrices, propagates a range of system uncertainties, allows for easy customization, and performs state-parameter assimilation. Second, the developed system was tested for a range of experimental sites across the U.S. Midwest where *in situ* soil moisture observations at the 10 and 20 cm depths were assimilated into the APSIM crop model. The system's constraint of several agricultural processes, including soil water dynamics, nitrate leaching, and crop development, was evaluated using available *in situ* observations, and the strengths and weaknesses of the developed system were investigated. Finally, to explore the generalizability of the system, 4 different RS surface soil moisture data products were assimilated within the developed data-assimilation system across the same experimental sites. System forecasts of soil water dynamics and crop development were again evaluated using *in situ* observations, and the impact of temporal resolution, spatial resolution, satellite sources, and estimated observation uncertainty for system performance were evaluated. The main findings of this project served to strengthen the results of past work and to provide new approaches and insights for the continued application and innovation of soil moisture data assimilation in agricultural forecasting.

As the first critical contribution of this project, the data-assimilation system developed in this work is unmatched in the literature. The established framework systematically merges observed data with a crop model using two different filtering approaches (i.e., EnFK and GEF) and can be easily adapted to assimilate a diverse range of observations at local to regional scales. By including the Miyoshi algorithm with the EnKF, the system leverages well-established tools from other forecasting disciplines to accurately estimate system error matrices, which are imperative for system performance but are difficult to define in practice. The GEF, on the other hand, allows for flexibility in both system configuration and model definition. Lastly, the system can efficiently propagate an array of uncertainties, such as those associated with climate, management, cultivar, soil, and initial conditions, to better account for variability in system components and improve the precision and accuracy of forecasts. In a single-site case study, the system was found to effectively constrain soil moisture, maintain high filter performance, and dynamically estimate soil properties in time using state-parameter assimilation. In fact, in broader applications, the system was able to mitigate the impact of soil moisture data quality on system performance by accurately estimating observation uncertainty through the Miyoshi algorithm and by leveraging the power of the EnKF to reduce overall uncertainty in the posterior distribution. Altogether, this system stands apart from previous assimilation efforts in crop modeling for its generalizability, its careful treatment of uncertainty, and its ability to accommodate a range of constraints and scales. Although there is still progress to be made, this innovative system can serve as a promising benchmark for further data-assimilation applications in crop modeling.

Another notable feature of the presented system is generalizability. Compared to traditional crop modeling studies (e.g., Malone et al., 2007; Ojeda et al., 2018) and other data-assimilation studies (e.g., Chakrabati et al., 2014; Liu et al., 2021), the approach for this work focused on the application of methods and data products that could be easily replicated/extracted for new sites or regions. The system employed an untouched APSIM model parameterization, gridded soil and climate drivers, globally-optimized cultivar parameter distributions, and a systematic method for estimating system error matrices. The use of these tools ensures a more realistic evaluation of soil moisture data assimilation for regional applications and will facilitate easier and faster future applications of the presented system for new locations. Even more importantly, the assimilation of soil moisture partially served to replace the role of site-level model calibration to improve site-level APSIM simulations, reducing the time required for model set-up and the possibility of overfitting the model. This difference in forecasting philosophy is important to consider when comparing the accuracy metrics reported in this study to the results of past crop modeling works (e.g., Malone et al., 2007; Martinez-Feria et al., 2019).

The last major novelty of this project is its multidimensional approach to system evaluation. To increase knowledge on the functionality of APSIM and of the system, the constraint of 5 different state variables, including yield, vegetative cover (i.e., NDVI and LAI), soil moisture, tile drainage, and annual NO_3 load, was evaluated against *in situ* observations for 19 site-years. In the past, many data-assimilation studies have focused their evaluation efforts exclusively on crop yield estimates (e.g., Launay and Guerif, 2005; Zhao et al., 2013; Jiang et al., 2014). However, this approach fails to leverage the full benefits of process-based crop models. Through its defined relationships in the model, the direct constraint of a state variable in a process-based crop model will impact, to some extent, all downstream model estimates that stem from the assimilated state variable. For example, by constraining soil moisture in the APSIM model in this study, the simulation of crop development and the soil water and N cycles, which are dependent on soil moisture estimates, typically diverged from the free model estimates. By investigating these changes, one can reveal important insights on the function of the assimilation system and the model processes. This study employed all available observations for the study sites to critically assess the value and implications of these downstream changes for future applications. In some cases, the changes indicated strong downstream constraint of model variables, and, in other cases, the changes were indicative of model biases.

One highlighted advantage of soil moisture data assimilation in this study is its capacity for improving model estimates in water-limited conditions. When assimilating *in situ* and RS soil moisture observations, the system demonstrated the ability to increase and improve crop yield estimates by reducing crop water stress and increasing crop water uptake at critical points in the growing season. The system was less effective in constraining yield when soil water availability was not a limiting factor for growth. This finding reflects the results of several past crop modeling studies (e.g., Chakrabati et al., 2014; Lu et al., 2021) and points to the utility of soil moisture SDA for constraining a crop model's understanding of soil-water-plant dynamics in drier years. Furthermore, *in situ* soil moisture data assimilation also demonstrated conditional constraint of annual tile drainage when experiencing drought conditions. Although more observations are needed on soil water cycle components to clearly identify the conditions where the system is valuable for drainage estimates, the constraint of drainage seems to be stronger when the model underestimates crop water uptake. This finding points to new ways that the presented data-assimilation system can be leveraged to improve agricultural forecasting more broadly. The constraint of tile drainage estimates with soil moisture data assimilation is a topic that has not been well investigated in the literature.

Next to improved simulation in drought years, soil moisture innovations from both the *in situ* and RS system applications point to another advantage of soil moisture data assimilation: the correction of precipitation input errors. Like many studies before (e.g., Ines et al., 2013; Lu and Steele-Dunne, 2019; Nair and Indu, 2019), the findings of this study reaffirm the capacity of soil moisture data assimilation to "localize" gridded weather estimates of precipitation to more accurately reflect observed values. Since cropping systems are highly sensitive to precipitation inputs and precipitation can be highly variable in space and time (Thaler et al., 2018), this is a strong advantage of soil moisture data assimilation for improving regional agricultural forecasts. Given high spatial resolution soil moisture data, the presented system can account for input error at fine spatial scales and improve model simulations when coarse resolution gridded weather datasets are applied.

In addition to highlighting opportunities, the comprehensive evaluation of the system in this study also identified several areas for improvement in both the APSIM model and the data-assimilation framework. First, the highlighted trend between weather conditions and soil water innovations indicates biases in the model's soil water processes. The model made larger errors in estimating surface and root-zone soil moisture after large precipitation events. This could indicate a number of missing or ill-defined processes in APSIM's SoilWat module related to vertical soil water flow, infiltration, and evaporation, among others. Second, the system's poor constraint of annual NO_3 leaching estimates demonstrates the need for more intensive evaluation of APSIM's SoilN module. In this study, soil moisture data assimilation consistently improved soil moisture forecasts throughout the soil profile, which, theoretically, should have improved the simulation of soil N processes via improved estimates of the soil moisture rate factors. The lack of constraint of NO_3 leaching estimates in this study highlights structural and/or parameter errors in APSIM's soil N processes. Lastly, the current method of assimilation adjustment in the system invalidates the model's water mass balance, removing or creating soil water at every time step. This approach can be beneficial when system inputs are biased (e.g., too much or too little precipitation), but, in any other case, these adjustments can lead to large errors in other model estimates. All three of these APSIM shortcomings need to be investigated and improved with further observation constraints.

The application of remote sensing soil moisture data products in the presented data-assimilation system could be a promising approach to improve regional agricultural forecasting capacity. However, this study highlighted important caveats for such applications. First, the assimilation of surface soil moisture is not as powerful as the assimilation of root-zone soil moisture values in terms of model constraint as it represents a smaller proportion of the soil profile and is not as closely related to other important state variables (De Lannoy et al., 2007). The PDA framework could be leveraged to overcome this limitation (Ines et al., 2013). Second, high temporal resolution was far more important to system performance than high spatial resolution as it helped to limit the magnitude of model errors. To reduce the assimilation interval for soil moisture data assimilation purposes, observations from several RS data products can be combined and assimilated after their individual values have been evaluated. Finally, accurate estimates of observation uncertainty are imperative for optimal system performance. Systematic approaches, like the Miyoshi algorithm or an adjusted GEF, can and should be used to ensure system uncertainties are well represented in the data-assimilation system.

REFERENCES

Abendroth, L. J., Herzmann, D. E., Chighladze, G., Kladivko, E. J., Helmers, M. J., Bowling, L., Castellano, M., Cruse, R. M., Dick, W. A., Fausey, N. R., Frankenberger, J., Gassmann, A. J., Kravchenko, A., Lal, R., Lauer, J. G., Mueller, D. S., Nafziger, E. D., Nkongolo, N., O'Neal, M., Sawyer, J. E., Scharf, P., Strock, J. S., & Villamil, M. B. (2017). Sustainable Corn CAP Research Data (USDA-NIFA Award No. 2011-68002-30190). National Agricultural Library - ARS - USDA. https://dx.doi.org/10.15482/USDA.ADC/1411953

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. *Agronomy Journal*, *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

Basak, A., Mengshoel, O. J., Kulkarni, C., Schmidt, K., Shastry, P., & Rapeta, R. (2017). Optimizing the decomposition of time series using evolutionary algorithms: Soil moisture analytics. *Proceedings of the Genetic and Evolutionary Computation Conference*, 1073–1080. https://doi.org/10.1145/3071178.3071191

Beck, H. E., Pan, M., Miralles, D. G., Reichle, R. H., Dorigo, W. A., Hahn, S., Sheffield, J., Karthikeyan, L., Balsamo, G., Parinussa, R. M., van Dijk, A. I. J. M., Du, J., Kimball, J. S., Vergopolan, N., & Wood, E. F. (2020). *Evaluation of 18 satellite- and model-based soil moisture products using in situ measurements from 826 sensors* [Preprint]. Global hydrology/Modelling approaches. https://doi.org/10.5194/hess-2020-184

Bernacchi, C. (2020). [Data on leaf area index for Energy Farm in 2018-2019] [Unpublished raw data]. University of Illinois.

Bijay-Singh, and 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

Birch, C. J., Hammer, G. L., & Rickert, K. G. (1998). Improved methods for predicting individual leaf area and leaf senescence in maize (*Zea mays*). Australian Journal of Agricultural Research 49: 249-62.

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. *Agronomy Journal*, *88*(5), 704–716. https://doi.org/10.2134/agronj1996.00021962008800050005x

Brilli, L., Bechini, L., Bindi, M., Carozzi, M., Cavalli, D., Conant, R., Dorich, C. D., Doro, L., Ehrhardt, F., Farina, R., Ferrise, R., Fitton, N., Francaviglia, R., Grace, P., Iocola, I., Klumpp, K., Léonard, J., Martin, R., Massad, R. S., … Bellocchi, G. (2017). Review and analysis of strengths and weaknesses of agro-ecosystem models for simulating C and N fluxes. *Science of The Total Environment*, *598*, 445–470. https://doi.org/10.1016/j.scitotenv.2017.03.208

Brummer, E. C. (1998). Diversity, Stability, and Sustainable American Agriculture. *Agronomy Journal*, *90*(1), 1–2. https://doi.org/10.2134/agronj1998.00021962009000010001x

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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *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. *European Journal of Agronomy*, *101*, 163–173. https://doi.org/10.1016/j.eja.2018.09.006

Chighladze, G., Abendroth, L. J., Herzmann, D., Helmers, M., Ahiablame, L., Allred, B., Bowling, L., Brown, L., Fausey, N., Frankenberger, J., Jaynes, D., Jia, X., Kjaersgaard, J., King, K., Kladivko, E., Nelson, K., Pease, L., Reinhart, B., Strock, J., & Youssef, M. (2021). Transforming Drainage Research Data (USDA-NIFA Award No. 2015-68007-23193). National Agricultural Library - ARS - USDA. https://doi.org/10.15482/USDA.ADC/1521092**.**

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. *Journal of Environmental Management*, *206*, 1072–1080. https://doi.org/10.1016/j.jenvman.2017.11.051

Curnel, Y., de Wit, A. J. W., Duveiller, G., & Defourny, P. (2011). Potential performances of remotely sensed LAI assimilation in WOFOST model based on an OSS Experiment. *Agricultural and Forest Meteorology*, *151*(12), 1843–1855. https://doi.org/10.1016/j.agrformet.2011.08.002

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*.

Dente, L., Satalino, G., Mattia, F., & Rinaldi, M. (2008). Assimilation of leaf area index derived from ASAR and MERIS data into CERES-Wheat model to map wheat yield. *Remote Sensing of Environment*, *112*(4), 1395–1407. https://doi.org/10.1016/j.rse.2007.05.023

Desroziers, G., Berre, L., Chapnik, B., & Poli, P. (2006). Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, *131*(613), 3385–3396. https://doi.org/10.1256/qj.05.108

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

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

Dietzel, R., Liebman, M., Ewing, R., Helmers, M., Horton, R., Jarchow, M., & Archontoulis, S. (2016). How efficiently do corn- and soybean-based cropping systems use water? A systems modeling analysis. *Global Change Biology*, *22*(2), 666–681. https://doi.org/10.1111/gcb.13101

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. *Agronomy Journal, 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. *Agricultural Water Management*, *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. *Environmental Research Letters*, *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

Dokoohaki H., Rai, T. S., & Kivi, M.[in prep]. Linking remote sensing and machine learning with APSIM model through Bayesian emulation and optimization to improve yield predictions in the U.S. Midwest.

Dorigo, W. A., Zurita-Milla, R., de Wit, A. J. W., Brazile, J., Singh, R., & Schaepman, M. E. (2007). A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. *International Journal of Applied Earth Observation and Geoinformation*, *9*(2), 165–193. https://doi.org/10.1016/j.jag.2006.05.003

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P. D., Hirschi, M., Ikonen, J., de Jeu, R., Kidd, R., Lahoz, W., Liu, Y. Y., Miralles, D., … Lecomte, P. (2017). ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions. *Remote Sensing of Environment*, *203*, 185–215. https://doi.org/10.1016/j.rse.2017.07.001

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). *Environmental Modelling and Software*, 62, 509–51.

Engman, E. T. (1991). Applications of microwave remote sensing of soil moisture for water resources and agriculture. *Remote Sensing of Environment*, *35*, 213–226. https://doi.org/10.1016/0034-4257(91)90013-V

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 Systems*, *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. *International Journal of Remote Sensing*, *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. *Global Change Biology*, *27*(1), 13–26. https://doi.org/10.1111/gcb.15409

Fisher, J. B., Lee, B., Purdy, A. J., Halverson, G. H., Dohlen, M. B., Cawse-Nicholson, K., Wang, A., Anderson, R. G., Aragon, B., Arain, M. A., Baldocchi, D. D., Baker, J. M., Barral, H., Bernacchi, C. J., Bernhofer, C., Biraud, S. C., Bohrer, G., Brunsell, N., Cappelaere, B., … Hook, S. (2020). ECOSTRESS: NASA's Next Generation Mission to Measure Evapotranspiration From the International Space Station. *Water Resources Research*, *56*(4). https://doi.org/10.1029/2019WR026058

Flathers, E., and Gessler, P. E. (2018). Building an Open Science Framework to Model Soil Organic Carbon. *Journal of Environmental Quality*, *47*(4), 726–734. https://doi.org/10.2134/jeq2017.08.0318

Gao, F., and Zhang, X. (2021). Mapping Crop Phenology in Near Real-Time Using Satellite Remote Sensing: Challenges and Opportunities. *Journal of Remote Sensing*, *2021*, 1–14. https://doi.org/10.34133/2021/8379391

Guerif, M., and Duke, C. L. (2000). Adjustment procedures of a crop model to the site specific characteristics of soil and crop using remote sensing data assimilation. *Agriculture, Ecosystems & Environment*, *81*(1), 57–69. https://doi.org/10.1016/S0167-8809(00)00168-7

Gurevich, H., Baram, S., & Harter, T. (2021). Measuring nitrate leaching across the critical zone at the field to farm scale. *Vadose Zone Journal*, *20*(2). https://doi.org/10.1002/vzj2.20094

Hachiya, T., and Sakakibara, H. (2016). Interactions between nitrate and ammonium in their uptake, allocation, assimilation, and signaling in plants. *Journal of Experimental Botany*, erw449. https://doi.org/10.1093/jxb/erw449

Hansen, S., Thorsen, M., Pebesma, E. J., Kleeschulte, S., & Svendsen, H. (2006). Uncertainty in simulated nitrate leaching due to uncertainty in input data. A case study. *Soil Use and Management*, *15*(3), 167–175. https://doi.org/10.1111/j.1475-2743.1999.tb00083.x

Helmers, M. J., Abendroth, L., Reinhart, B., Chighladze, G., Pease, L., Bowling, L., Youssef, M., Ghane, E., Ahiablame, L., Brown, L., Fausey, N., Frankenberger, J., Jaynes, D., King, K., Kladivko, E., Nelson, K., & Strock, J. (2022). Impact of controlled drainage on subsurface drain flow and nitrate load: A synthesis of studies across the U.S. Midwest and Southeast. *Agricultural Water Management*, *259*, 107265. https://doi.org/10.1016/j.agwat.2021.107265

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. *Quarterly Journal of the Royal Meteorological Society*, *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. *Journal of Atmospheric and Oceanic Technology, 30*, 8.

Hoffman, A. L., Kemanian, A. R., & Forest, C. E. (2020). The response of maize, sorghum, and soybean yield to growing-phase climate revealed with machine learning. *Environmental Research Letters*, *15*(9), 094013. https://doi.org/10.1088/1748-9326/ab7b22

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. *Environmental Modelling & Software*, *62*, 327–350. https://doi.org/10.1016/j.envsoft.2014.07.009

Hu, K., Li, B., Chen, D., Zhang, Y., & Edis, R. (2008). Simulation of nitrate leaching under irrigated maize on sandy soil in desert oasis in Inner Mongolia, China. *Agricultural Water Management*, *95*(10), 1180–1188. https://doi.org/10.1016/j.agwat.2008.05.001

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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *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. *Agricultural and Forest Meteorology*, *276–277*, 107609. https://doi.org/10.1016/j.agrformet.2019.06.008

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. *Journal of Hydrology*, *555*, 812–831. https://doi.org/10.1016/j.jhydrol.2017.10.061

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 Sensing of Environment*, *138*, 149–164. https://doi.org/10.1016/j.rse.2013.07.018

Jiang, Z., Chen, Z., Chen, J., Liu, J., Ren, J., Li, Z., Sun, L., & Li, H. (2014). Application of Crop Model Data Assimilation With a Particle Filter for Estimating Regional Winter Wheat Yields. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(11), 4422–4431. https://doi.org/10.1109/JSTARS.2014.2316012

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. In *European Journal of Agronomy* (Vol. 92, pp. 141–152). Elsevier B.V. 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. *European Journal of Agronomy*, *18*(3–4), 267–288. https://doi.org/10.1016/S1161-0301(02)00108-9

Keenan, T. F., Carbone, M. S., Reichstein, M., & Richardson, A. D. (2011). The model–data fusion pitfall: Assuming certainty in an uncertain world. *Oecologia*, *167*(3), 587–597. https://doi.org/10.1007/s00442-011-2106-x

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 Resources Research*, *43*(6). https://doi.org/10.1029/2006WR005100

Launay, M., and Guerif, M. (2005). Assimilating remote sensing data into a crop model to improve predictive performance for spatial applications. *Agriculture, Ecosystems & Environment*, *111*(1–4), 321–339. https://doi.org/10.1016/j.agee.2005.06.005

Li, H., Wang, L., Qiu, J., Li, C., Gao, M., & Gao, C. (2014). Calibration of DNDC model for nitrate leaching from an intensively cultivated region of Northern China. *Geoderma*, *223–225*, 108–118. https://doi.org/10.1016/j.geoderma.2014.01.002

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. *Agriculture, Ecosystems & Environment*, *294*, 106861. https://doi.org/10.1016/j.agee.2020.106861

Liang, X. Q., Harter, T., Porta, L., van Kessel, C., & Linquist, B. A. (2014). Nitrate Leaching in Californian Rice Fields: A Field- and Regional-Scale Assessment. *Journal of Environmental Quality*, *43*(3), 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. *Computers and Electronics in Agriculture*, *142*, 201–210. https://doi.org/10.1016/j.compag.2017.09.010

Lievens, H., Reichle, R. H., Liu, Q., De Lannoy, G. J. M., Dunbar, R. S., Kim, S. B., Das, N. N., Cosh, M., Walker, J. P., & Wagner, W. (2017). Joint Sentinel-1 and SMAP data assimilation to improve soil moisture estimates. *Geophysical Research Letters*, *44*(12), 6145–6153. https://doi.org/10.1002/2017GL073904

Linker, R., and Ioslovich, I. (2017). Assimilation of canopy cover and biomass measurements in the crop model AquaCrop. *Biosystems Engineering*, *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. *Journal of Hydrology and Hydromechanics*, *65*(2), 123–133. https://doi.org/10.1515/johh-2017-0011

Liu, Y., Wang, W., & Liu, Y. (2018). ESA CCI Soil Moisture Assimilation in SWAT for Improved Hydrological Simulation in Upper Huai River Basin. *Advances in Meteorology*, *2018*, 1–13. https://doi.org/10.1155/2018/7301314

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 Environmental Research and Risk Assessment*, *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. *Advances in Water Resources*, *34*(3), 395–406. https://doi.org/10.1016/j.advwatres.2010.12.005

Lu, Y., Dong, J., and Steele-Dunne, S. C. (2019). Impact of Soil Moisture Data Resolution on Soil Moisture and Surface Heat Flux Estimates through Data Assimilation: A Case Study in the Southern Great Plains. *Journal of Hydrometeorology*, *20*(4), 715–730. https://doi.org/10.1175/JHM-D-18-0234.1

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. *Agricultural Water Management*, *252*. https://doi.org/10.1016/j.agwat.2021.106884

Luce, G. A. "Optimum corn planting depth – 'Don't plant your corn too shallow.'" *University of Missouri Integrated Pest & Crop Management.* 6 Apr. 2016.

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. *Mathematical and Computer Modelling*, *58*(3–4), 634–643. https://doi.org/10.1016/j.mcm.2011.10.038

Machwitz, M., Giustarini, L., Bossung, C., Frantz, D., Schlerf, M., Lilienthal, H., Wandera, L., Matgen, P., Hoffmann, L., & Udelhoven, T. (2014). Enhanced biomass prediction by assimilating satellite data into a crop growth model. *Environmental Modelling & Software*, *62*, 437–453. https://doi.org/10.1016/j.envsoft.2014.08.010

Malone, R. W., Huth, N., Carberry, P. S., Ma, L., Kaspar, T. C., Karlen, D. L., Meade, T., Kanwar, R. S., & Heilman, P. (2007). Evaluating and predicting agricultural management effects under tile drainage using modified APSIM. *Geoderma*, *140*(3), 310–322. https://doi.org/10.1016/j.geoderma.2007.04.014

Marj, A. F., and Meijerink, A. M. J. (2011). Agricultural drought forecasting using satellite images, climate indices and artificial neural network. *International Journal of Remote Sensing*, *32*(24), 9707–9719. https://doi.org/10.1080/01431161.2011.575896

Martinez-Feria, R., Nichols, V., Basso, B., & Archontoulis, S. (2019). Can multi-strategy management stabilize nitrate leaching under increasing rainfall? *Environmental Research Letters*, *14*(12), 124079. https://doi.org/10.1088/1748-9326/ab5ca8

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. *European Journal of Agronomy*, *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 Problems in Science and Engineering*, *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. *Advances in Water Resources*, *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

Mourtzinis, S., & Conley, S. P. (2017). Delineating Soybean Maturity Groups across the United States. *Agronomy Journal*, *109*(4), 1397–1403. https://doi.org/10.2134/agronj2016.10.0581

Nair, A. S., & Indu, J. (2019). Improvement of land surface model simulations over India via data assimilation of satellite-based soil moisture products. *Journal of Hydrology*, *573*, 406–421. https://doi.org/10.1016/j.jhydrol.2019.03.088

Naz, B. S., Kurtz, W., Montzka, C., Sharples, W., Goergen, K., Keune, J., Gao, H., Springer, A., Hendricks Franssen, H.-J., & Kollet, S. (2019). Improving soil moisture and runoff simulations at 3 km over Europe using land surface data assimilation. *Hydrology and Earth System Sciences*, *23*(1), 277–301. https://doi.org/10.5194/hess-23-277-2019

Nearing, G. S., Crow, W. T., Thorp, K. R., Moran, M. S., Reichle, R. H., & Gupta, H. V. (2012). Assimilating remote sensing observations of leaf area index and soil moisture for wheat yield estimates: An observing system simulation experiment. *Water Resources Research*, *48*(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. *Agricultural Water Management*, *195*, 154–171. https://doi.org/10.1016/j.agwat.2017.10.010

Ouaadi, N., Jarlan, L., Khabba, S., Ezzahar, J., Le Page, M., & Merlin, O. (2021). Irrigation Amounts and Timing Retrieval through Data Assimilation of Surface Soil Moisture into the FAO-56 Approach in the South Mediterranean Region. *Remote Sensing*, *13*(14), 2667. https://doi.org/10.3390/rs13142667

Pablos, M., González-Zamora, Á., Sánchez, N., & Martínez-Fernández, J. (2018). Assessment of Root Zone Soil Moisture Estimations from SMAP, SMOS and MODIS Observations. *Remote Sensing*, *10*(7), 981. https://doi.org/10.3390/rs10070981

Park, S. K., and 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. *Environmental Research Letters*, *16*(6), 064063. https://doi.org/10.1088/1748-9326/abef8f

Pauwels, V. R. N., Verhoest, N. E. C., De Lannoy, G. J. M., Guissard, V., Lucau, C., & Defourny, P. (2007). Optimization of a coupled hydrology-crop growth model through the assimilation of observed soil moisture and leaf area index values using an ensemble Kalman filter: ASSIMILATION OF LAI AND SOIL MOISTURE. *Water Resources Research*, *43*(4). https://doi.org/10.1029/2006WR004942

de Paz, J. M., and Ramos, C. (2004). Simulation of nitrate leaching for different nitrogen fertilization rates in a region of Valencia (Spain) using a GIS–GLEAMS system. *Agriculture, Ecosystems & Environment*, *103*(1), 59–73. https://doi.org/10.1016/j.agee.2003.10.006

Peng, J., Albergel, C., Balenzano, A., Brocca, L., Cartus, O., Cosh, M. H., Crow, W. T., Dabrowska-Zielinska, K., Dadson, S., Davidson, M. W. J., de Rosnay, P., Dorigo, W., Gruber, A., Hagemann, S., Hirschi, M., Kerr, Y. H., Lovergine, F., Mahecha, M. D., Marzahn, P., … Loew, A. (2021). A roadmap for high-resolution satellite soil moisture applications – confronting product characteristics with user requirements. *Remote Sensing of Environment*, *252*, 112162. https://doi.org/10.1016/j.rse.2020.112162

Pradhan, P., Fischer, G., van Velthuizen, H., Reusser, D. E., & Kropp, J. P. (2015). Closing Yield Gaps: How Sustainable Can We Be? *PLOS ONE*, *10*(6), e0129487. https://doi.org/10.1371/journal.pone.0129487

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. *Frontiers in Plant Science*, *7*. https://doi.org/10.3389/fpls.2016.01630

Quine, T. A., and Zhang, Y. (2002). An investigation of spatial variation in soil erosion, soil properties, and crop production within an agricultural field in Devon, United Kingdom. *Journal of Soil and Water Conservation*, 11.

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. *Irrigation Science*, *37*(5), 561–581. https://doi.org/10.1007/s00271-019-00636-4

Reichle, R. H., Koster, R. D., Dong, J., & Berg, A. A. (2004). Global Soil Moisture from Satellite Observations, Land Surface Models, and Ground Data: Implications for Data Assimilation. *Journal of Hydrometeorology*, *5*(3), 430–442. https://doi.org/10.1175/1525-7541(2004)005<0430:GSMFSO>2.0.CO;2

Roelsma, J., and Hendriks, R. F. A. (2014). Comparative study of nitrate leaching models on a regional scale. *Science of The Total Environment*, *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? *European Journal of Agronomy*, *94*, 25–35. https://doi.org/10.1016/j.eja.2018.01.006

Sharp, J. M., Thomas, S. M., & Brown, H. E. (2011). A validation of APSIM nitrogen balance and leaching predictions. *Agronomy New Zealand*, 12.

Silva, J. V., and Giller, K. E. (2021). Grand challenges for the 21st century: What crop models can and can't (yet) do. In *Journal of Agricultural Science*. Cambridge University Press. https://doi.org/10.1017/S0021859621000150

Spijker, J., Fraters, D., & Vrijhoef, A. (2021). A machine learning based modelling framework to predict nitrate leaching from agricultural soils across the Netherlands. *Environmental Research Communications*, *3*(4), 045002. https://doi.org/10.1088/2515-7620/abf15f

Staton, M. "Pay close attention to soybean planting depth." *Michigan State University Extension*. 9 May 2012.

Stewart, L. K., Charlesworth, P. B., Bristow, K. L., & Thorburn, P. J. (2006). Estimating deep drainage and nitrate leaching from the root zone under sugarcane using APSIM-SWIM. *Agricultural Water Management*, *81*(3), 315–334. https://doi.org/10.1016/j.agwat.2005.05.002

Systems thinking, systems doing. Nat Food 1, 659 (2020). 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. *Monthly Weather Review*, *148*(10), 3973–3994. https://doi.org/10.1175/MWR-D-19-0240.1

Thaler, S., Brocca, L., Ciabatta, L., Eitzinger, J., Hahn, S., & Wagner, W. (2018). Effects of Different Spatial Precipitation Input Data on Crop Model Outputs under a Central European Climate. *Atmosphere*, *9*(8), 290. https://doi.org/10.3390/atmos9080290

Thorp, K. R., Hunsaker, D. J., & French, A. N. (2010). Assimilating Leaf Area Index Estimates from Remote Sensing into the Simulations of a Cropping Systems Model. *Transactions of the ASABE*, *53*(1), 251–262. https://doi.org/10.13031/2013.29490

de Valpine, P., Paciorek, C., Turek, D., Michaud, N., Anderson-Bergman, C., Obermeyer, F., Wehrhahn Cortes, C., Rodrìguez, A., Temple Lang, D., & Paganin, S. (2022). *NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling*. doi: 10.5281/zenodo.1211190, R package version 0.12.2, https://cran.r-project.org/package=nimble.

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., & Bodik, R. (2017). "Programming with models: writing statistical algorithms for general model structures with NIMBLE." *Journal of Computational and Graphical Statistics*, **26**, 403-413. doi: 10.1080/10618600.2016.1172487.

van der Laan, M., Annandale, J. G., Bristow, K. L., Stirzaker, R. J., Preez, C. C. du, & Thorburn, P. J. (2014). Modelling nitrogen leaching: Are we getting the right answer for the right reason? *Agricultural Water Management*, *133*, 74–80. https://doi.org/10.1016/j.agwat.2013.10.017

Verburg, K., and CSIRO Division of Soils (1996). *Methodology in soil-water-solute balance modelling: an evaluation of the APSIM-SoilWat and SWIMv2 models*. Division of Soils divisional report, no. 131.

Vergopolan, N., Chaney, N. W., Pan, M., Sheffield, J., Beck, H. E., Ferguson, C. R., Torres-Rojas, L., Sadri, S., & Wood, E. F. (2021). SMAP-HydroBlocks, a 30-m satellite-based soil moisture dataset for the conterminous US. *Scientific Data*, *8*(1), 264. https://doi.org/10.1038/s41597-021-01050-2

Wallach, D. (2011). Crop Model Calibration: A Statistical Perspective. Agronomy Journal, 103(4), 1144–1151. https://doi.org/10.2134/agronj2010.0432 Wang, X., and Bishop, C. H. (2003). A

Comparison of Breeding and Ensemble Transform Kalman Filter Ensemble Forecast Schemes. *American Meteorological Society 60, 1140-1158.*

"Water and Atmospheric Resources Monitoring Program. Illinois Climate Network. (2021). Illinois State Water Survey, 2204 Griffith Drive, Champaign, IL 61820-7495. http://dx.doi.org/10.13012/J8MW2F2Q.

Weiss, M., Jacob, F., & Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment*, *236*, 111402. https://doi.org/10.1016/j.rse.2019.111402

de Wit, A. J. W. and van Diepen, C. A. (2007). Crop model data assimilation with the Ensemble Kalman filter for improving regional crop yield forecasts. Agricultural and Forest Meteorology 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. *Advances in Meteorology*, *2016*. https://doi.org/10.1155/2016/4569218

Wu, S., Yang, P., Chen, Z., Ren, J., Li, H., & Sun, L. (2021). Estimating winter wheat yield by assimilation of remote sensing data with a four-dimensional variation algorithm considering anisotropic background error and time window. *Agricultural and Forest Meteorology*, *301–302*, 108345. https://doi.org/10.1016/j.agrformet.2021.108345

Zhao, Y., Chen, S., & Shen, S. (2013). Assimilating remote sensing information with crop model using Ensemble Kalman Filter for improving LAI monitoring and yield estimation. *Ecological Modelling*, *270*, 30–42. https://doi.org/10.1016/j.ecolmodel.2013.08.016

Zhou, H., Wu, J., Li, X., Geng, G., & Liu, L. (2016). Improving soil moisture estimation by assimilating remotely sensed data into crop growth model for agricultural drought monitoring. *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 4229–4232. https://doi.org/10.1109/IGARSS.2016.7730102

Zhu, P., Shi, L., Zhu, Y., Zhang, Q., Huang, K., & Williams, M. (2017). Data assimilation of soil water flow via ensemble Kalman filter: Infusing soil moisture data at different scales. *Journal of Hydrology*, *555*, 912–925. https://doi.org/10.1016/j.jhydrol.2017.10.078

APPENDIX A

SUPPLEMENTARY TABLES AND FIGURES

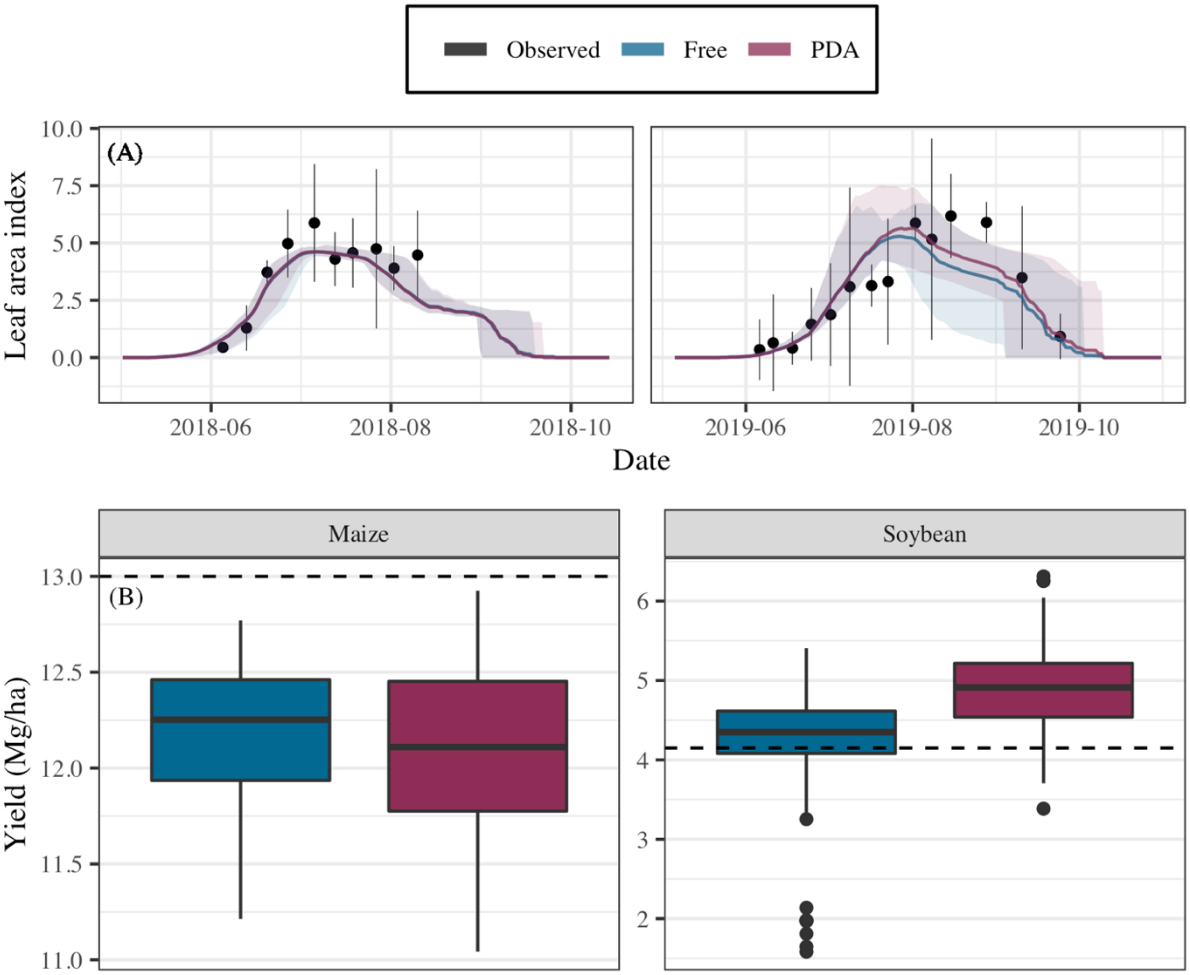

Figure A.1. (a) Time series of simulated and observed LAI estimates for study period. 95% credibility intervals are indicated by the shaded ribbon surrounding the mean lines for each simulation. Observed mean values are shown as points with bars demonstrating a 95% confidence interval (Student-t distribution, n = 3). (b) Boxplot summarizing the estimated distribution of yield for each scheme in both years. Dashed horizontal lines mark the observed yield value for both crops.

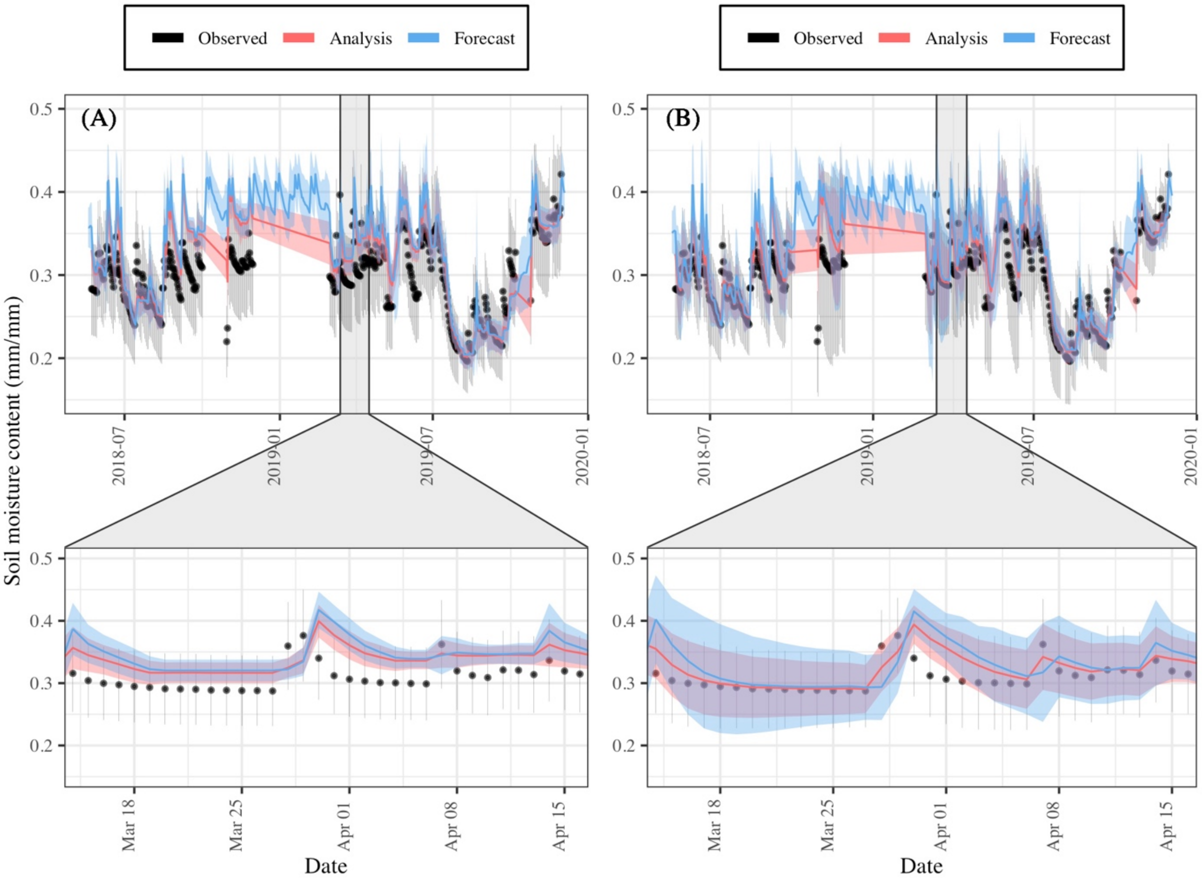

Figure A.2. Time series of SM4 estimates from the (a) SDA and (b) Miyoshi simulations with both the forecast and analysis distributions shown (95% confidence interval ribbon around the mean line). The observed means and 95% confidence intervals are shown with black points and lines, respectively. For both time series, the time period from 15 March 2019 to 15 April 2019 is highlighted to demonstrate a period of filter divergence in SDA (a) and its resolution in Miyoshi (b).

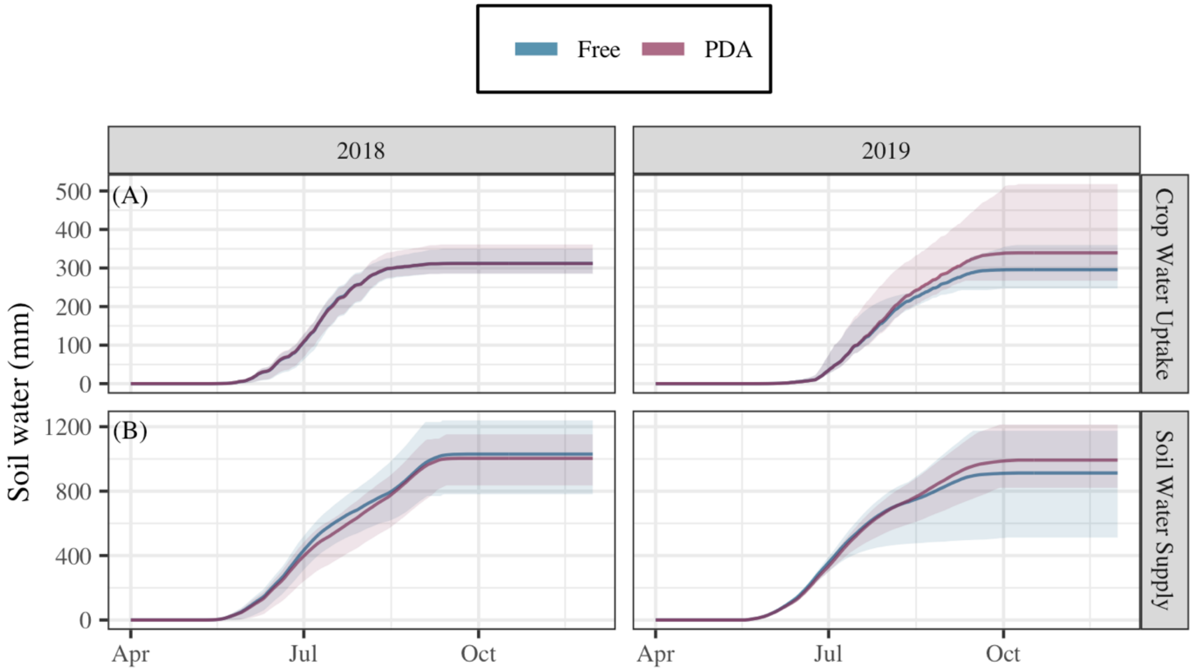

Figure A.3. Time series of simulated (a) cumulative crop water uptake (mm) and (b) cumulative soil water supply (mm) over the course of the study period at the Energy Farm. 95% credibility intervals are indicated by the shaded ribbon surrounding the mean lines for each simulation.

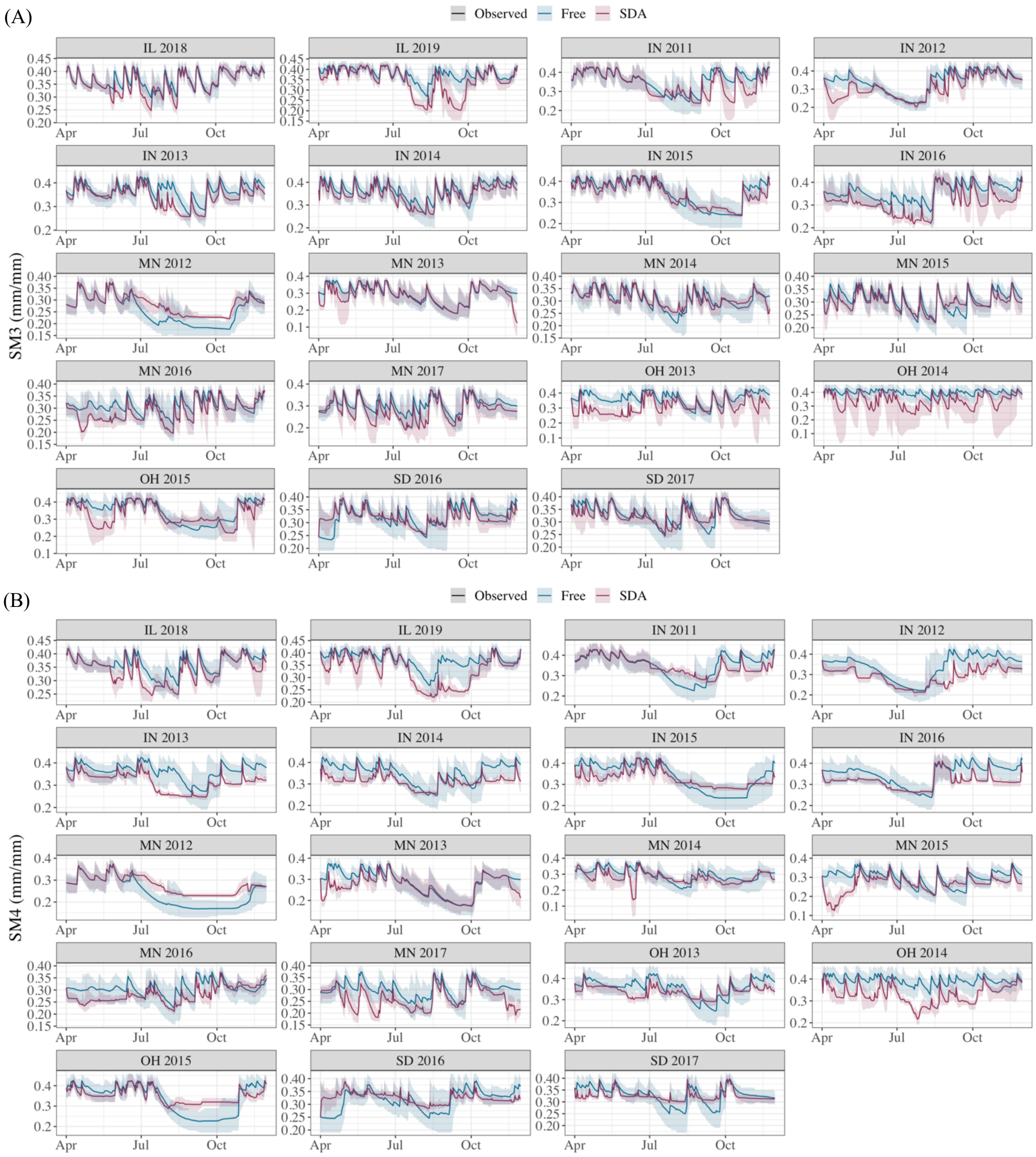

Figure A.4. Time series of soil moisture estimates at (a) 10 cm and (b) 20 cm from the two schemes with the mean daily estimates demonstrated with line graphs and the 95% credibility interval demonstrated by the shaded regions.

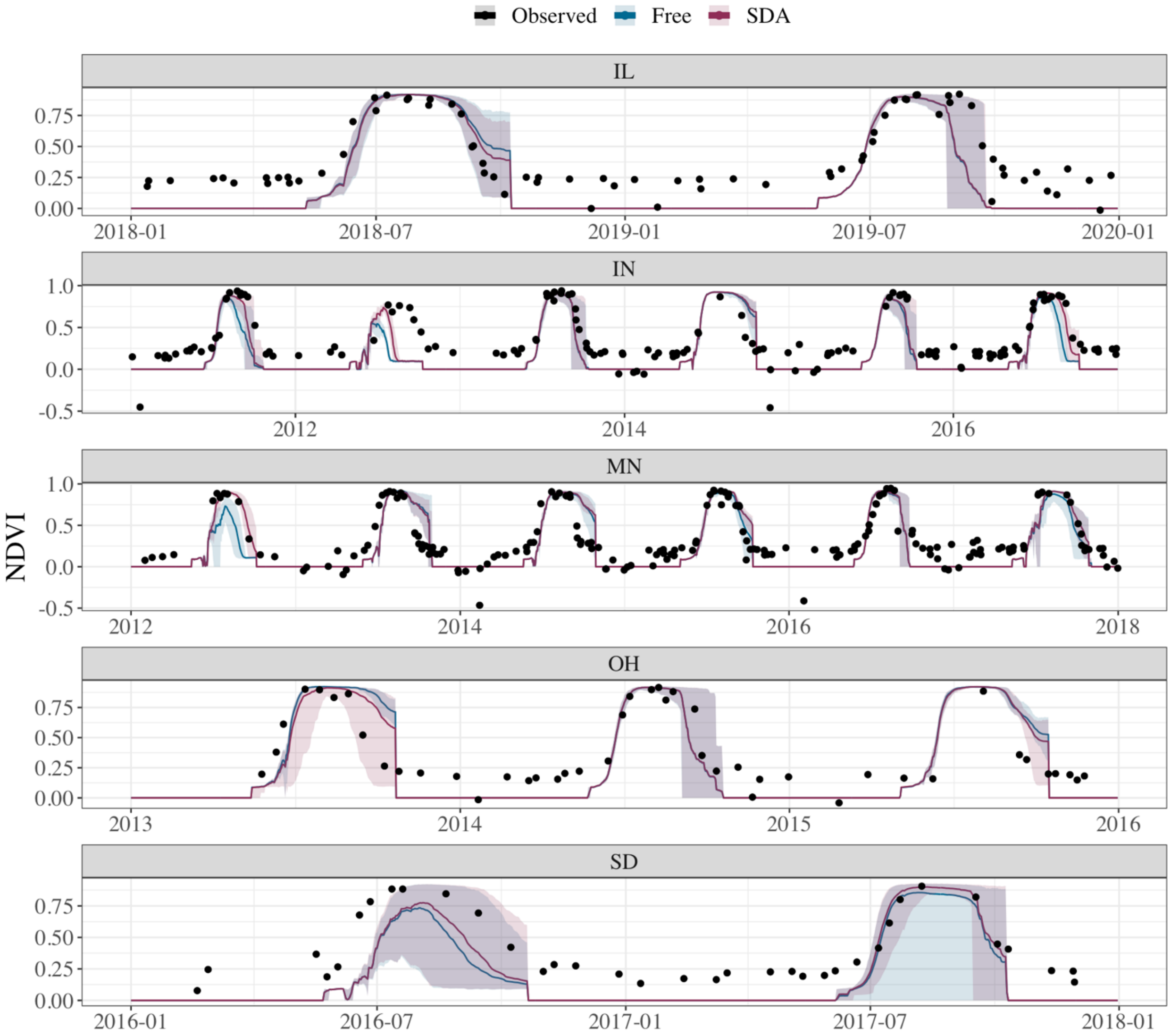

Figure A.5. Time series of NDVI estimates from the two schemes for each site-year with the mean daily estimates demonstrated with line graphs and the 95% credibility interval demonstrated by the shaded regions. Black points represent the observed values.

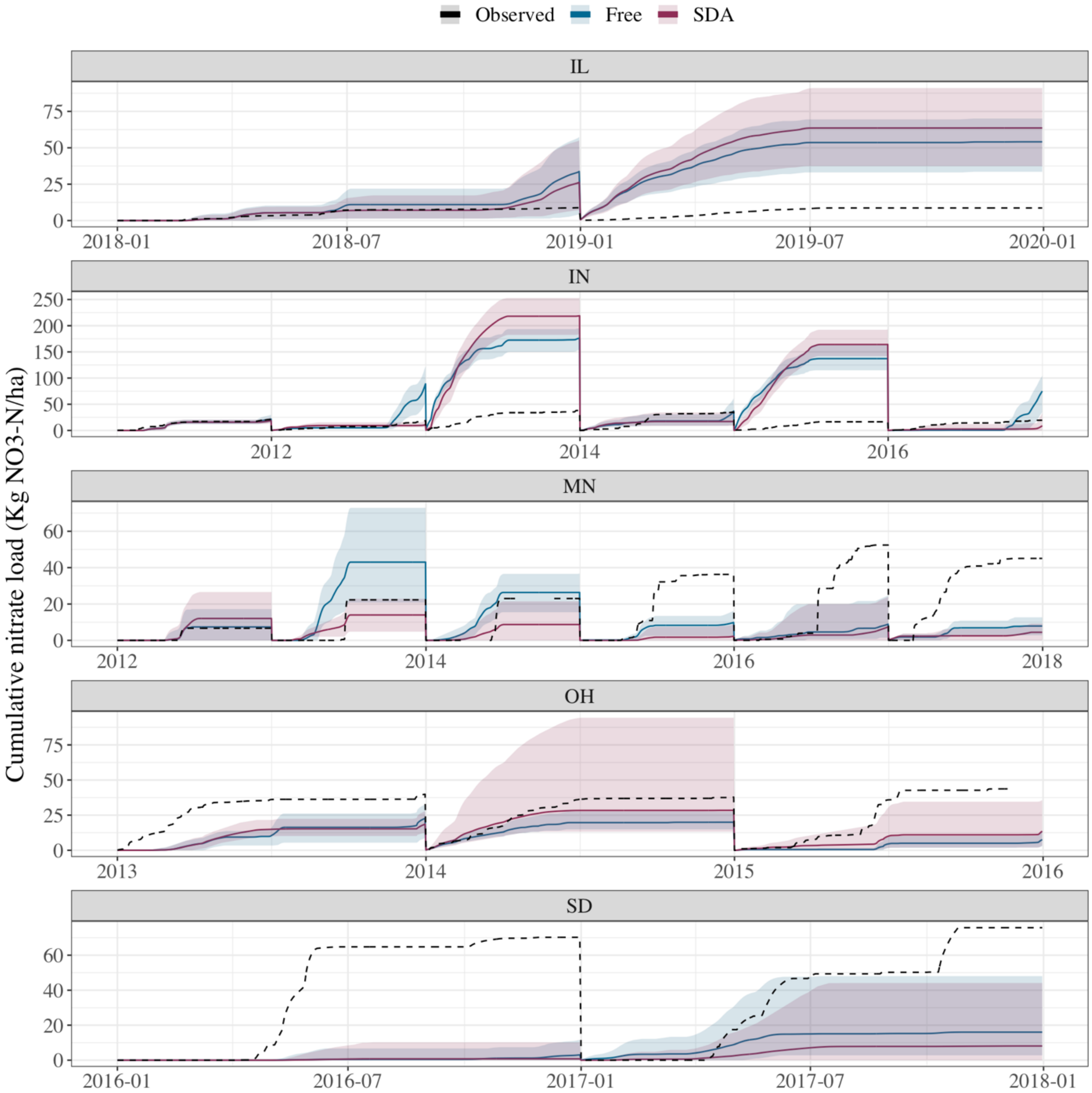

Figure A.6. Time series of cumulative NO_3 load estimates from the two schemes for each site-year with the mean daily estimates demonstrated with line graphs and the 95% credibility interval demonstrated by the shaded regions. Black dashed lines represent the observed cumulative value for each site-year.

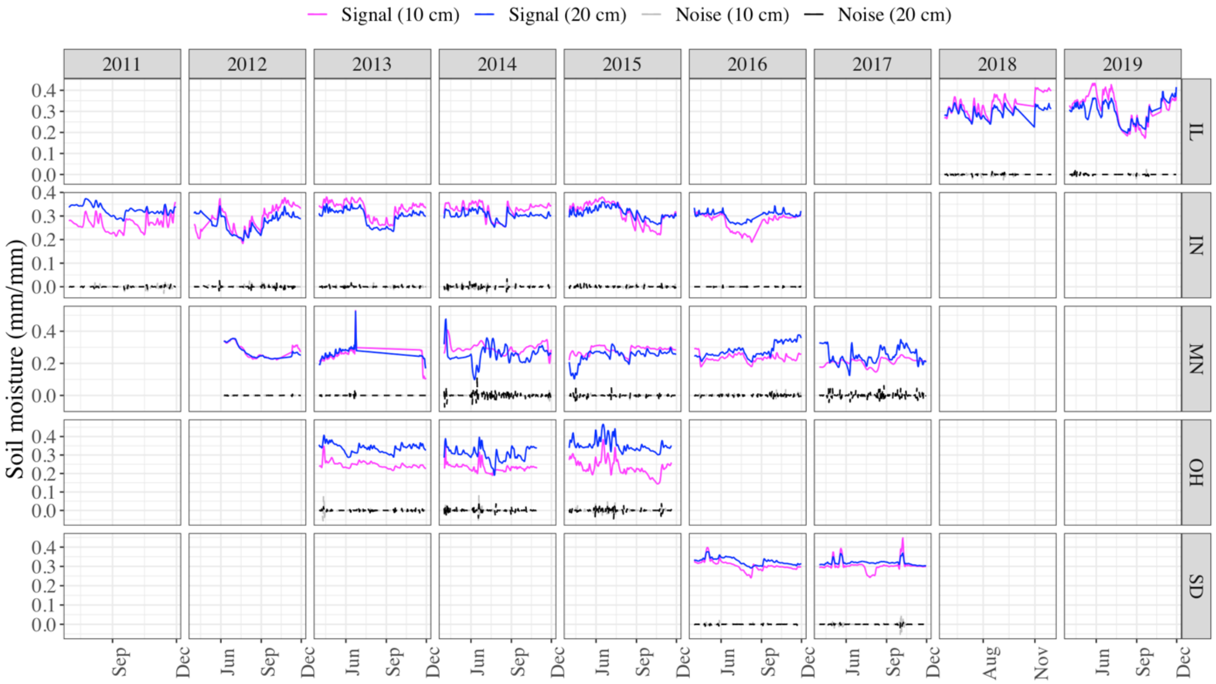

Figure A.7. Time series of the decomposed trend (i.e., fitted cubic spline) and noise (i.e., remainder) for each site-year soil moisture time series.

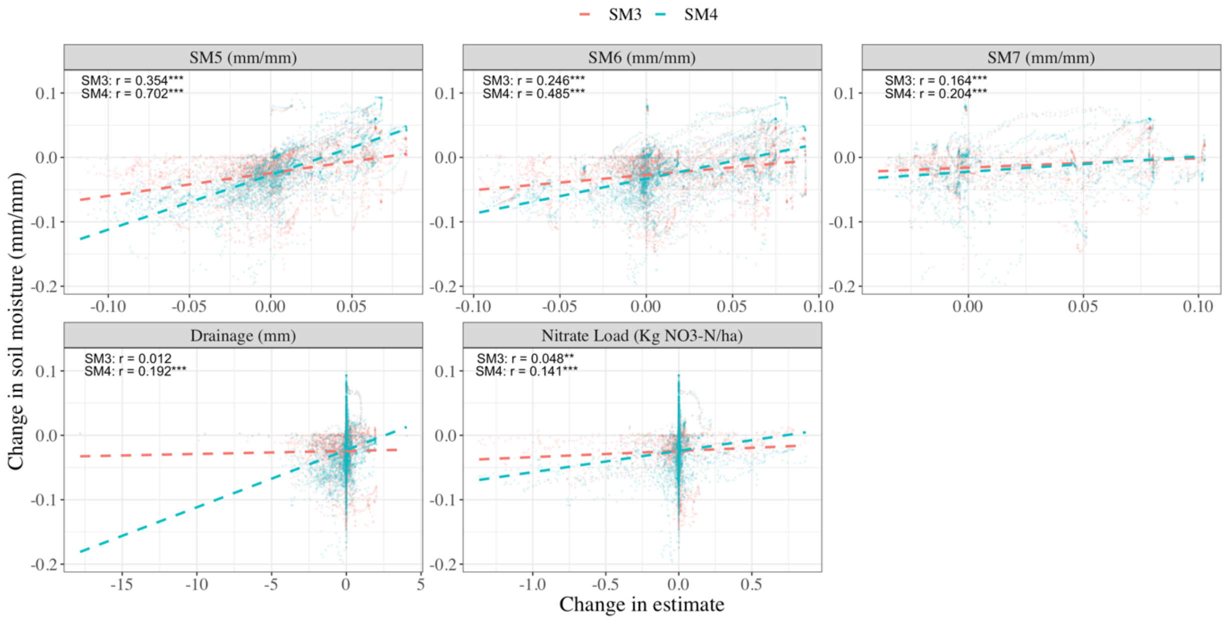

Figure A.8. Scatterplots comparing change (i.e., SDA – free model) in mean SM5, SM6, SM7, daily drainage, daily NO_3 leaching estimates with change in mean SM3 and SM4 estimates at each analysis time step. For each variable combination, the least squares regression line is demonstrated by a dashed line and the Pearson correlation coefficient is displayed. Asterisks denote significant coefficient values (** indicates p-value < 0.01 and *** indicates p-value ~ 0).

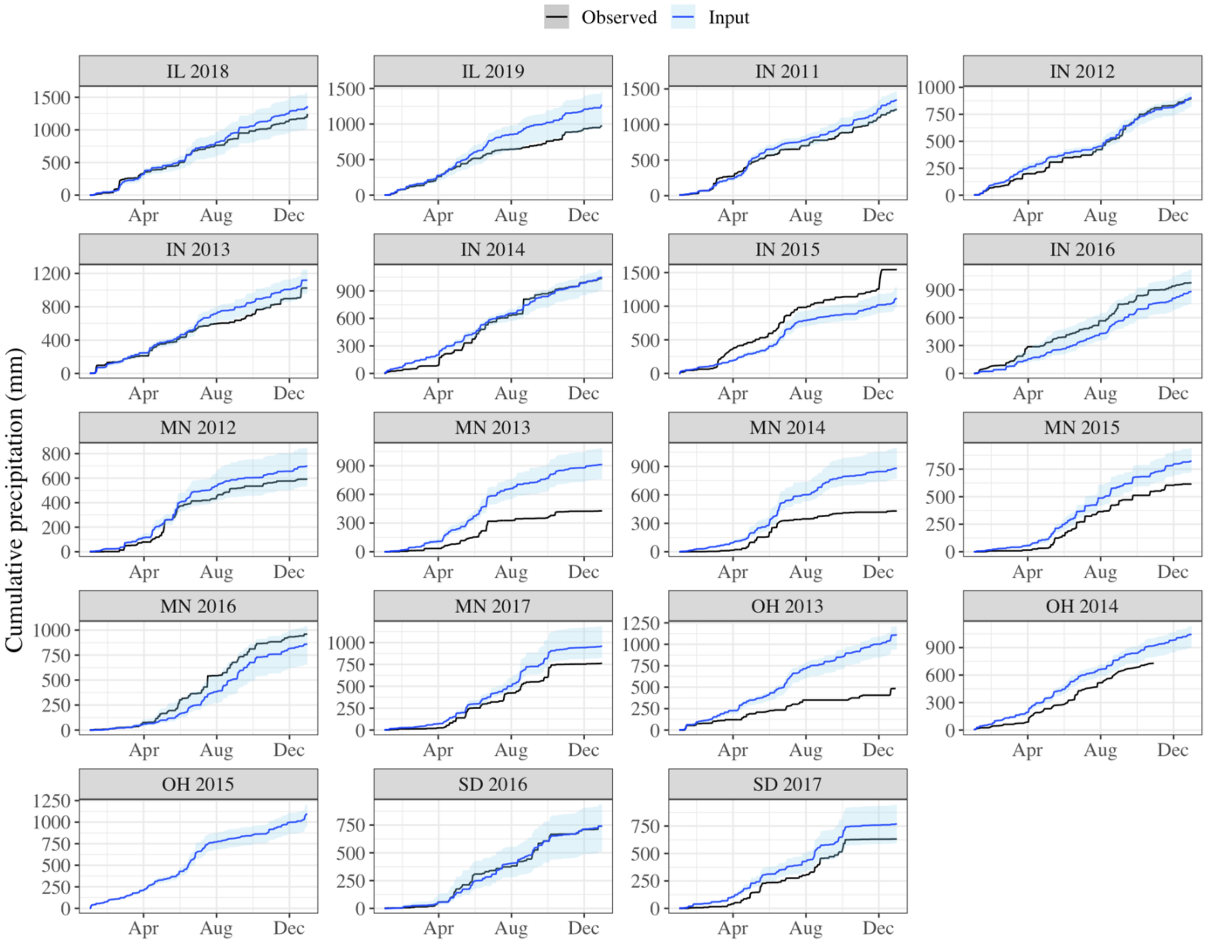

Figure A.9. Time series of cumulative precipitation (mm) for each site-year where the observed precipitation data (black line) is directly compared to the ensemble mean from the input weather ensembles (blue line) and the ensemble upper and lower limits (shaded blue region). Observed precipitation data were not available for the end of OH 2014, and no observations were available for OH 2015.

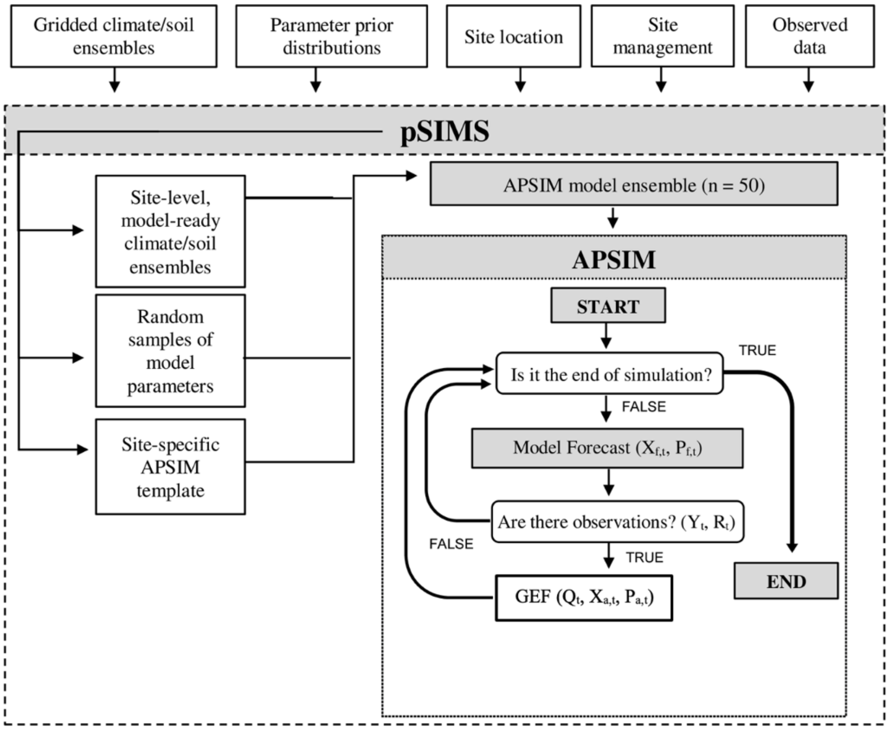
Gridded climate/soil ensembles
Parameter prior distributions
Site location
Site management
Observed data
pSIMS
Site-level, model-ready climate/soil ensembles
Random samples of model parameters
Site-specific APSIM template
APSIM model ensemble (n = 50)
APSIM
START
Is it the end of simulation?
TRUE
FALSE
Model Forecast ($X_{f,t}$, $P_{f,t}$)
Are there observations? (Y_t, R_t)
FALSE
TRUE
GEF (Q_t, $X_{a,t}$, $P_{a,t}$)
END

Figure A.10. Flowchart of full data assimilation workflow with the GEF instead of the EnKF-Miyoshi workflow.

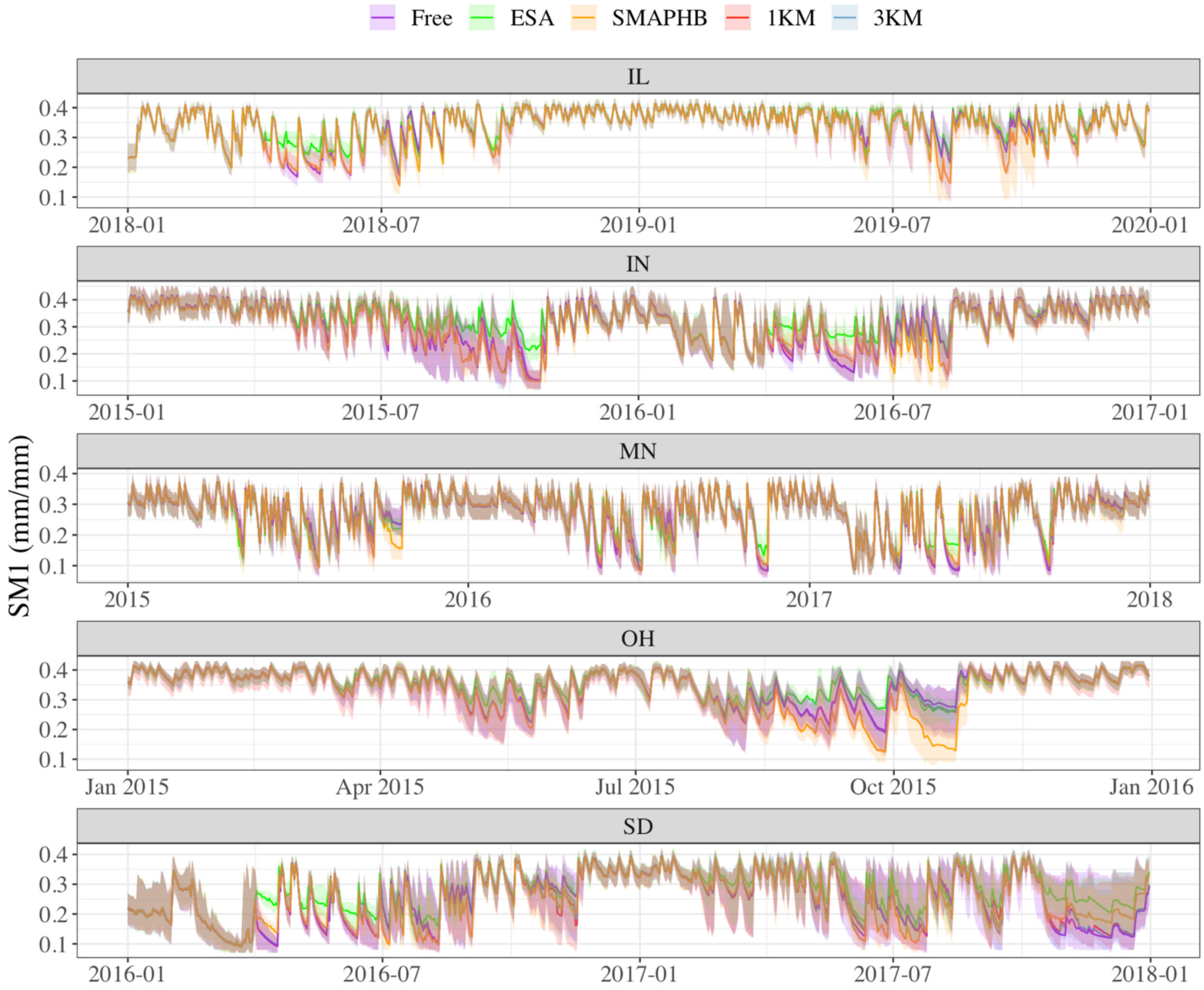

Figure A.11. Time series of SM1 estimates for different individual RS data product assimilation runs.

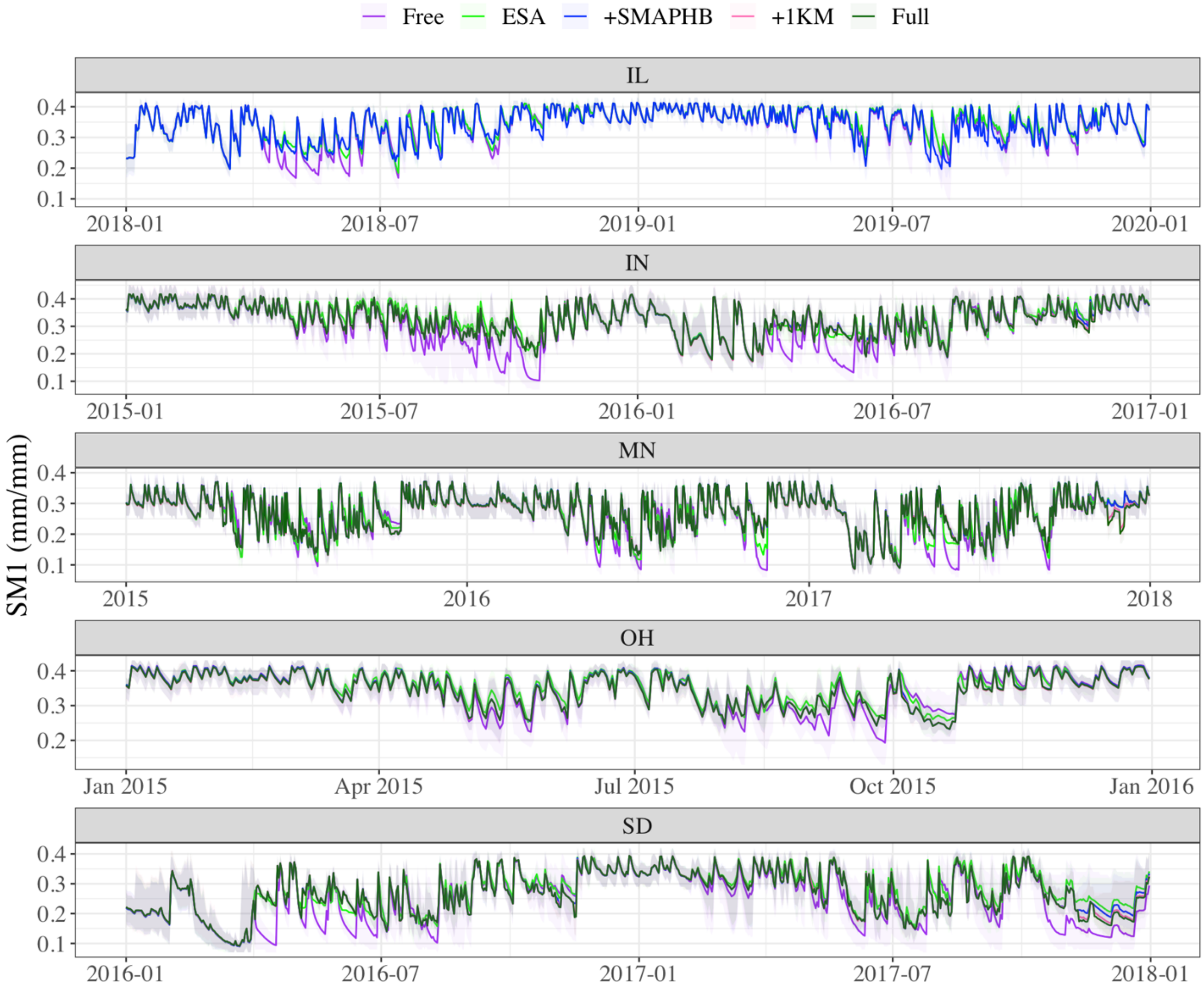

Figure A.12. Time series of SM1 estimates for the additive RS data product assimilation runs.

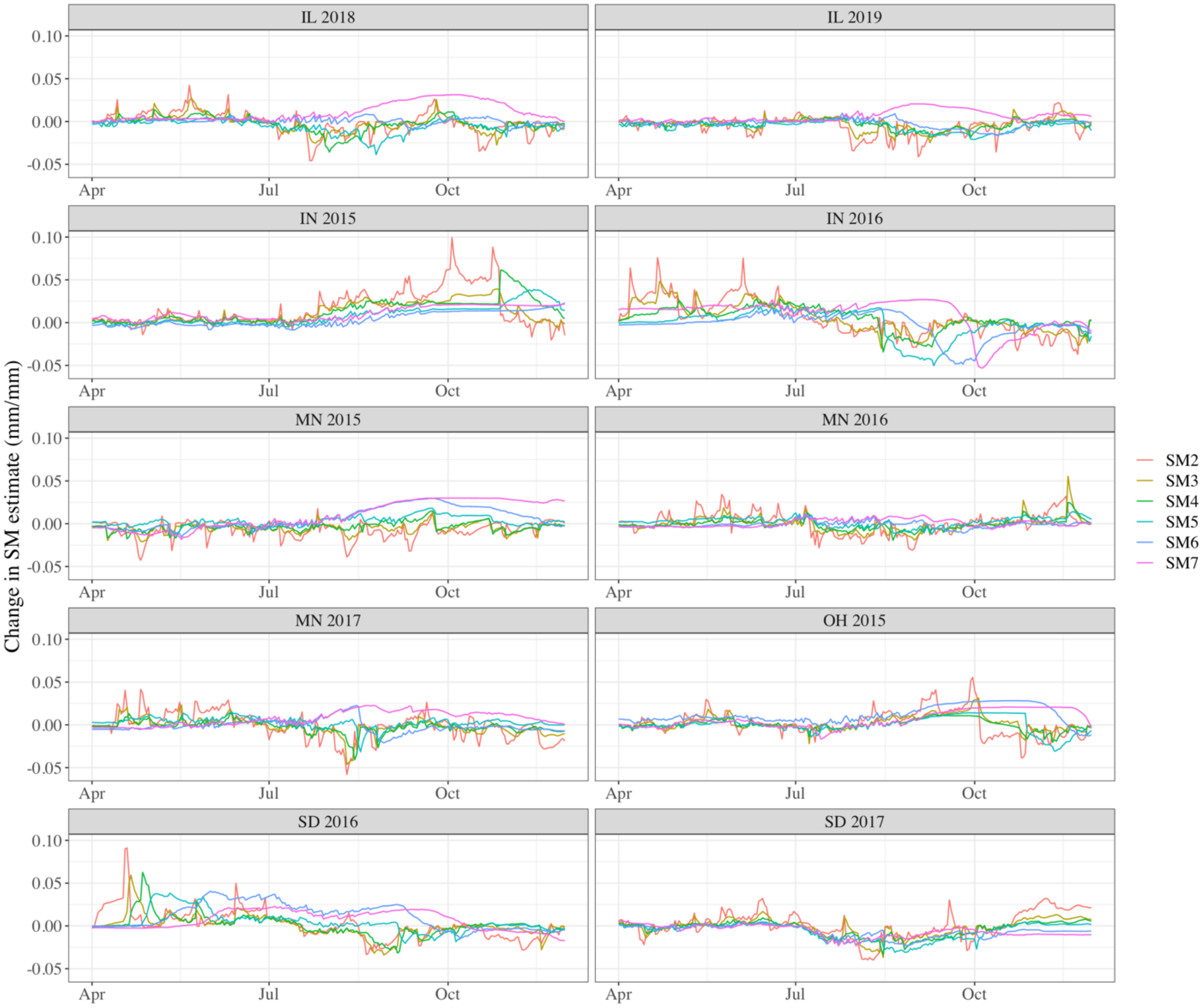

Figure A.13. Time series of differences in soil moisture mean estimates between the free model and RS-SDA (computed as RS-SDA – Free) for downstream soil layers.

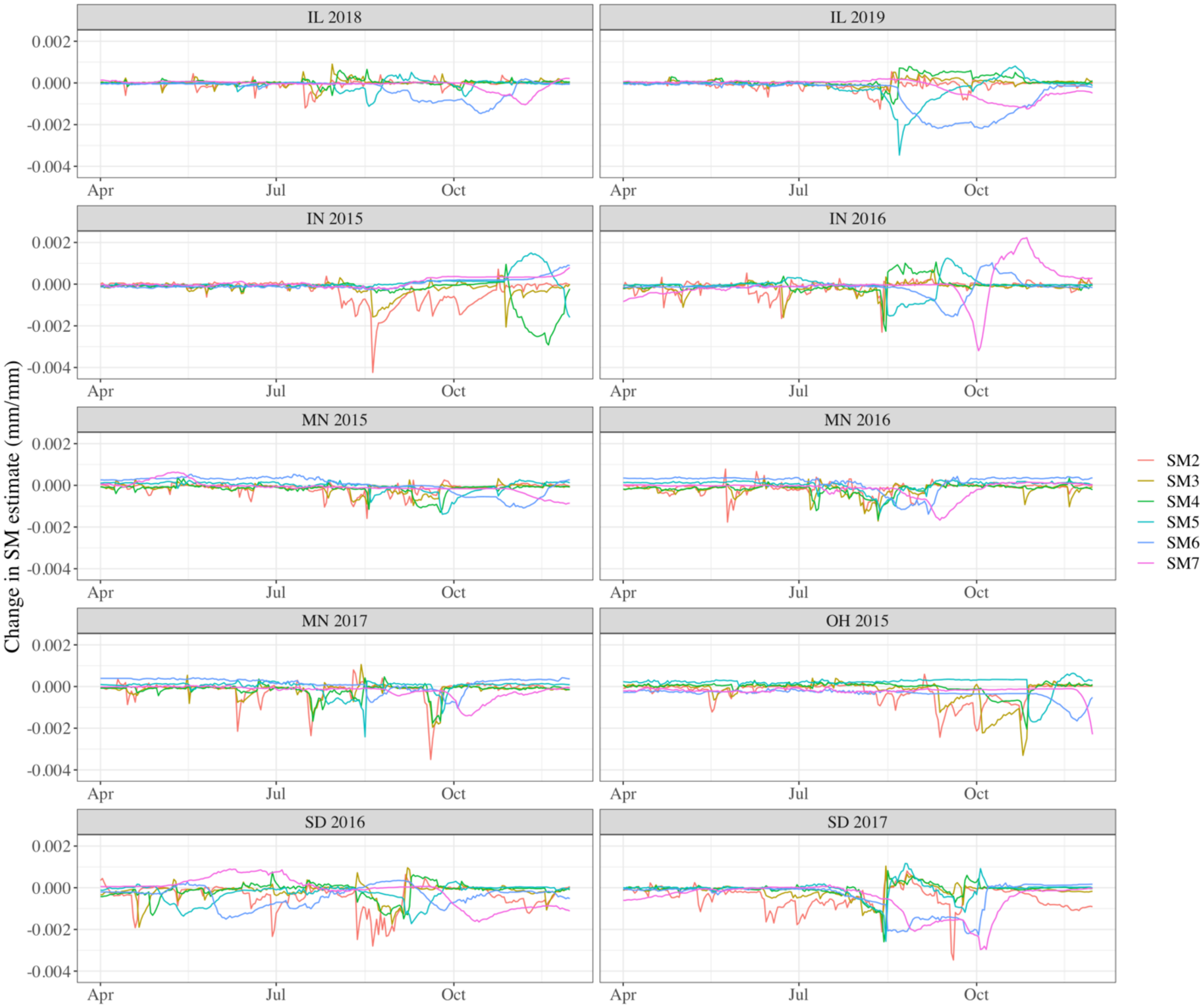

Figure A.14. Time series of differences in soil moisture weighted variance estimates between the free model and RS-SDA (computed as RS-SDA – Free) for downstream soil layers.

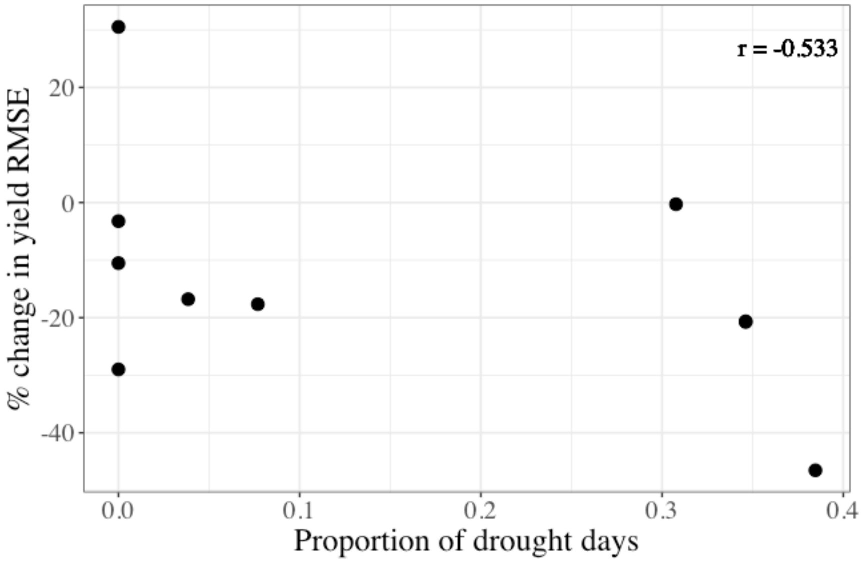

Figure A.15. Scatterplot of proportion of USDM drought days in the growing season and percent change in RMSE for yield estimates between RS-SDA and the free model for the 10 RS site-years. The monotonic relationship between the two variables was not significant at reasonable significance levels (p = 0.113). The points representing IN 2016 (0.35, -20.7) and MN 2015 (0.35, -20.6) are not distinguishable in this plot.

Table A.1. Fixed management parameters for the Energy Farm (2018-2019)

Variable	**Units**	**2018 Value**	**2019 Value**
Crop	crop type	maize	soybean
Residue type	crop type	maize	maize
Planting date	date	8 May 2018	17 May 2019
Row spacing	mm	762	762
Planting depth	mm	38	38
Sowing density	plants/m^2	8.4	34.6
Fertilizer date	date	8 May 2019	N/A
Fertilizer type(s)	APSIM fertilizer types	urea_N, NH4NO3	N/A
Fertilizer amount(s)	kg/ha	32.32, 92.3	N/A

Table A.2. Prior distributions for model ensembles

APSIM Variable	**Description**	**Units**	**Distribution**
icrag	Initial residue weight on the field	kg	Uniform (0, 2500)
water_fraction_full	Initial soil water fraction by volume	proportion	Uniform (0.05, 0.6)
tt_flower_to_maturity	Thermal time between flowering and maturity (maize cultivar)	ºC/day	Uniform (780, 860)
tt_flower_to_start_grain	Thermal time between flowering and start of grain fill (maize cultivar)	ºC/day	Uniform (150, 200)
tt_maturity_to_ripe	Thermal time between maturity and ripe stage (maize cultivar)	ºC/day	Uniform (150, 250)
tt_emerg_to_endjuv	Thermal time between emergence and end of juvenile stage (maize cultivar)	ºC/day	Uniform (240, 260)
head_grain_no_max	Maximum potential number of kernels per ear (maize cultivar)	Number of kernels/ear	Uniform (750, 900)
grain_gth_rate	Maximum potential growth rate of grain (maize cultivar)	Grain (g)/day	Uniform (7.1, 8.57)

Table A.3. Comparison of accuracy and precision in model forecasts between the Free and PDA schemes

Variable	Year	RMSE		Variance	
		Free	PDA	Free	PDA
Leaf Area Index *Unitless*	2018	0.869	0.870	0.074	0.064
	2019	1.165	1.000	0.411	0.370
Yield *Mg/ha*	2018	0.884	0.974	0.120	0.164
	2019	0.978	0.984	0.972	0.413
Daily Tile Flow *mm*	2018	1.308	0.997	0.774	0.418
	2019	1.214	0.945	0.397	0.253
	Both	1.262	0.972	0.586	0.335
Annual Tile Flow *mm*	2018	265.6	143.6	18405	8510.9
	2019	205.1	122.5	5673.5	4830.5
Daily NO_3 Load *Kg NO_3-N/ha*	2018	0.060	0.062	0.0016	0.0009
	2019	0.040	0.044	0.0007	0.0009
	Both	0.051	0.054	0.0011	0.0009
Annual NO_3 Load *Kg NO_3-N/ha*	2018	4.407	3.555	18.131	10.517
	2019	4.942	4.852	10.742	23.595