Remote Sensing of Environment 299 (2023) 113880

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

A flexible and efficient knowledge-guided machine learning data assimilation (KGML-DA) framework for agroecosystem prediction in the US Midwest

Qi Yang[a], Licheng Liu[a], Junxiong Zhou[a], Rahul Ghosh[b], Bin Peng[c,d], Kaiyu Guan[c,d,e], Jinyun Tang[f], Wang Zhou[c,d], Vipin Kumar[b], Zhenong Jin[a,*]

[a] Department of Bioproducts and Biosystems Engineering, University of Minnesota, Saint Paul, MN 55108, USA
[b] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA
[c] Department of Natural Resources and Environmental Sciences, University of Illinois Urbana- Champaign, Urbana, IL 61801, USA
[d] Agroecosystem Sustainability Center, Institute for Sustainability, Energy, and Environment, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[e] National Center for Supercomputing Applications, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[f] Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

ARTICLE INFO

Editor: Jing M. Chen

Keywords:
Knowledge-guided machine learning
Data assimilation
Carbon fluxes
Crop yield
Agroecosystem
U.S. Midwest

ABSTRACT

Process-based models are widely used to predict the agroecosystem dynamics, but such modeled results often contain considerable uncertainty due to the imperfect model structure, biased model parameters, and inaccurate or inaccessible model inputs. Data assimilation (DA) techniques are widely adopted to reduce prediction uncertainty by calibrating model parameters or dynamically updating the model state variables using observations. However, high computational cost, difficulties in mitigating model structural error, and low flexibility in framework development hinder its applications in large-scale agroecosystem predictions. In this study, we addressed these challenges by proposing a novel DA framework that integrates a Knowledge-Guided Machine Learning (KGML)-based surrogate with tensorized ensemble Kalman filter (EnKF) and parallelized particle swarm optimization (PSO) to effectively assimilate historical and in-season multi-source remote sensing data. Specifically, we incorporate knowledge from a process-based model, *ecosys*, into a Gated Recurrent Unit (GRU)-based hierarchical neural network. The hierarchical architecture of KGML-DA mimics key processes of *ecosys* and builds a causal relationship between target variables. Using carbon budget quantification in the US Corn-Belt as a context, we evaluated KGML-DA's performance in predicting key processes of the carbon cycle at three agricultural sites (US-Ne1, US-Ne2, US-Ne3), along with county-level (627 counties) and 30-m pixel-level (Champaign County, IL) grain yield. The site experiments show that updating the upstream variable, e.g., gross primary production (GPP), improved the prediction of downstream variables such as ecosystem respiration, net ecosystem exchange, biomass, and leaf area index (LAI), with RMSE reductions ranging from 9.2% to 30.5% for corn and 4.8% to 24.6% for soybean. Uncertainty in downstream variables was automatically constrained after correcting the upstream variables, demonstrating the effectiveness of the causality linkages in the hierarchical surrogate. We found joint use of in-season GPP and evapotranspiration (ET) products along with historical GPP and surveyed yields achieved the best prediction for county-level yields, while assimilating in-season LAI observations benefitted the prediction in extreme years. Uncertainty and error analysis of regional yield estimation demonstrated that KGML-DA could reduce prediction error by 26.5% for corn and 36.2% for soybean. Remarkably, the GPU-based tensor operation design makes this DA framework more than 7000 times faster than the PB model with a High-Performance Computing system, indicating the high potential of the proposed framework for in-season, high-resolution agroecosystem predictions.

* Corresponding author.
E-mail address: jinzn@umn.edu (Z. Jin).

https://doi.org/10.1016/j.rse.2023.113880
Received 30 May 2023; Received in revised form 7 October 2023; Accepted 19 October 2023
Available online 25 October 2023

1. Introduction

Cropland, covering 12–14% of the global ice-free surface, is a key link in the terrestrial carbon cycle (Shukla et al., 2019; Lal, 2011). Monitoring the dynamics of the carbon cycle in agroecosystems is essential for carbon budget quantification, in-season crop management, and nutrient management for sustainable agricultural production (Gan et al., 2014; Wang et al., 2023; Dold et al., 2017). In-situ measurements are the gold standard for ground-truthing carbon pools; However, the lack of scalability and high cost of this method call for earth observation (EO) technologies to quantify regional scale terrestrial carbon budgets, such as gross primary production (GPP) (Jiang et al., 2021), aboveground biomass (Liang et al., 2023), leaf area index (Kimm et al., 2020), and crop yield (Lobell et al., 2015; Jin et al., 2019). Although remote sensing-based methods provide possibilities for large-scale carbon budget monitoring, their limitations are also obvious – only a few components of the carbon cycle can be directly observed and the continuity of data is affected by cloud contamination and satellite revisit frequency (Weiss et al., 2020). Meanwhile, researchers have long been exploring simulations of agroecosystem carbon dynamics via process-based (PB) models. State-of-the-art PB models possess reasonably good extrapolation and transferability in time and space, and they have the capability to simulate comprehensive and continuous processes. From the first canopy photosynthesis model (Monsi, 1953) to more sophisticated agroecosystem models (e.g., DSSAT, APSIM, and *ecosys*) (Jones et al., 2003; Holzworth et al., 2014; Grant et al., 2020b), errors in the model structure have been reduced as field knowledge evolves. However, it remains challenging to retrieve many field-level crop parameters, aleatoric events (e.g., flooding, pests, and plant diseases), and the management schedules of individual farmers (e.g., the timing of planting, fertilization, and irrigation). Substantial parametric uncertainty and input uncertainty have therefore been introduced into simulations by default, so that models fail to capture the spatiotemporal variability of agroecosystems (Tao et al., 2018).

Data assimilation (DA), a well-known approach in model-data fusion (Guan et al., 2023), is among the most promising ways to address these uncertainties in simulations. Leveraging various readily accessible remote sensing products such as evapotranspiration (ET), leaf area index (LAI), and soil moisture (SM) (Jiang et al., 2020; Melton et al., 2022; Ma and Liang, 2022; Li et al., 2022a, 2022b), DA is widely used to constrain the predictive uncertainty in the agroecosystem (Huang et al., 2019; Jin et al., 2018; Ines et al., 2013; Hu et al., 2019; Yang et al., 2023). The DA approaches can be categorized into two groups: batch DA (for retrospective simulations) and sequential DA (for real-time simulations) (Markovich et al., 2022). The essential distinction between these two DA approaches lies in the manner of assimilating observations. Batch DA methods, such as variational methods and smoothing methods, perform global optimization for model parameters and initial states by fusing a period of historical time-series data all at once, while sequential DA (e.g., filter-based methods) ingests observations in real-time to update the model states (which is more suitable for dynamic systems) (Bertino et al., 2003). To maximize the value of historical and real-time data, methods that jointly use both approaches have been developed to estimate parameters and states simultaneously (Moradkhani et al., 2005b; Kang and Özdoğan, 2019).

However, three major stumbling blocks still stand in the way of large-scale high-resolution (e.g., 30 m) simulations of agroecosystems via PB model-based DA. First, the computational challenge of traditional DA frameworks: the large number of model runs (due to ensemble members, iterations, and large-scale pixel-level simulations) entail extremely high computation costs, especially for advanced PB models (Wood et al., 2011; Bauer et al., 2021). Second, the challenge of reducing structural error in PB model-based DA: a PB model is a synthesis of state-of-the-art knowledge, but once it is developed, its structural biases are hardwired (i.e., model structural error will be fixed once the model is developed), and correction requires comprehensive domain knowledge. Third, the challenge of developing a DA framework: coupling sequential DA algorithms with detailed PB models is highly intrusive (i.e., massive modifications for the PB model source code) and inflexible. For a detailed model in which one target variable may entangle with many intermediate variables, simply updating the target variable may break the model consistency and result in a non-convergence of estimates (Hu et al., 2017). Moreover, the DA framework needs to be scrutinized and redeveloped when new observation types are available to be assimilated, or an improved version of the PB model is released. The inflexibility of the traditional framework greatly hinders technological progress in fully mining information from multi-source remote sensing data (Table 1).

The rapid progress of ML-based surrogate models may overcome the first challenge by shifting large-scale simulation from CPU-intensive PB models to GPU-intensive deep neural networks (Bauer et al., 2021). Large-scale parameter calibration is becoming feasible thanks to the ultra-efficient inference of surrogate neural networks (Zhou et al., 2021a). A few preliminary attempts have been made to incorporate ML into a sequential DA framework. For example, Cintra and de Campos Velho (2018) and Wahle et al. (2015) surrogated an ensemble Kalman filter (EnKF) by neural networks to estimate the posterior covariance matrix of predictions for atmospheric and ocean models, in which the state vector is too large to calculate the Kalman gain. Certain works have investigated using ML to quantify observation uncertainty in the DA framework (Han et al., 2022). Notably, Brajard et al. (2020) developed an integrated ML-based DA framework, although it is an implementation of a toy model (an ordinary differential equation with one parameter). Although pure ML algorithms may outperform traditional methods in feature learning and inference speed, they have been criticized for their poor interpretability which may lead to poor extrapolation and generalization abilities (Shwartz-Ziv and Tishby, 2017). The low fidelity of the ML-based surrogate model also limited its ability to assimilate multi-source data (Table 1). As a result, a new technology that can mimic the comprehensive intermediate processes of PB models is needed to improve the fidelity of corresponding surrogates.

Knowledge-guided machine learning (KGML) is a new research paradigm that has the potential to surrogate a complex system and address the second challenge by introducing prior knowledge into the neural network (Karpatne et al., 2022; Willard et al., 2022; Shen et al., 2023). It leverages the strong feature-learning capability of machine learning (ML) and the interpretability of PB models. This method enables KGML-based surrogates to reproduce the causality and explicitness of the PB models, rendering it capable of consuming multi-source data to correct the model structural error. Pioneer achievements of KGML have demonstrated five possible approaches to integrating knowledge: 1) using a PB model-generated synthetic dataset to pre-train neural networks (Liu et al., 2022); 2) adding extra loss terms to the training objective function to ingest knowledge from physical laws, such as mass and energy balance (Jia et al., 2021) and kinetic partial differential equations (PDEs) (Cuomo et al., 2022); 3) hardcoding knowledge into network structures. For instance, concatenating a specific network layer with a physical equation (or model) to force that layer to output desired terms (Tsai et al., 2021; ElGhawi et al., 2022); PDEs can be hardcoded into the network via the Fourier approximation technique (Li et al., 2020); 4) hybrid modeling via residuals learning (Zhang et al., 2019), cascade coupling (e.g., using the output of ML as the input of a crop model, Han et al., 2022), and interactive connection (Yang et al., 2020); and 5) unsupervised representation learning and inverse modeling, for example, learning model parameters via embeddings derived from a self-supervised inverse learning neural network (Ghosh et al., 2022).

Although KGML sheds light on building high-fidelity surrogates, methods to address the third challenge (i.e., DA framework development challenge) remain uninvestigated. One reason is that most of the previous existing surrogates have only focused on limited target variables and specific processes (Zahura et al., 2020). Their non-hierarchical structure design outputs independent variables at the same level,

Table 1
Comparison of different DA methods for addressing the challenges in large-scale agroecosystem modeling.

Methods	Computational cost	Model structure		DA framework development	
		Fidelity	Potential for error reduction	Flexibility	Extensibility
PB model-based DA	High	High	Low (requires comprehensive domain knowledge for redevelopment)	Low (massive modifications for source code of a detailed PB model)	Low (redevelopment needed for new observation types or upgraded model)
ML-based DA	Low	Low (non-hierarchical; focus on limited variables or processes; lack of causality)	Medium (address structural error for specific process via surrogate fine-tuning)	Medium (without modifying model source code; independent outputs reduce DA performance)	Low (limited data interfaces)
KGML-based DA	Low	High (hierarchical; replicate multiple processes; causal linkage among these key processes)	High (Comprehensively adapt model structure via surrogate fine-tuning)	High (without modifying model source code; explicitly update the hidden states of RNN)	High (reserve possible interfaces for potential data types)

Notes: Fidelity refers to how comprehensively and accurately the model represents and reproduces the behavior of the real-world processes; flexibility means how flexibly the modeler can customize their own DA framework; extensibility refers to the ability that the DA framework adapt to new observation types that might not be available currently.

resulting in a lack of causality between target variables that is important to pass on information when an upstream variable is assimilated (illustrated in Fig. S1). Another critical issue in addressing this challenge is how to develop a surrogate neural network with the Markov property. This characteristic enables the neural network to assimilate real-time data, which means the surrogate should be able to explicitly inherit the previous model status into the next time step (i.e., model states at time $t + 1$ should depend only on the states from time t). It is a fundamental assumption of sequential DA that allows the model to carry the information assimilated at time t. For state-of-the-art recurrent neural networks (RNNs) that are used to tackle sequence data, the temporal model states are implicitly stored in a tensor named "hidden state" (Chung et al., 2014). Unfortunately, how to unlock the "hidden state" of RNN to provide an explicit causal simulation for data assimilation is still an unanswered question.

In this paper, we propose a KGML-based DA framework to simultaneously disentangle the aforementioned three challenges for large-scale, high-resolution DA (Table 1). For demonstration purposes, we used an advanced agroecosystem PB model, named *ecosys* (Grant, 2001; Grant et al., 2011, 2020b), to drive the KGML-DA. A hierarchical surrogate neural network with temporal awareness was designed to allow the network to carry over the assimilated information. We integrated this surrogate with tensorized EnKF and parallelized particle swarm optimization (PSO) to effectively assimilate historical and in-season observations. To investigate the contribution of in-season and historical data, we examined various DA strategies, including parameter calibration, state-updating, and the joint use of both techniques. Using the Midwestern US corn-soybean production system as a context, we tested the framework by predicting carbon budgets at three agricultural sites, along with county-level and pixel-level grain yield. Moreover, we investigated the value of assimilating multi-source remote sensing data, including SLOPE GPP, MODIS ET, and GLASS LAI. This framework possesses three notable features: 1) efficiency, i.e., the framework is capable of large-scale GPU-based parallel simulation; 2) flexibility and extensibility, i.e., designing the surrogate structure should be easy and flexible, and the DA framework should able to adapt to new observation types that might not be available currently; 3) interpretable and has an explicit calculation process for DA, i.e., we unlocked the implicit "hidden state" to provide the explicit updates of state variables. The KGML-DA framework is not limited to predicting the carbon cycle of agroecosystem, but has the potential to be applied to other ecosystems such as forests and grasslands for water, carbon, and nutrient cycles prediction.

2. Data and methodology

2.1. Study area and data

This study focuses on the major producing regions of corn and soybean, known as the Corn Belt, in the Midwestern US. A total of 627 counties from 14 states were involved (Fig. 1). The data used in this study can be categorized into four aspects (Table 2):

(1) Synthetic dataset generation. We randomly sampled 20,000 points from the corn or soybean fields over the study area to capture the responses of *ecosys* to different climate scenarios and soil conditions (Fig. 1). For each sampling point, corresponding hourly weather forcing of NLDAS-2 from 1980 to 2020 and soil properties from gSSURGO were extracted to drive the *ecosys* to produce synthetic data (data with a superscript "a" in Table 2).

(2) Site-scale experiments. The performance of the KGML-DA framework for estimating carbon cycle components was evaluated on three AmeriFlux sites (US-Ne1, US-Ne2, and US-Ne3 located in Mead, Nebraska) (Suyker and Verma, 2012; Jeffries et al., 2020) (Fig. 1). The data being assimilated is the SLOPE GPP, a daily 250 m remotely sensed GPP product that calculates GPP for C3 plants (e.g., soybean) and C4 plants (e.g., corn) separately, while other existing products neglect this distinction and consequently tend to underestimate GPP for corn and overestimate for soybean (Jiang et al., 2021). The 250 m pixels of the plot centers are not overlapped and have at least a two-pixel buffer. The validation data includes eddy-covariance-based fluxes, in-situ LAI and aboveground biomass. Data involved in site-scale experiments was marked with superscript "b" in Table 2.

(3) County-level yield estimation. Multisource remote sensing data, including SLOPE GPP, MODIS ET, and GLASS LAI, were assimilated to evaluate the proposed framework. The MODIS ET and GLASS LAI were centralized to the mean of the open-loop simulations prior to DA to mitigate their systematic underestimation in cropland (Chen et al., 2018). The yield estimates were compared against the NASS-reported yield. To aggregate the pixel-level data to county-level, we extract pure crop pixels based on Corn-Soy Data Layer (CSDL, Wang et al., 2020) (from 2000 to 2007) and USDA-Crop Data Layer (CDL, USDA, 2023) (from 2008 to 2020). Involved data were summarized in Table 2 with superscript "c".

(4) A 30-m version of SLOPE GPP was used for the 30-m yield mapping. The 30-m soil properties (gSSURGO) (Fig. S2) and 1-km gridded Daymet v4 climate data (Fig. S3) was used for driving the model. Since wind speed data is not included in the Daymet dataset, we downloaded the mean wind speed (2 m above

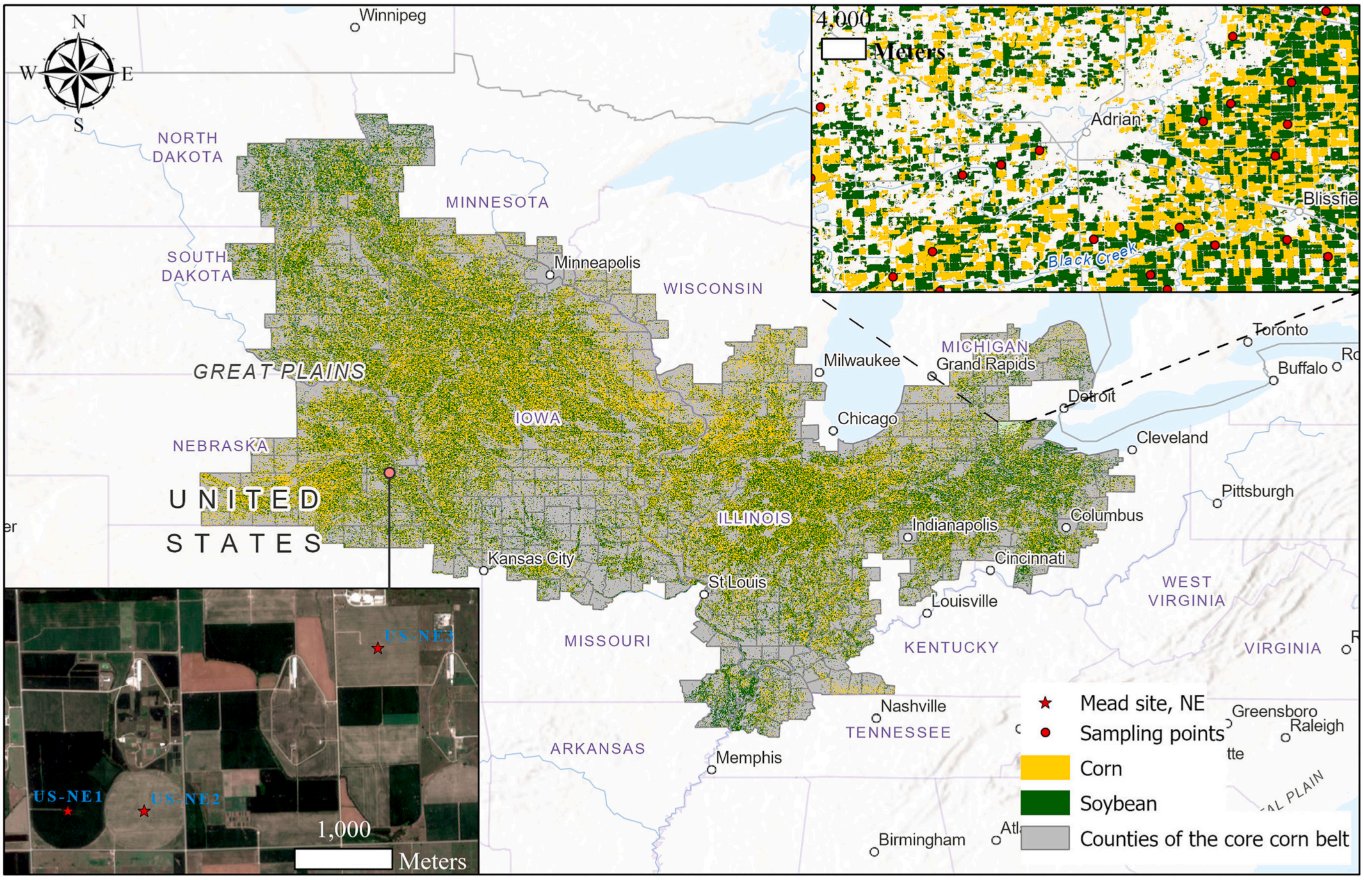

Fig. 1. Relative geo-location of the study area on the core Corn Belt area (USDA-Crop Data Layer of Year 2017 was used for demonstration, USDA, 2023). The top-right inset shows a zoomed-in example of the spatial distribution of the sampling points (red dots). The bottom-left is three validation sites (red stars) with flux measurements at Mead, Nebraska. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Datasets used in this study.

Datasets	Use	Descriptions	References
NLDAS-2	a, c	Hourly weather forcing (0.125°)	https://ldas.gsfc.nasa.gov/nldas/v2/forcing
gSSURGO	a, b, c	Grided soil properties (10 m)	https://nrcs.app.box.com/v/soils/folder/180112652169
FLUXNET2015	b	Eddy-covariance-based GPP, Reco, NEE, ET	https://fluxnet.org/data/fluxnet2015-dataset/
AmeriFlux BADM	b	In-situ LAI and aboveground biomass	https://ameriflux.lbl.gov/data/badm/
NASS yield	c, d	Counrt-level yield	https://quickstats.nass.usda.gov/
SLOPE GPP-250	b, c	Daily GPP (250 m)	Jiang et al., 2021
MODIS ET	c	8-day composite ET (500 m)	Mu et al., 2011
GLASS LAI	c	8-day composite LAI (250 m)	Ma and Liang, 2022
USDA-CDL	a, c, d	Crop data layer (30 m)	https://croplandcros.scinet.usda.gov/
CSDL	c, d	Corn-soy data layer (30 m)	Wang et al., 2020
Daymet-v4	d	Daily weather forcing (1 km)	https://daac.ornl.gov/DAYMET
SLOPE GPP-30	d	Daily GPP (30 m)	Luo et al., 2018; Jiang et al., 2021

Note: use of a, b, c, d represent synthetic data generation, site-scale experiment, county-level yield estimation, 30-m yield mapping, respectively.

ground) data from the NASA POWER Project (https://power.larc.nasa.gov/). Related data was indicated by the superscript “d”.

To ensure the temporal consistency in the assimilated data, we initially applied a Savitzky-Golay filter to smooth data with varying temporal resolutions (daily: SLOPE GPP; 8-day composite: MODIS ET and GLASS LAI), which was then sampled every 8 days to produce the temporal consistent observations.

2.2. The process-based ecosys

In this study, we aim to build an extensible DA framework that offers a wide array of data interfaces to assimilate possible observations, even if such data is not currently available. The prerequisite for achieving this goal is to utilize an advanced and holistic process-based model to guide the training procedure of the hierarchical surrogate neural network. *Ecosys* is one of the very few models dedicated to constructing comprehensive biophysical and biochemical processes and interactions for the soil–plant–atmosphere continuum (SPAC) system. It is an open-source hourly model that simulates detailed fluxes and pools of water, carbon, nitrogen, and phosphorus in the SPAC system, and has been well-examined for various ecosystems including crops, forests, and grassland (Grant, 2001; Grant et al., 2006; Mezbahuddin et al., 2020; Zhou et al., 2021a; Qin et al., 2021). Unlike traditional crop growth models (e.g., DSSAT, APSIM, and WOFOST) and soil biochemical models (e.g., DNDC and DayCent) that mainly focus on a specific domain, *ecosys* comprehensively integrates the crop/plant growth processes and soil microbial activities with elaborate descriptions for the exchanges of mass and energy and the chemical transformation of nutrients under diverse management practices. The sophisticated model structure makes *ecosys* capable of accurately simulating the detailed subprocesses in carbon, water, and nitrogen cycles (Grant et al., 2020a). For carbon cycle components, such as the dynamics of croplands GPP, net ecosystem exchange (NEE), ecosystem respiration (Reco), LAI, organ biomass, and methane, *ecosys* simulations showed high consistency with in-situ measurements (Zhou et al., 2021a; Chang et al., 2019). *Ecosys* also has been intensively validated in simulating the nitrogen (N) cycle dynamics, such as the N mineralization and plant uptake (Welegedara et al., 2020), soil inorganic nitrogen dynamics (Li et al., 2022a, 2022b), the nitri-denitrification processes and nitrous oxide (N_2O) emission (Yang et al., 2022). The main structure of *ecosys* was summarized in a book chapter of (Grant, 2001). More details about mechanistic process representations in *ecosys* can be found in the supplement of (Grant et al., 2020a).

2.3. Developing the KGML-DA framework

2.3.1. Generating synthetic data

The surrogate model of *ecosys* is the core of the KGML-DA framework. The requisite of establishing a high-fidelity surrogate neural network is to generate a large synthetic dataset that includes the responses of *ecosys* simulations to various parameter combinations and climate scenarios. To obtain model responses to different crop genotypes, six yield-sensitive crop parameters were selected (refers to Zhou et al., 2021a who conducted parameter sensitivity analysis based on a surrogate neural network) and randomly drawn from their respective uniform distributions to generate input files for the synthetic dataset. Additionally, management inputs, including planting date and nitrogen fertilizer (only for corn), are also perturbed. The ranges of parameters for corn and soybean are listed in Table 3. The generated input dataset was subsequently fed into *ecosys* to produce model responses for any possible scenarios (the first 20 years of data were used for model spin-up and the rest 21 years were used for generating a synthetic dataset). Finally, a synthetic dataset with 840,000 site-year data (20,000 sites × 21 years × 2 crops) was generated, which consists of 21 input variables and 9 output variables (detailed in Table S1).

Table 3
Yield-sensitive parameters selected for synthetic data generation. The sampling procedure doesn't adhere to a specific fixed interval, making it possible to sample any value within the range of variations.

Parameters	Descriptions	Corn		Soybean	
		Variation range	Default	Variation range	Default
CHL4	Fraction of leaf protein in bundle sheath chlorophyll	[0.02, 0.07]	0.05	–	0
VCMX	Rubisco Carboxylation Activity (umol g^{-1} s^{-1})	–	90	[20, 90]	45
GROUPX	Plant Maturity Group	[15, 21]	17	[16, 22]	18
STMX	Maximum number of fruiting sites per reproductive node	[2, 8]	5	[2, 8]	4
GFILL	Maximum rate of kernel filling (g C kernel h^{-1})	[0.0003, 0.0007]	0.0005	[0.0003, 0.0007]	0.0005
SLA1	Specific leaf area (m^2 kg^{-1})	[0.005, 0.025]	0.018	[0.005, 0.02]	0.01
PD	Planting date (DOY)	[105, 145]	121	[125,165]	140
NF	Total nitrogen fertilizer (g N m^{-2})	[0, 24]	18	–	0

Note: "-" denotes the parameter is not perturbed and the default value will be used.

2.3.2. Hierarchical GRU-based surrogate neural network

There are two neural network components in the KGML-DA framework: a main network for surrogating the process-based model, and an autoencoder network (which will be introduced in the next section) for conducting sequential data assimilation. The integration of knowledge into the surrogate encompasses three aspects: First, the key processes in the agroecosystem were hardcoded into the hierarchical structure but flexibility was left for the surrogate to explore the complex intermediate processes and interactions. We aimed to build a high-fidelity *ecosys* surrogate by balancing the "exploration (for underlying laws)" and "exploitation (for well-documented knowledge)" in neural network architecture design, where a high-fidelity surrogate model means it includes more intermediate processes and inherits more knowledge from the PB model. Second, a hierarchical network structural design was employed to establish a causal linkage among key agroecosystem processes. Specifically, we developed a hierarchical gated recurrent unit (GRU)-based neural network to mimic the calculation process of *ecosys* at a daily scale. Third, a large synthetic dataset described in the previous section was used to impart prior knowledge to the surrogate.

We chose GRU (Fig. S4a), one of the variants of the RNNs, as the building block of the surrogate for three reasons: 1) its high computational efficiency; 2) it has a similar performance with the state-of-the-art long short-term memory (LSTM) networks but a simpler structure (Gruber and Jockisch, 2020); and 3) its flexibility enables the construction of a hierarchical structure using GRU cells. In the proposed hierarchical GRU-based surrogate, the GRU cells can be grouped into three concatenate layers (i.e., the bottom, middle, and top) based on their functions (Fig. 2a). The bottom cells (cell-1 and cell-2) are sensitive to the climate forcing data and represent the primary processes in the SPAC system. Specifically, cell-1 simulates the crop phenology, a key upstream state variable that controls the clock of the whole system. Cell-2 models the carbon input (i.e., GPP) and aims to represent the light inception and photosynthesis processes of the agroecosystem. Both cell-1 and cell-2 are directly driven by the daily climate forcing, model parameters, and management information. The input and output of the GRU-based surrogate were summarized in Table S1. The hourly climate forcing in the synthetic dataset was aggregated into daily. Additionally, LAI from the previous simulation step (*t*-1) is fed into cell-2 to constrain the current photosynthetic capacity. Considering that evapotranspiration and assimilation of carbon dioxide are deeply coupled and are both controlled by the stomatal conductance associated with plant water stress (Ball et al., 1987), cell-2 also simulates the ET and topsoil moisture dynamics (0–30 cm) along with GPP.

The middle (cell-3) and top cells (cell-4) do not directly take the driving data as input; instead, they take the hidden state (i.e., the internal representation that encodes the memory of the cell) as input which is produced by the previous cells (Fig. 2a). This design ensures that the upper cells are impacted by the high-level features extracted by the previous cells (encoded accumulated environmental effects on the crop), rather than by low-level fluctuating driving data. Specifically, cell-3 takes the hidden states from cell-1 and cell-2 as input and models the processes of respiration that are related to the carbon pool and phenology. It estimates Reco and NEE fluxes and further predicts the aboveground biomass, which is a high-level component in the carbon cycle that integrates the assimilated photosynthate and carbon consumed by respiration. As the top cell of the surrogate, cell-4 learns the dry matter partition processes and then deduces LAI, which is determined by the leaf matter and the specific leaf area (SLA, related to phenology). This cell interprets all of the information flow distilled by lower-level cells and ultimately outputs the crop yield. For each cell, the daily values of the target variables (e.g., the GPP, ET, and SM for cell-2) are explicitly decoded from its hidden state.

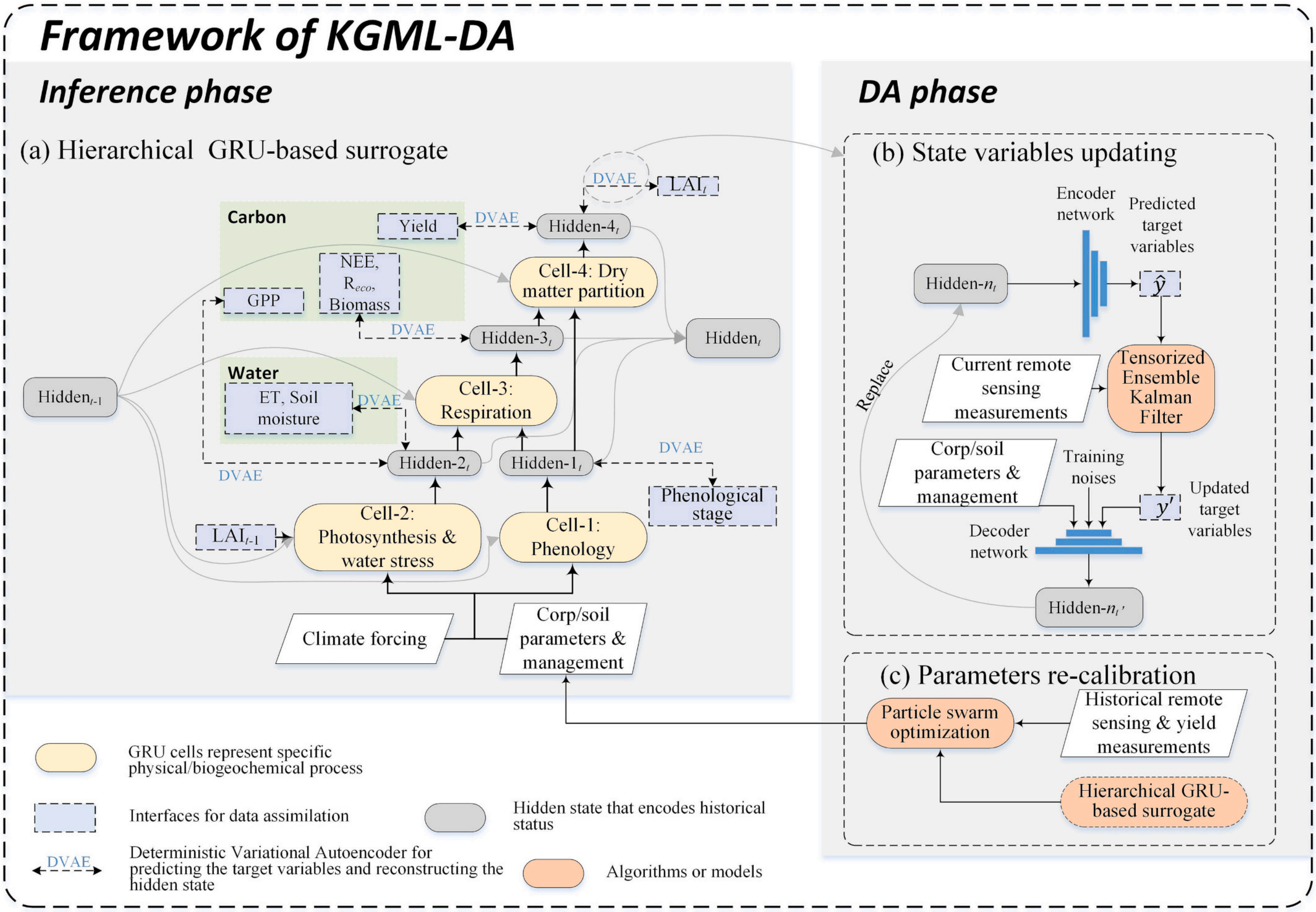
Framework of KGML-DA
Inference phase
(a) Hierarchical GRU-based surrogate
DA phase
(b) State variables updating
DVAE
LAI$_t$
Carbon
Yield
Hidden-4$_t$
NEE, R$_{eco}$, Biomass
Cell-4: Dry matter partition
GPP
Hidden-3$_t$
Hidden$_t$
Hidden$_{t-1}$
Water
ET, Soil moisture
Cell-3: Respiration
Hidden-2$_t$
Hidden-1$_t$
Phenological stage
LAI$_{t-1}$
Cell-2: Photosynthesis & water stress
Cell-1: Phenology
Climate forcing
Corp/soil parameters & management
Encoder network
Predicted target variables
Hidden-n$_t$
$\hat{y}$
Replace
Current remote sensing measurements
Tensorized Ensemble Kalman Filter
Corp/soil parameters & management
Training noises
y'
Updated target variables
Decoder network
Hidden-n$_t$'
(c) Parameters re-calibration
Particle swarm optimization
Historical remote sensing & yield measurements
Hierarchical GRU-based surrogate
GRU cells represent specific physical/biogeochemical process
Interfaces for data assimilation
Hidden state that encodes historical status
Deterministic Variational Autoencoder for predicting the target variables and reconstructing the hidden state
Algorithms or models

Fig. 2. Architecture of the KGML-DA framework: (a) the hierarchical GRU-based surrogate neural network that mimics the key processes of the *ecosys*; (b) details of the DVAE module to update state variables and hidden states; and (c) assimilating historical measurements for parameters re-calibration. A multi-task learning strategy with nine loss terms was used to train each cell. The gradient was detached between cells to make each cell learn the specific biophysical processes.

The fidelity of the GRU-based surrogate to *ecosys* was evaluated from two aspects: the accuracy of estimations and the surrogate responses to model parameters. Specifically, the estimated output variables by the surrogate (including phenological stage, GPP, ET, SM, aboveground biomass, Reco, NEE, LAI, and grain yield) were compared against the *ecosys* simulation in the test set of synthetic data. We evaluated the model responses of four yield-sensitive parameters (GROUPX, CHL4, VCMX, and SLA) between the surrogate and *ecosys* on a random site (from 2000 to 2020). To capture the response curve, we evenly discretized the variation ranges of investigated parameters into seven levels.

2.3.3. *Autoencoder for sequential data assimilation*

In a traditional PB model-based sequential DA framework, the prior state variables will be replaced by the updated values explicitly, and the information from observation at time *t* can be passed on to the rest of the simulation. However, as we mentioned above, the temporal patterns of state variables are encoded in the hidden state of the RNN-based neural networks, which means explicitly updating state variables is impracticable. A solution to address this issue is to reinitialize the hidden state at every time step and use the target variables from the previous time step as additional input (Xu et al., 2022). Nevertheless, reinitializing the hidden state to zero will lose its temporal dependencies and hence this method may not be suitable for variables that have a specific temporal pattern or trend such as biomass, LAI, and yield. In this study, we proposed a deterministic variational autoencoder (DVAE) to interpret the hidden state and to reconstruct the updated hidden state based on the updated target variables after assimilating measurements (Fig. 2b, Fig. S4b). As shown in Fig. 2b, the prior distributions of target variables are approximated by the ensemble members $\hat{y}$ that are inferred by the encoder. Then the posterior distributions of target variables ($y^{'}$) are estimated using the EnKF algorithm given the measurements (detailed in Section 2.3.5.1).

The encoder of VAE outputs the mean and standard deviation for each latent variable and the latent distribution is constrained to be a standard normal distribution. This strategy makes VAE a better extrapolation ability than autoencoder (AE) for which the latent space is discrete and irregulated (Kingma and Welling, 2013). However, the encoder of VAE may be hard to train because learning distribution is more difficult than learning the mean vector only. To address this issue, we deployed a DVAE without distribution learning but injecting noise into the decoder to smooth the latent space. This strategy makes the AE easy to train and keeps the advantages of VAE (Ghosh et al., 2019). Specifically, $y^{'} + \varepsilon$ is fed into the decoder to reconstruct the hidden state where the random noise ε follows a Gaussian distribution of $N(0, CV \times \mu_z)$. CV is the pre-set coefficient of variation (0.01 in this study) and μ_z is the expectation of the target variables. The temporal pattern encoded in the hidden state could be affected by crop and soil parameters and fertilizer information; as a result, these inputs are additionally fed into the decoder to reconstruct the updated hidden state (Fig. 2b).

2.3.4. *Training GRU-based surrogate and DVAE*

The training of neural networks took place on the Pytorch platform, utilizing an NVIDIA RTX 3090 GPU. The surrogate and four DVAE (for four GRU cells) were trained simultaneously using a shared batch of synthetic data. To ensure that the DVAE gradients would not interfere with the main network, the gradients of the output variables for each GRU cell were detached prior to feeding into the DVAE. A total of nine mean square error (MSE) losses corresponding to nine target variables (i.e., phenological stage, ET, SM, GPP, NEE, Reco, Biomass, LAI, yield) were used to train the main network, meanwhile, each DVAE was trained by the corresponding reconstruction MSE losses. The initial learning rate was set at 0.001, with a 4% decay rate for each epoch. During the developing phase, the dataset was split into a training set (81%), a validation set (9%), and a test set (10%) to monitor the overfitting (training and validation losses shown in Fig. S5). In the implementation phase, both the training and validation sets (90%) were utilized to train the surrogate neural network. The training was stopped after 30 epochs when the validation loss oscillated near the minimum value. The batch size for training is 256 and the Adam optimizer is used.

2.3.5. GPU-intensive data assimilation algorithms

2.3.5.1. Tensorized ensemble Kalman filter for state updating. We developed a three-dimension tensorized EnKF (t-EnKF) to update model states via assimilating in-season observations. EnKF, a variant of Kalman filter (KF), is the best-known sequential DA algorithm that uses the Monte Carlo method to estimate the prior distribution of state variables for nonlinear systems (Evensen, 1994). It automatically balances the confidence of observations and predictions via Kalman gain. There are two phases in EnKF: model prediction and state updating. For the prediction phase, the prior distribution of state variables, including the mean of state $\overline{x}_t^f$ and covariance matrices $\mathbf{P}_t^f$, are calculated from the predicted ensembles $x_{t,i}^f$. Assuming there is an observation vector $\boldsymbol{y}_t$ with noise covariance $\boldsymbol{R}_t$, the Kalman gain matrix ($\boldsymbol{K}_t$) is calculated as follows,

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

where $\boldsymbol{H}$ is the transformation matrix to project states from the state space to the measurement space. Then, the state vector $\boldsymbol{x}_t$ can be updated by $\boldsymbol{K}_t$ and $\boldsymbol{y}_t$, and $\boldsymbol{P}_t$ can be updated by $\boldsymbol{K}_t$ as follow,

$$\boldsymbol{x}_t^a = \boldsymbol{x}_t^f + \boldsymbol{K}_t\left(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_t^f + \boldsymbol{\varepsilon}\right) \tag{2}$$

$$\overline{\boldsymbol{x}}_t^a = N_e^{-1}\sum\nolimits_{i=1}^{N_e} \boldsymbol{x}_{t,i}^a \tag{3}$$

$$\boldsymbol{P}_t^a = (\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{H})\boldsymbol{P}_t^f \tag{4}$$

where superscript a and f represent posterior and prior estimates, respectively. $\boldsymbol{\varepsilon}$ is the random noise (in this study we assumed the $\boldsymbol{\varepsilon} = \boldsymbol{R}_t$) to mitigate the risk of the "filter divergence" issue. $\boldsymbol{I}$ is the unit matrix. Eqs. 1–4 are formulated without spatial subscripts, making them versatile expressions applicable to pixel, site, or county-level simulations. A total of 100 ensemble members were randomly generated by perturbing the sensitive model parameters (Table 3) with a CV of 0.1, and the observation frequency was once every eight days during the growing season. We assumed constant observation errors for SLOPE GPP (1 g C/m^2·day), MODIS ET (0.5 mm), and GLASS LAI (0.4 m^2/m^2), corresponding to 5%, 10%, and 10% respectively of their peak average values during the vegetation growth period (approximately 20 g C/m^{2}day, 5 mm, and 4 m^3/m^3). For ET and LAI, the ratio of error to peak value was assumed to be relatively higher (10%) due to the coarse resolution of MODIS ET (i.e., 500 m) and the systematic underestimation of both MODIS ET and GLASS LAI in cropland areas (Huang et al., 2015; Chen et al., 2018).

Traditionally, both the state matrix $\boldsymbol{x}_{t,i}$ and $\boldsymbol{P}_t$ have 2D shapes, that are (n_{sample}, n_{state}) and (n_{state}, n_{state}) respectively, where n_{sample} is the number of ensemble members and n_{state} is the length of the state vector. For simulation tasks involving a large number of sites or pixels, the 2-D matrix operation may have low efficiency when the parallel threads are limited. To tackle this issue, we upgraded the EnKF (2-D matrix operation) to the t-EnKF (3-D matrix operation) by adding the number of sites n_{site} as a new dimension. The tensor operation is backed by the Pytorch platform that allows the t-EnKF to enable GPU acceleration.

2.3.5.2. Parallelized particle swarm optimization for parameter estimation. Before conducting in-season simulations using t-EnKF, historical NASS yield and SLOPE GPP data were assimilated by PSO to reduce the uncertainty of the parameters. A total of seven parameters for corn and six parameters for soybean were selected as the uncertain parameters (Table 3). PSO is one of the evolutionary optimization algorithms (considered as batch DA method) to search the optimal parameter combinations for a large solution space (Kennedy and Eberhart, 1995), and the benefits of joint use of PSO and sequential DA method have been demonstrated in common land models to reduce predictive uncertainty (Zhang et al., 2021). It initializes a group of particles (25 particles in this study) with random locations and velocities to explore the solution space. The algorithm of PSO is relatively simple and intuitive and its effectiveness and efficiency for the parameter calibration of agroecosystem models have been widely validated (Jin et al., 2017; Guo et al., 2018). The individual particles of the population are independent, which means the optimization process can be run in parallel using the GRU-based surrogate. Specifically, the first dimension (i.e., the batch dimension) of the input tensor represents the particle population. Therefore, simulating a generation of PSO particles needed only one inference. The loss function of optimization is as follows,

$$loss = loss_{yield} + \alpha\, loss_{season\ GPP} + \beta\, loss_{monthly\ GPP} \tag{5}$$

where the $loss_{yield}$ is the MSE between the estimated final yield and NASS yield. $loss_{season\ GPP}$ and $loss_{monthly\ GPP}$ are the MSE calculated by the accumulated GPP during the growing season and the accumulated monthly GPP, respectively. α and β are the weight coefficients to normalize the loss terms. In this study, α is 0.0067 (the reciprocal of the total number of days in the growing season) and β is 0.033 (the reciprocal of 30 days).

2.4. Simulation experiment design

2.4.1. Site-scale experiments

The KGML-DA framework was evaluated at three agricultural sites (US-Ne1, US-Ne2, and US-Ne3) in the Midwest Corn Belt. Six target variables, including GPP, ET, Reco, NEE, aboveground biomass, and LAI, were evaluated against the ground truth data (the details of the ground truth data are described in Section 2.1). Two different compositions of the state vector of t-EnKF were investigated to test the effect of co-updating and the benefits of hierarchical structure. For the first case (Section 3.2.1), all the six target variables mentioned above were put into the state vector to calculate the covariance matrices and they were updated simultaneously after assimilating SLOPE GPP. For the second case (Section 3.2.2), the state vector only included the GPP, ET and SM from cell-2, and the rest target variables were supposed to be constrained by the information flow of the hierarchical structure. The crop parameters of GROUPX and SLA were manually tuned to match the magnitude of the maximum GPP and LAI for all three sites. For corn, nitrogen fertilizer with 18 g N/m^3 per year was applied before planting. For the soybean field, no nitrogen fertilizer was applied in the simulation.

2.4.2. Regional scale evaluation

2.4.2.1. Estimating county-level corn and soybean yield. A total of 627 counties in the US Midwest were selected to evaluate the KGML-DA performance for crop yield prediction from 2000 to 2020 (Fig. 1). To tackle the inconsistent spatial resolutions, we developed a sampling technique that aggregates pixel-level observations into county-level data. Specifically, for each county, a maximum of 200 sampling points (100 for corn and 100 for soybean) with a 240 m buffer (8 Landsat pixels) were randomly selected each year from CSDL and CDL. We assumed all the selected sampling points represent the pure crop pixels thus the spatial inconsistencies were eliminated. The time series of remote sensing products were extracted based on the sampling points, and the results were aggregated to the county scale. Besides the SLOPE GPP, we also investigated the utility of assimilating MODIS ET and GLASS LAI for improving model performance during the whole period (2000−2020) as well as during the extreme year (e.g., a severe drought hit the Corn Belt area in 2012).

Assimilation strategies combining the EnKF and PSO were designed to address the uncertainty from unknown input (e.g., management) and in-season events (e.g., pest and disease) and the model's parametric uncertainty. Fig. 3 demonstrates the strategies for parameter calibration, where the calibration nodes (yellow square) provide prior model parameters for the next several years and the updating nodes (grey circle) correct the current estimations based on the present in-season data. We assess two approaches to assimilate historical data for the calibration nodes. The first approach is to use a fixed interval (one-year and three years intervals were evaluated) between two calibration nodes, and the second approach uses all available historical data to retrieve the optimum parameter combinations. The state-updating node is deployed by default and it is always ready to assimilate observations. Common settings of observation uncertainty (e.g., a fixed value or a percentage of measurement) are non-dynamic and may lose representativeness for an outlier. For instance, an abnormal observation with low uncertainty may crash the simulation because the filter tends to accept the abnormal measurement and reject the model prediction. We proposed a dynamic observation uncertainty adaptation method to minimize the influence of the abnormal measurements by increasing the preset observation uncertainty when an observation significantly deviates from the model prediction (pseudo-code shown in Table S2). Specifically, this method detects the potential abnormal observation (i.e., the relative error of model predictions and observation exceeds a certain threshold) and uses an inflation factor to increase its uncertainty. Estimated county-level corn and soybean yield under different assimilation strategies (t-EnKF, PSO, and the joint use of both) was evaluated by NASS yield data. To consider the positive effect of technological advances on yield, the average annual increment of NASS yield over the past 20 years was derived to correct yield prediction starting at the midpoint of the calibration period. Except for the sensitive crop parameters, planting date (with a CV of 0.05) and nitrogen fertilizer amount for corn (with a CV of 0.1) were also perturbed to generate EnKF ensemble members. For cases without parameter calibration, the mean values of the perturbed parameters were initialized by default values (Table 3); for cases utilizing both historical data (PSO) and in-season data (t-EnKF), parameters calibrated by PSO (described in Section 2.3.5.2) were used.

2.4.2.2. Mapping 30-m crop yield. To examine the effectiveness of the proposed KGML-DA framework for pixel-level simulation of the agroecosystem, we mapped the 30-m crop yield for Champaign County (Illinois) from 2010 to 2013. The 30-m version of SLOPE GPP data was generated based on the 30-m fused daily reflectance data (Luo et al., 2018). The county-wise model parameters were calibrated by the historical 250 m SLOPE GPP and NASS yield data, and they were implemented to every pixel. And the state variables were updated by assimilating the 30 m SLOPE GPP with an 8-day interval.

2.4.3. Uncertainty analysis of county-level yield estimation

Accurate quantification of the source of the predictive errors and uncertainties is essential for modelers to better understand the weak spots of the simulation so as to facilitate the improvement of the DA schemes and model structure. In this study, we partition the total predictive error into the priori error and the residual error. We define the priori error as caused by inaccurate prior information, in other words, due to a lack of knowledge and awareness of the model state (e.g., unknown information about microclimate, irrigation, flooding and fertilization) and parameters. We define the best achievable performance (BAP) as produced by a well-calibrated model after assimilating available data. So the priori error = error of open-loop simulation based on an uncalibrated model - error of BAP. The residual error includes the model

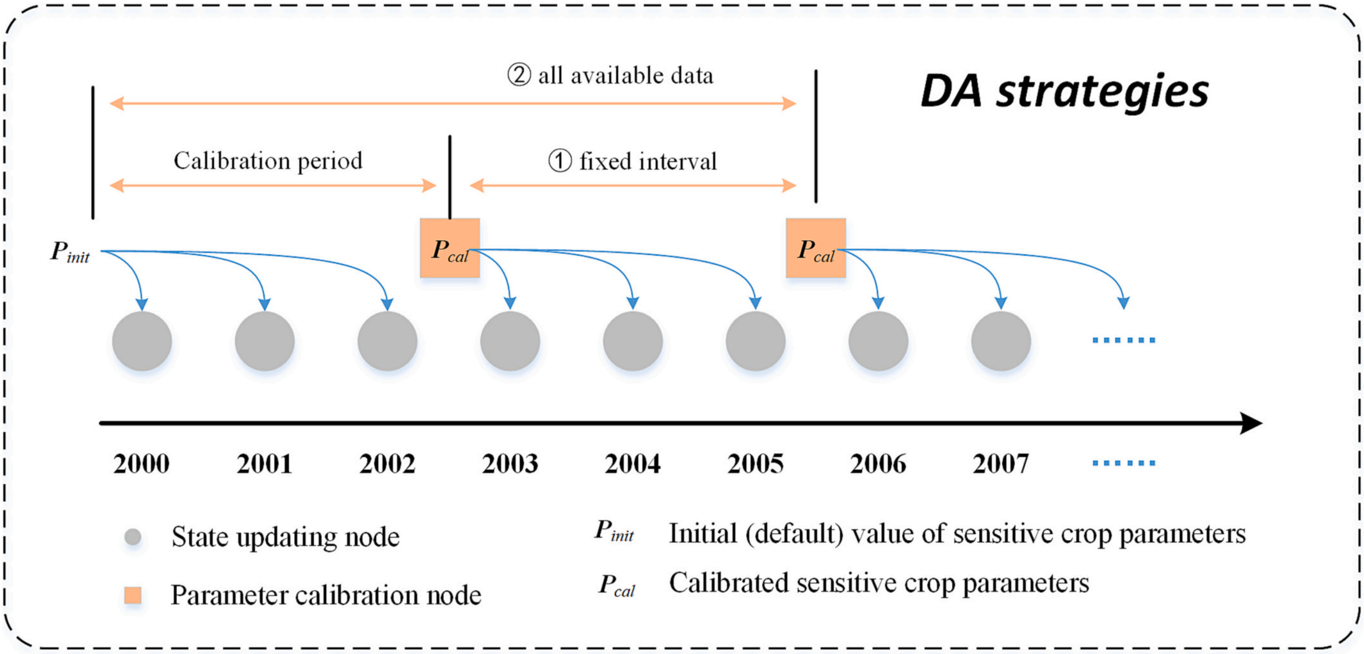

Fig. 3. Schematic diagram of the data assimilation strategies. The parameter is calibrated by the historical data (yellow square) using the PSO algorithm before the simulation of the current year. The current observations are assimilated to dynamically update the state variables (grey circle). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

structural error, the irreducible error (i.e., introduced by inherent randomness in the data), and the error introduced by limited observations. As a result, we quantified the priori error and the residual error by comparing the prediction error before and after DA.

We also evaluated the predictive uncertainties that were reduced by assimilating different combinations of remote sensing observations (i.e., SLOPE GPP, MODIS ET, and GLASS LAI). A hierarchical Bayesian model (Fig. S6) was fitted by Markov chain Monte Carlo (MCMC) to further partition the components of the prediction error and prediction variance into a global mean μ_g, spatial effect μ_s, and temporal effect μ_t as follows,

$$Y_{t,s} \sim N(\mu_o, \sigma) \tag{6}$$

$$\mu_o = \mu_g + \mu_s + \mu_t \tag{7}$$

$$\mu_g \sim N\left(\mu_g^{'}, \tau_g^{'}\right) \tag{8}$$

$$\mu_s \sim N(\mu_s^{'}, \tau_s^{'}) \tag{9}$$

$$\mu_t \sim N(\mu_t^{'}, \tau_t^{'}) \tag{10}$$

where $Y_{t,s}$ represents the samples. We assumed the prior $\mu_g^{'}$follows a uniform distribution $U(0, 60)$; $\mu_s^{'}$ and $\mu_t^{'}$ follow a normal distribution $N(0, 1)$; $\tau_g^{'}$, $\tau_s^{'}$ and $\tau_t^{'}$ follow an exponential distribution $Exp(10)$. The posterior distributions of μ_g, μ_s and μ_t were produced by MCMC with two chains and 2000 draws.

3. Results

3.1. Fidelity of the GRU-based surrogate

3.1.1. Evaluating the approximation error of the GRU-based surrogate

Model fidelity is commonly used to indicate how well a model mimics real-world processes. The higher fidelity of a surrogate indicates a lower approximation error to the PB model. The output variables include two components of the water cycle (ET and SM); three carbon fluxes (GPP, Reco, and NEE); two carbon pools (aboveground biomass and grain yield); phenological stage (DVS) and LAI. Fig. 4 evaluated the approximation error of the surrogate and showed good agreement between surrogate predictions and *ecosys* simulations, with R^2 of all target variables ranging from 0.85 to 0.99. No significant biases were observed for the surrogate predictions and the slopes of fitting equations were all beyond 0.96. The results indicated that the GRU-based surrogate well captured the *ecosys*' model responses of nine target variables in various simulation scenarios. To further assess the out-of-sample performance of the GRU-based surrogate neural network, we trained a surrogate using

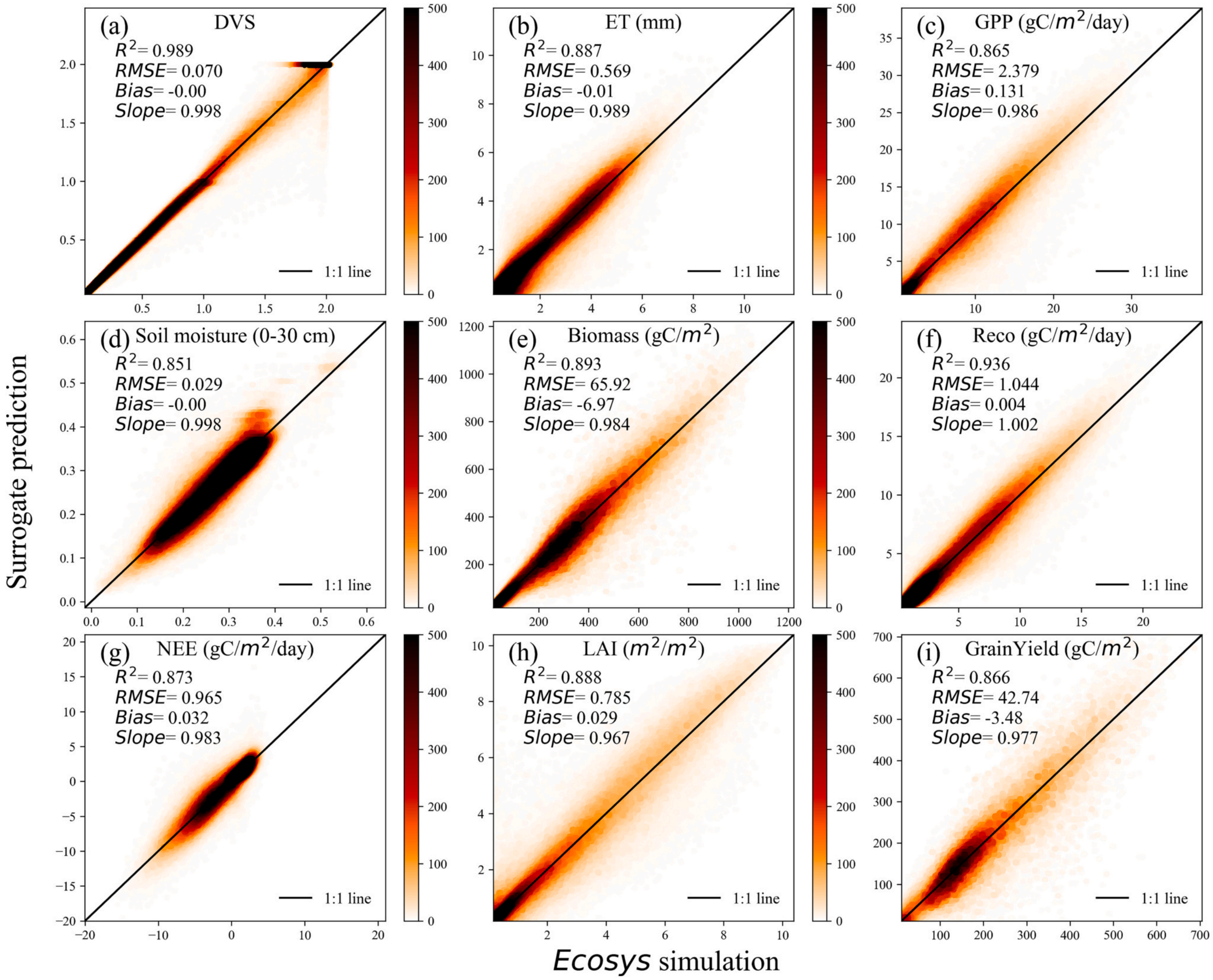

Fig. 4. Evaluation of the surrogate neural network for simulating (a) phenology (development stage); (b and d) components of the water cycle; (c, e, f, g, and i) carbon cycle; and (h) leaf area index on the test set.

synthetic data belonging to Iowa, Illinois and Indiana and then subjected the surrogate to testing across the entire Corn Belt region between 2000 and 2020. The distribution of prediction RMSE is illustrated in Fig. S7. Except for eastern North Dakota and northern Nebraska (both water-limited regions), no significant increase of RMSE was observed in regions beyond the "3I" states, indicating a credible out-of-sample performance for the surrogate.

3.1.2. Comparison of model responses to crop parameters

To achieve better spatial-temporal generalizability and transferability, the proposed GRU-based surrogate takes crop and soil parameters as input to learn the model responses to parameter variation. How the surrogate can reproduce parameter-induced responses consistent with the PB model quantifies the fidelity of this surrogate. Fig. 5 and Fig. S8 investigated the model responses to four crop parameters that are sensitive to GPP, LAI, and the final grain yield. CHL4 and VCMX are crop-specific sensitive parameters (Table 3), where CHL4 is only sensitive to corn and VCMX is only sensitive to soybean. Results show the responses of the GRU-based surrogate to the sensitive parameters agree well with the responses of *ecosys*. With the increase of CHL4 (for corn) and VCMX (for soybean), the yield responses of *ecosys* and its surrogate both sharply increased at first and then approached a flat, because higher leaf protein (CHL4 for corn) and rubisco activity (VCMX for soybean) increases the productivity of carbohydrates until photosynthesis is limited by light-dependent reactions. Plant maturity group (GROUPX: defined as the minimum number of vegetative nodes initiated before floral induction) influences the length of the growing season. Both *ecosys* and the GRU-based surrogate demonstrated that higher GROUPX postpones the phenological stage and produces higher GPP at the crop reproductive stage. Higher SLA means more leaf area with a fixed amount of leaf matter and thus significantly increases LAI. A discrepancy in the crop yield responses to SLA was observed between *ecosys* and the GRU-based surrogate (Fig. S8) due to the relatively low sensitivity to corn yield. These results indicate that the surrogate neural network learned the patterns of behavior and response to the changing environment and hence has high fidelity to the *ecosys*.

3.2. Site-level validation of the KGML-DA framework

3.2.1. Assimilating satellite-based GPP data

The performance of the KGML-DA framework for carbon budgets simulation (including GPP, ET, Reco, NEE, aboveground biomass, and LAI) was evaluated at three agricultural sites against the ground truth data (detailed in Section 2.4.1). To examine the effectiveness of the KGML-DA framework, we sequentially assimilated SLOPE GPP observations into the framework to update all target variables simultaneously (referred to as the full updating strategy). Compared with the open-loop simulation (i.e., the benchmark with no DA), the accuracies of almost all target variables were improved after assimilating in-season SLOPE GPP (Table 4). Specifically, the averaged root-mean-square error (RMSE) of the three sites for the six target variables decreased by 21.7%, 9.2%, 14.4%, 10.4%, 30.5% and 17.8% on average, and R^2 increased by 9.3%, 13.1%, 4.8%, 13.1%, 6.7% and 24.6%, respectively. The benefit of assimilating SLOPE GPP at US-Ne1 (i.e., continuous corn) is smaller than the other two sites, where the LAI prediction was not improved and the improvement for GPP, Reco, and biomass is slight. This is because the SLOPE GPP at US-Ne1 interfered with the signal from US-Ne2 (corn-soybean rotation) due to coarse spatial resolution, even if the pixels were not overlapped. Significant reduction in RMSE of biomass and LAI thanks to more accurate estimates of carbon fluxes. The updated LAI was fed into cell-2 of the neural network on the next simulation day to close this loop. In order to delve further into the behavior of priors (i.e., before DA) and posteriors (i.e., after DA), we conducted comparisons using predictions extracted on days when SLOPE GPP assimilation took place (Table S3). Results show the model performance of posteriors only slightly better than prior, indicating that there are no significant biases in the model's structure or the calibrated parameters (such biases could potentially result in prediction deviations when no assimilation takes place).

3.2.2. Passing information in the hierarchical structure

In Section 3.2.1, we listed all target variables in the state vector of t-EnKF to update them by their correlation with the observed GPP. However, updating a long list of state variables may jeopardize the simulation due to the poor or even spurious correlation between variables (Hu et al., 2019). One of the notable advantages of the hierarchical neural network is that the uncertainty of downstream variables will be automatically constrained after correcting the upstream variable. For example, assimilating GPP observations can benefit the downstream variables (e.g., Reco, biomass, and LAI) based on the hierarchical structure without putting the downstream variables into the state vector of the t-EnKF. Fig. 6 shows an example (from 2001 to 2009, US-Ne2 site) of the predicted trajectories of ensemble members after assimilating the upstream variable GPP. In this case, the Reco, biomass, LAI, and other variables simulated by cell-3 and cell-4 of KGML-DA were excluded from the state vector and not updated directly by the t-EnKF (referred to as partial updating strategy). Thus, the information for constraining the downstream variables only came from the hierarchical structure. Fig. 6 b-e show local details of the assimilation processes. Compared to the open-loop simulations (green lines), GPP simulated by KGML-DA (red lines) was significantly improved after assimilating SLOPE GPP product (purple cycles) (Fig. 6b). The updated hidden states of cell-2 (carrying the information of assimilated observation) were passed as the input to the cell-3 (the downstream cell) and made the predicted Reco and biomass closer to the ground truth (grey dots) (Fig. 6 c and d). Subsequently, the information carried by the hidden states of cell-3 was passed to cell-4 and the LAI predictions were improved (Fig. 6e). Notably, constraining the downstream variables via information flow from upstream variables makes the estimated trajectories smoother, whereas abrupt changes often observed when directly updating the target variables via covariance of the state vector (Fig. 6 d and e). We also compared the performance between the full updating strategy and the partial updating strategy (which excludes downstream variables of GPP). Results show that the partial updating strategy achieved better performance for the non-stressed US-Ne1 site (where the correlation between observations and state variables remains weak), as it propagates fewer observation errors by avoiding updating an extensive list of state variables via weak (even spurious) correlation. (Table S4).

3.3. Yield estimation of the US Corn Belt

The regional-scale performance of the proposed KGML-DA framework for yield estimation was evaluated over 600 counties in the Midwest of the US from 2000 to 2020 using the NASS surveyed county-level yield. The following sections investigated different DA strategies and the contribution of the in-season and historical data.

3.3.1. Updating model states via assimilating in-season data

For this scenario we assumed no historical data was available and the crop parameters for all counties were initialized by the default value. In other words, the crop parameters were homogeneous, and the variance of county-level yield estimations came from the climate forcing and the heterogeneity of soil properties (gSSURGO dataset). Fig. 7 a-d show the open-loop simulations that are only driven by climate data. The climate forcing explained 37.8% variation in yield estimation for corn and 31.9% for soybean (Fig. 7 a and c). After assimilating in-season GPP via t-EnKF, the R^2 improved to 0.521 for corn and 0.453 for soybean (Fig. 7 e and g). The accuracy of the multi-year averaged yield represents how well the model captured the spatial pattern of the yield. Although the R^2 of multi-year averaged yield was improved from 0.347 to 0.583 for corn and from 0.151 to 0.318 for soybean, the RMSE was not significantly reduced and the model overestimated yield for the low-yield county/

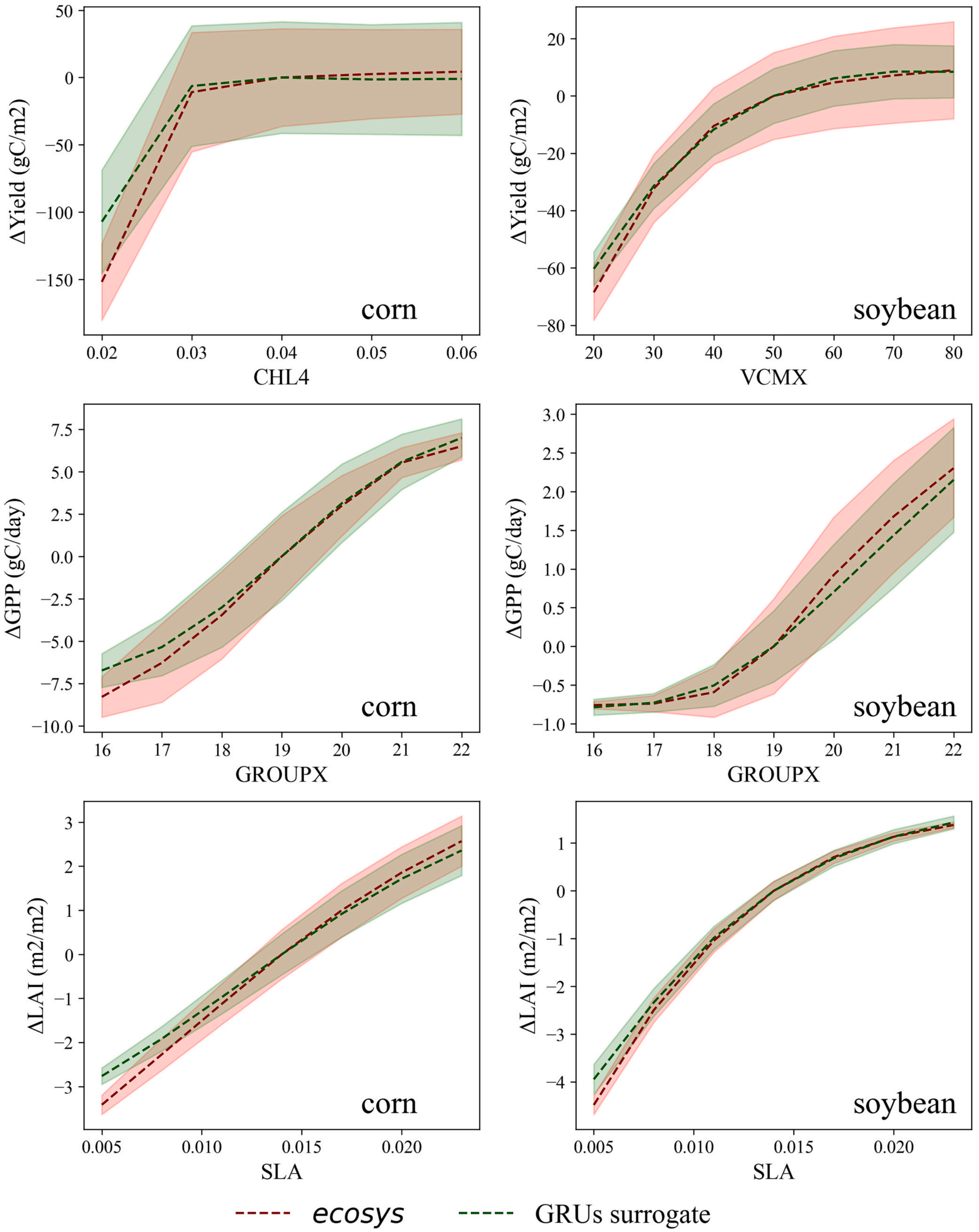

Fig. 5. Normalized responses of the surrogate neural network to the variation of four carbon-sensitive parameters. For plant maturity group (GROUPX), GPP responses were averaged during the reproductive stage (DOY 250 to 300); for the specific leaf area (SLA), LAI responses were averaged during the vegetative stage (DOY 190 to 220).

Table 4
Evaluating the proposed KGML-DA framework at three sites (US-Ne1, US-Ne2, and US-Ne3) with eddy-covariance fluxes observations and auxiliary ground truth data of aboveground biomass and LAI (from 2001 to 2007). For the t-EnKF cases, remotely sensed GPP was assimilated and all target variables are state variables in t-EnKF.

Target Variables	US-Ne1				US-Ne2				US-Ne3			
	Open-loop		DA		Open-loop		DA		Open-loop		DA	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
GPP (gC/m^2/day)	0.82	3.57	**0.85**	**3.19**	0.77	3.32	**0.88**	**2.34**	0.76	3.27	**0.83**	**2.45**
ET (mm)	0.57	1.60	**0.64**	**1.45**	0.53	1.56	**0.63**	**1.37**	0.48	1.49	**0.51**	**1.40**
Reco (gC/m^2/day)	0.86	1.43	**0.89**	**1.43**	0.80	1.84	**0.85**	**1.60**	0.79	2.15	**0.83**	**1.50**
NEE (gC/m^2/day)	0.64	3.10	**0.70**	**2.75**	0.61	2.67	**0.71**	**2.33**	0.61	2.33	**0.70**	**2.17**
Biomass (gC/m^2)	0.91	214.95	**0.92**	**162.35**	0.83	161.42	**0.93**	**117.56**	0.87	109.55	**0.93**	**66.01**
LAI (m^2/m^2)	**0.77**	**1.16**	0.74	1.40	0.58	1.85	**0.86**	**1.16**	0.60	2.33	**0.77**	**1.47**

Note: when a remotely sensed observation is assimilated on a given day, the updated value (i.e., posterior) is used. In cases where no assimilation occurs, the model's initial prediction is used.

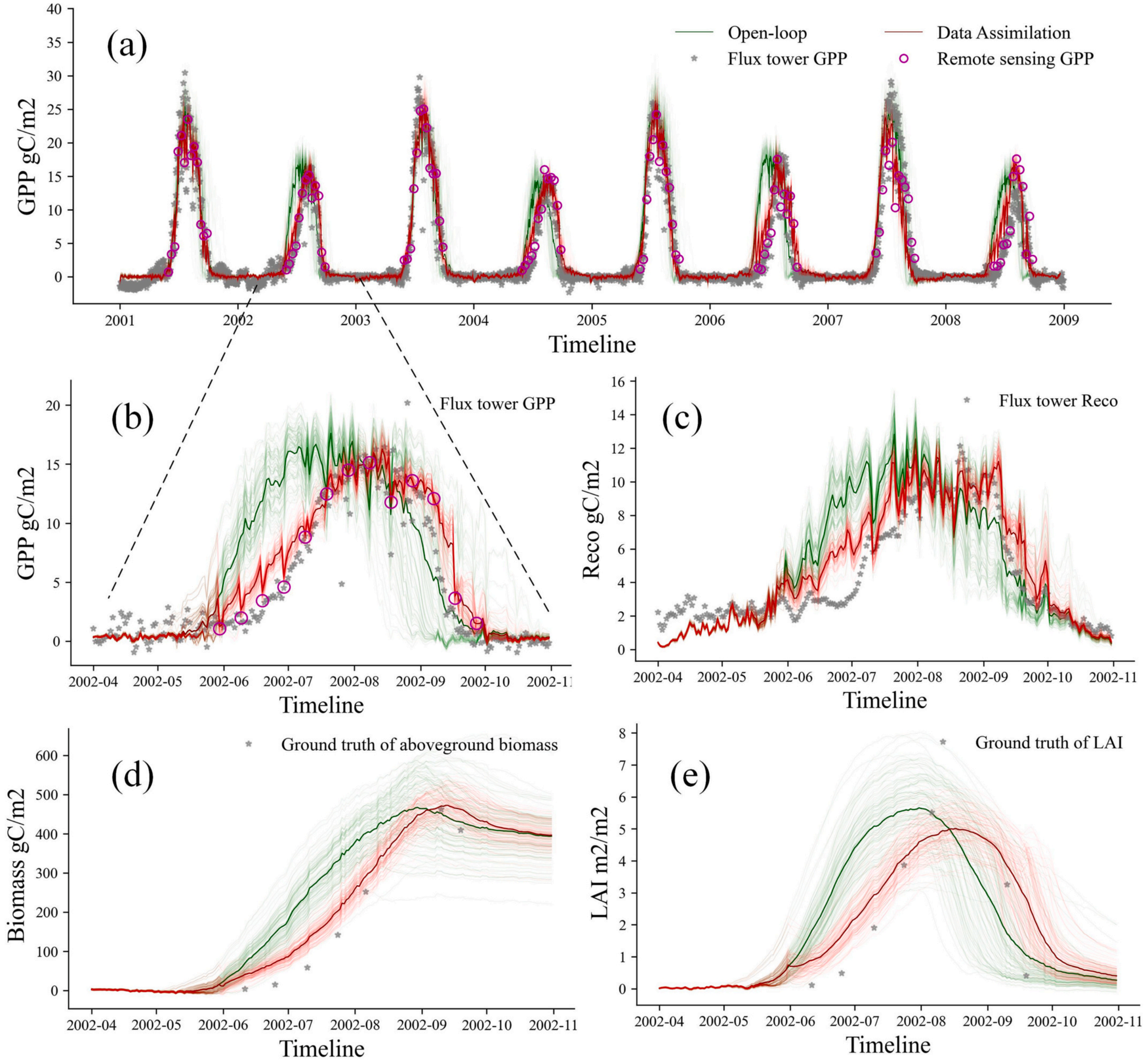

Fig. 6. Assimilating remotely sensed GPP observations improved the estimation of (a and b) GPP, (c) Reco, (d) aboveground biomass, and (e) LAI at the US-Ne2 site, Mead, NE.

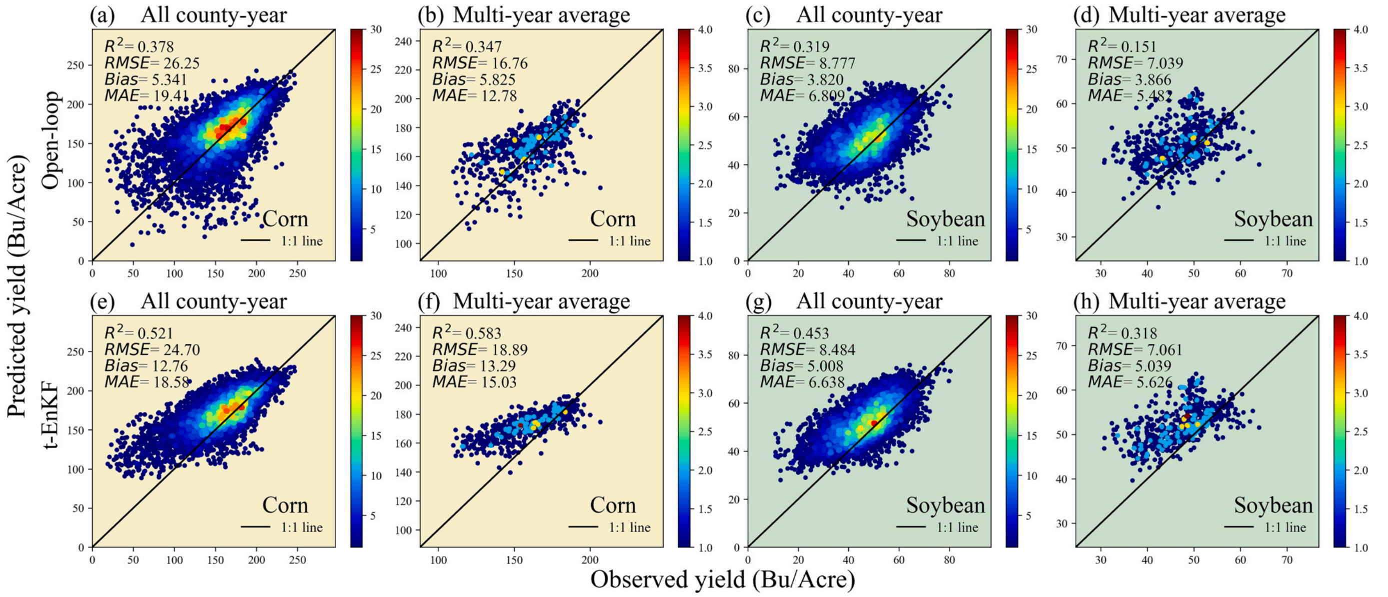

Fig. 7. Scatter plots of the predicted yield versus NASS yield using the default crop parameters: (a-d) open-loop simulations; (e-h) assimilating in-season GPP via t-EnKF. "All data" means all of the site-year data were evaluated. "Multi-year average" indicates that the temporal variation in yields has been averaged out, leaving only spatial patterns. The colour bar indicates the density of data points.

year (Fig. 7 b, d, f, and h). This result demonstrated that only assimilating in-season GPP mitigated the yield prediction error caused by uncalibrated parameters; however, the spatial yield pattern was not well captured.

3.3.2. Integrating prior knowledge via assimilating historical data

The spatial heterogeneity of model parameters can be derived from historical data and used as a prior for the in-season simulation. Fig. 8 evaluated the yield estimations using parameters calibrated by all available historical data (NASS yield and SLOPE GPP, described in Section 2.3.5.2) with a frequency of every three years. The parameter calibration is on top of the KGML-DA framework and prior to t-EnKF. Compared to the open-loop runs with default global parameters (Fig. 7 a-d), errors induced by parameter heterogeneity were addressed by assimilating historical observations (for corn and soybean, the R^2 of multi-year averaged yield was improved to 0.923 and 0.882; and the RMSE was decreased to 9.633 and 2.979, respectively). Additionally assimilating in-season GPP further improved model accuracy, with an R^2 of 0.620 and 0.632 and RMSE of 19.42 and 5.648 for corn and soybean, respectively (Fig. 8 e and g), suggesting that assimilating the in-season data can dynamically constrain the uncertainty induced by unknown or stochastic events during the growing season. However, assimilating in-season GPP did not improve the multi-year average (i.e., the spatial pattern) of the yield for the investigated counties (Fig. 8 f and h), likely because the spatial variation of yield is mainly interpreted by parameter spatial heterogeneity instead of the in-season aleatoric events.

Fig. 9 shows the multi-year average carbon budget predicted by the framework. The annual net biome productivity (NBP) was calculated by

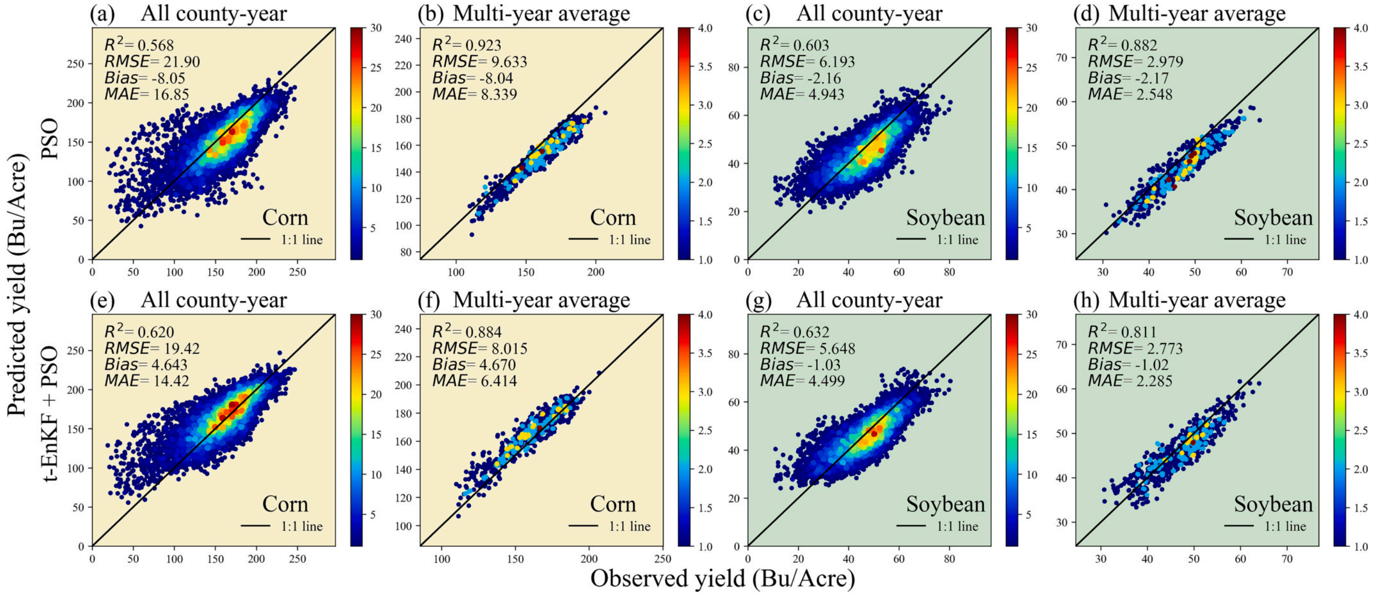

Fig. 8. Performance improved by the joint use of historical measurements (yield and GPP) and in-season GPP: (a-d) open-loop simulations with parameters calibrated by PSO algorithm using historical data; (e-h) utilizing all available data by integrating t-EnKF and PSO.

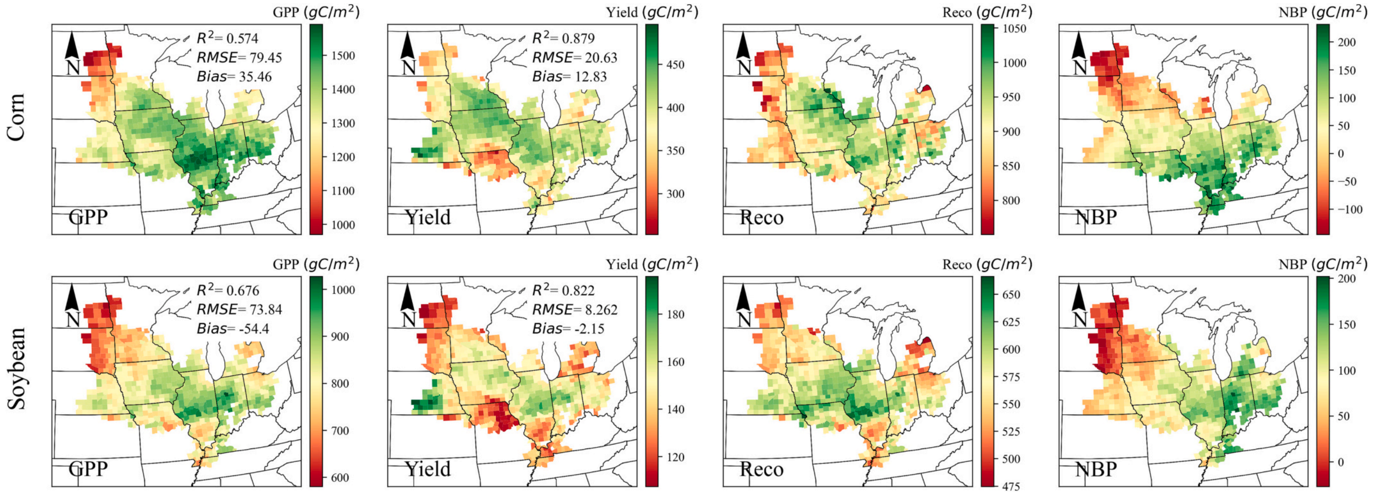

Fig. 9. Multi-year average carbon budget of the Corn belt area. NBP was calculated using the estimated GPP, Reco, and yield. Areas with higher NBP indicated more carbon sequestration. The metrics of GPP and yield were calculated based on the multi-year averaged SLOPE GPP and NASS yield.

deducing the crop yield and Reco from annually accumulated GPP. Negative NBP was observed in the northwest of the Corn Belt area (including east of North Dakota and South Dakota, and south of Minnesota), indicating that the agriculture in these areas is acting as a carbon source and thus there is huge potential to improve carbon sequestration through improving management.

3.3.3. *Effect of the frequency of parameter calibration on hybrid data assimilation*

Different lengths of parameter calibration period were investigated to evaluate the effect of parametric stability (Table 5). Lower bias was observed for cases with a short calibration period. This is caused by the linear yield trend correction described in Section 2.4.2.1. The adjusted yield increment increases as the calibration period extends, and thus the uncertainty introduced by the linear yield trend assumption is enlarged. Using parameters calibrated by the previous year's data to initialize the surrogate neural network produced the worst performance. This is because the calibrated parameters will be prone to overfit the specific scenario if only one year of data is available. In this circumstance, the parameters may fluctuate through time (yellow lines in Fig. 10) because the optimizer sacrificed the physical representation of these parameters to compensate for the model structure and input error. Cases utilizing all available historical data (progressively accumulated from the first year to the previous year, Fig. 3) to calibrate parameters produced the best results, with the R^2 higher than 0.57 for corn and 0.60 for soybean, and the RMSE lower than 21.91 for corn and 6.19 for soybean. The trends of parameters were more stable over time (black dashes in Fig. 10) in these cases, indicating that the influence of extreme events on parameters can be reduced by using long-term data to calibrate parameters. However, using long-term data to calibrate parameters that exhibit a temporal trend may cause a time-lag effect for the calibrated parameters (e.g., VCMX and STMX for soybean). Therefore, a hybrid method that uses different calibration periods for individual parameters should be investigated to improve the parameter calibration.

3.3.4. *Assimilating multi-source in-season remote sensing data*

The proposed DA framework reserved multiple interfaces to assimilate potential observations. Table 6 investigated the model performance with different combinations of multi-observations. For cases without considering parameter heterogeneity, assimilating only GPP or LAI sharply reduced the RMSE of yield estimation compared to the open-loop simulation for corn (reduced by 5.9% and 17.0%). Involving more types of observations tends to improve yield estimation. Assimilating in-season SLOPE GPP, MODIS ET, and GLASS LAI all together achieved the highest R^2 compared to other cases (0.53 for corn and 0.50 for soybean). Assimilating ET only slightly improved R^2 (0.42 and 0.37 for corn and soybean) but RMSE also increased (33.21 and 9.58 Bu/Acre for corn and soybean).

For cases using the calibrated parameters as initial, the historical ET and LAI data were not involved in the parameter calibration because they were systematically underestimated for the cropland. Assimilating in-season GPP and ET at the same time was only slightly better than assimilating in-season GPP with the R^2 of 0.63 for both corn and soybean and RMSE of 19.21 and 5.64 Bu/Acre for corn and soybean. Performance degradation was observed after only assimilating ET or LAI. This is probably because of the inconsistent pattern of LAI and ET between estimates and observations that may reverse the benefits from parameter calibration. GPP data already provides information related to photosynthetic rate, canopy pigments and water status. As a result, the information between SLOPE GPP, MODIS ET, and GLASS LAI overlapped and the benefits of additionally assimilating coarse ET and LAI are limited for yield estimation. However, the accuracy of yield estimation at the extreme year (2012) was significantly improved after additionally assimilating ET and LAI (Table S5). This is because the model tends to

Table 5
Effect of parameter calibration period on KGML-DA. Three calibration periods were investigated which are the previous year, the previous three years, and all historical data (depicted in Fig. 3).

DA method	Calibration period	Corn			Soybean		
		R^2	RMSE	Bias	R^2	RMSE	Bias
PSO	Previous year	0.36	28.31	**−3.66**	0.41	7.66	**−1.00**
PSO	Previous three years	0.49	23.78	−5.96	0.53	6.77	−2.16
PSO	All historical data	**0.57**	**21.91**	−8.05	**0.60**	**6.19**	−2.16
PSO + t-EnKF	Previous year	0.41	25.70	4.61	0.46	6.88	**−0.04**
PSO + t-EnKF	Previous three years	0.55	21.04	**3.99**	0.56	6.21	−1.31
PSO + t-EnKF	All historical data	**0.62**	**19.43**	4.64	**0.63**	**5.65**	−1.03

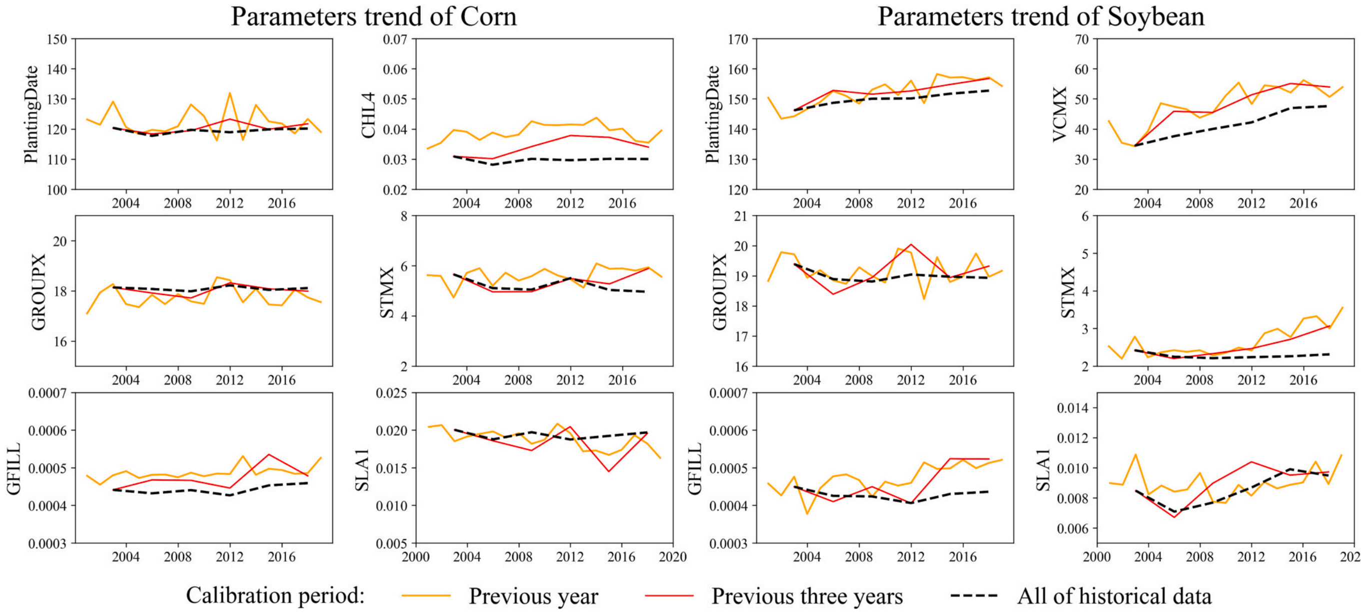

Fig. 10. Evolvement of the mean value of key crop parameters with different lengths of the calibration period. The yellow lines, red lines, and black dash represent the parameters that were calibrated by data from the previous year, the previous three years, and all available historical data, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
Model performance with different combinations of multi-observations for all county-year.

Initial parameters	Observation types for t-EnKF	Corn			Soybean		
		R^2	RMSE	Bias	R^2	RMSE	Bias
Default parameters	Open-loop	0.38	26.25	5.34	0.32	8.78	3.82
	GPP	**0.52**	24.70	12.77	0.45	8.48	5.01
	ET	0.42	33.21	23.62	0.37	9.58	6.01
	LAI	0.50	**21.80**	**3.11**	0.40	8.91	5.25
	GPP, ET	0.52	23.85	11.00	0.46	**7.70**	**3.69**
	GPP, ET, LAI	**0.53**	23.28	10.07	**0.50**	7.87	4.46
Parameter calibrated by PSO	Open-loop	0.57	21.91	−8.05	0.60	6.19	−2.16
	GPP	**0.62**	19.43	4.64	**0.63**	5.65	−1.03
	LAI	0.43	25.16	−5.07	0.54	6.25	−0.37
	ET	0.51	22.51	4.30	0.56	6.46	−1.92
	GPP, ET	**0.63**	**19.21**	**4.25**	**0.63**	**5.64**	**−1.17**
	GPP, ET, LAI	0.46	24.05	4.15	0.56	6.14	−1.03

overestimate yield under extreme stresses and the joint use of all available data can improve the estimation of covariance matrices of state variables. To further investigate the value of assimilating potential ET and LAI observation with higher resolution and accuracy, we conducted an Observing System Simulation (OSS) experiment (Curnel et al., 2011) using synthetic observations without error (Fig. S9). Fig. S10

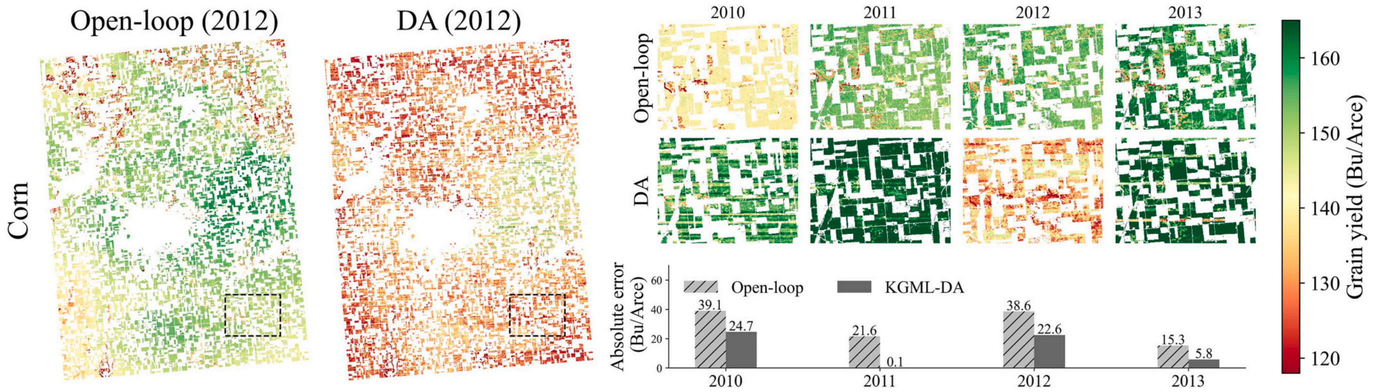

Fig. 11. 30-m yield map of corn estimated by open-loop and KGML-DA. The left subplots are the overall yield map of Champaign County, IL. The top right subplots are the local view of the temporal variation of yield estimates in a selected area (black dash rectangles). Bottom right is the absolute error of averaged yield estimation compared to the NASS-reported county scale yield. The result for soybean is shown in Fig. S12.

shows that the estimated yield agrees well with the observed yield even though only GPP was assimilated. Additional assimilation of ET did not significantly help improve the yield estimation but assimilating LAI helps correct the phenology and the yield in the early stage is significantly improved.

3.4. Pixel-level yield mapping

To demonstrate the effectiveness of the proposed KGML-DA framework for subplot-level simulation, we mapped the 30-m corn and soybean yield of Champaign County, Illinois, from 2010 to 2013 (Fig. S11–12 and Fig. 11). The left subplots of Fig. 11 show the spatial variability of the estimated yield in 2012, when the Midwest experienced an extreme drought. Although accumulated precipitation of Champaign County during the growing season of 2012 was not significantly reduced (Fig. S3b), the severe drought happened in the pre-season and July decreased the initial SM content and affected the pollination of corn (Fig. S3a), which resulted in a low reported corn yield (108.9 Bu/Arce). The open-loop yield map of 2012 is significantly overestimated due to the model parameters not calibrated for extreme events (the parameters calibrated by historical data), and its spatial variability only comes from weather and soil properties.

In contrast, the signal of the in-season crop condition and the extreme event could be captured by the 30-m SLOPE GPP data, and assimilating this data reduced the bias in yield estimation (Fig. 11 shows the absolute error of county-level corn yield was reduced by 9.5–21.5 Bu/Acre). To illustrate the details of the temporal and spatial variations of the estimated yield map, Fig. 11 also shows a zoomed-in view of a selected area in Champaign County. Compared to the open-loop simulation, the KGML-DA pulled up the underestimated corn yield (in 2010, 2011, and 2013) and suppressed the overestimated corn yield (in 2012) using the in-season information coded in the 30 m GPP. For soybean, the result shows less benefit from assimilating 30 m GPP for yield estimation (Fig. S12). In 2012, the reported soybean yield of Champaign County was not significantly affected by the drought. This is because the stomatal conductance and grain yield of soybean show less sensitivity to the atmosphere condition than corn (e.g., VPD) (Lobell et al., 2014; Gray et al., 2016), and the relatively late planting of soybean bypassed the period with severest drought (during May to Jul., of 2012). However, the observed soybean canopy greenness and GPP was reduced because the plant transited energy from leaves to root and grains to resist drought. As a result, assimilating the low GPP observation dragged down the estimated soybean yield in 2012 even if the real yield was not significantly affected (Fig. S12).

3.5. Uncertainty analysis of county-level yield prediction

3.5.1. Spatial-temporal effects of yield prediction error

The total yield prediction error (627 counties from 2003 to 2020) was partitioned into three components, including the global mean μ_g, spatial effect μ_s, and temporal effect μ_t, by a hierarchical Bayesian model (Dokoohaki et al., 2021). The posterior distribution of μ_g, μ_s, and μ_t was approximated by the MCMC algorithm. The sum of the expectation of μ_g and μ_s represent the spatial pattern of the prediction error (visualized in Fig. 12), which is important for modelers to improve the understanding of the model responses to various climate and environmental conditions. Compared to the open-loop simulations, the prediction error was reduced after assimilating in-season data (Fig. 12 b1 and f1) in the south of Minnesota, Iowa, Illinois, and Indiana (significantly reduced for corn and slightly reduced for soybean). However, the prediction error in the northwest and southwest region of the Corn Belt remains high mainly due to the poor consistency of spatial pattern between SLOPE GPP and NASS yield in these areas (e.g., the low yield was observed in north Missouri but GPP is moderate, Fig. S13). The northwest regions have a low annual air temperature and precipitation and the southwest regions may suffer water deficit due to high evaporation demand (Zhou et al., 2021b). As a result, the predicted yield was biased in these regions under the assumption of homogeneous crop parameters. Compared to the flat histograms in the cases with default parameters, calibrating parameters by the historical data eliminates the spatial effect and the prediction

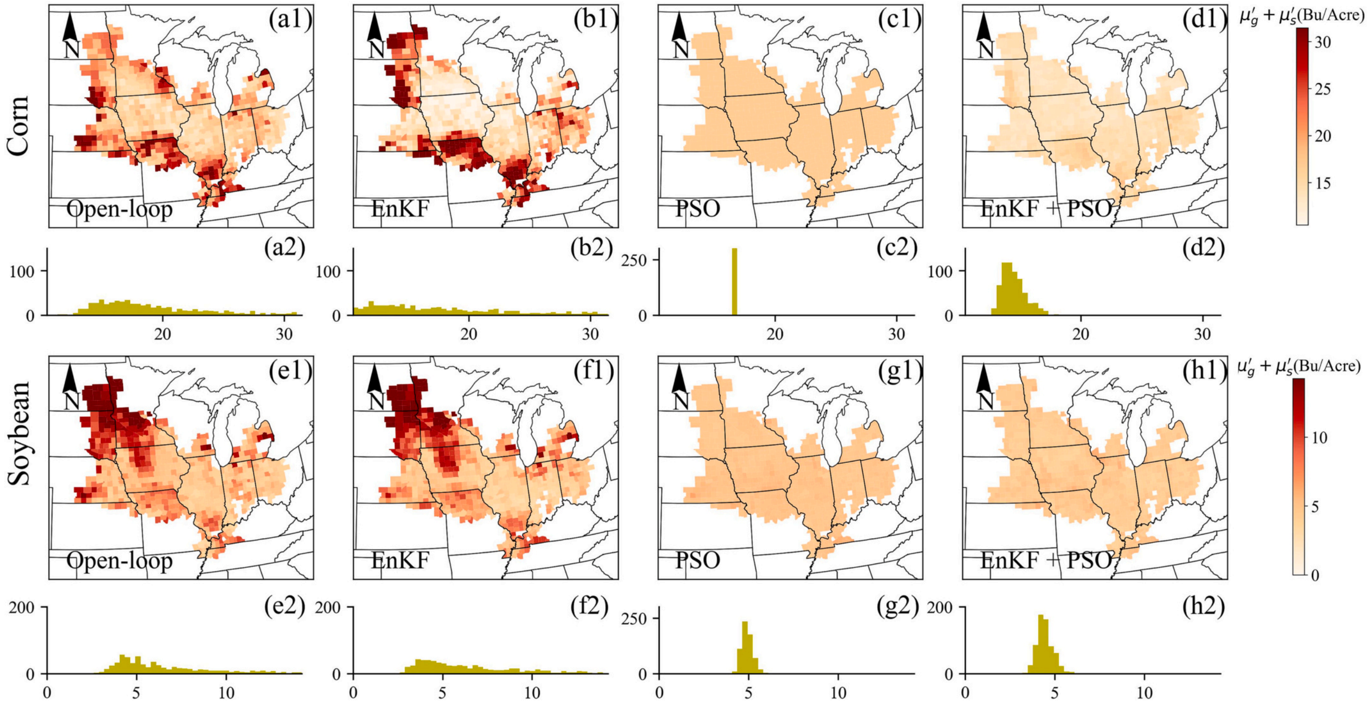

Fig. 12. The spatial distribution (a1-h1) and histogram (a2-h2) of the yield prediction errors (sum of the posterior global mean $\mu_g^{'}$ and spatial effect $\mu_s^{'}$): (a and e) open-loop simulations; (b and f) assimilating in-season GPP with default parameters; (c and g) open-loop simulations with calibrated parameters; (d and h) assimilating both in-season and historical data. $\mu_g^{'}$ and $\mu_s^{'}$ are the expectations of μ_g and μ_s, respectively.

errors of all counties are homogeneous, resulting in a more concentrated histogram (Fig. 12 c1–2 and g1–2). Assimilating both historical and in-season data via t-EnKF and PSO increased the variation of errors due to the introduction of the in-season observation uncertainty (Fig. 12 d2 and h2). However, it also produced the lowest mean prediction error and the hotspot areas with high prediction error were also addressed (Fig. 12 d1 and h1).

μ_t described the temporal pattern of the prediction error. Fig. 13a shows the reduction in temporal variability of the prediction error (standard deviation of $\mu_t^{'}$) compared to open-loop after DA. Assimilating in-season data significantly improves the temporal robustness of the model, with an 11.9% reduction of std. for corn and a 37.9% reduction for soybean. The reductions of $\mu_g^{'}$ (the expectation of μ_g) were depicted in Fig. 13b. For corn and soybean, the error was reduced by 12.5% and 26.9% after assimilating historical data and further reduced by 14.0% and 9.3% by assimilating in-season data. Leveraging the information from both historical and present in-season data, the reduced priori error was 26.5% for corn and 36.2% for soybean. The result validated the effectiveness of the proposed DA framework and indicated the necessity for assimilating both historical and in-season data to reduce uncertainty in yield prediction.

3.5.2. Uncertainty reduction after assimilating in-season remote sensing data

The variance of the predicted yield of the t-EnKF ensemble members was partitioned into global mean, spatial effect and temporal using the same Bayesian hierarchical model. The temporal effect of the uncertainty in yield estimation after assimilating different types of observations was visualized in Fig. 14. Compared to the open-loop simulation, the uncertainty of yield prediction was reduced by 6.8–11.3 Bu/Acre for corn and 1.1–2.0 Bu/Acre for soybean for the uncalibrated cases after assimilating in-season GPP. Parameter calibration altered the spatial pattern of yield prediction uncertainty (Fig. S14). This is because the model responses to the parameters are nonlinear and the initial values of the parameters significantly affect the prediction uncertainty. No significant difference was observed in the magnitude of the yield estimation uncertainty between only assimilating in-season GPP and assimilating GPP and ET, indicating that ET has a limited contribution to reducing uncertainty. In contrast, assimilating LAI further reduced the uncertainty especially for soybean, with a range of 2.3–3.0 Bu/Acre for the uncalibrated case and 2.6–3.7 Bu/Acre for the calibrated case. LAI is more influential for soybean yield because soybean tends to partition more dry matter to grow leaves. Although the uncertainty was reduced by assimilating remote sensing LAI, the yield prediction error was not effectively reduced (Table 6), likely because of the inconsistent magnitude between estimates and observation, the issue of mixed pixels, and the challenges to quantifying observation noise. Hence, there is a need to develop unbiased LAI products for croplands with a high spatial resolution to reduce the uncertainty in yield estimation.

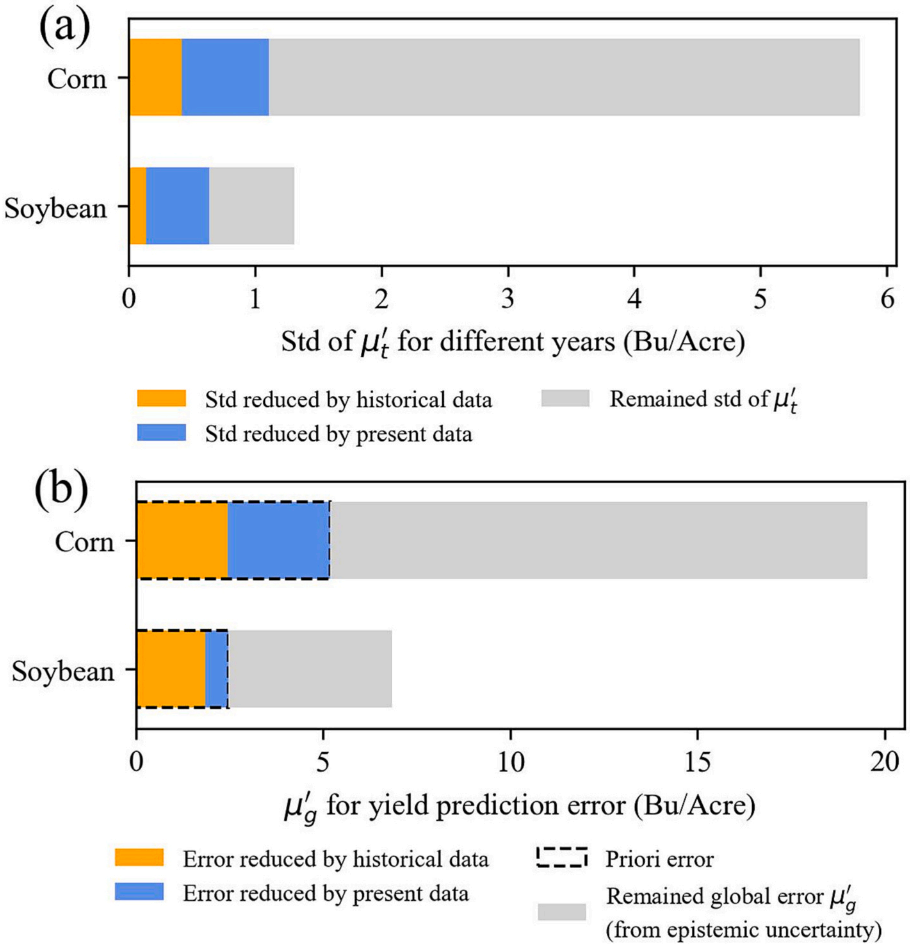

Fig. 13. Partitioning the contributions of historical and in-season data for the reduction of yield prediction error: (a) the standard deviation of $\mu_t^{'}$. A higher value indicated strong interannual fluctuation of prediction error; (b) $\mu_g^{'}$ for yield prediction error. $\mu_t^{'}$ is the expectation of μ_t.

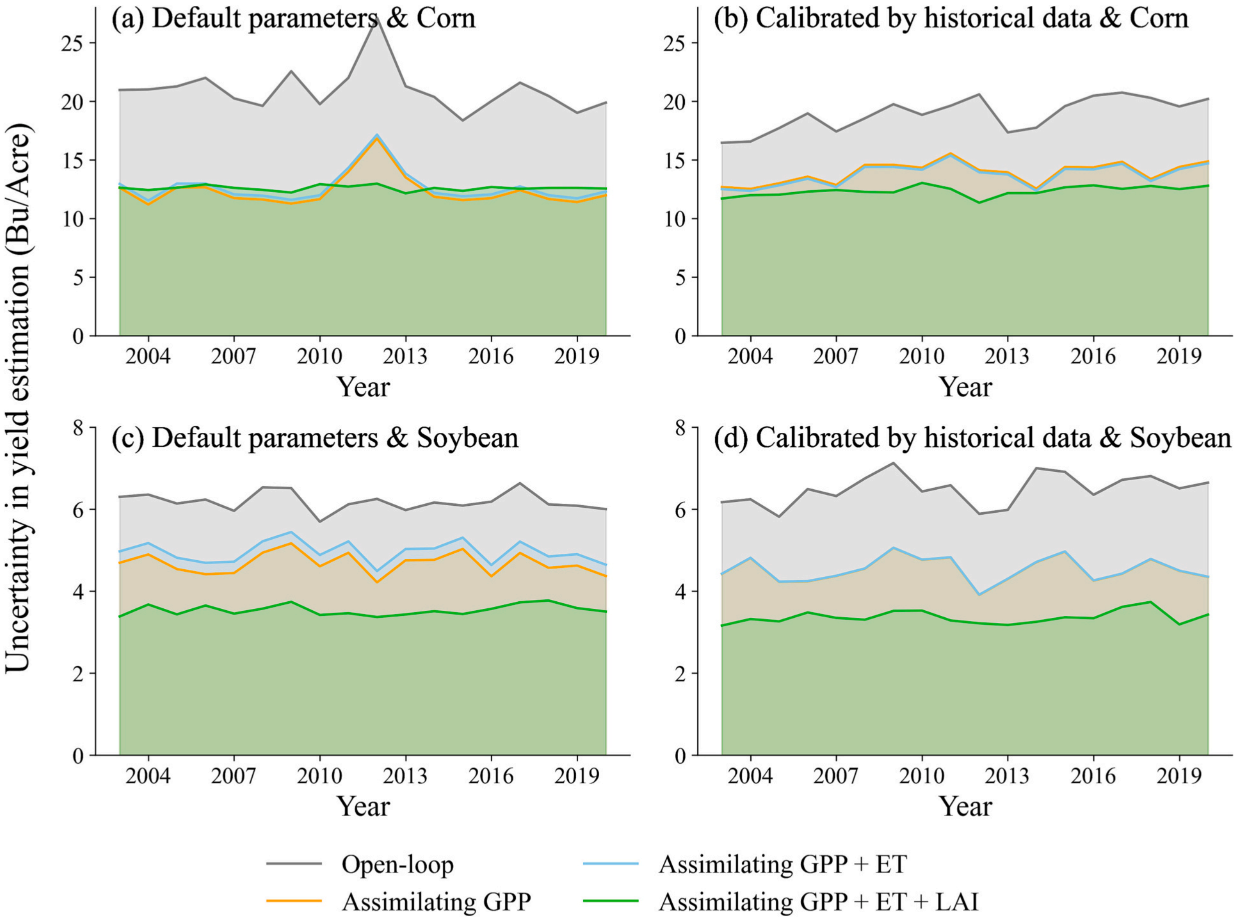

Fig. 14. Temporal evolution of the uncertainty in yield estimation with different assimilation strategies: (a and c) model initialized with default parameters; (b and d) model initialized with parameters calibrated by historical data.

4. Discussion

4.1. Data interfaces for multi-source observation

Nine target variables (including phenology, GPP, ET, SM, Reco, NEE, aboveground biomass, LAI, and yield) were simulated by the proposed surrogate neural network and each of them acts as a reserved data interface to ingest possible observations. In this study, we only investigated three remotely sensed observation types (i.e., GPP, ET, and LAI) for regional crop growth simulation. Results demonstrated that the improvement by assimilating in-season GPP was greater than ET and LAI. As an upstream variable in the hierarchical neural network, GPP is a compound variable associated with processes of light inception (affected by LAI) and the exchange of water and carbon (related to ET), thus GPP already includes information related to ET and LAI. Information from the assimilated GPP observation can be passed to every downstream variable to benefit the whole neural network. In addition to GPP, thriving earth observation technologies and novel algorithms offer possibilities to activate other upstream data interfaces, such as SM. Microwave-based remote sensing technologies are considered to be the most effective way to monitor global SM. However, the spatial resolution is coarse for SM products derived from the L-band passive radiometer (e.g., 36 km for the products of Soil Moisture Active Passive, SMAP; and the Soil Moisture and Ocean Salinity, SMOS) (Kerr et al., 2010; Entekhabi et al., 2010). Although the resolution can be downscaled to 1–10 km by integrating data from synthetic aperture radars (SAR) (Meyer et al., 2022), it is still too coarse to distinguish individual crop fields. Recent studies demonstrated that fine-scale (up to 30 m) SM with multiple layers could be predicted by ML using multi-source data such as weather, thermal, and topographic data (Zeng et al., 2019; Vergopolan et al., 2021). And inferring subgrid-scale SM profile from surface observation is promising via integrating PB models, ML, and DA techniques (Heathman et al., 2003; Kornelsen and Coulibaly, 2014; Feng et al., 2022). Meanwhile, the forthcoming NISAR (NASA-ISRO Synthetic Aperture Radar, planned to launch in 2024) mission is set to feature an advanced L-band instrument with a resolution of 3–10 m, making it particularly suitable for achieving high-resolution mapping of SM (Kellogg et al., 2020). Therefore, a comprehensive SM extension is worth being developed in future research to constrain the water cycle of the simulation by assimilating fine-scale surface SM data. Other than that, deriving management practices are critical for further constraining the predictive uncertainty. While extracting practices like planting and harvesting dates, tillage, and cover crop adoption from remote sensing data shows promise, the detection of certain crucial practices such as irrigation, drainage, and fertilizer use remains challenging (Guan et al., 2023).

4.2. Simulating hard-to-observe variables in the agroecosystem

The designed DA framework is general and it can be easily adapted to specific tasks without changing the framework structure. This advantage grants researchers high-level of flexibility to customize their own target variables (Especially variables that are difficult to observe at the regional scale) for each GRU cell. For instance, partition ecosystem respiration into autotrophic respiration (Ra) and heterotrophic respiration (Rh), and simulate evaporation (E) and transpiration (T) separately. Monitoring the components of Reco and ET is critical for understanding the natural ecosystem behaviors in the context of climate change (Xu et al., 2021). *Ecosys* provides hourly simulation for these components; however, in-situ measurement methods such as chamber (measure Rh) lysimeters (measure E), sap flow meters (measure T), and isotope (measure E, T, Ra, and Rh) are expensive and difficult to upscale to regional scale (Perez-Priego et al., 2017; Welp et al., 2008). Dynamic simulation of the components of Reco and ET can be achieved by PB ecosystem models (e.g., *ecosys*). However, there are still two main stumbling roadblocks in simulating sub-fluxes of Reco and ET: 1) large model structural uncertainty caused by a lack of understanding of the interactions in the SPAC system and the trade-off between computation efficiency and model complexity; 2) large uncertainty in input and parameter due to the scarcity of ground truth data and the heterogeneity of land surface. The proposed KGML-DA framework provides a promising path to address these challenges by balancing the prior knowledge and multi-source data. Specifically, the parametric uncertainty can be constrained by assimilating historical data and the overall uncertainty can be reduced by assimilating the in-season data. The mass balance of the decomposed components can be closed by a residual method (assuming one of the fluxes in the balance equation is a residual term) or additionally constrained by a mass balance loss. Except for assimilating remote sensing data, point-level ground truth data can be used to fine-tune the neural network and reduce the model structural uncertainty. Due to the developed framework being capable of assimilation of real-time data, the model performs at optimal and thus produces future projections (using forecasted weather data) with the lowest uncertainty. Therefore, it is also promising to deploy the proposed framework for projecting the future carbon budget and provide guidance to optimize field management.

4.3. Computational cost and large-scale spatial-temporal downscaling

The extremely high computation cost limits the implementation of a traditional PB model-based (e.g., *ecosys*) DA system for large-scale simulation. The computation demands increase exponentially if higher spatial resolution is requested, making regional simulation nearly impossible at the subplot level. This study upgraded the t-EnKF with the 3-D tensor operation so that the ensemble members for different sites (or pixels) can be simulated parallelly. The shift of computation from CPU-intensive to GPU-intensive greatly reduced the simulation time. A one-year run of *ecosys* (i.e., the process-based model we used in this study) takes 30 s for one site/pixel. While this run-time may be acceptable for simulations conducted at the county level (Zhou et al., 2021a, 2021b; Yang et al., 2022), the cumulative run-time will be astronomical for large-scale pixel-level simulations. For example, assuming the ensemble number is 100, at least 4 billion CPU hours are required for the traditional method to accomplish a twenty-year simulation on the cropland of 3I states with a 30 m resolution (about 250 million pixels). In contrast, the surrogate takes around 0.4 s per site for a one-year simulation. However, the magic is that ensemble members from different sites can be merged into the batch dimension, so a large number of cases can be run in parallel on the GPU. The simulation time could be sharply reduced by 3.75 million times on only one GPU with a batch of 500 × 100 (pixels×ensemble members), making it more than 7000 times faster than a traditional DA framework run on a High-Performance Computing (HPC) with 512 CPU cores. The computation time could be further reduced by deploying the framework on high-performance GPU clusters with a larger GPU memory so that a larger batch size can be implemented for one inference.

The remarkable computation efficiency makes the KGML-DA framework an effective downscaling tool for performing a fine-scale simulation with high spatial-temporal resolution. Multi-source remote sensing data with different resolutions can be tackled by hierarchical simulation at different scales. Specifically, simulation can be run at the county level, 500 m, 250 m, and 30 m respectively to assimilate multi-source data. The coarse-level simulation provides regional means of the target variables to the fine-level simulation as prior knowledge. The fine-level simulation explains the spatial variation on the basis of the mean value from coarse-level simulation. Observations with different time intervals constrain the temporal pattern of the target variables at different scales. For example, the value of fine-scale estimates at the observation date is directly constrained by high spatial resolution data but with relatively long-time intervals (e.g., 30 m, 16-day products). On the other hand, the temporal trend of the fine-scale simulations can also be indirectly constrained by the more frequent but coarse data. As a result, the downscaled estimates benefit from both the intertwined multiscale data.

4.4. Exploring more flexible data assimilation approaches for agroecosystem

The KGML-DA framework employed both PSO (i.e., batch DA method) and t-EnKF (i.e., sequential DA method) to maximize the utility of historical data and in-season present data. Historical data is valuable for predicting the future because a variable in the past and future may follow the same distribution or a particular pattern. Conversely, in-season data is essential in addressing uncertainties arising from unknown inputs and unforeseen events. The results of county-level yield estimation (Section 3.3.2) and error analysis (Section 3.5.1) both indicated that assimilating historical data outperforms open-loop simulation by a large margin, demonstrating that the batch DA approach is a fast solution to improve model performance. However, as the frequency of climate-induced extremes increases, assimilating historical data alone may bias predictions for extreme years, and assimilating both historical data and present data is the only way to keep the prediction on track.

For the sequential DA method, we opted for the widely used EnKF to assimilate multisource observations into the GRU-based surrogate due to its simplicity and effectiveness. However, inherent limitations of EnKF, such as its linear updating rule and the Gaussian assumption of state variables and observation errors, could potentially undermine its performance when applied to highly non-linear systems (Abbaszadeh et al., 2019). To address this challenge, the particle filter (PF), a non-parametric Bayesian filter, garnered attention among modelers due to its suitability for addressing non-linear and non-Gaussian systems (Moradkhani et al., 2005a; Jiang et al., 2014). Recent advancements have led to the development of enhanced PF algorithms, specifically tailored to mitigate the particle degeneracy issue found in traditional PF methods. For example, the Evolutionary Particle Filter with MCMC (EPFM) integrates a Genetic algorithm and MCMC to enhance particle diversity (Abbaszadeh et al., 2018; Gavahi et al., 2020). As a result, advanced DA algorithms that are more flexible to handle state variables and observations with arbitrary distributions are worth being integrated into our framework.

In this study, we designed a DVAE to assimilate in-season remote sensing data (elaborated in Section 2.3.3). Compared to previous work (Zhou et al., 2021a, 2021b) that utilizing historical and future data to calibrate parameters of the *ecosys* model (i.e., even years for parameter calibration and odd years for validation), the county-level yield predicted by our framework achieved similar accuracy without using the future data (we are doing temporal extrapolation instead of interpolation), indicating the effectiveness of assimilating in-season observations. We also conducted a comparison with the work of Kang and Özdoğan (2019), who integrated PB models with EnKF for yield estimation in the Corn Belt and validated by field-level yield data. Our approach exhibited lower RMSE values compared to theirs, as they reported RMSE values ranging from 1.4 ton/ha to 2.3 ton/ha (equivalent to 20.8 Bu/Acre to 34.2 Bu/Acre). One major limitation of the proposed method is the potential for information leakage once the hidden state has been reconstructed, owing to the presence of reconstruction errors. One possible strategy to bypass the hidden state reconstruction is to directly update the hidden state. For example, Guen and Thome (2020) devised a DA framework for video prediction which directly updated the hidden state of RNN. However, directly implementing this method into an agroecosystem model is not feasible due to its assumption of linear additivity in hidden state transitions and the limited number of state variables in their context. Consequently, the pursuit of more robust data assimilation approaches founded on deep neural networks warrants further investigation in forthcoming endeavors.

5. Conclusion

This study established a novel KGML-DA framework that is capable of assimilating historical and in-season multi-source data. As far as we know, this is the first attempt to implement both sequential and batch DA on a hierarchical surrogate neural network. This study demonstrated a paradigm of KGML-DA framework to reduce uncertainty in multi-variable simulations by leveraging knowledge from the PB model and multi-source data. The causal relationships between target variables were hardcoded into the hierarchical surrogate neural network to streamline the inference flow and automatically constrained all connected variables. The responses of the surrogate to driving data and parameters were examined to make sure it was competent for replacing the *ecosys*. This framework was first evaluated at three agricultural sites and then tested in the Midwest Corn Belt for county-level yield estimation and 30-m yield mapping. Additionally, uncertainty and error of estimated yield were analyzed. The result shows updating the upstream variable (e.g., GPP) improved the prediction of downstream variables (e.g., Reco, NEE, biomass, and LAI) at three agricultural sites, indicating the importance of the hierarchical structure for building the linkages between target variables. For county-level yield estimation, assimilating historical data addressed the parameter uncertainty and captured the heterogeneous pattern of crop parameters. Involving all historical data to calibrate parameters mitigated the effect of extreme events and thus produced stable parameters. Initializing the surrogate neural network with the calibrated parameters significantly improved the estimation of multi-year averaged yield (R^2 was improved from 0.367 to 0.923 for corn, and from 0.151 to 0.882 for soybean). Further assimilating in-season data on the basis of the calibrated parameters achieved the best performance by addressing the uncertainty induced by the stochastic (or unknown) events. Results indicated that the proposed KGML-DA framework is capable of accurately and dynamically simulating various variables in the agroecosystem, and it is potentially more than 7000 times faster than the PB model. This framework is not subject to a certain task (e.g., yield estimation) but can be extended to simulate the variables that are difficult to observe at the regional scale and to downscale the remote sensing observations to higher spatial-temporal resolution.

Declaration of Competing Interest

None.

Data availability

Data will be made available on request.

Acknowledgement

The authors acknowledge support from the Advanced Research Projects Agency–Energy (ARPA-E), US Department of Energy, SMART-FARM program (award no. DE-AR0001382), the National Science Foundation Signal in Soil Program (award no. 2034385), the National Science Foundation III: Medium (award no. 2313174), the Foundation for Food & Agriculture Research (FFAR) Seeding Solution Award, and the Director, Office of Science, Office of Biological and Environmental Research of the US Department of Energy (contract no. DE-AC02-05CH11231). We thank Jessica Till for helpful suggestions and comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2023.113880.

References

Abbaszadeh, P., Moradkhani, H., Yan, H., 2018. Enhancing hydrologic data assimilation by evolutionary particle filter and Markov chain Monte Carlo. Adv. Water Resour. 111, 192–204.

Abbaszadeh, P., Moradkhani, H., Daescu, D.N., 2019. The quest for model uncertainty quantification: a hybrid ensemble and variational data assimilation framework. Water Resour. Res. 55 (3), 2407–2431.

Ball, J.T., Woodrow, I.E., Berry, J.A., 1987. A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions. In: Biggins, J. (Ed.), Progress in Photosynthesis Research: Volume 4 Proceedings of the VIIth International Congress on Photosynthesis Providence, Rhode Island, USA, August 10–15, 1986. Springer Netherlands, Dordrecht, pp. 221–224.

Bauer, P., et al., 2021. The digital revolution of Earth-system science. Nat. Comp. Sci. 1 (2), 104–113.

Bertino, L., Evensen, G., Wackernagel, H., 2003. Sequential data assimilation techniques in oceanography. Int. Stat. Rev. 71 (2), 223–241.

Brajard, J., et al., 2020. Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the Lorenz 96 model. J. Comput. Sci. 44, 101171.

Chang, K.-Y., et al., 2019. Methane production pathway regulated proximally by substrate availability and distally by temperature in a high-latitude mire complex. J. Geophys. Res. Biogeosci. 124 (10), 3057–3074.

Chen, Y., et al., 2018. Improving regional winter wheat yield estimation through assimilation of phenology and leaf area index from remote sensing data. Eur. J. Agron. 101, 163-17.

Chung, J., et al., 2014. 'Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling', *arXiv [cs.NE]*.

Cintra, R.S., de Campos Velho, H.F., 2018. Data assimilation by artificial neural networks for an atmospheric general circulation model. Adv. Appl. Artificial Neural Networks 265.

Cuomo, S., et al., 2022. 'Scientific Machine Learning through Physics-Informed Neural Networks: Where we are and What's next', *arXiv [cs.LG]*.

Curnel, Y., et al., 2011. Potential performances of remotely sensed LAI assimilation in WOFOST model based on an OSS experiment. Agric. For. Meteorol. 151 (12), 1843–1855.

Dokoohaki, H., et al., 2021. A comprehensive uncertainty quantification of large-scale process-based crop modeling frameworks. Environ. Res. Lett. 16 (8), 084010.

Dold, C., et al., 2017. Long-term carbon uptake of agro-ecosystems in the Midwest. Agric. For. Meteorol. 232, 128–140.

ElGhawi, R., et al., 2022. 'Hybrid Modelling of Land-Atmosphere Fluxes: Estimating Evapotranspiration using Combined Physics-Based and Data-Driven Machine Learning', in. ui.adsabs.harvard.edu pp. EGU22–9890.

Entekhabi, D., et al., 2010. The Soil Moisture Active Passive (SMAP) Mission. Proc. IEEE 98 (5), 704–716.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. C: Oceans 99 (C5), 10143–10162.

Feng, D., Tan, Z., He, Q., 2022. 'Physics-informed neural networks of the Saint-Venant equations for downscaling a large-scale river model', *arXiv [physics.flu-dyn]*.

Gan, Y., et al., 2014. Improving farming practices reduces the carbon footprint of spring wheat production. Nat. Commun. 5, 5012.

Gavahi, K., et al., 2020. Multivariate assimilation of remotely sensed soil moisture and evapotranspiration for drought monitoring. J. Hydrometeorol. 21 (10), 2293–2308.

Ghosh, P., et al., 2019. 'From Variational to Deterministic Autoencoders', arXiv [cs.LG].

Ghosh, R., et al., 2022. Robust inverse framework using knowledge-guided self-supervised learning: an application to hydrology. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery (KDD '22), New York, NY, USA, pp. 465–474.

Grant, R.F., 2001. A review of the Canadian ecosystem model ecosys's. In: Modeling carbon and nitrogen dynamics for soil management.

Grant, R.F., et al., 2006. Intercomparison of techniques to model water stress effects on CO2 and energy exchange in temperate and boreal deciduous forests. Ecol. Model. 196 (3), 289–312.

Grant, R.F., et al., 2011. Controlled warming effects on wheat growth and yield: Field measurements and modeling. Agron. J. 103 (6), 1742–1754.

Grant, R.F., Dyck, M., Puurveen, D., 2020a. Nitrogen and phosphorus control carbon sequestration in agricultural ecosystems: modelling carbon, nitrogen, and phosphorus balances at the Breton Plots with ecosys under historical and future climates. Can. J. Soil Sci. 100 (4), 408–429.

Grant, R.F., Lin, S., Hernandez-Ramirez, G., 2020b. Modelling nitrification inhibitor effects on N_2O emissions after fall- and spring-applied slurry by reducing nitrifier NH_4^+ oxidation rate. Biogeosciences 17 (7), 2021–2039.

Gray, S.B., et al., 2016. Intensifying drought eliminates the expected benefits of elevated carbon dioxide for soybean. Nat. Plants 2 (9), 16132.

Gruber, N., Jockisch, A., 2020. Are GRU cells more specific and LSTM cells more sensitive in motive classification of text? Front. Artif. Intel. 3, 40.

Guan, K., Jin, Z., Peng, B., Tang, J., DeLucia, E.H., West, P.C., Jiang, C., Wang, S., Kim, T., Zhou, W., Griffis, T., 2023. A scalable framework for quantifying field-level agricultural carbon outcomes. Earth Sci. Rev. 243, 104462.

Guen, V.L., Thome, N., 2020. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In: Proceedings of the IEEE CVPR, pp. 11474–11484.

Guo, C., et al., 2018. Integrating remote sensing information with crop model to monitor wheat growth and yield based on simulation zone partitioning. Precis. Agric. 19 (1), 55–78.

Han, J., et al., 2022. Rice yield estimation using a CNN-based image-driven data assimilation framework. Field Crop Res. 288, 108693.

Heathman, G.C., et al., 2003. Assimilation of surface soil moisture to estimate profile soil water content. J. Hydrol. 279 (1), 1–17.

Holzworth, D.P., et al., 2014. APSIM – Evolution towards a new generation of agricultural systems simulation. Environ. Model Softw. 62, 327–350.

Hu, S., et al., 2017. Simultaneous state-parameter estimation supports the evaluation of data assimilation performance and measurement design for soil-water-atmosphere-plant system. J. Hydrol. 555, 812–831.

Hu, S., et al., 2019. Improvement of sugarcane crop simulation by SWAP-WOFOST model via data assimilation. Field Crop Res. 232, 49–61.

Huang, J., et al., 2015. Jointly assimilating MODIS LAI and ET products into the SWAP model for winter wheat yield estimation. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 8 (8), 4060–4071.

Huang, J., et al., 2019. Assimilation of remote sensing into crop growth models: current status and perspectives. Agric. For. Meteorol. 276-277, 107609.

Ines, A.V.M., et al., 2013. Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. Remote Sens. Environ. 138, 149–164.

Jeffries, G.R., et al., 2020. Mapping sub-field maize yields in Nebraska, USA by combining remote sensing imagery, crop simulation models, and machine learning. Precis. Agric. 21 (3), 678–694.

Jia, X., et al., 2021. Physics-guided machine learning for scientific discovery: an application in simulating lake temperature profiles. ACM/IMS Trans. Data Sci. 2 (3), 1–26.

Jiang, C., et al., 2020. BESS-STAIR: a framework to estimate daily, 30 m, and all-weather crop evapotranspiration using multi-source satellite data for the US Corn Belt. Hydrol. Earth Syst. Sci. 24 (3), 1251–1273.

Jiang, C., et al., 2021. A daily, 250 m and real-time gross primary productivity product (2000–present) covering the contiguous United States. Earth Syst. Sci. Data 13 (2), 281–298.

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 J. Selected Topics Appl. Earth Observ. Remote Sensing 7 (11), 4422–4431.

Jin, X., et al., 2017. Winter wheat yield estimation based on multi-source medium resolution optical and radar imaging data and the AquaCrop model using the particle swarm optimization algorithm. ISPRS J. Photogramm. Remote Sens. 126, 24–37.

Jin, X., et al., 2018. A review of data assimilation of remote sensing and crop models. Eur. J. Agron. 92, 141–152.

Jin, Z., et al., 2019. Smallholder maize area and yield mapping at national scales with Google Earth Engine. Remote Sens. Environ. 228, 115–128.

Jones, J.W., et al., 2003. The DSSAT cropping system model. Eur. J. Agron. 18 (3), 235–265.

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

Karpatne, A., Kannan, R., Kumar, V., 2022. Knowledge Guided Machine Learning: Accelerating Discovery using Scientific Knowledge and Data. CRC Press.

Kellogg, Kent, et al., 2020. NASA-ISRO synthetic aperture radar (NISAR) mission. In: 2020 IEEE Aerospace Conference. IEEE, p. 2020.

Kennedy, J., Eberhart, R., 1995. Particle swarm optimization. In: Proceedings of ICNN'95 - International Conference on Neural Networks, vol.4, pp. 1942–1948.

Kerr, Y.H., et al., 2010. The SMOS mission: new tool for monitoring key elements ofthe global water cycle. Proc. IEEE 98 (5), 666–687.

Kimm, H., et al., 2020. Deriving high-spatiotemporal-resolution leaf area index for agroecosystems in the US Corn Belt using Planet Labs CubeSat and STAIR fusion data. Remote Sens. Environ. 239, 111615.

Kingma, D.P., Welling, M., 2013. 'Auto-Encoding Variational Bayes', arXiv [stat.ML].

Kornelsen, K.C., Coulibaly, P., 2014. Root-zone soil moisture estimation using data-driven methods. Water Resour. Res. 50 (4), 2946–2962.

Lal, R., 2011. Sequestering carbon in soils of agro-ecosystems. Food Policy 36, S33–S39.

Li, Z., et al., 2020. 'Fourier Neural Operator for Parametric Partial Differential Equations', arXiv [cs.LG].

Li, X., et al., 2022a. A new SMAP soil moisture and vegetation optical depth product (SMAP-IB): Algorithm, assessment and inter-comparison. Remote Sens. Environ. 271, 112921.

Li, Z., et al., 2022b. Assessing the impacts of pre-growing-season weather conditions on soil nitrogen dynamics and corn productivity in the U.S. Midwest. Field Crop Res. 284, 108563.

Liang, M., et al., 2023. Quantifying aboveground biomass dynamics from charcoal degradation in Mozambique using GEDI Lidar and Landsat. Remote Sens. Environ. 284, 113367.

Liu, L., et al., 2022. KGML-ag: a modeling framework of knowledge-guided machine learning to simulate agroecosystems: a case study of estimating N_2O emission using data from mesocosm experiments. Geosci. Model Dev. 15 (7), 2839–2858.

Lobell, D.B., et al., 2014. Greater sensitivity to drought accompanies maize yield increase in the U.S. Midwest. Science 344 (6183), 516–519.

Lobell, D.B., et al., 2015. A scalable satellite-based crop yield mapper. Remote Sens. Environ. 164, 324–333.

Luo, Y., Guan, K., Peng, J., 2018. STAIR: A generic and fully-automated method to fuse multiple sources of optical satellite data to generate a high-resolution, daily and cloud-/gap-free surface reflectance product. Remote Sens. Environ. 214, 87–99.

Ma, H., Liang, S., 2022. Development of the GLASS 250-m leaf area index product (version 6) from MODIS data using the bidirectional LSTM deep learning model. Remote Sens. Environ. 273, 112985.

Markovich, K.H., White, J.T., Knowling, M.J., 2022. Sequential and batch data assimilation approaches to cope with groundwater model error: an empirical evaluation. Environ. Model Softw. 156, 105498.

Melton, F.S., et al., 2022. OpenET: filling a critical data gap in water management for the western United States. J. Am. Water Resour. Assoc. 58 (6), 971–994.

Meyer, R., et al., 2022. Exploring the combined use of SMAP and Sentinel-1 data for downscaling soil moisture beyond the 1 km scale. Hydrol. Earth Syst. Sci. 26 (13), 3337–3357.

Mezbahuddin, S., et al., 2020. Assessing effects of agronomic nitrogen management on crop nitrogen use and nitrogen losses in the western Canadian prairies. Front. Sustain. Food Syst. 4.

Monsi, M., 1953. The light factor in plant communities and its significance for dry matter production. Japan. J. Botany 14, 22.

Moradkhani, H., Hsu, K.L., Gupta, H., Sorooshian, S., 2005a. Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. Water resources research 41 (5).

Moradkhani, H., Sorooshian, S., et al., 2005b. Dual state–parameter estimation of hydrological models using ensemble Kalman filter. Adv. Water Resour. 28 (2), 135–147.

Mu, Q., Zhao, M., Running, S.W., 2011. Improvements to a MODIS global terrestrial evapotranspiration algorithm. Remote Sens. Environ. 115 (8), 1781–1800.

Perez-Priego, O., et al., 2017. Evaluation of eddy covariance latent heat fluxes with independent lysimeter and sapflow estimates in a Mediterranean savannah ecosystem. Agric. For. Meteorol. 236, 87–99.

Qin, Z., et al., 2021. Assessing the impacts of cover crops on maize and soybean yield in the U.S. Midwestern agroecosystems. Field Crop Res. 273, 108264.

Shen, C., et al., 2023. 'Differentiable modeling to unify machine learning and physical models and advance Geosciences', *arXiv [cs.LG]*.

Shukla, P.R., et al., 2019. IPCC, 2019: Climate Change and Land: an IPCC Special Report on Climate Change, Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse Gas Fluxes in Terrestrial Ecosystems.

Shwartz-Ziv, R., Tishby, N., 2017. 'Opening the Black Box of Deep Neural Networks via Information', arXiv [cs.LG].

Suyker, A.E., Verma, S.B., 2012. Gross primary production and ecosystem respiration of irrigated and rainfed maize–soybean cropping systems over 8 years. Agric. For. Meteorol. 165, 12–24.

Tao, F., et al., 2018. Contribution of crop model structure, parameters and climate projections to uncertainty in climate change impact assessments. Glob. Chang. Biol. 24 (3), 1291–1307.

Tsai, W.-P., et al., 2021. From calibration to parameter learning: harnessing the scaling effects of big data in geoscientific modeling. Nat. Commun. 12 (1), 5988.

USDA, 2023. Available at. USDA-National Agricultural Statistics Service-Research and Science - Cropland Data Layer Releases https://www.nass.usda.gov/Research_and_Science/Cropland/Release/index.php (accessed Otc. 2023).

Vergopolan, N., et al., 2021. SMAP-HydroBlocks, a 30-m satellite-based soil moisture dataset for the conterminous US. Sci. Data 8 (1), 264.

Wahle, K., Staneva, J., Guenther, H., 2015. Data assimilation of ocean wind waves using Neural Networks. A case study for the German Bight. Ocean Model. 96, 117–125.

Wang, S., et al., 2020. Mapping twenty years of corn and soybean across the US Midwest using the Landsat archive. Sci. Data 7 (1), 307.

Wang, Z., et al., 2023. Temperature effect on erosion-induced disturbances to soil organic carbon cycling. Nat. Clim. Chang. 13 (2), 174–181.

Weiss, M., Jacob, F., Duveiller, G., 2020. Remote sensing for agricultural applications: a meta-review. Remote Sens. Environ. 236, 111402.

Welegedara, N.P.Y., et al., 2020. Modelling nitrogen mineralization and plant nitrogen uptake as affected by reclamation cover depth in reclaimed upland forestlands of Northern Alberta. Biogeochemistry 149 (3), 293–315.

Welp, L.R., et al., 2008. deltaO of water vapour, evapotranspiration and the sites of leaf water evaporation in a soybean canopy. Plant Cell Environ. 31 (9), 1214–1228.

Willard, J., et al., 2022. Integrating scientific knowledge with machine learning for engineering and environmental systems. ACM Comput. Surv. 55 (4), 1–37.

Wood, E.F., et al., 2011. Hyperresolution global land surface modeling: meeting a grand challenge for monitoring Earth's terrestrial water. Water Resour. Res. 47 (5).

Xu, Z., et al., 2021. Evapotranspiration partitioning for multiple ecosystems within a dryland watershed: Seasonal variations and controlling factors. J. Hydrol. 598, 126483.

Xu, S., et al., 2022. 'Mini-Batch Learning Strategies for modeling long term temporal dependencies: A study in environmental applications', *arXiv [cs.LG]*.

Yang, Y., et al., 2020. Deep reinforcement learning-based irrigation scheduling. Trans. ASABE 63 (3), 549–556.

Yang, Y., et al., 2022. Distinct driving mechanisms of non-growing season N2O emissions call for spatial-specific mitigation strategies in the US Midwest. Agric. For. Meteorol. 324, 109108.

Yang, Q., et al., 2023. Regulating the time of the crop model clock: a data assimilation framework for regions with high phenological heterogeneity. Field Crop Res. 293, 108847.

Zahura, F.T., et al., 2020. Training machine learning surrogate models from a high-fidelity physics-based model: Application for real-time street-scale flood prediction in an urban coastal community. Water Resour. Res. 56 (10).

Zeng, L., et al., 2019. Multilayer soil moisture mapping at a regional scale from multisource data via a machine learning method. Remote Sens. 11 (3), 284.

Zhang, Q., et al., 2019. A dynamic data-driven method for dealing with model structural error in soil moisture data assimilation. Adv. Water Resour. 132, 103407.

Zhang, C., et al., 2021. A combined optimization-assimilation framework to enhance the predictive skill of community land model. Water Resour. Res. 57 (12), e2021WR029879.

Zhou, W., et al., 2021a. Quantifying carbon budget, crop yields and their responses to environmental variability using the ecosys model for U.S. Midwestern agroecosystems. Agric. For. Meteorol. 307, 108521.

Zhou, W., et al., 2021b. A generic risk assessment framework to evaluate historical and future climate-induced risk for rainfed corn and soybean yield in the U.S. Midwest. Weather Clim. Extrem. 33, 100369.