2023 INCITE Proposal Submission
Proposal

Title: Quantifying carbon outcomes over the U.S. Cropland using AI-based data-model fusion

Principal Investigator: Kaiyu Guan

Organization: University of Illinois

Date/Time Generated: 6/17/2022 7:53:51 PM

Section 1: PI and Co-PI Information

Question #1

Principal Investigator: *The PI is responsible for the project and managing any resources awarded to the project. If your project has multiple investigators, list the PI in this section and add any Co-PIs in the following section.*

Principal Investigator

First Name

Kaiyu

Last Name

Guan

Organization

University of Illinois Urbana-Champaign

Email

kaiyug@illinois.edu

Work Phone

609-647-1368

Address Line 1

W503, Turner Hall, 1102 S. Goodwin

Address Line 2

(No answer given.)

City

Urbana

State

Illinois

Zip Code

61801

Question #2

Co-PI (s)

First Name

Bin

Last Name

Peng

Organization

University of Illinois Urbana-Champaign

Email

binpeng@illinois.edu

Question #3

Institutional Contact: *For the PI's institution on the proposal, identify the agent who has the authority to review, negotiate, and sign the user agreement on behalf of that institution. The person who can commit an organization may be someone in the contracts or procurement department, legal, or if a*

university, the department head or Sponsored Research Office or Grants Department.

Institutional Contact

Institutional Contact Name

Laura Martin

Institutional Contact Phone

217-300-9954

Institutional Contact Email

laurarh@illinois.edu

Section 2: Project Information

Question #1

Select the category that best describes your project.

Research Category

Earth Science: Agricultural Sciences

Question #2

Please provide a project summary in two sentences that can be used to describe the impact of your project to the public (50 words maximum)

Project Summary

This project will use the advanced process-based models, satellite-based observations, and artificial intelligence to develop integrative solutions to quantify historical and future potential for soil organic carbon change and greenhouse gas emissions over every field of the US Midwestern cropland.

Section 3: Early Career Track

Question #1

Early Career

Starting in the INCITE 2022 year, INCITE is committing 10% of allocatable time to an Early Career Track in INCITE. The goal of the early career track is to encourage the next generation of high-performance computing researchers. Researchers within 10 years from earning their PhD (after December 31st 2012) may choose to apply. Projects will go through the regular INCITE Computational Readiness and Peer Review process, but the INCITE Management Committee will consider meritorious projects in the Early Career Track separately.

***Who Can Apply**: Researchers less than 10 years out from their PhD that need LCF-level capabilities to advance their overall research plan and who have not been a previous INCITE PI.*

How to Apply:

In the regular application process, there will be a check-box to self-identify as early career.

- *The required CV should make eligibility clear.*
- *If awarded, how will this allocation fit into your overall research plan for the next 5 years?*

Projects will go through the regular INCITE review process. The INCITE Program is targeting at least 10% of allocatable time. When selecting the INCITE Career Track, PIs are not restricted to just competing in that track.

- *What is the Early Career Track?*
  - *The INCITE Program created the Early Career Track to encourage researchers establishing their research careers. INCITE will award at least 10% of allocatable time to meritorious projects.*
- *Will this increase my chances of receiving an award?*
  - *Potentially, this could increase chances of an award. Projects must still be deemed scientifically meritorious through the review process INCITE uses each year.*
- *What do I need to do to be considered on the Early Career Track?*
  - *In the application process, select 'Yes' at 'If you are within 10 years of your PhD, would you like to be considered in the Early Career Track?' You will need to write a paragraph about how the INCITE proposal fits into your 5-year research and career goals.*
- *What review criteria will be used for the Early Career Track?*
  - *The same criteria for computational readiness and scientific merit will be applied to projects in the Early Career Track as will be applied to projects in the traditional track. The different will be manifest in awards decisions by the INCITE management committee.*

Early Career Track

If you are within 10 years of your PhD, would you like to be considered in the Early Career Track? Choosing this does not reduce your chances of receiving an award.

Yes

If ‘yes’, what year was your PhD? If 'no' enter N/A

2013

If ‘yes’, how will this allocation fit into your overall research plan for the next 5 years? If 'no' enter N/A.

Under several major federal grants like the DOE ARPA-E SMARTFARM program and the DOE CABBI center program, I have been using the computational resources from my campus at University of Illinois Urbana Champaign to develop prototype of a "system of systems" solution to quantify soil organic carbon change and greenhouse gas emissions over croplands. Now this research reaches to a critical stage that we are ready to scale up the technology to the whole U.S. if sufficient computation resources are made available. This research has strong implications to the scientific community (e.g. better quantification of carbon cycle) and to the society (e.g. better management and planning to reduce GHG emission to mitigate climate change). I will use this allocation together with other campus-level computational resources to generate the most rigorous estimations of historical soil organic carbon change and greenhouse gas emissions and carbon credit potentials under different hypothetical conservation management scenarios over the U.S. cropland. Thus the allocation from INCITE will allow me to scale up the technology and scientific discovery to an unprecedented scale to amplify both scientific and society impacts of the research. I greatly appreciate the possible support and consideration!

Section 4: INCITE Allocation Request & Other Project Funding/Computing Resources

Question #1

OLCF Summit (IBM / AC922) Resource Request - 2023

Question #2

OLCF Frontier (Cray Shasta) Resource Request – 2023

Question #3

OLCF Frontier (Cray Shasta) Resource Request – 2024

Question #4

OLCF Frontier (Cray Shasta) Resource Request – 2025

Question #5

ALCF Theta (Cray XC40) Resource Request - 2023

Question #6

ALCF Polaris Resource Request - 2023

Question #7

ALCF Polaris Resource Request - 2024

Question #8

ALCF Polaris Resource Request - 2025

Question #9

ALCF Aurora (Intel X^{e}) Resource Request – 2023

Question #10

ALCF Aurora (Intel X^{e}) Resource Request – 2024

Question #11

ALCF Aurora (Intel X^{e}) Resource Request – 2025

Question #12

List any funding this project receives from other funding agencies.

Funding Sources

Funding Source

DOE ARPA-E SMARTFARM Project, NSF Signals-in-Soil Project

Grant Number

DOE ARPA-E SMARTFARM Project, NSF Signals-in-Soil Project

Question #13

List any other high-performance computing allocations being received in support of this project.

Other High Performance Computing Resource Allocations

Resource

some limited resources are available to my research group through University of Illinois Urbana Champaign, through two sources: the Delta Supercomputer and the UIUC Campus Cluster

Allocation Agency

University of Illinois Urbana Champaign

Allocation

about 80K node hours per year

Allocation Year

2022

Section 5: Project Narrative and Supplemental Materials

Question #1

Using the templates provided here, please follow the INCITE Proposal Preparation Instructions to prepare your proposal. Elements needed include (1) Project Executive Summary, (2) Project Narrative, (3) Personnel Justification and Management Plan, (4) Milestone Table, (5) Publications Resulting from prior INCITE Awards (if appropriate), and (6) Biographical Sketches for the PI and all co-PI's. Concatenate all materials into a single PDF file. Prior to submission, it is strongly recommended that proposers review their proposals to ensure they comply with the proposal preparation instructions.

Concatenate all materials below into a single PDF file.

1. **Project Executive Summary (One Page Max)**
2. **Project Narrative (15 Pages Max)**
3. **Personnel Justification and Management Plan (1 Page Max)**
4. **Milestone Table**
5. **Publications resulting from prior INCITE Awards (if appropriate)**
6. **Biographical Sketches for the PI and all co-PI's.**

all.pdf
The attachment is on the following page.

PROJECT EXECUTIVE SUMMARY

Title: Quantifying carbon outcomes over cropland using AI-based model-data fusion

PI and Co-PI(s): Kaiyu Guan, Bin Peng

Applying Institution/Organization: University of Illinois Urbana-Champaign

Resource Name(s) and Number of Node Hours Requested: 1,421,955 CPU node hours + 234,763 GPU hours

Amount of Storage Requested: 960 TB

Executive Summary:

The agricultural sector of the U.S. has a significant potential to contribute to climate change mitigation by reducing greenhouse gas emissions and enhancing soil carbon sequestration (i.e. improved carbon outcome) over farmland. Recent surging of carbon insetting and offsetting needs from different entities to mitigate climate change offer an opportunity to provide direct financial incentives for farmers to adopt regenerative or climate-smart farming management practices for improved carbon outcome. However, we are currently lacking a field-level, accurate, scalable, and cost-effective quantification of carbon outcome under various management practices to support the national policy-design and serve the farming community for the potential agricultural carbon market.

This proposed INCITE project will fully leverage advanced artificial intelligence (AI), computing techniques and allocated computational resources to scale-up a first-of-its-kind, field-level, scalable and AI-based model-data fusion (MDF) solution to enable a large-scale agricultural carbon outcome quantification for every individual field over the US Midwest regions. The MDF system will efficiently and effectively integrate highly accurate and ubiquitous satellite observations of crop status, management practices, and environmental conditions with an advanced process-based ecosystem model *ecosys* to generate field-level carbon budgets under both historical and hypothetical management scenarios. This project will lead to breakthroughs by answering the following scientific questions: (1) What are the spatiotemporal variabilties of carbon budget over individual cropland parcels over the U.S. Midwest? (2) How do the environmental and management factors control those spatiotemporal variabilities of carbon budget? and (3) What is the climate change mitigation potential of the U.S. Midwestern cropland under different regenerative management scenarios? Ultimately, the outcome of this project will directly facilitate the design of policies related to climate-smart commodities and pathways towards science-based climate targets for different entities.

The proposal team includes two early-career scientists who have extensive HPC experience and have also developed the prototypes for the AI-based MDF solution through the DOE ARPA-E SMARTFARM projects and NSF Soil-in-Signal project, but currently lacks computational resources to scale over a broader region like the whole U.S. Midwest. This research has strong implications to the scientific community (e.g. better quantification of carbon cycle) and to society (e.g. better management and planning to reduce GHG emission to mitigate climate change). Thus the allocation from INCITE will enable to scale up the proposed technology and scientific discovery to an unprecedented scale to amplify both scientific and society impacts of the research. The team has demonstrated strong scaling performance in their prior research based on the Director's Discretionary allocation.

PROJECT NARRATIVE

1. Significance of Research

Agriculture sector contributes to 10% of the total greenhouse gas (GHG) emissions in the United States (U.S.), and more than 60% of agricultural GHG emissions are from U.S. farmland [1-2]. Effective strategies are urgently needed to reduce GHG emissions or enhance carbon sequestration over farmland [3-5], to help mitigate climate change. Strategic change of farming management practices has a great potential to improve carbon sequestration and mitigate GHG emissions from farmland, such as conservation tillage (reduce tillage/no-till), winter cover cropping, and smarter nitrogen fertilizer usage [6], and these practices have been also called "regenerative practices" or "climate-smart practices". However, adoption rates of these management practices over the U.S. Midwest, one of the world's major food baskets that produces one third of global corn and soybean production, are still very low despites recent substantial increases in government subsidy such as USDA's Pandemic Cover Crop Program and Partnerships for Climate-Smart Commodities. The surge of the public's perceived urgency in combating climate change and achieving sustainable development has recently spurred climate-pledges by individual entities to cut their carbon footprints and stimulate the growth of agricultural carbon markets. For example, a majority of the carbon footprints of consumer goods companies (CPGs) are related to agricultural production through their supply chains, and thus sourcing/procuring low-carbon-intensity or climate-smart commodities will help these companies reduce their overall carbon footprints or so-called Scope-3 emission in the supply chains. On the other hand, entities with climate targets may offset their carbon emissions through purchasing carbon credit from the agricultural carbon market. These emerging carbon insetting and offsetting needs provide direct financial incentives for farmers to adopt regenerative practices. **For either carbon insetting or offsetting in the context of agriculture, the foundation is built upon accurate quantification of carbon emission and carbon sequestration under various management practices. However, existing scientific literature is not yet conclusive as to where, when, if and by how much those regenerative practices might lead to genuine GHG reduction or carbon removal**.

From a carbon mass balance perspective, quantification of carbon outcome (carbon emission and sequestration) is about quantification of carbon budget over cropland, i.e. accounting carbon flowing in and out of a cropland under different management interventions. Though quantifying ecosystem-level carbon budgets is a classic problem in earth system and ecology studies, the requirements for quantifying carbon outcomes for farmland in the context of carbon insetting and offsetting programs have a much higher standard. To ensure the climate change mitigation can be assured, **effective quantification technology of carbon outcomes for the carbon insetting and offsetting programs should be at the field level, accurate, scalable, and cost-effective**. "Field-level accuracy" is needed if carbon outcome is associated with rewarding individual farmers' practice; it is also required for traceability of any aggregated carbon outcome in carbon footprint quantification. "Scalable" here means that the quantification solution must have an independently verified accuracy across all possible fields; in other words, showing that a solution works well at a few demonstration sites, as many existing Measurement-Reporting-Verification (MRV) efforts do, is not enough. Instead, true "scalability" means one method must demonstrate an acceptable accuracy of the solution at randomly selected 'real-world' sites. Aggregated-level accuracy, which is almost impossible to validate, must come from field-level accuracy. Finally, for any technology, there is a tradeoff between cost and accuracy, and the desired solution should be sufficiently cost-effective to achieve the needed accuracy [7].

The current research of carbon budget quantification in the literature cannot meet the above requirements for the carbon insetting and offsetting programs. There are three major categories of quantification methods in the literature about carbon budget quantification over cropland, including: (1) direct field measurements (such as soil sampling for SOC change [8-10], and eddy-covariance sensors to measure GHG emissions [11,12]); (2) emission factor estimation, in which a fixed linear factor is used to approximate the "outcome" based on different management practices [13]; and (3) process-based modeling [11,12,14,15]. **Direct field measurements** of changes in soil organic carbon (SOC) storage and carbon

fluxes like photosynthesis and respiration have significantly advanced our understanding of carbon cycling in the agroecosystems [16-18]. However, it is practically infeasible to collect field observations across every field of croplands due to the high financial and labor costs. Satellite observations can provide estimations of selected carbon fluxes, such as harvested yield [19-22] and gross primary production [23], but other cropland carbon budget components such as heterotopic respiration from soil are inadequately quantifiable from satellites. Moreover, it is challenging to use soil sampling to calculate soil carbon credit, not only because the high cost and relatively high uncertainty in soil sampling [24], and also because the soil carbon credit is the difference in SOC stock changes between counterfactual scenarios (e.g., with and without cover crops for the same field), which is hard to be measured directly unless paired experiments are properly implemented in the same field. For most all the commercial fields, once farmers adopt one particular scenario of practice, they rarely would simultaneously implement a counterfactual scenario in part of their farmland, leading to the inability to quantify the relative difference of carbon outcomes. **Emission factor methods**, as the most widely used approaches in past IPCC reports [13] and also the easiest method to use (usually a static linear factor multiplier for certain management practices, constant at large regional scale), suffer from their weakness of not able to capture spatial and temporal heterogeneity of environment, crop, and management conditions, and thus cannot comprehensively track the dynamics embedded in the interactions among these factors. Assuming the same (or a linear scaling of) emission or sequestration outcome based on a particular management practice across all the different fields is not only inaccurate, but also unfair for individual farmers in the agricultural carbon market. **Process-based models** have been widely used to calculate carbon budget over cropland [18,25,26] and also have been used to calculate agricultural carbon outcomes (e.g. COMET-FARM, DNDC). However, large uncertainties exist in model-simulated cropland carbon budgets mainly due to uncertainties in model structure, parameters, weather and soil inputs [27-30].

Systematic **model-data fusion (MDF)** has the potential to solve the above problem, by taking advantages from both readily available observations and process-based modeling [31]. MDF here refers to a set of techniques that constrains the uncertainty of states and parameters of process-based models or fine-tunes data-driven models (e.g. statistical model or neural networks) with local information to generate improved estimation of "carbon outcomes". When implemented properly, MDF can effectively reduce uncertainties in observations, model inputs, model parameters and model processes [32]. MDF also has the ability to evolve by incorporating new sensors/sensing data or new model developments to this framework, which is well aligned with the **DOE's iterative model-experiment (ModEx) integration approach**. Traditionally, MDF is conducted through model calibration at limited sites where in-situ observations are available and the calibrated models are then used for simulations over other places. Considering there is a large spatial heterogeneity caused by environmental conditions (soil and climate), management practices, crop conditions for the agricultural system, ensuring there is a local constraint for every field is critical to achieve high accuracy and realistic simulation of carbon budget at the field level. Besides accurate and high-resolution input data for the modeling, there are many location-specific parameters in the model that need to be optimized through MDF, which include: (1) plant physiological parameters that are varying across time and space and also genetically, but are generally not dynamically modeled in the current models, such as plant photosynthetic capacity, and grain-filling rate; (2) local soil properties, including soil hydrological, tile drainage efficiency, and some soil biogeochemical properties. Though in some cases we have an available soil database, it is well known that these soil data can have large uncertainties at a specific local area. Using observations to further constrain these soil related parameters can effectively reduce the uncertainties. Without such a local constraint, model simulations can be significantly deviating from reality, which should be avoided in the MRV for quantifying agricultural carbon outcomes.

However, several challenges exist when developing an operational and systematic MDF system for cropland carbon accounting at field level. **First**, there is a lack of an accurate and scalable method to identify field-level agricultural management practices at large areas; without such information, carbon budget could not be quantified accurately; **Second**, there is a lack of high-resolution field-level constraining observations for conducting MDF everywhere; **Third**, conventional model-data fusion methods (e.g. data assimilation,

Bayesian inference) have high computational cost to run even at a few sites, which prohibits scaling the applications to millions of individual fields over a broad region.

Recent advances in multi-scale remote sensing, artificial intelligence (AI), and computational techniques provide great promise to solve these challenges. PI Guan's lab at University of Illinois Urbana-Champaign has been a leader in agricultural remote sensing research and made significant contributions in the following aspects: (1) developed multi-source satellite data fusion algorithm STAIR to generate high resolution and daily cloud-/gap-free satellite data [33,34]; (2) developed advanced crop yield estimation algorithm by combing remote sensing, weather, and soil information [19,21,22,35-37]; (3) developed airborne-satellite integrative sensing framework to map cover crops and tillage practices and crop conditions [38,39]; and (4) developed satellite-based algorithms to estimate field-level biomass [40], crop photosynthesis [41] and water use [42]. With these technology developments, Guan's lab has solved the first two challenges and developed a comprehensive database of field-level management practices, crop biomass, photosynthesis, and water use across the whole U.S. Midwest, which can be directly used for developing an operational and systematic MDF system for cropland carbon accounting in the Midwest regions. For the third challenge, recent advances in AI-based surrogate modeling, physics-guided deep learning, and advanced GPU computing enable more efficient ways to integrate observations and models and thus achieve scalable and cost-effective model-data fusion down to every farmland at the continental scales [43-46]. PI Guan's lab has built AI-based MDF prototypes through several existing federal projects, including DOE ARPA-E SMARTFARM Phase I and II projects, and NSF Soil-in-Signal project, but currently lacks computational resources to scale over a broader region like the whole U.S. Midwest. We have previously tried Blue Waters and Delta supercomputers at National Center for Supercomputing Applications (NCSA) of University of Illinois Urbana-Champaign, Comet Supercomputer at the San Diego Supercomputer Center of University of California San Diego, and DOE ALCF's Theta through its Director's Discretionary allocation to PI Guan in 2022, which all supported the prototyping and early testing, but not the full scale-up of this project.

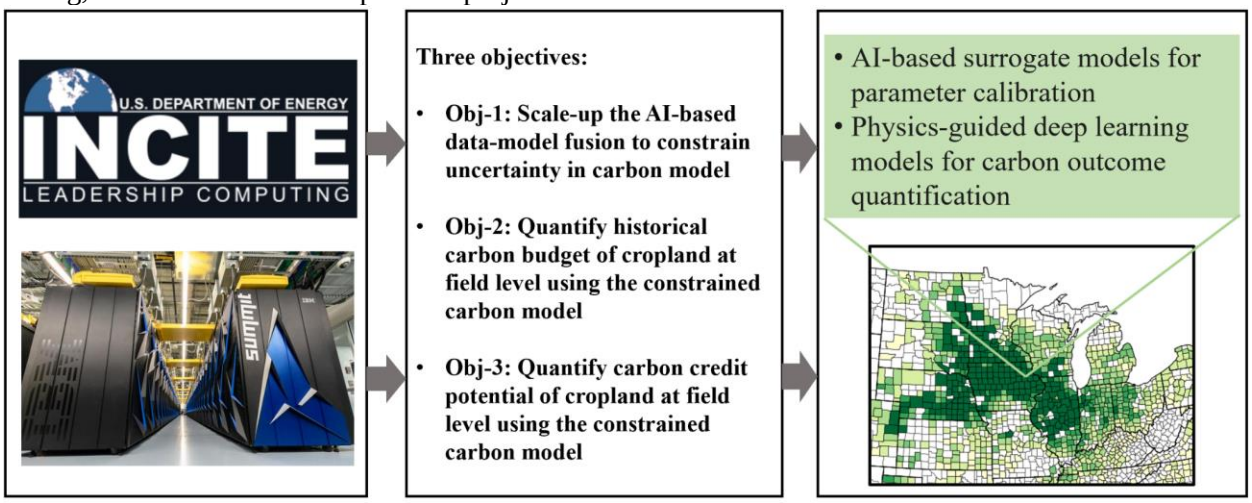


Fig. 1 Overview of the objectives of this proposal.

This project will fully leverage advanced AI techniques and INCITE computational resources to address the above challenge by developing a first-of-its-kind, field-level, scalable and AI-based MDF solution to enable a large-scale agricultural carbon outcome quantification for the US Midwest regions, which will ultimately contribute to climate change mitigation and improve long-term farmland sustainability of the U.S. This project aims to answer the following scientific questions (**Fig. 1**): (1) What are the spatiotemporal variabilties of carbon budget over individual cropland parcels over the U.S. Midwest? (2) How do the environmental and management factors control those spatiotemporal variabilities of carbon budget? and (3) What is the climate change mitigation potential of the U.S. Midwestern cropland under different regenerative management scenarios? To answer these questions, we will scale the AI-based

MDF across every individual cropland field over the U.S. Midwest, through which our highly accurate and ubiquitous satellite observations of crop, management practices, and environmental conditions are efficiently and effectively integrated with an advanced process-based ecosystem model *ecosys* [47] to generate field-level carbon budgets under both historical condition and hypothetical management scenarios.

2. Research Objectives and Milestones

In this project, we will have the following three objectives: **Obj-1:** Scale-up the AI-based model-data fusion to constrain uncertainty in carbon model; **Obj-2:** Quantify historical carbon budget of cropland at field level using the constrained carbon model; and **Obj-3:** Quantify carbon credit potential of cropland at field level using the constrained carbon model (**Fig. 1**). To achieve the first objective, we will build an **AI-based surrogate model** based on *ecosys* and use GPU computation to constrain the model parameters in *ecosys* with satellite-based observations at each individual farmland field. To achieve the second and third objectives, we will further build an **AI-based carbon outcome model** based on physics-guided deep learning (PGDL), trained with simulations from the observation-constrained *ecosys* model to quantify both crop productivity and carbon outcomes under historical and hypothetical management scenarios. Below we will provide specific details for each objective and how we plan to achieve them.

2.1 Obj-1: Scale-up the AI-based model-data fusion to constrain uncertainty in carbon model

This objective aims to build a carbon accounting system using AI-based model-data fusion. The model-data fusion technique integrates the strengths from both observations and process-based models by using observation data to systematically constrain parameters in the models. Model-data fusion within the Bayesian framework mainly relies on Monte-Carlo-based or evolutionary optimizations, which require thousands of sequential model simulations and thus is **computationally intensive**. PI-Guan's lab has built a prototype of the traditional model-data fusion system for several fields with the support from the DOE ARPA-E SMARTFARM project. Here we propose to use the AI technique and INCITE computational resources to solve the computation challenge and **scale-up the model-data fusion** for every field in the broad U.S. Midwest.

SYMFONI (a "System of Systems" Solution for Commercial Field-Level Quantification of SOC and N2O) is a novel Model-Data Fusion framework to quantify field-level carbon outcomes (including GHG, soil carbon change), applicable for broad agroecosystems. SYMFONI's development has been funded through the UIUC-led DOE ARPA-E SMARTFARM Phase 2 Project. SYMFONI integrates an advanced agroecosystem model, *ecosys* [47-49], with multi-stream observations (including remote sensing and ground-based sampling) at the individual field level, through the Model-Data Fusion (MDF). This MDF solution fully considers uncertainty propagation from both model simulation and observations into final estimations of carbon budget, including GHG emissions and SOC changes. The observations are used to either directly update inputs or constrain model parameters.

How SYMFONI works to quantify field-level carbon outcome: SYMFONI is a model-data fusion (MDF) framework to quantify the carbon footprints of climate-smart farming practices. There are three major components in SYMFONI: model engine, observation suite, and model-data fusion module. SYMFONI uses an advanced agroecosystem model, *ecosys*, as its **model engine**. *Ecosys* is an advanced mechanistic model constructed from basic scientific principles using parameters that may be determined independently of the model itself, therefore is widely applicable to different soils, climates and managements [47]. Importantly, *ecosys* employs complete physical and chemistry theories in simulating plants and soil-related processes. The model explicitly includes microbes' competitive and symbiotic nutrient interactions with plants, enabling a nutrient-based analysis of how various management practices could affect crop productivity. All major farming practices can be simulated by *ecosys* [50]. Previous work using *ecosys* has fully demonstrated its capabilities in simulating soil nitrogen cycle [50], N_2O and CH_4 emission [51,52], long-term soil organic matter trend [53], and impacts of different farming practices (i.e., tillage, cover crop fertilizer, and irrigation) [53]. The change of SOC at field level is holistically related to

the farmland carbon budget in *ecosys* (see equation in **Fig. 2**). The **observation suite** in SYMFONI provides capability to track the daily changes at each individual farmland parcel by using advanced sensing (ground, airborne hyperspectral imaging, and satellite remote sensing), including crop conditions (crop type, crop variety traits, phenology, maturity groups, photosynthetic capacity, carbon uptake, crop water use, and yield), management practices (planting/harvesting date, tillage practices, intercropping, crop rotation, cover cropping, fertilizer/pesticide applications), and environment conditions (weather information, soil temperature and moisture). The major terms in the carbon budget shown in **Fig. 2**, especially NEE, GPP, Reco, and Harvest, can be fully verified (and thus audited) at the annual scale using existing sensing technologies, in a cost-effective manner. In particular, NEE can be directly measured using the eddy-covariance (EC) technique, which has been a mature and widely-used tool across the world to measure terrestrial carbon budget [12]. Besides NEE, data from EC towers also provide robust measurements of GPP and Reco. EC towers have been established across global ecosystems to collect NEE, GPP, and Reco data, including agroecosystems for two decades. Most importantly, key carbon fluxes, such as GPP, can also be estimated from satellite data at field level over a broad region [23]. All these observations will be digested and integrated with the model engine in the **model-data fusion module** of SYMFONI. Some observations are directly used as model inputs (like crop type, management practices, and weather information), while others are used as model constraints to reduce the model uncertainties. The model-data fusion module uses advanced AI techniques to train surrogate models for original *ecosys* and GPU-based optimization to speed up the MDF such that it can be scaled up to millions of fields in an efficient way.

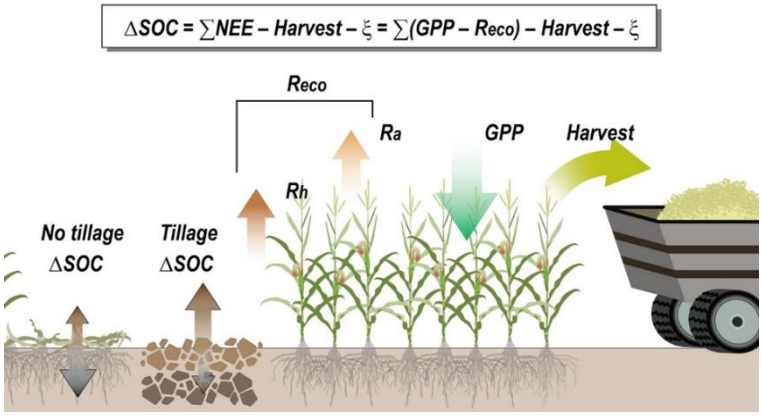


Fig. 2 Carbon budget of cropland. SOC is the change of SOC, NEE is net ecosystem exchange and can be derived as NEE=GPP-Reco, with GPP representing gross primary productivity (i.e. photosynthesis), and Reco representing ecosystem respiration; in particular, Reco = Ra + Rh, where Ra and Rh are autotrophic respiration (i.e. respiration from crop itself) and heterotrophic respiration (i.e. respiration from soil), respectively. Harvest represents carbon removed from the field via harvest (e.g. crop yield or other crop biomass). represents carbon loss from leaching, which is usually very small (<0.5%) and thus can be neglected in most cases.

Performance of SYMFONI solutions: In the prior work, we have demonstrated field-level performance of SYMFONI over many sites with flux tower observations as well as in sites with long-term soil carbon measurements (**Fig. 3**). Specifically, we used advanced satellite-based photosynthesis [23] and crop yield estimations [20] to constrain crop parameters in the *ecosys* model. Through systematic sensitivity analysis, we have identified key parameters to be constrained. For example, for corn, the constrained parameters in *ecosys* include fraction of leaf protein in bundle sheath chlorophyll, plant maturity group, maximum number of fruiting sites per reproductive node, and maximum rate of kernel filling. The existing SYMFONI results, when benchmarked with gold-standard eddy-covariance carbon flux data and the long-term SOC data (**Fig. 3**), show that the constrained model can significantly reduce uncertainty for the calculated carbon budget. For example, the SYMFONI-calculated photosynthesis, i.e. gross primary productivity (GPP), has only 15% relative error with benchmarked ground truth, compared with 57% relative error in the results of the default *ecosys* model without observational constraint. It is worth noting

that such a dramatic reduction in uncertainty has huge implications in the final calculated SOC change, as SOC change is only a small fraction of the total carbon budget [18], confirming that MDF must be used to achieve reliable carbon outcome quantification. In addition, SYMFONI also shows its ability to simulate long-term SOC change under different management practices (**Fig. 3b**).

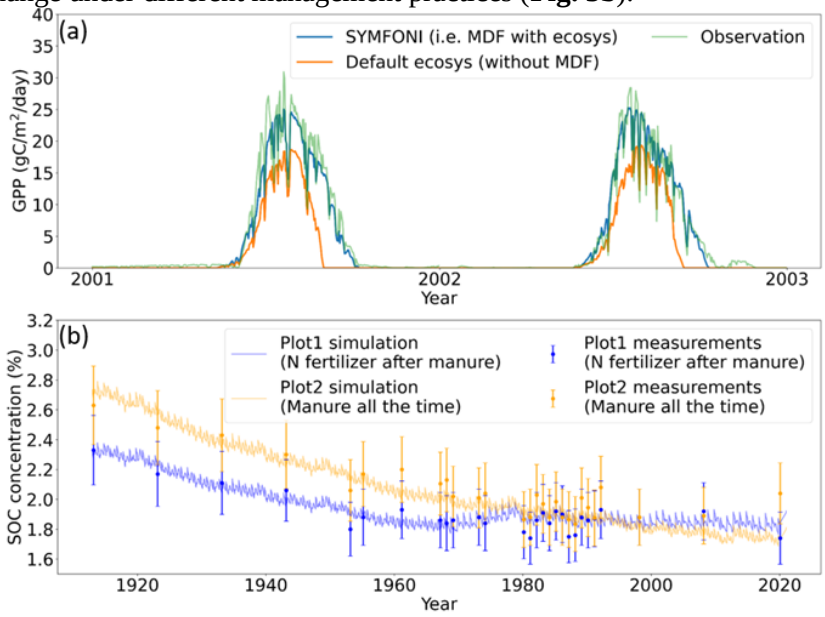


Fig. 3 SYMFONI demonstration. (a) Performance of GPP at the ARPA-E SMARTFARM site estimated by default (green line) and constrained (orange line, i.e. MDF in SYMFONI) *ecosys* models, benchmarked with flux tower observations (blue line). (b) SYMFONI estimated soil organic carbon (SOC) change in top layers at Morrow plot in Illinois in the continuous corn rotation systems with different fertilizer application practices over the last century (Plot 1 fertilizer application (N fertilizer after manure): 1912 - 1967 Manure of 4.5 Mg/ha; 1967 - 1998 N fertilizer of 300 lb N/ac ; 1999 - 2020 N fertilizer of 200 lb N/ac, Plot 2 fertilizer application (Manure all the time): 1912 - 2020 Manure of 4.5 Mg/ha).

Accelerate SYMFONI with AI surrogate model and GPU-based optimization: To scale up MDF in SYMFONI, we will use advanced AI and GPU to improve computation efficiency. Specifically, we will build an **AI-based surrogate model** for *ecosys*, significantly speeding up parameter calibration in SYMFONI's MDF. This calibration usually requires thousands of *ecosys* simulations — a 10-year *ecosys* simulation can take 0.5 hour in a normal CPU. An **AI-based surrogate model** can emulate *ecosys* responses to different model parameters without much accuracy loss, but can shorten a 10-year simulation from **0.5 hour in CPU** to **0.3 second in GPU** in our prior test at the Delta cluster at UIUC. To build this **AI-based surrogate model**, we used a recurrent neural network (RNN) architecture, with a typical example of Long Short Term Memory networks (LSTM). The inputs of the AI surrogate models include weather forecasting, management practices, and soil and vegetation parameters, while the outputs are target variables that can be observed. We will also use GPU-based optimization to further speed up the MDF process. In particular, we will use Nondominated Sorting Genetic Algorithm II (NSGA-II) [54,55]. NSGA-II uses a fast nondominated sorting approach to reduce the computational complexity of traditional multiobjective evolutionary algorithms from $O(MN^3)$ to $O(MN^2)$, where M is the number of objectives and N is the population size. More importantly, NSGA-II can leverage GPU for scalable and parallel implementation, further improving computational efficiency [56]. We will deploy our MDF system on the Summit GPU computing infrastructure. Our initial assessment has shown that by combining an AI surrogate model and GPU-based multi-objective optimization, we could achieve 10,000+ times more efficiency in the SYMFONI MDF. Specifically, to conduct a 10-year calibration experiment at a single site, we need to run the *ecosys* model for a total of 50,000 times (100 sampling seeds and 500 iterations per seed in NSGA-II), which would take ~0.5 CPU hr per 10 yr 50,000 = 25,000 CPU hr. While we can parallelize the model

evaluations over 100 sampling seeds, the 500 iterations need to run sequentially, requiring 250 node-hr on a 128-core CPU node. In contrast, with the AI surrogate and GPU-based optimization approach, the same calibration would only cost 2 minutes on a NVIDIA V100 Tensor Core GPU.

Tasks and Milestones: To achieve this objective, we will have the following three detailed tasks: (1) creating a synthetic dataset by running the *ecosys* model over randomly selected fields (10,000) in the U.S. Midwest; (2) training AI-based surrogate models for parameter calibration; and (3) calibrating the parameters using the AI-based surrogate models with satellite-based observations as constraints. For more details about these tasks, please check Step 1-3 in Fig. 5 and Section 3.1. The milestones are as follows: we will finish the first task at the end of Q1 of 2023 and finish the second task at the end of Q2 of 2023; we will finish parameter calibration for all the fields in 60% of the U.S. Midwestern States in Q3 and Q4 of 2023 and finish the remaining 40% fields of the U.S. Midwestern States in Q1 and Q2 of 2024 given the parameter calibration is most computationally intensive.

2.2 Obj-2: Quantify historical carbon budget of cropland at field level using the constrained carbon model

We will further build an **AI-based carbon outcome model** by learning from both observation-constrained *ecosys* model simulations and available observations of GHG emissions and carbon sequestration amounts. Multiple strategies of **physics-guided deep learning (PGDL)** will be implemented to build the **AI-based carbon outcome model** here, including pre-train AI models with *ecosys*-simulated database, reconstructing causal networks among different *ecosys* variables [57,58], mapping variable dependence structure (variable sequence) and variable nature (state or flux) in the *ecosys*, and adding constraints from real physical laws into the AI models (**Fig. 4**). Hierarchical multiscale recurrent neural networks (HM-RNNs) will be used to build the AI-based carbon outcome model, with carbon cycle related submodels for GPP, Ra, Rh, NEE, yield, and ΔSOC predictions, and other GHG emission related submodels for N_2O and CH_4 emission predictions (**Fig. 4**). In the AI-based carbon outcome model, we consider the connections and carbon balance between different carbon pools and fluxes explicitly (e.g., the direct connection of GPP with Ra and Yield, carbon inputs for Rh from plant residual after harvest, carbon balance equation of NEE=Ra+Rh-GPP and ΔSOC≈NEE-yield) (**Fig. 4**). To pre-train AI-based carbon outcome models, we will use the *ecosys* model with location-specific optimized parameters from the first objective to simulate a large database on GPP, Reco, harvested crop yield, CH_4, and N_2O emissions and all other intermediate variables. The pretrained models will be further fine-tuned with eddy-covariance observation data and remotely sensed yield data to improve the model performance. The AI models with a similar structure have been developed by PI Guan's group and collaborator Dr. Zhenong Jin's group from University of Minnesota, and have been successfully applied for predicting N_2O emission [59]. We will fully leverage the INCITE computational resources to build the AI-based carbon outcome models such that we can quantify historic carbon credit since 2000 accurately and efficiently at the field level.

Tasks and Milestones: To achieve this objective, we will have the following three detailed tasks: (1) generating the synthetic data for AI-based historical carbon outcome model training using the observation-constrained *ecosys* model; (2) building AI-based historical carbon outcome model based on the generated synthetic data; and (3) applying the AI-based carbon outcome model for historical field level carbon budget calculation. For more details about these tasks, please check Step 4-6 in Fig. 5 and Section 3.1. The milestones are as follows: we will finish the above three tasks for all the fields in half of the U.S. Midwestern States in Q3 and Q4 of 2023 and finish the remaining fields in another half of the U.S. Midwestern States in 2024 from Q1 to Q4.

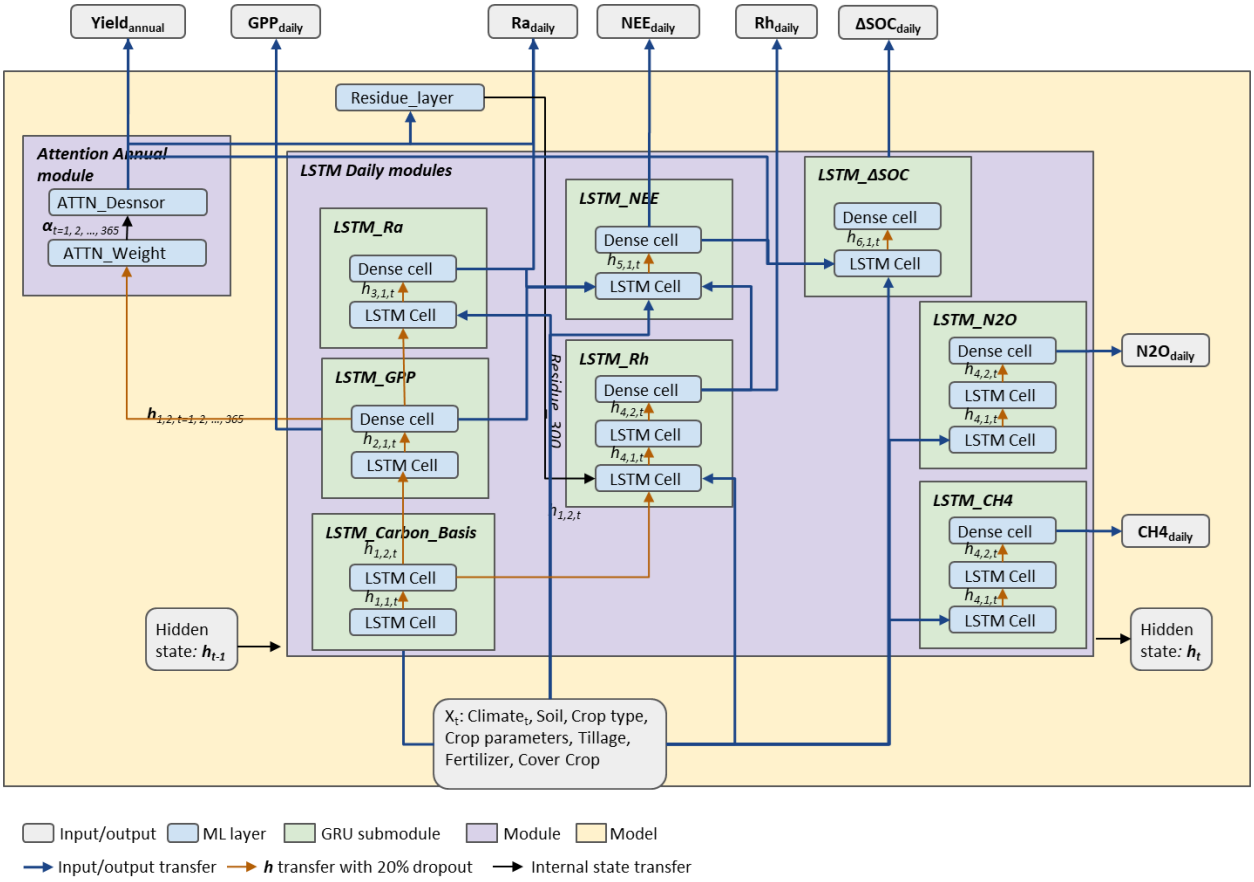


Fig. 4 An illustration of the **AI-based carbon outcome model** based on the physics-guided deep learning framework for quantifying crop productivity and carbon outcomes over farmland.

2. 3 Obj-3: Quantify carbon credit potential of cropland at field level using the constrained carbon model

We will further build AI-based carbon outcome models to quantify the carbon credit potential under different management scenarios. Specifically, we will consider the following types of scenarios over each field: (1) three tillage scenarios (conventional/reduced tillage/no-till), (2) two cover crop scenarios (with and without cover crops), (3) five fertilizer amount scenarios (current nitrogen fertilizer amount, reducing nitrogen fertilizer amount by 5%, 10%, 15%, and 20%), and (4) three fertilizer application window scenarios (100% fall application, 80% fall application with 20% sidedressing, and 50% fall application with 50% sidedressing). In total, we will test 3×2×5×3=90 scenarios for each field and get the potential GHG emissions and carbon sequestration amount for each scenario. The AI-based carbon outcome models will be used for large-scale simulations over every field in the U.S. Midwest. The output of scenario analysis will then be further analyzed for crop yield and environmental impacts of different management practice scenarios. Through this analysis, we can identify the best management practices to generate carbon credits and maintain crop yield for each individual field.

Tasks and Milestones: To achieve this objective, we will have the following three detailed tasks: (1) generating the synthetic data for AI-based carbon outcome model training using the observation-constrained *ecosys* model; (2) building AI-based historical carbon outcome model based on the generated synthetic data; and (3) applying the AI-based carbon outcome model for field level carbon budget calculation under different management scenarios. For more details about these tasks, please check Step 4-6 in Fig. 5 and Section 3.1. The milestones are as follows: we will finish the above three tasks for all the

fields in half of the U.S. Midwestern States in Q3 and Q4 of 2023 and finish the remaining fields in another half of the U.S. Midwestern States in 2024 from Q1 to Q4.

3. Computational Readiness

The proposed framework for field-level carbon budget and carbon outcome quantification contains six steps (**Fig. 5**), including: **[Step 1]** Generate the synthetic data for AI-based surrogate model training using the process-based model, *ecosys*; **[Step 2]** Built AI-based surrogate model based on the generated synthetic data from previous step; **[Step 3]** Calibrate the plant parameters using AI-based surrogate model for each individual field throughout the US Midwest; **[Step 4]** Generate synthetic data for AI-based carbon outcome model training using observation-constrained *ecosys*; **[Step 5]** Built AI-based carbon outcome model based on generated synthetic data; **[Step 6]** Quantifying carbon outcomes over the U.S. Midwest cropland fields using the AI-based carbon outcome model. CPU and GPU resources are used at different steps parallelly in this workflow.

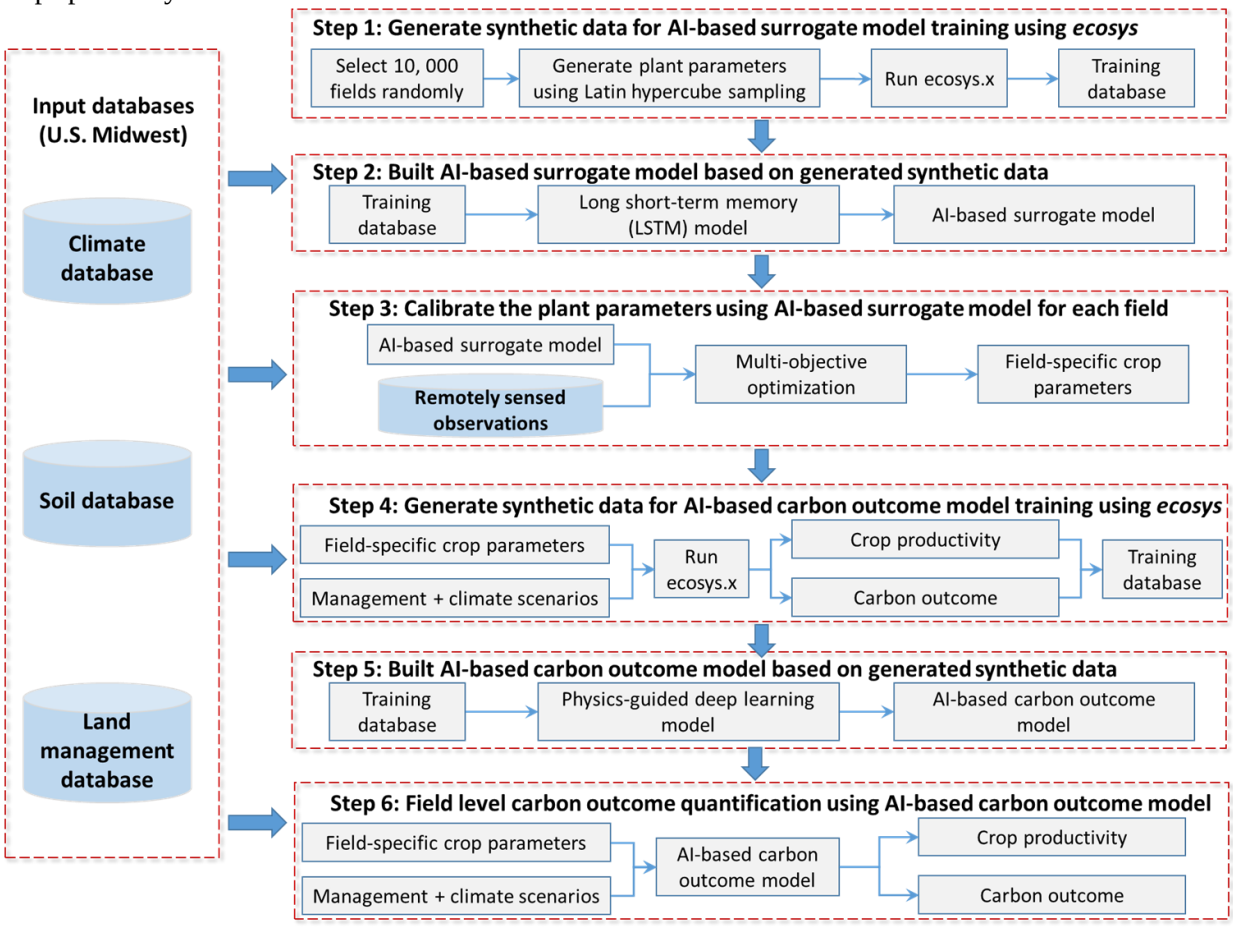


Fig. 5 Overview of the workflow.

3.1 Use of Resources Requested

The Obj-1 that aims to scale-up the AI-based MDF to constrain the process-based model throughout the U.S. Midwest cropland will be achieved by Step 1-3, and the Obj-2 and Obj-3 that aim to quantify historical carbon budget and carbon credit potential for each individual field throughout the U.S. Midwest cropland will be achieved by Step 4-6. The major computation resources needed by Step 1 and Step 4 are CPU resources, and those in Step 2-3 and Step 5-6 are GPU resources. Based on our detailed calculation and prior test on Theta through the ALCF Director's Discretionary allocation to PI Guan in 2022, we

conclude that **the total resources we will require for all the steps will be 1,421,955 CPU node hours, 234,763 GPU hours, 960 TB hard drive storage, 1 TB temporary local SSD storage.** The detailed information of computation resources and storage resources required for each step is as following:

Storage requirement for basic databases. The basic datasets that are used to drive the *ecosys* model and AI-based surrogate model include the hourly climate data from North American Land Data Assimilation System (NLDAS) with a spatial resolution of 0.125°, soil data from Gridded Soil Survey Geographic Database (gSSURGO) with a spatial resolution of 30 meter, crop type data from Cropland Data Layer (CDL) with a spatial resolution of 30 meter, field-level planting date, tillage date and intensity, and cover crop planting and terminate date derived from high resolution remotely sensed products (i.e., 30 meter daily STAIR satellite fusion product developed by PI Guan's group), and fertilizer rate for different crop types from USDA survey data. The data used to constrain the *ecosys* model include remotely sensed field level daily GPP, evapotranspiration (ET), leaf area index (LAI) and crop yield data averaged from 30 meter resolution. The storage needed is about 1.1 TB for climate data, about 0.3 TB for soil dataset, about 0.1 TB for CDL data, about 1.0 TB for field level land management data, and 2.0 TB for field-level GPP, ET, LAI, and yield data. Since the fertilizer data we used is at state level, the space needed can be ignored here. Thus, **the total storage needed for basic databases is about 4.5 TB.**

Computation resources and storage requirements for Step 1. In Step 1, we will generate the synthetic data for Step 2 to build the AI-based surrogate model of *ecosys* by running the *ecosys* model, which is written in Fortran 77. In this step, we will select 10,000 fields randomly throughout the US Midwest, and extract the climate, soil, management, and crop type data from the basic input databases. For the parameters of different crop types (i.e., corn, soybean, wheat), we will use the calibrated parameters based on the eddy-covariance observation from our previous studies as the baseline parameters, and sample the most sensitive plant parameters related to simulation of GPP, ET, LAI, and yield within the predefined ranges using the Latin hypercube sampling method. The total simulations we will run in this step is 10,000 simulations. From our experiments, one single simulation for one year using one core will take about 1 min. Thus, **the total computation resources of CPU** needed for this step is about 10,000 simulations * 1 core min/year simulations * 30 years / (64 cores/node) = **79 node hours**. For each 30 year simulations, the storage needed is about 4.5 MB, thus the **total storage** required for step 1 is about 4.5 MB/simulations * 10,000 simulations = **0.05 TB**. Since we will run the *ecosys* model and put the temporary outputs on local SSD to accelerate the program, the **required temporary local SSD in this step is 1 TB**.

Computation resources and storage requirements for Step 2. In Step 2, we will use a state-of-the-art artificial neural network, Long short-term memory (LSTM), to build the surrogate model of *ecosys* to predict daily GPP, ET, LAI, and yield simultaneously under different combinations of crop parameters using the synthetic dataset generated from Step 1. From our previous experiments on Delta cluster (managed by NCSA of UIUC) and Comet cluster (managed by UCSD/Scripps), **the computation resources needed in this step is about 20 GPU hours using the NVIDIA V100 GPU.** The storage required in this step is very low, which can be ignored here. The memory needed in this step is 64 GB.

Computation resources and storage requirements for Step 3. In Step 3, we will calibrate the plant parameters for each individual field using the AI-based surrogate model developed in Step 2 with the constraints of daily GPP, ET, LAI, and yield from the remote sensing observations. In the calibration process, we will use the multi-objective optimization method named NSGA-II, which could be paralleled during the calibration processes. From our previous experience, it will take about 50 seconds to calibrate a single field using the NVIDIA V100 GPU. Considering there are about 10 million of fields throughout the US Midwest, **the total GPU resources needed in this step is about 10,000,000 fields * 50 seconds V100 GPU/field = 138, 889 V100 GPU hours**. The storage needed for calibrated parameters is about 20 KB/field, thus the **total storage needed in this step is about 0.2 TB**.

Computation resources and storage requirements for Step 4. After calibrating the model at each field, we will conduct the forward simulation on 20% of select fields throughout the US Midwest using the *ecosys* model to generate the synthetic datasets to build **AI-based carbon outcome model** for the historical scenario using the historical management and climate information, and the hypothetical scenarios with difference farming management practices under both current climate and future climate change. The total

scenarios conducted in this step is 91 (1 for historical and 90 for hypothetical scenarios). Thus, **the total computation resources of CPU needed for this step is about 10,000,000 fields * 20% of total fields * 91 scenarios/field * 1 core min/year * 30 years/scenario / (64 cores/node) = 1,421,876 node hours.** The **total storage required** for step 4 is about 4.5 MB/scenario * 10,000,000 fields * 20% of total fields * 91 scenarios/field = **781 TB**. Since we will run the *ecosys* model and put the temporary outputs on local SSD to accelerate the program, the **required temporary local SSD in this step is 1 TB**.

Computation resources and storage requirements for Step 5. In this step, we will use the physics-guided deep learning model to build the surrogate model of *ecosys* to predict crop productivity and carbon outcomes simultaneously under different combinations of crop parameters, soil and climate conditions, farming management practices using the synthetic dataset generated from Step 4. We will train the carbon output model by state and scenario, to reduce the memory requirement and improve the model performance. From our previous experiments on Delta cluster and Comet cluster, the computation resources needed in this step is about 20 GPU hours using the NVIDIA V100 GPU for a single scenario and state, thus **the total GPU resources required for 91 scenarios and 11 states in Midwest is about 20 GPU hours/scenario/state * 91 scenarios * 11 states= 20,020 GPU hours**. The storage required in this step is very low, which can be ignored here. The memory needed in this step is 256 GB.

Computation resources and storage requirements for Step 6. We will apply the AI-based carbon outcome model built in Step 5 to all the fields throughout the US Midwest to predict field level crop productivity and carbon outcomes. From our previous experiments on Delta cluster and Comet cluster, the computation resources needed for one field and one scenarios is 0.3 second GPU, **thus the total GPU resources required for 91 scenarios and 10 millions in Midwest is about 0.3 second GPU/scenario/fields * 91 scenarios * 10,000,000 fields = 75, 834 GPU hours.** For one field and one scenarios, the storage needed is about 0.2 MB, thus the **total storage required is about 0.5 MB/scenario/fields * 91 scenarios * 10,000,000 fields = 174 TB.**

3.2 Computational Approach

The two major parts in the proposed framework are forward running the processes-based *ecosys* model (Step 1 and 4) and AI-based surrogate model of *ecosys* (Step 2-3 and Step 5-6). The *ecosys* model was developed using the language of Fortran 77. In our simulations, we treat each field as a single point, so that the simulations of each field can run independently from each other. We compiled the *ecosys* model using the intel ifort and icc compilers (version of 19.1.0.166 on Theta) with the optimization option of O2 to accelerate the computation process (we also tried to use the optimization option of O3 on other clusters, but the program is easily go crash, so we prefer to use O2 option for our simulations considering both the computation performance and stability).

For the forward running of the *ecosys* model, we use the language of Python with the library of mpi4py (compiled with OpenMPI) to do the parallel computation, by calling the external fortran program *ecosys.x* in each instance (compute core) (**Fig. 6**). For the simulations on each instance, we will first connect to the climate database (which has been preprocessed by converting from *.grd format to SQLite database format, and from space-before-time to time-before-space format to accelerate the data query speed using Python with the package of pygrib and sqlite3), soil and management databases, to extract the basic information for each field, and convert the extracted information to the standard input format of *ecosys* using the Python package of ecosystools developed by PI Guan's group to reduce the I/O burden. After that, we will distribute the tasks to each core using mpi4py, and run the simulations of each field independently by calling *ecosys.x*. After the external program finishes, the outputs of *ecosys* will be converted to SQLite database format and storage to the file system for each field independently to reduce the storage, I/O, inodes requirements and speed up the query process in the following analysis. To reduce the time on the communication of computation nodes and file system as well as reduce the I/O burden, we will run the *ecosys* model on the temporary local SSD, and also create and write the outputs to SQLite database to the local SSD first. When it is finished, we will move the SQLite databases to the file system for further analysis.

The AI-based surrogate model will be built using Python with the library of PyTorch, which can utilize the GPUs for deep learning model building and application with the help of cudatoolkit. The basic deep learning model we will use is the LSTM model with 2 recurrent layers and 64 features in the hidden state to consider the relationship of data in temporal space for Step 2 and 3. While in Step 5 and 6, we will build the AI-based carbon output model using the hierarchical structure with the basic unit of LSTM. Since the GPU resource used for surrogate model building is very small (~20 GPU hours), we will not do any further optimization here. For the application of AI-based surrogate model in the field-level crop parameters calibration, we will use the NSGII method with the help of Multi-objective Optimization in Python (pymoo) package (**Fig. 7**), which can obtain the optimized parameter using the genetic algorithm, and can using the GPU resource parallelly in each revolution step. From our previous testing on the Delta cluster, the calibration of a field using NSGII method could not fully utilize all the GPU computation cores, thus we will try to distribute the calibration tasks of four fields into one GPU with the help of mpi4py, and this will reduce half of the GPU time needed for each field calibration in average.

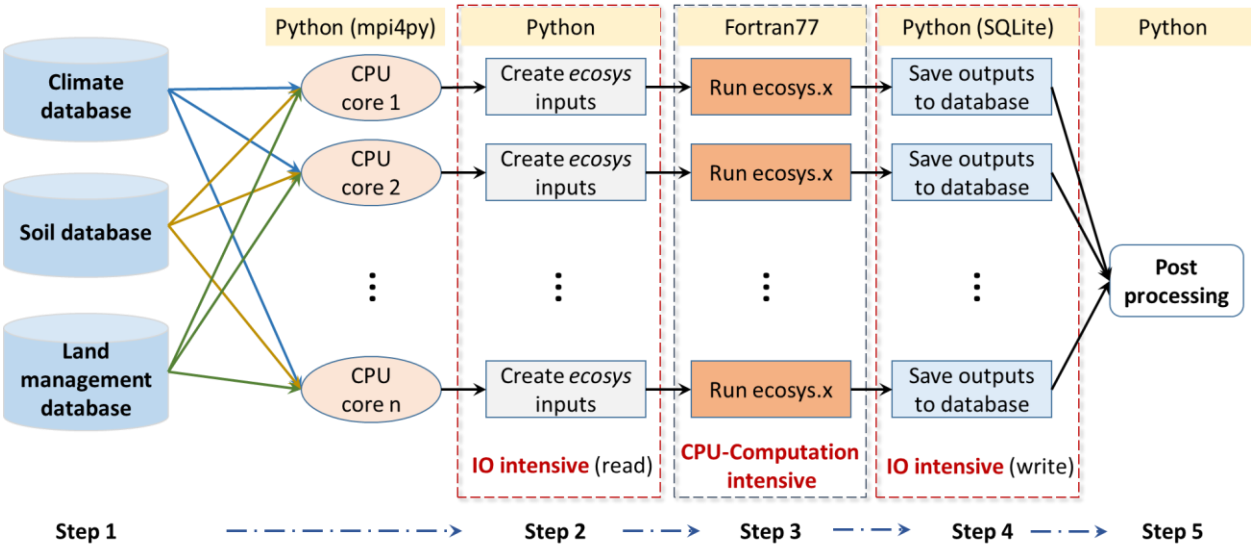


Fig. 6 Workflow for forward running the ecosys model.

While for the application of AI-based carbon outcome model, we will run the model county by county, and reorganize the inputs (e.g., build the lookup table of fields with the same climate data; input the soil information as static variable, and reshape it to time series with the same value) during the runtime of the program to reduce the IO burden. This will significantly reduce the GPU resource consumption from our previous experiment on using AI for 250 meter carbon budget prediction in the US Midwest.

The major software and packages we used in our workflow are as follows: **(1) Ecosys model**: the processed-based *ecosys* model written in Fortran 77 for forward simulation; **(2) Python scripts**: scripts used for data preprocessing and postprocessing, calling the external Fortran program, and distributing the tasks to each computation core; **(3) OpenMPI/mpi4py**: used for parallel computation, and distributing tasks to each computation core; **(4) Pytorch/CUDA**: used for building the application of the AI-based surrogate model on GPU; and **(5)** Other major Python packages used in the workflow:

- **NumPy,** the fundamental package for scientific computing in Python. It provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

- **Pandas,** a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python.
- **Torch,** an open source machine learning framework that accelerates the path from research prototyping to production deployment.
- **Subprocess,** call the external program.
- **Nass**, a wrapper around the public API for the USDA National Agricultural Statistics Service.
- **Osgeo/gdal,** a translator library for raster and vector geospatial data formats that is released under an MIT style Open Source License by the Open Source Geospatial Foundation. It presents a single raster abstract data model and single vector abstract data model to the calling application for all supported formats. It also comes with a variety of useful command line utilities for data translation and processing.
- **Rasterio**, a geographic information system that uses GeoTIFF and other formats to organize and store gridded raster datasets such as satellite imagery and terrain models. Rasterio reads and writes these formats and provides a Python API based on Numpy N-dimensional arrays and GeoJSON.
- **Geopandas**, an open source project to make working with geospatial data in Python easier. GeoPandas extends the datatypes used by pandas to allow spatial operations on geometric types. Geometric operations are performed by shapely. Geopandas further depends on fiona for file access and matplotlib for plotting.
- **Mpi4py,** a Python package builds on the MPI specification and provides an object oriented interface resembling the MPI-2 C++ bindings. It supports point-to-point (sends, receives) and collective (broadcasts, scatters, gathers) communication of any picklable Python object, as well as efficient communication of Python objects exposing the Python buffer interface (e.g. NumPy arrays and builtin bytes/array/memoryview objects).
- **Matplotlib,** a comprehensive library for creating static, animated, and interactive visualizations in Python.
- **Sqlite3**, a C library that provides a lightweight disk-based database that does not require a separate server process and allows accessing the database using a nonstandard variant of the SQL query language. Some applications can use SQLite for internal data storage. It's also possible to prototype an application using SQLite and then port the code to a larger database such as PostgreSQL or Oracle.
- **Lhsmdu**, a package that generates latin hypercube samples with multi-dimensional uniformity.
- **Shutil**, which offers a number of high-level operations on files and collections of files. In particular, functions are provided which support file copying and removal.
- **Pymoo**, an open-source framework offers state of the art single- and multi-objective algorithms and many more features related to multi-objective optimization such as visualization and decision making.

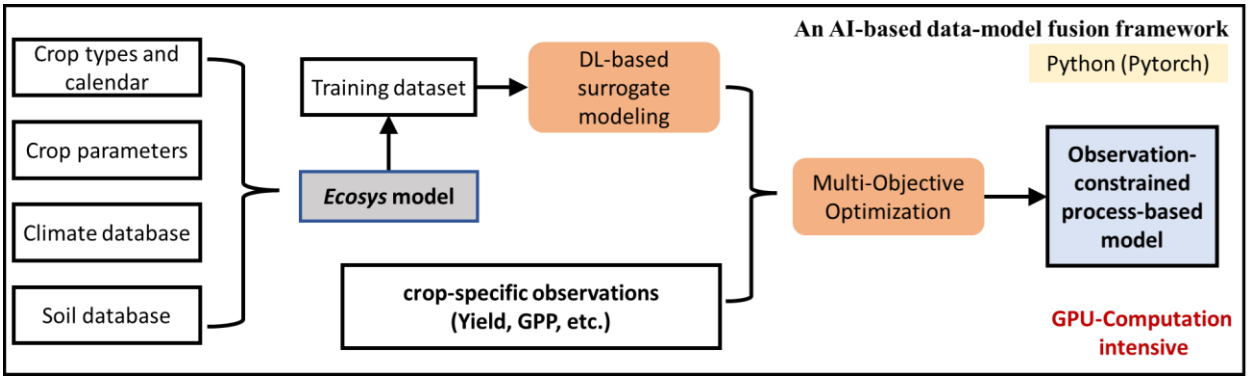
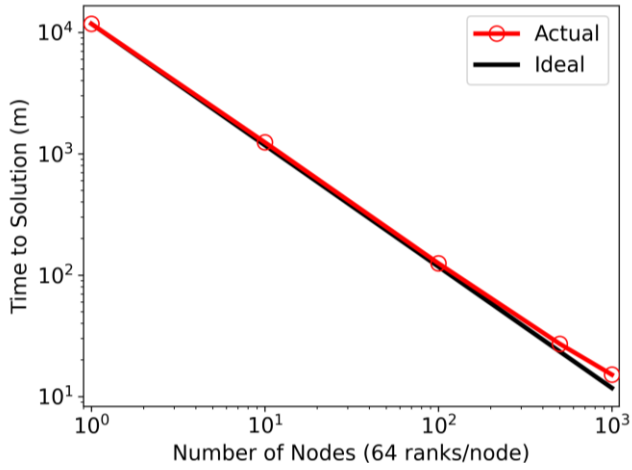


Fig. 7 AI-based model-data fusion framework.

3.3 Parallel Performance

To test the parallel performance of our workflow, we conducted a one-year forward simulation using the *ecosys* model on Theta through ALCF Director's Discretionary allocation to PI Guan in 2022. The forward simulation of the *ecosys* model is the most computation intensive part in our workflow. We tested the simulations using 1, 10, 100, 500, and 1000 nodes with 64 cores per node to evaluate the parallel performance. We treated the time that was used with 1 node as the ideal time to solve 64 one-year simulations, and calculated the time needed for the scenarios using different numbers of nodes assuming we have a total of 64, 000 simulations. The total actual time needed to solve the problem under specific node numbers was calculated by (64, 000 simulations)/(actual used node numbers)*(actual time used under specific node numbers).

From the simulations, we found that the forward simulations have satisfactory parallel performance, and the actual time used to solve the problem is very close to ideal time cost when the node numbers used are smaller than 100, while it is only slightly off from the ideal time cost when node numbers used are larger than 100 (**Fig. 8**). Since we used the processors to do the computations independently from each other (**Fig. 8**), the time that needed for each processor did not increase too much with the increasing of node numbers. With the increase of the node number, the I/O requirement will increase in the beginning and ending of the program, which will read the inputs and write the outputs to the filesystem server, thus the time needed to solve the problem only slightly increases with the increase of node numbers. To avoid the intensive I/O on slowing down the calculation efficiency problem in the actual production stage, we will run the simulations using every 500 nodes in one submit script separately.



nNodes (64 procs/node)	Time to Solution (m)	Ideal Time to Solution (m)
1	11743	11743
10	1239	1174
100	125	117
500	27	23
1000	15	12

Fig. 8 Examples of scaling performance tests on Theta, through the ALCF Director's Discretionary allocation to PI Guan in 2022.

Another time-consuming part of our workflow is using the AI-based surrogate model to derive the crop parameters of the process-based model and estimate the field-level carbon outcomes, which will use the GPU intensively. In this part of crop parameters calibration, the program will read daily climate, GPP, ET, LAI and yearly yield data of each field from the preprocessed databases stored on the filesystem. Given the data needed to read and write from/to the filesystem for each field is very small, the IO is not a limiting part of the parallel pipeline from our experience on the Delta cluster (managed by NCSA of UIUC). In the actual production stage, we will put the tasks of different fields on different GPU cores separately, and for each GPU core we will calibrate four fields parallelly to fully use the capability of each GPU core. For the application of AI-based carbon outcome models, the IO may be a limitation, especially from the input side. Considering there are 91 scenarios in total for Obj-2 and Obj-3 which have the same inputs and are independent from each other, we will combine the predictions of these two Objectives into a single code, which will significantly reduce the IO burden. Thus, the parallel of using GPU will not be a bottleneck of scaling from our experience on the Delta cluster.

3.4 Developmental Work

The PI Guan's group had already developed a Python package named **ecosystools** as the basic package by integrating all the functions needed to forward run the *ecosys* model, build the AI-based surrogate model, calibrate the crop parameters using the AI-based surrogate model, and store the inputs and outputs to the SQLite database, with the help of packages descripted in the Section 3.2. Although we have tested the performance of the workflow with thousands of fields throughout the Midwest on the Delta cluster, we will further improve the efficiency of the workflow on the Summit allocated in this proposal. Specifically, we will further test and develop the workflow on the Summit with 10, 000 fields distributed throughout the US Midwest. **The total resources needed at the development stages is about 7,188 CPU node hours, 1,891 GPU, and 4.4 TB hard drive storage.** The detailed information is as following:

- Computation resources and storage needed for the development of step 1 is about 10,000 simulations * 1 core min/year simulations * 30 years / (64 cores/node) = 78.2 node hours, and about 4.5 MB/simulations * 10,000 simulations = 0.05 TB.
- Computation resources needed for the development of step 2 is about 20 GPU hours using the NVIDIA V100 GPU.
- Computation resources and storage needed for the development of step 3 is about 10,000 fields * 50 seconds V100 GPU/field = 139 V100 GPU hours and 0.2 GB.
- Computation resources and storage needed for the development of step 4 is about 10,000 fields * 91 scenarios/field * 1 core min/year * 30 years/scenario / (64 cores/node) = 7,110 node hours, and about 4.5 MB/scenario * 10,000 fields * 91 scenarios/field = 3.9 TB.
- Computation resources needed for the development of step 5 is about 20 GPU hours/scenario * 91 scenarios = 1, 820 GPU hours (in the workflow development stage, we will train one model for all the states; while in the production stage, we will treat the model for different states separately).
- Computation resources and storage needed for the development of step 6 is about 0.3 second GPU/scenario/fields * 91 scenarios * 10,000 fields = 76 GPU hours, and about 0.5 MB/scenario/fields * 91 scenarios * 10,000 fields = 0.4 TB.

There is no further validation needed for development step 1 and 4; for step 2 and 5, we will validate the performance of developed AI-based surrogate model using *ecosys* generated synthetic data; for step 3, we will validate the performance of calibrated parameters by compared the predicted GPP, ET, LAI, and yield with remotely sensed observations; for step 6, we will validated the model performance using the meta-data analysis approach, by comparing the carbon outcome of different farming management practices derived by our approach with paired field experiments throughout the US Midwest, regarding both the direction and effect size of different farming management practices on crop productivity and carbon outcome [60-62].

REFERENCES

1. Hockstad L, Hanel L. Inventory of U.S. greenhouse gas emissions and sinks. Environmental System Science Data Infrastructure for a Virtual Ecosystem; 2018. doi:10.15485/1464240
2. Crippa M, Solazzo E, Guizzardi D, Monforti-Ferrario F, Tubiello FN, Leip A. Food systems are responsible for a third of global anthropogenic GHG emissions. Nature Food. 2021;2: 198–209.
3. Fargione JE, Bassett S, Boucher T, Bridgham SD, Conant RT, Cook-Patton SC, et al. Natural climate solutions for the United States. Sci Adv. 2018;4: eaat1869.
4. Bossio DA, Cook-Patton SC, Ellis PW, Fargione J, Sanderman J, Smith P, et al. The role of soil carbon in natural climate solutions. Nature Sustainability. 2020. pp. 391–398. doi:10.1038/s41893-020-0491-z
5. Griscom BW, Adams J, Ellis PW, Houghton RA, Lomax G, Miteva DA, et al. Natural climate solutions. Proc Natl Acad Sci U S A. 2017;114: 11645–11650.
6. Smith P, Martino D, Cai Z, Gwary D, Janzen H, Kumar P, et al. Greenhouse gas mitigation in agriculture. Philos Trans R Soc Lond B Biol Sci. 2008;363: 789–813.
7. DOE ARPA-E: DE-FOA-0002250. Systems For Monitoring And Analytics For Renewable Transportation Fuels From Agricultural Resources And Management (Smartfarm). 2020 Jan. Available: https://arpa-e-foa.energy.gov/Default.aspx?foaId=e7465e6e-cb9c-455c-9625-4c3db4386f00
8. Norman AG. Methods of Soil Analysis. Part 2. Chemical and Microbiological Properties. ASA-CSSA-SSSA; 1965.
9. Smith P. Monitoring and verification of soil carbon changes under Article 3.4 of the Kyoto Protocol. Soil Use Manage. 2006;20: 264–270.
10. Wendt JW, Hauser S. An equivalent soil mass procedure for monitoring soil organic carbon in multiple soil layers. Eur J Soil Sci. 2013;64: 58–65.
11. Baldocchi DD, Hincks BB, Meyers TP. Measuring Biosphere-Atmosphere Exchanges of Biologically Related Gases with Micrometeorological Methods. Ecology. 1988. pp. 1331–1340. doi:10.2307/1941631
12. Baldocchi D, Falge E, Gu L, Olson R, Hollinger D, Running S, et al. FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem-Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities. Bull Am Meteorol Soc. 2001;82: 2415–2434.
13. Buendia, Tanabe, Kranjc, Baasansuren. Refinement to the 2006 IPCC guidelines for national greenhouse gas inventories. IPCC: Geneva, Switzerland.
14. Ogle SM, Jay Breidt F, Easter M, Williams S, Killian K, Paustian K. Scale and uncertainty in modeled soil organic carbon stock changes for US croplands using a process-based model. Global Change Biology. 2010. pp. 810–822. doi:10.1111/j.1365-2486.2009.01951.x
15. Sándor R, Ehrhardt F, Brilli L, Carozzi M, Recous S, Smith P, et al. The use of biogeochemical models to evaluate mitigation of greenhouse gas emissions from managed grasslands. Science of The Total Environment. 2018. pp. 292–306. doi:10.1016/j.scitotenv.2018.06.020
16. Kucharik CJ, Brye KR, Norman JM, Foley JA, Gower ST, Bundy LG. Measurements and Modeling of Carbon and Nitrogen Cycling in Agroecosystems of Southern Wisconsin: Potential for SOC Sequestration during the Next 50 Years. Ecosystems. 2001;4: 237–258.
17. Luo Z, Baldock J, Wang E. Modelling the dynamic physical protection of soil organic carbon: Insights into carbon predictions and explanation of the priming effect. Glob Chang Biol. 2017;23: 5273–5283.
18. Zhou W, Guan K, Peng B, Tang J, Jin Z, Jiang C, et al. Quantifying carbon budget, crop yields and their responses to environmental variability using the ecosys model for U.S. Midwestern agroecosystems. Agric For Meteorol. 2021;307: 108521.
19. Guan K, Berry JA, Zhang Y, Joiner J, Guanter L, Badgley G, et al. Improving the monitoring of crop productivity using spaceborne solar-induced fluorescence. Glob Chang Biol. 2016;22: 716–726.
20. Guan K, Wu J, Kimball JS, Anderson MC, Frolking S, Li B, et al. The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields.

Remote Sens Environ. 2017;199: 333–349.
21. Peng B, Guan K, Pan M, Li Y. Benefits of seasonal climate prediction and satellite data for forecasting U.S. maize yield. Geophys Res Lett. 2018;45: 9662–9671.
22. Peng B, Guan K, Zhou W, Jiang C, Frankenberg C, Sun Y, et al. Assessing the benefit of satellite-based Solar-Induced Chlorophyll Fluorescence in crop yield prediction. Int J Appl Earth Obs Geoinf. 2020;90: 102126.
23. Jiang C, Guan K, Wu G, Peng B, Wang S. A daily, 250 m and real-time gross primary productivity product (2000–present) covering the contiguous United States. Earth Syst Sci Data. 2021;13: 281–298.
24. Potash E, Guan K, Margenot A, Lee D, DeLucia E, Wang S, et al. How to estimate soil organic carbon stocks of agricultural fields? Perspectives using ex-ante evaluation. Geoderma. 2022;411: 115693.
25. Huang Y, Yu Y, Zhang W, Sun W, Liu S, Jiang J, et al. Agro-C: A biogeophysical model for simulating the carbon budget of agroecosystems. Agric For Meteorol. 2009;149: 106–129.
26. Li C, Frolking S, Harriss R. Modeling carbon biogeochemistry in agricultural soils. Global Biogeochem Cycles. doi:10.1029/94GB00767
27. Shi Z, Crowell S, Luo Y, Moore B. Model structures amplify uncertainty in predicted soil carbon responses to climate change. Nature Communications. 2018. doi:10.1038/s41467-018-04526-9
28. Sulman BN, Moore JAM, Abramoff R, Averill C, Kivlin S, Georgiou K, et al. Multiple models and experiments underscore large uncertainty in soil carbon dynamics. Biogeochemistry. 2018;141: 109–123.
29. Mishra U, Drewniak B, Jastrow JD, Matamala RM, Vitharana UWA. Spatial representation of organic carbon and active-layer thickness of high latitude soils in CMIP5 earth system models. Geoderma. 2017;300: 55–63.
30. Jung M, Vetter M, Herold M, Churkina G, Reichstein M, Zaehle S, et al. Uncertainties of modeling gross primary productivity over Europe: A systematic study on the effects of using different drivers and terrestrial biosphere models. Global Biogeochem Cycles. 2007;21. doi:10.1029/2006gb002915
31. Paustian K, Collier S, Baldock J, Burgess R, Creque J, DeLonge M, et al. Quantifying carbon for agricultural soil management: from the current status toward a global soil information system. Carbon Management. 2019;10: 567–587.
32. Fer I, Kelly R, Moorcroft PR, Richardson AD, Cowdery EM, Dietze MC. Linking big models to big data: efficient ecosystem model calibration through Bayesian model emulation. Biogeosciences. 2018;15: 5801–5830.
33. Luo Y, Guan K, Peng J, Wang S, Huang Y. STAIR 2.0: A Generic and Automatic Algorithm to Fuse Modis, Landsat, and Sentinel-2 to Generate 10 m, Daily, and Cloud-/Gap-Free Surface Reflectance Product. Remote Sensing. 2020;12: 3209.
34. Luo Y, Guan K, Peng J. 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. 2018;214: 87–99.
35. Guan K, Good SP, Caylor KK, Medvigy D, Pan M, Wood EF, et al. Simulated sensitivity of African terrestrial ecosystem photosynthesis to rainfall frequency, intensity, and rainy season length. Environ Res Lett. 2018;13: 025013.
36. Cai Y, Guan K, Lobell D, Potgieter AB, Wang S, Peng J, et al. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric For Meteorol. 2019;274: 144–159.
37. Li Y, Guan K, Yu A, Peng B, Zhao L, Li B, et al. Toward building a transparent statistical model for improving crop yield prediction: Modeling rainfed corn in the U.S. Field Crops Res. 2019;234: 55–65.
38. Wang S, Guan K, Zhang C, Lee D, Margenot AJ, Ge Y, et al. Using soil library hyperspectral reflectance and machine learning to predict soil organic carbon: Assessing potential of airborne and spaceborne optical soil sensing. Remote Sens Environ. 2022;271: 112914.
39. Wang S, Guan K, Wang Z, Ainsworth EA, Zheng T, Townsend PA, et al. Airborne hyperspectral

imaging of nitrogen deficiency on crop traits and yield of maize by machine learning and radiative transfer modeling. Int J Appl Earth Obs Geoinf. 2021;105: 102617.
40. Kimm H, Guan K, Gentine P, Wu J, Bernacchi CJ, Sulman BN, et al. Redefining droughts for the U.S. Corn Belt: The dominant role of atmospheric vapor pressure deficit over soil moisture in regulating stomatal behavior of Maize and Soybean. Agric For Meteorol. 2020;287: 107930.
41. Jiang C, Ryu Y, Wang H, Keenan TF. An optimality-based model explains seasonal variation in C3 plant photosynthetic capacity. Glob Chang Biol. 2020;26: 6493–6510.
42. Jiang C, Guan K, Pan M, Ryu Y, Peng B, Wang S. BESS-STAIR: a framework to estimate daily, 30-meter, and allweather crop evapotranspiration using multi-source satellite data for the U.S. Corn Belt. Hydrol Earth Syst Sci Discuss. 2019; 1–36.
43. Daw A, Karpatne A, Watkins W, Read J, Kumar V. Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling. arXiv [cs.LG]. 2017. Available: http://arxiv.org/abs/1710.11431
44. Read JS, Jia X, Willard J, Appling AP, Zwart JA, Oliver SK, et al. Process-guided deep learning predictions of lake water temperature. Water Resour Res. 2019;55: 9173–9190.
45. Hanson PC, Stillman AB, Jia X, Karpatne A, Dugan HA, Carey CC, et al. Predicting lake surface water phosphorus dynamics using process-guided machine learning. Ecol Modell. 2020;430: 109136.
46. Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, Ganguly A, et al. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. IEEE Trans Knowl Data Eng. 2017;29: 2318–2331.
47. Grant R. A Review of the Canadian Ecosystem Model — ecosys. Modeling Carbon and Nitrogen Dynamics for Soil Management. 2001. doi:10.1201/9781420032635.ch6
48. Grant RF, Baldocchi DD, Ma S. Ecological controls on net ecosystem productivity of a seasonally dry annual grassland under current and future climates: Modelling with ecosys. Agricultural and Forest Meteorology. 2012. pp. 189–200. doi:10.1016/j.agrformet.2011.09.012
49. Grant RF, Juma NG, Robertson JA, Izaurralde RC, McGill WB. Long-Term Changes in Soil Carbon under Different Fertilizer, Manure, and Rotation. Soil Science Society of America Journal. 2001. pp. 205–214. doi:10.2136/sssaj2001.651205x
50. Grant RF, Dyck M, Puurveen D. 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. 2020;100: 408–429.
51. Metivier KA, Pattey E, Grant RF. Using the ecosys mathematical model to simulate temporal variability of nitrous oxide emissions from a fertilized agricultural soil. Soil Biol Biochem. 2009;41: 2370–2386.
52. Grant RF, Roulet NT. Methane efflux from boreal wetlands: Theory and testing of the ecosystem model Ecosys with chamber and tower flux measurements. Global Biogeochemical Cycles. 2002. pp. 2–1. doi:10.1029/2001gb001702
53. Grant RF. Changes in soil organic matter under different tillage and rotation: Mathematical modeling in ecosys. Soil Sci Soc Am J. 1997;61: 1159–1175.
54. Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput. 2002;6: 182–197.
55. Deb K, Jain H. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. IEEE Transactions on Evolutionary Computation. 2014. pp. 577–601. doi:10.1109/tevc.2013.2281535
56. Aguilar-Rivera A. A GPU fully vectorized approach to accelerate performance of NSGA-2 based on stochastic non-domination sorting and grid-crowding. Appl Soft Comput. 2020;88: 106047.
57. Runge J, Bathiany S, Bollt E, Camps-Valls G, Coumou D, Deyle E, et al. Inferring causation from time series in Earth system sciences. Nature Communications. 2019. doi:10.1038/s41467-019-10105-3
58. Runge J, Nowack P, Kretschmer M, Flaxman S, Sejdinovic D. Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances. 2019. doi:10.1126/sciadv.aau4996
59. Liu L, Xu S, Tang J, et al. KGML-ag: a modeling framework of knowledge-guided machine learning

to simulate agroecosystems: a case study of estimating N 2 O emission using data from mesocosm experiments. Geoscientific Model Development, 2022, 15(7): 2839-2858.
60. Jian J, Du X, Reiter M S, et al. A meta-analysis of global cropland soil carbon changes due to cover cropping. Soil Biology and Biochemistry, 2020, 143: 107735.
61. Luo Z, Wang E, Sun O J. Can no-tillage stimulate carbon sequestration in agricultural soils? A meta-analysis of paired experiments. Agriculture, ecosystems & environment, 2010, 139(1-2): 224-231.
62. Van Kessel C, Venterea R, Six J, et al. Climate, duration, and N placement determine N2O emissions in reduced tillage systems: a meta-analysis. Global change biology, 2013, 19(1): 33-44.

PERSONNEL JUSTIFICATION AND MANAGEMENT PLAN

PERSONNEL JUSTIFICATION

The project team consists of **PI Dr. Kaiyu Guan**, **Co-PI Dr. Bin Peng**, supervising a postdoctoral researcher (identified) and a graduate student (committed). **Dr. Guan** is an ecohydrologist and crop modeler who has rich expertise in developing and using process-based models for crop growth and soil biochemistry. He is an associate professor in the Department of Natural Resources and Environmental Sciences, founding director of the Agroecosystem Sustainability Center, and also a Blue Waters Professor with National Center for Supercomputing Applications at University of Illinois at Urbana-Champaign (UIUC). Guan has been the principal investigator for DOE SMARTFARM Phase I and II projects, multiple NASA, NSF, and USDA NIFA projects, and the co-investigator for multiple other federally-funded projects. **Dr. Peng** is a senior research scientist and research assistant professor at UIUC, with extensive expertise in crop modeling, ecosystem modeling, and remote sensing. Dr. Peng has been involved in several federally-funded projects as Co-PI and Co-Investigator.

Dr. Guan and Dr. Peng's team have extensive experiences in using supercomputers and other HPC-related resources in their research. They have been jointly awarded the **2020 HPC Innovation Excellence Award** by Hyperion Research, for their innovative work on using HPC to improve earth system modelings and satellite applications. In the past when Blue Waters Supercomputer was still in function, Guan and Peng had an annual allocation of ~1 million node hours of CPU, through both the Blue Waters Professorship allocation and Illinois Blue Waters Competition for allocation. Guan and Peng have developed a major research program at UIUC in using HPC for agroecosystem research in the past six years; the team has accumulated rich and extensive HPC experiences in handling large-scale computing efforts, with recent advancements in including more GPU for their AI-based applications.

Currently, this team has secured other federal funding to support the personnel of a postdoc and a graduate student to work on this project, funded from the DOE ARPA-E SMARTFARM project and the NSF Signal-in-Soil project. Both the postdoctoral researcher and graduate student have been identified for this project and can start the project anytime once the allocation is awarded. The postdoc will be mainly responsible for conducting AI-based model-data fusion and large-scale computational modeling of agroecosystem. The postdoctoral research will be secured with a two-year contract to cover the whole lifetime of this project, and the graduate student will work on this project and other related projects under supervision of Dr. Guan and Dr. Peng.

MANAGEMENT PLAN

Dr. Guan and Dr. Peng will lead this project together to manage the project deliverable and computing resource allocations/progress. Specifically, Dr. Guan will oversee the project progress and make sure the scientific goals will be achieved; Dr. Peng will be in charge of the specific research progress and also ensure the computing resources are used properly to achieve the research goals. Dr. Peng will also monitor the usage associated with our INCITE award to ensure that usage is only for the project being described herein and that all U. S. Export Controls are complied with.

Dr. Guan and Dr. Peng will supervise the postdoctoral researcher and graduate student on computational modeling, data analysis, and publishing. The project team will have weekly meetings to track progress on computational tasks, resource utilization, and scientific analysis. Dr. Guan will provide updates on the status of the work including publications, awards, and highlights of accomplishments from this project. Dr. Guan and Dr. Peng will guide the team manage the data generated from this project following all the data managing protocols specified by the INCITE Program. At the end of the project, we will move all the data and code out of the targeted HPC machine and get them archived into the Taiga data storage system hosted at NCSA of UIUC, where PI Guan has a dedicated 700TB space to use.

The project team aims to make data produced through the research publicly available as much as possible. We plan to use the data for generating publications and to publish them in high-impact and accessible journals.

Proposal Title: Quantifying carbon outcomes over cropland using AI-based model-data fusion

Year 1 hours	**Total number of node-hours for Year 1:** 711,017CPU node hours + 131,280 GPU		
Milestone:	**Details (as appropriate):**	**Dates:**	**Status: (renewals only)**
Obj-1: Scale-up the AI-based model-data fusion to constrain uncertainty in carbon model (Finish 60% of this Obj-1)	**Resource:** CPU + GPU **Node-hours:** 79 CPU node hours + 83,353 GPU hours **Production size runs (number of nodes):** 100 CPU nodes or 100 GPU nodes **Filesystem storage (TB and dates):** 0.17 TB **Archival storage (TB and dates):** 0.17 TB **Software Application:** *ecosys* (Intel Fortran), PyTorch (Python), ecosystools (Python) **Tasks:** (1) generate the synthetic data for AI-based surrogate model training using the process-based model, *ecosys*; (2) built AI-based surrogate model based on the generated synthetic data; (3) calibrate the plant parameters using the AI-based surrogate model for 60% fields throughout the US Midwest. **Dependencies:** (1) Task (2) of Obj-1 depends on Task (1) of Obj-1; (2) Task (3) of Obj-1 depends on Task (2) of Obj-1;	2023.01.01-2023.12.31	
Obj-2: Quantify historical carbon budget of cropland at field level using the constrained carbon model (Finish 50% of Obj-2)	**Resource:** CPU + GPU **Node-hours:** 7,813 CPU node hours + 527 GPU hours **Production size runs (number of nodes):** 500 CPU nodes or 100 GPU nodes **Filesystem storage (TB and dates):** 5.25 TB **Archival storage (TB and dates):** 5.25 TB **Software Application:** *ecosys* (Intel Fortran), ecosystools (Python) **Tasks:** (1) generate the synthetic data for AI-based historical carbon outcome model training using the process-based model, *ecosys* for half of the states in the US Midwest; (2) build AI-based historical carbon outcome model based on the generated synthetic data for half of the states in the US Midwest; (3) apply the AI-based carbon outcome model for historical field level carbon budget calculation for half of the states in the US Midwest. **Dependencies:** (1) Obj-2 depends on Task (3) of Obj-1. Since we will conduct the simulations county by county, this means that we can parallel conduct the computation of tasks of Obj-2 and Obj-1 for different counties. (2) Task (2) of Obj-2 depends on Task (1) of Obj-2, and Task (3) of Obj-2 depends on Task (2) of Obj-2. Since we plan to train and apply the carbon outcome model	2023.07.01-2023.12.31	

	state by state, this means that we can parallelly conduct the computation of tasks of Obj-2 for different states.		
Obj-3: Quantify carbon credit potential of cropland at field level using the constrained carbon model (Finish 50% of Obj-3)	**Resource:** CPU + GPU **Node-hours:** 703,125 CPU node hours + 47,400 GPU hours **Production size runs (number of nodes):** 500 CPU nodes or 100 GPU nodes **Filesystem storage (TB and dates):** 472 TB **Archival storage (TB and dates):** 472 TB **Software Application:** *ecosys* (Intel Fortran), ecosystools (Python) **Tasks:** (1) generate the synthetic data for AI-based carbon outcome model training using the process-based model, *ecosys* under different farming managements and climate scenarios for half of the states in the US Midwest; (2) build AI-based carbon outcome model based on the generated synthetic data under different farming managements and climate scenarios for half of the states in the US Midwest; (3) apply the AI-based carbon outcome model for field-level carbon budget calculation under different farming managements and climate scenarios for half of the states in the US Midwest. **Dependencies:** (1) Obj-3 depends on Task (3) of Obj-1. Since we will conduct the simulations county by county, this means that we can parallel conduct the computation of tasks of Obj-3 and Obj-1 for different counties. (2) Task (2) of Obj-3 depends on Task (1) of Obj-3, and Task (3) of Obj-3 depends on Task (2) of Obj-3. Since we plan to train and apply the carbon outcome model state by state, this means that we can parallel conduct the computation of tasks of Obj-3 for different states.	2023.07.01-2023.12.31	
Year 2 (if appropriate) **Total number of node-hours for Year 2:** 710,938 CPU node hours + 103,483 GPU hours			
Obj-1: Scale-up the AI-based model-data fusion to constrain uncertainty in carbon model (Finish the left 40% of this Obj-1)	**Resource:** CPU + GPU **Node-hours:** 55,556 GPU hours **Production size runs (number of nodes):** 100 CPU nodes or 100 GPU nodes **Filesystem storage (TB and dates):** 0.08 TB **Archival storage (TB and dates):** 0.08 TB **Software Application:** *ecosys* (Intel Fortran), PyTorch (Python), ecosystools (Python) **Tasks:** (1) calibrate the plant parameters using the AI-based surrogate model for the left 40% fields throughout the US Midwest. **Dependencies:** Depending on Task (2) of Obj-1 in Year 1;	2024.01.01-2024.6.30	

Obj-2: Quantify historical carbon budget of cropland at field level using the constrained carbon model (Finish the left 50% of Obj-2)	**Resource:** CPU + GPU **Node-hours:** 7,813 CPU node hours + 527 GPU hours **Production size runs (number of nodes):** 500 CPU nodes or 100 GPU nodes **Filesystem storage (TB and dates):** 5.25 TB **Archival storage (TB and dates):** 5.25 TB **Software Application:** *ecosys* (Intel Fortran), ecosystools (Python) **Tasks:** (1) generate the synthetic data for AI-based historical carbon outcome model training using the process-based model, *ecosys* for the left 50% states in the US Midwest; (2) build AI-based historical carbon outcome model based on the generated synthetic data for the left 50% states in the US Midwest; (3) apply the AI-based carbon outcome model for historical field level carbon budget calculation for the left 50% states in the US Midwest. **Dependencies:** (1) Obj-2 depends on the Task of Obj-1 in Year 2. Since we will conduct the simulations county by county, this means that we can parallel conduct the computation of tasks of Obj-2 and Obj-1 for different counties. (2) Task (2) of Obj-2 depends on Task (1) of Obj-2 in Year 2, and Task (3) of Obj-2 depends on Task (2) of Obj-2 in Year 2. Since we plan to train and apply the carbon outcome model state by state, this means that we can parallelly conduct the computation of tasks of Obj-2 in Year 2 for different states.	2024.01.01-2024.12.31	
Obj-3: Quantify carbon credit potential of cropland at field level using the constrained carbon model (Finish the left 50% of Obj-3)	**Resource:** CPU + GPU **Node-hours:** 703,125 CPU node hours+ 47,400 GPU hours **Production size runs (number of nodes):** 500 CPU nodes or 100 GPU nodes **Filesystem storage (TB and dates):** 472 TB **Archival storage (TB and dates):** 472 TB **Software Application:** *ecosys* (Intel Fortran), ecosystools (Python) **Tasks:** (1) generate the synthetic data for AI-based carbon outcome model training using the process-based model, *ecosys* under different farming managements and climate scenarios for the left 50% states in the US Midwest; (2) build AI-based carbon outcome model based on the generated synthetic data under different farming managements and climate scenarios for the left 50% states in the US Midwest; (3) apply the AI-based carbon outcome model for field level carbon budget calculation under different farming managements and climate scenarios for the left 50% states in the US Midwest. **Dependencies:**	2024.01.01-2024.12.31	

	(1) Obj-3 depends on Task (3) of Obj-1 in Year 2. Since we will conduct the simulations county by county, this means that we can parallelly conduct the computation of tasks of Obj-3 and Obj-1 for different counties. (2) Task (2) of Obj-3 depends on Task (1) of Obj-3 in Year 2, and Task (3) of Obj-3 depends on Task (2) of Obj-3 in Year 2. Since we plan to train and apply the carbon outcome model state by state, this means that we can parallel conduct the computation of tasks of Obj-3 in Year 2 for different states.		

PUBLICATIONS RESULTING FROM INCITE AWARDS

N/A as this is our first INCITE application.

Curriculum Vitae

Kaiyu Guan
Agroecosystem Sustainability Center, University of Illinois Urbana-Champaign
609-647-1368, kaiyug@illinois.edu

Education and Training

Princeton University	Environmental Engineering (Advisor: Eric Wood)	Ph.D.	2013
Princeton University	Environmental Engineering (Advisor: Eric Wood)	M.A.	2010
Nanjing University	Geography and Remote Sensing	B.Sc.	2008

Research and Professional Experience

Founding Director	Agroecosystem Sustainability Center, UIUC	2021 - Present
Blue Waters Professor	Natural Resources and Environmental Science & National Center for Supercomputing Applications, UIUC (Assistant: 2016-2021; Associate: 2021-present)	2016 - present
Quantitative Scientist	The Climate Corporation	2015 - 2016
Postdoc fellow	Stanford University	2013 - 2015

Five Publications Most Relevant to This Proposal

Guan has published 120+ papers in leading scientific journals, including Science, Nature, Nature Geoscience, Nature Food, Global Change Biology, Remote Sensing of Environment etc. His Google Scholar Page: https://scholar.google.com/citations?user=YLjpc_cAAAAJ&hl=en (by May 30, 2022, H-index: 43; Total citations: 6291). Selected Publications (Students or postdoctoral researchers supervised by Guan are in bold and underlined):

1. **Guan, K.***, Berry, J., Zhang, Y., Guanter, L., Joiner, J. and Lobell, D.B. (2016). Improving the monitoring of crop productivity using spaceborne solar-induced fluorescence. Global Change Biology 22(2): 716-726.
2. **Guan, K. ***, Wu, J., Kimball, J., Anderson, M., Li, B., and Lobell, D. (2017). The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields. Remote Sensing of Environment 199: 333-349.
3. **Luo, Y.**, **Guan, K.***, and 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 Sensing of Environment 214: 87-99.
4. **Jiang, C.**, **Guan, K.***, **Wu, G.**, **Peng, B.**, and **Wang, S**. (2020). A daily, 250 m, and real-time gross primary productivity product (2000–present) covering the Contiguous United States, Earth System Science Data.
5. **Zhou, W.*, Guan, K.***, **Peng, B.***, Tang, J., Jin, Z., **Jiang, C.**, Grant R., and Mezbahuddin S. (2021) Quantifying carbon budget, crop yields and their responses to environmental variability using the ecosys model for U.S. Midwestern agroecosystems, Agricultural and Forest Meteorology.

Research Interests and Expertise

I use satellite data, computational models, fieldwork, and machine learning approaches to address how climate and human practices affect crop productivity, water resource availability, and ecosystem functioning. I have keen interests in applying my knowledge and skills in solving real-life problems, such as large-scale crop monitoring and forecasting, water management and sustainability, and global food security. My lab closely works with scientists in computer science (deep learning, high performance computing), plant physiologists, agronomists, and economists in addressing the above real-world challenges. Since Feb 2016 joining UIUC, my group has received ~$20M for his own lab, ~$12M of which came from Guan's Lead PI projects. Majority of the funding comes from federal agencies, including NASA, NSF, USDA, DOE, etc.

Synergistic Activities

1. **Government Grant Proposal Review Panels**: Served in 20 panel reviews for various federal agencies, including NASA, NSF, USDA, DOE, and USAID.
2. **Editorial Service and Journal Reviewers**: Guest editor for Biogeoscience (2017) and Frontier in Big Data (2018), special editor for PNAS (2019). Serve as reviewers for leading domain journals (6-10 manuscripts/year).
3. **Service to Disciplinary and Professional Societies or Associations**: NASA Science Team of Carbon Monitoring System Science Team, 2017-present; NASA Science Team of NASA Harvest Program (NASA's Food Security and Agriculture Program)
4. **Conference session organizer**: Organized 17 sessions for the AGU fall meeting from 2012-2021.
5. **Undergraduate Mentorship:** Dr. Guan has served as a mentor for the NSF REU-INCLUSION and UIUC Students Pushing Innovation (SPIN) program targeting undergraduate students for gaining computational research experience. For the last six years, Guan has mentored >30 undergraduate students (five female students and ten students from minority groups) in the REUINCLUSION and SPIN programs.

Collaborators

Collaborator name	Current institution
Ainsworth, Elizabeth	UIUC
Anderegg, William	University of Utah
Anderson, Matha	USDA-ARS
Arkebauer, Timothy	University of Nebraska
Bernacchi, Carl	UIUC/USDA
Berry, Joseph	Carnegie Institute for Science
Caylor, Kelly	UCSB
Chen, Min	University of Wisconsin
DeLucia, Evan	UIUC
Elliott, Joshua	DAPAR
Funk, Chris	UCSB
Gao, Feng	USDA-ARS
Gentine, Pierre	Columbia University
Grayson, Badgley	Carbon Plan
Guanter, Luis	Universitat Politècnica de València
Hammer, Greame	The University of Queensland
Jain, Meha	University of Michigan
Joiner, Joanna	NASA Goddard
Kimball, John	University of Montana
Konings, Alexandra	Stanford University
Lawrence, David	NCAR
Li, Bo	UIUC
Li, Yan	Beijing Normal University
Li, Zhan	Canadian Forest Service
Lobell, David	Stanford University
Long, David G	Brigham Young University

Collaborator name	Current institution
Luo, Yunan	UIUC
McDowell, Nate	DOE
Meacham, Katherine	UIUC
Medvigy, David	University of Notre Dame
Meinzer, Frederick	Oregon State University
Miao, Guofang	UIUC
Moore, Caitlin	UIUC
Pan, Ming	UCSD
Peng, Bin	UIUC
Peng, Jian	UIUC
Pokhrel, Yadu	Michigan State University
Saleska, Scott	University of Arizona
Sato, Hisashi	Hokkaido University
Seifert, Christopher	Industry
Sheffield, Justin	University of Southampton
Shukla, Shrad	UCSB
Sultan, Benjamin	France Join institution
Suyker, Andrew	University of Nebraska
Urban, Daniel	Industry
Wang, Shaowen	UIUC
Wardlow, Brian	University of Nebraska
Wu, Jin	University of Hong Kong
Xu, Xiangtao	Cornell University
Yang, Xi	University of Virginia
Zhang, Yongguang	Nanjing University
Zhao, Lei	UIUC

Curriculum Vitae

Bin Peng
Institute for Sustainability, Energy, and Environment
University of Illinois at Urbana-Champaign
217-974-5389, binpeng@illinois.edu

Education and Training

Chinese Academy of Sciences	Geoinformatics	Ph.D.	2016
Nanjing University	Geography and Remote Sensing	B.Sc.	2011

Research and Professional Experience

Senior Research Scientist	ISEE, UIUC	2021 - Present
Research Assistant Professor	NRES, UIUC	2021 - Present
Research Scientist	NRES, UIUC	2020 - 2021
Postdoctoral Research Fellow	NCSA, UIUC	2016 - 2020

Five Publications Most Relevant to This Proposal

1. Peng, B.*, Guan, K. (2021). Harmonizing climate-smart and sustainable agriculture. Nature Food.
2. Zhou, W.*, Guan, K.*, Peng, B.*, Tang, J., Jin, Z., Jiang, C., Grant R., and Mezbahuddin S. (2021) Quantifying carbon budget, crop yields and their responses to environmental variability using the ecosys model for U.S. Midwestern agroecosystems, Agricultural and Forest Meteorology.
3. Peng, B.*, Guan, K.*, Tang, J., Ainsworth, E.A., Asseng, S., Bernacchi, C.J., Cooper, M., Delucia, E.H., Elliott, J.W., Ewert, F., Grant, R.F., Gustafson, D.I., Hammer, G.L., Jin, Z., Jones, J.W., Kimm, H., Lawrence, D.M., Li, Y., Lombardozzi, D.L., Marshall-Colon, A., Messina, C.D., Ort, D.R., Schnable, J.C., Vallejos, C.E., Wu, A., Yin, X., Zhou, W. (2020). Towards a multiscale crop modelling framework for climate change adaptation assessment. Nature Plants, 6, 338-348.
4. Peng, B.*, Guan, K.*, Pan, M., Li, Y. (2018). Benefits of seasonal climate prediction and satellite data for forecasting US maize yield. Geophysical Research Letters, 45, 9662-9671.
5. Peng, B.*, Guan, K.*, Chen, M., Lawrence, D.M., Pokhrel, Y., Suyker, A., Arkebauer, T., Lu, Y. (2018). Improving maize growth processes in the community land model: Implementation and evaluation. Agricultural and forest meteorology, 250–251, 64-89.

Research Interests and Expertise

My primary research interests are computational and process-based modeling (hydrological, crop, ecosystem and earth system modeling), remote sensing of soil moisture and crop productivity, and model-data integration. The overarching goal of my research is to deepen our understanding of terrestrial carbon-water-nutrient cycles and human impacts on ecosystems under climate change and land use intensification. The societal drivers for these research interests include climate change adaptation and mitigation, managing the risk of hydrometeorological and hydroclimatological disasters like droughts and floods, as well as ensuring water, energy and food securities and environmental sustainability. Over the past years, I have been focusing on the U.S. Midwestern agroecosystems.

Synergistic Activities

1. **Government Grant Proposal Review Panels**: NASA, FFAR
2. **Editorial Service and Journal Reviewers**: Associate Editor of the Agronomy Journal (IF=2.240) of American Society of Agronomy (ASA), 2020-present; Topic Editor and Review Board Member of Remote Sensing (IF=4.848), 2020-present; Guest Associate Editor of the Frontiers in Big Data, 2020-present; Young Editorial Board member of Journal of Remote Sensing, 2022-present Serve as reviewers for leading domain journals (6-10 manuscripts/year).
3. **Service to Disciplinary and Professional Societies or Associations**: NASA Science Team of Carbon Monitoring System Science Team, 2017-present.
4. **Conference session organizer**: Organized 7 sessions for the AGU fall meeting from 2019-2021.

5. **Mentors for graduate and undergraduate students**: 3 undergraduate interns/year, serves in multiple PHD committees.

Collaborators

Collaborator name	Current institution
Ainsworth, Elizabeth	University of Illinois at Urbana-Champaign
Asseng, Senthold	University of Florida
Arkebauer, Timothy	University of Nebraska
Bernacchi, Carl	University of Illinois at Urbana-Champaign
Chen, Min	DOE PNNL
Cooper, Mark	The University of Queensland
DeLucia, Evan	University of Illinois at Urbana-Champaign
Elliott, Joshua	DAPAR
Ewert, Frank	University of Bonn
Frankenberg, Christian	Caltech
Gentine, Pierre	Columbia University
Grant, Robert F.	University of Alberta
Guan, Kaiyu	University of Illinois at Urbana-Champaign
Gustafson, David I	CTIC
Hammer, Greame	The University of Queensland
Schnable, James C.	University of Nebraska-Lincoln
Jiang, Chongya	University of Illinois at Urbana-Champaign
Jin, Zhenong	University of Minnesota
Jones, James W.	University of Florida
Kimball, John	University of Montana
Kimm, Hyungsuk	University of Illinois at Urbana-Champaign
Köhler, Philipp	Caltech
Lawrence, David	NCAR
Li, Bo	University of Illinois at Urbana-Champaign
Li, Yan	Beijing Normal University
Lombardozzi, Danica L.	NCAR
Luo, Yunan	University of Illinois at Urbana-Champaign
Marshall-Colon, Amy	University of Illinois at Urbana-Champaign
Messina, Carlos D.	Corteva
Miao, Guofang	University of Illinois at Urbana-Champaign
Moore, Caitlin	University of Illinois at Urbana-Champaign
Ort, Donald R.	University of Illinois at Urbana-Champaign
Pan, Ming	Princeton University
Peng, Jian	University of Illinois at Urbana-Champaign
Pokhrel, Yadu	Michigan State University
Sun, Ying	Cornell University
Tang, Jinyun	LBNL
Vallejos, C. Eduardo	University of Florida
Wang, Shaowen	University of Illinois at Urbana-Champaign
Wardlow, Brian	University of Nebraska
Wu, Alex	The University of Queensland
Wu, Jin	University of Hong Kong
Yang, Xi	University of Virginia
Yin, Xinyou	Wageningen University & Research
Zhao, Lei	University of Illinois at Urbana-Champaign
Wang Zhou	University of Illinois at Urbana-Champaign

Section 6: Software Applications and Packages

Question #1

Please list any software packages used by the project, and indicate if they are on open source or export controlled.

Application Packages

Package Name

detailed information has been provided in the project narratives. We use a variety of software packages but all are open sourced.

Indicate whether Open Source or Export Controlled.

Open Source

Section 7: Wrap-Up Questions

Question #1

National Security Decision Directive (NSDD) 189 defines Fundamental Research as "basic and applied research in science and engineering, the results of which ordinarily are published and shared broadly within the scientific community, as distinguished from proprietary research and from industrial development, design, production, and product utilization, the results of which ordinarily are restricted for proprietary or national security reasons." Publicly Available Information is defined as information obtainable free of charge (other than minor shipping or copying fees) and without restriction, which is available via the internet, journal publications, textbooks, articles, newspapers, magazines, etc.

The INCITE program distinguishes between the generation of proprietary information (deemed a proprietary project) and the use of proprietary information as input. In the latter, the project may be considered as Fundamental Research or nonproprietary under the terms of the nonproprietary user agreement. Proprietary information, including computer codes and data, brought into the LCF for use by the project - but not for generation of new intellectual property, etc., using the facility resources - may be protected under a nonproprietary user agreement.

Proprietary Information

Are the proposed project and its intended outcome considered Fundamental Research or Publicly Available Information?

Yes

Will the proposed project use proprietary information, intellectual property, or licensing?

No

Will the proposed project generate proprietary information, intellectual property, or licensing as the result of the work being proposed?

If the response is Yes, please contact the INCITE manager, INCITE@doeleadershipcomputing.org, prior to submittal to discuss the INCITE policy on proprietary work.

No

Question #2

The following questions are provided to determine whether research associated with an INCITE proposal may be export controlled. Responding to these questions can facilitate - but not substitute for - any export control review required for this proposal.

PIs are responsible for knowing whether their project uses or generates sensitive or restricted information. Department of Energy systems contain only data related to scientific research and do not contain personally identifiable information. Therefore, you should answer "Yes" if your project uses or generates data that fall under the Privacy Act of 1974 U.S.C. 552a. Use of high-performance computing resources to store, manipulate, or remotely access any national security information is prohibited. This includes, but is not limited to, classified information, unclassified controlled nuclear information (UCNI); naval nuclear propulsion information (NNPI); and the design or development of nuclear, biological, or chemical weapons or of any weapons of mass destruction. For more information contact the Office of Domestic and International Energy Policy, Department of Energy, Washington DC 20585, 202-586-9211.

Export Control

Does this project use or generate sensitive or restricted information?

No

Does the proposed project involve any of the following areas?

i. Military, space craft, satellites, missiles, and associated hardware, software or technical data

ii. Nuclear reactors and components, nuclear material enrichment equipment, components (Trigger List) and associated hardware, software or technical data

iii. Encryption above 128 bit software (source and object code)

iv. Weapons of mass destruction or their precursors (nuclear, chemical and biological)

No

Does the proposed project involve International Traffic in Arms Regulations (ITAR)?

No

Question #3

The following questions deal with health data. PIs are responsible for knowing if their project uses any health data and if that data is protected. Note that certain health data may fall both within these questions as well as be considered sensitive as per question #2. Questions regarding these answers to these questions should be directed to the centers or program manager prior to submission.

Health Data

Will this project use health data?

No

Will this project use human health data?

No

Will this project use Protected Health Information (PHI)?

No

Question #4

The PI and designated Project Manager agree to the following:

Monitor Agreement

I certify that the information provided herein contains no proprietary or export control material and is correct to the best of my knowledge.

Yes

I agree to provide periodic updates of research accomplishments and to

acknowledge INCITE and the LCF in publications resulting from an INCITE award.

Yes

I agree to monitor the usage associated with an INCITE award to ensure that usage is only for the project being described herein and that all U. S. Export Controls are complied with.

Yes

I understand that the INCITE program reserves the right to periodically redistribute allocations from underutilized projects.

Yes

Section 8: Outreach and Suggested Reviewers

Question #1

By what sources (colleagues, web sites, email notices, other) have you heard about the INCITE program? This information will help refine our outreach efforts.

Outreach

Question #2

Suggested Reviewers

Suggest names of individuals who would be particularly suited to assess the proposed research.

The following scientists could be potential reviewers for our proposal due to their expertise in HPC & domain science, no prior collaborations but some familiarity of our work.

(1) Charlie Catlett, senior computer scientist, Argonne National Lab, email: cec@anl.gov, website: https://www.mcs.anl.gov/~catlett/

(2) Eliu Huerta, Lead for Translational AI, Argonne National Lab, email: elihu@anl.gov, webisite: https://www.anl.gov/profile/eliu-a-huerta

Section 9: Testbed Resources

Question #1

The ALCF and OLCF have test bed resources for new technologies, details below. If you would like access to these resources to support the work in this proposal, please provide the information below. (1 Page Limit)

The OLCF Quantum Computing User Program is designed to enable research by providing a broad spectrum of user access to the best available quantum computing systems, evaluate technology by monitoring the breadth and performance of early quantum computing applications, and Engage the quantum computing community and support the growth of the quantum information science ecosystems. More information can be found here: https://www.olcf.ornl.gov/olcf-resources/compute-systems/quantum-computing-user-program/quantum-computing-user-support-documentation.

The ALCF AI Testbed provides access to next-generation of AI-accelerator machines to enable evaluation of both hardware and workflows. Current hardware available includes Cerebras C-2, Graphcore MK1, Groq, Habana Gaudi, and SambaNova Dataflow. New hardware is regularly acquired as it becomes available. Up to date information can be found here: https://www.alcf.anl.gov/alcf-ai-testbed.

Describe the experiments you would be interested in performing, resources required, and their relationship to the current proposal. Please note, these are smaller experimental resources and a large amount of resources are not available. Instead, these resources are to explore the possibilities for these technologies might innovate future work. This request does not contribute to the 15-page proposal limit.

testbed.pdf
The attachment is on the following page.

We do not plan to use the test bed resources in this proposal.