agorocarbonalliance.com

Validation Report DayCentEVI

Prepared by:	Agoro Carbon Alliance
Reviewed by:	Dr. ir. Gerard H. Ros, Wageningen University / Nutrient Management Institute and supported by dr. Tessa van der Voort, Dr. ir. Tessa van der Voort
First submission:	June 27, 2023
Revised to address reviewer comments:	September 20, 2023
Model requirements version:	VMD0053 - Model Calibration, Validation and Uncertainty Guidance for the Methodology for Improved Agricultural Land Management

Agoro Carbon Alliance US, Inc.

100 N Tampa St - Tampa, FL 33602

Tel: 888-68-CARBON

CONTENTS

Tables

Figures

1. Overview

1.1 Report type

Project-specific, for Agoro Carbon USA Cropland Program, a grouped VCS ALM project implemented by Agoro Carbon Alliance. The Program's geographic area is delineated by political borders of the USA (excluding Alaska and Hawaii) and includes only cropland.

1.2 VCS version

Version 4.0, Sectoral Scope 14, accessed on 2 June 2023.

1.3 VCS model requirements version

Model Calibration, Validation, and Uncertainty Guidance for the Methodology for Improved Agricultural Land Management (ALM), Version 2.0, Sectoral Scope 14, accessed on 2 June 2023 (referred to hereafter as the "VMD53 Model Requirements").

1.4 Biogeochemical model version and confirmation materials

DayCent revision 279 (rev279) (Sept. 2022), a DDCentEVI version[1] (EVI signifies a version of DayCent with the option to use Enhanced Vegetation Index, known as EVI data).

This model version consists of the following components and is version-controlled:

1. The DDCentEVI model executable (DDCentEVI.exe), corresponding to the revision 279 of the DayCent source code repository previously downloaded from COMET-Farm[2].

2. The DayCent model parameters. These were originally derived from the distributed parameterizations, some of the model default parameter values are not adjusted during calibration and some have been calibrated and modified during the calibration process reported here.

3. Python scripts that were used to perform cross-validation for achieving the calibrated parameter sets.

The following materials can be provided for use by the reviewer of this report and project verifiers for the project:

---

[1] Copyright 2022 Colorado State University, Fort Collins, CO 80523 USA

[2] https://comet-farm.com/HelpPage, accessed on 12/19/2022.

- Copies of the validation dataset model input files used in the simulations for this report (original inputs before model calibration), Python and R code for analyzing results (model_validation_supporting_materials.7z)
- Appendix A. Documentation of Model Parameters
- Appendix B. Declaration of Practice Categories Requiring Evaluation
- Appendix C. Sampler diagnostics, posteriors of parameter distribution, and measurement uncertainty

1.5 Summary overview of changes according to reviewers' comments

Thanks to the meticulous reviews and insightful feedback provided by the Independent Modeling Experts (IMEs), we have implemented the following modifications in this version.

1. Elaboration regarding the handling of multiple observations within the same treatment over time for a specific experiment has been explicitly included. VMD53 (Section 5.1, page 10) states that "*calibration and validation data must be demonstrably independent. This requirement is met where datasets used for calibration and validation do not overlap in experimental research locations and are not taken from the same experimental study.*" For k-fold cross-validation, to account for the likelihood of high spatial and/or temporal correlation in repeated measurements from the same site and to ensure independence between calibration and validation datasets, it is imperative to maintain multiple observations over time for a specific experiment within the same fold, exclusively assigned to either calibration or validation, but not both. However, when datasets pertain to the same research location but are temporally separated due to different establishment times and distinct experimental designs or intentions, they are treated as distinct "sites" under different practice category and crop functional group combinations (PC/CFG). These sites can be randomly distributed across folds as allowed in VMD53 (Section 5.1, page 10) because data from these experiments may be correlated in space but are likely uncorrelated in management. We conducted a semi-variogram analysis to check correlations not addressed by this splitting of experimental data and found that SOC measurements from these experiments have no evident spatial correlation (Appendix C, Figure 3).

In this report, during k-fold cross-validation involving 30 study sites, we allocated 6 sites to the validation subgroup (constituting one-fold datasets) and assigned the remaining 24 sites to the calibration subgroup (comprising the remaining 4 folds of datasets) within each of the k=5 iterations. This approach ensures the segregation of calibration and validation datasets and avoids overlap between study sites in the two subsets. Such an arrangement effectively prevents any potential data leakage. Furthermore, by grouping all observations associated with a particular site within the same subgroup as the site itself, the chronological order of the data is conserved. This strategy becomes vital in addressing temporal autocorrelation. The biogeochemical model, tailored to simulate processes across extended time frames encompassing thousands of years, facilitates the capture of enduring trends and ecological interactions. Through the retention of temporal dependencies, the model was assessed to capture patterns within each site's SOC observations.

While the datasets with repeated measures from experiment sites have the likelihood of high correlations in both space and time in the same subgroup used for Bayesian calibration, some of this correlation could be explained by the model's simulation due to the interplay of variable inputs and model processes. However, it's possible that the model may not account for all aspects of this correlation for the type of datasets used in model calibration. Any remaining correlation structure not explained by the model's output was incorporated into the variance-covariance components σ^2_{site}, $\sigma^2_{time\_site}$, σ^2_{resid} for uncertainty estimation.

2. Added descriptions involving the use of repeated SOC measurements (soil sampling / replicates) per treatment and site to calculate measurement uncertainty. In this approach, the standard error (SE) is derived from the replicates reported for the measurements in the cited experimental studies. Not all studies, however, provide information about the SE or standard deviation (SD) of their measurements. An illustrative example data source is provided in Box 1 (Section 8), detailing measured SOC stocks and corresponding SEs for different treatments. Additional examples demonstrating the calculation of pooled measurement uncertainty (PMU), defined as the combined standard error of all measured values for a given practice change (VMD53, Section 5.2.4, page 23), are also included in this main report and Appendix C. Tables C1-C3 contain the complete set of reported SEs, paired treatments, and treatment descriptions for the validation datasets with reported measurement uncertainty.

3. To facilitate better comprehension, we have converted the units in some of the figures and tables to metric tons per hectare. While VMD53's protocol does not require it, we have chosen to incorporate the average yearly rate of change comparison. This comparison encompasses both short-term and long-term changes corresponding to the duration of the experiments. For instance, the histogram plots for the validation datasets, corresponding to each PC/CFG combination, now include mean, standard deviation (SD), and median values for both the measured rates of change in SOC stocks (t C ha^{-1} yr^{-1}) and the SOC stock changes up to the measurement year (g C m^{-2}, the unit used in DayCent) in response to the implemented practice changes. The histograms of model residuals for changes in SOC stocks are also presented in

both g C m^{-2} and t C ha^{-1} yr^{-1}. The mean squared error (MSE) and root mean squared error (RMSE) are presented in units of (t C $ha^{-1})^2$ and t C ha^{-1}, respectively. 4. We have excluded the Lethbridge-manure experiment from the ALL PCs and ALL CFGs comparison since this Cropland Project does not adopt the application of organic amendment practice (not in the 12 PC/CFG/SOC combinations requiring validation). It functions as a pre-existing practice in the baseline and carries over into the project phase. These data points (other experimental datasets comprise records of organic amendments application too) were included in the model calibration runs, so that the posterior parameter distribution to be influenced by these observations. However, they were not used for model validation in this report.

5. We have included treatment descriptions and examples where they enhance comprehension, a comprehensive detailed description of the 212 treatments across the 30 field experiments is provided in the Excel file titled "StudySite_Treatments.xlsx," available within the "model_validation_supporting_materials.7z" archive. Additionally, Tables C1-C3 provide the complete set of treatment descriptions for the validation datasets that include reported measurement uncertainty (SE).

6. In Sections 4, 5, 7, and 9, we have included additional descriptions, summaries, and discussions to provide insight into various aspects: elaboration on the duration and sampling events used for calibration and validation, details about PCs, the methodology employed to calculate treatment effects on SOC stock changes and average annual rate of change, highlighting constraints in dataset and needs for improving DayCent model processes.

2. Introduction

This report presents the validation of DayCent in modeling carbon (C) stock changes resulting from practice changes for C crediting for Agoro VCS ALM Cropland project. DayCent is a process-based ecosystem biogeochemical model which is publicly available. The model simulates C and nitrogen (N) dynamics, phosphorus, sulfur, and soil trace gas (N_2O, CH_4) fluxes in various managed ecosystems, including cropland, grassland, savanna, and forest (Parton et al., 2001; Del Grosso et al., 2006; Del Grosso et al., 2012; Zhang et al., 2018). The type and timing of management events, such as tillage, fertilization, irrigation, harvest, and grazing activities as management schedules in DayCent, can affect plant production and the retention of soil organic matter (SOM). The model operates on a daily time step and considers the exchange of C and N between the atmosphere, vegetation, and soil, which are influenced by climate, land use, land management, and other environmental factors. In the DDcentEVI version, plant production is updated daily, whereas in earlier versions of DayCent, it was updated weekly. Plant production is limited by water, light, and nutrient availability. The user manual and detailed model description for the DayCent model can be found in Hartman et al. (2021), Parts 1 and 2, respectively.

DDcentEVI has been used extensively to estimate net CO_2, N_2O, and CH_4 emissions from soils in the US national greenhouse house gas inventory submitted by the US Environmental Protection Agency (EPA) to the UN Framework Convention on Climate Change. This involves an annual set of simulations for total agricultural land use emissions for the EPA annual report (i.e. USEPA, 2013 and 2020), which requires model parameterization for standard crops, in addition to millions of model runs to estimate model uncertainty (Ogle et al., 2007 and 2010; Del Grosso et al., 2010).

The DayCent version evaluated in this report is based on the latest DDcentEVI version from the COMET-Farm platform, which implements USDA's entity-scale greenhouse gas inventory methods (Powers et al., 2014). This version was developed to simulate SOM dynamics to 30 cm soil depth and has successfully simulated changes in SOC resulting from changes in agricultural management, such as different N fertilization rates, different tillage intensities or no-till, and crop rotations including cover cropping (Gurung et al., 2020). These simulated management practices have also been implemented in this project (see Section 5 and Appendix B). For this project, a new option has been incorporated into the DDcentEVI version to initialize soil C pools by adopting an exponential curve and two parameters from the Environmental Policy Impact Climate (EPIC) model (Williams, 1995). Using directly measured SOC and tuning model parameters for the initialization of soil C pools has been used in EPIC (e.g., Wang et al., 2005), APEX (e.g., Wang et al., 2006 and 2012), and SWAT-C (Zhang et al., 2013). This approach has been increasingly used in process-based ecosystem models to initialize the conceptual SOC pools because it overcomes the difficulties present in the classical method of model spin-up initialization, especially for those sites where the previous land use history is unknown (Yeluripati et al., 2009; Dangal et al., 2022). This adoption for initializing SOC in DayCent modeling has been rigorously tested and shown to produce realistic results (Wang et al., 2023 under publication). In this report, we employ a Bayesian approach for model calibration. Once the model is appropriately calibrated and validated, the resulting probability distributions can be applied to comparable modeling domains that share similar soil types, climates, and management practices.

3. Responsible Parties

The DayCent model was calibrated and validated for this report by Agoro Carbon Alliance and Texas A&M Blackland Research Center. The following individuals contributed to this Validation Report:

- S. Wang added an alternative option to use direct soil measurements for the initialization of soil C pools. She also developed the DayCent-CUTE (auto-Calibration, sensitivity, and Uncertainty analysis ToolSet) tool for conducting Global Sensitivity Analysis (GSA) and played a key role in leading the model validation process. Additionally, she provided guidance on calibration and validation for this report, and performed analysis of the validation datasets, cross-validation, and model outputs. S. Wang also oversaw the alignment of datasets with the VMD53 Model Requirements, interpreted the criteria of the VMD53 Model Requirements, and led the writing of this report.
- N. Silvero, S. Wang, J. Shanahan, O. Olayemi, B. Munks, and S. Liu conducted literature searches.
- N. Silvero, and S. Wang assembled the validation dataset and created the input files required to run the model for each specific site.
- S. Park (from Texas A&M Blackland Research Center) and S. Wang implemented the Bayesian calibration approach, a script-based workflow for DayCent, utilized the **D**iffe**R**ential **E**volution **A**daptive **M**etropolis (DREAM) algorithm (Vrugt et al. (2009), and implemented k-fold cross-validation.
- The Agoro Science team, including J. Shanahan, Y. Zhao, and S. Liu, and the Carbon team, provided reviews of the report and all Validation Report materials.

4. Model Calibration and Validation

4.1 Description of the model calibration and validation approach

The calibration process involves the adjustment of the DayCent model parameters to enhance the model's ability to accurately simulate measured values. A Bayesian approach was utilized because Bayesian methods provide a coherent mathematical framework to effectively integrate diverse sources of information into model parameterization. Bayesian methods typically require implementing Markov Chain Monte Carlo (MCMC) methods for sampling probability distributions. The DREAM algorithm (Rug and Ter Braak, 2011; Vrugt, 2016), which is a type of MCMC algorithm, was utilized to estimate the joint posterior of DayCent parameters. For calibration and validation, we used k-fold cross-validation. DREAM has demonstrated efficiency in exploring high-dimensional parameter spaces (Bates and Campbell, 2001; Bikowski et al., 2012; Blasone et al., 2008; Hastings, 1970; Kow et al., 2012; Laloy et al., 2012; Minasny et al., 2011; Sadegh and Vrugt, 2014).

In benchmark experiments comparing various adaptive Markov chain Monte Carlo (MCMC) sampling techniques, DREAM has consistently outperformed commonly used optimization algorithms, offering superior solutions (Laloy et al., 2012; Vrugt et al., 2009, 2008). Consequently, the DREAM algorithm has gained widespread acceptance and is extensively utilized in diverse fields, including DayCent modeling (Gurung et al., 2020, Zhang et al., 2020), geostatistics (Minasny et al., 2011; Sun et al., 2013), hydrogeophysics (Hinnell et al., 2011), and hydrology (Vrugt et al., 2008a, 2009a; Shafii et al., 2014).

We have developed a script-based workflow for the calibration and validation of the DayCent model (Figure 1), which integrates the DREAM algorithm with k-fold cross validation. The core python script package, **Daycentpy**, has been developed to programmatically load, process, and manipulate DayCent model's parameterization. Several open-source software tools were used to implement the DayCent model's parameter estimation workflow. These tools include:

- The python package SpotPy (Houska et al., 2015) and PyMC (Salvatier et al., 2016) were used to perform computational optimization techniques for sensitivity analysis, calibration, and uncertainty analysis. SpotPy's implementation of DREAM includes several built-in features to help fine-tune and optimize the calibration process (e.g., set convergence criteria and control the number of chains used in the simulation to ensure accurate and efficient results).

- A web-based interactive development environment called Jupyter notebooks (Perez and Granger, 2007) was utilized along with the above-mentioned python packages to construct an interactive platform that allows for analyzing and documenting exploratory workflows. This technique can help improve transparency and reproducibility.

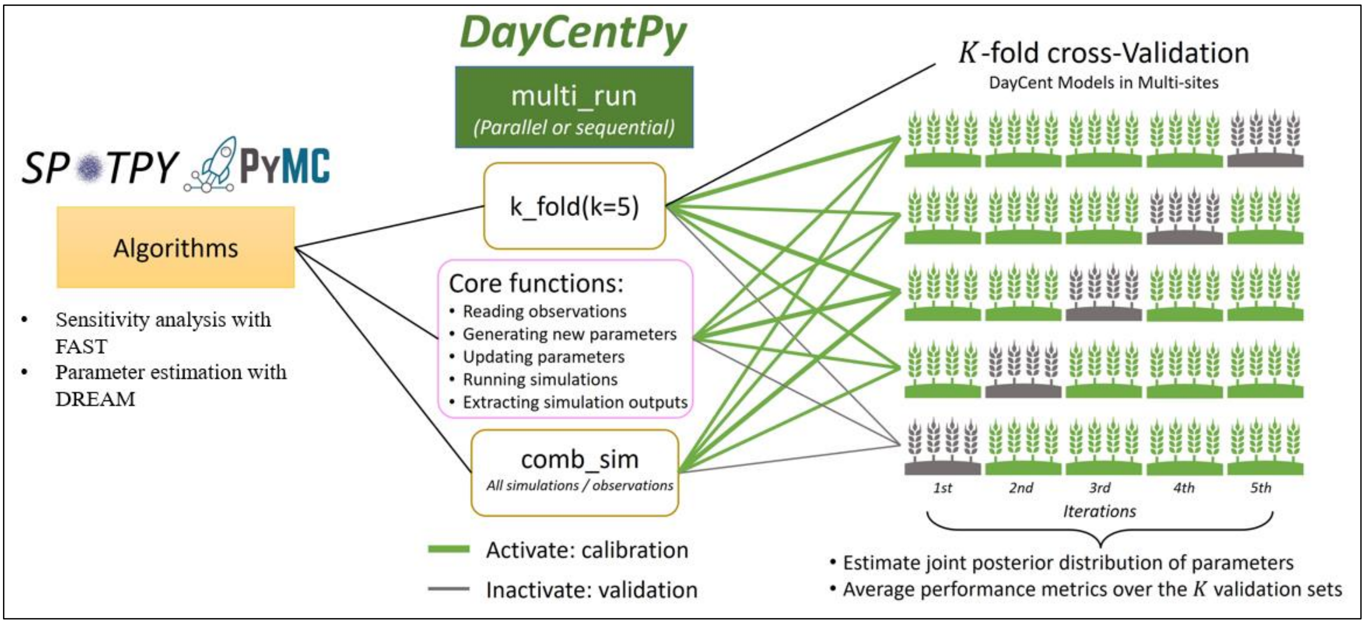

Figure 1. Flow chart of algorithms for DayCent calibration and validation. For each iteration of the k-fold cross-validation, the green-colored symbols of datasets are used for calibration, while the remaining (gray-colored symbols) are held out as the validation datasets for testing the model (as out-of-sample predictions).

In the SpotPy framework, DREAM parallel MCMC sampling method is used for model calibration. The algorithm employs differential evolution (DE) to generate new proposal samples, followed by an adaptive Metropolis-Hastings (AMH) acceptance-rejection step. In DREAM, multiple Markov chains are initialized, and DE is used to generate new candidate solutions by perturbing and recombining samples from different chains. The candidate solutions are then evaluated by the objective function, and the best candidate is accepted using the AMH algorithm. Over time, the algorithm adjusts its proposal distribution based on the chain history to improve sampling efficiency and convergence.

During the optimization process, we employed the gaussianLikelihoodMeasErrorOut() function for model parameterization. Given that the datasets included repeated measures from multiple experiment sites within the same subgroup for model calibration, some of the space and time correlation could be explained by the model simulation through the correlation in their variable inputs and model processes. Any remaining correlation structure not explained by the model output was integrated into the variance-covariance components (Cressie et al., 2009; Hoeting, 2009; Gurung et al., 2020) to estimate uncertainty. To address the spatiotemporal dependencies in the calibration dataset, the model incorporated two levels of nested random effects associated with site (σ^2_{site}) and year within site ($\sigma^2_{time\_site}$) (Pinheiro et al., 2010). We assumed that the model error followed a multivariate Gaussian distribution with a zero mean with these variance components and employed the restricted maximum likelihood (REML) estimator within the linear mixed-effect (LME) framework (Harville, 1977). This approach allowed us to estimate the most plausible values for σ^2_{site}, $\sigma^2_{time\_site}$, and residual error σ^2_{resid}. A summary of the posterior distributions of DayCent parameters and these associated hyperparameters were provided in Table A2 (Appendix A.) We used R-hat of Gelman and Rubin (1992) as a statistical diagnostic to assess the convergence of MCMC simulations. The calibration process used 24 MCMC chains, each of which was run until the R-hat dropped below 1.2.

The model was calibrated and validated simultaneously using a k-fold cross-validation procedure with k = 5 (Figure 1). In the k-fold cross-validation approach, the 30 experimental sites are randomly partitioned into k folds of equal size (by sites, namely, 6 sites for each fold). As illustrated in Figure 1, a single fold (gray-colored symbols) is held out as the validation dataset for testing the model (as out-of-sample predictions), while the remaining k-1 folds (green-colored symbols) are used for training. This process is iterated k times, with each fold serving as the validation set once. This k-folding approach used here satisfies the VMD53 Model Requirements (e.g., in a data-limited situation).

The following major steps were followed to calibrate and validate the DayCent model:

1) Initially, a test run was conducted for all sites to ensure successful execution of the model with model inputs and default parameter values. The result of the test run (as shown in Figure 2) helps visualize the observed SOC stocks in comparison to the simulated SOC stocks before model calibration.

2) After the initial run, the study sites were randomly divided into k = 5 folds using the create_k_fold_con() function in our Daycentpy tool. For the 30 experiments included in this validation report, each fold consisted of 6 experimental research sites for model validation, and the remaining 24 study sites were in a subgroup for model calibration. In VMD53, it is not permissible to employ the same experiment for both calibration and validation. Hence, for the report, it is imperative to maintain multiple observations over time for a specific experiment within the same fold, exclusively assigned to either calibration or validation, but not both. However, when datasets pertain to the same research location but are temporally separated due to different establishment times of experiments and distinct experimental designs or intentions, they are treated as distinct

"sites" under different practice category and crop functional group combinations (PC/CFG). More details on the justification for splitting the experimental data are provided in Section 4.4. A portion of the data points used in this report is shown in Table 1, where "0" indicates that the data is reserved for hold-out validation and "1" marks the data used for model training. The entire fold configuration of the data points was saved in a file named k_fold_con.csv, which is included in the model validation supporting materials. During the k-fold cross-validation iteration, the model was trained on 24 sites and validated on the remaining 6 sites (identified as "0" in Table 1), resulting in an approximately 80% - 20% split between the calibration and validation datasets, respectively. This process was repeated for each of the k folds, so that each fold, representing a distinct subset of the data, was used once as the validation set (as illustrated in Figure 1).

3) For each of the k iterations (indicated as #1, #2, #3, #4, #5 in Table 1), the “multi_run_dream()” function in Daycentpy was employed to perform k-fold cross-validation using the DREAM algorithm for MCMC sampling, as described earlier. In each iteration, the last 5000 posterior draws obtained from each of the 24 chains (208 to 209 per chain) were utilized for further analysis in the next step.
4) The fourth step involved using the out-of-sample SOC stock predictions from the validation dataset in each fold to calculate the differences in SOC stocks between paired experimental treatments. To evaluate the model's performance, several metrics were calculated. These metrics included model bias, MSE, RMSE, and the 90% prediction interval coverage of the validation data. These metrics were assessed separately in each fold, and then aggregated for calculating the average performance metric across all folds (k-fold iterations) to obtain an overall performance measure of the DayCent model for each combination of practice category (PC) and crop functional group (CFG). This provides a more robust estimation of the model's performance compared to evaluating it on a single validation set.

5) The five k-folds resulted in different sets of optimized parameters. We combined all 24 chains from the five k-folds to obtain a joint posterior parameter distribution that can be used to simulate new study sites. We drew 240 optimized parameter sets from this distribution to facilitate carbon crediting predictions. Figure A2 and Table A2 (refer to Appendix A) display the joint posterior parameter distributions obtained during cross-validation, as well as the selected posterior parameter distributions of the 240 parameter sets.

Table 1. Fold configuration used in the validation report [a]

Site Name	Treatment Name	Year with Observation	K = 5 model training and evaluation iterations [b]				
			#1	#2	#3	#4	#5
Broadbalk	Broadbalk_N3PK (Continuous wheat with 144 kg N ha^{-1} and 35 kg P ha^{-1})	1881	0	1	1	1	1
		1893	0	1	1	1	1
		1914	0	1	1	1	1
		1944	0	1	1	1	1
		1966	0	1	1	1	1
		1987	0	1	1	1	1
		1992	0	1	1	1	1
		1997	0	1	1	1	1
		2000	0	1	1	1	1
		2005	0	1	1	1	1
	Broadbalk_0N (Control: continuous wheat with 0 N fertilization)	1865	0	1	1	1	1
		1881	0	1	1	1	1
		1893	0	1	1	1	1
		1914	0	1	1	1	1
		1936	0	1	1	1	1
		1944	0	1	1	1	1
		1966	0	1	1	1	1
		1987	0	1	1	1	1
		1992	0	1	1	1	1
		1997	0	1	1	1	1
		2000	0	1	1	1	1
		2005	0	1	1	1	1
Brookings	HRR_CS_Co (CS: corn, soy rotation, Co: cover crop, Maximum residue remove)	2012	1	1	1	0	1
	HRR_CS	2012	1	1	1	0	1
	LRR_CS_Co (residue left)	2012	1	1	1	0	1
	LRR_CS	2012	1	1	1	0	1

[a]: only a portion of the data points are presented here as illustration. The full data contains 566 treatment observations and is saved in k_fold_con.csv (model_validation_supporting_materials.7z).

[b]: 1- data for training the model; 0- data in the hold-out fold (or out-of-sample) reserved for model validation. When conducting k-fold cross-validation for carbon modeling with multiple study sites and SOC observations over time, we assigned 6 sites to the validation subgroup (one-fold datasets) and the remaining 24 sites to the calibration subgroup (the remaining 4 folds of datasets) in each of the k=5 iterations. This ensures that the training and testing datasets do not overlap study sites, and observations from a specific site are exclusively used either for calibration or validation but not both, effectively preventing data leakage. Additionally, by including all observations associated with a site in the same subgroup as the site itself, we are preserving the temporal order of the data. This helps account for temporal auto-correlation, as observations from the same site that are close together in time will remain together in either the calibration or validation subgroup. By retaining these temporal dependencies, we enable the model to capture patterns within each site's SOC observations during both the training and evaluation phases. Any remaining correlation structure not explained by the model output was integrated into the variance-covariance components during calibration process.

4.2 Model setup

The datasets used for model calibration and validation were collected through literature reviews, and the DayCent model was then set up to run at all experimental sites for various treatments (refer to StudySite_Treatments.xlsx in model_validation_supporting_materials.7z for the description, e.g., as listed in Table 1, of each treatment). Site-level attributes were summarized in Tables 5-16 of **Section 7 "Description of Validation Dataset for Each PC/CFG/SOC Combination."** The input files for each study site were created using site-specific model input and management information, which is described in **Section 6 "Model Input Data."** This section also covers the procedures for dealing with missing data (refer to Section 6.2). Meteorological data, including daily minimum and maximum air temperature and precipitation, were extracted from the Parameter–Elevation Regressions on Independent Slopes Model (PRISM) database developed by the PRISM Climate Group[3] for the experimental studies conducted in the United States (US). Weather data for the US sites established before 1981 were extracted or gap-filled with DAYMET (Thornton et al., 1997) or local weather stations, which were assembled together with the experiments in the Agricultural Carbon Enhancement network (GRACEnet)[4] database. The nearest weather station was used for sites in the United Kingdom (Barré et al., 2010) and Canada (Environment and Climate Change Canada[5]). Temperature and precipitation data for sites in Brazil temperature were obtained from the SWAT global weather data or nearest weather station[6]. For Australian sites, temperature was obtained from SWAT and precipitation was obtained from the nearest weather station (Australian Government; Bureau of Meteorology[7].

The SOC sub-model in DayCent represents three conceptually defined soil C pools. Typically, the model requires a spin-up run lasting for thousands of years under grassland vegetation, historical weather data, and edaphic characteristics to reach an equilibrium state. This approach aims to mimic the site's history from native conditions to modern times. Following this, a simulation period of the base history is used to simulate the historical cropping or grazing practices prior to the time frame of interest to the user (Hartman et al., 2021). However, some

[3] https://prism.oregonstate.edu

[4] https://data.nal.usda.gov/dataset/gracenet-greenhouse-gas-reduction-through-agricultural-carbon-enhancement-network

[5] Historical Data - Climate - Environment and Climate Change Canada (weather.gc.ca)

[6] https://swat.tamu.edu/data/cfsr

[7] http://www.bom.gov.au/climate/data/

studies suggest that the initialization to equilibrium conditions may not provide high precision in the initial levels of SOC for process-based model simulations (Carvalhais et al., 2008; Wutzler and Reichstein, 2007). Therefore, an increasing number of applications of process-based ecosystem models have begun using measured SOC and tuning model parameters to initialize the conceptual SOC pools (Yeluripati et al., 2009; Dangal et al., 2022).

Considering the mentioned reality and the absence of reliable long-term meteorological and land-use history data preceding the modeling periods, we opted not to conduct long-term spin-up simulations for determining the initial fractions of soil C pools. Although terrestrial ecoregions, such as those with data from Olson et al. (2001) or Dinterstein et al. (2017), may be used in spin-up modeling for DAYCENT in the USA, we chose an alternative approach to employ an exponential curve and two parameters (FBM and FHP for fractions of OC in biomass pool and passive pool, respectively) from the EPIC model (Williams, 1995) and used directly measured SOC for initialization. Our initial DayCent simulation for the 30 experimental study sites, which were peer-reviewed and used for the validation report, shows that the model initialization setup is realistic. This is supported by the overall model performance of R^2 = 0.92 and RMSE = 5.25 t C ha^{-1} (Figure 2) for the dataset in their initial state prior to model calibration while using default model parameters. The model explains approximately 92% of the variance in the observed SOC. Considering the magnitude of the measured SOC stocks (ranging from 2040 to 15507 g C m^{-2}), the RMSE value of 5.25 t C ha^{-1} signifies a reasonable prediction error. The simulated mean of 49.22 t C ha^{-1} being close to the observed mean of 48.45 t C ha^{-1} suggests that the model realistically replicates the average values. Overall, these metrics collectively indicate a promising performance of the carbon model in predicting SOC stocks. Furthermore, using directly measured SOC stocks satisfies the VCS requirements, which stipulate that model simulations of SOC change for carbon credits must be initialized with the in-field directly measured SOC at t=0 or re-measured at every 5 years or less (VM0042 Section 8.2.1 Soil Organic Carbon Stocks, page 25).

Out of the thirty experimental sites used in our calibration and validation datasets, five sites lacked initial SOC measurements at the start of the experiment. In these cases, the model was executed from the year of the first SOC measurement (utilized to initialize soil C pools) until the end of the experiment with the last SOC measurement. The calibration and validation analyses were constrained to the period with SOC measurements at all experiment sites. The duration of the experimental studies varied, with two studies spanning 3 and 4 years, respectively, while the remaining experiments extended from 7 to 162 years. The treatments in each experimental study varied from 2 to 36, and the number of SOC measurements ranged from 3 to 67 instances during SOC measurement events (multiple years and treatments). As the effects of management practices on SOC may encompass both short- and long-term changes in soil biogeochemical processes, the repeated measurements of SOC stock changes were able to capture multi-year variations. Therefore, both calibration and validation comparisons between measured and simulated SOC changes were based on the extensive dataset derived from literature. It is important to note that these measurements and corresponding model predictions were at various time intervals. This confounds and complicates the model comparison when assessing the impact of management on SOC stocks.

The same model setup procedure will be used for the simulation of carbon crediting for the Agoro cropland project. This procedure includes the initialization of soil C pools using the initial SOC, soil physical and chemical properties, site latitude, longitude, and elevation (as described in **Section 6 “Model Input Data”**). Additionally, PRISM 4km gridded climate data will be utilized for the simulations of baselines and project scenarios.

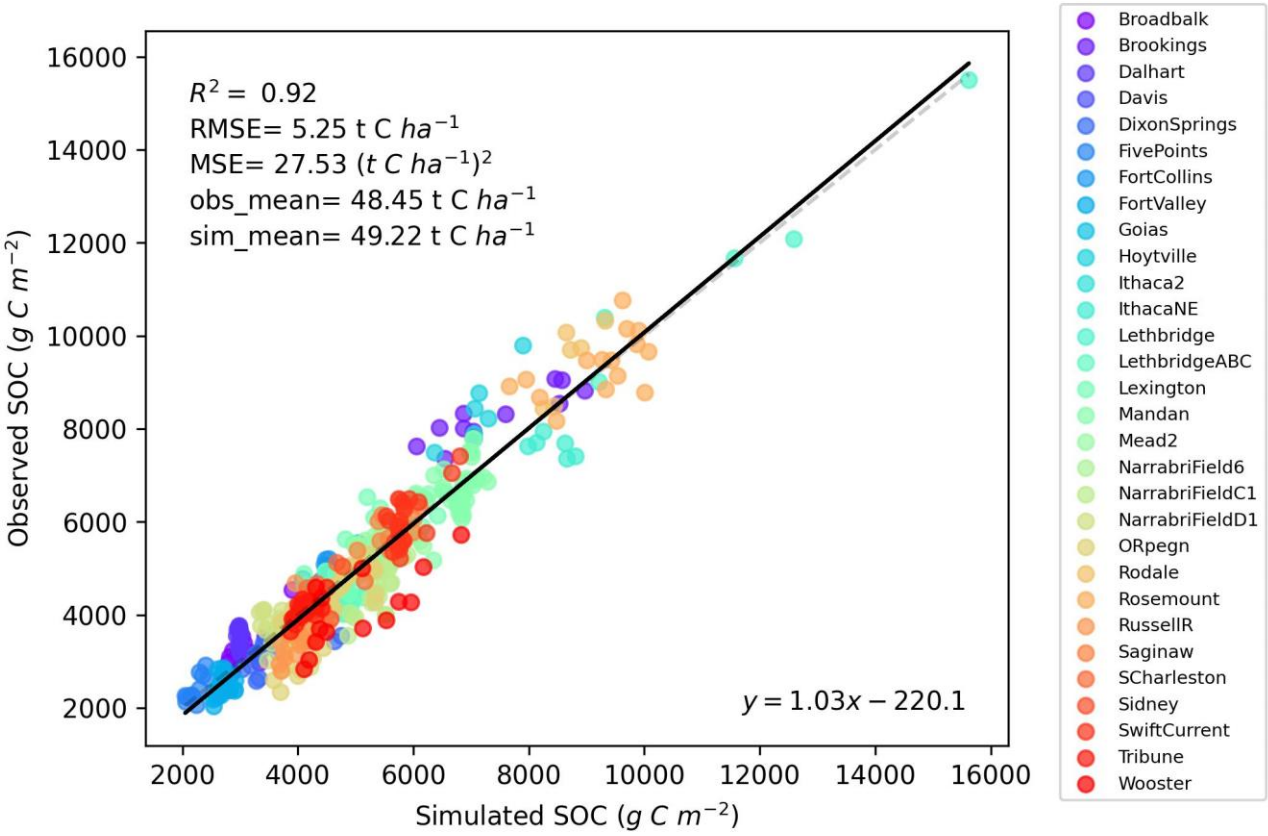

Figure 2. Initial simulation using DayCent default parameter values for the 30 peer-reviewed experimental study sites (experiment durations from 3 to 162 years) used in this validation report.

4.3 Documentation of internal model parameters

The DayCent model consists of hundreds of parameters, but typically only a small subset needs to be calibrated by users (Hartman et al., 2021). Many of the internal parameters have already been established and thoroughly evaluated through the model's evolution and applications. This was accomplished by fitting the model to long-term soil decomposition experiments that involved various types of plant materials added to soils with different textures. We conducted a variance-based Global Sensitivity Analysis (GSA) to identify the most influential model parameters for calibration purposes. We selected 23 parameters for this analysis and their descriptions and prior ranges are listed in Table A1. All 23 parameters were assumed to have independent uniform prior distributions with lower and upper bounds informed by the accumulated knowledge of the model development team (comprising scientists at Colorado State University). These bounds were informed by extensive testing of model algorithms and many previous studies (e.g., Necpálová et al., 2015; Gurung et al., 2020). Independence was assumed for simplicity. Constraining the prior range within its theoretical bounds is crucial to prevent any unexpected model behavior and ensures that the prior encompasses the entire range of the posterior distribution.

These parameters directly or indirectly affect SOC dynamics by influencing the decay rate of the soil C pools, C transfer efficiency, or crop productivity. We left the remaining parameters with the default values as distributed with the DayCent model files package. The full set of model parameters and model input files required to reproduce the validation results are included in the supporting materials as a zip file called "model_validation_supporting_materials.7z".

The extended Fourier amplitude sensitivity test (FAST) method (Saltelli et al., 1999) was used to estimate the sensitivity index of the DayCent model. The FAST method utilizes a periodic sampling approach and Fourier transformation to decompose a variance of a model output into variances contributed by model parameters. The total sensitivity index is calculated as follow (Saltelli et al., 2000)):

$$S_{Ti} = \frac{E_{X_{\sim i}}(V_{X_{\sim i}}(X_{\sim i}))}{V(Y)} = 1 - \frac{V_{X_{\sim i}}(E_{X_i}(X_{\sim i}))}{V(Y)}) \quad (1)$$

where S_{Ti} is the total sensitivity index, X_i is a factor being evaluated for its sensitivity, $X_{\sim i}$ is the sample size (N) times the number of factors (k-1) matrix of all factors, but X_i, Y is the generic

scalar model output equal to $Y=f(X_1, X_2, \ldots, X_k)$, $V_{X\sim i}$ is the variance of argument taken over X_i, $E_{X\sim i}$ is the mean of argument taken over X_i.
We used the total sensitivity indices for the 23 parameters (Figure A1) to identify the 12 most influential parameters for DREAM calibration (see column "Selected for calibration" in Table A1). During calibration, we treated the model parameters as population-level variables and conducted a single calibration for the entire project. Since we used the joint posterior from this single calibration in carbon crediting runs, the bias and uncertainty estimates presented in this validation report are generalizable to all crop types and management practices represented in the dataset.

4.4 Justification for splitting of experimental datasets

One of the key issues in model calibration is the acquisition of valid source information from the field experiments. Because only a limited number of published experiments have measured sufficient parameters over a long enough period to parameterize process-based ecosystem models (see **Section 6.3 "Data collection process"**), a way to calibrate with data scarcity is to take advantage of the peer-reviewed studies with high-quality measurements for both calibration and validation. According to VMD53 Model Requirements (**Section 5**), datasets for calibration and validation may be drawn from a single pool using k-folding, which ensures statistical independence (refer to VMD53 Section 5.1 Model Calibration).

Since the goal of calibration is to minimize prediction bias when data is limited, k-fold cross-validation is widely used for model evaluation. This method allows all data points to contribute to both calibration and validation processes. Cross-validation retains statistical independence of calibration and validation data by ensuring that each candidate model (associated with the set of fitted model parameters) is not evaluated against the same dataset that trained the model but is only used for out-of-sample predictions (Figure 1). Typically, 5 or 10 folds are used for k-fold cross validation, depending on the size of the dataset. In this project, we used a 5-fold cross-validation method, which was described in **Section 4.1 "Description of the model calibration and validation approach."** During each iteration of the k=5 folds cross-validation, only one-fold datasets (with the indicator "0" in Table 1) were used for validation, while the remaining 4 folds of datasets (with the indicator "1" in Table 1) were used for model calibration.

VMD53 (Section 5.1, page 10) states that "calibration and validation data must be demonstrably independent. This requirement is met where datasets used for calibration and validation do not overlap in experimental research locations and are not taken from the same experimental study." To account for the likelihood of high spatial and/or temporal correlation in repeated measurements from the same site and to ensure independence between calibration and validation datasets, the k-fold was configured such that: if an experiment site was assigned to a particular fold, all treatment observations associated with that experiment were also assigned to the same fold. This prevents data from the same site being used in both calibration and validation. For sites with multiple experiments but separate in space or time/duration, with separate experimental design or intention, such as differences at the level of PC/CFG combination, these experiments were considered as separate "sites" and were randomly allocated to folds, as allowed in VMD53 Section 5.1, because data are likely uncorrelated in management. We have one such case in the available dataset (see **Section 7**): IthacaNE and Ithaca2 at Ithaca, NE, where IthacaNE (2001-2010) was for tillage/residue management of corn, but Ithaca2 (1998-2011) was for 3-level inorganic N management.

Since these data would be permissible for inclusion in model calibration (VMD53 Section 5.1), they had been randomly divided into different folds. Another case in our collected dataset is for experiments conducted at Narrabri, New South Wales. NarrabriFieldC1 (1993-2012) was for cotton conventional till verse minimum tillage and cotton-wheat in minimum tillage management, while NarrabriField6 (1998-2008) and NarrabriFieldD1 (2002-2011) were for different cotton cropping systems of continuous cotton verse rotational management with wheat, bean, wheat-vetch, or fallow. We treated NarrabriFieldC1 as a separate site, but bound NarrabriField6 and NarrabriFieldD1 together, so that they were to be randomly allocated to the same fold. This was done to retain independent folds. To investigate potential correlations not accounted for by the described approach, we additionally constructed spatial variograms. However, our semi-variogram analysis revealed no apparent spatial correlation within our study sites (Appendix C, Figure 3C).

5. Practice Categories

This Section follows VMD53 Model Requirements Sections 5.2.1 to 5.2.3.

5.1 Practice categories (PCs)

The project's practices (see Table B1 in Appendix B) intended for crediting fall into three PCs:

1) Cropping practices, including single-crop rotation to double-crop rotation, cover crop etc. (CROPP)
2) Soil disturbance and/or residue management (DISTU)
3) Inorganic nitrogen fertilizer application (NFERT)

These PCs were identified from peer-reviewed and published experimental datasets with measurements of SOC stocks at different time intervals, which allow the comparisons (refer to Table 2, VMD53 Section 5.2.3, page 18) of: 1) continuous crop ("control"), different crop rotations, no cover crop ("control") and with cover crop (e.g., continuous corn, corn-wheat rotation, corn-wheat-sorghum-soybean rotation at the Tribune site in KS); 2) tillage effects such as conventional tillage ("control"), reduced till, and no till (e.g., corn-soybean rotation with conventional tillage and no till at the Wooster site in OH); and 3) different N application rates (e.g., continuous wheat without N application ("control") verse with 144 kg N ha^{-1} application at the Broadbalk site in England; continuous corn with 3 levels of N rates at 60, 120, and 180 kg N ha^{-1} at the Ithaca2 site in NE). Although VMD53 (Section 5.2.3, page 18) suggests that "ideally using control plots to test the PC", in some cases, the experiments do not have a traditional "control" scenario (e.g., only different treatment scenarios with varying levels of N application rates), and the purpose of the experiments is to investigate the effects of different treatments due to changes in management (e.g., effects of different N application rates of 60 vs. 120, 60 vs. 180, and 120 vs. 180 kg N ha^{-1} on SOC stock changes at the Ithaca2 site). The impact of N application rates on SOC is context-dependent and can vary based on the balance between increased plant productivity, organic matter input, microbial activity, and potential N saturation. Given that a comprehensive biogeochemical model is expected to capture the impact of different treatments on SOC stock changes, these comparative treatments were included in the model validation process. Treatment effects on SOC stock changes are calculated using the following equation:

$$\Delta SOC_{ij,j'}^{obs}(t) = SOC_{ij}^{obs}(t) - SOC_{ij'}^{obs}(t) \tag{2}$$

$$\Delta SOC_{ij,j'}^{sim}(t) = SOC_{ij}^{sim}(t) - SOC_{ij'}^{sim}(t) \tag{3}$$

where $\Delta SOC_{ij,j'}^{obs}(t)$ and $\Delta SOC_{ij,j'}^{sim}(t)$ are the measured and simulated SOC stock differences between treatment j and $j'(j \neq j')$ for the i^{th} site in year t, respectively. The observed and simulated average annual rate of SOC stock change is calculated as: $\Delta SOC_{ij,j'}^{obs}(t)/numOfyr$ and $\Delta SOC_{ij,j'}^{sim}(t)/numOfyr$, where numOfyr is the number of years from the SOC measurement to the starting year of the experiment

For this report, model bias (Section 8.1, Eq. 4) and model residuals (Section 9.1, e.g., Figure 21 histogram of model residuals) are based on SOC stock changes in response to practice changes, specifically, between different treatment scenarios. These treatments encompass PCs that are required evaluation of their impacts, but the treatment donated as j' is not necessarily (though it would be ideal) a "control" scenario due to limited data availability. In many ecological and agricultural systems, it is challenging to establish a true control treatment due to the dynamic and interconnected nature of ecosystem components even for experimental studies. Therefore, while having a control or baseline scenario is ideal, given the reality of data scarcity, our collected dataset allows the evaluation of a wider range of treatment scenarios (different practices) encompassing the PCs for testing their relative impacts on C sequestration.

In certain cropping systems, irrigation is a normal component of the management approach and is incorporated into both the project and baseline scenarios. Since it operates independently of practices aimed at emissions reduction, there is no necessity for the validation of water management/irrigation PC (VMD53, page 21) in this context. Similarly, the same principle applies to the application of organic amendments in this project, where it functions as a pre-existing practice in the baseline and carries over into the project phase. Our collected field experimental dataset comprises records of both irrigation and organic amendments application as management practices, as exemplified by the Lethbridge site. These data were included in the model calibration runs, enabling the posterior parameter distribution to be influenced by these observations. However, they are not used for model validation in this report.

5.2 Project domain

For each of the above practice categories, model evaluation begins with defining the project domain in terms of its biophysical attributes of unique crop functional groups, climate zones, and soil attributes.

5.2.1 Crop functional groups (CFGs)

In this project, individual crop types were grouped into the following four CFGs across crops sharing unique combinations of attributes (following VMD53 Section 5.2.2):

1) Non-N-fixing, annual, C4, herbaceous, non-flooded crops (NON_LEGUME_C4, e.g., corn, sorghum)

2) N-fixing, annual, C3, herbaceous, non-flooded crops (LEGUME_C3, e.g., soybeans, peas, alfalfa[8])
3) Non-N-fixing, annual, C3, herbaceous, non-flooded crops (NON_LEGUME_C3, e.g., wheat, rye, barley)
4) Non-N-fixing, annual, C3, shrubby, non-flooded crops (COTTON. This is only for cotton.)

5.2.2 Climate zones

For the validation report, the following climate zones as defined by IPCC (2019) were used (Table 2). All climate zones (CTD, WTD, WTM) occurring in the project area (see Table B3 in Appendix B) are represented by the validation dataset.

Table 2. Climate zones included in the validation datasets.

IPCC climate zone	Abbreviation
Warm Temperate Dry	WTD
Warm Temperate Moist	WTM
Cool Temperate Dry	CTD
Cool Temperate Moist	CTM
Tropical Moist	TrM

5.2.3 Soils

The validation dataset includes 8 soil texture classes (out of the 12 in the USDA NRCS soil texture classification) (Table 3). All 12 soil texture classes appear in the project area (Table B3 in Appendix B). The soil texture classes represented in the validation dataset meet minimum requirements of “at least three declared soil textural classes must be presented and the range in clay contents must span at least 15 percentage points” for each PC/CFG and emission source (ES) combination as outlined in VMD53 Model Requirements Section 5.3.2, Requirements 2 (see Section 7 “Description of Validation Dataset for Each PC/CFG/SOC Combination” for fulfilment of the requirements).

[8] It has the genetic potential to grow perennially but was included in this CFG when it was only grown for a single season and therefore it was qualifying as annual.

Table 3. USDA soil texture classes occurring in the project area and included in the validation dataset.

USDA NRCS soil texture class			**Represented in the validation dataset?**
Soil texture class	**Abbreviation**	**Midpoint clay %**	
Clay	Cl	70	Yes
Clay loam	ClLo	35	Yes
Loam	Lo	20	Yes
Loamy sand	LoSa	10	
Sand	Sa	5	
Sandy clay	SaCl	40	
Sandy clay loam	SaClLo	30	Yes
Sandy loam	SaLo	10	Yes
Silt	Si	5	
Silty clay	SiCl	45	Yes
Silty clay loam	SiClLo	35	Yes
Silt loam	SiLo	15	Yes

5.3 Domain covered by this validation

The project domain validated in this report includes 12 combinations of PC/CFG/SOC (Table 4). The dataset attributes for each PC/CFG combination validated for SOC meet minimum requirements per Requirement 2 in VMD53 Model Requirements Section 5.2.3, except for NFERT × COTTON. The validation data fails to meet the minimum requirements for the Inorganic N Fertilizer Application practice category for cotton due to data scarcity (only one site found, Table 5). However, the *Requirement* also recognizes the data scarcity issue and allows reasonable flexibility. To that end, we adopted the "Requirement 2: Special Rules for Practice Categories" in VMD53 Section 5.2.3 that the model may be considered validated for the COTTON CFG if (1) inorganic N fertilizer application has been successfully validated for another annual CFG (e.g., NFERT × NON_LEGUME_C3), and (2) this CFG has been successfully validated for the Cropping, Planting, and Harvesting PC (which is CROPP × COTTON). As listed in Table 5, considering that we have CROPP × COTTON, NFERT × NON_LEGUME_C3, plus NFERT PC across all annual CFGs (NFERT × ALL), which all cover the NFERT × COTTON climate zone and soil textures requirements, this validation report is usable for the NFERT × COTTON in the project.

Furthermore, all PC/CFG/SOC combinations pass the "stacked" practices requirement (VMD53 Model Requirements section 5.2.3, Requirement 1). This requirement means that when studies reporting the effects of changing multiple practices at once were used, the validation datasets used to validate a PC/CFG/SOC combination contain at least one experimental study that isolates the effect of the practice change being validated. For the counts of stacked and unstacked observation pairs, please refer to **Section 7 "Documentation of Validation Datasets for Each PC/CFG/SOC Combination."**

Table 4. Validated practice category and crop functional group combinations for SOC in this project.

Practice Category (PC)	Crop Functional Group (CFG)	PC × CFG × SOC
Cropping practices, planting, and harvesting (e.g., crop rotations, cover crops) (CROPP)	NON_LEGUME_C4 LEGUME_C3 NON_ LEGUME_C3 COTTON	CROPP × NON_LEGUME _C4 CROPP × LEGUME_C3 CROPP × NON_LEGUME e_C3 CROPP × COTTON
Soil disturbance and/or residue management (DISTU)	NON_LEGUME_C4 LEGUME_C3 NON_ LEGUME_C3 COTTON	DISTU × NON_LEGUME _C4 DISTU × LEGUME_C3 DISTU × NON_LEGUME _C3 DISTU × COTTON
Inorganic nitrogen fertilizer application (NFERT)	NON_LEGUME_C4 LEGUME_C3 NON_ LEGUME_C3 COTTON	NFERT × NON_LEGUME _C4 NFERT × LEGUME_C3 NFERT × NON_LEGUME _C3 NFERT × COTTON

Table 5. Summary of dataset attributes for each PC and CFG combination validated for SOC.

PC × CFG	Number of sites	Number of pairs of SOC observations	Climate zones	Soil textural classes	Clay % range
CROPP × NON_LEGUME _C4	12	210	CTD, CTM, TrM, WTD, WTM	Cl, Lo, SaLo, SiCl, SiClLo, SiLo	40
CROPP × LEGUME_C3	15	270	CTD, CTM, TrM, WTD, WTM	Cl, ClLo, Lo, SaLo, SiCl, SiClLo, SiLo	54
CROPP × NON_LEGUME_C3	16	292	CTD, CTM, WTD, WTM	Cl, ClLo, Lo, SaClLo, SaLo, SiCl, SiClLo, SiLo	54
CROPP × COTTON	6	163	TrM, WTD, WTM	Cl, ClLo, SaLo	54
DISTU × NON_LEGUME_C4	11	127	CTD, CTM, TrM, WTD, WTM	Cl, ClLo, SaLo, SiClLo, SiLo.	50
DISTU × LEGUME_C3	7	60	CTD, CTM, TrM, WTD, WTM	Cl, ClLo, SaLo, SiClLo, SiLo.	40
DISTU × NON_LEGUME_C3	6	99	CTD, WTD, WTM	Lo, ClLo, SaLo, SiClLo, SiLo	29
DISTU × COTTON	4	50	TrM, WTD, WTM	Cl, ClLo, SaLo	43
NFERT × NON_LEGUME_C4	9	139	CTD, WTD, WTM	Lo, SaLo, SiClLo, SiLo	22
NFERT × LEGUME_C3	4	73	WTD, WTM	Lo, SaLo, SiClLo, SiLo	21
NFERT × NON_LEGUME_C3	10	185	CTD, CTM, WTD, WTM	Lo, SaLo, SiClLo, SiLo	21
NFERT × COTTON [a] (NFERT × ALL)	1 (13)	36 (281)	WTM (CTD, CTM, WTD, WTM)	SaLo (Lo, SaLo, SiClLo, SiLo)	22

[a]: For the Inorganic N Fertilizer Application practice category for cotton (NFERT × COTTON), validation data fails to meet the minimum requirements (VMD53 Model Requirements Section 5.2.3, Requirement 2) due to data scarcity. However, the model may be considered validated for this CFG if (1) inorganic N fertilizer application has been successfully validated for another annual CFG (e.g., NFERT × NON_LEGUME_C3), and (2) this CFG has been successfully validated for the Cropping, Planting, and Harvesting PC (CROPP × COTTON).

6. Model Input Data

The DayCent model's inputs to be assembled can be divided into four categories: weather, soil, plant, and management information.

6.1 Site-specific model input and management information

- Daily weather data for the site, including long-term historical daily weather data for creating mean monthly weather statistics and data for the time-period to be simulated: daily maximum and minimum air temperature (°C), and daily precipitation (cm/day). Solar radiation (ly/day), relative humidity (%), and wind speed (m/h) are optional. When the optional weather inputs are not provided, the model can estimate them using internal calculations based on site latitude.
- Soil texture (sand%, silt%, clay%), bulk density, pH, wilting point, field capacity, root fraction, organic matter fraction, saturated hydraulic conductivity (Ksat) for each soil horizon from the surface to the first fully root-restrictive layer.
- Initial SOC stock (in g C m^{-2}) in the 0–30 cm soil layer.
- Site latitude and longitude (decimal degrees), elevation (m).
- Planting dates and methods, and crop names for all crops/rotations.
- Tillage dates, types, and intensities, such as implements used, depth, number of passes (if any)
- Harvest dates, methods, and types (e.g., grain, hay, etc.)
- Residue management (e.g., burning, straw/stover removal).
- Nitrogen fertilization dates and application rate (g N m^{-2})
- Irrigation starting and ending dates and amounts (cm).
- Organic amendments addition dates, amounts, types (e.g., manure, green manure, compost, straw amendments, N fraction, C:N ratio, the mass of the dry fraction)
- Herbicide application dates.

6.2 Procedures for missing data

Due to the lack of detail provided in some experiments, approximations were necessary for planting and harvesting dates in certain years of the study period. When published experimental studies report SOC measurement data to a depth other than 30 cm, SOC stocks were calculated using linear extrapolation to estimate the value to 30 cm based on the SOC measurements at the nearest depth increment. When not all required soil properties were available in the publications, soil information was obtained from the USDA Web Soil Survey for the reported soil series. The model has a supplemental program (mksitesoil.exe) which can be used to update site soil properties based on soil texture and bulk density, where the necessary soil information is missing from the publication.

6.3 Data collection process

All measured datasets used in this report were obtained from peer-reviewed and published long-term experimental studies that had measurements of SOC stock over time. Some of the datasets were contained in the Agricultural Collaborative Research Outcomes System (AgCROS)[9] and GRACEnet databases. To compile the necessary model input files for a specific experimental site, we combined information from multiple relevant publications, including journal articles and book chapters, as the long-term SOC measurements and the information required to parameterize DayCent for the study are not always reported in a single publication. All publications used in this compilation of model input files are listed in the citation columns of Tables 6-17 (Section 7) for the corresponding site.

Reviewed sites were excluded when they failed to meet one or more of the following criteria:

1) Studies must report sufficient information for modeling, or missing data can be inferred to simulate the site accurately (thus, the model inputs have low uncertainty.)
2) SOC was measured and re-measured and repeated measurements of SOC stock change must be able to capture multi-year changes, as practice effects on SOC may combine short and long-term changes in soil biogeochemical processes.
3) Reported SOC was at a depth of 30 cm, or to depths allowing a reasonable estimation to the 30 cm by interpolation across reported depths.

After reviewing and evaluating all long-term experiments, 30 sites (Figure 3) were selected for the calibration and validation process. The identified datasets comprise a total of 212 treatments and 566 measurements (excluding the initial/first-year measurements used for the initialization of soil C pools). Please refer to the StudySite_Treatments.xlsx file in model_validation_supporting_materials.7z for the count of treatments and measurements for each study site. These treatment measurements were combined into 873 pairs (i.e., corn-soybean no till verse corn-soybean conventional tillage) to get SOC stock changes in response to practice changes. Observation pairs for Irrigation and Organic Amendments Application were not required for validation in this report (see **Section 5 "Practice Categories Requiring Evaluation"**); however, these pairs were included in the calibration runs and for cross-validation of assessing ALL PCs and CFGs. This allows the calibrated parameter sets to be informed and evaluated by these observations.

---

[9] https://agcros-usdaars.opendata.arcgis.com/

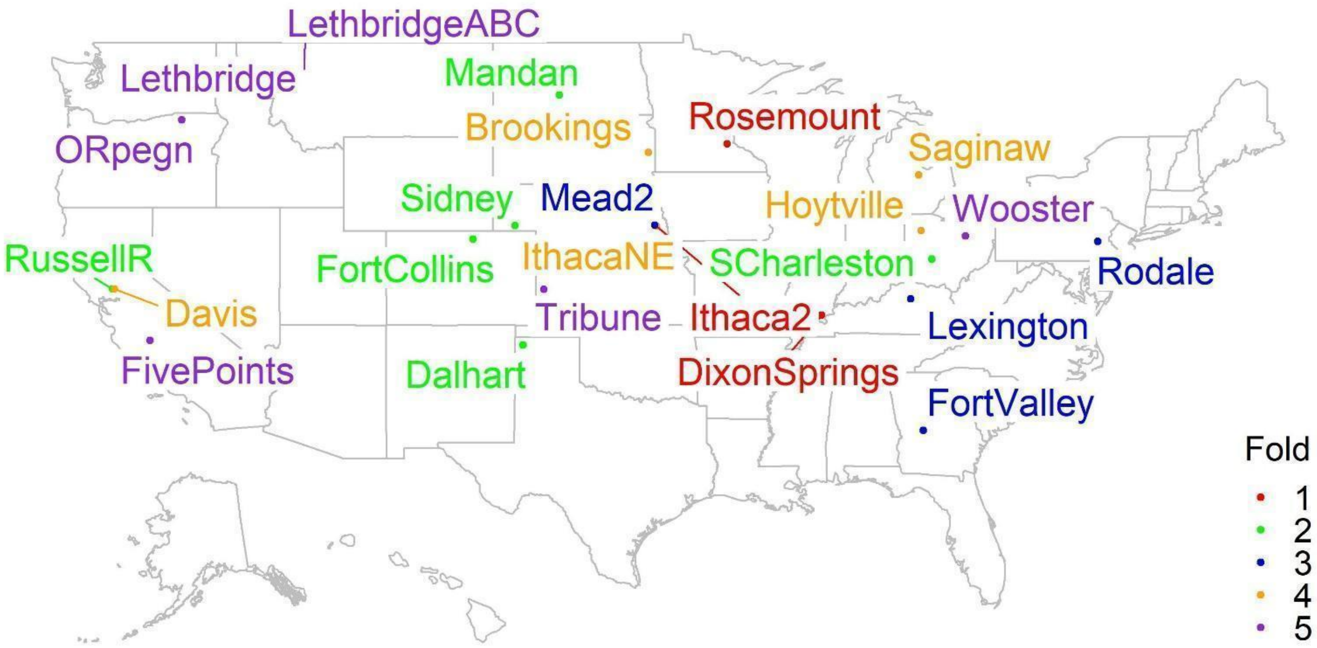

Figure 3. Locations of the experimental sites used for k-fold cross validation of the DayCent model. Sites not shown: Broadbalk at Rothamsted, England (fold 1); Goias at Goias, Brazil (fold 3); NarrabriField6 and NarrabrifieldD1 at Narrabri, New South Wales (Australia; fold 1); and NarrabriFieldC1 at Narrabri (fold 3).

If a study reported the effects of combined practices ("stacked" practice), then the study site was assigned to the composite of studies that contained at least one other validation study that isolated the same effect of practice change being validated. This was done to ensure that no practice category was validated solely against stacked practice studies. If the experimental studies reported the measurement uncertainty in their publications (e.g., standard deviation, or standard error for that treatment), then the uncertainty values were used to calculate pooled measurement uncertainty (PMU).

7. Documentation of Validation Datasets for Each PC/CFG/SOC combination

The following description and tables were organized to list the attributes of the peer-reviewed experimental studies comprising the validation datasets for each CFG/PC/SOC combination according to Section 5.2.3 Summary of Requirements in VMD53 Model Requirements. All SOC stocks are from the measurements at the 0-30 cm depth.

The SOC stock changes were quantified up to the measurement year in g C m^{-2} (the unit used in DayCent), and the corresponding rates of change in t C ha^{-1} yr^{-1} were graphically represented (with labelled mean, median, and SD) in the figures found in this section. However, when interpreting the yearly SOC change rates, it is essential to consider the following factors:

1) The experiments span various durations, ranging from 3 to 162 years. Consequently, the statistical metrics of mean, median, and SD of SOC stock change rates encompass both short-term and long-term changes.
2) Different experiment studies involve varying treatment levels, including different rates of inorganic N fertilization, and not all studies include a "control" scenario. The SOC stock changes observed from paired observations reflect the impacts of these diverse treatments (involving crops from the crop functional groups, specifically the PC/CFG combinations of interest) on SOC stocking rates.
3) Finally, it's important to recognize that carbon sequestration results from the entire cropping and operational management system, which can have enduring effects extending beyond the implementation year.

Detailed description of the 212 treatments across the 30 field experiments can be found in the Excel file named "StudySite_Treatments.xlsx," located in the model_validation_supporting_materials.7z archive. As elaborated in Section 5, these treatments encompass PCs that require evaluation of their impacts.

The measurements revealed more instances of SOC stock increases than declines resulting from management changes across different field experiment treatments. However, it also indicates that negative changes can be anticipated with specific environmental conditions, given that it is possible for changes in management to reduce SOC over time. Measurement errors also introduce another layer of complexity (refer to pooled measurement uncertainty in Section 8). Additionally, soil sampling locations for pairwise treatment plots can introduce variability over the measurement period, contributing to the observed noise in SOC change data.

7.1 CROPP x NON_LEGUME_C4

The validation datasets (Table 6), with their range of measured SOC changes from 210 paired observation differences shown in the histogram (Figure 4), are usable for this PC/CFG/SOC (CROPP x NON_LEGUME_C4 x SOC) combination within all declared climate zones and soil textures because:

- These experimental studies cover 5 IPCC climate zones (WTD, WTM, TrM, CTD, CTM; see full name for each Abbreviation in Table 2), and all 3 of the declared climate zones (WTD, WTM, CTD) in this project domain (see Appendix B) were presented.
- Six soil texture classes (Cl, Lo, SaLo, SiCl, SiClLo, SiLo; see full name for each Abbreviation in Table 3) are included. All of them are in the declared project area (≥ 3 declared soil textural classes must be represented as required in Requirement 2 in VMD53).
- Clay % contents range from 10% to 50%, which span 40 percentage points (≥ 15 percentage points as required in Requirement 2 in VMD53).
- At least one study isolates effects, e.g., only 11 out of the 210 observations compare stacks of PC changes.

Table 6. Dataset attributes of studies used in the validation of CROPP x NON_LEGUME_C4 x SOC.

Location	Study name	Citation	Year initiated (1st measure if not same yr)	Year measured (duration)	IPCC climate zone	Soil texture	Clay (%)	SOC measurement method	Number of pairs of observations
Brookings, SD	Brookings	Hammerbeck et al. (2012); Wegner et al. (2018)	2008	2012 (4 yrs)	CTD	SiClLo	35	Dry combustion	6
Dalhart, TX	Dalhart	Halvorson et al. (2009); Halvorson and Reule (2007)	1999	2004, 2005, 2006 (7 yrs)	WTD	SaLo	18	Dry combustion	3
Davis, CA	Davis	Clark et al. (1998)	1988	1992, 1996 (8 yrs)	WTD	Lo	17	Walkley-Black method	10 (8 stack PCs)
Fort Valley, GA	FortValley	Sainju et al. (2005)	1999	2002 (3 yrs)	WTM	SaLo	10	Dry combustion	54
Goias, Brazil	Goias	Ferreira et al. (2019)	2005	2014 (9 yrs)	TrM	Cl	50	Dry combustion	3 (1 stack PCs)
Hoytville, OH	Hoytville	Collins et al. (1999)	1963 (1980)	1993 (23 yrs)	CTM	SiClLo	40	Dry combustion	2
Mead, NE	Mead2	Varvel (2006)	1982 (1984)	1992, 1998, 2002 (18 yrs)	WTD	SiClLo	31	Dry combustion	54
Kutztown, PA	Rodale	Elliott et al. (1994); Pimentel et al. (2005)	1981	1994, 2002 (21 yrs)	WTM	SiLo	30	Not reported	2 (2 stack PCs)
Wooster, OH	Wooster	Collins et al. (1999) and Dick et al. (1997)	1962	1971, 1980, 1992 (30 yrs)	CTM	SiLo	15	Dry combustion, Walkley-BlackMethod	6
Saginaw, MI	Saginaw	Christenson (1997)	1972	1981, 1991 (19 yrs)	CTM	SiCl	47	Dry combustion	30
Tribune, KS	Tribune	Halvorson and Schlegel (2012)	2001	2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010 (9 yrs)	WTD	SiLo	26	Dry combustion	36
Winter, CA	RussellR	Kong et al. (2005); Wolf et al. (2018)	1993	2003, 2012 (19 yrs)	WTD	SiLo	18	Dry combustion	4
Total	**12**								**210**

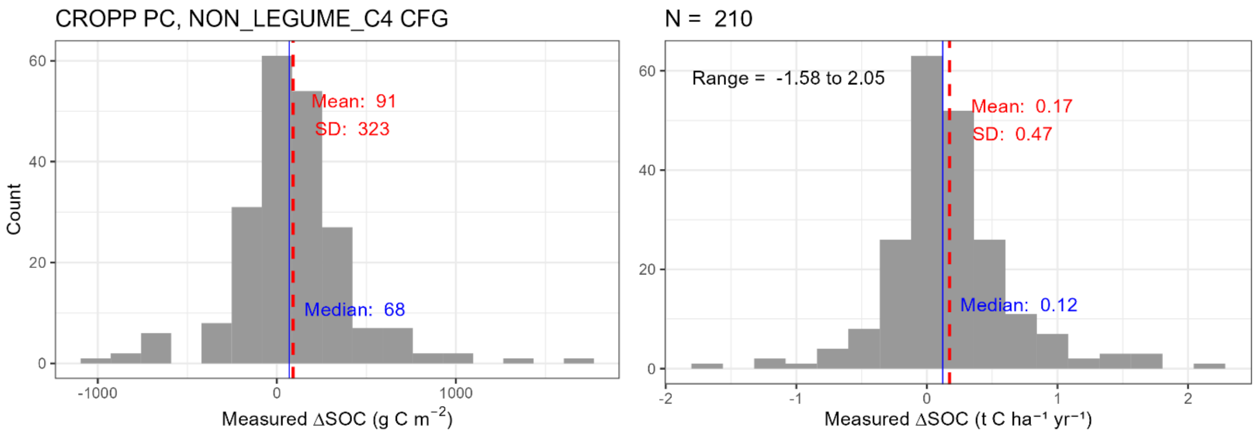

Figure 4. Histogram depicting observed changes in SOC in the studies (12 sites with experiment durations from 3 to 30 years) used for model validation in response to changed cropping practices involving crops from the NON_LEGUME_C4-type CFG.

7.2 CROPP x LEGUME_C3

The validation datasets (Table 7), with their range of measured SOC changes of 270 paired observation differences shown in the histogram (Figure 5), are usable for this PC/CFG/SOC (CROPP x Legume_C3 x SOC) combination within all declared climate zones and soil textures because:

- These experimental studies cover 5 IPCC climate zones (CTD, CTM, TrM, WTD, WTM), 3 of which (CTD, WTD, WTM) are in the declared project domain (Appendix B).
- Seven soil texture classes (Cl, ClLo, Lo, SaLo, SiCl, SiClLo, SiLo) are included. All of them are in the declared project area.
- Clay % contents range from 10% to 64%, which span 54 percentage points.
- At least one study isolates effects, e.g., only 11 out of the 270 observations compare stacks of PC changes.

Table 7. Dataset attributes of experimental studies used in validation of CROP x LEGUME_C3 x SOC.

Location	Study name	Citation	Year initiated (1st measure if not same yr)	Year measured (duration)	IPCC climate zone	Soil texture	Clay (%)	SOC measurement method	Number of observation pairs
Brookings, SD	Brookings	Hammerbeck et al. (2012); Wegner et al. (2018)	2008	2012 (4 yrs)	CTD	SiClLo	35	Dry combustion	6
Davis, CA	Davis	Clark et al. (1998)	1988	1992, 1996 (8 yrs)	WTD	Lo	17	Walkley-Black method	10 (8 stack PCs)
Five Points, CA	FivePoints	Mitchell et al. (2015); Mitchell et al. (2017); Veenstra et al. (2006)	1999	2000, 2004, 2007 (8 yrs)	WTD	ClLo	39	Dry combustion	6
Fort Valley, GA	FortValley	Sainju et al. (2005)	1999	2002 (3 yrs)	WTM	SaLo	10	Dry combustion	45
Goias, Brazil	Goias	Ferreira et al. (2019)	2005	2014 (9 yrs)	TrM	Cl	50	Dry combustion	4 (1 stack PCs)
Hoytville, OH	Hoytville	Collins et al. (1999)	1963 (1980)	1993 (23 yrs)	CTM	SiClLo	40	Dry combustion	2
Mead, NE	Mead2	Varvel (2006)	1982 (1984)	1992, 1998, 2002 (18 yrs)	WTD	SiClLo	31	Dry combustion	45
Narrabri, New South Wales	Narrabri Field6	Rochester (2011)	1995 (1998)	2000, 2002, 2004, 2006, 2008 (10 yrs)	WTD	Cl	56	Dry combustion	45
Narrabri, New South Wales	Narrabri FieldD1	Hulugalle et al. (2013)	2002	2005, 1998, 2004, 2006, 2008, 2011, 2012 (10 yrs)	WTD	Cl	64	Dry combustion	35
Kutztown, PA	Rodale	Elliott et al. (1994); Pimentel et al. (2005)	1981	1994, 2002 (21 yrs)	WTM	SiLo	30	Not reported	2 (2 stack PCs)
Winter, CA	RussellR	Kong et al. (2005); Wolf et al. (2018)	1993	2003, 2012 (19 yrs)	WTD	SiLo	18	Dry combustion	6
Swift Current, SK	SwiftCurrent	Campbell and Zentner (1997); Campbell et al. (2007)	1966	1976, 1990, 2009 (43 yrs)	CTD	SiLo	20	Dry combustion	12
Saginaw, MI	Saginaw	Christenson (1997)	1972	1981, 1991 (19 yrs)	CTM	SiCl	47	Dry combustion	28
Tribune, KS	Tribune	Halvorson and Schlegel (2012)	2001	2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010 (9 yrs)	WTD	SiLo	26	Dry combustion	18

Wooster, OH	Wooster	Collins et al. (1999); Dick et al. (1997)	1962	1971, 1980, 1992 (30 yrs)	CTM	SiLo	15	Dry combustion, Walkley-Black Method	6
Total	**15**								**270**

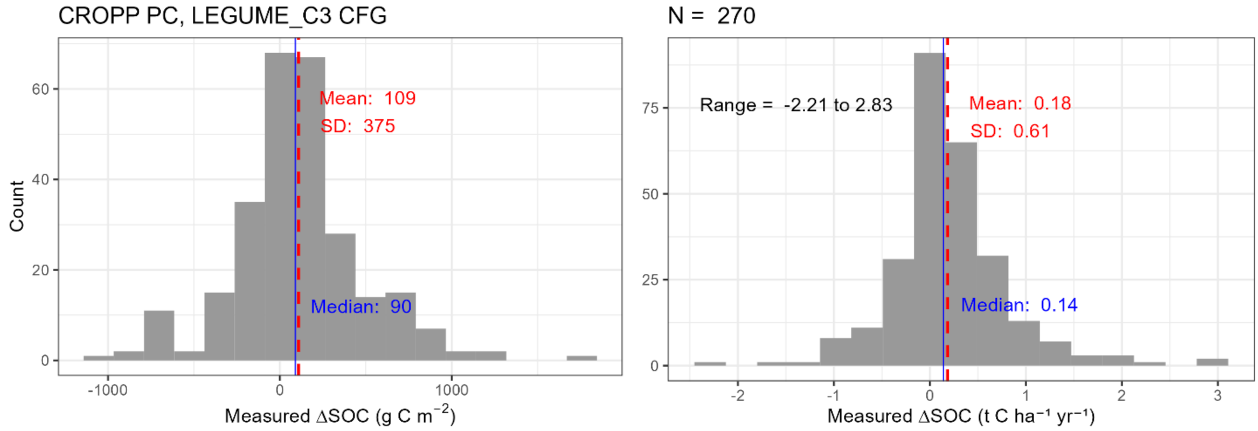

Figure 5. Histogram depicting observed changes in SOC in the studies (15 sites with experiment durations from 3 to 43 years) used for model validation in response to changes in cropping practices involving crops from the LEGUME_C3-type CFG.

7.3 CROPP x NON_LEGUME_C3

The validation datasets (Table 8), with their range of measured SOC changes of 292 paired observation differences shown in the histogram (Figure 6), are usable for this PC/CFG/SOC (CROPP x NON_LEGUME_C3 x SOC) combination within all declared climate zones and soil textures because:

- These experimental studies cover 4 IPCC climate zones (CTD, CTM, WTD, WTM), 3 of which (CTD, WTD, WTM) are in the declared project domain (Appendix B).
- Eight soil texture classes are included (Cl, ClLo, Lo, SaClLo, SaLo, SiCl, SiClLo, SiLo), and all of them are in the declared project area.
- Clay % contents range from 10% to 64%, which span 54 percentage points.
- At least one study isolates effects, e.g., only 10 out of the 292 observations compare stacks of PC changes.

Table 8. Dataset attributes of studies used in validation of CROPP x NON_LEGUME_C3 x SOC.

Location	Study name	Citation	Year initiated (1st measure if not same yr)	Year measured (duration)	IPCC climate zone	Soil texture	Clay (%)	SOC measurement method	Number of observation pairs
Davis, CA	Davis	Clark et al. (1998)	1988	1992, 1996 (8 yrs)	WTD	Lo	17	Walkley-Black method	10 (8 stack PCs)
Brookings, SD	Brookings	Hammerbeck et al. (2012); Wegner et al. (2018)	2008	2012 (4 yrs)	CTD	SiClLo	35	Dry combustion	6
Dalhart, TX	Dalhart	Halvorson et al. (2009); Halvorson and Reule (2007)	1999	2004, 2005, 2006 (7 yrs)	WTD	SaLo	18	Dry combustion	3
Five Points, CA	FivePoints	Mitchell et al. (2015); Mitchell et al. (2017); Veenstra et al. (2006)	1999	2000, 2004, 2007 (8 yrs)	WTD	ClLo	39	Dry combustion	6
Fort Valley, GA	FortValley	Sainju et al. (2005)	1999	2002 (3 yrs)	WTM	SaLo	10	Dry combustion	45
Lethbridge, AB	Lethbridge-ABC	Monreal and Janzen (1993)	1911	1922, 1940, 1953, 1967, 1980, 1990 (79 yrs)	CTD	SaClLo	31	Dry combustion (total C) and hot digestion with HCI to determine organic C indirectly	18
Mead, NE	Mead2	Varvel (2006)	1982 (1984)	1992, 1998, 2002 (18 yrs)	WTD	SiClLo	31	Dry combustion	18
Narrabri, New South Wales	Narrabri FieldC1	Senapati et al. (2014)	1993	1995, 1998, 2004, 2006, 2008, 2011, 2012 (19 yrs)	WTD	Cl	53	Dry combustion	7
Narrabri, New South Wales	Narrabri Field6	Rochester (2011)	1995 (1998)	2000, 2002, 2004, 2006, 2008 (10 yrs)	WTD	Cl	56	Dry combustion	35
Narrabri, New South Wales	Narrabri FieldD1	Hulugalle et al. (2013)	2002	2005, 1998, 2004, 2006, 2008, 2011, 2012 (10 yrs)	WTD	Cl	64	Dry combustion	35

Kutztown, PA	Rodale	Elliott et al. (1994); Pimentel et al. (2005)	1981	1994, 2002 (21 yrs)	WTM	SiLo	30	Not reported	2 (2 stack PCs)
Winter, CA	RussellR	Kong et al. (2005); Wolf et al. (2018)	1993	1997, 2003, 2012 (19 yrs)	WTD	SiLo	18	Dry combustion	8
Saginaw, MI	Saginaw	Christenson (1997)	1972	1981, 1991 (19 yrs)	CTM	SiCl	47	Dry combustion	30
Tribune, KS	Tribune	Halvorson and Schlegel (2012)	2001	2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010 (9 yrs)	WTD	SiLo	26	Dry combustion	36
Swift Current, SK	SwiftCurrent	Campbell and Zentner (1997); Campbell et al. (2007)	1966	1976, 1990, 2009 (43 yrs)	CTD	SiLo	20	Dry combustion	33
Total	**16**								**292**

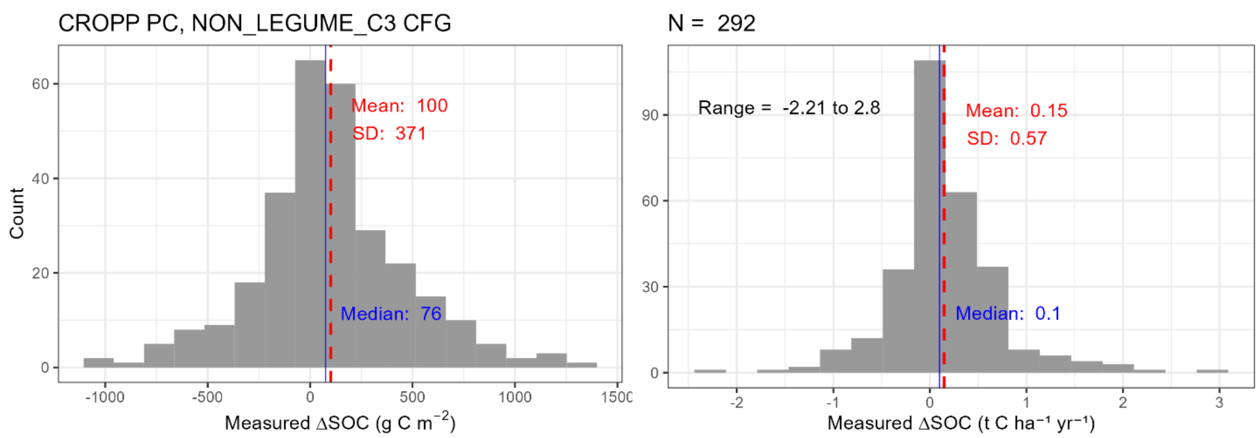

Figure 6. Histogram depicting observed changes in SOC in the studies (16 sites with experiment durations from 3 to 79 years) used for model validation in response to changed cropping practices involving crops from the NON_LEGUME_C3-type CFG.

7.4 CROPP x COTTON

The validation datasets (Table 9), with their range of measured SOC changes of 163 paired observation differences shown in the histogram (Figure 7), are usable for this PC/CFG/SOC (CROPP x COTTON x SOC) combination within all declared climate zones and soil textures because:

- These experimental studies cover 3 IPCC climate zones (TrM, WTD, WTM), and the declared climate zone WTD (Appendix B) is represented by the datasets.
- Three soil texture classes (Cl, ClLo, SaLo) are included. All of them are in the declared project area.
- Clay % contents range from 10% to 64%, which span 54 percentage points.
- At least one study isolates effects, e.g., only 1 out of the 163 observations compare stacks of PC changes.

Table 9. Dataset attributes of experimental studies used in the validation of CROPP x COTTON x SOC.

Location	Study name	Citation	Year initiated (1st measure if not same yr)	Year measured (duration)	IPCC climate zone	Soil texture	Clay (%)	SOC measurement method	Number of observation pairs
Five Points, CA	FivePoints	Mitchell et al. (2015); Mitchell et al. (2017); Veenstra et al. (2006)	1999	2000, 2004, 2007 (8 yrs)	WTD	ClLo	39	Dry combustion	6
Fort Valley, GA	FortValley	Sainju, Whitehead, and Singh (2005)	1999	2002 (3 yrs)	WTM	SaLo	10	Dry combustion	54
Goias, Brazil	Goias	Ferreira et al. (2019)	2005	2014 (9 yrs)	TrM	Cl	50	Dry combustion	4 (1 stack PCs)
Narrabri, New South Wales	Narrabri FieldC1	Senapati et al. (2014)	1985 (1993)	1995, 1998, 2004, 2006, 2008, 2011, 2012 (19 yrs)	WTD	Cl	53	Dry combustion	7
Narrabri, New South Wales	Narrabri Field6	Rochester (2011)	1995 (1998)	2000, 2002, 2004, 2006, 2008 (10 yrs)	WTD	Cl	56	Dry combustion	50
Narrabri, New South Wales	Narrabri FieldD1	Hulugalle et al. (2013)	2002	2005, 1998, 2004, 2006, 2008, 2011, 2012 (10 yrs)	WTD	Cl	64	Dry combustion	42
Total	**6**								**163**

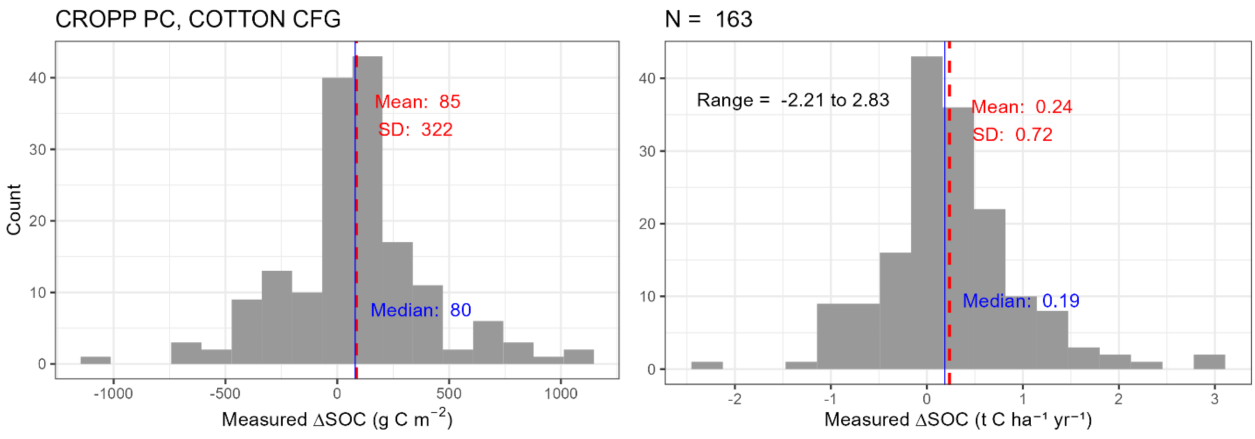

Figure 7. Histogram depicting observed changes in SOC in the studies (6 sites with experiment durations from 3 to 19 years) used for model validation in response to changed cropping practices involving crops from the COTTON-type CFG.

7.5 DISTU x NON_LEGUME_C4

The validation datasets (Table 10), with their range of measured SOC changes of 127 paired observation differences shown in the histogram (Figure 8), are usable for this PC/CFG/SOC (DISTU x NON_LEGUME_C4 x SOC) combination within all declared climate zones and soil textures because:

- These experimental studies cover 5 IPCC climate zones (CTD, CTM, TrM, WTD, WTM), all of which are in the declared project domain (Appendix B).
- Five soil texture classes are included, and all of them are in the declared project area: Cl, ClLo, SaLo, SiClLo, and SiLo.
- Clay % contents span 40 percentage points, ranging from 10% to 50%
- At least one study isolates effects, e.g., only 1 out of the 127 observations compare stacks of PC changes.

Table 10. Dataset attributes of studies used in the validation of DISTU x NON_LEGUME_C4 x SOC.

Location	Study name	Citation	Year initiated (1st measure if not same yr)	Year measured (duration)	IPCC climate zone	Soil texture	Clay (%)	SOC measurement Method	Number of observation pairs
Brookings, SD	Brookings	Hammerbeck et al. (2012); Wegner et al. (2018)	2008	2012 (4 yrs)	CTD	SiClLo	35	Dry combustion	8
Dixon Springs, IL	Dixonsprings	Olson et al. (2010)	1989	1991, 1992, 1995, 1996, 2000, 2003, 2007, 2009 (20 yrs)	WTM	SiLo	19	Walkley-Black method	24
Fort Collins, CO	FortCollins	Halvorson et al. (2009)	1999	2000, 2001, 2002, 2003, 2005, 2006, 2007 (8 yrs)	CTD	ClLo	34	Dry combustion	7
Fort Valley, GA	FortValley	Sainju et al. (2005)	1999	2002 (3 yrs)	WTM	SaLo	10	Dry combustion	36
Hoytville, OH	Hoytville	Collins et al. (1999)	1963 (1980)	1993 (23 yrs)	CTM	SiClLo	40	Dry combustion	3
Goias, Brazil	Goias	Ferreira et al. (2019)	2005	2014 (9 yrs)	TrM	Cl	50	Dry combustion	1 (1 stack PCs)
Ithaca, NE	IthacaNE	Jin and Varvel (2018)	2001	2010 (9 yrs)	WTD	SiLo	23	Dry combustion	9
Lexington, KY	Lexington	Blevins et al. (1983); Ismail et al. (1994)	1970 (1975)	1980, 1989 (14 yrs)	WTM	SiLo	23	Dry combustion, Sulfuric acid-permanganate method of Allison (1965)	8
Rosemount, MN	Rosemount	Clapp et al. (2000); Dolan et al. (2006)	1980	1993, 2002 (22 yrs)	CTD	SiLo	24	Dry combustion	18
South Charleston, OH	Scharleston	Collins et al. (1999); Jarecki and Lal (2005)	1962 (1971)	1992, 2003 (32 yrs)	WTM	SiLo	20	Dry combustion	6
Wooster, OH	Wooster	Collins et al. (1999) and Dick et al. (1997)	1962	1971, 1980, 1992; 2005 (43 yrs)	CTM	SiLo	15	Dry combustion, Walkley-Blackmethod	7
Total	**11**								**127**

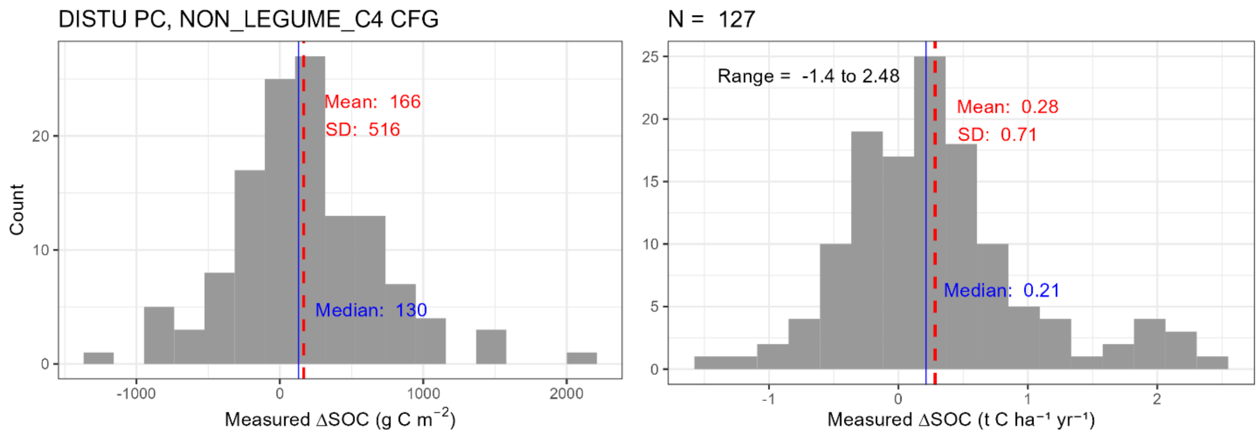

Figure 8. Histogram depicting observed changes in SOC in the studies (11 sites with experiment durations from 3 to 43 years) used for model validation in response to changed tillage practices involving crops from the NON_LEGUME_C4-type CFG.

7.6 DISTU x LEGUME_C3

The validation datasets (Table 11), with their range of measured SOC changes of 60 paired observation differences shown in the histogram (Figure 9), are usable for this PC/CFG/SOC (DISTU x LEGUME_C3 x SOC) combination within all declared climate zones and soil textures because:

- These experimental studies cover 5 IPCC climate zones (CTD, CTM, TrM, WTD, WTM), all of which are in the declared project domain (Appendix B).
- Five soil texture classes are included, and all of them are in the declared project area: Cl, ClLo, SaLo, SiClLo, and SiLo.
- Clay % contents range from 10% to 50%, which spans 40 percentage points.
- At least one study isolates effects, e.g., only 1 out of the 60 observations compare stacks of PC changes.

Table 11. Dataset attributes of studies used in the validation of DISTU x LEGUME_C3 x SOC.

Location	Study name	Citation	Year initiated (1st measure if not same yr)	Year measured (duration)	IPCC climate zone	Soil texture	Clay (%)	SOC measurement method	Number of observation pairs
Brookings, SD	Brookings	Hammerbeck et al. (2012); Wegner et al. (2018)	2008	2012 (4 yrs)	CTD	SiClLo	35	Dry combustion	8
Dixon Springs, IL	Dixonsprings	Olson et al. (2010)	1989	1991, 1992, 1995, 1996, 2000, 2003, 2007, 2009 (20 yrs)	WTM	SiLo	19	Walkley-Black method	24
Five Points, CA	FivePoints	Mitchell et al. (2015); Mitchell et al. (2017); Veenstra et al. (2006)	1999	2000, 2004, 2007 (8 yrs)	WTD	ClLo	39	Dry combustion	3
Fort Valley, GA	FortValley	Sainju, et al. (2005)	1999	2002 (3 yrs)	WTM	SaLo	10	Dry combustion	18
Hoytville, OH	Hoytville	Collins et al. (1999)	1963 (1980)	1993 (23 yrs)	CTM	SiClLo	40	Dry combustion	3
Goias, Brazil	Goias	Ferreira et al. (2019)	2005	2014 (9 yrs)	TrM	Cl	50	Dry combustion	1 (1 stack PCs)
Wooster, OH	Wooster	Collins et al. (1999); Dick et al. (1997)	1962	1971, 1980, 1992 (30 yrs)	CTM	SiLo	15	Dry combustion, Walkley-Blackmethod	3
Total	**7**								**60**

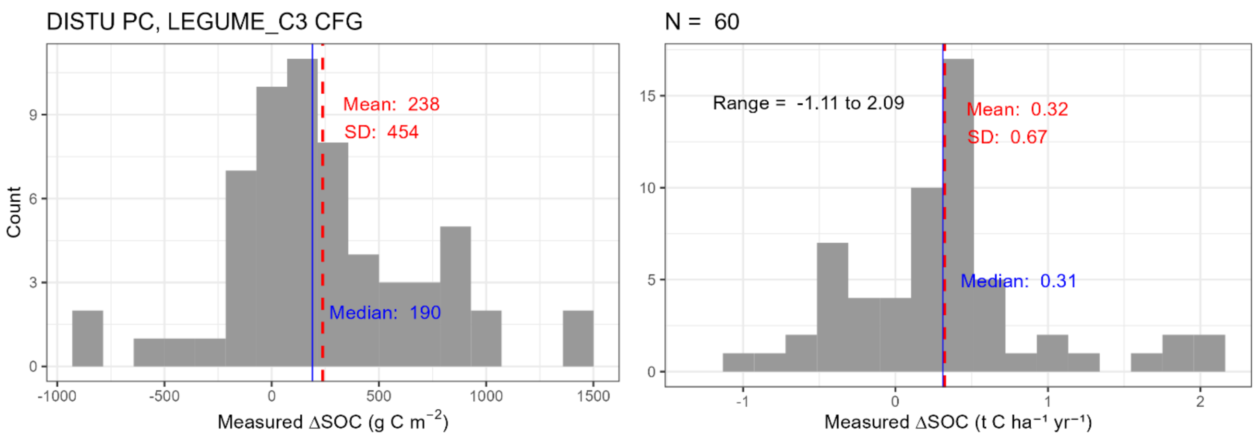

Figure 9. Histogram depicting observed changes in SOC in the studies (7 sites with experiment durations from 3 to 30 years) used for model validation in response to changed tillage practices involving crops from the LEGUME_C3-type CFG.

7.7 DISTU x NON_LEGUME_C3

The validation datasets (Table 12), with their range of measured SOC changes of 99 paired observation differences shown in the histogram (Figure 10), are usable for this PC/CFG/SOC (DISTU x NON_LEGUME_C3 x SOC) combination within all declared climate zones and soil textures because:

- These experimental studies cover 3 IPCC climate zones (CTD, WTD, WTM), all of which are in the declared project domain (Appendix B).
- Five soil texture classes (Lo, ClLo, SaLo, SiClLo, SiLo) are included, and all of them are in the declared project area.
- Clay % contents range from 10% to 39%, which span 29 percentage points.
- At least one study isolates effects, e.g., none of the 99 observations compare stacks of PC changes.

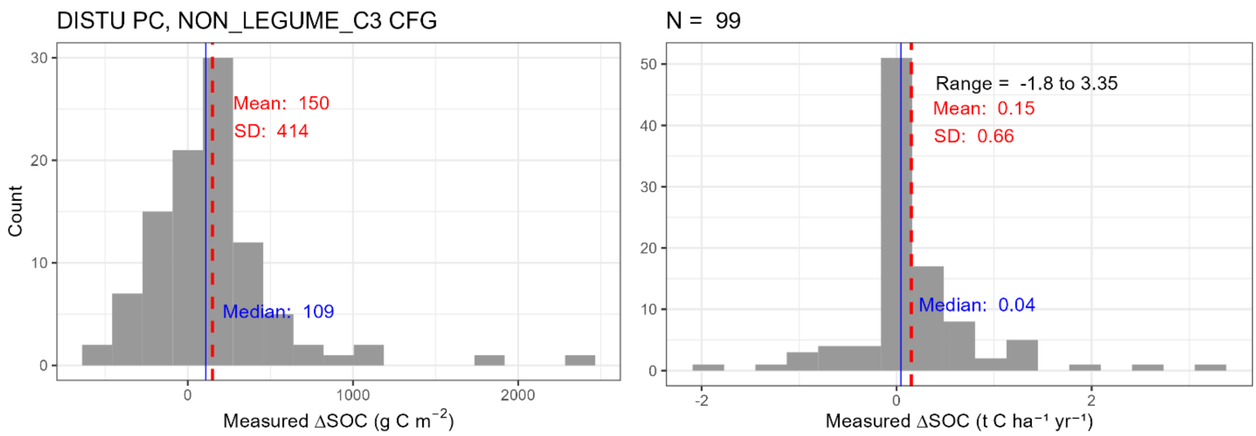

Figure 10. Histogram depicting observed changes in SOC in the studies (6 sites with experiment durations from 3 to 79 years) used for model validation in response to changed tillage practices involving crops from the NON_LEGUME_C3-type CFG.

Table 12. Dataset attributes of studies used in the validation of DISTU x NON_LEGUME_C3 x SOC.

Location	Study name	Citation	Year initiated (1st measure if not same yr)	Year measured (duration)	IPCC climate zone	Soil texture	Clay (%)	SOC measurement method	Number of observation pairs
Brookings, SD	Brookings	Hammerbeck et al. (2012); Wegner et al. (2018)	2008	2012 (4 yrs)	CTD	SiClLo	35	Dry combustion	3
Five Points, CA	FivePoints	Mitchell et al. (2015); Mitchell et al. (2017); Veenstra et al. (2006)	1999	2000, 2004, 2007 (8 yrs)	WTD	ClLo	39	Dry combustion	6
Fort Valley, GA	FortValley	Sainju et al. (2005)	1999	2002 (3 yrs)	WTM	SaLo	10	Dry combustion	18
Pendleton, OR	ORpegn	Bista et al. (2016); Ghimire et al. (2015); Rasmussen and Smiley (1997)	1931	1941, 1951, 1964, 1976, 1986, 1992, 1995, 2010 (79 yrs)	WTD	SiLo	22	Dry combustion, Walkley-Black	35
Mandan, ND	Mandan	Halvorson et al. (2002)	1983	1990, 1996 (13 yrs)	CTD	SiLo	20	Dry combustion	34
Sidney, NE	Sidney	Elliott et al. (1994)	1970	1993 (23 yrs)	CTD	Lo	25	Dry combustion	3
Total	**6**								**99**

7.8 DISTU x COTTON

The validation datasets (Table 13), with their range of measured SOC changes of 50 paired observation differences shown in the histogram (Figure 11), are usable for this PC/CFG/SOC (DISTU x COTTON x SOC) combination within all declared climate zones and soil textures because:

- These experimental studies cover 3 IPCC climate zones (TrM, WTD, WTM), and the declared climate zone WTD for DISTU x COTTON (see Appendix B) is represented by the datasets.

- Three soil texture classes (Cl, ClLo, SaLo) are included, all of which are in the declared project area.
- Clay % contents range from 10% to 53%, which span 43 percentage points.
- At least one study isolates effects, e.g., only 1 out of the 50 observations compare stacks of PC changes.

Table 13. Dataset attributes of experimental studies used in the validation of DISTU x COTTON x SOC.

Location	Study name	Citation	Year initiated (1st measure if not same yr)	Year measured (duration)	IPCC climate zone	Soil texture	Clay (%)	SOC Measurement method	Number of observation pairs
Five Points, CA	FivePoints	Mitchell et al. (2015); Mitchell et al. (2017); Veenstra et al. (2006)	1999	2000, 2004, 2007 (8 yrs)	WTD	ClLo	39	Dry combustion	6
Fort Valley, GA	FortValley	Sainju et al. (2005)	1999	2002 (3 yrs)	WTM	SaLo	10	Dry combustion	36
Goias, Brazil	Goias	Ferreira et al. (2019)	2005	2014 (9 yrs)	TrM	Cl	50	Dry combustion	1 (1 stack PCs)
Narrabri, New South Wales	Narrabri FieldC1	Senapati et al. (2014)	1985 (1993)	1995, 1998, 2004, 2006, 2008, 2011, 2012 (19 yrs)	WTD	Cl	53	Dry combustion	7
Total	**4**								**50**

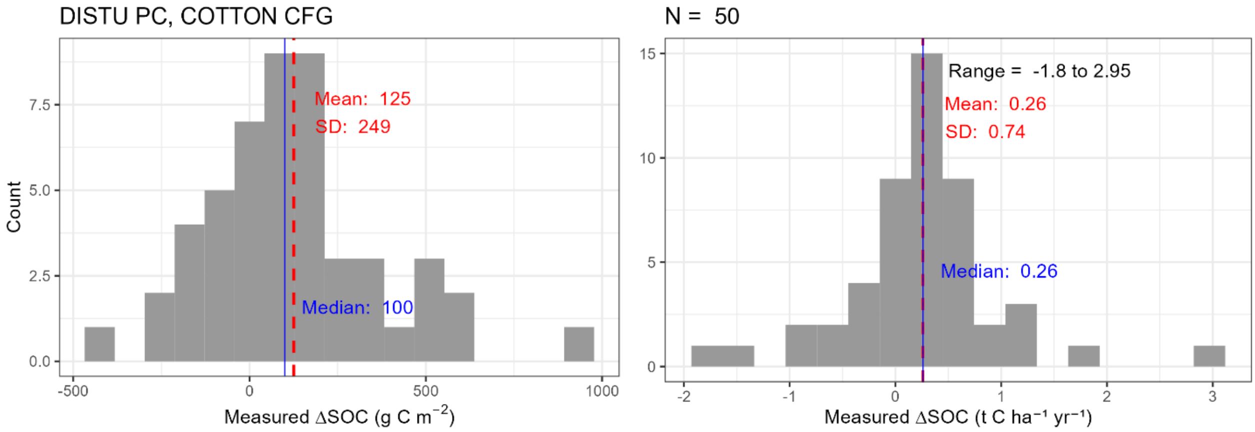

Figure 11. Histogram depicting observed changes in SOC in the studies (4 sites with experiment durations from 3 to 19 years) used for model validation in response to changed tillage practices involving crops from the COTTON-type CFG.

7.9 NFERT x NON_LEGUME_C4

The validation datasets (Table 14), with their range of measured SOC changes of 139 paired observation differences shown in the histogram (Figure 12), are usable for this PC/CFG/SOC (NFERT x NON_LEGUME_C4 x SOC) combination within all declared climate zones and soil textures because:

- These experimental studies cover 3 IPCC climate zones (CTD, WTD, WTM), all of which are in the declared project domain (Appendix B).
- Four soil texture classes (Lo, SaLo, SiClLo, and SiLo) are included, all of which are in the declared project domain.
- Clay % contents span 22 percentage points, ranging from 10% to 32%.
- At least one study isolates effects, e.g., only 12 out of the 139 observations compare stacks of PC changes.

Table 14. Dataset attributes of studies used in the validation of NFERT x NON_LEGUME_C4 x SOC.

Location	Study name	Citation	Year initiated (1st measure if not same yr)	Year measured (duration)	IPCC climate zone	Soil texture	Clay (%)	SOC measurement method	Number of observation pairs
Davis, CA	Davis	Clark et al. (1998)	1988	1992, 1996 (8 yrs)	WTD	Lo	17	Walkley-Black method	8 (8 stack PCs)
Dalhart, TX	Dalhart	Halvorson et al. (2009); Halvorson and Reule (2007)	1999	2000,2001, 2002, 2003 (7 yrs)	WTD	SaLo	18	Dry combustion	4
Mead, NE	Mead2	Varvel (2006)	1982 (1984)	1992, 1998, 2002 (18 yrs)	WTD	SiClLo	31	Dry combustion	54
Fort Valley, GA	FortValley	Sainju et al. (2005)	1999	2002 (3 yrs)	WTM	SaLo	10	Dry combustion	36
Ithaca, NE	Ithaca2	Jin and Varvel (2018); Jin et al. (2015)	1998	2011 (13 yrs)	WTD	SiClLo	32	Dry combustion	3
Lexington, KY	Lexington	Blevins et al. (1983); Ismail et al. (1994)	1970 (1975)	1980, 1989 (14 yrs)	WTM	SiLo	23	Dry combustion, Sulfuric acid-permanganate method of Allison (1965)	24
Kutztown, PA	Rodale	Elliott et al. (1994); Pimentel et al. (2005)	1981	1994, 2002 (21 yrs)	WTM	SiLo	30	Not reported	2 (2 stack PCs)
Rosemount, MN	Rosemount	Dolan et al. (2006)	1980	1993, 2002 (22 yrs)	CTD	SiLo	24	Dry combustion	6
Winter, CA	RussellR	Kong et al. (2005)	1993	2003, 2012 (19 yrs)	WTD	SiLo	18	Dry combustion	2
Total	**9**								**139**

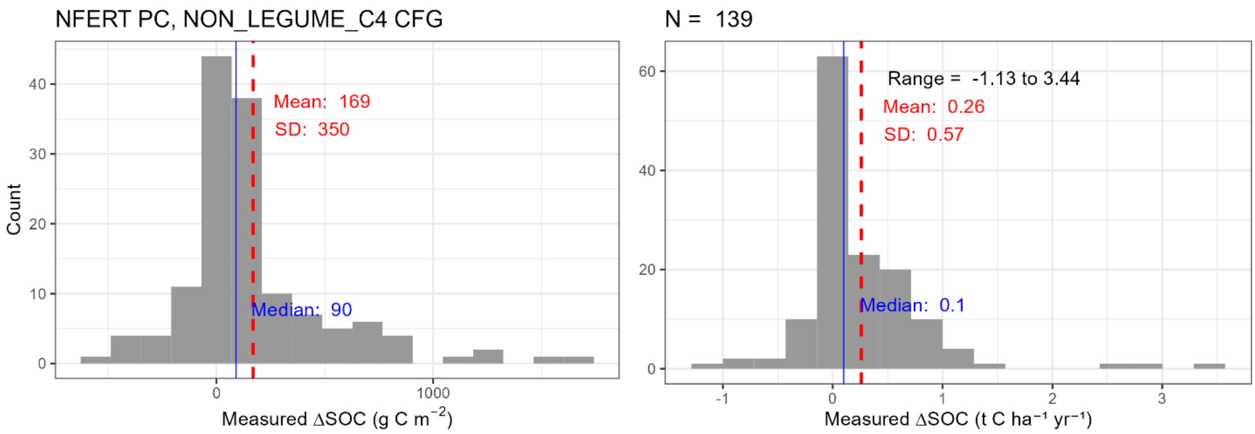

Figure 12. Histogram depicting observed changes in SOC in the studies (9 sites with experiment durations from 3 to 22 years) used for model validation in response to changed nitrogen management practices involving crops from the NON_LEGUME_C4-type CFG.

7.10 NFERT x LEGUME_C3

The validation datasets (Table 15), with their range of measured SOC changes of 73 paired observation differences shown in the histogram (Figure 13), are usable for this PC/CFG/SOC (NFERT x LEGUME_C3 x SOC) combination within all declared climate zones and soil textures because:

- These experimental studies cover 2 IPCC climate zones (WTD, WTM), both of which are in the declared project domain (Appendix B).
- Four soil texture classes (Lo, SaLo, SiClLo, SiLo) are included, and all of them are in the declared project area.
- Clay % contents range from 10% to 31%, which spans 21 percentage points.
- At least one study isolates effects, e.g., only 10 out of the 73 observations compare stacks of PC changes.

Table 15. Dataset attributes of studies used in the validation of NFERT x LEGUME_C3 x SOC.

Location	Study name	Citation	Year initiated (1st measure if not same yr)	Year measured (duration)	IPCC climate zone	Soil texture	Clay (%)	SOC measurement method	Number of observation pairs
Davis, CA	Davis	Clark et al. (1998)	1988	1992, 1996 (8 yrs)	WTD	Lo	17	Walkley-Black method	8 (8 stack PCs)
Fort Valley, GA	FortValley	Sainju et al. (2005)	1999	2002 (3 yrs)	WTM	SaLo	10	Dry combustion	18
Mead, NE	Mead2	Varvel (2006)	1982 (1984)	1992, 1998, 2002 (18 yrs)	WTD	SiClLo	31	Dry combustion	45
Kutztown, PA	Rodale	Elliott et al. (1994); Pimentel et al. (2005)	1981	1994, 2002 (21 yrs)	WTM	SiLo	30	Not reported	2 (2 stack PCs)
Total	**4**								**73**

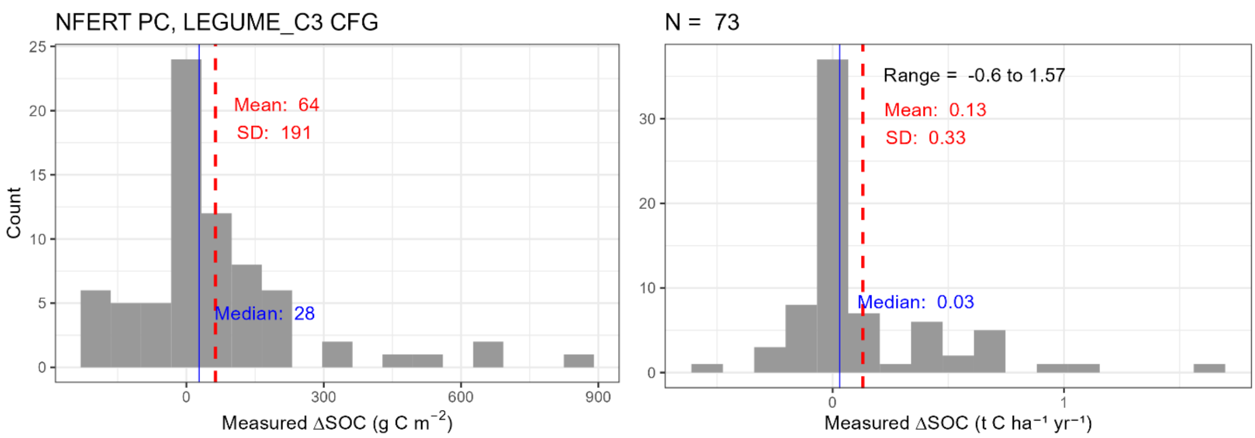

Figure 13. Histogram depicting observed changes in SOC in the studies (4 sites with experiment durations from 3 to 21 years) used for model validation in response to changed nitrogen management practices involving crops from the LEGUME_C3-type CFG.

7.11 NFERT x NON_LEGUME_C3

The validation datasets (Table 16), with their range of measured SOC changes of 185 paired observation differences shown in the histogram (Figure 14), are usable for this PC/CFG/SOC

(NFERT x NON_LEGUME_C3 x SOC) combination within all declared climate zones and soil textures because:

- These experimental studies cover 4 IPCC climate zones (CTD, CTM, WTD, WTM), 3 of which are in the declared project domain (Appendix B).
- Four soil texture classes (Lo, SaLo, SiClLo, SiLo) are included, all of which are in the declared project domain.
- Clay % contents span 21 percentage points, ranging from 10% to 31%.
- At least one study isolates effects, e.g., only 12 out of the 185 observations compare stacks of PC changes.

Table 16. Dataset attributes of studies used in the validation of NFERT x NON_LEGUME_C3 x SOC.

Location	Study name	Citation	Year initiated (1st measure if not same yr)	Year measured (duration)	IPCC climate zone	Soil texture	Clay (%)	SOC measurement method	Number of observation pairs
Rothamsted, England	Broadbalk	Research (2014)	1843	1865, 1881, 1893, 1914, 1936, 1944, 1966, 1987, 1992, 1997, 2000, 2005 (162 yrs)	CTM	SiClLo	25	Dry combustion	10
Dalhart, TX	Dalhart	Halvorson et al. (2009); Halvorson and Reule (2007)	1999	2000, 2001, 2002, 2003 (7 yrs)	WTD	SaLo	18	Dry combustion	4
Davis, CA	Davis	Clark et al. (1998)	1988	1992, 1996 (8 yrs)	WTD	Lo	17	Walkley-Black method	8 (8 stack PCs)
Lexington, KY	Lexington	Blevins et al. (1983); Ismail et al. (1994)	1970 (1975)	1980, 1989 (14 yrs)	WTM	SiLo	23	Dry combustion, Sulfuric acid-permanganate method of Allison (1965)	24
Mandan, ND	Mandan	Halvorson et al. (2002)	1983	1990, 1996 (13 yrs)	CTD	SiLo	20	Dry combustion	36
Mead, NE	Mead2	Varvel (2006)	1982 (1984)	1992, 1998, 2002 (18 yrs)	WTD	SiClLo	31	Dry combustion	18
Fort Valley, GA	FortValley	Sainju et al. (2005)	1999	2002 (3 yrs)	WTM	SaLo	10	Dry combustion	18

Pendleton, OR	ORpegn	Bista et al. (2016); Ghimire et al. (2015); Rasmussen and Smiley (1997)	1931	1941, 1951, 1964, 1976, 1986, 1992, 1995, 2010 (79 yrs)	WTD	SiLo	22	Dry combustion, Walkley-Black	74
Kutztown, PA	Rodale	Elliott et al. 1994); Pimentel et al. (2005)	1981	1994, 2002 (21 yrs)	WTM	SiLo	30	Not reported	2 (2 stack PCs)
Winter, CA	RussellR	Kong et al. (2005); Wolf et al. (2018)	1993	2003, 2012 (19 yrs)	WTD	SiLo	18	Dry combustion	6
Swift Current, SK	SwiftCurrent	Campbell and Zentner (1997); Campbell et al. (2007)	1966	1976, 1990, 2009 (43 yrs)	CTD	SiLo	20	Dry combustion	9
Total	**10**								**185**

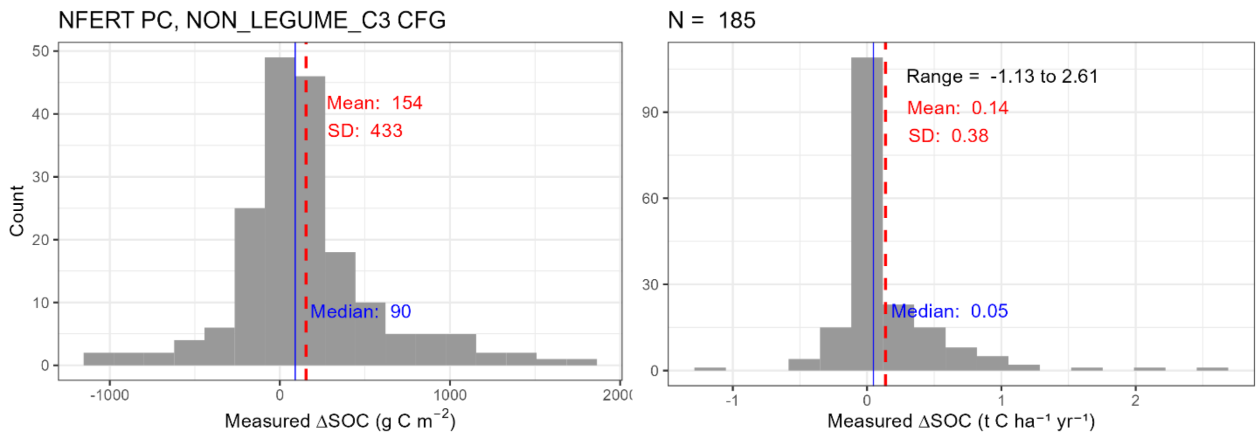

Figure 14. Histogram depicting observed changes in SOC in the studies (10 sites with experiment durations from 3 to 162 years) used for model validation in response to changed nitrogen management practices involving crops from the NON_LEGUME_C3-type CFG.

7.12 NFERT x ALL_CFGs

The combined experimental studies (Table 17) for the NFERT PC and all CFGs combinations for SOC, together with dataset list in Table 9 used in the validation of CROPP x COTTON x SOC, and/or Table 15 used in the validation of NFERT × NON_LEGUME_C3, support the use of the Validation Report for NFERT x COTTON (see **Section 5.3 "Domain covered by this validation"**, where the case has been addressed explicitly.) Specially, when considering observations of the NFERT PC across all annual CFGs:

- The selected studies cover 4 IPCC climate zones (CTD, CTM, WTD, WTM) and represent the declared project domain.
- Four soil texture classes (Lo, SaLo, SiClLo, SiLo) are included, all of which are in the declared project domain.
- Clay % contents span 21 percentage points, ranging from 10% to 31%.
- At least one study isolates effects, e.g., only 12 out of the 287 observations compare stacks of PC changes.

Table 17. Combined datasets of studies for NFERT x ALL_CFGs x SOC.

Location	Study Name	Citation	Year initiated (1st measure if not same yr)	Year measured (duration)	IPCC climate zone	Soil texture	Clay (%)	SOC measurement method	Number of observation pairs
Rothamsted, England	Broadbalk	Research (2014)	1843	1865, 1881, 1893, 1914, 1936, 1944, 1966, 1987, 1992, 1997, 2000, 2005 (162 yrs)	CTM	SiClLo	25	Dry combustion	10
Dalhart, TX	Dalhart	Halvorson et al. (2009); Halvorson and Reule (2007)	1999	2000, 2001, 2002, 2003 (7 yrs)	WTD	SaLo	18	Dry combustion	4
Davis, CA	Davis	Clark et al. (1998)	1988	1992, 1996 (8 yrs)	WTD	Lo	17	Walkley-Black method	8 (8 stack PCs)
Mandan, ND	Mandan	Halvorson et al. (2002)	1983	1990, 1996 (13 yrs)	CTD	SiLo	20	Dry combustion	36
Mead, NE	Mead2	Varvel (2006)	1982 (1984)	1992, 1998, 2002 (18 yrs)	WTD	SiClLo	31	Dry combustion	63
Fort Valley, GA	FortValley	Sainju et al. (2005)	1999	2002 (3 yrs)	WTM	SaLo	10	Dry combustion	36
Pendleton, OR	ORpegn	Bista et al. (2016); Ghimire et al. (2015); Rasmussen and Smiley (1997)	1931	1941, 1951, 1964, 1976, 1986, 1992, 1995, 2010 (79 yrs)	WTD	SiLo	22	Dry combustion, Walkley-Black	74
Rosemount, MN	Rosemount	Dolan et al. (2006)	1980	1993, 2002 (22 yrs)	CTD	SiLo	24	Dry combustion	6
Kutztown, PA	Rodale	Elliott et al. 1994);	1981	1994, 2002 (21 yrs)	WTM	SiLo	30	Not reported	2 (2 stack PCs)

		Pimentel et al. (2005)							
Winter, CA	RussellR	Wolf et al. (2018)	1993	2003, 2012 (19 yrs)	WTD	SiLo	18	Dry combustion	6 (2 stack PCs)
Swift Current, SK	SwiftCurrent	Campbell and Zentner (1997); Campbell et al. (2007)	1966	1976, 1990, 2009 (43 yrs)	CTD	SiLo	20	Dry combustion	9
Ithaca, NE	Ithaca2	Jin and Varvel (2018); Jin et al. (2015)	1998	2011 (13 yrs)	WTD	SiClLo	32	Dry combustion	3
Lexington, KY	Lexington	Blevins et al. (1983); Ismail et al. (1994)	1970 (1975)	1980, 1989 (14 yrs)	WTM	SiLo	23	Dry combustion, Sulfuric acid-permanganate method of Allison (1965)	24
Total	**13**								**281**

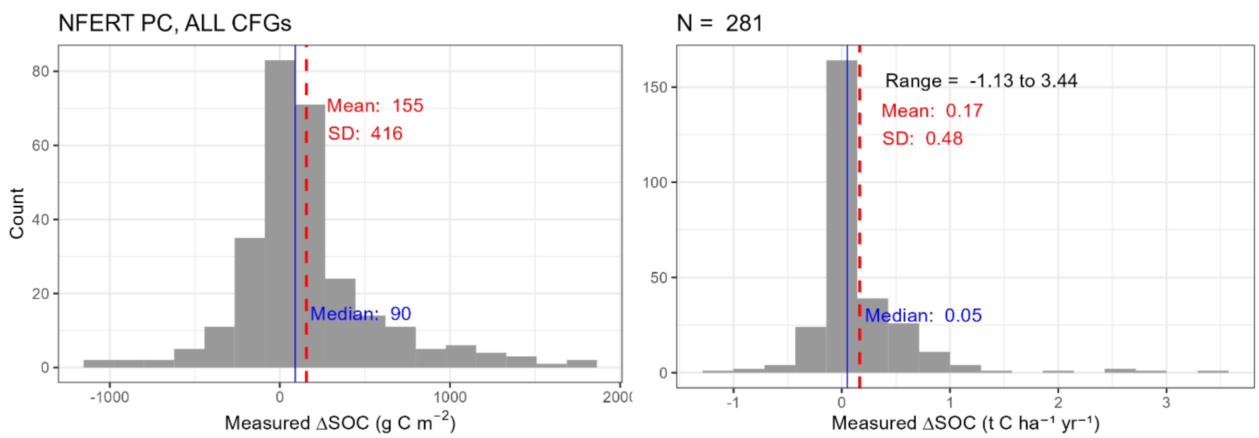

Figure 15. Histogram depicting observed changes in SOC in the studies (13 sites with experiment durations from 3 to 162 years) used for model validation in response to changed nitrogen management practices involving crops from ALL-type CFGs.

8. Assessment of Model Bias

This section follows *VMD53 Model Requirements Section 5.2.4 Assessment of Bias for Each Practice Category.*

8.1 Description of bias calculation

For all practice categories, model bias was calculated for each study PC × CFG combination as the average relative error (FAO, 2019) between model predicted and observed SOC changes in response to practice change using Eq. (4):

$$bias = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i) \tag{4}$$

where P_i is the modeled change in SOC for the i^{th} observation of the practice change (e.g., $SOC_{no\text{-}till}$- SOC_{till}); O_i is measured change in SOC for the i^{th} observation of the practice change; i is the index of the observation treatment pairs within the study; and n is the number of treatment pairs used from the study (per Equation 1 of the Model Requirements). When a study reported treatment pairs fitting multiple PC x CFG categories, only the paired observations matching the category of interest were included in the calculation.

The mean bias was then computed as the unweighted mean of all biases from individual studies (per VMD53, Section 5.2.4). The mean bias was compared to the pooled measurement uncertainty (PMU), which is defined as the pooled standard error of all the measured values for a practice change. Calculating PMU involves combining the uncertainties from multiple studies which report standard errors to obtain an overall estimate of uncertainty. The pooled measurement uncertainty was calculated as below:

$$PMU = \sqrt{\frac{\sum_{j=1}^{k}\sigma_j^2 \times (n_{j1}+n_{j2}-2)}{\sum_{j=1}^{k}(n_{j1}+n_{j2}-2)}} \tag{5}$$

where k is the number of observations examined across all studies with uncertainty reported; σ_j is the standard error of the j^{th} observation of difference between the treatments; n_{j1} and n_{j2} are the number of replicates (sample sizes) of the measurements in the two treatments, respectively. If the number of replicate measurements in the j^{th} observation are the same as $n_{j1} = n_{j2}$, then Eq (5) can be simplified as:

$$PMU = \sqrt{\frac{\sum_{j=1}^{k}\sigma_j^2 \times (n_j-1)}{\sum_{j=1}^{k}(n_j-1)}} \tag{6}$$

where n_j is the number of replicate measurements used in the j^{th} observation of the paired treatments. For the validation dataset in this report, uncertainty was expressed as standard error in all studies that reported it. If a site had reported standard deviation (SD), then the standard errors for the paired treatments 1 (σ_{Trt1}) and 2 (σ_{Trt2}) are calculated as:

$$\sigma_{Trt1} = \frac{SD_{Trt1}}{\sqrt{n_{Trt1}}}; \quad \sigma_{Trt2} = \frac{SD_{Trt2}}{\sqrt{n_{Trt2}}} \tag{7}$$

and σ_{diff} for this pair of observations would be:

$$\sigma_{diff} = \sqrt{\frac{SD_{Trt1}^2}{n_{Trt1}} + \frac{SD_{Trt2}^2}{n_{Trt2}}} \tag{8}$$

where Trt_1 and Trt_2 are the numbers of replicate measurements used in the paired treatment 1 and treatment 2, respectively.

8.2 Example calculation of model bias for a study

The Ithaca2 site has 3 levels of inorganic N application rates (NFERT PC with NON_LEGUME_C4 combination). In addition to the initial SOC measurement in 1998, SOC stocks were measured in 2011 (Table 18, refer to source of data in box 1 from Jin et al., 2015). Model bias was calculated using Eq. 4, resulting in the bias of 351.9 g C m^{-2} for this study site.

Table 18. Calculating model bias based on pairs of modeled and observed changes in SOC differences between treatments for one study site.

Treatments	SOC stock (g C m^{-2})		% PBIAS	Pairs	SOC stock change (g C m^{-2})		
	Observed	Modeled			Observed	Modeled	Modeled-observed
Ithaca2_60N	6020	5571.5	-7.4	Ithaca2_120N vs. Ithaca2_60N	-430	171.4	601.4
Ithaca2_120N	5590	5742.9	2.7	Ithaca2_180N vs. Ithaca2_120N	130	56.5	-73.5
Ithaca2_180N	5720	5799.4	1.4	Ithaca2_180N vs. Ithaca2_60N	-300	227.9	527.9
Sum							**1055.8**
Bias = 1055.8 /3 = 351.9 (g C m^{-2})							

8.3 Example derivation of PMU calculation of an observed practice change effect

The NFERT x NON_LEGUME_C4 validation dataset contained 2 sites with a total of 5 pairs of SOC observations from practice changes that had uncertainty reported for both observed treatments (Table 19), which allows computing the standard error of the difference, σ_{diff}, between the paired treatments. To compute the PMU for this PC/CFG category, Eq. 6 was used (Table 19). One example data source is presented in Box 1 from Jin et al., 2015, which reports standard errors (SEs) for various treatments. The complete collection of reported SEs, along with information about paired treatments, is available in Tables C1-C3 (in Appendix C). These tables include details such as an illustrative case demonstrating the calculation of PMU for the CROPP x NON_LEGUME_C4 validation dataset, as well as comprehensive treatment descriptions.

Table 19. Calculating pooled measurement uncertainty (PMU) of SOC change based on pairs of observations from the NFERT x NON_LEGUME_C4 validation dataset with SOC measurement uncertainty.

Site	Trt1_n	Trt2_n	g C m^{-2}			n-1	σ^2	σ^2x(n-1)
			σ_{Trt1}	σ_{Trt2}	σ_{diff}			
Ithaca2	3	3	410	520	662.2	2	438500	877000
	3	3	590	410	718.5	2	516200	1032400
	3	3	590	520	786.4	2	618500	1237000
RussellR	12	12	95	67	116.2	11	13514	148654
	6	6	261	120	287.3	5	82521	412605
Sum						**22**		**3707659**

$$PMU = \sqrt{3707659/22} = \mathbf{410.5\ (g\ C\ m^{-2})}$$

Box 1. One example of source data from Jin et al., 2015 (reported SOC stock measurements and standard errors in 0-30 cm depth; data in 1998 was used for the initialization of soil C pools).

Table 1. Selected properties for surface soils (0–30 cm) in 1998 and 2011 under continuous no-till corn at three N fertilizer rates (60, 120, 180 kg N ha^{-1}).

Depth	Baseline† 1998	2011 60 kg N ha^{-1}	2011 120 kg N ha^{-1}	2011 180 kg N ha^{-1}
cm				
		Bulk density, Mg m^{-3}		
0–5	1.25 ± 0.08‡***	1.42 ± 0.08	1.46 ± 0.06	1.45 ± 0.08
5–10	1.38 ± 0.08***	1.65 ± 0.05	1.64 ± 0.02	1.46 ± 0.07
10–30	1.39 ± 0.05*	1.38 ± 0.03	1.43 ± 0.01	1.40 ± 0.03
0–30	1.35 ± 0.05***	1.43 ± 0.03	1.47 ± 0.01	1.42 ± 0.02
		pH (1:1 water)		
0–5	6.52 ± 0.34***	5.97 ± 0.15 a§	5.75 ± 0.20 ab	5.45 ± 0.10 b
5–10	6.50 ± 0.31***	6.13 ± 0.13	6.17 ± 0.17	5.69 ± 0.19
10–30	6.60 ± 0.23*	6.38 ± 0.12 a	6.70 ± 0.11 b	6.01 ± 0.14 c
0–30	6.56 ± 0.24***	6.27 ± 0.13 a	6.45 ± 0.11 a	5.87 ± 0.12 b
		Soil organic C, Mg C ha^{-1}		
0–5	9.3 ± 2.0***	11.5 ± 1.0	12.6 ± 0.4	12.0 ± 0.7
5–10	9.1 ± 1.8***	11.7 ± 0.8	10.9 ± 0.2	10.0 ± 1.0
10–30	30.6 ± 8.4*	37.0 ± 3.6	32.4 ± 3.7	35.2 ± 4.5
0–30	49.4 ± 11.8***	60.2 ± 5.2	55.9 ± 4.1	57.2 ± 5.9
		ΔSoil organic C, Mg C ha^{-1}¶		
0–5	–	3.3 ± 0.7	3.7 ± 0.4	5.2 ± 0.9
5–10	–	2.5 ± 0.6	1.9 ± 0.5	2.1 ± 1.0
10–30	–	1.2 ± 0.3	0.9 ± 0.7	2.2 ± 0.9
0–30	–	7.0 ± 1.2	6.5 ± 1.1	9.5 ± 2.2

* For each depth, 1998 and 2011 are significantly different at $P \leq 0.05$.

*** For each depth, 1998 and 2011 are significantly different at $P \leq 0.001$.

† Baseline values for bulk density, soil pH, and soil organic C are from Follett et al. (2012) and Stewart et al. (2015). Note that 1998 values are averaged across all N fertilizer rates because no N fertilizer treatments had been imposed at the time of site establishment.

‡ Mean ± 1 SE.

§ Means followed by different letters indicate N fertilizer rate differences in 2011 when the N × year interaction was significant ($P \leq 0.10$). If no letter is shown, there were no statistical differences among N fertilizer rates.

8.4 Pooled measurement uncertainty used for each PC/CFG/SOC combination

PMUs of difference in SOC between treatments across entire observations with reported uncertainty were computed for each PC/CFG combination (Table 20) as demonstrated above in Section 8.3. Because not all studies report measurement standard error, Section 5.2.4 of the VMD53 Model Requirements allows PMU to be computed using all studies used in a Model Validation Report, which implies that studies evaluating different CFGs can be pooled together. In this report, the PMU value of NFERT x ALL_CFGs was used for NFERT x LEGUME_C3 and NFERT x COTTON. Table 20 also lists the number of sites (Num of sites) with reported measurement uncertainty, percent of observation pairs and percent of sites in the full dataset for the corresponding PC/CFG combination to show the degree of data coverage.

Table 20. Pooled measurement uncertainty values of SOC changes in response to practice changes for each PC/CFG/SOC combination validated and their respective coverage.

PC x CFG	Available measurement uncertainty			PMU	Coverage** (%)	
	Num of Obs	Num of sites	Citations*	(g C m^{-2})	Obs	Sites
CROPP x NON_LEGUME_C4	10	3	a; b	238.3	5	25
CROPP x LEGUME_C3	12	3	a; b	223.2	4	20
CROPP x NON_LEGUME_C3	14	2	e	181.8	5	12
CROPP x COTTON	6	1	e	240.3	4	17
DISTU x NON_LEGUME_C4	6	2	a	423.1	4	18
DISTU x LEGUME_C3	4	2	a	488.9	6	29
DISTU x NON_LEGUME_C3	5	1	d	87.6	5	17
DISTU x COTTON	6	1	e	209.9	12	25
NFERT x NON_LEGUME_C4	5	2	b; c	410.5	3	22
NFERT x NON_LEGUME_C3	16	2	b; d	178.5	8	20
NFERT x ALL_CFGs	19	3	b; c; d	243.3	7	23
ALL PCs x ALL_CFGs	59	6	a; b; c; d; e	231.44	7	20

*: a: Collins et al., 1999; b: Wolf et al., 2018 (with measurement uncertainty from the supplementary zip file[10]); c: Jin et al., 2015; Jin and Varvel, 2018; d: Ghimire et al., 2015; and e: Senapati et al., 2014

**: percentage of the observation pairs with uncertainty available over the observation pairs in the full dataset as listed in Table 5 for the corresponding PC x CFG category; and percentage of the sites with uncertainty available for at least one observation pairs over the total number of sites in the full dataset as listed in Table 4 for the corresponding PC x CFG category.

[10] https://esajournals.onlinelibrary.wiley.com/doi/full/10.1002/ecy.2105#support-information-section

8.5 Model bias across all categories

Each of the studies was validated in exactly one of the 5 k-folds (Figure 3), therefore, only one model bias could be calculated for one study. All values of study bias for each study in the validation dataset are ranked from the highest to the lowest in Table 21 (per Box 4, Summary of requirements in VMD53 Model requirements Section 5.2.4).

As listed in Table 20, the PMU for the entire validation dataset is 231.4 g C m^{-2}. The absolute mean model bias across all studies and PC/CFG combinations is 14.7 g C m^{-2} (Table 21), which is less than the PMU of 231.4 g C m^{-2}.

Table 21. Model bias for each study site across all PCs and CFGs.

Fold	Study	N treatment pairs	Model bias (g C m^{-2})
1	Rosemount	24	445.0
1	Broadbalk	10	384.0
1	Ithaca2	3	351.9
1	DixonSprings	24	100.6
1	NarrabriField6	50	-20.6
1	NarrabriFieldD1	42	-259.1
2	SCharleston	6	155.2
2	Mandan	70	-24.5
2	Dalhart	7	-141.1
2	FortCollins	7	-224.3
2	RussellR	14	-292.6
2	Sidney	3	-330.8
3	Rodale	2	318.9
3	NarrabriFieldC1	14	156.8
3	Lexington	32	11.0
3	Mead2	117	-4.2
3	FortValley	126	-113.4
3	Goias	4	-289.4
4	IthacaNE	9	372.2
4	Davis	10	179.4
4	Hoytville	5	-12.6
4	SwiftCurrent	42	-108.2
4	Brookings	14	-132.4
4	Saginaw	30	-141.6
5	Tribune	36	2.9
5	ORpegn	129	-7.1
5	FivePoints	12	-115.2
5	Wooster	13	-270.5

5	LethbridgeABC	18	-415.8
1	All studies	153	167.0
2	All studies	107	-143.0
3	All studies	295	13.3
4	All studies	110	26.2
5	All studies	210	-8.8
All dataset	Across all folds and studies	873	-14.7

Note: the full data contains 566 treatment observations, fold assignment is saved in k_fold_con.csv (model_validation_supporting_materials.7z).

8.6 Bias for CROPP x NON_LEGUME_C4

All values of model bias for each study in this validation dataset are ranked from the highest to the lowest (Table 22).

Table 22. Model bias for each study site in the CROPP x NON_LEGUME_C4 validation dataset.

Study	N_treatment_pairs	Model bias (g C m^{-2})
Hoytville	2	401.7
Davis	10	179.4
Brookings	6	53.3
Tribune	36	2.9
Mead2	54	-60.4
RussellR	4	-120.1
FortValley	54	-136.9
Saginaw	30	-141.6
Dalhart	3	-207.3
Goias	3	-308.6
Rodale	2	-318.9
Wooster	6	-736.8
Across all studies	210	-116.1

Note: Each of the studies was validated in exactly one of the 5 k-folds (Figure 3), model bias was calculated in that fold for this PC/CFG/SOC validation pairs.

Mean bias across all studies in this category is -116.1 g C m^{-2}.

Yes, the absolute bias value is smaller than the PMU of 238.3 g C m^{-2} (Table 20).

8.7 Bias for CROPP x LEGUME_C3

All values of model bias for each study in this validation dataset were ranked from the highest to the lowest (Table 23).

Table 23. Model bias for each study site in the CROPP x LEGUME_C3 validation dataset.

Study	**N_treatment_pairs**	**Model bias (g C m^{-2})**
Hoytville	2	401.7
Davis	10	179.4
Brookings	6	53.3
NarrabriField6	45	-20.4
SwiftCurrent	12	-43.8
Tribune	18	-45.3
Mead2	45	-72.7
FortValley	45	-120.0
Saginaw	28	-163.7
RussellR	6	-170.1
FivePoints	6	-261.1
Goias	4	-289.4
Rodale	2	-318.9
NarrabriFieldD1	35	-374.4
Wooster	6	-736.8
Across all studies	270	-132.1

Note: Each of the studies was validated in exactly one of the 5 k-folds (Figure 3), model bias was calculated in that fold for this PC/CFG/SOC validation pairs.

Mean bias across all studies in this category is -132.1 g C m^{-2}.

Yes, the absolute bias value is smaller than the PMU of 223.2 g C m^{-2} (Table 20).

8.8 Bias for CROPP x NON_LEGUME_C3

All values of model bias for each study in this validation dataset were ranked from the highest to the lowest (Table 24).

Table 24. Model bias for each study site in the CROPP x NON_LEGUME_C3 validation dataset.

Study	N_treatment_pairs	Model bias (g C m^{-2})
Davis	10	179.4
NarrabriFieldC1	7	122.5
Brookings	6	53.3
Tribune	36	2.9
NarrabriField6	35	-45.6
SwiftCurrent	33	-121.8
FortValley	45	-137.5
Saginaw	30	-141.6
Mead2	18	-142.4
NarrabriFieldD1	35	-201.3
Dalhart	3	-207.3
RussellR	8	-210.1
FivePoints	6	-261.1
Rodale	2	-318.9
LethbridgeABC	18	-415.8
Across all studies	292	-123.0

Note: Each of the studies was validated in exactly one of the 5 k-folds (Figure 3), model bias was calculated in that fold for this PC/CFG/SOC validation pairs.

Mean bias across all studies in this category is -123.0 g C m^{-2}.

Yes, the absolute bias value is smaller than the PMU of 181.8 g C m^{-2} (Table 20).

8.9 Bias for CROPP x COTTON

All values of model bias for each study in this validation dataset were ranked from the highest to the lowest (Table 25).

Table 25. Model bias for each study site in the CROPP x COTTON validation dataset.

Study	N_treatment_pairs	Model bias (g C m^{-2})
NarrabriFieldC1	7	122.5
NarrabriField6	50	-20.6
FortValley	54	-136.9
NarrabriFieldD1	42	-259.1
FivePoints	6	-261.1
Goias	4	-289.4
Across all studies	163	-140.8

Note: Each of the studies was validated in exactly one of the 5 k-folds (Figure 3), model bias was calculated in that fold for this PC/CFG/SOC validation pairs.

Mean bias across all studies in this category is -140.8 g C m^{-2}.

Yes, the absolute bias value is smaller than the PMU of 240.3 g C m^{-2} (Table 20).

8.10 Bias for DISTU x NON_LEGUME_C4

All values of model bias for each study in this validation dataset were ranked from the highest to the lowest (Table 26).

Table 26. Model bias for each study site in the DISTU x NON_LEGUME_C4 validation dataset.

Study	N_treatment_pairs	Model bias (g C m^{-2})
Rosemount	18	480.3
IthacaNE	9	372.2
SCharleston	6	155.2
Wooster	7	129.2
DixonSprings	24	100.6
FortValley	36	-134.0
Lexington	8	-202.9
FortCollins	7	-224.3
Brookings	8	-271.6
Hoytville	3	-288.7
Goias	1	-391.6
Across all studies	127	-25.1

Note: Each of the studies was validated in exactly one of the 5 k-folds (Figure 3), model bias was calculated in that fold for this PC/CFG/SOC validation pairs.

Mean bias across all studies in this category is -25.1 g C m^{-2}.

Yes, the absolute bias value is smaller than the PMU of 423.1 g C m^{-2} (Table 20).

8.11 Bias for DISTU x LEGUME_C3

All values of model bias for each study in this validation dataset were ranked from the highest to the lowest (Table 27).

Table 27. Model bias for each study site in the DISTU x LEGUME_C3 validation dataset.

Study	N_treatment_pairs	Model bias (g C m^{-2})
DixonSprings	24	100.6
Wooster	3	35.0
FivePoints	3	-92.1
FortValley	18	-118.8
Brookings	8	-271.6
Hoytville	3	-288.7
Goias	1	-391.6
Across all studies	60	-146.7

Note: Each of the studies was validated in exactly one of the 5 k-folds (Figure 3), model bias was calculated in that fold for this PC/CFG/SOC validation pairs.

Mean bias across all studies in this category is -146.7 g C m^{-2}.

Yes, the absolute bias value is smaller than the PMU of 488.9 g C m^{-2} (Table 20).

8.12 Bias for DISTU x NON_LEGUME_C3

All values of model bias for each study in this validation dataset were ranked from the highest to the lowest (Table 28).

Table 28. Model bias for each study site in the DISTU x NON_LEGUME_C3 validation dataset.

Study	N_treatment_pairs	Model bias (g C m^{-2})
Mandan	34	131.9
FivePoints	6	30.6
ORpegn	35	-89.9
Brookings	3	-90.7
FortValley	18	-159.1
Sidney	3	-330.8
Across all studies	99	-84.7

Note: Each of the studies was validated in exactly one of the 5 k-folds (Figure 3), model bias was calculated in that fold for this PC/CFG/SOC validation pairs.

Mean bias across all studies in this category is -84.7 g C m^{-2}.

Yes, the absolute bias value is smaller than the PMU of 87.6 g C m^{-2} (Table 20).

8.13 Bias for DISTU x COTTON

All values of model bias for each study in this validation dataset were ranked from the highest to the lowest (Table 29).

Table 29. Model bias for each study site in the DISTU x COTTON validation dataset.

Study	N_treatment_pairs	Model bias (g C m^{-2})
NarrabriFieldC1	7	191.1
FivePoints	6	30.6
FortValley	36	-134.0
Goias	1	-391.6
Across all studies	50	-76.0

Note: Each of the studies was validated in exactly one of the 5 k-folds (Figure 3), model bias was calculated in that fold for this PC/CFG/SOC validation pairs.

Mean bias across all studies in this category is -76.0 g C m^{-2}.

Yes, the absolute bias value is smaller than the PMU of 209.9 g C m^{-2} (Table 20).

8.14 Bias for NFERT x NON_LEGUME_C4

All values of model bias for each study in this validation dataset were ranked from the highest to the lowest (Table 30).

Table 30. Model bias for each study site in the NFERT x NON_LEGUME_C4 validation dataset.

Study	N_treatment_pairs	Model bias (g C m^{-2})
Ithaca2	3	351.9
Rosemount	6	339.1
Rodale	2	318.9
Davis	8	99.1
Lexington	24	82.3
Mead2	54	35.6
FortValley	36	-57.5
Dalhart	4	-91.4
RussellR	2	-833.3
Across all studies	139	27.2

Mean bias across all studies in this category is 27.2 g C m^{-2}.

Yes, the absolute bias value is smaller than the PMU of 410.5 g C m^{-2} (Table 20).

8.15 Bias for NFERT x NON_LEGUME_C3

All values of model bias for each study in this validation dataset were ranked from the highest to the lowest (Table 31).

Table 31. Model bias for each study site in the NFERT x NON_LEGUME_C3 validation dataset.

Study	N_treatment_pairs	Model bias (g C m^{-2})
Broadbalk	10	384.0
Rodale	2	318.9
Davis	8	99.1
ORpegn	74	-12.3
SwiftCurrent	9	-58.6
FortValley	18	-67.1
Dalhart	4	-91.4
Mead2	18	-111.1
Mandan	36	-172.2
RussellR	6	-402.6
Across all studies	185	-11.3

Note: Each of the studies was validated in exactly one of the 5 k-folds (Figure 3), model bias was calculated in that fold for this PC/CFG/SOC validation pairs.

Mean bias across all studies in this category is -11.3 g C m^{-2}.
Yes, the absolute bias value is smaller than the PMU of 178.5 g C m^{-2} (Table 20).

8.16 Bias for NFERT x ALL_CFGs

As indicated in Section 8.4, due to limited availability of reported measurement uncertainty (SD or SE), we utilized the NFERT x ALL_CFGs' PMU value of 243.3 g C m^{-2} (Table 20) for the model bias comparison for the NFERT x LEGUME_C3 combination.

Table 32. Model bias for each study site in the NFERT x ALL_CFGs validation dataset.

Study	N_treatment_pairs	Model bias (g C m^{-2})
Broadbalk	10	384.0
Ithaca2	3	351.9
Rosemount	6	339.1
Rodale	2	318.9
Davis	8	99.1
Lexington	24	82.3
Mead2	63	43.9
ORpegn	74	-12.3
FortValley	36	-57.5
SwiftCurrent	9	-58.6
Dalhart	4	-91.4
Mandan	36	-172.2
RussellR	6	-402.6
NFERT x ALL CFGs	281	63.4

Note: Each of the studies was validated in exactly one of the 5 k-folds (Figure 3), model bias was calculated in that fold for this PC/CFG/SOC validation pairs.

Mean bias across all studies in this category is 63.4 g C m^{-2}.

Yes, the bias value is smaller than the PMU of 243.3 g C m^{-2} (Table 20).

As explained in Section 5.3, the validation of the NFERT × COTTON is considered valid based on two essential criteria:

1) Successful validation of NFERT for another annual CFG: This criterion has been met, as demonstrated in Sections 8.14 and 8.15, where the absolute model biases for NFERT × NON_LEGUME_C4 (27.2 g C m^{-2}) and NFERT × NON_LEGUME_C3 (-11.3 g C m^{-2}) were found to be smaller than their respective PMU values (410.5 and 178.5 g C m^{-2}, respectively).
2) Successful validation of the COTTON CFG for the Cropping, Planting, and Harvesting PC (CROPP × COTTON): This condition has also been satisfied, as evidenced in Section 8.9, where the absolute model bias of -140.8 g C m^{-2} for CROPP × COTTON was smaller than the PMU value of 240.3 g C m^{-2} associated with its validation dataset.

Moreover, it's worth noting that the model bias for NFERT x ALL_CFGs, which encompasses the entire NFERT PC validation dataset (including one site with NFERT x COTTON), is smaller than the corresponding PMU. This comprehensive assessment of model biases collectively supports the validation of NFERT x COTTON.

Hence, the calculated absolute mean model bias is smaller than the PMU for all individual PC and CFG combinations, as well as for the entire dataset (with model bias of 14.7 < PMU of 231.4 g C m^{-2}). The calibrated DayCent model has passed the model bias test for the 12 PC/CFG/SOC combinations for this cropland SOC reporting.

9. Evaluation of Model Prediction Error

This section follows *VMD53 Model Requirements, Section 5.2.5 Using Data to Evaluation Model Prediction Error (p27)*.

9.1 Description of methods

We evaluated the accuracy of our model predictions using both graphical and statistical methods. For each combination of PC/CFG/SOC, we presented several graphical comparisons of measured versus modeled results. These included plots showing the 90% prediction intervals, which demonstrated that the measured values were contained within the intervals, scatterplots of the model predictions versus measurements, and histograms of residuals.

The bounds of model uncertainty associated with SOC stock changes were estimated for each k-fold iteration through joint probability draws from parameter distributions and hyperparameters. This method assigns uncertainties by propagating errors in DayCent calibration parameters and hyperparameters, which are influenced by uncertainties in model structure and measurement accuracy. To establish 90% posterior prediction intervals, we derived them from the posterior predictive distribution of SOC stocks by capturing the 5th and 95th percentiles. These prediction intervals comprehensively consider uncertainties in DayCent calibration parameters and errors in DayCent predictions. These errors stem from random effects, including variance across sites, variations across years within the same site, and unexplained errors in model predictions based on hyperparameters. The performance metric to gauge acceptable model uncertainty is whether 90% of measured SOC stock differences from out-of-sample validation data fall within the 90% posterior prediction interval. We formed the 90% posterior prediction intervals for each treatment pair to estimate differences in SOC stock. Coverage rates were then calculated as the proportion of posterior prediction intervals that encompassed the observed differences in SOC. In k-fold cross-validation, the model is tested on data independent of the training data, providing a more accurate estimate of its performance on new data. This process enhances the model's generalization capabilities and its ability to perform well when faced with new data.

In this validation report we utilized weakly informative independent priors with a uniform distribution defined as outlined in Table A1 and **Section 4.3 "Documentation of internal model parameters"**. These uniform distributions are sufficiently broad to extend beyond our current understanding of the parameters' range. For combinations of PC and CFG with little validation data or highly variable observations, the approach provides a conservative prediction error estimate.

Statistical measures of MSE and RMSE provide quantitative measures of the accuracy of the model predictions. The MSE and RMSE were calculated for each MCMC iteration as below:

$$MSE_{mc} = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2; \; RMSE_{mc} = \sqrt{MSE_{mc}} \tag{9}$$

where *mc* is the number of MCMC iteration. The corresponding mean $\pm\, SD$ ($\overline{MSE} \pm SD_{MSE}$, and $\overline{RMSE} \pm SD_{RMSE}$) were reported for each PC/CFG/SOC combination in the following sections.

9.2 Model prediction error across all categories

Model uncertainties for 90% prediction intervals were computed for each fold and subsequently compared to the observed effects of SOC changes (Figures 16). The coverage rates of posterior predictive intervals for every fold during the validation process consistently exceeded 90% (Table 33). The rates of SOC stock change and their corresponding residuals, measured in t C ha^{-1} yr^{-1}, were also assessed, with the respective values available in Figure 17 and Figure 18b. The per year SOC stock changes involve comparing pairwise treatment differences over the years elapsed. Since the model was calibrated with unified parameter sets for all treatments. These generalized models had no specific calibration for individual treatments. Consequently, to account for treatment differences effectively, it is crucial for the model to accurately capture the interactions between management practices and carbon sequestration processes within its underlying biogeochemical representations.

When comparing various pairwise combinations of cropping practices (including continuous cropping, rotations, and cover cropping), different rates of inorganic N fertilization, and various levels of soil disturbance (such as no-till, reduced till, chisel tillage, and conventional tillage), we observed that the model displayed a conservative tendency, resulting in underpredictions of SOC stock changes, where the model's central prediction (mean prediction) fell below the observed value. This is exemplified by the residue, which represents the difference between the mean of posterior predictions and the observed change rate, amounting to -0.1 t C ha^{-1} yr^{-1} (Figure 18b).

In the context of carbon crediting markets, the most critical metric is quantifying the change in response to management practices. While underestimation is preferable to avoid over-crediting in carbon-crediting programs, the underestimation of SOC change rates raises several implications and considerations. Firstly, the model may not fully capture the complex interactions between management practices and carbon cycling processes, including decomposition rates (which rely on decomposition parameters derived from the multi-site, multi-treatment SOC dataset used for model parameterization), vegetation growth, and carbon inputs. This lack of accuracy might not adequately represent real-world conditions. As scientific knowledge advances and our understanding of plant litter decomposition, soil organic matter formation, and microbial dynamics improves, the process-based representation of SOM dynamics can be enhanced. This, in turn, holds the potential to reduce uncertainties in model structure and parameters. Secondly, measurement errors, as indicated by the PMU (Table 20), introduce an additional layer of complexity and noise during model calibration. Additionally, variations in soil sampling locations for pairwise treatment plots over the measurement period can introduce uncontrolled variability that models cannot account for. In light of measurement errors and the variability in sampling locations, the implementation of standardized soil sampling protocols becomes crucial. These protocols ensure the uniformity of data across diverse locations and time intervals, thereby mitigating the influence of uncontrolled variability. Future improvements will hinge on a blend of enhanced measurement techniques, adherence to standardized protocols, and the refinement of modeling processes.

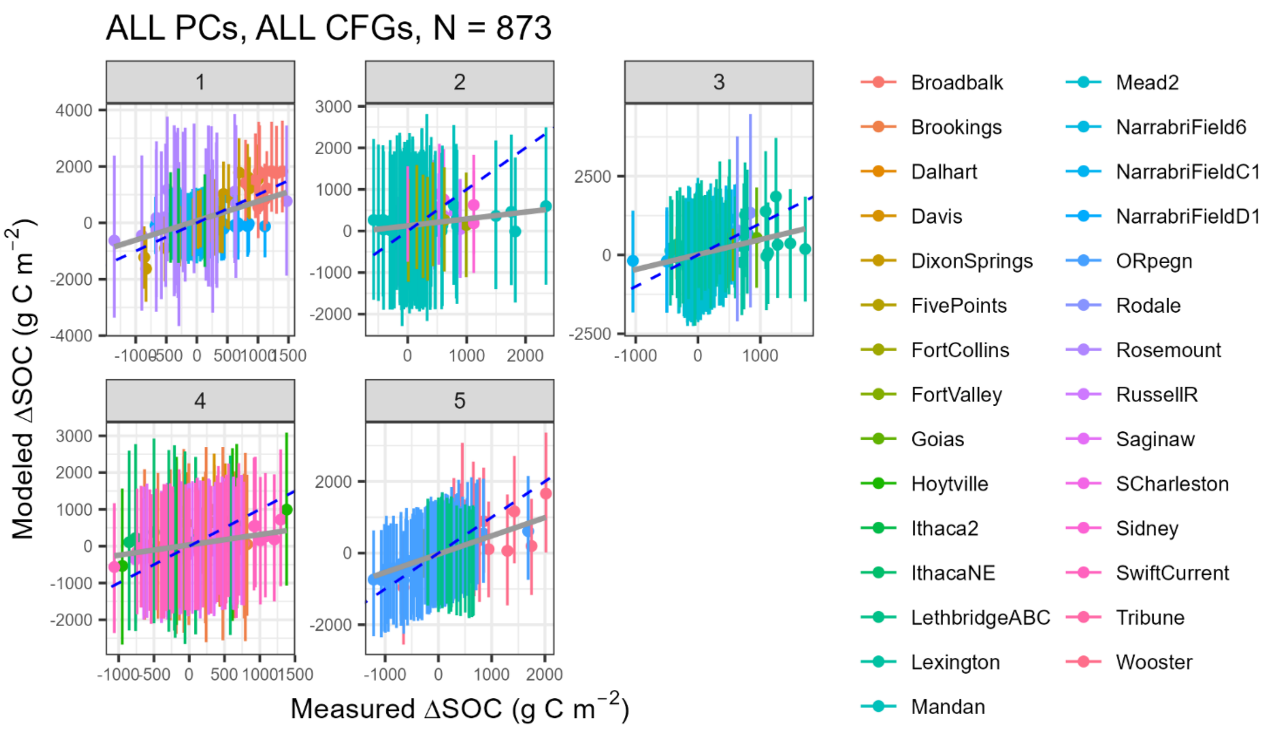

Figure 16. Modeled verse measured SOC change in all practice changes and crop types, faceted by fold. Error bars show 90% prediction intervals.

Table 33. Number of observed SOC stock changes falling inside and outside of 90% prediction intervals for each fold across all practices and crop types.

Fold	n_pair	n_in	n_out	% coverage
1	153	150	3	98
2	107	106	1	99
3	295	295	0	100
4	110	110	0	100
5	208	207	1	100
All	873	868	5	99

The MSE and RMSE for each MCMC iteration were calculated, with the means and corresponding standard deviations as below:

MSE ± SD = 15.89 ± 1.18 (t C ha^{-1})^2; RMSE ± SD = 3.98 ± 0.14 t C ha^{-1}.

Do prediction intervals cover observed data 90% of the time? Yes.

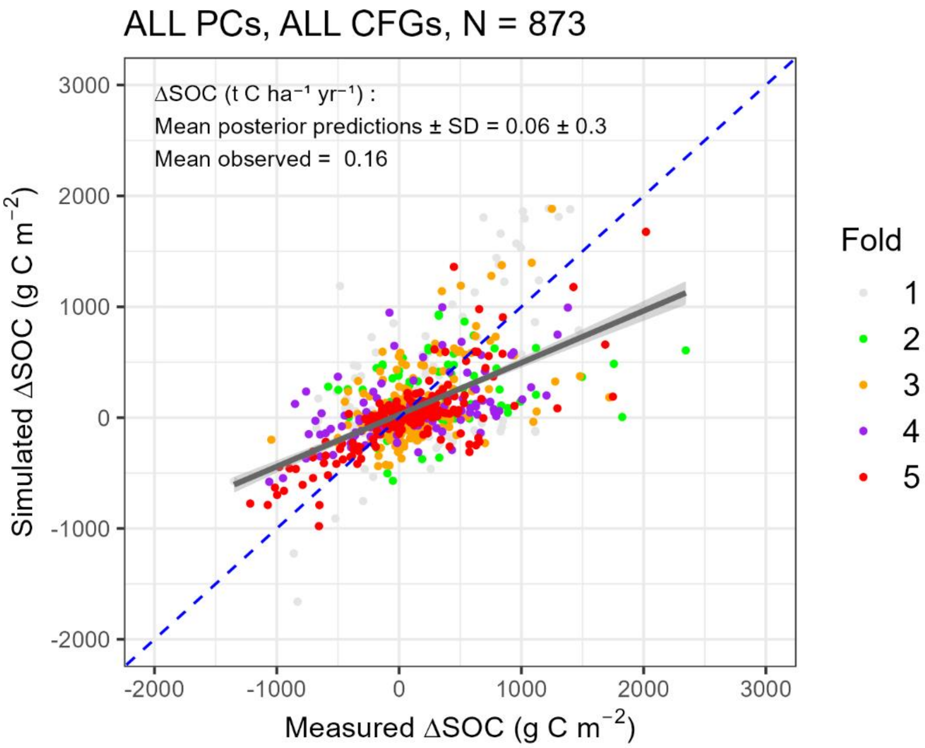

Figure 17. Scatterplot of <u>mean posterior predictions</u> verse measurements of SOC stock changes in all practice changes and crop types in the validation dataset (experiment durations ranging from 3 to 162 years). Shaded area is the 95% confidence interval around linear least-squares fit (a measure of the uncertainty associated with the estimated regression coefficients) and dashed line shows 1:1 relationship.

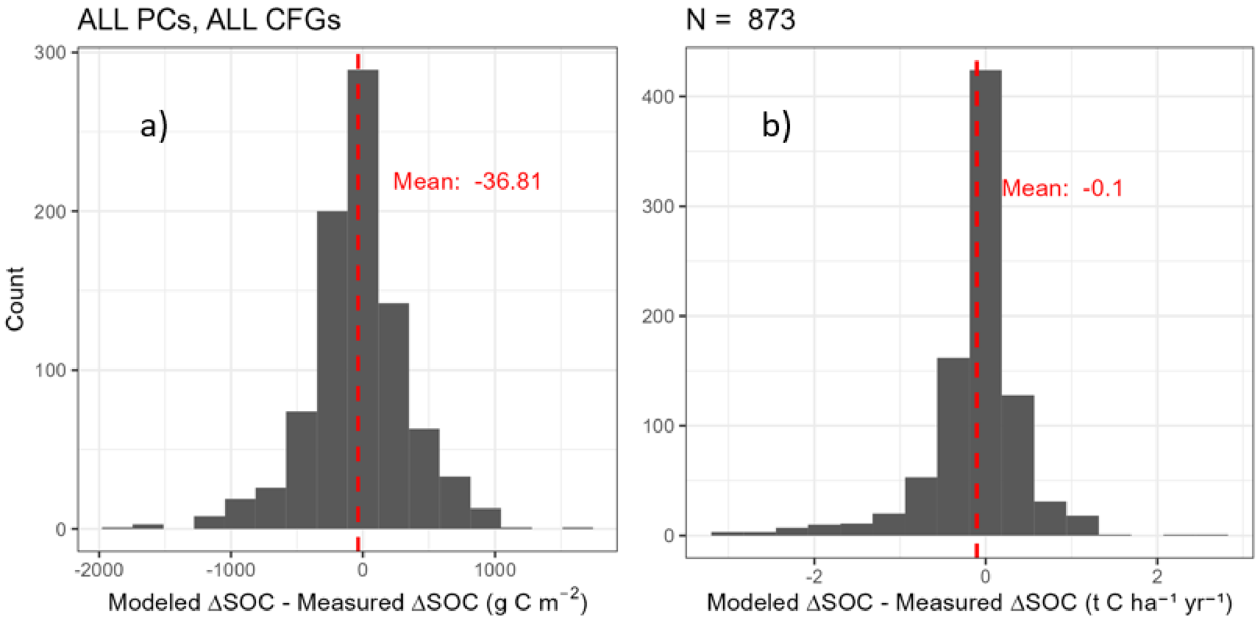

Figure 18. Histogram of model residuals (simulated - observed) for change in SOC in all studies used for model validation across all practices and crop types.

While the datasets used to validate model performance and uncertainty for each declared PC/CFG combination share observation points at the CFG level (Section 7), it's important to note that the paired PC is comprised of individual comparable practices rather than multiple practices combined. For example, let's examine the comparison between "Goias_CTCSM_2014" and "Goias_CTCM_2014," which both denote measurements of SOC stocks in the year 2014. These measurements were taken at the Goias site, where two different cropping practices were employed: a cotton-soybean-corn rotation with conventional tillage for "Goias_CTCSM_2014" and continuous cotton with conventional tillage for "Goias_CTCM_2014." In this particular case, the only comparable PC is the "cropping rotation" category, which compares "cropping rotation" to "single cropping," both under the same conventional tillage conditions. It's worth highlighting that this specific pair can be utilized for validating all three of the CFGs associated with NON_LEGUME_C4 (e.g., corn), LEGUME_C3 (e.g., soy), and COTTON, as long as the experimental data cover at least one complete rotation cycle, as outlined in VMD53 Section 5.2.3. Additionally, at the same Goias site, there was also a soybean-corn-cotton rotation with no-till, labeled as "Goias_NTS_2014." We won't pair "Goias_NTS_2014" with "Goias_CTCM_2014" due to the distinct differences in both tillage practices (DISTU PC) and cropping systems (CROPP PC) between the two treatments. However, this pairing between "Goias_NTS_2014" and "Goias_CTCSM_2014" had the same crop rotation and was utilized in the DISTU PC to validate all three CFGs, just as mentioned earlier. The similarities in observation/model performance between PC/CFG combinations presented here can be partly attributed to the shared observation points, particularly driven by the popularity of rotations involving corn and soybean. For access to the complete dataset, including treatment observations and fold assignment details, please refer to "k_fold_con.csv," and additional information about treatment descriptions can be found in "StudySite_Treatments.xlsx" within the "model_validation_supporting_materials.7z" archive.

CROPP

Crop rotation and cover cropping foster carbon sequestration in soil by bolstering organic matter inputs, stimulating microbial activity, ensuring a steady supply of carbon through root exudates, and enhancing soil structure. Both observed and modeled SOC stocks exhibit an overall positive impact. Notably, the modeling approach maintains a conservative stance, often under-predicting SOC changes when compared to counterfactual treatments. Among the cropping PC/CFGs, DayCent model performs less effectively for COTTON compared to other CFGs, such as NON_LEGUME_C4 (e.g., corn), LEGUME_C3 (e.g., soy), and NON_LEGUME_C3 (e.g., wheat). This discrepancy can be attributed, in part, to the limited availability of data for COTTON, with only six experiments available for analysis compared to 12 to 16 sites for the other three PC/CFG combinations. Furthermore, only two of the six experiments are situated within the United States. One significant challenge in accurately simulating cotton growth lies in the inherent differences in growth patterns and responses to environmental conditions that distinguish cotton from the other crops. These distinctions are not comprehensively addressed by the model's default parameters, which were primarily calibrated for the more extensively studied crops. To enhance the model's capacity to simulate cotton effectively, it is imperative to consider specific adaptations and adjustments tailored to cotton's growth characteristics. It is also important to acknowledge that cotton is highly susceptible to various pests and diseases, which can have a significant impact on its growth, yield, and, consequently, its biomass contribution. Unfortunately, the DayCent model does not currently incorporate mechanisms to account for these biotic stressors or their intricate interactions with environmental factors. Mitigating this constraint may necessitate the development of supplementary modules or components capable of more accurately representing the dynamics of pest and disease infestations, alongside efforts in data acquisition and parameterization adjustments.

9.3 CROPP x NON_LEGUME_C4

Model uncertainties for 90% prediction intervals were computed for each fold and subsequently compared to the observed effects of SOC changes (Figures 19). The coverage rates of posterior predictive intervals for every fold during the validation process consistently exceeded 90% (Table 34). Although VMD53's protocol doesn't mandate it, we've also included a yearly rate of change comparison. The model's mean prediction fell below the observed mean value (see Figure 20), resulting in a residual change rate of -0.13 t C ha^{-1} yr^{-1} (Figure 21b). As discussed in Section 9.2, this conservative tendency can prevent over-crediting in carbon crediting markets. However, it also underscores the potential for improvement in both the model's robust representation and SOC measurement protocols.

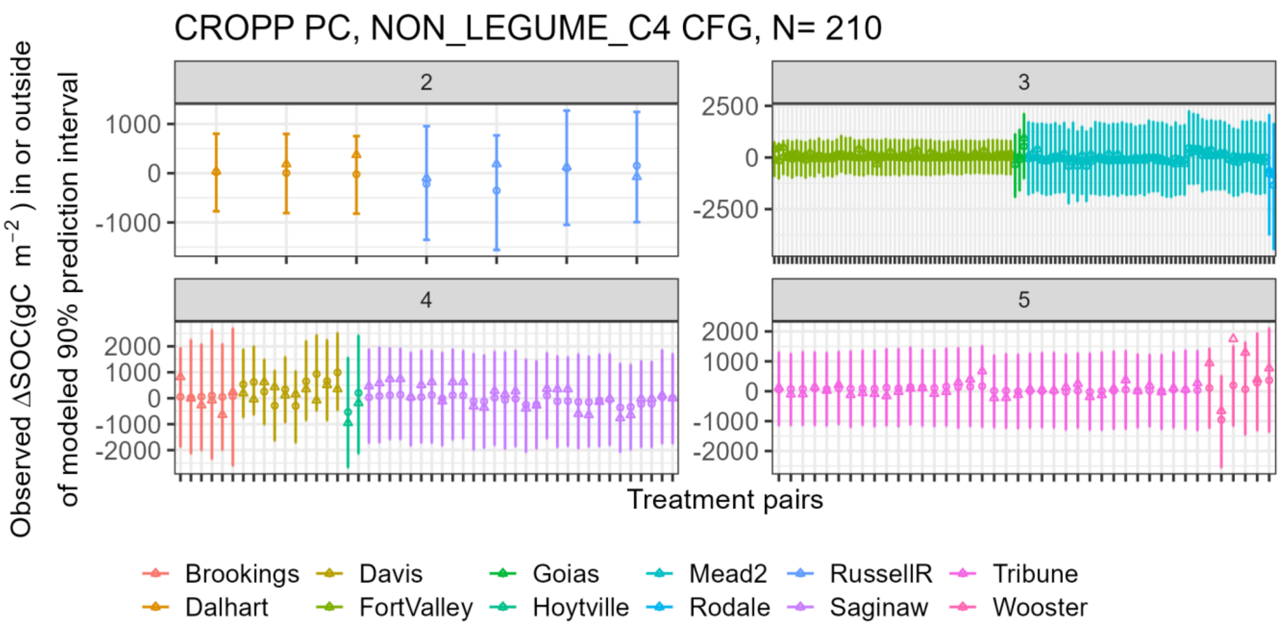

Figure 19. Modeled (dot) and measured SOC change (triangle) in response to changed cropping practices involving the NON_LEGUME_C4-type, faceted by fold, with the error bars showing 90% prediction intervals.

Table 34. Number of observed SOC stock changes falling inside and outside of 90% prediction intervals in each fold for SOC predictions in response to changed cropping practice involving NON_LEGUME_C4 – type CFG.

Fold	n_pair	n_in	n_out	% coverage
2	7	7	0	100
3	113	113	0	100
4	48	48	0	100
5	42	41	1	98
All	210	209	1	100

The MSE and RMSE for each MCMC iteration were calculated, with the means and corresponding standard deviations as:

MSE ± SD = 10.22 ± 0.83 (t C ha^{-1})^2; RMSE ± SD = 3.19 ± 0.12 t C ha^{-1}.
Do prediction intervals cover observed data 90% of the time? Yes.

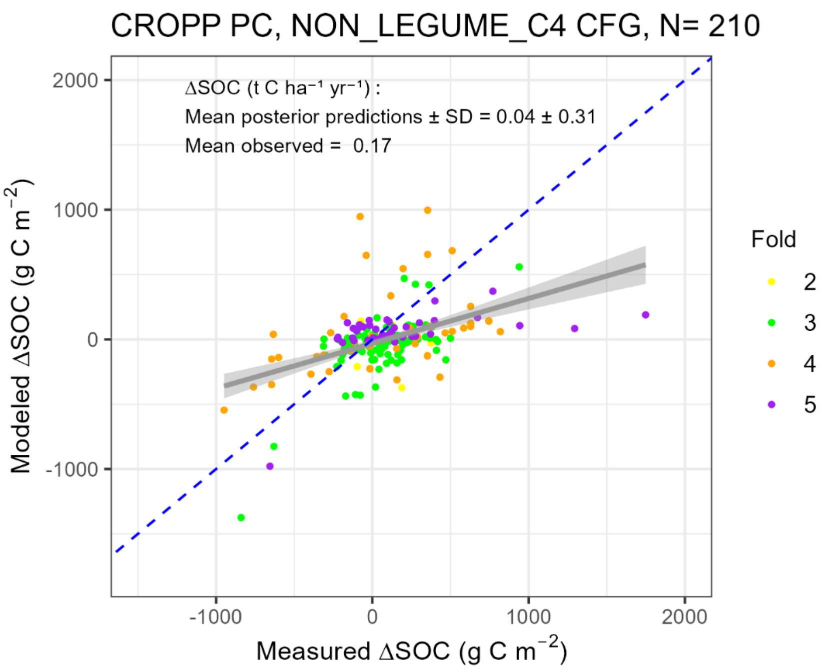

Figure 20. Scatterplot of mean posterior predictions verse measured SOC stock changes in response to changed cropping practices from the NON_LEGUME_C4 – type CFG (experiment durations ranging from 3 to 30 years). Shaded area is the 95% confidence interval around linear least-squares fit (a measure of the uncertainty associated with the estimated regression coefficients) and dashed line shows 1:1 relationship.

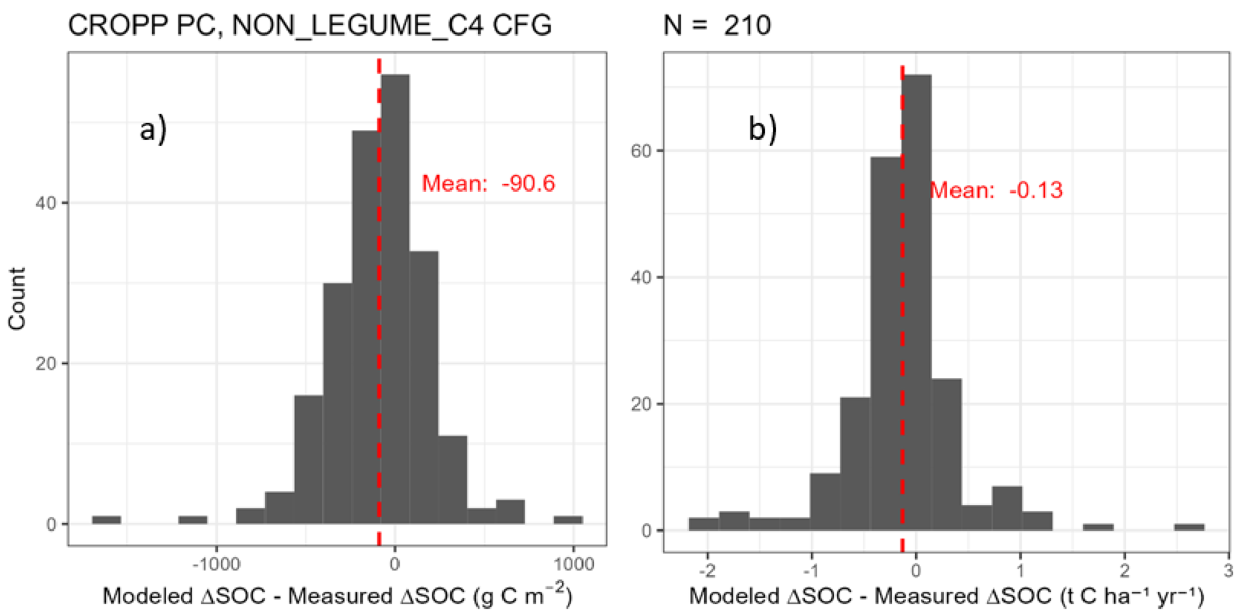

Figure 21. Histogram of model residuals (simulated - observed) for change in SOC in the validation dataset (12 sites with experiment durations from 3 to 30 years) used for evaluating changed cropping practices involving the NON_LEGUME_C4 – type CFG.

9.4 CROPP x LEGUME_C3

Model uncertainties for 90% prediction intervals were computed for each fold and subsequently compared to the observed effects of SOC changes (Figures 22). The coverage rates of posterior predictive intervals for every fold during the validation process consistently exceeded 90% (Table 35).

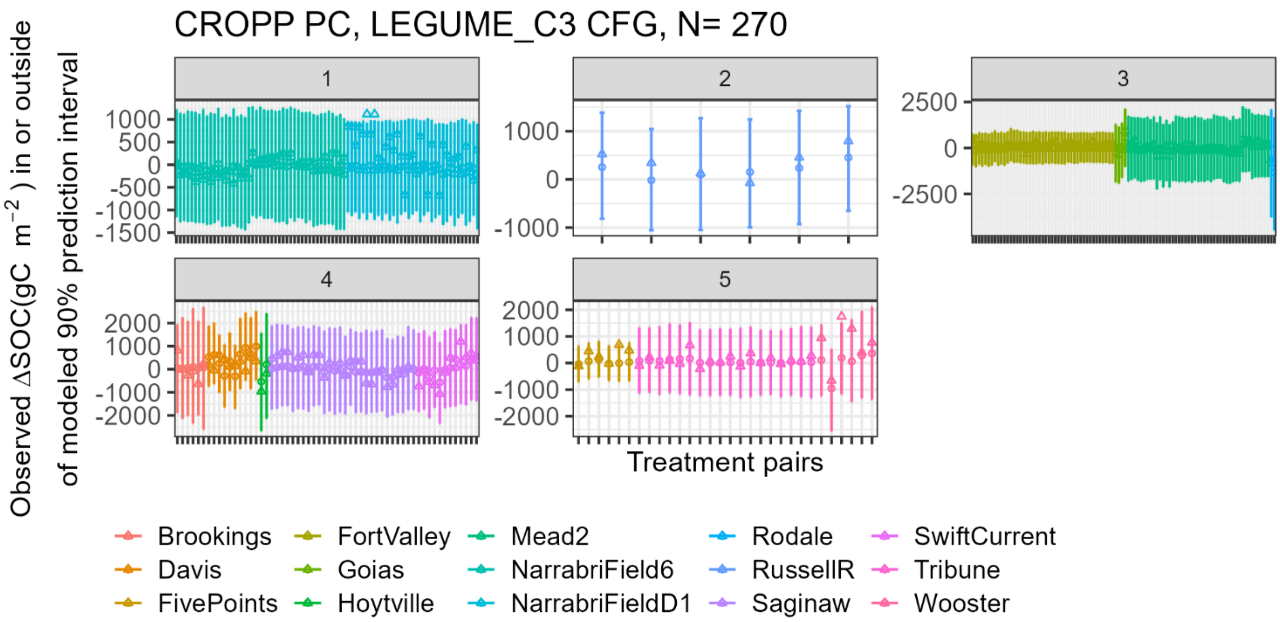

Figure 22. Modeled (dot) and measured SOC change (triangle) in response to changed cropping practices involving the LEGUME_C3-type, faceted by fold, with the error bars showing 90% prediction intervals.

Table 35. Number of observed SOC stock changes falling inside and outside of 90% prediction intervals in each fold for SOC predictions in response to changed cropping practice involving LEGUME_C3 – type CFG.

Fold	n_pair	n_in	n_out	% coverage
1	80	78	2	98
2	6	6	0	100
3	96	96	0	100
4	58	58	0	100
5	30	29	1	97
All	270	267	3	99

The MSE and RMSE for each MCMC iteration were calculated, with the means and corresponding standard deviations as:

MSE ± SD = 14.1 ± 1.16 (t C ha^{-1})^2; RMSE ± SD = 3.75 ± 0.15 t C ha^{-1}.
Do prediction intervals cover observed data 90% of the time? Yes.

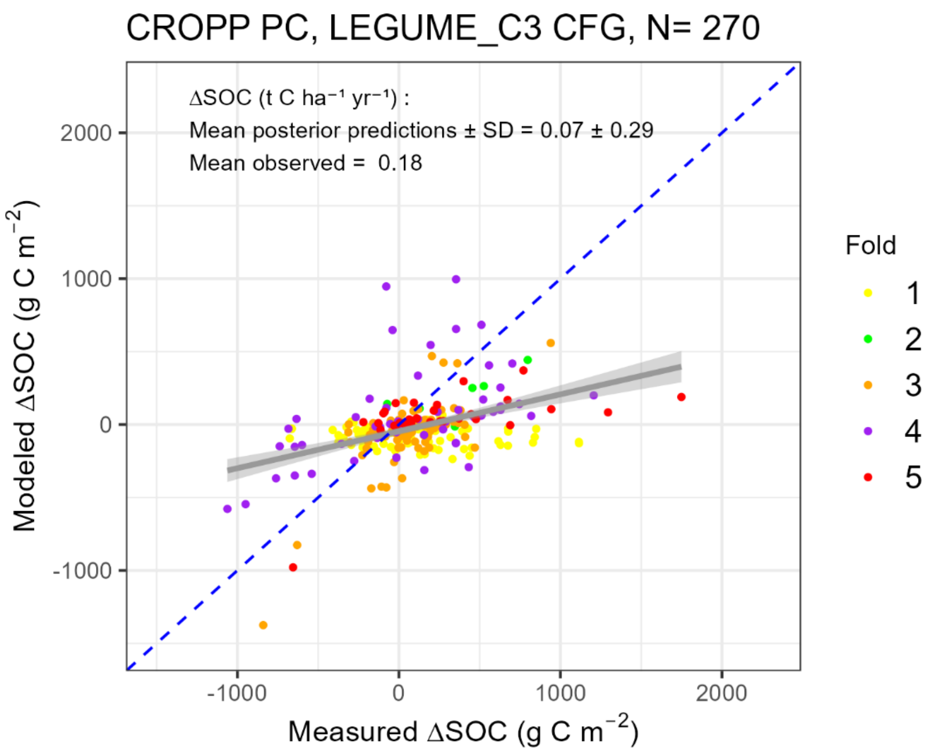

Figure 23. Scatterplot of mean posterior predictions verse measured SOC stock changes in response to changed cropping practices from the LEGUME_C3 – type CFG. Shaded area is the 95% confidence interval around linear least-squares fit (a measure of the uncertainty associated with the estimated regression) and dashed line shows 1:1 relationship.

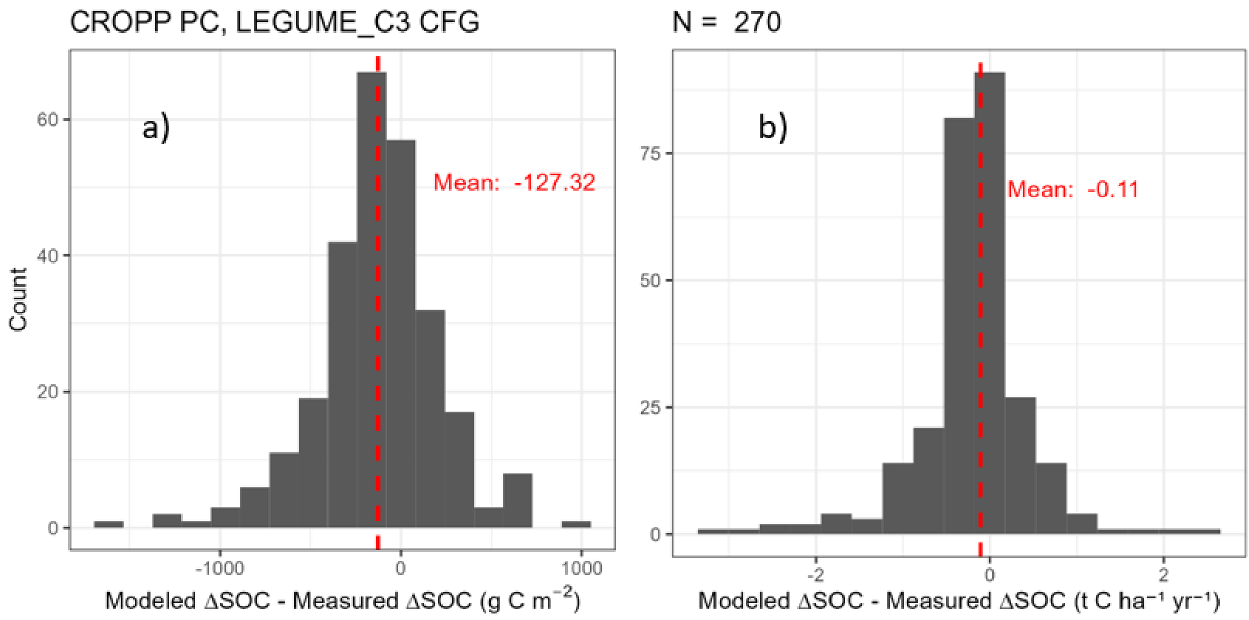

Figure 24. Histogram of model residuals (simulated - observed) for change in SOC in the validation dataset (15 sites with experiment durations from 3 to 43 years) used for evaluating changed cropping practices involving the LEGUME_C3 – type CFG.

9.5 CROPP x NON_LEGUME_C3

Model uncertainties for 90% prediction intervals were computed for each fold and subsequently compared to the observed effects of SOC changes (Figures 25). The coverage rates of posterior predictive intervals for every fold during the validation process consistently exceeded 90% (Table 36).

Table 36. Number of observed SOC stock changes falling inside and outside of 90% prediction intervals in each fold for SOC predictions in response to changed cropping practice involving NON_LEGUME_C3 – type CFG.

Fold	n_pair	n_in	n_out	% coverage
1	70	68	2	97
2	11	11	0	100
3	72	72	0	100
4	79	79	0	100
5	60	60	0	100
All	292	290	2	99

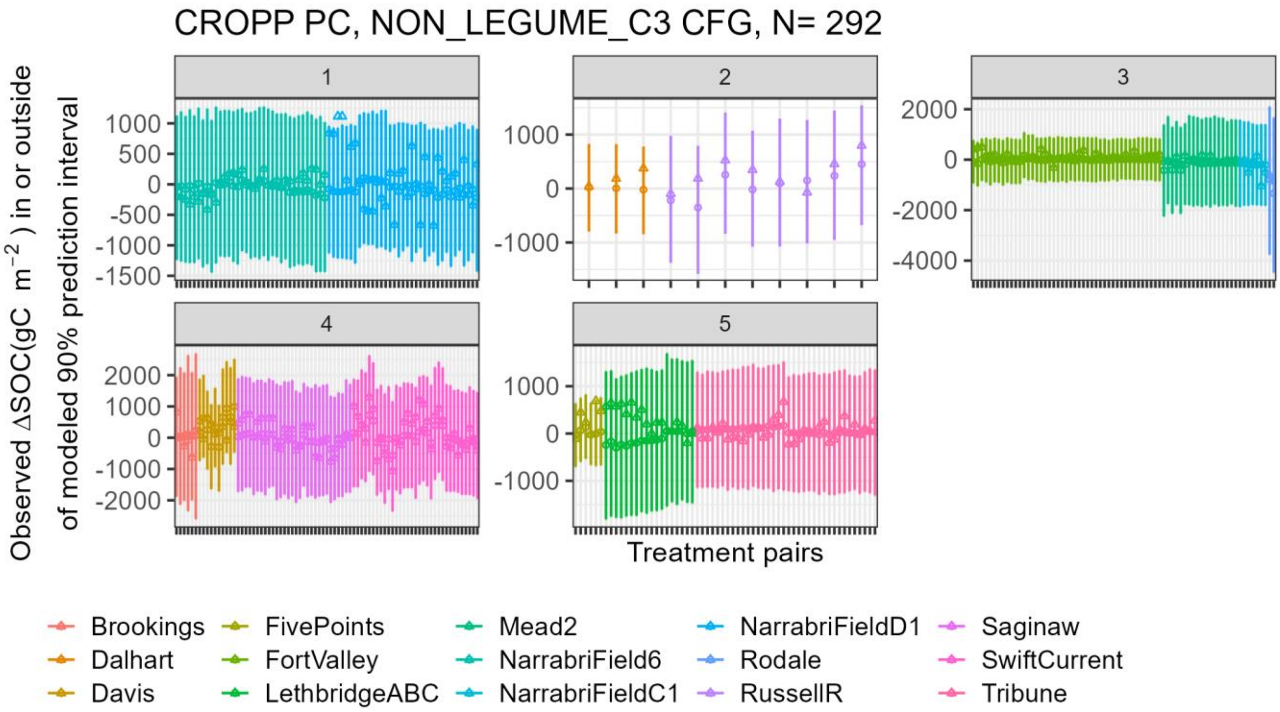

Figure 25. Modeled (dot) verse measured SOC change (triangle) in response to changed cropping practices involving the NON_LEGUME_C3-type, faceted by fold, with the error bars showing 90% prediction intervals.

The MSE and RMSE for each MCMC iteration were calculated, with the means and corresponding standard deviations as:

MSE ± SD = 14.4 ± 1.08 (t C $ha^{-1})^2$; RMSE ± SD = 3.79 ± 0.14 t C ha^{-1}.

Do prediction intervals cover observed data 90% of the time? Yes.

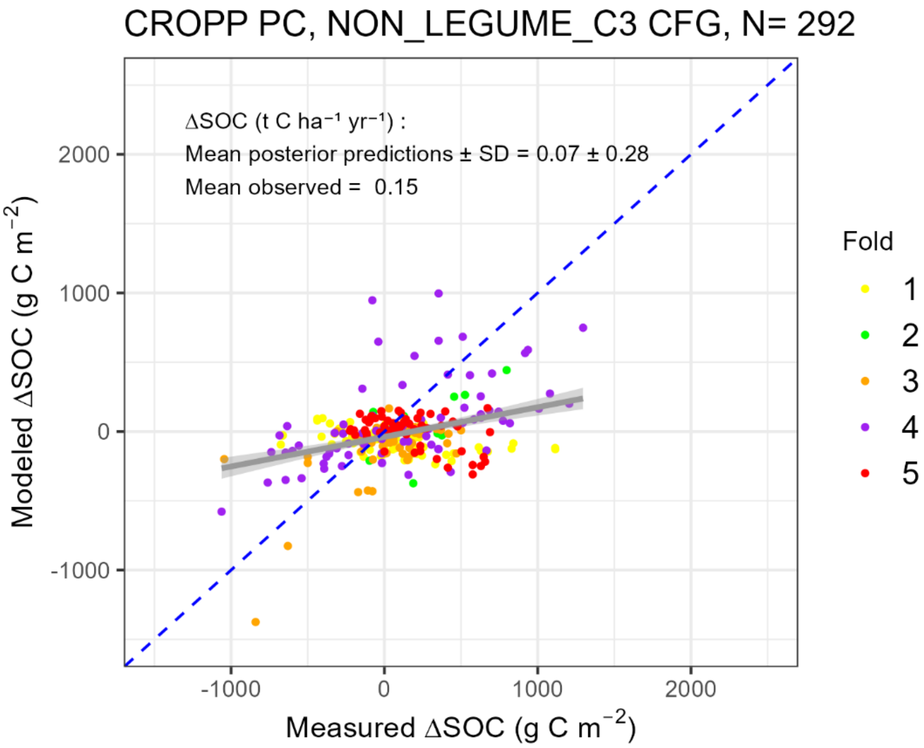

Figure 26. Scatterplot of mean posterior predictions verse measured SOC stock changes in response to changed cropping practices from the NON_LEGUME_C3 – type CFG. Shaded area is the 95% confidence interval around linear least-squares fit (a measure of the uncertainty associated with the estimated regression) and dashed line shows 1:1 relationship.

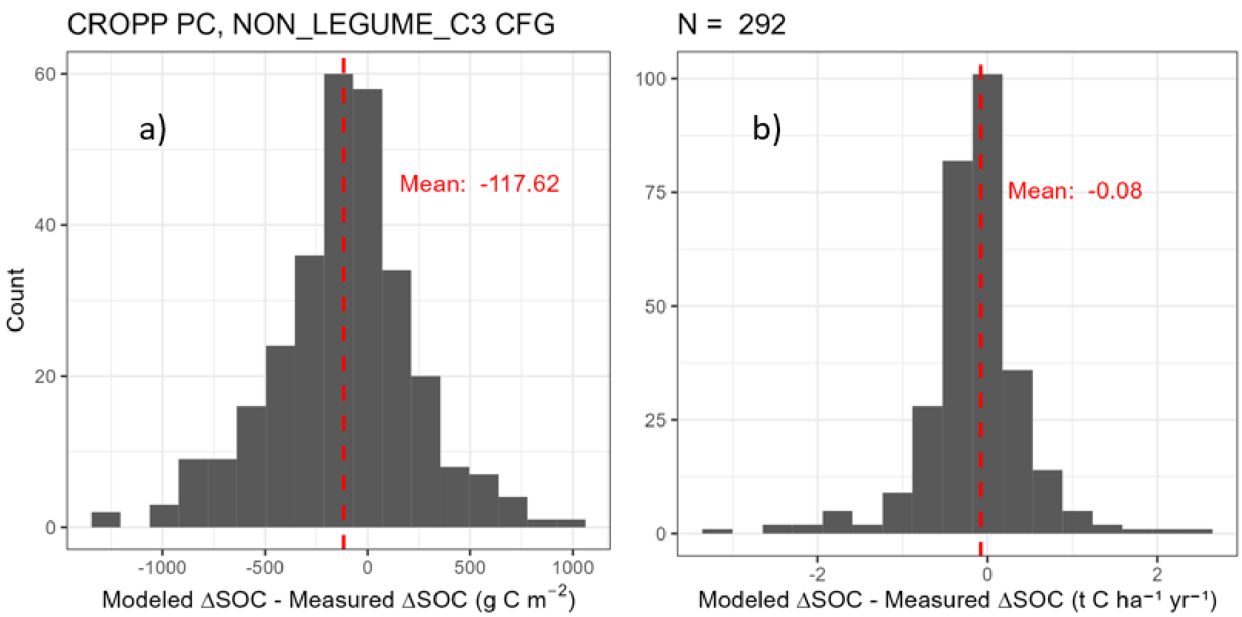

Figure 27. Histogram of model residuals (simulated - observed) for change in SOC in the validation dataset (16 sites with experiment durations from 3 to 79 years) used for evaluating changed cropping practices involving the NON_LEGUME_C3 – type CFG.

9.6 CROPP x COTTON

Model uncertainties for 90% prediction intervals were computed for each fold and subsequently compared to the observed effects of SOC changes (Figures 28). The coverage rates of posterior predictive intervals for every fold during the validation process consistently exceeded 90% (Table 37).

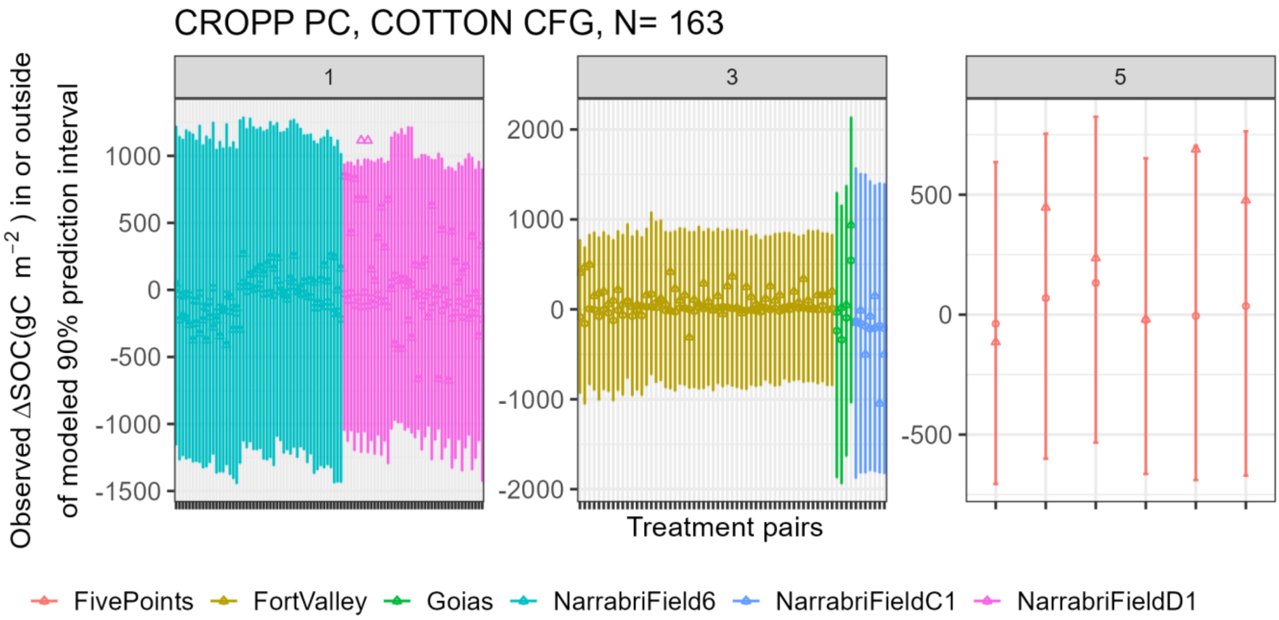

Figure 28. Modeled (dot) and measured SOC change (triangle) in response to changed cropping practices involving COTTON, with the error bars showing 90% prediction intervals.

Table 37. Number of observed SOC stock changes falling inside and outside of 90% prediction intervals in each fold for SOC predictions in response to changed cropping practice involving COTTON.

Fold	n_pair	n_in	n_out	% coverage
1	92	90	2	98
3	65	65	0	100
5	6	6	0	100
All	163	161	2	99

The MSE and RMSE for each MCMC iteration were calculated, with the means and corresponding standard deviations as:

MSE ± SD = 12.54 ± 1.82 (t C ha^{-1})^2; RMSE ± SD = 3.53 ± 0.24 t C ha^{-1}.

Do prediction intervals cover observed data 90% of the time? Yes.

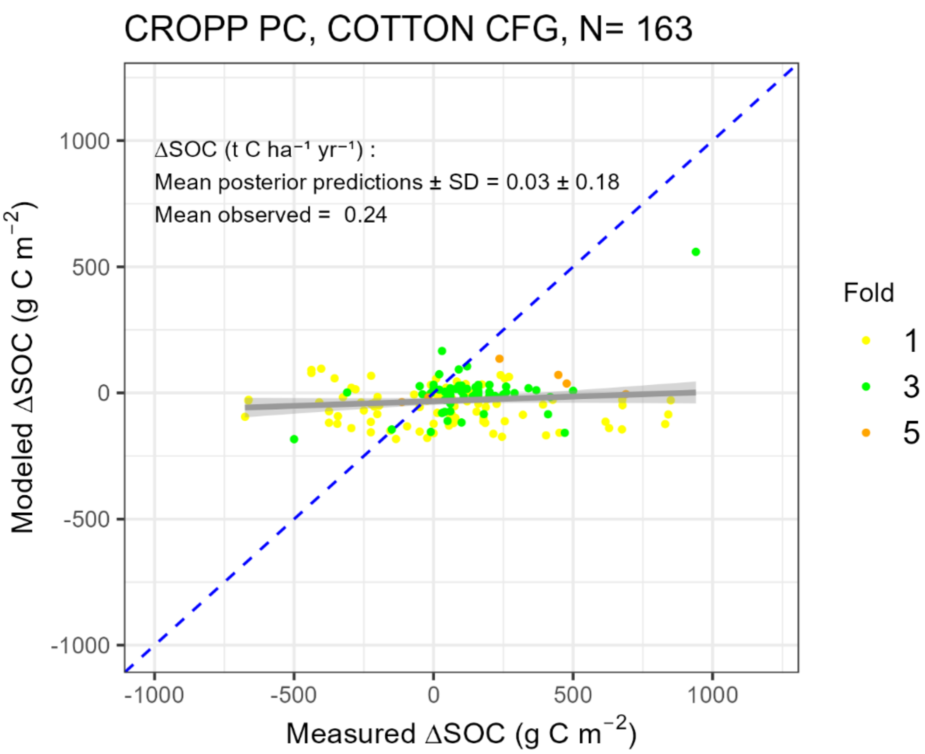

Figure 29. Scatterplot of <u>mean posterior predictions</u> verse measured SOC stock changes in response to changed cropping practices involving COTTON. Shaded area is the 95% confidence interval around linear least-squares fit and dashed line shows 1:1 relationship.

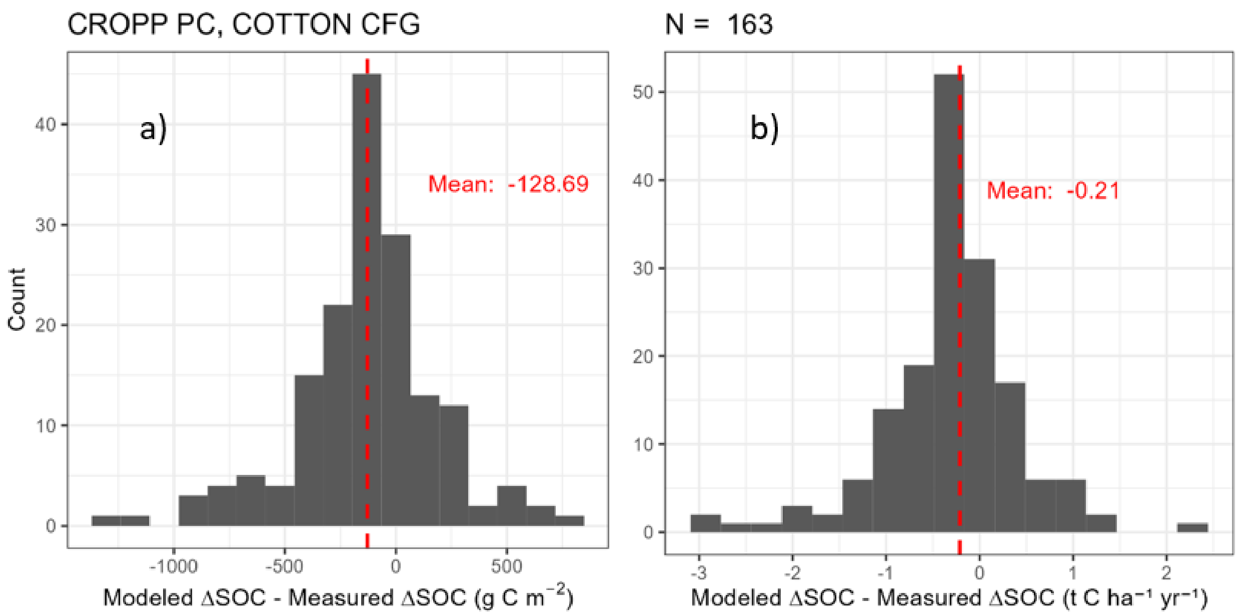

Figure 30. Histogram of model residuals (simulated - observed) for change in SOC in the validation dataset (6 sites with experiment durations from 3 to 19 years) used for evaluating changed cropping practices involving COTTON.

DISTU

Various tillage practices exhibit differing levels of soil disturbance, and the selection of crop residue management approaches can also vary across these strategies. Reduced or no-till practices, for instance, minimize soil disturbance, mitigate soil erosion, and retain organic matter near the surface. These practices foster cooler and more moisture-rich soil conditions, thereby decelerating the decomposition process and safeguarding SOC. Additionally, reduced or no-till practices uphold elevated soil carbon levels by restricting the infiltration of oxygen. In contrast, intensive tillage elevates soil oxygen levels, which can expedite the breakdown of organic matter. The model predicted increases in SOC stocks with adoption of reduced or no-till management in the top 30 cm of the profile. This effect does vary with deeper depths and depends on the climate and soil conditions (Ogle et al., 2019; Sun et al., 2020). Similar to the CROPP PC, in the case of DISTUR PCs, there were instances of both measured and modeled SOC decline. However, the simulated mean of posterior predictions of SOC changes exhibited less variability compared to the observed changes in SOC. Overall, the mean values of posterior predictions were consistently lower than the observed SOC changes. Since the same DayCent parameter sets were utilized for all treatments without specific calibration for individual treatments, it becomes imperative for the model to accurately capture the intricate interactions between management practices and carbon sequestration processes embedded within its underlying biogeochemical representations in order to effectively account for treatment differences. While it is more desirable for the model to underpredict rather than overpredict SOC changes, especially in the context of carbon-crediting programs to avoid over-crediting, it suggests that DayCent may not be adept at detecting SOC changes resulting from ploughing or crop residue incorporation. Gurung et al. (2020) and Mathers et al. (2023) introduced a new parameter, Till_Eff (referred to as "new" because it is not found in any of DayCent's input files but was used in calibration coding to update the DayCent cult.100 file). Till_Eff is defined as the "tillage multiplier on decomposition" and is used to scale tillage events from A to K, each corresponding to different cultivation factors. However, despite this adjustment, the issue of underprediction persists for DISTUR PCs as reported by Mathers et al. (2023). Additional model refinement within DayCent is necessary to accurately depict the influence of tillage on SOC dynamics (Gurung et al., 2020).

9.7 DISTU x NON_LEGUME_C4

Model uncertainties for 90% prediction intervals were computed for each fold and subsequently compared to the observed effects of SOC changes (Figures 31). The coverage rates of posterior predictive intervals for every fold during the validation process consistently exceeded 90% (Table 38).

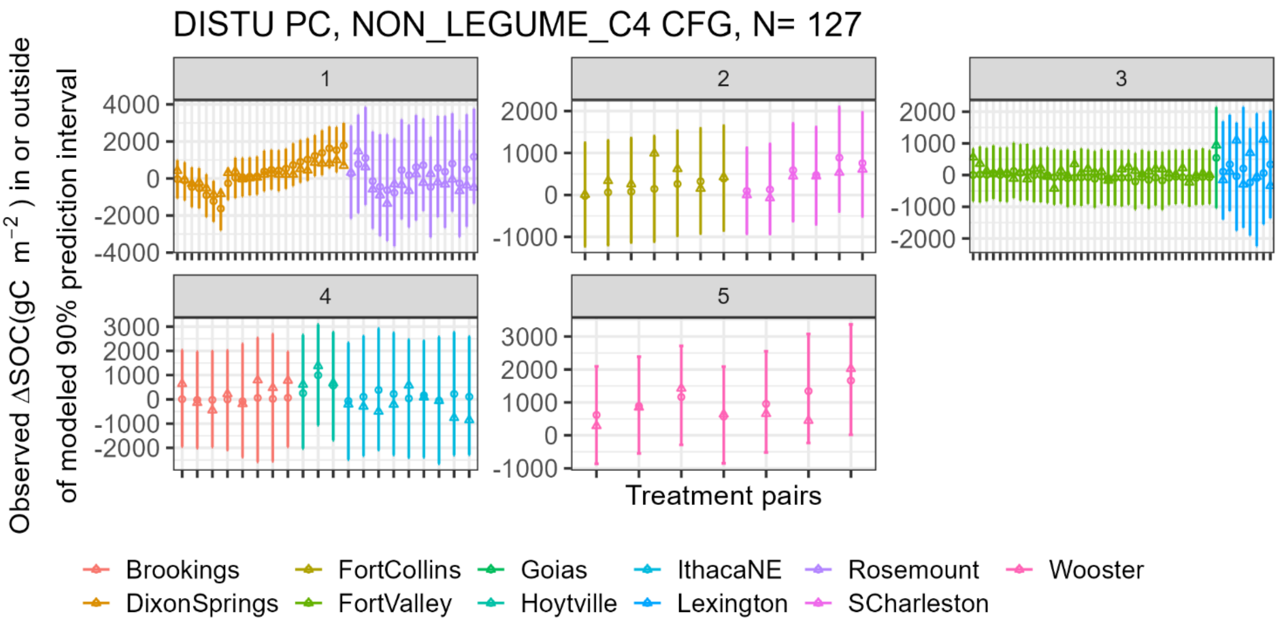

Figure 31. Modeled (dot) and measured SOC change (triangle) in response to changed tillage practices involving the NON_LEGUME_C4- type, with the error bars showing 90% prediction intervals.

Table 38. Number of observed SOC stock changes falling inside and outside of 90% prediction intervals in each fold for SOC predictions in response to changed tillage practice involving NON_LEGUME_C4-type.

Fold	n_pair	n_in	n_out	% coverage
1	42	41	1	98
2	13	13	0	100
3	45	45	0	100
4	20	20	0	100
5	7	7	0	100
All	127	126	1	99

The MSE and RMSE for each MCMC iteration were calculated, with the means and corresponding standard deviations as:

MSE ± SD = 26.22 ± 6.42 (t C ha^{-1})^2; RMSE ± SD = 5.09 ± 0.56 t C ha^{-1}.

Do prediction intervals cover observed data 90% of the time? Yes.

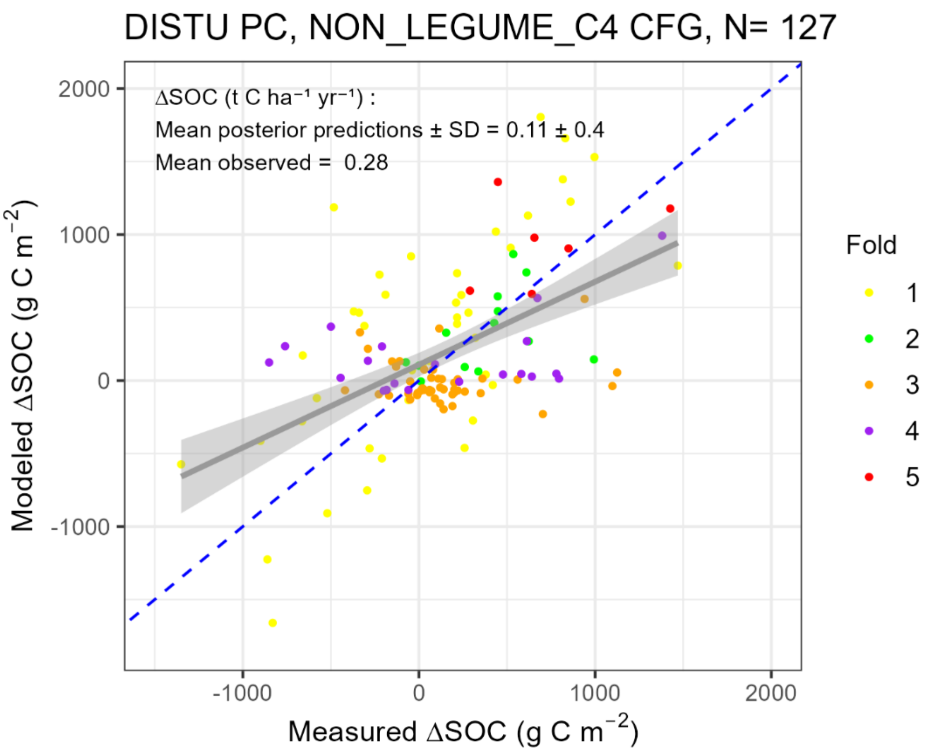

Figure 32. Scatterplot of mean posterior predictions verse measured SOC stock changes in response to changed tillage practices from the NON_LEGUME_C4 – type CFG. Shaded area is the 95% confidence interval around linear least-squares fit and dashed line shows 1:1 relationship.

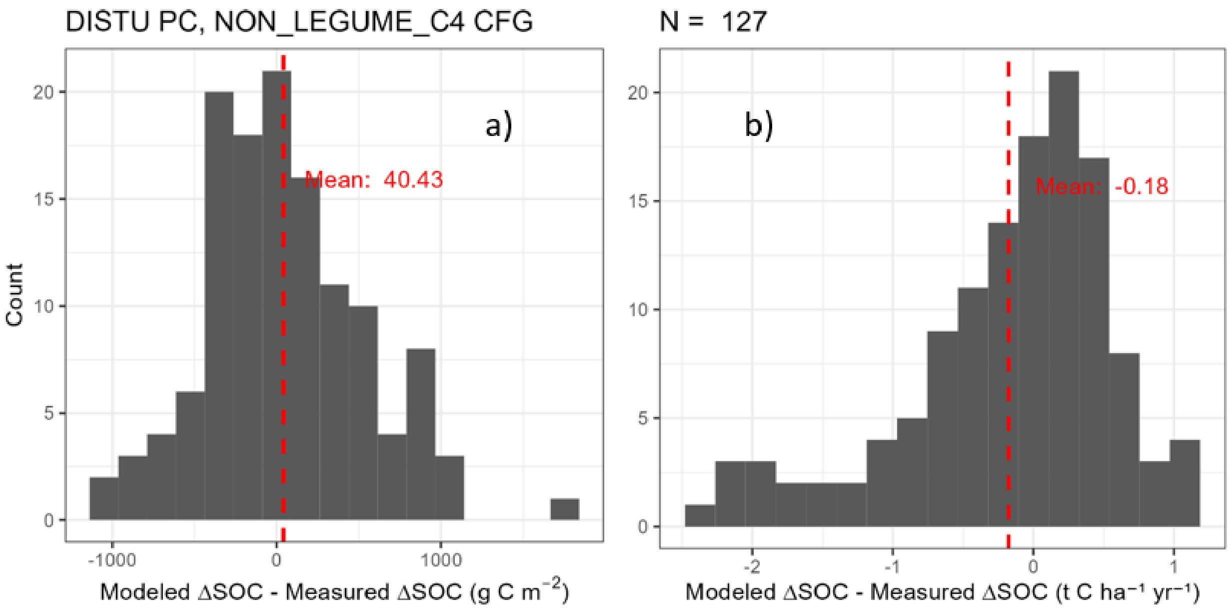

Figure 33. Histogram of model residuals (simulated - observed) for change in SOC in the validation dataset (11 sites with experiment durations from 3 to 43 years) used for evaluating changed tillage practices involving the NON_LEGUME_C4– type CFG.

9.8 DISTU x LEGUME_C3

Model uncertainties for 90% prediction intervals were computed for each fold and subsequently compared to the observed effects of SOC changes (Figures 34). The coverage rates of posterior predictive intervals for every fold during the validation process consistently exceeded 90% (Table 39).

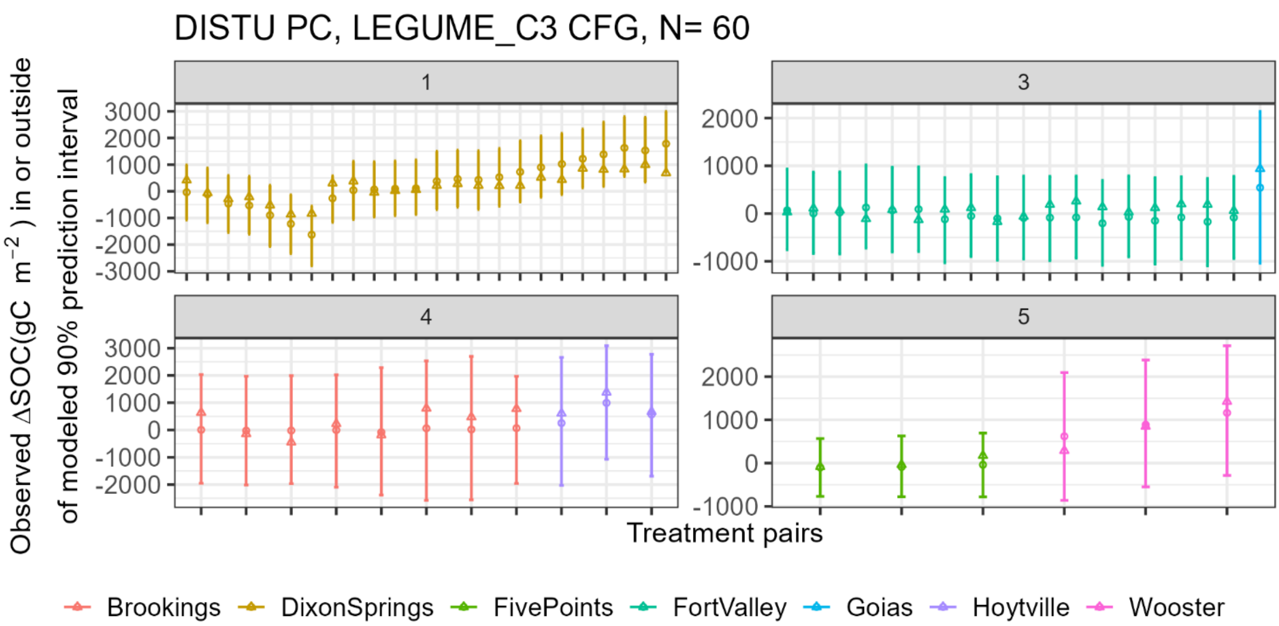

Figure 34. Modeled (dot) and measured SOC change (triangle) in response to changed tillage practices involving the LEGUME_C3-type, with the error bars showing 90% prediction intervals.

Table 39. Number of observed SOC stock changes falling inside and outside of 90% prediction intervals in each fold for SOC predictions in response to changed tillage practices involving LEGUME_C3-type CFG.

Fold	n_pair	n_in	n_out	% coverage
1	24	23	1	96
3	19	19	0	100
4	11	11	0	100
5	6	6	0	100
All	60	59	1	98

The MSE and RMSE for each MCMC iteration were calculated, with the means and corresponding standard deviations as:

MSE ± SD = 16.1 ± 4.19 (t C $ha^{-1})^2$; RMSE ± SD = 3.98 ± 0.5 t C ha^{-1}.

Do prediction intervals cover observed data 90% of the time? Yes.

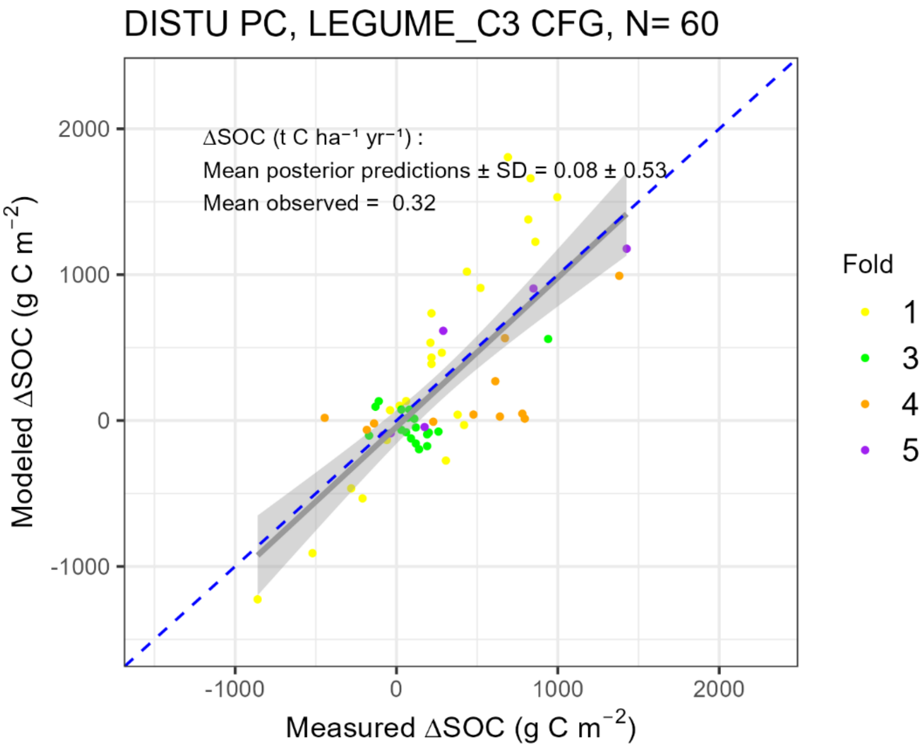

Figure 35. Scatterplot of mean posterior predictions verse measured SOC stock changes in response to changed tillage practices from the LEGUME_C3 – type CFG. Shaded area is the 95% confidence interval around linear least-squares fit and dashed line shows 1:1 relationship.

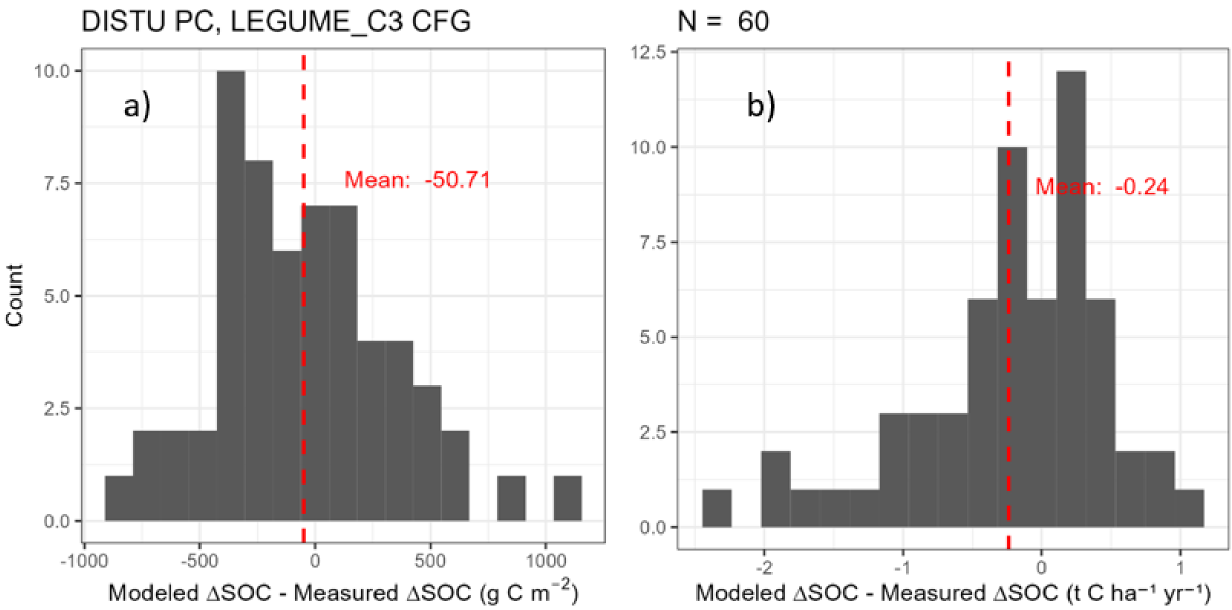

Figure 36. Histogram of model residuals (simulated - observed) for change in SOC in the validation dataset (7 sites with experiment durations from 3 to 30 years) used for evaluating changed tillage practices involving the LEGUME_C3 – type CFG.

9.9 DISTU x NON_LEGUME_C3

Model uncertainties for 90% prediction intervals were computed for each fold and subsequently compared to the observed effects of SOC changes (Figures 37). The coverage rates of posterior predictive intervals for every fold during the validation process consistently exceeded 90% (Table 40).

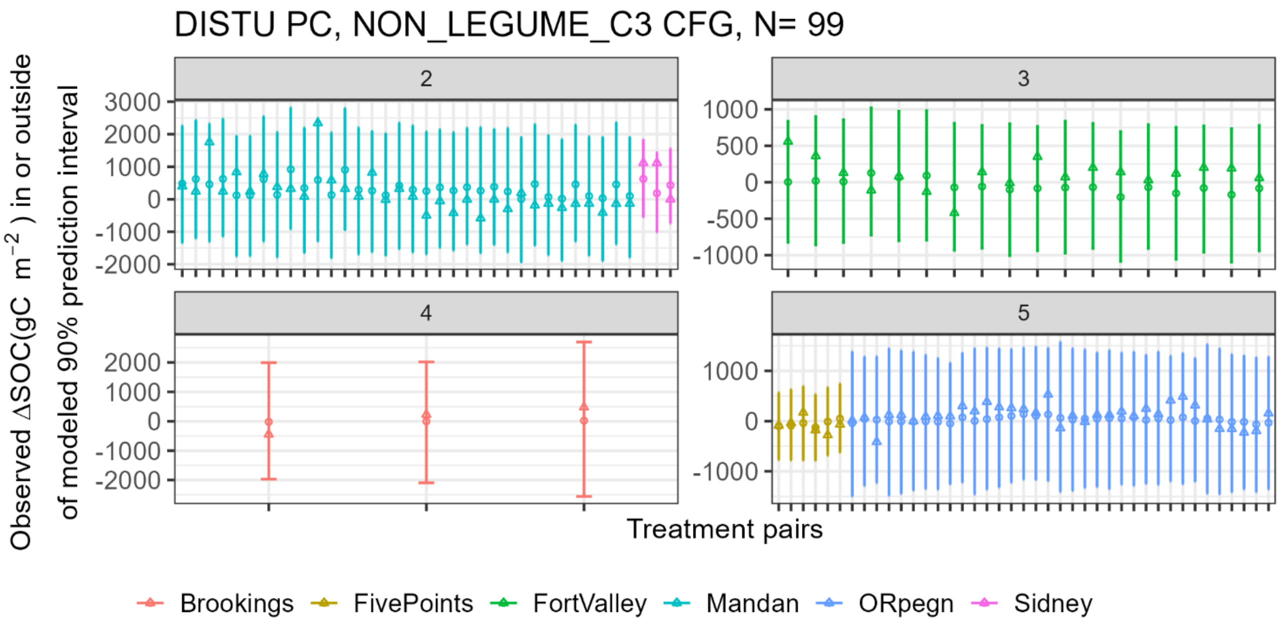

Figure 37. Modeled (dot) and measured SOC change (triangle) in response to changed tillage practices involving the NON_LEGUME_C3- type, with the error bars showing 90% prediction intervals.

Table 40. Number of observed SOC stock changes falling inside and outside of 90% prediction intervals in each fold for SOC predictions in response to changed tillage practice involving NON_LEGUME_C3-type CFG.

Fold	n_pair	n_in	n_out	% coverage
2	37	37	0	100
3	18	18	0	100
4	3	3	0	100
5	41	41	0	100
All	99	99	0	100

The MSE and RMSE for each MCMC iteration were calculated, with the means and corresponding standard deviations as:

MSE ± SD = 17.51 ± 4.24 (t C ha^{-1})^2; RMSE ± SD = 4.17 ± 0.4 t C ha^{-1}.

Do prediction intervals cover observed data 90% of the time? Yes.

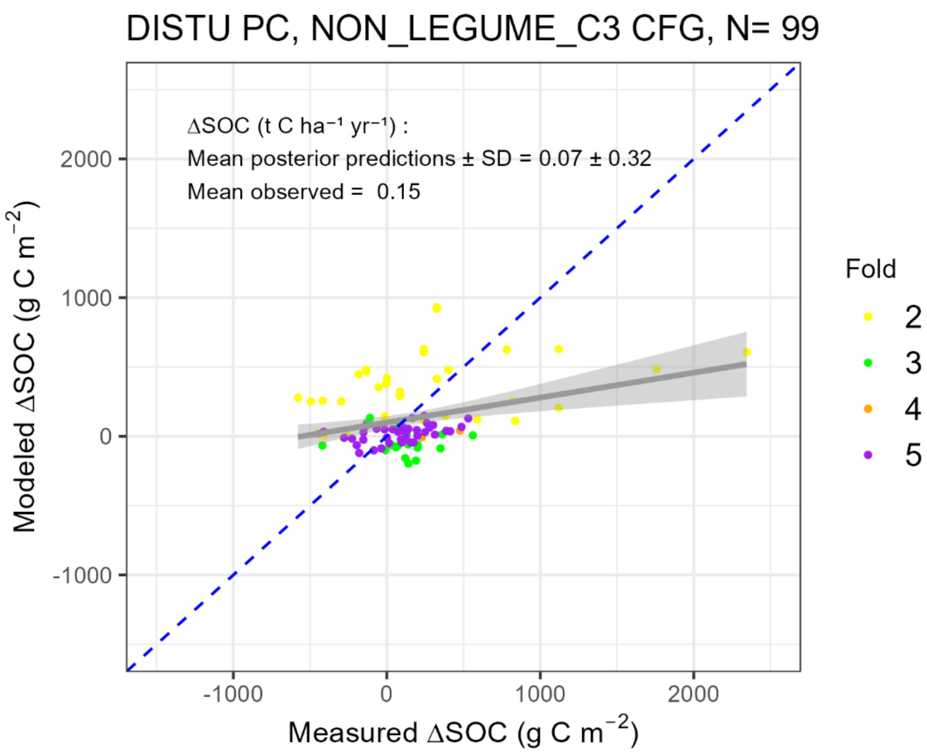

Figure 38. Scatterplot of mean posterior predictions verse measured SOC stock changes in response to changed tillage practices from the NON_LEGUME_C3 – type CFG. Shaded area is the 95% confidence interval around linear least-squares fit and dashed line shows 1:1 relationship.

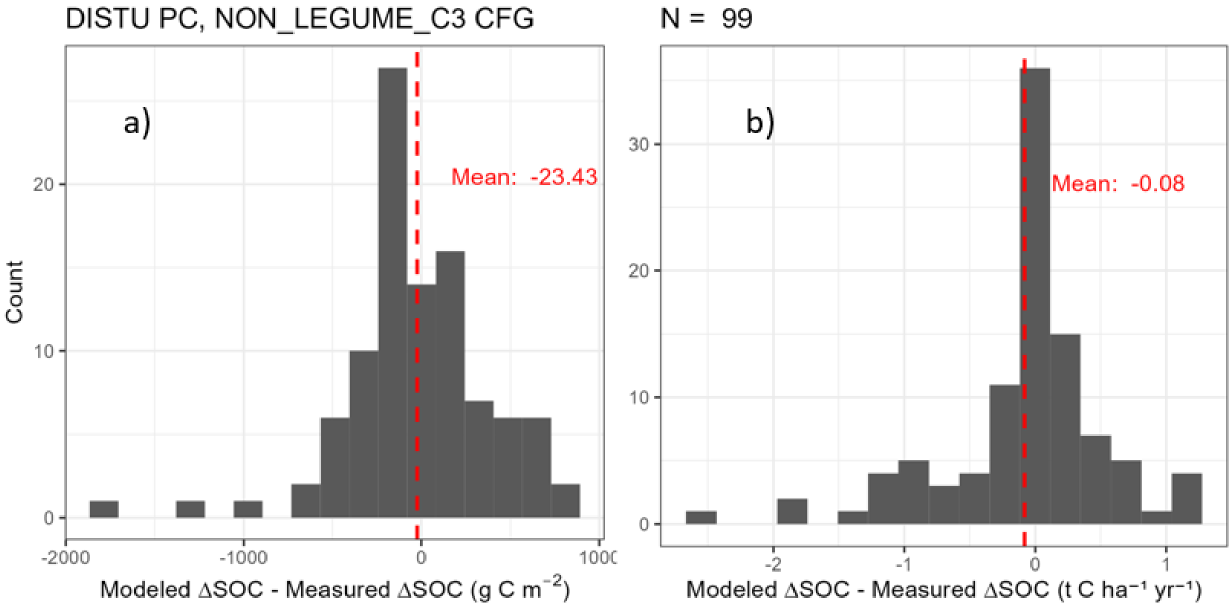

Figure 39. Histogram of model residuals (simulated - observed) for change in SOC in the validation dataset (6 sites with experiment durations from 3 to 79 years) used for evaluating changed tillage practices involving the NON_LEGUME_C3 – type CFG.

9.10 DISTU x COTTON

Model uncertainties for 90% prediction intervals were computed for each fold and subsequently compared to the observed effects of SOC changes (Figures 40). The coverage rates of posterior predictive intervals for every fold during the validation process consistently exceeded 90% (Table 41).

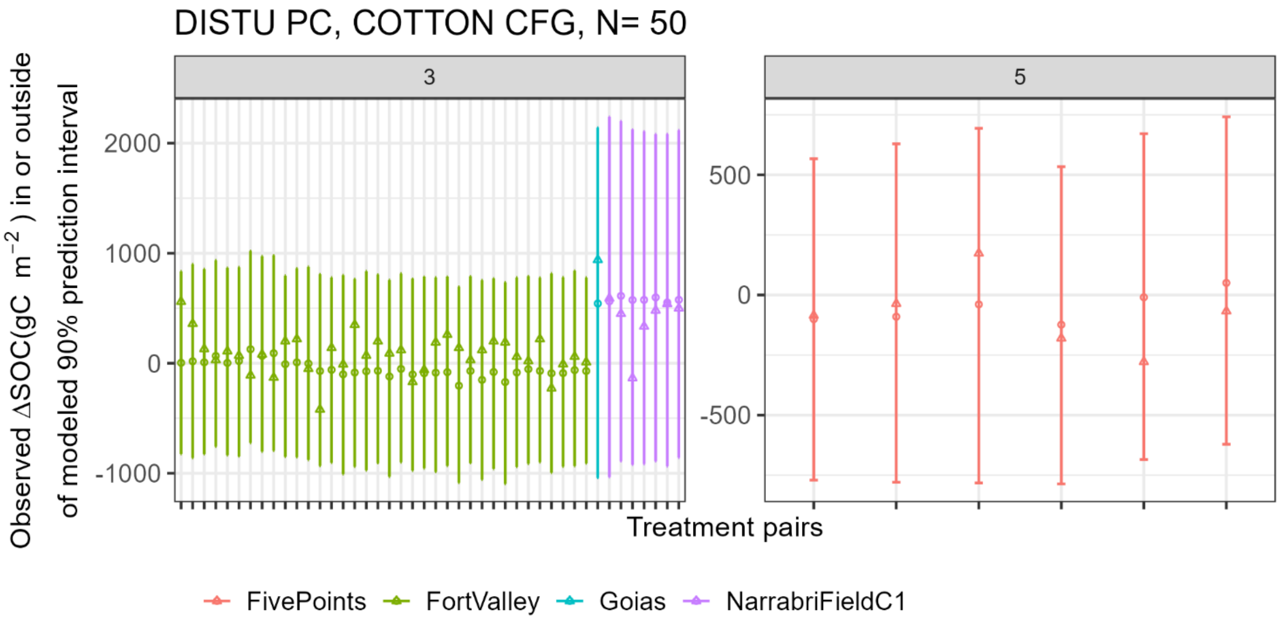

Figure 40. Modeled (dot) and measured SOC change (triangle) in response to changed tillage practices involving COTTON, with the error bars showing 90% prediction intervals.

Table 41. Number of observed SOC stock changes falling inside and outside of 90% prediction intervals in each fold for SOC predictions in response to changed tillage practice involving involving COTTON.

Fold	n_pair	n_in	n_out	% coverage
3	44	44	0	100
5	6	6	0	100
All	50	50	0	100

The MSE and RMSE for each MCMC iteration were calculated, with the means and corresponding standard deviations as:

MSE ± SD = 6.21 ± 0.79 $(t\ C\ ha^{-1})^2$; RMSE ± SD = 2.49 ± 0.15 $t\ C\ ha^{-1}$.

Do prediction intervals cover observed data 90% of the time? Yes.

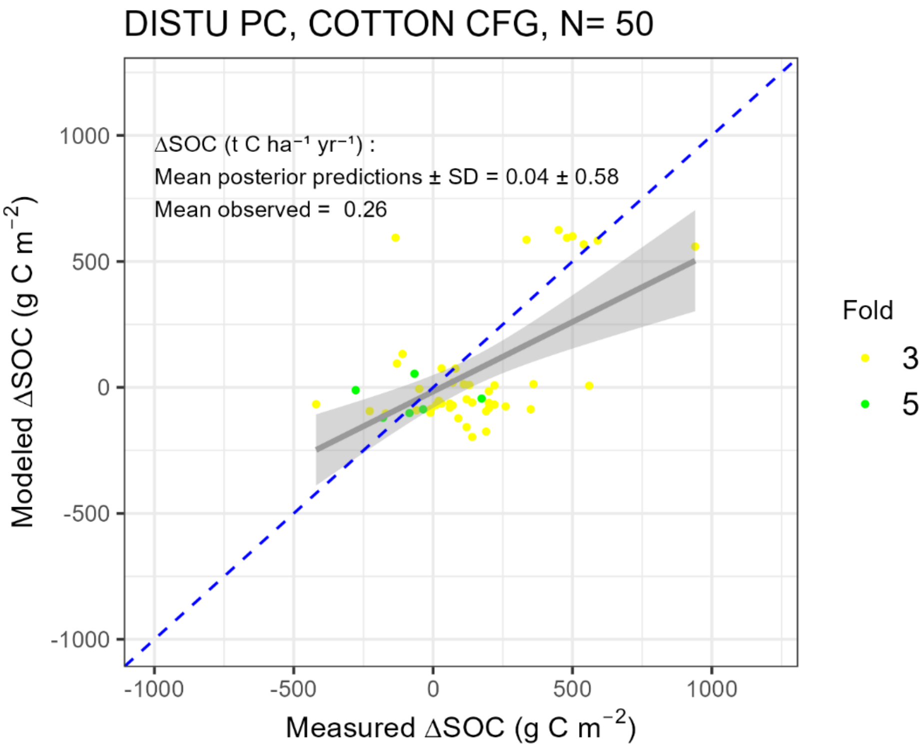

Figure 41. Scatterplot of mean posterior predictions verse measured SOC stock changes in response to changed tillage practices for COTTON. Shaded area is the 95% confidence interval around linear least-squares fit and dashed line shows 1:1 relationship.

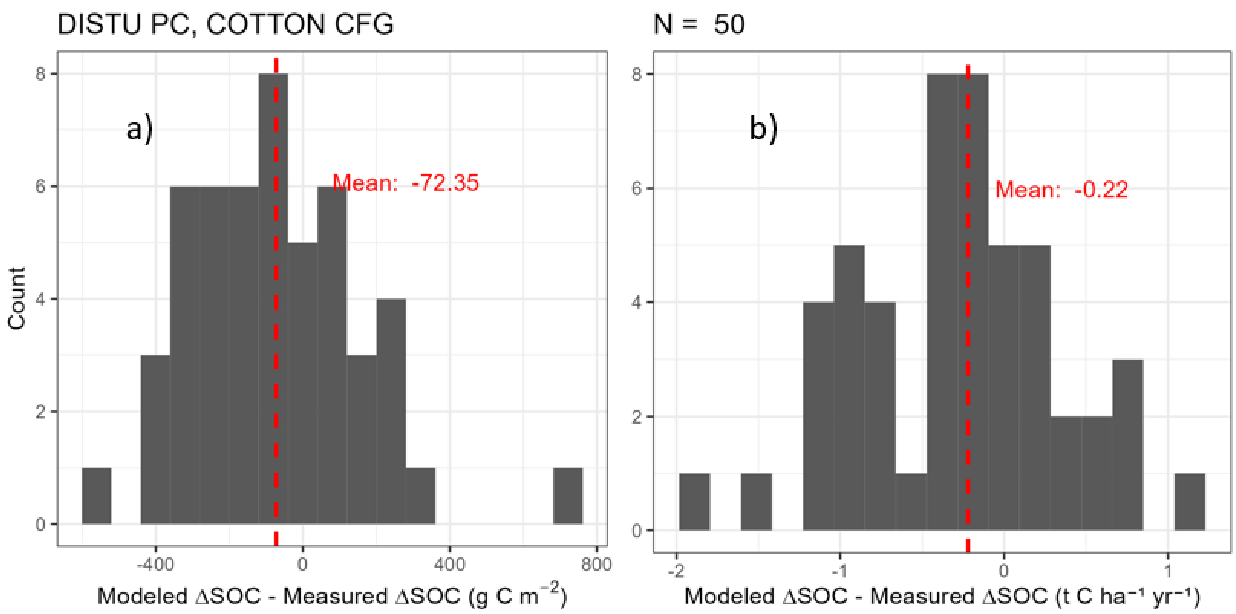

Figure 42. Histogram of model residuals (simulated - observed) for change in SOC in the validation dataset (4 sites with experiment durations from 3 to 19 years) used for evaluating changed cropping practices involving COTTON.

NFERT

Reduced N application, particularly in the context of agricultural practices, can significantly contribute to greenhouse gas emissions mitigation. This report focuses on calibrating and validating PC/CFG/SOC combinations due to data limitations for N_2O, with the quantification of N_2O reduction benefits linked to N reduction through the VM42 Quantification Approach 3. N management practices have a notable impact on soil carbon dynamics. When combined with sustainable agricultural practices such as cover cropping and reduced tillage, reduced N application can enhance C sequestration. For example, the inclusion of cover crop residues adds organic matter to the soil, while reduced soil disturbance during tillage helps prevent carbon loss. Existing research has indicated that mineral fertilization is less effective as a C sequestration strategy than alternative management approaches and may even result in SOC losses (Jiang et al., 2018; Salinas-Garcia et al., 1997). The use of mineral fertilizers, without due consideration for nutrient balance and ecosystem nutrient cycling, can lead to imbalances. Excessive N fertilization, in particular, can stimulate microbial activity, accelerating the decomposition of organic matter and releasing stored carbon as CO_2 through microbial respiration. This can negate potential gains in C sequestration.

Furthermore, the continuous and excessive application of inorganic N fertilizers can disrupt the composition and activity of soil microbial communities. Some microbial groups that thrive in the presence of high nutrient levels may become more efficient at breaking down organic matter, resulting in a reduction in SOC. Additionally, excessive N application can favor above-ground plant growth at the expense of below-ground root biomass. Since root biomass plays a significant role in SOC accumulation through root turnover and decomposition, a shift in resource allocation by plants can lead to reduced C input into the soil.

Therefore, the impact of inorganic N application rates on C sequestration is highly dependent on specific environmental contexts. This influence can fluctuate based on the intricate interplay among factors such as increased plant productivity, organic matter input, microbial activity, and the potential for nitrogen saturation within the ecosystem. For instance, the extreme annual SOC change rate observed in the NFERT PC validation dataset at the Lexington site in Kentucky was from the comparison between: "Lexington_NT_84N_1980" and "Lexington_NT_0N_1980." The former represents "continuous corn with no-till and an application of 84 kg N ha^{-1}," while the latter represents "continuous corn with no-till and no N application." Based on the 1980 measurements, the yearly SOC change rate was exceptionally high at 3.44 t C ha^{-1} yr^{-1}. This measurement was taken five years into the experiment, which initially started in 1975. However, by 1989, the average yearly SOC change rate had decreased significantly to 0.21 t C ha^{-1} yr^{-1}, marking the 14th year of the experiment. In a similar comparison between "Lexington_NT_168N_1980" and "Lexington_NT_0N_1980," the measurement in 1980 indicated an average yearly SOC change rate of 2.56 t C ha^{-1} yr^{-1}, which had reduced to 0.25 t C ha^{-1} yr^{-1} by 1989. This variability and context-dependent nature of inorganic N application rates on C sequestration have led to increased variations in SOC change rates (Figures 12-15) than that for other PCs.

The overall model performance adhered to the guidelines set forth by Verra VMD53, with more than 90% of model prediction intervals encompassing the measured values and a mean bias in this category lower than PMU (Section 8.6). However, it's worth noting that the measured ∆SOC changes generally exceeded the simulated values, which corresponds with the findings of Mathers et al. (2023). In their report, Mathers et al. observed that DayCent displayed a conservative behavior by consistently under-predicting SOC changes across 14 pairs of CFG and PC combinations.

9.11 NFERT x NON_LEGUME_C4

Model uncertainties for 90% prediction intervals were computed for each fold and subsequently compared to the observed effects of SOC changes (Figures 43). The coverage rates of posterior predictive intervals for every fold during the validation process consistently exceeded 90% (Table 42).

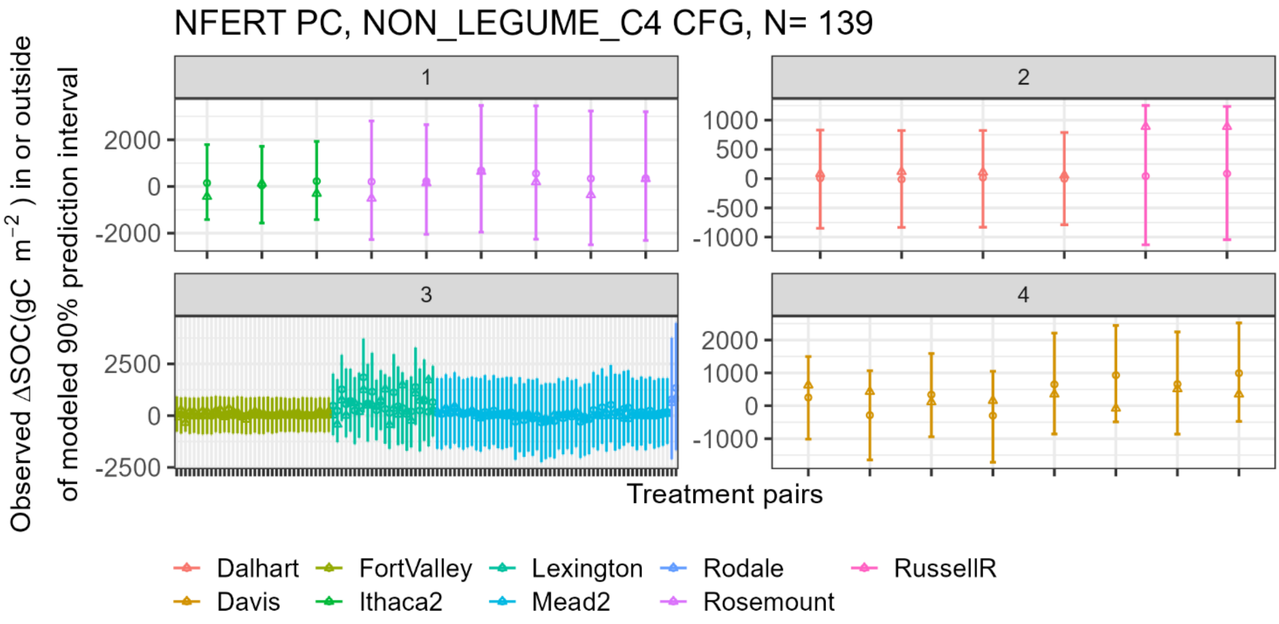

Figure 43. Modeled (dot) and measured SOC change (triangle) in response to changed nitrogen practices involving the NON_LEGUME_C4-type, with the error bars showing 90% prediction intervals.

Table 42. Number of observed SOC stock changes falling inside and outside of 90% prediction intervals in each fold for SOC predictions in response to changed nitrogen practices involving NON_LEGUME_C4-type CFG.

Fold	n_pair	n_in	n_out	% coverage
1	9	9	0	100
2	6	6	0	100
3	116	116	0	100
4	8	8	0	100
All	139	139	0	100

The MSE and RMSE for each MCMC iteration were calculated, with the means and corresponding standard deviations as:

MSE ± SD = 14.53 ± 2.64 (t C ha^{-1})^2; RMSE ± SD = 3.8 ± 0.33 t C ha^{-1}.

Do prediction intervals cover observed data 90% of the time? Yes.

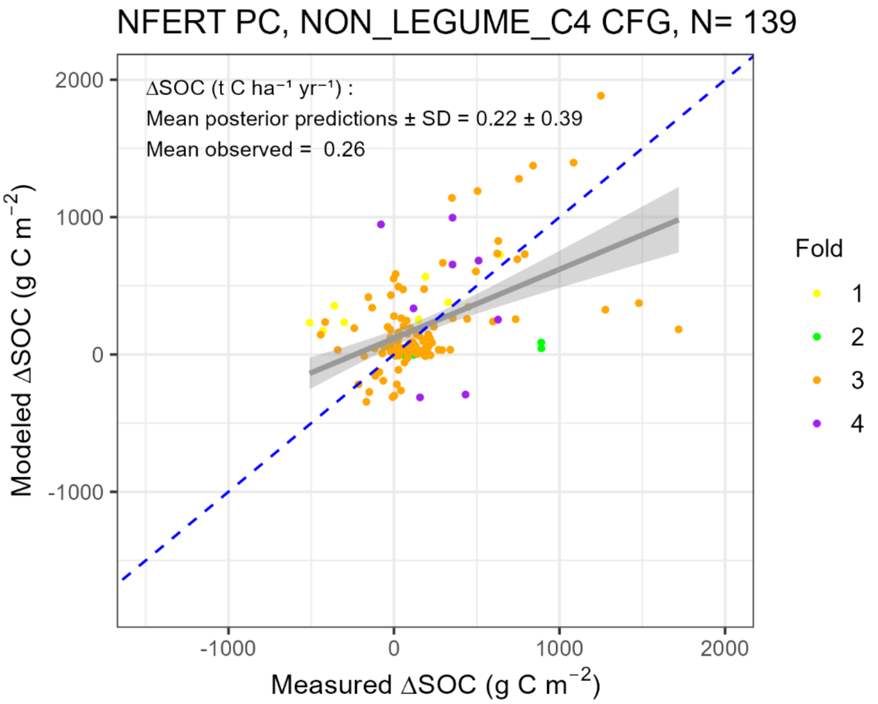

Figure 44. Scatterplot of <u>mean posterior predictions</u> verse measured SOC stock changes in response to changed nitrogen practices from the NON_LEGUME_C4 – type CFG. Shaded area is the 95% confidence interval around linear least-squares fit (a measure of the uncertainty associated with the estimated regression) and dashed line shows 1:1 relationship.

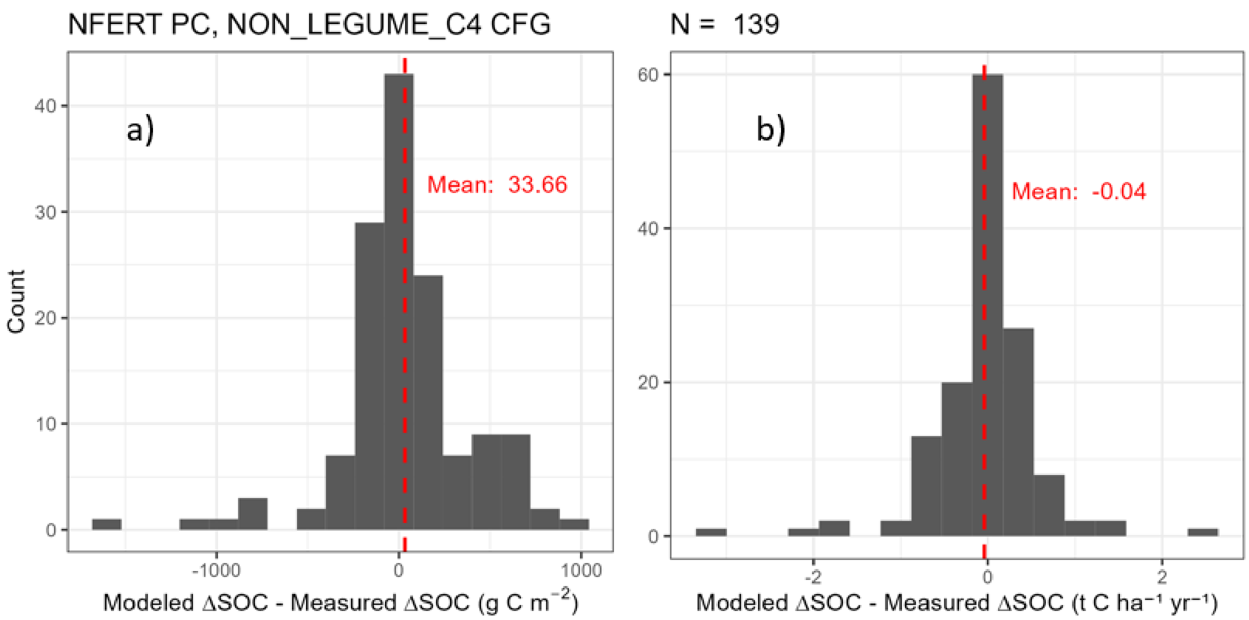

Figure 45. Histogram of model residuals (simulated - observed) for change in SOC in the validation dataset (9 sites with experiment durations from 3 to 22 years) used for evaluating changed nitrogen practices involving the NON_LEGUME_C4 – type CFG.

9.12 NFERT x LEGUME_C3

Model uncertainties for 90% prediction intervals were computed for each fold and subsequently compared to the observed effects of SOC changes (Figures 46). The coverage rates of posterior predictive intervals for every fold during the validation process consistently exceeded 90% (Table 43).

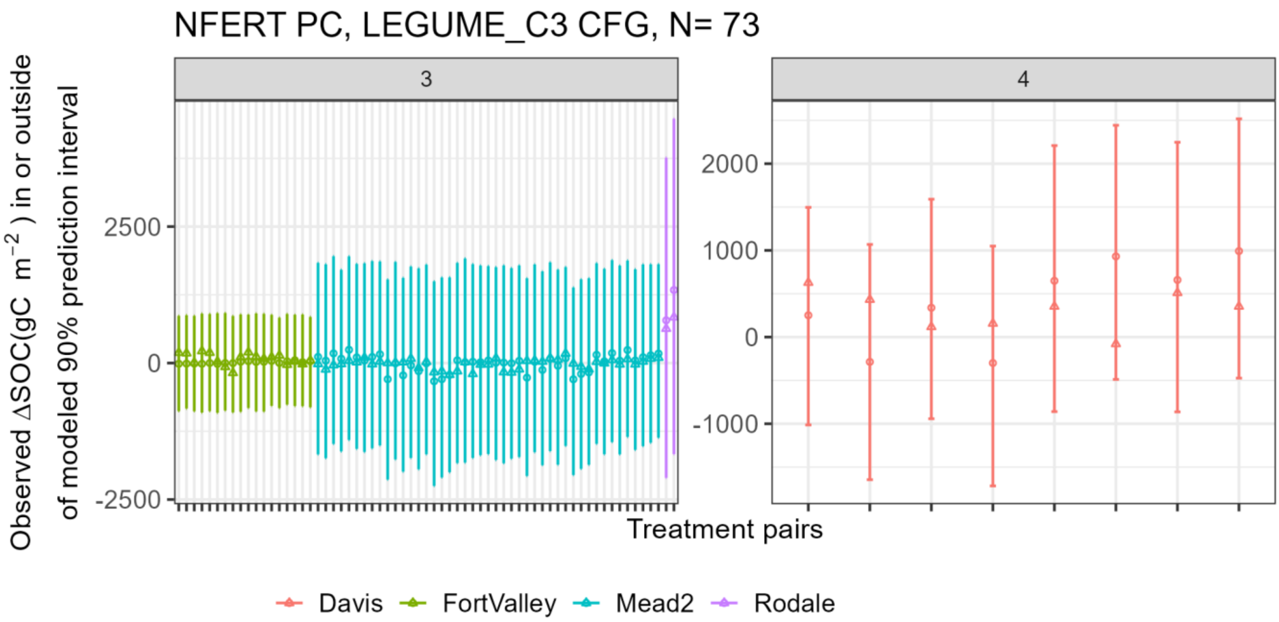

Figure 46. Modeled (dot) and measured SOC change (triangle) in response to changed nitrogen practices involving the LEGUME_C3-type, with the error bars showing 90% prediction intervals.

Table 43. Number of observed SOC stock changes falling inside and outside of 90% prediction intervals in each fold for SOC predictions in response to changed nitrogen practices involving LEGUME_C3-type CFG.

Fold	n_pair	n_in	n_out	% coverage
3	65	65	0	100
4	8	8	0	100
All	73	73	0	100

The MSE and RMSE for each MCMC iteration were calculated, with the means and corresponding standard deviations as:

MSE ± SD = 7.0 ± 1.99 (t C $ha^{-1})^2$; RMSE ± SD = 2.62 ± 0.34 t C ha^{-1}.

Do prediction intervals cover observed data 90% of the time? Yes.

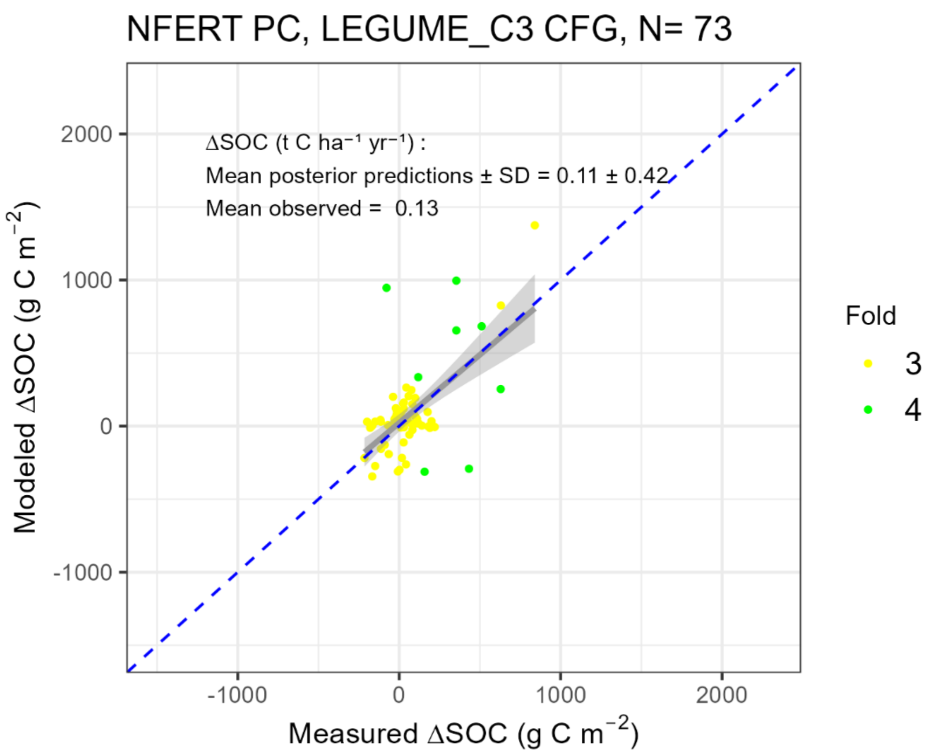

Figure 47. Scatterplot of mean posterior predictions verse measured SOC stock changes in response to changed nitrogen practices from the LEGUME_C3– type CFG. Shaded area is the 95% confidence interval around linear least-squares fit and dashed line shows 1:1 relationship.

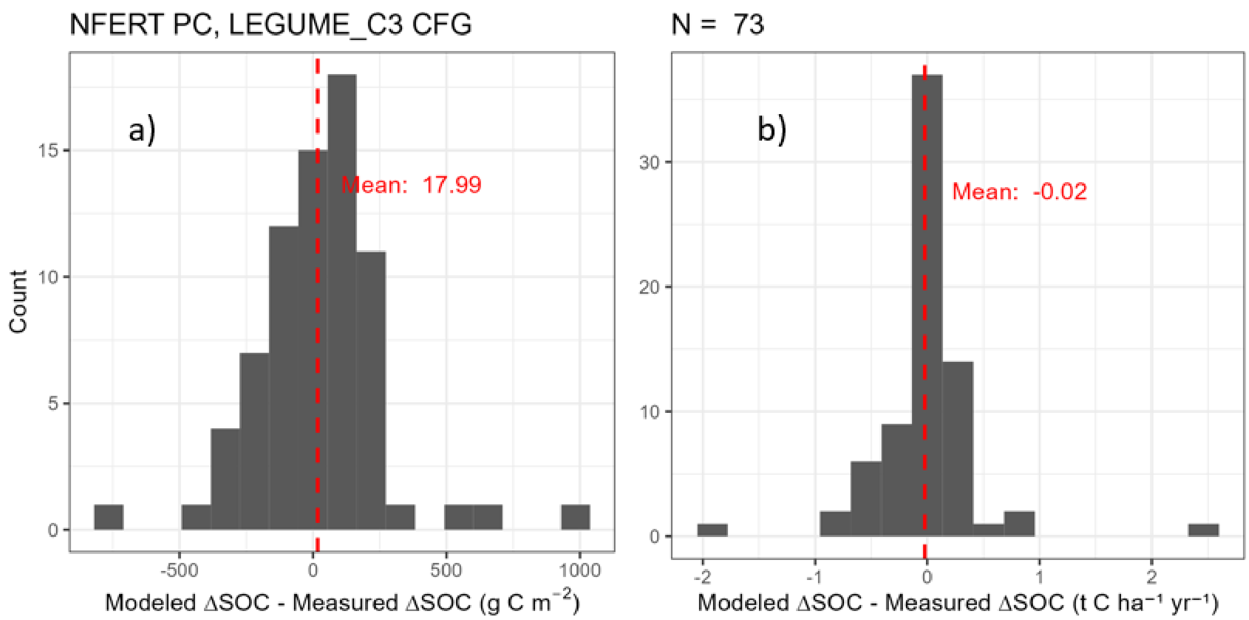

Figure 48. Histogram of model residuals (simulated - observed) for change in SOC in the validation dataset (4 sites with experiment durations from 3 to 21 years) used for evaluating changed nitrogen practices involving the LEGUME_C3– type CFG.

9.13 NFERT x NON_LEGUME_C3

Model uncertainties for 90% prediction intervals were computed for each fold and subsequently compared to the observed effects of SOC changes (Figures 49). The coverage rates of posterior predictive intervals for every fold during the validation process consistently exceeded 90% (Table 44).

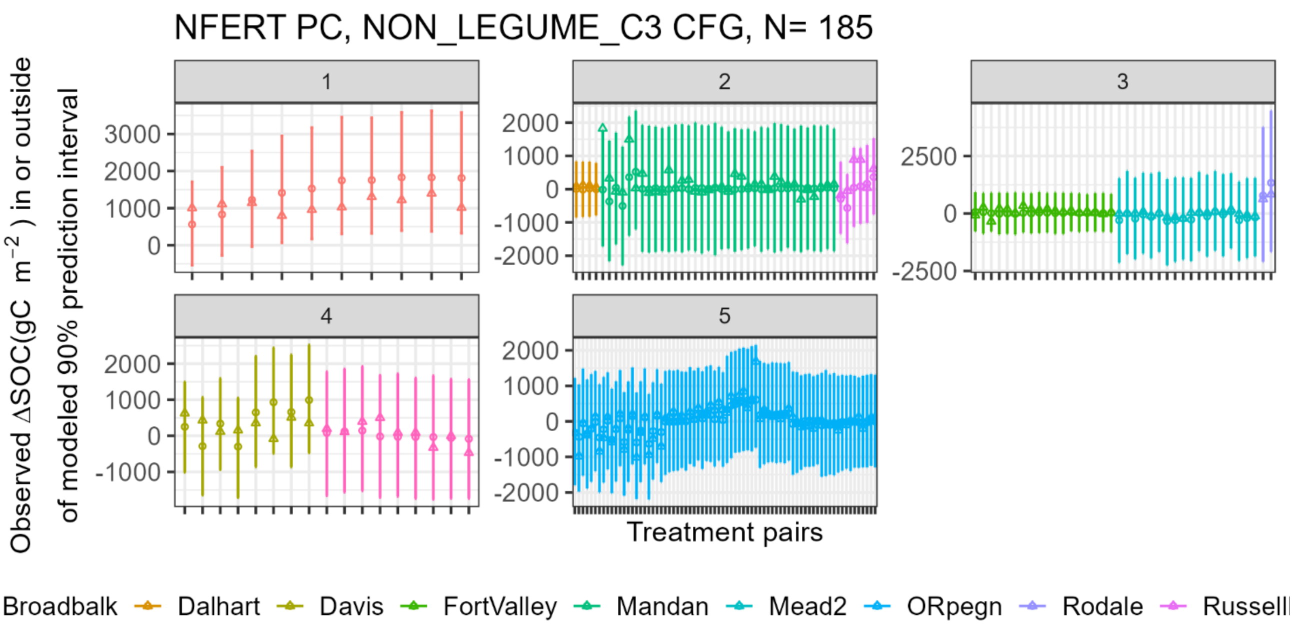

Figure 49. Modeled (dot) and measured SOC change (triangle) in response to changed nitrogen practices involving the NON_LEGUME_C3-type, with the error bars showing 90% prediction intervals.

Table 44. Number of observed SOC stock changes falling inside and outside of 90% prediction intervals in each fold for SOC predictions in response to changed nitrogen practices involving NON_LEGUME_C3-type CFG.

Fold	n_pair	n_in	n_out	% coverage
1	10	10	0	100
2	46	45	1	98
3	38	38	0	100
4	17	17	0	100
5	74	74	0	100
All	185	184	1	99

The MSE and RMSE for each MCMC iteration were calculated, with the means and corresponding standard deviations as:

MSE ± SD = 12.95 ± 2.65 (t C ha^{-1})^2; RMSE ± SD = 3.58 ± 0.36 t C ha^{-1}.

Do prediction intervals cover observed data 90% of the time? Yes.

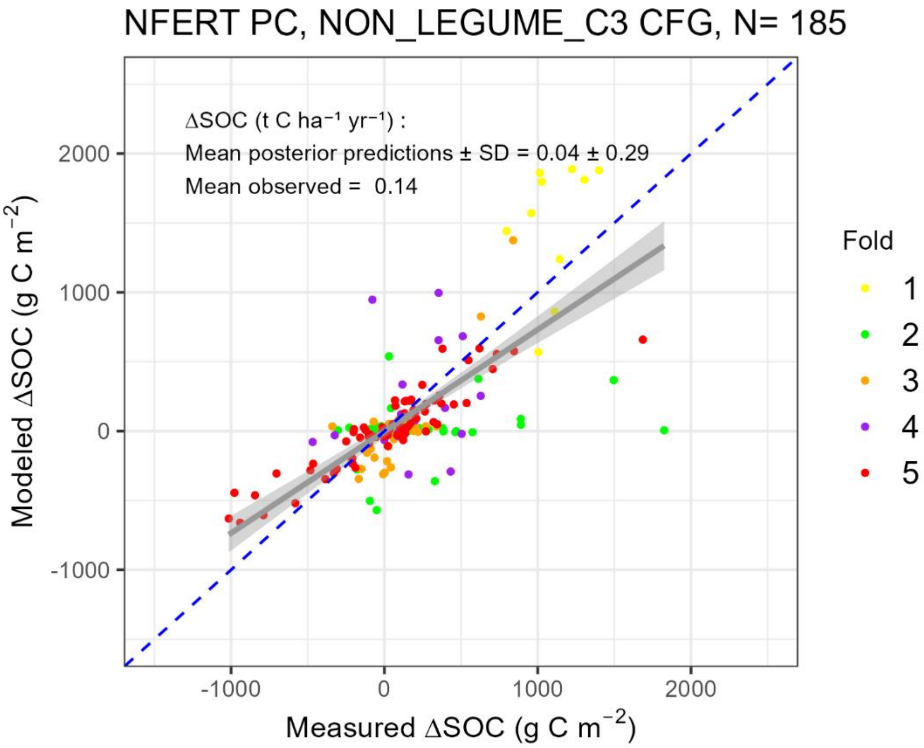

Figure 50. Scatterplot of <u>mean posterior predictions</u> verse measured SOC stock changes in response to changed nitrogen practices from the NON_LEGUME_C3 – type CFG. Shaded area is the 95% confidence interval around linear least-squares fit and dashed line shows 1:1 relationship.

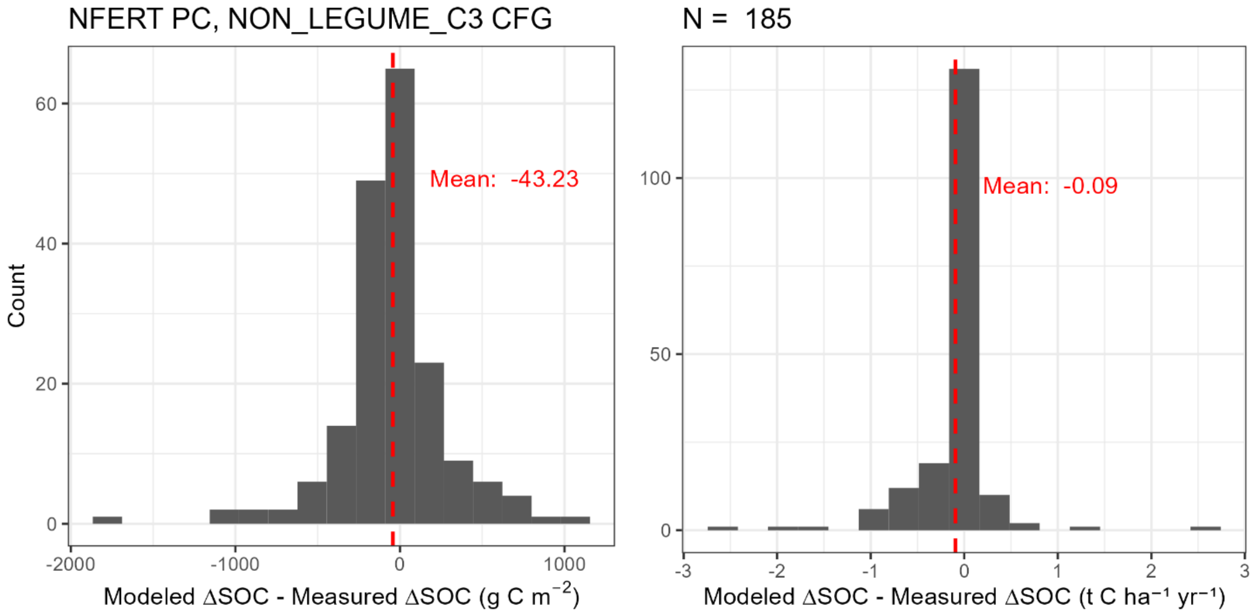

Figure 51. Histogram of model residuals (simulated - observed) for change in SOC in the validation dataset (10 sites with experiment durations from 3 to 162 years) used for evaluating changed nitrogen practices involving the NON_LEGUME_C3 – type CFG.

9.14 NFERT x ALL_CFGs

Model uncertainties for 90% prediction intervals were computed for each fold and subsequently compared to the observed effects of SOC changes (Figures 52). The coverage rates of posterior predictive intervals for every fold during the validation process consistently exceeded 90% (Table 45).

Table 45. Number of observed SOC stock changes falling inside and outside of 90% prediction intervals in each fold for SOC predictions in response to changed nitrogen management practices involving ALL_CFGs.

Fold	n_pair	n_in	n_out	% coverage
1	19	19	0	100
2	46	45	1	98
3	125	125	0	100
4	17	17	0	100
5	74	74	0	100
All	281	280	1	100

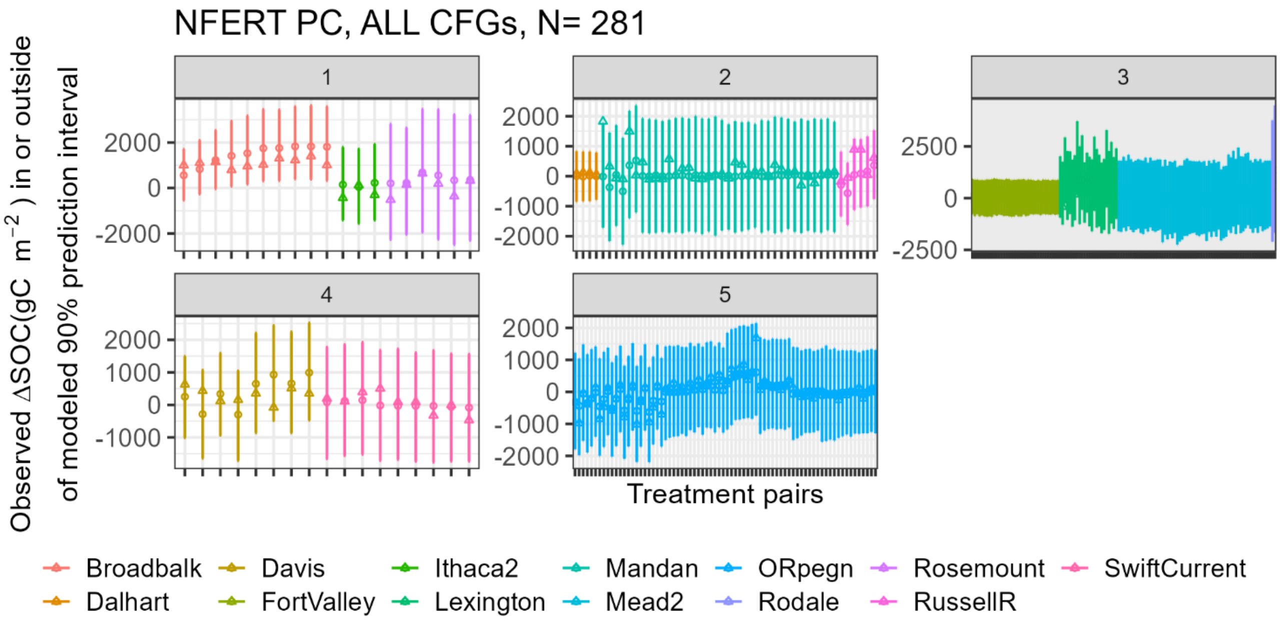

Figure 52. Modeled (dot) and measured SOC change (triangle) in response to changed nitrogen practices involving ALL_CFGs, with the error bars showing 90% prediction intervals.

The MSE and RMSE for each MCMC iteration were calculated, with the means and corresponding standard deviations as:

MSE ± SD = 13.58 ± 2.09 $(t\ C\ ha^{-1})^2$; RMSE ± SD = 3.67 ± 0.28 $t\ C\ ha^{-1}$.

Do prediction intervals cover observed data 90% of the time? Yes.

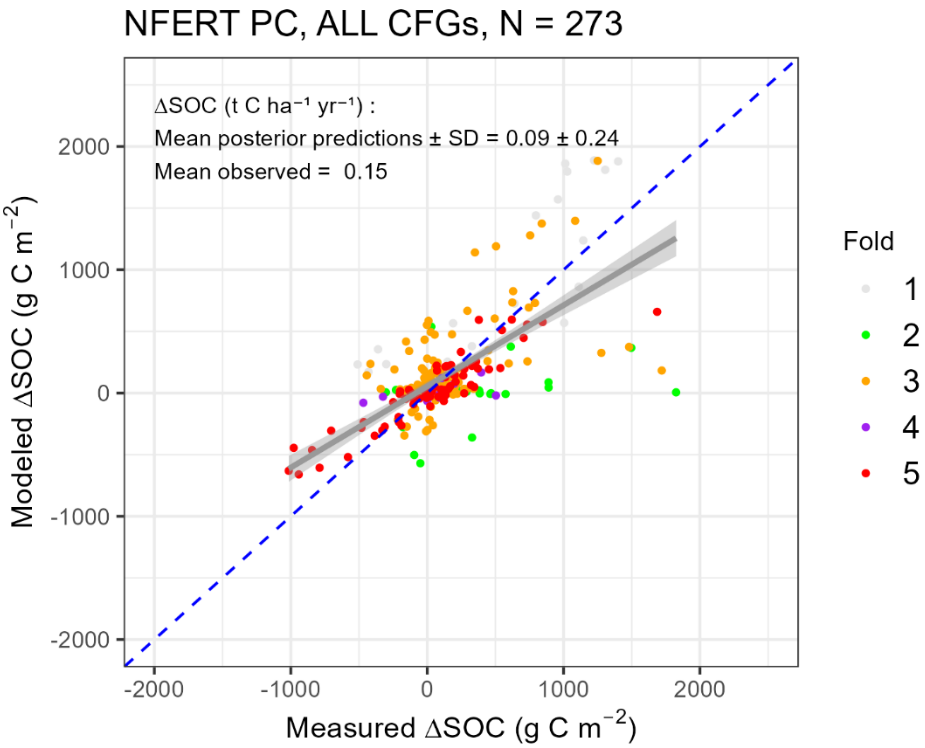

Figure 53. Scatterplot of <u>mean posterior predictions</u> verse measured SOC stock changes in response to changed nitrogen practices from ALL_CFGs. Shaded area is the 95% confidence interval around linear least-squares fit and dashed line shows 1:1 relationship.

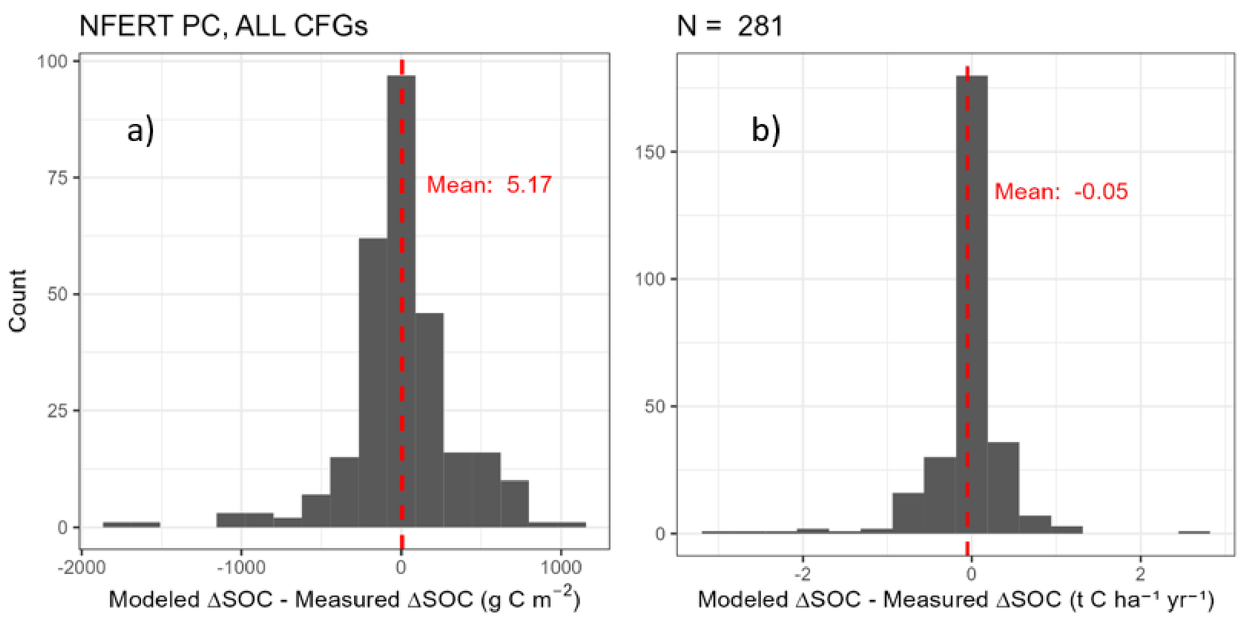

Figure 54. Histogram of model residuals (simulated - observed) for change in SOC in the validation dataset (13 sites with experiment durations from 3 to 162 years) used for evaluating changed nitrogen practices involving ALL_CFGs.

10. Model Validation Outputs for Use in Uncertainty Calculations

As per VCS VM42 Section 8.6 on Uncertainty, when using the model for project crediting, a deduction for uncertainty from modeling will be calculated using the methods outlined in VM42 Section 8.6.1. We have opted for a Monte Carlo approach for error propagation. The model error in predicting SOC stocks in baseline and project scenarios will be determined by sampling from the posterior distributions of the parameters that were adjusted during the k-fold cross-validation and from variance parameters capturing prediction error employed in this report. Since the five k-fold iterations resulted in different sets of optimized parameters, we used a total of 240 parameter sets from the k-fold cross-validation, with 48 parameter sets obtained from each individual k-fold iteration, for both baseline and project practice scenarios. The 240 parameter sets are representative of the parameter distribution observed in the joint posterior distribution during the k-fold calibration. Table A2 and Figure A2 illustrate the similarity between the marginal posterior distributions and indicate that the 240 parameter sets covered the parameter spaces. These parameter sets are records of the model's parameter values after they have been adjusted during the k-fold cross-validation process.

When utilizing DayCent for carbon crediting, the ensemble of simulations for a specific modeling unit will comprise a total of 480 simulations encompassing both the baseline and practice scenarios. The Monte Carlo prediction represents SOC stock estimate based on the range of input parameters and variability present in the modeling dataset. This process is repeated for each unique combination of parameter states in the stored posterior, which is a record of the model's parameter values after they have been adjusted during the k-fold cross-validation process. This approach is used to quantify uncertainty in the estimate of total emission reductions for the project.

11. Evaluation of Final Parameter Sets

The final 240 parameter sets for carbon crediting were obtained from the 5-fold cross-validation. This approach ensures computational efficiency when running the model for carbon crediting, which will be conducted at a large scale across numerous modeling units. These simulations can be time-consuming, and using a reduced number of optimized parameter sets reduces computation time while maintaining model accuracy. In this section, we compare the model's performance using the final parameter sets with the overall k-fold cross-validation performance.

11.1 Model performance across all PCs and CFGs

Table 46. Comparison of model's performance metrics of model bias, mean squared error, and root mean squared error across all PCs and CFGs during k-fold cross-validation and final calibration.

Performance statistics	During k-fold cross-validation	With final parameter sets
Mean bias (t C ha^{-2})	0.15	-0.09
MSE ± SD (t C ha^{-2})^2	15.89 ± 1.18	15.77 ± 1.02
RMSE ± SD (t C ha^{-2})	3.98 ± 0.14	3.97 ± 0.13

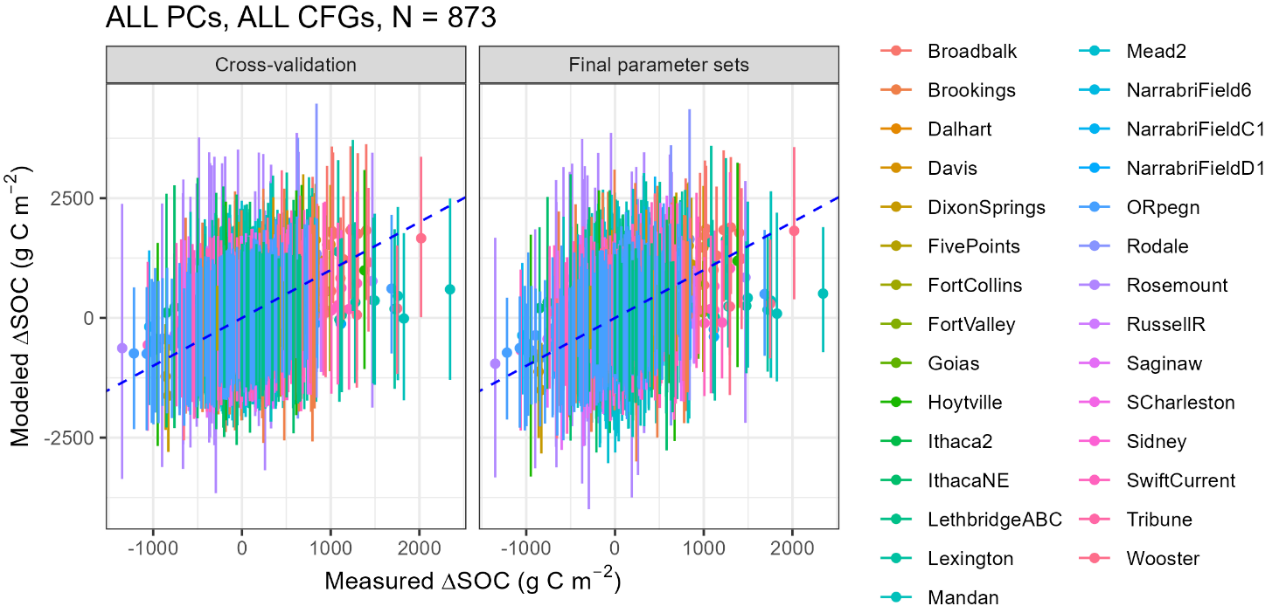

Figure 55. Modeled verse measured SOC stock changes in all practice changes and crop types, faceted by calibration stage. Error bars represent 90% prediction intervals.

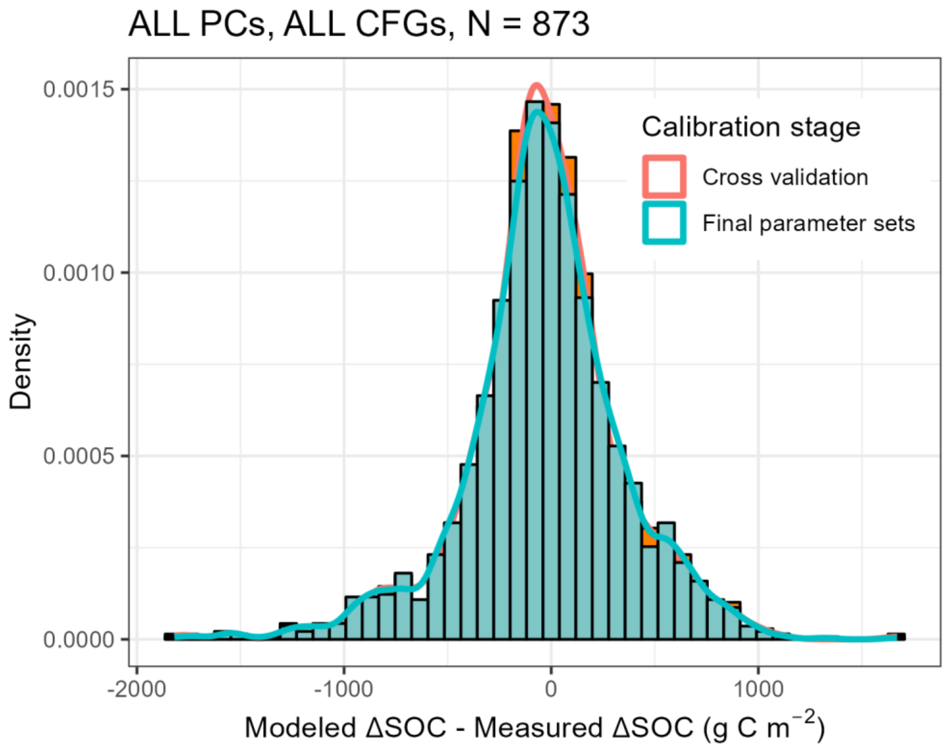

Figure 56. Histogram of model residuals (simulated - observed) for changes in SOC in all studies (experiment duration from 3 to 162 years) used for model validation across all practices and crop types. Orange bars represent residuals obtained during cross-validation, while greenish-blue bars represent residuals with the final parameter sets.

11.2 Model performance for each PC/CFG/SOC category

Table 47. Comparison of model bias in each PC x CFG category during cross-validation and with the final parameter sets.

PC x CFG	PMU	n sites	n obs	Bias k-fold	Bias final parameter sets	Final absolute bias value smaller?
NFERT x ALL CFGs	243.3	13	281	63.4	61.3	Yes
NFERT x NON_LEGUME_C4	410.5	9	139	27.2	27.8	No
NFERT x NON_LEGUME_C3	178.5	4	185	-11.3	-15.1	No
DISTU x NON_LEGUME_C4	423.1	11	127	-25.1	-25.1	Same
DISTU x COTTON	209.9	4	50	-76.0	-65.8	Yes
DISTU x NON_LEGUME_C3	488.9	7	99	-84.7	-81.3	Yes
CROPP x NON_LEGUME_C4	238.3	12	210	-116.1	-116.8	No
CROPP x NON_LEGUME_C3	223.2	15	292	-123.0	-123.8	No
CROPP x LEGUME_C3	181.8	16	270	-132.1	-132.8	No
CROPP x COTTON	240.3	6	163	-140.8	-139.9	Yes
DISTU x LEGUME_C3	87.6	6	60	-146.7	-131.5	Yes

Table 48. Comparison of MSE and RMSE in each PC x CFG category during cross-validation and with the final parameter sets.

PC × CFG	MSE_k-fold (t C ha^{-2})^2	RMSE_k-fold (t C ha^{-2})	MSE_final parameter sets (t C ha^{-2})^2	RMSE_final parameter sets (t C ha^{-2})	Final RMSE smaller?
DISTU.NON_LEGUME_C4	26.22 ± 6.42	5.09 ± 0.56	26.34 ± 6.55	5.1 ± 0.57	No
DISTU.NON_LEGUME_C3	17.51 ± 4.24	4.17 ± 0.4	17.04 ± 1.72	4.12 ± 0.2	Yes
DISTU.LEGUME_C3	16.1 ± 4.19	3.98 ± 0.5	16.25 ± 4.54	4 ± 0.52	No
NFERT.NON_LEGUME_C4	14.53 ± 2.64	3.8 ± 0.33	14.35 ± 1.94	3.78 ± 0.25	Yes
CROPP.NON_LEGUME_C3	14.4 ± 1.08	3.79 ± 0.14	14.26 ± 1.12	3.77 ± 0.14	Yes
CROPP.LEGUME_C3	14.1 ± 1.16	3.75 ± 0.15	14.09 ± 1.22	3.75 ± 0.16	Yes
NFERT.ALL	13.58 ± 2.09	3.67 ± 0.28	13.32 ± 1.67	3.64 ± 0.22	Yes
NFERT.NON_LEGUME_C3	12.95 ± 2.65	3.58 ± 0.36	12.56 ± 2.57	3.53 ± 0.35	Yes
CROPP.COTTON	12.54 ± 1.82	3.53 ± 0.24	12.66 ± 1.84	3.55 ± 0.25	No
CROPP.NON_LEGUME_C4	10.22 ± 0.83	3.19 ± 0.12	10.2 ± 0.72	3.19 ± 0.11	Yes
NFERT.LEGUME_C3	7 ± 1.99	2.62 ± 0.34	6.64 ± 1.4	2.56 ± 0.25	Yes
DISTU.COTTON	6.21 ± 0.79	2.49 ± 0.15	6.22 ± 0.77	2.49 ± 0.15	No
NFERT.COTTON	2.26 ± 0.1	1.5 ± 0.03	2.26 ± 0.09	1.5 ± 0.03	Yes

Table 49. Comparison of posterior 90% prediction interval coverage in each PC x CFG category during cross-validation and with the final parameter sets.

PC × CFG	**k-fold**		**Final parameter sets**	
	n in, n out	**% coverage**	**n in, n out**	**% coverage**
CROPP.NON_LEGUME_C4	209 in, 1 out	100	209 in, 1 out	100
DISTU.NON_LEGUME_C3	99 in, 0 out	100	98 in, 1 out	99
DISTU.COTTON	50 in, 0 out	100	50 in, 0 out	100
NFERT.NON_LEGUME_C4	139 in, 0 out	100	138 in, 1 out	99
NFERT.LEGUME_C3	73 in, 0 out	100	73 in, 0 out	100
NFERT.ALL	280 in, 1 out	100	280 in, 1 out	100
CROPP.LEGUME_C3	267 in, 3 out	99	265 in, 5 out	98
CROPP.NON_LEGUME_C3	290 in, 2 out	99	288 in, 4 out	99
CROPP.COTTON	161 in, 2 out	99	159 in, 4 out	98
DISTU.NON_LEGUME_C4	126 in, 1 out	99	127 in, 0 out	100
NFERT.NON_LEGUME_C3	184 in, 1 out	99	185 in, 0 out	100
DISTU.LEGUME_C3	59 in, 1 out	98	60 in, 0 out	100
All_PCs × ALL_CFGs	870 in, 5 out	99	866 in, 7 out	99

References

Bates, B.C., Campbell, E.P., 2001. A Markov Chain Monte Carlo Scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. Water Resour Res 37, 937–947. https://doi.org/10.1029/2000WR900363

Bikowski, J., Huisman, J.A., Vrugt, J.A., Vereecken, H., Van Kruk, J. Der, 2012. Integrated analysis of waveguide dispersed GPR pulses using deterministic and Bayesian inversion methods. Near Surface Geophysics 10, 641–652. https://doi.org/10.3997/1873-0604.2012041

Bista, P., Machado, S., Ghimire, R., Del Grosso, S.J., Reyes-Fox, M., 2016. Simulating soil organic carbon in a wheat–fallow system using the DAYCENT model. Agron J 108, 2554–2565. https://doi.org/10.2134/agronj2016.04.0202

Blasone, R.S., Vrugt, J.A., Madsen, H., Rosbjerg, D., Robinson, B.A., Zyvoloski, G.A., 2008. Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling. Adv Water Resour 31, 630–648. https://doi.org/10.1016/J.ADVWATRES.2007.12.003

Blevins, R.L., Thomas, G.W., Smith, M.S., Frye, W.W., Cornelius, P.L., 1983. Changes in soil properties after 10 years continuous non-tilled and conventionally tilled corn. Soil Tillage Res 3, 135–146.

https://doi-org.libproxy.unl.edu/10.1016/0167-1987(83)90004-1

Campbell, C.A., VandenBygaart, A.J., Zentner, R.P., McConkey, B.G., Smith, W., Lemke, R., Grant, B., Jefferson, P.G., 2007. Quantifying carbon sequestration in a minimum tillage crop rotation study in semiarid southwestern Saskatchewan. Can J Soil Sci 235–250. https://doi.org/10.4141/S06-018

Campbell, C.A., Zentner, R.P., 1997. Crop production and soil organic matter in long-term crop rotations in the semi-arid northern Great Plains of Canada, in: Paul, E.A., Paustian, K., Elliot, E.T., Cole, C.V. (Eds.), Soil Organic Matter in Temperate Agroecosystems - Long Term Experiments in North America. CRC Press, pp. 317–334. https://doi.org/10.1201/9780367811693

Campbell, C.A., Zentneri, R.P., Leyshoni, A.J., P LeysuoN A J, Z.R., 1983. First 12 years of a long-term crop rotation study in Southwestern Saskatchewan - Yields and quality of grain. Canadian Journal of Plant Science 63, 91–108.https://doi.org/10.4141/cjps83-009

Carvalhais, N., Reichstein, M., Seixas, J., James Collatz, G., Santos Pereira, J., Berbigier, P., Carrara, A., Granier, A., Montagnani, L., Papale, D., Rambal, S., José Sanz, M., Valentini, R., 2008. Implications of the carbon cycle steady state assumption for biogeochemical modeling performance and inverse parameter retrieval. Biogeochemical Cycles 22, 2007. https://doi.org/10.1029/2007GB003033

Clapp, C.E., Allmaras, R.R., Layese, M.F., Linden, D.R., Dowdy, R.H., 2000. Soil organic carbon and 13 C abundance as related to tillage, crop residue, and nitrogen fertilization under continuous corn management in Minnesota. Soil Tillage Res 55, 127–142. https://doi-org.libproxy.unl.edu/10.1016/S0167-1987(00)00110-0

Clark, M.S., Horwath, W.R., Shennan, C., Scow, K.M., 1998. Changes in soil chemical properties resulting from organic and low-input farming practices. Agron J 90, 662–671. https://doi.org/10.2134/agronj1998.00021962009000050016x

Collins, H.P., Blevins, R.L., Bundy, L.G., Christenson, D.R., Dick, W.A., Muggins, D.R., Paul, E.A., 1999. Soil carbon dynamics in corn-based agroecosystems: Results from carbon-13 natural abundance. Soil Science Society of America Journal 63, 584–591.https://doi-org.libproxy.unl.edu/10.2136/sssaj1999.03615995006300030022x

Christenson, D.R., 1997. Soil organic matter in sugar beet and dry bean cropping systems in Michigan, in: Paul E. A, Paustian, K., Elliot, E.T., Cole, C. V. (Eds.), Soil Organic Matter in Temperate Agroecosystems - Long-Term Experiments in North America. CRC Press, pp. 151–159. https://doi.org/10.1201/9780367811693

Cressie, N., Calder, C.A., Clark, J.S., Hoef, J.M. V, Wikle, C.K., 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecol. Appl. 19, 553–570. https://doi.org/https://doi.org/10.1890/07-0744.1

Dangal, S.R.S., Schwalm, C., Cavigelli, M.A., Gollany, H.T., Jin, V.L., Sanderman, J., 2022. Improving Soil Carbon Estimates by Linking Conceptual Pools Against Measurable Carbon Fractions in the DAYCENT Model Version 4.5. J Adv Model Earth Syst 14, e2021MS002622. https://doi.org/10.1029/2021MS002622

Del Grosso, S.J., Parton, W.J., Mosier, A.R., Walsh, M.K., Ojima, D.S., Thornton, P.E., 2006. DAYCENT National-scale simulations of nitrous oxide emissions from cropped soils in the United States. J Environ Qual 35, 1451–1460. https://doi.org/10.2134/JEQ2005.0160

Del Grosso, S.J., Ogle, S.M., Parton, W.J., Breidt, F.J., 2010. Estimating uncertainty in N2O emissions from U.S. cropland soils. Global Biogeochem Cycles 24, GB1009. https://doi.org/10.1029/2009GB003544

Del Grosso, S.J., Parton, W., Adler, P.R., Davis, S., Keogh, C., Marx, E., 2011. DayCent model simulations for estimating soil carbon dynamics and greenhouse gas fluxes from agricultural production systems, in: Liebig, M., Franzluebbers, A.J., Follett, R. (Eds.), Managing Agricultural Greenhouse Gases: Coordinated Agricultural Research through GRACEnet to Address Our Changing Climate. Elsevier Inc., New York, pp. 241–250. https://doi-org.libproxy.unl.edu/10.1016/C2010-0-66174-8

Dick, W.A., Edwards, W.M., McCoy, E.L., 1997. Continuous application of no-tillage to Ohio soils: Changes in crop yields and organic matter-related soil properties, in: Paul, E.A., Paustian, K., Elliot, E.T., Cole, C. V. (Eds.), Soil Organic Matter in Temperate Agroecosystems. Long-Term Experiments in North America. CRC Press, pp. 171–182. https://doi.org/10.1201/9780367811693

Dinerstein E, et al. 2017. An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm. BioScience 67(6): 534–545. doi:10.1093/biosci/bix014

Dolan, M.S., Clapp, C.E., Allmaras, R.R., Baker, J.M., Molina, J.A.E., 2006. Soil organic carbon and nitrogen in a Minnesota soil as related to tillage, residue and nitrogen management. Soil Tillage Res 89, 221–231. https://doi.org/10.1016/j.still.2005.07.015

Elliott, E.T., Burke, I.C., Monz, C.A., Frey, S.D., Paustian, K.H., Collins, H.P., Paul, E.A., Cole, C.V., Blevins, R.L., Frye, W.W., Lyon, D.J., Halvorson, A.D., Huggins, D.R., Turco, R.F., Hickman, M. V, 1994. Terrestrial carbon pools: Preliminary data from the Corn Belt and Great Plains regions, in: Doran, W., Coleman, D.C., Bezdicek, D.F., Stewart, B.A. (Eds.), SSSA Special Publications. Soil Science Society of America, pp. 179–191. https://doi-org.libproxy.unl.edu/10.2136/sssaspecpub35.c12

Ferreira, A.C. de B., Borin, A.L.D.C., Lamas, F.M., Bogiani, J.C., da Silva, M.A.S., Filho, J.L. da S., Staut, L.A., 2020. Soil carbon accumulation in cotton production systems in the Brazilian cerrado. Acta Sci Agron 42. https://doi.org/10.4025/actasciagron.v42i1.43039

Gelman, A. and D.B. Rubin. 1992. Inference from iterative simulation using multiple sequences. Statistical science 7.4: 457–472. https://www.jstor.org/stable/2246093

Ghimire, R., Machado, S., Rhinhart, K., 2015. Long-term crop residue and nitrogen management effects on soil profile carbon and nitrogen in wheat–fallow systems. Agron J 107, 2230–2240. https://doi.org/10.2134/agronj14.0601

Gurung, R.B., Ogle, S.M., Breidt, F.J., Williams, S.A., Parton, W.J., 2020. Bayesian calibration of the DayCent ecosystem model to simulate soil organic carbon dynamics and reduce model uncertainty. Geoderma 376, 114529. https://doi.org/10.1016/J.GEODERMA.2020.114529

Halvorson, A., Reule, C., 2007. How to enhance soil organic carbon sequestration. Fluid Journal 17–19. https://fluidfertilizer.org/wp-content/uploads/2016/05/57P17-19.pdf

Halvorson, A.D., Follett, R.F., Reule, C.A., Del Grosso, S., 2009. Soil organic carbon and nitrogen sequestration in irrigated cropping systems of the Central Great Plains, in: Lal, R., Follet, R.F. (Eds.), Soil Carbon Sequestration and the Greenhouse Effect - SSSA Special Publication. Wiley, pp. 141–157. https://doi.org/10.2136/sssaspecpub57.2ed.c9

Halvorson, A.D., Schlegel, A.J., 2012. Crop rotation effect on soil carbon and nitrogen stocks under limited irrigation. Agron J 104, 1265–1273. https://doi-org.libproxy.unl.edu/10.2134/agronj2012.0113

Halvorson, A.D., Wienhold, B.J., Black, A.L., 2002. Tillage, Nitrogen, and Cropping System Effects on Soil Carbon Sequestration. Soil Science Society of America Journal 66, 906–912. https://doi-org.libproxy.unl.edu/10.2136/sssaj2002.9060

Hammerbeck, A.L., S.J. Stetson, S.L. Osborne, T.E. Schumacher, and J.L. Pikul Jr. 2012. Corn residue removal impact on soil aggregates in a No-till corn/soybean rotation. Soil Sci. Soc. Am. J. https://doi-org.libproxy.unl.edu/10.2136/sssaj2011.0421

Hartman, M.D., W.J. William, S.J. Del Grosso, M. Easter, et. al. 2021. The Daily Century Ecosystem, Soil Organic Matter, Nutrient Cycling, Nitrogen Trace Gas, and Methane Model. User Manual, Scientific Basis, and Technical Documentation. Natural Resource Ecology Laboratory Colorado State University. Fort Collins, Colorado. https://storage.googleapis.com/public-sourcecode-access/comet-farm-API-documentation-repo.zip

Harville, D.A. 1977. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. Journal of the American Statistical Association, 72(358):320-338.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109. https://doi.org/10.1093/BIOMET/57.1.97

Hinnell, A.C., Ferr, T.P.A., Vrugt, J.A., Huisman, J.A., Moysey, S., Rings, J., Kowalsky, M.B., 2010. Improved extraction of hydrologic information from geophysical data through coupled hydrogeophysical inversion. Water Resour Res 46, 0–40. https://doi.org/10.1029/2008WR007060

Hoeting, J.A., 2009. The importance of accounting for spatial and temporal correlation in analyses of ecological data. Ecol. Appl. 19, 574–577. https://doi.org/https://doi.org/10.1890/08-0836.1

Houska, T., Kraft, P., Chamorro-Chavez, A., Breuer, L., 2015. SPOTting Model Parameters Using a Ready-Made Python Package. PLoS One 10, e0145180. https://doi.org/10.1371/JOURNAL.PONE.0145180

Hulugalle, N.R., Weaver, T.B., Finlay, L.A., Heimoana, V., 2013. Soil organic carbon concentrations and storage in irrigated cotton cropping systems sown on permanent beds in a Vertosol with restricted subsoil drainage. Crop Pasture Sci 64, 799–805. https://doi.org/10.1071/CP12374

Ismail, I., Blevins, R.L., Frye, W.W., 1994. Long-Term no-tillage effects on soil properties and continuous corn yields. Soil Science Society of America Journal 58, 193–198. https://doi-org.libproxy.unl.edu/10.2136/sssaj1994.03615995005800010028x

IPCC. 2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories. Intergovernmental Panel on Climate Change, Switzerland, 2019. isbn: 978-4-88788-232-4.

Jarecki, M.K., Lal, R., 2005. Soil organic carbon sequestration rates in two long-term no-till experiments in Ohio. Soil Sci 170, 280–291. https://doi.org/10.1097/00010694-200504000-00005

Jiang, G., Zhang, W., Xu, M., Kuzyakov, Y., Zhang, X., Wang, J., Di, J., Murphy, D.V., 2018. Manure and mineral fertilizer effects on crop yield and soil carbon sequestration: A meta-analysis and modeling across China. Global Biogeochem. Cycles. https://doi.org/10.1029/2018GB005960.

Jin, V.L., Schmer, M.R., Wienhold, B.J., Stewart, C.E., Varvel, G.E., Sindelar, A.J., Follett, R.F., Mitchell, R.B., Vogel, K.P., 2015. Twelve years of stover removal increases soil erosion potential without impacting yield. Soil Science Society of America Journal 79, 1169–1178. https://doi.org/10.2136/sssaj2015.02.0053

Jin, V.L. and G. Varvel. All; 2001–2016. ver. GPSR NATRES. Fort Collins, CO: USDA-ARS Natural Resources Database. Project: NEMEIRR. 2018. url: https://gpsr.ars.usda.gov/NEMEIRR.

Kong, A.Y.Y., Six, J., Bryant, D.C., Denison, R.F., van Kessel, C., 2005. The relationship between carbon input, aggregation, and soil organic carbon stabilization in sustainable cropping systems. Soil Science Society of America Journal 69, 1078–1085. https://doi.org/10.2136/sssaj2004.0215

Kow, W.Y., Khong, W.L., Chin, Y.K., Saad, I., Teo, K.T.K., 2012. “Enhancement of Markov chain Monte Carlo convergence speed in vehicle tracking using genetic operator. In: 2012 Fourth International Conference on Computational Intelligence, Modeling and Simulation (CIMSiM), pp. 270e275

Laloy, E., Linde, N., Vrugt, J.A., 2012. Mass conservative three-dimensional water tracer distribution from Markov chain Monte Carlo inversion of time-lapse ground-penetrating radar data. Water Resour. Res. 48, W07510.

https://ui.adsabs.harvard.edu/link_gateway/2012WRR....48.7510L/doi:10.1029/2011WR011238

Mathers, C., Black, C.K., Segal, B.D., Gurung, R.B., Zhang, Y., Easter, M.J., Williams, S., Motew, M., Campbell, E.E., Brummitt, C.D., Paustian, K., Kumar, A.A. 2023. Validating DayCent-CR for cropland soil carbon offset reporting at a national scale. Geoderma 438 116647. https://doi.org/10.1016/j.geoderma.2023.116647.

Minasny, B., Vrugt, J.A., McBratney, A.B., 2011. Confronting uncertainty in model based geostatistics using Markov chain Monte Carlo simulation. Geoderma 163, 150e622. http://dx.doi.org/10.1016/j.geoderma.2011.03.011.

Mitchell, J.P., Shrestha, A., Horwath, W.R., Southard, R.J., Madden, N., Veenstra, J., Munk, D.S., 2015. Tillage and cover cropping affect crop yields and soil carbon in the San Joaquin valley, California. Agron J 107, 588–596. https://doi.org/10.2134/agronj14.0415

Mitchell, J.P., Shrestha, A., Mathesius, K., Scow, K.M., Southard, R.J., Haney, R.L., Schmidt, R., Munk, D.S., Horwath, W.R., 2017. Cover cropping and no-tillage improve soil health in an arid irrigated cropping system in California's San Joaquin Valley, USA. Soil Tillage Res 165, 325–335. https://doi.org/10.1016/j.still.2016.09.001

Monreal, C.M., Janzen, H.H., 1993. Soil organic-carbon dynamics after 8O years of cropping a Dark Brown Chernozem. Can J Soil Sci 73, 133–136. https://doi.org/10.4141/cjss93-014

Necpálová, M., Anex, R.P., Fienen, M.N., Del Grosso, S.J., Castellano, M.J., Sawyer, J.E., Iqbal, J., Pantoja, J.L., Barker, D.W., 2015. Understanding the DayCent model: Calibration, sensitivity, and identifiability through inverse modeling. Environmental Modelling & Software 66, 110–130. https://doi.org/10.1016/J.ENVSOFT.2014.12.011

Ogle, S.M., Alsaker, C., Baldock, J., Bernoux, M., Breidt, F.J., McConkey, B., Regina, K., Vazquez-Amabile, G.G., 2019. Climate and soil characteristics determine where notill management can store carbon in soils and mitigate greenhouse gas emissions. Sci.Rep. https://doi.org/10.1038/s41598-019-47861-7.

Ogle, S.M., Breidt, F.J., Easter, M., Williams, S., Paustian, K., 2007. An empirically based approach for estimating uncertainty associated with modelling carbon sequestration in soils. Ecol Modell 205, 453–463. https://doi.org/10.1016/J.ECOLMODEL.2007.03.007

Ogle, S.M., Jay Breidt, F., Easter, M., Williams, S., Killian, K., Paustian, K., 2010. Scale and uncertainty in modeled soil organic carbon stock changes for US croplands using a process-based model. Glob Chang Biol 16, 810–822. https://doi.org/10.1111/J.1365-2486.2009.01951.X

Olson, K.R., Ebelhar, S.A., Lang, J.M., 2010. Cover crop effects on crop yields and soil organic carbon content. Soil Sci 175, 89–98. https://doi.org/10.1097/SS.0b013e3181cf7959

Olson DM et al. 2001. Terrestrial ecoregions of the world: a new map of life on Earth. Bioscience 51(11): 933-938. https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2

Parton, W.J., Holland, E.A., Del Grosso, S.J., Hartman, M.D., Martin, R.E., Mosier, A.R., Ojima, D.S., Schimel, D.S., 2001. Generalized model for NO x and N2O emissions from soils. Journal of Geophysical Research: Atmospheres 106, 17403–17419. https://doi.org/10.1029/2001JD900101

Pérez, F., Granger, B.E., 2007. IPython: A system for interactive scientific computing. Comput Sci Eng 9, 21–29. https://doi.org/10.1109/MCSE.2007.53

Pimentel, D., Hepperly, P., Hanson, J., Douds, D., Seidel, R., 2005. Environmental, energetic, and economic comparisons of organic and conventional farming systems. Bioscience 5, 573–582. https://doi.org/10.1641/0006-3568(2005)055[0573:EEAECO]2.0.CO;2

Pinheiro, J., Bates, D., Härdle, W., 2010. Mixed-Effects Models in S and S-PLUS. Springer, New York, GERMANY.

Powers, W., Auvermann, B., Cole. N. A., Gooch, C., Grant, R., Hatfield, J., Hunt, P., Johnson, K., Leytem, A., Liao, W., Powell, J.M., 2014. Quantifying greenhouse gas sources and sinks in animal production systems, in: Eve, M., Pape, D., Flugge, M., Steele, R., Man, D., Riley-Gilbert, M., Biggar, S. (Eds.), Quantifying Greenhouse Gas Fluxes in Agriculture and Forestry: Methods for Entity-Scale Inventory. United States Department of Agriculture, Washington, pp. 5–75. https://www.usda.gov/sites/default/files/documents/USDATB1939_07072014.pdf

Rasmussen, P.E., Smiley, R.W., 1997. Soil carbon and nitrogen change in long-term agricultural experiments at Pendleton, Oregon, in: Paul, E.A., Paustian, K., Elliot, E.T., Cole, C.V. (Eds.), Soil Organic Matter in Temperate Agroecosystems - Long Term Experiments in North America. CRC Press, pp. 353–360. https://doi.org/10.1201/9780367811693-26

Roberts, D.R., V. Bahn, S. Ciuti, M.S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, et al., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. In: Ecography 40.8, 913–929. https://doi.org/10.1111/ecog.02881

Rochester, I.J., 2011. Sequestering carbon in minimum-tilled clay soils used for irrigated cotton and grain production. Soil Tillage Res 112, 1–7. https://doi.org/10.1016/j.still.2010.10.012

Sadegh, M., Vrugt, J.A., 2013. "Approximate Bayesian Computation in hydrologic modeling: equifinality of formal and informal approaches. Hydrol. Earth Syst. Sci. - Discuss. 10, 4739e4797. http://dx.doi.org/10.5194/hessd-10-4739-2013.

Sadegh, M., Vrugt, J.A., 2014. Approximate Bayesian Computation using Markov Chain Monte Carlo simulation: DREAM(ABC). Water Resources Research 50, 6767–6787. https://doi.org/10.1002/2014WR015386

Sainju, U.M., Whitehead, W.F., Singh, B.P., 2005. Carbon accumulation in cotton, sorghum, and underlying soil as influenced by tillage, cover crops, and nitrogen fertilization. Plant Soil 273, 219–234. https://doi.org/10.1007/s11104-004-7611-9

Salinas-Garcia, J.R., Hons, F.M., Matocha, J.E., 1997. Long-term effects of tillage and fertilization on soil organic matter. Dynamics. Soil Sci. Soc. Am. J. https://doi.org/10.2136/sssaj1997.03615995006100010023x.

Saltelli, A., Tarantola, S., Chan, K.P.S., 1999. A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output. Technometrics 41, 39–56. https://doi.org/10.1080/00401706.1999.10485594

Saltelli, A., S. Tarantola, and F. Campolongo. 2000. Sensitivity analysis as an ingredient of modeling. *Statistical Science* 15(4): 377–395. https://www.jstor.org/stable/2676831

Salvatier, J., Wiecki, T. V., Fonnesbeck, C., 2016. Probabilistic programming in Python using PyMC3. PeerJ Comput Sci 2016, e55. https://doi.org/10.7717/PEERJ-CS.55/FIG-7

Senapati, N., Hulugalle, N.R., Smith, P., Wilson, B.R., Yeluripati, J.B., Daniel, H., Ghosh, S., Lockwood, P., 2014. Modelling soil organic carbon storage with RothC in irrigated Vertisols under cotton cropping systems in the sub-tropics. Soil Tillage Res 143, 38–49. https://doi.org/10.1016/j.still.2014.05.009

Shafii, M., Tolson, B., Matott, L.S., 2014. Uncertainty-based multi-criteria calibration of rainfall-runoff models: A comparative study. Stochastic Environmental Research and Risk Assessment 28, 1493–1510. https://doi.org/10.1007/S00477-014-0855-X/METRICS

Sindelar, A. J., Schmer, M. R., Jin, V. L., Wienhold, B. J., and Varvel, G. E. 2016. Crop rotation affects corn, grain sorghum, and soybean yields and nitrogen recovery. Agronomy Journal 108:1592-1602.https://doi-org.libproxy.unl.edu/10.2134/agronj2016.01.0005

Stetson, S.J., S.L. Osborne, T.E. Schumacher, A. Eynard, G. Chilom, J. Rice, K.A. Nichols, and J.L. Pikul Jr. 2012. Corn residue removal impact on topsoil organic carbon in a corn-soybean rotation. Soil Sci. Soc. Am. J. https://doi-org.libproxy.unl.edu/10.2136/sssaj2011.0420

Sun, W., Canadell, J.G., Yu, Lijun, Yu, Lingfei, Zhang, W., Smith, P., Fischer, T., Huang, Y., 2020. Climate drives global soil carbon sequestration and crop yield changes under conservation agriculture. Glob. Chang. Biol. https://doi.org/10.1111/gcb.15001.

Sun, X.L., Wu, S.C., Wang, H.L., Zhao, Y.G., Zhang, G.L., Man, Y.B., Wong, M.H., 2013. Dealing with spatial outliers and mapping uncertainty for evaluating the effects of urbanization on soil: A case study of soil pH and particle fractions in Hong Kong. Geoderma 195–196, 220–233. https://doi.org/10.1016/J.GEODERMA.2012.11.017

Thornton M, R. Shrestha, Y. Wei, P.E. Thornton, S. Kao, and B.E. Wilson. 2020. Daymet: Daily SurfSWFe Weather Data on a 1-km Grid for North America; Version 4, ORNL DASWF; Oak Ridge; Tennessee; USA,;https://doi,org/10,3334/ORNLDASWF/1840

U.S. Environmental Protection Agency. 2013. Inventory of U.S. greenhouse gas emissions and sinks: 1990–2011. 1200 Pennsylvania Ave., NW, Washington DC. http://www.epa.gov/climatechange/ghgemissions/usinventoryreport.html. Accessed 12 Oct. 2022

U.S. Environmental Protection Agency. 2020. Inventory of U.S. greenhouse gas emissions and sinks: 1990–2018. 1200 Pennsylvania Ave., NW, Washington DC. http://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks-1990-2018.. Accessed 12 Oct. 2022

Varvel, G.E., 2006. Soil Organic Carbon Changes in Diversified Rotations of the Western Corn Belt. Soil Science Society of America Journal 70, 426–433. https://doi.org/10.2136/sssaj2005.0100

Varvel, G.E., Riedell, W.E., Deibert, E., McConkey, B., Tanaka, D.L., Vigil, M.F., and Schwartz, R.C. 2006. Great Plains Cropping System Studies for Soil Quality Assessment. Renewable Agriculture and Food Systems. 21(1):3-14. https://doi.org/10.1079/RAF2005121

Veenstra, J.J., Horwath, W.R., Mitchell, J.P., Munk, D.S., 2006. Conservation tillage and cover cropping influence soil properties in San Joaquin Valley cotton-tomato crop. Calif Agric (Berkeley) 60, 146–153.

https://escholarship.org/content/qt38t4n0r5/qt38t4n0r5_noSplash_516a33a25e05de7e88ace3584e02d446.pdf?t=krnosa

Vrugt, J.A., Braak, C.J.F. ter, Clark, M.P., Hyman, J.M., Robinson, B.A., 2008. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. Water Resour Res 44, 0–09. https://doi.org/10.1029/2007WR006720

Vrugt, J.A., ter Braak, C.J.F., Gupta, H. V., Robinson, B.A., 2009. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? Stochastic Environmental Research and Risk Assessment 23, 1011–1026. https://doi.org/10.1007/S00477-008-0274-Y/METRICS

Vrugt, J.A., Ter Braak, C.J.F., 2011. DREAM(D): An adaptive Markov Chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems. Hydrol Earth Syst Sci 15, 3701–3713. https://doi.org/10.5194/HESS-15-3701-2011

Vrugt, J.A., 2016. Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. Environmental Modelling & Software 75, 273–316. https://doi.org/10.1016/J.ENVSOFT.2015.08.013

Wang, X., X. He, J.R. Williams, R.C. Izaurralde, J. D. Atwood. 2005. Sensitivity and uncertainty analysis of crop yields and soil organic carbon simulation with EPIC. Transactions of the ASAE. Vol. 48(3): 1041−1054. http://dx.doi.org/10.13031/2013.18515

Wang, X., J. Jeong, S. Parka , X. Zhang, J. Gao, N.S. Silvero. 2023. DayCent-CUTE: A Global Sensitivity, Auto-calibration, and Uncertainty Analysis Tool for DayCent. Environmental Modelling and Software. Under publication. ENSO_105832

Wang, X., J.R. Williams, P.W. Gassman, C. Baffaut, R.C. Izaurralde, J. Jeong, J.R. Kiniry. 2012. EPIC and APEX: model use, calibration, and validation Transactions of the ASAE. Vol. 55(4): 1447−1462. http://doi: 10.13031/2013.42253

Wang, X., S.R. Potter, J.R. Williams, J.D. Atwood, T. Pitts. 2006. Sensitivity analysis of APEX for national assessment. Transactions of the ASAE. Vol. 49(3): 679−688. http://dx.doi.org/10.13031/2013.20487

Wegner, B.R., Chalise, K.S., Singh, S., Lai, L., Abagandura, G.O., Kumar, S., Osborne, S.L., Lehman, R.M., Jagadamma, S., 2018. Response of soil surface greenhouse gas fluxes to crop residue removal and cover crops under a corn–soybean rotation. J Environ Qual 47, 1146–1154. https://doi.org/10.2134/jeq2018.03.0093

Williams, J.R., 1995. The EPIC model, in: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Colorado, pp. 909–1000. https://www-cabdirect-org.libproxy.unl.edu/cabdirect/abstract/19961904785

Wolf, K.M. , E. Torbert, D. Bryant, M. Burger, R. F. Denison, I. Herrera, J. Hopmans, W. Horwath, S. Kaffka, et al., 2018. The century experiment: the first twenty years of UC Davis' Mediterranean agroecological experiment. https://doi.org/10.1002/ecy.2105

Wutzler, T., Reichstein, M., 2007. Soils apart from equilibrium - Consequences for soil carbon balance modelling. Biogeosciences 4, 125–136. https://doi.org/10.5194/BG-4-125-2007

Yeluripati, J.B., van Oijen, M., Wattenbach, M., Neftel, A., Ammann, A., Parton, W.J., Smith, P., 2009. Bayesian calibration as a tool for initialising the carbon pools of dynamic soil models. Soil Biol Biochem 41, 2579–2583. https://doi.org/10.1016/J.SOILBIO.2009.08.021

Y. Zhang, M. Arabi, and K. Paustian. 2020. Analysis of parameter uncertainty in model simulations of irrigated and rainfed agroecosystems. Environmental Modelling & Software 126, 104642. https://doi.org/10.1016/j.envsoft.2020.104642

Zhang, X., Izaurralde, R.C., Arnold, J.G., Williams, J.R., Srinivasan, R., 2013. Modifying the Soil and Water Assessment Tool to simulate cropland carbon flux: Model development and initial evaluation. Science of The Total Environment 463–464, 810–822. https://doi.org/10.1016/J.SCITOTENV.2013.06.056

Zhang, Y., Suyker, A., Paustian, K., 2018. Improved crop canopy and water balance dynamics for agroecosystem modeling using DayCent. Agron J 110, 511–524. https://doi.org/10.2134/AGRONJ2017.06.0328