Modeling Soil Carbon and Greenhouse Gas Emissions

Identifying challenges and advancing guidance for using process-based models in soil emission reduction and removal projects

Environmental Defense Fund

Modeling Soil Carbon and Greenhouse Gas Emissions

Identifying challenges and advancing guidance for using process-based models in soil emission reduction and removal projects

Authors

Jocelyn M. Lavallee
Environmental Defense Fund

Brian G. McConkey
Viresco Solutions

Emily Kyker-Snowman
Carbon Direct

Michael C. Dietze
Boston University

Matthew Tom Harrison and Ke Liu
Tasmanian Institute of Agriculture, University of Tasmania

Ward Smith and Brian Grant
Agriculture and Agri-Food Canada

Senani Karunaratne
Commonwealth Scientific and Industrial Research Organisation

Alison Marklein
Terradot

How to cite this report: J.M. Lavallee, B.G. McConkey, E. Kyker-Snowman, M.C. Dietze, M.T. Harrison, W. Smith, B. Grant, S. Karunaratne, K. Liu, and A. Marklein. 2024. Modeling Soil Carbon and Greenhouse Gas Emissions: Identifying challenges and advancing guidance for using process-based models in soil emission reduction and removal projects. Environmental Defense Fund, New York, New York. https://www.edf.org/modeling-soil-carbon-emissions

Acknowledgments

This report was supported through gifts to Environmental Defense Fund from the Bezos Earth Fund, King Philanthropies, and Arcadia, a charitable fund of Lisbet Rausing and Peter Baldwin. We thank Emily Oldfield (Environmental Defense Fund); Jonathan Sanderman (Woodwell Climate Research Center); Dan Kane (TerraCarbon); Eric Potash (University of Illinois Urbana-Champaign); Viridiana Alcantara-Shivapatham (Verra); McKenzie Smith (Climate Action Reserve); and Hamze Dokoohaki and Michelle Schmidt (Indigo) for reviewing and providing feedback on this document.

About Environmental Defense Fund

Environmental Defense Fund is one of the world's leading environmental nonprofit organizations. Guided by science and economics, EDF finds practical and lasting solutions to the most serious environmental problems.

Table of Contents

Executive Summary

Agricultural soils globally represent an important opportunity for greenhouse gas (GHG) reductions and removals which can contribute to meeting climate goals. However, changes in GHG emissions and soil organic carbon (SOC) stocks in agricultural soils are costly to measure directly at scale, and GHG and SOC accounting frameworks for these types of projects increasingly rely on the use of process-based biogeochemical models.

The application of process-based models to measurement, monitoring, reporting, and verification (MMRV) schemes in large-scale contexts, such as carbon markets and supply chain accounting, has highlighted several challenges, and their use has raised important questions about — and in some cases mistrust in — the process-based model quantification process. Inconsistencies in approaches between modeling groups and individual projects, including modeling workflows and uncertainty quantification methods, contributes to the confusion and variation around modeled results. This report helps guide through these challenges by providing important context and recommendations for process-based model implementation in agricultural soil GHG and SOC projects to increase consistency, transparency, and confidence in this integral portion of the MMRV process.

First, the report describes the main project steps in which process-models are used: calibration, validation, project predictions, and true-up as well as the processes of scaling process-model predictions to the project level, estimating project-level uncertainty, and the modeling workflow (which includes initialization). In doing so the report shines light on the realities and shortcomings of process-model approaches for soil GHG and SOC projects that must be considered when designing protocols to guide rigorous and transparent process-model usage.

Second, the report highlights recommendations for improvements to existing protocols for process-model use in soil GHG and SOC projects. These recommendations are based on expert knowledge and direct experience of the authors of the report, as well as thorough review of existing protocols (including Verra's VM0042 and VMD0053; Climate Action Reserve's Soil Enrichment Protocol and related Requirements and Guidance for Model Calibration, Validation, Uncertainty, and Verification; and the "Estimating soil organic carbon sequestration using measurement and models method" under the Australian Carbon Credit Unit Scheme) and publicly-available model validation reports. In addition to these main recommendations, several other recommendations are made throughout the report and are summarized in a convenient table format at the end of the main report text.

Finally, the report includes two appendices with critical background information and further discussion of select modeling workflow components (including model initialization), and spatial and temporal relationships in measurements and model prediction errors. These topics were found to be missing or underrepresented in current protocols, current

applications, and discussions of process-model usage in soil GHG and SOC projects; yet both topics are critically important to ensuring consistency and rigor in process-model use in this context and many of the recommendations in the report relate directly to them.

Main recommendations for improved process-based model guidance

1. The modeling workflow, encompassing all aspects of how a model is set-up and run, and results are processed, should be kept as consistent as possible within a project. Differences in modeling workflow between validation and other project steps, such as project prediction, introduce additional uncertainty and present opportunities for gaming. Consistency across all project steps should be ensured as part of the validation process; critically, it is not only the model version and parameter set that must be validated, but the entire modeling workflow.
2. The data used for validation ultimately determines the context for which the model can be validated and reliably applied. Validation data must be sufficient to represent the project in scope and coverage, not only in terms of key biophysical variables such as soil texture, but also in the spatial distances and time spans of observations.
3. Model prediction error increases with the time span over which predictions are being made, and this can be accounted for during the uncertainty estimation process. However, it must be done carefully to ensure conservatism and that any assumptions made are supported by the validation data. For example, plots of error versus time along with distributions of timescales of the validation measurements could be provided and used by expert reviewers to assess whether claimed relationships between model prediction error and time are reasonable and conservative.
4. Current approaches and protocols often assume independence of measurement and model errors — which directly affects several project steps including how project-level uncertainty is calculated — without supporting evidence. Many factors can contribute to correlated measurement and model errors and because they are often spatially structured (e.g., at the field level or farm level), errors should be assumed to have spatial dependence unless sufficient evidence to support the assumption of independence is provided. Given that it may be hard to reach agreement on what "sufficient evidence" entails or how best to standardize requirements across projects, research on, and demonstrations of, potential approaches are needed.
5. Systematic model error (bias) is common for process-based models and must be handled thoughtfully during the validation process to avoid situations where it manifests differently in the project than during validation. If systematic error is not detected during validation but becomes highly relevant during project modeling, it could result in overestimates of GHG emission reductions or removals for the project. Categorical groupings of validation data to assess systematic error under different contexts is useful and required by most protocols, but more innovative approaches could be used to borrow strength across crop functional types, soil properties, climatic regions, and other key variables to assess systematic errors across continuous gradients.
6. A shared benchmarking platform could be used to validate models against a common dataset, with many benefits including increasing transparency in the process; improving confidence in the performance and utility of different models; and reducing potential for gaming via cherry-picking of data or using the same data or sites for calibration and validation (independence of calibration and validation data is a requirement of protocols, but in practice this can be difficult to enforce). Some efforts in this area are in their beginning stages and should be supported, coordinated, and potentially combined.

Introduction

Agricultural soils worldwide have been identified as having major potential for greenhouse gas (GHG) reductions and removals because they cover 40% of the earth's surface [1], are already actively managed, are major sources of the potent GHGs N2O and CH4 [2–4] and are depleted in soil organic carbon (SOC), some of which may be restored [5]. Changes to management practices such as efficient fertilizer use and water management can reduce emissions of N2O and CH4, both of which account for a significant component of current warming and need to be drastically reduced to meet climate targets [6]. And in many cases, SOC stocks can be increased through management changes such as cover cropping, perennialization, and agroforestry [7–14], many of which have important co-benefits for agricultural productivity [15–17], soil health [18], and ecosystem services [19]. With this in mind, various schemes such as voluntary carbon markets have emerged to encourage practice changes on agricultural soils while quantifying associated GHG benefits. This quantification, termed measurement, monitoring, reporting and verification (MMRV), must have the ability to isolate management-induced changes in GHG emissions and/or SOC stocks from other sources of variation such as climate and extreme weather events or changes that would have occurred regardless of the project initiation (i.e., the counterfactual). High-quality MMRV is critical to demonstrating the realized climate impacts of a given GHG mitigation project and is also the subject of considerable research, development, and debate [20–23].

Well-designed direct measurement over time remains the most reliable method of change quantification for agricultural soil GHG fluxes and SOC stocks, but it has several drawbacks [24]. Direct measurements are costly, time-intensive, and static in time (i.e., not forward-looking). At best, measurements can be used to assess past changes, but in reality this is often complicated or impossible because past measurements may not meet the current bar for high quality (especially considering that process understanding and measurement technology continuously evolve), or the goals of measurement campaigns may have shifted over time resulting in mismatched sampling designs that aren't directly comparable across time points. Further, the slow nature of SOC change coupled with its high spatial heterogeneity means that it is typically infeasible (sampling effort, cost) to detect any changes via direct measurement in under 5-10 years, especially at small scales [25]. Finally, the counterfactual "baseline" scenario, or what would have happened without the change in management practice spurred by the project, which is required for complete quantification of project impacts, cannot be directly measured.[1] That said, when direct sampling campaigns are well-designed and executed, they are irreplaceable and should remain an integral part of

[1] While one cannot directly measure the counterfactual scenario in a strict sense, measurements of representative areas under "business-as-usual" management can be used as baselines. However, finding these areas to use as baselines, which ideally match project sites in terms of key characteristics (e.g., soil type, climate, crop type, practices other than what the project focuses on etc.), can be difficult and time-intensive. See True-up section for further discussion.

any high-quality MMRV scheme. Yet the costliness, inability to look ahead, relatively long time-periods required to detect change of SOC in particular, and difficulty of capturing counterfactual baselines leave room for other tools besides direct sampling to help in sampling optimization, project planning, and project impact quantification (especially in the short term).

Box 1.

Guiding principles relevant for process-based modeling and GHG mitigation projects

Consistency

Standardization across protocols, between and within projects improves comparability and reliability of claims

Transparency

Publicly-accessible, comprehensive information enables scrutiny and instills confidence in claims

Conservatism

Decreases the risk of unanticipated adverse outcomes such as overcrediting and increases confidence in the reliability of claims

Robustness

Ensures model performance is consistent across different systems or applicaitons

Anti-gaming

Guards against purposeful manipulation of loopholes in standards, protocols, or oversight mechanisms to achieve a desired outcome

Process-based models can help fill this gap. Such models can make forward-looking predictions in a context-specific way that accounts for anticipated conditions on the ground including weather and management schedules. For SOC, they can also be used to make predictions for short timescales over which direct sampling would not detect changes, making them particularly useful for programs that make annual payments or otherwise rely on estimates of change in the short term. They can be used to simulate baseline scenarios in the absence of certain management changes while keeping all else consistent with the project simulation. Perhaps most enticingly of all, they can be applied at relatively low costs and are seen as a cost-effective means to lessen or optimize sampling for direct field validation. The benefits of process-models make them extremely popular for agricultural soil GHG emission reduction and removal projects; these benefits include being able to predict whether proposed management changes will be beneficial; inform farmer and land manager decision-making; inform on ideal project locations and scales; and determine payments to farmers and land managers on timescales that better align with their costs for implementing practice changes (e.g., annually).

However, models are imperfect and they must be used with caution and consideration. Increasingly, their use for agricultural soil GHG emission reduction or removal projects has come under scrutiny and many have criticized what they see as inadequate validation and overreliance on highly uncertain model results [26–29]. This has led to calls for increased transparency, rigor, and consistency in how models are being applied for these types of projects. It has also highlighted knowledge gaps around best practices for model use, especially in light of the need for conservatism and transparency in MMRV and GHG accounting to avoid perverse incentives, undesirable GHG outcomes (e.g., overcrediting), and loss of confidence in the various mechanisms being put forward to encourage agricultural soil GHG emission reduction or removal as a viable climate strategy.

In light of these issues, this report provides recommendations for improved approaches and research to address shortcomings and knowledge gaps in current modeling guidance. Specifically, we lay a foundation for shared understanding by providing background on key concepts and approaches in process-based modeling use in soil GHG emission reduction or removal projects; articulate difficulties and hurdles that currently limit modeling

capabilities and adherence to aspirational standards; and identify gaps and shortcomings in current guidance and protocols that are at odds with foundational principles of GHG mitigation projects (Box 1); and provide additional guidance and recommendations where possible to remedy these shortcomings, including for additional research[2].

Intended outcomes

The intended outcomes of this report are as follows:

1. Broaden understanding of key aspects of biogeochemical process-based model use in soil GHG MMRV protocols;
2. Improve consistency and comparability of implementations of biogeochemical process-based models for quantifying GHG changes across MMRV protocols;
3. Increase the quality of, and confidence in, process-based model implementation for producing scientifically credible GHG emission and SOC stock change estimates (and related uncertainty) using current data and knowledge; and
4. Encourage and enable research targeted toward knowledge gaps that currently limit abilities to achieve the above outcomes.

This is a rapidly developing area of research and development across academic, government, private and non-profit sectors with new discoveries, proposed approaches, and innovative ideas emerging at a dizzying pace. We see a clear need for accessible summation, explanation, and analysis of ongoing developments in process-modeling approaches and applications to assist interested stakeholders in understanding the current landscape, guide decision-makers in the agricultural sector, and generally serve as a starting point for constructive discussions. We hope that this work clarifies relevant issues and helps to focus the discourse, bring more interested parties into the discussion, illuminate urgent research needs, and ultimately improve the rigor and consistency of MMRV across all types of agricultural GHG mitigation projects.

Scope

This report focuses on process-based model usage in the context of agricultural GHG reduction or removal projects and related calculations of uncertainty (model prediction error and project-level uncertainty). The report does not explicitly consider statistical models, remote sensing applications, or machine-learning approaches, though any of these may be used in concert with process-based models to achieve the applications discussed here. For example, while there is currently no direct means of using remote sensing to measure changes in SOC, remote sensing can be invaluable in verifying that the management actions being credited are actually being undertaken and prescribing model inputs such as crop type, leaf traits, and phenology.

With regard to process-based model use in agricultural GHG reduction or removal projects, this report covers aspects of calibration, validation (evaluation in the context of MMRV protocols), uncertainty assessment, data considerations, modeling workflow (including initialization), and integration of direct measurements of a project after its initiation (i.e., "true-up" or "reconciliation"). We do not discuss in depth the broader aspects of soil GHG and SOC project design such as stratification, sampling schemes, measurement approaches, or baseline design, though these are critically important topics worthy of similar scrutiny and discussion. Further, we do not discuss data collection and quality assurance/control

[2] Discussions of foundational principles and high-quality carbon removals can be found at https://blogs.edf.org/climate411/2023/08/01/navigating-the-core-carbon-principles-and-the-landscape-of-guidance-toward-a-high-integrity-carbon-market/; https://icvcm.org/the-core-carbon-principles/; https://www.carbon-direct.com/solutions-remove

processes as the key considerations can be very application-specific and are typically covered in depth by protocols. Finally, this report focuses on the steps and key considerations in model application (i.e., after the model has been chosen), rather than the choice of which model to use, the strengths and weaknesses of different model structures, or the use of model ensembles. These are very important topics that are being actively explored elsewhere [e.g., 30–32].

The guidance and discussion herein is meant to be broadly applicable to any use of process-based models for soil GHG and SOC projects, including under different types of MMRV protocols, voluntary and regulatory markets, offsetting and within-supply-chain programs, emissions and removals, all appropriate biogeochemical process-based models, and all regions of the world.[3] However, given that there are significant limitations in data collection and availability in certain contexts (e.g., supply-chain accounting) and many areas of the world, adherence to these recommendations may not always be feasible. We do not necessarily recommend against approaches which may be the only option in such circumstances but aim to explain the limitations of those approaches to provide clarity and transparency in application, and avoid overconfidence in, or misuse of, modeled results.

This report is meant to complement existing protocols and guidance for process-based model application in soil GHG and SOC projects, rather than to act as a stand-alone protocol or guidance document. The recommendations in the report are based on expert knowledge and direct experience of the authors, as well as thorough review of existing protocols (such as Verra's VM0042 and VMD0053; Climate Action Reserve's Soil Enrichment Protocol and related Requirements and Guidance for Model Calibration, Validation, Uncertainty, and Verification; and the "Estimating soil organic carbon sequestration using measurement and models method" under the Australian Carbon Credit Unit Scheme) and publicly-available model validation reports.

Given that there has been significant progress and activity using process-based models for soil carbon crediting (i.e., "Scope 1"), we use these types of projects and the general characteristics of associated protocols [34, 35] as a starting place upon which this guidance is meant to build. While some of the details regarding protocol requirements and best practices for model use covered here may be seen as too onerous, data-intensive, or strict for other types of applications (e.g., "Insetting" or "Scope 3"), the overarching principles we discuss are applicable regardless of the implementation framework for GHG mitigation and SOC sequestration in agricultural soils. Similarly, we focus on SOC in several places as a way to provide more concrete examples and explanations; however, the principles generally apply to any process-based model output, including N_2O and CH_4.

[3] See [33] for details on different types of GHG emission reduction and removal programs.

The Project Steps

Background on the major steps in agricultural soil GHG projects involving process-based models

Soil GHG emission reduction or removal projects come in many different forms, from market-based projects that generate carbon credits, to pay-for-practice programs aiming to quantify any associated reductions in GHGs, and beyond [33]. Regardless of the broader goal or implementation mechanism of these projects, process-based models can be used to estimate GHG outcomes of changes in agricultural practices, and the basic steps in doing so are broadly consistent and are conceptualized here as the "project steps" (Fig. 1). These are calibration, validation, project modeling, and direct project measurements (i.e., "true-up").[4] During calibration, the set of model parameters is chosen which produces the best fit between model predictions and a set of observed values. Once calibrated, model performance is evaluated against a set of independent data, termed "validation" in many GHG mitigation protocols. Model validation is also the step during which model uncertainty (termed model prediction error) is assessed, based on the disagreement between model predictions and measurements at the same site(s). Once validated, the model is then used to make predictions for the project, including baseline scenarios where dynamic modeled baselines are used. Critically, the exact same model that was validated must be used for the project (note that later we recommend this be expanded to include the full modeling workflow). Finally, direct measurements may be made for the project to improve the model for use in the project and/or to verify project predictions.[5] These measurements will be conducted differently depending on the outcome of interest; in the case of SOC stocks, measurements may occur periodically after a prespecified amount of time (e.g., 5 years) has passed to allow detection of any changes. Data collection, selection, and processing is a critical component of all projects and supports multiple project steps, therefore we do not treat it as a separate step here. At any project step where the model is used, the modeling workflow (i.e., initialization, data inputs, pre- and post-processing) will be invoked and should be kept consistent throughout the steps (see Modeling Workflow section and Appendix A).

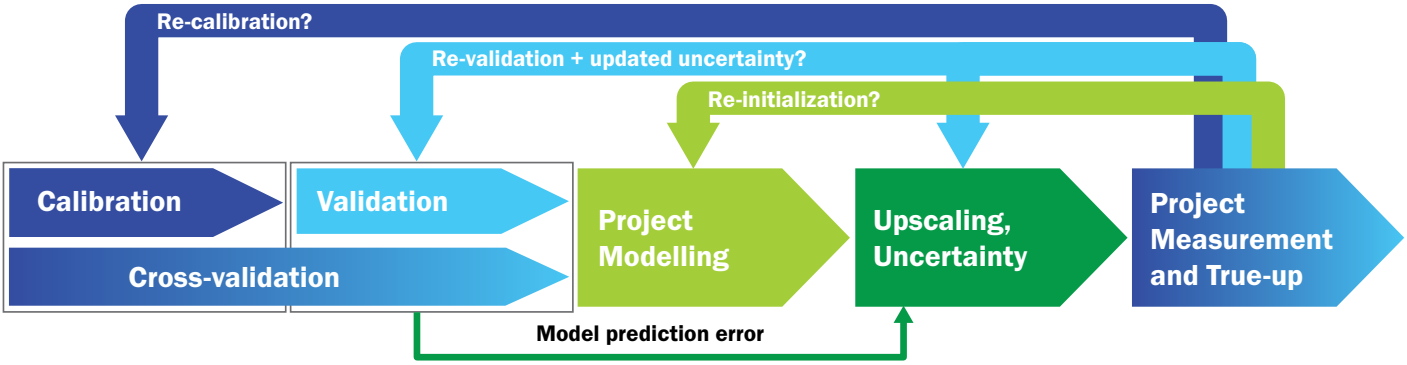
[4] The initial choice of which model(s) to use is extremely important, but not our focus here. Helpful discussion of model choice for agricultural GHG mitigation projects is provided by Tonitto et al. [32].

[5] Measurement of the project after initiation, or what some refer to as "true-up," is the least understood of the project steps as it has not yet been performed for a registered soil GHG emission reduction or removal project and there is no consensus on what the end goal of making subsequent direct measurements should be (e.g., model improvement, updated project-level uncertainty estimation, credit recalculation), or how exactly the direct measurements should be used to achieve any of those goals (e.g., re-initialization, recalibration, revalidation, etc.). Direct measurements can be periodic (e.g., of SOC stocks every few years) or constant (e.g., flux towers for trace gas measurement).

FIGURE 1.

Overview of the main project steps in which process-based modeling is used

At any project step where the model is used, the modeling workflow (including initialization, data inputs, pre- and post-processing) will be invoked and should be kept consistent throughout the steps. Further detail on each step is provided in the main text, including some of the current guidance and ideas for the true-up procedure.



Model parameters are adjusted to minimize prediction error against a calibration dataset.

Not typically required; an "off-the-shelf" model may be used without a new calibration if desired.

Calibrated model is used to simulate independent validation dataset.

Model prediction error is estimated based on model fit to validation dataset.

Model validation report is generated and submitted to a standard body and reviewed by an independent expert.

Cross-validation combines these steps, using data resampling to simultaneously calibrate and validate in an iterative process. Eventually, a single parameter set may be generated, e.g., the means across folds.

Using validated model, project and baseline* simulations are initialized.

Model is used to estimate GHG emissions or SOC stocks in project and baseline* scenarios. GHG emission reductions or removals are based on the difference.

*Not all projects use dynamic modelled baselines; static (not recommended) or measured dynamic baselines are options.

Model predictions are scaled to the project level.

Uncertainty of GHG emission or SOC stock changes for the project is calculated based on measurement, modeling, and scaling errors.

Model prediction error is calculated during the previous validation process.

Direct measurements of project GHG emissions or SOC stocks are conducted and may be used for different purposes, e.g., to ground-truth project-level predictions, update the model, and/or obtain a new uncertainty estimate for the project.

Current guidance on the "true-up" is inconsistent.

Each of these steps feeds into and depends upon the others, and all must be done in a way that represents the scale and character of the project. For example, if a project is going to be done at a large scale where many fields are aggregated and stratified to reduce sampling and modeling costs, then the model calibration, validation, project predictions, and any subsequent measurements must all be done in a way that reflects that project design. The model calibration would have to ensure generalizability across all of the fields or strata included in such a project, including ranges of soil types, climate conditions, the different crop types and management practice changes represented, and outcome variables of interest (e.g., SOC and N2O). The model would then have to be validated in a similar manner using data from sites spanning ranges of key input variables, with outcomes assessed and uncertainty calculated for each output variable of interest (e.g., each unique GHG). Later, if any direct project measurements (e.g., to inform a true-up) were performed, the sampling scheme would need to account for the project design, with the project scale and stratification in mind. On the other hand, if a project were specific to one field (this is unlikely but useful as an example),

each step would ideally be designed and performed to be relevant to that single field and its properties (e.g., soil type, weather, crop type, etc.). Though a more broadly applicable model could be used for a more specific application, it would likely be less accurate and precise than one calibrated in a more specific and relevant way.

Most MMRV and accounting use cases, such as for offset generation (i.e., voluntary carbon markets, carbon crediting), require additional steps that do not directly involve process-based modeling but inform or depend on the modeling activities. For example, in carbon crediting projects when the final carbon credits generated by a project are quantified, the uncertainty associated with the model predictions for the crediting period may be combined with other uncertainties to calculate an "uncertainty deduction" or "discount," effectively penalizing larger uncertainty and increasing confidence that any credits are reliable and conservative estimates of the climate benefits. If project uncertainty is not estimated in a rigorous way, for example, underestimated due to omission of important uncertainty sources, the deduction will be inadequate and confidence in the resulting credits will be undermined, thereby violating the principle of conservatism (Box 1). In most cases, these non-modeling steps are outside the scope of this work, however, we do discuss upscaling and project-level uncertainty because they have implications for how the modeling might be done, how uncertainties are calculated, and how we think about the effects of different approaches on outcomes. In the example given above in which a more broadly applicable model could be used for a project focused on only one field, using such a model would almost certainly result in a larger uncertainty deduction than using a model calibrated to that specific field. Hence, understanding credit generation from model results in this case would affect one's choice of model calibration and validation procedure.

While the project steps are broadly consistent across project types, exactly how each step is performed, and more specifically the level of rigor and detailed reporting ultimately required at each step, will depend on the type of project and end goal. For example, projects generating carbon credits for trading in a voluntary marketplace may be required to pass very high standards, while those meant to support claims of corporate supply chain GHG reductions (i.e., "Scope 3") may have more lenient standards to enable action where granular data collection is currently considered infeasible.

Model calibration

Calibration is the setting of parameter values that best enable the model to accurately simulate measured data. There are several methods of calibration available to the modeler which broadly fall into the categories of calibration by hand (i.e., tuning, trial and error) or statistical calibration which includes numerical optimization (e.g., maximum likelihood) and Bayesian numerical methods (e.g., Markov chain Monte Carlo, Sequential Monte Carlo [36]). The calibration process requires many decisions to be made by the modeler, such as which model parameters to focus on (it is common to focus on a subset of parameters for complex models with too many parameters to calibrate at once), some of which may be difficult to automate or standardize [30, 37]. In-depth discussions of common methods and key decisions made during calibration procedures are discussed in detail elsewhere [e.g., 38]. We do not intend to review those discussions here, rather only to highlight the diversity of calibration methods, which means that rigid requirements around calibration procedures across diverse models and GHG projects may be impractical and unnecessary at this point in time.[6] Perhaps more importantly, because the subsequent validation of the calibrated model serves as a checkpoint to assess the quality of the calibration regardless of the approach, the use of different calibration approaches across projects does not warrant concern

[6] An in-depth discussion of considerations for calibration approaches and potential to standardize them is provided by Wallach et al. [39].

so long as the same parameter sets are used throughout the project steps and the validation is handled appropriately (including that the implications of the calibration approach for uncertainty propagation are accounted for).

Aside from the calibration approach used, the data used for calibration are extremely important to determining the performance and applicability of the calibrated model. Calibration requires datasets that include all model inputs along with measurements of the variables being calibrated against (which does not necessarily have to be the variable that is the focus of the GHG mitigation project), ideally in the context of the practice change(s) of interest for the project. The quality of the measurements is extremely important because systematic measurement error or low precision (where there are few data points) can lead to poor parameter estimates and flawed model predictions. Further, whenever predictions are being made over time (e.g., changes in SOC stocks or fluxes of N_2O), calibration data should ideally include repeated measurements over similar time periods (that said, this point is even more important for validation, see Validation Data). Unfortunately, these requirements for calibration data represent a high bar and current datasets available for these uses are fairly sparse. For SOC stocks, high quality measurements over time to depths of at least 30 cm, that also include all of the required input data to run a complex model, are relatively rare. For trace gas emissions, there is a well-recognized measurement challenge of capturing "hot spots" in space and "hot moments" in time, as missed emissions pulses can cause cumulative field estimates to be biased low and can make model calibration challenging (e.g., if a model predicts a pulse during a period when no observations were made) [40, 41]. Currently, access to calibration data is a major limitation on model development and application across contexts, especially for estimating contributions of CH_4 and N_2O to farm-level GHG accounting.

The robustness of a model across contexts is an important consideration in the context of GHG mitigation projects. Rather than using a single parameter set across all areas being modeled in an agricultural GHG mitigation project, it may be appropriate to model changes in parameters under different conditions, for example for geographical areas such as Land Resource Regions of the U.S., assuming the necessary data is available. However, performing too many separate calibrations of a model to different regions and environmental conditions typically leads to a model that extrapolates poorly to anywhere other than exactly where it has been calibrated (i.e., overfitting). While robustness is a general modeling principle (Box 1), it is also possible that a model calibration could be too generalized (e.g., one global calibration across all regions and conditions) with mediocre performance everywhere. In such circumstances formal statistical approaches for model selection (e.g., AIC, wAIC, predictive loss) represent best practice for determining what levels of aggregation, and for which grouping variables, are most parsimonious with the observations. For example, the choice to calibrate a model by crop type versus by soil type essentially represents a hypothesis test, with model selection metrics trying to find the most parsimonious hypothesis by balancing model fit and model predictive error based on some form of penalty for model complexity. In addition, hierarchical models represent an alternative to fitting models independently for different groups, as these approaches allow one to borrow strength across different groups while simultaneously constraining the extent to which the calibrations of different groups can diverge from each other [42–44]. Hierarchical models also allow one to formally distinguish between predictions to known locations and the greater uncertainty encountered when predicting to new locations. As such it is important that model verification statistics for hierarchical models be calculated for out-of-sample locations, as this is how models are generally applied when calculating carbon credits. It should be noted that while hierarchical models are well established in statistics, they have thus far had limited applications in biogeochemical modeling, so this is an area with high potential for improvement and innovation [45, 46].

Overall, it is a well-known reality to modelers that we are limited in our collective ability to model all variables of interest under different management practices and across different environmental and geographical contexts equally well. We encourage the use of formal model selection to address over- and under-fitting and see hierarchical modeling as a research and innovation frontier. Rather than prescribing strict requirements around calibration, we suggest that protocols continue to leave room for innovative approaches but stress the critical need for transparency around data limitations and model calibration procedures, and to account for the resulting limitations on model accuracy, precision, and robustness during model validation.

Model validation

Once a model has been calibrated, its performance must be evaluated using a set of data independent from the calibration dataset. The term validation is currently used by GHG accounting or crediting programs to refer specifically to the formal performance evaluation procedure whereby a model is shown to pass predetermined requirements (e.g., thresholds for accuracy or precision) for approved use under a protocol.[7] Models which pass this testing for a particular project or application are deemed "validated" for use in those contexts. For consistency, we use the term validation under this definition throughout this guidance. For GHG accounting or crediting projects, model validation is an incredibly important step [24] because it is used to determine whether a model will be approved for use in a project, identify where a model performs better or worse, and estimate uncertainty for the project predictions (which can be used to adjust the final GHG emission reduction or SOC stock change estimate or credits generated, an "uncertainty deduction")[47].

The rigor of the validation process ultimately determines the perceived trustworthiness of any GHG emissions reductions or removal claims made, or quality of any carbon credits generated using a process-based model. Given recent calls for consistency across GHG mitigation protocols and between projects to improve overall trust in GHG emission reduction or removal claims [22], model validation procedures across registries and projects should be as consistent as possible. At the same time, the diversity of process-based models and projects they are applied to, as well as the difficulty of obtaining high-quality datasets with enough information to reliably drive models (especially in certain geographies such as the Global South [27]) means that validation procedures need to be adaptable. Further, given that this is a rapidly developing area of research and development, GHG mitigation protocols should allow room for innovation in modeling approaches. The validation process must reach a balance between flexibility to accommodate different approaches and consistency to ensure a base level of rigor by setting requirements that can be assessed under different project and modeling contexts. Further, the validation process must be designed to effectively serve as a quality control checkpoint whereby any decisions that decrease the accuracy or precision of project predictions are accounted for or penalized appropriately, to incentivize toward model improvement and away from gaming (see Box 1).

In practice, setting exact thresholds for acceptable model performance is complicated by the fact that there are no universal rules for what "sufficient" model performance looks like, especially in terms of precision (model performance is more typically seen as a sliding scale from better to worse). This continuum is currently recognized in existing protocol modeling guidance (e.g., Climate Action Reserve, Verra) that set floors on model accuracy (bias and goodness-of-fit) but then penalize carbon credits in proportion to model precision, allowing imprecise models to participate while creating a positive incentive to increase model precision.

[7] In this guidance, validation does not mean a comprehensive evaluation of the worthiness or correctness of the model. Instead, validation has the narrow definition that the model meets the specific acceptability criteria of the authority who approves the estimates made by the model, such as a registry. Hence, the term "validation" is used here rather than the more encompassing term "evaluation."

Validation data

While the scope and quality of the calibration data determine the applicability and accuracy of the model, the scope and quality of the validation data determine the level at which that applicability and accuracy can be assessed. If data for a given context are not included in the validation dataset, model performance can not be assessed for that context, and so the scope and quality of the validation data must be appropriate for the ultimate intended application of the model. This means that the requirements for validation data are particularly specific. For example, while model calibration could be performed based on data for different responses (e.g., crop yield), validation must be performed for the GHG emission or SOC stock and management action that is the focus of the project (e.g., SOC stock change over time in response to a change in cover crop). Protocols typically go a step further, requiring validation of the emission reduction or SOC stock, meaning the difference in GHG emission or SOC stock between the project and the baseline, which requires two separate data points (e.g., a treatment and a control) for validation.

As with calibration, the limitation on available data suited to model validation represents a major limiting factor to the level of detail over which model validation can be performed. For validating changes in SOC stocks, studies with paired treatment and control measurements (analogous to project activity versus business-as-usual baseline) that meet all of the previously suggested requirements for calibration data (e.g., depth, time series, precision) are very rare. Even though the critical need for this data has been widely acknowledged [e.g., 48], additional data is slow to emerge in part because SOC stocks change slowly and it often takes decades to measure detectable differences between field treatments. By contrast, for trace gas emissions treatment changes may be quicker to emerge, but the aforementioned issues with temporal resolution and gaps over time remain as hurdles to developing comprehensive validation datasets. Given these limitations, cross-validation, where one dataset is resampled with different portions of the data used for either calibration or validation on each iteration or "fold," has risen in popularity as a means to make the most of the limited data available [47, 49]. When cross-validation is used, the data for calibration and validation within each fold must still meet the requirement of independence from each other (e.g., time series measurements from a single site should not be split, different studies at the same site should not be split, etc.).

Given the limitations on acquiring validation data, particular attention must be paid to the nature of the validation data in terms of its quality and its representativeness of the project domain. If the project conditions are not well represented in the validation data, either because they make up a small proportion of the data, or are not included at all, poor model performance under those conditions is unlikely to be identified. An example of this would be a model that consistently underpredicts N_2O emissions in clayey soils (e.g., with >70% clay). If that model were validated against a dataset that contains no sites with high clay contents, it might pass testing with ease. Yet if the project area contained sites with high clay, the model performance would be misrepresented and the actual GHG emission reductions for the project would be less than claimed. For this reason, it is important that validation data are comprehensive in terms of representing ranges of key variables present in the project area, and that datasets are balanced with respect to those variables. Yet this is difficult to do in practice given the limited availability of high-quality validation data, and the latter suggestion might cause teams to discard validation data from over-represented conditions in efforts to compensate for small sample size under other conditions, which is clearly not ideal. Hence, current protocols tend not to have strict requirements in this area. Again, this highlights the importance of transparency, rigor, and acknowledgement of limitations during the validation process, as well as the potential utility of subsequent measurements for the project (see "True Up").

Bias (systematic error)

Bias is a measure of whether model predictions are consistently higher or lower than measurements (i.e., systematic error). It can be assessed in a general way by computing an average error of the model predictions versus the validation data, for example, using an unweighted mean of the biases for all individual observations or studies in the validation dataset [47]. If there is no systematic error, a model will be just as likely to over- and under-predict given any particular set of inputs, and the calculated bias should be near or at zero (Fig. 2a, left and right columns). If instead the bias is larger than a set threshold (e.g., the average or "pooled" measurement error [PMU] for the validation data), the model will be deemed inaccurate and fail validation.[8] An important implication of this approach is that a model can perform poorly for individual observations or sites (a few or even all of them) as long as it performs well on average (Fig. 2a, lower left panel and right column).

Because they are simplifications of reality, all models have some systematic errors due to misrepresentations of certain biological processes or biases in their calibration data [50], which tend to manifest in certain contexts where these shortcomings become relevant. For example, a model may not represent the effects of soil compaction on oxygen diffusion, resulting in flawed estimates of soil N_2O responses to tillage. Identifying and diagnosing systematic errors is extremely important because they will lead to over- or underestimates that do not cancel out or diminish as projects increase in size in the same way that random error will.

Importantly, systematic error will only be apparent for the certain situation(s) where a given model shortcoming becomes relevant, highlighting the need for validation data that represent good coverage of the project domain. Yet even where validation data do span the ranges of key variables, assessing bias based on the overall average might obscure model bias in specific situations that become relevant for the project predictions (Fig. 2c). To demonstrate this point, consider an example of a model that has been validated for predicting changes in SOC in a broad manner, using validation data spanning several practice change categories (e.g., reduced tillage, cover cropping, organic amendments) and crop functional types (e.g., corn, wheat and soy) across the entire U.S. That model might tend to overpredict changes in SOC in certain settings, for example, for implementing reduced tillage in pasture systems, while it might tend to underpredict in other settings, for example, for implementing cover cropping in corn-soy systems. Given that both could be included during a very broad model validation, the overall bias of the model predictions would be lower than the specific bias for reduced tillage in pasture systems in the Southeast. If that model were then used for a project focused primarily on reduced tillage in pasture systems in the Southeast, the estimated uncertainty of the GHG mitigation claim for that project would be misrepresented as unrealistically low (Fig. 2c).

Given that models commonly perform differently depending on the setting (especially for different practice changes, e.g., [51]), failing to evaluate model performance under different settings relevant to the project is very likely to give an inaccurate picture of model prediction uncertainty for the project. At the same time, separating validation data into too many categories for bias assessment can result in low sample sizes and correspondingly unreliable validation statistics, so a practical balance must be reached or other means must be used to check for systematic errors. In an effort to protect against undetected bias relevant to the project, some protocols require that bias be assessed separately for different combinations of key factors such as management practice type and crop functional type in addition to

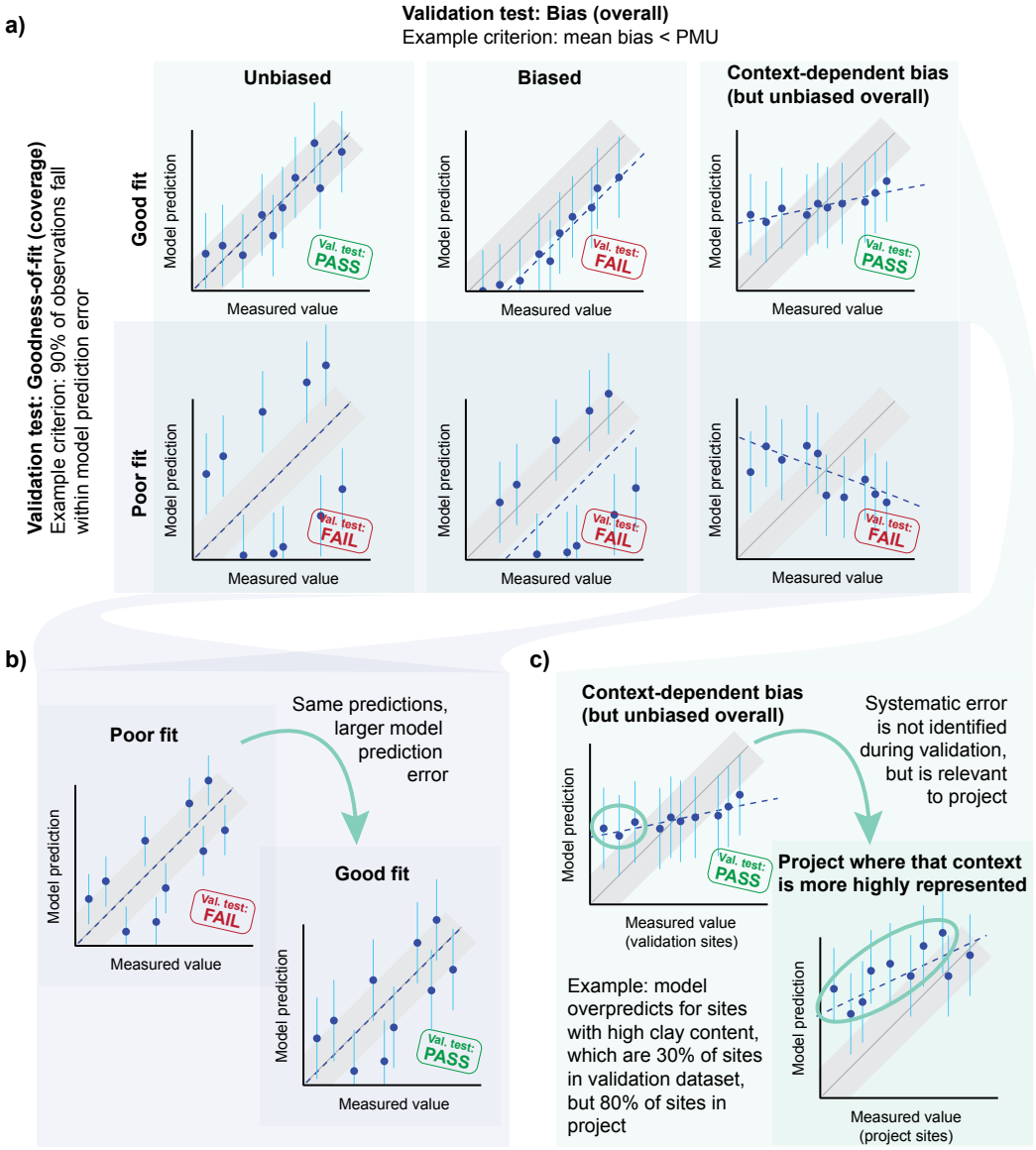
[8] One way to set a threshold on model accuracy is by requiring the average relative error be smaller than the validation data measurement error, sometimes termed the pooled measurement uncertainty (PMU). While this is a generally accepted method, its implication is that the less precise the measurements (larger PMU), the more likely a model is to be accepted as lacking bias. For this reason, it is important that protocols use other means to incentivize measurement precision, which we address below in "Estimating project-level GHG emission reductions and associated uncertainty."

showing that the validation data are distributed across relevant ranges of some biophysical variables (e.g., soil texture) represented in the project domain. These aspects are useful steps toward identifying relevant model bias but leave room for improvement and innovation which we discuss in our Recommendations for Improved Guidance and Future Research.

FIGURE 2.

Bias, goodness of fit, and model prediction error in the context of model validation

The figure demonstrates the interplay between validation criteria and hypothetical examples of passing or failing validation tests (a). The 1:1 line (perfect fit) between measurements and model predictions is shown with a thin gray line, with a shaded gray area representing a hypothetical PMU distance from the mean in both directions. The line of best fit for the points is shown as a dotted blue line and light blue error bars are the model prediction error. The left and right lower panels of (a) with no bias overall but poor fits could pass the goodness-of-fit test if the model prediction error were wider, demonstrated in (b). Unidentified bias in right panels of (a) may impact potential predictions for a project where that bias is more relevant than for the validation dataset, shown in (c).



Precision and model prediction error

Assessing a model's precision (i.e., random error), in addition to its accuracy (i.e., bias), is important for understanding the risk of applying it to a GHG mitigation project. A model may be accurate overall but imprecise (Fig. 2a, bottom left), resulting in a high risk of poor predictions for a given point. In the context of GHG mitigation projects, the aim of assessing model precision is to come to a generalized estimate of the model prediction error, or the uncertainty associated with predictions made for points within the associated project. Model prediction error is used in two distinct, but related, ways within mitigation projects. First, many registries have a validation criteria associated with model precision, often expressed in terms of model coverage (e.g., at least 90% of validation data must fall within the 90% predictive interval, see Goodness of Fit). Second, many current protocols impose an uncertainty deduction (e.g., reducing the final number of credits generated by a project) calculated based on model prediction error. Note that in both these cases it is essential that model prediction error account for both accuracy and precision — a model that is internally confident (high precision) but systematically biased (low accuracy) should have a high model prediction error and incur a large uncertainty penalty.[9] Even where average bias is lower than a threshold set by a model validation criterion (e.g., the PMU), the bias component of the model prediction error should not be ignored. Indeed, when averaging over a larger portfolio of sites, independent random errors will tend to cancel out quickly, while autocorrelated random errors cancel out slowly (see Appendix B) and systematic errors do not cancel at all.

Model prediction error can be estimated in different ways, but broadly speaking there are two general approaches: by assuming that the model's validation error in one context will apply to new predictions in another (validation-based) or by propagating model uncertainties into new predictions (propagation-based). When working with process-based models these two approaches often map on to frequentist and Bayesian approaches, respectfully.[10] Validation-based approaches use the difference between model predictions and observations to compute errors (e.g., root mean square error, RMSE), which can then be used to calculate model error intervals. Validation-based approaches typically produce a single fixed error estimate — more sophisticated approaches to modeling error across space, time, and conditions could be employed but would require the introduction of additional models and assumptions. By contrast, propagation-based approaches aim to simulate predictive probability distributions. In classic statistical modeling the most common sources of error being propagated are uncertainties in the model parameters (e.g., confidence interval) and the residual error, but process-based modelers may also be propagating the uncertainty in the model's initial conditions (e.g., SOC pool at the start of a simulation) and drivers (e.g., soil texture, meteorology), and hierarchical parameter variability (i.e., random effects)[52]. Such approaches capture that our confidence in predictions can vary depending on conditions, but it is critical that such predictive errors be validated to ensure their predictive coverage is not under- or over-confident.

As with accuracy, model prediction error will vary with the context and situation being modeled. For cases where model validation error is used as the sole estimate of model prediction error, the reliability of the prediction error estimate for the project depends on how well the validation data represents the project, as well as the amount of data available

[9] Note that some current protocols allow the use of an error calculated based on the standard deviation of the model residuals, not the RMSE. In this calculation, the mean of the residuals (i.e., the bias) is subtracted off and therefore not included in the error. In other words, as long as a model meets the bias validation criterion (e.g., mean bias < PMU), there is no penalty for model bias when issuing credits, despite the fact that biases are a riskier source of error, in the context of the portfolio effect, since they do not average out with increasing project size.

[10] The approaches are not completely exclusive and aspects can be combined. For example, there are numerous frequentist approaches to error propagation (e.g., bootstrapping) and error propagation is used in most classic statistical models (this is why a regression predictive interval is hourglass-shaped rather than two parallel lines).

to validate it under different contexts. One of the ways this becomes apparent is when comparing the duration of the experiments used in the validation dataset to the duration over which project predictions will be made. Validation data for SOC change typically span many different durations, from years to over a century. While matching the duration of validation data and project predictions might seem ideal, doing so can sometimes severely limit the amount of validation data available and be unnecessary because useful information on model performance can be gained from validating it over different time periods. That said, the uncertainty in model predictions changes over time, so the relationship between the modeled duration and the model prediction uncertainty should be taken into account and conservative principles applied (see Appendix B). We provide further detail and recommendations on this topic below (see Recommendation #3).

Goodness of fit

Once calculated, the model prediction error may be used to further assess model performance by comparing predictive probability distributions to observations (goodness of fit). A model which passes bias testing may still perform poorly at matching individual observations (Fig. 2a, upper row), so checking the discrepancy between model predictions and individual observations or studies is a useful exercise and some protocols set thresholds for goodness of fit. An example is the coverage requirement that 90% of observations in the validation dataset fall within the 90% confidence interval of the corresponding model prediction.[11] Similar to the bias threshold above, the implication is that the model can perform poorly in some cases (10% in this example) if it captures the observation the majority of the time (Fig 2a, upper right). Just as with bias, goodness of fit should ideally be investigated in relation to key variables and factors such as management practice categories to identify any specific shortcomings of the model that are relevant to the project. We discuss this further and provide recommendations below (see Recommendation #5).

Project prediction

Once a model has been validated for use in a given application, it can be used to produce predictions of changes in GHG emissions reductions or SOC stock changes for a project. Critically, the same model, parameter sets, initialization procedures, and model drivers used during validation must be used for project modeling, as a means to ensure that the accuracy and precision demonstrated in the validation report apply to the project predictions (we recommend that this include the entire modeling workflow, see Recommendation #1 and further description of the modeling workflow in Appendix A). Project modeling is typically done on a field or point basis, rather than at the landscape scale, and the predictions are scaled up to the landscape or project level later on.[12] The modeled points may represent every field or land unit enrolled in a project, or may represent a smaller sample of those units which is more typical for large projects (also known as "aggregated" projects). Where dynamic modeled baselines are used, two predictions are produced for each point: the project scenario where a practice change has occurred, and the business-as-usual scenario

[11] The implication of requiring a certain proportion of observed values to fall within the corresponding model prediction confidence interval is that the less precise the model prediction (larger confidence interval), the more likely a model is to be accepted as having a good fit. For this reason, it is important that protocols use other means to incentivize model precision, which we address below in "Estimating project-level GHG emission reductions and associated uncertainty."

[12] Though landscape scale models such as Landscape-DNDC are being developed, see https://ldndc.imk-ifu.kit.edu/

representing the counterfactual baseline to which project emissions are compared.[13] Each of these must be run with the same model inputs except those representing the management changes made. The input data used to drive the model can be forecast to obtain predictions for the future, but this introduces additional uncertainty that is not accounted for during the model validation and would need to be added to the total uncertainty in the predictions (see discussion on error propagation in the above section on Precision and Model Prediction Error). Alternatively, project predictions can be made by looking backward, typically in short time increments to allow for quantification of outcomes (and any payments) on convenient timescales. For example, after the first year of a project, observed input data can be used to produce predictions that are more reliable than forecasts because they are based on direct observations, which have much lower uncertainty.

Modeling workflow

Each of the steps described above requires running the model, and that in itself includes several steps that we refer to here as the "modeling workflow." While workflows may differ across models or projects, examples of common steps include initialization (which may include spin-up, see Appendix A: Model Initialization), prediction and post-processing (Appendix A, Fig. A1). At each step in the project when the model is run (calibration, validation, and project modeling, Fig. 1), these modeling workflow steps are performed. However, they could be performed in different ways, creating inconsistencies within the project. For example, if a different initialization method is used during validation versus project modeling, the validation modeling procedure would not fully represent the project modeling procedure and the model may perform differently for each of those applications. The same issue applies to data curation for model drivers (e.g., source of meteorological drivers, algorithms for gap-filling and down-scaling, soil texture, management history) and model parameters, though only the latter (model parameter constancy) is explicitly discussed in most current protocols. Critically, if one understood the implications of certain deviations in model workflow steps between project steps for predictions of GHG differences, one could make strategic choices to maximize crediting outcomes (i.e., "gaming," Box 1). An example of this would be using input data with lower uncertainty during model validation (e.g., site-level meteorology and soils data) than is available for project prediction (e.g., gridded soils and meteorology), and thus validating a lower model prediction error and bias than is achievable in practice. Changes in workflow could also be less intentional, such as a bug in pre- or post-processing that differs between validation and project prediction. Therefore, we see value in recognizing that the validation applies not only to a model version and parameter set, but also to the entire modeling workflow (see Recommendation #1). We provide further detail on key steps in the modeling workflow and a case study of differing approaches to model initialization in Appendix A.

Estimating project-level outcomes and associated uncertainty

Once project locations have been modeled individually, and predictions of the change in GHG emissions from the baseline for each point have been calculated, the predictions must be combined and scaled to come up with GHG mitigation estimates for the full project. Given that these points may represent only a portion of the project, scaling to the full project introduces additional uncertainty (scaling error or "sample error" in some protocols). Together, the scaling error combines with the observation/measurement errors and the model prediction error to produce the precision of estimated GHG emission reductions or SOC stock changes for the project.

[13] Different approaches to quantifying the counterfactual baseline scenario exist, including static measured baselines, dynamic measured baselines, and dynamic modeled baselines. We recommend dynamic baselines, and generally assume the use of dynamic modeled baselines in this report.

Depending on the application, the estimate of precision for the project may be used to calculate an uncertainty deduction to the final credits generated or allowable GHG emissions reduction or removal claim. This represents an extremely important opportunity to penalize imprecision and inaccuracy and guide toward more precise quantification in general, which may be lacking in other project steps. Remember that during model validation, there may be benefits to having lower measurement precision (less likely to identify model bias when the criterion for identifying excessive bias is whether it is larger than the measurement error, Fig. 2a), and higher model prediction error (more likely to pass goodness-of-fit testing when the criterion is that measurement values fall within the model prediction error bounds, Fig. 2b), so precision is not effectively encouraged or incentivized by that process. However, when total uncertainty (including observation error, model prediction error, and scaling error) leads to a penalty, precision is encouraged.

Given that the approach for estimating uncertainty for a given project has big implications for the ultimate claims or payments allowable under a protocol, it has been the subject of much scrutiny and debate. Ultimately, to be "reliable," the uncertainty estimate must not omit any relevant sources of uncertainty, and it must be properly propagated based on spatial and temporal considerations (e.g., non-independence). The latter aspect encompasses some of the biggest concerns among the scientific community with regards to how uncertainty has been treated in GHG accounting and MMRV thus far [53], and are thus the focus of this discussion, with related recommendations provided below (see Recommendation #2).

Combining errors in a robust way to produce a reliable estimate or project-level uncertainty may not be straightforward and requires consideration of the relationships between measured and modeled points for a given project. Specifically, how do model errors relate and can modeled points be considered independent? The answer to this question determines how the errors combine and has especially important implications for projects that rely on large numbers of modeled points to decrease their relative uncertainty deductions (i.e., the portfolio effect). Current modeling guidance for agricultural GHG mitigation projects lacks in-depth discussion of these issues. We therefore provide an overview of these considerations in Appendix B to support our recommendations (see Recommendation #4) for improvements to future guidance and areas for research on this topic.

True-up with direct project measurements

We have thus far highlighted the difficulties in obtaining rich validation datasets that can be used to rigorously test model performance and identify systematic errors across broad ranges of biophysical conditions. Further, even where validation data exist, they may have shortcomings that lead them to be imperfect representations of conditions within projects. One key shortcoming is that much of the existing validation data comes from research/experimental farms where treatments have been imposed and maintained in ways that may not be consistent with how real working farms are managed. Similarly, validation data from small, plot-scale experiments or strip trials may not be fully representative of effects that would manifest if practices were carried out at the larger field-scale. For trace gas fluxes, much existing validation data comes from chamber or other snapshot-like methods that (1) do not collect data consistently through time and are very likely to miss hot moments, and (2) are limited in their spatial coverage, likely missing hot spots, both of which have been shown to represent significant portions of total fluxes [54–56]. Added to this are the complications of adapting data from studies to match the needs for running the model; variations and mismatches in depth increments and time increments, missing input data, and other inconsistencies are common. These are only some examples of how available validation data may fall short of representing the actual effect sizes that will be realized when practice changes are implemented in projects on working farms under different biophysical conditions.

Given the shortage of appropriate validation data across geographies and biophysical contexts and the fact that there will always be unknown unknowns regarding mismatches between validation sites and the project area, validation with independent data can not guarantee accurate representation of model performance and uncertainty for a project. Because of this, there is general agreement that some form of direct measurement of the project after its initiation to "true-up" model predictions is necessary to instill confidence in project predictions. Beyond this however, the exact aim and best use of these measurements is subject of debate. Direct measurements of the project area could be used to (1) verify or "ground truth" the model predictions, or (2) improve the model for continued use in the project, for example, by performing an updated validation to obtain new uncertainty estimates for the project predictions, by recalibrating the model to improve subsequent predictions, or by using the measurements to reinitialize the model to base future predictions from a more realistic starting point. Some combination of (1) and (2) is also possible. The debate inevitably depends on a few considerations, including the reliability of measurements versus model predictions and the sampling effort needed to support different approaches (with additional considerations for the handling of the baseline scenario).

First, the idea of using direct measurements to ground truth model predictions or update a model introduces considerations of measurement quality, specifically accuracy and precision. It is commonly assumed that measurements are better reflections of the truth than model predictions, yet this may not be the case if the measurements are biased. Many common mistakes in soil sampling, soil sample preprocessing, and SOC measurement can contribute to measurement bias, thereby making measurements unreliable. Measurements may also be highly variable (imprecise), but this is less problematic so long as that error is known and accounted for. That said, measurement precision relates directly to the ability to detect changes over time, with lower precision leading to higher sampling requirements or failure to detect changes where they do exist. When it comes to reconciling measurements and model predictions, it is not prudent to assume that one or the other is an inherently better representation of the truth, and rather the uncertainties in each should be rigorously accounted for. Assuming that a biased measurement is the truth is dangerous, just as assuming a biased model prediction is the truth is dangerous. Overall, the utility of true-up measurements depends on their quality (bias and precision), and how consistently the measurements are collected through time.

The second consideration is the sampling effort and design required to carry out any given approach, but it is especially pertinent to ground-truthing GHG emission and SOC stock changes over time at the project scale. For SOC, detecting change over time is difficult due to its high spatial heterogeneity combined with its slow rate of change (low signal-to-noise ratio) leading many to dismiss altogether the possibility of ground-truthing model predictions with measurements, especially over timescales of less than 10 years. However, recent research shows that change detection for SOC is possible given dense enough sampling over large enough numbers of sites [25]. Yet, because detecting change over time requires repeated measurements performed in a comparable manner, projects that did not execute high quality sampling at the onset are limited in terms of what they can do with repeated measurements later on. Further, some current protocols limit the number of sites that can be revisited for SOC sampling in an effort to randomize which sites are visited (and avoid gaming, e.g., by ensuring exceptionally large SOC gains at farms that will be sampled). Finally, direct measurements are costly, and the value they may add to the project must be balanced with the cost of collecting them. These issues complicate the possibility of using direct measurements to ground truth model predictions, which is problematic given the importance of direct measurements and ground-truthing for instilling confidence in project predictions. This represents a major area in need of consensus: should true-up measure-

ments be used to ground truth model predictions, and how can they be designed to best accomplish that purpose? If ground-truthing is identified as a critical need, protocols should be designed and updated with change detection in mind and to encourage appropriate sampling campaigns with repeated measurements over time.

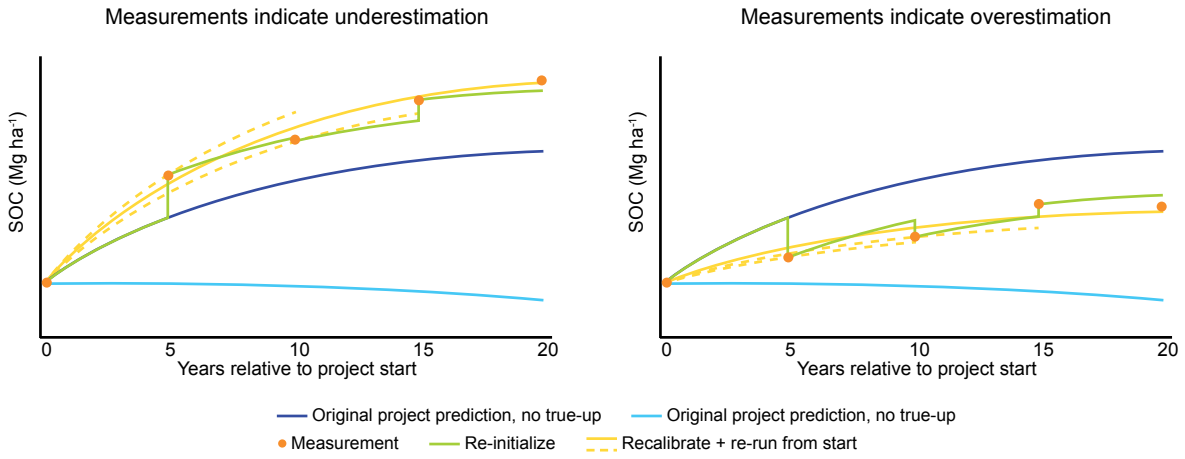
In addition to ground-truthing model predictions, direct measurements of the project area could be used to improve a model's ability to make predictions for the project, and/or update model validations to better represent the model's accuracy and precision for the project (including baseline scenarios if applicable), and/or to re-initialize the model for subsequent predictions (Fig. 3). One of these or different combinations could be achieved, for example, using cross-validation approaches that improve the model calibration and validate model predictions simultaneously. The choice of which approach is ideal for a given project depends on several factors, including (1) the project design, especially the total number of sites, number of sites sampled, and whether there are repeated measurements over time; (2) the labor required to carry out a particular approach; (3) the anticipated reduction in uncertainty of the model predictions for the project (which would require revalidation); and (4) whether the nature of the measurements allow for revalidation. Importantly, some approaches are more reliant on having repeated measurements at the same site than others, for example, revalidation is most likely to require measurements of changes over time, whereas recalibration or re-initialization could be done for a single time point.[14] It's currently unclear whether recalibrating, re-initializing, or a combination will lead to larger reductions in model prediction error and project-level uncertainty, but common sense among modelers suggests that the more updates (i.e., states and parameters) that can be made to the model based on sound data, the more the uncertainty will be reduced. Therefore, where repeated measurements are available, it may be advantageous to both recalibrate and revalidate the model to obtain an updated, more relevant (and possibly smaller), project-level uncertainty estimate. However, revalidation using project measurements could produce a higher project-level uncertainty estimate, in which case it may not be financially advantageous for project developers to do so, highlighting the potential importance of requirements in this area to maintain rigor and consistency across projects. Finally, regardless of whether measurements are repeated or not, it may be advantageous to use new measurements to initialize the model (re-initialize where they are resampled sites, initialize if they are sites measured for the first time; see Appendix A). However, re-initialization may prove problematic from a crediting perspective and may not be allowed under some protocols. State data assimilation could be a viable alternative (see Appendix A) but is generally not covered in current protocols.

[14] The need for repeated measurements also depends on what protocol is being followed as some may not require calibration or validation on changes over time, though we argue they should.

FIGURE 3.

Examples of possible true-up procedures

Many approaches to the true-up are possible. Two conceptual examples are demonstrated conceptually here: re-initialize and run model for following period or recalibrate and re-run from the start of the project (i.e., time 0). In this hypothetical scenario, SOC measurements are being taken at the start of the project and repeated every five years.



It must be noted that in the strictest sense, project outcomes, specifically the difference between what happened under the project versus what would have happened without an intervention (counterfactual baseline), can never actually be directly verified or "trued-up" because a counterfactual cannot be directly measured.[15] That said, measurements can be designed to approximate counterfactual baselines, e.g., by measuring sets of sites that are analogous to project sites aside from the changes in management spurred by the project. While it can be difficult to find analogs that match project fields exactly (the selection of baseline sites would be subject to the same potential shortcomings/mismatch as other validation sites discussed above), baselines could be measured at large (i.e., regional) scales and used in general ways, for example, to confirm whether directional trends in baseline GHGs and SOC predicted for the project area are correct. Another approach to measuring counterfactual baselines is to split individual fields into treatment (practice change) and control (business-as-usual) plots, though this can be onerous for the farmer and may also introduce bias, for example, if a non-random subset of farmers is willing to perform this type of splitting. Importantly, either of these potential measured baseline approaches would require strict guidance on the selection of baselines sites and/or regions to prevent gaming. Regardless of the approach taken with the measurement and modeling during the true-up, there is always the shortcoming that one can never know exactly how accurate the baseline predictions are or were. However, the amount of uncertainty this adds to a given project outcome can be managed in more or less reliable ways. In general, if a model is being recalibrated based on additional measurements from the project, it would be best practice not only to make updated predictions for project scenarios, but also for baseline scenarios. In theory, a model that is being improved in a general way (i.e., not overfitted via site-by-site calibration but over many sites) based on measurements should perform better for multiple types of scenarios, including baseline scenarios.

15 For this reason, some consider the term "true-up" to be a misnomer.

Recommendations for Improved Guidance and Future Research

1. Validation of the modeling workflow

The main purpose of model validation is to ensure a modeling approach is appropriate, accurate, and reliable for use in making predictions for a given project. Critically, it is not only the model structure and parameter set that are being validated, but the entire modeling workflow used to produce the model predictions that are then compared to observations. This includes data curation, model set-up, initialization, and any data assimilation methods that might be used. Two runs of the same model with the same parameter set but differing modeling workflows (e.g., initialization method) can give different predictions. While it may not be practical for protocols to prescribe exact approaches to any of these modeling steps, it is extremely important that they be documented in the validation report and kept as consistent as possible between the validation and project modeling. Otherwise, differences in modeling workflow between validation and project modeling will introduce additional uncertainty and present opportunities for gaming.

2. Validation data domain/coverage

The data used for validation ultimately determines the context for which the model can be validated, including the applicable geographies, biophysical environments, land management practices, and spatial and temporal extents. Ideally, datasets used for evaluation should span the full ranges of relevant input variables (soil texture, pH, aridity index, etc.) within the project domain, with relevance determined based on the process and geography being modeled. Yet it can be difficult or impossible to find validation data that match the project exactly in all of these aspects, and while this may not preclude model validation for use in a particular project, any mismatch between the character of the validation dataset and the project introduces additional uncertainty in the model predictions for the project. Additional research is needed to determine best practices for how to address this added uncertainty. Further, the extent of the validation data in terms of spatial and temporal ranges determines the level of rigor with which spatial and temporal relationships in model errors can be assessed, and consequently the assumptions that are justifiable during uncertainty estimation. Current protocols tend to focus on the "domain of applicability" as combinations of geography (e.g., land resource regions), biophysical properties (e.g., crop functional groups and soil texture) and land management practices, yet the additional aspects of space and time (discussed below) are very important and have begun to receive more attention, including in a recent peer-reviewed publication [28] and model validation report [57].

Considering time, validation data should ideally cover the maximum timescale of the project predictions or longer. This is because in general a model's variance will increase with time, so any estimate of model prediction error based on a shorter time period than the project application will likely be an underestimate. For example, if the time scale of interest is five years and the model assumes a constant variance (a common assumption and the default for some protocols), and that model is then validated based on data with time periods of < 5 years, the resulting variance will very likely be too low. In general, this should be achievable given that project prediction timescales are typically short (e.g., one to five years) relative to the duration of the long-term experiments available to be used for validation. However, where longer-term data are unavailable or where protocols use long project periods (e.g., 40 years), it may not be possible. In such cases, changes in model prediction error over time should be modeled (Recommendation #3), and any extrapolations of such time-dependent relationships into the future should be done in a conservative manner to ensure model prediction error is not underestimated.

Beyond covering the maximum timescale of the project application, there are benefits to additional temporal and spatial coverage of validation data. If validation data include a good representation of measurements on timescales similar to the project application, it allows for more informed modeling of temporal behavior (e.g., heteroskedasticity) of model prediction error, which will likely allow for tighter prediction intervals on shorter timescales. However, this is not guaranteed and the exact relationship of model error to time should not be assumed without formal analysis (which requires data at relevant timescales). Further, the spatial scale of validation data is important for understanding spatial relationships in model errors. Ideally, validation data should include points at the same spatial scale as what is being modeled for the project (minimum and maximum distances), including the subfield scale as appropriate. Where validation data do not provide adequate sample size across different temporal and spatial scales to allow these analyses, assumptions about model prediction error over time and about the independence of model error across space, should be conservative. We expand on this idea in the following two recommendations.

3. Accounting for changes in model prediction error over time

Given that GHG mitigation projects are modeled over time, temporal relationships in the data are very relevant for estimating the model prediction error. We include more in-depth discussion of these issues in Appendix B, and focus here on heteroskedasticity (i.e., change in variance) over time. It is important to consider the way this is handled during modeling and uncertainty assessment because the approach can cause bias in the model prediction error calculation.

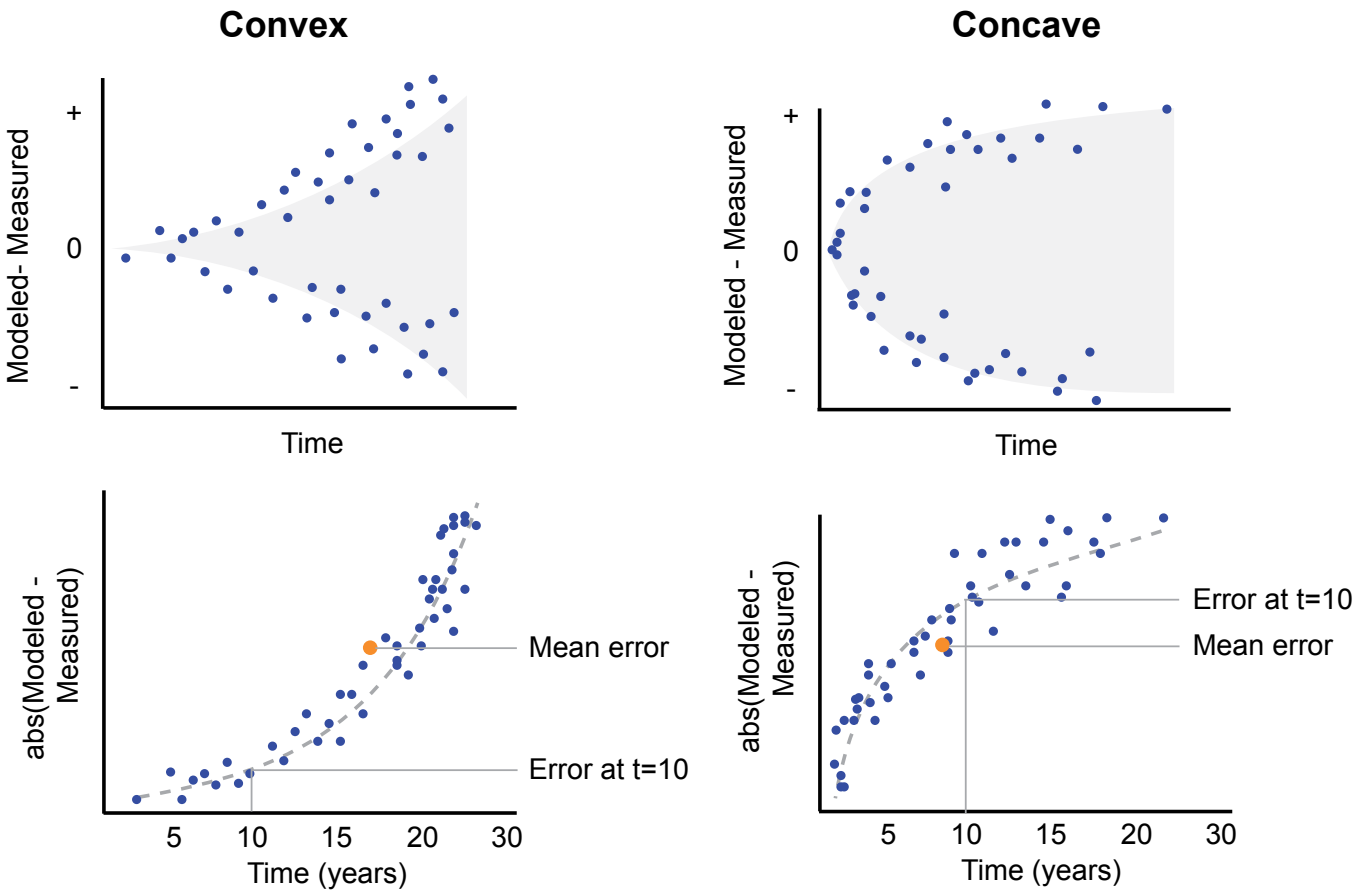
The most basic assumption (and the default in some protocols) is that model prediction error is constant over time, in which case the mean model prediction error over all of the timescales represented in the validation data could be used to represent model prediction error at any timescale for the project period (e.g., one year, five years, 10 years, etc.). A further assumption of some protocols is that this approach is conservative because model variance tends to increase with time, so using a mean that includes longer durations to represent error at shorter durations is likely an overestimate of the true error. However, this is not necessarily so. In reality, whether the mean model prediction error is a conservative estimate for a given time period depends on the nature of the relationship between model prediction error and time (Fig. 4 and see Appendix B). For example, the relationship could be concave or convex, each of which would give different estimates of model prediction error at shorter timescales (even with the same estimate at long timescales; Fig. 4). This may represent a relatively small issue for uncertainty accounting in most soil GHG and

SOC projects, but nonetheless should not be ignored, especially where non-conservative assumptions could be seen as gaming (Box 1). It is true that because model prediction error does tend to increase with time, project developers may benefit from modeling it as such rather than assuming it's constant, but reliably determining the nature of that increase requires adequate validation data spread across relevant timescales. Overall, current protocols generally do not include sufficient guidance or requirements on how to model variances over time or on what timescale the validation data must be relative to the application, which we see as an oversight.

FIGURE 4.

The relationship between model error and time dictates whether common assumptions are conservative

Example of potential relationships between model error (represented here as model prediction-measured value) and time for hypothetical validation data. The bottom panels show the difference between the mean error and the modeled error (line of best fit) at 10 years. Note that the mean error is the same in both of the bottom plots.



To better deal with model error over time, protocols could require plots of error versus time and/or absolute error versus time to assess the degree of heteroskedasticity and whether the error pattern is concave or convex along with a histogram showing the distribution of timescales of the validation measurements. A trained expert reviewer could use these plots to assess whether any claimed relationship between model prediction error and time is reasonable and conservative, especially if thresholds could be set to consistently guide their decision. Going a step further to model the error as a function of time would be better practice but may be too cumbersome for protocols to require it.

4. Spatial dependence of model errors

Current guidance for estimating project-level uncertainty often operates under the assumption that model errors are uncorrelated with the measurement values and are independent across samples. If this assumption is true, there is a strong expectation that errors will cancel out across many independent modeled points, and the relative uncertainty of the sum of modeled results for many independent points will be less than the mean relative uncertainty for a single point (see Appendix B). However, if model errors are not independent, they will cancel out much more slowly (if they are perfectly correlated they will not cancel at all). Many factors can contribute to correlated measurement and model errors and they are often spatially structured (e.g., at the field or farm level), manifesting as spatial autocorrelation of model errors. While assuming spatial independence of modeled points is beneficial for project developers, doing so without clear supporting evidence goes against both the principle of conservatism (Box 1) and standard practice in economic portfolio theory (regardless of the sampling design).[16,17] Rather, the most conservative approach, in the absence of an analysis of spatial autocorrelations, would be to assume perfectly correlated model errors (ρ=1) for modeled points when scaling up to the project level. In practice, however, the latter may be overly conservative and could render some projects economically unviable.

One potential solution would be to investigate spatial autocorrelation of model errors (e.g., based on the validation data or true-up measurements) to determine the minimum distance at which points can reasonably be assumed to be independent, and then formally account for correlated errors at distances below this threshold. Further explanation of this approach is provided in Appendix B. Alternatively (or in addition), spatial correlation of model errors could be investigated in a more general way using benchmarking datasets with control and treatment pairs, potentially on a regional basis given that relationships of model errors will depend on the spatial patterns in key controlling factors such as soil type, climate, and topography. Such studies could be used to inform on "best practice" assumptions of spatial relationships in a general way across protocols. While some studies have investigated spatial relationships in relevant environmental variables (e.g., SOC and bulk density; see Table B1 and [58]) with implications for modeling, we see a clear need for direct research on this topic and demonstration of practical approaches to assessing spatial relationships of model errors and the impacts of different assumptions on uncertainty calculations at various scales.

[16] Non-independence is especially likely for points within the same field, given likely overlaps in input data, similarities in soil types and management activities, etc. Additional discussion of this issue is provided in Appendix B.

[17] Importantly, even where measurements and model points are chosen based on random sampling designs (e.g., simple or stratified random sampling), spatial autocorrelation could still exist and should be assessed regardless of sampling design.

5. Context-dependence of systematic error

One of the pervasive concerns over using process-based models in GHG mitigation projects is whether the models are systematically overestimating GHG emissions reductions or removals (Box 1). The only way to address this is by rigorously identifying and diagnosing systematic model errors, and ensuring that there isn't systematic overprediction for any given project (Fig. 2). In practice, this is impossible to guarantee unless every part of a project is measured and modeled, which obviously can not be done. Therefore, instilling confidence that claimed project GHG emission reductions or removals are not overestimates requires thoughtful attention to potential sources of systematic model error and whether these might manifest differently in the project than during validation.

Some current protocols attempt to safeguard against likely sources of systematic model error by requiring bias be assessed separately for combinations of practice change category, crop functional type, and outcome (i.e., GHG or SOC stock). But within these categories, systematic error (e.g., at certain soil textures) could still exist and potentially go undiagnosed. Slicing the data into more and more categories to check for bias in each is impractical, because one could end up with seemingly endless combinations of key variables to divide the data by, and there is unlikely to be enough existing data to populate many such categories. While it does make sense to separate validation data according to outcome (i.e., GHG or SOC stock) and management practice because these generally represent different processes, it may be possible to use more innovative approaches that borrow strength across crop functional types, soil properties, climatic regions, and other key variables to assess systematic errors across continuous gradients rather than categorical groupings. For example, model bias (and also precision or goodness of fit) could be assessed using statistical approaches such as generalized linear/additive modeling or machine learning (e.g., using a Gaussian Process emulator/surrogate to model bias as a function of key covariates [59]).

Even the most sophisticated approaches can not guard against undiagnosed systemic error due to lack of relevant validation data. For example, a model might consistently underpredict winter N_2O emissions [e.g., 40], but this might not be apparent if the validation data do not contain many wintertime measurements. This issue also applies to time scales of measurements, for example, a model could be unbiased at longer time periods but be systematically biased at one year, which would not be apparent from a validation based on longer-term measurements. To a large extent, the potential for undiagnosed systematic error is simply the reality of current limitations in modeling capabilities and knowledge, which can only be overcome with additional research and model evaluation as more and better data become available. Yet it also highlights the potential issue of cherry-picking of validation data to avoid identification of model shortcomings, a major potential area for gaming (Box 1) which links to our final recommendation below. Regardless, this is an issue which deserves close attention from project developers, registries, and independent reviewers. Project developers must be encouraged to exercise due diligence and take reasonable steps to follow best practices, and independent model reviewers must be given clear instructions around systematic bias to ensure consistent interpretation and scrutinization of model validation reports.

6. Benchmark validation data

Right now, project developers are tasked with amassing their own datasets for model calibration and validation. In many ways, this is logical because the data must be relevant to the application, which the project developer is determining. However, this opens the door to gaming via cherry-picking of data (see above) or using the same data or sites for calibration and validation (independence of calibration and validation data is a requirement of protocols, but in practice this can be difficult to enforce). Further, it may preclude full transparency if project developers are unwilling or unable (e.g., for farmer privacy reasons) to share the complete details of the calibration and validation datasets, which has been common thus far.

One potential way to ameliorate some of these concerns is through a benchmarking platform that could be used to validate models against a common dataset. Doing so would help to increase transparency in the process and improve confidence in the performance and utility of different models. Such a platform could be designed to allow for splitting of the dataset by relevant conditions, for example, allowing a project focused in pastures in the Southeastern U.S. to validate against data relevant to those systems in terms of climate, plant types, soil types, etc. Further, such a platform could be designed to protect data privacy by allowing interaction without direct access to the data (i.e., a confidential consortium framework), greatly increasing both the potential amount of data and the diversity of their sources.

While the benefits of a hypothetical benchmarking platform are clear, the difficulties in creating one are also very apparent. First and foremost, it isn't clear who should spearhead the effort, host the platform or maintain it, all of which are monumental tasks that require funding, expertise, and dedicated personnel. It could potentially be led by a nonprofit, a public/private partnership or consortium, a registry, government, or international organization.[18] It is also not clear how to effectively encourage the sharing of existing data for use on the platform, especially where private datasets are seen as financial investments and assets. Finally, such a database would ideally need to be continually expanded and maintained as new data became available, and to support expansion into new applications and regions. Despite these challenges, we see this as a major priority to enable transparency, increase confidence, and generally advance process-based model use in soil GHG emission reduction and removal projects moving forward.

[18] Several efforts in this area are beginning or ongoing, see https://www.soilcarbonsolutionscenter.com/ecosystem-modeling-and-data-consortium; https://www.nasaharvest.org/initiatives/sustainable-and-regenerative-agriculture-sara; and https://cchange.research.iastate.edu/

Summary table of recommendations

Topic/Concept	Recommendation	Included in Current Guidance?	Relevant Section
Modeling Workflow Consistency (initialization, pre/post-processing, etc.)	The approach used should be clearly documented and performed consistently across the project steps, with any deviations justified Research is needed to understand how and why specific deviations lead to meaningful differences in model predictions	Unclear (may be included in validation review and/or verification process, but not explicit in public-facing guidance)	The Project Steps: Modeling workflow Recommendation #1
Calibration	Calibration data should include repeated measurements over the time period of interest for the intended application	No, but shortcomings of the calibration will ideally be apparent during validation. Further, if true-up measurements are used to recalibrate the model, this recommendation will be met at that time.	The Project Steps: Calibration
Parameter Set	The same parameter set used in the validation report must be used throughout the project	Yes	The Project Steps: Project Prediction
Validation Data	Validation data must be independent from what was used to calibrate the model, including for cross-validation methods (i.e., within folds)	Yes, though it can be difficult to fully confirm	The Project Steps: Validation data Recommendation #6
	Datasets used for evaluation should span the full ranges of relevant input variables (soil texture, pH, aridity index, etc.) within the project domain, with relevance determined based on the process and geography being modeled	Yes, but needs clarification (e.g. for which variables, how is coverage defined, etc.)	The Project Steps: Validation data Recommendation #2 Recommendation #5
	Validation data should include timescales that match or exceed the timescale of the project predictions Where the above is not possible, assumptions about how model error relates to time should be clearly justified and thoroughly reviewed	Unclear	Appendix 2 Recommendation #2 Recommendation #3
	Ideally, validation data should include points at the same spatial scale as what is being modeled for the project, including the sub-field scale if projects intend to model sub-field scale heterogeneity.	No	Appendix 2 Recommendation #2 Recommendation #4
Validation	At a minimum, models must be evaluated separately by management practice type and outcome (i.e., GHG or SOC stock)	Yes, and some require additional splitting by crop functional type. However, exceptions to this requirement (i.e. less granular validation) have also been allowed.	Recommendation #5

Topic/Concept	Recommendation	Included in Current Guidance?	Relevant Section
Model Prediction Error	Model prediction error should account for both accuracy and precision, including when used to test goodness-of-fit and calculate uncertainty deductions Changes in model prediction error over time should be modeled where possible, and unjustified assumptions avoided. What constitutes a conservative approach can not be generalized based on current knowledge. Research is needed to determine whether generalized guidance for modeling changed in model prediction error over time is possible	No	Appendix 2 Recommendation #3
Project-level Uncertainty	If spatial correlation of errors is not diagnosed, errors should not be assumed to be independent If spatial correlation of errors is diagnosed, correlated errors must be accounted for in calculations, and where modeled points are closer/ farther than the minimum/maximum distance tested, they can not be assumed independent Research is needed to better understand spatial dependence of model errors for these applications, identify practical approaches for assessment, and demonstrate the impacts of different assumptions on uncertainty calculations	No. While spatial correlation is mentioned, no requirements for diagnosing it are given, and independence is a pervasive assumption.	Appendix 2 Recommendation #4
True-up	Where repeated measurements are available, the model should be recalibrated and revalidated to obtain an updated project-level uncertainty estimate If ground-truthing is identified as a critical need, protocols should require resampling aimed at detecting and verifying changes in GHG emissions and SOC stocks over time (i.e. repeated measurements with adequate precision)	No. While some require measurement of a proportion of the project after initiation (e.g. after 5 years), the details of how the measurements are to be used are unclear. Further, some protocols forbid remeasurement of a certain proportion of points within the project.	The Project Steps: True-up

APPENDIX A

Background on Select Modeling Workflow Components

Model initialization

Here we define initialization to mean any method of establishing initial conditions of the model (including states). Initialization approaches can vary in several aspects, including whether and how a spin-up period is employed to inform or adjust initial values or conditions, whether and how measured data are used to set state variables [60], and whether or not equilibrium conditions are assumed. Differences in model structures and behavior mean that an effective initialization approach for one model may not work well for another, and even within a given model, the same initialization approach may work better for certain sites and scenarios than others.

Despite differences in initialization methods being commonplace [37, 61], existing protocols and model guidance documents typically do not cover initialization. This is problematic, for example, if one approach to model initialization is used during calibration and validation and a different approach is used when making project predictions. In the case of a carbon crediting project, it will no longer be clear whether the application will qualitatively meet the required accuracy and precision criteria, nor whether the correct uncertainty deduction is being applied. More generally, the relatively slow nature of soil carbon pool change means that initial condition uncertainty tends not to decline rapidly over time. Models will also be sensitive to how other model pools are initialized, such as soil moisture, soil nutrients, and vegetation carbon pools, though guidelines tend to provide even less information on how these pools should be initialized and updated (i.e., "true-up") than they do for soil carbon.[19]

Model initialization is especially crucial in the context of agricultural GHG mitigation projects because it can influence whether model-predicted GHGs and SOC increase or decrease in the early years of a project [62–64]. For example, incorrect assumptions during initialization can produce fallacious trends in SOC stocks as the state variables drift back towards the model's equilibrium [65]. In practice, this means that different initialization methods might generate meaningful differences in modeled SOC outcomes, particularly for the short-term projections (e.g., one to five years) that are used in many GHG mitigation projects, which can be highly sensitive to transient fluctuations post-initialization.

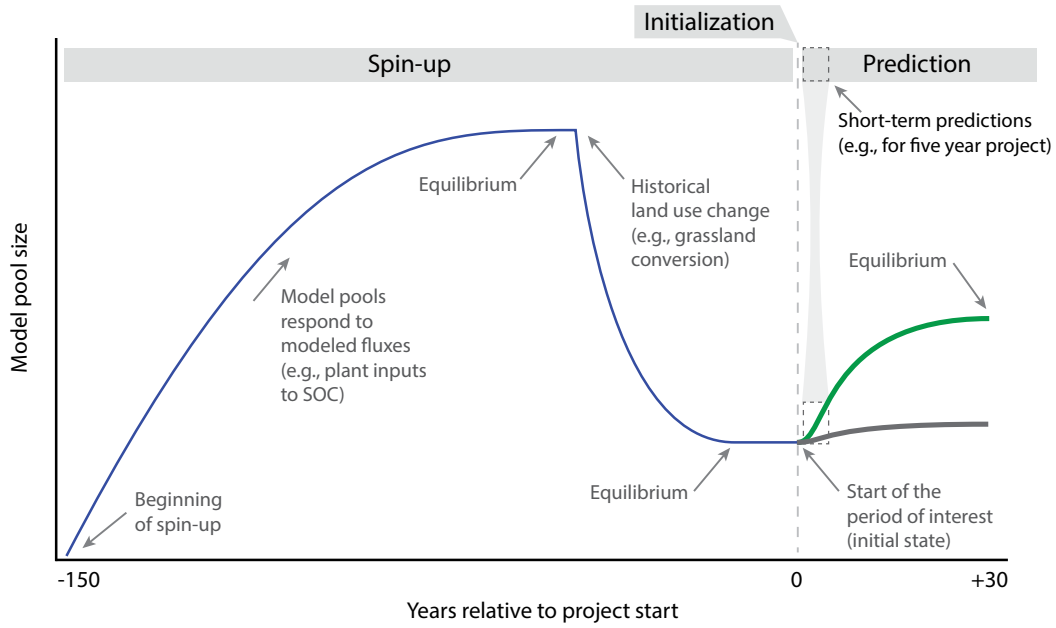
Here, we use the example of initialization of SOC pools to describe some important differences between approaches, including a demonstration of how results can differ depending on the approach used.

[19] Note however, that for a model that simulates these types of highly temporally variables with short carryover effects of no more than a few years, such as ecosystem models like DayCent and DNDC, inaccuracies in their pool values after several years would suggest that the model is not well calibrated.

FIGURE A1.

Different methods of running models may invoke spin-up periods or equilibrium assumptions

Representation of major steps in the modeling workflow (gray boxes) in relation to a conceptualization of a model run. Arbitrary time spans were chosen here; spin-up periods can be thousands of years long. In practice, each step can be attempted, evaluated, and iterated multiple times before a final approach is chosen and applied to the final model run in which all the steps are performed. Error propagation is not shown and is typically disregarded; if error is propagated through the spin-up process, the initial condition distribution can be very large.



Spin-up and equilibrium assumptions

Spin-up is a process of allowing a model to run over a period of time to reach a stable or "realistic" state prior to simulating the period of interest (e.g., the beginning of an experimental treatment or GHG mitigation project; Fig. A1). Spin-ups are frequently a part of the initialization process as they are used to set the model up to run well during the period of interest and may be used to inform initial pool sizes and/or the relative allocation of total SOC among different SOC pools. The choice of whether or how to use a spin-up depends primarily on the model; for some a spin-up is uncommon because it does not typically improve model performance, while for others a spin-up is considered a prerequisite to running the model [30]. When used, a spin-up can help to overcome transient and unstable model dynamics, which result when model pools do not match the model's tendency for that system, so that they do not affect the model run during the period of interest. For example, a spin-up for a model representation of a grassland system might entail beginning with no SOC — which is obviously not realistic — and running the model with typical climatic conditions and plant C inputs until a steady-state (presumably more realistic) SOC level is reached. Then, a treatment change can be imposed in the model from a more realistic starting point in terms of system dynamics. A spin-up can also be used to put the model on the correct directional trajectory prior to the start of the period of interest. For example, if a system was recently converted from grassland to cropland, SOC stocks will be decreasing through time. As this trajectory will impact the trajectories of other model pools and fluxes, representing it in a spin-up may improve later model predictions.

If a spin-up is used, the length of time represented during a spin-up is at the modeler's discretion, and several factors might influence the decision. Some may choose to run a relatively long (i.e., 1000s of years) spin-up to ensure model stability (e.g., [66]), while in other cases such a long spin-up may not be necessary to achieve acceptable model performance, or a lack of computing power or historical data might dissuade the modeler from running a long spin-up. Designing a spin-up raises the additional questions of whether the assumed scenario and model inputs during the spin-up period accurately reflect the history of the site, and whether the system (or pools of interest) can or should be allowed (or "forced") to reach equilibrium. Most approaches to model spin-up tend not account for the impact of these uncertainties in model parameters, drivers, scenarios, and process error, and those that explicitly propagate such uncertainties often produce extremely large initial condition uncertainties [67]. The shortcoming of unknown past conditions during the spin-up period, which may lead to inaccurate initial pool sizes, can be compensated for to some extent by initializing the model with direct site measurements (see Using Measurements to Set Initial States). For the assumption of equilibrium, there is considerable disagreement among modelers as to the best approach.

For long-term SOC simulations, modelers often assume that SOC pools must be in equilibrium or "steady-state" prior to application of experimental treatments or management changes [61, 68, 69]. But many argue this assumption is unrealistic because of widespread changes in climate, land use, agricultural practices, crop cultivars, and soil biota [70–72]. Soils that have been disturbed recently, or even centuries ago, are often in a transient state (rather than an equilibrium state) due to very slow rates of change of slower-cycling model pools (e.g., "passive SOM" in DAYCENT). Indeed, on multi-centennial timescales, historical climate is itself constantly changing rather than being at steady-state. Despite this, equilibrium conditions are often assumed at the outset by the modeler due to absence of requisite measured information to prove otherwise [73, 74], and its usefulness in avoiding spurious, transient model behavior. Assuming equilibrium when it is not the case in nature can result in (1) model calibration outcomes that overestimate decay of the slow pool and (2) simulated equilibrium situations that overestimate stocks of recently disturbed sites. However, clear evidence for or against equilibrium assumptions is elusive, and there are mixed opinions and evidence for any one method over another. For example, in a recent inter-model comparison study, models that started with a 5 to 10 year spin-up could do as well or better than those that were forced through equilibrium to simulate SOC dynamics in long-term bare fallow soils [30]. Another study on relatively undisturbed grassland sites showed that the assumption of equilibrium did not significantly impact model results [75].

Given that equilibrium assumptions are rarely, if ever, met, we caution against approaches that assume equilibrium for the reasons above. However, for some models a spin-up period run to equilibrium may be the best method of attaining more accurate model outputs, especially over shorter model prediction intervals. If equilibrium is assumed, so long as the same assumptions are used in the modeling done across all project steps including model validation, we do not see the assumption as reason enough to discount a particular approach.

Using measurements to set initial states (initialize pools)

When modeling to predict soil carbon stock changes, some of the most important targets of initialization are the sizes of the SOC pools, including the relative distribution of total SOC amongst model pools with different dynamics (e.g., active, slow, and passive pools in DAYCENT), as this has major impacts on model behavior [63].[20] Modelers can use direct measurements to initialize model pools at the start of the modeling period, and this is considered by many to be the best approach, as it may improve the accuracy of later predictions. Ideally, specific measurements corresponding to each SOC pool (rather than just total SOC) would be used to most accurately represent the starting distribution of total SOC among model pools. However, most common SOC models employ conceptual pools that don't correspond robustly to measurable entities, and the best way to use measurements to initialize model pools is not clear, especially across different models. There have been many efforts to understand whether and how soil samples can be partitioned (or "fractionated") to produce measured proxies for conceptual model pools [71, 76–83] with some successes for specific models, but it is not always clear that doing so improves model outcomes [60, 84–86]. Given this, there has been a push for the development of new SOM models and updated versions of existing models based on measurable fractions (i.e., model pools and their dynamics are based on measurable entities rather than theoretical concepts) as a means to overcome this issue, and increase model-measurement compatibility for all aspects of the model workflow from initialization to evaluation [87–89], though these have not yet been deployed for large-scale agricultural GHG mitigation projects. Further, soil organic matter fractionation is costly and time-intensive (though improvements in high-throughput quantification methods such as spectroscopy are being made), meaning that obtaining the necessary fraction data may not be feasible in many cases.

Given these issues, measurements for initialization are often limited to just total SOC and estimates of its distribution amongst its constituent pools must be determined by other means. One way is to spin-up the model until its SOC pools add to match an initial measurement of total SOC. This can be done in two ways: (1) the spin-up can be designed so that conditions are right for SOC pools to reach equilibrium levels that match the measurements (e.g., by altering the rate of plant C inputs; [90]) or (2) the total SOC after the spin-up can be manually adjusted to match the measurement (multiplying all SOC pools by the same scaling factor) under "relaxed equilibrium assumptions" [62, 64]. Doing so results in a distribution of the total C among the pools that matches model predictions for the spin-up scenario. The use of spin-ups in any capacity raises the issues posed above of accuracy of the spin-up conditions and assumptions of equilibrium, which are likely to represent shortcomings of the approach.

Alternatively, the initial distribution of SOC amongst model pools can be set manually using defaults based on site attributes such as coarse land use history (e.g., long-term cropland versus long-term grassland) [91]) but doing so would ignore effects of any recent changes in land use or specific site attributes (e.g., soil texture) that might cause deviations in the distribution of SOC from the default values.

Finally, iterative data assimilation approaches could be used to initialize SOC pools by statistically fusing ensemble model predictions with observations [92, 93]. This bears some similarity to the spin-up approaches described above but explicitly accounts for both model and data uncertainties, rather than treating what is usually a sparse sample of field measurements as "truth" and updates unobserved pools based on the strength of their correlation with observed pools. Assimilation approaches don't require an equilibrium assumption and would be particularly valuable for cases where repeat measurements are present.

[20] A common distinction is made between the term "pool," which is used to describe conceptual entities in a modeled context, while other terms may be used for measured entities. For example, for SOC, "fraction" is used to describe measured entities (which may be related to modeled pools).

The above approaches assume that some direct measurements of total SOC or SOC fractions (proxies of model pools) are available to the modeler for initialization purposes, however, this may not always be the case. Without direct measurements, the modeler has several options for estimating SOC, but the associated uncertainty might be significantly higher than if direct measurements are used (and it must be accounted for). One option would be to acquire estimates of total SOC from an empirical model product such as SSURGO [94]. However, given that SOC can vary significantly across sites with similar characteristics, these estimates might be quite far from the true values [95]. In such cases the initialization uncertainty should be inflated to account for this, for example, through spatial statistical models that formally account for interpolation uncertainty (e.g., quantile regression [96, 97], kriging). Alternatively, a spin-up simulation could be run without any adjustment based on measurements, resulting in predictions of SOC pool values based on historical rates of plant inputs to SOC and climate, though the resulting initialization uncertainty might be relatively large [67]. If measurements are a future possibility, spatial statistical models could be used to construct informative Bayesian priors that are updated using site-specific field data, which could reduce initialization uncertainty over field sampling alone or allow for a comparable level uncertainty with reduced sampling effort.

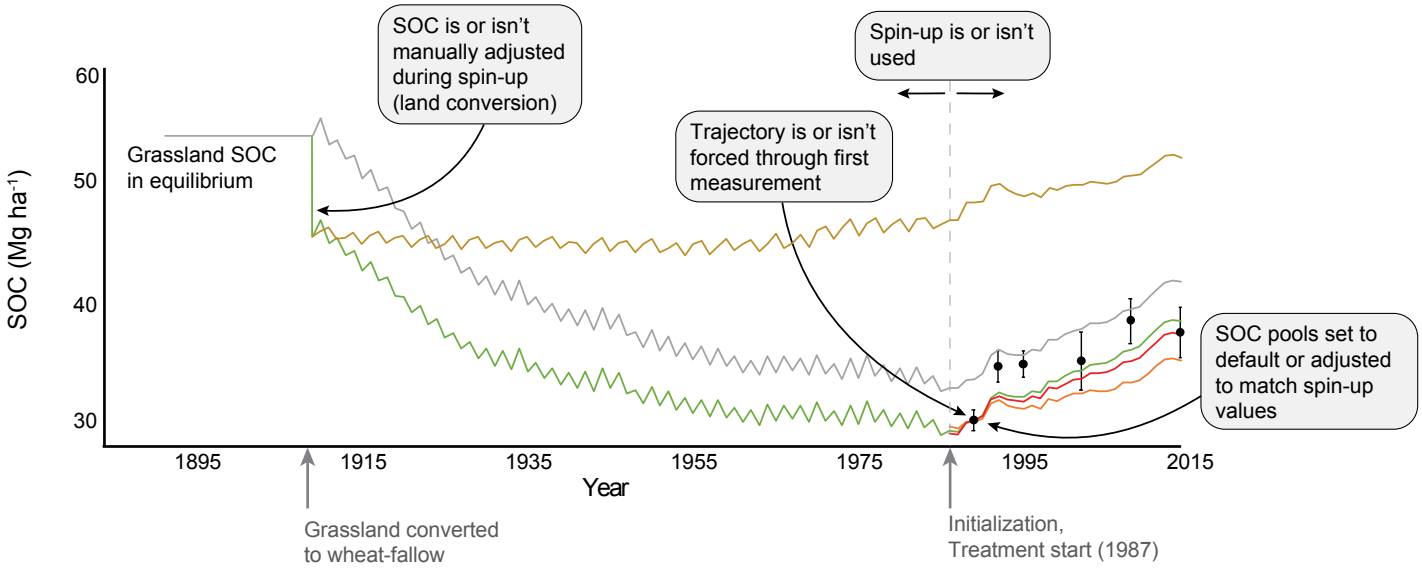
In theory, if a model is designed well (with realistic processes accounting for climate, decomposition, plant inputs, soil texture, etc.), and the input data is relatively accurate for the period of interest, then even if the model starts at an unrealistically low or high SOC level it will self-adjust appropriately through time to reach more realistic levels. That said, many argue equilibrium spin-up assumptions are rarely, if ever, met in nature [71, 72] as discussed above (Spin-up and Equilibrium Assumptions). Hence, wherever possible, we recommend the use of site-specific measurements of total SOC or SOC fractions for initialization of any project aimed at predicting changes in SOC and note that some current protocols require this. Further, because SOC is a major control on CH_4 and N_2O fluxes to and from soil, we maintain this recommendation for any project aimed at quantifying those fluxes.

Beyond recommending that measurements be used, more specific recommendations on ideal initialization approaches are not necessarily warranted at this point in time. There is currently a lack of scientific consensus on best practices, and thus there is a need for further analysis, development, and innovation. We do not yet know how different initialization approaches affect the accuracy of model predictions across sites, especially in the context of agricultural GHG project MMRV (i.e., for baseline-project comparisons, short timescales, etc.). That said, we can recommend that regardless of the initialization approach taken, it should be documented clearly and kept consistent across the project steps (to the extent possible, with any deviations clearly explained and justified), because it is clear that initialization approaches can affect outcomes. Figure A2 shows an example of modeling the same site using different initialization approaches using data from a continuous wheat plot at Swift Current Research and Development Centre in Saskatchewan as described by He et al., [98]. The predicted SOC stocks can be very different depending on the assumptions and approaches taken when initializing the model. However, if the same assumptions are made for both the practice change and baseline scenario, the difference in outcomes (i.e., practice change minus baseline) between different initialization approaches is often relatively small, on the order of 1-5 Mg ha^{-1} over 20 years (data not shown but available upon request; see also [94]). That said, it is not zero and even small differences could have large implications for the economic feasibility of a carbon crediting project, for example. Further, differences could very well be larger for different sites, especially those with more responsive plant communities and faster-changing SOC stocks (note that the Swift Current, Saskatchewan site shown here has a fairly dry climate). Future research could explore the effects of initialization approaches on modeled GHG emission reductions at additional sites, and with additional models, which could help to inform guidance on best practices and help to support rigorous review of model validation reports.

FIGURE A2.

Initialization approaches vary and affect final predictions

Demonstration of differences in SOC trajectories introduced by deviations in initialization approaches using the DayCent model. Black circles are field measurements with standard error bars. Colored lines are different model runs, each using a different initialization approach with variations in spin-up, manual adjustment of initial SOC pools, and cropping history. Such deviations in initialization approaches are common in practice depending on information available to the modeler (e.g., cropping history), which approach produces the best model performance after iterative attempts and evaluation, and other factors.

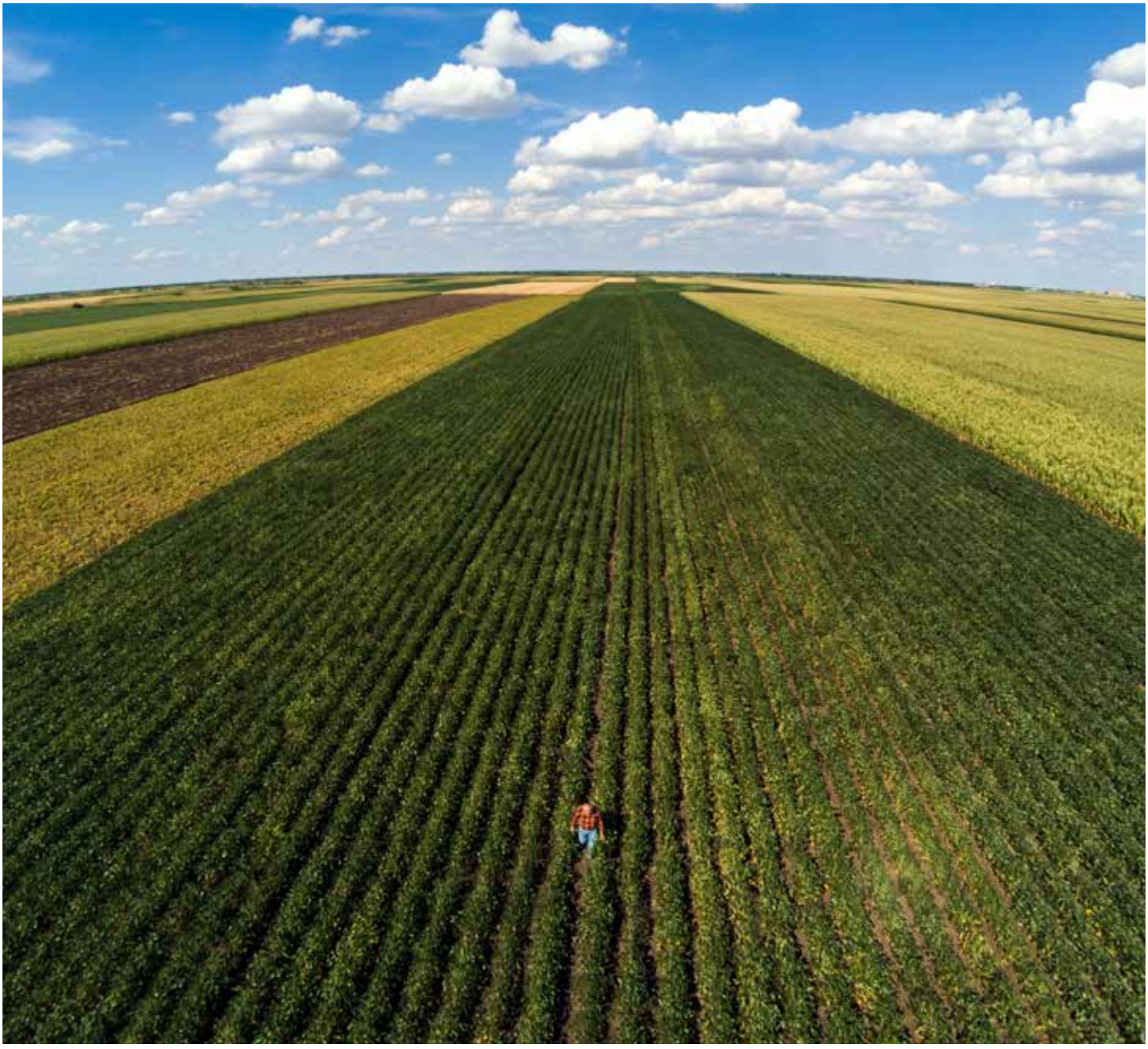


Data assimilation

Iterative data assimilation is an approach to model initialization and/or calibration that is used operationally in many disciplines (e.g., weather forecasting, navigation), but has yet to have much application in GHG MMRV outside of a research context. In an iterative data assimilation system, models are used to make probabilistic forecasts from one point in time into the future, and then statistically update (i.e., analysis) and re-forecast as new observations are encountered.[21] During the analysis step most data assimilation systems are designed to update model state variables (i.e., initial conditions) but some variants will also update model parameters or both parameters and states. Prior to model verification, the advantages of iterative data assimilation are related to computational efficiency and the ability to distinguish model process error from data observation error, but otherwise should give very similar parameter estimates as other methods (including the potential to include hierarchical variability). The real distinction with iterative methods is their ability to seamlessly continue to update states and parameters as one shifts from calibration to application. For modeled state variables, this provides a statistically optimal way of achieving true-up (e.g., updating SOC pools based on project verification measurements). On the parameter side, assimilation allows model parameters to continue to "learn" from verification measurements, including the potential to learn about local parameter deviations from

[21] Data assimilation has a forecast-analysis cycle. "Reanalysis" is used to describe a data product generated by a post-hoc run of a data assimilation system using cleaned data and observed drivers (i.e., it is a re-run of the analysis system).

the broader across-site calibration. While some protocols do allow the use of assimilation systems in concept, and such approaches have been used successfully in multiple regional- to global-scale land data assimilation systems, to our knowledge they have not yet been adopted in practice. The ability of assimilation systems to update model parameters post-validation (current protocols are currently unclear in their guidance on this) also raises questions about what additional validation requirements should be placed on such systems to ensure that the underlying workflows continue to meet that market's accuracy and precision requirements.[22]



[22] It is our understanding that state data assimilation might be permitted in some current protocols, but the guidance is vague and we have not seen it done under any existing protocols in practice. However, the guidance seems clear, as currently written, that parameter data assimilation would require the generation of a new validation report and independent model expert (IME) review, which can be a cumbersome and prohibitive process. A potentially valuable clarification is whether the assimilation algorithms could be verified in a way that would permit parameter learning without triggering a new review.

APPENDIX B

Spatial and Temporal Considerations for Model Prediction Error and Project-level Uncertainty

When scaling model predictions in space and time to come to estimates at the project level for a given project period, the project-level uncertainty must be estimated. Doing so requires appropriate handling of model prediction error, which introduces considerations of relationships between errors of different modeled points. Here we focus on spatial and temporal autocorrelation and temporal heteroscedasticity of model prediction error. If errors in measurement or model errors are correlated, they will not cancel as quickly as uncorrelated errors during spatial upscaling, and the resulting uncertainty estimate will be higher than if the correlation were ignored. Measurement errors, including errors in change estimates based on two measurements in time, can also be autocorrelated, and this possibility should be accounted for when handling soil sample data, for example, during any initial sampling or true-up. If errors are spatially autocorrelated, more measurements across space would be needed to reduce relative uncertainty to the same level as when errors are uncorrelated. Importantly, even where measurements and model points are chosen based on random sampling designs (e.g., simple or stratified random sampling), spatial autocorrelation could still exist and should be assessed regardless of sampling design.

Spatial dependence of model errors

Typical SOC models are designed for the point scale, representing biogeochemical processes as either two-dimensional (time, depth) or as depth-averaged time-series. In practice, the point scale can be considered equatable to the small plot scale (in the order of 10^1~10^3 m^2) or agricultural management research experimental area (i.e., the mean of multiple replicate plots within an agronomic field experiment interspersed over 10^3 to 10^6 m^2) since data collected at these scales are commonly used for calibration and validation of these models. The means for input data and modeled results are assumed to be representative for these practical point scales even though the plot or small experiment areas are not assumed to be completely homogenous. Modeling for a GHG mitigation project bigger than this scale (e.g., multiple fields or farms) requires upscaling the modeled results for the point scale to the scale of a project. A given project might involve many thousands or even millions of hectares and all land parcels within the project might not be contiguous.

The effect that spatial upscaling will have on application modeling uncertainty depends on how errors for modeled points within the project interact (i.e., their spatial dependence). If the errors of the model results (e.g., for SOC stock change) at different points in the project are independent (no spatial dependence), then they will often cancel out to some extent due to random variation in the uncertainty as the number of modeled points increases. In this case, if the predicted SOC stock change is overestimated at one point (positive error), there is an equal chance that it will be underestimated (negative error) at another point. Hence, there will be a strong chance that some of the errors will cancel out across many independent modeled points, and the relative uncertainty of the sum of modeled results for many independent points will be less than the mean relative uncertainty for a single point. Further, the relative uncertainty of the sum will decrease as the number of modeled points in the sum or project increases (this is true whether or not the points are modeled as independent, but if they are modeled as independent it will decrease more quickly)[52]. For example, if the independent points are an unbiased sample of size n from a population with a constant uncorrelated error σ, the uncertainty of the mean will rapidly decrease towards zero as $\sigma/\sqrt{n}$. Developers of large projects with many enrolled fields rely heavily on this cancellation of errors to reduce both their overall uncertainty penalties and their exposure to risk through what is known as the portfolio effect:

$$\textit{portfolio variance} = \sum_{i} w_i \sigma_i^2 + \sum_{i} \sum_{j \neq i} w_i w_i \rho_{ij} \sigma_i^2 \sigma_j^2$$

where w is the weight associated with a site (typically it's fractional area relative to the whole project), σ^2 is the site-specific variance, and ρ_{ij} is the correlation between sites *i* and *j*. The second term accounts for spatial correlations, and is a standard part of portfolio theory, but has often been dropped from GHG accounting projects.

In contrast, if the errors are systematic in space (i.e., dependent, correlated, or non-random), then the error for different points would tend to be similar, and during upscaling (if done correctly) errors would cancel out much more slowly. Indeed, if spatial errors are perfectly correlated (ρ=1), then spatial errors won't cancel out at all, leading to much greater uncertainty about GHG sequestration.

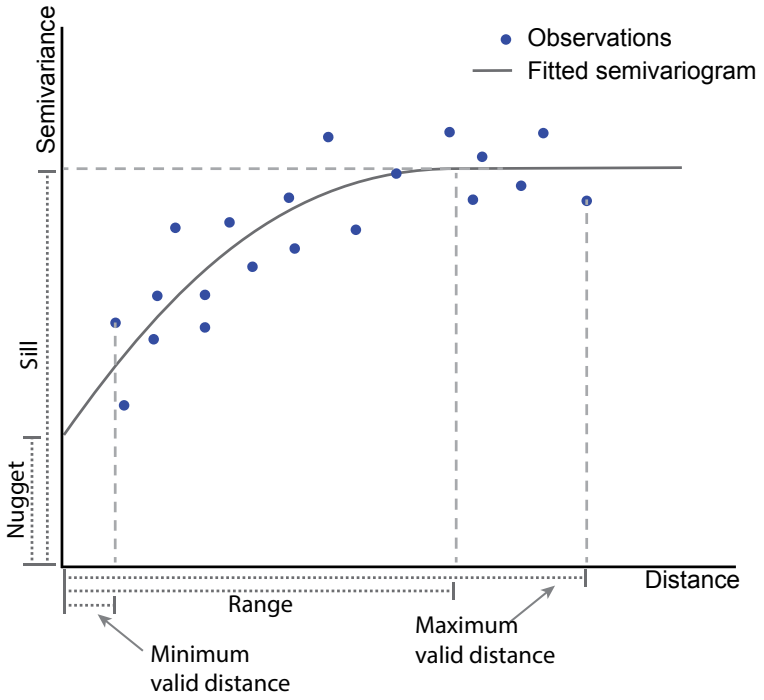
In relation to model errors, spatial autocorrelation is a measure of this degree of spatial dependence of model errors and in practice is often much closer to 1 than to 0. Spatial autocorrelation in a model's final predictions can arise from spatial autocorrelation in any (or all) of the uncertainties going into the modeling process, whether that be autocorrelated uncertainties in meteorological drivers, soil properties, carbon pool initial conditions, model parameter uncertainty (mean) and variability (random effects), model process error, and model structural assumptions. Indeed, in most cases all of these uncertainties have positive spatial autocorrelation. In practice it is not uncommon to attribute most (if not all) of the spatial error to the model residuals, though this can lead to some degree of scaling error (especially at larger scales) when the scales of autocorrelation are different for processes or have different spatial patterns.

Given the potentially large spatial scale of autocorrelation in key input variables in biogeochemical and ecosystem models used to predict soil GHG dynamics (e.g., climate, soil properties), it is inadvisable to assume without evidence that model errors will be independent across any scale, and therefore model errors should be assumed to be correlated until proven otherwise.

FIGURE B1.

A semivariogram approach to investigate spatial autocorrelation of model errors

Conceptualization of a semivariogram and its key properties. The nugget is the value of the semivariance when extrapolated to distance 0, which gives an estimate of fine-scale heterogeneity. The sill is the value of the semivariance at the asymptote, which is related to the "background" variance. The range is the pairwise distance at which the sill is reached, which is interpreted as the distance at which errors can be treated as independent. Maximum and minimum valid distances are defined in the main text.



Using a semivariogram to investigate spatial autocorrelation of model errors

One means of providing such evidence would be to investigate validation data for signs of spatial autocorrelation using a semivariogram approach during the validation process. A semivariogram plots semivariance (mean squared difference in response variable for any two points) as a function of the pairwise distance between locations. Typically, points closer together are more similar to each other (i.e., non-independent), and thus have lower semi-variance, and then semivariance increases to an asymptote determined by the background (uncorrelated) variance. In this case we are interested in knowing when either the residual (validation and/or project) or predictive model errors are autocorrelated.

For the sake of the current discussion the semivariogram helps us determine three key thresholds: the range, the minimum valid distance, and the maximum valid distance (Fig. B1). The range is the distance beyond which errors can be treated as independent (i.e., the distance where the semivariogram asymptotes). At distances shorter than the range, the pairwise covariances between points should be modeled explicitly, usually by fitting a parametric model to the semivariogram, as is described in textbooks on spatial statistics, geostatistics, or kriging [99]. This will primarily impact the estimated uncertainty around the sum or mean when combining predictions within or across sites.

The second key threshold is the minimum pairwise distance over which the semivariogram is valid (i.e., minimum valid distance). For example, if a validation dataset was well-separated in space, such that the mean minimum distance between points was 100 km, then it is not possible to calculate semivariance over distances shorter than this, and thus it is not possible to know how autocorrelated errors at shorter distances might be. Note also that adding in two validation points that are 10 m apart to a dataset that is otherwise 100 km apart does not permit one to make inferences about autocorrelation on scales between 10 m and 100 km. Validation data should ideally provide adequate sample sizes across different spatial scales to allow for reasonably well-constrained estimates of semivariance and basic assessments of stationarity and isotropy (i.e., that the semivariogram itself isn't dependent on location or direction). Fortunately, the number of pairwise distances among points increases quadratically with the number of points, so the number of data points required need not be huge. Given that spatial autocorrelation can only be checked for a minimum distance of the sites included in the validation data, it is not possible to assume that any points within the project that are closer than that minimum distance would not have autocorrelated errors. Where modeled points in the project are closer than the minimum distance of the validation data, they must be assumed to be perfectly correlated unless proven otherwise. This would be true even if no autocorrelation was detected at scales larger than the minimum distance threshold. When modeled points are larger than that minimum distance but closer than the range, then a formal accounting of spatially autocorrelated error should be included for those points in uncertainty calculations.

The third relevant threshold is the maximum pairwise distance over which a semivariogram is valid (i.e., maximum valid distance), which is directly analogous to the minimum distance threshold but is determined by the upper end of the pairwise distances. This threshold determines the scale of the "background" variance that the semivariogram converges to. While this threshold is often less of an issue, as existing protocols already require validation datasets to span the range of observed variability, it is nonetheless worth checking, especially if the data appears to be clustered, as estimates of the range are sensitive to the choice of maximum distance. For example, if a dataset sampled soils only within a single 1 km^2 block, one might conclude that the range of the model error is on the order of tens to hundreds of meters. One could not then apply this model over larger distances and argue that errors are independent, both because the model has not been validated for such extrapolations and because the model errors have not been shown to be independent when confronted with larger-scale environmental gradients (e.g., all points within the original 1 km^2 block may be systematically biased [i.e., possess a shared spatial error] relative to these larger gradients).

This raises a key point when interpreting Table B1, below: The ranges presented are limited by the maximum valid distances of those analyses and it cannot be assumed that errors at larger scales are uncorrelated. That model errors may be correlated at much larger scales than the maximum valid distance in Table B1 highlights the need for further research on error correlation at larger spatial scales. The relevant ranges in these cases may very likely be hundreds of kilometers rather than hundreds of meters. Further, it is currently unclear whether spatial error should be assessed separately for different contexts, (e.g., practice x region) or in a generalized way (e.g., for all of the continental U.S.). The answer may depend on how exactly the model was calibrated and validated, and the topic deserves further research.

Finally, it is worth noting that autocorrelated errors will not always result in a higher total uncertainty estimate. While positively autocorrelated errors do increase the uncertainty when calculating a sum, average, or spatial integral, they decrease the uncertainty when calculating a difference (e.g., change detection; [94]). For example, if the modeled error for two near-by sites is positively autocorrelated, that means that if one site is above or below average then the other is likely to be as well. In calculating the difference between sites, part of this shared error does cancel out, and indeed cancels out faster than independent

errors, with the magnitude of this effect proportional to the strength of the correlation. This concept also applies to temporal autocorrelation and detecting change over time. That said, in typical agricultural soil GHG projects the aim is to sum over space to estimate total project outcomes, rather than to detect change over space (change detection is generally limited to single sites or fields, which are then summed over space). Further, measurements over time may be subtracted, as when calculating changes in SOC stocks, or summed, as when calculating total emissions of N_2O or CH_4.

TABLE B1.

Examples of estimated spatial dependence of SOC and bulk density

Note that the statistics calculated in these studies are limited by the maximum valid distances of the studies, and it therefore cannot be assumed that spatial autocorrelation does not exist beyond these distances (i.e., extrapolation to larger spatial gradients is unwarranted without direct investigation of those spatial scales). Further note that spatial relationships in SOC stocks are related but not equivalent to spatial relationships in the changes of SOC stocks over time relative to a baseline, which is the focus of soil carbon removal projects. Properties and typical units are stocks (Mg C ha^{-1}), concentration (conc., g C g^{-1} soil), and bulk density (BD, g cm^{-3}). The Nugget:Sill ratio gives an indication of the relative amount of fine-scale heterogeneity that is unaccounted for.

Soil Property	Land Use	Location	Nugget:Sill Ratio	Range (m)	Max Valid Distance (m)	Reference
SOC stocks (0-20 cm)	improved pasture	Florida, USA	0.18	135	350	Xiong et al. [100]
BD (topsoil, 2-5-7.5 cm)	Cropland	Nottinghamshire, England	0.18	7	20	Lark et al. [101]
BD (subsoil, 32.5-37.5 cm)			0.72	27		
SOC conc., topsoil			0.57	43		
SOC conc., subsoil			0.18	204		
SOC stocks, topsoil			0.22	28		
SOC stocks, subsoil			0.48	31		
SOC mass, topsoil	Cropland	Manitoba, Canada	0.06	120	400	Bergstrom et al. [102]
SOC mass, A horizon			0.17	130		
BD	Grassland, site 1	Thuringia, Germany	0.76	131	200	Don et al. [103]
SOC conc.			0.22	199		
SOC stocks			0.24	514		
BD	Grassland, site 2		0.58	247		
SOC conc.			0.62	233		
SOC stocks			0.47	86		
SOC conc.	Average of 10 farms	New South Wales, Australia, 2013	0.47	215	1000	Singh and Whelan [104]
SOC conc.		2015	0.70	238		

Soil Property	Land Use	Location	Nugget:Sill Ratio	Range (m)	Max Valid Distance (m)	Reference
SOC conc.	Cropland	Maryland, USA	0.66	102	400	Lengnick [105]
SOC conc. (0-20 cm)	Watershed with 74% cultivated land (remainder grazing land, forest, and bush)	Ethiopia	0	381	375-490	Addise et al. [106]
SOC conc. (20-40 cm)			0	270		
SOC stocks (0-20 cm)			0.16	276		
SOC stocks (20-40 cm)			0	394		

Spatial relationships at small scales

When modeling GHG responses in a single field or farm, it is possible to model multiple points within that field or farm, rather than a single point assumed to be representative of the entire area, in an attempt to capture fine-scale heterogeneity. For a project developer, one benefit of doing so might be increasing the number of modeled points with relatively small increases in labor of soil sampling (e.g., for initialization purposes) or input data gathering (e.g., through farmer surveys). This has the potential advantage of capturing and explaining some of the variability attributed to random sampling, thus reducing overall residual error and hence the project-level uncertainty estimate (and uncertainty deduction if applicable). However, there is good reason to believe that (1) process-based models calibrated at larger scales may not be effective at predicting smaller-scale heterogeneity (i.e., $< 10^2$ to 10^3 m) and (2) points within the same field or farm should not be treated as spatially independent points for uncertainty purposes.

The first point highlights a shortcoming of many common validation procedures which do not test the ability of models to capture fine-scale heterogeneity. In part, this shortcoming is due to a lack of data to rigorously test model performance at the sub-field scale. Ideally, validation data should include points at the same spatial scale as what is being modeled for the project, including the sub-field scale as appropriate. As noted in the previous section on spatial autocorrelation, where this is not possible, modeled points closer than the minimum valid distance of the validation data should conservatively be treated as having perfectly correlated model error ($\rho=1$).

The second point is discussed in more detail in the above Spatial dependence of model errors section. Expanding on that discussion, this sort of fine-scale modeling needs to account for spatially correlated errors in models at the fine scale, which includes not only residual errors but also shared parameter and driver/input errors (it is very common to use climate, soil property or other input data from spatial databases with coarse resolutions). If modeling protocols do not account for this sort of fine-scale spatial separation or this sort of more sophisticated modeling, it is impossible to verify that models are predicting this fine-scale heterogeneity correctly or accounting for the relevant uncertainties correctly. Again, the solution to this issue is that correlated errors must be accounted for in calculations of within-site heterogeneity, or modeled points must not be closer than the minimum validated distance. or other input data from spatial databases with coarse resolutions). If modeling protocols do not account for this sort of fine-scale spatial separation or this sort of more sophisticated modeling, it is impossible to verify that models are predicting this fine-scale heterogeneity correctly or accounting for the relevant uncertainties correctly.

Again, the solution to this issue is that correlated errors must be accounted for in calculations of within-site heterogeneity, or modeled points must not be closer than the minimum validated distance.

Temporal relationships of errors

Given that GHG mitigation projects are modeled over time, temporal relationships in the data and model uncertainties are also relevant. Errors for a given point in space may display autocorrelation and/or heteroskedasticity (i.e., change in variance) over time.

Different models handle prediction variance over time differently or can be set to handle it in different ways. For example, in a random walk model with an independent process error, the variance of the predictions increases linearly with time (i.e., RMSE increases as the square root of time). If the model has internal stability, the increase is slower than linear (and in some cases will converge to a steady state), while for unstable/chaotic models the increase is faster than linear. The presence of autocorrelation in the model's process error will also affect the rate at which the predictive uncertainty grows, with positive autocorrelation increasing the rate and negative autocorrelation decreasing it. For example, a random walk model with a perfectly positive autocorrelation ($\rho=1$) in error would have a constant additive bias added each time step, causing the error to grow linearly (variance to increase quadratically), while with a perfect negative autocorrelation ($\rho=-1$) the error at each time point would change sign and cancel out the error from the previous time point.

It is important to consider the way that the relationship between model prediction error and time is handled during modeling and uncertainty assessment because the approach can cause bias in the uncertainty calculation. Consider a hypothetical example of a project with a time period of interest of five years, using a model validated against data covering a mix of time spans that bracket five years (e.g., 1-15 years, mean = five, median = three). Some current guidance suggests that because the reporting period is "shorter than the median length of experiments in the validation dataset, a single mixed-duration estimate of model error is a conservative estimate of model prediction error" (Climate Action Reserve Soil Enrichment Protocol, Verra VM0042). Due to Jensen's Inequality, the magnitude and direction of the bias in the calculation of model error for five-year predictions will depend on the curvature of the relationship between uncertainty and time (concave versus convex, Fig. 4). In most cases, the accumulation of model prediction error over time is concave (e.g., a random walk standard deviation grows as the square root of time) in which case the mean error < error at the mean time. In this example, if the relationship of RMSE versus time is concave and the mean time interval matches the period of interest (i.e., five years), then assuming a constant mean error will lead to an underestimated error (non-conservative). The use of the median (rather than mean) time introduces further complication for determining conservatism, because it requires knowing not only whether the error function is convex or concave, but also whether the sample of time points are left or right skewed. In this example the median time < mean time, so the error at the median time is lower than at the mean time, but it is not clear whether it is conservative or not. Furthermore, if the samples are skewed in the other direction (median time > mean time) then there is even higher potential for underestimation of the error. Determining what the mean or median sample period would need to be in order to be conservative depends on the exact curvature and how the sample points are distributed in time (variance, skew) and in ways that makes it difficult to make generalized predictions. Therefore, research is needed to determine whether generalizations regarding conservatism in handling model prediction error over time may be practical (e.g., if patterns in model prediction error through time are relatively consistent across applications).

References

[1] FAO, FAOSTAT, Land, Inputs and Sustainability: Land Use (2023). https://www.fao.org/faostat/en/#data/RL (accessed July 5, 2023).

[2] FAO, FAOSTAT, Climate Change: Agrifood Systems Emissions (2023). https://www.fao.org/faostat/en/#data (accessed June 12, 2023).

[3] B. Gu, X. Zhang, S.K. Lam, Y. Yu, H.J.M. van Grinsven, S. Zhang, X. Wang, B.L. Bodirsky, S. Wang, J. Duan, C. Ren, L. Bouwman, W. de Vries, J. Xu, M.A. Sutton, D. Chen, Cost-effective mitigation of nitrogen pollution from global croplands, Nature 613 (2023) 77–84. https://doi.org/10.1038/s41586-022-05481-8.

[4] H. Tian, R. Xu, J.G. Canadell, R.L. Thompson, W. Winiwarter, P. Suntharalingam, E.A. Davidson, P. Ciais, R.B. Jackson, G. Janssens-Maenhout, M.J. Prather, P. Regnier, N. Pan, S. Pan, G.P. Peters, H. Shi, F.N. Tubiello, S. Zaehle, F. Zhou, A. Arneth, G. Battaglia, S. Berthet, L. Bopp, A.F. Bouwman, E.T. Buitenhuis, J. Chang, M.P. Chipperfield, S.R.S. Dangal, E. Dlugokencky, J.W. Elkins, B.D. Eyre, B. Fu, B. Hall, A. Ito, F. Joos, P.B. Krummel, A. Landolfi, G.G. Laruelle, R. Lauerwald, W. Li, S. Lienert, T. Maavara, M. MacLeod, D.B. Millet, S. Olin, P.K. Patra, R.G. Prinn, P.A. Raymond, D.J. Ruiz, G.R. van der Werf, N. Vuichard, J. Wang, R.F. Weiss, K.C. Wells, C. Wilson, J. Yang, Y. Yao, A comprehensive quantification of global nitrous oxide sources and sinks, Nature 586 (2020) 248–256. https://doi.org/10.1038/s41586-020-2780-0.

[5] J. Sanderman, T. Hengl, G.J. Fiske, Soil carbon debt of 12,000 years of human land use, PNAS 114 (2017) 9575–9580. https://doi.org/10.1073/pnas.1706103114.

[6] D.E.H.J. Gernaat, K. Calvin, P.L. Lucas, G. Luderer, S.A.C. Otto, S. Rao, J. Strefler, D.P. van Vuuren, Understanding the contribution of non-carbon dioxide gases in deep mitigation scenarios, Global Environmental Change 33 (2015) 142–153. https://doi.org/10.1016/j.gloenvcha.2015.04.010.

[7] R.T. Conant, C.E.P. Cerri, B.B. Osborne, K. Paustian, Grassland management impacts on soil carbon stocks: a new synthesis, Ecological Applications 27 (2017) 662–668. https://doi.org/10.1002/eap.1473.

[8] M.-F. Dignac, D. Derrien, P. Barré, S. Barot, L. Cécillon, C. Chenu, T. Chevallier, G.T. Freschet, P. Garnier, B. Guenet, M. Hedde, K. Klumpp, G. Lashermes, P.-A. Maron, N. Nunan, C. Roumet, I. Basile-Doelsch, Increasing soil carbon storage: mechanisms, effects of agricultural practices and proxies. A review, Agron. Sustain. Dev. 37 (2017) 14. https://doi.org/10.1007/s13593-017-0421-2.

[9] M. Lessmann, G.H. Ros, M.D. Young, W. Vries, Global variation in soil carbon sequestration potential through improved cropland management, Global Change Biology 28 (2022) 1162–1177. https://doi.org/10.1111/gcb.15954.

[10] S.C. McClelland, K. Paustian, M.E. Schipanski, Management of cover crops in temperate climates influences soil organic carbon stocks: a meta-analysis, Ecol Appl 31 (2021) e02278. https://doi.org/10.1002/eap.2278.

[11] A.M. Prairie, A.E. King, M.F. Cotrufo, Restoring particulate and mineral-associated organic carbon through regenerative agriculture, Proceedings of the National Academy of Sciences 120 (2023) e2217481120. https://doi.org/10.1073/pnas.2217481120.

[12] J. Sanderman, R. Farquharson, J. Baldock, Soil carbon sequestration potential: A review for Australian agriculture, (2010). https://doi.org/10.4225/08/58518c66c3ab1.

[13] U. Stockmann, M.A. Adams, J.W. Crawford, D.J. Field, N. Henakaarchchi, M. Jenkins, B. Minasny, A.B. McBratney, V. de R. de Courcelles, K. Singh, I. Wheeler, L. Abbott, D.A. Angers, J. Baldock, M. Bird, P.C. Brookes, C. Chenu, J.D. Jastrow, R. Lal, J. Lehmann, A.G. O'Donnell, W.J. Parton, D. Whitehead, M. Zimmermann, The knowns, known unknowns and unknowns of sequestration of soil organic carbon, Agriculture, Ecosystems & Environment 164 (2013) 80–99. https://doi.org/10.1016/j.agee.2012.10.001.

[14] I. Virto, P. Barré, A. Burlot, C. Chenu, Carbon input differences as the main factor explaining the variability in soil organic C storage in no-tilled compared to inversion tilled agrosystems, Biogeochemistry 108 (2012) 17–26. https://doi.org/10.1007/s10533-011-9600-4.

[15] D.A. Kane, M.A. Bradford, E. Fuller, E.E. Oldfield, S.A. Wood, Soil organic matter protects US maize yields and lowers crop insurance payouts under drought, Environmental Research Letters 16 (2021). https://doi.org/10.1088/1748-9326/abe492.

[16] E.E. Oldfield, M.A. Bradford, A.J. Augarten, E.T. Cooley, A.M. Radatz, T. Radatz, M.D. Ruark, Positive associations of soil organic matter and crop yields across a regional network of working farms, Soil Science Society of America Journal 86 (2022) 384–397. https://doi.org/10.1002/saj2.20349.

[17] E.E. Oldfield, M.A. Bradford, S.A. Wood, Global meta-analysis of the relationship between soil organic matter and crop yields, SOIL 5 (2019) 15–32. https://doi.org/10.5194/soil-5-15-2019.

[18] S.A. Wood, M. Bowman, Large-scale farmer-led experiment demonstrates positive impact of cover crops on multiple soil health indicators, Nat Food 2 (2021) 97–103. https://doi.org/10.1038/s43016-021-00222-y.

[19] P. Smith, J. Adams, D.J. Beerling, T. Beringer, K.V. Calvin, S. Fuss, B. Griscom, N. Hagemann, C. Kammann, F. Kraxner, J.C. Minx, A. Popp, P. Renforth, J.L. Vicente Vicente, S. Keesstra, Land-Management Options for Greenhouse Gas Removal and Their Impacts on Ecosystem Services and the Sustainable Development Goals, Annual Review of Environment and Resources 44 (2019) 255–286. https://doi.org/10.1146/annurev-environ-101718-033129.

[20] D. Manning, F. Cotrufo, L. van der Pol, M. Machmuller, To make agriculture more climate-friendly, carbon farming needs clear rules, The Conversation (2021). http://theconversation.com/to-make-agriculture-more-climate-friendly-carbon-farming-needs-clear-rules-160243 (accessed December 10, 2023).

[21] E.E. Oldfield, J.M. Lavallee, E. Kyker-Snowman, J. Sanderman, The need for knowledge transfer and communication among stakeholders in the voluntary carbon market, Biogeochemistry (2022). https://doi.org/10.1007/s10533-022-00950-8.

[22] E.E. Oldfield, A.J. Eagle, R.L. Rubin, J. Rudek, J. Sanderman, D.R. Gordon, Crediting agricultural soil carbon sequestration, Science 375 (2022) 1222–1225. https://doi.org/10.1126/science.abl7991.

[23] A. Simmons, A. Cowie, B. Wilson, M. Farrell, M.T. Harrison, P. Grace, R. Eckard, V. Wong, W. Badgery, US scheme used by Australian farmers reveals the dangers of trading soil carbon to tackle climate change, The Conversation (2021). http://theconversation.com/us-scheme-used-by-australian-farmers-reveals-the-dangers-of-trading-soil-carbon-to-tackle-climate-change-161358 (accessed December 10, 2023).

[24] K. Guan, Z. Jin, B. Peng, J. Tang, E.H. DeLucia, P.C. West, C. Jiang, S. Wang, T. Kim, W. Zhou, T. Griffis, L. Liu, W.H. Yang, Z. Qin, Q. Yang, A. Margenot, E.R. Stuchiner, V. Kumar, C. Bernacchi, J. Coppess, K.A. Novick, J. Gerber, M. Jahn, M. Khanna, D. Lee, Z. Chen, S.-J. Yang, A scalable framework for quantifying field-level agricultural carbon outcomes, Earth-Science Reviews 243 (2023) 104462. https://doi.org/10.1016/j.earscirev.2023.104462.

[25] M.A. Bradford, L. Eash, A. Polussa, F.V. Jevon, S.E. Kuebbing, W.A. Hammac, S. Rosenzweig, E.E. Oldfield, Testing the feasibility of quantifying change in agricultural soil carbon stocks through empirical sampling, Geoderma 440 (2023) 116719. https://doi.org/10.1016/j.geoderma.2023.116719.

[26] M.J. Davoudabadi, D. Pagendam, C. Drovandi, J. Baldock, G. White, Modelling and predicting soil carbon sequestration: is current model structure fit for purpose?, (2022). http://arxiv.org/abs/2105.04789 (accessed September 6, 2022).

[27] A. Garsia, A. Moinet, C. Vazquez, R.E. Creamer, G.Y.K. Moinet, The challenge of selecting an appropriate soil organic carbon simulation model: A comprehensive global review and validation assessment, Global Change Biology (2023) 1–15. https://doi.org/10.1111/gcb.16896.

[28] J. Le Noë, S. Manzoni, R. Abramoff, T. Bölscher, E. Bruni, R. Cardinael, P. Ciais, C. Chenu, H. Clivot, D. Derrien, F. Ferchaud, P. Garnier, D. Goll, G. Lashermes, M. Martin, D. Rasse, F. Rees, J. Sainte-Marie, E. Salmon, M. Schiedung, J. Schimel, W. Wieder, S. Abiven, P. Barré, L. Cécillon, B. Guenet, Soil organic carbon models need independent time-series validation for reliable prediction, Commun Earth Environ 4 (2023) 1–8. https://doi.org/10.1038/s43247-023-00830-5.

[29] E. Bruni, C. Chenu, R.Z. Abramoff, G. Baldoni, D. Barkusky, H. Clivot, Y. Huang, T. Kätterer, D. Pikuła, H. Spiegel, I. Virto, B. Guenet, Multi-modelling predictions show high uncertainty of required carbon input changes to reach a 4‰ target, European Journal of Soil Science 73 (2022) e13330. https://doi.org/10.1111/ejss.13330.

[30] R. Farina, R. Sándor, M. Abdalla, J. Álvaro-Fuentes, L. Bechini, M.A. Bolinder, L. Brilli, C. Chenu, H. Clivot, M. De Antoni Migliorati, C. Di Bene, C.D. Dorich, F. Ehrhardt, F. Ferchaud, N. Fitton, R. Francaviglia, U. Franko, D.L. Giltrap, B.B. Grant, B. Guenet, M.T. Harrison, M.U.F. Kirschbaum, K. Kuka, L. Kulmala, J. Liski, M.J. McGrath, E. Meier, L. Menichetti, F. Moyano, C. Nendel, S. Recous, N. Reibold, A. Shepherd, W.N. Smith, P. Smith, J.-F. Soussana, T. Stella, A. Taghizadeh-Toosi, E. Tsutskikh, G. Bellocchi, Ensemble modelling, uncertainty and robust predictions of organic carbon in long-term bare-fallow soils, Global Change Biology 27 (2021) 904–928. https://doi.org/10.1111/gcb.15441.

[31] R. Sándor, F. Ehrhardt, P. Grace, S. Recous, P. Smith, V. Snow, J.-F. Soussana, B. Basso, A. Bhatia, L. Brilli, J. Doltra, C.D. Dorich, L. Doro, N. Fitton, B. Grant, M.T. Harrison, M.U.F. Kirschbaum, K. Klumpp, P. Laville, J. Léonard, R. Martin, R.-S. Massad, A. Moore, V. Myrgiotis, E. Pattey, S. Rolinski, J. Sharp, U. Skiba, W. Smith, L. Wu, Q. Zhang, G. Bellocchi, Ensemble modelling of carbon fluxes in grasslands and croplands, Field Crops Research 252 (2020) 107791. https://doi.org/10.1016/j.fcr.2020.107791.

[32] C. Tonitto, P.B. Woodbury, E.L. McLellan, Defining a best practice methodology for modeling the environmental performance of agriculture, Environmental Science & Policy 87 (2018) 64–73. https://doi.org/10.1016/j.envsci.2018.04.009.

[33] R. Parkhurst, L. Moore, R. Wright, M. Perez, Agricultural carbon programs: from program chaos to systems change, American Farmland Trust, Washington, DC, 2023. https://farmlandinfo.org/wp-content/uploads/sites/2/2023/08/AFT-SVS-Agricultural-Carbon-Programs.pdf (accessed December 12, 2023).

[34] Climate Action Reserve, Requirements and Guidance for Model Calibration, Validation, Uncertainty, and Verification For Soil Enrichment Projects, (2022). https://www.climateactionreserve.org/wp-content/uploads/2022/04/SEP_Model_Cal_Val_Guidance_4.2022.pdf (accessed January 20, 2024).

[35] Verra, VMD0053 Model Calibration, Validation, and Uncertainty Guidance for the Methodology for Improved Agricultural Land Management, v2.0, (2023). https://verra.org/methodologies/vmd0053-model-calibration-validation-and-uncertainty-guidance-for-the-methodology-for-improved-agricultural-land-management-v2-0/ (accessed January 20, 2024).

[36] M. Speich, C.F. Dormann, F. Hartig, Sequential Monte-Carlo algorithms for Bayesian model calibration – A review and method comparison , Ecological Modelling 455 (2021) 109608. https://doi.org/10.1016/j.ecolmodel.2021.109608.

[37] F. Albanito, D. McBey, M. Harrison, P. Smith, F. Ehrhardt, A. Bhatia, G. Bellocchi, L. Brilli, M. Carozzi, K. Christie, J. Doltra, C. Dorich, L. Doro, P. Grace, B. Grant, J. Léonard, M. Liebig, C. Ludemann, R. Martin, E. Meier, R. Meyer, M. De Antoni Migliorati, V. Myrgiotis, S. Recous, R. Sándor, V. Snow, J.-F. Soussana, W.N. Smith, N. Fitton, How Modelers Model: the Overlooked Social and Human Dimensions in Model Intercomparison Studies, Environ. Sci. Technol. (2022) acs.est.2c02023. https://doi.org/10.1021/acs.est.2c02023.

[38] D. Wallach, D. Makowski, J.W. Jones, F. Brun, Chapter 7 - Calibration of System Models, in: D. Wallach, D. Makowski, J.W. Jones, F. Brun (Eds.), Working with Dynamic Crop Models (Third Edition), Academic Press, 2019: pp. 251–274. https://doi.org/10.1016/B978-0-12-811756-9.00007-1.

[39] D. Wallach, T. Palosuo, P. Thorburn, Z. Hochman, E. Gourdain, F. Andrianasolo, S. Asseng, B. Basso, S. Buis, N. Crout, C. Dibari, B. Dumont, R. Ferrise, T. Gaiser, C. Garcia, S. Gayler, A. Ghahramani, S. Hiremath, S. Hoek, H. Horan, G. Hoogenboom, M. Huang, M. Jabloun, P.-E. Jansson, Q. Jing, E. Justes, K.C. Kersebaum, A. Klosterhalfen, M. Launay, E. Lewan, Q. Luo, B. Maestrini, H. Mielenz, M. Moriondo, H. Nariman Zadeh, G. Padovan, J.E. Olesen, A. Poyda, E. Priesack, J.W.M. Pullens, B. Qian, N. Schütze, V. Shelia, A. Souissi, X. Specka, A.K. Srivastava, T. Stella, T. Streck, G. Trombi, E. Wallor, J. Wang, T.K.D. Weber, L. Weihermüller, A. de Wit, T. Wöhling, L. Xiao, C. Zhao, Y. Zhu, S.J. Seidel, The chaos in calibrating crop models: Lessons learned from a multi-model calibration exercise, Environmental Modelling & Software 145 (2021) 105206. https://doi.org/10.1016/j.envsoft.2021.105206.

[40] K. Fuchs, L. Merbold, N. Buchmann, D. Bretscher, L. Brilli, N. Fitton, C.F.E. Topp, K. Klumpp, M. Lieffering, R. Martin, P.C.D. Newton, R.M. Rees, S. Rolinski, P. Smith, V. Snow, Multimodel Evaluation of Nitrous Oxide Emissions From an Intensively Managed Grassland, Journal of Geophysical Research: Biogeosciences 125 (2020) e2019JG005261. https://doi.org/10.1029/2019JG005261.

[41] R.K. Gaillard, C.D. Jones, P. Ingraham, S. Collier, R.C. Izaurralde, W. Jokela, W. Osterholz, W. Salas, P. Vadas, M.D. Ruark, Underestimation of N2O emissions in a comparison of the DayCent, DNDC, and EPIC models, Ecological Applications 28 (2018) 694–708. https://doi.org/10.1002/eap.1674.

[42] J.S. Clark, Why environmental scientists are becoming Bayesians, Ecology Letters 8 (2005) 2–14. https://doi.org/10.1111/j.1461-0248.2004.00702.x.

[43] M.J. Davoudabadi, D. Pagendam, C. Drovandi, J. Baldock, G. White, Advanced Bayesian approaches for state-space models with a case study on soil carbon sequestration, Environmental Modelling & Software 136 (2021) 104919. https://doi.org/10.1016/j.envsoft.2020.104919.

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

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

[46] I. Fer, A. Shiklomanov, K.A. Novick, C.M. Gough, M.A. Arain, J. Chen, B. Murphy, A.R. Desai, M.C. Dietze, Capturing site-to-site variability through Hierarchical Bayesian calibration of a process-based dynamic vegetation model, (2021) 2021.04.28.441243. https://doi.org/10.1101/2021.04.28.441243.

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

[48] K. Paustian, S. Collier, J. Baldock, R. Burgess, J. Creque, M. DeLonge, J. Dungait, B. Ellert, S. Frank, T. Goddard, B. Govaerts, M. Grundy, M. Henning, R.C. Izaurralde, M. Madaras, B. McConkey, E. Porzig, C. Rice, R. Searle, N. Seavy, R. Skalsky, W. Mulhern, M. Jahn, Quantifying carbon for agricultural soil management: from the current status toward a global soil information system, Carbon Management 10 (2019) 567–587. https://doi.org/10.1080/17583004.2019.1633231.

[49] Y. Yang, S. Huang, Suitability of five cross validation methods for performance evaluation of nonlinear mixed-effects forest models – a case study, Forestry: An International Journal of Forest Research 87 (2014) 654–662. https://doi.org/10.1093/forestry/cpu025.

[50] D. Cameron, F. Hartig, F. Minnuno, J. Oberpriller, B. Reineking, M. Van Oijen, M. Dietze, Issues in calibrating models with multiple unbalanced constraints: the significance of systematic model and data errors, Methods in Ecology and Evolution 13 (2022) 2757–2770. https://doi.org/10.1111/2041-210X.14002.

[51] D.L. Liu, K.Y. Chan, M.K. Conyers, Simulation of soil organic carbon under different tillage and stubble management practices using the Rothamsted carbon model, Soil and Tillage Research 104 (2009) 65–73. https://doi.org/10.1016/j.still.2008.12.011.

[52] M.C. Dietze, Prediction in ecology: a first-principles framework, Ecological Applications 27 (2017) 2048–2060. https://doi.org/10.1002/eap.1589.

[53] R. Kennedy, S. Serbin, CMS Uncertainty Working Group, Characterizing and communicating uncertainty: lessons from NASA's Carbon Monitoring System, Environmental Research Letters (in review).

[54] T.L. Anthony, W.L. Silver, Hot spots and hot moments of greenhouse gas emissions in agricultural peatlands, Biogeochemistry (2023). https://doi.org/10.1007/s10533-023-01095-y.

[55] E.D. Roy, C.R.H. Wagner, M.T. Niles, Hot spots of opportunity for improved cropland nitrogen management across the United States, Environ. Res. Lett. 16 (2021) 035004. https://doi.org/10.1088/1748-9326/abd662.

[56] S. Waldo, E.S. Russell, K. Kostyanovsky, S.N. Pressley, P.T. O'Keeffe, D.R. Huggins, C.O. Stöckle, W.L. Pan, B.K. Lamb, N20 Emissions From Two Agroecosystems: High Spatial Variability and Long Pulses Observed Using Static Chambers and the Flux-Gradient Technique, Journal of Geophysical Research: Biogeosciences 124 (2019) 1887–1904. https://doi.org/10.1029/2019JG005032.

[57] Indigo Ag, Validation Report DayCent-CR Version 1.0.2, 2022. https://www.climateactionreserve.org/wp-content/uploads/2022/11/CAR1459_model_val_DayCentCR_1.0.2.pdf.

[58] A. Chappell, J. Baldock, R.V. Rossel, Sampling soil organic carbon to detect change over time, CSIRO, 2013.

[59] R.B. Gramacy, Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences, Chapman & Hall, 2021. https://www.routledge.com/Surrogates-Gaussian-Process-Modeling-Design-and-Optimization-for-the/Gramacy/p/book/9781032242552 (accessed February 19, 2024).

[60] Nemo, K. Klumpp, K. Coleman, M. Dondini, K. Goulding, A. Hastings, Michael.B. Jones, J. Leifeld, B. Osborne, M. Saunders, T. Scott, Y.A. Teh, P. Smith, Soil Organic Carbon (SOC) Equilibrium and Model Initialisation Methods: an Application to the Rothamsted Carbon (RothC) Model, Environ Model Assess 22 (2017) 215–229. https://doi.org/10.1007/s10666-016-9536-0.

[61] R. Farina, R. Sándor, M. Abdalla, J. Álvaro-Fuentes, L. Bechini, M.A. Bolinder, L. Brilli, C. Chenu, H. Clivot, M. De Antoni Migliorati, C. Di Bene, C.D. Dorich, F. Ehrhardt, F. Ferchaud, N. Fitton, R. Francaviglia, U. Franko, D.L. Giltrap, B.B. Grant, B. Guenet, M.T. Harrison, M.U.F. Kirschbaum, K. Kuka, L. Kulmala, J. Liski, M.J. McGrath, E. Meier, L. Menichetti, F. Moyano, C. Nendel, S. Recous, N. Reibold, A. Shepherd, W.N. Smith, P. Smith, J. Soussana, T. Stella, A. Taghizadeh-Toosi, E. Tsutskikh, G. Bellocchi, Ensemble modelling, uncertainty and robust predictions of organic carbon in long-term bare-fallow soils, Glob. Change Biol. 27 (2021) 904–928. https://doi.org/10.1111/gcb.15441.

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

[63] P.D. Falloon, P. Smith, Modelling refractory soil organic matter, Biol Fertil Soils 30 (2000) 388–398. https://doi.org/10.1007/s003740050019.

[64] S. Hashimoto, M. Wattenbach, P. Smith, A new scheme for initializing process-based ecosystem models by scaling soil carbon pools, Ecological Modelling 222 (2011) 3598–3602. https://doi.org/10.1016/j.ecolmodel.2011.08.011.

[65] J.B. Yeluripati, M. Van Oijen, M. Wattenbach, A. Neftel, A. Ammann, W.J. Parton, P. Smith, Bayesian calibration as a tool for initialising the carbon pools of dynamic soil models, Soil Biology and Biochemistry 41 (2009) 2579–2583. https://doi.org/10.1016/j.soilbio.2009.08.021.

[66] S.M. Ogle, F.J. Breidt, M. Easter, S. Williams, K. Killian, K. Paustian, Scale and uncertainty in modeled soil organic carbon stock changes for US croplands using a process-based model, Global Change Biology 16 (2010) 810–822. https://doi.org/10.1111/j.1365-2486.2009.01951.x.

[67] A. Raiho, M. Dietze, A. Dawson, C.R. Rollinson, J. Tipton, J. McLachlan, Towards understanding predictability in ecology: A forest gap model case study, (2020) 2020.05.05.079871. https://doi.org/10.1101/2020.05.05.079871.

[68] F. Bilotto, M.T. Harrison, M.D.A. Migliorati, K.M. Christie, D.W. Rowlings, P.R. Grace, A.P. Smith, R.P. Rawnsley, P.J. Thorburn, R.J. Eckard, Can seasonal soil N mineralisation trends be leveraged to enhance pasture growth?, Science of The Total Environment 772 (2021) 145031. https://doi.org/10.1016/j.scitotenv.2021.145031.

[69] B. Henry, R. Dalal, M.T. Harrison, University of Tasmania, Australia, B. Keating, The University of Queensland, Australia, Creating frameworks to foster soil carbon sequestration, in: CNRS, France, C. Rumpel (Eds.), Burleigh Dodds Series in Agricultural Science, Burleigh Dodds Science Publishing, 2022. https://doi.org/10.19103/AS.2022.0106.25.

[70] M.T. Harrison, B.R. Cullen, D.E. Mayberry, A.L. Cowie, F. Bilotto, W.B. Badgery, K. Liu, T. Davison, K.M. Christie, A. Muleke, R.J. Eckard, Carbon myopia: The urgent need for integrated social, economic and environmental action in the livestock sector, Global Change Biology 27 (2021) 5726–5761. https://doi.org/10.1111/gcb.15816.

[71] M. Herbst, G. Welp, A. Macdonald, M. Jate, A. Hädicke, H. Scherer, T. Gaiser, F. Herrmann, W. Amelung, J. Vanderborght, Correspondence of measured soil carbon fractions and RothC pools for equilibrium and non-equilibrium states, Geoderma 314 (2018) 37–46. https://doi.org/10.1016/j.geoderma.2017.10.047.

[72] T. Wutzler, M. Reichstein, Soils apart from equilibrium - consequences for soil carbon balance modelling, Biogeosciences 4 (2007) 125–136. https://doi.org/10.5194/bg-4-125-2007.

[73] M.J. Hill, Generating generic response signals for scenario calculation of management effects on carbon sequestration in agriculture: approximation of main effects using CENTURY, Environmental Modelling & Software 18 (2003) 899–913. https://doi.org/10.1016/S1364-8152(03)00054-9.

[74] M.T. Harrison, P.P. Roggero, L. Zavattaro, Simple, efficient and robust techniques for automatic multi-objective function parameterisation: Case studies of local and global optimisation using APSIM, Environmental Modelling & Software 117 (2019) 109–133. https://doi.org/10.1016/j.envsoft.2019.03.010.

[75] N. Senapati, P. Smith, B. Wilson, J.B. Yeluripati, H. Daniel, P. Lockwood, S. Ghosh, Projections of changes in grassland soil organic carbon under climate change are relatively insensitive to methods of model initialization, European Journal of Soil Science 64 (2013) 229–238. https://doi.org/10.1111/ejss.12014.

[76] J.A. Baldock, J. Sanderman, L.M. Macdonald, A. Puccini, B. Hawke, S. Szarvas, J. McGowan, Quantifying the allocation of soil organic carbon to biologically significant fractions, Soil Res. 51 (2013) 561–576. https://doi.org/10.1071/SR12374.

[77] B.T. Christensen, Matching Measurable Soil Organic Matter Fractions with Conceptual Pools in Simulation Models of Carbon Turnover: Revision of Model Structure, in: D.S. Powlson, P. Smith, J.U. Smith (Eds.), Evaluation of Soil Organic Matter Models, Springer Berlin Heidelberg, Berlin, Heidelberg, 1996: pp. 143–159.

[78] E. Kanari, L. Cécillon, F. Baudin, H. Clivot, F. Ferchaud, S. Houot, F. Levavasseur, B. Mary, L. Soucémarianadin, C. Chenu, P. Barré, A robust initialization method for accurate soil organic carbon simulations, Biogeosciences 19 (2022) 375–387. https://doi.org/10.5194/bg-19-375-2022.

[79] Z. Luo, E. Wang, I.R.P. Fillery, L.M. Macdonald, N. Huth, J. Baldock, Modelling soil carbon and nitrogen dynamics using measurable and conceptual soil organic matter pools in APSIM, Agriculture, Ecosystems & Environment 186 (2014) 94–104. https://doi.org/10.1016/j.agee.2014.01.019.

[80] J.O. Skjemstad, L.R. Spouncer, B. Cowie, R.S. Swift, Calibration of the Rothamsted organic carbon turnover model (RothC ver. 26.3), using measurable soil organic carbon pools, Soil Res. 42 (2004) 79–88. https://doi.org/10.1071/sr03013.

[81] J.U. Smith, P. Smith, R. Monaghan, A.J. MacDonald, When is a measured soil organic matter fraction equivalent to a model pool?, European Journal of Soil Science 53 (2002) 405–416. https://doi.org/10.1046/j.1365-2389.2002.00458.x.

[82] S.P. Sohi, N. Mahieu, J.R.M. Arah, D.S. Powlson, B. Madari, J.L. Gaunt, A Procedure for Isolating Soil Organic Matter Fractions Suitable for Modeling, Soil Science Society of America Journal 65 (2001) 1121–1128. https://doi.org/10.2136/sssaj2001.6541121x.

[83] M. Zimmermann, J. Leifeld, M.W.I. Schmidt, P. Smith, J. Fuhrer, Measured soil organic matter fractions can be related to pools in the RothC model, European Journal of Soil Science 58 (2007) 658–667. https://doi.org/10.1111/j.1365-2389.2006.00855.x.

[84] C. Cagnarini, G. Renella, J. Mayer, J. Hirte, R. Schulin, B. Costerousse, A. Della Marta, S. Orlandini, L. Menichetti, Multi-objective calibration of RothC using measured carbon stocks and auxiliary data of a long-term experiment in Switzerland, European Journal of Soil Science 70 (2019) 819–832. https://doi.org/10.1111/ejss.12802.

[85] K.M. Coelli, S.B. Karunaratne, J.A. Baldock, S. Ugbaje, A.J.V. Buzacott, P. Filippi, S. Cattle, T.F.A. Bishop, A nationally scalable approach to simulating soil organic carbon in agricultural landscapes, in: MODSIM2021, 24th International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand, 2021. https://doi.org/10.36334/modsim.2021.B6.coelli.

[86] J. Leifeld, M. Zimmermann, J. Fuhrer, F. Conen, Storage and turnover of carbon in grassland soils along an elevation gradient in the Swiss Alps, Global Change Biology 15 (2009) 668–679. https://doi.org/10.1111/j.1365-2486.2008.01782.x.

[87] A.D. Robertson, K. Paustian, S. Ogle, M.D. Wallenstein, E. Lugato, M.F. Cotrufo, Unifying soil organic matter formation and persistence frameworks: the MEMS model, 2018 (2018) 1–36. https://doi.org/10.5194/bg-2018-430.

[88] Y. Zhang, J.M. Lavallee, A.D. Robertson, R. Even, S.M. Ogle, K. Paustian, M.F. Cotrufo, Simulating measurable ecosystem carbon and nitrogen dynamics with the mechanistically defined MEMS 2.0 model, Biogeosciences 18 (2021) 3147–3171. https://doi.org/10.5194/bg-18-3147-2021.

[89] S.R.S. Dangal, C. Schwalm, M.A. Cavigelli, H.T. Gollany, V.L. Jin, J. Sanderman, Improving Soil Carbon Estimates by Linking Conceptual Pools Against Measurable Carbon Fractions in the DAYCENT Model Version 4.5, Journal of Advances in Modeling Earth Systems 14 (2022) e2021MS002622. https://doi.org/10.1029/2021MS002622.

[90] S.B. Karunaratne, T.F.A. Bishop, J.S. Lessels, J.A. Baldock, I.O.A. Odeh, A space–time observation system for soil organic carbon, Soil Res. 53 (2015) 647–661. https://doi.org/10.1071/SR14178.

[91] H. Clivot, J.-C. Mouny, A. Duparque, J.-L. Dinh, P. Denoroy, S. Houot, F. Vertès, R. Trochard, A. Bouthier, S. Sagot, B. Mary, Modeling soil organic carbon evolution in long-term arable experiments with AMG model, Environmental Modelling & Software 118 (2019) 99–113. https://doi.org/10.1016/j.envsoft.2019.04.004.

[92] Qianyu Li, Dongchen Zhang, Alexis Helgeson, Michael Dietze, Shawn P. Serbin, Soil carbon assimilation effectively constrains carbon-cycle model forecasting, (in review).

[93] T. Viskari, M. Laine, L. Kulmala, J. Mäkelä, I. Fer, J. Liski, Improving Yasso15 soil carbon model estimates with ensemble adjustment Kalman filter state data assimilation, Geoscientific Model Development 13 (2020) 5959–5971. https://doi.org/10.5194/gmd-13-5959-2020.

[94] W. Zhou, K. Guan, B. Peng, A. Margenot, D. Lee, J. Tang, Z. Jin, R. Grant, E. DeLucia, Z. Qin, M.M. Wander, S. Wang, How does uncertainty of soil organic carbon stock affect the calculation of carbon budgets and soil carbon credits for croplands in the U.S. Midwest?, Geoderma 429 (2023) 116254. https://doi.org/10.1016/j.geoderma.2022.116254.

[95] R. Sándor, F. Ehrhardt, P. Grace, S. Recous, P. Smith, V. Snow, J.-F. Soussana, B. Basso, A. Bhatia, L. Brilli, J. Doltra, C.D. Dorich, L. Doro, N. Fitton, B. Grant, M.T. Harrison, M.U.F. Kirschbaum, K. Klumpp, P. Laville, J. Léonard, R. Martin, R.-S. Massad, A. Moore, V. Myrgiotis, E. Pattey, S. Rolinski, J. Sharp, U. Skiba, W. Smith, L. Wu, Q. Zhang, G. Bellocchi, Ensemble modelling of carbon fluxes in grasslands and croplands, Field Crops Research 252 (2020) 107791. https://doi.org/10.1016/j.fcr.2020.107791.

[96] L. Poggio, L.M. de Sousa, N.H. Batjes, G.B.M. Heuvelink, B. Kempen, E. Ribeiro, D. Rossiter, SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, SOIL 7 (2021) 217–240. https://doi.org/10.5194/soil-7-217-2021.

[97] K. Vaysse, P. Lagacherie, Using quantile regression forest to estimate uncertainty of digital soil mapping products, Geoderma 291 (2017) 55–64. https://doi.org/10.1016/j.geoderma.2016.12.017.

[98] W. He, B.B. Grant, Q. Jing, R. Lemke, M. St. Luce, R. Jiang, B. Qian, C.A. Campbell, A. VanderZaag, G. Zou, W.N. Smith, Measuring and modeling soil carbon sequestration under diverse cropping systems in the semiarid prairies of western Canada, Journal of Cleaner Production 328 (2021) 129614. https://doi.org/10.1016/j.jclepro.2021.129614.

[99] M.A. Oliver, R. Webster, Basic Steps in Geostatistics: The Variogram and Kriging, Springer International Publishing, Cham, 2015. https://doi.org/10.1007/978-3-319-15865-5.

[100] X. Xiong, S. Grunwald, R. Corstanje, C. Yu, N. Bliznyuk, Scale-dependent variability of soil organic carbon coupled to land use and land cover, Soil and Tillage Research 160 (2016) 101–109. https://doi.org/10.1016/j.still.2016.03.001.

[101] R.M. Lark, B.G. Rawlins, D.A. Robinson, I. Lebron, A.M. Tye, Implications of short-range spatial variation of soil bulk density for adequate field-sampling protocols: methodology and results from two contrasting soils, European Journal of Soil Science 65 (2014) 803–814. https://doi.org/10.1111/ejss.12178.

[102] D.W. Bergstrom, C.M. Monreal, E. St. Jacques, Spatial dependence of soil organic carbon mass and its relationship to soil series and topography, Can. J. Soil. Sci. 81 (2001) 53–62. https://doi.org/10.4141/S00-016.

[103] A. Don, J. Schumacher, M. Scherer-Lorenzen, T. Scholten, E.-D. Schulze, Spatial and vertical variation of soil carbon at two grassland sites — Implications for measuring soil carbon stocks, Geoderma 141 (2007) 272–282. https://doi.org/10.1016/j.geoderma.2007.06.003.

[104] K. Singh, B. Whelan, Soil carbon change across ten New South Wales farms under different farm management regimes in Australia, Soil Use and Management 36 (2020) 616–632. https://doi.org/10.1111/sum.12590.

[105] L.L. Lengnick, Spatial variability of early season nitrogen availability indicators in corn, Communications in Soil Science and Plant Analysis 28 (1997) 1721–1736. https://doi.org/10.1080/00103629709369912.

[106] T. Addise, B. Bedadi, A. Regassa, L. Wogi, S. Feyissa, Spatial Variability of Soil Organic Carbon Stock in Gurje Subwatershed, Hadiya Zone, Southern Ethiopia, Applied and Environmental Soil Science 2022 (2022) e5274482. https://doi.org/10.1155/2022/5274482

Environmental Defense Fund

Headquarters
257 Park Avenue South
New York, NY 10010
T 212 505 2100
F 212 505 2375

Austin, TX
301 Congress Avenue
Austin, TX 78701
T 512 478 5161
F 512 478 8140

Bentonville, AR
1116 South Walton Boulevard
Bentonville, AR 72712
T 479 845 3816
F 479 845 3815

Boston, MA
18 Tremont Street
Boston, MA 02108
T 617 723 2996
F 617 723 2999

Boulder, CO
2060 Broadway
Boulder, CO 80302
T 303 440 4901
F 303 440 8052

Raleigh, NC
4000 Westchase Boulevard
Raleigh, NC 27607
T 919 881 2601
F 919 881 2607

Sacramento, CA
1107 9th Street
Sacramento, CA 95814
T 916 492 7070
F 916 441 3142

San Francisco, CA
123 Mission Street
San Francisco, CA 94105
T 415 293 6050
F 415 293 6051

Washington, DC
1875 Connecticut Avenue, NW
Washington, DC 20009
T 202 387 3500
F 202 234 6049

Beijing, China
C-501, Yonghe Plaza
28 East Andingmen East Road
Dongcheng District
Beijing 100007, China
T +86 10 6409 7088
F +86 10 6409 7097

La Paz, Mexico
Revolución No. 345
E/5 de Mayo y Constitución
Col. Centro, CP 23000
La Paz, Baja California Sur, Mexico
T +52 612 123 2029

London, UK
3rd Floor, 41 Eastcheap,
London EC3M 1DT
T +44 203 310 5909