CLIMATE ACTION RESERVE

Requirements and Guidance for Model Calibration, Validation, Uncertainty, and Verification

For Soil Enrichment Projects

Version 1.1a

April 2022

Table of Contents

Definitions

Calibration

Any process involving the adjustment of parameters and constants within a model so that the model more accurately estimates measured values.

Validation

The process of evaluating model performance relative to measured values, with a validated model having demonstrated satisfactory performance in terms of goodness of fit and characterization of model prediction error. Model validation typically uses datasets independent of datasets used in model calibration, unless (e.g., in a data-limited situation) a statistical approach like k-folding is applied.

Goodness of fit

A characterization of the discrepancy between measured and modeled values.

Model prediction error

The uncertainty in a model's prediction as determined from comparison to measurements, required in this document to be the same measurements used to validate the model.

Parameter sets

The set of mathematical values and constants contained in a model that characterizes the biophysical and biogeochemical system being represented. Depending on the model and the application, the value of an individual parameter within a given parameter set might be defined as a single number, a set of numbers with declared rules for choosing from them, and/or a probability distribution.

Model-driving input data

Numeric data provided to the model needed to execute a model run, such as meteorological time series data, or rates of fertilizer application.

Model version

A uniquely traceable record of source code, internal parameters, and ancillary inputs that are needed to reproduce a given model output (collectively the "model files"). A model version must change any time there is a change in any of the model files.

Overfitting

Model parameterization that is too specific to the training data, yielding unrealistically high model performance under the exact conditions used for testing but typically much lower performance on new data. Detecting overfitting is one key goal of a robust validation method.

Pooled measurement uncertainty

An estimate of the typical uncertainty associated with experimental measurements of the emissions change resulting from a given practice change. It is computed from the observed variation between replicate measurements.

1 Introduction

The purpose of this document is to provide a standardized approach to test model performance as a component of credit quantification in a SEP project. This document must be used for any gas or pool for which models are employed for quantification outside of the protocol equations. A Validation Report generated by following this document is designed to support independent expert review of a model proposed for use in a SEP project, grouping model testing by combinations of geographic regions, crop types and practice changes where a model may be used to issue credits. The Validation Report is also designed to support project verification, demonstrating that a model is valid and used appropriately to quantify changes to greenhouse gas sources, sinks, and/or reservoirs for a specific project.

The requirements and guidance presented in this document fall into two main categories:

1. Standardizing best practices for SEP-appropriate use of peer-reviewed observed experimental data to test a model and determine model prediction error; and,
2. Standardizing demonstration of acceptable model fit and a lack of bias when a model is being used to estimate soil organic carbon (SOC) stock change and, if applicable to the project, flux change of N_2O and CH_4.

Requirements falling into category 1 are meant to address the importance of using high-quality observed experimental data of soil emissions reductions as the basis of evaluating model performance. Changes in agricultural practices have a great diversity of impacts on soil emissions. Soil emissions are also highly variable, with rapid growth of new studies and experimental methods to capture variance, increase precision, and reduce uncertainty. Further, the format of experimental data can be highly variable across published studies. Requirements described in this document are meant to ensure that appropriate and consistent methods are followed to locate, aggregate, and use observed data for model improvement and testing.

Requirements falling into category 2 are meant to provide SEP-specific guidance for using the above datasets for model calibration, validation, and the determination of a model's prediction error, in the context of measurement uncertainties. These are highly technical processes that vary widely across areas of scientific research. The Validation Report aims to ensure that model validation is specific to the model being proposed for use in a SEP project, is appropriate for the cropping system and biophysical conditions occurring in the project, and that validation requirements related to the assessment of model bias and fit have been met.

A discussion regarding verification requirements for the proper implementation of this guidance can be found in Section 5 of this document, as well as Section 8.3.1 of the Soil Enrichment Protocol (SEP).[1]

All stakeholders making use of this guidance should contact the Reserve to ensure they are using the most up-to-date version of this guidance. Project developers and verifiers must use the version of this document that is in place as of the first date of the relevant reporting period. In the cases where multiple reporting periods are being verified at once, project developers and verifiers should seek Reserve guidance and approval as to which version of this guidance should be applied to which reporting period.

[1] The current version of the SEP may be accessed at: http://www.climateactionreserve.org/how/protocols/soil-enrichment/.

Options and requirements for submitting Validation Reports are outlined in Section 3.6. Validation Reports must show that all requirements for a specific project have been met, including proof that the same model version and parameter sets are used, and that all project domain and crop functional group/practices category combinations have met minimum requirements for model validation. Validation Reports must either be independently assessed by an approved third-party entity or accepted for publication in one of the peer-reviewed publications listed in Section 3.6. Model Validation Reports will be public documents.

For each subsequent monitoring report, as long as a project area remains constant, or is only expanded to include new fields that already fit within the validated project domain, the existing Validation Report can be used. If the project is expanded to new practice effects, new crop functional groups, or the model is changed (using new model calibrations, changed parameters/parameter sets, or a different model, following guidance described in Section 6.5 of the SEP protocol), the Validation Report needs to be revised, reviewed by an approved third-party entity, and re-submitted.

The overall flow of the requirements for the use of models is illustrated in Figure 1.1., below.

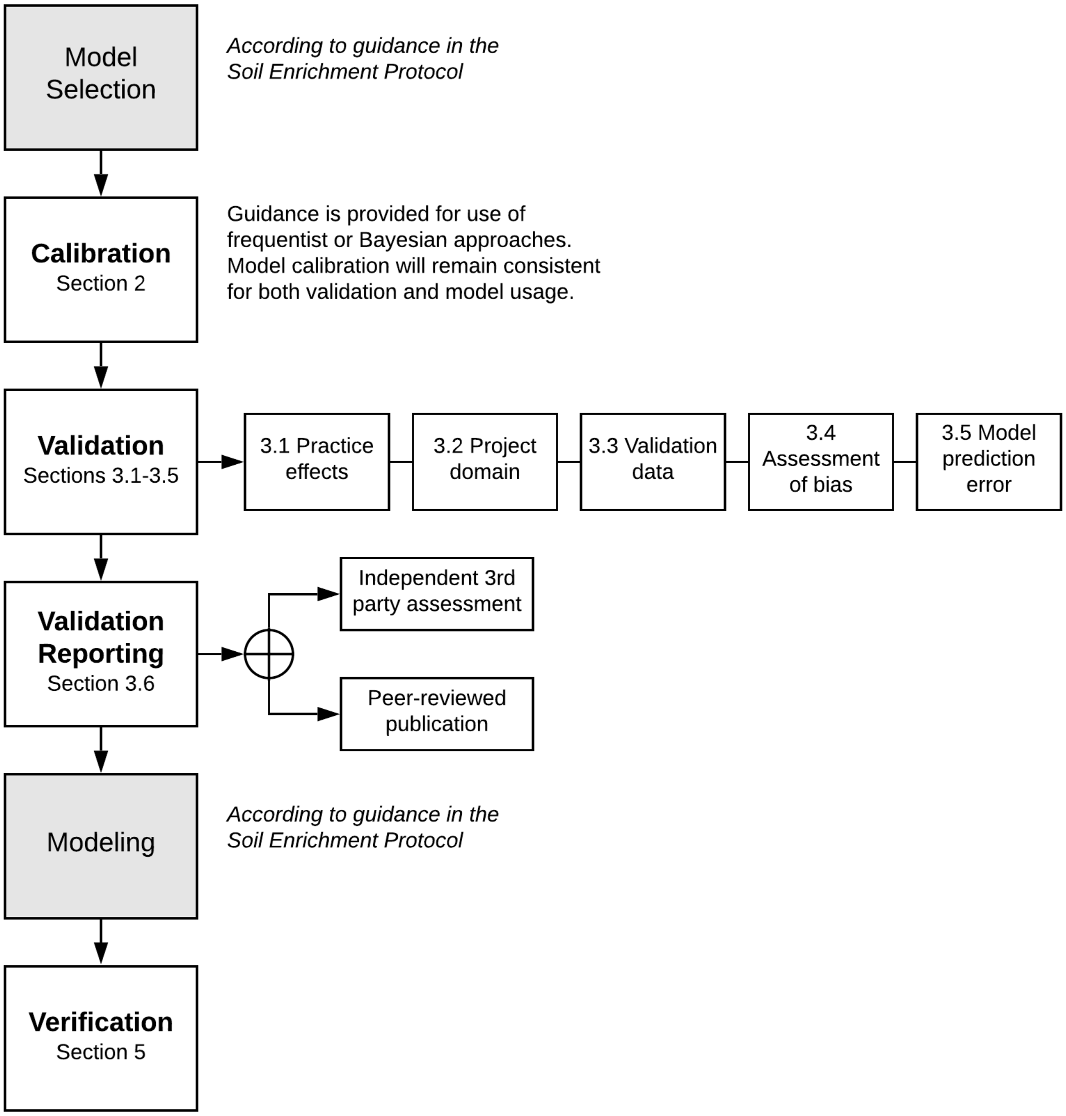

Model Selection
According to guidance in the Soil Enrichment Protocol
Calibration
Section 2
Guidance is provided for use of frequentist or Bayesian approaches. Model calibration will remain consistent for both validation and model usage.
Validation
Sections 3.1-3.5
3.1 Practice effects
3.2 Project domain
3.3 Validation data
3.4 Assessment of bias
3.5 Model prediction error
Validation Reporting
Section 3.6
Independent 3rd party assessment
Peer-reviewed publication
Modeling
According to guidance in the Soil Enrichment Protocol
Verification
Section 5

Figure 1.1. Steps Related to the Use of Models for Quantification in SEP Projects

2 Model Calibration

Model calibration, parameterization, and validation are often poorly defined concepts. For the purposes of this document, calibration is defined as any process involving the adjustment of parameters and constants within a model so that the model more accurately simulates measured values.

Model calibration is a variable and model-specific set of processes. Some examples include:

- Statistical procedures to optimize rates of mass flow and the simulation of internal model pools (e.g., optimizing the allocation of daily net primary production to root growth to more accurately simulate observed root growth for a given crop);

- Adjusting model parameters with directly measured values (e.g., setting the simulated fraction of plant residue left on the soil surface after a method of harvest using an average of observed values);
- 'Tuning' a set of model parameters that may not be able to be measured directly using overall model performance and an understanding of model sensitivities (e.g., adjusting a constant downregulating the rate of soil biological processes under moisture-limited conditions using measures of soil respiration).

Deterministic models, where the same inputs always result in the same outputs, may have different calibration processes than stochastic models, which include random variability. Mechanistic models, which are based on mathematical representations of mechanisms within the modeled system, are more generalizable with fewer data than empirical models, which are based on statistical synthesis of observations and cannot be extended outside of where observations are available.

For any model used in a SEP project, model calibration must be independent from model validation (i.e., using a separate process and separate datasets). Further, for either process, the quality of measured datasets (i.e., rigor of the experimental design, accuracy of observations, applicability to the system that a model is being calibrated or validated to simulate) will determine the quality of the model (aka, "garbage in, garbage out"). Separation of datasets for model calibration and validation may be done manually, if it can be shown that both datasets are comparable in their applicability to the system that is being simulated. Separation may also be done statistically using cross-validation methods (Section 2.3). Within independent datasets for model calibration and validation, it is acceptable to use statistical approaches like bootstrapping to make use of datasets that are more limited.

For the purposes of this protocol, calibration and validation data should be demonstrably independent. This requirement can be met if datasets used for calibration and validation do not overlap in experimental research locations and are not taken from the same experimental study. If calibration and validation datasets for SOC change or trace gas flux do overlap in either experimental study or research location, independence between the datasets used for calibration and validation should be demonstrated at the crop functional group/practice category combination level (Section 3.2). For example, if root measurements and N_2O flux measurements from a subset of treatments in a tilled soybean/corn rotation experiment were used for model calibration, the N_2O flux measurements from the remaining treatments in the same study could not be used as validation data for either the corn or the soy crop functional group and tillage practice effect combinations. However, if at the same research facility N_2O flux was measured in a demonstrably separate corn/soy rotation experiment (separate in space or time, with separate experimental design or intention), those data would be permissible for inclusion in model validation. Depending on the model, in some cases it may be defensible to use cultivar-specific measurements of crop growth to calibrate modeled crop growth, while using SOC change or trace gas flux change measurements from the same study to validate model performance. Such cases should be clearly explained and presented for review in the Validation Report.

The protocol does not prescribe a model calibration procedure. However, the calibration procedure must be reported to ensure model parameters and parameter sets were generated appropriately (see SEP protocol Section 6.5, item 3), as well as meet the requirements that:

1. The parameter sets used when validating the model are the same used when the model is applied to simulate baselines and project practices; and,
2. That the data used for model calibration and validation are separate.

‘Parameter sets’, in this context, mean all values internal to a model that determine how input data drive model performance and behavior, and that are changed using processes independent of model-driving input datasets. This means model parameters that are not dependent on input datasets when the model is run (for example through a Bayesian statistical procedure) must be declared and shown to be set appropriately following the above calibration requirements. It is encouraged for model parameters to be as generalizable as possible across the project domain, with minimal use of different parameter sets. However, it is acceptable for different parameter sets to be used as long as they are defined at scales no smaller than LRRs, (i.e., the same parameter set is used for all simulations within a given LRR). Parameter sets must be declared for each project LRR and should be used to simulate all crop functional groups and practice categories within that LRR. An exception may be made for crop growth parameters to be set at scales finer than LRRs, for example to reflect different maturity groups within an LRR. The use of varying crop growth parameters must be clearly defined in the Validation Report and presented as parameter sets specific to each LRR where the crop is simulated, to ensure their appropriate use in model validation and SEP project simulations.

Because verifying that the model behaves as validated requires knowing all values that drive model behavior are set as validated, the definition of ‘parameter set’ above means it is not acceptable to validate a model and then adjust declared model parameters when using the model to simulate project baselines and practices. If a model approach uses data to adjust internal model parameters dynamically while running simulations—such as through the use of data assimilation for components of the model—this should be clearly specified as part of the model’s structure and required inputs, such that it is clear how this model component is used in the same way for model validation as it will be used to model simulations in an SEP project. These components should be clearly differentiated from declared model parameters for the purposes of model validation, such that both can be easily verified. All declared parameter sets must be validated following the guidance in section 3. If the minimums described in Section 3 do not result in all parameter sets being validated, and additional steps are not taken to validate all parameter sets, unvalidated parameter sets cannot be approved for use in the project.

Because biogeochemical models often contain a large number of parameters, different strategies can be employed to perform calibration. General guidance for frequentist and Bayesian approaches are provided in Sections 2.1 and 2.2.

2.1 Guidance on Model Calibration using Frequentist Approaches

Wallach et al. (2019) provide helpful guidance on common approaches to frequentist model calibration, including how to decide how many parameters to estimate, which parameters to estimate, whether to calibrate in stages, and how to avoid over-parameterization (i.e., where the model fits the data well but has poor predictive ability). Examples of model calibration are abundant in the peer-review literature and span a wide range of complexity and automation in their approaches (e.g., Bruun et al. 2003, Yeluripati et al. 2009, Liang et al. 2009).

2.2 Guidance on Bayesian Methods for Calibration, Validation, and Error

Model calibration can also be completed using Bayesian statistical methods, which apply a probabilistic approach to integrating existing knowledge and observed data (Wikle & Berliner, 2007). Bayesian statistical approaches are an emerging area of development in soil biogeochemical modeling. They typically require implementing Markov Chain Monte Carlo methods for sampling probability distributions. This can be computationally demanding with soil biogeochemical models, which can have dozens to hundreds or more parameters. Parameter values in these types of models can also be difficult to constrain (i.e., use data or existing knowledge to set limits on the range of values that a parameter may have, and define its probability distribution across that range). When there is little prior knowledge about a parameter value, 'uninformative priors' or 'weakly informative priors' are used to represent what is known or believed about the parameter. The resulting posterior distribution, or the distribution that represents the integration of prior knowledge and observed data, can be wide unless the observed data are strongly informative, i.e., have highly accurate and precise values. The following figure illustrates a strong prior belief (A) versus a weak prior belief (B).

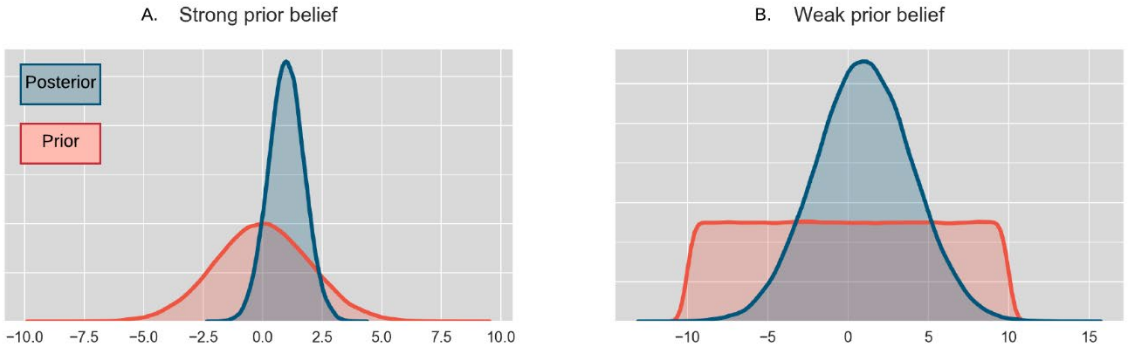


Figure 2.1. Comparison of Prior and Posterior Distributions when there is Strong Prior Belief Versus Weak Prior Belief

(E.g., strong and consistent evidence and prior analyses (A) versus weak or variable evidence or no prior analyses (B).)

Across dozens or hundreds of parameters, Bayesian methods can be complex to implement and require large quantities of data. Despite these challenges, Bayesian methods provide a coherent mathematical framework to integrate diverse sources of information into model parameterization, as evidenced by their central role in the developing field of Ecological Forecasting (Dietze, 2017), as well as in the Predictive Ecosystem Analyzer Project data-model integration system.[2] A Bayesian approach is encouraged for model validation and model prediction error, as the confidence intervals around model predictions will be directly based on the availability and variance of observed data. Figure 2.2 presents a conceptual workflow for a Bayesian approach to these analyses. A Bayesian approach also allows predictive error to be calculated from the posterior distributions of model parameters and of hyperparameters associated with structural uncertainty, and thus to be propagated to the posterior predictive intervals of individual observations.

[2] Accessible at: pecanproject.org

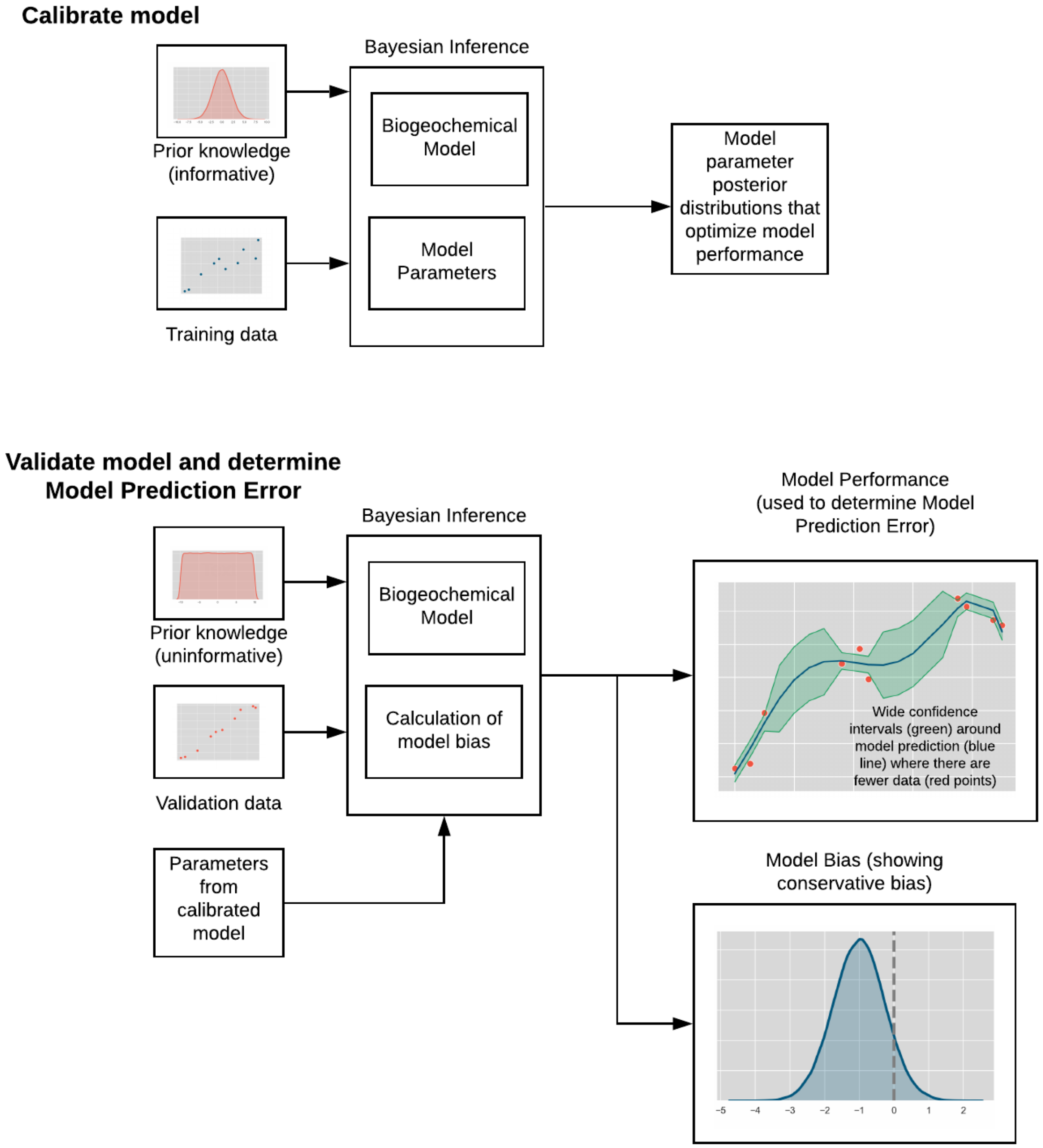

Calibrate model
Bayesian Inference
Prior knowledge (informative)
Biogeochemical Model
Training data
Model Parameters
Model parameter posterior distributions that optimize model performance
Validate model and determine Model Prediction Error
Bayesian Inference
Model Performance (used to determine Model Prediction Error)
Prior knowledge (uninformative)
Biogeochemical Model
Validation data
Calculation of model bias
Wide confidence intervals (green) around model prediction (blue line) where there are fewer data (red points)
Parameters from calibrated model
Model Bias (showing conservative bias)
−5
−4
−3
−2
−1
0
1
2

Figure 2.2. Conceptual Framework for Bayesian Approach to Model Calibration and Validation

In this example model calibration is a separate analytical process from validating model performance and determining model prediction error. In a fully integrated analysis, informative posteriors from model calibration might be used as priors in model validation.

2.3 Guidance on Cross-Validation Approaches

When evaluating statistical models in the presence of limited data, it is common to use cross-validation to evaluate the predictive performance of a given model against all available datapoints while guarding against overfitting. Traditionally, model validation is performed by dividing the available data into two independent, non-overlapping groups of datasets; one for training the model and the other “hold-out” group for model validation (Larson, 1931; Mosteller and Wallace, 1963). This traditional approach does not use all the available data for training, resulting in loss of information (Browne, 2000) and potentially poorer prediction performance. To improve information efficiency and prediction performance, several methods of using the same data for both training and validation have been proposed, including *k*-fold cross-validation (Mosier, 1951; Mosteller and Tukey, 1968; Refaeilzadeh et al., 2009), leave-one-out cross-validation (Geisser, 1975; Stone, 1976), and resampling validation via the “.632 bootstrap” (Efron and Tibshirani, 1997). All of these methods improve prediction performance by repeating hold-out validation multiple times across different subsets of the data and provide estimates of model uncertainty from the distribution of prediction performance across all results. This list is only illustrative; whatever approach is used, a validation report that uses a cross-validation approach must follow the guidelines below to demonstrate independence of calibration and validation data, correctly estimate the uncertainty of the model when applied to observations not in the validation set and validate the model using the same parameter values that will be used during credit estimation.

2.3.1.1 Demonstrate Independence

Because cross-validation makes use of the same data for both calibration and validation, when using k-fold cross-validation the independence between calibration and validation data must be demonstrated for each fold following the same principles discussed elsewhere in section 2. The demonstration should also explain how fold assignment accounted for any statistical dependencies that were induced by data structure, such as temporal or spatial or phylogenetic correlations between sites, soils or climate factors. For more details including recommended methods for constructing cross-validation folds in structured data, see Roberts et al. (2017).

2.3.1.2 Use All the Cross-Validated Parameters

Because cross-validation produces a range of results rather than a single output, it is understood that the parameter values within the declared set will be updated throughout the cross-validation process (e.g. each fold of a k-folding method will validate against a different calibration and therefore different parameter values). However, the method of choosing the final parameter set must be a prespecified part of the cross-validation method, and the parameter values identified as the final validated set (whether these are point values or distributions) must be the ones used to simulate project baselines and practices.

For example, a Bayesian modeling approach might perform ensemble simulations that draw their parameter values from the joint posterior distribution of all the calibrated parameters along with hyperparameters associated with structural model uncertainty (Gurung et al., 2020; Kennedy and O’Hagan, 2001). As a second example, a model operator might choose to collapse the cross-validation results to a single point value for each final parameter but to compute model uncertainty from model residuals across all validation folds. In this case additional demonstration would be needed to show that the resulting parameterization is scientifically reasonable (e.g. where all parameters are uncorrelated it may be sufficient to set them to the means of their marginal distributions, but if pairs of parameters are correlated this may produce incorrect model behavior) and that it does not understate model uncertainty.

2.3.1.3 Report Which Parameters are Calibrated

Many soil biogeochemical models have parameters numbering in the hundreds, and a model developer might choose to focus calibration on updating a selection of the most influential parameters while leaving other parameters as constants (Gurung et al., 2020). In this case the parameters held fixed can be viewed as a static part of the model structure with any miscalibration in their values contributing to the whole-model error term, while the parameters adjusted during cross-validation will have their uncertainty explicitly estimated. Because this affects interpretation of the cross-validation results, if cross-validation is performed on only a subset of model parameters the declared parameter set must indicate which parameters are updated during the validation procedure and which are left as constants. As for all validation approaches, all parameters must still be included in the declared final parameter set (whether they were calibrated or held constant) and the values used during crediting must be the same ones used during validation.

2.3.1.4 Document Variance Terms

For models that partition model uncertainty into multiple error terms, it must be clearly stated how the parameters and hyperparameters for these terms were estimated and how they were used in the cross-validation. For example, when validating a model that attributes some of the observed variance to a “location effect”, the report must specify whether the values used for each location in the validation data are the same values estimated for that location during calibration or if they are new draws from an estimated distribution of location effects.

Summary of Requirements Described in Section 2

Required for the Validation Report:

- **Model version**, defined as a uniquely traceable record of source code, internal parameters, and ancillary inputs that are needed to reproduce a given model output (collectively the “model files”). A model version must change any time there is a change in any of the model files.
- **Description** of the model calibration process
- **Documentation** of all internal model parameter sets, including proof that parameter sets are defined at a resolution no finer than one LRR.
- If there is justification to claim an allowance for crop growth parameter sets to vary within LRRs (e.g., varying maturity groups), crop growth parameter sets and their use must be documented per each LRR where the crop will be simulated. If using methods such as cross-validation that link the calibration and validation processes, also document: (1) the process for determining final parameter sets from the range of validation results, (2) which parameters are updated during the validation procedure and which are held constant, and (3) how variance terms are estimated and used during calibration and validation.
- **Justification** for splitting of experimental data between calibration and validation (where applicable), clearly described at the crop functional group/practice category/emissions source combination level. If using methods such as cross-validation that link the calibration and validation processes, demonstrate independence between folds.

Required Upon Request of the Verification Team:

- **Datasets** used for model calibration, including but not limited to full citation, experimental locations, specific crops and practices studied, LRRs and IPCC climate zones, soil textures and clay contents, and number of observations.

3 Validating and Reporting Model Performance and Uncertainty

3.1 Declare Practice Categories Requiring Evaluation

For every practice considered additional within the project (as determined by application of Section 3 of the SEP), the model must be shown to have an acceptable goodness of fit and unbiased representation of the underlying biogeochemical process governing the effect of that practice. To do so, each practice must be binned into the Practice Categories (PCs) shown in Table 3.1 to demonstrate the domain of practice effects and the categories requiring evaluation. Validating model performance and uncertainty within a practice category can be accomplished using any practice effect in the category domain, evaluated using appropriate experimental data meeting requirements described below. Projects are encouraged to evaluate a range of practice effects in each practice category domain. If the project employs an additional practice change which does not fit obviously into one of the practice categories in Table 3.1, guidance must be sought from the Reserve.

Table 3.1. Practice Categories and their Associated Practice Effects Requiring Biogeochemical Performance Evaluation

Practice Category Requiring Evaluation	Domain of Practice Effects
Inorganic nitrogen fertilizer application	Magnitude, form, timing, or method for nitrogen fertilizer applied, with form encompassing inorganic N fertilizers, and method encompassing surface, subsurface, or irrigation-based application
Organic amendments application	Magnitude, form, timing, method or variation in C:N ratio for organic amendments applied. Forms include and are not limited to biochar, mulch, compost, and manure, and methods encompass surface, subsurface, or irrigation-based application
Water management/irrigation	Magnitude, timing, source or method of irrigation water applied
Soil disturbance and/or residue management	Soil disturbance including tillage and compaction, and residue management encompassing soil exposure after harvest and physical incorporation of green manure
Cropping practices, planting and harvesting (e.g., crop rotations, cover crops)	Variety of crops grown, increasing crop rooting depth, may include cover crops and soil preparations such as changing soil pH through liming
Grazing practices	Any of the following: presence/absence of grazing, stocking density, forage type or quality, species of grazers, mixed or single species herds, loading weight, grazing time, and rest/recovery periods

A project developer must declare all practice effects requiring evaluation for the project.

3.2 Define the Project Domain

For each practice category declared in the project description, the model must be evaluated in terms of its fit and bias in estimating emissions reductions. Evaluation of each category begins with defining the project domain in terms of its biophysical attributes. Specifically, the project developer must declare the unique crop functional groups, land resource regions, and soil attributes associated with each declared practice category.

3.2.1 Declare Project Crop Functional Groups

Crop functional groups (CFGs) for each practice category must be declared. Individual crop types can be grouped into functional groups across crops sharing unique combinations of the following attributes:

- N fixation (Y/N),
- Annual/perennial (A/P) (defined in accordance with the NRCS Conservation Compliance categorization of crops[3]),
- Photosynthetic pathway (C3/C4/CAM),
- Growth form (tree/shrub/herbaceous) (trees and shrubs have woody plant growth, versus herbaceous species that do not grow woody plant material),
- Flooded/not flooded

3.2.2 Declare Project Land Resource Regions

The full list of land resource regions (LRRs) associated with each practice category must be declared.[4] LRRs represent distinct combinations of climate, land resource use, and geographic features. To support acquisition of studies to validate model performance and uncertainty outside the US, IPCC climate zones must be declared for each practice category following the climate zone definitions given in the 2006 IPCC Guidelines for National Greenhouse Gas Inventories.

3.2.3 Declare Project Soils

Soils are to be declared for each practice category in terms of (1) soil textural class and (2) the associated clay content[5] of that class. NRCS soil texture classes include sand, loamy sand, sandy loam, loam, silt loam, silt, sandy clay loam, clay loam, silty clay loam, sandy clay, silty clay, and clay.

Summary of Section 3.2

Required for the Validation Report (Types 1 and 3- described Section 3.6)

- List of combinations of PCs and CFGs occurring in the Project
- List of combinations of PCs, CFGs, and emissions sources (ESs) validated
- List of LRRs and IPCC climate zones included in the project domain
- List of soil texture classes and associated clay contents in the project domain
- (Type 3 only) Corresponding Type 2 report containing referenced model validation

Required for the Validation Report (Type 2- described Section 3.6)

[3] Resource can be found here: https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/programs/farmbill/?cid=stelprdb1262733

[4] Resource can be found here: https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/technical/nra/nri/?cid=nrcs143_013721

[5] See Table A-1 for clay contents of NRCS soil textural classes. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/?cid=nrcs143_014055

- List of combinations of PCs, CFGs, and ESs validated
- LRRs and IPCC climate zones for which each combination is validated
- List of soil texture classes and associated clay contents for which each combination is validated

Required Upon Request of the Verification Team

- List of specific crops and practices occurring in the Project, and a description of how these were binned into the PCs and CFGs validated.

3.3 Gather Data to Validate Model Performance and Uncertainty

Requirement 1: Generalized Dataset Attributes

Datasets to validate model performance and uncertainty for each declared PC/CFG/ES combination from Section 3.1 must include measurements for each modeled quantity, where the modeled quantity is the change in the flux of emissions to the atmosphere for SOC, N_2O, and/or CH_4 that results from the adoption of any practice associated with that effect. Datasets may include individual practice categories as well as combinations of practice categories (e.g., “stacked” practices), provided the practice category in question is experimentally varied and measured within the study. Some hypothetical examples of acceptable experimental treatments to evaluate practice categories are given in the following table:

Table 3.3. Examples of Some Permissible Experimental Treatments to be Used in Evaluating Practice Categories

(see Table 3.1 for the complete list of practice effects that can be included in each Practice Category)

Experimental Treatment	Practice Category
Comparison of two different application rates of urea	Inorganic nitrogen fertilizer application
Comparison of conventional tillage using moldboard plow to strip tillage	Soil disturbance and/or residue management
Comparison of single-crop rotation to double-crop rotation; comparison of bare fallow to cover crop	Cropping practices, planting and harvesting (e.g., crop rotations, cover crops)

Datasets to validate model performance and uncertainty must adhere to the following guidelines:

- Measured datasets must be drawn from peer-reviewed and published experimental datasets with measurements of the emissions source(s) of interest (SOC stock change and/or N_2O and CH_4 change, as applicable), ideally using control plots to test the practice category.
- Studies must report sufficient information to be modeled, i.e., providing enough information that model inputs have low uncertainty relative to modeled results, and the model can be appropriately initialized. ‘Enough’ information to initialize and run a model accurately is model- and emissions source-specific. Therefore, the reported information required to initialize and model a study accurately should be described for the model version and parameter sets being validated, and any processes used to address unreported information fully described in the Validation Report.

- Studies reporting the effects of changing multiple practices at once ("stacked" practice changes) may be used provided that the composite of all studies used to validate a PC/CFG/ES combination contains at least one study that isolates the effect of the practice change being validated.
- All validation dataset sources must be reported. The same measurement dataset sources can be used for validating multiple practice categories, when appropriate. Datasets may be used from studies outside of the US. However, the associated IPCC climate zone where these datasets were collected should correspond to the declared IPCC climate zones of the project.
- In the case of SOC stocks, repeat measurements of SOC stock change must be able to capture multi-year changes, as practice effects on SOC may combine short and long-term changes in soil biogeochemical processes. Measurements from paired fields leveraging space-for-time analysis methods that approximate multi-year changes may also be used for SOC validation. Newer methods for SOC stock monitoring are becoming available that can observe changes with greater precision at shorter time intervals. These methods will be acceptable if there is peer-reviewed support or independent expert support approved by the Reserve for their use in SOC monitoring and if the methods demonstrate accurate measurement of multi-year impacts on SOC stock changes. Measured datasets of SOC stock change may be made at any depth, but the model must also predict SOC stock change at the corresponding depth. Thus, a fully compiled dataset for validating model performance and uncertainty may contain different depths for SOC stock change measurements as long as the model is predicting SOC stock change at each corresponding depth.
- In the case of N_2O and CH_4 flux, any combination of measurements from chambers and/or eddy covariance flux towers are acceptable. Methods of temporal aggregation should be documented in the Validation Report (e.g., Mishurov & Kiely, 2011; Turner et al., 2016), as well as the portions of the calendar year covered, in aggregate, by all N_2O and/or CH_4 measurements. Justification should be provided when portions of the year are missing.
- Datasets can be drawn from a benchmark database maintained by a third party, if approved by the Registry. The use of datasets from a benchmark database should include full citation of the database as well as a description of how datasets were extracted, including exclusion criteria for any records not used in the validation.
- Project developers are expected to use a process for selecting data for validating model performance and uncertainty that results in the assembly of datasets that are representative of the range of peer-reviewed observed results. Project developers must describe the methods, selection process, and data manipulations used to create the dataset applied in the model validation process. This includes describing search terms and databases used to identify available datasets, criteria used to select dataset sources, origin of extracted data (e.g., figures, tables, databases with DOI), original units of data and data uncertainty, and data manipulations used to convert original units into the units described above. The project developer should report the number of validation data measurements of each data type (SOC, N_2O and CH_4) for each project domain combination of PC and CFG, and include a histogram showing the range of validation data values (e.g., measured SOC change). In the case where validation data are unevenly distributed across the project domain (e.g., almost all validation data are reported in sandy soils, with only a few in soils with higher clay content), the method used to link validation data to model structural error (described in more detail in Section 3.5 below) should demonstrate that it addresses the discrepancy.

Requirement 2: Specific Dataset Requirements to Validate Model

The specific requirements for validating model performance and uncertainty for a PC/CFG/ES combination are set based on the geographic extent of a project (i.e., the number of declared LRRs), as well as the soil attributes encountered within the project (i.e., the declared soil textural classes and clay contents).

The following logic is used to determine the number of declared LRRs required, in total, in the combination's dataset. If the number of declared LRRs is equal to:

- 1, then the validation dataset for that combination must include that (declared) LRR;
- 2, then the validation dataset for that combination must include those two (declared) LRRs;
- 3 or more, the validation dataset for that combination must include at least 3 (declared) LRRs.

For all PC/CFG/ES combinations, at least three declared soil textural classes must be represented in the validation dataset, and the range in clay contents must span at least 15 percentage points. When the number of declared soil textural classes is less than three, all textural classes that do occur within the project's geographic extent must be included in the dataset, and there must be a range in clay contents spanning at least 15 percentage points. Once validated, a PC/CFG/ES combination will be approved for crediting within all declared LRRs and for all declared soil textures.

The purpose of these minimums is to ensure testing for generalized model performance, i.e., that a model is not hyper-calibrated for a specific combination of factors that leads to poor model performance in other contexts. It is in a project's interest to exceed these minimums and validate the model across more LRRs, soil texture classes, and clay contents, because model prediction error must use the same dataset as model validation and will penalize the use of few data points (see Section 3.5). If the available data fail to meet one of these minimums but exceeds the others in a way that supports a demonstrable test of generalized model performance, a case may be made for a valid exception to Requirement 2. This should be addressed explicitly in the Validation Report and will need to be approved by the Registry and by the external reviewer.

Note that all model parameter sets used in crediting must be validated for each PC/CFG/ES combination (see Section 2). If model parameter sets vary by LRR this may require additional measurement datasets beyond the minimum described above to ensure all parameter sets are validated.

3.3.1 Special Rules for Practice Categories

For studies used in validating model performance and uncertainty in the Cropping practice category, any CFG occurring within the experimental period of measurements may be counted toward validation. For example, if two rotations were compared where one had a repeating corn-soy rotation, and the other introduced a cover crop between corn and soy, the study could be used to validate all three of the CFGs associated with corn, soy, and the cover crop for the Cropping practice category, provided that experimental measurements spanned at least one full rotation.

If grazing practices have been validated on pasture, and a CFG has been validated for either the Cropping or Soil Disturbance practice categories, the model can be considered validated for grazing on residue for that CFG. For grazing practices, pasture can be defined as any perennial grass or legume. C3 and C4 grasses do not need to be validated separately for pasture grazing.

For rice cropping systems, inorganic sulfur fertilizer application can be considered an extra practice category eligible for crediting due to its effects in reducing methane emissions. Validation of the inorganic sulfur fertilizer application practice category would be analogous to the inorganic nitrogen fertilizer application practice category, and would encompass the same domain of practice effects to be used in validation (i.e., magnitude, form, timing, or method for nitrogen fertilizer applied, with form encompassing inorganic S fertilizers, and method encompassing surface, subsurface, or irrigation-based application).

For studies focused on grass blends that include a mixture of C3 and C4, or N-fixing and non-N-fixing, all CFGs represented in the blend may be considered represented in that study.

When validating a model for the Organic Amendments Application practice category, data from all CFGs classified as “annual” may be pooled together for validation, and the validation result may be considered applicable for crediting of organic amendment practices for all included annual CFGs. Each perennial CFG must still be validated separately.

When validating a model for the Inorganic N Fertilizer Application practice category, it is expected that validation data may be scarce for CFGs that fix N (e.g. soybean), because these crops are often grown without N fertilization. Therefore the model may be considered validated for annual, herbaceous, C3, N-fixing crops if (1) inorganic N fertilizer application has been successfully validated for another annual CFG, and (2) the annual, herbaceous, C3, N-fixing CFG has been successfully validated for the Cropping, Planting, and Harvesting PC.

Cropping systems using irrigation as a normal part of management separate from practices intended to reduce emissions, i.e. where irrigation is present in both project and baseline, are not required to have the Water Management/Irrigation PC validated, provided that irrigation is represented in at least one study in the validation dataset.

Summary of Section 3.3

Required for the Validation Report

- Full description of data requirements to initialize and run the model version and parameter sets accurately, as well as the process for addressing missing information
- A full accounting of the studies comprising the validation dataset for each CFG/PC/ES combo, for each emissions source. Study attributes should include:
  - Citation
  - LRR and IPCC climate zone
  - PC and CFGs being studied
  - Soil texture(s) and clay contents being studied
  - Experimental time period
  - Depths of SOC measurements
  - Measurement technique, e.g., dry combustion for SOC, or chambers for N_2O
  - Methods of temporal aggregation used for observations of N_2O and CH_4
  - Portions of the calendar year covered by all N_2O and/or CH_4 measurements, with justification provided when portions are missing.
  - Number of observations used in validation
  - Measurement uncertainty associated with replicates, where reported
  - Experimental location (only when split between calibration and validation)

Required Upon Request of the Verification Team

- Additional details for validation studies including, but not limited to:
  - Experimental location and corresponding LRR and IPCC climate zone
  - Specific crops and practices being studied
  - Original units of measurements
  - Mathematical transformations performed on measurement data

Study-specific use of data to initialize and run the model, as well as a record for the filling of missing information using process described in Validation Report

3.4 Assessment of Bias for Each PC/CFG/ES Combination

For each PC/CFG/ES being validated, the model must be shown to be unbiased in estimating the change in SOC, N_2O, or CH_4 pools for the project domain defined in Section 3.2, using measured data that meet the requirements of Section 3.3. This is done using the calculation of *bias*, a simplified version of average relative error (FAO, 2019), calculated between measured data and model predictions. Bias indicates the average tendency of the modeled estimates to be larger or smaller than their observed counterparts (Moriasi et al., 2007). Positive values indicate model overestimation bias, meaning that the model overestimates the practice effect. A negative value indicates model underestimates the practice effect.

The calculation of bias is defined as:

Equation 3.1

$$\boldsymbol{bias = (\sum_{i=1}^{n} P_i - O_i)/n}$$

Where,

P_i = Predicted (modeled) change in SOC, N_2O, or CH_4 with the practice

O_i = Observed change in SOC, N_2O, or CH_4 with the practice

n = Number of observations for a given study

Model bias should be calculated for each study and a mean bias should be computed as the unweighted mean of all biases from individual studies. The mean bias should be less than or equal to an estimate of pooled measurement uncertainty (PMU). Pooled measurement uncertainty (PMU) is defined as the pooled standard error of all the measured values for a practice change, where standard error is derived from replicates of the measurements (Figure 3.1.). Because not all studies will report measurement standard error, PMU may be computed using all studies used in a Validation Report using the same measurement technique. When PMU cannot be reasonably obtained, a default replacement value may be used for PMU that is based on typical measurement error for a given measurement technique, per approval of the Registry.

Equation 3.2

$$\boldsymbol{\sigma_{meas}} = \sqrt{\frac{\sum_{j=1}^{k} \boldsymbol{\sigma_j^2 (n_j - 1)}}{\sum_{j=1}^{k} \boldsymbol{(n_j - 1)}}}$$

Where,

k = Number of observations examined

σ_j = Standard error of the j^{th} observed change in SOC, N_2O, or CH_4

n_j = number of replicate measurements used in the j^{th} observation

Evaluation of bias may be done for studies comparing more than two treatments, for example three fertilizer rates, by pairing observations so that each of the three treatments is isolated as a ”control” in turn. This will result in three sets of comparisons, once for each combination of treatments.

A model is judged as valid if mean model bias is less than PMU, and model predictive uncertainty is determined as described in Section 3.5. However, per-study bias should be reported, ranked from highest to lowest. The intention of reporting per-study bias, as well as evaluating mean model bias compared to PMU, is to avoid penalizing any one study in terms of measured data or model performance (i.e., where there are few or variable measured data, or the model is biased in its prediction).

However, it should be recognized that there may be circumstances where a model may be performing reasonably well even if mean bias is greater than PMU, for example due to limited availability of measured datasets or poor reporting of measured uncertainties. A project developer is allowed to petition for validating the model for use, if it can be clearly justified that the model is showing reasonable overall performance given available measured data. Such a petition will need to be approved by the Registry, following expert review.

In this model evaluation framework, large model biases result in large residuals. Following guidance for model predictive error in section 3.5, this means large model bias in either direction (positive bias or negative bias) will result in large predictive uncertainty, and thus increase credit deductions. High model prediction error will therefore be yielded in two circumstances- 1) through low precision of an accurate model or 2) high precision of an inaccurate model. Figure 3.1. and Figure 3.2 walk through an illustrative process to meet the requirements described above.

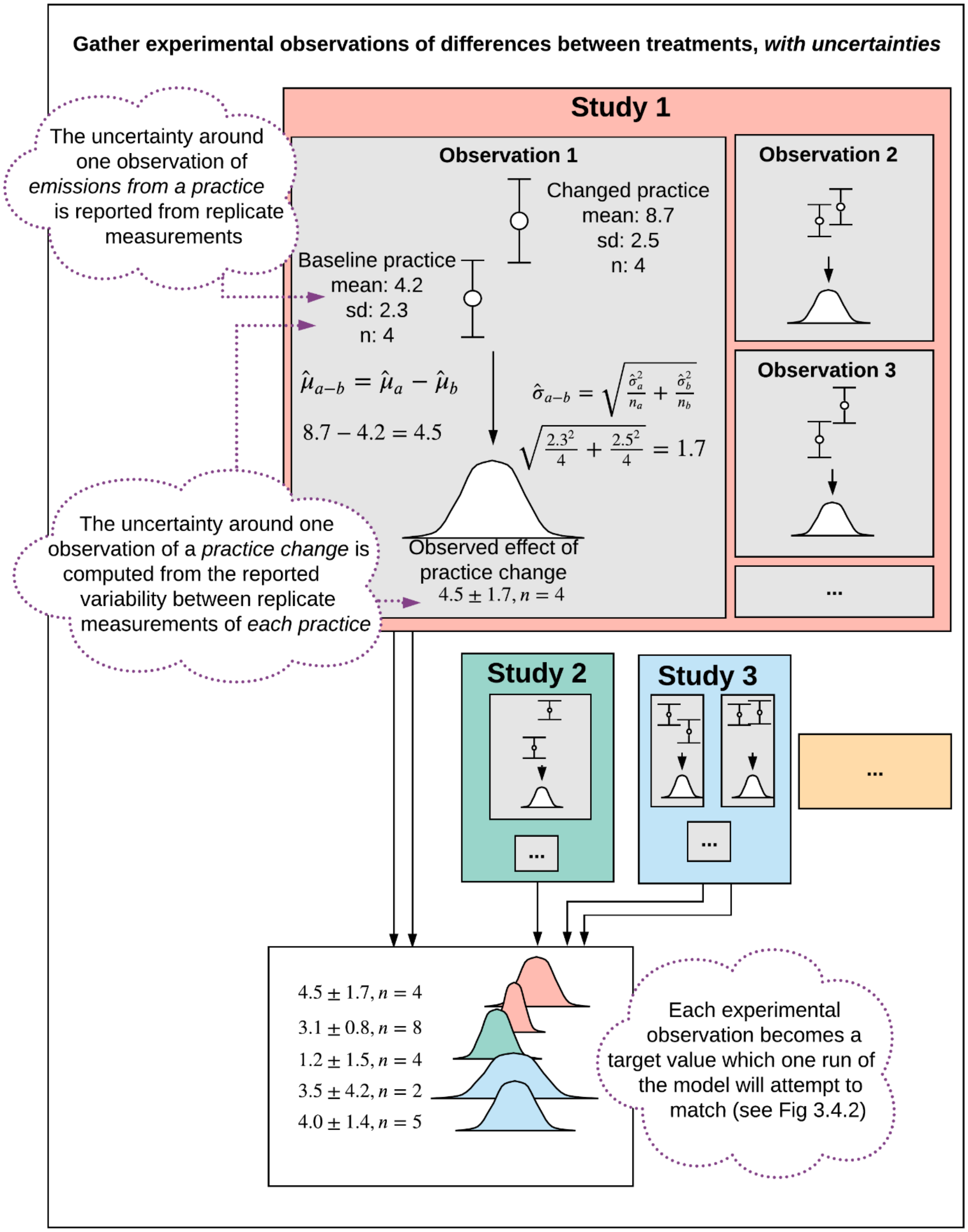
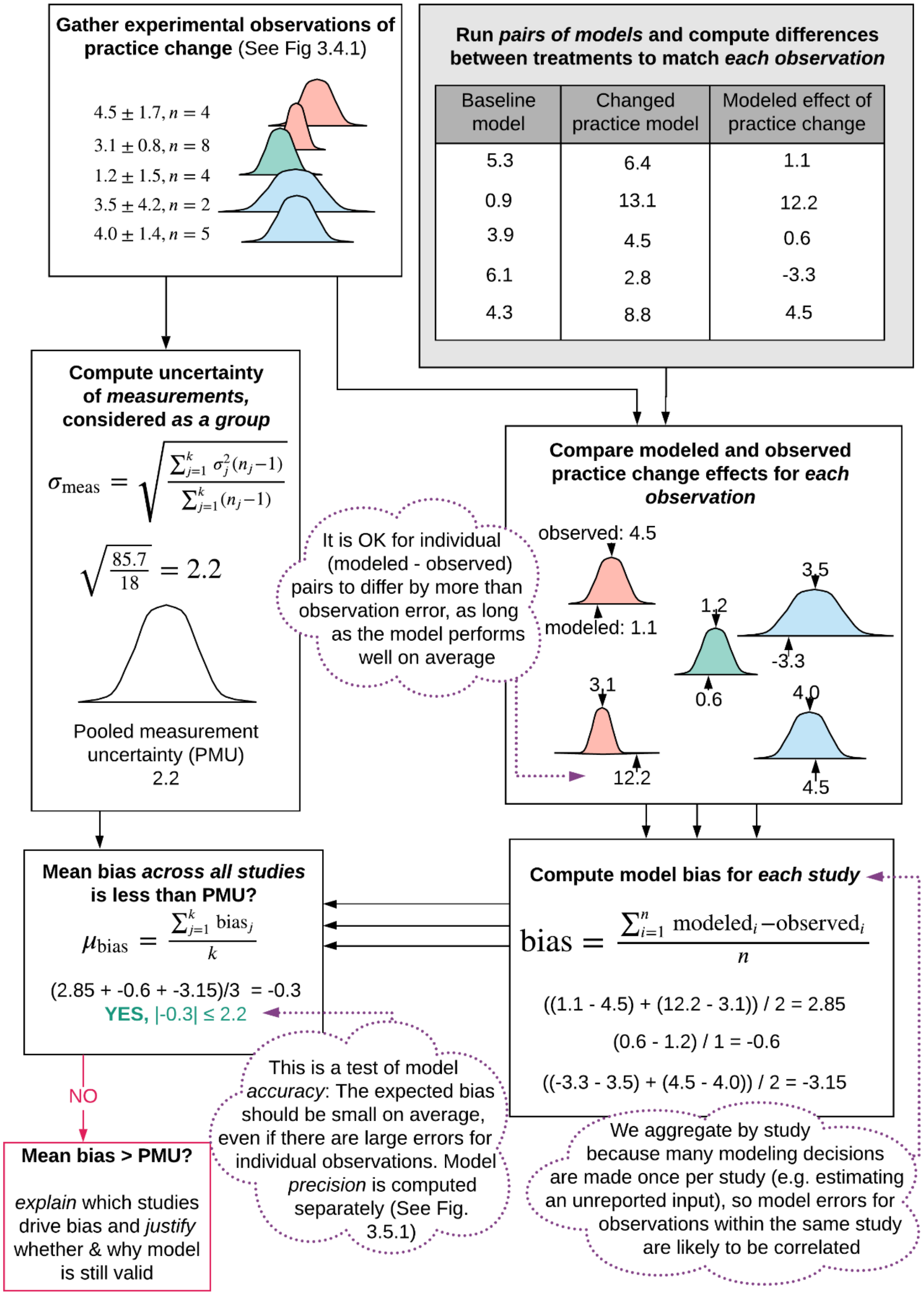

Gather experimental observations of differences between treatments, *with uncertainties*
Study 1
Observation 1
Changed practice
mean: 8.7
sd: 2.5
n: 4
Baseline practice
mean: 4.2
sd: 2.3
n: 4
$\hat{\mu}_{a-b} = \hat{\mu}_a - \hat{\mu}_b$
$8.7 - 4.2 = 4.5$
$\hat{\sigma}_{a-b} = \sqrt{\frac{\hat{\sigma}_a^2}{n_a} + \frac{\hat{\sigma}_b^2}{n_b}}$
$\sqrt{\frac{2.3^2}{4} + \frac{2.5^2}{4}} = 1.7$
Observed effect of practice change
$4.5 \pm 1.7, n = 4$
Observation 2
Observation 3
...
The uncertainty around one observation of *emissions from a practice* is reported from replicate measurements
The uncertainty around one observation of a *practice change* is computed from the reported variability between replicate measurements of *each practice*
Study 2
...
Study 3
...
...
$4.5 \pm 1.7, n = 4$
$3.1 \pm 0.8, n = 8$
$1.2 \pm 1.5, n = 4$
$3.5 \pm 4.2, n = 2$
$4.0 \pm 1.4, n = 5$
Each experimental observation becomes a target value which one run of the model will attempt to match (see Fig 3.4.2)

Figure 3.1. Visual Summary of One Possible Approach to Calculations for Determining Measurement Uncertainty of an Observed Practice Change Effect



Gather experimental observations of practice change (See Fig 3.4.1)

Run *pairs of models* and compute differences between treatments to match *each observation*

Baseline model	Changed practice model	Modeled effect of practice change
5.3	6.4	1.1
0.9	13.1	12.2
3.9	4.5	0.6
6.1	2.8	-3.3
4.3	8.8	4.5

Compute uncertainty of *measurements*, considered *as a group*

$$\sigma_{\text{meas}} = \sqrt{\frac{\sum_{j=1}^{k} \sigma_j^2 (n_j - 1)}{\sum_{j=1}^{k} (n_j - 1)}}$$

$$\sqrt{\frac{85.7}{18}} = 2.2$$

Pooled measurement uncertainty (PMU) 2.2

Compare modeled and observed practice change effects for *each observation*

It is OK for individual (modeled - observed) pairs to differ by more than observation error, as long as the model performs well on average

Compute model bias for *each study*

$$\text{bias} = \frac{\sum_{i=1}^{n} \text{modeled}_i - \text{observed}_i}{n}$$

((1.1 - 4.5) + (12.2 - 3.1)) / 2 = 2.85

(0.6 - 1.2) / 1 = -0.6

((-3.3 - 3.5) + (4.5 - 4.0)) / 2 = -3.15

We aggregate by study because many modeling decisions are made once per study (e.g. estimating an unreported input), so model errors for observations within the same study are likely to be correlated

Mean bias *across all studies* is less than PMU?

$$\mu_{\text{bias}} = \frac{\sum_{j=1}^{k} \text{bias}_j}{k}$$

(2.85 + -0.6 + -3.15)/3 = -0.3

YES, |-0.3| ≤ 2.2

This is a test of model *accuracy*: The expected bias should be small on average, even if there are large errors for individual observations. Model *precision* is computed separately (See Fig. 3.5.1)

Mean bias > PMU?

explain which studies drive bias and *justify* whether & why model is still valid

Figure 3.2. Visual Summary of Calculations for Demonstrating that Model Bias is on a Similar Scale as Measurement Error

Summary of Section 3.4

Required for the Validation Report

- One complete example derivation of:
  - Calculation of model bias for a study, per Figure 3.1.
  - Calculation of PMU for a single measurement technique, per Figure 3.2.
- All values of PMU used for each PC/CFG/ES combination validated.
- All values of study bias for each study in a PC/CFG/ES's validation dataset, ranked highest to lowest
- Average bias across all studies in a PC/CFG/ES's validation dataset.

Required Upon Request of the Verification Team

- Complete derivations and/or calculations made of PMU, study bias, and average model bias for each PC/CFG/ES combination.

3.5 Using Data to Evaluate Model Prediction Error

In order to validate the model for performance and uncertainty, the same datasets should be used to estimate the uncertainty of a model's predictions, i.e., the model prediction error (SEP, Appendix D) and evaluate model fit. The calculation of model uncertainty bounds associated with a particular prediction (i.e., the prediction interval) should account for where there are few validation data (e.g., by using a weakly informative prior if using a Bayesian framework, Fig 2.1.B) as well as account for data variability (i.e., with a wider posterior when data are more variable if using a Bayesian framework; Fig 2.2). These features enable the model to adequately estimate the confidence in its predictions, as described next.

The SEP allows model prediction errors [to be computed by either of two methods][6]:

- [Analytical error propagation (SEP Appendix D.1), which requires only the mean squared error of validation simulations; or][6]
- [Monte Carlo simulation (SEP Appendix D.2), which requires declaration of the distributions of model parameters and/or inputs (including any hyperparameters for model structural uncertainty) obtained from validation simulations.][6]

The Validation Report may support [either or both approaches][6] by reporting the values needed to compute model prediction errors using the chosen Appendix. Thus if parameter distributions are not reported, then the model may not be considered validated for use with a Monte Carlo approach to uncertainty calculation. If using a Monte Carlo approach, simulations used to generate estimates of emissions reductions for an SEP project will need to use the same parameter values reported in the validation report, as well as include random draws from any hyperparameters.

Whatever approach is used to compute model uncertainty bounds, the Validation Report must demonstrate that they have been set appropriately. Specifically, measured versus modeled results should be compared for each PC/CFG/ES combination for changes in SOC, N_2O, and CH_4 (if relevant), and demonstrate a minimum confidence coverage of 90% for 90% prediction intervals (i.e., the 90% prediction intervals should contain the measured value for at least 90% of the validation data). It should be recognized that there may be circumstances where model

[6] Text in brackets aligns with a revision to SEP Appendix D that is currently under consideration by CAR but not yet fully approved. If this revision undergoes further edits before finalization, the bracketed wording may require updating.

uncertainty bounds are appropriately set even if 90 % conference coverage is not achieved, for example due to limited availability of measured datasets. A project developer is allowed to petition for validating the model for use with such error bounds, if it can be clearly justified that the model prediction error is appropriately set given available measured data (for example, error bounds that cover 6 out of 7 observations, or 7 out of 8 observations where missing one drops the confidence coverage below 90%). Such a petition will need to be approved by the Registry, following expert review.

Validation data come from experiments that range in duration from a few years to many decades, and model prediction error at each point is derived from simulations that match the durations of those experiments. This means that the prediction intervals needed for the 90% confidence coverage check will necessarily represent the accumulated model error over varying time intervals. By contrast, the quantity of interest when estimating model prediction error for SEP reporting is always the model error from a single reporting period (generally on the order of one year) and therefore likely to be smaller than the raw mean squared validation error.

For reporting periods 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 and is acceptable to use as such for SEP reporting. For example, if a model is validated against a dataset containing experiments with lengths of (2, 2, 3, 5, 9, 48) years, the error from this validation may be applied to any simulation of length 4 years or shorter.

It is also acceptable to include an explicit timescale component in the model's uncertainty calculations, such that the time since measurement is taken into account and shorter intervals are assigned commensurately less timecourse uncertainty. Exact methods of performing such a correction are not yet specified here, and any such correction must be fully described in the Report and approved by the Registry, following expert review.

In the model Validation Report, the following should be also included for each PC/CFG/ES combination and for changes in SOC, N_2O, and CH_4:

- Scatterplot of the model predictions versus measurements
- Histogram of residuals (the differences between predictions and measurements)
- Mean squared error

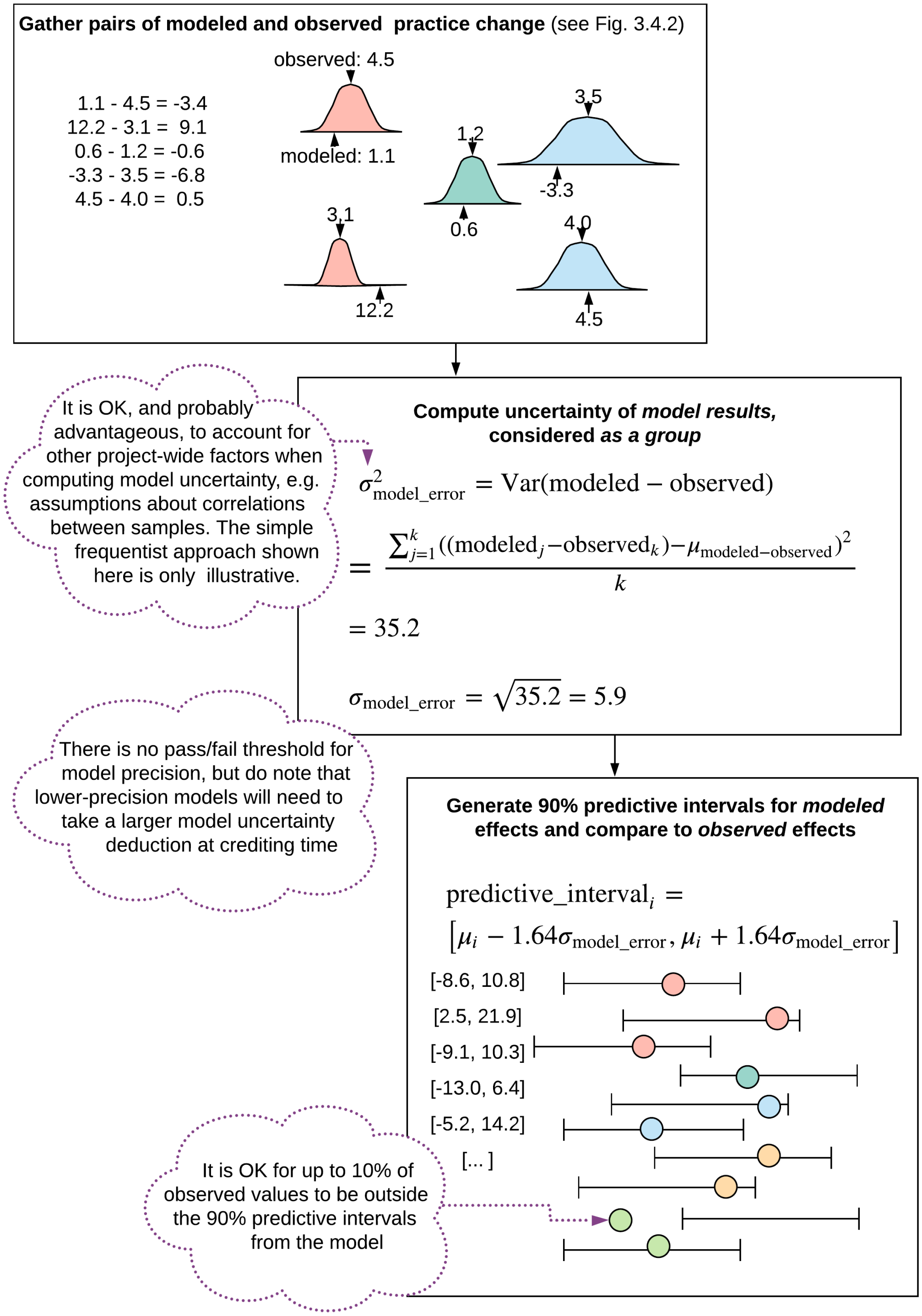

Gather pairs of modeled and observed practice change (see Fig. 3.4.2)
1.1 - 4.5 = -3.4
12.2 - 3.1 = 9.1
0.6 - 1.2 = -0.6
-3.3 - 3.5 = -6.8
4.5 - 4.0 = 0.5
observed: 4.5
modeled: 1.1
3.5
1.2
-3.3
0.6
3.1
4.0
12.2
4.5
Compute uncertainty of *model results, considered as a group*
$\sigma^2_{\text{model\_error}} = \text{Var}(\text{modeled} - \text{observed})$
$= \frac{\sum_{j=1}^{k}((\text{modeled}_j - \text{observed}_k) - \mu_{\text{modeled}-\text{observed}})^2}{k}$
$= 35.2$
$\sigma_{\text{model\_error}} = \sqrt{35.2} = 5.9$
It is OK, and probably advantageous, to account for other project-wide factors when computing model uncertainty, e.g. assumptions about correlations between samples. The simple frequentist approach shown here is only illustrative.
There is no pass/fail threshold for model precision, but do note that lower-precision models will need to take a larger model uncertainty deduction at crediting time
Generate 90% predictive intervals for *modeled* effects and compare to *observed* effects
$\text{predictive\_interval}_i =$
$\left[\mu_i - 1.64\sigma_{\text{model\_error}}, \mu_i + 1.64\sigma_{\text{model\_error}}\right]$
[-8.6, 10.8]
[2.5, 21.9]
[-9.1, 10.3]
[-13.0, 6.4]
[-5.2, 14.2]
[...]
It is OK for up to 10% of observed values to be outside the 90% predictive intervals from the model

Figure 3.3. Illustrative Example of One Possible Approach to Computing Model Prediction Error and Demonstrating that Model Predictions are Consistent with Validation Data

Summary of Section 3.5

Required for the Validation Report

- For each PC/CFG combination and emissions source:
  - Graphs of measured versus modeled results demonstrating that the 90% prediction intervals contain the measured value at least 90% of the time, per Figure 3.3.
  - Scatterplot of the model predictions versus measurements
  - Histograms of residuals (the differences between predictions and measurements)
  - Mean squared error
- Documentation of final parameter sets and model prediction error for each emissions source
  - If using SEP Appendix D or D.1]: Document the model prediction error for emissions reductions as estimated using mean squared error
  - If using SEP Appendix D.2][7]: Document the distributions of model parameters and hyperparameters

3.6 Reporting on Model Validation

A model Validation Report following the above requirements and guidance should be submitted to the Reserve either by the expert team member or a third-party expert employed by the project during a given reporting period. To be accepted by the Reserve, the Report will need to have received review and approval by an independent expert entity as having followed guidance in this document. The independent expert entity will need to be approved by the Reserve, following an assessment of conflict of interest between the reviewing entity and the project developer. Model validation requirements must be satisfied and confirmed prior to the completion of verification activities. This validation report will need to be revised, re-reviewed, and re-submitted at any point changes are made to the model.

There are 3 options for model Validation Reports:

1. Project-specific, that includes demonstration of model validation for a specific project's domain and combinations of crop functional groups, practices categories, and emissions sources;
2. Generalized to demonstrate overall performance of a given model, i.e., demonstrating where model performance is valid over a range of possible project domains and crop functional group/practices category combinations; or
3. Project-specific and referencing an existing model Validation Report (type 1 or 2).

The above guidelines are written for type 1 model Validation Reports. A type 2 Report would follow the same guidelines, but clearly identify all project domains and crop functional group/practice category/emissions source combinations where a given model version and parameter set/s have been calibrated and validated for model performance and uncertainty. A type 3 Report would follow the same guidelines, but clearly demonstrate that the referenced model validation report (type 1 or 2) meets model validation requirements for the specific project, including proof that the same model version and parameter sets are used, and that all project domain and crop functional group/practices category combinations have met minimum requirements for model validation. All types of model Validation Reports must either be independently assessed by an approved third-party entity or accepted for publication in one of the peer-reviewed publications listed in Section 3.6.

For each subsequent monitoring report, as long as a project area remains constant, or is only expanded to include new fields that already fit within the validated project domain, the existing Validation Report can be used. If the project is expanded to new practice effects, new crop functional groups, or the model is changed (using new model calibrations, changed parameters/parameter sets, or a different model, following guidance described in Section 6.5 of the SEP protocol), the Validation Report needs to be revised, reviewed by an approved third-party entity, and re-submitted. All model Validation Reports will be public documents.

It is also acceptable that the Validation Report is submitted as a peer-reviewed journal article, provided that the journal is on the pre-approved list provided below and the publication is approved for use by the Registry. It is acceptable that the journal article has not yet been printed as long as it has passed peer review and has been accepted for publication with revisions that do not change any aspects of model validation following the guidance in this document. In this circumstance, the project should submit the peer reviewed publication and responses to all revisions that clearly demonstrate revisions do not impact model validation. Where the peer-reviewed publication option in pursued, it is additionally acceptable that model validation is completed using a different method than explicitly evaluating bias and goodness of fit as described above. The publication must demonstrate that separate datasets were used for model calibration and model validation (unless qualifying for a special exception; see Section 3.1). The model validation must demonstrate the model was found acceptable for use by the peer reviewers for a given biophysical domain and a given set of practices. Additionally, the biophysical domain and practices used in the publication must be shown to completely meet the same domain requirements laid out in Sections 3.2 and 3.3, as well as cover the practice categories and crop functional groups identified in Section 3.1. The same datasets used in the peer-reviewed model validation should be used to calculate model prediction error used in the project and evaluate model uncertainty. The same model version and model parameter values/parameter set values must be used in the peer-reviewed publication as are used in the project. Lastly, as a means of enhancing transparency with the peer reviewers, the authors must clearly state the purpose of the paper as being to validate the model for use in generating verifiable carbon credits.

Pre-approved Journals:

- Agricultural and Forest Meteorology
- Agricultural Systems
- Agriculture, Ecosystems and Environment
- Agronomy Journal
- Atmospheric Environment
- Biogeochemistry
- Biogeosciences
- Ecological Applications
- Ecological Modeling
- Ecosystems
- Environmental Modelling and Software
- Environmental Pollution
- Field Crops Research
- Frontiers in Ecology and the Environment
- Geoderma
- Global Biogeochemical Cycles
- Global Change Biology
- Journal of Environmental Quality

- Journal of Geophysical Research - Biogeosciences
- Nutrient Cycling in Agroecosystems
- Plant & Soil
- PLoS ONE
- Science of the Total Environment
- Soil & Tillage Research
- Soil Biology & Biochemistry
- Soil Science Society of America Journal
- Soil Use & Management
- Vadose Zone Journal

4 Substitution for Missing Crops

If during the calibration and validation process no sufficient data are available for a specific crop grown in the project, an alternative crop from the same (validated) CFG may be used as a substitution in both the baseline and with-project simulations. If an entire CFG is not validated, substitutions may be made that entail specific replacements be made for the baseline and with-project simulations. This method depends on the availability of alternative, conservative CFGs for a given practice category that meet all of the above criteria; without any alternatives no substitution can be made.

- Baseline:
  - Replace the missing CFG with an unfertilized perennial grass
- Project:
  - Replace the missing CFG with an alternative CFG that is demonstrably more conservative in generating credits than the missing CFG. Conservatism may be demonstrated with literature showing that crops in the replacement CFG sequester less carbon on average, and/or emit more trace gas emissions on average, than crops in the missing CFG.

5 Verification of Model Usage

Each verification team must include a person or persons who are expert in the particular biogeochemical model used to quantify emission reductions in that reporting period (if any). Guidance is provided in Sections 2-4 for requirements that models must meet to be eligible. Verifiers will be required to confirm the requirements of Sections 2-4 of this document are met.

Expert guidance is needed to ensure the given biogeochemical model is appropriately calibrated and validated for model performance and uncertainty for each reporting period. If the project employs the use of a third-party expert to undertake calibration, validation, and/or running a biogeochemical model in a given reporting period, then there will be no need for the verification team to independently verify such activities have been done appropriately, provided the verification team: confirms that the use of such third-party has been approved by the Reserve, that the party in question has the requisite expertise, that all requisite steps as set out in Section 2 of this document have been followed, and provided the expert provides the verification team with a sensitivity analysis regarding the requisite data inputs for the given model.

In other words, the verifier is simply required to confirm approval from the Reserve, confirm the qualification of the third-party, and confirm the requisite validation steps have been followed, but the verifier does not independently need to run the model themselves to confirm results appear reasonable. The verification team will still be required to confirm the reasonableness of all data input into the given biogeochemical model, following the requirements for baseline modeling in Section 3.4.1.1 of the SEP, and following expert guidance on the sensitivity of the given model to the requisite data inputs.

6 References

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*. https://doi.org/10.1006/jmps.1999.1279

Bruun, S., Christensen, B. T., Hansen, E. M., Magid, J., & Jensen, L. S. (2003). Calibration and validation of the soil organic matter dynamics of the Daisy model with data from the Askov long-term experiments. *Soil Biology and Biochemistry*, *35*(1), 67–76.

Climate Action Reserve. Expected adoption: June 10, 2020. Soil Enrichment Protocol. Available at http://www.climateactionreserve.org/how/protocols/soil-enrichment/.

Dietze, M. (2017). *Ecological Forecasting*. Princeton University Press

Efron, B., & Tibshirani, R. (1997). Improvements on Cross-Validation: The .632+ Bootstrap Method. *Journal of the American Statistical Association*. https://doi.org/10.2307/2965703

Food and Agriculture Organization of the United Nations FAO. (2019). Measuring and modelling soil carbon stocks and stock changes in livestock production systems. Available at http://www.fao.org/3/CA2934EN/ca2934en.pdf.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1975.10479865

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

Intergovernmental Panel on Climate Change. (2006). 2006 IPCC Guidelines for National Greenhouse Gas Inventories. Available at https://www.ipcc-nggip.iges.or.jp/public/2006gl/.

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 63(3), 425–464. https://doi.org/10.1111/1467-9868.00294

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*. https://doi.org/10.1037/h0072400

Liang, Y., Gollany, H. T., Rickman, R. W., Albrecht, S. L., Follett, R. F., Wilhelm, W. W., … Douglas, C. L. (2009). Simulating soil organic matter with CQESTR (v. 2.0): Model description and validation against long-term experiments across North America. *Ecological Modelling*, *220*(4), 568–581

Mosier, C. I. (1951). The need and means of cross validation. I. Problems and designs of cross-validation. *Educational and Psychological Measurement*.

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. *Handbook of Social Psychology*, 2, 80–203.

Mosteller, F., & Wallace, D. L. (1963). Inference in an Authorship Problem: A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed Federalist Papers. *Journal of the American Statistical Association*.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *American Society of Agricultural and Biological Engineers*, *50* (3), 885−900.

Mishurov, M., & Kiely, G. (2011). Gap-filling techniques for the annual sums of nitrous oxide fluxes. *Agricultural and Forest Meteorology*, *151*(12), 1763–1767.

Natural Resources Conservation Service. 2014 Farm Bill - Conservation Compliance Crop List. United States Department of Agriculture. Available at https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/programs/farmbill/?cid=stelprdb1262733.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In L. LIU & M. T. ÖZSU (Eds.), *Encyclopedia of Database Systems* (pp. 532–538). Springer US. https://doi.org/10.1007/978-0-387-39940-9_565

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography*, 40, 913–929. https://doi.org/10.1111/ecog.02881

Stone, M. (1976). Cross-Validatory Choice and Assessment of Statistical Predictions (With Discussion). *Journal of the Royal Statistical Society: Series B (Methodological).* https://doi.org/10.1111/j.2517-6161.1976.tb01573.x

Turner, P. A., Baker, J. M., Griffis, T. J., & Venterea, R. T. (2016). Impact of Kura Clover Living Mulch on Nitrous Oxide Emissions in a Corn-Soybean System. *Journal of Environmental Quality*, *45*(5), 1782–1787.

Wallach, D., Makowski, D., Wigington, J., Brun, F. (2014). Working with Dynamic Crop Models: Methods, Tools and Examples for Agriculture and Environment, second ed. Elsevier Science, Oxford, UK.

Wikle, C. K., & Berliner, L. M. (2007). A Bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena*, *230*(1–2), 1–16. https://doi.org/10.1016/j.physd.2006.09.017

Yang, J. M., Yang, J. Y., Liu, S., & Hoogenboom, G. (2014). An evaluation of the statistical methods for testing the performance of crop models with observed data. *Agricultural Systems*, *127*, 81‑89. Available at https://doi.org/10.1016/j.agsy.2014.01.008.

Yeluripati, J. B., van Oijen, M., Wattenbach, M., Neftel, A., Ammann, A., Parton, W. J., & Smith, P. (2009). Bayesian calibration as a tool for initialising the carbon pools of dynamic soil models. *Soil Biology and Biochemistry*, *41*(12), 2579–2583.