Summary of Revisions to SEP Model Guidance

August 25, 2020

This document summarizes proposed changes to the Soil Enrichment Protocol (SEP) Model Guidance, written and revised by M. Motew, C. Black, and N. Campbell of Indigo Ag, addressing comments received during external review from model experts at Dagan, as well as S. Wood, B. McConkey, K. Paustian, M. Easter, and Y. Zhang. Comments were received during in-person review on August 19, 2020 as well as via line-edit comments to a revision version from August 17, 2020 and emailed comments sent on dates listed below. Responses to emailed comments are at the bottom of this document.

Overview of Revisions

1. **Changed main document title to "Requirements and Guidance for Model Calibration, Validation, Uncertainty, and Verification".**
    - Change aimed to be more comprehensive of what is included. The title phrase 'requirements and guidance' is clearer than simply 'guidance', as this document sets standards for SEP model testing. 'uncertainty' was added because model uncertainty requirements presented in this document are technically challenging to achieve but were not obvious to the reader. The document was edited throughout to make model uncertainty requirements and guidance more prominent.
2. **We added a Definitions section at the top of the document as well as several paragraphs to the Introduction in order to more clearly explain the rationale for the standards that the SEP model requirements and guidelines impose.**
    - Statement describing how the design of the Validation Report supports independent 3rd party expert verification.
    - SEP model requirements and guidance were also described as falling into 2 categories; 1) standardized use of data, 2) standardized model testing. This highlights the data component of this document's standards, which is the most challenging initial 'lift' required to achieve SEP model requirements.
    - The SEP protocol allows any model to be used. These model standards, while technically challenging to achieve the first time for any model, will bring multiple models into the same assessment framework and allow for 1:1 comparisons when validating and verifying model appropriateness to issue credits.
3. **Section 4 on "Substitution for Missing Crops" has been revised to allow alternative crops from within the same CFG to be used in both baseline and with-project simulations.**
    - But, in the case of missing CFGs (I.e. incompletely validated), only a perennial grass may be substituted in baseline simulations, and a demonstrably more conservative CFG may be substituted in with-project simulations.
    - The previous version allowed an entirely different CFG to be substituted for a missing CFG based on similar attributes, but this is fundamentally at odds with the requirement to validate CFGs.
4. **Requirements 2 and 3 were combined into a single Requirement 2: Specific Dataset Requirements to Validate Model to more clearly scale from small to large projects, and additional edits made to clarify requirements.**

- Requirement 2 now indicates that: “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 3. This should be addressed explicitly in the Validation Report and will need to be approved by the Registry and by the external reviewer.”
    - **i.** This allows, at the discretion of the Registry and external reviewer, to make an exception for cases where generalized model performance can be demonstrated using the available data. For example, a validation dataset may have more soil types and clay contents required, span a wide geographic range, but only technically occur in 2 LRRs.

5. **Edits to clarify the use of IPCC climate zones in place of LRRs. Specifically**,
    - “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.”
6. **Added a new section (3.3.1. Special Rules for Practice Categories) which addresses:**
    - The difficulty in validating each CFG for the grazing PC due to a lack of data for grazing on crop residue, and lack of specificity in grazing studies to identify C3 and C4 species.
    - Alleviation of inevitable confusion about which CFGs would count when binning rotation-based studies according to CFG.
    - The addition of sulfur fertilizer for rice, and associated requirements
    - Clarity for handling blends of CFGs grown together.
7. **Change made to Table 3.1**
    - Added green manure to the Soil disturbance and/or residue management category in Table 3.1 since it is best associated with the plowing in of cover crop residue.
8. **Fixed a typo in the equation 3.4.2 for pooled measurement uncertainty.**
    - It previously had a sum in the numerator from j = i through k, but should have been from j = 1 through k.
9. **We updated Requirement 1 (for generalized validation dataset attributes)**
    - Removed the term “statistically robust” since it does not align with its true statistical connotation (i.e. robust to outliers).
    - New guidance for use of studies focused on “stacks” of practices
    - We removed reference to the Global Soils Partnership as a benchmark database because it does not approve datasets, only collates them. Instead we recommend that a 3rd party benchmark database can be used if approved by the Registry, and as long as exclusion criteria for records not used are included: “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.”
10. **To improve clarity on reporting requirements, we have added short bulleted lists summarizing the reporting requirements to the end of each section in the document.**
11. **Significant changes made to evaluation of bias, removing measure of conservatism and instead clarifying evaluation of goodness of fit, in response to external reviewers (see detailed responses below and in original revision comment responses).**

Dagan Questions and Comments
Provided by Pete Ingraham on 8/21/2020

Relevant documents provided by Indigo, 8/17/2020:
- CAR-SEP-Model-Guidance-Response-to-Dagan.docx
- rev_SEP Model Cal_Val_Ver Guidance_8_17_20.docx
- Synthesis of revisions SEP Model_cal_val8_17_20.docx

Comment:
On Pooled Measurement Uncertainty (Q1-3)
The response stated that "standard deviation is derived from replicates of the measurements" -- this language and Figure 3.4.1 suggests that standard deviation applies to the difference between two measurements (treatment 1, "baseline", and treatment 2, "changed practice") -- and, in fact, that is how PMU is calculated in Figure 3.4.2. I am fairly certain that Equation 3.4.2 is described incorrectly: standard deviation is still stated as of a study, however, I think it should read standard deviation of the difference between two observations AND *k* should be "observation differences" not studies. In addition, *n* is not explicitly stated as the number of practice changes, but should be. Here's a table that follows the calculations in 3.4.2 with the example where rows are *j* which must be practice changes and not studies...

Study	SD of a practice change	n (replicates)	n-1	SD^2 * (n-1)
1	3.40	4	3	34.68
1	2.20	8	7	33.88
2	2.90	4	3	25.23
3	6.00	2	1	36.00
3	3.20	5	4	40.96
		PMU:		**170.75**

Response:
Good catch! *k* should have been observations not studies, and this is now corrected in Eqn 3.4.2

Comment:
On standard deviation of a measurement:
Most studies do not report standard deviation (SD), the report standard error. We record standard error (SE) in our database. SD could be derived from SE if we knew the number of replicates in a measurement, but this is not always reported and, if it is, that is not recorded in our database. So, currently, the number of studies with reported SD in our calibration and validation datasets is zero -- it would be a non-trivial amount of work to make a data and database change, but could be done. Can you comment on whether the intent here was to use standard deviation or standard error?

When do you anticipate that you will have guidance on what to do for cases with missing standard deviation?

Response:
When standard error is available, it can be used (note that the calculation for standard error of a practice change can be rewritten as the sum of the standard errors of the treatments), but the difficulty is how to weight it when calculating PMU. We propose that:

- For studies where n is known, use it
- For studies that report SE but not N of individual observations, assume N is the same as the design of the experiment being reported
- For studies that report SE but do not report N at all, weight the SE as if N=2 when computing PMU.

Comment:
On table A3 response (Q11):
Response indicates that each row applies to a study ("In this case, 'n' refers to the number of practice change observations made…"). However, the response also says that "reported uncertainty would correspond to the measurement uncertainty of the observation." This seems to simultaneously say that this table's rows are per study and per observation. I guess really the point here is that there is necessary per-study information to report on and necessary per-observation data to report on and perhaps that should be done in two tables...

Validation scales:
In section 3.3 of the Guidance document, I am confused by the titles of requirement 2 and requirement 3. Both include the word validation in the title and then seem to differ by the scale of the validation. We understand validation to mean "the process of evaluating model performance relative to measured values." We have measured values at the field scale, we don't have measured values at a project scale. What are some examples where 2 vs 3 would apply?

Response:
This confused a number of other reviewers as well. In the latest revision we have merged the two requirements and clarified the differences that apply only to scale.

Comment:
Prediction intervals:
In section 3.5 of Guidance document, it says "measured versus modeled results should be compared for each crop functional group/practice category 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." Does this mean at least 90% of measured values from the validation specific to a crop functional group/practice category fall in a 90% prediction interval? Or is the 90% of measured value from the entire validation dataset? If one needs to check for each crop functional group/practice category, one might come across a group where there are for example 7 observations. If only one is missed, this is a percent of 6/7 or ~85% coverage which is not at least 90%.

Response:
The entire validation procedure applies within a CFG/PC combination, so the requirement is for 90% of the values compared to generate the error bounds in question. We have also added language clarifying that there is some room for flexibility (with Reserve approval needed) in cases where coverage is reasonable but slightly below 90%.

Brian McConkey Comments
Submitted 8/24/20

Comment:
Since there will be a comparison of the modeled estimate of SOC with project with the measured value the measured SOC values, the ability of the models to estimate SOC for one treatment is a critical validation. The guidance only requires the difference between two treatments but that is only important if one is only interested in modelling the difference, i.e. you can live with inaccuracy of modelling stock but not modelled difference. But for the protocol, inaccuracy of stock as important because only the one treatment is measured, and the modelled estimates will be trued-up to the measurements (albeit the true-up is conceptual only without an exact method). So this means that the validation will have to seek out data that have measured SOC for at least two times. The requirement that only studies with multiple treatments is not necessary as a single treatment with such time series is perfectly suitable This allows NEE measurements from CO_2 fluxes also be used, they sometimes only have one treatment.

In fact, many studies with SOC measurements only have the difference of ending stocks. It is still important to validate for difference, since that is important since the BAU SOC will be modelled only so the difference is important. However, validating only for difference would not be optimal.

So the SOC model must be validated for both ability to predict absolute SOC stocks and for the difference. If observations have only ending SOC stocks, they can only be used for validating the modeled difference as initial SOC used to model the difference is unknown. I suggest that if a time series, those data only used for validating absolute SOC stocks as that is most important and, frequently, more challenging the modelling the difference.

For N_2O and CH_4, there would be no pressing reason to validate the ability to model absolute emissions since both the with-project and BAU will be modelled. Nevertheless, it adds confidence if the model is validated that it models the absolute emissions for each treatment accurately.

Response:
We agree with the importance of accuracy in SOC stocks, in addition to accurately modeling SOC change. This is a central reason for the required use of measured SOC to begin with-project and baseline model simulations, so that the model is essentially forced to start in the correct place. Given the model is not required to simulate initial SOC stocks accurately, and instead must used measured SOC as an initial input, we feel that the focus of Model Validation on SOC change is reasonable. That said, we strongly agree with the importance of measure-remeasure SOC datasets, both from the standpoint of datasets to validate model performance as well as in the context of true-ups during an SEP project. Regarding the model validation component of this, we did not want to limit validation to these datasets alone, mainly from the practical standpoint that so few of them exist. We have been considering putting together a guidance document for expert verifiers of a Validation Report on ‘what to look for’ in datasets used for model validation.

Comment:
Documentation for verification report.

I am not sure if the guidance considered that initialization for a modeling run is included in "calibration". In practice there has to be a reproducible way to initialize the C stocks for the model run. In my experience, model initialization is more critical than parameter calibration. Therefore, it is important that the initialization for the validation runs is clearly specified and also important that the project owner documents in for verifier, that the initialization method used in the actual estimation for the project was validated.

Response:
We agree with this point. Responding to this comment and another, in section 3.3, first bullet, we added the requirement to declare model-specific needs for reported information to initialize and run the model accurately, as well as the process used to address missing information. In the summary box at the end of section 3.3. we added the requirement that this is included in the Validation Report, and that upon request the per-study use of data to initialize and run the model is made available.

Longer term we believe this is an area that can be further strengthened through, for example, a metric to evaluate the amount of data reported in a study compared to the amount of data required to initialize and run the model accurately. It would also be valuable to have standard approaches to fill missing information in this context. However, it is not clear yet how this is best achieved and we did not want to add an unnecessarily burdensome requirement. We believe reporting is a step forward, and will at least ensure transparency. This is suggested as an area for future development.

Comment:
Mean error, bias and conservativeness

I don't see any model goodness of fit testing that emphasized the mean error as an important indicator of fit. The problem is that if the model estimate is within the confidence limits of the observations, then the sign of the difference is meaningless. For the latter situation, you can't know if the model is actually over- or underestimating or over- or underestimating from the observed mean by chance alone. So it is dangerous to assume that a +ve and -ve difference have some cancelling out - such apparent cancelling out decreases the information on deviation, and that could be by chance alone. However, the fact that there are deviations is important. The mean absolute error, which is the mean of absolute value of each difference between model and observation is much more meaningful and is not prone to providing wrong information about deviations due to chance effects on the sign of the difference. Therefore, I strongly suggest using this rather than mean error, that is, incorrectly, called bias in the model guidance (bias would be a characteristic of the model, the measure of the mean error).

If the mean absolute error is within the pooled confidence limits of observations, then you have to conclude that the model estimates are good and no need to test further about conservativeness. However, if the mean absolute error falls outside the confidence limits, then obviously, the mean deviation is large enough there is concern about accuracy. . Inspection may show that the model does not work for some practices. Perhaps those practices then have to be excluded from the project. Or go to other models or use an ensemble of models or something. But as long as that MAE is within the confidence limit there is no valid test to determine if there is a bias. Considering the sign makes as much sense as claiming that particular dice are overestimating dice if their mean roll is larger than 7 and underestimating

dice if their mean roll is less than 7. My own experience is more as a measurer than a modeler and have learned to be adamant that it is a fool's game to try to interpret any differences within the confidence limits. Therefore, I can't see a valid scientific way to assess conservativeness and still have a validated model. If the mean error of modelled vs observed relevant treatment pairs is less than the lower confidence limit, there would be statistical justification to say that the model is underestimating the difference. But then it has to be more inaccurate at estimating SOC stocks of one of the two treatments in the pair - so the model can't be described as a validated model. So to have a validated model and conservativeness is an oxymoron.

Response:
Given this comment and others received during external review, we decided to remove conservativeness as a metric of model validation.

Instead we only focus on model goodness of fit, with accuracy of overall model behavior evaluated in the context of mean model bias relative to pooled measurement uncertainty. Model precision was also more clearly linked to requirements around model prediction error evaluation, and the design of linking model precision directly to the quantification of credits- I.e. an accurate model with low precision can be used to issue credits, but will penalize credit generation via high model predictive uncertainty.

In aggregate, we revised the model bias assessment to better meet the overall intention of giving an external reviewer the necessary tools to determine if model behavior is reasonable for crediting a PC/CFG/ES. A model is now be judged as valid if mean model bias is less than PMU, and model predictive uncertainty is determined as described in the Model Guidance. However, we do not want to penalize 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), so we made changes to improve the ability to diagnose problematic behavior. This includes adding a requirement for per-study model bias to be reported, ranked in order from largest to smallest, in order to allow an external expert to examine per-study model performance in the context of this aggregate measure of model accuracy. We also recognize that there may be special circumstances where a model may be performing reasonably 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.

In the current approach, large model biases will always mean large residuals. Therefore, in either direction (positive bias or negative bias) large bias will mean larger predictive uncertainty, and thus increase credit deductions. With the approach now described in the Model Guidance document, high model prediction error will be yielded in two circumstances- 1) through low precision of an accurate model or 2) high precision of an inaccurate model. We feel this is a reasonable approach for the purposes of an SEP project.

We evaluated the use of MAE but did not feel that it was as informative as the use of bias as now applied, as, for example, bias directionality can support diagnosis of behavioral patterns. We also did not feel it had clear advantage as an overall metric of model performance, given that the current approach is explicit evaluating overall accuracy, and penalizes low model precision through credit uncertainty deductions.