Comments: Address “accuracy” and “scalability” of process-based modeling in the Soil Enrichment Protocol by Climate Action Reserve

(a joint comment from scientists from University of Illinois, University of Minnesota, and Lawrence Berkeley National Lab)

Authors: Kaiyu Guan[1,2], Zhenong Jin[3], Evan DeLucia[4,5,6], Bin Peng[1,2], Jinyun Tang[7], Andrew Margenot[1], Wang Zhou[1], Ziqi Qin[1], DoKyoung Lee[1], and Yuxin Wu[7]

[1]College of Agricultural, Consumer and Environmental Sciences, University of Illinois, Urbana, IL, USA
[2]National Center for Supercomputing Applications, University of Illinois, Urbana, IL, USA
[3]Department of Bioproducts and Biosystems Engineering, University of Minnesota, St. Paul, MN, USA
[4]Department of Plant Biology, University of Illinois, Urbana, IL, USA
[5]Institute for Sustainability, Energy, and Environment, University of Illinois, Urbana, IL, USA
[6]Carl R. Woese Institute for Genomic Biology, University of Illinois, Urbana, IL, USA
[7]Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Summary:

We appreciate the idea of using a market-driven approach to sequestrate soil carbon, which could benefit both farmers and the environment. The protocol by Climate Action Reserve (CAR) promotes a hybrid approach combining soil sampling and other *in-situ* measurements with “process-based modeling” to quantify field-level soil carbon credit. We agree with the general potential of this approach. **However, in the current form, the CAR protocol did not satisfactorily address how to rigorously validate models to ensure model accuracy and achieve scalability**. Especially when a model’s result will be used to trade for credits, no one would trust a model without robust and consistent performance.

We point out that **model accuracy, characterized by “model uncertainty”, plays the most essential role here, as it directly relates to the final estimated carbon credit from a model**. The only way to quantify uncertainty of a model is through model validation. We further highlight that **model validation is the only judgement criteria of a model’s merit**. Any model to be used in a carbon credit system should publicly report its “uncertainty” in a reproducible format, which is derived from its validation performance benchmarked with a high-quality ground truth dataset following the standard model validation protocol. **No exemption should be permitted for any model, even if it is widely used, peer-reviewed, or has a long history.** To enable such objective assessment, we strongly recommend developing and compiling an **open-source and high-quality dataset through community efforts to make the model validation results transparent and intercomparable**.

We further identify a major missing point in this CAR protocol - **model scalability**. **In the current context, a method that works well at one or a few demonstration sites is not enough; the consistent performance with the accepted “uncertainty” is also required when applying to randomly selected sites.** The current protocol has no discussion regarding how to ensure the model scalability. **Again, instead of based on a model’s history or reputation, we**

should design an appropriate protocol of model validation to address this requirement to test "model scalability". We provide a detailed pathway of how to conduct model validation (**Figure 1**), including to develop a two-tier validation system, and use community effort to develop open-source data to enable objective model validation, in particular, to test model's performance at many random fields (Tier 2 sites), which is the key metrics to determine the extent of model scalability.

In this comment, we first provide detailed rationale and reasonings in **Section A**. We then explicitly identify where the CAR protocol falls short, followed by our suggested revision, **in Section B**. **We aim to bring in unbiased, science-based recommendations to ensure the best practices in modeling soil carbon stock and changes, also hoping to help this protocol to have a just and long-lasting value**.

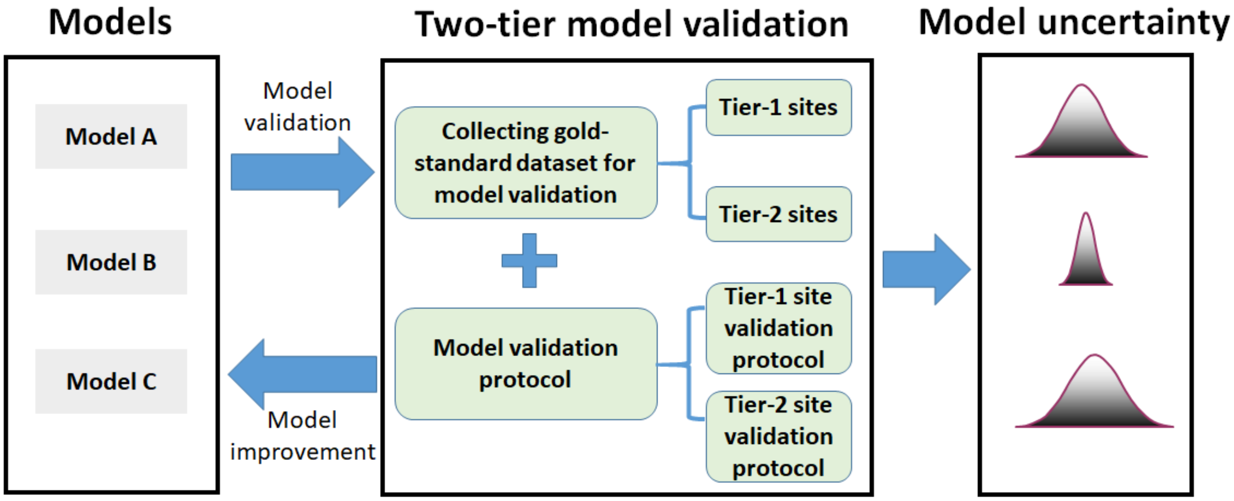
Models
Model A
Model B
Model C
Model validation
Model improvement
Two-tier model validation
Collecting gold-standard dataset for model validation
Tier-1 sites
Tier-2 sites
Model validation protocol
Tier-1 site validation protocol
Tier-2 site validation protocol
Model uncertainty

Figure 1. Conceptual diagram to illustrate our comments for how to **rigorously validate models to ensure model accuracy and achieve scalability.** The two-tier model validation approach can be found in detail in Section A.

Section A: Rationale and Reasoning

We appreciate and applaud the effort of using market-driven approaches to improve environmental sustainability; in this case, we refer to quantifying soil carbon sequestration from various management practices for “additionality” with certain tenure, on which a carbon credit market will be created. This market-driven approach is supposed to provide farmers incentives to adopt conservation practices, and to include farmers as a part of the solution for improving environmental sustainability and mitigating climate change. This effort has good intentions.

However, we would like to point out that whether this market-driven approach can be successful or not, as proposed by the Climate Action Reserve (CAR), is deeply dependent on the **accuracy** and **scalability** of the methods to quantify field-scale soil carbon stock and its change. Especially when a model’s result will be used to trade for credits, no one would trust a model without robust and consistent performance. **These two criteria are not fully addressed in the current protocol, constraining its feasibility and long-term value.** Let’s first define what we mean by “accuracy” and “scalability” in this context.

“**Accuracy**” of a method is characterized by **uncertainty that a method has in quantifying field-level soil carbon credit**. In layman’s language, one has greater trust in a method that has a smaller uncertainty, and less confidence in a value produced from a method with higher uncertainty. **Uncertainty level is used to directly discount the calculated credit, and thus knowing uncertainty of a method is very critical.** There is a well-established statistical framework to illustrate this concept, which is reported in the recent DOE ARPA-E SMARTFARM document (DOE ARPA-E SMARTFARM Program, 2020), of using the measurement/model uncertainty to discount the estimated credits. The standard deviation of a measurement system was used as an example to discount the value of credits generated in this framework (**Figure 2**). For example, if a field that has 15 g CO_2e/MJ carbon emission reduction credit measured with three different methods that have a standard deviation of 1, 3, and 10 g CO_2e/MJ, the estimated credit for the 15 g reduction would be discounted (adjusting the value to the 90th percentile) to a 14, 10, or 2 g CO_2e/MJ for the three methods, respectively. Thus, **any method should clearly report its uncertainty before putting into use, as the reported uncertainty is needed to assess the final carbon credit for any field**. Such quantification of uncertainty should be done at the same spatial scale as the intended application. Specifically, if the field level is where soil carbon credit is quantified, **uncertainty of any suggested method should be reported at the field level.**

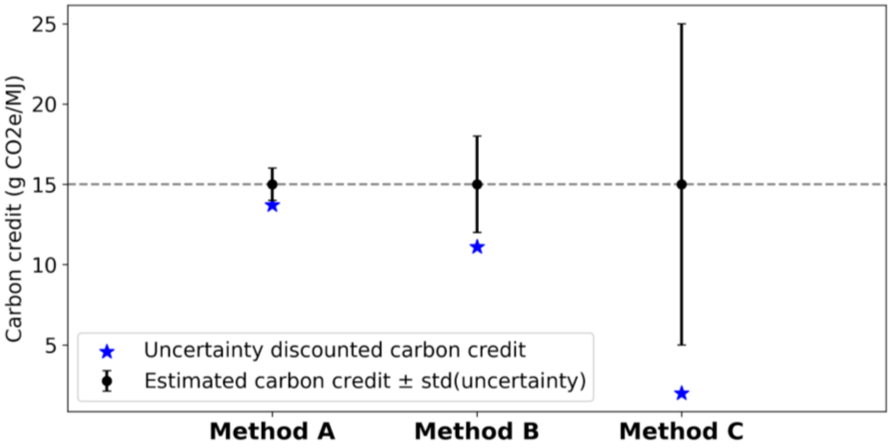

Figure 2. An example of carbon credit discounted based on the measurement uncertainty (DOE ARPA-E SMARTFARM, 2020).

"**Scalability**" here refers to **maintain the minimum accepted "accuracy" performance to quantify field-level soil carbon credit across all possible fields** with sufficiently low cost**; in other words, a method that works well at one or a few demonstration sites is not enough; the consistent performance with the accepted "uncertainty" is also required when applying to randomly selected sites.** For example, intensive soil sampling offers high "accuracy" but is cost prohibitive, and thus lacks "scalability". The CAR protocol promotes a combination of soil sampling plus "process-based modeling" to quantify field-level soil carbon credit. We agree with the general potential of this hybrid approach combining soil sampling and other *in-situ* measurements with "process-based modeling" as a possible solution, and it is probably the only viable solution in the near term, given that sensor technology for SOC may still take a long time to be robust and commercially available. **However, "accuracy" and "scalability" for "process-based modeling" lack necessary details or largely missing in the CAR protocol.** Below, we will elaborate two key points to address the above issue.

1. Model validation is the only criteria by which a model's merit can be evaluated

As we claim above, **any method should clearly report its uncertainty before its operational use**; the same applies to any proposed process-based models. Most importantly, we emphasize that **model validation, a procedure to benchmark model simulation with independent and high-quality observational data, is the only way to quantify model uncertainty**. A reliable protocol for field-level soil carbon sequestration should include the following two aspects:

(i) Which model variables should be validated?

(ii) What qualifies as benchmark ground truth data for validation?

For (i), we believe the soil carbon quantification in this protocol requires quantification of both carbon pools and fluxes of the agroecosystem at a field scale. **Table 1** provides a minimum list of the carbon related variables for this purpose, as well as a high-level list. Inclusion of variables from the high-level list, in more ideal cases could include measurements of soil microbial activities and biogeochemical transformation rates, for validation. The high-level list is recommended because they allow models to get the right answer for the right reasons; otherwise, good model performance may be achieved because of over-fitting at validation sites, limiting spatial scalability. But the minimum variable list provides the basic requirements.

Table 1. List of the carbon related variables that should be validated for soil carbon quantification.

Variable	Physical meaning	Measuring technique or data sources
Minimum variable list for model validation		
NPP	Net Primary Productivity is the rate at which energy is stored as biomass by plants. it is equal to the difference between Gross Primary Productivity (GPP) and autotrophic respiration (Ra).	Biomass destructive sampling, or remote sensing
Crop yield	Harvested grain yield	Grain harvesting (from farmer self-report or harvester machine), or remote sensing
SOC stock	The absolute value of soil organic carbon stock	Soil core sampling, or hyperspectral remote sensing
SOC change	The changes of soil organic carbon stock after adopting certain management practices for a period of time	Soil core sampling, or remote sensing
The impacts of management practices on the above terms	The responses of the above factors to different management practices for a given period of time	Field experiments under different management practices
High-level or advanced variable list for model validation		
NEE	Net ecosystem exchange is the net exchange of carbon between an ecosystem and the atmosphere (per unit ground area)	Eddy-covariance flux towers
GPP	Gross primary productivity is the rate at which solar energy is captured by photosynthesis (energy captured per unit area per unit time)	Estimated from eddy-covariance flux measurements, or remote sensing
Ra	Autotrophic respiration is the total amount of organic carbon that is respired (oxidized to CO_2) by plants per unit time	Stable or radioactive isotope labelling
Rh	Heterotrophic respiration refers to the carbon lost by organisms in ecosystems other than the plants	Soil flux chambers, or eddy-covariance flux towers
Soil moisture	Soil water content of soil in different layers	Point sensor measurement, or cosmic ray neutron sensing
Soil temperature	Physical temperature of soil in different layers	Point sensor measurement

For (ii) about model validation benchmark data, **to make the model validation results transparent and intercomparable, high-quality observational dataset should be compiled through community efforts**. This dataset should ensure site representativeness to include different environmental conditions (e.g. climate, soil properties) and management practices (e.g. different tillage practices, cover crop uses), all at the field level. We should use this standard benchmark data and the same protocol to evaluate different models, and this derived uncertainty metrics should be reported. **Thus, instead of debating which model is "better" or "worse", the most objective solution is to validate a model's simulation performance based on the benchmark data. Using this objective way to benchmark different models enables new models to join the available model list, and also motivate them to improve existing models.**

Any model to be used in a carbon credit system should publicly report its "accuracy" in a reproducible format, which is derived from its validation performance benchmarked with a high-quality ground truth dataset following the standard model validation protocol. **No exemption is available for any model, even if it is widely used, peer-reviewed or developed by a reputable group or institute.**

It is worth noting that there have been several model intercomparison (MIP) efforts in the research communities for climate models (CMIP) (Eyring et al. 2016) and crop models (AgMIP) (Rosenzweig et al. 2013), which set guiding examples for agroecosystem or soil biogeochemistry modeling efforts in agricultural carbon sequestration programs. It is also worth noting that the new SMARTFARM program by DOE ARAP-E is developing such a gold-standard and open-source data for benchmarking field-level soil carbon change and GHG emissions (DOE ARPA-E SMARTFARM Program 2020).

2. Two-tier model validation is needed to demonstrate model scalability

Once we all agree on the importance of model validation, we need to understand how to conduct model validation to test whether a **model has scalability**. As stated before, 'scalability' in this context is defined as the **ability of a model to perform robustly with accepted accuracy on all targeted fields**. All models have their own structures and parameters, and almost all credible models have been calibrated and validated at their specific study sites, which usually have collected rich measurements of different variables; however, generally no validations are conducted to test the model performance at random sites that only have limited measurements. **"Model scalability"** should not only be demonstrated by model performance at data-rich sites, where parameter calibration is allowed; a true test of "model scalability" should be also demonstrated at many random sites, where only limited measurements are available to truly test the transferability of the model. The latter is what a real-world application entails - we need to quantify soil carbon credit at any given field. **Only models that can reproduce the accepted 'accuracy' extent at any random fields can be used as an accepted method for a carbon credit system.**

Before we illustrate what a validation system should look like, we should understand the basic principle of validation - the validation data should always be split into two independent parts: **in-sample and out-of-sample data** (Hastie, Tibshirani, and Friedman 2009). In-sample validation allows one to calibrate and train a model with observations; then the calibrated model should be tested against the "out-of-sample" data. Whereas "in-sample" performance informs where improvements may be possible, we should note that **"out-of-sample" model performance is what counts** - this basic principle has been frequently violated in the literature, as many work mixed "in-sample" and "out-of-sample" model performance (**Figure 3**).

To achieve the above goal to fully validate the **model scalability**, we believe a two-tier validation approach is needed (**Figure 3**). Both tiers of validation data and the usage protocol should be prepared and used by the broader community, and these results should be reported to the community for fair and transparent comparison.

Tier 1: This tier includes sites that have collected a complete suite of measurements data, and can be regarded as gold-standard sites. Example includes ARPA-E SMARTFARM Phase 1 sites (https://arpa-e.energy.gov/?q=arpa-e-programs/smartfarm), USDA LTAR sites (https://ltar.ars.usda.gov/, e.g. Kellogg Biological Station), some NEON sites (https://preview.neonscience.org/), and AmeriFlux sites at cropland and pasture land (https://ameriflux.lbl.gov/). Tier 1 sites enable detailed model calibration and out-of-sample validation. Usually, for the Tier 1 sites, they have the ability to measure whole ecosystem flux (e.g. NEE, GPP), soil carbon flux and stock, plant biomass etc.

Tier 2: This tier includes an extensive number of sites to test the model scalability performance. These sites in general only have limited amounts of ground measurements (for example, most sites may only have reported crop yield and SOC), but they represent the real-world situation for operational use. Validation can be directly made to compare the simulated crop yield, SOC stock and changes with observations. When doing model validation over tier 2 sites, only basic information about site location and management history will be provided, and the modeling team should report their simulation results for independent comparison with observations.

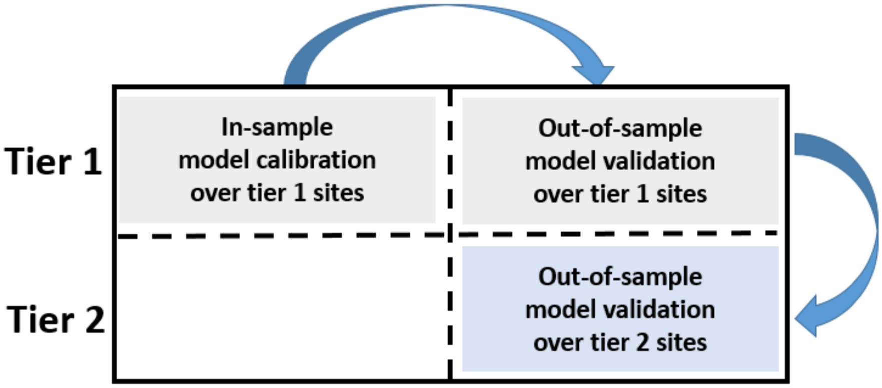
Tier 1
Tier 2
In-sample model calibration over tier 1 sites
Out-of-sample model validation over tier 1 sites
Out-of-sample model validation over tier 2 sites

Figure 3. A conceptual illustration of the two-tier model validation approach.

Section B: Suggested Revisions

Based on our reasoning in Section A, we identified the following three major parts that the current CAR protocol is required to revise. **The Red color provides our suggested revision.**

1. Model validation is the only criteria by which a model's merit can be evaluated.

The requirements for the model described at P73 - 74 in "*Soil Enrichment Protocol"* is **far from sufficient.** The quoted text provides the criteria of the CAR "*Soil Enrichment Protocol"*:

"Models used to estimate stock change/emissions may be empirical or process-based, and must meet the following conditions:

1. Publicly available;
2. Shown in at least one peer-reviewed study to successfully simulate changes in soil organic carbon and, where modelling is used for non-reversible emissions impacts, trace gas emissions resulting from changes in agricultural management included in the project description;
3. Able to support repeating the project model simulations. This includes clear versioning of the model use in the project, stable software support of that version, as well as fully reported sources and values for all parameters used with the project version of the model. In the case where multiple sets of parameter values are used in the project, full reporting includes clearly identifying the sources of varying parameter sets as well as how they were applied to estimate stock change/emissions in the project. Acceptable sources include peer-reviewed literature and appropriate expert groups, and must describe the data sets and statistical processes used to set parameter values (i.e., the parameterization or calibration procedure, see guidance described in 5);
4. Incorporate one or more input variables that are monitored ex-post;
5. Validated according to the guidance contained in the external document titled Model Calibration, Validation, and Verification Guidance for Soil Enrichment Projects, using the same parameters or sets of parameters applied to estimate SOC/trace gas emissions in the project."

We think the above conditions could not enable fair and transparent model validation and model uncertainty quantification. We point out that **model accuracy, characterized by "model uncertainty", plays the most essential role here, as it directly relates to the final estimated carbon credit from a model**. The only way to quantify uncertainty of a model is through model validation.

We further highlight that **model validation is the only judgement criteria of a model's merit**. Any model to be used in a carbon credit system should publicly report its "uncertainty" in a reproducible format, which is derived from its validation performance benchmarked with a high-quality ground truth dataset following the standard model validation protocol. **No exemption**

should be permitted for any model, even if it is widely used, peer-reviewed, or has a long history.

To enable such objective assessment, we strongly recommend developing and compiling an **open-source and high-quality dataset through community efforts to make the model validation results transparent and intercomparable**.

2. Validation data needs to be the same standard data for all the models, to ensure apple-to-apple comparison.

Requirements for validation dataset described P8 - 10 in "*Model Calibration, Validation, and Verification Guidance For Soil Enrichment Projects"* did not require the same standard data for model validations.

"Measured datasets must be drawn from peer-reviewed and published experimental datasets with measurements of SOC stock change (and annual measures of N2O and CH4 change if applicable) using control plots to test the practice category. All dataset sources must be reported.

Project developers are expected to use a process for selecting data for model validation that results in the assembly of validation 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, N2O and CH4) for each project domain combination of practice category and crop functional group, and include a histogram showing the range of validation data values."

As we claim above, "**any method should clearly report its uncertainty before its operational use**"; the same applies to any proposed process-based models. Most importantly, we emphasize that **model validation, a procedure to benchmark model simulation with independent and high-quality observational data, is the only way to quantify model uncertainty**. A reliable protocol for field-level soil carbon sequestration should include the following two aspects:

(i) Which model variables should be validated?

(ii) What qualifies as benchmark ground truth data for validation?

For (i), we believe the soil carbon quantification in this protocol requires quantification of both carbon pools and fluxes of the agroecosystem at a field scale. **Table 1** provides a minimum list of the carbon related variables for this purpose, as well as a high-level list.

For (ii) about model validation benchmark data, **to make the model validation results transparent and intercomparable, high-quality observational dataset should be compiled**

through community efforts. This dataset should ensure site representativeness to include different environmental conditions (e.g. climate, soil properties) and management practices (e.g. different tillage practices, cover crop uses), all at the field level. We should use this standard benchmark data and the same protocol to evaluate different models, and this derived uncertainty metrics should be reported. **Thus, instead of debating which model is "better" or "worse", the most objective solution is to validate a model's simulation performance based on the benchmark data. Using this objective way to benchmark different models enables new models to join the available model list, and also motivate them to improve existing models.**

Any model to be used in a carbon credit system should publicly report its "accuracy" in a reproducible format, which is derived from its validation performance benchmarked with a high-quality ground truth dataset following the standard model validation protocol. **No exemption is available for any model, even if it is widely used, peer-reviewed or developed by a reputable group or institute.**

It is worth noting that there have been several model intercomparison (MIP) efforts in the research communities for climate models (CMIP) (Eyring et al. 2016) and crop models (AgMIP) (Rosenzweig et al. 2013), which set guiding examples for agroecosystem or soil biogeochemistry modeling efforts in agricultural carbon sequestration programs. It is also worth noting that the new SMARTFARM program by DOE ARAP-E is developing such a gold-standard and open-source data for benchmarking field-level soil carbon change and GHG emissions (DOE ARPA-E SMARTFARM Program 2020).

3. Model scalability should be addressed through the two-tier validation approach.

Requirements for model validation for the entire Project Domine is described at P10 in "*Model Calibration, Validation, and Verification Guidance For Soil Enrichment Projects*". **The requirements below only show that a model could work at a few sites, which is not enough to show the model's capacity of scalability. The consistent performance with the accepted "accuracy" is required for any randomly selected site.**

"Requirement 3: Validating a practice category / crop functional group combination for the entire Project Domain can only be completed if there are measurements of SOC stock and annual N2O and CH4 flux change (if applicable) that in total cover:

▪ At least three declared LRRs for projects within the US (or two IPCC climate zones per each required LRR for projects outside of the US)

▪ At least three declared soil textural classes

▪ A range in declared clay amount per unit of soil spanning at least 15 percentage points"

We identify a major missing point in this CAR protocol - **model scalability**. **In the current context, a method that works well at one or a few demonstration sites is not enough; the consistent performance with the accepted "uncertainty" is also required when applying to randomly selected sites.** The current protocol has no discussion regarding how to ensure the model scalability. **Again, instead of based on a model's history or reputation, we should design an appropriate protocol of model validation to address this requirement to test "model scalability".** We provide a detailed pathway of how to conduct model validation (**Figure 1**), including to develop a two-tier validation system, and use community effort to develop open-source data to enable objective model validation, in particular, to test model's performance at random fields, which is the key metrics to determine the extent of model scalability.

Specifically, "model scalability" should not only be demonstrated by model performance at data-rich sites, where parameter calibration is allowed; a true test of "model scalability" should be also demonstrated at many random sites, where only limited measurements are available to truly test the transferability of the model. The latter is what a real-world application entails - we need to quantify soil carbon credit at any given field. **Only models that can reproduce the accepted 'accuracy' extent at any random fields can be used as an accepted method for a carbon credit system.**

Before we illustrate what a validation system should look like, we should understand the basic principle of validation - the validation data should always be split into two independent parts: in-sample and out-of-sample data (Hastie et al. 2009). In-sample validation allows one to calibrate and train a model with observations; then the calibrated model should be tested against the "out-of-sample" data. Whereas "in-sample" performance informs where improvements may be possible, we should note that **"out-of-sample" model performance is what counts** - this basic principle has been frequently violated in the literature, as many work mixed "in-sample" and "out-of-sample" model performance (**Figure 3**).

To achieve the above goal to fully validate the **model scalability**, we believe a two-tier validation approach is needed (**Figure 3**). Both tiers of validation data and the usage protocol should be prepared and used by the community, and these results should be reported to the community for fair and transparent comparison.

Tier 1: This tier includes sites that have collected a complete suite of measurements data, and can be regarded as gold-standard sites. Example includes ARPA-E SMARTFARM Phase 1 sites (https://arpa-e.energy.gov/?q=arpa-e-programs/smartfarm), USDA LTAR sites (https://ltar.ars.usda.gov/, e.g. Kellogg Biological Station), NEON sites (https://preview.neonscience.org/), and some AmeriFlux sites at cropland and pasture land (https://ameriflux.lbl.gov/). Tier 1 sites enable detailed model calibration and out-of-sample validation. Usually, for the Tier 1 sites, it has the ability to measure whole ecosystem flux (e.g. NEE, GPP), soil carbon flux and stock, plant biomass etc.

Tier 2: This tier includes an extensive number of sites to test the model scalability performance. These sites in general only have limited amounts of ground measurements (for example, most sites may only have reported crop yield and SOC), but they represent the real-world situation for operational use. Validation can be directly made to compare the simulated crop yield, SOC stock and changes with observations. When doing model validation over tier 2 sites, only basic information about site location and management history will be provided, and the modeling team should report their simulation results for independent comparison with observations.

References:

DOE ARPA-E SMARTFARM Program. 2020. “SYSTEMS FOR MONITORING AND ANALYTICS FOR RENEWABLE TRANSPORTATION FUELS FROM AGRICULTURAL RESOURCES AND MANAGEMENT (SMARTFARM).” *Https://arpa-E-foa.energy.gov/FileContent.aspx?FileID=e60d7816-0cad-4585-9971-310ffa278590*, 16–18.

Eyring, Veronika, Sandrine Bony, Gerald A. Meehl, Catherine A. Senior, Bjorn Stevens, Ronald J. Stouffer, and Karl E. Taylor. 2016. “Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) Experimental Design and Organization.” *Geoscientific Model Development* 9 (5): 1937–58.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media.

Rosenzweig, C., J. W. Jones, J. L. Hatfield, A. C. Ruane, K. J. Boote, P. Thorburn, J. M. Antle, et al. 2013. “The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and Pilot Studies.” *Agricultural and Forest Meteorology* 170 (March): 166–82.