Reviewer #1

The study by Gautam et al. presents an ensemble framework based on three established, process-based, agroecosystem models. This is used to simulate yield of corn and soybean and SOC across USA over a relatively long period. The study makes use of large datasets for model calibration and clearly shows the improvement provided by the ensemble compared to the use of each single model, while discussing the results not only in terms of model reliability but also introducing environmental issues and policies implications. Also, the manuscript is rather concise and provides, for the most part, precise information. I think that this work is relevant for SOIL and its readers, and it fits the special issue "Advances in dynamic soil modelling across scales" well.

Author's response: We thank the reviewer for the thorough and constructive evaluation of our manuscript. We appreciate the positive assessment of the relevance, clarity, and contribution of the ensemble modeling framework. We have carefully addressed all comments and believe the revisions have significantly improved the clarity, robustness, and impact of the manuscript.

However, I have some comments, especially about how the results are shown/discussed and how certain datasets are described. I am sure that the authors can address my comments and I think taht this will not modify the main findings in a substantial way. But I believe that this will provide a stronger manuscript that is clearer for the readers of SOIL and will have more impact. Please find below my main concerns, followed by minor points, and by more detail comments.

Main points:

- Lines 23-26 Page 6 state: "Each of the three models was validated" and "Model parameterization was conducted to minimize the root mean square error (RMSE)". If I understand correctly, minimization of RMSE for yield and SOC was performed. I think this is rather calibration and not validation as, if I again understand correctly, all available data were used to minimize RMSE, leaving no independent dataset for model validation. This paragraph is rather short, and I may have misinterpret it. I thus suggest that the authors revise the text and the wording. Also, model validation through 80/20 or split sampling would be a nice addition to the manuscript, but it is true that validation is not mentioned elsewhere in the current text, so it may not be one of the objectives of this work. Alternatively, to this, I would include additional metrics like Nash-Sutcliffe Efficiency and R2, and discuss the results and the possibly different outcomes of these metrics. Also, the RMSE and respective measurement units are consistently provided throughout the manuscript, which is well done. But I think that the addition of the RMSE in % of the maximum value (or an alternative strategy that provides the same benefit) would help the reader in better understanding the value of these RMSE values. At least in some key points in which results are discussed.

Author's response: We thank the reviewer for pointing out this lack of clarity. We completely agree that the original wording was ambiguous, and we have thoroughly revised Section 2.3 to explicitly define our calibration, metrics selection, and data-handling workflows.

Specifically, we have updated the text from "validation" to "calibration" to accurately reflect that parameter tuning minimized error across the target datasets. To address the underlying concern regarding model overfitting, we tested a rigorous 50/50 stratified data-splitting protocol. Because the calibration and independent validation subsets yielded nearly identical error metrics (e.g., Corn yield RMSE of 3.0 Mg/ha for both subsets), we verified that the framework is structurally stable and robust against overfitting, justifying pooling the datasets to preserve geographic continuity across CONUS. Furthermore, we have embraced the suggestion to include a relative error metric by integrating Normalized RMSE (%RMSE) alongside absolute RMSE. We intentionally omitted R^2 and NSE because, in broad regional applications, these metrics can be misleadingly inflated by the strong regional observation mean and fail to accurately characterize localized structural errors. The manuscript text has been modified into a unified, transparent paragraph to reflect these improvements. Line 265-286

"Each of the three models was calibrated using county level crop yield data from the National Agricultural Statistics Service (NASS) (NASS, 2025) and site-level SOC data from the Rapid Carbon Assessment (RaCA) dataset (Wills et al., 2014) across CONUS. Model parameterization was optimized to minimize the root mean square error (RMSE) and normalized RMSE (%RMSE). Relative metrics like the Nash-Sutcliffe Efficiency (NSE) and R2 were intentionally excluded, as they can be misleadingly inflated in large-scale regional contexts where model predictions closely align with a strong regional observation mean.

To explicitly evaluate the risk of model overfitting and assess framework stability prior to final execution, a stratified data-splitting protocol was implemented using the createDataPartition framework in the caret package. Continuous datasets (RaCA SOC and NASS yields) were partitioned into equal-sized numeric bins based on quantiles, and 50% of the data points were randomly sampled for calibration, leaving the remaining 50% as an independent validation subset. This cross-validation protocol yielded identical error distributions between the independent groups specifically, an Ensemble SOC RMSE of 4.7 kg/m2 (calibration) vs. 4.6 kg/m2 (validation), a corn yield RMSE of 3.0 Mg/ha (calibration) vs. 3.0 Mg/ha (validation), and a soybean yield RMSE of 1.0 Mg/ha (calibration) vs. 1.0 Mg/ha (validation). Because this identical performance demonstrated that the modeling framework was structurally stable and robust against overfitting, the data subsets were pooled for the final simulations. This approach maximizes geographic continuity across the CONUS domain and fully utilizes the available high-resolution datasets. This approach provides a robust upper bound of current process-based modeling capabilities at a 4 km2 scale, representing the collective central tendency of diverse mechanistic assumptions."

I am unfortunately not very familiar with the datasets that were used to calibrate the models and calculate RMSE values. I think that the current text does not provide the reader with sufficient information about the yield and SOC datasets mentioned at section 2.3. For example, the approximate number of points, their distribution, and their spatial patterns across the modelled domain are not clear. This has also repercussions on the readability of figures 4, 5, and 6 and on those parts of text (e.g., P08L12-14) where th espatial distribution of "more coherent and robust estimates of the ensemble across diverse agroclimatic zones" are mentioned. It is true that the text discusses some key local spatial patterns, like in the Mid-West corn belt and the Mississippi river basin. But the reader cannot currently see the spatial match between the models and the yield/SOC datasets, and I think a spatial understanding of the model performance is important given the large-scale of the study and the conclusions that are provided. I do not know what the best way is to show the spatial patterns of the measurements and the agreement between models and calibration. It is possible that maps showing measured values with the same 4km scale would not be sufficiently readable or informative. An alternative/addition could be to show maps of the RMSE distribution, maybe only for the ensemble results. I am however sure that the authors can find a suitable way to address this point.

Author's response: The manuscript has been revised to include additional details on the yield (Figure S1) and soil organic carbon (SOC) datasets (Figure S2), including their spatial distributions across the conterminous United States (CONUS). To improve spatial interpretability, we have incorporated additional analyses to evaluate model performance (%RMSE). Specifically, maps illustrating the spatial distribution of root mean square error (RMSE) for the ensemble model have been added. The Results and Discussion sections have been updated accordingly to reflect these enhancements.

- In my view, the colours in figures 4, 5, and 6 are not intuitive. I agree with the use of colour classes instead of a continuous colour ramp, and with the use of one colour scale for yield (figures 4 and 5) and one for SOC (Figure 6). Also, current scales are readable for colour blind readers, which is nice indeed. But I find the current scales difficult to read. For example, 3-5 Mg/ha and 11-13 MG/ha have similar colours in figures 4 and 5, although they represent very different values. Same applies to 0-0.2 and 1-1.2 SOC intervals. I would suggest ramping the scale with more continuous colours.

Author's response: The color scales in Figures 4-6 have been revised to use more continuous and perceptually uniform gradients. These changes improve the differentiation between value ranges while maintaining accessibility for colorblind readers, enhancing overall figure readability.

- Data availability: I am not sure that all data are contained within the article, as stated in the data availablility statement. While one can obtain the datasets used for yield and SOC, other data and results like simulation results cannot be obtained from the text or from citations, if I did not overlook something in the text. Copernicus journals require that data that correspond to journal articles are deposited in reliable (public) data repositories. I would argue that simulations results, at least, should be made available in a readable format. However, it is sometimes the case that these datasets involve also codes for results (e.g., to get plots and statistics from simulations) and materials for figures. Please check the Copernicus data policy for this.

Author's response: The Data Availability section has been updated to clarify the availability of datasets used and generated in this study. Simulation outputs and supporting data have been made available through a DRYAD and the corresponding link/DOI (http://datadryad.org/share/LINK_NOT_FOR_PUBLICATION/NcqhUbKeeSv3GIP0ea3KUkvlLjEAjGVJDKAFaJPs43o) has been included in the revised manuscript to ensure compliance with journal data policies.

Minor points:

1. The spatial resolution was set to 4 km. While it is not my intent to challenge this choice, I believe the readers would benefit from a brief explanation of the reasons behind it. Being it computation time, spatial distribution of calibration data, or policy-related, I think this is an interesting detail.

Author's response: A clarification has been added to explain the rationale for selecting the 4 km spatial resolution. This resolution reflects a balance between computational feasibility and the need to represent large-scale spatial variability across the study domain. L183-188

"We selected a 4 km spatial resolution to balance between computational efficiency and the need to adequately capture large-scale spatial variability across the CONUS study domain. This resolution is appropriate for regional to national-scale assessments of crop productivity and soil carbon dynamics, including broad carbon accounting applications. However, it may not fully capture fine-scale heterogeneity in soil properties, management practices, and microclimatic conditions that influence carbon fluxes at the field level."

2. I am not good with names, and I had troubles pinpointing states in the US map. While recognising states on the US map is trivial for some, I think that labelling specific states that are mentioned in the text would be useful.

Author's response: We appreciate the reviewer's suggestion. While labeling specific states can aid geographic interpretation, adding state names across CONUS-scale maps substantially reduces visual clarity and leads to overcrowding, particularly given the high spatial resolution of the data. To maintain readability, we have retained clean maps and instead ensured that key regions referenced in the text are clearly described.

Detailed comments:

The use of page-wise line numbers complicated the referencing to parts of the text. I will use PxxLxx to refer to page and line. I suggest using a continuous line numbering in the revised manuscript.

Author's response: Thank you for the comments we have now continuous line numbering in revised manuscripts

P03L07: what management of agroecosystem? Sustainable management? Or something more specific was intended here to then transition to SOC management and sequestration?

Author's response: The wording has been revised to clarify the type of agroecosystem management being referenced, with a more specific description provided to improve clarity. L96-102

"As anthropogenic pressures intensify through consumption patterns, resource demand and weather variability, sustainable management of agriculture practices has become essential to maintain productivity and ecological balance. Enhancing soil organic carbon (SOC) sequestration is one of the key benefits of improved management, contributing to climate solutions and long-term soil health maintenance (Paustian et al., 2016)."

P03L11: I think that a comma after "crop rotation" and after "Merr.)" would help readability.

Author's response: The sentence has been revised to improve readability by adding the suggested punctuation.

P03L13: better to use rotations ""can improve" instead of "improve"? I feel that pointing at a possibility would be more appropriate here.

Author's response: The wording has been revised to "can improve" to better reflect the conditional nature of the statement.

P03L20-34: this paragraph offers a well-made description of the three models used in the manuscript. I do not think it is necessary to mention other models here. However, in the discussion part, a brief mention of the possible benefits and drawback of including additional models in the ensemble (or even substituting a current model), withpossible examples, would benefit the reader and the discussion.

Author's response: A brief discussion has been added to the manuscript to acknowledge the potential benefits and limitations of including additional models in the ensemble framework, providing context for future extensions of this approach. L485-488

"Finally, the limited ensemble size in this study may not capture the full structural uncertainty found in larger initiatives like AgMIP, as additional models could introduce different sensitivities that would likely widen the projected uncertainty bounds."

P04L11: Is the 4km resolution sufficient for carbon accounting? Also, in the conclusions, carbon credits quantification is mentioned. I do not challenge the value of the presented methodology and results. But I wonder if the scale is sufficient for accurate carbon crediting. Especially given the fact that a lot of attention in this topic is currently shifting towards small scales like field-scales.

Author's response: Additional discussion has been added in method section to clarify the implications of the 4 km spatial resolution for carbon accounting applications. The manuscript now emphasizes that while the current resolution is suitable for large-scale assessments, finer spatial resolution may be required for field-scale applications such as carbon crediting. L183-188

"We selected a 4 km spatial resolution to balance between computational efficiency and the need to adequately capture large-scale spatial variability across the CONUS study domain. This resolution is appropriate for regional to national-scale assessments of crop productivity and soil carbon dynamics, including broad carbon accounting applications. However, it may not fully capture fine-scale heterogeneity in soil properties, management practices, and microclimatic conditions that influence carbon fluxes at the field level."

P04L14-16: I think this study has the potential to "emphasize the spatial variability" but the main and minor review points that I have listed above, currently hinder this possibility.

Author's response: The manuscript has been revised to improve the clarity of spatial variability representation (added spatial map of RMSE for yield (fig S1) and SOC (figure S2), including additional description and supporting analyses to better illustrate regional patterns. Page 23-24

P04L31-37: In describing these datasets, their scales and resolutions are not provided but they would be useful to the reader.

Author's response: Supplemental table (table S1) with details of the data has been included to describe the spatial resolution and scale of the datasets used in the study, improving transparency and clarity for the reader. Page 22

Table 1 Summary of biogeochemical process representations and model structures for the three-model ensemble.

	DAYCENT	DNDC	ECOSYS
Conceptual Approach	Semi-empirical / Pool-based	Mechanistic / Microbial-kinetic	Mechanistic / Thermodynamic
Yield	Potential biomass limited by moisture and Nitrogen (N)	Photosynthesis-driven; N and water restricted	Full canopy-root coupling; hourly CO2 fixation
Belowground Process	Empirical root distribution by soil depth	Dynamic root growth based on C-allocation	Detailed rhizosphere; ion diffusion and mass flow
SOC Turnover	First-order pool based (Active, Slow, Passive)	Double Monod kinetics; microbial biomass-driven	Microbial functional groups; substrate-specific decay

Note: Daycent = Daily Century; DNDC = Denitrification-Decomposition; Ecosys = Ecosystem model.

Section 2.2: the vertical discretization of the soil layers is not discussed but should be mentioned when relevant in one or more of the three models.

Author’s response: A description of soil layer discretization has been added where relevant to clarify how vertical soil processes are represented in the modeling framework. L260-261

“The models handle vertical soil discretization according to their distinct structural designs, aligned with our primary evaluation depth of 0-30 cm.”

P05L12: missing a space after the comma.

Author’s response: The space has been removed

P05L23: which kind of stress conditions? Mentioning the most important owuld be useful.

Author’s response: The manuscript has been revised to specify the key stress conditions considered in the simulations, improving clarity of model assumptions. L216-220

“This harvest index (HI) is crop-specific and sensitive to environmental and stress conditions, including water limitation (drought), nutrient stress (particularly nitrogen deficiency), and temperature extremes (heat or cold stress). This sensitivity enables DAYCENT to represent realistic interactions among biophysical processes and management practices.”

P06L37: This sentence similar to other sentences across the text that mention the higher performance of the model ensemble compared to single models. Although I wonder how likely it is that this model ensemble would perform worse than the single models it is built on (maybe a short addition on this in the discussion could be beneficial), I agree that this improved performance should be mentioned in the appropriate parts of the text, as already done by the authors. The following discussion section does a good job in illustrating the strengths and weaknesses of each model and how they converge in the higher performance of the ensemble. What is missing is a discussion on the value of this higher performance, both generally and spatially. On the one hand because PP09L33-34 is a bit generic. On the other because the spatial distribution of the improvements of the ensemble is not explicit in the figures.

Author’s response: Additional results and discussion has been included to better explain the value of the ensemble approach, including its contribution to reducing uncertainty and improving robustness across regions. The spatial distribution of performance improvements has also been clarified in the revised manuscript. L385-388, L459-455

"Figure S2 illustrates the spatial distribution of RMSE between ensemble model-simulated soil organic carbon (SOC) and RACA observations across the conterminous United States, highlighting regional variability in model performance."

"The ensemble's mean SOC and reduced RMSE illustrate its effectiveness in mitigating structural uncertainty and reproducing the SOC patterns captured in the RaCA database. These results underscore the promise of ensemble-based approaches for improving national scale carbon accounting and guiding sustainable land management strategies.

The divergent SOC trajectories across the models shows how different biogeochemical model interpret the management impacts differently."

P08L29-31: I agree that the ensemble suggest this, but I think that the capacity of conventional corn-soybean systems to enhance carbon sequestration should either be supported by references or be presented as a possibility.

Author's response: The statement has been revised to more clearly reflect that the observed carbon sequestration potential represents a modeled outcome and is presented with appropriate caution. Supporting references have also been added where applicable. L388-392

"Overall, the ensemble projection suggested a modest net increase in SOC across much of the Corn Belt, underscoring the capacity of conventional corn-soybean systems to enhance carbon sequestration under region-specific environmental and management conditions. However, these estimates represent modeled outcomes and should be interpreted with caution, as they may vary depending on site-specific conditions and management practices."

P08L43-35: I would suggest improving the readability of this sentence.

Author's response: The sentence has been revised to improve readability and clarity. L409-411

"ECOSYS employs a detailed rhizosphere sub-model where N-uptake is governed by ion diffusion and mass flow this can lead to localized N-stress that is more extreme than observation (Grant et al., 2001)."

P08L46 to P09L1-3: also of this sentence.

Author's response: The sentence has been revised to improve clarity and readability. L411-414

"Regional anomalies, such as DNDC's higher corn yield projections in the Southeast, are likely influenced by its representation of water stress and nutrient availability under less favorable climatic conditions, as well as its relatively linear response of crop yield to precipitation (Kaur et al., 2023)."

P09L15: this is another case in which an explicit description of data distribution would help because, up to now, spatial density of calibration was not discussed.

Author's response: Additional figure with spatial model preference for yield and SOC (figure S1 and Figure S2) has been added regarding the spatial distribution and density of calibration data to improve interpretation of the results. P23-24

P09L15-18: also here, please check sentence readability.

Author's response: The sentence has been revised to improve clarity and readability. L427-430

"These model-specific biases highlight the ongoing challenges in accurately representing complex interactions among soil texture, weather variability, and cultivar adaptability, further reinforcing the value of ensemble approaches for improving yield estimation."

P09L20-23: I think Grace&Robetrson indicate such potential when certain regenerative farming practices are adopted. Also, Wu mentions scenarios with shiftings in crop rotation. Does this compare well to the rotations used in the ensemble of this study? Also, the area of the corn belt, as of figure 6, seems to show both increases (up to 2t/ha per year if I am correct) and decreases (down to -2t/ha per year) in SOC. Providing an average value of this increase for the area that is discussed would be beneficial for this discussion. At the same time, it would be interesting to discuss how this increase interacts with the discrepancy between the ensemble results and the RaCA dataset shown in figure 3 (where the ensemble overestimates RaCA median by some 15 %). While I agree that the ensemble offers the best comparison in this figure, I think it is necessary to provide a clear discussion about how significant this increase per year is, in the entire corn belt, with respect to the accuracy of the ensemble itself.

Author's response: The discussion has been expanded to better contextualize SOC changes within the corn belt, including clarification of variability in both increases and decreases. Additional context has been provided to relate these findings to previous studies and to highlight uncertainties relative to observational datasets. L427-441

"These model-specific biases highlight the ongoing challenges in accurately representing complex interactions among soil texture, weather variability, and cultivar adaptability, further reinforcing the value of ensemble approaches for improving yield estimation.

Ensemble projections of SOC stock changes (0-30 cm) across CONUS under a 10-year corn–soybean rotation indicate a modest net sequestration. In the Corn Belt, SOC changes exhibit substantial spatial variability, with both gains and losses (approximately -2 to +2 t CO_2 ha^{-1} yr^{-1}), resulting in a relatively small positive average across the region. This pattern is broadly consistent with previous studies that report SOC sequestration potential under specific management or rotation scenarios (Grace and Robertson, 2021; Wu et al., 2025), although those studies often consider alternative practices or crop shifts not explicitly represented here. The magnitude of the modeled SOC increase should be interpreted in the context of model *uncertainty, as reflected by the ensemble's tendency to overestimate SOC relative to the RaCA dataset. Therefore, while the results suggest a potential for modest carbon gains, their significance at regional scale remains sensitive to both spatial heterogeneity and model accuracy."*

P09L36-46: This paragraphs nicely discuss some limitations of the study. However, I think the spatial resolution should also be addressed. 4km is likely a good trade-off, given the necessity of a supercomputer to run the simulations. But the implication of this resolution depending on applications is not discussed and, in cases like carbon credit quantification (mentioned in the conclusions), the scale becomes important as such activity likely needs higher spatial resolutions.

Author's response: The limitations section has been expanded to include discussion of spatial resolution and its implications for different applications, including carbon crediting and field-scale assessments. L482-485

"Additionally, the 4 km spatial resolution, while appropriate for regional-scale assessments, may not capture fine-scale variability in soils and management practices, which could limit its applicability for field-scale applications such as carbon credit quantification."

P09L40: please double check the grammar.

Author's response: The sentence has been revised to correct grammatical issues.

P10L15: I think that, for carbon crediting and regenerative agriculture initiatives, it has the potential to do these things. However, this should be presented more clearly as an outlook of this ensemble that will probably requires further model testing or development. For example, because regenerative practices are not included in the current simulations and, probably, the resolution of 4km is not sufficient for accurate carbon crediting.

Author's response: The statement has been revised to present carbon crediting applications as a potential future application, with clarification that further model development and higher spatial resolution may be required. L501-510

"Overall, this study highlights the value of an ensemble modeling framework for enhancing regional to national-scale assessments of crop productivity and soil carbon dynamics. By characterizing structural uncertainty and utilizing the ensemble mean to mitigate individual model biases, this approach provides a more robust scientific basis for carbon measurement, reporting, and verification. However, its application to farm-level carbon credit quantification and specific regenerative agriculture initiatives should be considered preliminary, as these applications may require further model refinement and higher-resolution spatial data to capture local-scale variability. Ensemble agroecosystem modeling provides a critical foundation for advancing evidence-based land management and sustainable agricultural policy in a changing environmental condition."

P12L28: In reference “FAO. (2025). Agricultural land (% of land area). Retrieved from: https://ourworldindata.org/grapher/agricultural-land-percent-land-area”, the hyperlink does not work.

Author’s response: The reference link has been corrected.

Figure 2 and Figure 3: please use consistent number of decimals. Also, in the second panel of Figure 2, is the RMSE of Ecosys 1.1? It seems to deviate most from NASS and has smaller RMSE than Caycent. Having something more than RMSE, like model efficiency and RMSE as % of max value of measurements (or something similar), would help in better reading these results.

Author’s response: The figures have been revised to ensure consistent use of decimal precision. Additional clarification has been provided to improve interpretation of RMSE values, including inclusion of normalized metrics in the result text and the spatial map of RMSE for yield. Page 17-18

Figure 6: Most values are concentrated between -02 and 0.2 km C m2 yr-1. Would it be more readable if the colour ramp is stretched in this range, maybe maintaining different colours for negative and positive values? Or it would get too complicated to read?

Author’s response: The color scale in Figure 6 has been revised to improve readability, with greater emphasis on the range where most values are concentrated while maintaining clear distinction between positive and negative values. Page 21

Reviewer #2

This manuscript egusphere-2026-1094 presents a multi-model ensemble framework combining three process-based agroecosystem models (DAYCENT, DNDC, ECOSYS) to simulate corn and soybean yields as well as soil organic carbon (SOC) stock changes across the continental United States (CONUS) from 2014–2023. The study is spatially comprehensive (4 km^2 resolution) and leverages harmonized environmental, soil, and management datasets. The ensemble median generally outperforms individual models when compared against NASS yield data and RaCA SOC measurements, showing reduced RMSE and improved central tendency. The authors conclude that ensemble modeling reduces structural uncertainty and provides a more robust basis for carbon accounting and sustainable agricultural policy. While the topic is timely and relevant to SOIL, the manuscript contains several conceptual, methodological, and presentational issues that need substantial revision before publication. Please find my comments below.

Author’s response: We would like to thank the reviewer for their time and thoughtful critique of this manuscript. We have addressed all the comments and believe that the article has been improved because of the valuable feedback. Please find our point-by-point responses below.

Major comments:

1. The manuscript oscillates between claiming the ensemble “represents” uncertainty (e.g., lines 38–39, Page 3) and “reduces” uncertainty (e.g., lines 17–18, Page 2). These are different scientific goals. Please clarify: is the ensemble intended to characterize the range of plausible outcomes from structural variability, or to improve predictive accuracy via error cancellation? The framing affects the interpretation of RMSE improvements and the value of including biased models like ECOSYS.

Author’s response: We appreciate the reviewer’s request for clarification. We have revised the manuscript (specifically in the abstract, introduction & conclusion section) to consistently frame that the ensemble framework can be used as a tool to characterize structural uncertainty to capture the range of plausible outcomes resulting from the different modeling assumptions inherent in DAYCENT, DNDC, and ECOSYS. While the ensemble average is used as an estimator of the mean condition (which empirically improves predictive accuracy when compared to NASS and RaCA data), we clarify that the primary value of the multi-model approach lies in its ability to quantify the spread of model responses across diverse CONUS environments. We acknowledge that including a model with different sensitivities (eg., ECOSYS) is not about correcting the mean, but about ensuring the structural variability of agroecosystem processes is fully represented in our SOC and yield prediction. Line 63-69, Line 128-132, Line 501-508

"Overall, the ensemble framework characterizes structural uncertainty by capturing the range of plausible model responses, while the ensemble mean provides a more reliable estimate of system behavior than any individual model. By integrating diverse model perspectives, this approach enhances predictive robustness and reduces dependence on model-specific assumptions. Consequently, it offers a scalable and data-driven framework for assessing soil carbon dynamics and crop productivity across U.S. agroecosystems, while explicitly characterizing and communicating uncertainties associated with model variability."

"Ensemble modeling leverages the complementary strengths of individual models to characterize a range of plausible system responses under diverse climate and management conditions. This framework provides a more robust basis for environmental assessments by quantifying structural uncertainty while utilizing the ensemble mean to provide a stable estimate of mean agroecosystem conditions (Basso et al., 2025; Martre et al., 2015)."

"Overall, this study highlights the value of an ensemble modeling framework for enhancing regional to national-scale assessments of crop productivity and soil carbon dynamics. By characterizing structural uncertainty and utilizing the ensemble mean to mitigate individual model biases, this approach provides a more robust scientific basis for carbon measurement, reporting, and verification. However, its application to farm-level carbon credit quantification and specific regenerative agriculture initiatives should be considered preliminary, as these applications may require further model refinement and higher-resolution spatial data to capture local-scale variability."

2. ECOSYS consistently underperforms for both corn and SOC. The manuscript acknowledges this but does not convincingly justify why ECOSYS should be retained in the ensemble beyond "complementary strengths." If ECOSYS is mechanistically more detailed but poorly calibrated for CONUS, its inclusion may degrade rather than enhance ensemble performance. Please provide a clearer rationale, or consider a sensitivity analysis excluding ECOSYS.

Author's response: We agree with the reviewer that ECOSYS's performance metrics are lower than those of DAYCENT and DNDC in this specific application. However, we have chosen to retain ECOSYS to ensure that the ensemble framework captures a biologically plausible range of structural uncertainties. In response to the comment, we have revised the limitations section to provide a clearer rationale. We now explicitly discuss how ECOSYS's mechanistic complexity makes it more sensitive to the uniform phenotypic parameters (such as maturity groups) used for computational efficiency at the CONUS scale. We characterize this as a sensitivity penalty where a high-complexity model is constrained by simplified regional inputs rather than a structural failure. By utilizing the ensemble mean, we mitigate the impact of this underperformance on our central estimates while allowing the inter-model spread to provide a more conservative quantification of the uncertainty inherent in agroecosystem modeling. We believe that excluding ECOSYS would result in an artificially narrow uncertainty range that overlooks known mechanistic sensitivities. *L467-482*

"This study uses a uniform set of crop phenotypic and parameter values for corn and soybean across the entire spatial domain. While this standardization enabled large-scale simulations and computational efficiency, it inherently omitted genotype-environment interactions and region-specific cultivar adaptations. In ECOSYS, for example, crop phenology is highly sensitive to thermal time and maturity group parameters, both of which vary significantly across U.S. regions. Consequently, the underperformance of ECOSYS relative to DAYCENT and DNDC likely reflects a sensitivity penalty where the model's mechanistic complexity was constrained using simplified, uniform maturity groups. Previous work have demonstrated that regional calibration of maturity group is essential for accurately reproducing phenological development (Li et al., 2022); thus, using a single maturity group may have misrepresented region-specific growth dynamics and yield responses. Despite this, ECOSYS was retained in the ensemble to ensure the structural uncertainty range accounts for high-complexity mechanistic feedback such as plant-water relations and microbial kinetics that simpler, pool-based models might overlook. Although this approach supports continental scale consistency, it introduces limitations in representing localized cultivar adaptation and environmental responses, highlighting an important direction for future ensemble model refinement."

3. Section 2.3 uses only RMSE for model evaluation and parameter fine-tuning. RMSE does not penalize model complexity, nor does it account for systematic bias or pattern similarity. Please add additional metrics (e.g., Nash-Sutcliffe efficiency, percent bias, or Akaike Information Criterion as you suggested) to better characterize model performance. Also clarify whether the same data were used for calibration and validation to avoid overfitting.

Author's response: We thank the reviewer for this important clarification and completely agree that the distinction between calibration and independent validation is critical for interpreting model performance.

To evaluate the potential for model overfitting and assess framework stability across both SOC and crop yields, we tested a stratified data-splitting protocol using the createDataPartition framework from the caret package. This algorithm protects against sampling bias in continuous datasets by first dividing the data (RaCA SOC or NASS yields) into equal-sized numeric bins based on quantiles, and then randomly sampling exactly 50% of the data points from each individual bin to form the calibration set, leaving the remaining 50% for independent validation. This stratification ensures that both subsets mirror the exact same underlying mean, variance, and outlier distribution. This approach resulted into nearly identical error distributions and performance metrics between the independent groups: an Ensemble RMSE of 4.7 kg/m2 (calibration) vs. 4.6 kg/m2 (validation) for soil organic carbon, 3.0 Mg/ha (calibration) vs. 3.0 Mg/ha (validation) for corn yield, and 1.0 Mg/ha (calibration) vs. 1.0 Mg/ha (validation) for soybean yield. Because of this identical performance explicitly shows that modeling framework is structurally stable and free from overfitting, we opted to pool the subsets and present the full dataset in the final manuscript figures to fully utilize the available datasets and maintain geographic continuity across the CONUS. We intentionally opted not to report R2 or Nash-Sutcliffe Efficiency (NSE), as these metrics can be misleadingly inflated in regional contexts where models are closely aligned with the regional observation mean; instead, we rely on RMSE and normalized RMSE (%RMSE) to provide an interpretable measure of error magnitude across the CONUS. To further enhance spatial transparency, we have incorporated additional analyses in the supplementary material, including maps showing the spatial distribution of the RMSE of yield and SOC datasets (Figures S1 and S2) show structural uncertainty, providing a more rigorous assessment of the framework's current capabilities while acknowledging that independent validation on entirely external temporal datasets remains a priority for future research. Line 256-286

"Each of the three models was calibrated using county level crop yield data from the National Agricultural Statistics Service (NASS) (NASS, 2025) and site-level SOC data from the Rapid Carbon Assessment (RaCA) dataset (Wills et al., 2014) across CONUS. Model parameterization was optimized to minimize the root mean square error (RMSE) and normalized RMSE (%RMSE). Relative metrics like the Nash-Sutcliffe Efficiency (NSE) and R2 were intentionally excluded, as they can be misleadingly inflated in large-scale regional contexts where model predictions closely align with a strong regional observation mean.

To explicitly evaluate the risk of model overfitting and assess framework stability prior to final execution, a stratified data-splitting protocol was implemented using the createDataPartition framework in the caret package. Continuous datasets (RaCA SOC and NASS yields) were partitioned into equal-sized numeric bins based on quantiles, and 50% of the data points were randomly sampled for calibration, leaving the remaining 50% as an independent validation subset. This cross-validation protocol yielded identical error distributions between the independent groups specifically, an Ensemble SOC RMSE of 4.7 kg/m2 (calibration) vs. 4.6 kg/m2 (validation), a corn yield RMSE of 3.0 Mg/ha (calibration) vs. 3.0 Mg/ha (validation), and a soybean yield RMSE of 1.0 Mg/ha (calibration) vs. 1.0 Mg/ha (validation). Because this identical performance demonstrated that the modeling framework was structurally stable and robust against overfitting, the data subsets were pooled for the final simulations. This approach maximizes geographic continuity across the CONUS domain and fully utilizes the available high-resolution datasets. This approach provides a robust upper bound of current process-based modeling capabilities at a 4 km2 scale, representing the collective central tendency of diverse mechanistic assumptions."

4. In the method section (Section 2.3), it appears that the same datasets (NASS crop yield data and RaCA SOC data) have been used for both model parameterization (i.e., fine-tuning to minimize RMSE) and subsequent model validation. This practice risks overfitting and can lead to overly optimistic performance metrics, including the reported RMSE improvements for the ensemble. Please clarify whether any form of data separation (e.g., temporal or spatial holdout, cross-validation) was applied. If not, the reported agreement between simulated and observed values may reflect calibration rather than true predictive skill, and the ensemble's apparent advantage could be overstated.

Author's Response: We thank the reviewer for this important clarification. We acknowledge the concern regarding the potential for overfitting when utilizing the same datasets for both model parameterization and performance evaluation.

To rigorously test for regional sampling bias and evaluate the potential for model overfitting across the CONUS, we executed a diagnostic stratified data-splitting protocol using the createDataPartition framework. This algorithm protects against sampling imbalances in continuous datasets by dividing the complete range of empirical observations (RaCA SOC or NASS yields) into equal-sized numeric bins based on quantiles. It then randomly samples exactly 50% of the data points from each individual bin to form a calibration set, reserving the remaining 50% as an independent validation holdout. This stratification ensures that both subsets share identical statistical properties (mean, variance, and outlier distributions).

As detailed in our response to Comment 3, data split yielded nearly identical error distributions and performance metrics between the independent calibration and validation groups. As RMSE values were mathematically indistinguishable, it explicitly demonstrated that our regional multi-model framework is structurally stable and entirely free from over-calibration. Consequently, to fully utilize all datasets, and maintain complete geographic continuity across the CONUS, we opted to pool the data for the final manuscript figures.

We have revised Section 2.3 to explicitly detail this stratified split diagnostic and clarify that the final metrics represent a globally optimized calibration across the full domain, ensuring each model architecture is evaluated at its best achievable regional fit. We contend that the ensemble's advantage remains highly significant: even across the pooled domain, the ensemble mean consistently smooths out individual model discrepancies. This structural diversity effectively mitigates the systematic biases of any single model architecture, providing a robust, transparent baseline of current process-based modeling capabilities at this scale while independent validation on entirely external temporal datasets remains a priority for future research.

5. The manuscript states that "model parameterization was conducted to minimize RMSE between observed and predicted values" but does not describe the spatial/temporal splitting of NASS and RaCA data. If the same years and locations used for calibration are also used for evaluation, the reported RMSE values likely underestimate true prediction uncertainty. Please discuss how this might affect the apparent ensemble improvement.

Author's Response: We agree with the reviewer that because the models are evaluated across the same spatial and temporal domains used for calibration, the reported RMSE reflects the error bound of a globally optimized regional calibration rather than a fully independent prediction uncertainty.

As detailed in our responses to Comments 3 and 4, we explicitly evaluated this potential underestimation of uncertainty by executing a diagnostic stratified split using the createDataPartition framework. This technique uses 50% independent spatial holdout across identical quantile bins to mirror the full dataset's distribution. As shown by the metrics reported in our response to Comment 3, the ensemble error bounds were identical between the calibration and independent validation subsets across all soil carbon and crop yield. This mathematical convergence directly demonstrates that the apparent ensemble improvement is not due to localized overfitting or over-calibration. While we acknowledge that a fully independent temporal holdout would capture a broader range of true predictive uncertainty which remains a priority for future work as more data becomes available, the relative advantage of the ensemble mean over individual models remains a highly robust finding of this study. We have expanded our discussion in the revised manuscript to explicitly contextualize these error values as an optimized baseline, outlining limitation of our current uncertainty estimates and the clear mathematical advantage of the multi-model ensemble.

6. The finding that the Midwest and Southeast show SOC gains while the Great Plains and West show losses under corn-soybean rotation is important but under-discussed. Why would the same rotation cause SOC losses in drier/western regions? Is this due to lower baseline SOC, different decomposition rates, or management differences? Similarly, DNDC projected SOC losses exceeding 0.01 kg C m^{-2} yr^{-1} in high-yielding zones. Please expand the mechanistic interpretation.

Author's response: We thank the reviewer for this insightful comment. We agree that the divergent SOC trends between the humid regions (Midwest and Southeast) and the semi-arid regions (Great Plains and West) need more detailed mechanistic explanation in the manuscript.

The observed SOC losses in the Great Plains and West under corn-soybean rotations are primarily driven by a negative carbon mass balance. In these regions, water-limited Net Primary Productivity (NPP) results in lower biomass returns, while periodic moisture and high temperatures maintain high rates of heterotrophic respiration. Furthermore, the low-residue incorporation due to low yield in these drier environments often fails to meet the minimum carbon input required to maintain SOC equilibrium.

Regarding the high-yielding zones, DNDC projections suggest that the combination of nitrogen availability and optimal soil moisture and temperature profiles create a high carbon turnover. Here, accelerated microbial activity can mineralize existing soil organic matter faster than new residues are humified. We have expanded the discussion to incorporate these biogeochemical mechanisms. Note that DNDC use Double Monod kinetic equation where microbial growth is very sensitive to availability of C and N (Li et al., 1992), whereas model like DAYCENT use conservative first-order pool based approach where N primarily limits plant productivity rather than acting as a direct catalyst for the mineralization of stabilized SOM. Line 454-464

"*The divergent SOC trajectories across the models shows how different biogeochemical model interpret the management impacts differently. The widespread SOC decrease projected by DNDC, particularly in high-yielding zones, is driven by its Double Monod kinetic framework, which simulates microbial growth as a function of simultaneous carbon and mineral nitrogen availability (Li et al., 1992). In this framework, abundant nitrogen leads to mineralization of* *existing organic matter to satisfy stoichiometric demands faster than humification of crop residue. Conversely, DAYCENT employs a conservative pool-based approach where nitrogen primarily boosts plant productivity and residue inputs rather than accelerating the decay of stabilized pools (Parton et al., 1998). Ecosys uses microbial carbon use efficiency (CUE) and physical protection within soil aggregates, assuming that high-input carbon is stabilized as microbial biomass rather than respired (Grant, 1998). "*

7. The caption and title of Figure 6 currently read "ECOSYS model projected SOC change..." but the figure shows (and text describes) multiple models and the ensemble. This must be corrected to "Agroecosystem model projected..." or similar.

Author's response: We thank the reviewer for identifying this oversight. We have corrected the title and caption of Figure 6 as suggested

8. You state that the ensemble "captures uncertainty", but with only three models, it is unlikely that their spread represents the full range of uncertainty across all existing agroecosystem models. Please add a discussion of how the three-model spread compares to published multi-model ensemble studies (e.g., AgMIP) and whether the results are robust to inclusion of other models (e.g., APSIM, EPIC).

Aurthor's response: We thank the reviewer for this constructive comment. We acknowledge that a three-model ensemble does not represent the range of structural uncertainty found in larger initiatives like AgMIP. The selection of Ecosys, Daycent, and DNDC for this study was driven by their widespread use in regional carbon assessments and their direct alignment with our project objectives. While these three models capture distinct biogeochemical processes ranging from pool-based to microbial-kinetic approaches we agree that the inclusion of models such as APSIM or EPIC would likely widen the ensemble spread. We have added a statement to the method and limitations section to further clarify the choice made in this study. L480-488

"Although this approach supports continental scale consistency, it introduces limitations in representing localized cultivar adaptation and environmental responses, highlighting an important direction for future ensemble model refinement. Additionally, the 4 km spatial resolution, while appropriate for regional-scale assessments, may not capture fine-scale variability in soils and management practices, which could limit its applicability for field-scale applications such as carbon credit quantification. Finally, the limited ensemble size in this study may not capture the full structural uncertainty found in larger initiatives like AgMIP, as additional models could introduce different sensitivities that would likely widen the projected uncertainty bounds."

Minor comments:

1. Line 13–14 (Page 2): Add units to RMSE values

Aurthor's response: Thank you for the comment, we have added unit in updated manuscript.

2. Lines 17-18 (Page 4): The claim of "actionable information for policymakers" is unsupported by the current results (baseline only, no management comparisons). Please tone down or reframe.

Aurthor's response: We acknowledge the reviewer's comment that without explicit management scenario comparisons, actionable solution for specific policy interventions is limited. We have updated this statement to emphasize that the results provide a spatial baseline map and identify regional risk zones for SOC decline. L152-154

"Furthermore, the results provide a spatially explicit baseline that can inform policymakers, startups, and stakeholders in prioritizing regions for carbon MRV and identifying areas at higher risk for SOC decline under current crop rotations."

3. Lines 24 (Page 4): Change "projections" to "estimates" or "simulations" since the study does not forecast future conditions.

Aurthor's response: Thank you for the comment, we have updated the text in revised manuscript.

4. Line 27 (Page 5): Typo: "where grown" → "were grown."

Aurthor's response: Thank you for finding the typo, we have updated the text in revised manuscript.

5. Section 2.2 (Page 5–6): The model descriptions are dense and lack comparative synthesis. Please add a summary table or paragraph comparing key differences: process formulation (mechanistic vs. semi-empirical), required inputs, treatment of belowground processes, and yield determination logic. Also explicitly state why these three models were selected over others.

Aurthor's response: We thank the reviewer for the suggestion to provide a more comparative synthesis. We have added a rationale for our model selection to the methods section and expanded the limitations section to address the scope of our ensemble. Additionally, we have included a summary table (Table 1) that synthesizes the key structural differences between the models, specifically focusing on process formulations, belowground dynamics, and yield determination logic. These changes clarified why these three specific models were chosen to represent a broad spectrum of biogeochemical modeling approaches.

Table 1 Summary of biogeochemical process representations and model structures for the three-model ensemble.

	DAYCENT	DNDC	ECOSYS
Conceptual Approach	Semi-empirical / Pool-based	Mechanistic / Microbial-kinetic	Mechanistic / Thermodynamic
Yield	Potential biomass limited by moisture and Nitrogen (N)	Photosynthesis-driven; N and water restricted	Full canopy-root coupling; hourly CO2 fixation
Belowground Process	Empirical root distribution by soil depth	Dynamic root growth based on C-allocation	Detailed rhizosphere; ion diffusion and mass flow
SOC Turnover	First-order pool based (Active, Slow, Passive)	Double Monod kinetics; microbial biomass-driven	Microbial functional groups; substrate-specific decay

Note: Daycent = Daily Century; DNDC = Denitrification-Decomposition; Ecosys = Ecosystem model.

6. Figure 2: Use a color-blind friendly palette. Also explain why the ensemble shows many outliers despite lower RMSE.

Aurthor's response: Thank you for suggestion, we have updated figure 2 and figure 3 to color blind palette. Regarding the outliers, the ensemble achieves a lower RMSE by averaging out systematic biases, which mathematically tightens the interquartile range. This narrower distribution causes localized extreme values (outliers) in ensemble plot. Page 17-18

7. Line 17 (Page 7): Delete the extra "a" before "an RMSE of 2.7."

Aurthor's response: Thank you for finding the typo, we have updated the text in revised manuscript.

8. Line 43–44 (Page 8): You suggest ECOSYS bias may come from coupled root-canopy processes. Could you quantify/support this (e.g., from sensitivity analyses) using your model output?

Aurthor's response: We thank the reviewer for this constructive suggestion. Based on our model outputs and previous literature, we have clarified that the underprediction stems from the model's high sensitivity to nitrogen (N) availability and rhizosphere uptake kinetics rather than general root-canopy coupling. Specifically, ECOSYS utilizes a detailed rhizosphere sub-model where N-uptake is explicitly governed by ion diffusion and mass flow. In large-scale simulations driven by standardized, coarse regional input data, this detailed mechanics can induce artificial, localized N-stress that is more severe than actual field observations. We note that conducting a new, dedicated sensitivity analysis of these highly localized soil-crop processes is outside the scope of this study, as our primary objective centers on macro-level, regional-scale ensemble modeling across CONUS. However, we have added a sentence to the discussion (Lines 407-414) citing established literature (Grant et al., 2001) and clarifying how regional inputs interact with ECOSYS's underlying nitrogen-water coupling to support this claim.

"In our large-scale application, the use of standardized N-input data likely triggered disproportionate N-stress within the model, leading to lower-than-observed yields. ECOSYS employs a detailed rhizosphere sub-model where N-uptake is governed by ion diffusion and mass flow this can lead to localized N-stress that is more extreme than observation (Grant et al., 2001)"

9. Line numbers are renewed in every page.

Aurthor's response: Thank you for the comment, we have now updated it to continuous line number.

Citation: https://doi.org/10.5194/egusphere-2026-1094-RC2

Reference

Basso, B., Tadiello, T., Millar, N., Robertson, G. P., Paustian, K., Maureira, F. S., Albarenque, S., Baer, B., Price, L., and Sharma, P. (2025). A multi model ensemble reveals net climate benefits from regenerative practices in US Midwest croplands. *Scientific Reports* **15**, 24881.

Grace, P., and Robertson, G. P. (2021). Soil carbon sequestration potential and the identification of hotspots in the eastern Corn Belt of the United States. *Soil Science Society of America Journal* **85**, 1410–1424.

Grant, R. (1998). Simulation of methanogenesis in the mathematical model ecosys. *Soil Biology and Biochemistry* **30**, 883–896.

Grant, R., Jarvis, P., Massheder, J., Hale, S., Moncrieff, J., Rayment, M., Scott, S., and Berry, J. (2001). Controls on carbon and energy exchange by a black spruce–moss ecosystem: testing the mathematical model ecosys with data from the BOREAS experiment. *Global biogeochemical cycles* **15**, 129–147.

Kaur, N., Hui, D., Riccuito, D. M., Mayes, M. A., and Tian, H. (2023). Response patterns of simulated corn yield and soil nitrous oxide emission to precipitation change. *Ecological Processes* **12**, 17.

Li, C., Frolking, S., and Frolking, T. A. (1992). A model of nitrous oxide evolution from soil driven by rainfall events: 1. Model structure and sensitivity. *Journal of Geophysical Research: Atmospheres* **97**, 9759–9776.

Li, Z., Guan, K., Zhou, W., Peng, B., Jin, Z., Tang, J., Grant, R. F., Nafziger, E. D., Margenot, A. J., and Gentry, L. E. (2022). Assessing the impacts of pre-growing-season weather conditions on soil nitrogen dynamics and corn productivity in the US Midwest. *Field Crops Research* **284**, 108563.

Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J. W., Rötter, R. P., Boote, K. J., Ruane, A. C., Thorburn, P. J., and Cammarano, D. (2015). Multimodel ensembles of wheat growth: many models are better than one. *Global change biology* **21**, 911–925.

NASS (2025). Quick Stats Database (Crop Yield Data). (N. A. S. S. U.S. Department of Agriculture, ed.).

Parton, W. J., Hartman, M., Ojima, D., and Schimel, D. (1998). DAYCENT and its land surface submodel: description and testing. *Global and planetary Change* **19**, 35–48.

Paustian, K., Lehmann, J., Ogle, S., Reay, D., Robertson, G. P., and Smith, P. (2016). Climate-smart soils. *Nature* **532**, 49–57.

Wills, S., Loecke, T., Sequeira, C., Teachman, G., Grunwald, S., and West, L. T. (2014). Overview of the US rapid carbon assessment project: sampling design, initial summary and uncertainty estimates. *In* "Soil carbon", pp. 95–104. Springer.

Wu, Y., Davis, E. C., and Sohngen, B. L. (2025). Crop rotation and the impact on soil carbon in the US Corn Belt. *Carbon Balance and Management* **20**, 6.