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 agree that the original wording may have been unclear. The model evaluation presented in this study reflects calibration rather than independent validation, as the available datasets were used to optimize model performance by minimizing RMSE for yield and SOC. We note that R^2 and Nash–Sutcliffe Efficiency (NSE) were not used in this study, as they can be misleading when applied to calibration-only analyses. Instead, model performance is evaluated using RMSE and normalized RMSE (%RMSE) to provide a consistent and interpretable assessment. The manuscript has been revised to clarify this distinction, and a

statement acknowledging the absence of independent validation due to data limitations has been included.

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. 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.

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.

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.

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.

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 and SOC), including additional description and supporting analyses to better illustrate regional patterns.

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.

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.

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.

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 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.

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.

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

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

P08L46 to P09L1-3: also of this sentence.

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

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.

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

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

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.

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.

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.

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.

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.