Response to referee comments

Referee #3

This manuscript conducts a meta-analysis to explore the uncertainty of simulating NEE using comparing machine learning techniques, time-scale and spatial-scale changes, and input variables. This is an important topic to solve the difference between observed and predicted NEE. However, this manuscript doesn't clarify the objectives and detail of data processing. Oversimple descriptions in the Methods section makes readers confusing. Additionally, the usage of too many speculative explanations in the discussion section is hard to draw universal conclusions. This manuscript doesn't clarify the motivation of the work, especially in the advantages and potential of ML. In summary, the paper needs to be substantially revised, and some parts need further elaboration.

Response: We would like to thank the reviewer for the positive comments and the time invested to review our manuscript. The revised manuscript will follow the reviewer's recommendations.

L32, This sentence hardly reflects the scientific value of this paper.

Response: Thank you for the insightful comments. We will delete this sentence.

L40-L43, The advantages and the current situation applied to ML need to be further reviewed, which is beneficial for readers to understand the purpose of introducing ML in this paper.

Response: Thank you for the insightful comments. We will further state the advantages and state-of-the-art of using machine learning for NEE simulations compared to process-based models. Previous process-based and empirical models are limited in their performance in NEE prediction due to our poor understanding of the mechanisms of NEE. Compared to process-based models, machine learning-based NEE prediction models can improve accuracy by establishing complex relationships between observed NEE and environmental variables in a data-driven manner.

L45, The sentence, "a synthesis evaluation is …limited", needs to be further explained otherwise it is hardly understood.

Response: We will provide more descriptive text, such as the limitations of the existing local multi-model evaluation.

L49-50, need references, preferably with 2 examples

Response: We will provide the appropriate references here.

L52-54, There is a logical gap between this sentence and the previous statements.

Response: We will emphasize the importance of evaluation studies from 'local' to 'global', thus making the logic smooth.

L88-93, The uncertainty caused by spatio-temporal heterogeneity cannot be confused with the volume of data sets. Because large-data volume does not equate to higher heterogeneity. Big data provides more opportunities to build balanced-training data. This section may need to be rewritten.

Response: Thank you for the insightful comments. Indeed, data volume does not always correspond to large spatiotemporal heterogeneity. For example, a site with a long year span and small interannual climate variability may have a large data volume, but it may not contain large spatiotemporal heterogeneity. We will rewrite this paragraph to avoid this confusion.

L107-108, need references

Response: We will provide the reference (Marcot and Hanea, 2021) here.

L116, "Other Features" needs to be clarified. The purpose of this manuscript may be to explore: the uncertainty of NEE evaluation results caused by ML techniques, spatio-temporal resolution remote sensing data, and verification methods according to the introduction?

Response: We will revise it as ‘machine learning algorithms, Spatio-temporal resolution of remote sensing data, and validation methods’.

L144, An oversimplified description of the workflow, please give an overview and detailed sub-steps of data processing and simulation. It is hard to know the objectives of each analysis for readers.

Response: We will provide more details on data extraction.

L150, Abbreviations in the figure need to be clarified

Response: We will provide the meaning of these abbreviations in the figure caption.

L178, Need scale bar and north arrow in figure 3

Response: We will add a scale bar and north arrow.

L198, It is difficult to find the differences among algorithms using simple comparisons in figure 5a, and needs more statistically testing. Additionally, this figure confuses me. Why are MLR, RF, SVM, and ANN separately compared? Please provide explanations. Why is PLSR with high R2 removed? Finally, there are also some problems with the image. The caption does not explain the details of the box. Does the line in the box represent the mean or the median?

Response: Thank you for the insightful comments. MLR, RF, SVM, and ANN are the more commonly used methods. Other algorithms such as PLSR may have too small a sample size. Since cross-study comparisons of algorithm accuracy include differences in data used in model construction, we perform a pairwise comparison of these four algorithms in figure 5b. Multiple models are developed for uniform training data in a single study so that the interference of data variation is removed. The line in the box shows the median. We will revise this figure (and the caption of the figure) in detail to provide more clarification.

L205, Avoid using the word “significant” without statistically testing

Response: We will replace the word ‘significant’.

L206-210, It is hard to read the trend in Figure 6. Recommend adding a line chart to demonstrate the decreasing trend.

Response: We will add trend lines.

L212, There are no details of the boxplot. Are all models incorporated into the time-scales comparison, or only RF, SVM, and ANN? Please add the details of data processing.

Response: All models have been included in the assessment of time-scales variations.

L223, Also, use these words carefully without statistically testing.

Response: We will replace the word ‘significantly’.

L263, Need to reorder the y-axis text in figure 8. Furthermore, a serious question is whether the comparison analysis of these variables keeps other variables constant? If not, conclusions based on comparisons of R2 may not hold water.

Response: We will readjust the order. Indeed, in the assessment of the impacts of variables, the interference between variables is not eliminated (and indeed it is difficult to keep the other variables constant). Therefore, the subsequent Bayesian network-based analysis can be considered a multivariate analysis with the elimination of the interference between the variables.

L299, Lacking the in-depth discussion of the uncertainty of NEE prediction resulting from time-scale change.

Response: We will improve the discussion section by discussing in more depth the impact of factors affecting NEE prediction (especially the change in the time scale you mentioned). Since this study only provides findings based on statistics obtained, we will compare some of the explanations of possible effects on time scales at the mechanistic level in previous studies.

L308, There are too many speculative parts and insufficient supporting materials in section 4.1 of discussed.

Response: We will add more references to support our discussion in section 4.1.

For example, we will add three references for the following discussion text:

‘In addition, the impacts of lagged effects (Hao et al., 2010; Cranko Page et al., 2022) of covariates are not considered in most models, which may underestimate the degree of explanation of NEE for some predictor variables (e.g. precipitation). Most of the machine learning-based models use only the average Ta and do not take into account the maximum temperature, minimum temperature, daily difference in temperature, etc., as in the process-based ecological models (Mitchell et al., 2009).’

L321-323, The discussion of model accuracy difference caused by satellites needs careful. This sentence needs further support. Are you implying that the time scale compensates for the uncertainty caused by the spatial scale?

Response: The discussion on this in the current version was not careful although it is true that the temporal availability of MODIS data and Landsat data differs greatly. We will consider adding references.

L326-330, This sentence is too long

Response: We will simplify this sentence.

L330-332, The time-scale discussion containing spatial-scale matching will confuse readers.

Response: Thank you for the insightful comments. We will separate the discussion of time-scale and the discussion of spatial-scale matching. In the current version, these two parts are placed in one paragraph and it may confuse readers. We will revise this.

L349, Does "coarse-resolution" here note spatial resolution or temporal resolution?

Response: This refers to spatial resolution. We will clarify this in the manuscript.

References

Cranko Page, J., De Kauwe, M. G., Abramowitz, G., Cleverly, J., Hinko-Najera, N., Hovenden, M. J., Liu, Y., Pitman, A. J., and Ogle, K.: Examining the role of environmental memory in the predictability of carbon and water fluxes across Australian ecosystems, 19, 1913–1932, https://doi.org/10.5194/bg-19-1913-2022, 2022.

Hao, Y., Wang, Y., Mei, X., and Cui, X.: The response of ecosystem CO2 exchange to small precipitation pulses over a temperate steppe, Plant Ecol, 209, 335–347, https://doi.org/10.1007/s11258-010-9766-1, 2010.

Marcot, B. G. and Hanea, A. M.: What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?, Comput Stat, 36, 2009–2031, https://doi.org/10.1007/s00180-020-00999-9, 2021.

Mitchell, S., Beven, K., and Freer, J.: Multiple sources of predictive uncertainty in modeled estimates of net ecosystem CO2 exchange, Ecological Modelling, 220, 3259–3270, https://doi.org/10.1016/j.ecolmodel.2009.08.021, 2009.