Reply to the comments
RC1: 'Comment on egusphere-2025-4920', Anonymous Referee #1, 23 Dec 2025

We thank Anonymous Referee #1 for reviewing the manuscript (egusphere-2025-4920). We appreciate the supporting words. Here we answer the questions and issues raised by the Reviewer in detail.

Note that the comments of Anonymous Referee are shown below *in italic*. Our responses to the comments are presented below in normal font style.

This manuscript evaluates machine learning-based ensemble approaches (stacking meta-models) for improving carbon flux predictions in agricultural systems. The authors compare Multiple Linear Regression, Random Forest, XGBoost, and XGBoost with environmental covariates (XGB+) against traditional multi-model median (MMM) approaches across four sites (two croplands, two grasslands). The study reports substantial improvements and uses SHAP analysis to provide interpretability. While the topic is relevant and the interpretability focus is commendable, fundamental methodological flaws in the validation strategy undermine the reliability of the results. The inappropriate use of random train/test splits on temporally autocorrelated data, combined with limited site coverage and missing analysis of temporal structure, prevent acceptance in the current form.

We thank the Reviewer for the evaluation of the manuscript. Similar issues were raised by Reviewer 2. We performed additional calculations to address the issues and provide solid evidence on the utility of our novel approach. We also extended the applicability of the method. We hope that the Reviewer will accept the improved manuscript.

Major comments
1. Validation strategy: The authors use random 70/30 train/test splits (line 238-244) on daily time-series carbon flux data without any consideration of temporal autocorrelation structure. This approach could bring potentially serious issues. Daily carbon fluxes exhibit strong temporal autocorrelation due to weather persistence, phenological continuity, and soil moisture memory. Splitting such data could be quite risky.

We agree that 70/30 train/test splits are not correct for temporal generalization or for prediction due to temporal autocorrelation of the fluxes. However, for within-regime flux reconstruction, which was the primary scope of the study, random splitting remains a useful benchmark (e.g. Papale et al., 2006; Moffat et al., 2007) for evaluating how well the meta-model captures the complex, non-linear relationships within the existing data manifold. By removing 30% of the data across the entire temporal range, we force the model to reconstruct complex fluxes (like NEE during a drought or a peak growing season) using only the structural relationships it has learned from the remaining 70%, without the benefit of knowing the specific daily sequence. While it does

not test temporal extrapolation to the same extent as leave-one-year-out validation, it remains a critical diagnostic for model fidelity to the observed data manifold.
Nevertheless, to address the issue we made an alternative validation using the complete leave-one-year-out (LOYO) approach (averaging validation results for all years separately). The results will be included in the revised manuscript. Not surprisingly, in croplands the approach was not suitable, due to the presence of crop rotation, meaning that different crop types are grown in consecutive years. In this context, since the training dataset covers only a few years for crops, leaving an entire year out leads to underperforming model construction. For grasslands the method worked remarkably well, due to the same (or slowly changing) community structure/plant functional type. This method implicitly resolved the temporal autocorrelation issue.
We extended the materials and methods, plus the results section with tables quantifying the LOYO validation results. This dual-validation approach provides a transparent quantification of the "autocovariance bonus" in the random split versus the "generalization penalty" in the LOYO split. Looking ahead, model training stratified by crop type may become feasible if more ample datasets become available. LOYO based validation could be more effective, as the LOYO structure would better reflect the temporal generalization challenges within each crop-specific subset.

2. Inconsistent NEE handling: It was stated that NEE is modeled independently from GPP and RECO. No justification was provided. More importantly, it could bring some inconsistent among variables.

While modelling NEE independently bypasses the formal constraint NEE=RECO-GPP, NEE represents the primary observable in eddy-covariance measurement, whereas GPP and RECO are derived through partitioning functions. In theory, these variables should satisfy the conservation equation NEE=RECO-GPP. However, in practice, enforcing this constraint during the ML stacking process involves a significant trade-off. To prioritize predictive accuracy, we thus modelled NEE independently. We also highlight the relevance of independent NEE estimation (like NEE is the prior for atmospheric inversions). Nevertheless, to address the question, we performed a consistency check by reconstructing NEE from the GPP and RECO model stacks, allowing us to quantify the trade-off between predictive performance and strict adherence to the physical mass-balance constraint. The results indicated very strong performance for the independent validation that is also a novel result.

3. Persistent bias at C1: At this site, all models produce persistent biases. Yet, no reason was provided. It is difficult to accept that as exception.

Thank you for raising this point regarding the persistent bias at site C1 (Ottawa, Canada). We agree that the persistent bias at C1 is noteworthy, although its origin is not straightforward to diagnose. The bias is likely rooted in the input features themselves (namely, the individual process-based models that form the ensemble). Because all meta-models exhibit a similar shift, the evidence points to a systematic bias in the underlying process-based models, which appear unable to capture certain site-specific characteristics at C1 (e.g. phenological patterns or management-related dynamics). In such cases, the meta-model is necessarily constrained by the collective behaviour of its inputs. Importantly, despite this inherited bias, the stacking framework substantially reduces the magnitude of the error relative to the Multi-Model Median and individual process-based models. As shown in Figure 3, the meta-models yield a noticeably tighter alignment

along the 1:1 line, indicating that the approach extracts the maximum available information even when the ensemble is systematically shifted. A full diagnosis of the site-level drivers of this bias would require a dedicated, site-specific analysis that extends beyond the scope of the present study. Given the already expanded manuscript - particularly due to the addition of the LOYO validation - we opted not to include a more detailed investigation here. Nonetheless, we agree that this remains an interesting direction for future work.

Overall, the methodological questions seem too important. Thus, only a few minor comments are given here.

Minor comments

1. Line 190-192: Provide quantitative justification for excluding Indian site ("relatively poor temporal coverage" is vague)

The Indian site was excluded from the analysis primarily due to methodological inconsistencies. Specifically, fluxes were measured using chamber-based techniques rather than eddy covariance, which characterizes the remainder of the dataset. Furthermore, the absence of concurrent GPP and NEE observations at this site precluded its use in our multi-variable validation framework. We extended the text to explain the decision.

2. Line 244: Specify: "randomly" - with what seed? Same split across meta-models?

We used the random number generator routine of base R. Seed was not specified explicitly. In such cases, according to the documentation of R, the system time and process ID of R is used internally. It is random in this sense, as much as it is according to computer science. We aimed methodological reproducibility instead of raw numerical reproducibility. The results is typically independent from the seed selection.

3. Line 245: Specify complete XGBoost hyperparameter tuning procedure

The hyperparameter optimization was performed by latin hypercube sampling. The exact ranges will be seen in the code after refactoring the codebase to easier replication.
For the random split strategy random sampling-based optimization was used, with bounds defined in the to be completed appendix.

4. Line 275: "site-level observations" - clarify this means temporal, not spatial, validation

We extended the previous sentence to make it clear that we performed temporal validation. We also added a statement in the "4.4 Limitations and future research" section clarifying that while the models show high temporal fidelity, spatial extrapolation to unobserved sites was not the objective of this work.

5. Lines 295-296: Provide specific methodological justification for independent NEE modeling or change approach

This issue was also raised by the other Reviewer. While we recognize that modelling NEE independently bypasses the formal constraint of NEE=RECO-GPP, NEE represents the primary measurement in eddy covariance, whereas GPP and RECO are products of partitioning functions. To maximize predictive accuracy, we modelled NEE independently. We then performed a consistency validation by reconstructing NEE from the GPP and RECO stacks to quantitatively demonstrate the trade-off between model performance and adherence to physical consistency. Overall, modelling NEE is also justified in many situations (e.g. to provide priors to atmospheric inversions), which means that is is not wise to omit modelling NEE. We are not aware of any approach where the consistency of GPP, RECO and NEE is retained after multimodel construction (MMM already breaks the consistency). Nevertheless, the results indicate that the independent NEE calculation (when NEE is consistent with GPP and RECO) is a good alternative of the meta-model based NEE.

6. Tables 3-5: Add statistical significance indicators comparing meta-models to MMM

As there is a clear tendency in the statistical indicators as we move to more sophisticated methods, we believe that there is no need to provide significance. It would not change the results, but perhaps move the focus from the main point to some other direction. We hope that the Reviewer accepts our decision.

7. Figure 11: Add legend explaining color intensity mapping and improve interpretability

Done

REFERENCES

Papale, D., Reichstein, M., Aubinet, M., Canfora, E., Bernhofer, C., Kutsch, W., Longdoz, B., Rambal, S., Valentini, R., Vesala, T., Yakir, D. (2006). Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: Algorithms and uncertainty estimation. Biogeosciences 3, 571-583. https://doi.org/10.5194/bg-3-571-2006

Moffat, A.M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A.D., Barr, A.G., Beckstein, C., Braswell, B.H., Churkina, G., Desai, A.R., Falge, E., Gove, J.H., Heimann, M., Hui, D., Jarvis, A.J., Kattge, J., Noormets, A., Stauch, V.J. (2007). Comprehensive comparison of gap-filling techniques for eddy covariance net carbon flux data. Agricultural and Forest Meteorology 147, 209-232. https://doi.org/10.1016/j.agrformet.2007.08.011

Reply to the comments
RC2: 'Comment on egusphere-2025-4920', Anonymous Referee #2, 03 Jan 2026

We thank Anonymous Referee #2 for reviewing the manuscript (egusphere-2025-4920). Thank you for the positive words. Here we answer the questions and issues raised by the Reviewer in detail.

Note that the comments of Anonymous Referee are shown below *in italic*. Our responses to the comments are presented below in normal font style.

The manuscript proposes a stacking meta-modelling framework to combine outputs from multiple process-based ecosystem models for prediction of GPP, RECO and NEE using advanced machine learning approach in comparisons to the classical approach of the multi-model median, MMM. The idea is interesting, because a stacking approach can improve predictive skill and can also provide diagnostic insight into model strengths and weaknesses.
I think that in its current form, however, the manuscript requires major revisions. The two most important issues are: a) lack of conceptual clarity and terminology (the narrative is currently difficult to follow for a broad audience), and b) a week validation strategy for time-series data (I'm not completely sure that I have understood well), but a random split on daily time series risks to increase the performance of the stacked metamodel due to the strong autocorrelation in time, and c) I'm not fully convinced that T and P can be used efficiently in the stacking proves, because these variables are already used as (very important) driving variables in all the models that are a part of the ensemble.
I think that with a clearer framing and a time-aware evaluation protocol (or a robustness analysis demonstrating that conclusions hold), the paper could become a relevant contribution to the creation of high-quality simulations.

Thank you very much for the notes. Although we tried to present the manuscript in a straightforward way, it seems that some adjustments are truly needed. In the revised manuscript we changed the text and added clarifications for better readability. Below we also address the issues raised concerning the random split method. We performed additional calculations to demonstrate the applicability of the methods and the value for future studies.

Recommendation: *Major revisions.*
My main concern is that (or at least I have a strong A-priori) that the validation on daily time series must be time-aware and random splitting is not adequate on its own. If training/validation is done via random 70/30 splits on daily data (sorry if I'm wrong, I'm not really sure that validation is done in this way), the validation set is not independent due to strong temporal autocorrelation and seasonality. This can inflate performance and may mislead readers about generalisation to new periods.
So, I invite the authors to replace or complement random splitting with a time-aware strategy: blocked cross – validation, contiguous hold-out blocks or similar. Due to the hyperparameters tuning, please ensure the protocol avoids optimistic bias

Thanks for the relevant note. Indeed, the validation was done using the 70/30 split logic (see lines 238-239 in the original manuscript). Indeed, this validation approach is not correct for temporal generalization or for prediction. However, it is still acceptable for within regime flux reconstruction, that was the scope of the study. We highlighted this issue in the revised manuscript.
In order to address the issue we made an alternative validation using the leave-one-year-out (LOYO) approach. The results will be included in the revised manuscript. Not surprisingly, in croplands the approach was not suitable, due to the presence of crop rotation, meaning that different crop types are grown in consecutive years. In this context, since the training dataset covers only a few years for crops, leaving an entire year out leads to underperforming model construction. For grasslands the method worked remarkably well, due to the same (or slowly changing) community structure/plant functional type. This method implicitly resolved the temporal autocorrelation issue.
We extended the materials and methods, plus the results section with tables quantifying the LOYO validation results.
In a future study, perhaps we can train the models for different crop types if ample data is available. In that case the LOYO based validation might work better.

The Interpretation of the "Stacking+met" in my opinion needs a stronger rationale. Adding meteorological covariates may be useful, but process models already embed meteorological forcing internally. I would like to understand if Stacking+met is still primarily an ***aggregation method*** *or whether it becomes a broader statistical correction that partially bypasses process constraints. Please, could you explain if "Stacking+met" represents conceptually: regime-dependent weighting? residual correction? hybrid modelling? Thanks, even the question seem to be theoretical only I think that is very relevant for interpretation: is it improved the combination of models, or learning a shortcut from meteorology that compensates for shared structural biases? Where possible, may be very interesting to explant the consideration about possible improvements in specific regimes (e.g., dry/wet, warm/cold, extremes).*

We thank the reviewer for this thoughtful and theoretically important question. We agree that distinguishing between model aggregation and residual correction is essential for interpreting the behaviour of the stacking+met approach.
Although process-based models incorporate meteorological forcing, they typically do so through simplified functional forms that cannot fully capture the diversity of temperature- and moisture-driven processes - particularly under extreme conditions or specific phenological stages. As documented in recent benchmarking (Bellocchi et al., 2023; Sándor et al., 2017; 2023) and model-specific (Sándor et al., 2018) studies, these simplifications lead to systematic, regime-dependent errors across models.
Within this context, XGBoost with environmental covariates (stacking+met) can be viewed as a hybrid modelling framework. By adding meteorological covariates to the stacking layer, the meta-model XGB+ performs two complementary functions:
- Regime identification: it detects environmental conditions (e.g. high VPD, low soil moisture) under which certain process-based models consistently over- or under-perform.
- Regime-dependent weighting: rather than applying a single global weight, the meta-model adjusts the influence of individual process-based models based on the prevailing meteorological state, thereby improving predictions in regimes where structural process-based model limitations are most pronounced.

We acknowledge that this behaviour could be interpreted as learning a “shortcut” that compensates for shared structural biases. However, we view it more appropriately as an empirical error-correction mechanism that uses meteorological context to address systematic residual patterns that process-based models do not resolve. This does not bypass process constraints, but rather complements them by correcting predictable deviations.
We have added text to the Discussion (Section 4.1) clarifying that Stacking+met represents a hybrid, regime-dependent integration framework rather than a purely statistical aggregation. While a detailed regime-specific analysis (e.g. dry/wet, warm/cold, extremes) is beyond the scope if this baseline study, we now explicitly highlight this as a promising direction for future research.

In my opinion modelling NEE independently requires justification. If NEE is modelled separately rather than derived from RECO–GPP, this breaks a key consistency relationship that many readers will expect, so, there is the need to clearly justify the methodological reasons for independent NEE modelling, discussing the trade-off between improved NEE estimation at the cost of the loss of consistency among GPP/RECO/NEE.

This issue was also raised by the other Reviewer. While we recognize that modelling NEE independently bypasses the formal constraint of NEE=RECO-GPP, NEE represents the primary measurement in eddy covariance, whereas GPP and RECO are products of partitioning functions. To maximize predictive accuracy, we modelled NEE independently. We then performed a consistency validation by reconstructing NEE from the GPP and RECO stacks to quantitatively demonstrate the trade-off between model performance and adherence to physical consistency.

There is the need to Clarify concepts and use consistent terminology (ensemble / stacking / meta-modelling). In my opinion the Introduction and M&M mix related concepts in a way that makes it hard to understand what is exactly being proposed and what is new relative to existing practice. In my opinion, early in the Introduction, there is the need to provide a single, clear definition of what is the baseline (MMM), what is the stacking here (there is a confusion between techniques that use stacking (e.g. RF) and the creation of a model stacking using that techniques). This approach will allow to use after an uniform and simple naming scheme.

We thank the reviewer for highlighting the need for clearer and more consistent terminology. We agree that a well-defined conceptual hierarchy is essential for understanding what is proposed and what is new relative to existing practice.
To address this, we have revised the Introduction and Materials & Methods to establish a unified terminology:
- Baseline (MMM): The Multi-Model Median is defined as the unweighted central tendency of the process-based model outputs and serves as the benchmark for evaluating all subsequent methods.
- Base models: These are the individual process-based ecosystem models (e.g. DNDC or APSIM) that generate the initial flux estimates.
- Stacking (Meta-Modelling): This refers to the ensemble-learning framework in which the outputs of the base models are used as predictors for a second-level learner.
- Meta-Models: These are the machine-learning algorithms (MLR, RF, XGB, XGB+) that implement the stacking framework.
To avoid the ambiguity noted by the reviewer, we now explicitly distinguish between RF as a standalone predictive model and RF used as a meta-model within the stacking architecture. In this

study, RF is used exclusively in the latter role - to learn the optimal combination of process-based model outputs.
The new text in the Introduction in the context reads as: “The results of this meta-model are compared with the multi-model ensemble median and other meta-models to demonstrate the improvement, interpretability and reliability of ensemble predictions (where multi-model median is considered as the baseline).”
Considering the terminology of stacking we modified the text to make it clear and avoid ambiguity. In the context of the study, stacking only means a specific ensemble modelling technique frequently used in machine learning. In this study it used to construct a new meta-model from the output of ecosystem models which are the “base models” in the stacking terminology.
In this sense stacking is an alternative expression for combining multiple models and getting a new estimation. The first sentence of the original Abstract already pointed in this direction.
Nevertheless, we added new text to the Materials and methods to put down a clear and straightforward terminology: “Thus, throughout the study, stacking means an ensemble learning technique, where the multi-model framework is used to construct a new estimation for the target variables.”
Considering the novelty of the study we added a sentence to the end of the Introduction stating that “The novelty of the study is the interpretability of the multi-model system and the linkage with environmental factors to improve performance of the stacking method. Additional novelty is the application of multiple meta-modelling approaches that is an abstract framework to provide improved modelling exercises.”

Please consider removing/shortening conceptual digressions that are not needed for the core message (e.g., broad “no free lunch” statements), unless directly tied to the study design and results, or put a detailed mathematical support in supplemental materials.

It was deleted.

Please consider too these minor comments: 1) consider a clearer separation into process-model ensemble generation, meta-modelling approaches, validation protocol, end interpretability and diagnostics. 2) Be uniform in Acronyms and naming: define each acronym once and use consistently throughout (MMM, stacking, XGB, etc.). 3) when using SHAP, specify exactly what dataset it refers to (site/flux/season, random selection). 4) ensure that the validation scheme is explicit in figure captions

We addressed all four comments. 1) we presented a clear methodology in the MS. 2) we adjusted the text to define all acronyms only once. 3) done. 4) it is also adjusted. Note that we added LOYO validation as well, and now the main text focuses on that.

REFERENCES

Bellocchi, G., Barcza, Z., Hollós, R., Acutis, M., Bottyán, E., Doro, L., Hidy, D., Lellei-Kovács, E., Ma, S., Minet, J., Pacskó, V., Perego, A., Ruget, F., Seddaiu, G., Wu, L., Sándor, R., 2023. Sensitivity of simulated soil water content, evapotranspiration, gross primary production and biomass to climate change factors in Euro-Mediterranean grasslands. Agricultural and Forest Meteorology 343, 109778. https://doi.org/10.1016/j.agrformet.2023.109778

Sándor, R., Barcza, Z., Acutis, M., Doro, L., Hidy, D., Köchy, M., Minet, J., Lellei-Kovács, E., Ma, S., Perego, A., Rolinski, S., Ruget, F., Sanna, M., Seddaiu, G., Wu, L., Bellocchi, G., 2017. Multi-model simulation of soil temperature, soil water content and biomass in Euro-Mediterranean grasslands: Uncertainties and ensemble performance. European Journal of Agronomy 88, 22–40. https://doi.org/10.1016/j.eja.2016.06.006

Sándor, R., Picon-Cochard, C., Martin, R., Louault, F., Klumpp, K., Borras, D., Bellocchi, G., 2018. Plant acclimation to temperature: Developments in the Pasture Simulation model. Field Crops Research 222, 238–255. https://doi.org/10.1016/j.fcr.2017.05.030

Sándor, R., Ehrhardt, F., Grace, P., Recous, S., Smith, P., Snow, V., Soussana, J.-F., Basso, B., Bhatia, A., Brilli, L., Doltra, J., Dorich, C.D., Doro, L., Fitton, N., Grant, B., Harrison, M.T., Skiba, U., Kirschbaum, M.U.F., Klumpp, K., Laville, P., Léonard, J., Martin, R., Massad, R.S., Moore, A.D., Myrgiotis, V., Pattey, E., Rolinski, S., Sharp, J., Smith, W., Wu, L., Zhang, Q., Bellocchi, G., 2023. Residual correlation and ensemble modelling to improve crop and grassland models. Environmental Modelling & Software 161, 105625. https://doi.org/10.1016/j.envsoft.2023.105625