Dear Editor,

Thank you for your positive assessment and for recognizing the substantial revisions we made to address the referees' comments. We are particularly grateful for the additional constructive comments and especially for the request to conduct statistical tests to verify the reliability of the results. Conducting these tests will further enhance the value of the manuscript.

We provide the following responses (in a larger, regular font) to the reviewers' comments (in a smaller, bold font):

1. Statistical significance tests (RC1 Minor #6 — not adequately addressed). RC1 requested statistical significance indicators in Tables 3–5 comparing meta-models to MMM. The authors declined, arguing that "there is a clear tendency in the statistical indicators." This justification is insufficient. With only four sites, a definitive statement about the superiority of meta-models over MMM requires either formal testing or an explicit acknowledgement that the sample is too small for reliable inference. Given the small number of independent replicates, some of the reported improvements may not be statistically distinguishable from noise. Thus, the reviewer's (and my) concern is legitimate. The authors should either: (a) add appropriate tests to approve that the meta-model's RMSE is systematically lower than MMM's (e.g., paired t-test or better Wilcoxon signed-rank tests on RMSE or correlation values across years within each site), or (b) provide a substantive statistical argument for why such tests are uninformative or misleading in this specific context. Dismissing the request with "it would not change the results" is not an adequate response to a reviewer's scientific concern.

We thank the editor for raising this important concern and fully acknowledge that our original response was inadequate. We have now added comprehensive statistical testing using one-sided Wilcoxon signed-rank tests comparing each meta-model against the Multi-Model Median baseline (new Tables 4, 6, and 8 for GPP, RECO, and NEE). The methodology is described at the end of Section 2.3 (in yellow in the 'track changes' version): we performed separate tests for each LOYO fold and aggregated p-values using the median across years to avoid pseudoreplication, then applied Benjamini-Hochberg correction per site to control for multiple testing. The results (new texts in green in the 'track changes' version) strongly support our conclusions: 43 of 60 model-site-variable combinations show statistically significant improvements ($p < 0.05$, 72%), with particularly compelling evidence from the grassland sites where all five meta-models significantly outperform MMM for all three variables. Site C1 shows no significant differences, which we attribute to limited statistical power with only 2 test years, but the consistent and robust results from the other three sites provide strong evidence for the effectiveness of meta-modeling. We have also added discussion explaining cases where RMSE and Wilcoxon results diverge, noting that these metrics are complementary: RMSE quantifies error magnitude while Wilcoxon tests for consistent improvement across years.

2. Cropland LOYO limitations in the Abstract and Conclusions. The primary validation strategy presented in the abstract and conclusions is now LOYO-based, yet LOYO is effectively inapplicable to cropland sites C1 and C2, where it produced substantially worse results compared to the 70/30 split (e.g., RF+ RMSE LOYO = 4.72 vs. 1.95 for 70/30 at C1 in Table 3). Readers encountering the

LOYO-centred summary in the abstract or conclusions will not have this crucial context. Please add a concise caveat in both the abstract and conclusions clarifying that LOYO results apply primarily to grassland sites, and that 70/30 splits remain the reference for cropland performance assessment.

It is greatly appreciated that this important nuance was pointed out. We have now added caveats in both the abstract and conclusions to clarify that LOYO results are primarily applicable to grassland sites. See changes in green in the 'track changes' version.

3. Persistent bias at C1: The authors' response is acceptable, but it may be helpful to add one more sentence in Conclusions explicitly noting that the meta-modelling framework's performance is bounded by ensemble-level systematic biases at data-limited sites, or one interpretive sentence to use existing SHAP results to state whether C1's bias is driven by a few dominant models or distributed across all inputs.

We are grateful that we were given two options. We chose the first one and added an extra sentence in yellow in the 'track changes' version.

4. Table and figure numbering consistency. The track-changes version shows inconsistencies in table numbering (the RECO table is numbered Table 4 but referred to as Table 5 in text, and the NEE table is referenced as Table 5 but referred to as Table 7 in text). Please verify that all table and figure numbers, in-text cross-references, and captions are fully consistent and correct in the final clean version.

The identification of these numbering inconsistencies is appreciated. All table and figure numbers, in-text cross-references, and captions have been thoroughly checked and corrected to ensure full consistency in the final clean version.

5. Hyperparameter tuning appendix. The response to RC1 Minor #3 states that XGBoost hyperparameter ranges "will be seen in the code after refactoring" and are "to be completed in the appendix." The appendix must be complete in the submission — not deferred to future code commits. Please confirm that the full hyperparameter tuning procedure is documented in the manuscript or its appendix before resubmission.

The importance of including the full hyperparameter tuning procedure within the manuscript is acknowledged. To address this, a new Appendix B has been created (highlighted in yellow), which contains the comprehensive details of the XGBoost hyperparameter tuning procedure, including the full ranges evaluated, ensuring complete documentation in the current resubmission. We have also corrected a related sentence (in yellow, above Figure 1.) in Section 2.3.

6. LOYO NEE at C2 — brief clarification needed. Table 5 shows that for NEE at C2, the independent NEE consistency validation (INDEP) approach (NEE = meta-model GPP – meta-model RECO) achieves lower RMSE than the direct LOYO NEE prediction for the most advanced meta-models (RF+ RMSE: 2.06 vs. 2.22; XGB+ RMSE: 1.96 vs. 2.25), suggesting that the physically consistent reconstruction from separate GPP and RECO stacks can outperform direct NEE prediction in terms of random error. However, this comes at the cost of a systematic positive bias (~0.28–0.40 g C m^{-2} d^{-1}) that is essentially absent in the direct LOYO meta-models. This trade-off — improved precision but

degraded accuracy — deserves explicit discussion to help readers understand when INDEP is preferable to direct NEE stacking.

We are thankful for this insightful observation about the precision-accuracy trade-off between the INDEP and direct LOYO approaches at site C2, which indeed deserves explicit discussion. We have added clarification to the Results section (in blue) explaining that the INDEP approach achieves better precision (lower RMSE: 1.96 vs. 2.25 g C m^{-2} d^{-1} for XGB+) but introduces systematic bias (~0.3-0.4 g C m^{-2} d^{-1}), whereas direct LOYO maintains near-zero bias with slightly higher random error, and that this trade-off makes INDEP preferable for applications prioritizing precision and physical consistency (e.g., gap-filling, carbon budgets) while direct stacking is better suited when unbiased estimates are essential (e.g., model validation, trend analysis). This addition helps readers understand when each approach is most appropriate for their specific applications.

Minor Editorial Notes

• Line 410-415 / Table 3 (track changes version): The statement that correlation increases "by a maximum of 0.11 for the best-performing meta-model compared with the MMM" refers to the 70/30 strategy. Ensure this is clearly labelled, as LOYO gains are generally smaller. Currently, "In case of within-regime validation (70/30)" started a new sentence.

The incomplete sentence in question was left in the manuscript from an earlier version. We have deleted it in the revised version.

• Ensure writing clarify: after a quick look at the first sentence of the abstract, I noticed a typo: "We evaluated five stacking-based meta-models - Multiple Linear Regression, Random Forest, XGBoost, and also Random Forest and XGBoost" — XGBoost appeared twice. I also noticed several other errors in the text and figure numbering (some mentioned above). Please conduct a careful examination of the manuscript and supplement to eliminate typos and errors throughout the manuscript.

We have conducted a thorough and careful proofreading of both the main manuscript and the supplementary materials. All identified typos, grammatical issues, and minor formatting errors have been corrected to ensure the highest level of clarity.

Typos:

"include e measurements" (Section 2.2) was corrected to "include measurements"

"70/30at G4" (Section 3.1, under Fig. 3 analysis) was corrected to "70/30 at G4"

"the70/30" (Section 4.2) was corrected to "the 70/30".

"in all three metric" (Section 3.3, Table 7 analysis) was corrected to "in all three metrics"

"method to goes beyond" (Section 5, Conclusions) was corrected to "method that goes beyond"

Grammatical Issues:

Abstract: "In case of the independent validation strategy correlation increase was..."

was changed to:

"In the case of the independent validation strategy correlation increase was..."

Section 1, Introduction: "Additionally, the application of multiple meta-modelling approaches providing an abstract framework for …"

was changed to:

"Additionally, the application of multiple meta-modelling approaches provides an abstract framework for …"

Section 2.3: "RECO can be erroneously simulated if one of the two major components are misrepresented..."

was changed to:

"RECO can be erroneously simulated if one of the two major components is misrepresented..."

Conclusions: "...new combinations of models are tested that we call here as meta-models."

was changed to:

"...new combinations of models are tested that we call meta-models."

Conclusions: "...those multi-model ensemble techniques are still at their infancy..."

was changed to:

"...those multi-model ensemble techniques are still in their infancy..."

Section 3.3: "...scatterplots comparing the observations and the models The figures were constructed..." A period was inserted after "models".

Conclusions: "Without the environmental variables continuous retraining is needed to maintain accuracy." A comma was added after "variables".

Section 3.3: "meta-model based" was changed to "meta-model-based".

In Table 4, "gC m^{-2} day^{-1}" was changed to "g C m^{-2} day^{-1}" to be consistent with Table 3.

• Please also don’t forget to address the editorial system’s “Notification to the authors:

1. Coloured or marked text in *.pdf manuscript is not allowed. For the next revision, please provide clean version of the manuscript (with the black text) 2. For the next revision, please use the initials instead of the full names of authors in the section "Author`s contribution".”

A clear version is uploaded and the "Author`s contribution" section was corrected accordingly.