In the submission by Wangari et al., the authors present the development of a random forest model to predict GHG fluxes at high spatial resolution for a study area in central Germany. The authors apply state of the art methods on a comprehensive and interesting dataset. The topic is of interesting to the readership of Biogeosciences. I have a few concerns regarding the RF model development that I wish to see addressed before the article can be considered for publication.

In my opinion, the data does not substantiate the development of a model at 1 m spatial resolution. The only variables that truly convey information at that scale are the ones derived from the DEM and they are not dominant in the important predictor variables in the CD models. The soil properties are interpolated using a simple interpolation routine. There exist a large body of literature on soil mapping and interpolation (also using ML based approaches like RF or geostatistics) and I find the applied approach too simplified to support a 1 m resolution.
I would recommend to conduct the modelling at 10 m spatial resolution instead; meaning applying the IDW interpolation of the soil properties at 10 m and aggregating the DEM to 10 m as well.
Response: Thank you for your critical comment and suggestion. We agree that predicting landscape GHG fluxes at 10 m resolution would best reflect our available 10 m resolution remotely sensed data for the most important predictor variables (Sentinel-2 datasets). To take up your suggestion, we remodeled the GHG fluxes to a 10 m spatial resolution. Additionally, we also compared the measured versus the predicted flux values at 1 m and 10 m resolutions (Figures 1 and 2, respectively). We found that the resolution does matter, i.e., a finer resolution of 1 m is better in representing the measured fluxes, particularly for N_2O and CH_4 fluxes that are either sinks or sources over short distances than the coarser 10 m resolution. This finding is surprising as we expected (probably you did too) that the latter coarser resolution would better model the measured fluxes because of the much lower uncertainties linked to the downscaling of the predictor variables. We attribute this finding to the fine-scale heterogeneities that have been extensively reported on soil GHG fluxes, which are better represented by the finer-scale resolution despite the uncertainties introduced by the statistical downscaling. Based on these results, we have decided to stick with the 1 m resolution. We hope that you also share the same opinion after seeing these results.

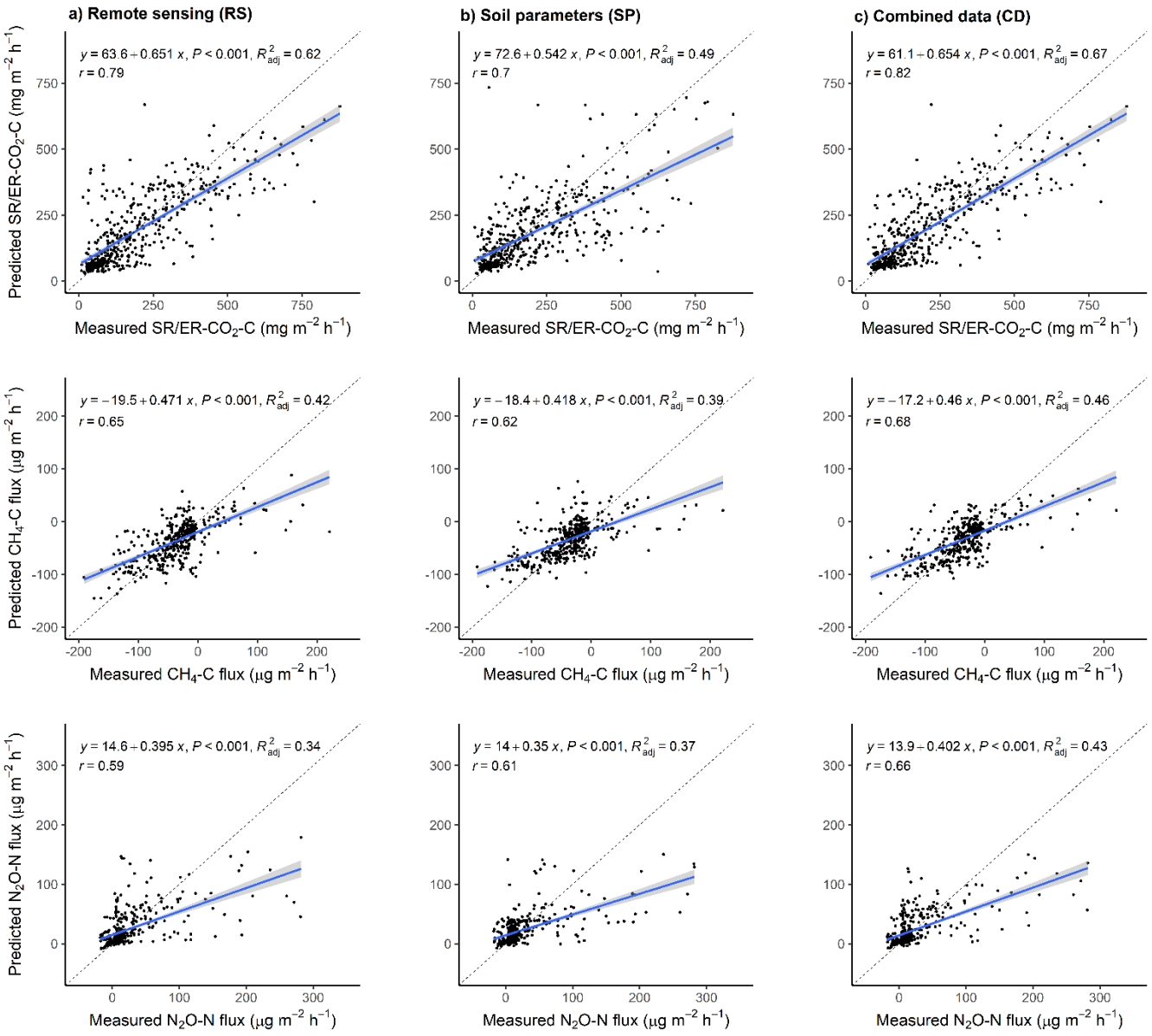

Figure 1: Linear regressions (with 95% confidence bands) of the measured and predicted GHG fluxes at 1 m resolution using remotely sensed data (RS), soil physico-chemical parameters (SP), and combined data (CD). The GHG fluxes from all the sampling point locations were included in this regression analysis.

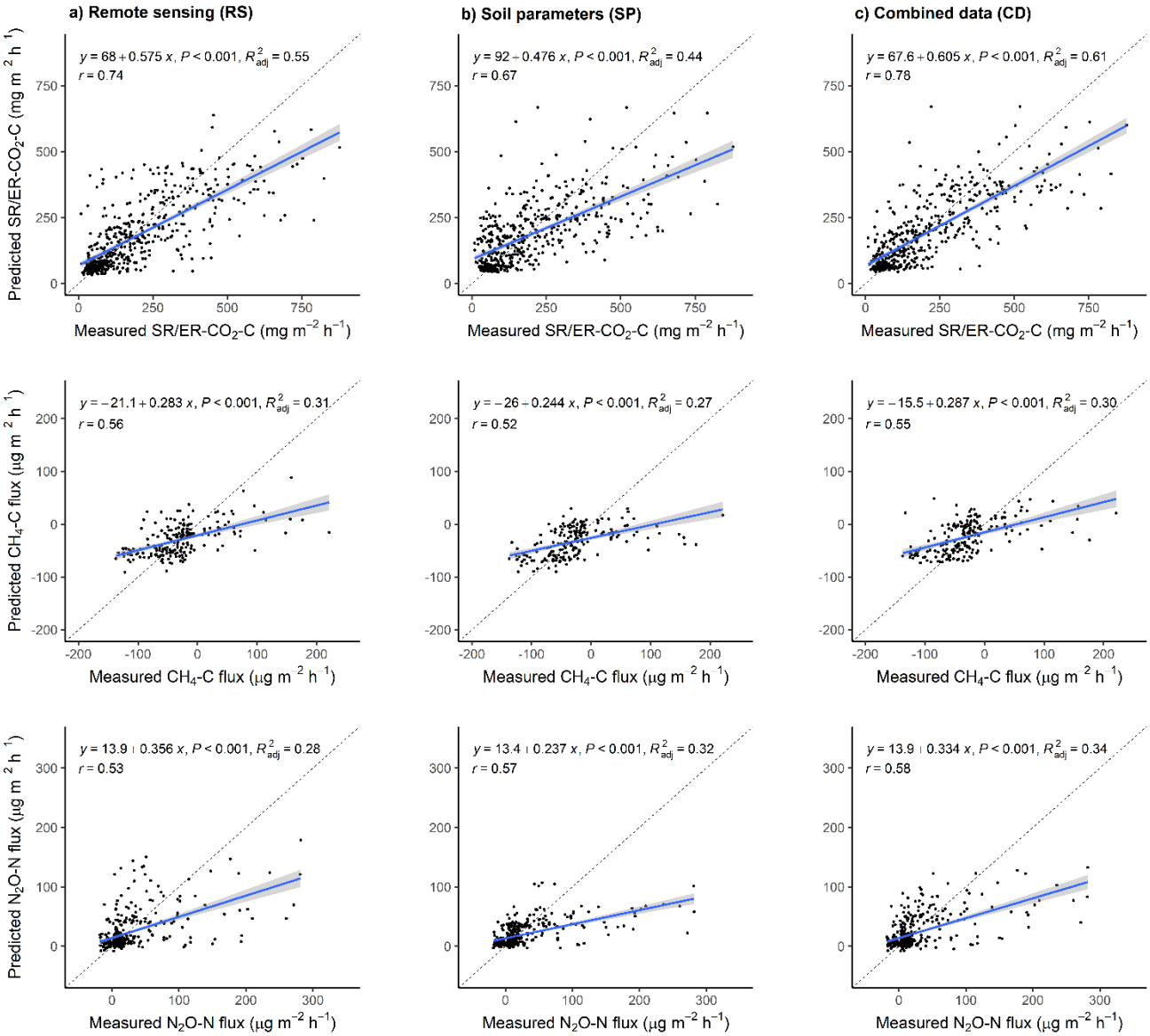

Figure 2: Linear regressions (with 95% confidence bands) of the measured and predicted GHG fluxes at 10 m resolution using remotely sensed data (RS), soil physico-chemical parameters (SP), and combined data (CD). The GHG fluxes from all the sampling point locations were included in this regression analysis.

The RF models are trained individually for the three land use classes while the summer and autumn data are treated jointly. I would expect that a RF model could easily utilize information from a land use map as additional predictor variable. Especially if the GHG fluxes show significantly different distributions across the three land use classes.

For each of the three fluxes one model could be trained using data jointly from the three land use classes as well as both seasons. In line with the argument of the authors that joining summer and autumn trains more robust models, I would expect the same for including data from diverse land use classes.

Response: Thank you for your critical comment. We did consider adding land use as an extra predictor variable and then using a single-trained model to model the entire landscape, similar to what we did for our seasonal data. However, we trained models for each land use to separately investigate the main predictor variables of the fluxes in the individual land uses. This approach was made to identify the underlying drivers of spatial heterogeneities of soil GHGs in each land use, which would have been lost if we built models for the entire landscape. While we see the advantage of having one robust model for the whole landscape, we believe the benefits of separating by land use are more significant. For starters, having separate predictors by land use means that one can infer different process mechanisms for each land use. Secondly, the land use-specific information on best predictors can also be used as a benchmark by other people interested in using a similar modeling framework to model homogenous landscapes regarding land use. We have added this rationale to the methods section.

"…..Modeling land use-specific GHG fluxes also enabled the identification of the best remotely-sensed predictors as the dominance of individual GHG production, consumption, and processes may vary in dependence of land use. These best predictors can also be used as benchmark parameters in future studies that use a similar modeling framework to model GHG fluxes in single land-use landscapes."

It is unclear whether the data plotted in Fig 3 are the 10-fold CV or the 30% test data. It should be the 30% test data that is being evaluated here. Also, it would be very interesting to see the model's performance for the 30% test data reported in a similar way as the 70% used in the 10-fold cv evaluation in Table 2. In this way, the model's robustness can be evaluated.

Response: Thank you for your critical comment. In this study, we used two main methods for validating our model. The first one is the traditional hold-out method. In this method, the data set was split into the training set (70%) and the testing set (30%). However, after further research into ML models, we realized that this validation method is biased as it depends heavily on which data points end up in the training set and which end up in the test set, and the evaluation may be significantly different depending on how the split is made.

To address this limitation, we switched our main validation method to a more sophisticated repeated k-fold cross-validation method. In this method, the data set was automatically divided into 10 subsets, and the hold-out method was repeated 10 times. Each time, one of the 10 subsets is used as the test set, and the other 9 subsets are put together to form a training set. Then the average error across all 10 trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once and gets to be in a training set 9 times. The variance of the resulting estimate is reduced as the K is increased, and the final model contains modeling parameters that are independent of the individual training or test datasets.

However, we still used this sophisticated method of k-fold cross-validation on 70% of our data and not 100% of the data because we were also interested in lowering the amount of data we used for our model building in the hope that future studies with lower sample numbers could still apply this model setup. The remaining 30% of the data was then used only as an extra external validation step for comparing the means of the measured and the predicted values and was also included in the supplementary material (Figure A1). Plotting this 30% dataset on a 1:1 line will not be appropriate as it

will introduce biases in the comparison due to the lower number of data points and lack of representativeness due to the simple random split approach. Therefore, Table 2 was only used to represent the results from the k-fold cross-validation method, which represents the true robustness of the model and is free from the biases discussed above.

Based on these explanations, Fig 3 includes all our data (100%) since the model training was independent of any single data point because of the sophisticated k-fold repeated cross-validation methodology, and the additional 30% data was totally not used in model training. We have edited the Fig 3 caption to make it clear.

"Figure 3: Linear regressions (with 95% confidence bands) of the measured and predicted GHG fluxes using remotely sensed data (RS), soil physico-chemical parameters (SP), and combined data (CD). GHG fluxes from all the sampling locations (both the 70% training data and the 30% test data) were considered in this regression analysis. The dotted line represents the 1:1 line."

Also, please discuss the limitations of a simple random split sample strategy taking inspiration in the following articles:

Bjerre, E., Fienen, M. N., Schneider, R., Koch, J., & Højberg, A. L. (2022). Assessing spatial transferability of a random forest metamodel for predicting drainage fraction. Journal of Hydrology, 612, 128177.

Meyer, H., & Pebesma, E. (2022). Machine learning-based global maps of ecological variables and the challenge of assessing them. Nature Communications, 13(1), 2208.

Response: Thank you for your critical comment and also the references shared. As mentioned above, and also referred to by the above articles, we realized that our initial simple random split sample strategy or the hold-out validation methodology was biased as it depends heavily on which data points end up in the training set and which end up in the test set, and the evaluation may be significantly different depending on how the division is made. To address this, we switched our main validation method to a more sophisticated repeated k-fold cross-validation method. Based on the description we have given of the method in our previous response, we do believe that the limitation related to the traditional simple and random test/training dataset split was minimized. We have added this information to the materials and methods section and cited the relevant papers.

"In addition to this the hold-out approach of model validation, we defined a ten-fold (K=10) repeated cross-validation scheme on the 70% dataset using the 'trainControl' function to internally validate our trained models and prevent model overfitting (Berrar, 2018). This model validation strategy also minimized the limitation of the initial hold-out approach, providing a more spatially robust model validation step (Meyer & Pebesma, 2022)."

It is unclear how random forest hyperparameters were set and if a sensitivity analysis or tuning has been carried out.

Response: Thank you for your critical comment. The random forest's most important hyperparameters (mtry= = number of variables at each tree, and n.tree = the number of trees) were tuned automatically within the caret package. Tuning was done automatically after a sensitivity analysis (based on MAE values) was performed ten times to choose the best mtry and n.tree resulting in the optimal trained model i.e., the one with the lowest MAE. We have added this information to the methods section.

"The random forest's most important hyperparameters (mtry = number of variables at each tree, and n.tree = the number of trees) were tuned automatically within the CARET package. Tuning was done automatically after a sensitivity analysis (based on assesing the mean absolute error: MAE) was performed 10 times to choose the best mtry and n.tree resulting in the optimal trained model, i.e., the one with the lowest MAE"

I lack a discussion on how many chamber measurements are needed for the proposed upscaling approach. This would be interesting for future design of upscaling experiments.

Response: Thank you for your critical comment. We have added text in the discussion section to reflect on this more.

"It is note worthhty that the applicability of this upscaling approach largely depends on the availability of spatially extensive chamber measurements. In this study, the 70% modeling dataset represented data from ~20 stratified chamber locations per km^2 on the arable land and ~16 chambers per km^2 in the forest. These number of chamber measurement locations are within the range of those recommended by Wangari et al., 2022 (29 for arable and 13 for forest) for accurate quantification of landscape GHG fluxes. Based on these findings, these chamber numbers may be adoptable to other studies looking to upscale GHG fluxes using a combination of chamber measurements and remotely-sensed data, but this will highly depend on the level of similarities in landscape properties with our study."

Moreover, the measurement campaigns were carried out over a little more than a week. How does day-to-day and diurnal variability introduce uncertainty to the dataset? I also assume that the predictor variable soil temperature can be affected by temporal variability. How did the authors account for that in their analysis?

Response: Thank you for your comment. We totally agree that day-to-day and diurnal variability can be misinterpreted as spatial variability based on our sampling strategy. To minimize the effect of day-to-day variability, we limited the duration of each campaign to 10 days. The diurnal effect was minimized by conducting measurements at random sites spread across different land uses and different parts of the landscape. We showed in our earlier publication (Figure S2, S3, S4, S5, and Table S2 in supporting information; Wangari et al. 2022) that day-to-day or diurnal variabilities were negligible on our datasets from each campaign. We have added this point in the methods and referenced it using our earlier publication.

"The day-to-day or diurnal variabilities related to our sampling strategy had a negligible effect on our data, with most of the variability in the data linked to spatial heterogeneities. Details of this finding as well as soil sampling, analysis, and flux measurement methods are described in Wangari et al. (2022)."

I think the authors should broaden their discussion up for alternative upscaling methods. Such a discussion should include process-based modelling and water table depth-based upscaling using empirical response functions.

Tiemeyer, B., Freibauer, A., Borraz, E. A., Augustin, J., Bechtold, M., Beetz, S., ... & Drösler, M. (2020). A new methodology for organic soils in national greenhouse gas inventories: Data synthesis, derivation and application. Ecological Indicators, 109, 105838.

Koch, J., Elsgaard, L., Greve, M. H., Gyldenkærne, S., Hermansen, C., Levin, G., ... & Stisen, S. (2023). Water-table-driven greenhouse gas emission estimates guide peatland restoration at national scale. Biogeosciences, 20(12), 2387-2403.

Response: Thank you for your comment. As requested, we have broadened our discussion on alternative methods of upscaling based on your recommendations and the references provided.

"This approach represents a Tier 3 approach of upscaling landscape GHG fluxes, as it provides spatially explicit GHG fluxes at a high resolution comparable to modeled fluxes using either process-based models or statistical functions (e.g., Haas et al., 2013; Tiemeyer et al., 2020; Koch et al., 2023)."

Maybe I missed it, but SR/ER_CO2 needs a clear definition.

Response: Thank you for your comment. We forgot to add the definition in the methods section of this study, as we had described the definitions in our earlier publication. We have now updated our current methods section to include a clear definition of SR and ER.

"The CO_2 fluxes quantified using the opaque chambers represented either soil respiration (SR) (root and microbial respiration) or ecosystem respiration (ER) (root, microbial, and plant respiration). The CO_2 measurements in autumn across the entire landscape were SR since above-ground biomass was not included in the chambers during measurements. In contrast, the summer CO_2 measurements on arable and grasslands were ER since the above-ground vegetation was incorporated using chamber extensions while the forest measurements remained as SR due to minimal above-ground vegetation on the forest floor."

I enjoyed reading the manuscript and look forward to seeing a revised version.

Response:

Thank you for your kind words and for taking the time to review our manuscript. We do appreciate the constructive feedback given.