# Referee 1

**General remarks**

This paper presents a multifaceted approach for modelling and upscaling spot data to landscape level. It is a follow-up of a paper (Wangari et al. 2022, JGR: Biogeosciences, 127, e2022JG006901) that showed the measured GHG fluxes in more detail. It is noteworthy that the data from the spring campaign (March 14-15) was left out of this modelling exercise.
I generally like the setup and believe it brings useful information to landscape and land use type assessments. There is not much to space for critics. The methods seem sound, text is well written and easily readable. However, it could benefit from more clarity in showing the improvement in upscaling prediction performance and restrictions of the measured GHG data in seasonal and spatial scales in Ch. 4.1. *That would result in moderate changes only.*
**Response**: Thank you for your overall comments and suggestions. We have addressed most of them in the current version of our manuscript. Please find below the detailed responses. Your comments were beneficial in helping us to provide clarity on the strengths and limitations of our study, which collectively offered the basis for further discussions on future research gaps in the modeling of landscape GHG fluxes.

Upscaling of GHG fluxes measured from micro to macro scale is hampered by spatial and temporal uncertainties of varying biological and physical origins. At the same time, an adequate increase of the frequency of chamber measurements is hard. This paper uses flux data, soil physical and chemical characteristics from different types of ground vegetation-soil systems for high resolution upscaling to landscape level with help of remote sensed parameters and indices, and DEM for Random Forest modelling by different land use types separately. Measurements of soil characteristics and GHG fluxes included two rather short campaigns in 2020. GHG's were measured daytime using opaque chamber and "fast box" techniques during late June-early July and September 8-17. Those probably yielded estimates relative non-winter emission strengths rather than annual fluxes for the sites.
**Response**: Thank you for your comments. We agree on the limitation of the study in terms of temporal measurements throughout the year. However, our study primarily focused on capturing the spatial heterogeneities in soil GHG fluxes. For that reason, we had to limit our campaigns to only a few days to reduce the risk of intra and inter-daily variations being mistaken for spatial variability. Regarding the representation of the annual fluxes, we agree that our summer and autumn measurements do not represent the fluxes from other seasons, e.g., winter. For that same reason, we only talked about the spatial variability of GHG fluxes during the summer and autumn seasons throughout our manuscript and refrained from talking about annual fluxes from seasonally resolved measurements, which was not the focus of our study. To make clear the limitation of our measured GHG data in terms of representing the seasonal trends of soil GHG fluxes in our landscape, we have added the following sentences to the conclusion, which is a potential research gap for the future.

"While we identified hot and cold spots of soil GHG flux across the Schwingbach landscape through RF modeling, the entire exercise was limited to two measuring campaigns of a few days in two seasons (summer and autumn). For this reason, it is still unclear whether these hot and cold spots persist throughout the year and their overall contribution to the annual landscape GHG flux estimates. Future studies should, therefore, aim at increasing the temporal resolution of similar spatially extensive measurements to at least monthly scales, which, when combined with remotely-sensed data, may be able

to create similar landscape flux maps and identify the contribution of GHG hot and cold spots to annual estimates."

In forest land, the trees may contribute to the measured soil $CO_2$ emissions through root respiration and add to uncertainties.

**Response**: Thank you for your comment. We agree that the contribution of root respiration to total soil respiration might be substantial in forests. We have now added a description in the methods section to reflect this view.

"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."

For $CO_2$, the opaque chamber flux represents a somewhat artificial sum of heterotrophic and autotrophic gas release, but not ecosystem net $CO_2$ exchange that could be measured using e.g. using transparent chamber or eddy covariance over a landscape or within separate land use types. Thus, the $CO_2$ fluxes could be hard to compare with literature data.

**Response**: Thank you for your comment. We agree that comparing net $CO_2$ fluxes to only those from respiration is misleading. However, extensive $CO_2$ flux comparisons to past literature values were done in our earlier publication and were limited to only fluxes quantified using opaque chambers (See Wangari et al., 2022). In this study, comparisons were only made on the prediction performance of RF models, which we also noted to be uncertain because of the different validation methods. These reflections were also in the discussion.

"Nevertheless, caution has to be taken when interpreting any conclusions from these study comparisons due to the limitations of different model validation techniques, different predictor variables used for modeling, and the different ecosystems and spatial scales of measurement and predictions."

The authors should elaborate in discussion how their results could be applicable e.g., in land use planning or mitigation efforts, given the representativeness of their data.

**Response**: Thank you for your critical comment. We agree that the $CO_2$ fluxes we measured with opaque chambers were not net $CO_2$ ecosystem exchange and, therefore, do not represent the net fluxes for $CO_2$, which would be suitable for mitigation measures. However, the common hotspot regions of all three gasses, including $N_2O$ and $CH_4$, representing net values in our study, were primarily within arable lands. Therefore, these common hot spot regions can be a target for mitigation strategies, considering that around 60% of anthropogenic $N_2O$ emissions come from arable lands. At the same time, agricultural land use also lowers the ability of soils to sink atmospheric $CH_4$. Land use planning measures such as targeted fertilizer regimes or expansion of local forests can play a vital role as local GHG mitigation solutions. These reflections were also included in the discussion.

"Identifying common patches with elevated emissions of the three GHGs can inform priority areas for implementing localized mitigation measures within a landscape. These common patches covered only 1.5% of our landscape (~0.2 $km^2$) and had the highest GHG fluxes contributing around 5%, 1%, and 8% of the landscape $CO_2$, $CH_4$, and $N_2O$ emissions. The location of these patches primarily (99.9%) on arable land emphasized the significant role of focusing on mitigating GHG fluxes from arable soils. Because

most of the common GHG hotspots in the arable soils were also in areas with high water content, mitigation strategies that aim at adjusting the fertilizer application rates at specific areas that hold more water may be successful in lowering the emissions (e.g., Hassan et al., 2022). In contrast to hot spot regions of elevated GHG emissions, $CH_4$ uptake hotspots inform future mechanisms for leveraging the GHG sink ability of soils, such as expanding local forests. This finding is supported by uptake hot spots identified on forest soils in this study, offsetting 8% of the total landscape $CH_4$ flux. The expansion of forested areas will also likely have a much higher mitigation impact via $CO_2$ sequestration."

The paper claims an improved prediction performance compared to other approaches in upscaling the mosaic of landscape GHG fluxes. Table 3 shows the approach of this study compared to that of other published studies using RF. It is however difficult to evaluate the *performance differences thereof with other types of approaches.*

**Response**: Thank you for your critical comment. We included Table 3 to compare with other studies that have used a similar RF approach. We agree that comparisons are difficult even with studies that have a similar RF approach due to differences in model validation, predictor variables, amount of data, etc. This limitation was also discussed in lines 380-383.

"Nevertheless, caution has to be taken when interpreting any conclusions from these study comparisons due to the limitations of different model validation techniques, different predictor variables used for modeling, and the different ecosystems and spatial scales of measurement and predictions."

Please explain clearly why the present approach is an improvement over others. Are there any relative qualitative or quantitative indices for such evaluation?

**Response:** Thank you for your critical question. Qualitatively, compared to the other studies, ours included more spatially well-distributed sites over a larger area that accounts for landscape GHG heterogeneities. In addition to remotely sensed data, we used more measured soil parameters to model the landscape GHG fluxes. This approach differed from previous studies focusing on remotely sensed data and a few soil parameters, such as soil moisture. All these points were highlighted in the discussion.

"Compared with other studies that have upscaled GHG fluxes using the random forest algorithm, we considered more site-measured data on soil parameters, all three GHG fluxes, and different land uses (Table 3). Moreover, point selections for measurements were done by implementing a stratified sampling plan that represented the spatial variability of several landscape characteristics, specifically land use, soil type, and slope (Wangari et al., 2022)."

Hot and cold landscape spots of emissions were identified and their contribution to overall GHG fluxes was evaluated. This is very useful for using the results in GHG mitigation.

**Response**: Thank you for the word of encouragement. It was indeed our hope that the information provided in our work could at least provide a starting work framework for the generation of detailed spatial maps, which in the future may be used to inform mitigation strategies.

**Specific remarks**

Lines 24-25 Please complete the comparison sentence since you make comparisons between approaches in Ch. 3.5 (Fig. 6) and in discussion Table 3 and elsewhere. Be more specific.

**Response**: Thanks for pointing out the comparison confusion. We have added Random Forest to the statement to avoid confusion with the area-weighted mean approach in Figure 6.

"Based on these field-based measurements and remotely-sensed data on landscape and vegetation properties, we used the Random Forest (RF) algorithm to predict GHG fluxes at a landscape scale (1 m resolution) in summer and autumn. The RF results showed improved GHG flux prediction performance when combining field-measured soil parameters with remotely-sensed data."

## Referee 2

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.

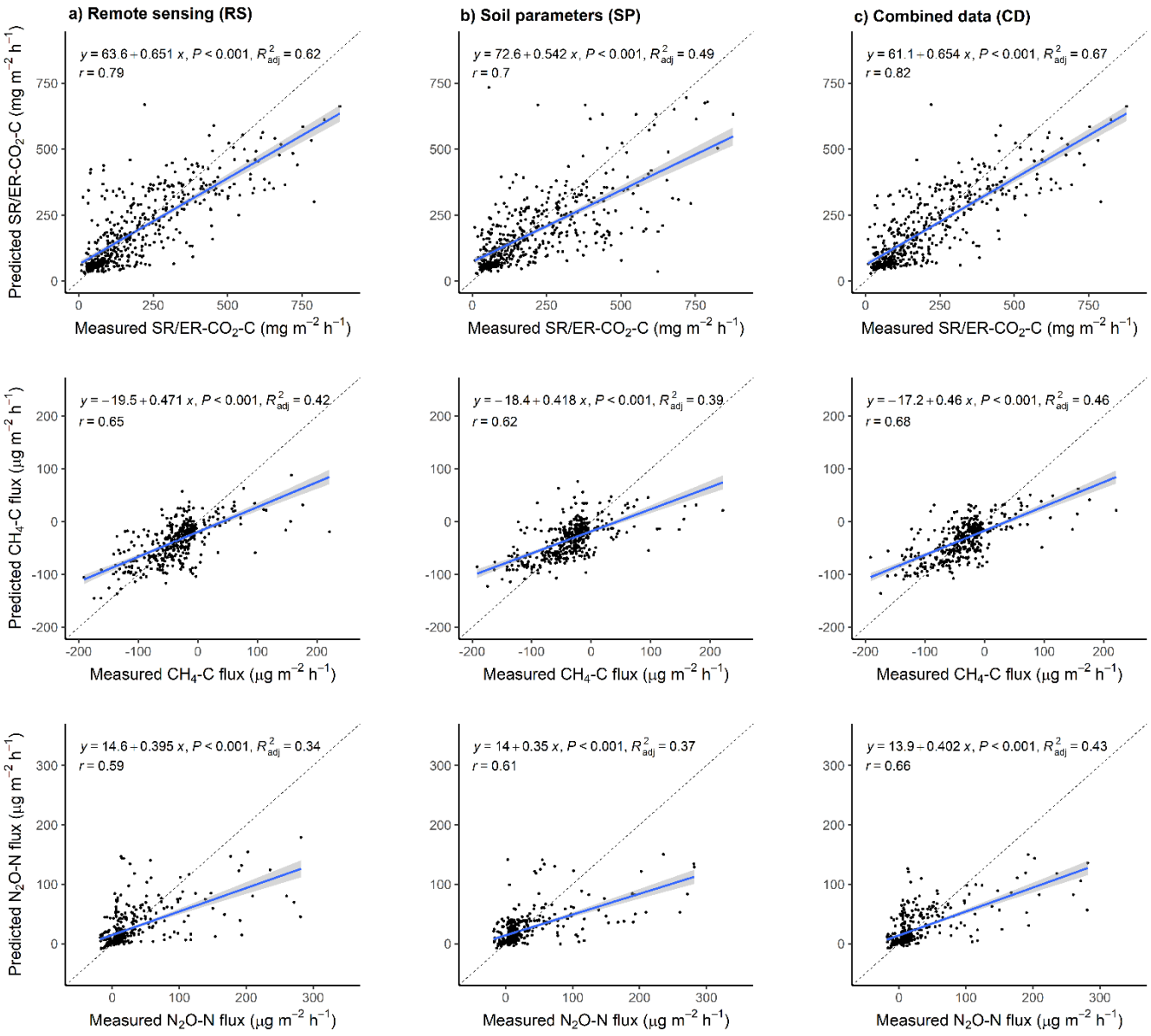


**Figure 1:** Linear regressions (with 95% confidence bands) of the measured and predicted GHG fluxes at 1 m resolution using remotely sensed data (RS), soil physico-chemical parameters (SP), and combined data (CD). The GHG fluxes from all the sampling point locations were included in this regression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

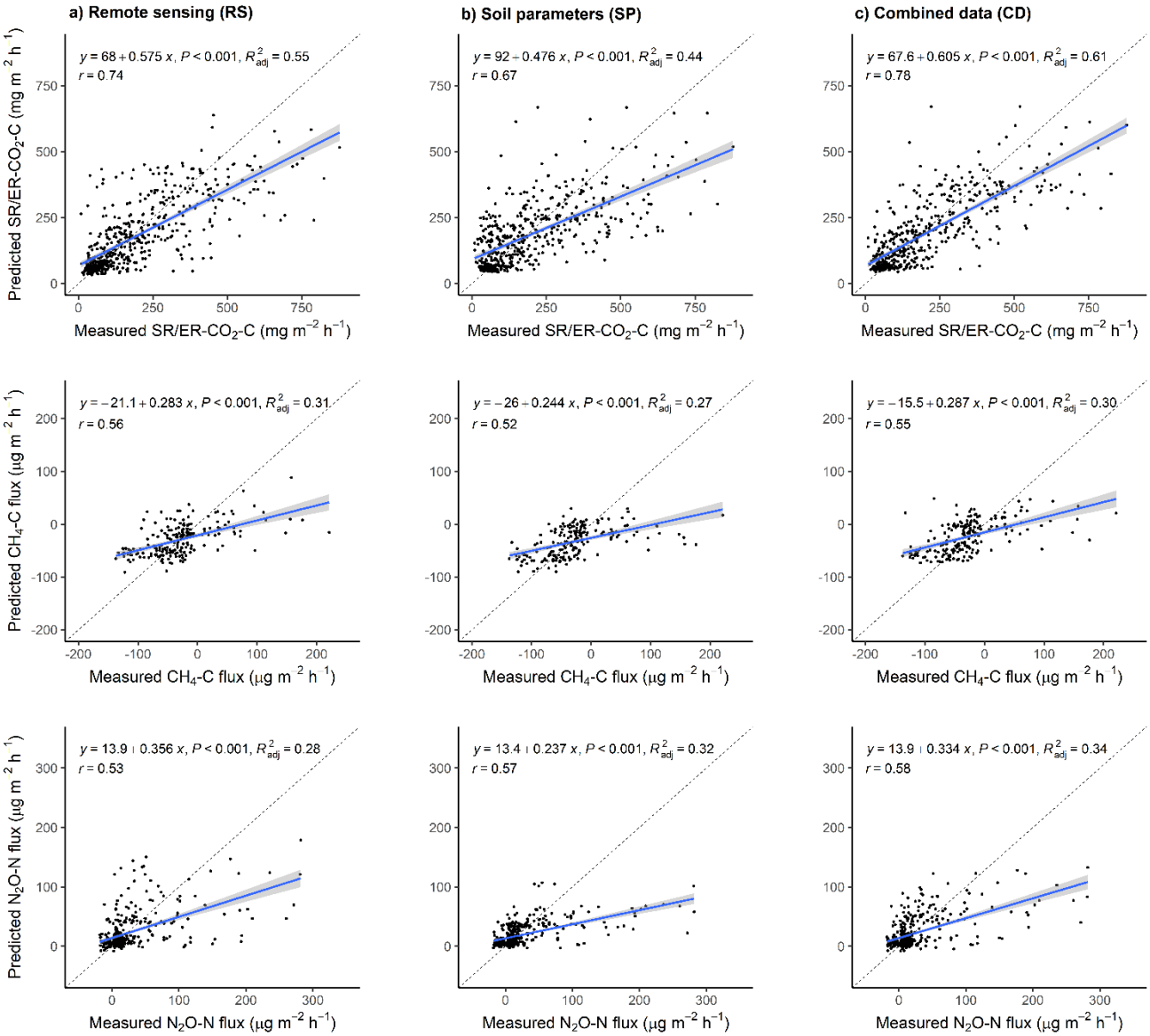


**Figure 2:** Linear regressions (with 95% confidence bands) of the measured and predicted GHG fluxes at 10 m resolution using remotely sensed data (RS), soil physico-chemical parameters (SP), and combined data (CD). The GHG fluxes from all the sampling point locations were included in this regression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

## Community Referee

**L24-25:** The results showed improved GHG flux prediction performance when combining field-measured soil parameters with remotely-sensed data.
**Comment:** "improved" compared with what?
**Response:** Thank you for the comment. We have made the changes to indicate that performance was better than models trained with isolated field-measured soil parameters and remotely sensed data only.
"The RF models combining field-measured soil parameters and remotely-sensed data outperformed those with field-measured predictors or remotely-sensed data alone."

**L28-30:** Similar seasonal patterns of higher soil/ecosystem respiration (SR/ER-$CO_2$) and nitrous oxide ($N_2O$) fluxes in summer and higher methane ($CH_4$) uptake in autumn were observed in both the measured and predicted landscape fluxes.
**Comment:** Are you really measuring ecosystem respiration with a "fast-box" technique? Aren't you missing above-ground respiration? Particularly for the forests.
**Response:** Thanks for raising this issue. We forgot to define the SR/ER-$CO_2$ fluxes in this study. The forest $CO_2$ fluxes were measured only on the forest floor with little or no above-ground biomass; thus, they were termed soil respiration (SR). The $CO_2$ fluxes measured on grassland and arable land in autumn were also categorized as soil respiration (SR) since the grass was mowed and the arable fields were harvested and plowed. However, the arable and grassland measurements in summer were termed ecosystem respiration (ER) since we incorporated the above-ground biomass using chamber extensions. We have added these details in the methods section.
"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."

**L59-62:** Nevertheless, the practicability of increasing the number of chamber measurement locations to quantify landscape fluxes is constrained by extensive human and technical resource requirements, hence there is a need for alternative ways of estimating GHG landscape fluxes.
**Response:** Thanks for the grammar correction of practicability to practicality. We have made the changes.

**L69-71:** The RF algorithm has been widely applied to gap-fill and upscale soil GHG fluxes in temperate ecosystems from point measurements to larger scales, with relatively better prediction accuracies (e.g., Philibert et al., 2013; Räsänen et al., 2021; Vainio et al., 2021).
**Comment:** Better compared with what?
**Response:** Thanks for the comment. We have adjusted the statement since the comparison of the ML algorithms was already made in the previous statement (L67-69).
"The RF algorithm has been widely applied to gap-fill and upscale soil GHG fluxes in temperate ecosystems from point measurements to larger scales (e.g., Philibert et al., 2013; Räsänen et al., 2021; Vainio et al., 2021)."

**L88-95**: In this study, we aimed to determine the potential of applying the RF algorithm to predict the spatial and seasonal variability of soil $CO_2$, $CH_4$, and $N_2O$ fluxes using a high number of stratified sampling locations (n = 268) spread across a relatively large (~5.8 $km^2$) landscape with heterogeneous land uses (forest, grassland, and arable land). Specifically, we aimed to: (a) evaluate the effectiveness of high-resolution RS data and relatively low-resolution data on soil physico-chemical parameters in predicting soil GHG fluxes across different land uses; (b) predict high-resolution soil GHG fluxes at a landscape scale and detect GHG hot spots and cold spots; and (c) compare landscape GHG fluxes upscaled from RF-predicted high-resolution maps with aggregated landscape flux estimates from averaged (point) fluxes multiplied by landscape area.
**Response:** Thanks for the editorial changes.
"In this study, we determined the potential of applying the RF algorithm to predict the spatial and seasonal variability of soil $CO_2$, $CH_4$, and $N_2O$ fluxes using a high number of stratified sampling locations (n = 268) spread across a relatively large (~5.8 $km^2$) landscape with heterogeneous land uses (forest, grassland, and arable land). Specifically, we: (a) evaluated the effectiveness of high-resolution RS data and relatively low-resolution data on soil physico-chemical parameters in predicting soil GHG fluxes across different land uses; (b) predicted high-resolution soil GHG fluxes at a landscape scale and detected GHG hot spots and cold spots; and (c) compared landscape GHG fluxes upscaled from RF-predicted high-resolution maps with aggregated landscape flux estimates from averaged (point) fluxes multiplied by landscape area."

**L88-97**: We hypothesized improved prediction accuracies using a combination of RS datasets that act as proxies of key drivers of soil GHG fluxes (e.g., vegetation cover and water content) and the site-measured soil parameters representing the actual field conditions.
**Comment:** improved compared with what?
**Response:** Thanks for the comment. We compared the prediction accuracies of models trained with soil parameters only and remotely-sensed data only to those with combined predictors. We have revised the statement to make it clear.
"We hypothesized that combining RS data that act as proxies of key drivers of soil GHG fluxes (e.g., vegetation cover and water content) and site-measured soil parameters representing the actual field

conditions would yield improved GHG flux prediction accuracies in our models than using either RS data or site-measured soil parameters in isolation."

**L97-101:** We also hypothesized that the high-resolution upscaled fluxes from the RF approach, which better captures hot and cold spot regions across the landscape, would avoid possible under- or overestimations of landscape fluxes derived from land use specific area-weighted averages calculated from few point chamber measurement locations.
This seems a bit out of place here. You haven't really mentioned why the RF better captures hot and cold spots on the landscape, so it seems odd to put it here in your hypotheses.
**Response:** Thank you for your observation. We have removed the RF to make the hypothesis more general.
"We also hypothesized that the high-resolution upscaled fluxes, which represent most GHG hot and cold spot regions across the landscape, would avoid possible under- or overestimations of landscape fluxes derived from land use specific area-weighted averages calculated from few point chamber measurement locations."

**L106-107:** Land uses within the landscape are mainly forests (57%) and arable lands (34%). Grasslands cover about 8% and are primarily located in riparian zones (Figure 1).
**Comment:** What type of forests? Beech? coniferous? a bit more precision here would be helpful.
**Response:** Thanks, this was expounded on the other publication. We have added the details now.
"The forest is mainly covered with mixed (44%) trees, 32% deciduous, and 23% coniferous trees (Figure 1a). The common species in the forest include European beech (*Fagus sylvatica*), spruce (*Picea abies*), European oak (*Quercus robur)*, and Scots Pine (*Pinus sylvestris*) (Wangari et al., 2022)."

**L107-109:** The dominant soil types are cambisol (69%, forest and arable), stagnosol (23%, mainly arable), and gleysol (5%) which are found along grassland riparian zones (Wangari et al., 2022).
**Comment:** please indicate the classification system. I assume that this uses the WRB classification system?
**Response:** Thank you for your comment. The system is indeed the WRB classification. We have indicated this in the manuscript.
"The dominant soil types (World Reference Base classification) are cambisol (69%, forest and arable), stagnosol (23%, mainly arable), and gleysol (5%), which are found along grassland riparian zones (Wangari et al., 2022)."

**L185-186:** The optimal trained model was automatically selected using the mean absolute error (MAE) metric with the least value.
**Response:** Thanks for the grammar correction: least to lowest. We have corrected in the text.

**L235-237:** The performance of the final models selected for the prediction of landscape fluxes varied across input datasets (RS, SP, and CD), GHG fluxes (SR/ER_$CO_2$, $CH_4$, and $N_2O$), and land use (forest, grassland, and arable land) (Table 2).
**Comment:** is it really "ER" if you aren't measuring the respiration from the above-ground biomass (at least not in the forested area).

**Response:** Thanks again for the comment. As mentioned earlier, we have added the definition of the SR/ER $CO_2$ fluxes measured in this study.
"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."

**L287-288:** The remaining landscape area (24%) had higher $N_2O$ fluxes in autumn than in summer, particularly in forested areas.
**Comment:** I am not sure what you are trying to say here. Possibly that the majority of the landscape area with higher autumn $N_2O$ fluxes were forests? Please clarify.
**Response:** We have modified the statement for clarity.
"Around 24% of the landscape, primarily on the forested areas, had higher $N_2O$ fluxes in autumn than in summer."

**L401-402:** To illustrate, parts of the landscape (24% and 37%) showed even opposite trends of higher $N_2O$ fluxes and lower $CH_4$ uptake rates in autumn, and these areas were predominantly in the forested ecosystem.
**Comment:** Were these the same types of forests as the rest? Were there different tree species?
**Response:** Thank you for your comment. These findings were predominately in the mixed forest area. We have added this information to the text.
"To illustrate, parts of the landscape (24% and 37%) showed even opposite trends of higher $N_2O$ fluxes and lower $CH_4$ uptake rates in autumn, and these areas were predominantly in the mixed forest ecosystem."

**L403-405:** For example, decaying fallen leaves during autumn can favor denitrification in forest soils but not in grassland or arable ecosystems.
**Comment:** Can you explain why? and a citation here would be very useful.
**Response:** Thank you for your comment. We have added an explanation of the mechanism based on the increase of carbon and nitrogen availability through the mineralization of the leaves when decaying.
"For example, decaying fallen leaves during autumn can favor denitrification in forest soils by increasing carbon and mineral N availability (e.g., Groffman & Tiedje, 1989), which may not be true for grassland or arable ecosystems due to harvesting and mowing."

**L405-406:** The higher $CH_4$ uptake rates in summer could be due to the increased exposure of some forest soils to the sun leading to drier and warmer soils that promote $CH_4$ oxidation (Steinkamp et al., 2000).
Wouldn't there be less sun in the summer? Weren't the forests predominantly deciduous cover?
**Response:** Thank you for your critical comment. The landscape was mainly dominated by mixed forests. We were motivated to include the sun because aspect was a key driver of $CH_4$ trends within the landscape. We have, however edited the text to link our findings to warmer temperatures rather than sun exposure.
"The higher $CH_4$ uptake rates in summer could be due to warmer summer temperatures leading to drier, more aerated forest soils that promote $CH_4$ oxidation (Steinkamp et al., 2000)."

**L418-419:** Increased soil moisture values, a key characteristic of the riparian regions, has also been reported to drive elevated soil GHG fluxes (Kaiser et al., 2018; Vainio et al., 2021).

I'm pretty sure that soil C content tends to be quite high in riparian areas as well. Which could also lead to higher SR_CO2.

**Response:** Thank you for your suggestion. We have added it in the discussion.

"Increased soil moisture values and higher soil C contents, key characteristics of the riparian regions, have also been reported to drive elevated soil GHG fluxes (Kaiser et al., 2018; Vainio et al., 2021)."

**L424-425:** This finding emphasizes the importance of capturing the $N_2O$ hot spots and improving the spatial coverage of $N_2O$ measurements, as it can introduce enormous uncertainty in landscape fluxes.

What do you mean by "capturing"? Do you mean both measuring emissions from these and determining how much of these are spread across the landscape? This may require a bit of clarification.

**Response:** Thank you for your comment and suggestion. We have rephrased the statement and made it clear.

"This finding emphasizes the importance of increasing the spatial coverage of $N_2O$ measurements to include more hot spot areas, as they can introduce enormous uncertainty in landscape fluxes if not quantified."

**L429-430:** Identifying common patches with elevated emissions of the three GHGs can inform priority areas for implementing localized mitigation measures within a landscape.

**Response:** Thanks for the grammar correction.

**L433-435:** The mitigation strategies may include adjusting the fertilizer application rates, especially in specific areas that hold more water, probably due to topographical or soil conditions (e.g., Hassan et al., 2022).

Maybe mention above that these "common" hot spots were in arable soils with high water? Otherwise this comment seems a bit out of place

**Response:** Thank you for your comment. We have rephrased the text to give context to the discussion point.

"Because most of the common GHG hotspots in the arable soils were also in areas with high water content, mitigation strategies that aim at adjusting the fertilizer application rates at specific areas that hold more water may successfully lower the emissions (e.g., Hassan et al., 2022)."

**L439-440:** The expansion of forested areas will also likely have a much higher mitigation impact via $CO_2$ sequestration.

Much higher than what? Perhaps just use "high".

**Response:** Thank you for your comment. We have rephrased the statement as advised.

"The expansion of forested areas will also likely have a high mitigation impact via $CO_2$ sequestration".

**L442-444:** We also found significant shifts in the geo-locations of hotspot regions between summer and autumn, suggesting that seasonal changes in land management and soil conditions may also lead to a temporal expansion or contraction of the hot spot regions.

Is there really a lot of "seasonal changes in land management"?

**Response:** Thank you for your questions. Yes, there is. For example, synthetic fertilizer application is only limited to periods before the growing season, i.e., early and late spring, while harvesting mainly

occurs at the end of summer. Both these land management practices can have an effect on the temporal trends of GHGs. We supplemented the sentence for clarification: "We also found significant shifts in the geo-locations of hotspot regions between summer and autumn, suggesting that seasonal effects of land management (e.g., fertilization, harvesting, and residue management) and soil conditions may also lead to a temporal expansion or contraction of the hot spot regions."

**L452-453:** In agreement with our hypotheses, the landscape fluxes were either over or under-estimated by the area-weighted average approach compared to the RF modeling approach.
According to Figure 3, your predicted fluxes were biased towards underestimation. Wouldn't that suggest that the RF is underestimating landscape fluxes rather than that the area-weighted average approach over-estimates?
**Response:** Thank you for your critical comment. While we acknowledge that some of the overestimation we found was due to the general trend where the RF models underestimated high fluxes, we were convinced that the number of averaged sampling sites, biased with either high or low values, was responsible for the differences in the two approaches. For example, we found no significant differences when we compared one-to-one means between the measured and the RF-predicted fluxes (for the sampling sites) within the same season (See Figures 1 and 2 here). However, if our model underestimation had a strong effect, one would expect the area-weighted average from the measured to be higher than the RF predicted fluxes.

When we compared the area-weighted approach to the cumulative landscape fluxes from the RF-generated maps, which in theory includes fluxes from most cold and hot spots, we found biases in the former approach due to seasonality. The area-weighted approach tended to overestimate during the summer and underestimate during autumn. These findings could mean that the simple area-weighted approach failed to represent cold spots in the summer due to biases toward measuring high-flux regions and hot spots in the autumn due to biases toward measuring low-flux regions. We had added some explanations in the discussion.

"An alternative explanation of the differences in landscape flux estimates from both approaches could be the underestimation of high fluxes by the RF models, which we also found in our study. However, the landscape means of RF predicted and measured fluxes from 30% of our sampled sites were primarily similar (Figure A1 in Appendices), suggesting that the lack of spatial representation of all hot and cold spots by the area-weighted mean approach rather than the inability of the RF models to reproduce high values accounted for the findings above."

**L453-455:** The overestimated landscape $CO_2$ and $N_2O$ fluxes by up to 50% during the peak summer season suggest an overrepresentation of the high fluxes measured at most of the sampling points, resulting in elevated mean and upscaled fluxes.
is this the overestimate by the "area-weighted average approach"? because you say that this approach both over- and under-estimated the fluxes (compared with RF).
**Response:** Thank you for your critical comment. We have rephrased the statement to make it clearer that it is an overestimate from the area-weighted approach.
"The overestimated landscape $CO_2$ and $N_2O$ fluxes by the area-weighted average approach of up to 50% during the peak summer season suggest an overrepresentation of the high fluxes measured at most of the sampling points, resulting in elevated mean and upscaled fluxes."

**L460-461:** An alternative explanation of the differences in landscape flux estimates from both approaches could be the underestimation of high fluxes by the RF models, which we also found in our study.
Wouldn't this be a bigger problem when trying to calculate annual fluxes than underestimating the low fluxes?
**Response:** Thank you for your question. We think missing out on cold or hot spots may be a bigger problem in estimating the annual fluxes of an entire landscape. This conclusion is motivated by the fact that when we compared one-to-one means between the measured and the predicted fluxes, we found no significant differences, which could mean that the error from the RF underestimation may not be that important when the fluxes are averaged. In addition, as we have shown for $N_2O$ and $CO_2$, missing out on hot spots, for example, will result in significant uncertainties in calculating the final landscape fluxes.

**L461-464:** However, the landscape means of RF predicted and measured fluxes from 30% of our sampled sites were primarily similar (Figure A1 in Appendices), suggesting that the lack of spatial representation of all hot and cold spots by the area-weighted mean approach rather than the inability of the RF models to reproduce high values accounted for the findings above.
What about the other 70%? Were they randomly distributed? or was there some bias that could be noted?
**Response:** Thank you for the question. The split was done randomly; hence, the distribution of the sites was also random. The mean comparison results for both the test and the training dataset were primarily similar (See Figure 1: test data and Figure 2: training data).

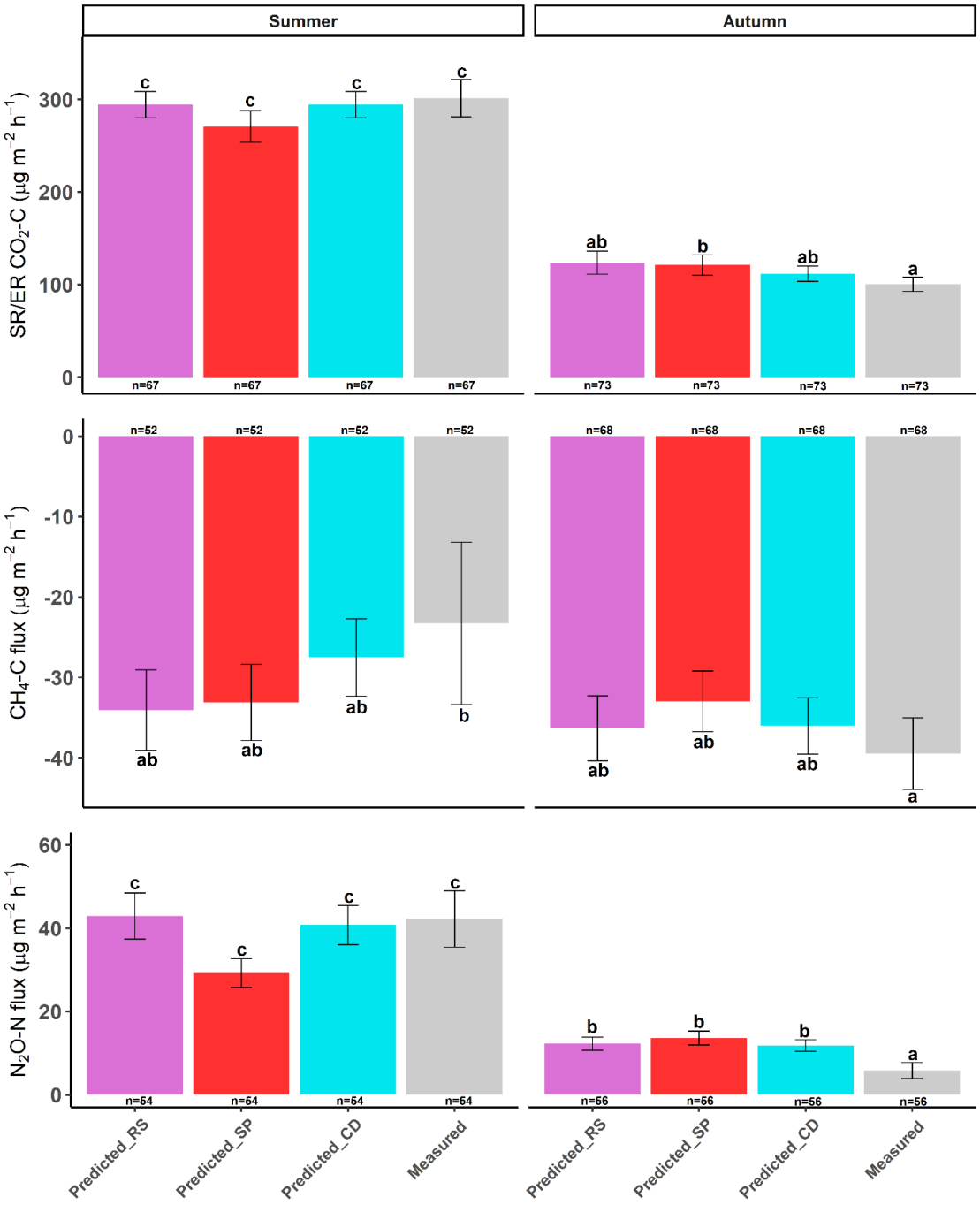


**Figure 1:** Bar graphs showing the mean fluxes (±SE) predicted using remote sensing (RS), soil properties (SP), and combined data (CD) and the measured fluxes at the sampling sites in the 30% model test dataset. The upper-case and lower-case letters indicate significant differences ($p<0.05$) in the mean fluxes in the different seasons and across the measured and predicted fluxes.

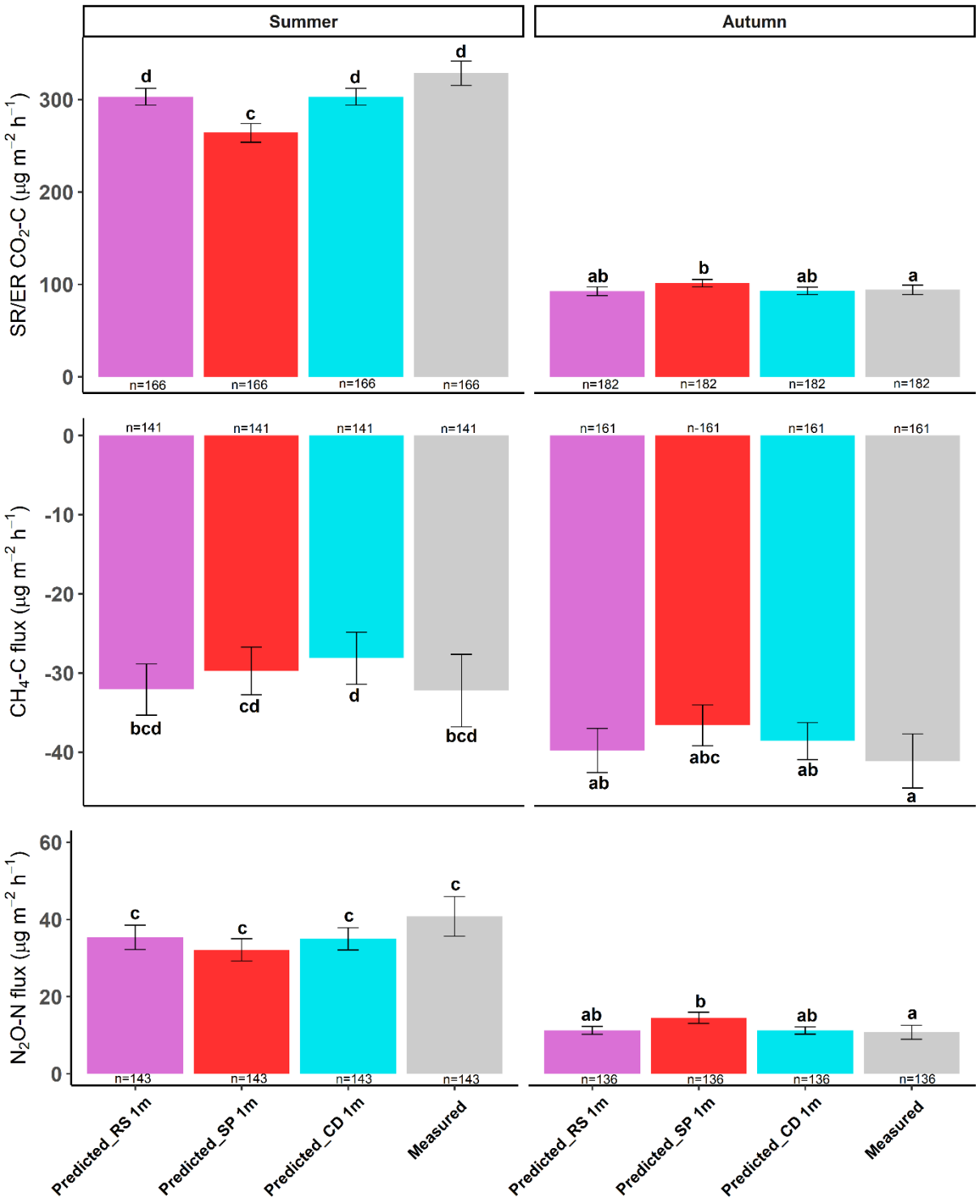


**Figure 2:** Bar graphs showing the mean fluxes (±SE) predicted using remote sensing (RS), soil properties (SP), and combined data (CD) and the measured fluxes at the sampling sites in the 70% model training dataset. The upper-case and lower-case letters indicate significant differences ($p<0.05$) in the mean fluxes in the different seasons and across the measured and predicted fluxes.

**L468-469:** The high (50%) overestimation of landscape $N_2O$ fluxes suggested the higher sensitivity of reliably estimating $N_2O$ fluxes using the (aggregated means) conventional method.

you keep mentioning that the aggregated means over (or under) estimates landscape fluxes. But that is only when compared to the RF method. Do we really know that the RF method is more accurate? For me, the only way to know for sure would be to compare with a tall flux tower that actually measures the landscape flux.

**Response:** Thank you for your critical comment. You are correct that there is no way of exactly validating the results from the RF maps in our study to determine how accurate they are in representing total landscape fluxes. However, on purely methodological grounds, an average that better represents the heterogeneity of GHGs across an entire landscape, such as that computed by the RF models, offers improved estimates than only a few measured points. We also showed this in our earlier publication, where the mean flux uncertainties decrease logarithmically with the number of measurements done (Wangari et al., 2022). The next steps are to use the results from the RF maps to guide field measurements by chambers or flux towers and check the validity of the model. We plan to work on this in a follow-up project.

**L479-481:** This study's high spatial resolution upscaling (1 m pixel) enabled capturing small-scale variabilities in GHG fluxes within short distances, which would have been missed out with coarser resolution upscaling.

**Response:** Thank you for the grammar correction. We have made the changes.

**L497: Table 3:** Comparison of other that have upscaled landscape fluxes using the random forest algorithm.

other what? Other "studies"?

**Response:** Thanks for the grammar correction. We have rephrased the statement to make it clearer. "Comparison with other studies that have upscaled landscape fluxes using the random forest algorithm."

**L510-511: Figure A4:** Maps showing the hot and cold spots of the (a) summer and (b) autumn seasons. These regions were defined using each season's specific threshold.

I'm not sure of this "hot" and "cold" spot designation. To me, hot spots are places in the landscape that have high annual emissions. And I think it may be difficult to determine a hot spot from two measurements across an entire year. Areas that had relatively high emissions during both campaigns could probably be considered hot spot, but I'm not sure if we should consider a site a "hotspot" if it had relatively high emissions during only one of the campaigns.

**Response:** Thank you for your critical comment. You are right that it will be very interesting to see if these spatial hot or cold spots in our study are persistent throughout the year, which would clearly designate them as such. However, due to our study's temporal limitation, we only designated them as summer/autumn hot or cold spots. We have added this clarity in the materials and methods section where we calculated the hot and cold spots to indicate that these are only for summer and autumn. We have also added a reflection of this in the conclusion.

Materials and methods:

"2.6 Identification of summer and autumn GHG 'hot' and 'cold' spots from predicted landscape fluxes"

Results:

"3.4 Summer and autumn hot spots and cold spots"

Conclusion:
"While we identified hot and cold spots of soil GHG flux across the Schwingbach landscape through RF modeling, the entire exercise was limited to two seasons (summer and autumn). For this reason, it is still unclear whether these hot and cold spots persist throughout the year and their overall contribution to the annual landscape GHG flux estimates. Future studies should, therefore, aim at increasing the temporal resolution of similar spatially extensive measurements to at least monthly scales, which, when combined with remotely-sensed data, may be able to create similar landscape flux maps and identify the contribution of GHG hot and cold spots to annual estimates."

**L513-514: Table B1 a, b, c:** Cross-validation results of different models developed for SR/ER-$CO_2$ fluxes in 1a) forest, 1b) grassland and 1c) arable land using different predictors in the training dataset. Stepwise elimination of the least important predictors was implemented.
This does not agree with Table 2. I think that this is for the calibration data and Table 2 is for the validation data, but that is not clear with the Table captions.
**Response:** Thank you for your critical comment. Table 2 and Tables B1-B5 show the cross-validation results of the trained models. We have seen the issue of why Table 2 is different: i.e., Tables B1, B3, and B5 have the log-transformed RMSE and MAE values for $CO_2$ and $N_2O$ fluxes. We have now adjusted Tables B1, B3, and B5 to have retransformed values of RMSE and MAE to align with Table 2.

540 Table B6: The minimum, maximum, mean, standard deviation, and standard error of the measured fluxes at all the sampling p
541 and the predicted landscape fluxes using remote sensing (RS), soil properties (SP), and combined data (CD).

| Measured fluxes at sampling points | | Summer | | | | | Autumn | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Land use | Flux type | Min | Max | Mean | STDEV | SE | Min | Max | Mean | STDEV | SE |
| Forest | | 60 | 589 | 210 | 111 | 12.0 | 10 | 446 | 74 | 53 | 5.5 |
| Grassland | SR/ER-$CO_2$-C (mg $m^{-2}$ $h^{-1}$) | 136 | 693 | 350 | 123 | 14.1 | 9 | 419 | 131 | 82 | 8.6 |
| Arable | | 78 | 877 | 431 | 192 | 23.3 | 14 | 238 | 84 | 51 | 6.1 |
| Forest | | -201 | 176 | -62 | 47 | 5.1 | -214 | 7 | -68 | 48 | 4.9 |
| Grassland | $CH_4$-C (µg $m^{-2}$ $h^{-1}$) | -84 | 221 | -9 | 43 | 5.2 | -100 | 28 | -23 | 21 | 2.4 |
| Arable | | -133 | 157 | 8 | 74 | 12.3 | -43 | 11 | -17 | 10 | 1.4 |
| Forest | | -13 | 117 | 14 | 24 | 2.9 | -17 | 78 | 5 | 11 | 1.3 |
| Grassland | $N_2O$-N (µg $m^{-2}$ $h^{-1}$) | -17 | 281 | 32 | 57 | 7.0 | -18 | 154 | 12 | 30 | 3.7 |
| Arable | | 13 | 282 | 84 | 65 | 8.4 | -15 | 54 | 12 | 12 | 1.6 |
| **Predicted landscape fluxes (RS data)** | | | | | | | | | | | |
| Forest | | 37 | 327 | 171 | 51 | 0.03 | 38 | 288 | 74 | 26 | 0.01 |
| Grassland | SR/ER-$CO_2$-C (mg $m^{-2}$ $h^{-1}$) | 59 | 484 | 294 | 70 | 0.10 | 39 | 477 | 186 | 89 | 0.13 |
| Arable | | 35 | 668 | 324 | 111 | 0.08 | 28 | 559 | 102 | 86 | 0.06 |
| Forest | | -147 | 65 | -70 | 21 | 0.01 | -148 | 65 | -72 | 25 | 0.01 |
| Grassland | $CH_4$-C (µg $m^{-2}$ $h^{-1}$) | -60 | 50 | -15 | 17 | 0.02 | -64 | 32 | -18 | 11 | 0.02 |
| Arable | | -60 | 89 | -5 | 23 | 0.02 | -60 | 75 | -16 | 11 | 0.01 |
| Forest | | -8 | 38 | 7 | 5 | 0.003 | -6 | 27 | 4 | 4 | 0.002 |
| Grassland | $N_2O$-N (µg $m^{-2}$ $h^{-1}$) | -8 | 144 | 26 | 34 | 0.05 | -9 | 69 | 12 | 8 | 0.01 |
| Arable | | 0 | 190 | 60 | 33 | 0.02 | -1 | 183 | 18 | 17 | 0.01 |
| **Predicted landscape fluxes (SP data)** | | | | | | | | | | | |
| Forest | | 55 | 343 | 194 | 34 | 0.02 | 41 | 214 | 70 | 14 | 0.01 |
| Grassland | SR/ER-$CO_2$-C (mg $m^{-2}$ $h^{-1}$) | 72 | 470 | 320 | 38 | 0.05 | 52 | 319 | 128 | 44 | 0.06 |
| Arable | | 36 | 733 | 266 | 90 | 0.06 | 28 | 733 | 124 | 60 | 0.04 |
| Forest | | -123 | 54 | -51 | 11 | 0.01 | -138 | -29 | -51 | 10 | 0.01 |
| Grassland | $CH_4$-C (µg $m^{-2}$ $h^{-1}$) | -65 | 37 | -8 | 8 | 0.01 | -65 | 13 | -10 | 6 | 0.01 |
| Arable | | -87 | 85 | -7 | 26 | 0.02 | -67 | 85 | -13 | 17 | 0.01 |
| Forest | | -9 | 49 | 9 | 7 | 0.00 | -9 | 23 | 6 | 4 | 0.00 |
| Grassland | $N_2O$-N (µg $m^{-2}$ $h^{-1}$) | -6 | 124 | 20 | 8 | 0.01 | -7 | 54 | 7 | 7 | 0.01 |
| Arable | | 12 | 157 | 45 | 10 | 0.01 | 0 | 150 | 19 | 9 | 0.01 |
| **Predicted landscape fluxes (CD data)** | | | | | | | | | | | |
| Forest | | 82 | 325 | 185 | 31 | 0.02 | 42 | 195 | 66 | 14 | 0.01 |
| Grassland | SR/ER-$CO_2$-C (mg $m^{-2}$ $h^{-1}$) | 155 | 496 | 322 | 47 | 0.07 | 52 | 349 | 145 | 61 | 0.09 |
| Arable | | 68 | 694 | 321 | 105 | 0.08 | 29 | 568 | 110 | 59 | 0.04 |
| Forest | | -125 | 55 | -57 | 18 | 0.01 | -136 | -27 | -59 | 19 | 0.01 |
| Grassland | $CH_4$-C (µg $m^{-2}$ $h^{-1}$) | -69 | 36 | -6 | 9 | 0.01 | -69 | 13 | -11 | 6 | 0.01 |
| Arable | | -72 | 78 | 0 | 24 | 0.02 | -72 | 53 | -17 | 11 | 0.01 |
| Forest | | -9 | 49 | 9 | 7 | 0.00 | -9 | 23 | 6 | 4 | 0.00 |
| Grassland | $N_2O$-N (µg $m^{-2}$ $h^{-1}$) | -9 | 152 | 25 | 31 | 0.05 | -8 | 83 | 6 | 7 | 0.01 |
| Arable | | 16 | 168 | 58 | 21 | 0.02 | 1 | 128 | 16 | 12 | 0.01 |

542

Is this really "ecosystem respiration"? I would guess that your chamber was not big enough to measure respiration from the forest above-ground biomass.
**Response:** Thanks for raising this issue. As mentioned earlier, we have added these details in the methods section.