Authors response to reviews, paper "Does increased spatial replication above heterogeneous agroforestry improve the representativeness of eddy covariance measurements?", submitted to *Biogeosciences*, 10.5194/egusphere-2025-810

José Ángel Callejas-Rodelas[1], Alexander Knohl[1,2], Ivan Mammarella[3], Timo Vesala[3,4], Olli Peltola[5], and Christian Markwitz[1]

[1]University of Göttingen, Bioclimatology, Göttingen, Germany
[2]Centre for Biodiversity and Land Use, University of Göttingen, Göttingen, Germany
[3]Institute for Atmosphere and Earth System Research (INAR)/Physics, Faculty of Science, University of Helsinki
[4]Institute for Atmosphere and Earth System Research (INAR)/Forest Science, Faculty of Agriculture and Forestry, University of Helsinki
[5]Natural Resources Institute Finland (LUKE), Latokartanonkaari 9, Helsinki, 00790, Finland

The reviewers' comments are named as R1 (reviewer 1) and R2 (reviewer 2) followed by _C1, _C2, _C3, etc., numbering in order the comments. The authors' response is numbered in a similar way, using AR_C1, AR_C2, etc. The new figures crafted for this author's response are numbered AR1, AR2, etc., to distinguish them from the figures in the submitted manuscript.

1 Reviewer 1

1.1 General comments

R1_General comment. This manuscript describes the results of a field experiment with three low-cost eddy-covariance systems over a patchy agroforestry system and a patchy monocropping system. By analyzing these data from two growing seasons, the authors attempt to answer the question that is raised in the title? The topic of agroforestry is also highly relevant. Overall, the manuscript is well written and clearly structured. The data processing is described in detail. The figures are also clear and easy to read. However, I see major deficits in the experiment design which is not really suited to address the title question, at least not in a general sense as it is

formulated. Moreover, I cannot agree with some of the data-processing choices that were made and transparently communicated in the manuscript. As a consequence, data of poor quality and consequently large uncertainty are included in the analysis as the underlying assumptions of the EC-method are compromised. Moreover, I find that gap-filled fluxes should not be included in such an analysis as these modelled data are inherently much smoother than actual measurements. These choices in the data processing limit the ability to draw valid conclusions regarding the hypothesis that is posed by the authors in the introduction section. However, I believe this can still be corrected and the formulation of the objectives can be adjusted. Hence, I recommend major revisions before this manuscript can be accepted.

AR_General comment. We appreciate the reviewer's comment about our manuscript. We are thankful for bringing out the main novelty of the study and key points, and also for the recommendations regarding changes that the manuscript should undergo. These major points are addressed throughout the comments in the following section.

1.2 Specific comments

R1_C1. L37: Since the topic is surface heterogeneity, it would make sense to put this specific type of heterogeneity of and agroforestry system in a more general context of heterogeneity, also stressing that the effects depend on the type of heterogeneity and the scale of heterogeneity (Bou-Zeid et al. 2020)

AR_C1. The text will be changed accordingly, not only focusing on the nature of sources and sinks of CO_2 and H_2O, but also on how ecosystem heterogeneity affects eddy covariance measurements in general. We thank the reviewer for this comment and for the literature recommendation. A paragraph on the topic will be added to the introduction and the discussion sections.

R1_C2. L95: The random uncertainty if low cost sensors is not necessarily larger than for conventional EC. This is certainly the case for a systematic error.

AR_C2. Indeed this is something shown for example in Markwitz and Siebicke (2019). We also found this during the intercomparison campaign, where the random error of the LC-EC setups was similar to the conventional eddy covariance setup (Callejas-Rodelas et al., 2024). However, in the current study the random error at the 30-min time scale was similar to the spatial standard deviation across AF1, AF2 and AF3, also at 30-min time scale. Please find attached a plot related to this topic (Fig. AR7), with its explanation in the response to comment R1_C14.

R1_C3. L96: In my mind, the statistical robustness could only be improved through more sampling points (i.e. EC towers) if the surface can be considered homogeneous and footprints are comparable in nature. Otherwise you measure the spatial variability over a heterogeneous surface but you cannot really average those into an overall estimate that would then possibly have a lower uncertainty.

AR_C3. The aim of this study was to investigate whether the spatial variability across a heterogeneous agroforestry site was larger than the variability between two distinct ecosystems, e.g. AF and open cropland, using a distributed network of three stations equipped with LC-EC setups. The flux and meteorological data gathered from the three stations gives a more complete picture of the exchange processes at the ecosystem, compared to the typical situation in which only one station would be installed at the AF. The ecosystem heterogeneity affects the reliability of fluxes if measured with only one station, however with three systems, different patches of the ecosystem can be attributed to different fluxes. Nonetheless, we will change that sentence accordingly in the text that sentence and the related information, to make the statement more clear. It would be risky to average the measurements from the three stations, so instead of saying that the spatial replication is improved over heterogeneous sites with the distributed network, we will make clear that the spatial variability can be addressed and flux differences across a heterogeneous site can be understood from the different footprints. Moreover, the spatial replication can be better achieved with lower-cost setups installed at the stations, due to the reduced cost in instrument acquisition and its comparable performance to standard EC (Callejas-Rodelas et al., 2024; van Ramshorst et al., 2024).

R1_C4. L98: The third objective is not really related to the overarching hypothesis and the title.

AR_C4. We think it is important to keep it, since the idea of the paper is not only to study spatial variability of fluxes within the agroforestry, but also to compare whether the spatial variability within AF is larger or smaller than between AF and MC. The intercomparison paper of Callejas-Rodelas et al. (2024) demonstrated that the differences between AF and MC were larger than differences between lower-cost and conventional EC setups, which is a premise to trust the lower-cost measurements. However, given the heterogeneity of the AF system, from the intercomparison campaign we cannot know if a single EC station at the AF is sufficiently representative of the ecosystem. That is why we installed a network of three stations. Comparing them to the MC again is related to the same concept of testing whether the spatial variability within the AF is larger than the ecosystem difference. However, we appreciate the comment and will revise the corresponding text to keep the storyline across the manuscript, so then it is consistent with the findings of the first intercomparison campaign and with the objectives stated in the introduction.

R1_C5. Figure 1: I would not call it a monocropping system if the EC tower is located at the edge of a field between two different crops, and, hence, is measuring fluxes from both crops to a certain extent (or even another crop) depending on the specific footprint.

AR_C5. The term monocropping is changed to open cropland (OC) across the whole text and in the figures.

R1_C6. L144: Was the flow turbulent inside the tubes for this flow rate, which depends on the Reynolds number and hence the diameter? This would be important to minimize diffusion along the tube. The given flow rates seem to be rather low. What are reasons for this choice and what are the consequences for the frequency response characteristics of these measurements?

AR_C6. The flow inside the tube of the LC-EC was not turbulent as we used a low-energy consumption pump. This was a trade-off we had to make given that the stations were run on solar energy only. In Callejas-Rodelas et al. (2024) we tested the system with this setup and found good agreement with a LI-7200 (Licor Biosciences, USA). Nevertheless, we see here an opportunity to further improve the system.

R1_C7. L173: How does the RH-dependent fit look like? Could you please also give some indicators on the quality of the fit?

AR_C7. In Figure AR1 you can find an example fit of the time response vs. RH, as a direct output from the processing software EddyUH. Figure AR2 shows the fits for all the stations in Wendhausen, for the three years 2022, 2023 and 2024. The year 2022 was not used for the data analysis, but in some cases part of the data from 2023 were processed together with 2022, to have a more complete dataset. The fit corresponds to the equation: $y=a+b\cdot(\frac{RH}{100}^c)$. In general, the fits are good, with large r^2 coefficients (above 0.9), showing a exponential dependency of the time response with RH. In some situations, however, and especially in 2024, the fits are not good and the time responses estimated for high RH were too large to be realistic. In those situations, the coefficients corresponding to the previous good fit were used. For AF1 in 2023 and 2024, the fit of 2022 was used. For AF2 in 2024, the fit of 2023 were used. In the original preparation of the manuscript, AF1 was already processed this way, but AF2 was not, therefore the fluxes from 2024 were re-calculated for this station and the whole processing and gap filling was done again. The figures and text will be updated in the manuscript.

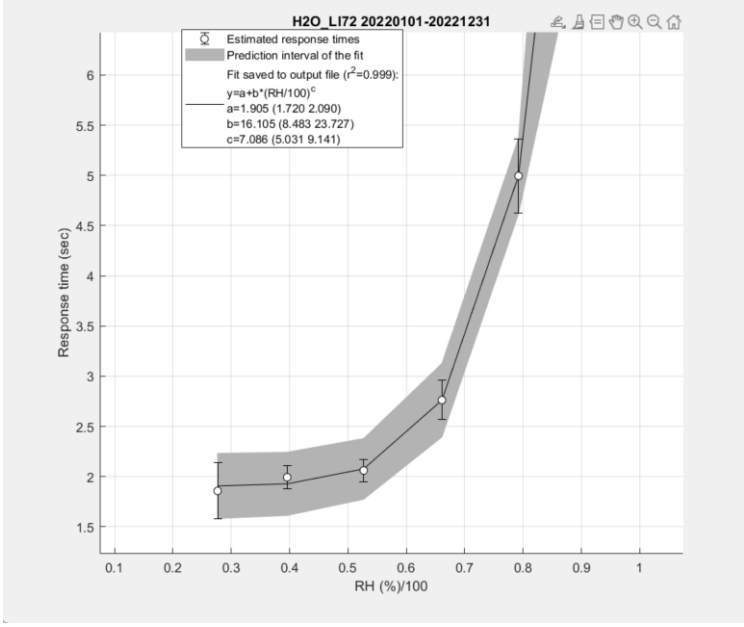

Figure AR1: Example of fit of the time response with the equation $y=a+b\cdot(\frac{RH}{100}^c)$. The example corresponds to Wendhausen MC station in 2022.

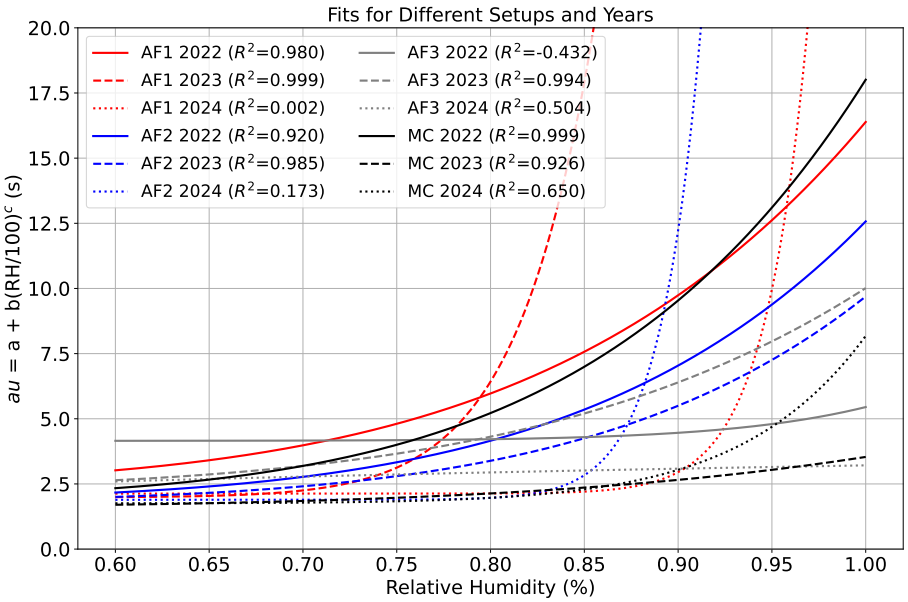

Figure AR2: Fit of the time response dependent on RH for all the stations (AF1 in red, AF2 in blue, AF3 in grey and MC in black), for the years 2022 (solid line), 2023 (dashed line) and 2024 (dotted line). Next to the legend labels the coefficients of determination (r^2) for the fit are displayed.

R1_C8. L175: What is the reasoning behind this threshold of quality flags <7? Normally, only data with flags ≤3 are considered high quality and flags 4-6 are only suitable for calculating annual or monthly sums as they are at least better than gap filling as they have deviations of up to 100%. If data are restricted to flags 1-3, the test on well-developed turbulence can for example ensure that measurements are conducted above the RSL, and hence are not influenced by single roughness elements, i.e. single trees, and the steady state test can ensure that the footprint does not vary too much within a 30-min averaging interval due to variable wind conditions, so that the time series becomes non-stationary and a covariance calculation or any other calculation of Gaussian statistics are not meaningful anymore. Hence, I highly recommend to use only data with flags 1-3 for this study.

AR_C8. Thanks for the comment and the suggestion. We tested the variability in fluxes, friction velocity ($USTAR$) and other variables, depending on wind direction and stability, and it seems that the results are quite similar independently of the quality flag level that is selected. Attached there are some example plots, not included in the paper, to demonstrate this. The first plot (Fig. AR3), shows the standard deviation of FC and LE with respect to $USTAR$. The standard deviation was calculated across the three stations at the AF for a given 30-min period. It illustrates how similar different levels of quality flags with respect to $USTAR$ are. The behavior of the data is similar under different levels of filtering, just the magnitude of the standard deviation is increasing slightly when using data with quality flags from 1 to 6. The second plot (Fig. AR4) shows the same but depending on wind direction (WD). We observe a similar variability across stations for all quality check levels, as the standard deviation of FC and LE does not change for different $USTAR$ and WD. Because the standard deviation across the three stations at the AF does not change for different $USTAR$ or WD, this explains a similar variability across stations for all these levels of quality checks. Therefore, we would keep the data filtering and gap-filling as in the

original manuscript for Figures 4 and 7, which need weekly and daily sums, respectively. Figures 5 and 6 will only include filtered data but not gap-filled data. This distinction will be clarified in the Methods section.

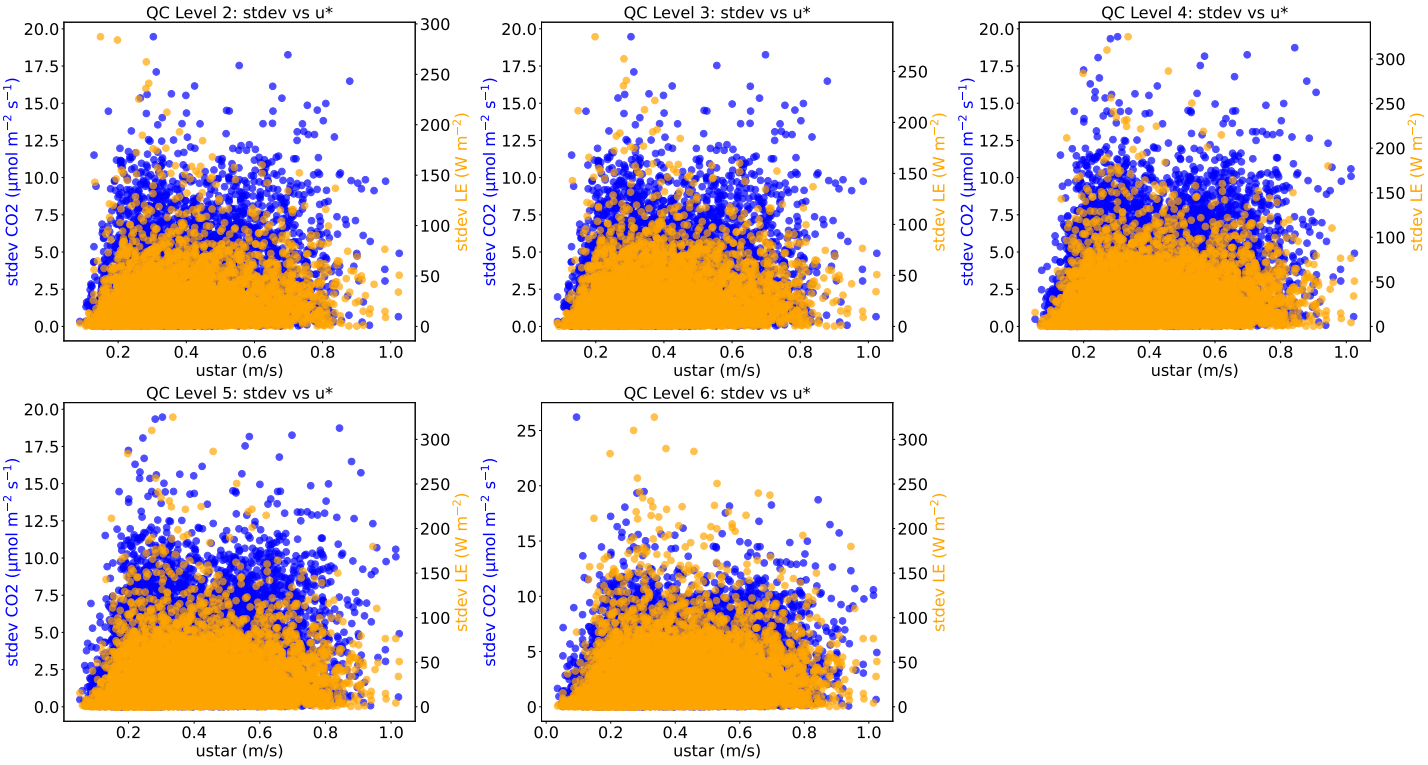

Figure AR3: Spatial standard deviation of FC (blue, left y-axis) and LE (orange, right y-axis) across the three AF stations for different levels of filtered data (quality flags ranging from 1 to 6, 1 to 5, 1 to 4, 1 to 3 and 1 to 2), depending on friction velocity.

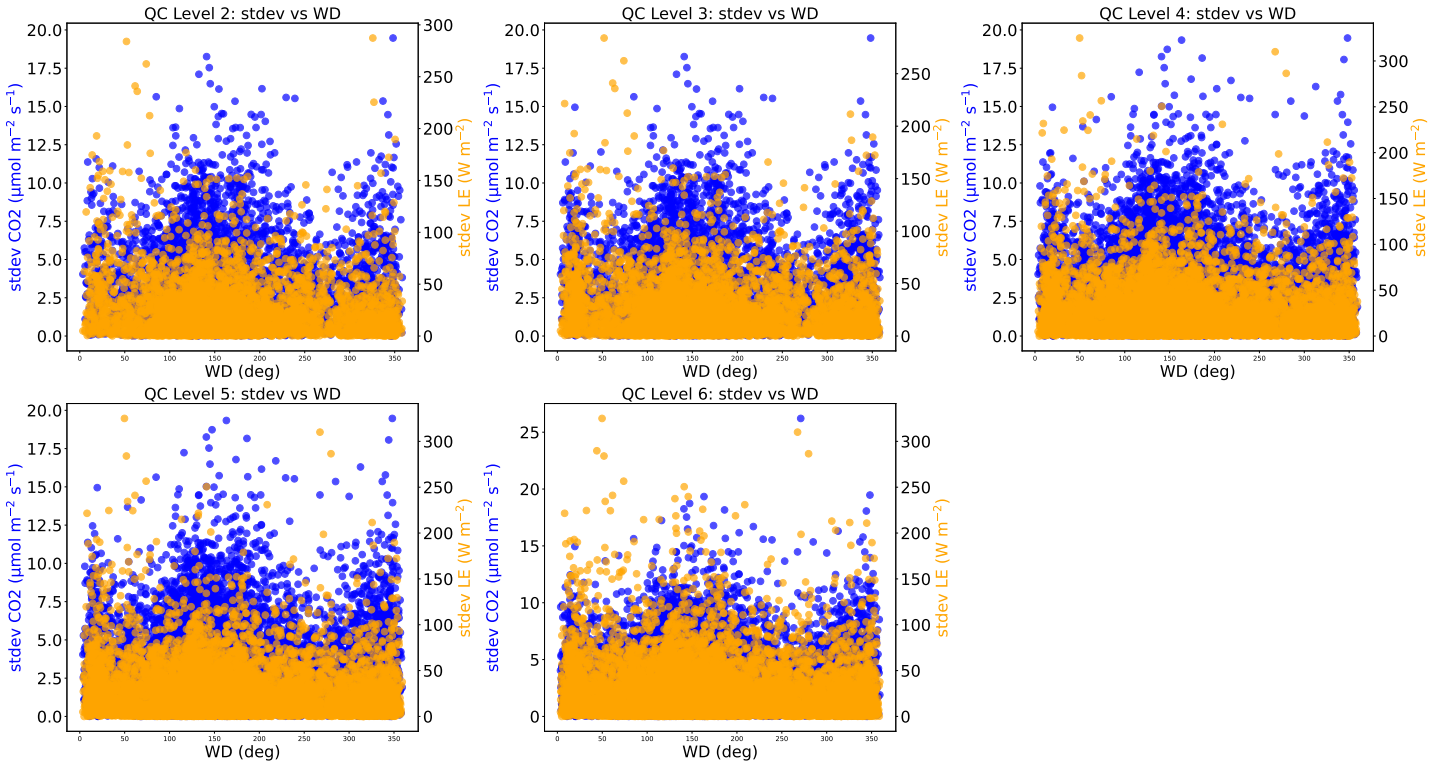

Figure AR4: Spatial standard deviation of FC (blue, left y-axis) and LE (orange, right y-axis) across the three AF stations for different levels of filtered data (quality flags ranging from 1 to 6, 1 to 5, 1 to 4, 1 to 3 and 1 to 2), depending on wind direction.

R1_C9. L210: In my mind, it does not make sense to apply gap filling for the objectives of this study. Only actual measurements should be used to analyse the spatiotemporal variability and heterogeneity effects, no modelled data, which are inherently much smoother that actual flux measurements.

AR_C9. Thanks for the comment and the suggestion. We will make some changes in the manuscript. Figures 4 (weekly sums of carbon and ET) and 7 (effect size comparing daily sums) will keep gap-filled data, otherwise sums cannot be calculated. On the other hand, Figures 5 and 6 will be re-plotted using only measured and filtered data, to address the spatial variability in wind sectors and weeks without the bias of the gap filling process. This comment also relates to the reply on Figure 4, R1_C13. Therein the attached figure (Fig. AR5) related to this topic is explained.

R1_C10. Table 1: Be aware that these number represent just the error of the gap-filling and not the error of the EC measurements. These can be estimated based on other methods (e.g. Lenschow et al. 1994, Finkelstein and Sims 2001, Billesbach 2011, Richardson et al. 2012).

AR_C10. Thank you for your comment. The data from Table 1 were only used to assign an error to the modeled data using XGBoost. In section 2.5 we explain how the error is attributed to individual 30-min fluxes. The caption in Table 1 was modified accordingly in the text to make this more clear.

R1_C11. L241: How was the zero-plane displacement height calculated for the towers between two adjacent fields with different canopy height?

AR_C11. Displacement height was calculated as 0.6 times the aerodynamic canopy height. The aerodynamic canopy height was calculated, following the explained procedure in lines 233 to 243 of the manuscript, according to Chu et al. (2018), at the 30-min time scale. The aerodynamic canopy height accounts for the effect of different surfaces and canopy heights on the wind profile under neutral conditions. In order to provide full time series for the footprint modeling, a running mean of the aerodynamic canopy height was calculated with a hundred 30-min intervals, for 8 different wind sectors, to fill all gaps which include non-neutral conditions. This was also clarified in the manuscript.

R1_C12. L289: In principle, it would be fine to determine the uncertainty from an intercomparison experiment. But then, it should be guaranteed that the underlying surface is homogeneous and the footprints are overlapping. This was clearly not the case in the study of Callejas-Rodelas et al. (2024) and hence this study cannot be used for this purpose. Moreover, other measures than the slope of a regression are better suited to describe the uncertainty based on an intercomparison experiment, for example comparability (RMSD) and bias.

AR_C12. In relation to comments R1_C2 and R1_C14, the random error calculated according to Finkelstein and Sims (2001) was included in the calculations of the uncertainty of the daily sums. Additionally, we re-run the figure another time using as the individual error in measured data the sum of the random error and the uncertainty from the intercomparison experiment, but taking this time the RMSE instead of the slope. The largest RMSE was used, with values of 3.1 μmol m^{-2} s^{-1} for FC and 44.1 W m^{-2} for LE.

Nonetheless, the effect size was calculated in a wrong way, due to a mistake in the units and magnitudes. Figure AR9 displays the corresponding new plot, in the left the error for the measurements was considered as the random error, while in the right, it was considered as the sum of the random error plus the RMSE mentioned in the previous paragraph. The new values are more reasonable, considering the Fig. 3 in the paper by Hill et al. (2017) where they did a simulation of the effect size, the statistical power and the number of stations necessary to detect ecosystem differences. Please see more information on the reply to comment R2_C8.

R1_C13. Figure 4: Which of these data are actually measured and which are gap-filled? How do the measurements compare for 30 min flux estimates?

AR_C13. Please find below two figures related to this. The first figure, Fig. AR5, shows the time series of FC, LE and H, with all the data that were used in the paper, that is, measured data plus filled data for a maximum gap duration of 2 weeks. These two figures correspond to station AF1, just to illustrate the time series from one of the stations. The third figure, Fig. AR6, shows the 1-1 plots with linear fits of the modelled vs. measured data, for the XGBoost gap-filling and all the stations. The RMSE values for the test datasets from the XGBoost gap-filling are displayed in Table 1. Looking at the figures, the filled data reproduce quite well the dynamics of the measured data. A visible effect of the gap-filling, indeed, is that it smooths down the time series. Therefore, as written above

in the reply to comment R1_C9, we will change Figures 5 and 6 to use only measured data, while Figures 4 and 7, that need weekly and daily sums respectively, will be kept as they are with gap filled data.

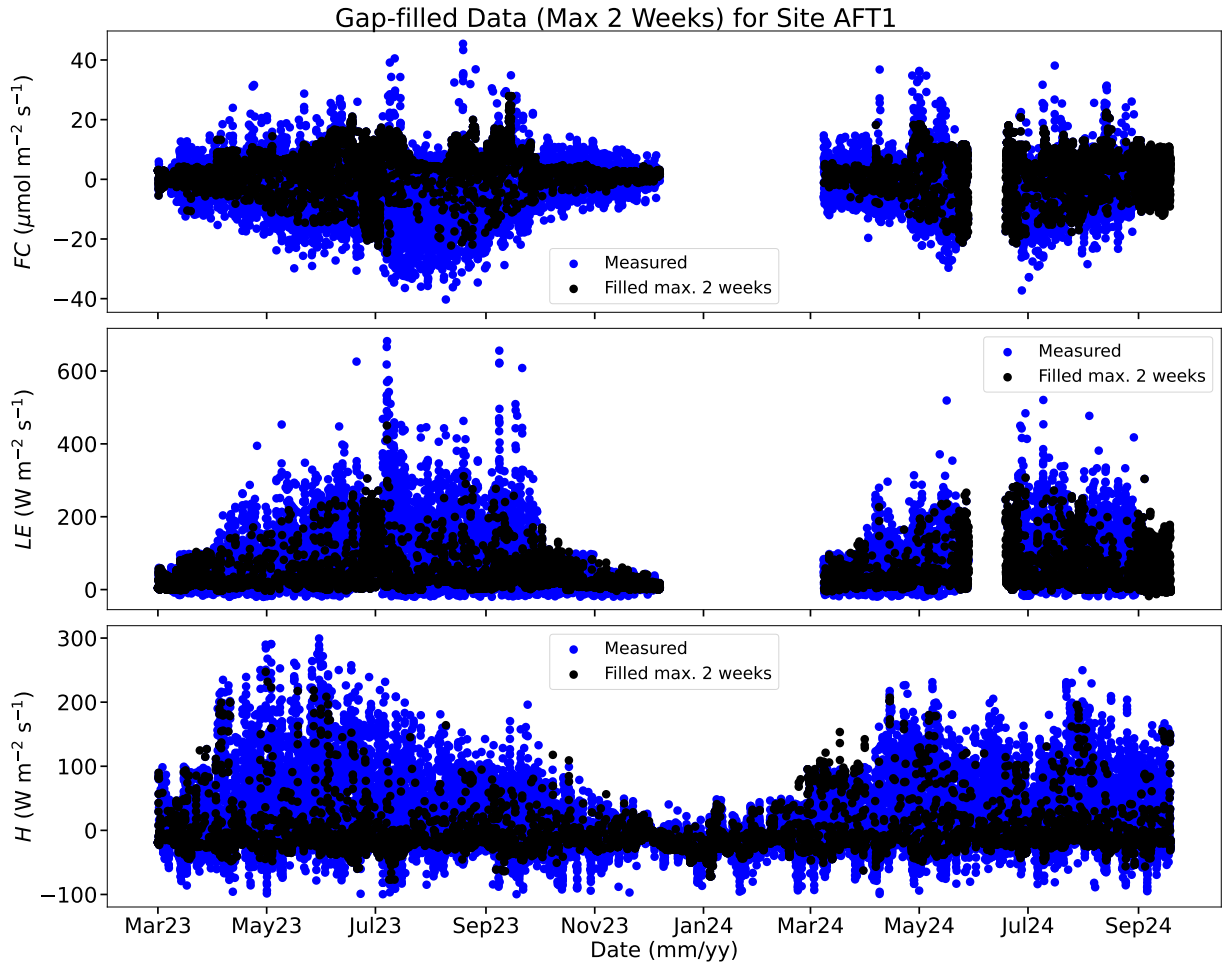

Figure AR5: Time series of measured (blue) and gap-filled (black) data considering gaps with a maximum duration of 2 weeks, for one example station (AF1). Those are the actual data used in the study.

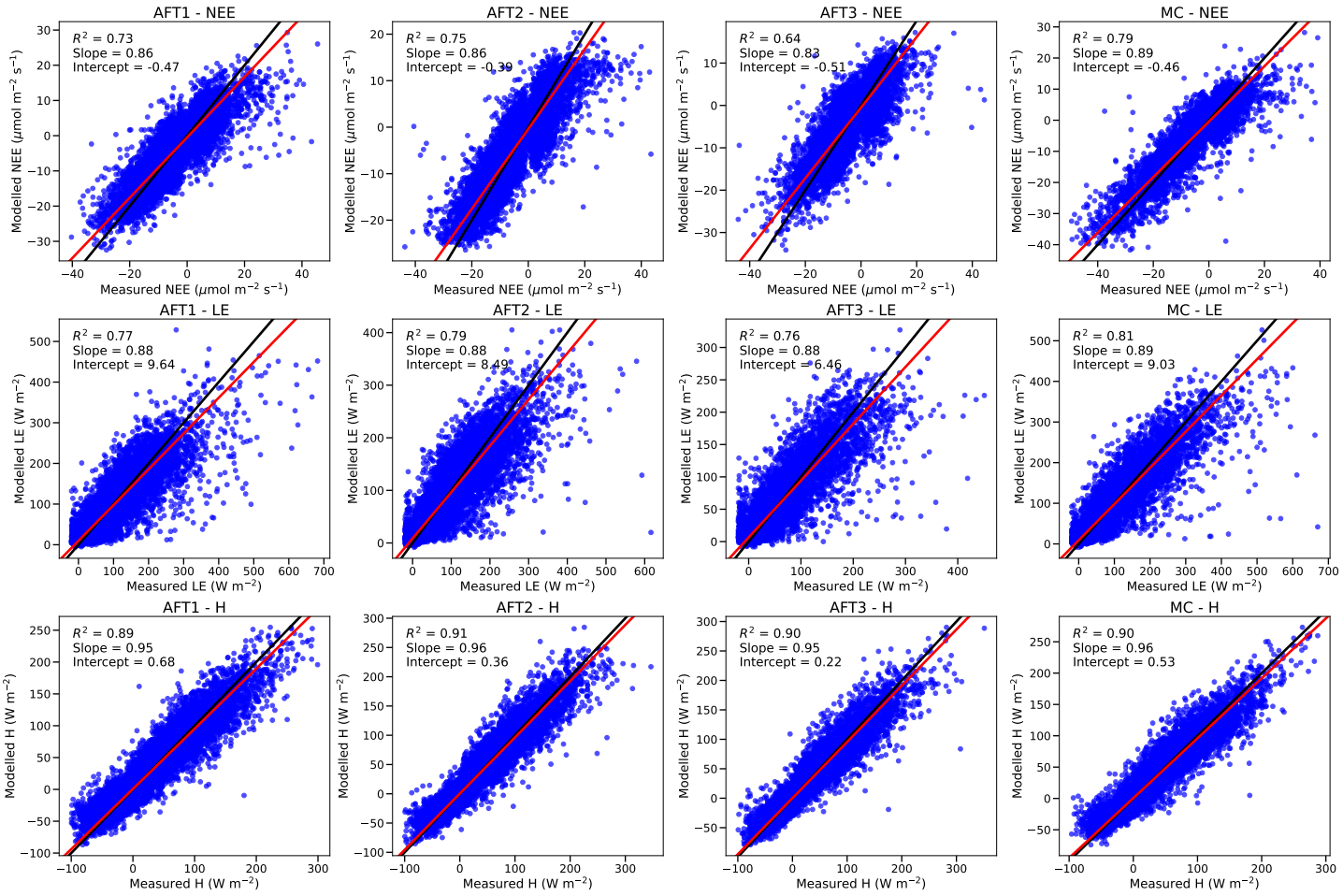

Figure AR6: Scatter plots of modeled vs. measured data, for FC, LE and H for the four stations used in the study (AF1, left column, AF2, second column, AF3, third column, MC, fourth column). The linear models were fitted on all the measured and corresponding modeled data.

R1_C14. L738: The random error should be considered for this study as it is necessary to assess whether the spatio-temporal variability is actually larger than the measurement error.

AR_C14. We included the random error in the calculations of the effect size (Fig. 7 in the original submitted manuscript) with the error propagated for the daily sums of CO2 and LE. The new plots can be seen in Figure AR9. Please also see below a plot of the different errors considered in this study (Fig. AR7). The first column shows the double exponential fit of the random error distributions for FC and LE, for all the stations (AF1, AF2, AF3 and MC). Random error was calculated according to Finkelstein and Sims (2001). The second column shows the histogram of the spatial standard deviation across the three stations at the AF with the exponential fit, at the 30-min time scale. The third column shows the spatial and temporal standard deviations, calculated according to equation 2 in the submitted manuscript, but at the daily time scale instead of weekly as shown in Figure 6 in the original manuscript. Finally, the fourth column shows the histogram and exponential fit of the distribution of the difference between random error and spatial standard deviation for FC and LE at the 30-min time scale. The

main outcome is that the random error, at the 30-min time scale, is of similar magnitude as the spatial standard deviation. However, when data are aggregated, the random error reduces as shown by some authors (e.g. Moncrieff et al., 1996 or Rannik et al., 2016), while the spatial standard deviation can be important in such a heterogeneous site, as displayed in Fig. 4 (weekly sums) and Fig. 5 (coefficients of spatial variation) in the original manuscript.

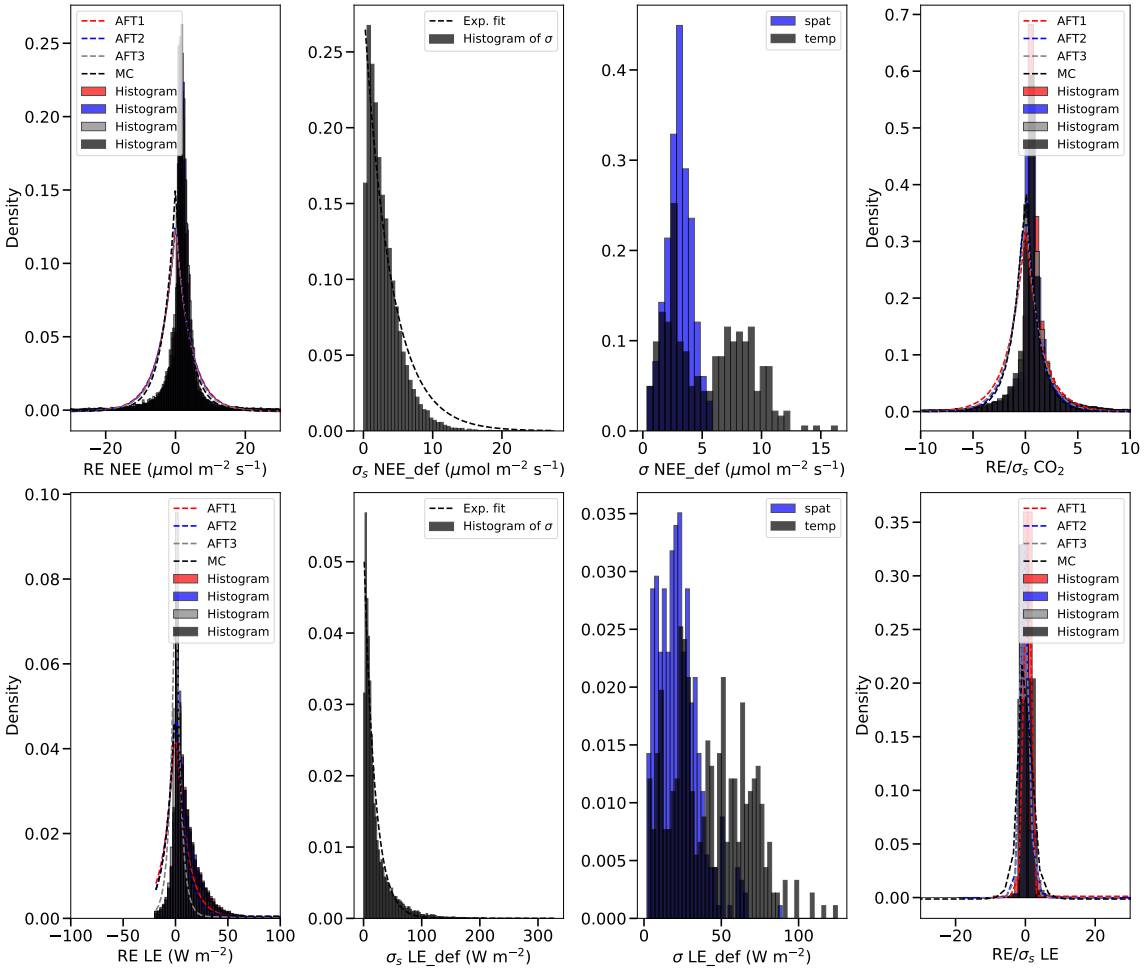

Figure AR7: (Left column) Random error of FC (top) and LE (bottom) for all the stations, with the double exponential fit to the histograms. (Second column) Histogram with exponential fit of the spatial standard deviation calculated at the 30-min time scale. (Third column) Histograms of the spatial and temporal standard deviations calculated at the daily time scale. (Fourth column) Histograms and double exponential fits of the fraction of random error and spatial standard deviation, at the 30-min time scale, for all the stations.

R1_C15. L763ff: This statement is too simplistic and does consider the enormous complexity of this question. Homogeneous conditions within the footprint are still main prerequisite for eddy-covariance measurements. Otherwise, additional transport terms become relevant which are usually neglected and almost impossible to measure. Please also consider that this kind of thermal surface heterogeneity induces secondary circulations and local advection. As a consequence, dispersive fluxes can develop, so that the eddy-covariance system measuring only the temporal covariance with the w-component severely underestimates the actual surface flux.

AR_C15. Thanks for pointing this out. We will address the corresponding changes in the text, to put more in context this topic, and stating that the lower-cost eddy covariance setups might help to reduce uncertainty induced by heterogeneity in ecosystems, but that fluxes could not be averaged and the impact of advection and non-diffusive fluxes cannot be neglected and should somehow be accounted for.

References

Bou-Zeid E, Anderson W, Katul GG, Mahrt L (2020) The Persistent Challenge of Surface Heterogeneity in Boundary-Layer Meteorology: A Review. Boundary-Layer Meteorol. https://doi.org/10.1007/s10546-020-00551-8

Billesbach DP (2011) Estimating uncertainties in individual eddy covariance flux measurements: A comparison of methods and a proposed new method. Agric For Meteorol 151:394–405

Finkelstein PL, Sims PF (2001) Sampling error in eddy correlation flux measurements. J Geophys Res 106:3503–3509. https://doi.org/10.1029/2000JD900731

Lenschow DH, Mann J, Kristensen L (1994) How Long Is Long Enough When Measuring Fluxes and Other Turbulence Statistics? J Atmos Ocean Technol 11:661–673.
https://doi.org/10.1175/1520-0426(1994)011¡0661:HLILEW¿2.0.CO;2

Richardson AD, Aubinet M, Barr AG, et al (2012) Uncertainty quantification. In: Aubinet M, Vesala T, Papale D (eds) Eddy Covariance: A Practical Guide to Measurement and Data Analysis. Springer, Dordrecht, pp 173–210

2 Reviewer 2

2.1 General comments

R2_General comment. The manuscript reports the results from the monitoring of CO2, H2O and sensible heat fluxes applying the eddy covariance method over a heterogeneous agroforestry field and a conventional cropping field. The authors deployed three low-cost eddy covariance tower in the agroforestry field to assess if the representativeness of fluxes due to the heterogeneity of the surface can be improved by increasing the number of measurement points, as stated in the title.

The application of the eddy covariance method over heterogenous surfaces, especially in terms of canopy structure (height, density, etc.) is challenging because the basic requirements for the application of the method are not fulfilled and other terms, besides the measured turbulent fluxes, should be taken into account. In my opinion, the authors do not give the right importance to this issue and only focus on the spatial representativeness, in terms of footprint area, of the vertical turbulent fluxes. I am well aware that accounting also for advection fluxes would have required a completely different and more demanding instrument setup, so I kindly ask the authors to at least acknowledge in more details the challenging aspects of making eddy covariance measurements over heterogeneous surface, as mentioned at lines 49-51, not just in terms of heterogeneity of scalar sources and sinks.

The manuscript is generally well written, but some sections are very dense and difficult to read. In particular, I think that the results section reports in too much details the patterns of the different variables observed. I suggest

to include only the main and significant results so that reading might be easier.

Overall, the manuscript try to characterize carbon and water fluxes over agroforestry systems that are not yet well studied and the analysis approach and findings might be important also for studies on other heterogeneous ecosystems, so I consider that the manuscript should be considered for publication but before that some minor revisions are necessary.

AR_General comment. We appreciate the comments from reviewer 2 and are thankful for the general recommendations regarding the manuscript. We will address these changes in the text, expanding the discussion on the heterogeneity and representativeness of measurements further beyond the heterogeneous distribution of sources and sinks of carbon and water vapour. The results section will be revised and re-written to make it shorter and more reader friendly, removing details that are not completely necessary for the story of the paper.

2.2 SPECIFIC COMMENTS

R2_C1. L104: is "monocropping" the right term for this site? 3 different crops were grew in the same field, not just one. Maybe "conventional cropping system" might be more appropriate for this specific site. Please consider this comment and change the term accordingly throughout the manuscript.

AR_C1. The term "monocropping" will be changed by "open cropland (OC)" throughout the text and in the figures and tables.

R2_C2. L165: is there a particular reason why you decided to perform the sectorial planar fit with 8 different sectors? The sector of the planar fit should be determined based on the topography or characteristics of the surface. Why did you opt for this rotation method instead of "normal" planar fit or double rotation? Please add a sentence in the text explaining the reasons for your choice.

AR_C2. We chose the planar fit dividing in wind direction sectors because of the surface heterogeneity. The 8-sectors division is based on the default recommendation by ICOS, as noted in Sabbatini et al. (2018). The information was added to the text.

R2_C3. L194: based on which criteria did you reject the data?

AR_C3. The sentence was intended to explain that gaps introduced in the original dataset were due to quality check filters. We will remove this sentence.

R2_C4. L236: it is not clear to me if you used only daytime data to assess the aerodynamic parameters or if you calculated footprints only for daytime periods. If this is the case, I think that you should consider also nighttime periods because they contribute to an important part of the C flux.

AR_C4. In the original data analysis, we only used daytime data for the footprint calculation, but for the aerodynamic canopy height we used all data. However, the footprint model was re-run for all the stations to consider nighttime periods as well, in agreement to this comment. The text will be checked to clarify this. Figure 3 in the original manuscript was changed to include the new footprints and the 50 % lines as well, in relation to comment R2_C7 below (see Fig. AR8 below). The 80 % contour lines did not change much by considering nighttime data. The 50 % line shows a small footprint around the stations, confirming what is discussed in the corresponding section of the manuscript: the major contributions to the footprint come from an area surrounding the station, hence, explaining differences in fluxes.

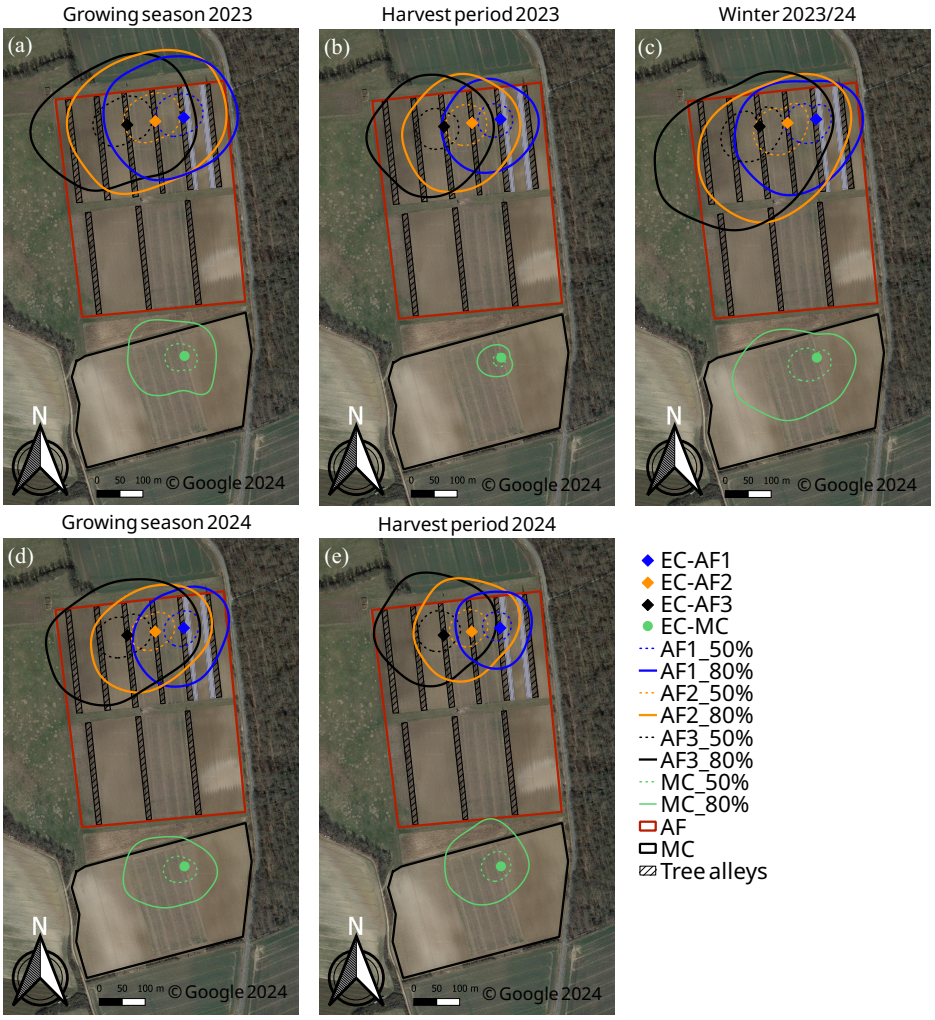

Figure AR8: 80 % (solid lines) and 50 % (dashed lines) contour of the footprint climatology for all the stations (blue, AF1, orange, AF2, black, AF3, green, MC) and periods considered in the study, (a) (growing season 2023), (b) harvest period 2023, (c) winter 2023/2024, (d) growing season 2024, (e) harvest period 2024.

R2_C5. L246: why did you aggregate the data in wind sectors of 30°? This is not consistent with the 45° sector of the planar fit.

AR_C5. We followed the reviewer recommendation and binned data in sectors of 45° to be consistent with the planar fit. The reason to use 30° was to achieve a better resolution in the spatial division of data. The results are almost the same, just slightly different values of the coefficients of variation due to a higher number of data points per wind sector. The figure (Fig. 5 in the manuscript) style and presentation was kept similarly, with the exception of the logarithmic y-axis in the coefficients of variation plots, for a better visualization.

R2_C6. L322-323: I do not think you can define "large" a value of 0.5 kPa, I would delete this sentence.

AR_C6. Done.

R2_C7. Figure3: I think it would be interesting to show also line of 50 or 60 % contribution to fluxes so one can have an idea of the location of the area contributing more to fluxes.

AR_C7. Figure 3 now shows also the 50 % line. In the corresponding explanation of the results this will also be changed. Please check the comments in R2_C4 for more details on this.

R2_C8. L460: could you please explain better the meaning of "effect size" in terms of flux spatial variability here or in the discussion session?

AR_C8. The effect size, as calculated in the paper, relates the difference between two ecosystems, or two stations, to the ensemble standard deviation. A large effect size implies that the differences between ecosystems being compared are large, compared to the pooled standard deviation. According to the paper by Hill et al. (2017), where this definition was introduced, a very large effect size, of around 10, implies that only 1 or 2 EC systems are enough to have a confidence of more than 95 % that the reported fluxes are certain. This can be visualized in the figure 3 from the paper of Hill et al. (2017).

Now we realized that there was a mistake in the effect size calculation. The error values were reported in units of μmol m^{-2} s^{-1} and W m^{-2} for FC and LE, respectively, but the daily sums were reported in g C m^{-2} and in mm (when converting LE into evapotranspiration). Therefore, the obtained values were wrong. In the left Figure AR9, the error for the measurements was considered as the random error, while in the right Figure AR9, as the sum of the random error plus the RMSE mentioned in the previous paragraph. The new figure will be added to the manuscript instead of the previous Figure 7.

The new values are much larger, reaching magnitudes of 40. This, put in the context of Fig. 3 in the paper by Hill et al. (2017), as mentioned before, means that we can be confident to resolve ecosystem differences at the daily time scale, between AF and MC, but also within the AF. However, our values are calculated using daily sums, which we assume are more noisy than yearly values as used in Hill et al. (2017). We will also modify the text so then it is more clear and the statistical power according to the paper of Hill et al. (2017) will be reported to put our effect size values in context.

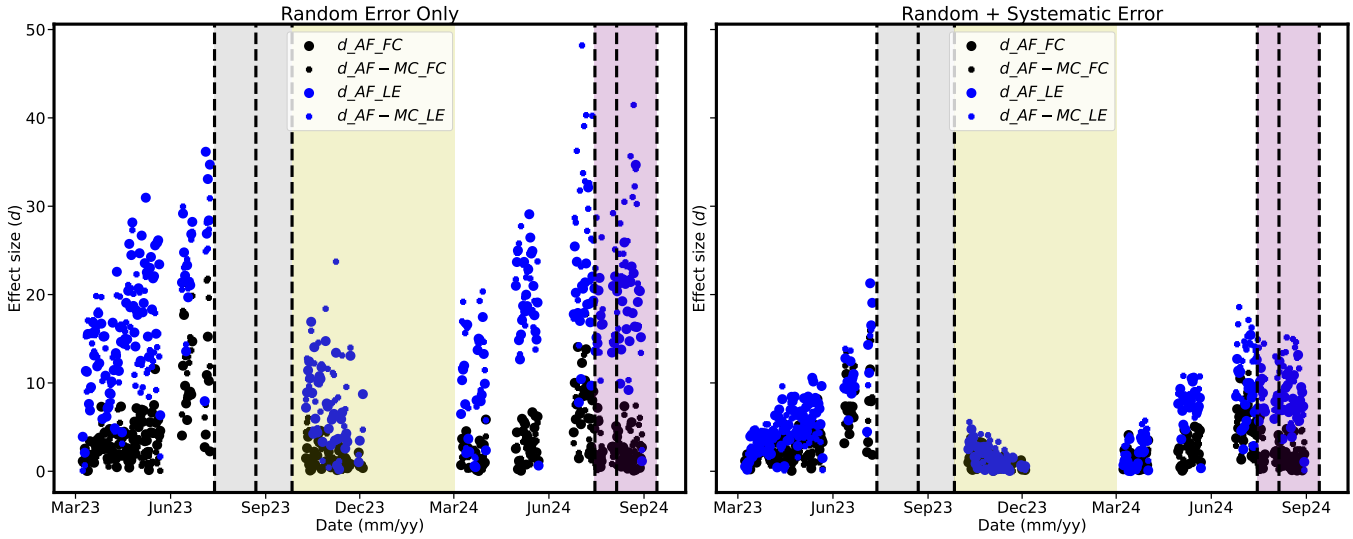

Figure AR9: Effect size (d) calculated for the comparison of AF1 vs. the average of the three stations at the AF (d_AF_FC and d_AF_LE), and for the comparison of the average of the three stations at the AF and the MC ($d\_AF\text{-}MC\_FC$ and $d\_AF\text{-}MC\_LE$). The left plot shows the case in which the measurement errors were taken as the random error. The right plot shows the case in which the measurement errors were taken as the sum of the random error plus the RMSE of the previous intercomparison campaign in 2022.

In order to not extend excessively this document, we did not include the figures corresponding to the effect size calculation for monthly and annual sums of FC and LE. The extra figures could be included in the Appendix, to backup the discussion, or maybe a panel plot showing the different effect size time series calculated at different temporal scales. Additionally, we calculated the effect size for the comparison of individual stations.

R2_C9. L700-702: I think this is a very important point that could lead to misinterpretation of results. In such heterogenous surfaces, the development of the homogeneous surface layer is not obvious and turbulent and mean flux divergence in the horizontal and vertical directions might be important. Please add a comment on how missing information on these processes could have affected your results.

AR_C9. We will add an extra paragraph to the discussion on this subject.

2.3 TECHNICAL CORRECTIONS

R2_C10. L10-11: please rephrase this sentence, I cannot find the subject of "contributed".

R2_C11. L66: Markwitz and Siebicke (2019) should be in parentheses

R2_C12. L127: close the parentheses after "NETRAD".

R2_C13. L209: 2 should be 3 instead.

R2_C14. L218: “developed” instead of “developped”

R2_C15. L258: should U be WS instead?

R2_C16. L274-275: should C be FC instead?

R2_C17. L305: please add “total” to “monthly values of P”

AR_Technical corrections. All the small technical corrections are addressed and changed.

References

Callejas-Rodelas, J. Á., Knohl, A., van Ramshorst, J., Mammarella, I., and Markwitz, C.: Comparison between Lower-Cost and Conventional Eddy Covariance Setups for CO_2 and Evapotranspiration Measurements above Monocropping and Agroforestry Systems, Agricultural and Forest Meteorology, 354, 110086, https://doi.org/10.1016/j.agrformet.2024.110086, 2024.

Chu, H., Baldocchi, D. D., Poindexter, C., Abraha, M., Desai, A. R., Bohrer, G., Arain, M. A., Griffis, T., Blanken, P. D., O'Halloran, T. L., Thomas, R. Q., Zhang, Q., Burns, S. P., Frank, J. M., Christian, D., Brown, S., Black, T. A., Gough, C. M., Law, B. E., Lee, X., Chen, J., Reed, D. E., Massman, W. J., Clark, K., Hatfield, J., Prueger, J., Bracho, R., Baker, J. M., and Martin, T. A.: Temporal Dynamics of Aerodynamic Canopy Height Derived From Eddy Covariance Momentum Flux Data Across North American Flux Networks, Geophysical Research Letters, 45, 9275–9287, https://doi.org/10.1029/2018GL079306, 2018.

van Ramshorst, J. G. V., Knohl, A., Callejas-Rodelas, J. Á, Clement, R., Hill, T. C., Siebicke, L., and Markwitz, C.: Lower-Cost Eddy Covariance for CO_2 and H_2O Fluxes over Grassland and Agroforestry, https://doi.org/10.5194/amt-2024-30, 2024.