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.

In this document, the specific changes introduced after the major reviewer's comments are explained and shown. Minor changes are not included to reduce the extension of the document.

1 Reviewer 1

1.1 General comments

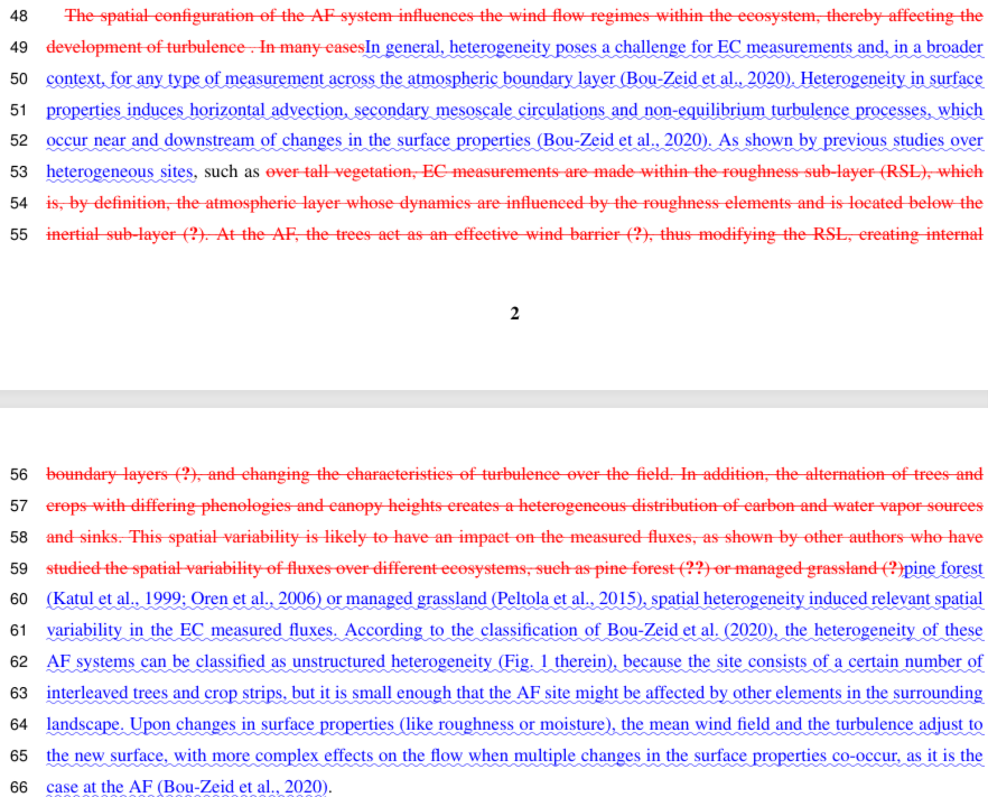
R1_General comment. This manuscript describes the results of a field experiment with three low-cost eddy-covariance systems over a patchy agroforestry system and a patchy monocropping system. By analyzing these data from two growing seasons, the authors attempt to answer the question that is raised in 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 was 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 was added to the introduction and the discussion sections, as shown in figures AR1 and AR2.

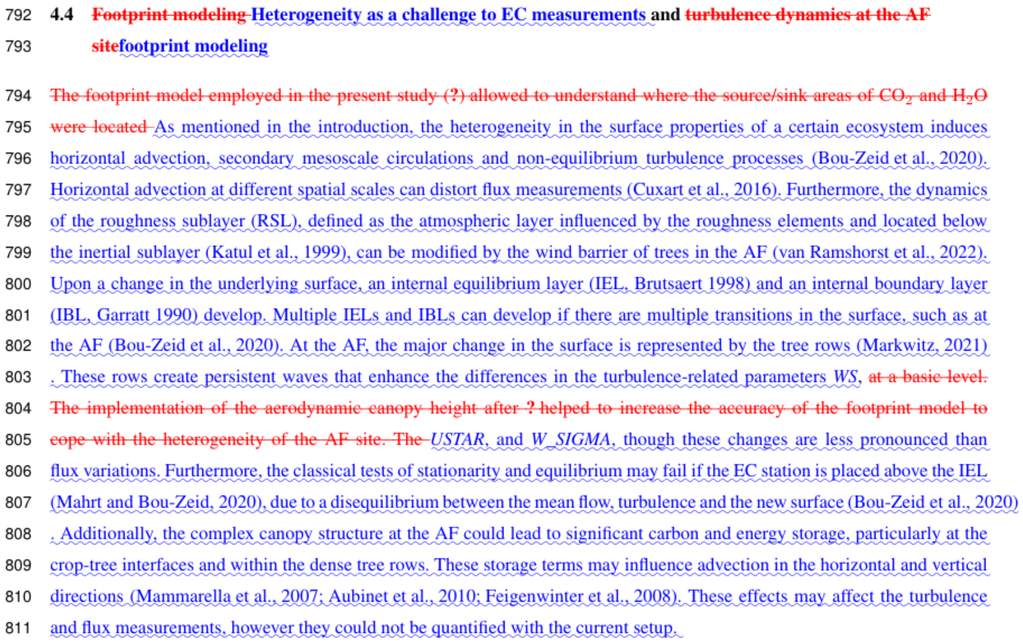


~~The spatial configuration of the AF system influences the wind flow regimes within the ecosystem, thereby affecting the development of turbulence . In many cases~~In general, heterogeneity poses a challenge for EC measurements and, in a broader context, for any type of measurement across the atmospheric boundary layer (Bou-Zeid et al., 2020). Heterogeneity in surface properties induces horizontal advection, secondary mesoscale circulations and non-equilibrium turbulence processes, which occur near and downstream of changes in the surface properties (Bou-Zeid et al., 2020). As shown by previous studies over heterogeneous sites, such as ~~over tall vegetation, EC measurements are made within the roughness sub-layer (RSL), which is, by definition, the atmospheric layer whose dynamics are influenced by the roughness elements and is located below the inertial sub-layer (?). At the AF, the trees act as an effective wind barrier (?), thus modifying the RSL, creating internal~~

2

~~boundary layers (?), and changing the characteristics of turbulence over the field. In addition, the alternation of trees and crops with differing phenologies and canopy heights creates a heterogeneous distribution of carbon and water vapor sources and sinks. This spatial variability is likely to have an impact on the measured fluxes, as shown by other authors who have studied the spatial variability of fluxes over different ecosystems, such as pine forest (??) or managed grassland (?)~~pine forest (Katul et al., 1999; Oren et al., 2006) or managed grassland (Peltola et al., 2015), spatial heterogeneity induced relevant spatial variability in the EC measured fluxes. According to the classification of Bou-Zeid et al. (2020), the heterogeneity of these AF systems can be classified as unstructured heterogeneity (Fig. 1 therein), because the site consists of a certain number of interleaved trees and crop strips, but it is small enough that the AF site might be affected by other elements in the surrounding landscape. Upon changes in surface properties (like roughness or moisture), the mean wind field and the turbulence adjust to the new surface, with more complex effects on the flow when multiple changes in the surface properties co-occur, as it is the case at the AF (Bou-Zeid et al., 2020).

Figure AR1: New paragraph in introduction related to heterogeneity.



4.4 ~~Footprint modeling~~ Heterogeneity as a challenge to EC measurements and ~~turbulence dynamics at the AF site~~footprint modeling

~~The footprint model employed in the present study (?) allowed to understand where the source/sink areas of CO_2 and H_2O were located~~ As mentioned in the introduction, the heterogeneity in the surface properties of a certain ecosystem induces horizontal advection, secondary mesoscale circulations and non-equilibrium turbulence processes (Bou-Zeid et al., 2020). Horizontal advection at different spatial scales can distort flux measurements (Cuxart et al., 2016). Furthermore, the dynamics of the roughness sublayer (RSL), defined as the atmospheric layer influenced by the roughness elements and located below the inertial sublayer (Katul et al., 1999), can be modified by the wind barrier of trees in the AF (van Ramshorst et al., 2022). Upon a change in the underlying surface, an internal equilibrium layer (IEL, Brutsaert 1998) and an internal boundary layer (IBL, Garratt 1990) develop. Multiple IELs and IBLs can develop if there are multiple transitions in the surface, such as at the AF (Bou-Zeid et al., 2020). At the AF, the major change in the surface is represented by the tree rows (Markwitz, 2021) . These rows create persistent waves that enhance the differences in the turbulence-related parameters *WS*, ~~at a basic level. The implementation of the aerodynamic canopy height after ? helped to increase the accuracy of the footprint model to cope with the heterogeneity of the AF site. The~~ *USTAR*, and *W_SIGMA*, though these changes are less pronounced than flux variations. Furthermore, the classical tests of stationarity and equilibrium may fail if the EC station is placed above the IEL (Mahrt and Bou-Zeid, 2020), due to a disequilibrium between the mean flow, turbulence and the new surface (Bou-Zeid et al., 2020) . Additionally, the complex canopy structure at the AF could lead to significant carbon and energy storage, particularly at the crop-tree interfaces and within the dense tree rows. These storage terms may influence advection in the horizontal and vertical directions (Mammarella et al., 2007; Aubinet et al., 2010; Feigenwinter et al., 2008). These effects may affect the turbulence and flux measurements, however they could not be quantified with the current setup.

Figure AR2: New text in discussion related to heterogeneity.

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. AR15), 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 changed in the text that sentence and the related information accordingly, 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 clarified 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). No screenshot of the tracked changes manuscript is attached because the changes correspond to different sections of the text.

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

AR_C4. We appreciate the comment, but 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 revised 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 was 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 AR3 you can find an example fit of the time response vs. RH, as a direct output from the processing software EddyUH. Figure AR4 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 were updated in the manuscript (some of the screenshots in this author's response already include the new plots).

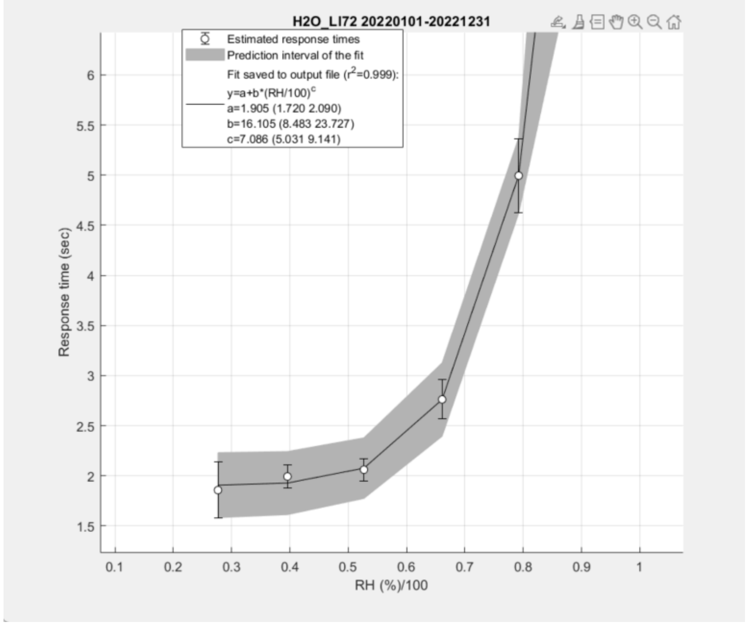


Figure AR3: Example of fit of the time response with the equation y=a+b·($\frac{RH}{100}^c$). The example corresponds to Wendhausen MC station in 2022.

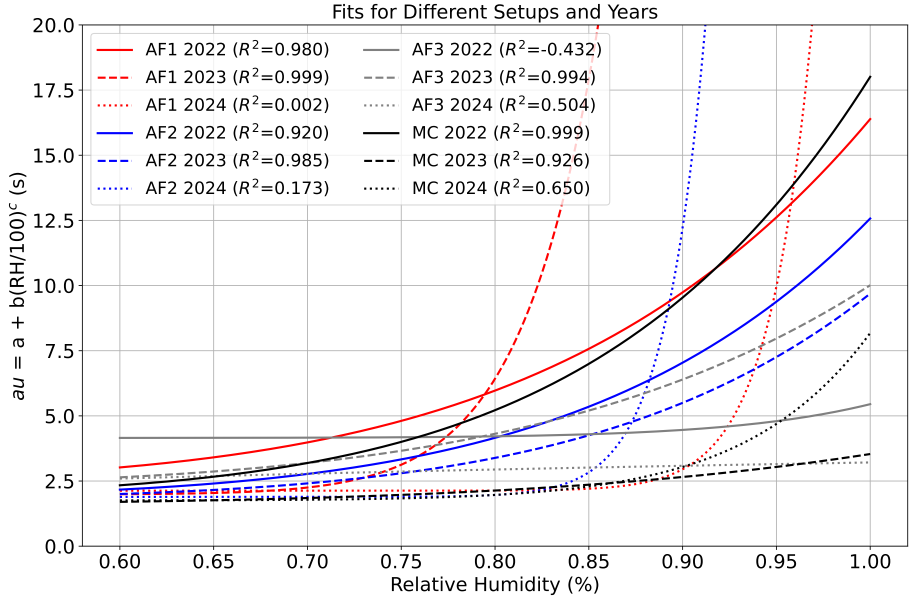


Figure AR4: Fit of the time response dependent on RH for all the stations (AF1 in red, AF2 in blue, AF3 in grey and MC in black), for the years 2022 (solid line), 2023 (dashed line) and 2024 (dotted line). Next to the legend labels the coefficients of determination (r^2) for the fit are displayed.

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. AR5), 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. AR6) 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 kept the data filtering and gap-filling as in the original manuscript for Figures 4 and 7, which need weekly and daily sums, respectively. The new figures 5 and 6 only included filtered data but not gap-filled data. This distinction was clarified in the Methods section (screenshot in Fig. AR7).

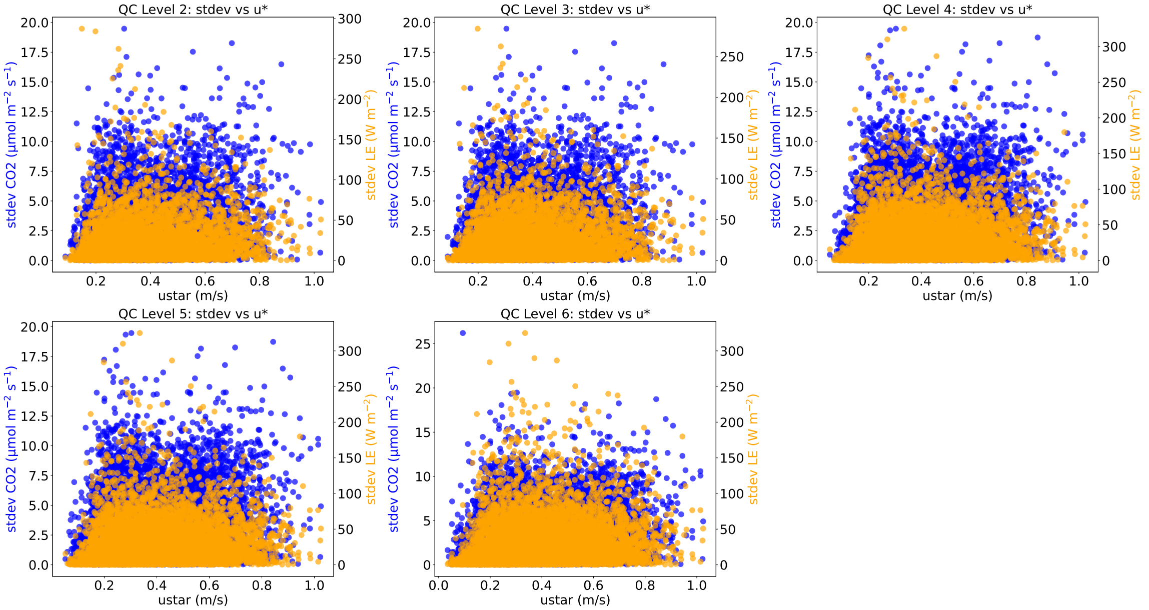


Figure AR5: Spatial standard deviation of FC (blue, left y-axis) and LE (orange, right y-axis) across the three AF stations for different levels of filtered data (quality flags ranging from 1 to 6, 1 to 5, 1 to 4, 1 to 3 and 1 to 2), depending on friction velocity.

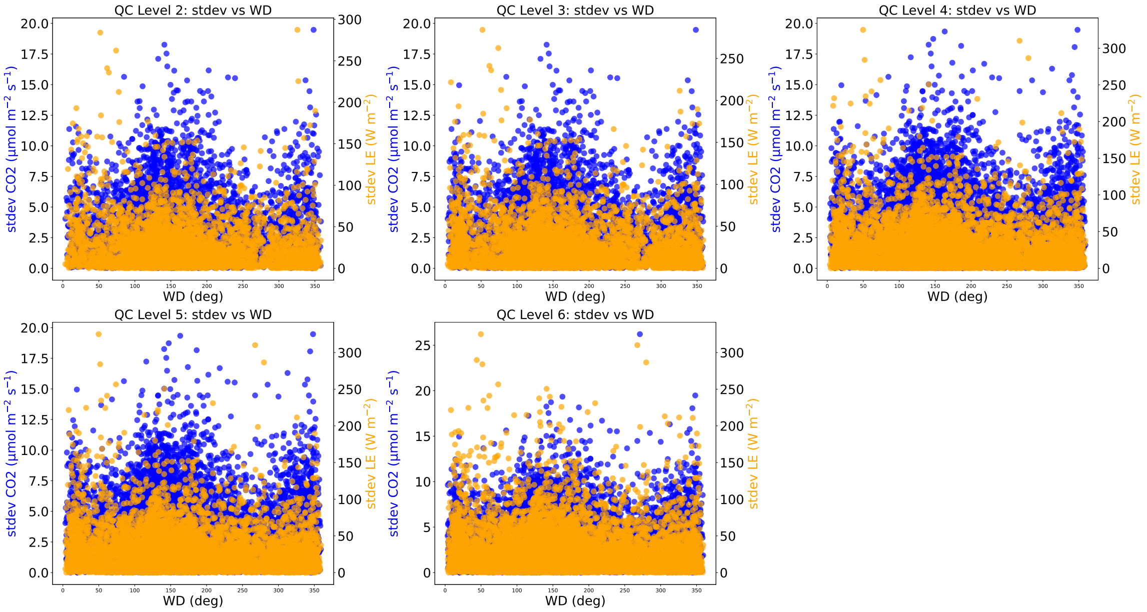
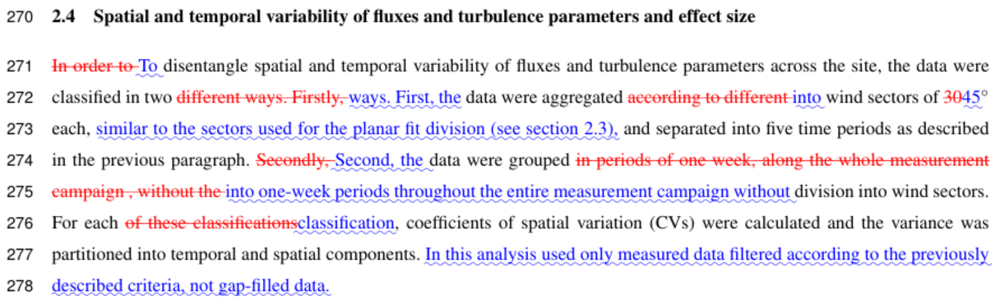


Figure AR6: Spatial standard deviation of FC (blue, left y-axis) and LE (orange, right y-axis) across the three AF stations for different levels of filtered data (quality flags ranging from 1 to 6, 1 to 5, 1 to 4, 1 to 3 and 1 to 2), depending on wind direction.



2.4 Spatial and temporal variability of fluxes and turbulence parameters and effect size

~~In order to~~ To disentangle spatial and temporal variability of fluxes and turbulence parameters across the site, the data were classified in two ~~different ways. Firstly,~~ ways. First, the data were aggregated ~~according to different~~ into wind sectors of ~~30~~45° each, similar to the sectors used for the planar fit division (see section 2.3), and separated into five time periods as described in the previous paragraph. ~~Secondly,~~ Second, the data were grouped ~~in periods of one week, along the whole measurement campaign , without the~~ into one-week periods throughout the entire measurement campaign without division into wind sectors. For each ~~of these classifications~~classification, coefficients of spatial variation (CVs) were calculated and the variance was partitioned into temporal and spatial components. In this analysis used only measured data filtered according to the previously described criteria, not gap-filled data.

Figure AR7: Change in methods section to clarify which data were used to classify in wind sectors and weekly intervals.

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 made some changes in the manuscript. Figures 4 (weekly sums of carbon and ET) and 7 (effect size comparing daily sums) kept gap-filled data, otherwise sums cannot be calculated. On the other hand, Figures 5 and 6 were 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. AR13) related to this topic is explained. Below you can find the new figures (5 and 6 in the manuscript) included in the manuscript as screenshots. The clarification that no gap-filled data were used was included in the caption. The text was also changed accordingly in the manuscript (screenshots AR10 and AR11).

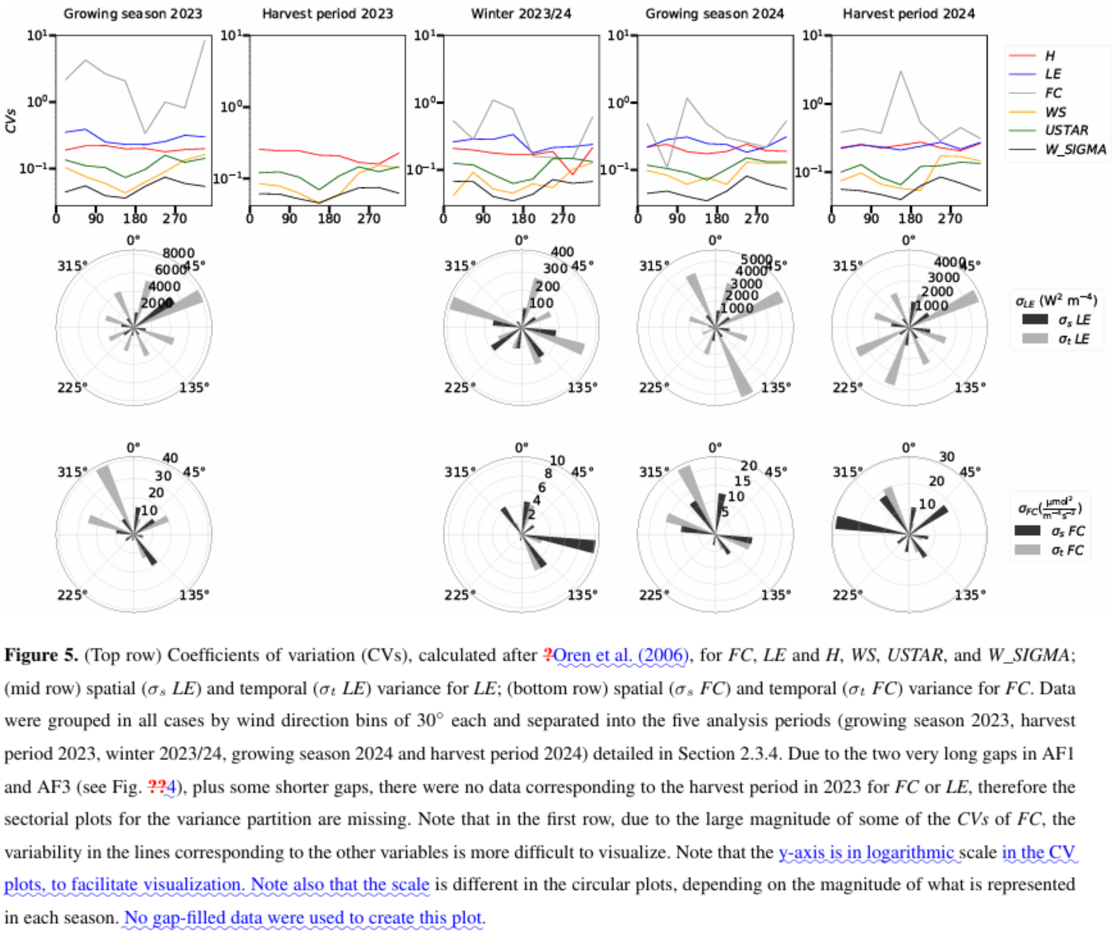


Figure 5. (Top row) Coefficients of variation (CVs), calculated after ~~?~~Oren et al. (2006), for *FC*, *LE* and *H*, *WS*, *USTAR*, and *W_SIGMA*; (mid row) spatial (σ_s *LE*) and temporal (σ_t *LE*) variance for *LE*; (bottom row) spatial (σ_s *FC*) and temporal (σ_t *FC*) variance for *FC*. Data were grouped in all cases by wind direction bins of 30° each and separated into the five analysis periods (growing season 2023, harvest period 2023, winter 2023/24, growing season 2024 and harvest period 2024) detailed in Section 2.3.4. Due to the two very long gaps in AF1 and AF3 (see Fig. ~~??~~4), plus some shorter gaps, there were no data corresponding to the harvest period in 2023 for *FC* or *LE*, therefore the sectorial plots for the variance partition are missing. Note that in the first row, due to the large magnitude of some of the *CVs* of *FC*, the variability in the lines corresponding to the other variables is more difficult to visualize. Note that the y-axis is in logarithmic scale in the CV plots, to facilitate visualization. Note also that the scale is different in the circular plots, depending on the magnitude of what is represented in each season. No gap-filled data were used to create this plot.

Figure AR8: Changes in Figure 5.

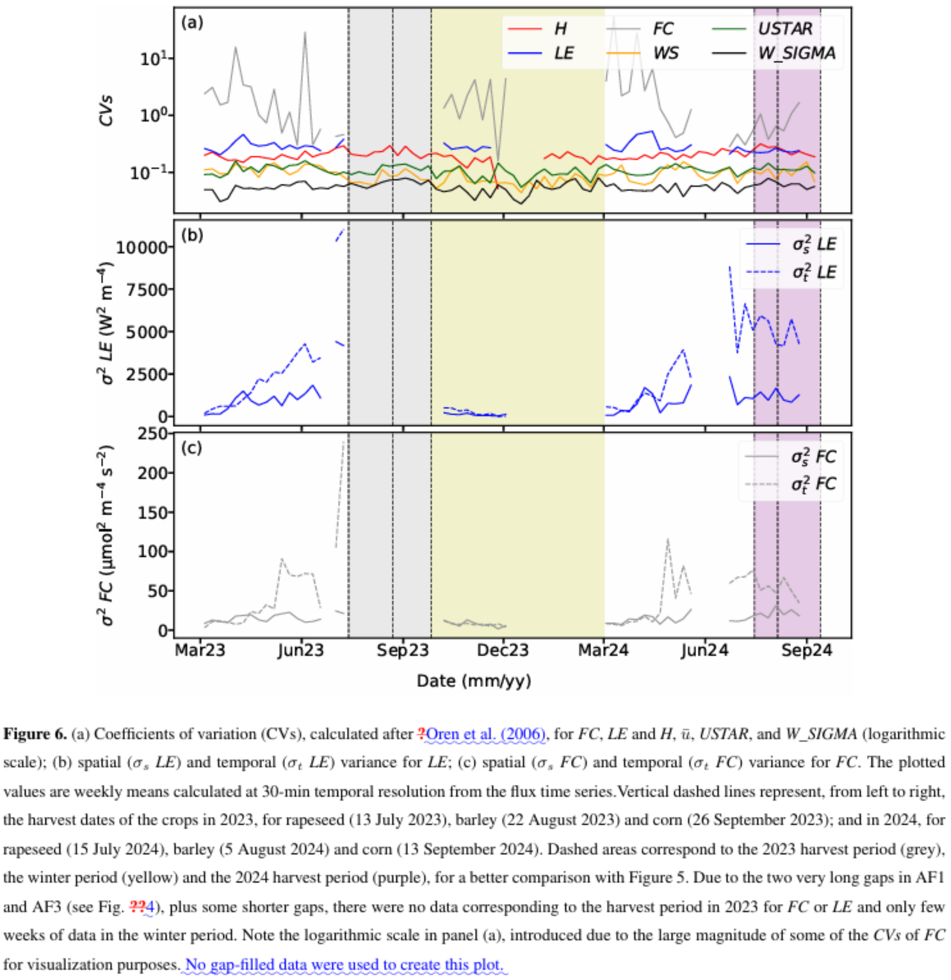
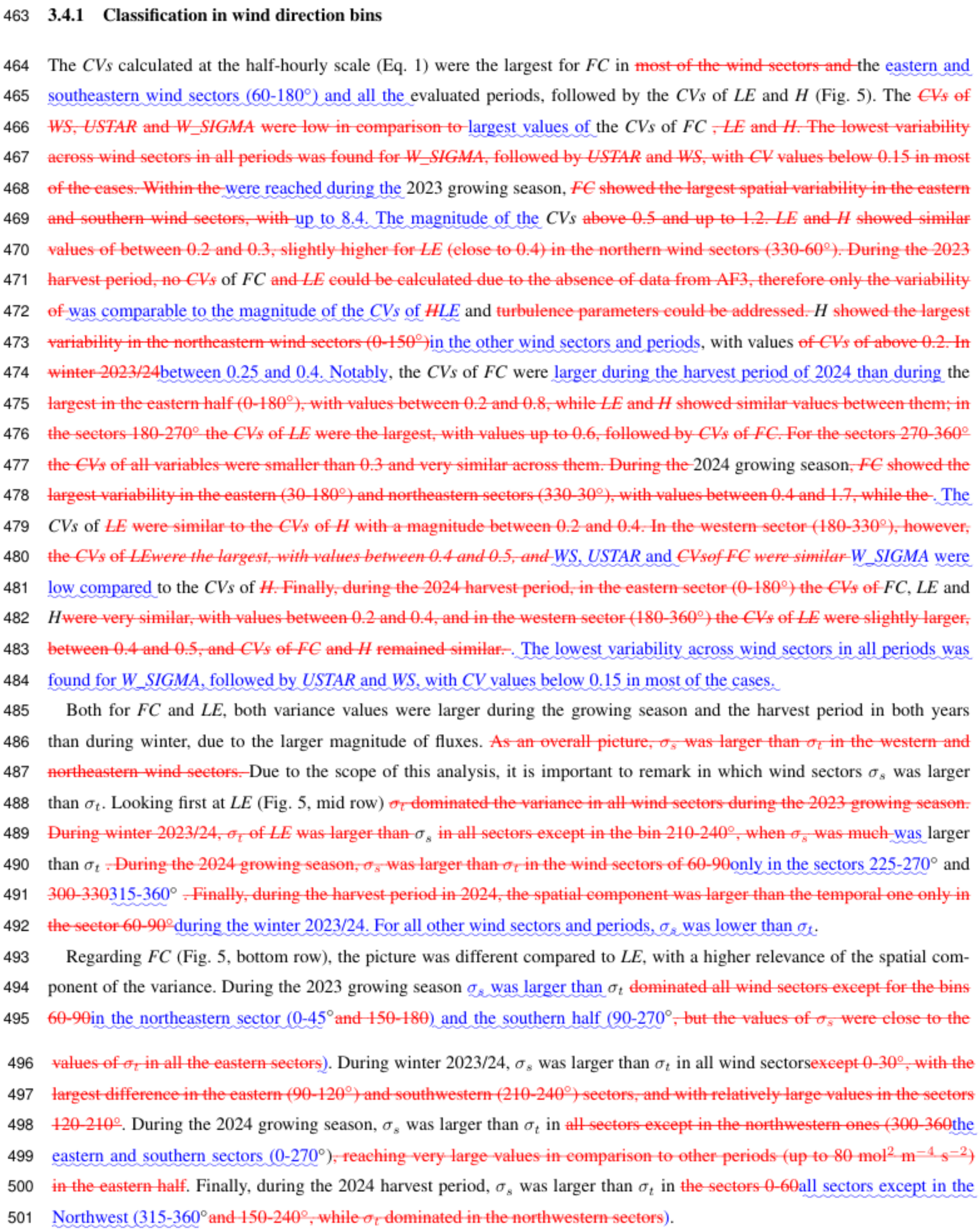


Figure 6. (a) Coefficients of variation (CVs), calculated after ~~?~~Oren et al. (2006), for *FC*, *LE* and *H*, $\bar{u}$, *USTAR*, and *W_SIGMA* (logarithmic scale); (b) spatial (σ_s *LE*) and temporal (σ_t *LE*) variance for *LE*; (c) spatial (σ_s *FC*) and temporal (σ_t *FC*) variance for *FC*. The plotted values are weekly means calculated at 30-min temporal resolution from the flux time series.Vertical dashed lines represent, from left to right, the harvest dates of the crops in 2023, for rapeseed (13 July 2023), barley (22 August 2023) and corn (26 September 2023); and in 2024, for rapeseed (15 July 2024), barley (5 August 2024) and corn (13 September 2024). Dashed areas correspond to the 2023 harvest period (grey), the winter period (yellow) and the 2024 harvest period (purple), for a better comparison with Figure 5. Due to the two very long gaps in AF1 and AF3 (see Fig. ~~??~~4), plus some shorter gaps, there were no data corresponding to the harvest period in 2023 for *FC* or *LE* and only few weeks of data in the winter period. Note the logarithmic scale in panel (a), introduced due to the large magnitude of some of the *CVs* of *FC* for visualization purposes. No gap-filled data were used to create this plot.

Figure AR9: Changes in Figure 6.



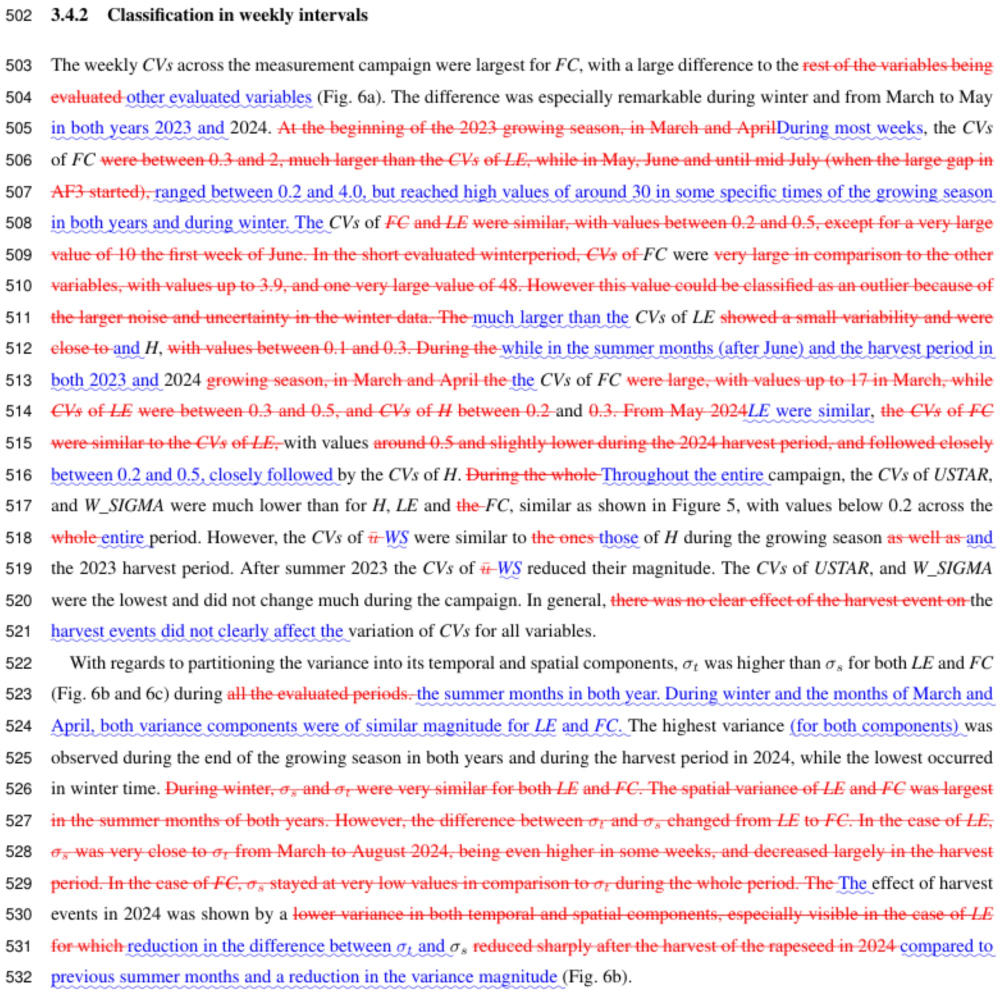
3.4.1 Classification in wind direction bins

The *CVs* calculated at the half-hourly scale (Eq. 1) were the largest for *FC* in ~~most of the wind sectors and~~ the eastern and southeastern wind sectors (60-180°) and all the evaluated periods, followed by the *CVs* of *LE* and *H* (Fig. 5). The ~~*CVs* of *WS*, *USTAR* and *W_SIGMA* were low in comparison to~~ largest values of the *CVs* of *FC* ~~, *LE* and *H*. The lowest variability across wind sectors in all periods was found for *W_SIGMA*, followed by *USTAR* and *WS*, with *CV* values below 0.15 in most of the cases. Within the~~ were reached during the 2023 growing season, ~~*FC* showed the largest spatial variability in the eastern and southern wind sectors, with~~ up to 8.4. The magnitude of the *CVs* ~~above 0.5 and up to 1.2. *LE* and *H* showed similar values of between 0.2 and 0.3, slightly higher for *LE* (close to 0.4) in the northern wind sectors (330-60°). During the 2023 harvest period, no *CVs*~~ of *FC* ~~and *LE* could be calculated due to the absence of data from AF3, therefore only the variability of~~ was comparable to the magnitude of the *CVs* of ~~*H*~~*LE* and ~~turbulence parameters could be addressed.~~ *H* ~~showed the largest variability in the northeastern wind sectors (0-150°)~~in the other wind sectors and periods, with values ~~of *CVs* of above 0.2. In winter 2023/24~~between 0.25 and 0.4. Notably, the *CVs* of *FC* were larger during the harvest period of 2024 than during the ~~largest in the eastern half (0-180°), with values between 0.2 and 0.8, while *LE* and *H* showed similar values between them; in the sectors 180-270° the *CVs* of *LE* were the largest, with values up to 0.6, followed by *CVs* of *FC*. For the sectors 270-360° the *CVs* of all variables were smaller than 0.3 and very similar across them. During the~~ 2024 growing season~~, *FC* showed the largest variability in the eastern (30-180°) and northeastern sectors (330-30°), with values between 0.4 and 1.7, while the~~. The *CVs* of ~~*LE* were similar to the *CVs* of *H* with a magnitude between 0.2 and 0.4. In the western sector (180-330°), however, the *CVs* of *LE*were the largest, with values between 0.4 and 0.5, and~~ *WS*, *USTAR* and ~~*CVs*of *FC* were similar~~ *W_SIGMA* were low compared to the *CVs* of ~~*H*. Finally, during the 2024 harvest period, in the eastern sector (0-180°) the *CVs* of~~ *FC*, *LE* and *H*~~were very similar, with values between 0.2 and 0.4, and in the western sector (180-360°) the *CVs* of *LE* were slightly larger, between 0.4 and 0.5, and *CVs* of *FC* and *H* remained similar.~~. The lowest variability across wind sectors in all periods was found for *W_SIGMA*, followed by *USTAR* and *WS*, with *CV* values below 0.15 in most of the cases.

Both for *FC* and *LE*, both variance values were larger during the growing season and the harvest period in both years than during winter, due to the larger magnitude of fluxes. ~~As an overall picture, σ_s was larger than σ_t in the western and northeastern wind sectors.~~ Due to the scope of this analysis, it is important to remark in which wind sectors σ_s was larger than σ_t. Looking first at *LE* (Fig. 5, mid row) ~~σ_t dominated the variance in all wind sectors during the 2023 growing season. During winter 2023/24, σ_t of *LE* was larger than~~ σ_s ~~in all sectors except in the bin 210-240°, when σ_s was much~~ was larger than σ_t ~~. During the 2024 growing season, σ_s was larger than σ_t in the wind sectors of 60-90~~only in the sectors 225-270° and ~~300-330~~315-360° ~~. Finally, during the harvest period in 2024, the spatial component was larger than the temporal one only in the sector 60-90°~~during the winter 2023/24. For all other wind sectors and periods, σ_s was lower than σ_t.

Regarding *FC* (Fig. 5, bottom row), the picture was different compared to *LE*, with a higher relevance of the spatial component of the variance. During the 2023 growing season σ_s was larger than σ_t ~~dominated all wind sectors except for the bins 60-90~~in the northeastern sector (0-45°~~and 150-180~~) and the southern half (90-270°~~, but the values of σ_s were close to the values of σ_t in all the eastern sectors~~). During winter 2023/24, σ_s was larger than σ_t in all wind sectors~~except 0-30°, with the largest difference in the eastern (90-120°) and southwestern (210-240°) sectors, and with relatively large values in the sectors 120-210°~~. During the 2024 growing season, σ_s was larger than σ_t in ~~all sectors except in the northwestern ones (300-360~~the eastern and southern sectors (0-270°)~~, reaching very large values in comparison to other periods (up to 80 mol^2 m^{-4} s^{-2}) in the eastern half~~. Finally, during the 2024 harvest period, σ_s was larger than σ_t in ~~the sectors 0-60~~all sectors except in the Northwest (315-360°~~and 150-240°, while σ_t dominated in the northwestern sectors~~).

Figure AR10: Changes in description of results from Figure 5.



3.4.2 Classification in weekly intervals

The weekly *CVs* across the measurement campaign were largest for *FC*, with a large difference to the ~~rest of the variables being evaluated~~ other evaluated variables (Fig. 6a). The difference was especially remarkable during winter and from March to May in both years 2023 and 2024. ~~At the beginning of the 2023 growing season, in March and April~~During most weeks, the *CVs* of *FC* ~~were between 0.3 and 2, much larger than the *CVs* of *LE*, while in May, June and until mid July (when the large gap in AF3 started),~~ ranged between 0.2 and 4.0, but reached high values of around 30 in some specific times of the growing season in both years and during winter. The *CVs* of ~~*FC* and *LE* were similar, with values between 0.2 and 0.5, except for a very large value of 10 the first week of June. In the short evaluated winterperiod, *CVs* of~~ *FC* were ~~very large in comparison to the other variables, with values up to 3.9, and one very large value of 48. However this value could be classified as an outlier because of the larger noise and uncertainty in the winter data. The~~ much larger than the *CVs* of *LE* ~~showed a small variability and were close to~~ and *H*, ~~with values between 0.1 and 0.3. During the~~ while in the summer months (after June) and the harvest period in both 2023 and 2024 ~~growing season, in March and April the~~ the *CVs* of *FC* ~~were large, with values up to 17 in March, while *CVs* of *LE* were between 0.3 and 0.5, and *CVs* of *H* between 0.2~~ and ~~0.3. From May 2024~~*LE* were similar, ~~the *CVs* of *FC* were similar to the *CVs* of *LE*,~~ with values ~~around 0.5 and slightly lower during the 2024 harvest period, and followed closely~~ between 0.2 and 0.5, closely followed by the *CVs* of *H*. ~~During the whole~~ Throughout the entire campaign, the *CVs* of *USTAR*, and *W_SIGMA* were much lower than for *H*, *LE* and ~~the~~ *FC*, similar as shown in Figure 5, with values below 0.2 across the ~~whole~~ entire period. However, the *CVs* of ~~$\bar{u}$~~ *WS* were similar to ~~the ones~~ those of *H* during the growing season ~~as well as~~ and the 2023 harvest period. After summer 2023 the *CVs* of ~~$\bar{u}$~~ *WS* reduced their magnitude. The *CVs* of *USTAR*, and *W_SIGMA* were the lowest and did not change much during the campaign. In general, ~~there was no clear effect of the harvest event on~~ the harvest events did not clearly affect the variation of *CVs* for all variables.

With regards to partitioning the variance into its temporal and spatial components, σ_t was higher than σ_s for both *LE* and *FC* (Fig. 6b and 6c) during ~~all the evaluated periods.~~ the summer months in both year. During winter and the months of March and April, both variance components were of similar magnitude for *LE* and *FC*. The highest variance (for both components) was observed during the end of the growing season in both years and during the harvest period in 2024, while the lowest occurred in winter time. ~~During winter, σ_s and σ_t were very similar for both *LE* and *FC*. The spatial variance of *LE* and *FC* was largest in the summer months of both years. However, the difference between σ_t and σ_s changed from *LE* to *FC*. In the case of *LE*, σ_s was very close to σ_t from March to August 2024, being even higher in some weeks, and decreased largely in the harvest period. In the case of *FC*, σ_s stayed at very low values in comparison to σ_t during the whole period. The~~ The effect of harvest events in 2024 was shown by a ~~lower variance in both temporal and spatial components, especially visible in the case of *LE* for which~~ reduction in the difference between σ_t and σ_s ~~reduced sharply after the harvest of the rapeseed in 2024~~ compared to previous summer months and a reduction in the variance magnitude (Fig. 6b).

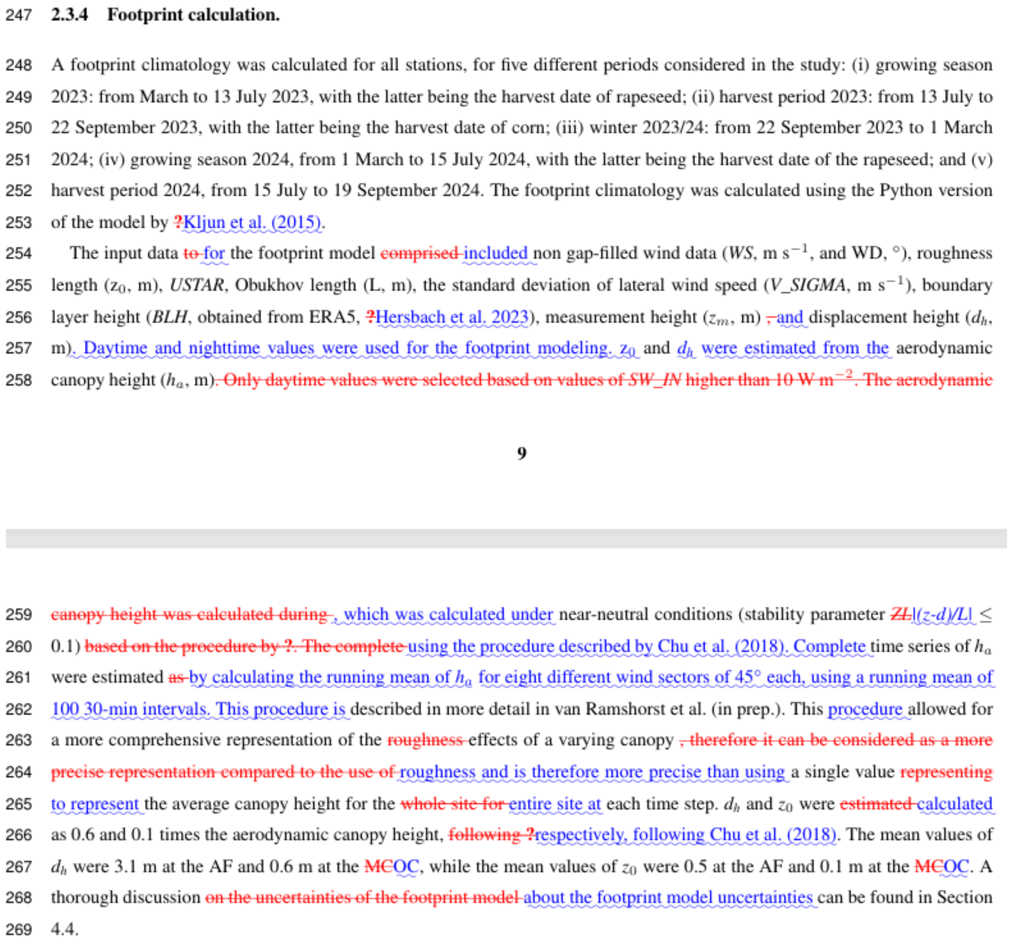
Figure AR11: Changes in description of results from Figure 6.

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 (see Fig. AR12 below).



2.3.4 Footprint calculation.

A footprint climatology was calculated for all stations, for five different periods considered in the study: (i) growing season 2023: from March to 13 July 2023, with the latter being the harvest date of rapeseed; (ii) harvest period 2023: from 13 July to 22 September 2023, with the latter being the harvest date of corn; (iii) winter 2023/24: from 22 September 2023 to 1 March 2024; (iv) growing season 2024, from 1 March to 15 July 2024, with the latter being the harvest date of the rapeseed; and (v) harvest period 2024, from 15 July to 19 September 2024. The footprint climatology was calculated using the Python version of the model by ~~?~~Kljun et al. (2015).

The input data ~~to~~ for the footprint model ~~comprised~~ included non gap-filled wind data (*WS*, m s^{-1}, and WD, °), roughness length (z_0, m), *USTAR*, Obukhov length (L, m), the standard deviation of lateral wind speed (*V_SIGMA*, m s^{-1}), boundary layer height (*BLH*, obtained from ERA5, ~~?~~Hersbach et al. 2023), measurement height (z_m, m) ~~,~~ and displacement height (d_h, m). Daytime and nighttime values were used for the footprint modeling. z_0 and d_h were estimated from the aerodynamic canopy height (h_a, m)~~. Only daytime values were selected based on values of *SW_IN* higher than 10 W m^{-2}. The aerodynamic canopy height was calculated during~~ , which was calculated under near-neutral conditions (stability parameter ~~ZL~~ $|(z\text{-}d)/L| \leq$ 0.1) ~~based on the procedure by ?. The complete~~ using the procedure described by Chu et al. (2018). Complete time series of h_a were estimated ~~as~~ by calculating the running mean of h_a for eight different wind sectors of 45° each, using a running mean of 100 30-min intervals. This procedure is described in more detail in van Ramshorst et al. (in prep.). This procedure allowed for a more comprehensive representation of the ~~roughness~~ effects of a varying canopy ~~, therefore it can be considered as a more precise representation compared to the use of~~ roughness and is therefore more precise than using a single value ~~representing~~ to represent the average canopy height for the ~~whole site for~~ entire site at each time step. d_h and z_0 were ~~estimated~~ calculated as 0.6 and 0.1 times the aerodynamic canopy height, ~~following ?~~respectively, following Chu et al. (2018). The mean values of d_h were 3.1 m at the AF and 0.6 m at the ~~MC~~OC, while the mean values of z_0 were 0.5 at the AF and 0.1 m at the ~~MC~~OC. A thorough discussion ~~on the uncertainties of the footprint model~~ about the footprint model uncertainties can be found in Section 4.4.

Figure AR12: Changes in description of aerodynamic canopy height and footprint calculation.

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 **??** 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. AR13, 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. AR14, 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 changed Figures 5 and 6 to use only measured data, while Figures 4 and 7, that need weekly and daily sums respectively, were kept as they are with gap filled data.

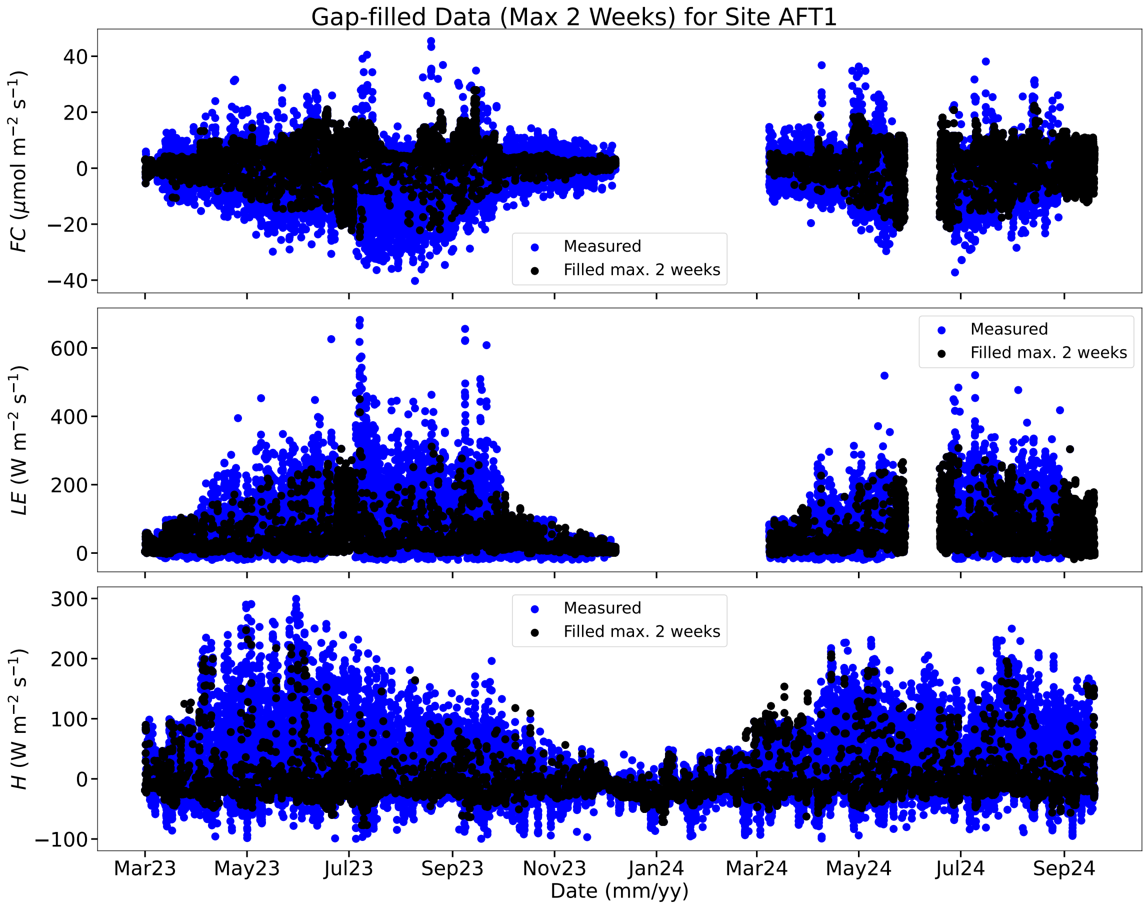


Figure AR13: Time series of measured (blue) and gap-filled (black) data considering gaps with a maximum duration of 2 weeks, for one example station (AF1). Those are the actual data used in the study.

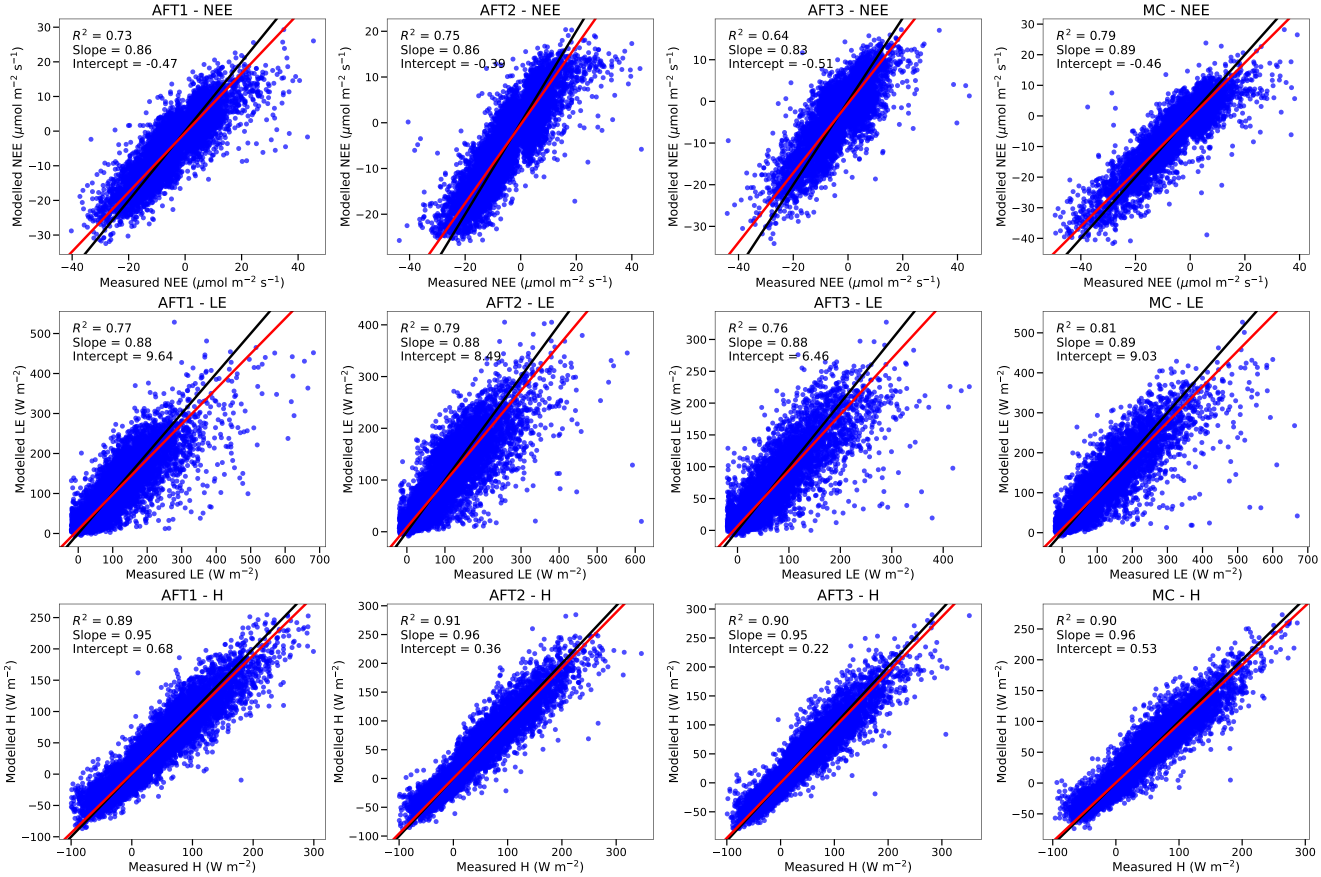


Figure AR14: Scatter plots of modeled vs. measured data, for FC, LE and H for the four stations used in the study (AF1, left column, AF2, second column, AF3, third column, MC, fourth column). The linear models were fitted on all the measured and corresponding modeled data.

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 AR19. Please also see below a plot of the different errors considered in this study (Fig. AR15). 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.

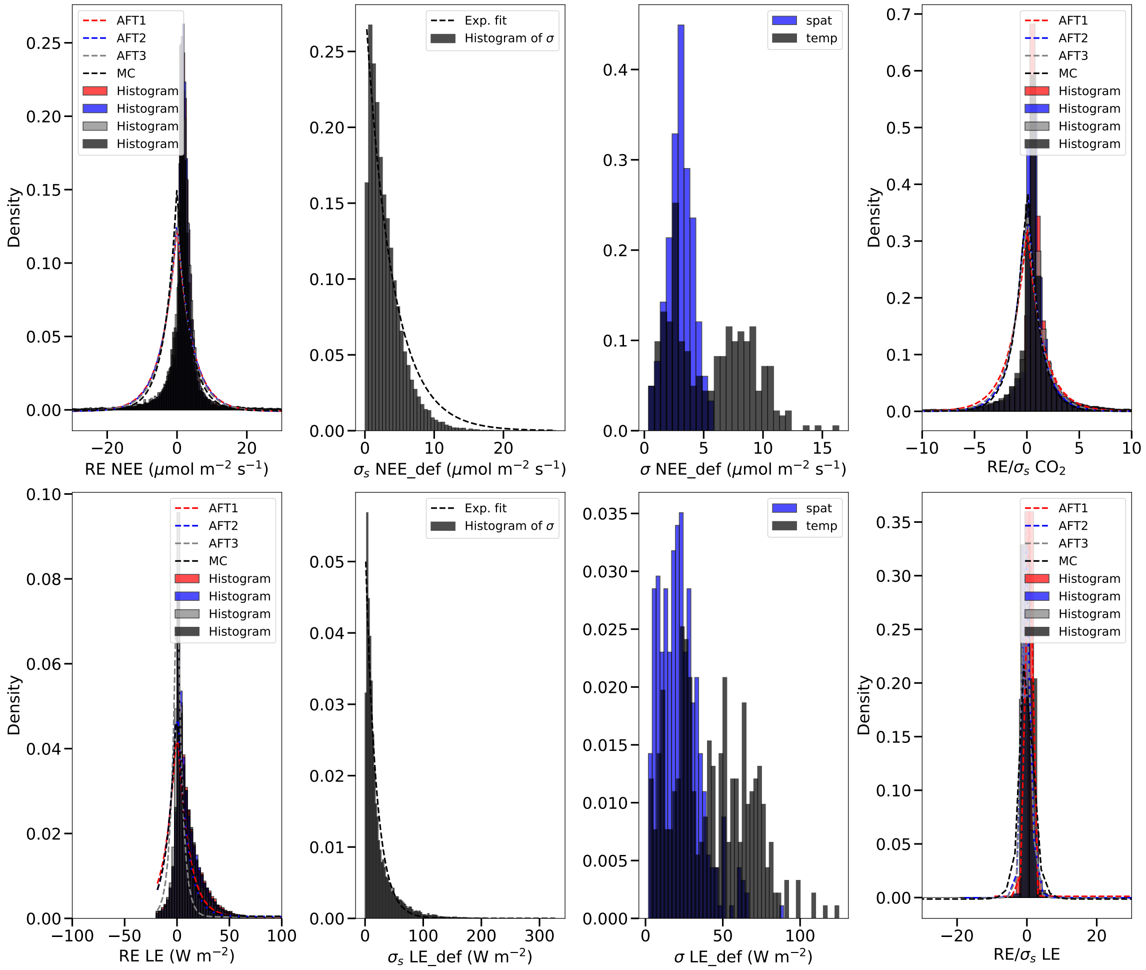


Figure AR15: (Left column) Random error of FC (top) and LE (bottom) for all the stations, with the double exponential fit to the histograms. (Second column) Histogram with exponential fit of the spatial standard deviation calculated at the 30-min time scale. (Third column) Histograms of the spatial and 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 addressed 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. Please refer to figures AR1 and AR2 to see the major changes in the text related to this. There were other minor changes as well that are not included here.

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 addressed 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 was 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. All the changes are shown in the tracked changes document.

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” was 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.

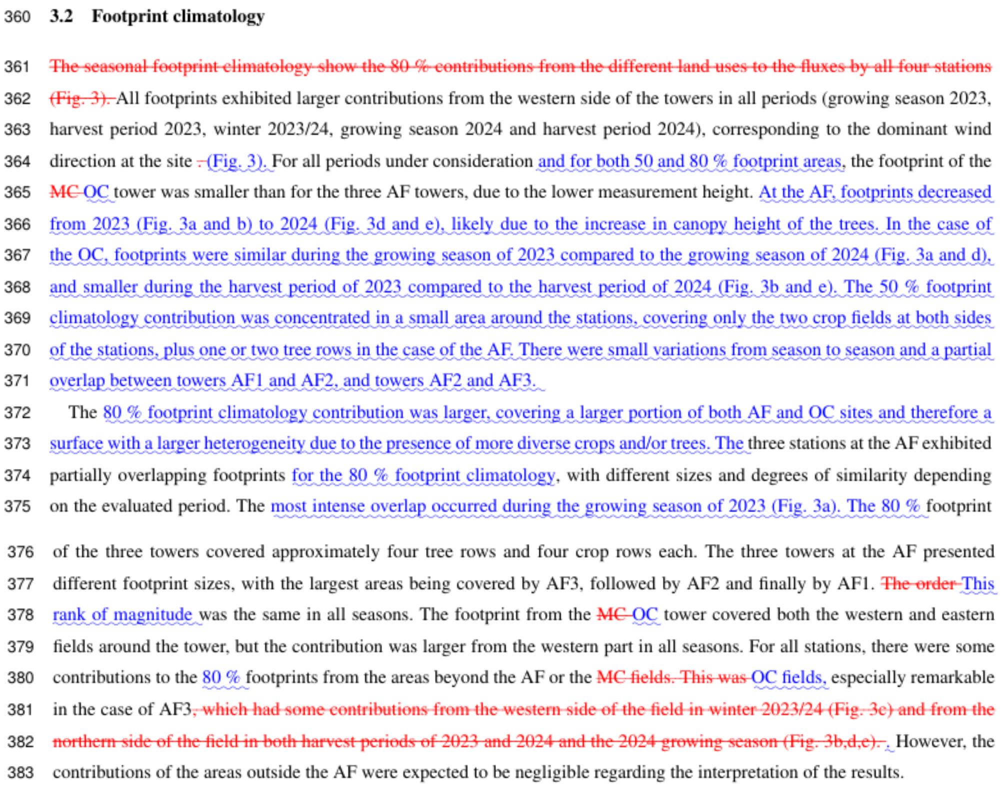
AR_C2. We chose the planar fit dividing in wind direction sectors because of the surface heterogeneity. The 8-sectors division is based on the default recommendation by ICOS, as noted in Sabbatini et al. (2018). The information was added to the text.

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 removed 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 was 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. AR16, AR17 and AR18 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. The corresponding text in the results section was changed accordingly (screenshots below).



3.2 Footprint climatology

~~The seasonal footprint climatology show the 80 % contributions from the different land uses to the fluxes by all four stations (Fig. 3).~~ All footprints exhibited larger contributions from the western side of the towers in all periods (growing season 2023, harvest period 2023, winter 2023/24, growing season 2024 and harvest period 2024), corresponding to the dominant wind direction at the site ~~.~~ (Fig. 3). For all periods under consideration and for both 50 and 80 % footprint areas, the footprint of the ~~MC~~ OC tower was smaller than for the three AF towers, due to the lower measurement height. At the AF, footprints decreased from 2023 (Fig. 3a and b) to 2024 (Fig. 3d and e), likely due to the increase in canopy height of the trees. In the case of the OC, footprints were similar during the growing season of 2023 compared to the growing season of 2024 (Fig. 3a and d), and smaller during the harvest period of 2023 compared to the harvest period of 2024 (Fig. 3b and e). The 50 % footprint climatology contribution was concentrated in a small area around the stations, covering only the two crop fields at both sides of the stations, plus one or two tree rows in the case of the AF. There were small variations from season to season and a partial overlap between towers AF1 and AF2, and towers AF2 and AF3.

The 80 % footprint climatology contribution was larger, covering a larger portion of both AF and OC sites and therefore a surface with a larger heterogeneity due to the presence of more diverse crops and/or trees. The three stations at the AF exhibited partially overlapping footprints for the 80 % footprint climatology, with different sizes and degrees of similarity depending on the evaluated period. The most intense overlap occurred during the growing season of 2023 (Fig. 3a). The 80 % footprint of the three towers covered approximately four tree rows and four crop rows each. The three towers at the AF presented different footprint sizes, with the largest areas being covered by AF3, followed by AF2 and finally by AF1. ~~The order~~ This rank of magnitude was the same in all seasons. The footprint from the ~~MC~~ OC tower covered both the western and eastern fields around the tower, but the contribution was larger from the western part in all seasons. For all stations, there were some contributions to the 80 % footprints from the areas beyond the AF or the ~~MC fields. This was~~ OC fields, especially remarkable in the case of AF3~~, which had some contributions from the western side of the field in winter 2023/24 (Fig. 3c) and from the northern side of the field in both harvest periods of 2023 and 2024 and the 2024 growing season (Fig. 3b,d,e).~~. However, the contributions of the areas outside the AF were expected to be negligible regarding the interpretation of the results.

Figure AR16: Changes in footprint climatology plot and results description, part 1.

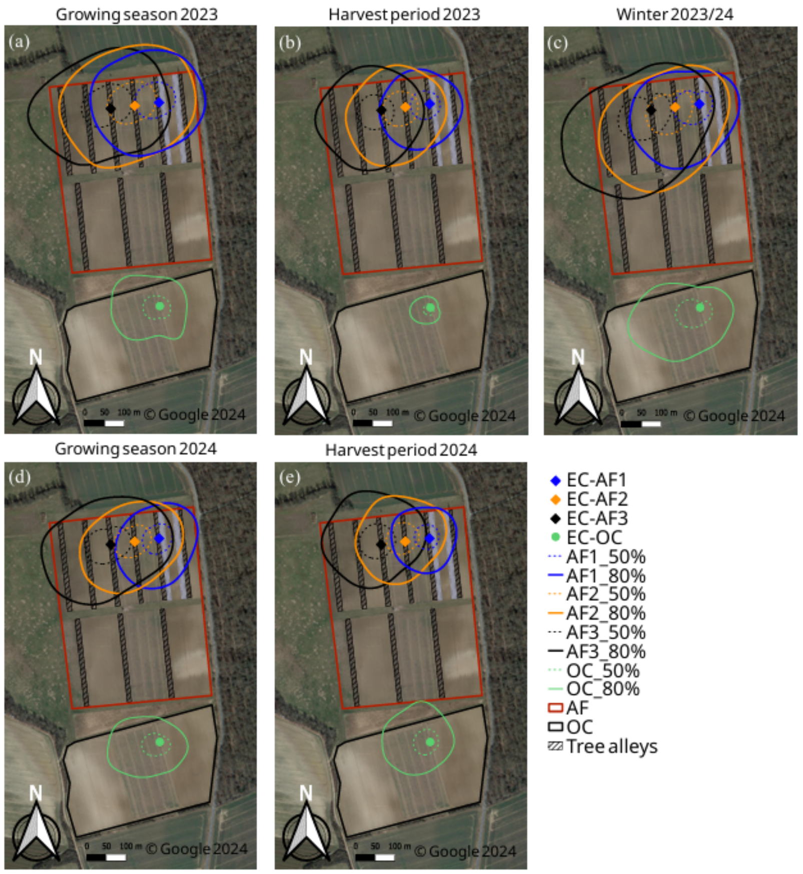
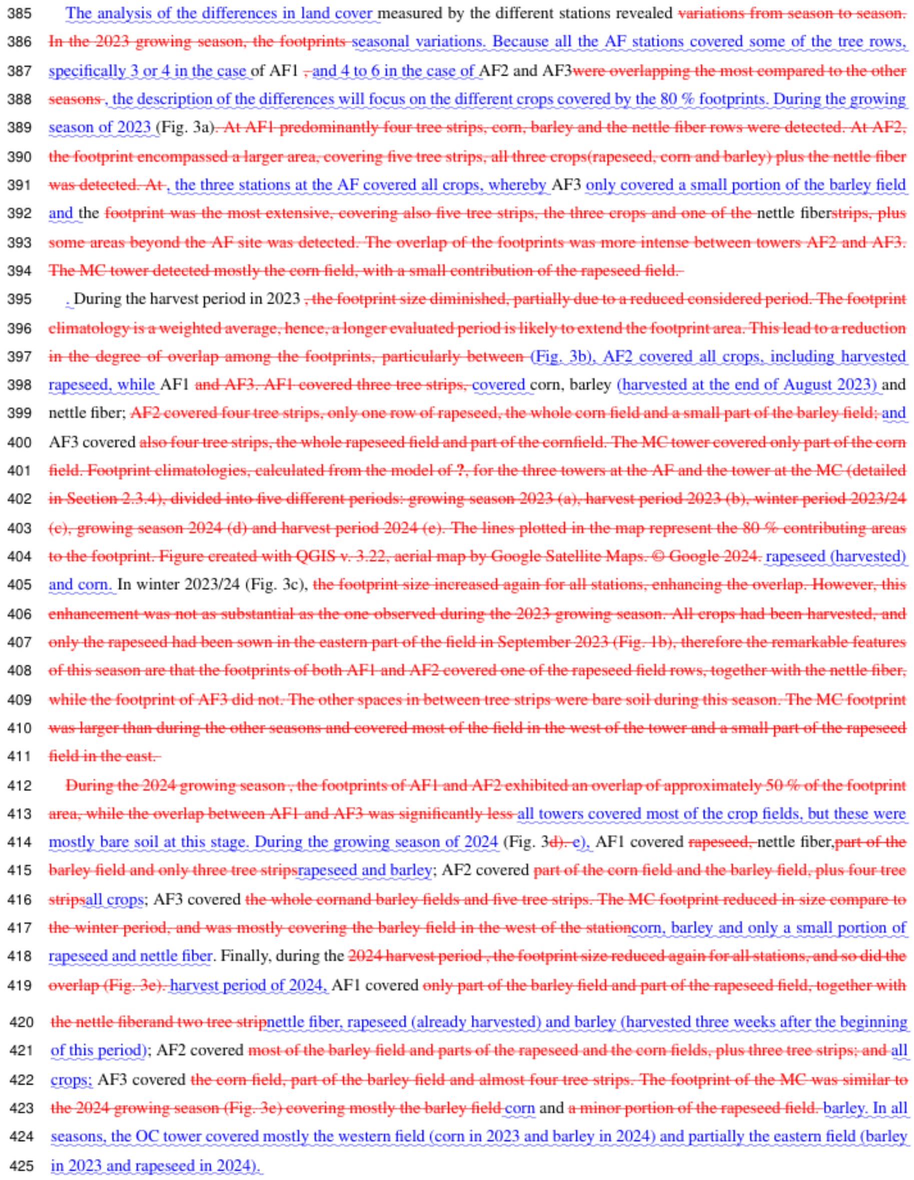


Figure 3. Footprint climatologies, calculated from the model of Kljun et al. (2015) as detailed in section 2.3.4, for the three towers at the AF (AF1, blue; AF2, orange; AF3, black) and the tower at the OC (green), divided into five different periods: growing season 2023 (a), harvest period 2023 (b), winter period 2023/24 (c), growing season 2024 (d) and harvest period 2024 (e). The lines plotted in the map represent the 80 % (solid line) and 50 % (dashed line) contributing areas to the footprint. The station locations are marked with diamonds for the AF stations and a circle for the OC station. Figure created with QGIS v. 3.22, aerial map by Google Satellite Maps. © Google 2024.

Figure AR17: Changes in footprint climatology plot and results description, part 2.



The analysis of the differences in land cover measured by the different stations revealed ~~variations from season to season. In the 2023 growing season, the footprints~~ seasonal variations. Because all the AF stations covered some of the tree rows, specifically 3 or 4 in the case of AF1 ~~,~~ and 4 to 6 in the case of AF2 and AF3~~were overlapping the most compared to the other seasons~~, the description of the differences will focus on the different crops covered by the 80 % footprints. During the growing season of 2023 (Fig. 3a)~~. At AF1 predominantly four tree strips, corn, barley and the nettle fiber rows were detected. At AF2, the footprint encompassed a larger area, covering five tree strips, all three crops(rapeseed, corn and barley) plus the nettle fiber was detected. At~~ , the three stations at the AF covered all crops, whereby AF3 only covered a small portion of the barley field and the ~~footprint was the most extensive, covering also five tree strips, the three crops and one of the~~ nettle fiber~~strips, plus some areas beyond the AF site was detected. The overlap of the footprints was more intense between towers AF2 and AF3. The MC tower detected mostly the corn field, with a small contribution of the rapeseed field.~~

. During the harvest period in 2023 ~~, the footprint size diminished, partially due to a reduced considered period. The footprint climatology is a weighted average, hence, a longer evaluated period is likely to extend the footprint area. This lead to a reduction in the degree of overlap among the footprints, particularly between~~ (Fig. 3b), AF2 covered all crops, including harvested rapeseed, while AF1 ~~and AF3. AF1 covered three tree strips,~~ covered corn, barley (harvested at the end of August 2023) and nettle fiber; ~~AF2 covered four tree strips, only one row of rapeseed, the whole corn field and a small part of the barley field;~~ and AF3 covered ~~also four tree strips, the whole rapeseed field and part of the cornfield. The MC tower covered only part of the corn field. Footprint climatologies, calculated from the model of ?, for the three towers at the AF and the tower at the MC (detailed in Section 2.3.4), divided into five different periods: growing season 2023 (a), harvest period 2023 (b), winter period 2023/24 (c), growing season 2024 (d) and harvest period 2024 (e). The lines plotted in the map represent the 80 % contributing areas to the footprint. Figure created with QGIS v. 3.22, aerial map by Google Satellite Maps. © Google 2024.~~ rapeseed (harvested) and corn. In winter 2023/24 (Fig. 3c), ~~the footprint size increased again for all stations, enhancing the overlap. However, this enhancement was not as substantial as the one observed during the 2023 growing season. All crops had been harvested, and only the rapeseed had been sown in the eastern part of the field in September 2023 (Fig. 1b), therefore the remarkable features of this season are that the footprints of both AF1 and AF2 covered one of the rapeseed field rows, together with the nettle fiber, while the footprint of AF3 did not. The other spaces in between tree strips were bare soil during this season. The MC footprint was larger than during the other seasons and covered most of the field in the west of the tower and a small part of the rapeseed field in the east.~~

~~During the 2024 growing season , the footprints of AF1 and AF2 exhibited an overlap of approximately 50 % of the footprint area, while the overlap between AF1 and AF3 was significantly less~~ all towers covered most of the crop fields, but these were mostly bare soil at this stage. During the growing season of 2024 (Fig. 3~~d).~~ e), AF1 covered ~~rapeseed,~~ nettle fiber,~~part of the barley field and only three tree strips~~rapeseed and barley; AF2 covered ~~part of the corn field and the barley field, plus four tree strips~~all crops; AF3 covered ~~the whole cornand barley fields and five tree strips. The MC footprint reduced in size compare to the winter period, and was mostly covering the barley field in the west of the station~~corn, barley and only a small portion of rapeseed and nettle fiber. Finally, during the ~~2024 harvest period , the footprint size reduced again for all stations, and so did the overlap (Fig. 3e).~~ harvest period of 2024, AF1 covered ~~only part of the barley field and part of the rapeseed field, together with the nettle fiberand two tree strip~~nettle fiber, rapeseed (already harvested) and barley (harvested three weeks after the beginning of this period); AF2 covered ~~most of the barley field and parts of the rapeseed and the corn fields, plus three tree strips; and~~ all crops; AF3 covered ~~the corn field, part of the barley field and almost four tree strips. The footprint of the MC was similar to the 2024 growing season (Fig. 3e) covering mostly the barley field~~ corn and ~~a minor portion of the rapeseed field.~~ barley. In all seasons, the OC tower covered mostly the western field (corn in 2023 and barley in 2024) and partially the eastern field (barley in 2023 and rapeseed in 2024).

Figure AR18: Changes in footprint climatology plot and results description, part 3.

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. In figures AR8 and AR10 the changes in Figure 5 and its description are shown.

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 (see screenshot AR16) now shows also the 50 % line. In the corresponding explanation of the results this was also 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 AR19, the error for the measurements was considered as the random error, while in the right Figure AR19, as the sum of the random error plus the RMSE mentioned in the previous paragraph. The new figure was added to the manuscript instead of the previous Figure 7, and the description of the results was changed accordingly (see screenshots below).

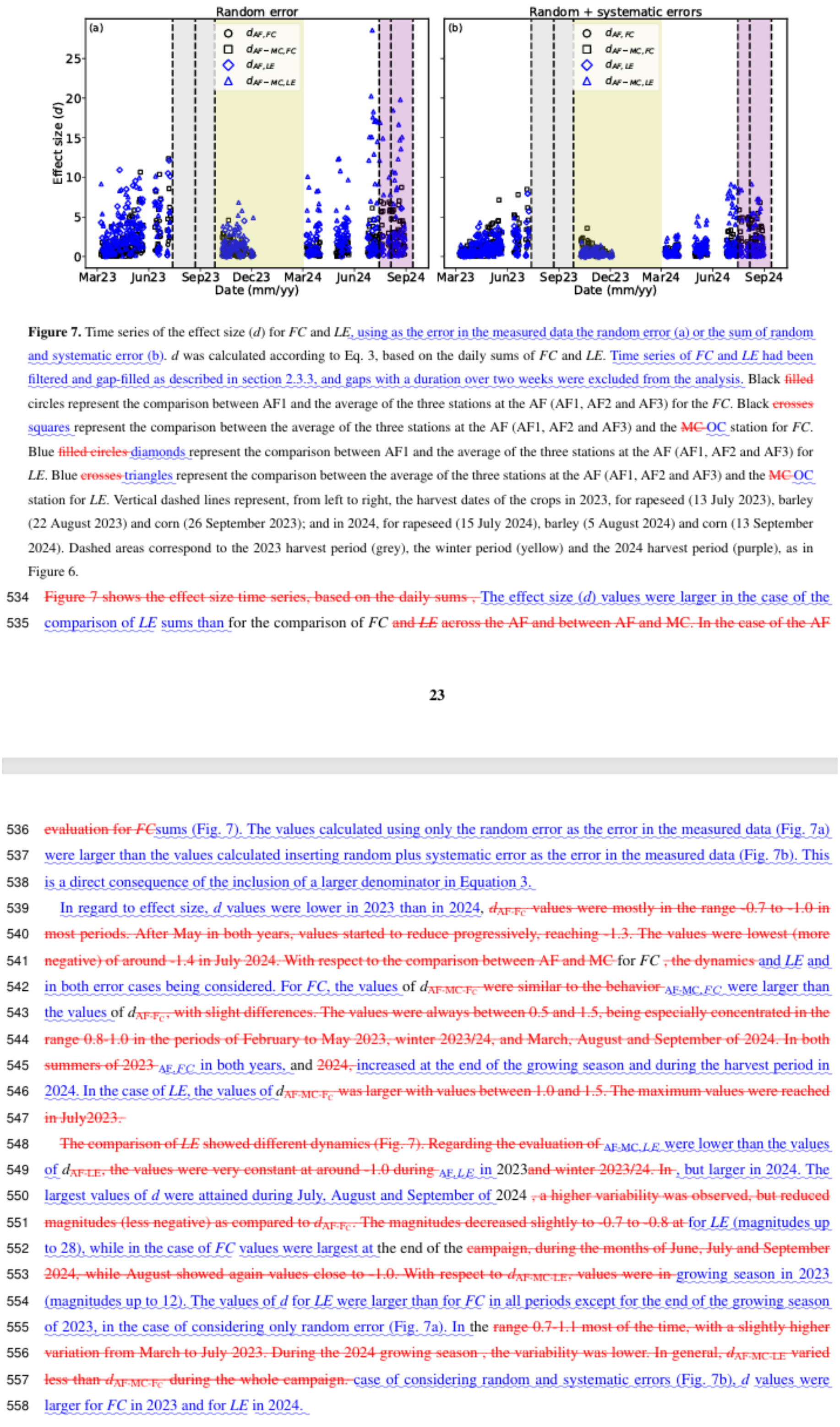


Figure 7. Time series of the effect size (d) for *FC* and *LE*, using as the error in the measured data the random error (a) or the sum of random and systematic error (b). d was calculated according to Eq. 3, based on the daily sums of *FC* and *LE*. Time series of *FC* and *LE* had been filtered and gap-filled as described in section 2.3.3, and gaps with a duration over two weeks were excluded from the analysis. Black ~~filled~~ circles represent the comparison between AF1 and the average of the three stations at the AF (AF1, AF2 and AF3) for the *FC*. Black ~~crosses~~ squares represent the comparison between the average of the three stations at the AF (AF1, AF2 and AF3) and the ~~MC~~ OC station for *FC*. Blue ~~filled circles~~ diamonds represent the comparison between AF1 and the average of the three stations at the AF (AF1, AF2 and AF3) for *LE*. Blue ~~crosses~~ triangles represent the comparison between the average of the three stations at the AF (AF1, AF2 and AF3) and the ~~MC~~ OC station for *LE*. Vertical dashed lines represent, from left to right, the harvest dates of the crops in 2023, for rapeseed (13 July 2023), barley (22 August 2023) and corn (26 September 2023); and in 2024, for rapeseed (15 July 2024), barley (5 August 2024) and corn (13 September 2024). Dashed areas correspond to the 2023 harvest period (grey), the winter period (yellow) and the 2024 harvest period (purple), as in Figure 6.

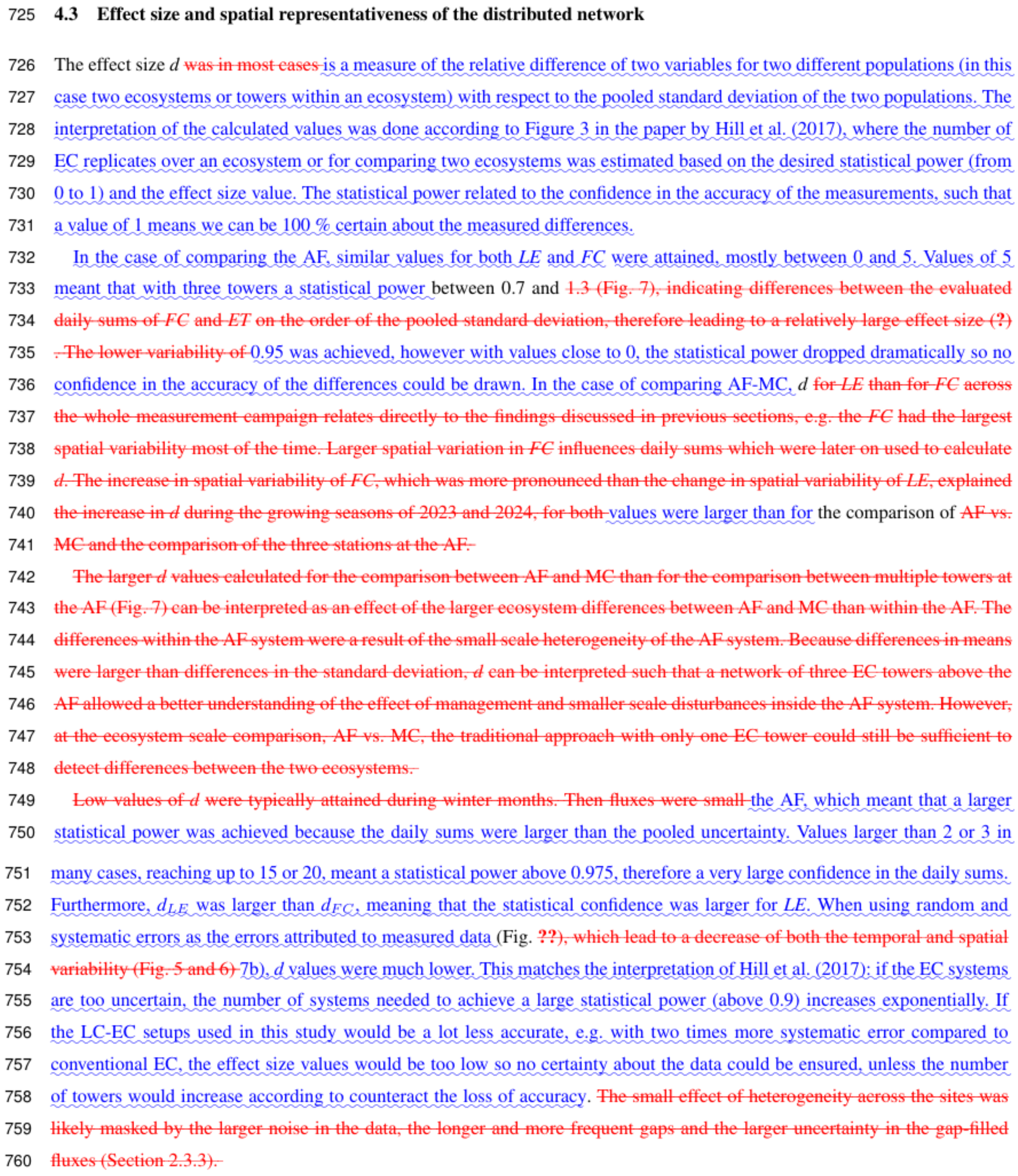
~~Figure 7 shows the effect size time series, based on the daily sums ,~~ The effect size (d) values were larger in the case of the comparison of *LE* sums than for the comparison of *FC* ~~and *LE* across the AF and between AF and MC. In the case of the AF evaluation for *FC*~~sums (Fig. 7). The values calculated using only the random error as the error in the measured data (Fig. 7a) were larger than the values calculated inserting random plus systematic error as the error in the measured data (Fig. 7b). This is a direct consequence of the inclusion of a larger denominator in Equation 3.

In regard to effect size, d values were lower in 2023 than in 2024, ~~$d_{\text{AF-Fc}}$ values were mostly in the range -0.7 to -1.0 in most periods. After May in both years, values started to reduce progressively, reaching -1.3. The values were lowest (more negative) of around -1.4 in July 2024. With respect to the comparison between AF and MC~~ for *FC* ~~, the dynamics~~ and *LE* and in both error cases being considered. For *FC*, the values of d~~$_{\text{AF-MC-Fc}}$ were similar to the behavior~~ $_{\text{AF-MC},FC}$ were larger than the values of d~~$_{\text{AF-Fc}}$, with slight differences. The values were always between 0.5 and 1.5, being especially concentrated in the range 0.8-1.0 in the periods of February to May 2023, winter 2023/24, and March, August and September of 2024. In both summers of 2023~~ $_{\text{AF},FC}$ in both years, and ~~2024,~~ increased at the end of the growing season and during the harvest period in 2024. In the case of *LE*, the values of d~~$_{\text{AF-MC-Fc}}$ was larger with values between 1.0 and 1.5. The maximum values were reached in July2023.~~

~~The comparison of *LE* showed different dynamics (Fig. 7). Regarding the evaluation of~~ $_{\text{AF-MC},LE}$ were lower than the values of d~~$_{\text{AF-LE}}$, the values were very constant at around -1.0 during~~ $_{\text{AF},LE}$ in 2023~~and winter 2023/24. In~~ , but larger in 2024. The largest values of d were attained during July, August and September of 2024 ~~, a higher variability was observed, but reduced magnitudes (less negative) as compared to $d_{\text{AF-Fc}}$. The magnitudes decreased slightly to -0.7 to -0.8 at~~ for *LE* (magnitudes up to 28), while in the case of *FC* values were largest at the end of the ~~campaign, during the months of June, July and September 2024, while August showed again values close to -1.0. With respect to $d_{\text{AF-MC-LE}}$, values were in~~ growing season in 2023 (magnitudes up to 12). The values of d for *LE* were larger than for *FC* in all periods except for the end of the growing season of 2023, in the case of considering only random error (Fig. 7a). In the ~~range 0.7-1.1 most of the time, with a slightly higher variation from March to July 2023. During the 2024 growing season , the variability was lower. In general, $d_{\text{AF-MC-LE}}$ varied less than $d_{\text{AF-MC-Fc}}$ during the whole campaign.~~ case of considering random and systematic errors (Fig. 7b), d values were larger for *FC* in 2023 and for *LE* in 2024.

Figure AR19: Changes in results of effect size.

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 modified the text so then it is more clear, including a better explanation of the meaning of the effect size and the statistical power according to the paper of Hill et al. (2017). Please see screenshots below for the respective changes.



4.3 Effect size and spatial representativeness of the distributed network

The effect size *d* ~~was in most cases~~ is a measure of the relative difference of two variables for two different populations (in this case two ecosystems or towers within an ecosystem) with respect to the pooled standard deviation of the two populations. The interpretation of the calculated values was done according to Figure 3 in the paper by Hill et al. (2017), where the number of EC replicates over an ecosystem or for comparing two ecosystems was estimated based on the desired statistical power (from 0 to 1) and the effect size value. The statistical power related to the confidence in the accuracy of the measurements, such that a value of 1 means we can be 100 % certain about the measured differences.

In the case of comparing the AF, similar values for both *LE* and *FC* were attained, mostly between 0 and 5. Values of 5 meant that with three towers a statistical power between 0.7 and ~~1.3 (Fig. 7), indicating differences between the evaluated daily sums of *FC* and *ET* on the order of the pooled standard deviation, therefore leading to a relatively large effect size (?) . The lower variability of~~ 0.95 was achieved, however with values close to 0, the statistical power dropped dramatically so no confidence in the accuracy of the differences could be drawn. In the case of comparing AF-MC, *d* ~~for *LE* than for *FC* across the whole measurement campaign relates directly to the findings discussed in previous sections, e.g. the *FC* had the largest spatial variability most of the time. Larger spatial variation in *FC* influences daily sums which were later on used to calculate *d*. The increase in spatial variability of *FC*, which was more pronounced than the change in spatial variability of *LE*, explained the increase in *d* during the growing seasons of 2023 and 2024, for both~~ values were larger than for the comparison of ~~AF vs. MC and the comparison of the three stations at the AF.~~

~~The larger *d* values calculated for the comparison between AF and MC than for the comparison between multiple towers at the AF (Fig. 7) can be interpreted as an effect of the larger ecosystem differences between AF and MC than within the AF. The differences within the AF system were a result of the small scale heterogeneity of the AF system. Because differences in means were larger than differences in the standard deviation, *d* can be interpreted such that a network of three EC towers above the AF allowed a better understanding of the effect of management and smaller scale disturbances inside the AF system. However, at the ecosystem scale comparison, AF vs. MC, the traditional approach with only one EC tower could still be sufficient to detect differences between the two ecosystems.~~

~~Low values of *d* were typically attained during winter months. Then fluxes were small~~ the AF, which meant that a larger statistical power was achieved because the daily sums were larger than the pooled uncertainty. Values larger than 2 or 3 in many cases, reaching up to 15 or 20, meant a statistical power above 0.975, therefore a very large confidence in the daily sums. Furthermore, d_{LE} was larger than d_{FC}, meaning that the statistical confidence was larger for *LE*. When using random and systematic errors as the errors attributed to measured data (Fig. ~~??), which lead to a decrease of both the temporal and spatial variability (Fig. 5 and 6)~~ 7b), *d* values were much lower. This matches the interpretation of Hill et al. (2017): if the EC systems are too uncertain, the number of systems needed to achieve a large statistical power (above 0.9) increases exponentially. If the LC-EC setups used in this study would be a lot less accurate, e.g. with two times more systematic error compared to conventional EC, the effect size values would be too low so no certainty about the data could be ensured, unless the number of towers would increase according to counteract the loss of accuracy. ~~The small effect of heterogeneity across the sites was likely masked by the larger noise in the data, the longer and more frequent gaps and the larger uncertainty in the gap-filled fluxes (Section 2.3.3).~~

Figure AR20: Changes in discussion of effect size.

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 added an extra paragraph to the discussion on this subject. Please refer to screenshots in figures AR1 and AR2 at the beginning of this document.

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.