**Authors' response to Anonymous referee #1**

On behalf of the authors, I would like to thank the Reviewer 1 for their time and effort to improve our manuscript. Please find my response below with the Reviewer's original comments marked in *grey italic*. The line numbers refer to the clean version (without the track changes).

While the MS Word automatic track changes document marked all figures as new, we updated only 2 figures:
- Figure 9 was remade for the altered indices (see the explanation in the corresponding section)
- Figure 8 has now corrected the y-axis labels, which were previously mislabelled as "modelled"

*Following the first round of review, the authors have made a tremendous effort to revise the manuscript and to address my comments as well as those of the second reviewer, which I highly appreciate. In my view, the manuscript has improved substantially and can now be considered for publication after the authors address a few remaining minor comments, many of which are only suggestions aimed at improving readability and accessibility.*

**Response:** We sincerely thank you for your thoughtful comments, which greatly helped improve the manuscript. The study has already improved considerably compared to the original version.

*I found one issue regarding the new analysis around Figure 9, where it appears that the authors may, perhaps inadvertently, have changed the definition of certain metrics and applied a different resampling technique than in an earlier part of the manuscript, resulting in diverging results for the same dataset. This is somewhat confusing and should be clarified.*

**Response:** Thank you for identifying this inconsistency arising from the new analysis. Your comment helped us carefully reconsider the methodology and what the analysis results actually tell us about our forest. We provide a detailed response and clarification under the specific comment further below.

*Despite this, given the overall high quality of the revised manuscript, I recommend to the editor that it be accepted after the authors address the comments at their own discretion, without another round of review. I have high confidence in the authors' ability to do so and look forward to seeing the final version of the manuscript, which I consider a valuable contribution to the field. Their study skillfully explores the carbon dynamics of a relatively underexplored ecosystem, offering insights that will benefit future assessments.*

*Below are some specific comments, listed roughly in order of appearance. All line numbers are referring to the manuscript version WITHOUT track changes.*

*Regarding the whole ms: for carbon fluxes, you use both µmol C m-2 time-1 and g C m-2 time-1 throughout the manuscript. Unless you have a good reason to do so, why not make it consistent?*

**Response:** We intentionally used both µmol C $m^{-2}$ $s^{-1}$ and g C $m^{-2}$ (period$^{-1}$) to match the conventions of flux data presentation at different temporal scales. Parameters, calculated from half-hourly carbon fluxes (GPPsat - canopy photosynthetic capacity; and ERref – respiration at reference temperature) usually retain the standard flux unit of µmol C $m^{-2}$ $s^{-1}$, while daily and cumulative (seasonal or annual) sums are expressed in g C $m^{-2}$ (period$^{-1}$).

*l.63: minor and optional comment: you could change the wording here from "did not address" to "was beyond the scope" as currently it sounds a bit negative towards your own work, which I find unnecessary.*

**Response:** We appreciate the suggestion. Although we do not consider "did not address" to be negative, we agree that "was beyond the scope" provides a smoother phrasing. The text has been revised accordingly (line 63-64)

*l.66: "We utilise three years of EC flux measurements, representing a "wet" year (2017), a "drought" year (2018), and a "recovery" year (2019)."; My suggestion is that you add something in the tone of: "…three years of EC flux measurements with contrasting environmental conditions: 2017 being an anomalous wet (add numbers here), 2018 an anomalous dry (add numbers here) and 2019 being a recovery year (add numbers here)." Regarding what I mean here by numbers see my comment regarding l.260*

**Response:** We agree that highlighting the contrasting environmental conditions adds clarity, however as this sentence is part of the introduction, we prefer to keep it descriptive. See modified sentence in lines 66-67

*l.104 & l.107: The tower was 21 m high, but PAR was measured at 25 m — might this be a typo? Also, in Appendix A you state that measured incoming shortwave radiation was used, whereas in the instrumentation paragraph only a quantum sensor is mentioned. Perhaps that information is just missing?*

**Response:** Thank you for noticing this inconsistency! This text remained from an earlier version of the manuscript and was not updated. We have now corrected it (lines 105-107).

*l.132: I agree with that choice. Later on, I added a comment suggesting that you likely overestimate Rn, which further complicates any potential attempts at closure.*

**Response:** While the approach is simplified, it allowed checking if the difference in evapotranspiration between the years would be biased by the EBC variability.

*l.241: "filtered to maximise the share of T" This probably means that you filtered based on radiation > X or GPP > 0 values? Maybe you can say what you did here.*

**Response:** The filtering criteria have now been specified in the Methods text (lines 250-252)

*l.260: This comment is regarding the whole paragraph and also relates to my comment in l.66: I think you could briefly present an argument for the wet/dry/recovery classification. Currently, you describe that the drought and recovery are different because of longer dry periods in the drought year but that none of the years is significantly drier in comparison to the long-term average precipitation. I agree with your classification looking at the values in Table 1 and Fig.2e but I think some kind of quantitative measure would strengthen the manuscript and make this more intuitively. A simple way to do so would be to use ETCCDI indices, which might better show differences between the years. Also, you could mention the above average temperature during the drought period. Taken together, while I don't think it is strictly necessary to do so, I think that currently it's not super easy to figure out by briefly looking at Fig2e together with you saying that the precipitation input is not different and that SWC is mainly governed by precipitation how this classification is justified. In other words, the classification made only sense for me once I understood all the results and it seems not very intuitive yet.*

**Response:** We thank the reviewer for this suggestion. We agree that the difference in precipitation is not immediately apparent from the figure. While the total precipitation in 2018 remained within one standard deviation of the multiyear average, it was still lower than in the other years. Its temporal distribution, extended consecutive dry days combined with above-average temperatures ("heatwave"), likely drove the drought conditions. We edited this part in the manuscript (lines 274-288). We hope that this explanation clarifies the wet/dry/recovery classification without adding additional metrics to our already lengthy manuscript.

*l.284: Maybe add "cumulative" to the bracket such as: (cumulative annual NEE < 0) and add to the sentence whether the number is the average of all years or the number for 2017 as the brackets have the numbers for the other two years.*

**Response**: We agree. The revised sentence in l.290

*Regarding the values shown Table 1: For SWC you note that measurements started only in July, but Fig. 3 shows that all fluxes also start in May. I think for consistency you could do the same for the carbon fluxes in the table simply referring to what you explain starting in l.183*

**Response**: Thank you for this comment. We improved the clarity of the Table 1 caption text and marked the annual 2017 values in italic.

*l.300f. minor comment just for consistency: You present very detailed p-values as smaller than, greater than or equals here but in l.306 you say (not significant) which would, for example, be also the case for l.303 (…while GPP declined only marginally). This statement in particular, you could also say something like "GPP did not significantly decline" to match the wording in l.305 where you say that there was "no differences between the daily*

*values" based on your statistical test. I recommend you check that throughout the manuscript to have more consistency.*

**Response**: Thank you for this comment. We edited the paragraph (lines 307-312) and checked for the consistency

*Fig. 8: The caption could be slightly improved by more clearly stating that solid lines represent measured data, whereas dashed lines show modelled data. This is also is true for the text, for example in line.450: where you say "GPP in 2018 was reduced from July onwards" add precisely what you mean for easier understanding. Here is another example of what I mean: In the drought year 2018, measured GPP (Fig.3c, Fig.8a) was reduced from July onwards […] When applying the model parameters from 2018 […], this GPP suppression persisted in the modelled GPP for both the other years (Fig.8b). I think if you guide the reader a bit more along your results together with the figures it's much easier to follow. In the next sentence ("The difference between the observed GPP..."), I am also not sure if you are referring to all results from this analysis or just to the what we see in Fig 8a and c. This goes on for a bit in this paragraph, the results are very interesting but with small changes it could be much easier to read and understand.*

**Response:** We agree that both the Figure 8 caption and the corresponding text in the Results section needed clarification. We have revised the figure caption and rewritten the related paragraph to explicitly state that solid lines represent observed fluxes and dashed lines represent cross-year modelled fluxes. In addition, we corrected the y-axis labels in Figure 8, which were previously mislabelled as "modelled." We would like to note that, as daily aggregated GPP and ER are, strictly speaking, not directly measured, we use the terms "observed" and "cross-year modelled" to distinguish between fluxes derived from gap-filled EC data and those modelled using parameter sets from other years.

*Fig.9 and corresponding texts parts (from l. 239 and l.460): I tried a while to understand what the benefit to the overall story of the manuscript of this new analysis is but to be honest couldn't quite figure it out. Some points which I thought some time about*

*- I looked at the Lloret et al. (2011) paper to understand it and found that what you define as Recovery is equivalent to their Resilience. I don't know if this happened by accident or if you had a certain motivation to do so, if that is the case, please address in the methods section why you made new definitions here. Your resilience term I find quite hard to interpret with the text you provide and I don't know exactly what I should take from it.*

**Response:** We very much appreciate this comment! The previous modification of the metrics was a result of our overthinking, which did not improve clarity as intended. After careful consideration, we have reverted to the more conventional definitions: Rt = dry/ref; Rc=rec/dry; Rs = rec/ref. (lines 241-248). Corresponding updates have been made to Figure 9 and throughout the text. Importantly, we note that this revision does not change the results or the conclusions of our study.

*- Also, I don't understand the reasoning behind your analysis that you calculating these terms based on daily values. As for example the resistance term you define is just comparing the drought to the reference term that value is equivalent to what you show in and after Table 1, where you, for example, find that the GPP term is insignificantly different between the years based on annual (or growing season) sums. In the new analysis based on the daily values you now say that the difference is around 14% due to the different calculations. I find that contradicting and you should make clear why you do so and differentiate with which analysis and why you want to address shorter-term carbon dynamics or carbon sink functions of your ecosystem.*
*- I also want to say that I do not understand the logic behind using the daily values here but I might miss something.*

**Response:** We would like to thank the reviewer for bringing this apparent contradiction to our attention! Indeed, the seasonally accumulated values of GPP for 2017 and 2018 were similar, whereas the resistance index calculated from daily GPP was 0.85, meaning a 15% decline. The discrepancy comes from the difference in temporal scales. Seasonal sums smooth out short-term variability: in 2018, higher GPP during the early growing season partly compensated for suppression in mid- to late summer (Figure 3), resulting in comparable seasonal totals. In contrast, the resistance index, calculated from the bootstrapped averages of daily values, captures these episodic declines more accurately, reflecting the stronger suppression of photosynthesis during the peak drought period.

*Then, the bootstrapping approach is likely shifting around the day of year between the years randomly to explore what you call uncertainty here. What kind of uncertainty would that be? Accounting for different phenological development between different years.*

**Response:** Since the variability of the daily ecosystem parameters can be rather high, the averages used in indices calculations could be affected by single outstanding days. To avoid this, we chose to use the bootstrapping approach.

*I believe that if you open the pandoras box of uncertainty here you would rather start a step earlier and explore the uncertainties related to the eddy covariance method itself, more importantly the flux partitioning and finally your ML-based gap-filling approach. Note that I do not think you should do so for the manuscript but the bootstrapping approach based on daily values gives the reader a false impression that the uncertainties were understood while the real source of uncertainty lies elsewhere. From experience, I would go as far as speculate that a full analysis of the uncertainties related to flux partitioning might even result in the not being able (in a statistical sense) to differentiate between the individual years.*

**Response:** We agree that the term "uncertainty" may have been misleading in this context. While the reported confidence intervals reflect the variability of the indices derived from bootstrapping, they, of course, do not represent the full uncertainty of the flux measurements themselves. We have rephrased the corresponding section in the Methods (lines 252-254) to avoid potential misinterpretation.

*- The interpretation of the outcome is relatively brief and, in my opinion, does not go beyond what you already discussed effectively in your previous analysis. Consider better explaining what added benefit the analysis has and what arguments which have not been presented before can be drawn from it.*

*- Then, from l.594 you discuss, again this being an example for the general issue, the reduction in GPP which you previously have shown to be not significant based on the values in Table 1 (e.g., l.302). I would make sure to remove this contradiction and if there is some added value of these two different calculations discuss their importance and what added benefit they bring to your analysis.*

**Response:** We hope that the added explanation above cleared the discrepancy

*l.543 and Figure A1: Very creative approach to estimate the energy balance closure in the absence of data from a net radiometer and ground heat flux plates. Neglecting the longwave term will overestimate the true net radiation, as net longwave radiation is usually negative on an annual scale. This means that if you had included that term, your energy balance closure would actually appear better than shown in Fig. A1. From the turbulent flux perspective, this might therefore less problematic as interpreted. Regarding the approximation of G=0.05*Rn , this relationship is probably more complex. I quickly checked that assumption for one year at an evergreen needleleaf forest site, where I just had the data at hand and found that 0.05*Rn slightly overestimates the actual G with a linear regression suggestion 0.03*Rn - 5 (with units being Wm-2, relevant for the intercept). Obviously, that's a very different site with different soil and moisture regime but one could argue that this is close enough in the broader picture. My recommendation would be that you briefly mention what you expect from neglecting the longwave term in Rn and leave the simple G model as it is or briefly say what errors you expect from it. Overall, in my opinion, you miss a chance here to say that your closure is actually better than you can show based on your limited instrumentation based on the neglection of the longwave term.*

**Response:** Thank you for the positive and constructive feedback! We added the mention of the missing longwave radiation and the limitation of the simplified G calculation approach to the corresponding appendix text

*l.544, l.551: Avoid statements as "sufficient GPP" and "adequate water supply". There are probably more of these kind of wordings throughout the manuscript. I recommend to double check carefully.*

**Response**: We carefully reviewed the manuscript and revised instances of vague wording that we could identify.

*l.563: in brackets: I suggest removing the brackets and adding "representing" [the maximum stomatal aperture]*

**Response:** We agree. Rephrased, see line

*l.648: I'm not sure here, but could it also be that a large amount of volatile carbon was available after ER was so low in 2018? You do indirectly suggest that in the sentence starting from line 653 anyways but not implicitly for your study.*

**Response:** We agree that this is a plausible explanation. However, without direct measurements of heterotrophic respiration, we cannot confirm it, so it remains speculative. We added a sentence in the Discussion to clarify this possibility (lines 662-664)

*l.687: "High EWUE and reduced Gc…" That sentence seems to lack something. Could go: In consequence, high EWUE and reduced Gc….*

**Response:** High EWUE and reduced Gc were the indicators of stomatal regulation, rather than the consequences of declined GPP and ET

*l.696: Completely optional and just my own opinion: I obviously haven't been to the forest but from my interpretation from all of your results I wouldn't think that this 2018 drought would cause a delayed tree mortality. While this is a more general statement and you say so, I find this final sentence a bit underwhelming given the results you produced. You already mention the call for continued long-term monitoring at the end of the discussion and your data just consists of 3 years before the station was moved. Maybe you could think of a better fitting last sentence suitable for your specific manuscript.*

**Response:** We rephrased the final sentence: see lines 705-708