Major concerns:
(1) The authors make several claims that are not supported by data. a. The authors make conclusions regarding collar artefacts, particularly the alteration of root growth by the 8100A collars. This is even mentioned as one of the "main reasons for the "observed differences in the performance of the two systems" (iv; line 17) in the abstract. However they present no data to support this claim. A single photo in supplementary material does not constitute data.

We partly disagree with the reviewer on this matter. We have indeed no quantitative data to support our claim. We did not expect this phenomenon to occur, and at the end of the study, it was not possible anymore to invest additional time into the collection, preparation and quantification of the roots. We have therefore rephrased the respective sentence in the abstract. However, we are still convinced that the picture in the supplementary material is very meaningful. It is an example picture of a phenomenon which we observed at all eight LI-8100 collars. As pointed out in the manuscript, root respiration is a very important component contributing to soil respiration, and its alteration by chamber measurements is an on-going discussion. Thus, we consider it would not be okay to abstain from mentioning this phenomenon in the paper only because we were not able to provide quantitative data. Such qualitative field observations are important for the flux community for the improvement of chamber designs and measurement protocols.

b. The authors say that "the impact of the automated chamber systems on the environmental conditions increased with the size of the chamber itself and additionally with the size of the frame…" (line 14, page 14713). Presumably the authors are referring to the soil temperature and moisture data presented in Figure 2, as the bulk density, C content, and DOC content were not different (line 6, page 14707). However the authors present no in-situ (i.e., nonchamber affected) measurements of soil temperature or moisture, as such it is not possible to conclude whether the AGPS or the Licor systems altered these values relative to in-situ conditions. The authors can only compare the temperature and soil moisture from the AGPS and Licor systems (e.g, Fig. 2c-d). Remarkably, the authors do not quantitatively or statistically compare these data in any way.

The interpretation of the reviewer is not entirely correct. The main purpose of this paragraph is to point out the direct large effect that permanently installed chamber structures might have on vegetation growth. Changes in the vegetation structure/growth can lead to different temperature and moisture regimes at the soil surface. In our specific case, we clearly observed that the soil surface in the LI8100 plots was shaded earlier by the regrowing poplar canopy than the soil surface in the AGPS collars. In the revised manuscript we replaced 'environmental conditions' with 'vegetation structure' and we have rephrased the paragraph to clarify the message. We can supply pictures in the supplemental material which document the growth of the vegetation around the chambers.

A detailed quantitative comparison of the soil moisture data was not feasible because the majority of the observed differences were within the accuracy range of the soil moisture sensor of the AGPS (SPADE sensor, ±4 %). For soil temperature, we have added a quantitative comparison to the Results section to meet the reviewer's concern. However, we refrain from testing the statistical significance of the soil temperature differences between the AGPS and the LI-8100A. Regardless of the result of a statistical test, this would not help to explain the biological significance of these differences.

(2) The authors make no statistical comparison of the flux estimates provided by the two automated systems apart from the integrated temporal sums reported on line 14, and this

statistical test appears to be based on the 95% confidence intervals of integrated predictions from a Loyd and Tylor equation. I find the lack of other statistical comparisons of the two methods to be a striking omission for a manuscript purporting to compare the two methodologies. I suggest the authors consider a robust statistical comparison, possibly such as repeated-measures ANOVA on daily mean flux estimates, with fixed effects of method, date, and method x date and a random chamber (collar) term. Other methods such as time-series analyses, spectral analyses, or generalized additive models may also be appropriate (see comment below for generalised additive model information). Such methods could identify particular dates or periods when the flux estimates diverged, which could usefully focus the manuscript around methodological issues specific to those periods.

We have tried to visualize differences in the mean daily flux estimates between the two chamber systems throughout the monitoring period with generalized additive models. However, the 95% confidence interval for the AGPS was so wide that it overlapped with the LI-8100A for the majority of the days. Therefore, unfortunately, this analysis did not provide new information. However, we appreciate the suggestion to include the information about the time series analysis in the revised text.

Neither autochamber method is quantitatively compared to the CO_2 concentration gradient method. I suggest the authors consider removing this method from the manuscript.

The two methods cannot be directly quantitatively compared because they measure on different time scales. The gradient method provides, however, valuable information about changes in soil CO_2 concentration dynamics and their potential for altering soil CO_2 flux dynamics. If we remove this method from the manuscript, we have to start speculating why we see certain differences in the CO_2 flux rate between the two chamber systems. However, the soil CO_2 concentration measurements clearly illustrated the impact of the soil moisture conditions on the soil CO_2 flux dynamics at the site and thus significantly contributed to understanding the flux chamber measurements.

Other concerns:
(3) It is difficult to compare the methods in the time series plots (Figs. 4-5) given the issue of overplotted points. I suggest the authors explore heat maps, density clouds, or even simple running averages to visualize the central tendency of these datasets. Better yet, generalized additive models (GAMs) with random chamber effects could be used to display the estimated mean and 95% confidence intervals of these datastreams over time, and any statistical difference between the methods could be inferred via the 95% confidence intervals. See the "mgcv" R package and associated articles (Wood, 2011).

We are grateful for the reviewers' suggestion and added the results of generalized additive models showing the daily trend in the data for the two chamber systems. We can of course provide daily means separately for nighttime and daytime conditions. We think, however, that it would be difficult to find a suitable (biologically meaningful) averaging window for the atmospheric CO_2 concentration conditions (constant/fluctuating) since the duration of these conditions for a period of time was highly variable.

(4) The authors should more fully illustrate how well the Lloyd and Taylor models describe the observations, particularly if the summed predictions of these models will be used for inference of methodological differences, as is currently done. Model predictions plotted on top of the data vs. temperature, or observed vs. predicted plots, would be useful to assess potential bias. The

authors do present the residual standard errors and the parameter standard errors in Table 3, but these numbers are of limited utility to assess bias.

We have provide the information in the revised manuscript as requested by the reviewer. We are convinced that this clarifies the use of the model predictions.

(5) The units in Table 3 for "Average cSR" appear to be incorrect. Efflux rates of 897 μmol CO2 m-2 s-1 (for example, AGPS, Wide, not filtered) are a bit high. The units for these cumulative sums are likely incorrect. I also hope and expect the authors intended to refer to the 95% confidence interval, rather than the 5% confidence interval.

We thank reviewer #3 for pointing out these typos. They have been corrected.

Reference Wood S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society Series B-Statistical Methodology, 73, 3-36.