Replies to Referee #1

1. ***My only one remark concerns conclusion that researchers should use an established software only to get good results (p. 2108, lines 17-21 and p.2120, lines 1-3). I do not understand either why authors warn against software intercomparisons and usage in-house scripts. It was not discussed in the paper. An in-house script can be much more convenient and effective if it is oriented on specific measurement system. The established software packages could be a good tool to verify it. The paper rather show that even these popular packages must be very carefully tuned (including code modifications!) to get comparable results. So, I think that the main conclusion is that as long as there is no only one widely accepted software (if ever will be) the very detailed information on 'processing scheme' and 'processing steps' (in the sense introduced in the paper) is essential. Moreover, the comparison of the results given by a software used by any group with the results given by the 'reference' software could be very helpful in flux comparisons.***

We agree with the general sense of the Referee's comment. In particular, we agree that documenting the processing scheme can help ensure comparability of flux results from different sources and we revised the last paragraph of the conclusions accordingly. We also agree that established software packages can be used as a tool to validate 'in-house scripts'. Our point, though, concerns the *quality* of the validation – in our own experience too often overlooked. This is why in our paper we want to show how complex and nuanced the cross-validation is.

We do not conclude that established software is the *only* means to get good results. What we maintain is that using established software is a more secure way to assure comparability of fluxes collected from different sources, such as in regional or global databases. Also, we do not warn against software inter-comparisons in general, but we do warn against 'quick and dirty' inter-comparisons, especially when carried out by individuals who are not experts of the either software package under consideration.

Furthermore, while it is true that an in-house script can be much more convenient if it is oriented on a specific system (and we now stress this point in the revised manuscript), it is also true that maintaining in-house scripts to catch up with the latest developments of the method (or with modifications to the EC system in use) is difficult or sometimes impossible (e.g. if the main developer of the scripts left the group), without mentioning the waste of researchers time that this ubiquitously replicated effort implies. For this reasons, we maintain that relying on software developed by dedicated groups is in general a better strategy, on the long term. For further comments on this point, please see also reply to Comment 6 of Referee 2.

2. ***I also suggest to reconsider sentences in lines 4-6, p. 2112. Even if EddyPro is free, the phrases: "rapidly increasing world-wide", "comprehensive", "user-friendly" sounds a bit commercial for me.***

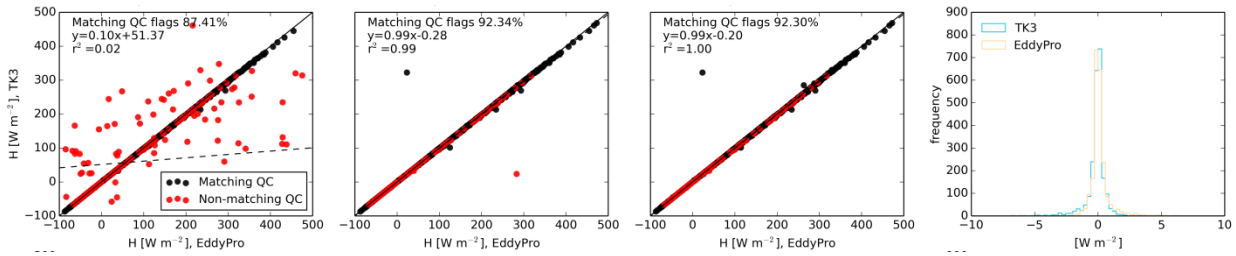
We accepted the Referee's comment and eliminated that sentence completely. We replaced it with the following: "EddyPro is free of charge, open-source software released under the GPL license".

REPLIES TO REFEREE #2

1. ***My main, and only major, issue is that some of the conclusions drawn (not necessarily in that section but across the whole text) are not supported by the results, and occasionally even in conflict with them or with each other (see detailed comments). This may in part result from a somewhat unclear motivation and scope. An optional improvement related to this, from which the value of the paper as a citation reference could greatly benefit, might be to stress that users of either of the two software packages (ensuring a certain version number, processing scheme setting and dataset properties similar to one of the two tested datasets) may expect a closer agreement than between arbitrarily chosen software packages. Ideally, this would be supported by a probability density function or confidence interval of the flux differences after the last "round"***

We believe we had a clear motivation for our paper and hopefully we satisfied the doubts of the Referee in the following replies to detailed comments.

We appreciate the suggestion "*to stress that users of either of the two software packages […] may expect a closer agreement than between arbitrarily chosen software package"* and added a sentence on this line in the Abstract and in the Conclusions. We did not include confidence intervals in the plots because, especially at Round 3, r^2 values were generally very high and thus the confidence intervals turned out to be very narrowed, virtually undistinguishable. In the revised manuscript, we instead tentatively added plots with the distribution of regression residuals, as in the example below.



2. ***P2110L15-19: The difference between C1 and C4 requires some more explanation, or maybe it is better to bin them together. C4 also does not seem to be mentioned any more throughout the rest of the manuscript.***

C1 and C4 are actually totally distinct, and cannot be binned together. C1 represents a *user* error: among the various options available, the user selects one that is unsuitable for the data being processed. An example would be applying the angle-of-attack correction of Nakai and Shimoyama (2012) to a CSI-CSAT3 sonic anemometer, or applying the 'self-heating' correction of Burba et al. (2012) to a closed-path gas analyzer such as the LI-7000. Usually, software packages do not provide "fool-proof" graphical interfaces that avoid such mistakes. Both EddyPro and TK3, among the most advanced in this sense, only provide a partial safety against them. C4 (implementation errors) are

instead plain bugs, or the result of the misinterpretation of an algorithm, or any other type of *developer* error, which affects all users.

C4 is not discussed in the paper because our comparison did not highlight any such type of mistake. Nonetheless, it appears to us that, for completeness, causes C4 should be mentioned in the list of potential reasons for disagreement between different software packages. In the revised manuscript, we clarified better the definition of C1 and C4.

3. P2111L05: Since only C2 and C3 are given as alternatives, did you mean by C1 "C1 and C4" ?.

The initial part of that paragraph ("*Assuming no bugs in the software […]*") eliminates C4 from following consideration. We now make it more explicit: "*Assuming no bugs (causes C4) in the software […]*"

4. P2112L05: I could not verify the claim that EddyPro is open source. Please note that the expectations of the average article reader (e.g. that open source software enables him to inspect and modify any part of the source code) should determine the classification of the software type; if you rely on a different definition of "open source" it would be more instructive to clearly mention if the source code is open or not, rather than use this term. In my previous (and maybe wrong) perception, ECO2S was and is open source, while EddyPro and TK3 are both not. If this is correct, state it clearly for all three software packages, or for none of them at all.

EddyPro is open source in the most classic sense: it is released under the GPL (version 3) license, which – among the rest – implies the full source code be available to anyone who requests it. The way LI-COR Biosciences distributes the software installer and the source code fully comply the GPL license. The release license can be found in the "Help >> About…" menu.

TK3 is technically not open-source, but parts of the code can be made available upon request as specified in the replies to the Referees' comments during the review for AMTD. We now made this clear in the revised manuscript.

5. P2112L04-24: Please check this passage carefully for irrelevant advertising. For example, giving reasons why the use of EddyPro is rapidly increasing is speculation not withholding the criteria of scienitific publication, unless there is a section about how user questionnaires on their reasons to use EddyPro have been evaluated.

Following a similar comment by Referee #1, we eliminated that sentence (see reply to the second comment of Referee #1).

6. P2108L15, P2113L17, P2118L19, P2118L25, P2119L23: The fact that the authors -and not the user - have full control of the source code, is mentioned at least five times across the manuscript

- I wonder why? It is in conflict with the claim at P2112L05 (see comment above) of EddyPro being "open source". Also, it suggests that additionally having self-written software is of vital importance to anyone who aims at research grade eddy-covariance processing and contributing to its further improvement. But this conclusion is not mentioned anywhere; instead in two places (see last comment) the wording even seems to suggest an inferiority of in-house software.

Perhaps there is a misunderstanding in the meaning of 'having full control'. What we mean with that is a complete knowledge of the code and the ability to meaningfully and safely modify it. In this sense, being able to access the source code of such complex software as EddyPro or TK3, does not automatically assure having 'full control' of it.

We do not say that we as the authors – *and not the users* – have full control of the source code. Any user can, in theory, get to the same level of knowledge that we have with our code. In practice, however, this is seldom or never the case. Note that this consideration applies also to in-house software: only the developer is assured to have full control of the source code. Most likely, any other user will learn how to set up the software, but won't have 'full control' of the code.

Regarding the 'inferiority' of in-house software: again, we need to resort to heuristic, experience-based argumentations. In principle, in-house software can of course be perfectly valid alternatives to 'official' EC software, if not better (e.g. when tailored on the EC system of interest). In principle, the user of such in-house software knows that code perfectly and knows how to best set it up to get research-grade results. In practice though, this is seldom the case due to the complexity of the algorithms, and that's one of the main motivations of our paper. In practice, in-house scripts tend to be a less suitable choice because they tend at least to: (1) stagnate (not follow latest developments); (2) age (e.g. being developed for a given platform, they may not be easily ported to new OS); (3) not be sufficiently documented; (4) be comprised of difficult-to-read code; (5) not be user-friendly; (6) not be easily extensible to new EC systems (e.g. analyzers for new gases). Of course, this is not a rule (and there are notable exceptions), and cannot be demonstrated. It is however the result of our experience, which is why we think that, on the long term, relying on 'official' EC software is a better strategy for a large number of researchers.

In the revised manuscript, we provide a clearer definition of what we mean by 'in-house' vs 'official' software and make some of these considerations more explicit, while avoiding absolute statements that cannot be demonstrated.

7. ***P2114L11: RMA is indeed one of the most reasonable options for this kind of comparison without a "true" reference. But since according to some definitions, RMA is rather an alternative to regression than a type of it (e.g. Webster, 1997, Eur. J. Soil Sci. 48: 557), this choice should also be mentioned in the figure caption to avoid misunderstandings (keeping the term "RMA regression" while doing so is fine with me). To avoid meaningless regression parameters caused by far outliers in the first round (Fig. 1), you might also want to check alternative methods from robust statistics. Or maybe use a significance test and dont show regression lines with an insignificant slope parameter, just the R2. But on the other hand, as long as these parameters are only used to demonstrate that there are still a lot of far outliers at***

that stage (as currently done), I guess things can stay as they are. Note that one interesting conclusion is not yet mentioned: Since RMA slopes, unlike linear regression slopes, are not low-biased towards the chosen X variable but depend on the ration of the variances of both variables, the three low slopes in Fig. 1 indicate that in all these cases the variance of EddyPro results must have been several times larger than the one of the TK3 results.

Concerning the first part of the comment, in the revised manuscript we now mention 'RMA regression' whenever appropriate. We decided to keep the current plots with the 'insignificant' slopes because our aim is in fact '*only to demonstrate that there are still a lot of far outliers at that stage*', as the Referee stated it. Please note that the regression coefficients indicated in the plots published on AMTD are actually slightly wrong, as they refer to the OLS regression. In the revised manuscript we use instead the RMA coefficients, which however do not change much for Rounds 2 and 3, and certainly do not affect any discussions or conclusions.

Concerning the conclusion on EddyPro variance, we do appreciate the suggestion. We did mention a similar concept when we wrote '*Since the comparison showed that for most of the scattering data points, EddyPro results were less realistic than TK3 results […]*'. However, quantifying it in terms of flux variances adds rigor to the manuscript.

8. ***P2115L23-P2116L02: Difficult to understand from this passage: Which software package supported which interpretation of the data initially, which one (if any) was then identified as erroneous by the authors, and which adaptations to which software package were then performed?***

We did not specify those details because the cause was merely accidental. TK3 was set to treat concentration data as mixing ratios, while EddyPro was set to treat it as mole fraction. Strictly speaking, because we were not interested in the accurate estimation of fluxes but in the comparison of our implementations, we could have chosen to change the setting of either software (i.e., either change EddyPro to interpret data as mixing ratio, or change TK3 to interpret data as mole fraction), the result being the same. In actuality, we changed the settings in EddyPro because we could verify that data were, indeed, mixing ratios. We now changed that paragraph to better clarify this.

9. ***P2116L20: Maybe an example would be helpful here.***

In the revised manuscript, we added the following example: "*(e.g. in EddyPro the stationarity test is evaluated after coordinate rotation for tilt correction, while in TK3 it is evaluated before that processing step)*".

10. ***P2117L05: The term "possibly" raises some uncertainty as to whether and how iterations were performed (also does not seem to occurr in table 2).***

Iterations were not performed in our comparison. The whole sentence "*For closed-path data, the spectral correction is the last step in the chain (possibly before an iteration of the corrections) thus*

[...]" is valid in general, it is not specific to our comparison, that's why the *possibility* of an iteration. We revised the sentence to clarify that it is not related (only) to our comparison.

11. P2117L10: The decision to use the open-path dataset for ANOTHER three rounds rather than a repetition, though doubtlessly based on the experience of the authors and beneficial for the total agreement after the analysis, is a bit awkward and requires extra care in the interpretation and documentation. For example: Would the results of the first round on the open-path dataset have looked similar as the ones of the closed path dataset, if the order of both datasets would heve been reverted? Has it been ensured that the comparison results from the last round with the first dataset could still be reproduced (or at least didnt deteriorate) after the last round with the second dataset?

The Referee rises an interesting point that we didn't address in our paper, but which can help clarify (here and in the revised manuscript) the aim and the - somewhat peculiar - workflow in our analysis. A couple of things must be considered: first, by design of the comparison, the differences found are specific to the way things actually went. Repeating our exercise today, we would not necessarily make things the same way, and thus we would not necessarily fall into the same issues and the resulting discrepancies. Second, processing open-path data tests rather different parts of the codes than processing closed-path data, which is why we performed again three rounds (and the choice of 3 rounds was arbitrary, anyway). Thus, in our opinion, whether results of the first round with the open-path dataset would have been different if it was performed before the closed-path dataset is not really a relevant question.

More relevant is instead whether the closed-path comparison could be reproduced after modifying the software in the open-path comparison. We did verify that this was the case, although we did not report that in the paper. We anyway expected things to go this way, because the modifications performed during the open-path comparison do not concern the parts of the codes that are relevant when processing closed-path data.

12. P2120L01: In fact the study demonstrates that the name of the software used will be of little help without the exact version number. By the way, providing here or in the abstract the version numbers of EddyPro and TK3 from which the "last round" agreement can be expected, would be highly useful.

We agree with the Referee, and – also following the first suggestion of Referee #1 – we changed the last sentence of the Conclusions to stress that documenting the software version *and* the processing scheme would improve the quality of flux databases, allowing for example the extraction of flux data from different sources, which were processed in comparable, mutually consistent ways.

13. P2120L02: The role of "ad-hoc" intercomparisons is somewhat unclear. Do you refer to what would have happened if a researcher was willing or able to only perform round 1 of the comparison? I agree that it would have been a pity not to take the chance to improve both

software packages, but nevertheless that researcher would have obtained the correct quantification of the agreement of the used version numbers of both softwares, rather than a misconception to warn against.

We argue the conclusion drawn by the Referee ("*[…] that researcher would have obtained the correct quantification of the agreement of the used version numbers of both softwares, rather than a misconception to warn against*"), exactly because disagreement can arise due to C1, i.e. to "*inaccuracies in software setup that lead to unintended differences*", which – in our experience - happens to be a very common source of disagreement. As specified in the text, when we refer to 'ad-hoc intercomparisons' we intend those cases in which a user performs a comparison between a software to which she/he is accustomed, and a new one. In this situation, it is really common to incur in errors while setting the new software (C1), as we witness continuously with users of our software packages referring to us for support.

14. P2120L03: Any software (inlcuding e.g. any new version release of an otherwise highly proven software) not undergoing systematic quality assurance will be a risk, no matter whether it is in-house or not (see also comment on P2108L15 ff.). Similarly, at the end of the abstract it would be sufficient to demand that researchers rely on extensively validated software, no matter whether it is "established" or not.

We agree with the Referee in theory, but also speculate that – in practice - it is much more likely to find systematic quality assurance and validation procedures applied for 'official software' than for 'in-house' software. Because the eddy covariance method is not yet fully consolidated and new corrections and procedures are still being developed, it is reasonable to expect that structured software projects can catch up with these news in a more timely and quality-assured manner. As explicitly stated in the introduction this is, in essence, the message we want to give with our paper, beyond documenting the agreement between EddyPro and TK3.

References

Nakai, T., K. Shimoyama, 2012, Ultrasonic anemometer angle of attack errors under turbulent conditions, Agr. Forest Meteorol., 162–163, 14–26, doi: 10.1016/j.agrformet.2012.04.004.

Burba, G., A. Schmidt, R.L. Scott, T. Nakai, J. Kathilankal, G. Fratini, C. Hanson, B. Law, D. McDermitt, R. Eckles, M. Furtaw, M. Velgersdyk, 2012, Calculating CO2 and H2O eddy covariance fluxes from an enclosed gas analyzer using an instantaneous mixing ratio, Glob. Change Biol., 18, 385–399, doi: 10.1111/j.1365-2486.2011.02536.x.