

Author's response to the reviews

Suggestions for revision or reasons for rejection (will be published if the paper is accepted for final publication)

Submitted on 16 Feb 2021

Reviewer #1: Thomas Wutzler, twutz@bgc-jena.mpg.de

General comments

The paper by Mahabbati et al. presents an updated comparison of gap-filling algorithm, which are an important tool in the analysis of data from eddy-covariance sensors and understanding the ecosystem functioning. Their methodology is oriented at the Australian version of the data processing chain taking into account information in addition to the eddy-stations from weather forecasting models and from BIOS2 model data integration environment. They corroborate previous findings of complex methods being not much better than simple methods for gap-filling of meteorological drivers. Contrary, for the carbon fluxes itself they find a better performance of the machine learning (ML) based approaches.

The current revision took into account two of my three major concerns (gaps vs. seasonality and adding MDS). However, their comparison of gap-filling still only uses a combination of drivers that is quite specific to the Australian setup (weather-forecasting model, and the Australian BIOS2 model-data-integration environment) and hard to transfer to other sites and setups. The inclusion of the ancillary datasets is (no doubt) valueable for the performance of the gapfilling. But including an additional comparison-scenario with a constraint set of drivers to my opinion would greatly help the transferability of conclusions on the choice of methods to other readers. Hence, I still encourage the authors to include such a scenario. Nevertheless, given that the usage of quite the specific set of drivers is made clear enough, the study is worth publishing without this additional scenario.

We thank the first reviewer for his constructive comments. Using the common drivers for gap-filling the fluxes would definitely increase the transferability of the conclusions. Since the corresponding author needs to write two other papers for his PhD completion, we would consider the third concern in the upcoming papers.

I congratulate the authors to achieve running all these various types of approaches in the same setup on the same dataset.

We appreciate the time and effort the first reviewer has put in reviewing this paper. Moreover, the corresponding author wants to thank the first reviewer for introducing REddyProc online tool, which ha has found easy-to-use and convenient.

Specific comments

My concerns about “Australian setup” did not concern the selection of sites, but rather the selection of the set of inputs to the gap-filling which maybe not available at other sites.

Now that we understand the concern we would consider it for the next paper of the corresponding author's series of papers for his PhD. That is a key point for generalising the conclusions to the global community. Besides, it is worth mentioning the study undertaken by

Moore et al. (2020) where the authors used the “Australian methodology” to gap-fill the flux data (e.g. using ERA-Interim and local weather station data) in the Midwest region of the United States that provided satisfying results.

To me, the current setup of “gap-filling” of environmental drivers reads more like a downscaling or interpolation/integration of various sources. The same variable from various sources is used as a driver for the prediction this variable.

The authors claim in their reply to my specific comments: “it is less likely that changing the input features makes any change in the comparative performance of the models.” I am not in a position to assess this claim.

However, this claim together with summarizing the specific drivers should be placed prominently in the discussion together with the citations given in their reply to my concern. From my previous report I repeat my suggestion of an additional scenario “using only data commonly available at eddy-covariance sites, which are the measurements themselves (Fc, Fh, Fe) together with ancillary measurements (Rg, VPD, rH, Tair, Tsoil, Ustar, precip, wind speed, and wind direction)”. Then you can also compare the very common case of using MDS of filling Tair.

First, we would like to clarify that in the PyFluxPro, the suite of scripts whereby the EC data are processed in the OzFlux Network, just one of the ancillary sources is used to gap-fill each meteorological driver depending on availability of the data based on a priority. In this study, however, more than one source of data are used together to fill the drivers’ gaps since the outcome provided lower values for the RMSE. The sentence “it is less likely that changing the input features makes any change in the comparative performance of the models.” refers to gap-filling of fluxes based on the drivers some of which might commonly been used as input data for the gap-filling process.

We have included the suggestion in the discussion part, and the corresponding author would consider using the commonly available drivers in his following paper of the series.

Thanks for adopting the suggestion of the distribution of the consecutive gaps and the fitting to the entire data. Please, also state this also in the manuscript (section 2.3?). Currently, that way of training the model (with the data from 2012 and the data from 2013 excluding the artificial gaps, correct?), does not become clear in the current version of the manuscript.

That has been addressed in section 2.3.

Minor comments / Technical corrections

Tables: I found it hard to keep associating values within the same row. Please, consider adding some horizontal guiding lines.

Horizontal guiding lines have been added.

Table 3: Please, link to the text where the data-sources are described and maybe provide a summary in the table caption.

The table has been linked to the text.

I found it hard to always switch back to Table 2. Please, consider repeating the meaning of some acronyms at the relevant paragraphs, e.g. heading 3.1.1 “CO₂ flux (FC)”. I would prefer some slightly longer acronyms, e.g. Tair, Tsoil compared to Ta and Ts.

The meaning of the acronyms are added in part 3.

In the version I got, some reference in parenthesis are missing, e.g. P15L339, or P16L458.
The references have been added accordingly.

The references have been added.

*Moore, C. E., Berardi, D. M., Blanc-Betes, E., Dracup, E. C., Egenriether, S., Gomez-Casanovas, N., Hartman, M. D., Hudiburg, T., Kantola, I., Masters, M. D., Parton, W. J., Van Allen, R., von Haden, A. C., Yang, W. H., DeLucia, E. H. and Bernacchi, C. J.: The carbon and nitrogen cycle impacts of reverting perennial bioenergy switchgrass to an annual maize crop rotation, *GCB Bioenergy*, 12(11), 941–954, doi:10.1111/gcbb.12743, 2020.*

Suggestions for revision or reasons for rejection (will be published if the paper is accepted for final publication)

Submitted on 25 Feb 2021

Anonymous Reviewer #2

GENERAL COMMENTS

The two main changes in this version of the paper are the addition of the MDS method for gap-filling fluxes and a randomized selection of introduced gaps. These are both considerable improvements and make the results from the methods comparison more robust. However, the short records of data (only 1-2 years) was also a key concern from both reviewers and was not addressed in this revision. The paper, as it is, represents an interesting contribution showing methods were mostly comparable with the short records and the traditional MDS algorithm still performs reasonably well, and potentially the more complex methods might not lead to improvements that are worth the extra costs. However, these are not conclusive results from the paper. The paper could have been a key contribution to the literature, and although the contributions seem to be technically sound, they do not advance the state of the art in gap-filling of eddy covariance data.

SPECIFIC COMMENTS

In an answer to a reviewer comment, the authors state: "...since the main goal of the study was to compare different gap-filling algorithms, we do not believe changing the input data leads to a difference in the relative performance of the algorithm". For the comparison of these algorithms, the only factor changing their performance will be the input data. The input data is even more important for methods such as ANNs and RF, which are entirely dependent on relationships between the data variables.

As for the relative performance of the models, it is more likely that as long as all models using the same input features, the relative performance does not change considerably. For instance, Kim et al. (2020) used ANNs, RF, SVR, and MDS to fill the data gaps of methane, and the relative performance of the models almost matches the current study. Additionally, the relative performance of the methods in this study remains similar amongst the five towers, which reinforces the assumption. Finally, for the next paper of this series, we would train the developed gap-filling models using the commonly used drivers, which would address your concern.

Still in the answers, about ancillary datasets: "Even though this is true, the ancillary data used in the current study have been used to gap-fill the drivers' data, and not the fluxes directly. As such, it might not be a concern." It might be good to clarify in the methods that only the measured values for the drivers were used to gap-fill fluxes. Although it is fair to assume no gap-filled driver data was used to fill the fluxes, I couldn't find this statement in the paper.

Just to clarify: the ancillary datasets are used to gap-fill the drivers. For gap-filling the fluxes, those gap-filled drivers are used, otherwise the regression methods could not fill the gaps for which they do not have the corresponding drivers. In other words, the drivers used to fill the gaps of fluxes had included both the measured values and the gap-filled values. I wonder how it would be possible for an ANNs model to fill the gaps of a flux while it does not have access to the corresponding input data.

The argument that many previous research results use only single years for evaluation omits that most of these had limited access to long and uniform records. With record spanning over 20 years of data available from most regional flux networks, this is not a limitation any longer and should have been integral to the paper. Seasonal patterns can be correctly identified by many of the methods used, but only if using multi-year data. Using single year limited the results of the paper, which could have been a considerable contribution to both the eddy covariance and machine learning scientific communities.

The point is generally true and quite useful. But, we have had some considerations to use a limited years of data as follows:

- a) The results of this study would still be useful for the newly established EC sites and for the sites which had been active for a short period of time. It would be a valuable knowledge to know how different gap-filling algorithms perform when the training data is timely limited.*
- b) Amongst the algorithms, panel data turned out to be memory-hungry. That being said increasing the training period would have increased the memory demand at a level which it becomes impossible for us to apply it based on the hardware we had access to.*
- c) As the corresponding author is writing down three papers for his PhD thesis, the idea of using longer training periods would be included in the following papers, which would address the concern.*

In Moffat 2007 the RMSE values for the best performing algorithms (mainly ANN variants but also MDS) were consistently under 3.0 gC m⁻² d⁻¹. Since these were consistently higher in this manuscript, this might support the argument that there was too little data to train the runs presented in this paper. Since the year selected to perform the tests was very complete, if the short record is not an limitation, as argued by the authors, one could expect these results to be better.

The point might be true. Although the RMSE values depend on a variety of factors, including the magnitude of the flux values. For instance, in this study, the RMSE values of Alice Springs Mulga (Tropical and Subtropical Desert Climate) were significantly lower than those of Tumbarumba (Oceanic climate). Moreover, the gap lengths here have been mostly by far longer than those were applied by Moffat 2007 (unlike Moffat 2007 here we had gap lengths of 20 days and longer). The longer gaps have been another factor for larger RMSE values. We will check the effects of longer training data out by applying three different training periods (1, 3, and 5 years) in the next paper of the series.

The introduction of randomized gaps improves the soundness of the results. However, in the methods, it is a bit unclear how all the many realizations of the random gaps were aggregated for the final results. This could be explained in more detail. As an example, it is curious that the RMSE values for Fc at Alice Springs Mulga are so low, yet the R² values for the site are also low, while for Tumbarumba, the RMSE values are more within the expected ranges while R² values are also higher.

We have edited the final paragraph of 2.4. so that the reader clearly understand the way by which we aggregated the results and reported the performance metrics. As for the later point, since we have not included the site-by-site results in the main manuscript, it might sound a bit irrelevant to mention the point and explain the reason (smaller magnitude of Alice Springs' carbon flux has led smaller RMSE, which does not necessarily mean the performance superiority of the models for the Alice Springs data).

Finally, I will note that I disagree with the last recommendation in the conclusions.

Ensembles are useful when there isn't a "true" value against which one can compare an estimation value. In gap-filling, artificially introducing gaps (original true values) for comparisons allow precise estimations of uncertainty. Using ensembles for gap-filling would introduce unnecessary uncertainty. However, playing to the strengths of each method one can procedurally combining them (e.g., one method for short and one for long gaps) to improve final results without mixed uncertainties.

This a reasonable concern. The idea of using an [multistage] ensemble model comes from the idea that in the final stage an averaging would be taken of the outcomes of the chosen methods, which makes the results smoother and prevents large fluctuations (Yang and Browne. (2004). As such, we believe that there is a chance of declining the uncertainty of the gap-filling by using an ensemble model. Moreover, there are some researches which claims such advantage for the ensemble models, e.g. (Bormann, H. et al, 2006). However, we are not in a position to reject the second reviewer's opinion. Thus, we believe that it is worth applying the method, calculating the uncertainties, and see whether the pros of using the ensemble models outweigh the cons or not. We are already testing the idea for the second paper of the series.

TECHNICAL CORRECTIONS

- Net ecosystem exchange (NEE) is usually defined as the sum of CO₂ turbulent fluxes (commonly represented as F_c) and CO₂ storage fluxes (commonly represented as S_c); so the definition in the paper for F_c as equivalent to NEE can be misinterpreted.

Since we have mentioned "NEE" as an alternative abbreviation for F_c, and considering the context and other explanations, we believe that it would not be a serious problem, particularly for the readers who are familiar with the EC.

- It might be good to harmonize formatting for Figures 2, 3, and 4.

The formats have been harmonised.

- page 15, L449: missing reference "()"

The reference has been added.

- page 24, L703: "3)" -> "4)"

The number has been corrected.

- From previous review, in the abstract: The acronyms RF and CLR were referenced before being defined

Full names have been added.

References:

Bormann, H., Breuer, L., Croke, B., Gräff, T., Hubrechts, L., Huisman, J. A., ... & Seibert11, J.: Reduction of predictive uncertainty by ensemble hydrological modelling of discharge and land use change effects., Uncertainties 'monitoring-conceptualisation-modelling' sequence catchment Res., 133 [online] Available from: https://www.researchgate.net/publication/234016348_REDUCTION_OF_PREDICTIVE_UNCERTAINTY_BY_ENSEMBLEHYDROLOGICAL_MODELLING_OF_DISCHARGE_AND_L

ANDUSE_CHANGE_EFFECTS (Accessed 26 March 2021), 2006.

Kim, Y., Johnson, M. S., Knox, S. H., Black, T. A., Dalmagro, H. J., Kang, M., Kim, J. and Baldocchi, D.: Gap-filling approaches for eddy covariance methane fluxes: A comparison of three machine learning algorithms and a traditional method with principal component analysis, Glob. Chang. Biol., 26(3), 1499–1518, doi:10.1111/gcb.14845, 2020.

Yang, S. and Browne, A.: Neural network ensembles: combining multiple models for enhanced performance using a multistage approach, Expert Syst., 21(5), 279–288, doi:10.1111/j.1468-0394.2004.00285.x, 2004.