Comments on the first reviewer's comments:

Interactive comment on "A comparison of gap-filling algorithms for eddy covariance fluxes and their drivers" *by* Atbin Mahabbati et al.

Thomas Wutzler (Referee) twutz@bgc-

jena.mpg.de

Received and published: 2 October 2020

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. For gap-filling of meteorological drivers, they corroborate previous findings of complex methods being not much better than simple methods. Contrary, for the carbon fluxes itself they find a better performance of the machine learning (ML) based approaches. This study is a valuable contribution to the Australian setup. However, their findings are difficult to transfer to other processing setups and other sites. Hence, the paper is a quite special application and in the current form better suited for an Australian journal.

Even though the data used in this paper came from Australia, the focus was to find out whether ML algorithms other than ANNs can provide more robust results regarding gap-filling of drivers and fluxes. That being said, the towers are selected just as samples to compare the performance of different algorithms. In that sense, the paper is algorithm-oriented rather than Australian-style oriented, and the output is suitable for all members of the FLUXNET. Please note that the diversity amongst the towers has been wide, and there is less likely that an algorithm like RF, which has consistently provided a robust performance in all the five different climates and sites, perform poorly in other parts of the world, or with different input features. The initiation of this study is to compare different algorithms.

I encourage the authors for a major revisions to extent their study to setups that are comment also applicable at other sites for submission to GI.

I have several major concerns, which I state here and explain below. First, I propose to add a comparison with fitting the models to only data that are commonly available at other sites. Second, the methodology needs to be updated to introduce gaps at random positions in time instead of all starting at 1st of January to avoid confounding of gap-length with seasonality. Third, I propose to include the MDS algorithm that was simple but well performing at previous gap-filling comparisons and a "business as usual" for gap-filling NEE at many sites.

For the first suggestion, 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 algorithms. Moreover, as mentioned in the materials and methods, a variety of climates is involved in this study (Beringer et al. 2016), which makes the results useful for different types of audiences. Australia's area is almost as twice as big as Western Europe, and it has a large variety of climates.

For the second suggestion, we accepted and changed the gap-filling scenario. The gaps have now been selected randomly during 2013. The gap scenario part has been changed to: *"In order to find out the effect of gap size on the performance of our gap-filling algorithms,* the data *of nine different gap windows (i.e. 1, 5, 10, 20, 30, 60, 90, 180 and 365 consecutive days) were removed randomly from the datasets during 2013. Afterwards, the data from 2011 to 2013 were used to train the algorithms. Finally, the trained algorithms were used to fill the artificial gaps superimposed to the datasets. The entire process permutated five times in each scenario to ensure the performance was not sensitive to the gap period. As such, 15 variables, 9 window lengths, 8 gap-filling methods (MDS excluded), and 5 permutations across 5 towers resulted in 27000 computations for the meteorological features. Similarly, 3 fluxes, 9 window lengths, 9 gap-filling methods, and 5 permutations across 5 towers resulted in 6075 computations for the major fluxes, overall."*

For the third suggestion, we have included the MDS method for the major fluxes.

Specific comments

In order to be usable at other sites, the methods should be compared in addition to the presented setup by 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), and maybe another comparison using in addition more detailed radiation measurements and ground heat flux and soil water storage (Table 2).

We believe that it is a useful suggestion. However, as mentioned earlier, the main goal was to compare different gap-filling algorithms and it is less likely that changing the input features makes any change in the comparative performance of the models. For instance, Kim (Kim et al 2019) compared ANNs, RF, SVR and MDS to fill the gaps of Methane flux with different input features than this study, and the performance ranking amongst the ML methods was quite similar to this paper: RF outperformed the rest, and ANNs outperformed SVR. Besides, the data used in that research came from North America. Nonetheless, we included the MDS method using the commonly used input features: Fsd, Ta, and VPD.

In the current comparison setup, the larger gap-lengths comprise a larger proportion of other seasons, while the short gap-lengths only comprise summer records. Hence, the conclusions on gap-lengths are confounded with seasonality. I suggest to randomly distribute gaps in the portion of the entire data series with sufficiently high proportion of non-missing original data. Moreover, most data-processing setups will not fit a model for each gap tailored at the gap-length. Hence, I suggest to introduce several gaps (of a given length) across the entire dataset (say of proportions of 40% and 70% of the data according to p6L215) and let each methods fill all these gaps and compute the statistics across all the gaps but also of the aggregated annual value. In this way a recommendation can be presented that is closer to the gap-filling as applied at many sites. The decision to adjust the training window to the gap-length is very difficult to compare to other gap-filling of real time series where gap-lengths vary. Most investigators will not effort to fit a model around each gap. I suggest training the methods on a shifting window and filling all gaps inside this window, and for efficiency use only few increasing window lengths of the training.

This is a good suggestion, and this is a better approach in general for a realistic gap-filling process. As such, we changed the gap-filling scenario to the following: for each gap length, we randomly picked out a period and removed the data. Then we trained the algorithms with the rest of data, and filled the gaps. The entire process permutated five times in each scenario to ensure the performance was not sensitive to the gap period. However, the gaps were chosen consecutively to be more challenging for the algorithms. Short gaps have not been considered a concern, overall.

Moffat et al. (2007) concluded that the quite simple and widely applied MDS algorithm for filling Fc, i.e. NEE time series, which is using only the common variables NEE, Rg, Tair, and VPD as predictors. What are the reasons to omit this for many sites “business as usual”-algorithm? The computation can even be outsourced to the online tool provided by the MPI-BGC Jena.

We accepted the suggestion and included the MDS.

P7L226: Were all the eight drivers used or a subset of them, maybe different by method? What is q? The formulation “by trial and error” needs more explanation.

All the eight drivers were used for all methods, except for the FBP and MDS. Symbol q is the specific humidity, which has now been mentioned on table 2. Here “trial and error” was made based on applying feature importance analysis using random forest, and then feeding the algorithms with the different combinations of the suggested features to find out which combination provide the best performance metrics. The sentence has been edited like this: *“… based on a combination of RF feature selection and testing out a series of feature combinations.”*

P10L308: Here it does not become clear what cross-sections have been used. I imaged some categories based on similar environmental conditions or day/night time. This only becomes clear in the discussion, in that data from other sites have been used with site as cross-section. This cross-site gap-filling is hard to transfer to other studies. In what respect does the PD model differ from a classical mixed effects models?

For each tower, we used the four rest towers as its cross-sections. Now that we know how much important the similarity of the cross-sections are, it is obvious that the method can be used for the regions where the density of towers are high enough, e.g. central Europe. Nonetheless, the computational problem is also a big concern, making the method not feasible, at least as long as our computational power has not been dramatically changed. Regarding the difference of PD from classical mixed effect models, it should be noticed that PD can be considered as a combination of a classical mixed effect model with a time series model ,e.g. ARIMA models. The additional cross-sections information has been provided in the methods, accordingly.

P27L720 Conclusions 4 and 5 are mere speculations given the results presented in the paper. They should be moved to the discussion. Contrary to the suggestion 4, I hypothesize that using net radiation as a predictor should handle this case already well (at least with RF). Otherwise, I suggest first trying to add a nighttime/daytime flag to the set of predictors before splitting the dataset.

We merged the conclusions 4 and 5, and moved them to the discussion. As for the reviewer’s hypothesis, it is a good idea to be tested out. However, as mentioned earlier, this study is the first ongoing series of papers the corresponding author is going to prepare for his thesis. Thus, it is a good idea to include the reviewer’s hypothesis in the second paper, since the later should be the logical consequence of what has been found in the first paper.

P1L35: Currently, I was confused reading the abstract. It was hard for me to spot the distinction between filling of environmental drivers and filling of fluxes. This can be formulated more clearly.

The abstract has been revised thoroughly to address the issue.

Technical corrections

P2L43: This formulation does not become clear to me.

Right point. The sentence has been edited.

Tab 2: I suggest indicating the commonly used abbreviation for the fluxes in parentheses in addition to the notation of the paper (NEE, LE, H). P11L331: typo: “non-periodic” eq 12: one bar too much.

All suggestions have been done.

P23L583: I suggest to provide another table with method abbreviations or repeat the abbreviations at the beginning of the discussion. By this way you do not force your readers to study the methods section first.

Sounds useful. This has been done.

The Second Reviewer:

Anonymous Referee #2

Received and published: 20 October 2020

GENERAL COMMENTS

The paper presents a detailed evaluation of eight algorithms for gap-filling time series data, using eddy covariance data as a target for the comparisons. The content about the algorithms and the metrics for comparisons are a strong feature of the paper. However, it is more limited in advancing the knowledge of best practices for eddy covariance and micro-meteorological data gap-filling. In other words, the evaluation of the algorithms against each other is of interest, but the chosen test domain is not clearly impacted. It seems that to really benefit the knowledge of methods for gap-filling eddy covariance data, longer time series and more representative gap scenarios would be necessary, as well as a clear comparison to more established methods. Multi-year datasets are key to properly evaluate these algorithms. Such datasets are now widely available, so it is unclear why only 2013 was used. With this aspect in mind, it seems clear that in the evaluation of the first objective of the paper longer gaps led to disproportional increases in uncertainty. This might not have happened if other years without gaps for the same season were available, for instance. More direct comparisons to "classic" gap-filling algorithms would have helped in this evaluation. Implementations of algorithms such as MDS are now widely available, including as part of OZFlux's own OzFluxQC software package. The comparison of newer methods is informative, but unless compared to currently used solutions, it's hard to assess the improvement. Although the authors are correct, and performance of the MDS algorithm was shown to be comparable to ANNs before, parameterizing MDS is much simpler (no choices in layers, nodes, iterations, or window sizes) and would lead to a more robust and clear comparison.

Please note that as a PhD student whose thesis is based on a series of papers, the current paper is the very first one that has mainly provided as the initial attempt to find out how different algorithms would perform against each other. As such, almost all the points mentioned in the general comment, which are helpful, would be covered in the second paper, e.g. including multiple-year datasets, and applying different random gap scenarios. However, as the second referee has mentioned, we accept the idea of adding the results of the MDS in the current study. Last but not least, the year 2013 was chosen for the fact that the data during the period had less missing data, and that year was a common year of available data amongst all five towers that their data were used. Besides, most of the researches have been done in the field includes just one or two years of data, so the results of this paper can be compared with the majority of similar previous researches. Besides, some researchers still fill the annual gaps by using only the data of that year, thus using a year of data for training the algorithms can still be justified.

Should the authors choose to really focus on the comparison among the methods presented, I would suggest adding all the comparison metrics RMSE, R2, MBE, etc., for all sites individually and combinations thereof as supplementary materials, making this a valuable and thorough comparison of methods, and reducing the focus from the application to eddy covariance. If the intention really is to show the impact on EC, longer time series and more direct comparisons to current methods would be necessary.

We are happy with adding all the comparison metrics for all sites as supplementary materials. Besides, the intention was to make a comparison between different algorithms, and as such, in case using a year of data is insufficient, it would be equally insufficient for all algorithms.

SPECIFIC COMMENTS

On the ancillary datasets, it seems they introduce some entanglement to this evaluation. One of the key advantages of purely empirical methods, such as the ones presented in the paper, is that they will not be biased by predefined models (like the reanalysis datasets) or atmospheric interferences (like the MODIS data). After an evaluation without datasets such as these, adding them to improve the methods would be a natural choice. However, without the unbiased evaluation it is hard to qualify the sources of uncertainty in the paper's evaluation.

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.

Although the performance criteria selected for the paper work well, it is curious to see that the methods all seem to represent high variabilities but fail to capture the extremes, as the authors point out for CO2 and latent heat fluxes – and this doesn't seem to be the case for sensible heat flux. Could this be an issue of the underlying data requiring further quality control before the gap-filling methods are applied? Or maybe this is an artifact of the period selected in the examples?

This was one of the surprising things raised during the study, and to be honest, we do not have a solid answer to that yet. However, estimating the sensible heat flux is an easier task as against the two others. This can justify the exception of sensible heat flux. For Fc, and Fe, our best guess is that the issue happens due to lack of information (hidden features). We will try to figure that out in the second paper of this series.

The following claim requires either more details or a reference, otherwise it's not possible to know what concerns/challenges the authors are referring to and what aspects of gap-filling the paper is aiming to address: "...there are some serious concerns regarding the challenges associated with the technique, e.g. data gaps and uncertainties."

Those concerns have been explained in the following paragraphs.

The +/-25gCm-2y-2 (Moffat et al. 2007) and +/-30gCm-2y-2 (Richardson & Hollinger 2007) are dependent on the underlying datasets used for the evaluation. These numbers should not be taken as general benchmarks.

That is the right point. The point has been emphesised in order to not mislead the reader.

In the sentence "Nevertheless, one of the concerns regarding this algorithm is that the independent variables, here meteorological drivers, might be auto-correlated." it is unclear why this would be a concern, since the meteorological drivers being autocorrelated is one of the assumptions that allow the MDS method to work.

The comment is true. We have deleted the sentence.

The sentence "This challenge becomes acute when the gaps happen within a period when the ecosystem behaviour is changing and thereby showing different response under similar meteorological conditions." is another reason why multi-year datasets should be used to compare these algorithms.

Firstly, the gap-filling scenario has been changed in such a way that the data of up to two years (2012 and 2013) have been used now. Secondly, as mentioned earlier, we have not used multiple years of data because: (a) the focus had been on algorithm comparison, (b) most previous researches used a year or two, so our results could be more comparable with them, and (c) 2013 was the year during which the datasets for all five towers were smaller proportions of gaps. Finally, in the second paper of the series, we are using up to five years of data for training and testing. Hence, the concern would be considered in the bigger picture.

The gap scenarios and training windows selected are highly structured and rigid. It's unclear how the evaluation over these scenarios would translate into real-world examples, which have both structured gaps (e.g., from sensor failures) and arbitrary gaps (e.g., from data filtering). It seems it would be important to use are least one scenario with gaps and training data both randomized, and also combinations of lengths for gap windows and training windows.

This is a good and constructive suggestion. we accepted and changed the gap-filling scenario. The gaps have now been selected randomly during 2013. The gap scenario part has been changed to: *"In order to find out the effect of gap size on the performance of our gap-filling algorithms,* the data *of nine different gap windows (i.e. 1, 5, 10, 20, 30, 60, 90, 180 and 365 consecutive days) were removed randomly from the datasets during 2013. Afterwards, the data from 2011 to 2013 were used to train the algorithms. Finally, the trained algorithms were used to fill the artificial gaps superimposed to the datasets. The entire process permutated five times in each scenario to ensure the performance was not sensitive to the gap period. As such, 15 variables, 9 window lengths, 8 gap-filling methods (MDS excluded), and 5 permutations across 5 towers resulted in 27000 computations for the meteorological features. Similarly, 3 fluxes, 9 window lengths, 9 gap-filling methods, and 5 permutations across 5 towers resulted in 6075 computations for the major fluxes, overall."*

TECHNICAL CORRECTIONS

— Abstract -

The acronyms RF and CLR were referenced before being defined

Thank you for letting us know that. Those acronyms have been predefined in the revised version.

- "...RF provided more consistent results with less bias, relatively." It would be clearer to describe "relatively" to what in this sentence.

That is a helpful suggestion. The authors mean related to the other ML algorithms used in the study. The sentence has been edited.

- This sentence is a bit unclear "In each scenario, the gaps covered the data for the entirety of 2013 by consecutively repeating them, where, in each step, values were modelled by using earlier window data." Were measured and modelled data used simultaneously in training? — Introduction

The scenario has changed, and so as the mentioned quotation.

- "...and not measured at the point." Maybe could be "not measured at a point scale"?

That is right. We edited the sentence.

- A more classic reference for FLUXNET is: Baldocchi et al. 2001. FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem-Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities. BAMS, 11: 2415-2434.

Thank you for reminding that reference. We would use include it in the introduction.

- And more appropriate references for EUROFLUX and AmeriFlux are: Aubinet, M. et al. 1999. Estimates of the Annual Net Carbon and Water Exchange of Forests: The EUROFLUX Methodology. Advances in Ecological Research, pp. 113–175. Law, B. 2007. AmeriFlux Network aids global synthesis. Eos, 88, 286–286. Novick, K. A. et al 2018. The AmeriFlux network: A coalition of the willing. AFM, 249:444-456.

We used one of the references in the revised version.

- "Despite the capability of EC to frequently validate process modelling analyses..." might be more precisely phrased as something like "Despite EC data being frequently used to validate process modelling analyses..."

The suggestion has been considered.

- "[...] Moffat et al. (2007) compared a couple of different commonly-used gap-filling algorithms"; in fact, Moffat et al. 2007 compared 15 gap-filling techniques.

Right. We have replaced "15" instead of "a couple".

- **Materials and Methods**

- "and Tumbarumba form 2011 to 2013..." form -> from

Thank you for mentioning the mistake.

- "Each algorithm was tuned up individually using gird search,..." gird -> grid

Thank you for mentioning the mistake.

— Results

- Even with a maximum zoom in the PDF file, it is rather hard to read the axis for Figures 3 and 4

Since the scenario has changed in the revised version, the mentioned figures could not be plotted anymore. They have been removed.

— Discussion

- This sentence is unclear: "That is because ANNs have been checking out for a long time in different locations and considered as one of the most reliable algorithms in the field for more than a decade"

The authors mean occasional superiority of random forest algorithm, needs to happen in several future studies to convince us to suggest RF instead of ANNs, or identify the algorithm as another standard method. We will add a sentence to clarify the point.