In the following response, the original review is shown in grey and our response in green.

RC2: 'Review of "CLIMFILL: A Framework for Intelligently Gap-filling Earth Observations"', Rene Orth, 30 Aug 2021

Review of Bessenbacher et al., gmd-2021-164

"CLIMFILL: A Framework for Intelligently Gap-filling Earth Observations"

This study introduces a sophisticated procedure to gap-fill Earth observation time series while benefitting from independently and concurrently observed related variables. The authors showcase the method with reanalysis data where some parts are intentionally masked, and the reconstructed estimates are finally compared with the original data. Thereby, they consider ground temperature, terrestrial water storage, surface layer soil moisture and precipitation and discuss the results both in terms of reconstucted individual time series, and for the interactions between reconstructed variables compared with respective estimates from the original data.

Recommendation:

I think the paper requires major revisions.

This is a useful and timely contribution for the Earth science community, and interesting for the readership of the Geoscientific Model Development. Benefitting from a growing suite of Earth observations,complex statistical tools and machine learning applications are increasingly employed in Earth science research.Mostly, these analysis tools require gap-free data which is often derived through gap-filling procedures.In this context, improving the quality of the gap-filling by exploiting the relationships between the independent Earth observations is a promising avenue.

However, I have some concerns regarding the description of the method and the benchmarking of the results, as detailed below.

We thank Rene Orth for this useful feedback and the helpful comments below.

General comments:

(1) Comparing the results from the plain interpolation with that at the end of all four steps of the gap-filling procedure is interesting to understand the method and the relevance of the various steps. However, it is not a suitable benchmarking exercise as it is to be expected that the results after four steps are closer to the original ERA5 data than the result after the first relatively crude interpolation step. Instead, an established univariate

gap-filling technique should be employed here as a benchmark to illustrate under which circumstances the presented methodology offers benefits over previous approaches. Also, this could reveal to which is extent the gap filling can be improved by (i) complete exploration of uni-variate time series beyond neighbors, versus (ii) a multivariate approach.

Thank you for pointing this out, a very similar issue was raised by Anonymous Referee #1 and we have copied the answer here for your convenience:

“We thank Anonymous Referee #1 for this valuable comment. Indeed the proposed interpolation method which is used in step 1 of CLIMFILL as well as as a benchmark for the CLIMFILL algorithm is simple. In step 1 a simple method was chosen deliberately since the primary focus is to initialize the subsequent iterative optimisation process. Moreover, the method was chosen to ensure that the problem stays computationally feasible. For the revised manuscript however, we will consider the application of other interpolation methods in step 1 of CLIMFILL as well as in benchmarking the CLIMFILL algorithm with other interpolation approaches such as kriging, Gaussian processes, or derivatives thereof.

Note also, that the high spatial and temporal resolution of the data and the relatively large number of observed values (as compared to an interpolation based on station data, like in E-OBS (Haylock et al, 2008)) make kriging computationally very expensive. Kriging has a cubic computational complexity, which means the computational expense increases with O(n**3), where n is the number of observed values. Standard kriging is therefore not an option for our dataset which consists of several hundred thousand observed points per day.

To counteract this issue we are currently exploring recent literature on Gaussian Processes for large data, focussing e.g. on divide-and-conquer approaches in which the data are strategically split into more manageable chunks.”

(2) I think it would be useful for future CLIMFILL users to give more guidance on the methods to use in each step of the algorithm. Table 2 offers many possible choices, but in addition some recommendations would be needed on when to use which method and why. Also, the selection of employed variables is important as their inter-relations are a key source for the gap reconstructions, so also some additional advice on this would be helpful.

Thank you for this suggestion. We will include additional text discussing the merits of methodological choices, while also acknowledging that the perfect setup will likely depend on the application. For the selection of employed variables, please see our reply to your next comment (3).

(3) I think that the feature selection is a bit arbitrary and dependent on expert knowledge. To somewhat address this issue, maybe several features could be used by default, such as the 34 features used in the presented example and maybe even additional time lags and windows. Then, the random forest model can be employed to rank the features by their importance (e.g. using SHAP value importance) to make a more informed decision

on the useful features. Finally, the gap-filling could be re-run with only retaining relevant features.

This is a very helpful comment. Anonymous Referee #1 also suggested adding more time lags and windows, especially in the future. The initial choice indeed was indeed guided by expert knowledge. For the revised version we aim to quantitatively quantify feature importance and explore the effects of adding more features or removing features from the dataset.

(4) There is advanced statistical and data science language used across the manuscript and I recommend to clarify this with additional information to allow a broader geoscientific audience to follow this manuscript. Please see my respective suggestions in the specific comments below.

We will revise the mentioned sections of the text to make it more accessible to a broader geoscientific audience.

I do not wish to remain anonymous - Rene Orth.

Specific comments:

line 2: estimates for what?

changed or “estimates of statistical moments”

line 5: remove "up"

The wording is changed accordingly.

line 7: I agree that technically the algorithm does not require a gap-free donor variable; however if all variables have gaps at the same time and if this period is longer, then the final gap-fill estimate will naturally have a low quality

Yes, it is true that if all variables have gaps at the same time this will likely affect the quality of the estimate. Note, however, that this does not have to result in poor overall performance in case of a good first guess in step 1.

line 15: "profit", maybe rephrase as "are improved by"

The wording is changed accordingly.

lines 45, 144 & Table 1: Jung et al. 2019 and O & Orth 2021 are relevant studies in this context and could be mentioned here

We thank the reviewer for pointing us towards these publications.

line 46: please clarify "scale somewhere between"

We suggest replacing the mentioned text with “are positioned on a gradient between”

line 84: please clarify "difficult observational record"

We change the wording from “difficult” to “fragmented” in coherence with line 28, 43 and 51

lines 108/109 and 111 are in contrast to each other

A good point. We will either describe “faulty sensor pixel” scenario as MCAR scenario or remove sentence “All these three missingness patterns can be observed in earth observation data:”

line 151: this is unclear, please rephrase

Sentence changed to “There is a growing body of literature of different methods that are originally equipped with dealing with only spatial or temporal relations *that* are expanded and altered to take into account the information from the other dimension as well (von Buttlar et al., 2014; Gerber et al., 2018).”

line 154: "another" should be "other" I guess

The wording is changed accordingly.

Table 2, caption: "other" should be "another" I guess

The wording is changed accordingly.

Table 2, right column: "or more complex interpolation methods", "Guided by ...", these are not exactly examples as the column title suggests

The wording is changed to “Possible alternative methods”

line 170: remove "on"

The wording is changed accordingly.

line 171: feels a bit random which letters are capitalized here and which are not

The section title will be changed to “CLIMFILL: A Generalised Framework for Infilling Missing Values in Multivariate Spatio-Temporal Geoscientific Data”

line 173: "the highly structured nature", please explain

We suggest replacing “the highly structured nature” with “the spatial, temporal and cross-variable dependence structure” to make the meaning more clear.

Figure 2, caption: The framework is divided into four steps, not three.

This will be corrected.

line 178: Abbreviation CLIMFILL is mentioned earlier and should be explained at the first occasion

The abbreviation CLIMFILL will only be explained in the abstract at the first mention.

line 181: please clarify "correlation structure"

We will replace “correlation structure” with “dependence structure” to ensure consistency across the text. (for example, lines 189, 366, 450,...)

lines 203, 311: please clarify "constant"

The wording will be changed from “constant” to “time-independent”

line 216: quotation marks not needed

The sentence is changed accordingly.

lines 229: please clarify "stabilising the results"

The sentence now reads: “For reducing the risk of discontinuities at the cluster edges, the clustering procedure is repeated E times with different numbers of terminal clusters on copies of the data X(1),...,X(E).”

line 231: please clarify "terminal clusters"

We will remove the word “terminal” for clarification.

line 243: I think this should be "to overwrite the former estimates"

The wording is changed accordingly.

lines 250/251: "learns different weights", please clarify

We replace “weights” with “model parameters” for clarification.

Figure 3, caption: replace "substracting" with "subtracting"

The wording is changed accordingly.

line 272: How are deserts defined and detected?

We change the sentence from “Permanently glaciated areas and deserts are masked” to “Permanently glaciated areas and deserts (defined as areas with less 50 mm average yearly precipitation in the years 2003-2012) are masked”.

line 311: It should be 4 and not 3 additional features I guess?

We replace “3” with “four”.

line 314: please clarify "non-normality"

We change “treat its inherent non-normality” to “transform the values into a gaussian distribution”.

line 316: How does this add up to 34?

four variables + precipitation additional variable = 5

four embedded features per variable = 4*5 = 20

6 constant maps

lat, lon time = 3

5 + 20 + 6 + 3 = 34 features

We will add in the text brackets to ensure the number of features is clarified. Please also note that these might change due to comments of both reviewers about feature selection and feature importance (see above).

line 319: "respectively" should be added after "clusters" I guess

The wording is changed accordingly.

line 326: I wonder if and how different spatial resolutions can affect the accuracy of the gap filling, it would be great if the authors could shortly discuss this.

This is a good question. The different spatial resolutions indeed have similar accuracy in gapfilling, and we will add a discussion and a figure in the appendix to show this.

line 326: "where one fold is one year", please clarify

Thanks for spotting this example of convoluted machine-learning lingo. We change the sentence from “The hyper-parameters 325 of the supervised learning functions are determined via leave-one-out cross-validation on clustered ERA5 data between 2015 and 2020 downscaled to 2.5 degrees resolution, where one fold is one year.” to “The hyper-parameters 325 of the supervised learning functions are determined via leave-one-year-out cross-validation on clustered ERA5 data between 2015 and 2020 downscaled to 2.5 degrees resolution.”

Figure 7, caption: what is "CLIMPUTE-RF"?

Thanks for spotting this typo. We change it to “CLIMFILL-RF”

line 351: please clarify "det"

$\det$ is the determinant of the respective covariance matrix. Note that sentence however might disappear when the Bhattacharyya distance is replaced in response to a suggestion of Anonymous Referee #1 .

Figure 8, caption: sentences should not end with "with" and "create".

The wording is changed accordingly.

line 361: "This" should be added before "leads".

The wording is changed accordingly.

line 367, section 3.4: I very much like the idea of studying the performance of the gap-filling across missingness patterns and different severity of the gaps.

We thank the reviewer very much for this feedback.

Figure 10, caption: the B-distance is not actually displayed in this figure

We change from “pearson correlation, root mean square error (RMSE) and B-distance (for more details see text)” to “pearson correlation and root mean square error (RMSE)”

line 373: How exactly are the satellite swaths imitated?

We have added an explanation in caption of figure 8 as a response to Referees #1 and #2: “In swaths-only missingness we create long ellipses centered around the equator to simulate characteristic satellite swath missingness patterns”.

line 401: I do not quite understand the point on the bias correction.

The mentioning of the bias reduction step will be revised.

line 427: similar in "remotely sensed" data but underestimated in "satellite observations", this should be the same thing?

We agree this sentence reads a bit confusing and will revise it accordingly.

Figure 2: The figure is rather small now and should be enlarged to make it easier to see all details.

The figure is changed accordingly.

Figure 4: The months axis should not go to 12.5

The figure is changed accordingly.

References:

Jung, M., et al., The FLUXCOM ensemble of global land-atmosphere energy fluxes, Sci. Data 6, 74 (2019).

O, S. and R. Orth, Global soil moisture data derived through machine learning trained with in-situ measurements, Sci. Data 8, 170 (2021).

References:

Haylock, M.R., Hofstra, N., Klein Tank, A. M. G., Klok, E .J, Jones, P. D. and New, M. (2008): A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006. Journal of Geophysical Research, doi:10.1029/2008JD010201