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

RC1: 'Comment on gmd-2021-164', Anonymous Referee #1, 18 Aug 2021

Review of « CLIMFILL: A Framework for Intelligently Gap-filling Earth Observations » by V. Bessenbacher et al.

This manuscript addresses an important problem: the gap-filling of global observations and the generation of continuous spatial and temporal data. It is well written and the methodology is mostly clear. This said, I have some reservation regarding the justification of the methodology and the validation approach. These are detailed below.

We thank Anonymous Referee #1 for this useful feedback and the helpful comments below.

Major comments:

- The method is well described, but involves a number of modeling choices (initial interpolation method, clustering approach, random forest estimation and averaging) that are not always justified, except by the experimental results showing that “it works”. The problem is that I cannot make sure that it works with the current benchmarking. Indeed, the proposed method is compared only against the interpolation of step 1 of the proposed method itself. It is not in my opinion a sufficient benchmark. While it shows that steps 2-4 do have some added value compared to the extremely simple interpolation of step 1, added value against other interpolation methods is not demonstrated. By construction, it is expected that steps 1-4 perform than step 1 alone. I suggest to demonstrate the performance of the proposed approach is to compare it against something slightly more sophisticated, and already known to work in such contexts, for example (co-)kriging, possibly with a separate variogram model in each of the clusters defined in step 3.

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.

- The introduction stresses, with reason, that the reproduction of the dependencies between variables is critical, and that these dependencies are complex. However, the evaluation metric used relies on the assumption that these distributions are Gaussian, whereas it is clearly not the case, as seen in figure 6. Instead of eq. 2, I suggest using a metric that considers a numerical description of the joint distribution, such as for example the Jensen-Shannon divergence (or many other possible divergences available in the literature). Applied to the distributions in figure 6, the computational cost would be minimal.

For the revised manuscript, we will seek to replace the Bhattacharrya distance with a metric that does not assume Gaussian distributions. We will particularly look into options that make use of the distributions in Figure 6, once they are normalised (as suggested in the next comment). We thank the reviewer for suggesting the Jensen-Shannon divergence/distance, which we will consider.

- Some elements in figure 6 do not allow me to fully evaluate the results of the proposed method. I see significant differences between the distributions, e.g. in d) there is an important bias towards values of soil moisture around 0.3, which seems more important than in c). More generally, the distribution in d) looks globally like c) but smoothed (comparable to the smudging effect of adding a random noise). Similarly, in figure A1 there are important artifacts in the reproduction of the marginal distributions by CLIMFILL, which imply that the joint distribution is also inaccurate. Visually interpreting these effect is difficult because the joint distributions are presented as histograms, with counts of data instead of densities of probability. As a result, the integral of the joint distributions is not the same, especially for b). These histograms should be normalized by their integrals to reflect probabilities rather than counts.

We agree that normalising the histograms is a good idea and will do this for the revised manuscript. For the other issues mentioned in this part, please see the summarised response to a related comment below (comment starting with “While I see the logic in separating ….”).

- In figure 11 as well as in the supplementary figures, I cannot see that CLIMFILL is systematically better than the simple interpolation. A quantitative assessment might help highlighting such differences. Why not using the same regions in figure 12 as in figure 11? Is the focus on randomly chosen regions or on the areas of larger discrepancies?

In Figure 11, the current SREX regions have been chosen to highlight exemplary how CLIMFILL performs in different climates across the globe. In Figure 12, the aim is to

focus on the regions where the largest difference is seen between original data, benchmark and CLIMFILL. We will revise these two figures to also quantitatively reflect the differences between benchmark and CLIMFILL.

- While I see the logic in separating the interpolation of a global trend (step 1) and detailed data-driven smaller scale features (steps 2-4), step 1 is a spatial KNN, which inherently assumes smoothness. This is a modeling decision having implications that are not evaluated. For example, the distributions in figure 6c and 6d present features that are absent from the dataset used to interpolate from (figure 6b). It means that some unobserved statistical properties have been created. It seems to me that the large peak in figure 6c and its smoother version in figure 6d are typical of nearest-neighbor algorithms that propagate a single nearest value far from observations.

We thank the reviewer for this helpful feedback. Indeed, the distribution in Figure 6c presents a feature that is absent in Figure 6b, namely a large number of previously missing surface layer soil moisture values that have been gap-filled with a value of roughly 0.3 m^3/m^3. The feature is a consequence of large gaps, where the initial gapfill by interpolation because no spatially or temporally close points are found to inform about the missing value. In these instances missing values are initialized with the (constant) global mean, which is causing this issue.

As mentioned above we will consider other interpolation techniques in the revised submission, which may alleviate this issue. In case this issue remains present (note e.g. that kriging would also converge to constant values in the presence of large gaps) we will ensure that it is properly discussed in the revised text.

- The interpolation approach is largely driven by the large number of features, which is fine and quite usual, but this means that it may not perform well in case of large gaps, or when some of these covariates are unknown. How does it perform when no covariates are present (or only e.g. topography and lat/lon)? Furthermore, it is mentioned in l. 155 and below that a shortcoming of existing gap-filling approaches is that they heavily rely on covariates and not on spatial relationships. As I understand it, CLIMFILL also relies largely on covariates (steps 2 and 3), and very little on spatial relationships (spatial dependence is only considered in step 1, and in a very loose way as through a nearest neighbor approach).

We thank the reviewer for pointing out that this part of the algorithm needs clarification in the text.

For our response below we assume that “the interpolation approach” refers to the whole CLIMFILL framework. For understanding the properties of CLIMFILL (and the imputation methods it is based on (Stekhoven and Bühlmann, 2012)), it is necessary to recall that the gap-filling happens in two distinct steps:

First, an ad-hoc estimate for missing values is derived for each variable separately. Here this is done using spatial interpolation and hence does not depend on the presence of gap-free covariates.

Second, the ad-hoc estimate of all considered variables is optimised using an iterative procedure by accounting for the covariance of all participating variables.

Consequently, the entire procedure is designed to operate on multivariate datasets in which gaps may be present in all variables. In particular, the second step can help to overcome limits of simple interpolation in the presence of large gaps, since information of other - possibly gap-free - variables is exploited. Nonetheless, we acknowledge that this may not always be the case and will revise the article to better communicate the merits and limitations of the presented approach.

- One problem I see with the proposed approach is that it does not consider or attempt to quantify uncertainty. The values in the middle of a large gap are given with the same confidence as for a single pixel gap. Similarly, the uncertainty should be larger when few covariates are present. Furthermore, on l.232 it is mentioned that the different clusterings obtained (which may in some sense convey a sense of variability) are averaged, thus collapsing any uncertainty into a mean value.

We fully agree with Anonymous Referee #1 that uncertainty quantification is an important aspect of statistical inference and that the confidence in the estimate is likely proportional to the density of available observations. However, with the structure of the considered data and the used methodology, this uncertainty cannot be treated using standard methods. For example,the algorithm consists of 4 steps, each carrying its specific methodological uncertainties and it is not straight-forward to combine these uncertainties. Furthermore, the complex dependence structure of the data, exhibiting temporal, spatial, and cross-variable dependence needs to be considered.

Consequently the development of well-behaved uncertainty estimates for CLIMFILL would require a significant and independent research effort which is beyond the scope of this study.

Minor/editorial comments:

- Some of the references are quite outdated, such as Rubin (1976) that is mentioned repeatedly, whereas the literature on spatial statistics and geostatistics, which is precisely concerned with interpolation in similar spatio-temporal applications, is quite incomplete. Some starting points could be:

Cressie, N. and K. Wikle (2011). Statistics for Spatio-Temporal Data. New Jersey, Wiley.

Chilès, J. P. and P. Delfiner (2012). Geostatistics: Modeling Spatial Uncertainty: Second Edition.

We thank the reviewer for these suggestions and will adapt the literature overview of spatial statistics and geostatistics to reflect the recent research, starting with the publications mentioned.

- Limitations of Gaussian processes are mentioned in l. 136, but with no details. There are several applications in the literature where Gaussian processes (or

other forms of random processes) have been successfully used with large datasets.

This is a valid point, we will revise this section to reflect this.

- 139: are well suited

The wording is changed accordingly.

- 153: provided by other variables

The wording is changed accordingly.

- Table 2, caption: it is not clear to me what is meant by “method class” in this table

We changed the Table caption to: “Main CLIMFILL settings per step. Each task can be performed using alternative methods mentioned in the last column”

- 170: outlook for possible future work

The wording is changed accordingly.

- Caption of figure 2: the framework is divides into four steps

The wording is changed accordingly.

- 203: constant maps describing properties

The wording is changed accordingly.

- 206: Please develop the motivation for the Takens Theorem. I do not see the link to the present approach, especially given the observational uncertainties considered here.

This section will be revised so that the link becomes more clear.

- 217: built from variables

The wording is changed accordingly.

- 216-218: This sentence seems to refer to the data format in the specific implementation described. Probably not needed in this methodological description.

We omit this sentence.

- 244: I understand that the subscript “updated” stands for estimated value. A more common notation would be to use a hat.

Thanks for this really useful suggestion. We will implement the notation with the hat.

- 244: the meaning of subscript m is unclear to me. Is it the same as in l. 217?

The meaning of the subscript “m” is the same in line 244 and line 217. It refers to the number of features, as described in the caption of Algorithm 1. We thank the reviewer for pointing out that this could be communicated a bit clearer. We will make sure to mention the definition of the subscripts m and n in the text before line 217.

- 266: it is mentioned that the proposed approach could be used to interpolate sparse in-situ measurements. This could be expanded upon or removed, as I do not see how it could be achieved easily in the present form because the model is heavily data-driven. Same comment for l.489-490.

This is a good idea. We will explore whether these two sections can be removed.

- Figure 5: why id the temporal window smaller in the future period than in the past?

We thank Anonymous Referee #1 for this helpful feedback. As a response to this comment and a comment from Referee #2 (Rene Orth), we will systematically explore the impact of different versions of feature selection, i.e. step 2 of the algorithm. We aim to quantitatively explore the effects of adding more features or removing features from the dataset.

- 305: the point is filled

The wording is changed accordingly.

- 323: are scalable

The wording is changed accordingly.

- 339: what is the criterion allowing to state that the shape of the distribution is well recovered?

When introducing quantitative measures of the reconstruction of the original distribution in the benchmark and CLIMFILL results, for example by using the Jensen Shannon divergence as proposed on the histograms in Figure 6. The text will be revised accordingly

- Legend of Figure 8: a) is described twice in the legend and c) is missing.

We replaced the respective sentence with “Comparison of artificial (a) random and (b) swaths-only missingness and (c) missingness in the real data“

- Legend of Figure 8: “In swaths-only…”: incomplete sentence

We have added the missing part such that the sentence now reads “In swaths-only missingness we create long ellipses centered around the equator to simulate characteristic satellite swath missingness patterns.”

- 361: Leads to…: incomplete sentence.

changed to “This leads…”

- 401: This last sentence is intriguing, especially during the presentation of results. It could be expanded upon in the discussion.

The mentioning of the bias reduction step will be revised.

- 468: recovering the physical

The wording is changed accordingly.

- 498-500: “closing the largest gaps first”: it is not immediately clear to me that this would be the best strategy. One potential drawback of this approach might

be (but I am not sure) to artificially reduce the uncertainty related to the large gaps, precisely where uncertainty is important. One could also argue that a strategy could be to start with areas that are fairly certain (i.e. small gaps).

We acknowledge that the wording of this sentence is not optimal and will revise the text with focus on communicating uncertainties of the estimates.

- In figure A1, I recommend showing the precipitations on a log axis, and normalizing the joint distributions rather than displaying counts (same comment as for figure 6).

The figure is changed accordingly.

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

Stekhoven, D. J. and Bühlmann, P. (2012): MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics. 28, 112–118, doi:10.1093/bioinformatics/btr597