Reply to Reviewer Comments to the Manuscript

CLIMFILL v0.9: A Framework for Intelligently Gap-filling Earth Observations

Verena Bessenbacher, Sonia I. Seneviratne, Lukas Gudmundsson
Geophysical Model Development, doi:10.5194/gmd-2021-164

RC: ***Reviewer Comment***, AR: *Authors response*, ☐ Manuscript text

1. Letter to the editor

Dear Prof. Tomomichi Kato,

Please find enclosed the revised version of the article entitled "CLIMFILL v0.9: A Framework for Intelligently Gap filling Earth Observations" (Paper GMD-2021-164).

Overall, the reviewers did support the methodological basis of the article. However, both reviewers suggested some modifications to parts of the framework. We have addressed them below in our point-by-point response letter. Following these comments, changes have been made to the individual steps of the gap-filling process as well as to its evaluation. In particular, the following main revisions were made to the manuscript:

- In the first step of the algorithm, the simple spatiotemporal interpolation using nearest neighbours was replaced by a more sophisticated, state-of-the art interpolation that is based on a well established approach from Haylock et al., (2008) with modifications to reduce computational expense. It is a combination of thin-plate-spline interpolation and kriging. Consequently, also the benchmark used for comparing CLIMFILL has changed to this interpolation method.
- We replaced the Bhattacharyya-distance with a non-parametric measure for computing the distance of two multivariate distributions, the Jenson-Shannon distance, as suggested by Anonymous Referee #1, to account for the non-normality of the data.
- In the second step of the algorithm, several "flavours" of feature engineering have been explored to analyse the impact of considering different sets of variables to the gap-filling. We evaluate which set of features gives the best results at the beginning of the results section.

With these modifications triggered by the suggestions of the reviewers, the computational expense grew outside of what was capable for our computational resources. Therefore, we had to decrease the amount of data used in this study from ten years to one year. Most prominently, trying different sets of features requiring running CLIMFILL several times as compared to before, only once. One CLIMFILL run on our machines with 128 cores takes approx 24h with the newly reduced dataset. Furthermore, computing the Jenson-Shannon distance is approx. 100 times more expensive than the Bhattacharyya-distance. The caveat of this decision is that the analysis of interannual variability is inhibited, which triggered changes in Figure 11 and 12 consequently.

Note, however, that neither the modifications to the methodology suggested by the referees nor decrease in the amount of data have changed the overall conclusions of the manuscript.

The valuable comments from the reviewers changed the structure of article substantially throughout the text. These changes can be viewed in the accompanyed difference-document. We are confident that the above-mentioned revisions of the paper together with the updated supporting information have increased the value of the submitted manuscript.

Yours Sincerely,

Verena Bessenbacher (on behalf of all co-authors)

2. Astrid Kerkweg

2.1. Comment

RC: ***The main paper must give the model name and version number (or other unique identifier) in the title.***

AR: *We thank the Executive Editor for this comment and adjusted the title of the manuscript according to the GMD guidelines.*

> CLIMFILL v0.9: A Framework for Intelligently Gap-filling Earth Observations

3. Anonymous Referee #1

RC: ***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.***

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

RC: ***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.***

AR: *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. Due to the high spatial and temporal resolution of the data and the relatively large number of observed values, kriging is 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. For the revised manuscript however, we have now replaced both the interpolation in step 1 and the benchmark with a state-of-the-art spatial interpolation method borrowed from Haylock et al., (2008) adapted with a sparse, computationally efficient kriging method Das et al., (2018) to make the problem computationally feasible. We described the procedure in the methods section:*

> The interpolation step creates initial estimates based on the spatial or spatiotemporal context of the gap using interpolation. Following the approach of Haylock et al. (2008), the data is first divided into monthly climatology maps and anomalies. The climatology maps are gap-filled using thin-plate-spline interpolation to represent the spatial trends in the data. Subsequently, the daily anomalies from the monthly climatology are gap-filled using kriging. In contrast to the E-OBS dataset created in (Haylock et al., 2008) from in-situ observations, satellite data has a much larger number of observed values, making a direct implementation of this approach computationally infeasible. For the interpolation of the monthly climatology maps we there- fore restrict the thin-plate-spline interpolation to the 50 closest neighbors of each point. The interpolation of the daily anomalies follows Das et al. (2018), who suggest reducing complexity of kriging/Gaussian Process regression by repeated interpolations on random sub-samples of all available data points and averaging the resulting estimates. In particular, the missing values in the anomalies are estimated by randomly selecting 1000 observed points per month over which the interpolation is calculated. This is repeated five times and the mean of all interpolations for each missing point is taken as the gap-fill estimate. Finally, monthly maps and anomalies are summed up to form the inital gap-fill estimate from step 1.

Please also note changes to the framework regarding to the clustering step described in the answers below.

RC: ***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.***

AR: *We thank Anonymous Referee #1 for this valuable comment. For the revised manuscript, we have replaced the B-distance with the Jenson-Shannon distance based on the distributions in Figure 6. We describe this in the text:*

We additionally examine which subset of features is most descriptive for the problem at hand and settle on one of the propositions. To allow for a quantitative assessment of the similarity of the multivariate distributions of observed and simulated variables~~. To overcome this issue~~ , we apply a scalar measure of multivariate similarity. In this study, we use the Jenson-Shannon distance (JS-distance). This measure compares the ~~multivariate~~ distance between two ~~datasets or~~ multivariate distributions, where a value of one means that the two samples are from the same distribution, and ~~a positive value indicates~~ one indicates that the distributions are not overlapping. We apply the JS-distance on the four-dimensional histograms computed of the ~~relative distance between two distributions .~~ four variables using 50 bins for each variable.

RC: ***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.***

AR: *We have normalised the histograms by dividing the counts by the number of datapoints in the original ERA5 data to ensure comparability among the plots (a) through (d). 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 ...."). *

RC: ***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?***

AR: *We have changed Figure 11 as follows: we have added the RMSE of original and CLIMFILL values to allow for an overview across regions and highlight several regions to show exemplarly how CLIMFILL works there. Figure 12 has changed completely and therefore this comment is not applicable anymore.*

RC: ***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.***

AR: *We thank the reviewer for this helpful feedback. Indeed, the distribution in Figure 6c presented a feature that was 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* $\frac{m^3}{m^3}$*. The feature was 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. Replacing the interpolation step with the new method described above has now removed these statistical artefacts.

RC: ***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).***

AR: *When no covariates are present, the initial interpolation gap fill is not changed by the iterative, multivariate procedure in step 4 and therefore the inital gap fill is used. We have rewritten the literature review to stress that the shortcomings of existing gap filling approaches is that they cannot ingest spatial, temporal and multivariate dependencies once missing values are present in more than one of the included observations and hope this clarifies this issue. Please see the new introduction text for the respective changes, as there are too many to quote them here directly.*

RC: ***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.***

AR: *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. Consequently, we have removed the part where different clusterings are obtained: instead of 3 different clusterings where each of them is gapfilled with 100 trees in step 4, we now have only one clustering, but each of them with 300 trees. We have verified that the results are similar with this changed method.*

RC: ***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.***

AR: *We thank the reviewer for the provided sources and have added literature on spatial statistics in the introduction:*

> In the geoscientific literature, among the most commonly used approaches for estimating unobserved points are spatial and temporal interpolation methods, including nearest neighbour regression as well as kriging and derivatives thereof (Liu et al., 45 2018; Cowtan and Way, 2014; Haylock et al., 2008; Cressie et al., 2006) (for an overview see Cressie and Wikle 2015; Allard et al. 2013). Spectral methods are used as well (Zhang et al., 2018; von Buttlar et al., 2014; Brooks et al., 2012).

RC: ***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.***

AR: *Gaussian processes and their applications with large datasets are now reviewed in the introduction:*

> A wide range of algorithms that make use of cross-variable dependence to estimate missing values exist in statistical liter- ature. In the following, we are highlighting two common approaches: On one hand, Gaussian processes are a natural choice for gap filling problems (Gelfand and Schliep, 2016) and are mathematically identical to kriging, if the predictors are latitude and longitude. Gaussian processes however have limitations when moving to large data (Heaton et al., 2019) as is the case in Earth observation data. In recent years, some applications of Gaussian processes have been shown to work in settings with too much data to estimate the co-variance matrix between all datapoints precisely. They estimate the co-variance matrix via sophisticated sampling techniques (Wang and Chaib-draa, 2017; Das et al., 2018), pre-process the data via dimension reduction methods (Banerjee et al., 2008) or apply the Gaussian Process to local subsets of the data (Gramacy and Apley, 2015; Datta et al., 2016).

RC: ***139: are well suited***

AR: *The wording is changed accordingly.*

> On the other hand, iterative procedures like the MICE-Algorithm ("Multiple imputation by chained equation", van Buuren (2018)) are ~~suited well~~ well suited for multivariate imputation and scale to large data, but cannot account for neighborhood relations.

RC: ***153: provided by other variables***

AR: *The text has changed, therefore this comment is not applicable anymore.*

RC: ***Table 2, caption: it is not clear to me what is meant by "method class" in this table***

AR: *The table has been removed, therefore this comment is not applicable anymore.*

RC: ***170: outlook for possible future work***

AR: *The wording is changed accordingly.*

> Finally, Sect. 4 discusses the results and provides a conclusion and an outlook ~~on~~ for possible future work.

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

AR: *The wording is changed accordingly.*

> The framework is divided into ~~three~~ four steps.

RC: ***203: constant maps describing properties***

AR: *The wording is changed accordingly.*

> For example, gap-free ~~constant maps of~~ time-independent maps describing properties of the land surface such as topography or land cover can be included.

RC: ***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.***

AR: *The reference to the Takens Theorem has been removed.*

RC: ***217: built from variables***

AR: *This sentence has been omitted as a response to the next comment.*

> ~~For the application of data science methods, the data need to be rearranged in a table **X** build from variables $v_1, ...v_n$ and derived features $v_1^*, ..., v_m^*$ as columns and space-time points as rows.~~

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

AR: *We omit this sentence*

> ~~For the application of data science methods, the data need to be rearranged in a table **X** build from variables $v_1, ...v_n$ and derived features $v_1^*, ..., v_m^*$ as columns and space-time points as rows.~~

RC: ***244: I understand that the subscript "updated" stands for estimated value. A more common notation would be to use a hat.***

AR: *Thanks for this really useful suggestion. We have implemented the notation with the hat in the text:*

Subsequently both $\mathbf{y}_v^{(e,k)}$ and $\mathbf{X}_{-v}^{(e,k)}$ are divided into two sets of data points: (1) all data points where $\mathbf{y}_v^{(e,k)}$ was originally observed are used to fit the supervised learning method $\mathbf{y}_{v,o}^{(e,k)} = f(\mathbf{X}_{-v,o}^{(e,k)})$ and (2) all data points where $\mathbf{y}_v^{(e,k)}$ was missing $\mathbf{y}_{v,m}^{(e,k)}$ are predicted from the fitted function and to overwrite the former estimates: ~~$\mathbf{y}_{v,m}^{(e,k),updated} = f(\mathbf{X}_{-v,m}^{(e,k)})$~~ $\hat{\mathbf{y}}_{v,m}^{(e,k)} = f(\mathbf{X}_{-v,m}^{(e,k)})$.

and Algorithm 1:

Create an updated estimate with the fitted regression model ~~$\mathbf{y}_{v,m}^{(e,k),updated} = f(\mathbf{X}_{-v,m}^{(e,k)})$~~ $\hat{\mathbf{y}}_{v,m}^{(e,k)} = f(\mathbf{X}_{-v,m}^{(e,k)})$.
Replace $\mathbf{y}_{v,m}^{(e,k)}$ with the new updated ~~$\mathbf{y}_{v,m}^{(e,k),updated}$~~ $\hat{\mathbf{y}}_{v,m}^{(e,k)}$ in $\mathbf{X}^{(e,k)}$.

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

AR: *The meaning of the subscript m is not the same in line 244 and line 217. We thank the reviewer for pointing out this inconsistency in our notation. In line 217, we are referring to the number of embedded features, in line 244 we refer to the part of the datapoints where the predictor variable is missing. We have changed the notation such that n_v is the number of variables (previously n) and n_f is the number of features (previously m) and m solely refers to the part of the datapoints where the predictor variable is missing. This changes the caption of Algorithm 1:*

Pseudo-code algorithm of the CLIMFILL clustering and learning step (step 3 and 4), where E is the number of epochs, K is the number of clusters, ~~n~~ n_v is the number of variables and ~~m~~ n_f the number of features. $\mathbf{X}_{-v}$ refers to the data table with all variables except v.

and line 1 in Algorithm 1:

$\mathbf{X}$ is a matrix containing all variables and features as ~~$n + m$~~ $n_v + n_f$ columns and all data points as rows.

RC: ***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.***

AR: *This is a very valid point. Indeed using the presented framework for interpolating sparse in-situ measurements is not immediately straight forward and some design choices would need to be made which are outside of the scope of this paper. The sections are therefore removed from the text.*

~~We note however that the framework is naturally extendable to include more satellite observations, and in situ observations that can be treated as a very sparse gridded product.~~

~~in situ data could possibly be included as well if treated as a very sparsely gridded data where the area of representation for the point measurement is accessed (see e.g. Nicolai-Shaw et al., 2017).~~

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

AR: *We thank Anonymous Referee #1 for this helpful feedback. As a response to this comment and a related comment from Referee #2 (Rene Orth), we have created symmetric embedded features for both the future and the past and systematically explore the impact of different versions of feature selection. For details, please refer to the detailed reply to the comment from Referee #2 below.*

RC: ***305: the point is filled***

AR: *This paragraph was rewritten due to the new interpolation method. This comment is therefore not applicable anymore.*

RC: ***323: are scalable***

AR: *The wording is changed accordingly.*

Random Forests have have favorable properties for gap filling applications: they can handle mixed types of data, are ~~scale-able~~ scalable to large amounts of data and non-parametric, i.e. adaptive to linear and non-linear relationships (Tang and Ishwaran, 2017).

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

AR: *Quantitative measures of the recovery of the original distribution are included by using the Jensen Shannon divergence as proposed by the reviewer above. Please see Figure 6 and the accompanying text.*

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

AR: *The caption of Figure 8 is changed as follows:*

Comparison of ~~(a)~~ artificial (a) random and (b) swaths-only missingness and (c) missingness in the real data in example snapshot of ERA5 ground temperature on 1st of August ~~2003 with.~~ 2003.

RC: ***Legend of Figure 8: "In swaths-only...": incomplete sentence***

AR: *The caption of Figure 8 is changed as follows:*

In swaths-only missingness we create long ellipses centered around the equator to simulate characteristic satellite swath missingness patterns. Note that the two missingness patterns are not exactly the same for each day and variable to allow for mutual learning.

RC: ***361: Leads to...: incomplete sentence.***

AR: *This paragraph was rewritten due to the new distance measure. This comment is therefore not applicable anymore.*

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

AR: *Thank you for this comment. As a reaction to this comment and the one of Referee #2 regarding bias correction, we have decided to remove the mentioning of bias correction in the manuscript. It is not straight-forward to apply and this would be out of the scope of this manuscript.*

> ~~Introducing an additional bias correction step could help alleviate these problems.~~

RC: ***468: recovering the physical***

AR: *The wording is changed accordingly.*

> Since the framework is targeted at recovering ~~the~~ the physical dependence structure across variables, the improvement in univariate measures like correlation and bias tend to be improved at a smaller scale than the multivariate dependence structure.

RC: ***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).***

AR: *This part of the text has been removed, therefore this comment is not applicable anymore.*

RC: ***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).***

AR: *The joint distributions are normalised. However, we decided to not plot the precipitation in log axis because then the comparison between the distributions for no rain would not be possible.*

4. Rene Orth, Referee #2

RC: ***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.

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

RC: ***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.***

AR: *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:*

> The interpolation step creates initial estimates based on the spatial or spatiotemporal context of the gap using interpolation. Following the approach of Haylock et al. (2008), the data is first divided into monthly climatology maps and anomalies. The climatology maps are gap-filled using thin-plate-spline interpolation to represent the spatial trends in the data. Subsequently, the daily anomalies from the monthly climatology are gap-filled using kriging. In contrast to the E-OBS dataset created in (Haylock et al., 2008) from in-situ observations, satellite data has a much larger number of observed values, making a direct implementation of this approach computationally infeasible. For the interpolation of the monthly climatology maps we there- fore restrict the thin-plate-spline interpolation to the 50 closest neighbors of each point. The interpolation of the daily anomalies follows Das et al. (2018), who suggest reducing complexity of kriging/Gaussian Process regression by repeated interpolations on random sub-samples of all available data points and averaging the resulting estimates. In particular, the missing values in the anomalies are estimated by randomly selecting 1000 observed points per month over which the interpolation is calculated. This is repeated five times and the mean of all interpolations for each missing point is taken as the gap-fill estimate. Finally, monthly maps and anomalies are summed up to form the inital gap-fill estimate from step 1.

RC: ***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.***

AR: *We thank Rene Orth for this helpful feedback. As for the selection of employed variables, please see our reply to the next comment. As for the possible choices, we have adapted the manuscript to give CLIMFILL users more guidance on how to apply the framework: We acknowledge that testing all the possible choices in Table 2 and provide recommendations in which settings they should be used is out of scope for this work. We therefore have removed the Table 2. To nevertheless make the framework easily applicable to users, we have detailed the hyper-parameter choices and their specific reasoning in Appendix Table A3.*

We believe this way with the settings described in this manuscript CLIMFILL is easily adaptable to the individual gap-filling needs of users.

RC: ***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.***

AR: *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 have now added time lags and windows such that they are symmetric for past and future. Additionally, we have tried 3 flavors of feature selection which we discuss in the first part of the results and then settle on the best performing feature subset for the rest of the results. The 3 flavours are:*

- only the four variables that have missing values
- the four variables plus all embedded features
- the four variables plus all embedded features plus all constant maps

Please note the respective changes in the text:

> The above procedure thus results in a set of 34 features: The four variables, the six embedded features of each of the four variables, totalling in 24 embedded features, the six maps and latitude, longitude and time information. All data are standardized to have zero mean and a standard deviation of one. We perform feature selection experiments (only the four variables, all embedded features, all embedded and constant features) to find the most descriptive subset of these 34 features, which we then use for computing the results.

RC: ***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.***

AR: *Thank you for pointing this out. We have done our best to remove or introduce data science language where possible, starting with your specific suggestions below.*

RC: ***line 2: estimates for what?***

AR: *The sentence has been changed to make this more clear:*

> Their abundance and often complex patterns can be a barrier for combining different observational datasets and may cause biased estimates of derived statistics .

RC: ***line 5: remove "up"***

AR: *the wording is changed accordingly.*

> Here we propose CLIMFILL (CLIMate data gap-FILL), a multivariate gap-filling procedure that builds ~~up~~ upon simple interpolation by additionally applying a statistical imputation method which is designed to account for dependence across variables.

RC: ***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***

AR: *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. In other words, if all variables have gaps at the same time and if this period is longer, the final gap fill estimate will be the gap fill estimate of the first step, the spatial interpolation, which is a state-of-the-art interpolation method.*

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

AR: *The text has changed, therefore this comment is not applicable anymore.*

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

AR: *We thank the reviewer for pointing us towards this very relevant literature and have included them in our introduction:*

> Several data products gap-fill one or more observations to a spatially or temporally complete data sets using auxiliary vari- ables (Huffmann et al., 2019; Brocca et al., 2014) or estimate variables that are only observed through sparse station networks 55 through statistical up-scaling (O. and Orth, 2021; Zhang et al., 2021; Ghiggi et al., 2019; Jung et al., 2019; Martens et al., 2017; Gudmundsson and Seneviratne, 2015; Jung et al., 2011, 2009).

RC: ***line 46: please clarify "scale somewhere between"***

AR: *The text has changed, therefore this comment is not applicable anymore.*

RC: ***line 84: please clarify "difficult observational record"***

AR: *We have changed the wording from "difficult" to "fragmented" in coherence with line 28, 43 and 51 of the originally submitted manuscript*

> For example, in the state-of-the-art atmospheric reanalysis product ERA-5 the ~~difficult~~ fragmented observational record of soil moisture is used only sparsely (Hersbach et al., 2020), although the added value of assimilating remote sensing soil moisture has been shown for weather forecast models (Zhan et al., 2016) and flood forecasting (Brocca et al., 2014; Sahoo et al., 2013).

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

AR: *A good point. We have removed the sentence "All these three missingness patterns can be observed in earth observation data:" such that consistency is ensured.*

~~. All these three missingness patterns can be observed in Earth observation data~~

RC: ***line 151: this is unclear, please rephrase***

AR: *The text has changed, therefore this comment is not applicable anymore.*

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

AR: *The text has changed, therefore this comment is not applicable anymore.*

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

AR: *Table 2 has been omitted. This comment is therefore no longer applicable.*

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

AR: *Table 2 has been omitted. This comment is therefore no longer applicable.*

RC: ***line 170: remove "on"***

AR: *The wording is changed accordingly.*

Finally, Sect. 4 discusses the results and provides a conclusion and an outlook ~~on~~ for possible future work.

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

AR: *The section title is changed to:*

CLIMFILL v1: A Generalised Framework for Infilling Missing Values in Multivariate ~~spatio-temporal geoscientific data~~Spatio-Temporal Geoscientific Data

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

AR: *The sentence is changed to:*

We aim for a multivariate gap-filling framework that exploits the ~~highly structured nature~~spatial, temporal and cross-variable dependence structure of Earth system observations to produce estimates for missing values.

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

AR: *The sentence is changed to:*

> Overview on the structure of the gap-filling framework. The framework is divided into ~~three~~ four steps.

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

AR: *We remove the explanation of the abbreviation here:*

> The framework CLIMFILL ~~(CLIMate data gap-FILL)~~ works mutually, i.e. information available in each of the variables is used for filling the gaps of all the other variables.

and insert it in the abstract at the first occurrence:

> Here we propose CLIMFILL (CLIMate data gap-FILL), a multivariate gap-filling procedure that builds ~~up~~ upon simple interpolation by additionally applying a statistical imputation method which is designed to account for dependence across variables.

RC: ***line 181: please clarify "correlation structure"***

AR: *We replaced "correlation structure" with "dependence structure" to ensure consistency across the text. (for example, lines 189, 366, 450,...)*

> With this design we implicitly assume that if one variable is not observed at a certain space-time point, a subset of the other variables might be observed and can reconstruct the missing value while conserving the ~~correlation~~ dependence structure among all variables.

RC: ***lines 203, 311: please clarify "constant"***

AR: *We replace the word "constant" with "time-indepentent in the first instance, and remove the word "constant" from "constant maps" in the second instance because "maps" already implies the fact that these features do not change with time.*

> For example, gap-free ~~constant maps of~~ time-independent maps describing properties of the land surface such as topography or land cover can be included.

> ~~Constant maps~~ Maps of altitude, topographic complexity, land cover class and land cover height from ERA5 as well as latitude, longitude and time are added to the list of features and copied for each time step.

RC: ***line 216: quotation marks not needed***

AR: *The sentence is changed accordingly.*

> Earth observations often inform about time dependent processes like seasonal effects, weather persistence or soil moisture memory effects that act from daily to monthly or subseasonal time scales (Nicolai-Shaw et al., 2016).

RC: ***lines 229: please clarify "stabilising the results"***

AR: *Since the Algorithm has been changed such that only one clustering is done, this paragraph is removed.*

> ~~For stabilising the results and to reduce 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 $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(E)}$.~~

RC: ***line 231: please clarify "terminal clusters"***

> ~~For stabilising the results and to reduce 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 $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(E)}$.~~

AR: *Since the Algorithm has been changed such that only one clustering is done, this paragraph is removed.*

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

AR: *The wording is changed accordingly.*

> (1) all data points where $\mathbf{y}_v^{(e,k)}$ was originally observed are used to fit the supervised learning method $\mathbf{y}_{v,o}^{(e,k)} = f(\mathbf{X}_{-v,o}^{(e,k)})$ and (2) all data points where $\mathbf{y}_v^{(e,k)}$ was missing $\mathbf{y}_{v,m}^{(e,k)}$ are predicted from the fitted function ~~and~~ to overwrite the former estimates:

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

AR: *We replace "weights" with "model parameters" for clarification.*

> Note that the framework is set up such that each cluster applies the same supervised learning method but learns different ~~weights~~ model parameters.

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

AR: *The text has changed, therefore this comment is not applicable anymore.*

RC: ***line 272: How are deserts defined and detected?***

AR: *We add the definition for deserts for clarification:*

> Permanently glaciated areas and deserts (defined as areas with less 50 mm average yearly precipitation in the years 2003-2012) are masked.

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

AR: *This paragraph has been rewritten as a response to comments on feature engineering above. This comment is therefore not applicable anymore.*

RC: ***line 314: please clarify "non-normality"***

AR: *the log-scaling of the preciptation has been removed, therefore this sentence has been removed.*

> ~~Furthermore, precipitation is divided into a log-scaled precipitation-amount variable and a binary precipitation-event variable to treat its inherent non-normality.~~

RC: ***line 316: How does this add up to 34?***

AR: *As a response to comments above, the number of features has been changed. The text now explains the number of features here:*

> The above procedure thus results in a set of 34 features: The four variables, the six embedded features of each of the four variables, totalling in 24 embedded features, the six maps and latitude, longitude and time information.

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

AR: *the clustering algorithm has been changed, therefore this comment is not applicable anymore.*

RC: ***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.***

AR: *We have changed the cross-validation procedure, which is now run on the original resolution. Therefore this comment is not applicable anymore.*

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

AR: *We have changed the cross-validation procedure, which is now run on the original resolution. Therefore this comment is not applicable anymore.*

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

AR: *Thanks for spotting this typo. We change it to:*

> Map of B-distance of univariate interpolation (a) and CLIMFILL-RF (b) as well as B-distance per land cover type (c) and altitude (d) for interpolation gap-fill and ~~CLIMPUTE-RF~~ CLIMFILL gap-fill in real missingness case.

RC: ***line 351: please clarify "det"***

AR: *The B-distance has been replaced, therefore this paragraph has been rewritten and this comment is not applicable anymore.*

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

AR: *Thank you. We have changed the caption to incorporate the comments:*

> Comparison of ~~(a)~~ artificial (a) random and (b) swaths-only missingness and (c) missingness in the real data in example snapshot of ERA5 ground temperature on 1st of August ~~2003 with.~~ 2003. Random missingness was created by randomly sampling without replacement from the pool of all gridpoints on land at all timesteps in the desired fraction of missing values. In swaths-only missingness we create long ellipses centered around the equator to simulate characteristic satellite swath missingness patterns. Note that the two missingness patterns are not exactly the same for each day and variable to allow for mutual learning.

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

AR: *The B-distance has been replaced, therefore this paragraph has been rewritten and this comment is not applicable anymore.*

RC: ***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.***

AR: *We thank the reviewer very much for this feedback.*

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

AR: *Thank you for spotting this. We have changed the sentence to:*

> Median performance of gap-filling with CLIMFILL-RF on different missingness patterns and fractions of missingness expressed in three metrics: pearson correlation ~~,~~ and root mean square error (RMSE) ~~and B-distance (for more detail see text)~~ per variable

RC: ***line 373: How exactly are the satellite swaths imitated?***

AR: *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.

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

AR: *Thank you for this comment. As a reaction to this comment and the one of Referee #1 regarding bias correction, we have decided to remove the mentioning of bias correction in the manuscript. It is not straight-forward to apply and this would be out of the scope of this manuscript.*

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

AR: *The text has changed, therefore this comment is not applicable anymore.*

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

AR: *We have enlarged the Figure.*

RC: ***Figure 4: The months axis should not go to 12.5***

AR: *The figure is changed accordingly.*