Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2021-164-RC1, 2021

Comment on gmd-2021-164

Anonymous Referee #1

Referee comment on "CLIMFILL v0.9: a framework for intelligently gap filling Earth observations" by Verena Bessenbacher et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-164-RC1, 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.

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.

- 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.

- 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.

- 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?

- 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.

- 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).

- 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.

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.

- 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.
- 139: are well suited
- 153: provided by other variables
- Table 2, caption: it is not clear to me what is meant by "method class" in this table
- 170: outlook for possible future work
- Caption of figure 2: the framework is divides into four steps
- 203: constant maps describing properties
- 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.
- 217: built from variables
- 216-218: This sentence seems to refer to the data format in the specific implementation described. Probably not needed in this methodological description.
- 244: I understand that the subscript "updated" stands for estimated value. A more common notation would be to use a hat.
- 244: the meaning of subscript m is unclear to me. Is it the same as in l. 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.
- Figure 5: why id the temporal window smaller in the future period than in the past?
- 305: the point is filled
- 323: are scalable
- 339: what is the criterion allowing to state that the shape of the distribution is well recovered?
- Legend of Figure 8: a) is described twice in the legend and c) is missing.
- Legend of Figure 8: "In swaths-only…": incomplete sentence
- 361: Leads to…: incomplete sentence.
- 401: This last sentence is intriguing, especially during the presentation of results. It could be expanded upon in the discussion.
- 468: recovering the physical
- 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).
- 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).