Reply to the questions and comments of Reviewer #2 for the MS "Technical note: Mind the gap – benchmarking of various imputation approaches for precipitation stable isotope time series" by Kern and Hatvani.

We would sincerely like to thank the reviewer for his/her efforts and constructive comments, all of which have been of great assistance in preparing the improved version of the manuscript. The whole manuscript has been revised. The reviewers pointed out week points in the text and it is our hope that these have now been strengthened. The questions of the reviewer are in bold.

General comments:

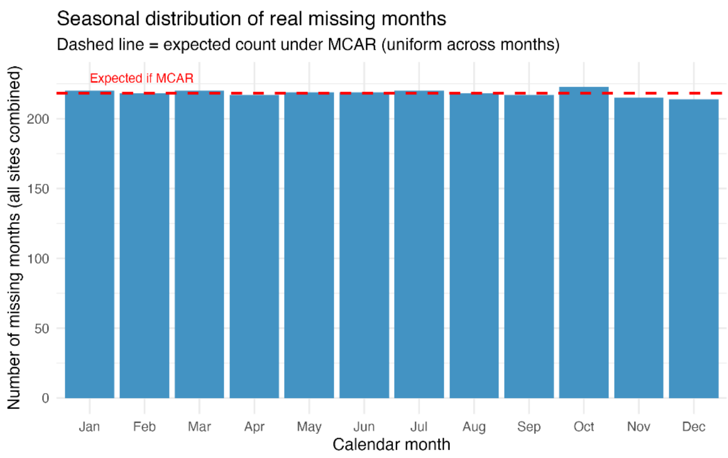
However, several aspects require substantial revision. In particular, the artificial missingness design, the reproducibility of SPbI, the justification of the 250 km threshold, the representativeness of the focus sites, and the interpretation of d-excess errors need to be strengthened.

Recommendation: Major revision.

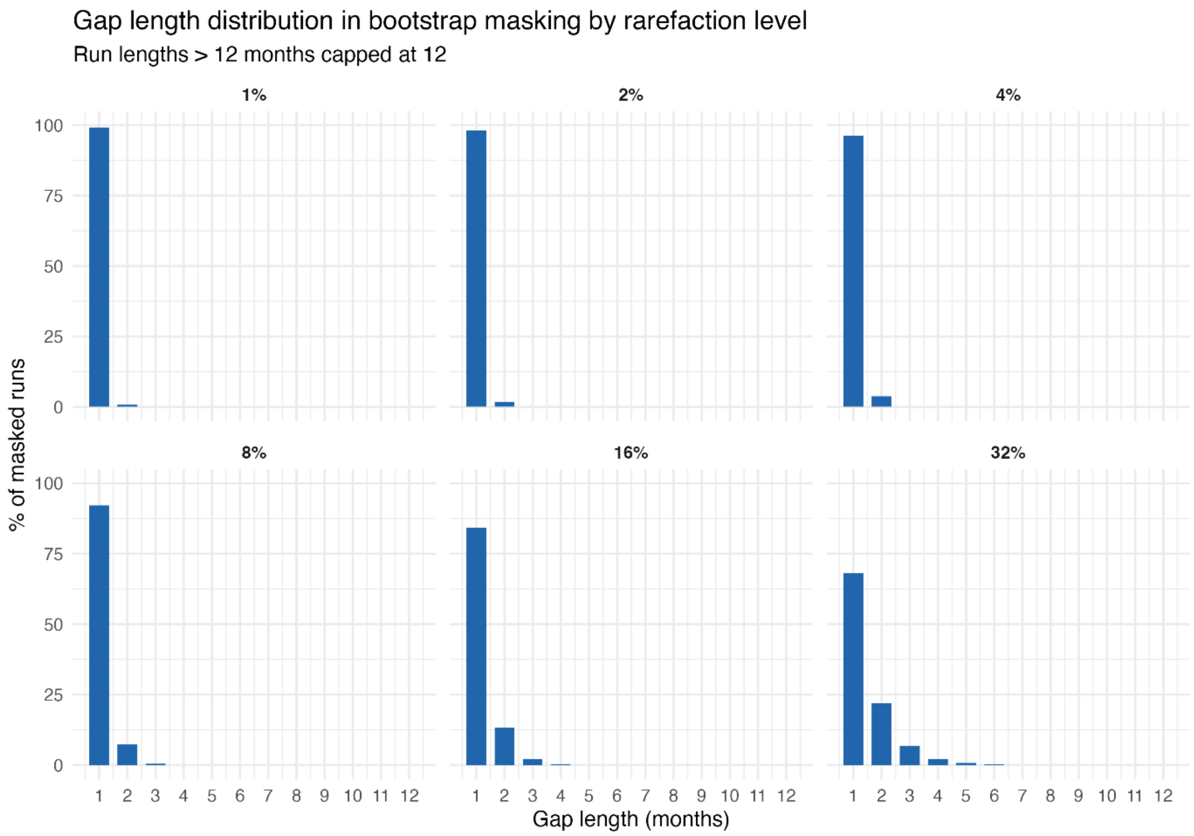
Major comments
1. The artificial missingness mechanism is too idealized
Location: Lines 15–17, 73–75, 105–108
The authors randomly masked monthly data at 1%, 2%, 4%, 8%, 16%, and 32%. This is useful for controlled benchmarking, but real gaps in precipitation isotope records are often not completely random. Missing data may cluster in winter, during low-precipitation months, after extreme events, or during instrument failure periods. The authors should discuss or test:

- **whether the masking scheme corresponds to MCAR;**
- **how seasonal or consecutive missingness may affect the results;**
- **whether 2–3 month or 6-month gap scenarios could be evaluated;**
- **whether missing values near isotopic extremes influence method performance.**

Thank you for your suggestions. In the revision multiple additional tests have been run, metrics calculated and comparisons performed with e.g. modelled data. The questions raised above formed one segment of the refined approach applied in the revised study.
We fully agree with the line of thought of the Reviewer, the presented MS deals with a masking MCAR scheme. We ran some tests to look into the matter. As seen from the figure below the masked data correspond to each month mostly uniformly.



At low rarefaction levels (1–2%), virtually all artificial gaps are single months (~99%). Two-month gaps begin to appear at 4% masking (~3%) and grow to ~22% at 32%. At 32% rarefaction approximately one third of masked positions form multi-monthly runs, while at all lower masking levels single-month gaps strongly dominate the distribution."
These findings will be incorporated into the first paragraph of Sect. 2.2.1.



2. The SPbI method needs more detailed description
Location: Lines 128–134
SPbI is the central methodological contribution, but the current description is not sufficiently detailed for reproducibility. Please clarify:

- **the minimum number of neighbouring stations required;**
- **whether imputation is performed when only one neighbour is available;**

Thank you for the comment. No fixed minimum number of neighboring stations was imposed a priori. Imputation was performed whenever at least one neighboring station had an available observation for the target month; if no neighboring value was available, the imputation was recorded as unsuccessful/fallback.
The second sentence was taken from the last paragraph of Sect. 2.2.1, while the minimum number of stations required has now been added to the MS.

- **whether the regional mean is unweighted;**

The imputed value was then calculated as the unweighted arithmetic mean of the available elevation-corrected neighboring observations. The text has been extended in the last paragraph of Sect. 2.2.1 stating: unweighted arithmetic means were derived.

- **why distance weighting or correlation weighting was not considered;**
- **whether the method is applicable across different moisture-source or climatic regimes.**

Thank you for the two questions above. We agree that distance/correlation weighting and testing across different climatic or moisture-source regimes are valuable future directions. However, these extensions would require additional methodological development and rigorously processed datasets from other regions and are therefore beyond the scope of this technical note.

- **whether the isotope lapse rates are seasonally and spatially constant;**

The Isotopic lapse rate is an adjustable field, even seasonally or regionally in the IsoQC app https://sapps.geochem.hu/isoqc/

We used the regional default as stated in the MS in the last paragraph of Sect. 2.2.1.

I recommend adding a formula, pseudo-code, or workflow diagram for SPbI.
The whole code of the procedure presented in the MS is documented here https://github.com/istvan60/SPbI as declared in the Code availability section in the original version of the MS. Note, the GitHub repo contains the updated and revised version of the codes. A new Readme file provides an overview https://github.com/istvan60/SPbI/blob/main/README.md

3. The 250 km threshold requires stronger quantitative support
Location: Lines 25–28, 157–158, 230–234, 251–252; Fig. 4
The manuscript recommends SPbI when neighbouring data are available within 250 km. This is an important practical conclusion, but the threshold appears to be based mainly on visual interpretation of Fig. 4. Please provide:

- **a clear criterion for defining the 250 km threshold;**
- **confidence intervals for MAD/RMSE in each distance band;**
- **statistical tests among distance bands;**

Thank you for the comment, we would like to provide an answer to multiple comments together: To evaluate the distance range over which SPbI remained advantageous over non-spatial imputation methods, SPbI was recalculated using non-overlapping 50 km distance bands from 0–50 km to 350–400 km. In the previous version of the results there was an error in the calculations, which has now been corrected.
For each isotope, masking percentage, bootstrap replicate and distance band, imputation error was summarized using MAD and RMSE. To compare SPbI with the Linear and Sinusoidal benchmark methods, errors were paired using identical station–masking–bootstrap combinations. For each comparison, the paired error difference was calculated as the benchmark error minus the SPbI error. Confidence intervals were calculated for the mean paired difference, and one-sided Wilcoxon signed-rank tests were used to test whether SPbI errors were significantly lower than benchmark errors. P-values were adjusted using the Holm correction. A distance band was considered to support SPbI only when SPbI was significantly better than the relevant benchmark method according to the selected criterion. The practical distance threshold was defined as the outer limit of the most distant non-overlapping band in which SPbI remained significantly better than both Linear and Sinusoidal imputation. https://github.com/istvan60/SPbI/blob/main/STAGE%202%20CI%20detection.R

Based on the new evaluation motivated by the Reviewer's questions the previously suggested 250 km distance was reduced to 200 km and will be changed in the MS in every relevant place. E.g. ". The benchmarking results recommend the use of SPbI where station data are available within 200 km distance, and the sinusoidal or linear approach otherwise."

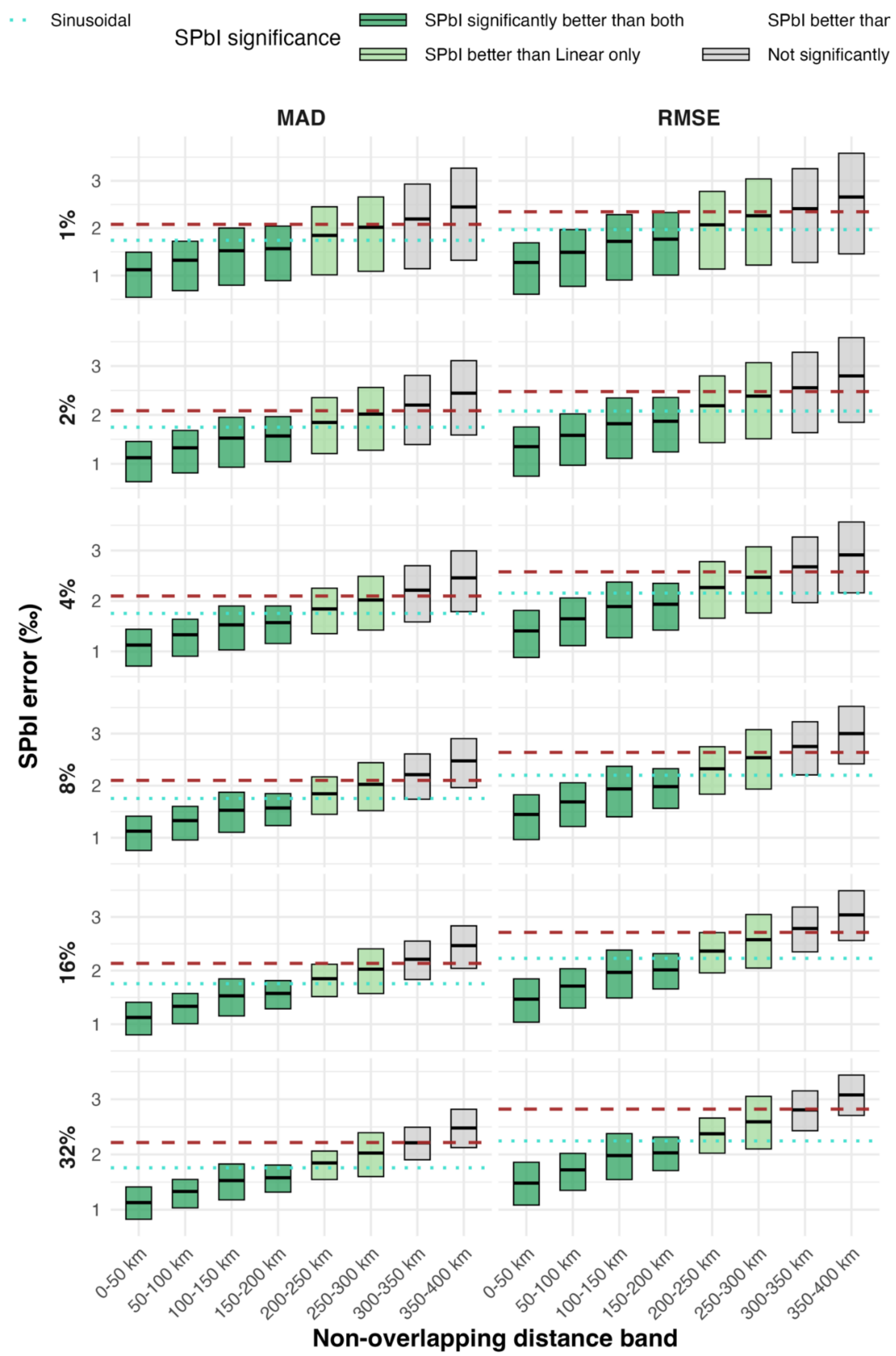


Fig. 3. (*draft*) Distance-dependent $\delta^{18}O$ imputation error of the spatial proximity-based imputation method (SPbI) compared with the Linear and Sinusoidal benchmark methods. SPbI errors were calculated for non-overlapping 50 km distance bands. Boxes show the interquartile range of SPbI errors, and the horizontal line within each box represents the mean error. Dashed and dotted horizontal lines indicate the mean errors of the Linear and Sinusoidal benchmark methods, respectively. Box colors indicate whether SPbI was significantly better than both benchmark methods, only one of them, or neither, based on paired comparisons using identical station–masking–bootstrap combinations. Results are shown separately for masking percentage (X%) and error metric.

 separate distance-dependent results for $\delta^{18}O$ and $\delta^{2}H$;
The counterpart-figure showing the results for $\delta^{2}H$ can be found in the OSM, Fig, S2, which was included in the original submission already. An updated figure will be provided in the revised version.

 a discussion of whether the threshold is region-specific.
The statement should be made more cautiously, as the 250 km threshold likely depends on station density, topography, and regional isotope spatial coherence.
Please see our answer to your question following your Q.#4.

4. Representativeness of the focus sites should be clarified
Location: Lines 79–94; Fig. 1
Although 132 stations were used, the benchmarking relies on focus sites with at least seven years of uninterrupted records. The manuscript states that most focus sites are from Austria, with only two from Slovenia and one from Hungary. The exact number of focus sites is not clearly stated. Please add:
 the total number of focus sites;
As now stated in the last paragraph of Sect. 2.1, $n = 53$
 the number of focus sites by country;
It is now stated in the last paragraph of Sect. 2.1; 50 + 2 + 1.
 their elevation range;
From 2 to 3106 m asl for the full dataset and 117 to 2245 m asl for the focus sites.

 their spatial and climatic representativeness;
 potential bias caused by the dominance of Austrian stations.
The distribution of the focus sites is proportional to that of the full station network not only across the Köppen–Geiger zones (covering mostly the Cfa, Cfb, Dfa, and Dfb zones), but also along the elevation gradient. As shown by the inset histogram, the hypsographic distribution of the focus sites closely mirrors that of the complete assessed station network. Below 2500 m a.s.l., focus sites represent approximately 40% of the stations in each elevation band, indicating a balanced selection across the main hypsographic range of the study area.

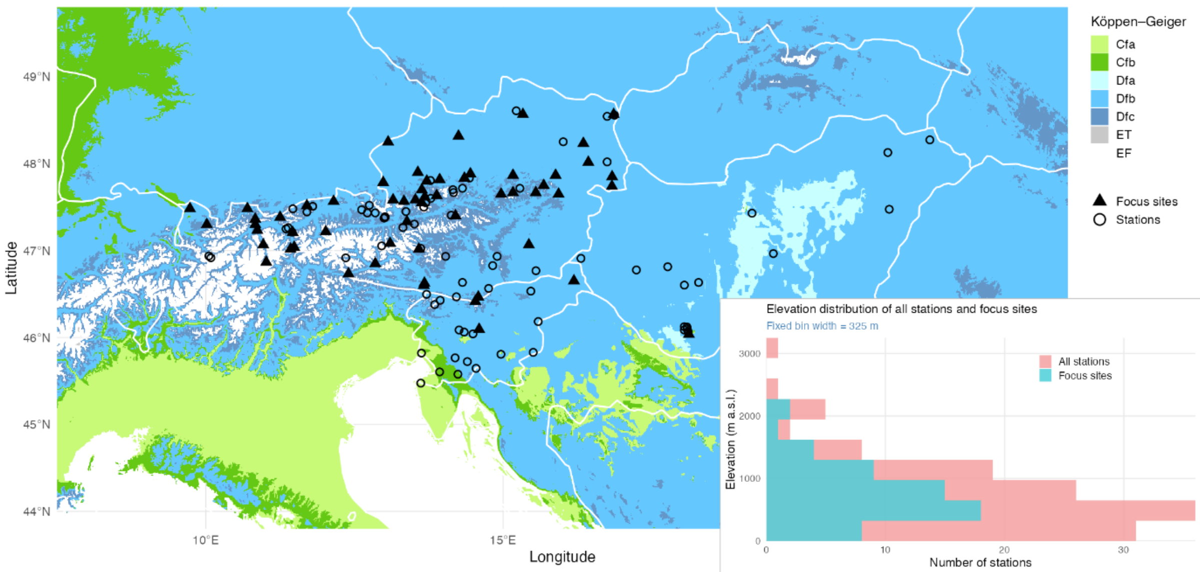
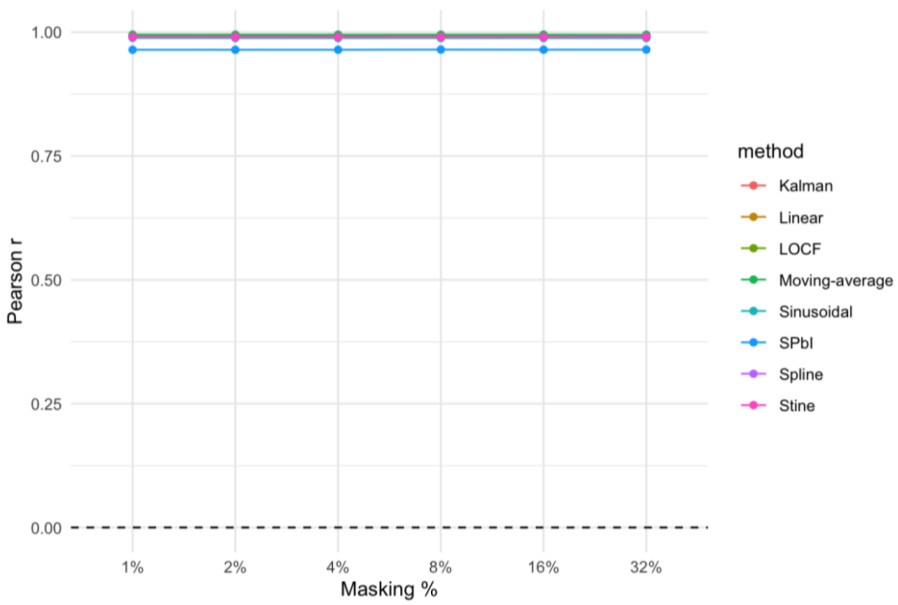


Fig. 1. Spatial distribution of the precipitation stable isotope monitoring sites across the study area. Circles denote the locations with δ_p data, while triangles mark the sites with at least 7 years uninterrupted monthly $\delta^{18}O$ and $\delta^{2}H$ observations between 1973 and 2022. Background colors indicate Köppen–Geiger climate classification zones (Beck et al., 2018). The inset histogram shows the hypsographic distribution of all stations (pink) and focus sites (teal) in 325 m elevation bands, demonstrating that the focus sites are representative of the full station network across the sampled elevation range.

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., Wood, E. F. 2018: *Present and future köppen-geiger climate classification maps at 1-km resolution*. **Scientific Data** 5-1. https://doi.org/10.1038/sdata.2018.214

5. This is important for assessing the transferability of the conclusions. The d-excess analysis should be deepened
Location: Lines 136–144, 192–216; Fig. 3
Including d-excess is a strength of the paper. However, the current interpretation is mostly descriptive. Please discuss:
how $\delta^{18}O$ and $\delta^{2}H$ errors propagate into d-excess;
whether errors in $\delta^{18}O$ and $\delta^{2}H$ are correlated;
Thank you for the question, errors are highly correlated, $r > 0.95$; see figure below.



whether Sinusoidal imputation damps d-excess seasonality;
Yes, it inherently does.

whether SPbI preserves anomalous d-excess events;
whether d-excess errors vary seasonally.
Additional metrics such as bias, correlation, or extreme-value preservation would strengthen the analysis.
Besides the original metrics, Lin's CCC and the slope of the imputed-on-observed regression were added to Fig. 3 to quantify, respectively, the agreement with the 1:1 relationship and the degree of amplitude retention in the imputed d-excess; see figure below. Damping of the seasonal loss of amplitude in the d-excess signal also affects data interpolation using the SPbI method; however, according to the results, among the methods that provide the most accurate estimates, SPbI is capable of producing estimates that suffer the least loss of amplitude. In other words, SPbI method is best able to create imputed records preserving the amplitude of the seasonal fluctuation as well as the magnitude of the extreme values with the highest fidelity i.e. (closest match with the actual values). This will be incorporated into the conclusions section of the MS.
Additional tests were also run to answer the question, the general observation is, d-excess imputation errors peak for all methods in Jan-Feb and reach a minimum in Jul-Sep. This reflects the pattern seen in the regression slopes annotated on the figure below, thus a separate figure is not provided to avoid repetition.

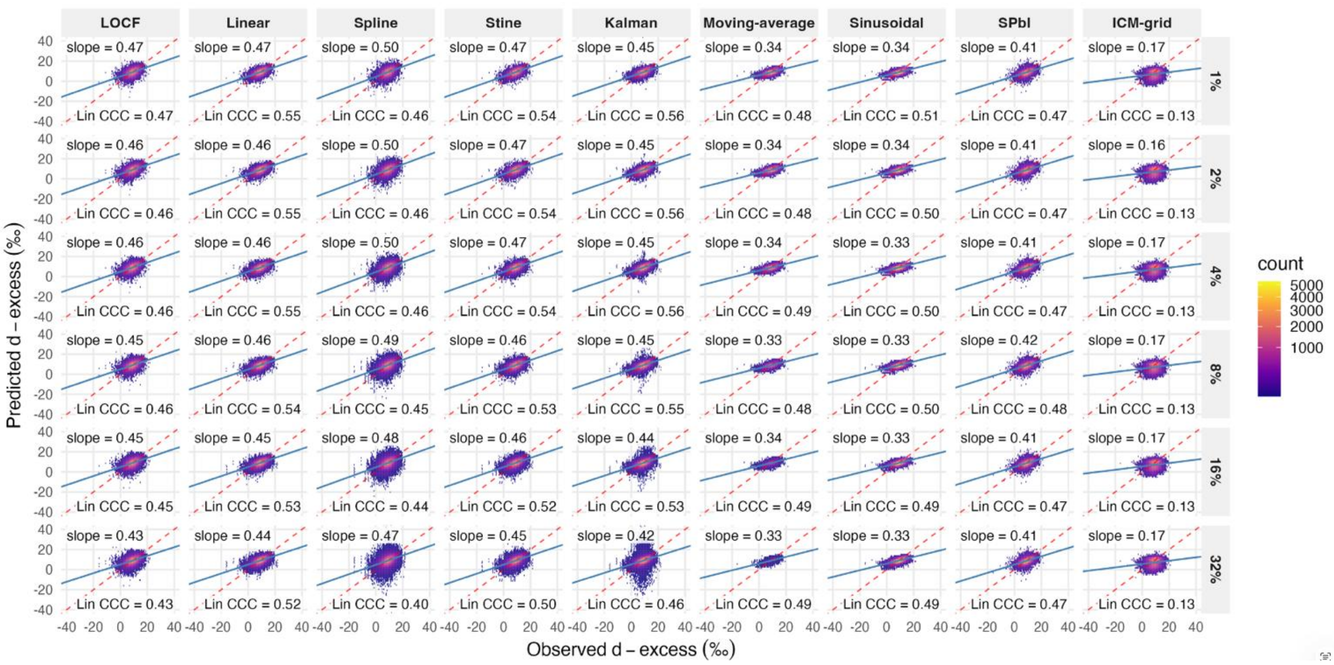


Fig. 3 (draft). Comparison of observed and imputed *d-excess* values across all masking percentages (*X* %). Color scale = point density. Hex-binning is used to visualize point density. The dashed red line indicates the 1:1 relationship, while the blue line shows the fitted relationship between observed and imputed values. The slope of this fitted line is shown in the upper-left corner of each panel. Lin's concordance correlation coefficient (Lin CCC), shown in the lower-right corner. Results were aggregated across stations and bootstrap replicates for each rarefaction level *X*. Predictions include the tested station-based imputation methods and the external ICM-grid monthly isotope product (Adachi and Yamanaka, 2026). ICM-grid was evaluated against the same withheld station-month observations

Adachi, I., & Yamanaka, T. (2026). *Global dynamic precipitation isoscapes over three-quarters of a century*. **Earth System Science Data Discussions**. https://doi.org/10.5194/essd-2026-52.
For the dataset see https://doi.org/10.5281/zenodo.17301045

6. The evaluation metrics are somewhat limited
Location: Lines 137–156
MAD, RMSE, and Bland–Altman plots are useful, but precipitation isotope time series also require evaluation of:
□ isotopic extremes;
□ δ^2H–$\delta^{18}O$ relationships;
□ local meteoric water line properties;

Besides the original metrics, Lin's CCC and the slope of the predicted vs. observed regression were added to Fig. 3 to quantify, respectively, the agreement with the 1:1 relationship and the degree of amplitude retention in the imputed d-excess. Please see the response above.

□ precipitation-weighted annual or seasonal means.
Since amplitude loss is discussed in Lines 180–183, a quantitative amplitude retention metric would be helpful.

We would like to refrain from the evaluation of precipitation-weighted annual or seasonal means, simply because the imputation performing better at the monthly scale will perform better at the annual and seasonal scales.

7. The choice of benchmark methods should be better justified
Location: Lines 109–123, 128–134

The manuscript compares eight methods, but does not explain why other relevant methods were excluded, such as regression models using meteorological covariates, kriging, regression kriging, Bayesian models, random forest, gradient boosting, isoscape-based imputation, or bivariate methods preserving δ^2H–δ^{18}O covariance. The authors do not need to implement all these methods, but they should clarify the scope and limitations of the benchmark.

Thank you for the comment. The benchmark was deliberately restricted to methods that reconstruct missing values from precipitation isotopic data, without the need for external predictors.

Regression models using meteorological covariates were not considered because our aim was to impute δ_p from the isotopic records alone. Introducing covariates such as temperature or precipitation amount would risk circular reasoning. When the imputed series is used for with the given environmental variable due to circular reasoning.

Kalman smoothing – which was in fact applied in the present study can be considered mathematically equivalent to optimal linear prediction under an implied temporal covariance structure, making it the functional analogue of temporal kriging within the tested set. Based on the suggestions of Reviewer #1 a model was incorporated – as you have read above – thus the ICM grid's isoscapes were considered.

The most straightforward approach in preserving δ^2H–δ^{18}O covariance is indeed imputation using the LMWL. However, as stated in the MS, this can only be done when at least one of the two primary parameters is available for a given month. In addition, according to a global assessment at least 48 months are required to calculate the robust LMWL (Putman et al. 2019), hence in case of observations spanning a rather short time period, the application of bivariate methods preserving δ^2H–δ^{18}O covariance risks imputation-bias, due to unreliable LMWL.

The present benchmark deliberately targets the case in which both δ^{18}O and δ^2H are missing simultaneously, where the covariance relationship cannot be exploited. Nevertheless, the question is interesting and should be explored in a follow-up study.

We will extend the end of the Introduction reflecting on the matter discussed above clarifying the scope further.

Putman A. L., Fiorella, R. P., Bowen, G. J., Cai, Z. (2019). A global perspective on local meteoric water lines: Meta-analytic insight into fundamental controls and practical constraints. *Water Resources Research*, 55, 6896–6910. https://doi.org/10.1029/2019WR025181

Minor comments

1. Line 12: "in wide range" should be "in a wide range". Accepted and corrected.

2. Lines 23–24: "has shown a tendency of reducing extremes" should be "showed a tendency to reduce extremes". Accepted and corrected.

3. Line 71: "would outperforms" should be "would outperform" or "that outperforms". Accepted and corrected. Has been changed to which outperforms…

4. Lines 83–85: Please format the d-excess equation consistently. Apologies, we do not see the inconsistency, the font and size and general formatting is in harmony with Eq. 2.

5. Lines 89–92: State how many focus sites remain after applying the 84-month criterion. Accepted and corrected.

6. Line 94: Discuss the imbalance among countries. It is now discussed in detail how many sites come from a given country and Fig. 1 has also been redrawn.

7. Lines 105–108: "~20o,000" appears to be a typo. Accepted and corrected.

8. Line 121: "by sine curve" should be "by a sine curve". Accepted and corrected.

9. Line 123: "A 12-month sinusoidal was fit" should be "A 12-month sinusoidal curve was fitted". Accepted and corrected.

10. Line 132: Use "km^{-1}" instead of "km-1". Accepted and corrected.

11. Lines 146–151: RMSE and MAD equations appear incorrectly rendered. This bit has been removed as suggested by Reviewer #1.

12. Line 162: "In the case of both δp" is unclear. Specify δ^{18}O and δ^2H. Accepted and corrected.

13. Lines 175–177: Consider reporting fallback rates in a table. We collected the fallback rates for the imputation methods. The suggested table will be added into the Supplementary material as Table S1 with a brief assessment.

14. Fig. 2 and Fig. 3: Some labels are difficult to read. The text size will be increased in the final revision.
15. Line 203: Use "Fig. 3" rather than "Fig 3." Accepted and corrected.
16. Lines 207–210: The sentence appears incomplete and should be checked. Due to the new figure, the sentence in question is already changed.
17. Fig. 4: The x-axis label "Radius" is misleading; "Distance band" would be clearer. Accepted and corrected as you can already see from the inserted fig. 4 above.
18. Line 249: Remove the duplicated period in "isotopes..". Accepted and corrected
19. Lines 255–259: Add punctuation between "sources" and "Umweltbundesamt". This was formatted using EndNote, but it will be checked at the final submission. Thank you for pointing that out.
20. Lines 272–273: The AI-use statement should be checked against journal policy. It was slightly extended, but met the journal standards in the first place already.

Yours sincerely,

Budapest, 30.06.2026 István Gábor Hatvani