LegacyClimate 1.0: A dataset of pollen-based climate reconstructions from 2594 Northern Hemisphere sites covering the last 30 ka and beyond

Response to comments of Anonymous Referee #1 (second round, 25.11.2022)

1. General comments

Reviewer comment: (1) *In general the authors have done a good job incorporating the comments raised in the previous round. However, I would recommend that more of the response to the review is included in the main text as I think this would make the reasoning for several analyses steps easier to follow.*

Response: We added more text of the previous response letter to the main part of the manuscript where appropriate to not too much defocus the text. We followed the reviewer's suggestion for a better integration of the LegacyAge 1.0 chronologies (Li et al., 2022) and added an ensemble of 1000 realizations of the age-models for each record to the database of reconstructions together with some explanations how users can assess chronological errors in addition to reconstruction errors (section 2.1). We described in more detail why we focused our reconstructions to temperature and precipitation (section 2.1) and revised the discussion to address the three different reconstruction methods that we used in the study (section 5.2). In addition, we extended our quality measures (section 2.2) and discussed the influence of human land use to the reconstruction results (section 4.6).

Reviewer comment: (2) *There are also some points that require additional attention. Most of these I already raised in the previous round of review and I think the manuscript would still benefit from some more discussion and guidance for the user of these data.*

Response: Thank you for the suggestions. However, some of the comments do not address the main purpose of the manuscript which are 1) presenting a data environment & R routine for pollen-based climate reconstructions, 2) presenting a dataset of climate reconstructions, 3) providing data that allow assessment of the reconstructions. Some comments rather address questions related to the potential usages of either of the three types of information provided in the manuscript. We provide some information on how the datasets can be used and will improve this text (see comments below), however, addressing all (or many) potential usages of the datasets in detail would blow up the text ending up with an entirely new manuscript reviewing "usage of pollen-based climate reconstruction". Accordingly, we try our best to, at the same time, address the reviewer comments but not defocus the manuscript.

2. Major issues

Reviewer comment: (1) *The choice of reconstruction methods remains a bit vague and could go beyond stating that these are the most commonly used methods. Surely the authors have good scientific reasons to use these. The argument that with providing the raw pollen data, everyone can make their own reconstruction seems to either downplay the importance of carefully evaluating such reconstructions or make the current manuscript superfluous. As such, I find that a bit of a slippery slope and I encourage the authors to clearly state why they chose these methods.*

Response: We a priori selected MAT and WA-PLS as main reconstruction methods because 1) they are most commonly used and the obtained results are thus comparable and the method-related biases are best understood and because 2) they represent two different families of reconstruction methods, i.e. an analogue-based and a regression-based reconstruction method. Our evaluation showed that despite the differences in the reconstruction method the overall results are very similar (Figure 11; i.e. the high correlation values when comparing the reconstruction methods). Accordingly, we assume that the provided results are suitable for most applications. However, as we provide the data and R environment for climate reconstruction, users can perform their own reconstructions for specific needs. In the discussion part we provide some information on future options on using the data with CREST representing a third family of bioproxy-based reconstructions (i.e. using taxa distribution ranges).

Reviewer comment: (2) *Towards the end of the revised manuscript the authors mention other reconstruction methods and that they could help to "explore a larger fraction of the "method uncertainty" space". Exactly how this could be done remains unclear and the authors should more clearly outline an approach on how the difference between the methods can be used to evaluate the reconstructions (see e.g. Kucera et al. 2005).*

Response: Thank you for your comment. We adopted the metrics proposed in Kucera et al. (2005) to the discussion on the comparison of different reconstructions in a multi-technique approach.

New text: Kucera et al. (2005) propose several metrics for a multi-technique approach to assess the uncertainty space: correlations between the residuals (observed minus reconstructed values) between pairs of techniques are used to investigate the similarity in the reconstructions among different techniques. The correlation between the residuals in seasonal reconstructions (e.g. summer and winter temperatures, summer and annual temperatures) can be used to investigate the degree of independence of different seasonal reconstructions. Error rate estimates (RMSEP) determined by cross validation of the calibration data sets and the leaving-one-out method can be used to compare the calibration of individual transfer function techniques, though it should be considered that error estimates may vary with the choice of the cross-validation procedure (Kucera et al., 2005).

Reviewer comment: (3) *The authors have carried out several analyses which can be used to evaluate the reconstructions and they should be complimented for this rigorous approach. They provide three different reconstructions, information on the transfer function performance (from the CCA), information about the analogue quality, on the significance of the reconstruction etc. and indicate that all this information can be used to assess the reconstructions. However, I miss clear guidelines of how this*

should be done, or what the authors think is the best approach. All these quality measures are provided but, in the end, not used (e.g. figure 10 contains all time series from a single reconstruction method), so what is the point exactly? I suggest that the authors provide instructions on how to use the data, e.g. indicate which lambda ratios should be omitted, at which analogue distance researchers should ignore the reconstruction, how to use the results significance test, how to interpret the difference between the methods, what to do with sites that show marked human influence, etc. Alternatively, they should provide a single reconstruction for each site (with uncertainty) that they think is best (with of course an explanation of the reason why).

Response: Reviews on how the criteria that we provide can be used and how results can be interpreted are provided in Juggins (2012) and in Chevalier (2019), among others. We think it is impossible to provide a clear guidance for selection of reconstructions. For example, a threshold for filtering of the reconstructions would strongly depend on the purpose of the study for which the user will use the climate reconstructions provided. Even more, likely no threshold at all will be used for many studies, but our assessment results can be used as quantitative input information for the likelihood that a reconstruction is reliable.

- Juggins, S.: Quantitative reconstructions in palaeolimnology: new paradigm or sick science?, Quaternary Science Reviews, 64, 20–32, https://doi.org/10.1016/j.quascirev.2012.12.014, 2013.

- Chevalier, M.: Enabling possibilities to quantify past climate from fossil assemblages at a global scale, Glob. Planet. Change, 175, 27–35, https://doi.org/10.1016/j.gloplacha.2019.01.016, 2019.

Reviewer comment: (4) *The authors explicitly state that the purpose of these reconstructions is to evaluate climate models, but without clear instructions it is unclear how these data can be used for this purpose. The data certainly cannot be used "out of the box" but require processing. This is fine of course and almost always the case with complex data like paleoclimate data, but it means that users need to be provided with clear guidelines and examples.*

Response: We agree that evaluating model outputs using proxy-based reconstructions is a complex task and strongly depends on the purpose of the evaluation. For example, the purpose of an evaluation can target (1) the mean or site-specific changes or it can target (2) evaluating relative or absolute values or it can target (3) spatial or temporal climate variability at specific scales. All these types of evaluation require a specific handling of the proxy-data. It is out of the focus of this manuscript to provide detailed guidance how the dataset should be handled as it strongly depends on the purpose of the model-proxy-comparison study. We clarified this in the new text.

Reviewer comment: (5) *Finally, the authors assess the influence of human land use on their reconstructions by looking at the proportion of certain pollen types in the time series. This seems a reasonable approach, as far as I can judge, but the influence of humans not only affects the time series,*

but also the pollen data set and the climatic variables used for calibration. Perhaps the calibration even more? This is alluded to on page 34, but not really explored. Is it not possible to remove sites with high human influence from the training set? Or are there other approaches, such as (perhaps an outrageous suggestion) calibrating using deeper/older samples assuming relatively stable late Holocene conditions?

Response: Thank you for this comment. It is not the main intention of the manuscript to set up new frameworks of climate reconstruction considering human impact. This would be a challenging task, deserving a study on its own. However, we agree that human impact may have impacted the reconstruction. Accordingly, we provided some information to assess this posteriori.

Reviewer comment: (6) *And a related question, to what degree are the climate variables averaged over the period between 1970 and 2000 representative of the conditions during deposition of the pollen?*

Response: Most of the samples in the modern dataset are collected between 1980 and 2010; mostly from peat, lake or soil surface (Top 2 cm) that integrate the signals of a surface sample over several years up to a few decades. So, using a 30-year mean from 1970-2000 seems to be appropriate.

Reviewer comment: (7) *I also imagine that human influence on the core top data also affects the identification of analogues and wonder if it could be the reason for the low analogue quality in Western Europe and the British Isles (page 35)?*

Response: This is well possible. It illustrates the problem particularly of analogue-based reconstructions methods. We addressed this by the following sentence:

New text: MAT is often recommended for large-scale studies, but it is highly sensitive to the quality of analogues (Chevalier et al. 2020). Low analogue situations can arise from two causes: climate conditions that differ strongly from today (e.g., the low atmospheric CO_2 concentration during the LGM; Jackson and Williams, 2004), or in regions with limited modern samples (e.g., extratropical Asia). Furthermore, growing human influence on the landscape since the Middle to Late Holocene especially in densely settled regions in Europe contributed to gaps within the potential bioclimatic space of taxa and probably also led to extinction events, especially for disturbance-dependent taxa (Zanon et al., 2018). We report the analogue distance for each sample to help identify such situations. From our assessments, we revealed that analogues quality is overall rather good at least for the Holocene and except for Western Europe in particularly the British Isles (Fig. 4).

3. Specific comments

Reviewer comment: (1) *Title: …"and beyond." Why not state the exact duration of the time covered by the synthesis?*

Response: Because the longest record covers the last 300.000 but this period is only covered by few records. For convenience we rounded the exact duration to 30 ka.

Reviewer comment: (2) *Page 4: Holocene conundrum. Is there still a conundrum, or is there mechanistic value in comparing global mean temperatures. The debate has progressed since Liu et al. (Osman et al. 2021; Cartapanis et al. 2022; Kaufman et al. 2020).*

Response: Thank you for your contribution. We address the progression of the conundrum debate in the discussion (section 5.4).

New text: The debate has since progressed and hints to discrepancies in data-model comparisons due to spatiotemporal dynamics related to heterogeneous responses to climate forcing and feedbacks (i.e. the timing of a Holocene Thermal Maximum in the Northern Hemisphere extra-tropics between reconstructions from continental and from marine proxy records; Cartapanis et al., 2022) and sometimes poor spatial averaging due to unevenly distributed proxies. Proxy-only reconstructions often rely on latitudinal binning and weighting, which makes this approach particularly sensitive to latitudinal bands that contain only sparse spatial coverage and thus do not represent a true global average (Osman et al., 2021). Those spatiotemporal dynamics should be considered in data-model comparison.

Reviewer comment: (3) *Page 4: "Pollen data are the only land-derived proxy … Quaternary period". I suggest to be a bit more specific and tone this down. There is no a priori reason why one could not evaluate a model based on a single or a few observations. High spatio-temporal coverage allows one to investigate different aspects, but I don't think there is a reason to be offensive to other proxy types.*

Response: we rephrased the sentence.

New text: Among land-derived proxy data pollen are particularly suitable for temporarily and spatially high-resolution evaluation of climate model simulations of the late Quaternary period.

Reviewer comment: (4) *Page 4-5: "MAT and WA-PLS rely on extensive collections of modern training data." Does that not hold for any transfer-function based reconstruction?*

Response: Methods that make use of the modern distribution of taxa e.g. climate range method or CREST do not use modern spectra, but make use of modern taxa distributions.

Reviewer comment: (5) *Page 6: PANGAEA is a data publisher, or repository, not a data base.*

Response: Thank you, we changed the phrasing in the text.

Reviewer comment: (6) *Page 6: "We restricted the analyses to the 70 most common taxa to reduce computational power…" this seems odd wording, why would you like to reduce computational power. Is demand meant? More importantly, how was this tested? Please provide details.*

Response: Yes, demand was meant. The modern pollen training data from North America contains 70 taxa. To keep it comparable we applied this to the datasets from Europe and Asia.

New text: We restricted the analyses to the 70 most common taxa on each continent to reduce computational power after making sure that higher taxa number would not substantially improve model statistics in climate reconstructions. The number of taxa is limited by the modern training dataset from North America, which contains 70 taxa after applying our taxa harmonization routine (see details in

Herzschuh et al., 2022c) To keep the taxa comparable for the reconstructions, we restricted the number of taxa in the fossil datasets.

Reviewer comment: (7) *Page 8: "... to measure how well the target environmental variable is strongly related…" delete strongly.*
Response: We deleted the word "strongly".

Reviewer comment: (8) *Page 8: please provide details of what the minDC function actually does. What are these probability thresholds and what are they based on?*
Response: the minDC function calculates the smallest dissimilarity between a record sample and the training set samples. The probability thresholds indicate the analog quality: a minimum dissimilarity <1% is considered to have very good analogs, <2.5% good analogs and a minimum dissimilarity <5% to have poor analogs.
New text: To infer the analogue quality as an indicator of no-analogue situations we calculated the minimum dissimilarity (squared chord distance) between modern pollen assemblages and fossil pollen assemblages with probability thresholds of 1% (indicating very good analogs), 2.5% (good analogs) and 5% (poor analogs) using the *minDC* function from the analogue R-package (version 0.17-6, Simpson et al., 2021).

Reviewer comment: (9) *Page 9: significance test. I asked before, how were the random environmental fields generated? Was it by simple permutation or was spatial autocorrelation taken into account. If the environment is spatially auto- correlated, which I imagine is likely the case, a red-noise null should be used instead of the default white noise null.*
Response: We used the *randomTF*-function with the default setting, which uses a white-noise null distribution and does not take spatial autocorrelation into account. However, we have indeed to deal with spatially autocorrelated environments. We therefore prepared surrogate fields for the climate variables we want to reconstruct. We implemented those surrogate climatology fields via the "autosim" argument into the *randomTF* function as the red noise null distribution and recalculated the significance values for our reconstructions. We will re-upload an updated version of the significance data set to PANGAEA and an updated version of the R code to Zenodo.
New text: A statistical significance test (Telford and Birks, 2011) was applied using the *randomTF* function in the *palaeoSig* R-package (version 2.0-3, Telford, 2019). In this test, the proportion of variance in the fossil pollen data explained by the reconstructed environmental variable is estimated from redundancy analysis (RDA) and tested against a null distribution generated by replacing the modern training dataset with randomly generated surrogate fields. The surrogate fields were simulated to have realistic spatial autocorrelation by fitting variograms to the WorldClim temperature and precipitation data; 1000-member ensembles were simulated for each variable.

Reviewer comment: (10) *Page 9: "In addition, we calculated the correlation between WA-PLS reconstruction of …" please provide a rationale for this analysis that helps to understand why this has been done and how the results should be interpreted.*
Response: the analysis was done in order to assess potential biases in the dataset with regard to human influence. Correlations, both positive or negative, would indicate potential biases in the reconstruction. We reported this in section 4.6, i.e. that correlations indicate that too low temperatures become reconstructed (see also reviewer comment 19).

Reviewer comment: (11) *Page 9: "To ease data handling, the dataset files are separated into…" I disagree that splitting a single file into more than one eases data handling. I would argue that it is easier to filter a data set than combine different ones as I don't need to load multiple files.*
Response: The files are organized in the same way as the LegacyPollen 1.0 datasets (https://doi.org/10.1594/PANGAEA.929773) to provide consistency between the data products and to keep the datasets to a reasonable size. We think that it makes it easier for users to load only a smaller data set if they want to work on a continental or regional scale, especially in combination with the LegacyPollen 1.0 data product.

Reviewer comment: (12) *Page 11: "Minimum dissimilarities between modern pollen assemblages and fossil pollen assemblages for each site for MAT" Why not provide this information for each sample, rather than for each site? Like this it does not allow for meaningful filtering of the data since part of a time series may have poor analogues.*
Response: We do provide minimum dissimilarities for each sample for each site. We clarified this in the text now.

Reviewer comment: (13) *Page 11: Located in instead of located from.*
Response: We changed the phrasing in the text.

Reviewer comment: (14) *Page 12/Fig 1: some of the fossil samples seem to come from marine sites (e.g. the Bay of Biscay or the Caspian Sea). Is that correct and if so, what do these time series tell us about vegetation and climate at that location?*
Response: yes, it is correct that there are few records that come from marine sites which were taken from the continental shelf. It contains information from source areas from the nearby continents (e.g. fluvially transported material). If users want to focus on terrestrial-only records, those marine sites could be filtered out by the *archive type* provided in the metadata. We indicated this in the text.

Reviewer comment: (15) *Page 14/Fig 3: and what do values below 1 mean, after all the scale goes down to 0, implying that for some sites none of the variance is constrained?*
Response: We provide the λ_1/λ_2 ratios as an assessment tool to determine ecologically important determinants as proposed by Juggins (2013). Values <1 suggest that the variable of interest doesn't explain a significant proportion of the variance in the modern pollen assemblage data. However, most

training data sets encompass multiple environmental gradients, derived from environmental variables that are often correlated and additional requirements to such variables would be necessary to explain a significant and independent portion of the variation in the training data set. Depending on the purpose of their studies users can decide for themselves whether they want to apply more strict criteria to the training data.

Reviewer comment: (16) *Page 21/Fig 6: better to show the full range of p values. Or somehow indicate where the sites are that did not pass the test as we can assess if there are any spatial patterns.*
Response: We updated the figure with the significance values that take spatial autocorrelation into account scaling from p = 0-0.2. In addition, we now also added those records that did not pass the significance level (p>=0.2) to the plots and indicated them with grey rectangles.

Reviewer comment: (17) *Page 21/Table 2: there are no values for MAT.*
Response: We added values for MAT to the table now.

Reviewer comment: (18) *Page 22: "only in single records" please rephrase. Single means one.*
Response: We replaced the word "single" with "individual".

Reviewer comment: (19) *Page 22: "High Plantaginaceae correlate with low TJuly in Central Europe indicating potential biases" please explain why only negative correlations are a problem, the reasoning is unclear to the non-pollen specialist.*
Response: Not only negative correlations are a problem, but also positive correlations, as they indicate potential biases in the reconstruction. While in Central Europe Plantaginaceae correlate negatively with T_{July}, positive correlations can be found between Plantaginaceae and P_{ann}. We specified this in the text now.
New text: High Plantaginaceae correlate with low T_{July} and high P_{ann} in Central Europe indicating potential biases in the temperature reconstructions i.e. too low temperatures become reconstructed. Similar correlations are found for Rumex, especially in Northern Europe (Fig. 8).

Reviewer comment: (20) *Page 23/Fig 7: the colour scale is not colourblind friendly.*
Response: We adapted the color scale to a color-blind friendly palette derived from CARTOcolor palettes in R (https://github.com/Nowosad/rcartocolor).

Reviewer comment: (21) *Page 28/Fig 10: which method was used for the reconstructions shown here?*
Response: The figure is based on the reconstruction with WA-PLS. We clarified this in the caption now.

Reviewer comment: (22) *Page 34: "We a priori selected … Tann and TJuly" it would be good to move this section to earlier in the discussion.*
Response: We shifted the sections to paragraph 5.2.

Reviewer comment: (23) *Page 36: "Climate reconstruction data sets like LegacyClimate 1.0 thus…" this is not a logical conclusion, it is only valid for reconstructions with a similar coverage as LegacyPollen 1.0. Reword.*

Response: We rephrased this sentence to clarify that LegacyClimate 1.0 is derived from LegacyPollen 1.0 and therefore has the same hemispheric coverage as LegacyPollen 1.0.

New text: Our LegacyPollen 1.0 fossil pollen synthesis (Herzschuh et al., 2022c) contains records from all over the Northern Hemisphere extratropics. We used this synthesis to produce our LegacyClimate 1.0 reconstruction data set, which thus can be used to infer spatio-temporal patterns in climate reconstructions that are not only limited to a local or regional scale.

Reviewer comment: (24) *Page 37: "Temperature reconstructions from proxy data indicate peak temperatures during the Holocene Thermal Maximum around 6000 years BP followed by a pronounced cooling trend toward the late Holocene (Kaufman et al., 2020b)" I think this is an oversimplification of the Kaufman reconstruction. They also highlight spatial variability in the Holocene temperature trends. As do Osman et al and Cartapanis et al.*

Response: Globally or hemispherically averaged temperature reconstructions often show a pronounced Holocene Thermal Maximum, followed by a decline in temperatures towards the Late Holocene, hence it is correct, that spatial variability is reported. We rephrased the sentence to point to those spatial variabilities in the cited studies.

New text: Globally or hemispherically averaged temperature reconstructions from proxy data indicate peak temperatures during the Holocene Thermal Maximum around 6000 years BP followed by a pronounced cooling trend toward the late Holocene, which is also visible in our pollen-based reconstructions (Fig. 10). Hence, spatial variability in the Holocene temperature trends (e.g. missing of a pronounced maximum for certain latitudinal bands; delayed thermal maximum on land compared to the ocean) indicate a more local rather than a global Holocene Thermal Maximum (Kaufman et al., 2020b; Osman et al., 2021; Cartapanis et al., 2022).

Reviewer comment: (25) *Page 37: "Temperature reconstructions are often derived from sea-surface temperatures as either mean annual temperatures (Birks, 2019; Bova et al., 2021) or global mean surface temperature (Marcott et al., 2013; Marsicek et al., 2018; Kaufman et al., 2020a and 2020b)." That seems an oversimplification of the cited literature. Some of the cited studies included seawater temperature estimates, but not all are exclusively based on SST.*

Response: It is true that not all of the cited studies are based on sea surface temperatures. We wanted to point out that often annual mean temperatures (no matter if derived from STT- or near-surface air temperature) are used for reconstructions. For multi-proxy comparisons it might be useful to not fully rely on annual mean temperatures, but rather also take seasonal mean temperatures into account. We rephrased the sentence.

Reviewer comment: (26) *Page 37: "In this respect, it might help…" how confident are the authors about the independence of the Tann and TJuly reconstructions? In other words, can we really reconstruct seasonality?*

Response: In most regions T_{July} and T_{ann} are not independent from each other. We decided to provide both T_{July} and T_{ann} because regionally T_{July} (or in general summer temperature) is more important - in particular in high latitudes. However, T_{ann} is more commonly used in multi-proxy comparisons.

New text: Despite T_{ann} being more commonly used in multi-proxy comparisons, it might be useful to also consider T_{July}, as regionally the mean July temperature (or in general summer temperature) is more important in particular in high latitudes. However, it is argued that proxy-based climate reconstructions are seasonally biased and therefore might be the reason for the observed proxy-model divergence (Liu et al., 2014; Rehfeld et al., 2016; Kaufman et al., 2020b). In this respect, it might help that we provide T_{July} along with T_{ann} reconstructions derived from our tailoring approach, which provides the opportunity to assess seasonal impacts on the reconstruction (especially in the high latitudes) in addition to a consistent reconstruction synthesis.

Reviewer comment: (27) *Page 37: "So far … hemispheric scale" is that statement still true after publication of https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2022GL099730 by the same lead author.*

Response: This paper has been cited in this paragraph as Herzschuh et al (2022a). However, this study focuses only on the Holocene and evaluates only model output of single transient run. We assume that there is much more potential.

Reviewer comment: (28) *Page 37: The last two paragraphs of section 5 seem to stand on their own and should be better integrated with the remaining text.*

Response: We shifted the last paragraph about potential biases in the pollen-based reconstruction to section 5.2 following our discussion about the assessment of our modern dataset used for reconstruction. The penultimate paragraph about the reconstruction of precipitation highlights its potential use for i.e. the evaluation of climate models, as precipitation reconstructions have not been implemented on a hemispheric scale so far. We therefore kept this sentence in paragraph 5.4.