Response to reviewers for "**More is not always better: downscaling climate model outputs from 30 to 5-minute resolution has minimal impact on coherence with Late Quaternary proxies"**

Reviewer 1:

RC1: This paper looks into the comparison between climate models and proxies and to what extent the differences between them could be reduced. The authors use statistical methods to increase the resolution of the model data to make it more comparable to proxy data, which represent local conditions. The conclusion is that even though the downscaled model data has more details the comparison with proxies is not really improved.

Considering the assumptions made and the methods used in the paper I wonder why anyone should expect an improvement of the model data. I suppose the paper can be a valuable contribution if these methods are commonly used in their part of the field. In any case, I think the authors should make it clear that their results apply to one particular type of statistical downscaling. It's not possible to draw any general conclusions about downscaling from these findings. Especially since the authors completely fails to mention dynamical downscaling.

Dynamical downscaling is known to improve the description of processes in the climate system and improve the description of local climate (e.g. Rummukainen, 2016). Dynamical downscaling is not very common within the field of palaeoclimate, but there are studies, e.g. Strandberg et al., 2011; Russo and Cubash, 2016; Velasquez et al., 2021; Strandberg et al., 2022; Strandberg et al., 2023.

Statistical downscaling is also known to improve local climate data and successfully minimize biases in climate models (e.g. Francois et al., 2020, Berg et al., 2022)
Bias adjustment methods (also more advanced methods like quantile mapping) build on the assumption that the relationship between model and observations is constant. This works for the present and future (coming 100 years or so) climate because climate change is not that large. For palaeoclimates, however, you cannot expect this relationship to hold. You can't expect the model biases to be the same in the present climate as in the LGM or in the early Eemian. In a climate different from today, and with different topography the weather regimes are not the same as today – and therefore you can't expect the model biases to be the same as today. If you in addition to these faulty assumptions use a very simplified method that only gives an offset of the model data, then I wonder why you at all expect your method to improve anything. Figure 2 clearly shows that your methods only slightly shifts model data. But you would like your method to also correct trends and variability.

> **AC1:** We would like to thank the reviewer for drawing attention to our lack of discussion around dynamical downscaling, and for providing useful references. As the reviewer themselves suggests, dynamical downscaling is not very common within the field of palaeoclimate and associated fields (i.e. archaeology, palaeoecology etc.) who consume model outputs. This is because

this methodology is not accessible to researchers working with a large number of time steps due to the computational costs and time involved, particularly when exploring climatic variability over extended temporal or geographic spans. Yet tackling questions in archaeology, palaeoecology etc. often require finer levels of spatial resolution than typically provided by publicly available climatic model time series. We do not provide an overly extended discussion of these other methods of downscaling (i.e. dynamical downscaling), given they are not very relevant to our field, but add specific reference to them on Lines 71-75

“High resolution simulations of multiple time slices are often desired by consumers of model output yet difficult to obtain due to computational costs. For example, dynamical downscaling allows for the detailed description of processes in the climatic system and can improve the capturing of localised climatic conditions (Rummukainen, 2016; Strandberg et al., 2023), however this method is rarely applied in fields like palaeoecology and archaeology due to the computational costs, particularly when a large number of time steps are required.”

And reiterate this point in the discussion on Lines 536-550:

“Our results suggest that using statistical methods of downscaling simulated time series to much higher resolutions does not significantly improve the agreement between model output and pollen-proxy reconstructions, yet we note that there is a trade-off between enhancing spatial resolution and increasing potential error. Such error in a given location could either be caused by using too coarse a resolution on the one hand or by unreliable interpolation on the other. For this reason, there are likely to be many circumstances in which it is still better to use downscaled models (with caveats), particularly when variability within 30-min cells (~55km on each side) is important (e.g. Boisard et al. 2025). For example, the identification of conditions at specific locations within climatic extremes may be overlooked when using a model at a broader scale, such as at Late Pleistocene archaeological site Fincha Habera in the Bale Mountains of southern Ethiopia (Groos et al. 2021). Here, lower annual temperatures predicted by delta-downscaled models may better characterise the on-site environment than that also incorporating environmental trends in surrounding lower altitude landscape (Timbrell et al. 2022). Other methods of increasing model output, such as dynamical downscaling, may be better equipped for more localised applications, yet these are largely inaccessible for consumers of model output in fields like palaeoecology and archaeology where the computational costs are impractical. Overall, we present a streamlined pipeline for delta-downscaling climate model time series within the pastclim R package (Leonardi et al. 2023), though we stress that careful consideration is required to select the optimal method and spatial resolution, based on the scope of the research question at hand.”

We have also stressed the importance of testing the delta method as one of the most accessible methods of downscaling for consumers of palaeoclimatic model outputs on Lines 61-69:

"Models additionally offer much wider spatial coverage of the landscape that can be directly related to specific study sites and the palaeoclimatic differences between them. However, the integration of modelled climate with proxy data is not straightforward. For example, using simulations at a coarse resolution can produce biases when compared to on-site proxies due to the underlying complexity of the physical landscape, particularly in coastal and topographically diverse regions (Maraun and Widmann, 2018). Resultant differences can be in the order of several degrees for temperature and tens of percent for precipitation, which could lead to substantially different biome classifications and estimations of ecologies (Kottek et al., 2006). Such variations can have important implications for the diverse fields employing model output for the reconstruction of past and present species distributions, dispersal and extinction processes, and biogeographic patterns."

RC2: My point here is that the conclusions drawn in the paper are far too general. Statements like: "our results imply that downscaling to a very fine scale has minimal to no effect on the coherence of model data with pollen records." (l 28-29) are simply wrong. Your conclusions only apply to the methods used in this study, not all varieties of downscaling and bias adjustment.

AC2: We thank the reviewer for pointing out that some of our statements are too generalised. We have corrected this throughout the manuscript, for example on Lines 507-510:

"Our results highlight that further downscaling models via statistical methods to much higher resolutions (5-minute) fails to *consistently* capture more of the climatic trend from pollen proxy records. Indeed, we were unable to demonstrate any statistically significant differences in model-data coherence between 30-min and 5-min model resolutions in any subset of this large dataset."

And added further specification that we are testing the delta method *specifically,* including in the title:

"More is not always better: delta-downscaling climate model outputs from 30 to 5-minute resolution has minimal impact on coherence with Late Quaternary proxies"

RC3: I think that the authors could be a bit more critical towards proxies. It's a bit much to call it "golden standard", and this comes from a modeller who is used to see all problems in models, and less so in proxies. Remember that proxies also have uncertainties. For example, Strandberg et al. (2011) come to the conclusion that the comparison between climate model and proxy data is mostly limited by the large errors bars in proxy data.

AC3: We have added further critique of proxies using the reference suggested by the reviewer, although we retain our stance that proxies are typically considered to be 'gold standard' by archaeologists, palaeontologists etc. when looking at climatic conditions at specific locations in the past:

Lines 99-102: "Proxies offer a more localised account of climate in certain places, yet they too can be associated with high degrees of uncertainty, arising from multiple sources. Nonetheless, determining model agreement with empirical reconstructions from proxies remains a widely applied method for ground-truthing downscaled climatic output."

Lines 107-113: "A recent meta-analysis by Laepple et al. (2023) found that studies in the Northern Hemisphere (where data are more abundant) have mixed results, suggesting potential areas of mismatch at local and regional scales. These authors suggest that shortcomings in both model simulations and proxy reconstructions may contribute to this divergence with models being less efficient at simulating local and regional temperature variability at relatively long timescales and methods of temperature reconstruction from proxies facing systematic deficiencies, though stronger emphasis is placed on the former. Strandberg et al. (2022) conversely suggest that comparisons between models and proxies are mostly limited by the large errors associated with proxy data."

RC4: I would also like you to think about the distribution between figures in the paper and the supplementary material. The paper doesn't include so many figures, and some of them are, to be honest, not that informative.

AC4: We have reworked all of the figures based on your specific suggestions (see below). Thank you.

RC5: At the same time the paper is quite heavy on reference to the supplementary. Perhaps you would like to lift something from the supplementary to the main text? And while you're at it rework some of the existing figures.

AC5: Thank you for this suggestion. We also apologise for the missing tables in the SOM. It was requested upon submission that that four tables from the manuscript be moved from the main text into the SOM due to CoP formatting issues. A new version of the SOM was submitted, including these 4 tables, but is unfortunate that this version was not shared with the reviewers nor uploaded online. We have however moved these large tables to an Appendix (Appendix A) so they are more easily accessible within the manuscript itself.

RC6: In conclusion, this paper has a very shallow description and discussion of downscaling and bias adjustment methods. This should be expanded. The conclusions should be reformulated to only apply to the methods used in the study, instead of all methods. If this is done, I think that the paper could be accepted (assuming that the methods are actually used in other projects). Otherwise I will recommend rejection.

AC6: Thank you for this summary. We believe we have sufficiently addressed all of your comments (see below) and would like to stress that accessible methods (i.e. that can be easily applied within a workflow, require manageable processing and accessible computational power) to downscale a large number of reconstructions are indeed very sought after in our field, who tend to be consumers of climatic model output as opposed to modellers.

Comments

RC7: L56-57 It could also be worth to mention that climate models also offer a picture that is also consistent across variables, thus giving a more complete picture of the climate.

AC7: We have amended Lines 59-61:

"Model output have the potential to overcome these shortfalls, providing tangible values for parameters such as temperature, precipitation, and a range of derived bioclimatic indices (e.g., Hijmans *et al.,* 2005), that are consistent across variables for a more complete account of climatic conditions."

RC8: L60 what do you mean by "observational data" here? Do you mean proxies? In that case, say so. Proxies and observations are different things. If you mean observations, explain why it is relevant to mention here. The rest of the paragraph is about proxies.

AC8: We have changed this to say 'proxy data' for clarity.

RC9: L63 "errors" Perhaps it's better to talk about "differences" since proxies also have errors.

AC9: Thank you for this suggestion; we have changed this to differences.

RC10: L71 "Different methods" -> "Different statistical methods". Otherwise you should also mention dynamical downscaling.

AC10: We have edited the manuscript accordingly and added more discussion about dynamical downscaling, as suggested on Lines 71-81:

"High resolution simulations of multiple time slices are often desired by consumers of model output yet difficult to obtain due to computational costs. For example, dynamical downscaling allows for the detailed description of processes in the climatic system and can improve the capturing of localised climatic conditions (Rummukainen, 2016; Strandberg et al., 2023), however this method is rarely applied in fields like palaeoecology and archaeology due to the computational costs, particularly when a large number of time steps are required. Most of the recently produced time series of palaeoclimate outputs have been downscaled from the native resolution of the models (usually in the order of 2 or 3 arc-degrees) to a higher resolution of 30 arc-minutes using

statistical methods (Fordham et al. 2017; Beyer *et al.* 2020a; Krapp *et al.* 2021; Zeller and Timmerman 2024; Mondanaro et al. 2025) as these approaches can be more easily applied to several time periods. Within statistical downscaling, different methods exist to increase the spatial resolution of model simulations; these include the delta method, generalised additive models (GAMs), and quantile mapping. These are all aimed at minimising biases in models, characterised as differences in statistical distributions between observed and simulated series."

RC11: Section 2.1 Here, I would like you to explain a bit more. It's difficult to follow what is done and in which order. Consider a more linear description, like GMC run, bias adjustment, downscaling etc. For example I don't understand what the Beyer et al simulation is. Is it a GCM run, a modification of the HadCM3 run or something else? Please also give some details about the HadCM3 run, for example regarding resolution and time span.

AC11: We now provide a detailed description of the output from Beyer et al. (2020), and the original HadCM3 model output (Huntley et al. 2022) we have subsequently added upon request from Reviewer 2 on Lines 129-170:

"2.1 Climate models

To test the impact of delta-downscaling at different resolutions, we used two time series of model simulations. The first one is a set of raw temperature and precipitation outputs from the HadCM3 GCM, at their native resolution of 3.275 x2.5 arc-degrees taken from Huntley et al. (2022). We consider a set of simulations in which the HadCM3 was run with appropriate boundary conditions for the last 120k years at 2,00 years intervals (the original set in that paper covered the last 800k years). The second series comes from Beyer et al. (2020a) within the pastclim R package (Leonardi et al. 2023). These reconstructions are based on an older series of runs of the HadCM3 Global Circulation Model (Singarayer and Valdes 2010, Singarayer and Burrough, 2015; Valdes et al. 2017) for the last 120k years, in 72 snapshots (2,000-year time steps between 120,000 BP and 22,000 BP; 1,000-year time steps between 22,000 BP and the pre-industrial modern era). As in the other set, the original model output of HadCM3 had a grid resolution of 3.75 x 2.5 arc-degrees.

These outputs were first downscaled using a series of runs of the higher resolution HadAM3H model, available at 1.25 x 0.83 arc-degrees for the last 21,000 years in 9 snapshots (2,000-year time steps between 12,000 BP and 6,000 BP; 3,000-year time steps otherwise) using an approached termed dynamic delta downscaling by Beyer et al (2020a). This method consists of generating a set of delta matrices based on the few time steps for which outputs were available from both HadCM3 and HadAM3H, and then using these matrices to downscale each time step in the full set by using a weighted interpolation of the two closest delta matrices based on CO2 (see Beyer et al, 2020a, for details). This approach takes advantage of the higher resolution of local dynamics

captured by HadAM3H, which is computationally too expensive to be run for all time steps. These outputs were then debiased and downscaled in Beyer et al. (2020a) to 0.5 x 0.5 arc-degrees with the delta method using the Climate Research Unit Global Climate Dataset (CRU) as the modern climatic reference (Mitchell and Jones, 2005).

We delta downscaled and debiased these two model outputs to a resolution of both 30 arc-minutes and 5 arc-minutes using modern observation from WorldClim2 (Fick and Hijmans, 2017). For the Beyer et al (2020a) model, as it was already at 30 arc-minutes, the delta downscaling at this resolution gives us a debiased version based on WorldClim2 rather than CRU. We used a global relief map from ETOPO2022 (NOAA National Center for Environmental Information, 2022) to reconstruct past coastlines following sea level change (Spratt and Lisiecki, 2016). We selected WorldClim2 as the modern reference as the transfer functions used in the LegacyClimate1.0 dataset were also derived from this dataset (at 30-minute resolution), allowing us to control for the effects of the modern data used for debiasing on our results. All data manipulations were done using the R package pastclim (Leonardi et al. 2023).

Downscaling was performed one monthly variable at a time (i.e., January temperature) by taking the coarse simulations from Beyer et al. (2020a) with the corresponding set of high-resolution modern simulations from WorldClim2 (Fick and Hijmans, 2017) and equally high-resolution global relief map (NOAA National Centres for Environmental Information, 2022). Through integrating both bathymetric and topographic values for masking sea level changes, a delta raster was computed, adding the difference between past and present-day simulated climate to present-day observed climate, following Beyer et al. (2020a) and Krapp et al. (2021) The delta method therefore assumes that local (i.e. grid-cell-specific) model biases are constant over time (Maraun and Widmann, 2018). The resulting matrix only covers the land extent at the present. We then expanded this matrix to reach the largest land-extent in any of the times-steps under consideration using an inverse-distance-weighted interpolation. For most of the world, at the resolution of 30 and 5 arc-minutes, this only requires interpolating a small number of cells away from the coastline; for higher resolutions, other interpolating algorithms might be more appropriate. We note that the delta-downscaling can also be obtained by creating first the difference between model outputs, which is then applied to the observational model. However, such a direction is more computationally expensive, as the interpolation outside the coastlines would have to be repeated for each time step."

RC12: L123-124 Is this the same simulation as in lines 112-113.

AC12: Yes, here we were referring to the Beyer et al. (2020a) output. We have adjusted the method sections to improve the clarity of our workflow (see above).

RC13: Eq. 1 Please explain what "DM", "sim", "raw" and "obs" denotes.

AC13: We have amended this section (Lines 172-190) as follows:

"For temperature variables, the bias in a geographical location x (a cell with a given latitude and longitude) is given by the difference between present-day observed $T_{obs}(x, 0)$ and simulated $T_{sum}^{\oplus}(x, 0)$ temperature, interpolated to the desired higher resolution grid via bilinear interpolation. Downscaled temperature (T_{sim}^{DD}) in x at time t is thus estimated as

$$T_{sim}^{DD}(x,t) \coloneqq T_{sim}^{\oplus}(x,t) + \left(T_{obs}(x,0) - T_{sim}^{\oplus}(x,0)\right)$$

Precipitation is lower bounded by zero and covers different orders of magnitude across different regions compared to temperature. Multiplying rather than adding the bias correction is common when applying the delta method for precipitation, which corresponds to applying the simulated relative change to the observations (Maraun and Widmann, 2018). However, this method can therefore be hypersensitive in drylands, leading to overprediction of precipitation (and thus exacerbating the 'drizzling' bias of GCM). We have therefore adopted an additive approach for precipitation, analogous to the one used for temperature, with clamping within the range of observed maximum and minimum for current climate (see Beyer et al. 2020a). Like temperature, downscaled precipitation is estimated as

$$P_{sim}^{DD}(x,t) \coloneqq P_{sim}^{\oplus}(x,t) + \left(P_{obs}(x,0) - P_{sim}^{\oplus}(x,0)\right)$$"

RC14: L161 Why do you use "bio01" here and "Tann" elsewere? Use a consistent terminology. I would prefer abbreviations like Tann instead of bio01, because they are easier to understand.

AC14: We use 'bio01' and 'Tann' etc. as this is how mean annual temperature are abbreviated in the climatic model and proxy dataset respectively. We retain bio01, bio12 and bio10 when describing the model output in the Methods and in Figures of the modelled climatic layers, however we use the full variable names (e.g. mean annual temperature) throughout the manuscript when discussing our results to ensure consistency.

We have added an additional sentence on Lines 216-221 explaining that these terms are equivalent variables:

"Our use of a single database reconstructing climate based on a single proxy reduces inter-site variability resulting from the type of data utilised and allows the generation of analogous climatic parameters with direct relevance to bioclimatic variables available in the Beyer et al. (2020a) model; T_{ann}, T_{july} and P_{ann} from LegacyClimate1.0 are the equivalent bioclimatic variables to bio01, bio10 and bio12 from HadCM3 GCM (Huntley et al. 2022) and Beyer et al. (2020a) model time series, which are standardly used in climatic modelling. "

Moreover, we have provided an account of the equivalent climatic variables extracted in Table 1, and have added an explanation of their abbreviations in Table 1.

"Table 1. Summary of the proxy records selected from the LegacyClimate 1.0 (Herzschuh *et al.,* 2023) and the model outputs (Beyer *et al.,* 2020a; Huntley et al. 2022) selected for analysis of mean annual temperature (bio01, T_{ann}), mean July temperature (bio10, T_{july}) and total annual precipitation (bio12, P_{ann})."

RC15: L211-213 If this sentence is the only thing you write about Fig 2, why show it at all? I think it would be worth to describe also the differences between WAPLS and MAT.

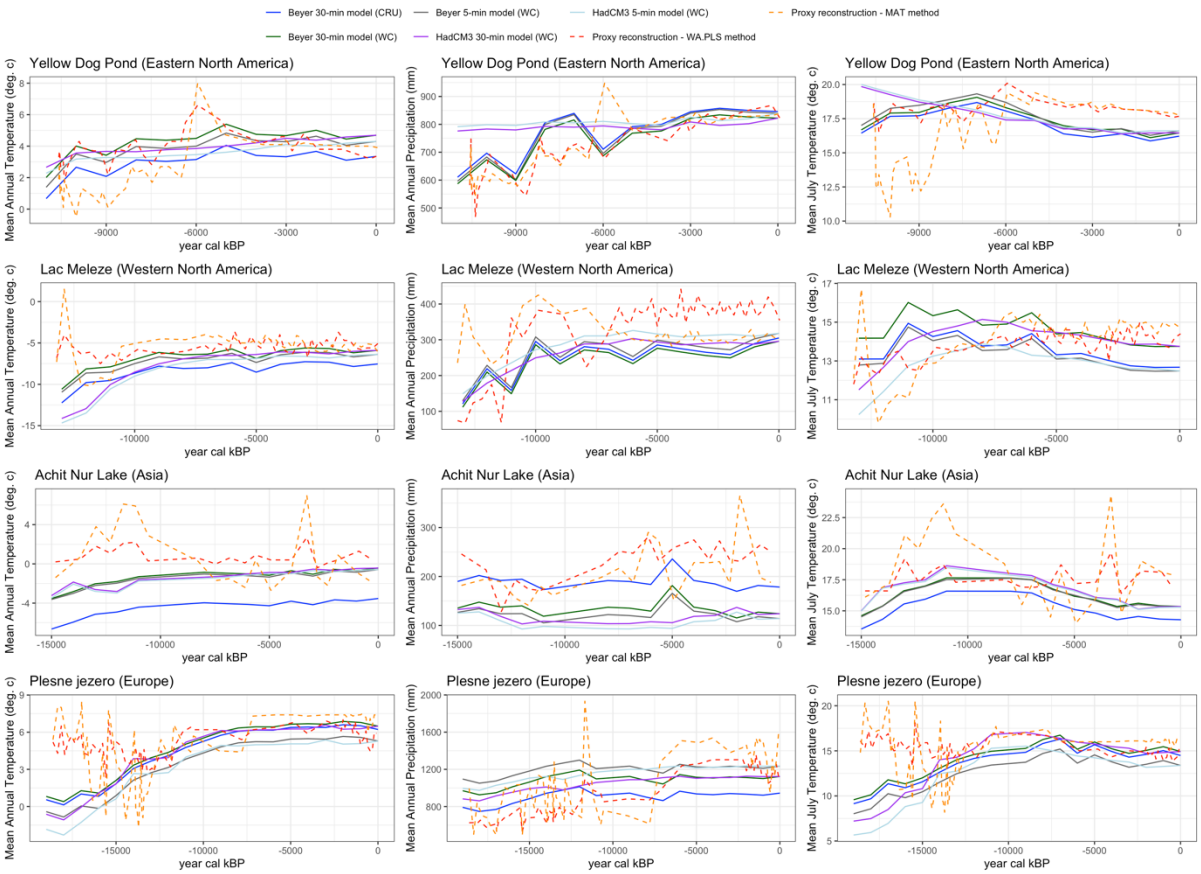
AC15: We show Figure 2 as it visually captures the comparisons between time series that we are quantifying in this paper. We have added an additional sentence on Lines 259-261:

"Figure 2 highlights a sample of non-interpolated time series from proxy sites across the geographic span of the LegacyClim1.0 dataset, highlighting the coherence through time between different models and empirical reconstructions (WA-PLS and MAT) of the three climatic parameters (annual temperature, July temperature and annual precipitation)."

We do not think it is relevant to this paper to extensively describe the differences between the WA-PLS and MAT methods. These two state-of-the-art analytical methods have been commonly used in the field for over 3 decades, and there is ample documentation on how they work and how they perform in different situations. We feel that entering into technicalities would not add anything significant to the paper. However, and to guide interested readers, we have added three important references that correspond to extensive reviews of the field of pollen-based climate reconstructions that clearly highlight that the relative strengths and weaknesses of each of the methods (Sweeney et al., 2018; Birks et al. 2010; Chevalier et al., 2020). If the reviewer is referring here to the differences in *results* between WA-PLS and MAT, these are reported throughout Section 3, with limited variations between methods.

RC16: Fig 2 It's difficult to see the difference between the lines representing models. Consider using colours that are more different from each other, and to use dashes and dots to separate them even more.

AC16: We have made these suggested amendments by changing to a divergent colour scheme and using line representations to differentiate proxy from model time series in Figure 2.



RC17: Fig 2 How large are the areas shown here? How is the comparison between model and proxies made? Is it one model grid point vs. One proxy data point? If you average model data over a larger area some of the point of downscaling will disappear.

AC17: Figure 2 shows the climatic time series produced by the proxy reconstruction and the model output at the coordinates of the proxy sites. We have added further description in Lines 223-226 of the comparison in the methods section:

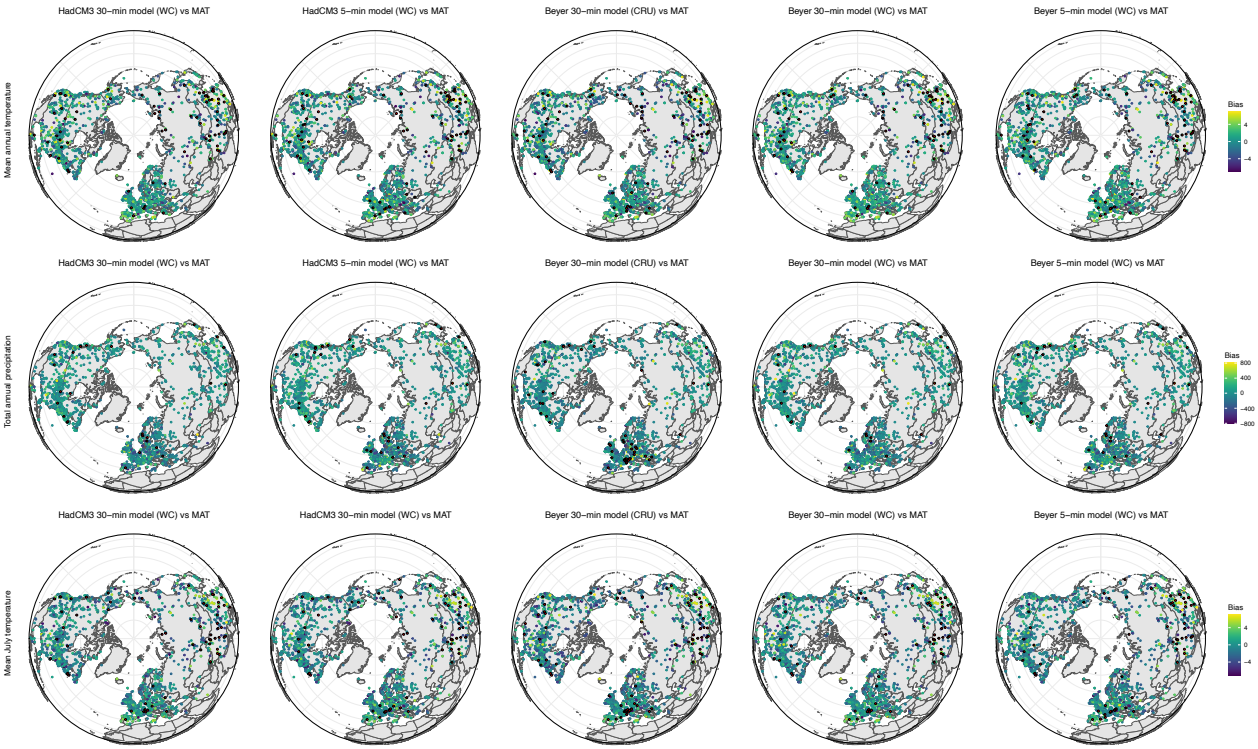
"To facilitate comparison between the proxy reconstructions and the model outputs, we interpolate each proxy record via bilinear interpolation to the equivalent chronological resolution of the climatic models to enable quantification of differences between the time series; interpolating to regular time intervals ensures that periods of particularly dense sampling in the original cores do not exert undue influence on the results. For this, we extracted the climatic values from the model at the coordinates of the proxy site for the time steps captured in the proxy record."

RC18: Fig 3 Add units to the panels. Add temperature, precipitation etc to the leftmost panel in every row.

AC18: We have made these suggested amendments to Figure 3 (see AC19).

RC19: Fig 3 This could be presented much better. The panels are small, the data only covers a part of the panels, the colours are difficult to distinguish. I cannot draw any conclusions from looking at Fig 3. Think about alternative ways to show this. Perhaps you could collect the point in regions and do boxplots show the differences per region. That would give you a quantitative comparison.

AC19: Thank you for this suggestion. We have made edits to Figure 3 to improve the readability of this figure (namely cropped the map so the data fill the frame, and highlighted outliers in red). Boxplots are however a good suggestion, and we have added these for the regional subgroups and landscape subgroups to the SOM as alternative ways of displaying the results presented in the tables in Appendix 2 and Figure 3.



RC20: L297 Is “predict” the right word here? The proxy data do not predict temperatures.

AC20: We have changed this to ‘indicate’.

RC21: Fig 4 It’s obvious that Fig 4 shows the effect of the resolution. I’m, however, not sure that it shows the “effects of landscape dynamics”. What do you mean by that. Furthermore, I think you could make your point by showing just one region in one line. This is a lot of figure space for little information.

AC21: This figure demonstrates how increasing the resolution of the model better captures more fine-scale detail of the landscape, such as coastlines and topographic differences. We believe that this figure effectively highlights the impact that downscaling can have in different types of landscapes (i.e. in the

Pittsburg Basin where it is very flat and inland, there is little change, whereas in South Italy there is much more detail captured in localised climate at coastlines and areas of diverse topography). We have added further detail to this effect on Lines 382-387:

"Downscaling model outputs to a very high resolution is often performed to account for smaller-scale landscape features that can locally impact climatic conditions, such as topography and coastlines (Fig. 4). Figure 4 highlights these effects of increasing model resolution in different areas of varying landscape complexity; for example, in the Pittsburg Basin (which is inland and flat) there is little change in the climate signal captured at proxy sites (white circles) following downscaling, whereas, in southern Italy and the Qillian Mountains, downscaling captures more localised details in climates associated with landscape-level variations. Proxy records at higher elevations and topographic complexity may therefore be expected to show stronger coherence with the higher resolution models compared to those at relatively lower resolution."

RC22: Fig 4 What do the dots represent?

AC22: We have added to the caption of Figure 4:

"Figure 4. Three regional examples of modelled mean annual temperature for the present day (bio01), demonstrating how downscaling increases spatial resolution by capturing the effects of landscape dynamics through space on climate depending on the underlying topography. Geographic variability in temperature is shown, as simulated by the Beyer et al. (2020a) 30-min model output (CRU), Beyer et al. (2020a) 30-min model output (WC), and Beyer et al. (2020a) 5-min model output (WC), Locations of proxy locations from LegacyClimate 1.0 are shown as white circles."

RC23: L322 Is it correct to refer to Fig 4 here?

AC23: Thank you for pointing this out – we were referring to Figure 5 here. This has been amended.

RC24: L334 "Models are also inherently calibrated …" This is a very general statement that doesn't apply to all climate models. Pleas specify which models you refer to.

AC24: We have specified that here we are referring to delta-downscaled models on Lines 454-457:

"Delta-downscaled models are also inherently tuned to replicate current rather than past climate patterns, and proxy reconstructions rely on the identification of modern analogue species that may have a different link to climate than palaeoecological communities, likely further contributing to higher divergence in older time periods (Chevalier *et al.* 2020)."

RC25: L364 I don't think this is a question well posed. How do you know that the downscaling is the problem, and not the methods you used to do the downscaling. Again, this is a very general statement that doesn't apply to all downscaling techniques.

AC25: We have edited the phrasing of Lines 502-507:

"Increasing the spatial resolution of model time-series is often thought to be required to more accurately capture the climatic conditions of specific places at specific times. But what is the optimal spatial resolution for adequately detailing finer-scale signals? We tackle this question by testing the agreement between different model outputs and empirical reconstructions from pollen proxies from the Late Quaternary for annual and July temperatures and annual precipitation. Ground-truthing modelled climate in this way is common, as proxies are considered to be the 'gold standard' for capturing more localised variations in climatic conditions in specific places"

We have also specified that we are referring to the methods that we tested in the paper on Lines 507-508:

"Our results highlight that further downscaling models via the delta method to much higher resolutions (5-minute) fails to *consistently* capture more of the climatic trend from proxy records."

RC26: L364-369 I think this is a testament of the poor methods you use.

AC26: It may be the case that other methods, such as dynamical downscaling, would produce better results, however unfortunately, these are not accessible methods to many researchers who use climatic models. We have stressed this on Lines 546-550.

"Other methods of increasing model output, such as dynamical downscaling, may be better equipped for more localised applications, yet these are largely inaccessible for consumers of model output in fields like palaeoecology and archaeology where the computational costs are impractical. Overall, we present a streamlined pipeline for delta-downscaling climate model time series within the pastclim R package (Leonardi et al. 2023), though we stress that careful consideration is required to select the optimal method and spatial resolution, based on the scope of the research question at hand."

RC27: L376 You have note mentioned that Beyer et al is a climate emulator. Please add this to section 2.1.

AC27: This was a mistake and has been removed.

RC28: L401-403 This is simply wrong. You only show that the downscaling method used in this paper fails. Based on that you should not dismiss all different ways to do

downscaling. It would be unfortunate if the community thought that all downscaling is pointless.

> **AC28:** We have amended Lines 556-558 specify that we are referring to the method we have tested in the paper:
>
> "We show that downscaling via the delta-method fails to consistently capture more signal from temperature and precipitation proxy reconstructions, though model time series at both median (30-arc minutes) and fine-grained (5-arc minutes) spatial resolutions characterise climatic variables in broadly similar ways to pollen proxies."
>
> As highlighted in AC26, we have added further discussion of other methods that may be better equipped than the ones tested in this paper, albeit more inaccessible to most consumers of model time series.

Minor comments
RC29: L49 missing "(" somewhere before this ")"

> **AC29:** We have removed this error.

Reviewer 2:

RC1: This is a disappointing paper, because the issue of whether and how to downsccale climate-model output is an important one, and even as models achieve ever higher resolutions, the demand for even higher resolution data will remain. This paper attempts to assess the match between a collection of pollen-derived reconstructions and climate-model output downscaled to 30-min and 5-min resolutions. However, the climate-model output is represented by the Beyer et al. (2020a) 30-min data set which itself was produced by debiasing and downscaling HadCM3 model output. There is therefore a big assumption here, then, that the Beyer et al. data is sound, and there were no artefacts generated in the process of its creation.

> **AC1:** We agree that the issue of downscaling is a very important one, and indeed we are frequently asked to include further downscaling in our workflow as it is a 'more accurate' representation of past climatic conditions in specific places. In our paper, we seek to use relatively simple methods (as are typical for consumers of climate model outputs) to downscale a large number of reconstructions to test whether this is the case. Of course, accuracy is difficult to ascertain due to error potentially arising from multiple sources in both models and proxies, however assessing the agreement with empirical reconstructions from proxies is an important starting point to encourage discussion and is a widely used approach to ground-truth model output.
>
> We thank the reviewer for suggesting that we include a comparison with directly downscaled HadCM3 outputs. We have done so, using a model time series from Huntley et al (2022), which is an updated version of that used to generate Beyer

et al. (2020a). The conclusions of our paper do not change. Because Beyer et al. (2020a) used a more complex downscaling approach which involved integrating information from a higher resolution model (now better described in the methods, see response below), and users of the pastclim R package (Leonardi et al. 2023) are likely to use it as a possible starting point (given that it is easily accessible along with our functions for downscaling), we have kept the previous comparisons with Beyer et al. (2020a) as well. Those comparisons also show the importance of using different modern day observational data to downscale and debias and compare to proxies, which might in turn have been calibrated against such observations. Overall, we show that the conclusions are not linked to the processing that was done for Beyer et al (2020a).

RC2: I think a better experimental design would have been to start with actual model output, and to spend more time focusing on the performance of the downscaling and debiasing routines for present-day data.

AC2: We thank the reviewer for the suggestion to start with the actual model output. We have added the HadCM3 GCM output to our analyses (using a recent time series from Huntley et al. (2022), which is supposed to be a slight improvement on the original set of runs used in Beyer et al. (2020a), and report highly comparable results to that previously presented. Indeed, like with the Beyer et al. (2020a) model time series, we find little net difference of downscaling with the HadCM3 model output from the 30-min and the 5-min resolution, with no statistically significant differences in coherence between the proxy records and the model outputs at different resolutions for any subset tested.

Although an interesting idea to focus on downscaling and de-biasing routines for the present day, this is not the focus of our analysis. We are interested in testing whether delta-downscaling (a method routinely used to downscale large time series of palaeoclimate reconstructions) can be used on model time series to improve the output's coherence with proxy records during the Late Pleistocene and Holocene. This is important because consumers of climate model outputs are increasingly interested in performing continuous-time analyses at a high spatial resolution across a wide range of climatic and ecological applications, such as (palaeo) species distribution modelling and empirical analyses of the effects of climate on spatiotemporally disparate samples. As a field, we are becoming increasingly aware of issues related to optimising resolution, yet there is currently no consensus as to when downscaling may be important nor how one should accurately increase the resolution of model output to capture climate in the past at a sufficient level of detail. Delta-downscaling is often suggested as a solution, due to its practicality when applied to tens or hundreds of time steps.

We have added further discussion to this effect on Lines 39-49:

"Recently, the production of high-resolution simulations, characterising climatic variables across vast time periods, have allowed for the production and analyses of time series similar to those produced using proxy data (e.g., Fordham et al., 2017; Armstrong et al., 2019; Holden et al., 2019; Beyer et al., 2020; Brown et al., 2020; Karger et al., 2021; Krapp et al., 2021; Timmerman et al., 2022). Openly accessible simulated datasets, such as those published by Beyer et al. (2020a), Krapp et al. (2021), Yun et al. (2023) and Barreto et al. (2023), and associated analytical packages toolkits (e.g., the analytical packagetool pastclim for manipulating and extracting modelled data; Leonardi et al., 2023), are particularly useful for scientists interested in Middle-Late Pleistocene and Holocene timescales (e.g. Beyer et al., 2021; Padilla-Iglesias et al., 2022; Blinkhorn et al., 2022; Leonardi et al., 2022), facilitating continuous-time analyses at a high spatial resolution across a wide range of applications, such as habitat and species distribution modelling (SDM) and the quantitative analysis of climate change in relation to spatiotemporally diverse biological and behavioural phenomena (e.g. Beyer et al., 2021; Padilla-Iglesias et al., 2022; Blinkhorn et al., 2022; Timmerman et al. 2022; Leonardi et al., 2022; Zeller and Timmerman 2024; Mondanaro et al. 2025)."

"...As a community, we are becoming increasingly aware of issues related to the scale and resolution of climate variables, yet it is currently unclear what is a desirable level of downscalinglevel of downscaling is desirable for applications like SDM. Indeed, the ODMAP (Overview, Data, Model, Assessment, Prediction) protocol stresses the importance of spatial resolution and extent of environmental predictors, as well as processing and scaling (Fitzpatrick et al. 2021), yet there is still no universally agreed upon pipeline for SDM to help determine when downscaling may be important." (Lines 90-94).

RC3: The paper also completely avoids even commenting on other approaches for downscaling, such as dynamic downscaling, and the take-home message, that the target resolution doesn't matter, could be taken to say "why bother?"

AC3: We certainly do not want to imply that downscaling does not matter. we have added in further discussion of dynamic downscaling on Lines 69-78, as requested also by R1:

"High resolution simulations of multiple time slices are often desired by consumers of model output yet difficult to obtain due to computational costs. For example, dynamical downscaling allows for the detailed description of processes in the climatic system and can improve the capturing of localised climatic conditions (Rummukainen, 2016; Strandberg et al., 2023), however this method is rarely applied in fields like palaeoecology and archaeology due to the computational costs, particularly when a large number of time steps are required. Most of the recently produced time series of palaeoclimate outputs have been downscaled from the native resolution of the models (usually in the order of 2 or 3 arc-degrees) to a higher resolution of 30 arc-minutes using statistical methods (Fordham et al. 2017; Beyer et al. 2020a; Krapp et al. 2021;

Zeller and Timmerman 2024; Mondanaro et al. 2025) using statistical downscaling, as these approachesis method can be more easily applied to several time periods."

To make sure that the take-home message of the paper could not be taken to say 'why bother', we have expanded the final paragraph (Lines 536-550):

"Our results suggests that using statistical methods of downscaling simulated time series to much higher resolutions does not necessarily significantly improve the agreement between model outputs and pollen-proxy reconstructions, yet we note that there is a trade-off between enhancing spatial resolution and increasing potential error. Such error in a given location could either be caused by using too coarse a resolution on the one hand or by unreliable interpolation on the other. For this reason, there are likely to be many circumstances in which it is still better to use downscaled models (with caveats), particularly when variability within 30-min cells (~55km on each side) is important (e.g. Boisard et al. 2025). For example, the identification of conditions at specific locations within climatic extremes may be overlooked when using a model at a broader scale, such as at Late Pleistocene archaeological site Fincha Habera in the Bale Mountains of southern Ethiopia (Groos et al. 2021). Here, lower annual temperatures predicted by delta-downscaled models may better characterise the on-site environment than that also incorporating environmental trends in surrounding lower altitude landscape (Timbrell et al. 2022). Other methods of increasing model output, such as dynamical downscaling, may be better equipped for more localised applications, yet these are largely inaccessible for consumers of model output in fields like palaeoecology and archaeology where the computational costs are impractical. Overall, we present a streamlined pipeline for delta-downscaling climate model time series within the pastclim R package (Leonardi et al. 2023), though we stress that careful consideration is required to select the optimal method and spatial resolution when using models, based on the scope of the research question at hand."

We also point the reviewer to the final sentence of the abstract (Lines 30-33):

"Optimal spatial resolution is therefore likely to be highly dependent on specific research contexts and questions, with careful consideration required regarding the trade-off between highlighting local-scale variations and increasing potential error via unreliable interpolation."

And our (edited) conclusion (Lines 553-560):
"Paleoclimatic proxies and climate models constitute two contrasting yet complementary sources of information on past climates. Demand for high-resolution climatic simulations that characterise landscape-scale heterogeneities come from the multitude of fields that employ ecological data, such as those that wish to map species distributions through time and space or quantitatively test hypotheses about the impact of climatic change and/or variability on various biological or behavioural phenomena. We show that

downscaling via the delta-method fails to consistently capture more signal from temperature and precipitation proxy reconstructions, though model time series at both median (30-arc minutes) and fine-grained (5-arc minutes) spatial resolutions characterise climatic variables in broadly similar ways to pollen proxies. Utilising model output for analyses of past climate therefore involves a careful balancing act between accentuating variations relevant to the study questions and the potential introduction of error by unreliable interpolation."

Based on this, we do not believe that the take home message is 'why bother' but that careful consideration should be required to determine *when* downscaling is important, given that coherence between proxy records and model outputs does not change significantly. We understand that the reviewer is 'disappointed' with the results, however if we only publish positive results these important issues will be overlooked. Given that 'the demand for even higher resolution data will remain', encouraging debate about this issue can only benefit any field that employs climatic model output in their research.

RC4: The paper is not well written or produced. The figures don't work very well, there are missing tables, and it lacks even first-order attempts to explain patterns in the results.

AC4: We apologise for the missing tables – they were removed from the manuscript to SOM upon request of CoP after submission and the reformatting was subsequently incomplete. The four tables are now included in Appendix A so that they are still easily accessible within the manuscript itself.

We have reworked all of the figures following the feedback by both reviewers and added further discussion of our results throughout the manuscript, which we highlight specifically below. We highlight that our paper does not seek to determine the source of the discrepancies between models and proxies (which is impossible from our study design) but rather to explore the influence of downscaling on model-data coherence across different scenarios in order to make recommendations about *when* downscaling might be useful.

RC5: Terms like "estimation," "prediction," "reconstruction" are used interchangeably, and applied both to the model output and reconstructions.

AC5: Thank you for pointing out these inconsistencies; we have standardised our terminology throughout the manuscript.

RC6: Line 16: Models also provide physically consistent simulations of multiple climate variables.

AC6: We have amended Lines 14-16:
"While proxies are thought to provide the 'gold standard' in reconstructing the local environment, they only provide point estimates for a limited number of

locations. On the other hand, models have the potential to afford more extensive and standardised geographic coverage of multiple bioclimatic variables."

And reiterated this point later in the manuscript on Lines 59-61:
"Model output have the potential to overcome these shortfalls, providing tangible values for parameters such as temperature, precipitation, and a range of derived bioclimatic indices (e.g., Hijmans *et al.,* 2005), that are consistent across variables for a more complete account of climatic conditions."

RC7: Line 21: "model output"

AC7: We have made this correction.

RC8: Line 22: I know this the Abstract, but I think the delta method needs to be described in a bit more detail. It's not the interpolation to a finer spatial resolution that's important, but the application of the long-term mean differences (present minus paleo usually) to high resolution observed modern data that produces results with greater spatial variability than that provided by the model.

AC8: We have edited Line 21-23:

"Here, we explore the impact of increasing the resolution of model output from 30 to 5 arc-minutes using the delta-downscaling method, which interpolates and applies the long-term difference between past and present model datasets to a higher resolution grid of observed present-day climate."

RC9: Line 20: Sufficient for what?

AC9: We have added further detail on Line 18-21:
'Most publicly available model time-series have been downscaled to 30 or 60 arc-minutes, but it is unclear whether such resolution is sufficient for certain applications like species distribution models, or whether this may homogenise environments and mask the spatial variability that is often the primary subject of analysis."

RC10: Line 49: I'm not sure what "an absolute, linear, and standardized representation" is.

AC10: We have edited this paragraph (Lines 49-59) to improve clarity on this point:

"Proxy data, while allowing for detailed reconstructions of climatic conditions through time, are rarely in direct association with archaeological or palaeontological sites, nor do they consistently provide an absolute, linear, and standardised representation of past climate across large geographic areas. In this sense, they often provide relative estimates of past climate, an issue highlighted in a synthesis of eastern African Late-Middle Pleistocene climate

records by Timbrell *et al.* (2022), demonstrating that different proxy records – even from within a relatively spatiotemporally restricted region – can provide alternate ideas of relative 'humidity'. This is the result of the diverse nature of the data employed (i.e., pollen, lake sediments, ice cores etc.), which record climate in an inconsistent way that typically cannot be articulated as the bioclimatic indicators and environmental parameters that are routinely in species distribution models (SDMs) (e.g. Beyer *et al.* 2021; Blinkhorn *et al.* 2022; Leonardi *et al.* 2022)."

RC11: Line 53: "variable nature" Variable in what sense? And I'm not sure what "data … cannot be articulated" means.

AC11: We have edited Line 56-59 to make it clearer:
"This is the result of the diverse nature of the data employed (i.e., pollen, lake sediments, ice cores etc.), which record climate in an inconsistent way that typically cannot be articulated as the bioclimatic indicators and environmental parameters that are routinely in species distribution models (SDMs) (e.g. Beyer *et al.* 2021; Blinkhorn *et al.* 2022; Leonardi *et al.* 2022)."

RC12: Line 57: Replace "Modelled data" by "Model output" or "Model simulations".

AC12: We have made this correction.

RC13: Line 64: I'm not sure what "estimation of ecologies experienced on the ground" means. Are you perhaps referring to applying model output to a species distribution model?

AC13: We have edited Line 65-69 to clarify:
"Resultant differences can be in the order of several degrees for temperature and tens of percent for precipitation, which could lead to substantially different biome classifications and estimations of ecologies experienced (Kottek et al., 2006). Such variations can have important implications for the diverse fields employing model output for the reconstruction of past and present species distributions, dispersal and extinction processes, and biogeographic patterns."

RC14: Line 65: This sentence essentially says that the spatial variation of simulated climate is lower than that of real-world climate, which has already been said several times.

AC14: We agree that this is repetitive and so have removed it as suggested.

RC15: Line 69: These two sentences don't follow. The cost of high-spatial resolution simulations don't have anything to do with the interpolation approaches discussed in the rest of the paragraph.

AC15: We have amended Lines 69-88 to highlight that we are referring to the production of a large number of time slices (which is what we tend to use for our

analyses in archaeology and palaeoecology), and add further information regarding dynamical downscaling based on the above suggestion:

"High resolution simulations of multiple time slices are often desired by consumers of model output yet difficult to obtain due to computational costs. For example, dynamical downscaling allows for the detailed description of processes in the climatic system and can improve the capturing of localised climatic conditions (Rummukainen, 2016; Strandberg et al., 2023), however this method is rarely applied in fields like palaeoecology and archaeology due to the computational costs, particularly when a large number of time steps are required. Most of the recently produced time series of palaeoclimate outputs have been downscaled from the native resolution of the models (usually in the order of 2 or 3 arc-degrees) to a higher resolution of 30 arc-minutes using statistical methods (Fordham et al. 2017; Beyer et al. 2020a; Krapp et al. 2021; Zeller and Timmerman 2024; Mondanaro et al. 2025) as these approaches can be more easily applied to several time periods."

RC16: Line 77: "delta-downscaling uses as map of local differences …" This would work, but in practice what is usually done is to calculate "experiment minus control" long-term mean differences on the model grid, which are then interpolated and applied to a higher resolution grid of observed present-day climate.

AC16: When applying the delta downscaling to a large time series of simulations (as described here for time series), we found it practical to define a single matrix of local differences that can then be applied to all the model outputs. The advantage of this approach is that the delta matrix can then be extended beyond the coastal boundaries of observations with a small amount of idw interpolation, which is only performed once, and then applied directly to the individual model time-steps. As the reviewer points out, for the current land-cover, the result is the equivalent whichever direction we approach the correction from. We now point out the two approaches on Lines 164-170:

"The resulting matrix only covers the land extent at the present. We then expanded this matrix to reach the largest land-extent in any of the times-steps under consideration using an inverse-distance-weighted interpolation. For most of the world, at the resolution of 30 and 5 arc-minutes, this only requires interpolating a small number of cells away from the coastline; for higher resolutions, other interpolating algorithms might be more appropriate. We note that the delta-downscaling can also be obtained by creating first the difference between model outputs, which is then applied to the observational model. However, such a direction is more computationally expensive, as the interpolation outside the coastlines would have to be repeated for each time step."

RC17: Line 83: I would refer to these as "interpolations" rather than "predictions".

AC17: We have made this amendment.

RC18: Line 88: This is the third "gold standard" invocation. Reconstructions can have considerable uncertainty attached to them, arising from multiple sources.

> **AC18:** We have adjusted and varied our language throughout the paper to highlight that proxies being the 'gold standard' reflects the general view of the field rather than our personal opinion, yet proxies are still associated with considerable uncertainty. We have added single quotation marks around 'gold standard' to indicate this, as well as made edits to the following:
>
> Lines 14-16: "While proxies are thought to provide the 'gold standard' in reconstructing the local environment, they only provide point estimates for a limited number of locations. On the other hand, models have the potential to afford more extensive and standardised geographic coverage of multiple bioclimatic variables."
>
> Lines 33-39: "Proxy records, such as those derived from pollen or other biomarkers, tend to be the preferred method for characterising past environments at specific locations; however, in order to extrapolate beyond the individual core sites and across wider regions, often it is necessary to rely on modelled or simulated climatic conditions."
>
> Lines 99-102: "Proxies offer a more localised account of climate in certain places, yet they too can be associated with high degrees of uncertainty, arising from multiple sources. Nonetheless, determining model agreement with empirical reconstructions from proxies remains a widely applied method for ground-truthing downscaled climatic output."

RC19: Line 101: "further downscaling" Further from what?

> **AC19:** The model output published by Beyer et al. (2020) has already been downscaled hence it was appropriate to say 'further' downscaling here. We note the resolutions targeted later in the section (i.e. we are further downscaling from 30 min to 5 min). To make this clearer, we have restructured Lines 115-117:
>
> "Given the ever-increasing demand to produce more accurate models of past climate across extended timeframes, we tested whether downscaling climatic models from a relatively coarser (30-min) to a higher resolution (5-min) leads to increased agreement with empirical reconstructions of past climate from proxies."

RC20: Lines 112-121: If I understand this correctly, you're using already downscaled model output (Beyer et al., 2020a) as the starting point, and further downscaling it. Wouldn't it be better to begin with the original HadCM3 output?

> **AC20:** We thank the reviewer for this suggestion. We have added the HadCM3 GCM from Huntley et al. (2022) to our analysis and find highly similar results with

that of the Beyer et al. (2020a) output. Pertinently, we also find no statistically significant differences in coherence with proxy records between the HadCM3 GCM model output at 30-min and at 5-min resolution. We have retained Beyer et al (2020a) since it is an easily accessible product that includes more sophisticated initial downscaling that takes advantage of a few runs of a high resolution GCM, and it is likely to be used by others in the future (particularly consumers of climatic models) as a starting point for further delta-downscaling.

RC21: Line 118: "National Center".

AC21: We have made this correction.

RC22: Line 127: See line 77 comment.

AC22: See AC16

RC23: Line 131: The terms in the equation should be defined. The equation reads like the Line 77 description of the delta method as opposed to the line 127 version. If all of the data were on the same grid, the approaches are in fact identical (as can be seen by rearranging the terms), but what did you actually do? Another issue is that the geographical location, x, is presumably a two-dimensional variable (in longitude and latitude), and so all the equation is illustrating is de-biasing, and not downscaling.

AC23: We have now defined the terms more clearly on Lines 172-190, and we hope that clarifies our approach.

"For temperature variables, the bias in a geographical location x (a cell with a given latitude and longitude) is given by the difference between present-day observed $T_{obs}(x,0)$ and simulated$T_{sim}^{\oplus}(x,0)$ temperature, interpolated to the desired higher resolution grid via bilinear interpolation. Downscaled temperature (T_{sim}^{DD}) in x at time t is thus estimated as

$$T_{sim}^{DD}(x,t) \coloneqq T_{sim}^{\oplus}(x,t) + \left(T_{obs}(x,0) - T_{sim}^{\oplus}(x,0)\right)$$

Precipitation is lower bounded by zero and covers different orders of magnitude across different regions compared to temperature. Multiplying rather than adding the bias correction is common when applying the delta method for precipitation, which corresponds to applying the simulated relative change to the observations (Maraun and Widmann, 2018). However, this method can therefore be hypersensitive in drylands, leading to overprediction of precipitation (and thus exacerbating the 'drizzling' bias of GCM). We have therefore adopted an additive approach for precipitation, analogous to the one used for temperature, with clamping within the range of observed maximum and minimum for current climate (see Beyer et al. 2020a). Like temperature, downscaled precipitation is estimated as

$$P_{sim}^{DD}(x,t) \coloneqq P_{sim}^{\oplus}(x,t) + \left(P_{obs}(x,0) - P_{sim}^{\oplus}(x,0)\right)\text{“}$$

RC24: Lines 134-139: How is “GCM drizzle” handled?

> **AC24:** To partially account for the drizzle problem, we have now adopted an additive approach for precipitation, analogous to the one used for temperature. As discussed in Beyer et al. (2020a), using an additive approach with clamping within the range of observed maximum and minimum for current climate, can help for avoiding extreme dampening of precipitation. We now mention this clearly in the text (see above).

RC25: Lines 152-158: The interpolation method needs to be better described. It’s implied that an inverse-distance weighted method was used, and that this can induce artefacts. Why was this method used, and not something else, like conservative remapping from the SCRIP package (https://github.com/SCRIP-Project/SCRIP)?

> **AC25:** The interpolation only has to deal with a few cells that emerge when sea level changes. We had explored different interpolation algorithms when we designed the approach that we used for Beyer et al. (2020a) and Krapp et al. (2021), but found very little difference in estimates, arguably due to the small number of cells that are interpolated. We agree that, if we were to go for even higher resolution, it might be better to consider other approaches, and have now pointed the reader to that possibility on Lines 164-167:
>
> “The resulting matrix only covers the land extent at the present. We then expanded this matrix to reach the largest land-extent in any of the times-steps under consideration using an inverse-distance-weighted interpolation. For most of the world, at the resolution of 30 and 5 arc-minutes, this only requires interpolating a small number of cells away from the coastline; for higher resolutions, other interpolating algorithms might be more appropriate.”

RC26: Line 198: “Considering that downscaling to higher resolutions is thought to capture localized climate dynamics…” Statements like this appear several times. I’m not sure that it’s “climate dynamics” that is being captured, but instead just simply spatial (mainly topographic) variations in climate.

> **AC26:** We have made this amendment on Lines 245-251 to clarify the hypothesis being tested in our statistical analyses:
>
> “Considering that downscaling to higher resolutions is thought to capture spatial variations in climate, we tested the statistical significance of differences in model-data coherence between lower resolution (30-min) and higher resolution (5-min) models, using a standard significance threshold of $p < 0.05$ via the Kruskal-Wallis non-parametric test.”

RC27: Line 204: “These analyses allow us to evaluate both the output of the climate models and the reliability of the proxy data in predicting specific climatic parameters in the past.” How is that possible. To evaluate the climate-model output, one would have to regard the proxy-based reconstructions as true, and to evaluate reliability of the proxy-based reconstructions, the model output would have to be regarded as true. Neither are.

> **AC27:** We thank the reviewer for making this distinction, which is an important one. We have amended Lines 152-153:
>
> “These analyses allow us to evaluate the coherence between the output of the climate models and the reconstructions of specific climatic parameters from proxy data…”

RC28: Line 213: “the most divergent variable on average is reconstructed mean annual temperature” This is somewhat of a surprise, given the global scope of the analysis. How does the performance here compare with other large-scale studies that examine present-day climate reconstructed using pollen data.

> **AC28:** We have added some discussion to this effect on Lines 264-270:
>
> “Considering the NRMSE, the most divergent variable on average is mean annual temperature, particularly for the output of the HadCM3 30-min model (Appendix A Tables A1-3). This result contrasts with other large-scale studies (Bartlein et al. 2011; Chevalier et al. 2021), potentially due to the assumptions made for the proxy reconstructions employed that modern analogues should be utilised from within 2000km around each site. Precipitation should be less affected given that it is more variable through space however temperature tends to be much more autocorrelated, meaning that much colder/warmer temperatures occurring in the past may not occur within these geographic limits.”

RC29: Line 220: “tends to estimate” But Beyer et al. (2020a) are downscaled simulations.

> **AC29:** We have clarified on Lines 149-155 that the problem is that Beyer et al. (2020a) was downscaled, and thus debiased, based on CRU, but the proxies that we use were calibrated with Worldclim2. The difference between these two observational databases can lead to a mismatch between the two, which is resolved by using the same observational dataset for both.
>
> “For this study, we delta downscaled and debiased these two datasets to a resolution of both 30 arc-minutes and 5 arc-minutes using modern observation from WorldClim2 (Fick and Hijmans, 2017). For the Beyer et al (2020a) dataset, as it was already at 30 arc-minutes, the delta downscaling at this resolution gives us a debiased version based on WorldClim2 rather than CRU. We used a global relief map from ETOPO2022 (NOAA National Center for Environmental Information, 2022) to reconstruct past coastlines following sea level change

(Spratt and Lisiecki, 2016). We select WorldClim2 as the modern reference as the transfer functions used in the LegacyClimate1.0 dataset were also derived from this dataset (at 30-minute resolution), allowing us to control for the effects of the modern data used for debiasing on our results."

RC30: Lines 220-245: I would expect to see here, or in the very short Section 4, some discussion of the source of the differences.

AC30: We try and keep Section 4 short and concise, however agree that the manuscript is lacking in discussion around the sources of the differences we find. We have added discussion about the potential reasons for differences between climatic parameters (see AC28), between regions, depending on landscape properties and chronology:

Lines 358-361: "Fig. 3 and Supplementary Fig. S2 highlight these spatial heterogeneities in bias across the Northern Hemisphere, which could have many potential different sources, i.e. geographic variation in the performance of the model outputs, the quality of the present-day calibration data for LegacyClimate 1.0 or the modern reference used for de-biasing, and/or the impact of confounding variables on the pollen-climate relationships."

Lines 403-407: "Our results also show that proxy reconstructions tend to indicate warmer temperatures at higher elevations and/or in areas of higher topographic roughness compared to model outputs and colder temperatures at lower elevations and/or lower topographic roughness (Appendix A Table A2). This is a known bias of transfer functions when constructing more 'extreme climates' from proxies, given that elevation negatively correlates with temperature and these functions rely on averages of data from modern calibration data sets (Chevalier et al., 2020)."

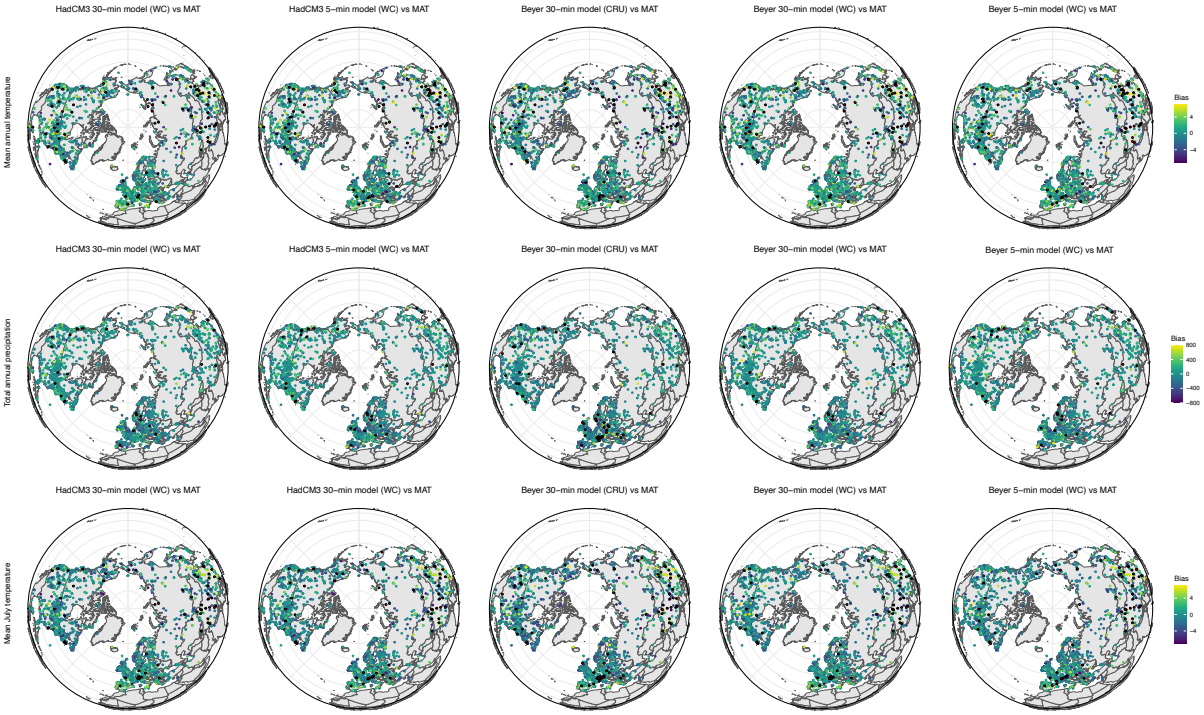
Lines 452-457: "Chronological uncertainties in the proxy age model may complicate the comparison between climate simulations and pollen-based records, as well as the process of signal smoothing via interpolation to facilitate analysis. Delta-downscaled models are also inherently tuned to replicate current rather than past climate patterns, and proxy reconstructions rely on the identification of modern analogue species that may have a different link to climate than palaeoecological communities, likely further contributing to higher divergence in older time periods (Chevalier *et al.* 2020)."

RC31: Section 3.1: Again, I would expect some attempt to explain the spatial variations. There are several sources that I imagine could play a role: spatial variations in the performance of the GCM, variations in the quality of the present-day calibration data for LegacyClimate, variations in the quality of the CRU and WorldClim data, impacts of confounding variables on the pollen-climate relationships.

AC31: Thank you for this comment. We have added in some discussion to this effect, based on the reviewers' helpful suggestions (see AC30):

RC32: Fig. 3: The figure is extremely difficult to read. There is a lot of useless white space between panels, and scales are unnecessarily duplicated. Also, I don't see any data from the Southern Hemisphere (or south of 20N?), which results in even more useless white space. What happened to the graticule over the Pacific? I think a polar-centered projection is fine, but it should fill the frame.

AC32: We have made these edits to Figure 3 to improve readability (by colouring outliers in red), as well as reduced white space and duplicated scales.



RC33: Line 292: "higher resolution models compared to those at relatively lower resolution" This implies multiple models, but line 114 refers to a single HadCM3 model.

AC33: We have amended this to specify that we are dealing with the equivalent model outputs at different resolution.

RC34: Fig. 4: What are the dots? What do you mean by "landscape dynamics"? Is the landscape changing in some way?

AC34: We have added to the figure caption that the dots are locations of proxy records studied in the analysis. The landscape dynamics are the spatial complexities revealed with increasingly high-resolution model, which you can clearly see in the figure. We have made the following amendments to the caption of Figure 4:

"Figure 4. Three regional examples of modelled mean annual temperature for the present day (bio01), demonstrating how downscaling increases spatial resolution by capturing the effects of landscape dynamics through space on

climate depending on the underlying topography. Geographic variability in temperature is shown, as simulated by the Beyer et al. (2020a) 30-min model output (CRU), Beyer et al. (2020a) 30-min model output (WC), and Beyer et al. (2020a) 5-min model output (WC), Locations of proxy locations from LegacyClimate 1.0 are shown as white circles."

RC35: Line 299: "… a known bias of transfer functions…" In addition to topographic effects, this bias also arises from "compression" in regression-based calibrations—the fact that the fitted values from less-than-perfect regressions always have lower amplitude than the observed values.

AC35: We have edited Lines 405-407 accordingly:

"This is a known bias of transfer functions when constructing more 'extreme climates' from proxies, given that elevation negatively correlates with temperature and these functions rely on averages of data from modern calibration data sets (Chevalier et al., 2020)."

RC36: Line 314: "time slice" I think a better term would be "time interval".

AC36: We refer to time slices or time steps as this is regular terminology used in our field when time series of climate reconstructions are used.

RC37: Line 326: The supplemental material I downloaded only contains Table S1.

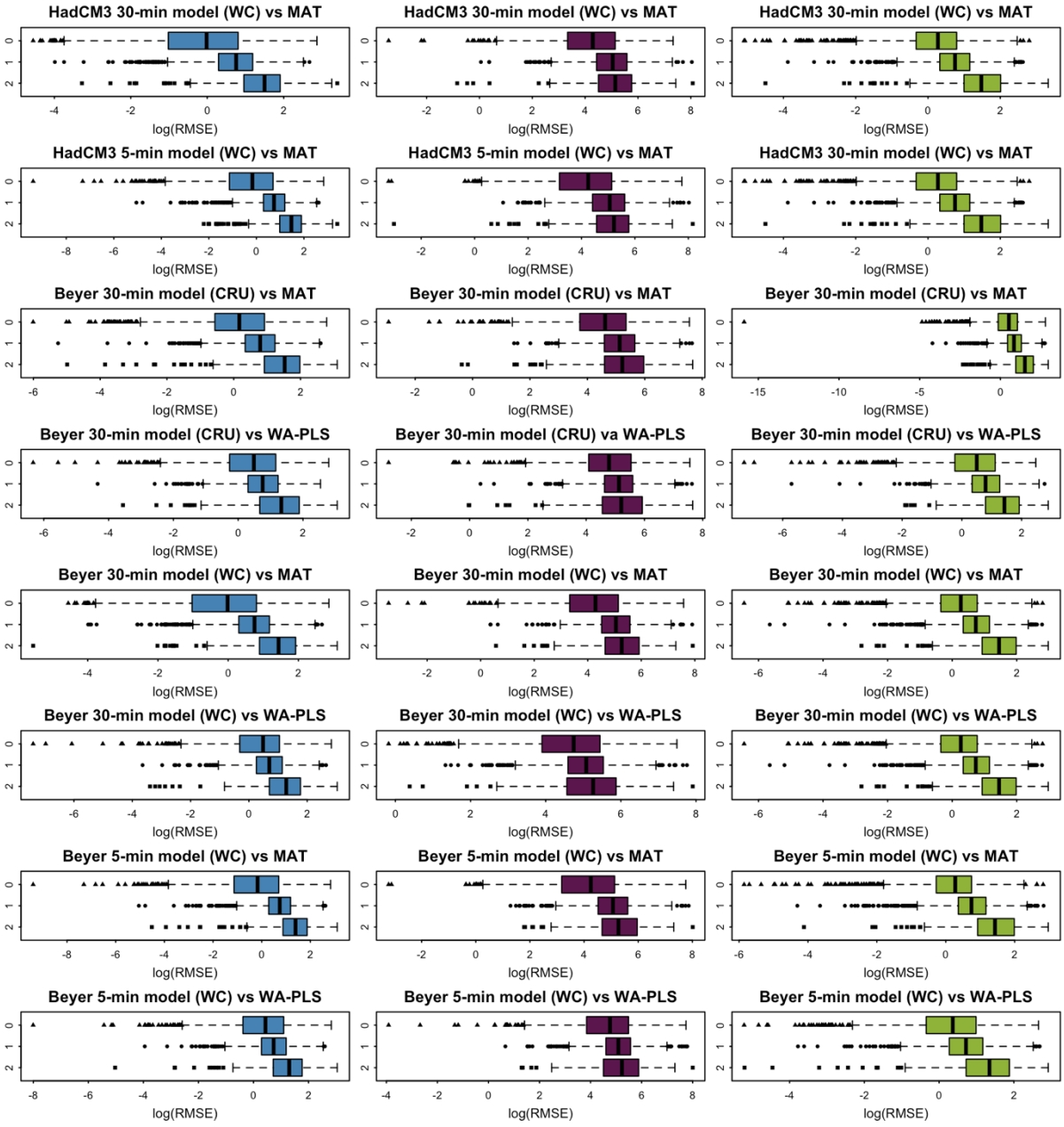
AC37: We apologise for the missing tables. The CoP editorial team requested that four tables from the manuscript were moved from the main text into the SOM due to formatting issues. A new version of the SOM was submitted, including these 4 supplementary tables, however it is unfortunate that this version was not shared with the reviewers nor uploaded online. We have now moved these tables to an Appendix (Appendix A), so that they are more easily accessible within the manuscript.

RC38: Line 334: "Models are also inherently calibrated…" If you're referring to GCMs, they are most definitely not calibrated in the sense that the term is used elsewhere in this paper.

AC38: We have edited Line 454-456:
"Delta-downscaled models are also inherently designed to replicate current rather than past climate patterns…"

RC39: Fig. 5: Labels are unreadable.

AC39: We have increased the size of the axis labels on this figure to improve readability.



RC40: Line 347: "Table 2" No Table 2.

> **AC40:** We apologise for this error; this reference was left over from the initially submitted manuscript (before we were requested to move tables to the SOM). This should now be "Appendix A Table A4", and has been changed accordingly.

RC41: Lines 353-362: There is no way to evaluate these statements without the supplementary tables. Also, there's no attempt to explain the results. An obvious candidate for poor performance of the reconstructions in the MIS 2 interval is low CO2, which, to my understanding was not considered in LegacyClimate.

> **AC41:** We were unfortunately unaware that the incorrect SOM had been uploaded, and have now submitted them in Appendix A for direct reference in the

manuscript. We thank the reviewer for this comment, and we have added further discussion of the results on Lines 485-499:

"Our results highlight that records spanning into MIS 2 consistently exhibit significantly higher proportions of divergent time series across all variables (Appendix A Table A4). The later may specifically be a consequence of low CO2 during MIS 2, which was not considered in LegacyClimate1.0, although this would mainly have an effect on moisture-related variables rather than temperature. Another potential source of divergence, leading to warmer reconstructions by proxies compared to the model outputs as well as significant deviations in precipitation, could derive from the geographic limits imposed on the LegacyClimate1.0 proxies for the modern samples used to perform reconstructions. This is particularly problematic for the LGM as comparable signals should be present within the modern climate space within the limit defined (2000km around each site), which is likely unreasonable for some areas (e.g. northerly areas of Europe, see Figure 1). Similarly, we find sites in Asia and higher altitude areas, where modern calibration data tend to be more limited, also have more divergent time series than expected given the sample size of this subset for all three variables (Appendix A Table A4). Sites in flatter areas exhibit significantly higher proportions of divergent time series for annual and July temperatures than expected by random chance, whereas sites in higher roughness locations and West North America are more highly divergent than expected in precipitation (Appendix A Table A4). Interestingly, we find that proxy records that capture the present day also occur in the most divergent subset more often than expected for annual temperature and precipitation, however this is because many of these records also span into later time periods (Appendix A Table A4)".

RC42: Line 366: "capture more signal" Jargon.

AC42: We are not sure why the reviewer refers to this as 'jargon' but have changed this to 'climatic trend' to vary the terminology with other sentences.

RC43: Line 376: "Beyer et al. (2020a) climate emulator" I don't understand. Beyer et al. is just downscaled and debiased data. "Climate emulators" are a different thing altogether.

AC43: This was an error, and we have changed this to 'climate simulations'.