

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 2:

RC1: This is a disappointing paper, because the issue of whether and how to downscale 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 in the introduction:

“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 downscaling. A level 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."

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, 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 approaches 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 in the Discussion:

"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:

“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:

“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 added this suggestion to this sentence:

“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

“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 this line in the abstract:

“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 here:

‘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 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 this sentence 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 this sentence 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 this section 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 in the text:

“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:

“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.”

“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.”

“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 the sentence:

“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, 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

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

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

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

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:

“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 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 this sentence:

“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:

“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 in the methods 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 Worldclim. 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.

“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:

“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.”

“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).”

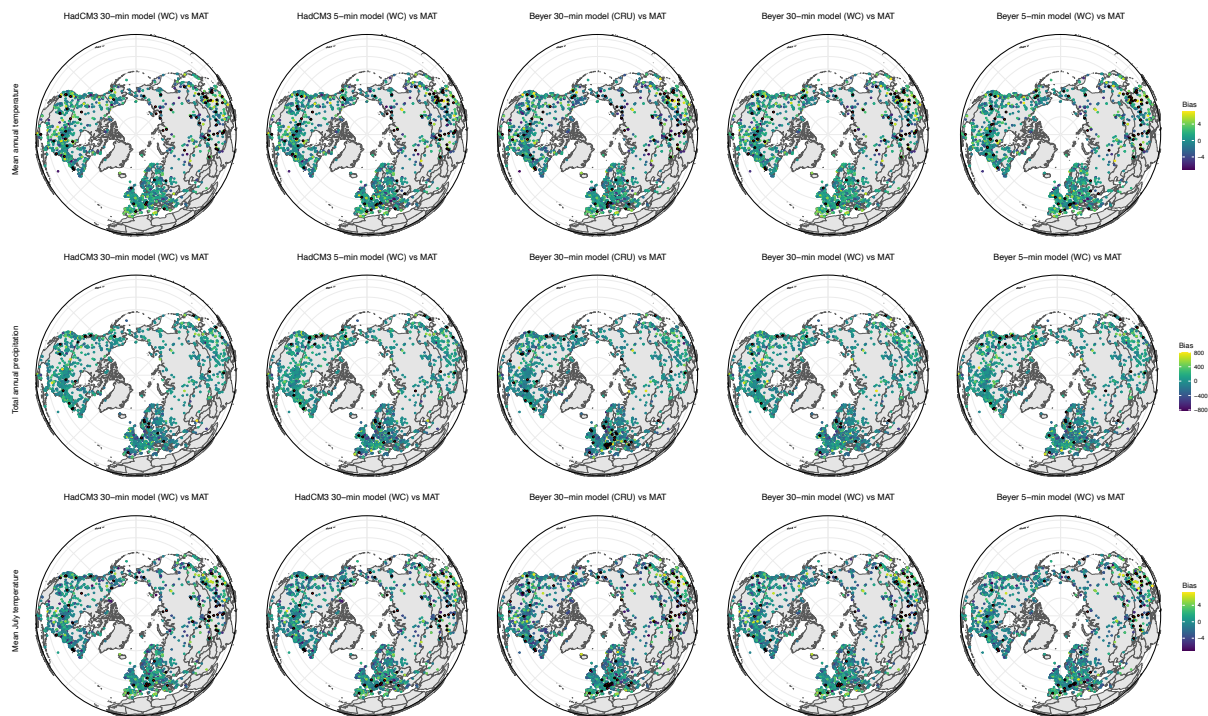
“Chronological uncertainties in the proxy age model may complicate the comparison between climate simulations and pollen-based records, as well as the process of signal smoothing via interpolation to facilitate analysis. Delta-downscaled models are also inherently tuned to replicate current rather than 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 this sentence 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.

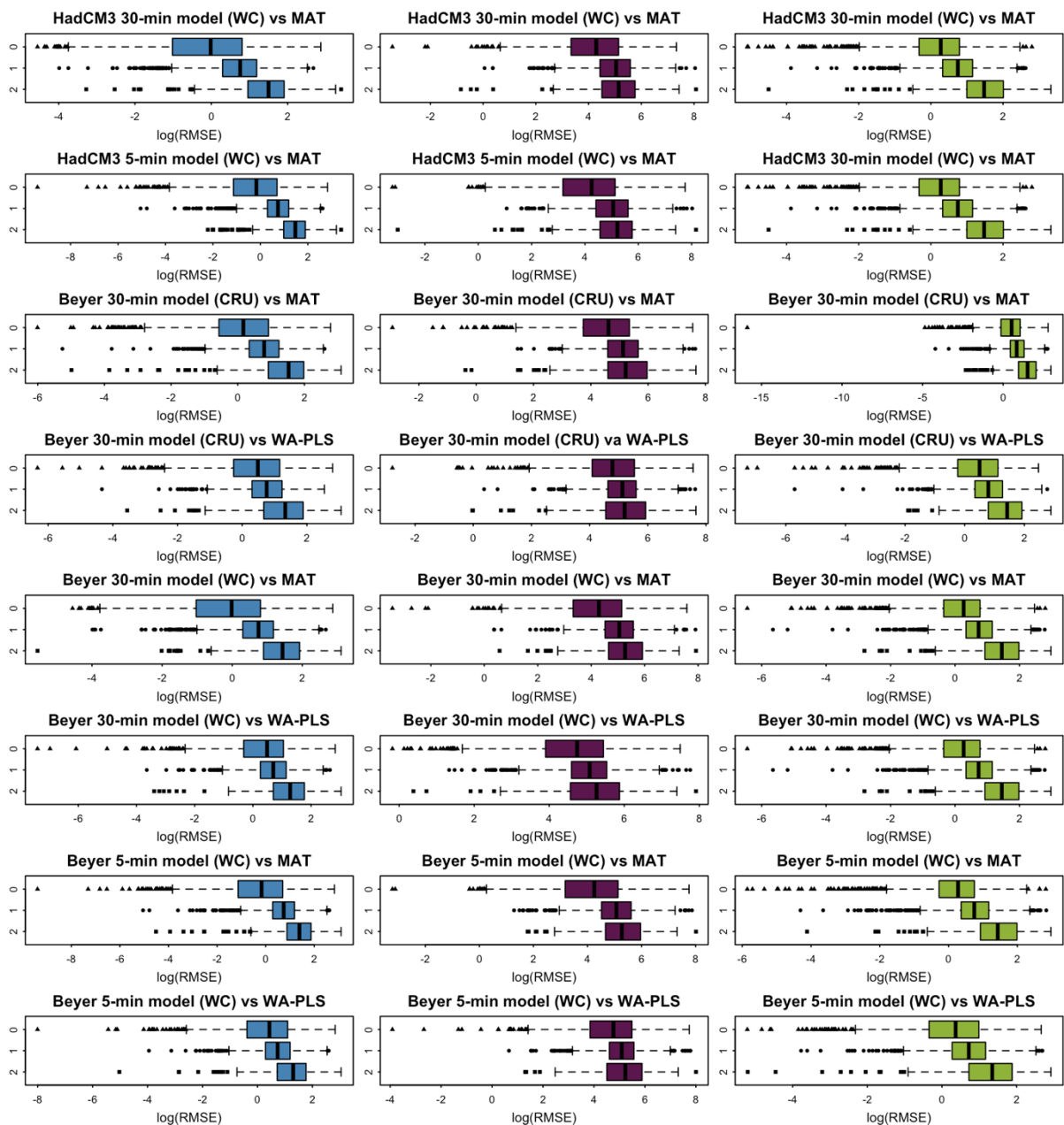
RC38: Line 334: “Models are also inherently calibrated...” If you’re referring to GCMs, they are most definitely not calibrated in the sense that the term is used elsewhere in this paper.

AC38: We have edited this sentence to read:

“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:

“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’.