Testing the reliability of global surface temperature reconstructions of the last glacial cycle with pseudo-proxy experiments

Jean-Philippe Baudouin[1], Nils Weitzel[1], Maximilian May[2], Lukas Jonkers[3], Andrew M. Dolman[4], and Kira Rehfeld[1,5]

[1]Department of Geosciences, University of Tübingen, Tübingen, Germany
[2]Institute of Environmental Physics, Heidelberg University, Heidelberg, Germany
[3]MARUM Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany
[4]Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar-und Meeresforschung, Potsdam, Germany
[5]Department of Physics, University of Tübingen, Tübingen, Germany

Correspondence: Jean-Philippe Baudouin (jean-philippe.baudouin@uni-tuebingen.de)

Abstract.

~~Reconstructing~~ Reconstructions of past variations of the global mean surface temperature ~~is~~ (GMST) are used to characterise the Earth system response to perturbations as well as validate Earth system simulations. ~~Reconstructing GMST beyond~~ Beyond the instrumental period~~relies on algorithms aggregating~~, reconstructions rely on local proxy temperature records and algorithms aggregating these records. Here, we propose to establish standards for ~~the evaluation of~~ evaluating the performance of such reconstruction algorithms. Our framework relies on pseudo-proxy experiments. That is, we test the ability of ~~the~~ an algorithm to reconstruct a simulated GMST, using artificially generated proxy data created from the same simulation. We apply the framework to an adapted version of the GMST reconstruction algorithm used in Snyder (2016), and the metadata of the synthesis of marine proxy records for temperature of the last $130\,\mathrm{kyr}$ from Jonkers et al. (2020). We use an ensemble of 4 transient simulations of the last glacial cycle or the last $25\,\mathrm{kyr}$ for the pseudo-proxy experiments.

~~We find the algorithmto be able to reconstruct~~ Given the dataset and the algorithm, we find that the reconstruction is reliable for timescales longer than $4\,\mathrm{kyr}$ ~~over~~ during the last $25\,\mathrm{kyr}$. However, beyond $40\,\mathrm{kyr}$ BP, age uncertainty limits the ~~algorithm capability~~ reconstruction reliability to timescales longer than $15\,\mathrm{kyr}$. ~~The main sources of uncertainty are~~ For the long timescales, uncertainty on temperature anomalies is caused by a factor, that rescales near global mean sea surface temperatures to GMST, the proxy ~~measurement~~ measurements, the specific set of record locations, and potential seasonal ~~bias~~ biases. Increasing the number of records significantly reduces all sources of uncertainty but the scaling. We also show that a trade-off exists between the inclusion of ~~a large number of~~ many records, which reduces the uncertainty on long time scales, and of only records with low age uncertainty, high accumulation rate, and high resolution, which improves the reconstruction of the short timescales.

Finally, the method and the quantitative results presented here can serve as a basis for future evaluations of reconstructions. We also suggest future avenues to improve reconstruction algorithms and discuss the key limitations arising from the proxy data properties.

1 Introduction

The global mean surface temperature (GMST) is a fundamental quantity to describe climate change. It is a major component of the Earth's energy balance and describes the ~~Earth system response to~~ response of the Earth system to external perturbations. Furthermore, the GMST is one of the variables characterising Earth's habitability ~~with consequences on ecosystems~~ and how ecosystems develop. Hence, GMST is the main target measure of international efforts to limit anthropogenic global warming (UNFCCC, 2015). Past evolution of GMST is used to evaluate Earth System Models (ESMs, e.g. PAGES 2k Consortium, 2019; Brierley et al., 2020; Kageyama et al., 2021; Lunt et al., 2021) and to estimate the equilibrium climate sensitivity, i.e. the response of GMST to a doubling of atmospheric CO_2 concentrations (Snyder, 2016; Friedrich and Timmermann, 2020).

~~In order~~ Comprehensive ESMs are being developed to project climate change for the end of the century and beyond~~, comprehensive ESMs are being developed~~. Some of these ESMs include an atmosphere-ocean general circulation model along with modules for ice sheets (Nowicki et al., 2016; Ziemen et al., 2019; Muntjewerf et al., 2020), iceberg melting (Rackow et al., 2017; Erokhina, 2020), dynamic vegetation (Song et al., 2021), and the carbon cycle (Brovkin et al., 2012; Arora et al., 2020; Kleinen et al., 2021). During the Last Glacial Cycle (LGC, the last $130\,\mathrm{kyr}$, which includes the Last Interglacial (LIG), the last glacial period, and the Holocene), large changes in these Earth system components occurred in response to radiative forcing changes of comparable magnitude to those projected in future scenarios. Therefore, the LGC is an important evaluation period for ESMs.

Reconstructing past GMST relies on proxy records extracted for example from sediments or ice. Many methods for reconstructing the GMST over the LGC use a bottom-up approach, where ~~a large number of~~ many local proxy temperature records are combined to compute an area-averaged temperature (e.g. Shakun et al., 2012; Snyder, 2016; Friedrich and Timmermann, 2020; Osman et al., 2021; Clark et al., 2024). This approach aims ~~at suppressing~~ to suppress noise and regional influences in these time series but remains sensitive to the spatial and temporal coverage of the data. ~~It also relies on a homogeneous dataset for consistency between the records. For the LGC, a new database has recently been published: the *PalMod 130k marine palaeoclimate data synthesis* (Jonkers et al., 2020, hereafter J20). This database provides numerous proxy records of near-surface sea temperature from marine sediments. The records are subject to uncertainties from dating and proxy-to-temperature relationship (Bradley, 2014). In particular, several processes influence the relationship between the Sea Surface Temperature (SST) and the measured proxy value, such as other variables than SST, bioturbation, and sampling biases. These limitations lead to uncertainties that need to be accounted for in any GMST reconstruction . The J20 database is metadata-rich, which enables the characterisation of these uncertainties.~~

~~Compared to the Common Era or the Holocene, for which many methods have been developed and reconstructions computed (see for example Consortium, 2017; PAGES 2k Consortium, 2019; Kaufman et al., 2020), the LGC presents the challenge of a non-stationary climate state and sparser records. Snyder (2016, hereafter S16) developed an algorithm to overcome this challenge and provides a GMST reconstruction of the Pleistocene to investigate glacial cycles. It is a robust algorithm which , compared to others, accounts for data sparsity, quantifies uncertainties and has a limited reliance on model output. Therefore, this algorithm is a good basis for developing evaluation standards and assessing reconstruction algorithm performances and~~

~~limitations~~To our knowledge, a proper evaluation of such reconstruction methods hasn't been performed, which limits their use for model evaluations. In particular, we note several open questions regarding the ~~performances of algorithms such as S16 to reconstruct the GMST during~~ quality of GMST reconstructions of and beyond the LGC:

60
- How reliable are GMST reconstructions of the LGC~~based on SST proxies~~?

- Does the non-uniform spatiotemporal distributions of proxy samples lead to biased GMST reconstructions?

- ~~What are the contributions of the various~~ Which sources of uncertainty impact reconstructions, and in what way?

- What is the shortest timescale on which amplitudes and timing of GMST variations can be accurately reconstructed, for a given algorithm and dataset?

65
- What are the sources of the loss of accuracy on short timescales?

- Which limiting factors should be prioritised for improving the GMST reconstruction quality (uncertainty, resolution)?

~~We study these questions using~~ This study tests a reconstruction method similar to the one published in Snyder (2016, hereafter S16). This algorithm was used to compute a GMST reconstruction of the Pleistocene based on SST proxy records to investigate glacial cycles. It is robust enough to tackle and account for the non-stationarity of climate states and the sparsity of the records.

70 Compared to other algorithms proposed to reconstruct GMST on glacial-interglacial timescales, it quantifies uncertainties more comprehensively and has a limited reliance on model output. Therefore, we use it as the basis of our work, although most of the findings will likely hold for other algorithms using similar principles for the reconstruction.

This study does not provide a new reconstruction, but an evaluation method based on pseudo-proxy experiments (PPEs) ~~based on~~ and transient climate simulations. PPEs enable the test of ~~algorithms~~ an algorithm in controlled, idealised environ-

75 ments, where the underlying truth is known (i.e. the simulation outputs). They rely on pseudo-proxies that are synthetic time series imitating proxy records, based on the spatiotemporal climate state of the simulation. The computation of pseudo-proxies is performed using a proxy system model (*sedproxy*, Dolman and Laepple, 2018), which simulates the processes from the fixation of the climate-dependent quantity to the measurement of the proxy in the lab, including the entailed uncertainties. ~~The use of metadata from the J20 database to compute pseudo-proxies provides a realistic example for GMST reconstruction.~~

80 ~~By applying the reconstruction algorithm to various sets of pseudo-proxy data, we can evaluate the performance of the reconstruction method, as well as the quality of the potential LGC reconstruction based on the J20 dataset~~To create realistic pseudo-proxies, we use the metadata of a recently published database: the *PalMod 130k marine palaeoclimate data synthesis* (Jonkers et al., 2020, hereafter J20). This database provides numerous proxy records of near-surface sea temperature from marine sediments. It is metadata-rich, which enables us to test the impact of various uncertainty sources on the reconstruction,

85 and whether these are well accounted for in the algorithm's uncertainty estimates.

In the following, we first describe the J20 database and the transient climate simulations employed to compute the pseudo-proxies. Then, we present the GMST reconstruction algorithm that we evaluate and the design of the PPEs. The results from the PPEs follow it. Finally, we discuss the ~~performance of the algorithmwhen applied to~~ algorithm's performance when using the J20 ~~database~~metadata, and future avenues to improve GMST reconstructions.

2 Data

2.1 Temperature reconstructions

Our study relies on one of the largest ~~synthesis~~ syntheses of marine proxy records for temperature of the LGC available: the extended *PalMod 130k marine palaeoclimate data synthesis* ~~(Jonkers et al., 2020, hereafter J20)[1].~~, beta version 2.0.0 (Jonkers et al., 2020, her . This dataset is a multi-proxy compilation of globally distributed marine proxy records spanning the last $130\,\mathrm{kyr}$ and was developed within the *PalMod* initiative ~~(Latif et al., 2016; Fieg et al., 2023)[1]~~ (https://www.palmod.de; Latif et al., 2016; Fieg et al., 2023) as a comprehensive reference for transient climate simulations. We use various metadata from this database: core locations, proxy types, species (for Mg/Ca), sedimentation rates, and age ensembles. We prefer this database over the one assembled for S16 ~~as it enables a more comprehensive uncertainty assessment, needed to develop our evaluation standards~~ where only locations, proxy types and mean chronologies are available. In particular, ~~it includes richer metadata and~~ the effect of bioturbation can be computed using sedimentation rates or the impact of habitat preferences using species. The J20 database also provides a unified framework for the age models. ~~The age-depth models~~ These are constrained using a blend of radiocarbon dates (between 0 and $\sim 40\,\mathrm{kyr}$ BP) and δ^{18}O benthic stratigraphy based on Lisiecki and Stern (2016). Chronological uncertainty is assessed using the Bayesian framework BACON (Blaauw and Christen, 2011). Specifically, the database contains 1000 age ensemble members for each sediment core. We do not use the temperature records in J20, as our study does not aim to present a new GMST reconstruction.

We selected all proxy records for SST located within the latitudes $60°$N and $60°$S, following S16, with at least 10 values. This leaves us with 265 time series at 189 locations. These records rely on temperature dependencies of the chemical composition, chemical product, and the temperature preferences of various organisms living near the sea surface. Here, this includes alkenone indices (U^{k}_{37}, $U^{k'}_{37}$, 89 time series, Prahl et al., 1988), a lipid-based index (TEX$_{86}$, 6, Schouten et al., 2002), the long chain diol index (LDI, 4, Rampen et al., 2012), Mg/Ca ratios in planktonic foraminifera (103, Nürnberg et al., 1996), and microfossil assemblages (planktonic foraminifera, radiolaria, and diatoms, 63, Imbrie, 1971). In addition, we only kept the time series for which the representative season was indicated as annual (105 time series) or unspecified (112). In some cases, assemblages from the same core have been used to estimate winter and summer temperatures. If the timestep coincides, we average the two to form 48 "pseudo-annual" time series.

Record metadata show some distinctive features (Fig.1 and 2). First, record locations tend to better cover the Atlantic than the Indian and Pacific oceans, and to be close to the coast (Fig.1-A). Second, the number of proxy records is maximum for the deglaciation (>200), while it is relatively constant between ~~40~~ 40 and $120\,\mathrm{kyr}$ BP (82, Fig. 2-A). Yet, substantial data gaps occur in the Southern Hemisphere for periods prior to the LGM (Fig.1-B). Third, the lowest age uncertainty is achieved in the last $30\,\mathrm{kyr}$ (<$3\,\mathrm{kyr}$), where radiocarbon dating is reliable and frequently employed. For the rest of the LGC, median age

---

[1] ~~We have used the beta version 2.0.0 of the database, that was made available to us on https://seafile.zfn.uni-bremen.de/d/9339a9ae27e44b208e85/. It incorporates many new records, an updated age model, as well as a few bug fixes compared to the original version published in Jonkers et al. (2020)~~

[1] ~~https://www.palmod.de/~~

120 uncertainty is slightly below $10\,\mathrm{kyr}$. Finally, the highest temporal resolution and the highest accumulation rate [1] are also found in the last $30\,\mathrm{kyr}$ ($< 1\,\mathrm{kyr}$). Note that a high accumulation rate leads to a bioturbated layer covering a shorter period of time.

2.2 Climate model output

To compute the pseudo proxy time series used in the PPEs, we use transient simulation outputs from four different climate models: FAMOUS, LOVECLIM, CESM, and MPI-ESM. This ~~model-model~~ multimodel approach is used to test the robustness
125 of the result.

FAMOUS, the Fast Meteorological Office and UK Universities Simulator, is a version of the Hadley Centre Coupled Model version 3 (HadCM3, Gordon et al., 2000) AOGCM with reduced resolution (Smith et al., 2008). The simulation used here is the version ALL-5G produced in the QUEST project, and covers the last $120\,\mathrm{kyr}$ (Smith and Gregory, 2012). The simulation is forced with transiently changing orbital parameters (Berger, 1978), greenhouse gas concentrations (Spahni et al., 2005; Lüthi
130 et al., 2008), and Northern Hemisphere ice sheet extents and topographies from the ICE-5G v1.2 dataset (Peltier, 2004)[1]. The time-varying external boundary conditions are accelerated by a factor of ten in the simulation, such that the last $120\,\mathrm{kyr}$ are effectively simulated in $12\,\mathrm{kyr}$ in the model (Smith and Gregory, 2012).

LOVECLIM (LOch–Vecode-Ecbilt-CLio-agIsm Model) is an Earth system model of intermediate complexity (Goosse et al., 2010), that is, with coarser spatial resolution, and simpler representation of physical processes than in general circulation mod-
135 els. The simulation used here (Loveclim 800k, Timmermann and Friedrich, 2016) uses version 1.1 of LOVECLIM (Goosse et al., 2007), with only the atmospheric, oceanic, and vegetation modules activated. Greenhouse gas concentrations are prescribed following the measurements from the EPICA DOME C ice core (Lüthi et al., 2008), and the evolution of ice sheets (surface elevation and albedo, land-sea mask) are prescribed following a simulation with the CLIMBER-2 model (Ganopolski and Calov, 2011). Finally, orbital configurations are derived from Berger (1978). Similarly to FAMOUS, the external forcings
140 are accelerated by a factor of five.

Several simulations of the last $26\,\mathrm{kyr}$, using MPI-ESM 1.2 (Mauritsen et al., 2019), have been published in Kapsch et al. (2022). We select two simulations from this paper. The first simulation (called GLAC1D-P2 in Kapsch et al. (2022), hereafter MPI-G1D) is forced using the GLAC-1D ice sheet reconstruction (Tarasov et al., 2012; Briggs et al., 2014). By contrast, the second simulation (ICE6G-P2 in Kapsch et al. (2022), hereafter MPI-I6G) uses ICE-6G (Peltier et al., 2015). Forcings are
145 otherwise identical (Köhler et al., 2017; Berger, 1978, for greenhouse gases and insolation respectively). These simulations also include meltwater flux forcing from ice sheet melt through a dynamical routing.

Finally, we use a simulation of the last three million years (Yun et al., 2023), based on CESM 1.2. (Hurrell et al., 2013). For our period of interest, it uses as boundary conditions LGM land-sea mask, greenhouse gas concentrations from EPICA DOME C ice core (Lüthi et al., 2008), orbital configurations from Berger (1978), and the ice sheet elevation and albedo from
150 a CLIMBER-2 simulation (Willeit et al., 2019). The time-varying boundary conditions are accelerated by a factor of five. Open-access data availability is limited to millennial averages, which restricts the use of this simulation in our analysis.

---

[1]~~a high accumulation rate lead to a bioturbated leading covering a shorter period of time~~

[1]The land-sea mask is that of the pre-industrial condition in this simulation, so the ice sheet topography is only applied to the pre-industrial land surface

3 Methods

We first present the GMST reconstruction algorithm, which is an adaptation of S16 (sect. 3.1). Then, we describe the design of PPEs to evaluate the reconstruction quality (sect. 3.2).

3.1 Global mean surface temperature reconstruction

3.1.1 The S16 algorithm

The original S16 algorithm reconstructs GMST over the Pleistocene following four main steps.

First, all the irregular raw time series are brought to a common, equidistant ($1\,\mathrm{kyr}$) time axis. Then, the time series are smoothed using a Gaussian kernel which represents the effect of age uncertainty (a constant $10\,\mathrm{kyr}$, as the 95% interval). The algorithm also estimates the signal's standard deviation arising from age, measurement, and calibration uncertainty. The measurement and calibration uncertainty is defined as a standard deviation of $1.5\,\mathrm{K}$, a value similar to other reconstruction studies (Shakun et al., 2012; Tierney et al., 2020). The SST anomalies are finally computed with respect to the late Holocene (0 to $5\,\mathrm{kyr}$ BP).

Second, the considered domain between $60°S$ and $60°N$ is subdivided into latitudinal bands (with 9 different configurations), used to cluster the records. The objective is to calculate SST anomalies for each band. Therefore, this latitudinal subdivision is based on the assumption that temperature anomalies are approximately homogeneous within each band (Rohling et al., 2012). This method is a trade-off to still account for the heterogeneity of the record locations, despite the sparsity of the available SST records. At this stage, a Monte Carlo approach is used to propagate the signal uncertainty from each record to the zonal band averages. Specifically, independent realisations of the signal uncertainty are used for each record to compute one thousand zonal band averages, for each of the 9 band configurations. Then, n timeseries are randomly selected from all the time series available (the number of records times the one thousand realisations of the previous sampling), where n is the number of records. That way, the algorithm can evaluate the uncertainty related to the selection of records, and in particular to their location.

Third, the zonal band averages are aggregated to derive a $60°S$ to $60°N$ SST anomaly.

Fourth, the mean SST anomaly is scaled to a GMST anomaly. Similar to other studies (Bereiter et al., 2018; Friedrich and Timmermann, 2020), a linear scaling coefficient is used to account for the stronger cooling of polar and terrestrial regions compared to the spatial mean SST change between $60°S$ and $60°N$. The coefficient is derived from PMIP2 and PMIP3 LGM and pre-industrial simulations (Braconnot et al., 2007, 2012). A Monte Carlo approach is again used to quantify the uncertainty: five thousand different realisations of the scaling factor (defined as a Gaussian distribution with a mean of 1.9 and a standard deviation of 0.2) are applied to randomly selected realisations of the mean SST anomaly.

3.1.2 Adaptation of S16 algorithm

We decide to adapt the S16 algorithm to make full use of the J20 dataset (age ensembles, records with no data in the last 5 kyr), and improve the uncertainty quantification by the algorithm.

First, we apply the measurement and calibration noise before the interpolation of the raw time series. That way, the uncertainty depends on the number of measurements performed: averaged over the same time span, a high-resolution record should have a reduced uncertainty compared to a low-resolution one. We consider a thousand different samples of this noise to quantify the uncertainty.

Second, we make use of the age ensembles available for each record in the J20 database: each time series is interpolated using one of the 1000 age ensemble members. A different realisation of the measurement and calibration noise is used for each of these time series. This method lends itself to the Monte Carlo approach used later in the S16 algorithm and avoids any assumption on the distribution of the age uncertainty. We also use a centennial timestep for the interpolation, so that the higher frequency variability of the reconstructed GMST can be investigated.

Third, 96 records from the J20 database do not cover the reference period $0\text{-}0\text{-}5$ kyr BP, from which the SST anomaly is computed[2]. Including them increases the number of available records for the period $40\text{-}40\text{-}120$ kyr BP by 60%. Therefore, we implement a zonal iterative offsetting procedure. We assume that the SST anomalies are homogeneous enough within each latitudinal band, so that the offset needed to compute the SST anomaly of one record can be derived from the others. In each of the zonal bands, the time series are adjusted iteratively, with respect to all previously adjusted time series, according to the following steps:

1. Preliminary step: all records in the considered zonal band are sorted by the age of their most recent data point, starting with the closest to the present.

2. First offsetting: the temperature anomaly of the record including the most recent age is computed with respect to the most recent 5 kyr of that same record.

3. Iterative offsetting: we now consider the interval consisting of the most recent 5 kyr of the next non-adjusted record with the most recent age. We compute the mean of all the already adjusted time series in the zonal band over this interval. The offset of the new record is then calculated such that this mean equals the average temperature anomaly of the new record over the same interval.

Fourth, all latitudinal bands are averaged using an area-weighted mean where the weight is proportional to the oceanic surface[2], instead of the entire surface of the band in S16. For this, we use the ICE-6G reconstruction (Peltier et al., 2015) of the LGM for the entire time period, assuming that changes in oceanic surfaces in each band are small enough compared to other uncertainty sources to be neglected.

---

[2]S16 also considers 8 records which have data between 5 and 8 kyr BP. That concerns 26 records in J20

[2]We use the ICE-6G reconstruction (Peltier et al., 2015) of the LGM for the entire time period, assuming that changes in oceanic surfaces in each band are small enough compared to other uncertainty sources to be neglected.

Finally, as in S16, the resulting Monte Carlo realisations are used to compute summary statistics of the GMST reconstruction, namely mean and 95% confidence interval. However, in our case, each GMST time series does not span exactly the same time period, because they are based on different sets of age models. Therefore, the statistics are computed only when values from at least half of the realisations are available.

3.2 Pseudo-proxy experiments

We evaluate the algorithm using pseudo-proxy experiments (PPEs). The general design of a PPE is presented in fig. 3. Pseudo-proxies are artificially generated proxy data created from a reference climate evolution, in our case, temperature outputs from transient simulations. Thus, we apply the GMST reconstruction algorithm to these pseudo-proxies to analyse how well the reference GMST evolution, i.e. the GMST of the transient simulations, is recovered by the reconstruction algorithm. In this idealised setting, influences of individual aspects of the algorithm or characteristics of the data can be studied using sensitivity experiments. In the following, we first describe the strategy to construct pseudo-proxies before presenting the design of the PPEs we perform.

3.2.1 Construction of pseudo-proxies with *sedproxy*

Pseudo-proxies are created by simulating the processes that contribute to the recording of a climate signal in an archive. Here, we use the *R* package *Sedproxy sedproxy* to simulate these processes for marine sediment records (Dolman and Laepple, 2018, 2023). More specifically, the package provides a forward model that converts an ocean temperature time series with monthly resolution, into pseudo-proxy time series with the desired resolution. Table 1 summarises the processes simulated for each proxy type and the associated metadata. The algorithm implemented in *sedproxy* consists of four steps.

First, the input temperature data is converted to the relevant proxy unit using well-established calibrations (forward model). This forward model represents the dependency of a process to ocean temperature (the sensor), such as the calcite formation in foraminiferal tests, which uses both Ca or Mg ions, in a proportion depending on the temperature (Nürnberg et al., 1996). This model corresponds to the inverse of the calibration function, used to infer temperature from proxy measurements. This step is needed to consider the structure of the calibration error. Note that the calibration for TEX_{86} and LDI is not yet implemented in *sedproxy*. Given that it concerns only ten records, the impact on our analysis should not be significant. In addition, calibration methods used for assemblages in J20 are heterogeneous, and no established forward models, as needed in this first step, exist. Hence, no conversion is performed for these 63 records either. This limits insight into the error structure, but it cannot be overcome with our current framework.

Second, for each time point (in fact, sediment depth) for which the proxy record is to be modelled, the algorithm computes monthly weights. These weights represent the chance for the sensor to have recorded the temperature of that specific month and year. The weights combine effects from seasonal preferences of species, bioturbation, and sediment thickness of the sample (for an exhaustive description, see Dolman and Laepple, 2018). In *sedproxy*, seasonal preferences overrepresentation of high abundance periods, such as particular seasons, are only implemented for Mg/Ca ratios, although models for alkenones do exist (Tierney and Tingley, 2018). These preferences are based on the recording species indicated in J20, and computed in *sedproxy*

using the FAME v1.0 model (Roche et al., 2018). For 16 of the records, no seasonal preference is applied, as no growth model is provided for the associated species (namely N. incompta, G. inflata, G. truncatulinoides, P. obliquiloculata and G. crassaformis). Bioturbation is modelled as a low-pass filter (Berger and Heath, 1968). It supposes a constant bioturbated layer width, which is fixed at $10\,\mathrm{cm}$ in our study ~~(default value, cf. Zuhr et al., 2022)~~(default value, cf. Boudreau, 1998; Zuhr et al., 2022). While other studies have suggested lower values (Teal et al., 2008; Zhang et al., 2024), we consider this a conservative approach to our analysis. It also requires sedimentation rate as input, which we derive from the depth and mean age available in J20[2]. The impact of the sediment thickness of the sample is also modelled at this stage. In our study, it is computed from the difference between two sample depths, but not higher than $2\,\mathrm{cm}$. Finally, for Mg/Ca ratios, *sedproxy* samples 30 pseudo-proxy values whose temporal distribution follows that given by the monthly weights. This sampling corresponds to the number of foraminifera needed to produce a measurement ~~(default value, Dolman and Laepple, 2018).~~(30 is the default value in Dolman and Laepple, ~~sedproxy~~*Sedproxy* also adds to these values an error corresponding to the inter-individual variability (standard deviation of $2\,\mathrm{K}$, Dolman and Laepple, 2018), and eventually takes the average over the 30 samples. This sampling process is irrelevant for both $U^{K'}_{37}$ and TEX_{86}: the sample size is considered infinite, and the pseudo-proxy values are based on a weighted mean using the monthly weights. For assemblages, this sampling would be relevant but what is implemented in *sedproxy* is too simple to represent it and dependent on a forward model (cf. previous paragraph), and therefore not considered.

The third stage of the algorithm consists of applying a random Gaussian noise depending on the measurement error. This one is set to $0.26\,\mathrm{K}$ and $0.23\,\mathrm{K}$ for Mg/Ca ratios and $U^{K'}_{37}$ respectively (Dolman and Laepple, 2018). Unfortunately, no standard measurement errors are available for TEX_{86} and assemblages, so we arbitrarily set it to $1\,\mathrm{K}$. This large uncertainty aims to quantify the impact of not using forward models for the temperature to proxy calibration.

The fourth step of the algorithm consists of converting the time series from proxy unit back to temperature. Calibration uncertainty, which arises from using empirical relationships inferred from imperfect data, is only considered at this step. This is done by sampling the calibration parameters using a Gaussian distribution.

In addition to, and before applying *sedproxy*, we add a step to account for age uncertainty. Specifically, we randomly select one age member among the one thousand available, so that *sedproxy* computes the pseudo-proxy values at these specific time points. The resulting time series is then associated with all the age ensemble members available.

Finally, the input data for *sedproxy* corresponds to the temperature of the oceanic grid point nearest to the proxy location. For LOVECLIM, MPII6G and MPIG1D, we use the topmost level of ocean temperature. ~~In addition, we~~ We use air temperature for FAMOUS and surface temperature for CESM instead, with a screen for land mass and values below -2°C, so that the seasonality can be investigated.

3.2.2 Specification of the pseudo-proxy experiments

Different sets of PPEs are designed, first to evaluate the performance of the algorithm, and then to test the influence of various data properties on the reconstruction. The characteristics of the pseudo-proxies computed for each PPE are summarised in table 2. In the following, we further describe the method to compute the pseudo-proxies and discuss the rationale of each experiment.

---

[2]~~Individual age members cannot be used due to a too skewed distribution of the sedimentation rates in BACON~~

Table 1. Processes simulated in *sedproxy* for each type of proxies. The values are those suggested in Dolman and Laepple (2018). The parameter for the bioturbation is the bioturbated layer width.

Proxy type	Mg/Ca	$U^{K'}_{37}$	TEX_{86}	LDI	Assemblages
Number of records	103	89	6	4	63
Transfer function	Anand et al. (2003)	Müller et al. (1998)	None used	None used	None used
Seasonal bias	Yes (for supported species)	Not considered			
Bioturbation	10 cm				
Sediment layer width	distance between each measurement (if <2 cm)				
Sampling size	30	Infinite sample size			
Inter-individual variations	2	Not defined			
Measurement error (K)	0.26	0.23	1	1	1

The *Full PP* experiment is used to evaluate the performance of the algorithm under normal conditions. The pseudo-proxies are computed as described in sect. 3.2.1. The computation of the pseudo-proxies is performed ten times to test the robustness of our results to the random elements in the methodology (measurement, calibration, and age uncertainty when applying *sedproxy*). The individual contributions of the five sources of uncertainty quantified by the algorithm are also determined. We indicate in the following in parentheses how each uncertainty quantification is switched off: the measurement and calibration noise (no noise is added), the age model ensemble (only the mean age is considered), the selection of records (all time series are used), the latitudinal band configurations (only one configuration is considered: 60°N - 40° - 20°N - 0° - 20°S - 40°S - 60°S), and the scaling factor (no uncertainty is applied).

The *measurement noise* experiment only includes the measurement uncertainty from *sedproxy* (along with the calibration uncertainty), and its estimate from the algorithm. The experiment is used to compare the representation of measurement noise in the algorithm with the one introduced in the pseudo-proxy, as well as to quantify the effect of the latter on the signal quality. ~~In order to~~ To investigate this effect, we perform 30 iterations of the PPE.

The experiment *SST at proxy locations* is designed to evaluate the ability of the algorithm to recover the 60°S-60°N mean SST signal, given the number and locations of the proxy data. None of the processes modelled by *sedproxy* is included, that is, the GMST algorithm is directly provided with the modelled SST timeseries at the grid point nearest to the proxy location.

Full PP at random locations corresponds to a series of PPEs investigating the influence of the number of proxy records on the uncertainty. Four numbers are tested: 61 is the number of records selected in S16, 82 is the average number of records during MIS 3 to 5 in our dataset (Fig. 2-A), 200 is the minimal number of records during the deglaciation, and 400 is an example in case a much larger dataset could be obtained. Oceanic locations are randomly generated for each pseudo-proxy. To each of these locations, we assign metadata (proxy type and species), that corresponds to a random record in J20 located within 5° of latitude. To construct age models, we start by selecting one hundred age ensemble members from each of the 44 (for the MPI

Pseudo-Proxy Experiments	Simulations	Proxy characteristics				Seasonality	*Sedproxy*		Uncertainties quantified in the GMST algorithm	Members	Figures
		Number	Locations	Age resolution	Age uncertainty		Biotur-bation	Measure-ments			
Full PP	MPIs, FAMOUS, LOVECLIM	225 - 265	J20	J20	J20	only *Mg/Ca* and not LOVECLIM	J20	Yes	First all, then each separately	10	4, 5,8,9
Measurement noise	MPIs, FAMOUS, LOVECLIM	225 - 265	J20	100 yr	-	-	-	Yes	Measurement	30	9
SST at proxy locations	MPIs, FAMOUS, LOVECLIM, CESM	225 - 265	J20	100 yr	-	-	-	-	resampling and latitudinal bands	1	6-A
Full PP at random locations	MPIs, FAMOUS, LOVECLIM	61, 82, 200, 400	Random	100 yr	J20*	only *Mg/Ca* and not LOVECLIM	-	Yes	First all, then each separately	100	~S1
No offsetting	MPIs, FAMOUS, LOVECLIM	225 - 265	J20	J20	J20	only *Mg/Ca* and not LOVECLIM	J20	Yes	-	10	7-A
Warmest month	MPIs, FAMOUS, CESM	225 - 265	J20	100 yr	-	Warmest month	-	-	-	1	7-B
Bioturbation	MPIs, FAMOUS, LOVECLIM	225 - 265	J20	100 yr	-	-	J20*	-	-	1	9
Age uncertainty	MPIs, FAMOUS, LOVECLIM	225 - 265	J20	100 yr	J20*	-	-	-	-	1	9
Age uncertainty and bioturbation	MPIs, FAMOUS, LOVECLIM	225 - 265	J20	100 yr	J20*	-	J20*	-	-	1	9
Age resolution	MPIs, FAMOUS, LOVECLIM	225 - 265	J20	J20*	-	-	-	-	-	100	9
Measurement noise at age resolution	MPIs, FAMOUS, LOVECLIM	225 - 265	J20	J20*	-	-	-	Yes	-	100	9

Table 2. Characteristics of the PPEs. The first set is used to analyse the uncertainty sources. The second set, with the addition of *Full PP* and *Measurement noise*, is used to investigate the representation of the time continuum. Apart from *Full PP at random locations*, the number of timeseries considered depends on the simulation time span: 225 (MPI), 263 (FAMOUS), 265 (LOVECLIM, CESM). J20 corresponds to the use of metadata from J20. J20* refers to special cases described in the text. Some PPEs are computed several times to form an ensemble. It allows the sampling of measurement noise, random location, or reconstructed age ensemble when applied. The figures where the corresponding PPE is used are indicated in the last column. The results of *Full PP at random locations* are only discussed in the text.

simulations) or 24 (for FAMOUS and LOVECLIM) sediment cores that fully span respectively the deglaciation $(2-25\,\text{kyr})$ and the LGC $(5-130\,\text{kyr})$. The age models are then interpolated so that their mean resolution is a hundred years. This creates 44 or 24 age ensembles that are randomly assigned to a location. Finally, these metadata are used to compute pseudo-proxy time series with *sedproxy*. Effects from bioturbation and sediment thickness are removed, as they do not impact the uncertainty estimates of the algorithm. This procedure is iterated 100 times for each number of pseudo-proxy records tested, so that an uncertainty range can be drawn from the ensemble.

The *no offsetting* experiment is used to evaluate the impact of the offsetting procedure: this one is removed from the algorithm and the correct temperature anomaly of the simulation at each location is used instead. The experiment is otherwise identical to the *Full PP* experiment. The *warmest month* experiment is used to estimate an upper bound of seasonality bias on the proxy records. The pseudo-proxies are computed by only considering the warmest month within each year.

The next three experiments evaluate the individual impact on the signal of bioturbation, age uncertainty and age resolution respectively, using a simple and idealised setup. In the *bioturbation* experiment, the accumulation rate is identical at each location and is equal to the average across the entire J20 dataset (but still time-dependent, cf. Fig. 2-D). Note that we consider bioturbation together with the sediment thickness, as the two are intertwined in *sedproxy*. For the *age uncertainty* experiment, we use the same 44 or 24 age ensembles computed for the experiment *Full PP at random locations* to consider age uncertainty independently from other parameters. We consider this selection of age ensembles to closely represent the age uncertainty as a function of time for the entire proxy dataset (as given in Fig. 2-B). Finally, for the *age resolution* experiment, we construct age models for the entire time period which resemble the original distribution of temporal resolutions across time (Fig 2-C). The experiments *~~Age~~ age uncertainty and bioturbation* and *~~Measurement~~ measurement noise at age resolution* investigate combined effects.

Finally, note that CESM data has a too low temporal resolution to be able to consider the smoothing factors applied by *sedproxy*, and that LOVECLIM data doesn't enable the investigation of seasonality.

4   Results

The various pseudo-proxy experiments are analysed in this section. We start with the *Full PP* experiment, whose results are presented in Fig. 4. This experiment gives a general overview of the ability of the GMST reconstruction algorithm. In particular, the algorithm successfully reconstructs the orbital timescale variations $(> 10\,\text{kyr})$ of the four simulations. The mean pseudo-proxy reconstruction presents however a strong smoothing of ~~sub-orbital~~ shorter timescales. Finally, the temperature anomaly of the glacial period in the pseudo-proxy reconstructions is, for most simulations, slightly too cold, although the targeted simulated GMST is within the reconstruction uncertainty range. In light of these preliminary results, we arrange the analysis of the results around three main topics, addressed in the following subsections: quantification and origin of the uncertainty, amplitude and timing of orbital timescale variations, and representation of the timescale continuum.

4.1 Quantification and origin of the uncertainty on the GMST

4.1.1 Uncertainty range as estimated by the algorithm

The GMST reconstruction algorithm estimates uncertainties using a Monte Carlo approach. Five different sources of uncertainty are estimated at different stages of the algorithm: measurement and calibration noise, age model uncertainty, selection of records (which we interpret as a location resampling), latitudinal band configurations, and scaling factor. The uncertainty is propagated to the GMST reconstruction by considering an ensemble of realisations.

The width of the total 90% confidence interval (CI) is presented in Fig. 5, for the various simulations, alongside an estimation of the individual contributions of the considered uncertainty sources. Note that the uncertainty estimates do not depend on the uncertainty sources accounted for when creating the pseudo-proxies. We first notice a large uncertainty increase at both ends of the time series. The Holocene has otherwise the lowest uncertainty ($\sim 1\,\mathrm{K}$ for all simulations). The uncertainty is maximum during the LGM ($2-4\,\mathrm{K}$). We also find that the root sum square of the individual contributions is very close to the total CIconfidence interval. This implies that reducing the uncertainty from the highest sources of uncertainty is the most impactful.

One of the most important contributors to the uncertainty estimate is the scaling factor. By definition, it is proportional to the temperature anomaly, and exhibits, therefore, large fluctuations. It is negligible during the Holocene and the LIG, but accounts from for between 62% (FAMOUS) to and 80% (LOVECLIM) of the uncertainty estimate during the LGM, depending on how cold the simulated LGM is. On average, the most important contributor to the uncertainty estimate is however the measurement and calibration noise, with a CI width mostly between 1 and $1.5\,\mathrm{K}$, and an average contribution from 38% (LOVECLIM) to 50% (FAMOUS, MPI-I6G). It also increases sharply at both ends of the time series and is the main contributor to the same behaviour in the total CI. The CI width confidence interval of the location resampling and the age model are below 1 and around $0.5\,\mathrm{K}$ respectively (or, in terms of average contribution, around 15% and 6% for all simulations). Finally, the configuration of the latitudinal bands has the smallest impact on the uncertainty estimate (below $0.5\,\mathrm{K}$, or 1%).

4.1.2 Ability of the algorithm to estimate the uncertainty

We further the analysis by investigating whether the GMST reconstruction algorithm correctly estimates the uncertainty sources, starting . We can use the PPEs that only include one uncertainty source, and compare it to the estimates from the algorithm. We start with the measurement and calibration noise. By definition, the 90% CI should have a coverage frequency of the targeted signal of 90%. However, using the *measurement noise* experiment, we find that the 90% coverage frequency is already reached by a CI, smaller than the 90% CI by a factor of 2 (FAMOUS) to 2.3 (All the others). The overestimation of the CI is not unexpected: the measurement noise introduced in the pseudo-proxies by *sedproxy* is between 0.23 and $1\,\mathrm{K}$ (cf. tab. 1), while the noise introduced in the reconstruction algorithm is $1.5\,\mathrm{K}$. This overestimation is a conservative approach in case the measurement and calibration uncertainty is higher in reality than what is used to compute the pseudo-proxies.

Next, we evaluate the ability of the reconstruction algorithm to estimate the 60°S-60°N mean SST (MSST, before applying the scaling factor) based on the specific set of proxy locations made available in the J20 dataset. To that end, we compare the MSST reconstructed from the *SST at proxy locations* experiment to the simulated MSST (Fig. 6-A). We find that the

reconstruction for all simulations ~~but CESM~~ exhibits a cold bias during the LGM (up to $0.5\,\mathrm{K}$) and most of the glacial period. In addition, location resampling and latitude band configurations, which aim to account for it, are not large enough to cover the bias. Yet, the pseudo-proxy experiments using random proxy locations can reproduce the simulated MSST (Fig. S1). Therefore, the bias is caused by the specific set of locations in the J20 dataset: there is an over-representation of regions with strong LGM cooling (e.g. ~~North-West~~ North Atlantic and origin of the ~~Kurushio~~ Kuroshio extension current), compared to regions with weaker to no cooling (e.g. ~~North polar gyre and eastern part of the subtropical gyres in the Pacific~~ Pacific gyres) in the same latitudinal band (cf. Fig. 1-A, Fig. S3). In addition, how the uncertainty from the location sampling of records is estimated relies on the hypothesis that the temperature anomalies are well-distributed, which is not the case, explaining the underestimation of the uncertainty range.

The last uncertainty estimate we evaluate is the scaling factor (Fig. 6-B). We find the errors constrained within $1\,\mathrm{K}$. The coverage frequency of the 90% CI ranges between 46% (MPI-I6G) and 79% (LOVECLIM), which denotes a small underestimation. By definition, the uncertainty of the scaling factor is proportional to the temperature anomaly. We suggest that this overly strong constraint leads to an overestimation (underestimation) of the uncertainty for large (small) temperature anomalies. This can be tackled by considering in addition an additive error independent of the SST anomaly.

4.1.3 Factors influencing the uncertainty estimates

We continue the analysis by investigating the dependency of the uncertainty estimates. Using the experiments *Full PP at random locations*, we find the number of records considered to strongly impact the uncertainty. Using 61 proxies as a baseline (number of records in S16), we find that 83 records (average number of records between MIS3 to 5) reduce the uncertainty range of the reconstruction by ∼15% (See also Fig. S1). With 200 (as during the deglaciation) the reduction reaches 46%. In case 400 records were available, the reduction would be 62%. This reduction mostly ~~affect~~ affects the band configurations and the measurement noise, followed by the location resampling and age uncertainty (in relative values). The proxy number has ~~of course~~ by definition no effect on the scaling factor.

The variation of the number of records through time in the *full PP* experiment (Fig. 2-A) explains most of the variations in the uncertainties shown in figure 5, and in particular the increase at both ends of the time series[2]. Note that the number of records considered in the PPE at both ends of the timeseries is not only related to the number of records available in J20 (cf. Fig. 2-A), but also to the fact that some pseudo-proxy values cannot be computed as age uncertainty spans beyond the timespan of the simulations. This is most evident in the MPI simulations (Fig. 5-B and D). In addition, measurement noise also depends on the age resolution, as a higher number of measurements reduces the noise introduced in the reconstruction algorithm. The location sampling uncertainty depends on the homogeneity of the temperature anomaly field, that is, to the first order, on the GMST anomaly. Finally, uncertainty due to the age model is dependent on the rate of change of the GMST: abrupt

---

[2]~~The number of records considered in the PPE at both ends of the timeseries is not only related to the number of records available in J20 (cf. Fig. 2-A), but also to the fact that some pseudo-proxy values cannot be computed as age uncertainty spans beyond the timespan of the simulations. This is most evident in the MPI simulations (Fig. 5-B and D)~~

changes such as the deglaciation, or the millennial-scale variability exhibited in the MPI simulation or LOVECLIM increases the uncertainty. However, these effects only become noticeable once the number of records is fixed.

4.1.4 Potential sources of uncertainty non-quantified by the algorithm

395 Finally, we investigate other sources of uncertainty, which are not or only partly estimated by the reconstruction algorithm. We quantify, in particular, the effect of the offsetting procedure and potential seasonal ~~bias~~ biases (Fig. 6). The offsetting procedure generates errors generally below $0.2\,\mathrm{K}$ on the reconstructed GMST, except for FAMOUS, where it reaches $0.5\,\mathrm{K}$. We suggest this bias is reasonable given the number of records it enables the algorithm to additionally consider (96).

We also estimate an upper bound to the effect of seasonality bias on the pseudo-proxy reconstruction. If all records were not 400 recording the mean annual temperature, but that of the warmest month, it would generate biases up to $0.75\,\mathrm{K}$. This bias mostly follows the ~~northern hemisphere insolation curve~~summer solstice insolation at $65°\mathrm{N}$: warm during the early Holocene, cold during the LGM and MIS4, and warm again during MIS5. These ~~orbital-scale~~ precession-scale biases can significantly impact the evaluation of ~~orbital timescalevariations~~this timescale, in particular regarding the LGM cooling which it could overestimate, or an early Holocene warm period which would only appear for the summer months.

405 In conclusion, we reckon that the estimation of the uncertainty range by the algorithm is realistic, but that the estimations of the individual contributions are biased and should be used with caution. We suggest that the measurement and calibration noise, the set of record locations, and the scaling factor contribute overall similarly to the uncertainty.

4.2 Amplitude and timing of orbital timescale variations

Given the J20 dataset, one of the key characteristics we expect the algorithm to reconstruct accurately is the amplitude and 410 timing of orbital timescale variations of the simulated GMST. While these characteristics can be qualitatively assessed in Fig. 4, we also provide quantitative metrics with uncertainty and robustness (Tab. 3). Note that we use slightly different definitions for the LGM and the end of the LIG to accommodate the period and behaviour of the simulations (cf. caption).

As already discussed, a cold bias is evident for LGM and end of LIG temperature in all simulations but FAMOUS, but the true anomalies remain included in the 90% CIs. This bias is mostly caused by the location bias (cf. Fig. 6-A), while the other 415 factors (offsetting procedure, scaling to GMST and seasonality) either reduce or amplify it.

We also investigate the timing of various events: Glacial onset, LGM, and deglaciation end. These metrics are only designed to compare the result of the pseudo-proxy reconstruction to the simulations. We find the reconstructions to match very closely the timing of the simulations. We further suggest that the 90% CI as calculated by the GMST algorithm largely overestimates the uncertainty. While the overestimated measurement noise helps to account for non-quantified sources of temperature bias in 420 the GMST algorithm, it may also lead to an overestimation of the timing uncertainty, where no biases are evident. Interestingly, the PPE including only bioturbation shows a lag of about half the bioturbated layer, which is not evident in the table. This bias is caused by the non-symmetric shape of the function characterising the sediment surface mixing.

Finally, we find our result robust to different realisations of pseudo-proxies. The measure of robustness also describes a limit under which temperature and timing differences depend on the pseudo-proxy realisation: about $0.5\,\mathrm{K}$ and from $0.7\,\mathrm{kyr}$

Table 3. Quantitative assessment of the reconstruction of the amplitude and timing of orbital timescale variations. We use the GMST reconstructed for the *Full PP* experiment. The LGM temperature anomaly is computed for LOVECLIM and FAMOUS as the coldest $4\,\mathrm{kyr}$ later than $40\,\mathrm{kyr}$ BP with respect to the last $5\,\mathrm{kyr}$. The same definition is used for LGM occurrence. Due to the difficulty of defining the coldest period in the MPI simulations, we use the average between 19 and $23\,\mathrm{kyr}$ BP instead. The ~~Deglaciation~~ deglaciation end is defined as the last time the temperature anomaly reaches one-tenth of the LGM temperature anomaly. Similarly, glacial onset is defined as the first time the temperature anomaly falls below one-tenth of the LGM temperature anomaly. Finally, we also compute a temperature anomaly for the end of the LIG, as the average between 119 and $123\,\mathrm{kyr}$ for LOVECLIM and between 110 and $114\,\mathrm{kyr}$ BP for FAMOUS. "Mean" is the mean value across the reconstruction ensembles. "Robustness" is the range of the mean value for each of the ten ensembles computed for the experiment. It characterises the robustness of the mean reconstruction to random factors in the proxy data. "90% CI" is the mean 90% confidence interval, as calculated by the GMST algorithm, within each reconstruction ensemble. The target is the actual simulated value.

	LOVECLIM				FAMOUS			
	mean	robustness	90% CI	target	mean	robustness	90% CI	target
End of LIG anomaly (K)	-0.4	-0.68/-0.05	-1.32/0.56	-0.09	-0.04	-0.18/0.17	-1.03/0.93	-0.03
Glacial onset (kyr BP)	118.2	116.8/119.7	113.5/122.9	118.3	106	102.5/108.4	93.1/113.2	109.9
LGM anomaly (K)	-6.65	-6.89/-6.39	-7.95/-5.40	-5.97	-3.6	-3.75/-3.43	-4.39/-2.85	-3.57
LGM occurrence (kyr BP)	23.3	22.6/24	20.9/27	22.4	25.7	24.4/26.8	19.8/32.2	25.3
Deglaciation end (kyr BP)	8.7	8.1/9.2	6.5/10	9.9	5	4.7/5.2	4/6.2	5.4
	MPI-I6G				MPI-G1D			
LGM anomaly (K)	-6.02	-6.13/-5.83	-7.24/-4.85	-5.69	-7.67	-7.82/-7.46	-9.17/-6.25	-6.72
Deglaciation end (kyr BP)	8.9	8.6/9.3	7.1/10.2	9.2	8.8	8.5/9.2	7.4/9.9	9.2

in the Holocene to $5\,\mathrm{kyr}$ during the LIG. Interestingly, these values are below the uncertainty associated with individual proxy measurements. In addition, we suggest that this measure characterises the uncertainty better than the ~~CI~~ confidence interval provided by the GMST algorithm, when comparing values to one another.

4.3 Representation of the timescale continuum

As noted in Fig. 4, the algorithm does not reconstruct the high-frequency variability of the simulated GMST. We further investigate this loss of signal using spectral densities and coherence. We focus on the MSST as the scaling factor does not influence the high-frequency variability. Figure 8 compares the spectrum of variability of the simulated MSST to the one reconstructed in several PPEs. We first focus on the *Full PP* experiment. The individual Monte Carlo realisations of the reconstruction (light grey lines) appear to overestimate the power spectral density (PSD) for most timescales but the longest ones. In particular, their PSD exhibits a straight slope for the shorter timescales ($\sim -2\,\mathrm{K}^2$). The overestimated spectrum is mostly related to the large measurement noise introduced in the GMST algorithm. By contrast, the spectrum of the ensemble mean (darker grey lines) underestimates variability at most timescale, with a steeper slope (between $-2.7\,\mathrm{K}^2$ for FAMOUS and $-3.3\,\mathrm{K}^2$ for the MPI simulations). The PSD is properly represented only for ~~timescale~~ timescales longer than $10\,\mathrm{kyr}$ for the

entire LGC, and longer than $4\,\mathrm{kyr}$ for the last $25\,\mathrm{kyr}$ (the latter holds regardless of the simulation). None of the pseudo-proxy reconstructions are able to represent all the characteristics of the simulated PSD, such as the high variability at short timescale in FAMOUS, or the high variability at millennial timescale in the MPI simulations. The loss of PSD between the ensemble mean and each member's spectrum is due to a lack of coherence between reconstruction members (i.e. loss of signal).

We further investigate the cause of the underestimated variability of the ensemble mean with sensitivity PPEs. We identify four potential causes: the bioturbation, which smooths the local SST time series, the age uncertainty, which reduces the coherence between the local SST time series, the age resolution, which permits resolving the shortest timescale, and the measurement and calibration noise in *sedproxy*, which adds non-climatic variability (cf. coloured lines in Fig. 8). The PSD from the *measurement noise* experiment (yellow) is the closest to the simulated one with just slight overestimation ~~, but lose~~ but loses the simulated PSD characteristics below $1\,\mathrm{kyr}$. In this PPE, the noise is injected at the centennial scale, resulting in increased variability at this timescale and upward propagation. Over the LGC, none of the other PPEs is able to represent the simulated spectrum below $10\,\mathrm{kyr}$, with the strongest loss of PSD due to age uncertainty (blue). For the last $25\,\mathrm{kyr}$, the PPEs are still able to reproduce the large multi-millennial variability, which is only lost by the joint consideration of bioturbation and age uncertainty (purple).

The loss of signal at specific timescales can be further characterised using coherence. Coherence, or the squared coherency spectrum, is a statistical method to analyse the dependency between two time series through the timescales (Von Storch and Zwiers, 1999). It can be seen as a timescale-dependent squared correlation. A coherence of 0 implies that the two time series vary independently at this timescale. A coherence of 1 means that the time series vary synchronously, although a lag can exist. The coherence also does not characterise the amplitude of the variation, which we have already investigated with the spectral densities. Fig. 9 presents the coherence between the simulation and reconstructed MSST from the same PPEs as in Fig. 8. For the *Full PP* experiment, coherence is very high (> 0.9) for the longest timescales in all simulations. The coherence decreases however sharply for timescales shorter than $15\,\mathrm{kyr}$ for the LGC simulations (FAMOUS and LOVECLIM), which corresponds to the start of the PSD loss in Fig. 8-A and C. The drop happens at shorter timescales (below $4\,\mathrm{kyr}$) for the last $25\,\mathrm{kyr}$, again corresponding to the timescale ~~on~~ from which the PSD loss starts. A similar timescale has been found for the Holocene (Dolman et al., 2021).

We further the analysis with the sensitivity PPEs. We find the age uncertainty (blue) to be the main cause of the loss of coherence on the LGC. For the last $25\,\mathrm{kyr}$, bioturbation also plays an important role, but neither can individually explain the coherence loss exhibited by the *Full PP* experiment. In addition, using a smaller bioturbation depth further limits the influence of bioturbation (Fig. S2). Measurement noise (yellow) and age resolution (brown) have the smallest impact on the coherence for either the LGC or the last $25\,\mathrm{kyr}$. However, the measurement noise is in reality linked to the age resolution. The PPE considering both (orange line) exhibits a coherence loss at a timescale close to the age uncertainty and bioturbation, suggesting it also has a key role for the last $25\,\mathrm{kyr}$. Note that, for the number of records considered, the set of record locations does not impact the coherence significantly. Simulated GMST and MSST are also significantly coherent at all timescales (~~not shown~~Fig. S5).

Our results show that the difference in results between the LGC and the last 25 kyr is directly related to the dataset characteristics (Fig. 2-B to D). In the last 25 kyr, the average 90% CI of the age uncertainty and the average bioturbated layer have similar values, corresponding to the timescale when coherence is lost. Age uncertainty, resolution, and bioturbated layer increase beyond this time period, but the increase in age uncertainty is the most important. This one reaches ~ 10 kyr beyond 40 kyr BP, again similar to the timescale when coherence drops.

5   Discussion

The ~~thorough investigations of~~ results from the previous section allow us to discuss the performance of the reconstruction algorithm applied to the J20 dataset (Sect. 5.1). Nevertheless, future avenues for these pseudo-proxy analyses are possible (Sect. 5.2). We also discuss which limiting factors from either the algorithm (Sect. 5.3) or the dataset (Sect. 5.4) should be addressed in priority to improve the GMST reconstruction quality.

5.1   Quality of the reconstruction

According to our analysis of the PPEs, the timing of GMST variations can be reconstructed for timescales longer than 4 kyr and a timing uncertainty of ± 0.5 to 1 kyr for the last 25 kyr. Beyond 40 kyr BP, only timescale longer than 15 kyr can be reconstructed, with an uncertainty of up to ± 3 kyr. This precision of the reconstruction, and its variation through time, is directly related to the underlying dataset characteristics. The reconstruction of shorter timescales beyond 40 kyr BP is limited by the age uncertainty of the records. For the last 25 kyr, the limitation originates from a combination of age uncertainty, bioturbation, and measurement and calibration noise.

Most reconstructed pseudo-proxy GMSTs also exhibit a cold bias of up to 1 K during the glacial period, which we attribute to the specific set of record locations in J20. Uncertainty on the temperature anomaly is otherwise strongly dependent on the number of proxy records and the scaling factor. The GMST algorithm estimates the uncertainty range (90% CI) from less than 1 K during the Holocene, to up to 3 K for a cooling of 6 K during the LGM. Our results suggest, however, that the Holocene uncertainty is overestimated because of an overestimated measurement noise uncertainty. Similarly, the uncertainty for large temperature anomalies is also likely overestimated due to the definition of the scaling factor uncertainty. For the LIG, overestimated measurement noise and underestimated scaling uncertainty seem to cancel each other out and lead to reasonable ~~results~~ values. These overestimations help nevertheless to account for uncertainty sources not properly accounted for, such as locations or potential seasonality bias. In particular, an underestimation of the seasonality bias could lead to a deterioration of the performance, both in terms of amplitude and timing of variations, especially for the period between the LGM and the early Holocene, where uncertainty is otherwise the lowest.

The analysis of the uncertainty of the timing and amplitude of variations shows that there exists a trade-off between including ~~a large number of~~ many records, which helps to reduce the biases and uncertainty on the amplitude, and including only the records of the highest quality (low age uncertainty, high accumulation rate, high resolution), which help better resolve the

shorter timescale variability. While a finer selection of records could balance the two, we also suggest that the algorithm could handle records of different characteristics better (cf. Section 5.3).

505   These results are robust across the set of simulations, and we assume them to hold for the real-world reconstruction. In addition, while these results are specific to the adapted S16 algorithm applied to ~~the~~ J20, we expect that the general concepts discussed here hold for any aggregation algorithm and dataset. In particular, the algorithm could be used for regional reconstructions (as done in Clark et al., 2024), as long as a sufficiently high number of records are considered. We suggest, for example, that the accurate reconstruction of the timing of orbital timescale variations ($> 10\,\mathrm{kyr}$) offers a good opportunity to
510 investigate the synchronicity of regional temperature anomalies and forcings.

5.2 Advances in pseudo-proxy experiments

Our results rely on advances in pseudo-proxy experiments, which were enabled by the availability of proxy system models and climate simulation output.

5.2.1 Model data

515 In addition to the proper availability of simulated climate fields for the production of ~~PEE~~PPE, the PPE results can only be transferred to real proxy data if these fields are realistic enough. The recent advances in computing capacities, modelling, and our understanding of climate drivers have enabled the production of more realistic climate simulations over longer periods of time (Ivanovic et al., 2016). Yet, the PPEs we conduct require data only available for a handful of simulations. However, a multi-model framework for PPEs is key to assessing our results' robustness.

520   One of the requirements is of course the simulated time period, which needs to cover more than the period of interest, ~~in order~~ to take into account the age uncertainty of the records. For example, properly investigating the LIG requires data from the penultimate deglaciation. Another requirement is that these simulations must include transient forcings, so that realistic variations of the GMST can be produced at sufficiently high resolution. In particular, a centennial or lower resolution is needed here to resolve the effect of bioturbation. The addition of other forcings such as meltwater in the MPI simulations also increases
525 the degree of realism of the simulated GMST (Weitzel et al., 2024). These additions make more refined ~~analysis~~ analyses of the results possible, which can more easily be transferred to real proxy data. For instance, we can assess that the GMST algorithm is not able to reconstruct the effect of melt-water forcing, such as the abrupt events during the deglaciation.

    Finally, the spatio-temporal variability of the simulated climate ~~field~~ fields affects our results. For example, the amplitude of the temperature bias due to the proxy locations directly depends on the spatial pattern of the simulated temperature field.
530 Differences between the simulations lead to differences in the estimation of the bias in the PPE reconstruction (cf. Fig. 6-A). Therefore, despite some agreements between simulations on the sign and amplitude of the bias, proper quantification of this bias for ~~the~~ real-world ~~reconstruction~~ reconstructions would first require an evaluation of the degree of realism of the spatio-temporal patterns. In support of this, the temperature field of simulations is known to be more homogeneous at the regional scale compared to proxy records (Jonkers et al., 2023). This has been demonstrated for the multidecadal to multimillennial-scale
535 (Laepple et al., 2023; Weitzel et al., 2024). For the orbital timescale ($> 10\,\mathrm{kyr}$), Paul et al. (2021) suggested that upwelling

regions, where many records come from, can have a different variability while being too small to be properly resolved by the models considered here. The use of too homogeneous fields can lead to an underestimation of location biases on the reconstructed temperature and overconfidence in the algorithm.

5.2.2 Proxy system models

540 In PPEs, proxy system models are used to produce as realistic as possible pseudo-proxy values. However, proxy system models only represent our current understanding of the proxy system, which limits the extrapolation of our results to the real-world reconstruction. Yet, the use of *sedproxy* presents the advantage, compared to more conceptual approaches (e.g. Wang et al., 2014; Jaume-Santero et al., 2020; Nilsen et al., 2021; Weitzel et al., 2024), to model specific processes, such as the measurement and calibration noise and bioturbation, as a function of ~~the~~ a record's metadata. Sensitivity PPEs can therefore be computed to

545 evaluate the impact of these processes.

Yet, our results are also limited by the proxy system model used. For example, the impact of the processes occurring during the *sensor stage* for TEX_{86}, LDI or assemblages has only been crudely accounted for within a generic measurement error. We therefore did not investigate in depth the influence of the proxy types on the reconstruction. While increased uncertainty from the measurement or the calibration will reduce the ability of the reconstruction to resolve the shorter timescale, other

550 biases affecting the longer timescale would require a more thorough analysis. These potential biases include for example the effect of any variables other than the SST, such as depth, salinity or seasonality, on the temperature signal recorded (Telford et al., 2013; Timmermann et al., 2014; Ho and Laepple, 2015; Jonkers and Kučera, 2017). The seasonality bias is considered for most species recording Mg/Ca ratios in *sedproxy* ~~,~~ but assumed negligible for the other proxy types. For this reason, we design a PPE to estimate an upper bound of the effect of seasonality, where all proxies record the warmest month of the year.

555 This warm season bias has, for example, been one of the main hypothesised reasons for the model-data discrepancies during the early Holocene (Liu et al., 2014; Marsicek et al., 2018; Bova et al., 2021), although there is also evidence for cold season bias for other periods and species (Steinke et al., 2008; Timmermann et al., 2014).

5.3 Improvement of the algorithm

One of the primary objectives of this paper is to provide a framework to evaluate the performance of GMST reconstruction

560 algorithms. We focus on an adaptation of S16 as an example. The adaptations are minor ~~,~~ and only required to improve the use of the J20 dataset and the characterisation of the algorithm's performance (cf. sect. 3.1.2). Yet, our evaluation makes evident avenues of improvement for the algorithm which we discuss here.

We have already discussed how the algorithm is sensitive to the spatial distribution of proxy records. This sensitivity relates to the assumption that temperature anomalies are similar within a latitudinal band. Both the simulations used here and other

565 proxy analyses show that this is not the case (e.g. Judd et al., 2020; Tierney et al., 2020; Paul et al., 2021). Some studies have used more complex methods and relied on either present SST observations (e.g. Paul et al., 2021), or climate simulations (e.g. Osman et al., 2021; Annan et al., 2022). These methods rely however on the assumption that the spatial covariance of present-

day does not change through time, or that of the simulated SST is similar to the reality. The design of another aggregating method could help to reduce the influence of areas with a large density of records, without relying on external datasets.

570 Our refined algorithm relies on the stacking of proxy records to increase the signal-to-noise ratio. However, records are stacked together, regardless of their quality (age uncertainty, accumulation rate, resolution). This stacking method leads to a trade-off between data quality which improves the reconstruction of the shorter timescales, and a high number of records, which reduces the overall uncertainty (cf. 5.1). A new stacking method could be researched to limit the impact of this trade-off by, for example, taking into account the timescale resolved by each record.

575 Finally, the scaling of the MSST to the GMST was introduced by S16 due to the limited availability of reliable local temperature reconstruction over land and ice-covered areas for the investigated time period. Similarly, Clark et al. (2024) uses a scaling factor, although with a different definition. This scaling is ~~however~~ in both cases a critical source of uncertainty, especially concerning the amount of glacial cooling or LIG temperature anomaly. This ~~has direct consequences on~~ directly affects our ability to constrain Earth system sensitivity or the global response to climate forcings. The transient simulations

580 used here show that the variation of GMST and MSST is not completely ~~linear~~collinear (Fig. S4). This behaviour suggests going beyond the simple LGM-to-PI ratio used to compute the scaling factor in S16, or even the one-to-one match between MSST and GMST supposed in Clark et al. (2024). Despite their scarcity, the use of terrestrial proxies for temperature could still help to better characterise the scaling. Yet, large areas ~~are devoid of~~ lack proxy information, ~~in particular~~ particularly over sea ice and ice sheets that have melted. Physics-based constraints remain needed for these areas.

585 ### 5.4 Limiting factors from the dataset and future developments

Expanding the dataset with new records, in particular for the periods and areas where fewer records are available, will evidently improve the reconstruction. Nevertheless, ~~spatial-temporal~~ spatio-temporal inhomogeneity will always remain due to geological constraints. Other avenues can improve the quality of and the confidence in the reconstruction, from the perspective of the proxy dataset. First, the age uncertainty is the limiting factor preventing the reconstruction of multi-millennial timescale beyond

590 30 kyr BP, and ~~the main focus should be put on reducing this uncertainty (e.g. Waelbroeck et al., 2019; Peeters et al., 2023)~~ better characterisation of and constraints on age uncertainty are required. Progress has been made for example using visual matching and quantifying the associated uncertainty or considering the impact of bioturbation on age model (e.g. Waelbroeck et al., 2019; L . In addition to improving the confidence in the absolute age, information on the relative age between records can also be provided, if properly accounted for by the reconstruction algorithm.

595 Second, measurement and calibration noise and bioturbation are ~~another source~~ other sources of loss of signal on short timescales, particularly in the last 30 kyr. Here, we suggest better quantifying these uncertainties for each record, so that the algorithm can sort the records depending on the timescale they can resolve. In particular, bioturbation depth could be quantified for each record, and measurement uncertainty could depend on the number of replications or sample size for each data point.

Finally, we find that potential seasonality bias can significantly decrease the accuracy of the reconstruction. For many of 600 the records available, it is unclear whether the reconstructed temperature suffers ~~for~~ from a seasonal bias. Corrections from

seasonally biased records have been suggested (e.g. Bova et al., 2021), but rely on assumptions that must be verified beforehand (Laepple et al., 2022).

6 Conclusions

In this study, we design a framework for the evaluation of algorithms reconstructing spatial mean temperatures. Our framework relies on recent advances in pseudo-proxy experiments, which include the availability of realistic proxy system models and long transient climate simulations. We apply the framework to an adapted version of the GMST reconstruction algorithm used in S16, and the synthesis of marine proxy records for ~~temperature~~ temperatures of the LGC from J20. The quantitative results presented here can serve as a basis for future evaluations of LGC reconstructions, while the framework can be applied to other aggregation-based reconstruction algorithms and other datasets.

Our results are based on PPEs computed from an ensemble of 4 transient simulations of the LGC or ~~of~~ the last 25 kyr. We find that the pseudo-proxy reconstructions based on the S16 algorithm and the J20 dataset perform differently over time. For the last 25 kyr, the temperature variations and their timing are accurately reconstructed for timescales of at least 4 kyr. Sensitivity PPEs show that age uncertainty, bioturbation, and measurement noise smooth the reconstruction, leading to a loss of signal below this timescale. Uncertainties remain large on the amplitude of temperature variations, even for longer timescales, although we find the algorithm to overestimate it. The reconstructions exhibit in particular a cold bias for most model simulations, which is related to the non-uniform distribution of record locations. Other sources of uncertainty include the scaling of the mean SST to the GMST, the measurement and calibration noise, and ~~potential seasonality~~ a potential seasonal bias. The number of records plays a critical role in reducing all uncertainty sources but that of the scaling. The decreasing proxy number and increasing uncertainty on the scaling are the two main factors explaining the increased uncertainty from the Holocene to the LGM. Beyond 40 kyr BP, only timescales longer than 15 kyr can be reconstructed, due to a sharp rise in age uncertainty. We assume these results to hold for real-world reconstructions of the LGC.

Our results show also the existence of a trade-off between the inclusion of ~~a large number of~~ many records, which overall reduces the uncertainty, and of only the highest quality records (low age uncertainty, high accumulation rate, high resolution), which improves the reconstruction of the short timescale. The reconstruction could be improved by a better filtering of the input record data, or by a better handling of the varying record quality by the algorithm. We also suggest other avenues of improvement for the algorithm, to handle better the spatial aggregation and the scaling to GMST. From the proxy record perspective, reducing the age uncertainty is the most critical challenge to tackle.

Code and data availability. The R code and the proxy metadata (subset of the PalMod database) to reproduce the results and plots of this study are available at https://doi.org/10.5281/zenodo.14025763. All simulation datasets are also available online: https://www.wdc-climate.de/ui/entry?acronym=PMMXMCRTDIP122 (MPII6G), https://www.wdc-climate.de/ui/entry?acronym=PMMXMCRTDGP122 (MPIG1D), https://data.ceda.ac.uk/badc/quest/data/quaternaryq/famous_glacial_cycle (FAMOUS), http://climatedata.ibs.re.kr:9090/dods/public-data/loveclim-784k (LOVECLIM), http://climatedata.ibs.re.kr:9090/dods/public-data/pcesm-3ma/pcesm-3ma-jan (CESM).

Author contributions. KR, NW and JPB designed the study. JPB processed the data and implemented the pseudo-proxy experiments with assistance from MM and input from AMD, LJ and NW. All authors discussed the results. JPB wrote the manuscript. All authors commented on earlier drafts of the manuscript and approved its final version.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We ~~thank~~ are grateful to Marie-Luise Kapsch and Uwe Mikolajewicz for ~~their help in providing the data of~~ sharing the MPI simulations ~~before~~ data ahead of their publication, as well as their insightful comments and support ~~through the development of the study~~throughout the study's development. We also thank the teams responsible for the production of the FAMOUS, CESM and Loveclim simulations, and to have provided the output online. We are deeply appreciative of our two reviewers, Bryan C. Lougheed and Kaustubh Thirumalai, who took the time to thoroughly read our paper and provided thoughtful comments that clarified and enriched its content. We extend our gratitude to our editor, Stephen Obrochta, who ensure a smoothed publication process.

This research was funded by the German Federal Ministry of Education and Research (BMBF) within the Research for Sustainability initiative through the PalMod project (grant no. 01LP1509C, 01LP1922A, 01LP1926C, 01LP2308A, and 01LP2311C), and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; project no. 395588486).

AMD was also supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; Project no. 468685498; SPP 2299/Project no. 441832482). KR is also a member of the Machine Learning Cluster of Excellence, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC number 2064/1; Project no. 390727645).

Finally, we acknowledge support from the Open Access Publication Fund of the University of Tübingen.

References

Anand, P., Elderfield, H., and Conte, M. H.: Calibration of Mg/Ca thermometry in planktonic foraminifera from a sediment trap time series, Paleoceanography, 18, https://doi.org/10.1029/2002PA000846, 2003.

Annan, J. D., Hargreaves, J. C., and Mauritsen, T.: A new global surface temperature reconstruction for the Last Glacial Maximum, Climate of the Past, 18, 1883–1896, https://doi.org/10.5194/cp-18-1883-2022, 2022.

Argus, D. F., Peltier, W. R., Drummond, R., and Moore, A. W.: The Antarctica component of postglacial rebound model ICE-6G_C (VM5a) based on GPS positioning, exposure age dating of ice thicknesses, and relative sea level histories, Geophysical Journal International, 198, 537–563, https://doi.org/10.1093/gji/ggu140, 2014.

Arora, V. K., Katavouta, A., Williams, R. G., Jones, C. D., Brovkin, V., Friedlingstein, P., Schwinger, J., Bopp, L., Boucher, O., Cadule, P., Chamberlain, M. A., Christian, J. R., Delire, C., Fisher, R. A., Hajima, T., Ilyina, T., Joetzjer, E., Kawamiya, M., Koven, C. D., Krasting, J. P., Law, R. M., Lawrence, D. M., Lenton, A., Lindsay, K., Pongratz, J., Raddatz, T., Séférian, R., Tachiiri, K., Tjiputra, J. F., Wiltshire, A., Wu, T., and Ziehn, T.: Carbon–concentration and carbon–climate feedbacks in CMIP6 models and their comparison to CMIP5 models, Biogeosciences, 17, 4173–4222, https://doi.org/10.5194/bg-17-4173-2020, 2020.

Bereiter, B., Shackleton, S., Baggenstos, D., Kawamura, K., and Severinghaus, J.: Mean global ocean temperatures during the last glacial transition, Nature, 553, 39–44, https://doi.org/10.1038/nature25152, 2018.

Berger, A. L.: Long-Term Variations of Daily Insolation and Quaternary Climatic Changes, Journal of Atmospheric Sciences, 35, 2362 – 2367, https://doi.org/10.1175/1520-0469(1978)035<2362:LTVODI>2.0.CO;2, 1978.

Berger, W. H. and Heath, G. R.: Vertical mixing in pelagic sediments, Journal of Marine Research, 26, 1968.

Blaauw, M. and Christen, J. A.: Flexible paleoclimate age-depth models using an autoregressive gamma process, Bayesian Analysis, 6, 457 – 474, https://doi.org/10.1214/11-BA618, 2011.

Boudreau, B. P.: Mean mixed depth of sediments: The wherefore and the why, Limnology and Oceanography, 43, 524–526, https://doi.org/https://doi.org/10.4319/lo.1998.43.3.0524, 1998.

Bova, S., Rosenthal, Y., Liu, Z., Godad, S. P., and Yan, M.: Seasonal origin of the thermal maxima at the Holocene and the last interglacial, Nature, 589, 548–553, https://doi.org/10.1038/s41586-020-03155-x, 2021.

Braconnot, P., Otto-Bliesner, B., Harrison, S., Joussaume, S., Peterchmitt, J.-Y., Abe-Ouchi, A., Crucifix, M., Driesschaert, E., Fichefet, T., Hewitt, C. D., Kageyama, M., Kitoh, A., Laîné, A., Loutre, M.-F., Marti, O., Merkel, U., Ramstein, G., Valdes, P., Weber, S. L., Yu, Y., and Zhao, Y.: Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum – Part 1: experiments and large-scale features, Climate of the Past, 3, 261–277, https://doi.org/10.5194/cp-3-261-2007, 2007.

Braconnot, P., Harrison, S. P., Kageyama, M., Bartlein, P. J., Masson-Delmotte, V., Abe-Ouchi, A., Otto-Bliesner, B., and Zhao, Y.: Evaluation of climate models using palaeoclimatic data, Nature Climate Change, 2, 417–424, https://doi.org/10.1038/nclimate1456, 2012.

Bradley, R.: Paleoclimatology: Reconstructing Climates of the Quaternary (3rd Edition), Academic Press/Elsevier, ISBN 9780123869135, https://doi.org/10.1016/c2009-0-18310-1, 2014.

Brierley, C. M., Zhao, A., Harrison, S. P., Braconnot, P., Williams, C. J. R., Thornalley, D. J. R., Shi, X., Peterschmitt, J.-Y., Ohgaito, R., Kaufman, D. S., Kageyama, M., Hargreaves, J. C., Erb, M. P., Emile-Geay, J., D'Agostino, R., Chandan, D., Carré, M., Bartlein, P. J., Zheng, W., Zhang, Z., Zhang, Q., Yang, H., Volodin, E. M., Tomas, R. A., Routson, C., Peltier, W. R., Otto-Bliesner, B., Morozova, P. A., McKay, N. P., Lohmann, G., Legrande, A. N., Guo, C., Cao, J., Brady, E., Annan, J. D., and Abe-Ouchi, A.: Large-scale features and

evaluation of the PMIP4-CMIP6 *midHolocene* simulations, Climate of the Past, 16, 1847–1872, https://doi.org/10.5194/cp-16-1847-2020, 2020.

Briggs, R. D., Pollard, D., and Tarasov, L.: A data-constrained large ensemble analysis of Antarctic evolution since the Eemian, Quaternary Science Reviews, 103, 91–115, https://doi.org/10.1016/j.quascirev.2014.09.003, 2014.

Brovkin, V., Ganopolski, A., Archer, D., and Munhoven, G.: Glacial CO_2 cycle as a succession of key physical and biogeochemical processes, Climate of the Past, 8, 251–264, https://doi.org/10.5194/cp-8-251-2012, 2012.

Clark, P. U., Shakun, J. D., Rosenthal, Y., Köhler, P., and Bartlein, P. J.: Global and regional temperature change over the past 4.5 million years, Science, 383, 884–890, https://doi.org/10.1126/science.adi1908, 2024.

Consortium, P.: A global multiproxy database for temperature reconstructions of the Common Era, Scientific Data, 4, 170 088, https://doi.org/10.1038/sdata.2017.88, 2017.

Dolman, A. and Laepple, T.: sedproxy: Simulation of Sediment Archived Climate Proxy Records, https://cran.r-project.org/web/packages/sedproxy/sedproxy.pdf, r package, 2023.

Dolman, A. M. and Laepple, T.: Sedproxy: a forward model for sediment-archived climate proxies, Climate of the Past, 14, 1851–1868, https://doi.org/10.5194/cp-14-1851-2018, 2018.

Dolman, A. M., Kunz, T., Groeneveld, J., and Laepple, T.: A spectral approach to estimating the timescale-dependent uncertainty of paleoclimate records – Part 2: Application and interpretation, Climate of the Past, 17, 825–841, https://doi.org/10.5194/cp-17-825-2021, 2021.

Erokhina, O.: A new Eulerian iceberg module for climate studies: Formulation and application to the investigation of the sensitivity of the AMOC to iceberg calving, Ph.D. thesis, Universität Hamburg Hamburg, https://pure.mpg.de/rest/items/item_3270743/component/file_3270744/content, 2020.

Fieg, K., Latif, M., Ilyina, T., and Schulz, M.: From the last interglacial to the future – new insights into climate change from the PalMod Earth System modelling framework, in: EGU General Assembly 2023, pp. EGU23–10 048, Vienna, Austria, https://doi.org/10.5194/egusphere-egu23-10048, 2023.

Friedrich, T. and Timmermann, A.: Using Late Pleistocene sea surface temperature reconstructions to constrain future greenhouse warming, Earth and Planetary Science Letters, 530, 115 911, https://doi.org/10.1016/j.epsl.2019.115911, 2020.

Ganopolski, A. and Calov, R.: The role of orbital forcing, carbon dioxide and regolith in 100 kyr glacial cycles, Climate of the Past, 7, 1415–1425, https://doi.org/10.5194/cp-7-1415-2011, 2011.

Goosse, H., Driesschaert, E., Fichefet, T., and Loutre, M.-F.: Information on the early Holocene climate constrains the summer sea ice projections for the 21st century, Climate of the Past, 3, 683–692, https://doi.org/10.5194/cp-3-683-2007, 2007.

Goosse, H., Brovkin, V., Fichefet, T., Haarsma, R., Huybrechts, P., Jongma, J., Mouchet, A., Selten, F., Barriat, P.-Y., Campin, J.-M., Deleersnijder, E., Driesschaert, E., Goelzer, H., Janssens, I., Loutre, M.-F., Morales Maqueda, M. A., Opsteegh, T., Mathieu, P.-P., Munhoven, G., Pettersson, E. J., Renssen, H., Roche, D. M., Schaeffer, M., Tartinville, B., Timmermann, A., and Weber, S. L.: Description of the Earth system model of intermediate complexity LOVECLIM version 1.2, Geoscientific Model Development, 3, 603–633, https://doi.org/10.5194/gmd-3-603-2010, 2010.

Gordon, C., Cooper, C., Senior, C. A., Banks, H., Gregory, J. M., Johns, T. C., Mitchell, J. F. B., and Wood, R. A.: The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments, Climate Dynamics, 16, 147–168, https://doi.org/10.1007/s003820050010, 2000.

Ho, S. L. and Laepple, T.: Glacial cooling as inferred from marine temperature proxies TEXH86 and UK′37, Earth and Planetary Science Letters, 409, 15–22, https://doi.org/10.1016/j.epsl.2014.10.033, 2015.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J.-F., Large, W. G., Lawrence, D., Lindsay, K., Lipscomb, W. H., Long, M. C., Mahowald, N., Marsh, D. R., Neale, R. B., Rasch, P., Vavrus, S., Vertenstein, M., Bader, D., Collins, W. D., Hack, J. J., Kiehl, J., and Marshall, S.: The Community Earth System Model: A Framework for Collaborative Research, Bulletin of the American Meteorological Society, 94, 1339 – 1360, https://doi.org/10.1175/BAMS-D-12-00121.1, 2013.

Imbrie, J.: A new micropaleontological method for quantitative paleoclimatology: application to a late Pleistocene Caribbean core, The late Cenozoic glacial ages, pp. 71–181, 1971.

Ivanovic, R. F., Gregoire, L. J., Kageyama, M., Roche, D. M., Valdes, P. J., Burke, A., Drummond, R., Peltier, W. R., and Tarasov, L.: Transient climate simulations of the deglaciation 21–9 thousand years before present (version 1) – PMIP4 Core experiment design and boundary conditions, Geoscientific Model Development, 9, 2563–2587, https://doi.org/10.5194/gmd-9-2563-2016, 2016.

Jaume-Santero, F., Barriopedro, D., García-Herrera, R., Calvo, N., and Salcedo-Sanz, S.: Selection of optimal proxy locations for temperature field reconstructions using evolutionary algorithms, Scientific Reports, 10, 7900, https://doi.org/10.1038/s41598-020-64459-6, 2020.

Jonkers, L. and Kučera, M.: Quantifying the effect of seasonal and vertical habitat tracking on planktonic foraminifera proxies, Climate of the Past, 13, 573–586, https://doi.org/10.5194/cp-13-573-2017, 2017.

Jonkers, L., Cartapanis, O., Langner, M., McKay, N., Mulitza, S., Strack, A., and Kucera, M.: Integrating palaeoclimate time series with rich metadata for uncertainty modelling: strategy and documentation of the PalMod 130k marine palaeoclimate data synthesis, Earth System Science Data, 12, 1053–1081, https://doi.org/10.5194/essd-12-1053-2020, 2020.

Jonkers, L., Cartapanis, O., Langner, M., McKay, N., Mulitza, S., Strack, A., and Kucera, M.: PALMOD 130k marine palaeoclimate data synthesis V1_0_1, https://doi.org/10.1594/PANGAEA.914466, 2020.

Jonkers, L., Laepple, T., Rillo, M. C., Shi, X., Dolman, A. M., Lohmann, G., Paul, A., Mix, A., and Kucera, M.: Strong temperature gradients in the ice age North Atlantic Ocean revealed by plankton biogeography, Nature Geoscience, 16, 1114–1119, https://doi.org/10.1038/s41561-023-01328-7, 2023.

Judd, E. J., Bhattacharya, T., and Ivany, L. C.: A Dynamical Framework for Interpreting Ancient Sea Surface Temperatures, Geophysical Research Letters, 47, e2020GL089 044, https://doi.org/10.1029/2020GL089044, e2020GL089044 2020GL089044, 2020.

Kageyama, M., Harrison, S. P., Kapsch, M.-L., Lofverstrom, M., Lora, J. M., Mikolajewicz, U., Sherriff-Tadano, S., Vadsaria, T., Abe-Ouchi, A., Bouttes, N., Chandan, D., Gregoire, L. J., Ivanovic, R. F., Izumi, K., LeGrande, A. N., Lhardy, F., Lohmann, G., Morozova, P. A., Ohgaito, R., Paul, A., Peltier, W. R., Poulsen, C. J., Quiquet, A., Roche, D. M., Shi, X., Tierney, J. E., Valdes, P. J., Volodin, E., and Zhu, J.: The PMIP4 Last Glacial Maximum experiments: preliminary results and comparison with the PMIP3 simulations, Climate of the Past, 17, 1065–1089, https://doi.org/10.5194/cp-17-1065-2021, 2021.

Kapsch, M.-L., Mikolajewicz, U., Ziemen, F., and Schannwell, C.: Ocean Response in Transient Simulations of the Last Deglaciation Dominated by Underlying Ice-Sheet Reconstruction and Method of Meltwater Distribution, Geophysical Research Letters, 49, e2021GL096 767, https://doi.org/10.1029/2021GL096767, e2021GL096767 2021GL096767, 2022.

Kaufman, D., McKay, N., Routson, C., Erb, M., Dätwyler, C., Sommer, P. S., Heiri, O., and Davis, B.: Holocene global mean surface temperature, a multi-method reconstruction approach, Scientific Data, 7, 201, https://doi.org/10.1038/s41597-020-0530-7, 2020.

Kleinen, T., Gromov, S., Steil, B., and Brovkin, V.: Atmospheric methane underestimated in future climate projections, Environmental Research Letters, 16, 094 006, https://doi.org/10.1088/1748-9326/ac1814, 2021.

Köhler, P., Nehrbass-Ahles, C., Schmitt, J., Stocker, T. F., and Fischer, H.: A 156 kyr smoothed history of the atmospheric greenhouse gases CO_2, CH_4, and N_2O and their radiative forcing, Earth System Science Data, 9, 363–387, https://doi.org/10.5194/essd-9-363-2017, 2017.

Laepple, T., Shakun, J., He, F., and Marcott, S.: Concerns of assuming linearity in the reconstruction of thermal maxima, Nature, 607, E12–E14, https://doi.org/10.1038/s41586-022-04831-w, 2022.

765   Laepple, T., Ziegler, E., Weitzel, N., Hébert, R., Ellerhoff, B., Schoch, P., Martrat, B., Bothe, O., Moreno-Chamarro, E., Chevalier, M., Herbert, A., and Rehfeld, K.: Regional but not global temperature variability underestimated by climate models at supradecadal timescales, Nature Geoscience, 16, 958–966, https://doi.org/10.1038/s41561-023-01299-9, 2023.

Latif, M., Claussen, M., Schulz, M., and Brücher, T.: Comprehensive Earth system models of the last glacial cycle, Eos, 97, https://doi.org/10.1029/2016EO059587, published on 23 September 2016, 2016.

770   Lisiecki, L. E. and Stern, J. V.: Regional and global benthic $\delta18O$ stacks for the last glacial cycle, Paleoceanography, 31, 1368–1394, https://doi.org/10.1002/2016PA003002, 2016.

Liu, Z., Zhu, J., Rosenthal, Y., Zhang, X., Otto-Bliesner, B. L., Timmermann, A., Smith, R. S., Lohmann, G., Zheng, W., and Timm, O. E.: The Holocene temperature conundrum, Proceedings of the National Academy of Sciences, 111, E3501–E3505, https://doi.org/10.1073/pnas.1407229111, 2014.

775   Lougheed, B. C.: Using sedimentological priors to improve 14C calibration of bioturbated sediment archives., Radiocarbon, 64, 135–151, https://doi.org/10.1017/RDC.2021.116, 2022.

Lunt, D. J., Bragg, F., Chan, W.-L., Hutchinson, D. K., Ladant, J.-B., Morozova, P., Niezgodzki, I., Steinig, S., Zhang, Z., Zhu, J., Abe-Ouchi, A., Anagnostou, E., de Boer, A. M., Coxall, H. K., Donnadieu, Y., Foster, G., Inglis, G. N., Knorr, G., Langebroek, P. M., Lear, C. H., Lohmann, G., Poulsen, C. J., Sepulchre, P., Tierney, J. E., Valdes, P. J., Volodin, E. M., Dunkley Jones, T., Hollis, C. J., Huber, M.,

780   and Otto-Bliesner, B. L.: DeepMIP: model intercomparison of early Eocene climatic optimum (EECO) large-scale climate features and comparison with proxy data, Climate of the Past, 17, 203–227, https://doi.org/10.5194/cp-17-203-2021, 2021.

Lüthi, D., Le Floch, M., Bereiter, B., Blunier, T., Barnola, J.-M., Siegenthaler, U., Raynaud, D., Jouzel, J., Fischer, H., Kawamura, K., and Stocker, T. F.: High-resolution carbon dioxide concentration record 650,000–800,000 years before present, Nature, 453, 379–382, https://doi.org/10.1038/nature06949, 2008.

785   Marsicek, J., Shuman, B. N., Bartlein, P. J., Shafer, S. L., and Brewer, S.: Reconciling divergent trends and millennial variations in Holocene temperatures, Nature, 554, 92–96, https://doi.org/10.1038/nature25464, 2018.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Brovkin, V., Claussen, M., Crueger, T., Esch, M., Fast, I., Fiedler, S., Fläschner, D., Gayler, V., Giorgetta, M., Goll, D. S., Haak, H., Hagemann, S., Hedemann, C., Hohenegger, C., Ilyina, T., Jahns, T., Jimenéz-de-la Cuesta, D., Jungclaus, J., Kleinen, T., Kloster, S., Kracher, D., Kinne, S., Kleberg, D., Lasslop, G., Kornblueh, L.,

790   Marotzke, J., Matei, D., Meraner, K., Mikolajewicz, U., Modali, K., Möbis, B., Müller, W. A., Nabel, J. E. M. S., Nam, C. C. W., Notz, D., Nyawira, S.-S., Paulsen, H., Peters, K., Pincus, R., Pohlmann, H., Pongratz, J., Popp, M., Raddatz, T. J., Rast, S., Redler, R., Reick, C. H., Rohrschneider, T., Schemann, V., Schmidt, H., Schnur, R., Schulzweida, U., Six, K. D., Stein, L., Stemmler, I., Stevens, B., von Storch, J.-S., Tian, F., Voigt, A., Vrese, P., Wieners, K.-H., Wilkenskjeld, S., Winkler, A., and Roeckner, E.: Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and Its Response to Increasing CO2, Journal of Advances in Modeling Earth Systems,

795   11, 998–1038, https://doi.org/10.1029/2018MS001400, 2019.

Muntjewerf, L., Petrini, M., Vizcaino, M., Ernani da Silva, C., Sellevold, R., Scherrenberg, M. D. W., Thayer-Calder, K., Bradley, S. L., Lenaerts, J. T. M., Lipscomb, W. H., and Lofverstrom, M.: Greenland Ice Sheet Contribution to 21st Century Sea Level Rise as Simu-

lated by the Coupled CESM2.1-CISM2.1, Geophysical Research Letters, 47, e2019GL086 836, https://doi.org/10.1029/2019GL086836, e2019GL086836 10.1029/2019GL086836, 2020.

800   Müller, P. J., Kirst, G., Ruhland, G., von Storch, I., and Rosell-Melé, A.: Calibration of the alkenone paleotemperature index U37K′ based on core-tops from the eastern South Atlantic and the global ocean (60°N-60°S), Geochimica et Cosmochimica Acta, 62, 1757–1772, https://doi.org/10.1016/S0016-7037(98)00097-0, 1998.

Nilsen, T., Talento, S., and Werner, J. P.: Constraining two climate field reconstruction methodologies over the North Atlantic realm using pseudo-proxy experiments, Quaternary Science Reviews, 265, 107 009, https://doi.org/10.1016/j.quascirev.2021.107009, 2021.

805   Nowicki, S. M. J., Payne, A., Larour, E., Seroussi, H., Goelzer, H., Lipscomb, W., Gregory, J., Abe-Ouchi, A., and Shepherd, A.: Ice Sheet Model Intercomparison Project (ISMIP6) contribution to CMIP6, Geoscientific Model Development, 9, 4521–4545, https://doi.org/10.5194/gmd-9-4521-2016, 2016.

Nürnberg, D., Bijma, J., and Hemleben, C.: Assessing the reliability of magnesium in foraminiferal calcite as a proxy for water mass temperatures, Geochimica et Cosmochimica Acta, 60, 803–814, https://doi.org/10.1016/0016-7037(95)00446-7, 1996.

810   Osman, M. B., Tierney, J. E., Zhu, J., Tardif, R., Hakim, G. J., King, J., and Poulsen, C. J.: Globally resolved surface temperatures since the Last Glacial Maximum, Nature, 599, 239–244, https://doi.org/10.1038/s41586-021-03984-4, 2021.

PAGES 2k Consortium: Consistent multidecadal variability in global temperature reconstructions and simulations over the Common Era, Nature Geoscience, 12, 643–649, https://doi.org/10.1038/s41561-019-0400-0, 2019.

Paul, A., Mulitza, S., Stein, R., and Werner, M.: A global climatology of the ocean surface during the Last Glacial Maximum mapped on a
815   regular grid (GLOMAP), Climate of the Past, 17, 805–824, https://doi.org/10.5194/cp-17-805-2021, 2021.

Peeters, F. J. C., van der Lubbe, H. J. L., and Scussolini, P.: Age-Depth Models for Tropical Marine Hemipelagic Deposits Improve Significantly When Proxy-Based Information on Sediment Composition Is Included, Paleoceanography and Paleoclimatology, 38, e2022PA004 476, https://doi.org/10.1029/2022PA004476, e2022PA004476 2022PA004476, 2023.

Peltier, W.: global glacial isostasy and the surface of the ice-age earth: The ICE-5G (VM2) Model and GRACE, Annual Review of Earth and
820   Planetary Sciences, 32, 111–149, https://doi.org/https://doi.org/10.1146/annurev.earth.32.082503.144359, 2004.

Peltier, W. R., Argus, D. F., and Drummond, R.: Space geodesy constrains ice age terminal deglaciation: The global ICE-6G_C (VM5a) model, Journal of Geophysical Research: Solid Earth, 120, 450–487, https://doi.org/10.1002/2014JB011176, 2015.

Prahl, F. G., Muehlhausen, L. A., and Zahnle, D. L.: Further evaluation of long-chain alkenones as indicators of paleoceanographic conditions, Geochimica et Cosmochimica Acta, 52, 2303–2310, https://doi.org/0.1016/0016-7037(88)90132-9, 1988.

825   Rackow, T., Wesche, C., Timmermann, R., Hellmer, H. H., Juricke, S., and Jung, T.: A simulation of small to giant Antarctic iceberg evolution: Differential impact on climatology estimates, Journal of Geophysical Research: Oceans, 122, 3170–3190, https://doi.org/10.1002/2016JC012513, 2017.

Rampen, S. W., Willmott, V., Kim, J.-H., Uliana, E., Mollenhauer, G., Schefuß, E., Sinninghe Damsté, J. S., and Schouten, S.: Long chain 1,13- and 1,15-diols as a potential proxy for palaeotemperature reconstruction, Geochimica et Cosmochimica Acta, 84, 204–216,
830   https://doi.org/10.1016/j.gca.2012.01.024, 2012.

Roche, D. M., Waelbroeck, C., Metcalfe, B., and Caley, T.: FAME (v1.0): a simple module to simulate the effect of planktonic foraminifer species-specific habitat on their oxygen isotopic content, Geoscientific Model Development, 11, 3587–3603, https://doi.org/10.5194/gmd-11-3587-2018, 2018.

Rohling, E. J., Medina-Elizalde, M., Shepherd, J. G., Siddall, M., and Stanford, J. D.: Sea Surface and High-Latitude Temperature Sensitivity to Radiative Forcing of Climate over Several Glacial Cycles, Journal of Climate, 25, 1635 – 1656, https://doi.org/10.1175/2011JCLI4078.1, 2012.

Schouten, S., Hopmans, E. C., Schefuß, E., and Sinninghe Damsté, J. S.: Distributional variations in marine crenarchaeotal membrane lipids: a new tool for reconstructing ancient sea water temperatures?, Earth and Planetary Science Letters, 204, 265–274, https://doi.org/10.1016/S0012-821X(02)00979-2, 2002.

Shakun, J. D., Clark, P. U., He, F., Marcott, S. A., Mix, A. C., Liu, Z., Otto-Bliesner, B., Schmittner, A., and Bard, E.: Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation, Nature, 484, 49–54, https://doi.org/10.1038/nature10915, 2012.

Smith, R. S. and Gregory, J.: The last glacial cycle: transient simulations with an AOGCM, Climate Dynamics, 38, 1545–1559, https://doi.org/10.1007/s00382-011-1283-y, 2012.

Smith, R. S., Gregory, J. M., and Osprey, A.: A description of the FAMOUS (version XDBUA) climate model and control run, Geoscientific Model Development, 1, 53–68, https://doi.org/10.5194/gmd-1-53-2008, 2008.

Snyder, C. W.: Evolution of global temperature over the past two million years, Nature, 538, 226, https://doi.org/10.1038/nature19798, 2016.

Song, X., Wang, D.-Y., Li, F., and Zeng, X.-D.: Evaluating the performance of CMIP6 Earth system models in simulating global vegetation structure and distribution, Advances in Climate Change Research, 12, 584–595, https://doi.org/10.1016/j.accre.2021.06.008, special Issue on Arctic Rapid Change, 2021.

Spahni, R., Chappellaz, J., Stocker, T. F., Loulergue, L., Hausammann, G., Kawamura, K., Flückiger, J., Schwander, J., Raynaud, D., Masson-Delmotte, V., and Jouzel, J.: Atmospheric Methane and Nitrous Oxide of the Late Pleistocene from Antarctic Ice Cores, Science, 310, 1317–1321, https://doi.org/10.1126/science.1120132, 2005.

Steinke, S., Kienast, M., Groeneveld, J., Lin, L.-C., Chen, M.-T., and Rendle-Bühring, R.: Proxy dependence of the temporal pattern of deglacial warming in the tropical South China Sea: toward resolving seasonality, Quaternary Science Reviews, 27, 688–700, https://doi.org/10.1016/j.quascirev.2007.12.003, 2008.

Tarasov, L., Dyke, A. S., Neal, R. M., and Peltier, W.: A data-calibrated distribution of deglacial chronologies for the North American ice complex from glaciological modeling, Earth and Planetary Science Letters, 315-316, 30–40, https://doi.org/10.1016/j.epsl.2011.09.010, sea Level and Ice Sheet Evolution: A PALSEA Special Edition, 2012.

Teal, L. R., Bulling, M. T., Parker, E. R., and Solan, M.: Global patterns of bioturbation intensity and mixed depth of marine soft sediments, Aquatic Biology, 2, 207–218, https://doi.org/10.3354/ab00052, 2008.

Telford, R. J., Li, C., and Kucera, M.: Mismatch between the depth habitat of planktonic foraminifera and the calibration depth of SST transfer functions may bias reconstructions, Climate of the Past, 9, 859–870, https://doi.org/10.5194/cp-9-859-2013, 2013.

Tierney, J. E. and Tingley, M. P.: BAYSPLINE: A New Calibration for the Alkenone Paleothermometer, Paleoceanography and Paleoclimatology, 33, 281–301, https://doi.org/10.1002/2017PA003201, 2018.

Tierney, J. E., Zhu, J., King, J., Malevich, S. B., Hakim, G. J., and Poulsen, C. J.: Glacial cooling and climate sensitivity revisited, Nature, 584, 569–573, https://doi.org/10.1038/s41586-020-2617-x, 2020.

Timmermann, A. and Friedrich, T.: Late Pleistocene climate drivers of early human migration, Nature, 538, 92–95, https://doi.org/10.1038/nature19365, 2016.

Timmermann, A., Sachs, J., and Timm, O. E.: Assessing divergent SST behavior during the last 21 ka derived from alkenones and G. ruber-Mg/Ca in the equatorial Pacific, Paleoceanography, 29, 680–696, https://doi.org/10.1002/2013PA002598, 2014.

UNFCCC: Paris Agreement, https://unfccc.int/sites/default/files/english_paris_agreement.pdf, 2015.

Von Storch, H. and Zwiers, F. W.: Statistical analysis in climate research, Statistical analysis in climate research, ISBN Statistical analysis in climate research, https://doi.org/10.1017/CBO9780511612336, 1999.

875   Waelbroeck, C., Lougheed, B. C., Vazquez Riveiros, N., Missiaen, L., Pedro, J., Dokken, T., Hajdas, I., Wacker, L., Abbott, P., Dumoulin, J.-P., Thil, F., Eynaud, F., Rossignol, L., Fersi, W., Albuquerque, A. L., Arz, H., Austin, W. E. N., Came, R., Carlson, A. E., Collins, J. A., Dennielou, B., Desprat, S., Dickson, A., Elliot, M., Farmer, C., Giraudeau, J., Gottschalk, J., Henderiks, J., Hughen, K., Jung, S., Knutz, P., Lebreiro, S., Lund, D. C., Lynch-Stieglitz, J., Malaizé, B., Marchitto, T., Martínez-Méndez, G., Mollenhauer, G., Naughton, F., Nave, S., Nürnberg, D., Oppo, D., Peck, V., Peeters, F. J. C., Penaud, A., Portilho-Ramos, R. d. C., Repschläger, J., Roberts, J., Rühlemann, C.,

880   Salgueiro, E., Sanchez Goni, M. F., Schönfeld, J., Scussolini, P., Skinner, L. C., Skonieczny, C., Thornalley, D., Toucanne, S., Rooij, D. V., Vidal, L., Voelker, A. H. L., Wary, M., Weldeab, S., and Ziegler, M.: Consistently dated Atlantic sediment cores over the last 40 thousand years, Scientific Data, 6, 165, https://doi.org/10.1038/s41597-019-0173-8, 2019.

Wang, J., Emile-Geay, J., Guillot, D., Smerdon, J. E., and Rajaratnam, B.: Evaluating climate field reconstruction techniques using improved emulations of real-world conditions, Climate of the Past, 10, 1–19, https://doi.org/10.5194/cp-10-1-2014, 2014.

885   Weitzel, N., Andres, H., Baudouin, J.-P., Kapsch, M.-L., Mikolajewicz, U., Jonkers, L., Bothe, O., Ziegler, E., Kleinen, T., Paul, A., and Rehfeld, K.: Towards spatio-temporal comparison of simulated and reconstructed sea surface temperatures for the last deglaciation, Climate of the Past, 20, 865–890, https://doi.org/10.5194/cp-20-865-2024, 2024.

Willeit, M., Ganopolski, A., Calov, R., and Brovkin, V.: Mid-Pleistocene transition in glacial cycles explained by declining CO_2 and regolith removal, Science Advances, 5, eaav7337, https://doi.org/10.1126/sciadv.aav7337, 2019.

890   Yun, K.-S., Timmermann, A., Lee, S.-S., Willeit, M., Ganopolski, A., and Jadhav, J.: A transient coupled general circulation model (CGCM) simulation of the past 3 million years, Climate of the Past, 19, 1951–1974, https://doi.org/10.5194/cp-19-1951-2023, 2023.

Zhang, S., Solan, M., and Tarhan, L.: Global distribution and environmental correlates of marine bioturbation, Current Biology, 34, 2580–2593.e4, https://doi.org/10.1016/j.cub.2024.04.065, 2024.

Ziemen, F. A., Kapsch, M.-L., Klockmann, M., and Mikolajewicz, U.: Heinrich events show two-stage climate response in transient glacial

895   simulations, Climate of the Past, 15, 153–168, https://doi.org/10.5194/cp-15-153-2019, 2019.

Zuhr, A. M., Dolman, A. M., Ho, S. L., Groeneveld, J., Löwemark, L., Grotheer, H., Su, C.-C., and Laepple, T.: Age-Heterogeneity in Marine Sediments Revealed by Three-Dimensional High-Resolution Radiocarbon Measurements, Frontiers in Earth Science, 10, https://doi.org/10.3389/feart.2022.871902, 2022.
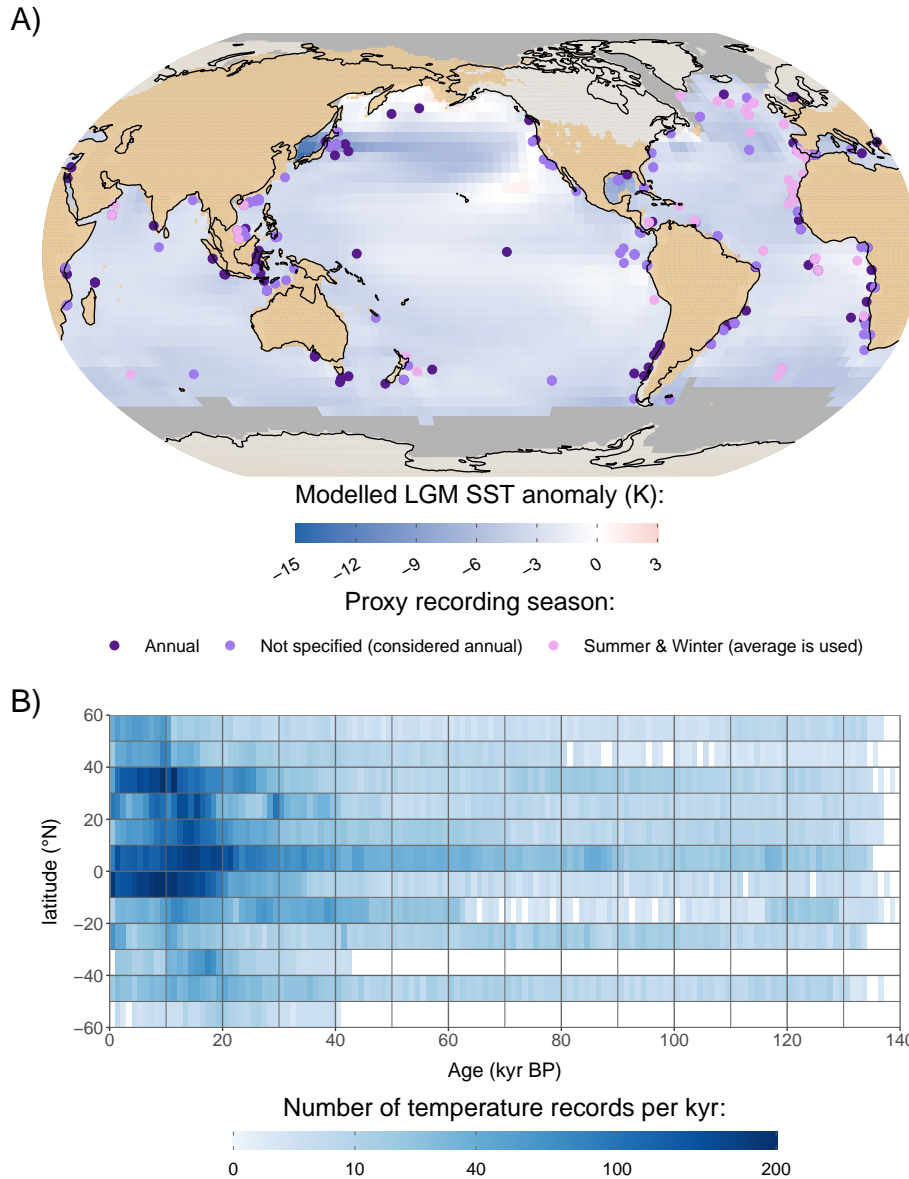
Figure 1. Spatial and temporal distribution of the selected temperature proxy records. Panel A shows the locations of inferred SST time series between 60°N and 60°S available in version 2.0 of the *PalMod 130k marine palaeoclimate data synthesis* and used for this study. Colours indicate whether the recording seasons of the time series are labelled as annual (violet), not indicated (purple), or whether both summer and winter time series are available (lilac). The background shows, over the ocean, the LGM (19-23 kyr BP) anomaly of sea surface temperature (SST) from MPI-I6G as an example (see sect. 2.2). Ice-free land (brown) and ice sheets (light grey) are taken from ICE-6G (Argus et al., 2014; Peltier et al., 2015). Darker grey in the ocean corresponds to sea-ice-covered areas in the simulation. Panel B represents the latitudinal and temporal distribution of the SST proxy time series from the above panel. The number of proxy data points is given per latitudinal bands of 10° and 1 kyr.
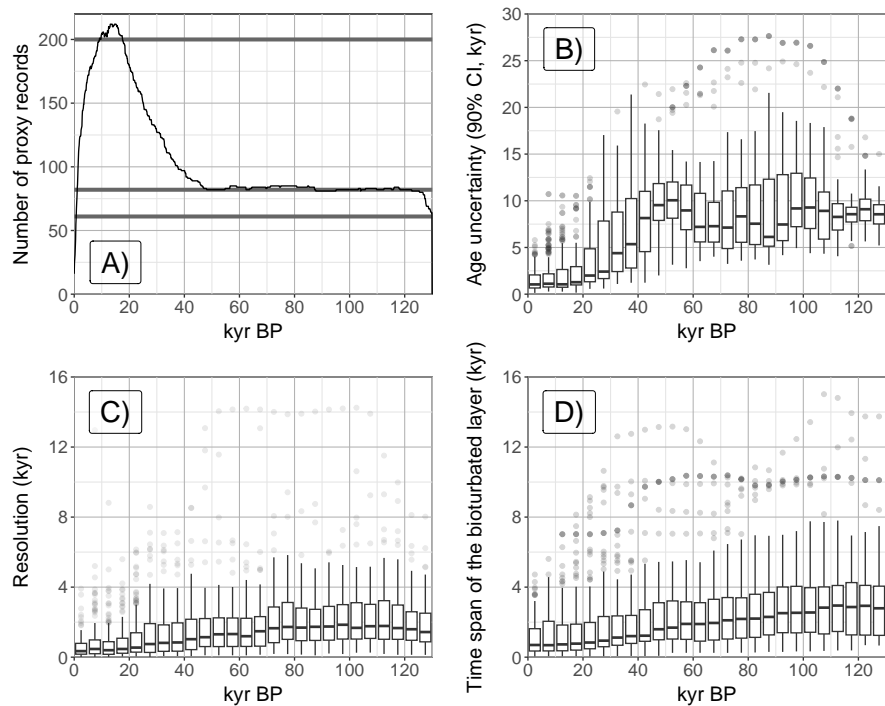
Figure 2. Characteristics of the proxy records in the J20 database. In panel A, the number of records is computed using the entire time span of the time series. The grey horizontal lines correspond to the values 61, 82, and 200, used in the pseudo-proxy experiments (cf. Tab. 2). In panels B, C and D, the investigated characteristic is averaged over a $5\,\mathrm{kyr}$ bin for each ~~timeseries~~time series, so that the box plots represent the distribution over the ensemble of records. The bioturbated layer in panel D is computed as the bioturbated layer width (fixed at $10\,\mathrm{cm}$) divided by the accumulation rate, so that the layer width can be interpreted as a period of time.
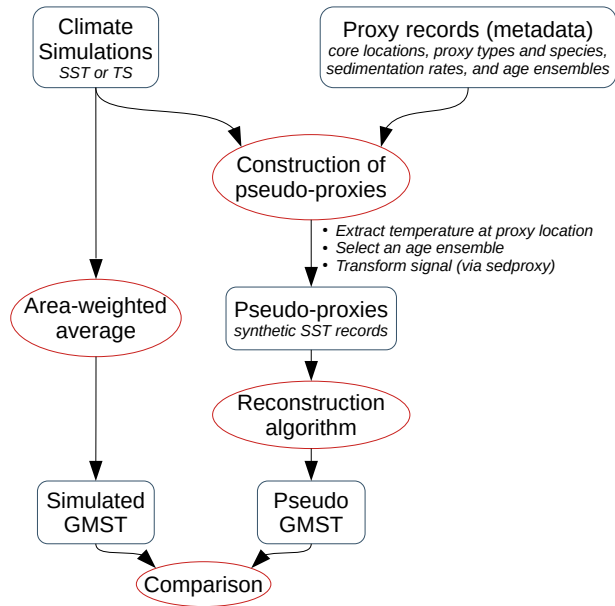
Figure 3. Flowchart summarising the data processing for a pseudo-proxy experiment. Rectangles are for data, while ellipses are for data processing. The metadata proxy records are from the J20 database. The reconstruction algorithm is the modified version of S16. The signal transformation from *sedproxy* includes: habitat preferences, bioturbation, and measurement and calibration noise
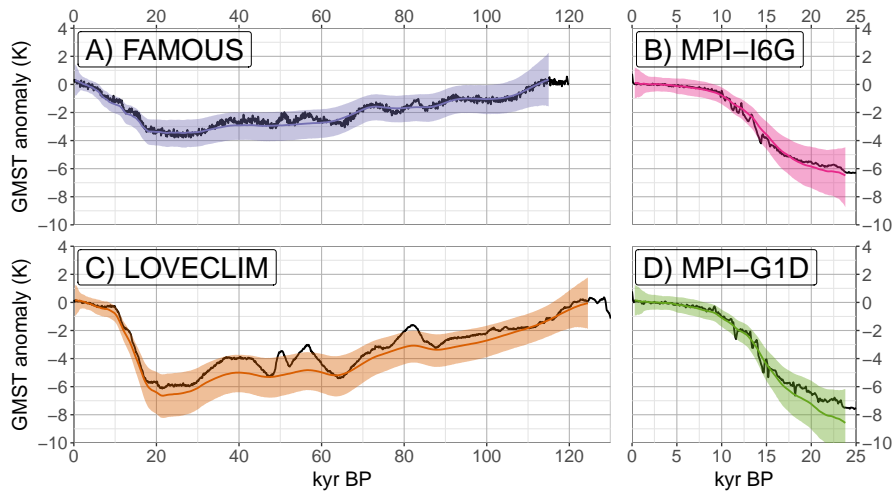
Figure 4. ~~Reconstructed~~ Comparison of the simulated GMST (black lines) and the corresponding reconstructed GMST (coloured line) for each simulation. The reconstructed GMST is based on the pseudo-proxies from the *Full PP* experiment. The ~~coloured lines are the mean of the reconstructed GMST ensemble, while the~~ shading corresponds to the 90% confidence interval of the reconstruction. We consider here the mean across the ten computed members. The ~~black lines are the simulated GMSTs. The~~ anomaly is computed with respect to the late Holocene (0 to 5 kyr BP).
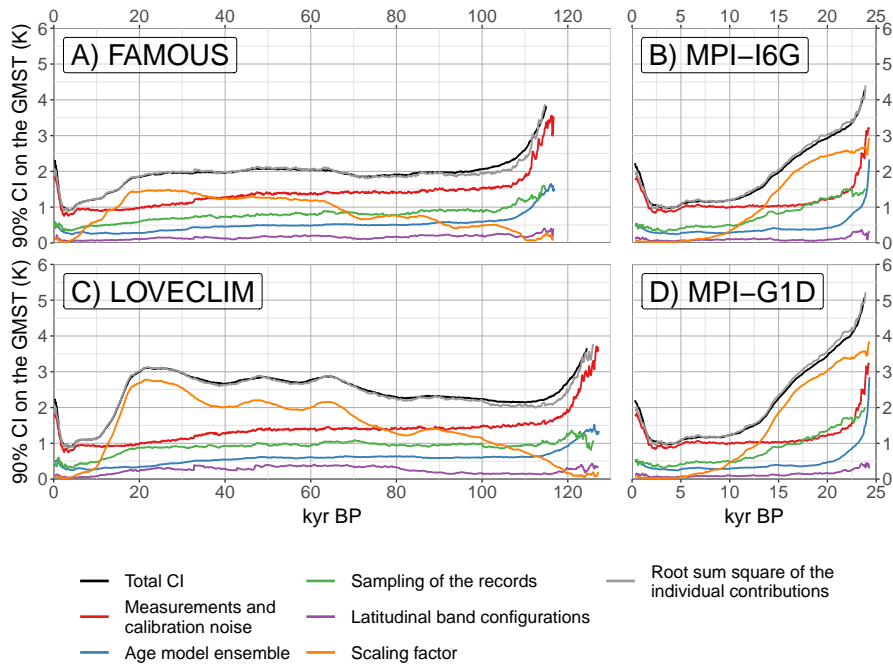
Figure 5. Estimation of various sources of uncertainty for the *Full PP* experiment. The uncertainty is quantified as the width of the 90% confidence interval (CI) of the reconstructed GMST ensemble. The black lines correspond to the CI in fig 4. The coloured lines are the individual contributions of each source of uncertainty quantified in the GMST reconstruction algorithm. Note that we interpret the CI arising from the selection of record as a location resampling. The grey line is the root sum square of the individual contribution. It should correspond to the black line if the individual contributions are independent of each other.
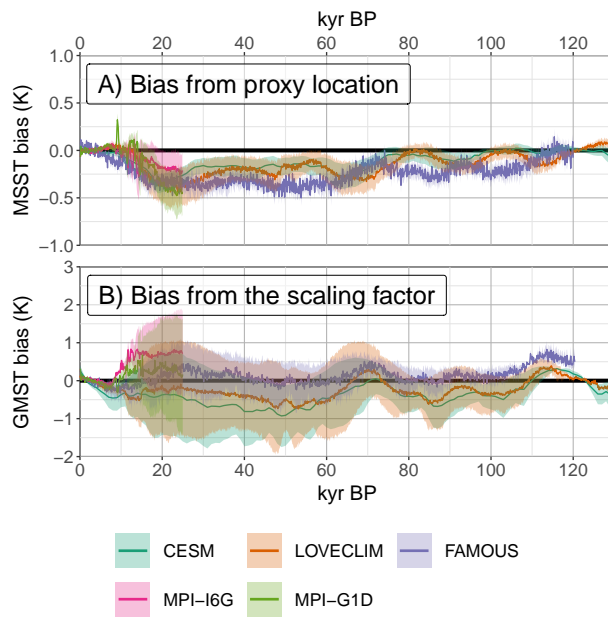
Figure 6. Potential reconstruction biases and evaluation of the confidence interval. In panel A, the $60°$S-$60°$N mean SST (MSST) reconstruction from the experiment *SST at proxy locations* is compared to the simulated MSST. The shaded area is the 90% CI, only estimating uncertainty from location resampling and the latitudinal band configuration. A properly quantified confidence interval should overlay the 0 line 90% of the time. In panel B, we apply the scaling factor with uncertainty to the simulated MSST. We then ~~compare the result to~~ subtract the GMST from the result. As above, a properly quantified confidence interval should overlay the 0 line 90% of the time.
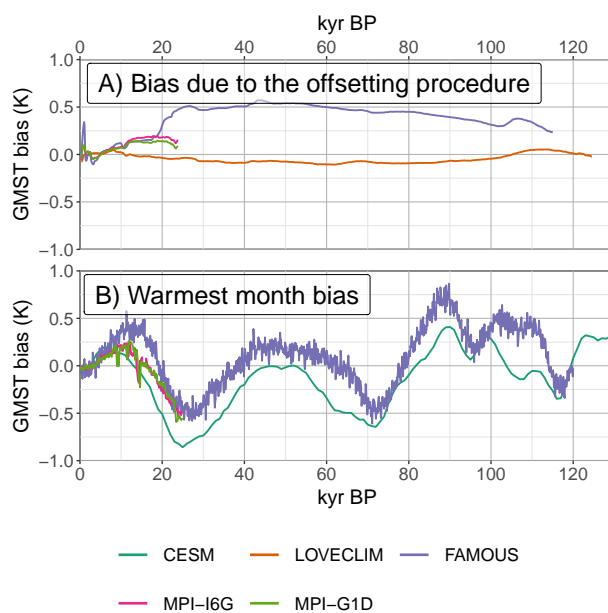
Figure 7. Other potential sources of biases, non-quantified in the uncertainty range. In panel A, the GMST reconstruction from the *Full PP* experiment is ~~compared to~~ subtracted from the one of the *no offsetting* experiment. Positive values mean that the offsetting procedure produces a too warm reconstruction. In panel B, the GMST reconstruction from the ~~*warmest month*~~ *SST at proxy locations* experiment is ~~compared to~~ subtracted from the one of the ~~*SST at proxy locations*~~ *warmest month* experiment. Positive values mean that reconstructions based on proxies biased towards the warm season are too warm.
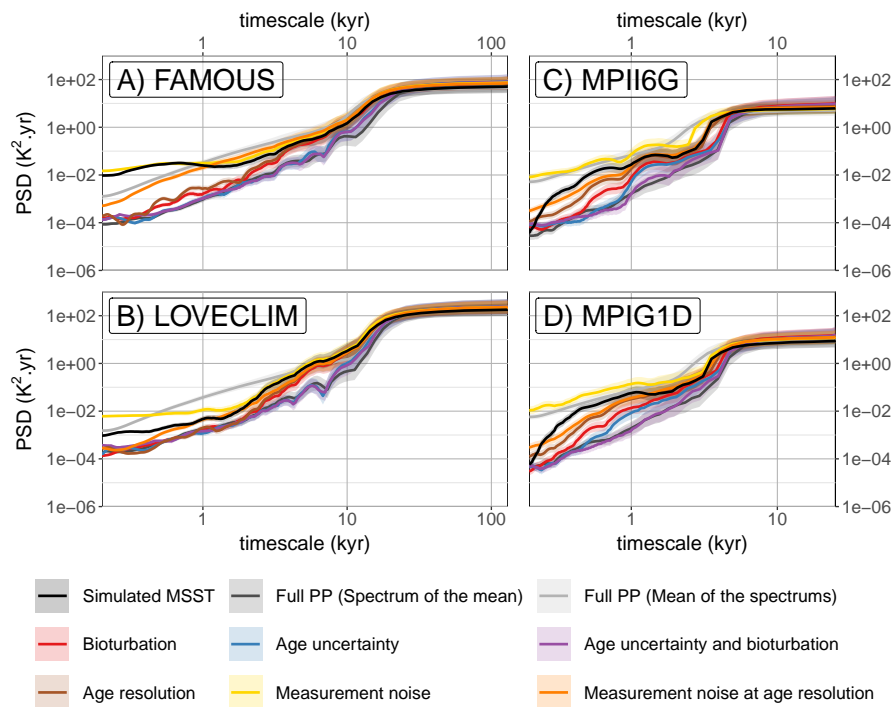
Figure 8. Loss or gain of variability at specific timescales between reconstructed and simulated MSST. The black line is the Power Spectral Density (PSD) of the simulated MSST. The other lines correspond to various PPEs defined in tab. 2. For the *Full PP* experiment, we consider both the PSD of the ensemble mean, and the PSD of each ensemble member. Only the ensemble mean is considered for the other PPEs. Finally, the shading corresponds to the confidence interval at the level 0.05 of the PSD estimates.
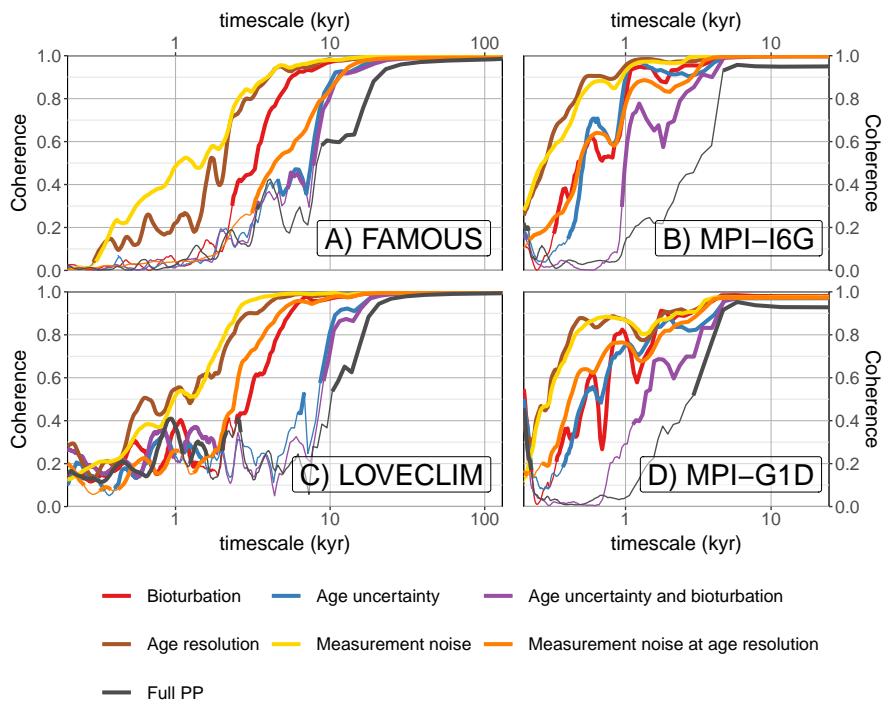
Figure 9. Disentangling the origin of the loss of signal at high frequencies. The loss of signal is quantified by the coherence between the simulated and reconstructed MSST for various PPEs. For the latter, we use the ensemble mean. The thicker lines correspond to significant values at the level of 0.05. Specifically, they are values above the 95 percentile of the coherences between the simulated MSST and randomly phase-shifted time series of the same characteristics.