Clim. Past, 20, 865–890, 2024
https://doi.org/10.5194/cp-20-865-2024

Climate of the Past
Open Access
EGU

Towards spatio-temporal comparison of simulated and reconstructed sea surface temperatures for the last deglaciation

Nils Weitzel[1], Heather Andres[2], Jean-Philippe Baudouin[1], Marie-Luise Kapsch[3], Uwe Mikolajewicz[3], Lukas Jonkers[4], Oliver Bothe[a], Elisa Ziegler[1,5], Thomas Kleinen[3], André Paul[4], and Kira Rehfeld[1,5]

[1]Department of Geosciences, University of Tübingen, Tübingen, Germany
[2]Northwest Atlantic Fisheries Centre, Fisheries and Oceans Canada, St. John's, Newfoundland, Canada
[3]Max Planck Institute for Meteorology, Hamburg, Germany
[4]MARUM Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany
[5]Department of Physics, University of Tübingen, Tübingen, Germany
[a]formerly at: Institute of Coastal Systems – Analysis and Modelling, Helmholtz-Zentrum Hereon, Geesthacht, Germany

Correspondence: Nils Weitzel (nils.weitzel@uni-tuebingen.de)

Received: 11 May 2023 – Discussion started: 24 May 2023
Revised: 13 February 2024 – Accepted: 22 February 2024 – Published: 8 April 2024

Abstract. An increasing number of climate model simulations is becoming available for the transition from the Last Glacial Maximum to the Holocene. Assessing the simulations' reliability requires benchmarking against environmental proxy records. To date, no established method exists to compare these two data sources in space and time over a period with changing background conditions. Here, we develop a new algorithm to rank simulations according to their deviation from reconstructed magnitudes and temporal patterns of orbital and millennial-scale temperature variations. The use of proxy forward modeling allows for accounting for non-climatic processes that affect the temperature reconstructions. It further avoids the need to reconstruct gridded fields or regional mean temperature time series from sparse and uncertain proxy data.

First, we test the reliability and robustness of our algorithm in idealized experiments with prescribed deglacial temperature histories. We quantify the influence of limited temporal resolution, chronological uncertainties, and non-climatic processes by constructing noisy pseudo-proxies. While model–data comparison results become less reliable with increasing uncertainties, we find that the algorithm discriminates well between simulations under realistic non-climatic noise levels. To obtain reliable and robust rankings, we advise spatial averaging of the results for individual proxy records.

Second, we demonstrate our method by quantifying the deviations between an ensemble of transient deglacial simulations and a global compilation of sea surface temperature reconstructions. The ranking of the simulations differs substantially between the considered regions and timescales, which suggests that optimizing for agreement with the temporal patterns of a small set of proxies might be insufficient for capturing the spatial structure of the deglacial temperature variability. We attribute the diversity in the rankings to more regionally confined temperature variations in reconstructions than in simulations, which could be the result of uncertainties in boundary conditions, shortcomings in models, or regionally varying characteristics of reconstructions such as recording seasons and depths. Future work towards disentangling these potential reasons can leverage the flexible design of our algorithm and its demonstrated ability to identify varying levels of model–data agreement. Additionally, the algorithm can be applied to variables like oxygen isotopes and climate transitions such as the penultimate deglaciation and the last glacial inception.

1 Introduction

Major boundary condition changes make the transition from the Last Glacial Maximum (LGM, ∼ 21 ka, where ka stands for "kilo-annum", i.e., thousands of years ago) to the current warm period, the Holocene interglacial (starting at ∼ 11.65 ka), an important period for understanding past global warming episodes and a valuable period for testing climate models. This transition, called the last deglaciation, is the most recent period with natural radiative forcing variations of comparable magnitude to projected anthropogenic emissions. During the deglaciation, the configuration of orbital parameters changed, resulting in a minimum in Northern Hemisphere summer insolation around 24 ka and a maximum around 11 ka (Berger, 1978). The CO_2 concentration increased from ∼ 185 to ∼ 280 ppm (Köhler et al., 2017), and sea level rose by ∼ 130 m (Lambeck et al., 2014) because large ice sheets over North America (the Laurentide and Cordilleran ice sheets) and Europe (the Fennoscandian and British ice sheets) retreated entirely (Batchelor et al., 2019).

In recent years, the last deglaciation has been simulated with an increasing number of climate models that apply transiently changing boundary conditions (Ivanovic et al., 2016). Proxy-based temperature reconstructions suggest that (near-)surface temperatures increased at most places since the LGM (Cleator et al., 2020; Paul et al., 2021) and by 3.6–6.5 K in the global mean (Annan et al., 2022; Tierney et al., 2020). Most climate models simulate LGM global mean surface air temperature (GMSAT) anomalies in this range (Kageyama et al., 2021). However, proxy evidence suggests that considerable regional differences exist in the magnitude and temporal pattern of the deglacial temperature changes (Clark et al., 2012). So far, it has not been quantitatively assessed whether climate models can not only reproduce the reconstructed GMSAT changes but also the spatial fingerprint of the temperature evolution when forced with appropriate boundary conditions. This assessment is challenging because it relies on sparse and indirect observations of past climate and uncertain boundary conditions (Ivanovic et al., 2016).

Previous model–data comparison efforts involving global databases of proxy records focused on the Common Era (e.g., PAGES 2k Consortium, 2019; PAGES 2k-PMIP3 group, 2015) or on time slices such as the LGM and the mid-Holocene (e.g., Hargreaves et al., 2013; Harrison et al., 2014). They quantify either differences between two distinct states (e.g., LGM vs. pre-industrial) or fluctuations during a stationary climate state (e.g., magnitude of temperature variability). So far, transient simulations of the last deglaciation have only been compared against a small number of selected proxy records or large-scale mean reconstructions (e.g., Dallmeyer et al., 2022; He et al., 2021; Liu et al., 2009; Menviel et al., 2011). Here, we develop a model–data comparison algorithm that compares last deglaciation simulations with temperature reconstructions in space and time. In particular, our algorithm allows to quantitatively assess the following four questions.

1. Is the magnitude of simulated deglacial warming in agreement with reconstructions?
2. Is the temporal pattern of the glacial-to-interglacial (called orbital-scale) warming trend accurately simulated?
3. Are the magnitudes of simulated millennial-scale variations modulating the warming trend similar to reconstructions?
4. How much does the temporal pattern of simulated millennial-scale variations deviate from reconstructions?

We analyze the four components of the deglacial temperature evolution associated with these questions separately because the robustness of their reconstruction varies, and they are potentially controlled by different mechanisms and uncertain boundary conditions. In the following, we call these four components the "orbital magnitude" (magnitude of orbital-scale temperature variations), "orbital pattern" (temporal pattern of orbital-scale variations), "millennial magnitude" (magnitude of millennial-scale variations), and "millennial pattern" (temporal pattern of millennial-scale variations). Note that throughout this paper we use the term "orbital" to describe climate variations occurring on similar timescales (∼ 6 kyr and longer) to variations in the Earth's orbital configuration, although changes in greenhouse gas (GHG) concentrations and ice sheets are the main contributors to radiative forcing on these timescales during the deglaciation.

To illustrate our model–data comparison algorithm, we use a global database of sea surface temperature (SST) reconstructions and an ensemble of last deglaciation simulations (Sect. 2). SSTs are reconstructed from geochemical indices and species assemblages extracted from marine sediment cores. Both reflect the climate state at the time of deposition (Jonkers et al., 2020). However, the reconstructed temperatures are also influenced by non-climatic processes during the recording of the temperature signal, the archival of the sensors in the sediment, and the measurement of the proxy. These include imperfect calibrations to temperature, biases from confounding environmental variables, deviations from mean annual SST through seasonal and habitat depth preferences, temporal smoothing by bioturbation, noise from using a small number of short-living replicates, measurement errors, and chronological uncertainties (Dolman and Laepple, 2018; Jonkers and Kučera, 2017, 2019; MARGO Project Members, 2009; Osman et al., 2021). Here, and in the following, we refer to sensors as the organisms recording the temperature signal (e.g., planktonic foraminifera) and to proxies as the measured temperature-sensitive quantities (e.g., Mg / Ca ratios, species compositions).

The influence of non-climatic processes creates a challenge for model–data comparison: whether a simulation produces a more realistic climate evolution than others is not necessarily the same as finding the simulation that minimizes the difference to a set of reconstructions, since reconstructions are an imperfect representation of the actual climate evolution. To obtain a representation of the simulated climate that is comparably disturbed by non-climatic processes as reconstructed SSTs, we use proxy system models (PSMs). PSMs are mathematical descriptions of the processes involved in the recording, archiving, and measurement of the response of an environmental proxy to the climate (Evans et al., 2013). PSMs are applied to climate simulation output to create forward-modeled proxy time series which mimic the properties of real proxies. A comparison of these forward-modeled proxy time series against proxy-based reconstructions facilitates a more consistent comparison under the assumption that real and modeled proxies are subject to comparable modifications (Bühler et al., 2021; Dee et al., 2017). In particular, using PSMs can account for biases in reconstructions of timescale-dependent climate variability from proxy data and determine the significance of reconstructed temperature patterns in the presence of non-climatic noise (e.g., Jonkers and Kučera, 2017; Laepple and Huybers, 2014). PSMs can be employed in a forward or inverse manner. In forward approaches, a PSM is applied to simulation output. Inverse approaches infer gridded fields or time series with regular time steps by inverting PSMs with Bayesian statistics (Tingley et al., 2012). We choose the forward approach because it follows the natural process chain from the climate signal to the sample measurements (Evans et al., 2013) and it avoids the estimation of spatio-temporal temperature correlation structures, which are hard to estimate from sparse proxy data (Tingley et al., 2012).

A second challenge in model–data comparison is to separate mismatches between simulations and reconstructions due to uncertain boundary and initial conditions, poorly constrained model parameters, and imperfect or missing representations of relevant processes by climate models (Braconnot et al., 2012). This challenge could in principle be assessed through large model ensembles, but computational resources are insufficient to produce them. Therefore, we focus here on incorporating methods to account for uncertainties from imperfect reconstructions.

The goals of this paper are threefold. First, we motivate and present our proposed model–data comparison algorithm (Sect. 3.1). Second, we test our algorithm with pseudo-proxy experiments (PPEs; von Storch et al., 2004), in which the deglacial climate evolution is prescribed by a reference simulation (Sects. 3.3, 4.1). These experiments help us to understand the characteristics of our algorithm and to assess its reliability and robustness under limited temporal resolutions, chronological uncertainties, and non-climatic modulations of the proxy records. To our knowledge, model–data comparison algorithms have never been systematically tested with PPEs. Third, we demonstrate our method by quantifying the deviations between forward-modeled proxy time series derived from 10 last deglaciation simulations and the global compilation of SST reconstructions (Sect. 4.2). Finally, we discuss implications and limitations of our results and outline future work (Sect. 5).

2 Data

2.1 Transient simulations

We use 10 previously published simulations from three climate models which all simulate the period 22 to 6 ka (Fig. 1, Table 1). Six simulations employ MPI-ESM-CR (Kapsch et al., 2022; Kleinen et al., 2023a, b). In these simulations, GHG concentrations and orbital parameters are updated transiently. Ice sheet topographies are changed according to the GLAC-1D or ICE-6G reconstructions (see Table 1). Meltwater from ice sheets is either transported into the ocean using dynamic river routing (Riddick et al., 2018), distributed uniformly over all grid cells, or removed from the system (see Table 1). MPI_Glac1D_PTK uses a parameter configuration that leads to a smaller LGM-to-Holocene temperature difference than in the other MPI-ESM simulations. Furthermore, atmospheric parameters in the "P3" simulations are slightly different from those in the "P2" simulations to correct a pre-industrial cold bias (Kapsch et al., 2022).

We further include three CCSM3 simulations from the TraCE-21ka project (Liu et al., 2009). In TraCE-ALL, orbital parameters, GHG concentrations, ICE-5G ice sheet topographies, and manually prescribed meltwater fluxes are adapted transiently. In TraCE-GHG, all boundary conditions except for GHG concentrations are fixed at the 22 ka state of TraCE-ALL. Similarly, only orbital parameters are changed in TraCE-ORB. Finally, we use the ALL-5G simulation from the QUEST FAMOUS last glacial cycle ensemble (Smith and Gregory, 2012). Orbital parameters, GHG concentrations, and Northern Hemisphere ICE-5G ice sheet topographies are updated transiently. In contrast to the other simulations, the Antarctic ice sheet topography and land–sea mask are fixed to pre-industrial values, and the transient boundary conditions are applied with an acceleration factor of 10. No meltwater fluxes are applied in FAMOUS.

In the following, we denote the six MPI-ESM simulations and TraCE-ALL as the "main set of simulations" and TraCE-ORB, TraCE-GHG, and FAMOUS as "sensitivity experiments". The latter three simulations either change only one boundary condition transiently or employ boundary conditions faster than they occurred in reality. Therefore, we do not expect them to cover all changes in the climate system with the same degree of realism as the other seven simulations. More information on the simulations is provided in the Supplement (Sect. S2).

The simulation ensemble has a large spread in the four components of the deglacial temperature evolution described

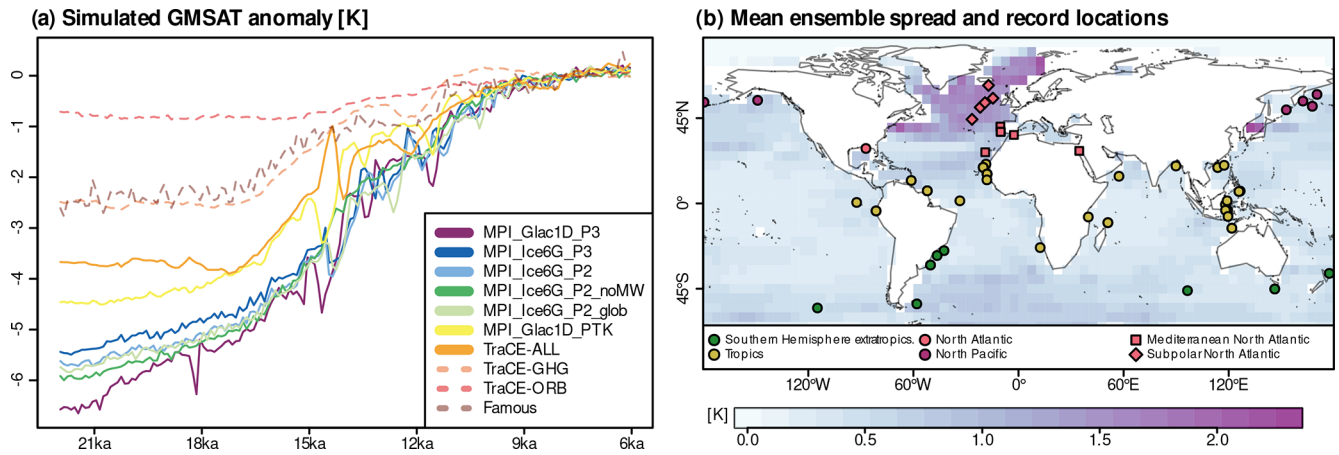

Figure 1. **(a)** GMSAT anomalies of the transient simulation ensemble members. Anomalies were computed with respect to the mean in the window 9 to 6 ka. **(b)** Locations of SST reconstruction records employed in the model–data comparison (dots) and simulation ensemble spread as measured by the standard deviation at each location and time step, averaged over all time steps (colors in the background). The colors of the dots indicate the regions considered in Sect. 4.2 and the shape of the dots in the North Atlantic mark the records used for the separation into Mediterranean and subpolar North Atlantic in Sect. 4.2 and Fig. 9. Ocean grid cells are selected based on the ICE-6G history (Peltier et al., 2015).

Table 1. Properties of the 10 transient simulations of the last deglaciation included in the simulation ensemble: the name used throughout the paper, the employed climate model, whether orbital and GHG forcings were varied transiently or fixed at LGM values, the employed ice sheet reconstructions, how meltwater fluxes were applied (local input through dynamical river routing, local input according to a manually defined scheme, distributed equally across all grid cells, or no meltwater input), and the main reference of the simulation.

Name	Model	Orbital	GHG	Ice sheets	Meltwater	Reference
MPI_Glac1D_P3	MPI-ESM-CR	yes	yes	GLAC-1D	river routing	Kapsch et al. (2022)
MPI_Ice6G_P3	MPI-ESM-CR	yes	yes	ICE-6G	river routing	Kapsch et al. (2022)
MPI_Ice6G_P2	MPI-ESM-CR	yes	yes	ICE-6G	river routing	Kapsch et al. (2022)
MPI_Ice6G_P2_noMWF	MPI-ESM-CR	yes	yes	ICE-6G	none	Kapsch et al. (2022)
MPI_Ice6G_P2_glob	MPI-ESM-CR	yes	yes	ICE-6G	global	Kapsch et al. (2022)
MPI_Glac1D_PTK	MPI-ESM-CR	yes	yes	GLAC-1D	river routing	Kleinen et al. (2023a)
TraCE-ALL	CCSM3	yes	yes	ICE-5G	local (manual)	Liu et al. (2009)
TraCE-GHG	CCSM3	no	yes	fixed at LGM	none	Liu et al. (2009)
TraCE-ORB	CCSM3	yes	no	fixed at LGM	none	Liu et al. (2009)
FAMOUS	FAMOUS	yes	yes	ICE-5G	none	Smith and Gregory (2012)

in Sect. 1 (Fig. 1). In the main set of simulations, the deglacial GMSAT increase is between ∼ 4 K in TraCE-ALL and ∼ 6.5 K in MPI_Glac1D_P3. With ∼ 1 K in TraCE-ORB and ∼ 3 K in TRACE-GHG and FAMOUS, the deglacial warming is lower in the three sensitivity experiments. Deglacial warming starts later in TraCE-ALL than in the MPI-ESM simulations, and the warming trend is smoother in MPI-ESM than in TraCE-ALL. Two different aspects of meltwater injection appear to play an important role in the GMSAT histories of these runs: the method of application and the progression through time. Simulations without meltwater fluxes feature weak millennial-scale fluctuations (e.g., MPI_Ice6G_P2_noMWF), and simulations with locally applied meltwater fluxes (e.g., MPI_Ice6G_P2) generate stronger GMSAT fluctuations than the simulation with global injection (MPI_Ice6G_P2_glob). Differing meltwater histories lead to an abrupt warming at ∼ 14.5 ka in TraCE-ALL but cooling events in the MPI-ESM experiments with meltwater input.

2.2 Sea surface temperature reconstructions

We use temperature reconstructions from the PalMod 130k marine paleoclimate data synthesis v1.1.1 (Jonkers et al., 2023), which is a compilation of published proxy records derived from marine sediment cores. V1.1.1 is an update from Jonkers et al. (2020) with 252 published (near-)surface temperature time series covering various periods of the last glacial cycle. As described in Jonkers et al. (2020), age models are harmonized using the Bayesian age modeling algorithm BACON (Blaauw and Christen, 2011). For each sedi-

ment core, 1000 iterations of the age–depth model are saved in the database to quantify chronological uncertainties. The database combines temperature reconstructions from multiple proxies which are taken unchanged from the original publications. For some proxy records, reconstructions from different original publications are included in the database. We retain all records from the same sediment cores if they are based on different proxies. We average reconstructions originating from the same sediment core and proxy if all sample depths coincide. If the depths differ, we select the time series covering the longest period during the deglaciation. Reconstructions from the same proxy data but calibrated for different seasons are averaged to obtain pseudo-annual temperatures. More details on the preprocessing of the proxy records are provided in the Supplement (Sect. S3).

We select all (near-)surface temperature samples in the interval 22–6 ka from the database. Most of these records reflect surface or mixed-layer temperatures (Kucera et al., 2005; Rebotim et al., 2017; Tierney and Tingley, 2018). While the used sensors occupy a range of depths, we denote all samples as sea surface temperature (SST) reconstructions in the following. To compute robust statistics, we use only time series with at least 10 samples, which cover more than 8 kyr and have a mean temporal resolution of at least 1 kyr. 74 temperature records from 50 unique sediment cores satisfy these conditions (Fig. 1b, Table 2). Most of them are located on continental margins with the biggest clusters located in the North Atlantic and the Indo-Pacific Warm Pool. A total of 38 temperature records are reconstructed from Mg / Ca, 17 from U^{k}_{37}, 17 from planktonic foraminifera assemblages, 1 from TEX_{86}, and 1 from diatom assemblages. Unlike some recent studies focusing on either assemblage-based temperature reconstructions (e.g., Paul et al., 2021) or geochemical proxies (e.g., Osman et al., 2021), we employ a multi-proxy approach using the calibrations proposed by the original authors for assemblages and geochemical proxies, respectively. We make this choice because the number of records in the database is too small to focus on specific proxy types, and proxy types tend to be regionally clustered, which makes a systematic assessment of differences between them unfeasible within our study design. For more discussion on the differences between proxy types, see Paul et al. (2021), and the references therein.

3 Methods

This section first presents our model–data comparison algorithm (Sect. 3.1). The algorithm employs a simple PSM with two parameters that we estimate in Sect. 3.2. Section 3.3 describes the PPEs for assessing the reliability and robustness of our algorithm.

3.1 Model–data comparison algorithm

As visualized in Fig. 2, our model–data comparison algorithm consists of four main steps which we present in the following. An enhanced description of the algorithm with computational details is provided in the Supplement (Sect. S4).

3.1.1 Compute forward-modeled proxy time series from simulation output.

To compare simulations and reconstructions, we have to bridge the gaps between the two types of data in terms of spatio-temporal coverage and non-climatic influences on the proxy measurements. This is done in a forward approach, in which a PSM is applied to simulation output. The PSM output, which we call "forward-modeled proxy time series", is compared to the measured proxies. We perform this comparison in temperature units instead of measured proxy units because it allows for averaging deviations from different proxies and no established forward calibrations exist for assemblage-based reconstructions. Our PSM takes simulated 3D (long × lat × time) mean annual SST fields (T_{Sim}) as input and modifies them to resemble a reconstructed SST record (T_{FM}, where FM stands for forward-modeled). The PSM consists of three steps: spatial interpolation to the proxy location (P_{space}), temporal downsampling to the proxy time axis (P_{time}), and a Gaussian additive noise process ε with a specified signal-to-noise ratio (SNR) and temporal autocorrelation structure. The noise process summarizes the effects of inherent uncertainties of the SST reconstructions (see Sect. 1) and uncertainties in the formulation of the PSM. Thus, the PSM is defined as follows:

$$T_{\mathrm{FM}} = P_{\mathrm{time}}\left(P_{\mathrm{space}}(T_{\mathrm{Sim}})\right) + \varepsilon. \quad (1)$$

Accounting for reconstruction uncertainties, which is done in the temporal downsampling and additive noise components of the PSM, requires a probabilistic comparison framework. We implement such a framework using a Monte Carlo approach, which propagates uncertainties through the algorithm.

3.1.2 Decompose time series into magnitudes and temporal patterns of timescale-dependent variations.

We decompose each temperature time series into four components, orbital magnitude, orbital temporal pattern, millennial magnitude, and millennial temporal pattern, each of which is designed to assess one of the four questions posed in Sect. 1. For the timescale decomposition, we use Gaussian smoothers (Fig. 2, second row; see Figs. S2–S9 in the Supplement for further examples of timescale decompositions) as they are a robust method for the analysis of irregularly spaced time series in the time and frequency domain (Rehfeld et al., 2011). For each timescale, we define the magnitude of variations as the standard deviations of the filtered

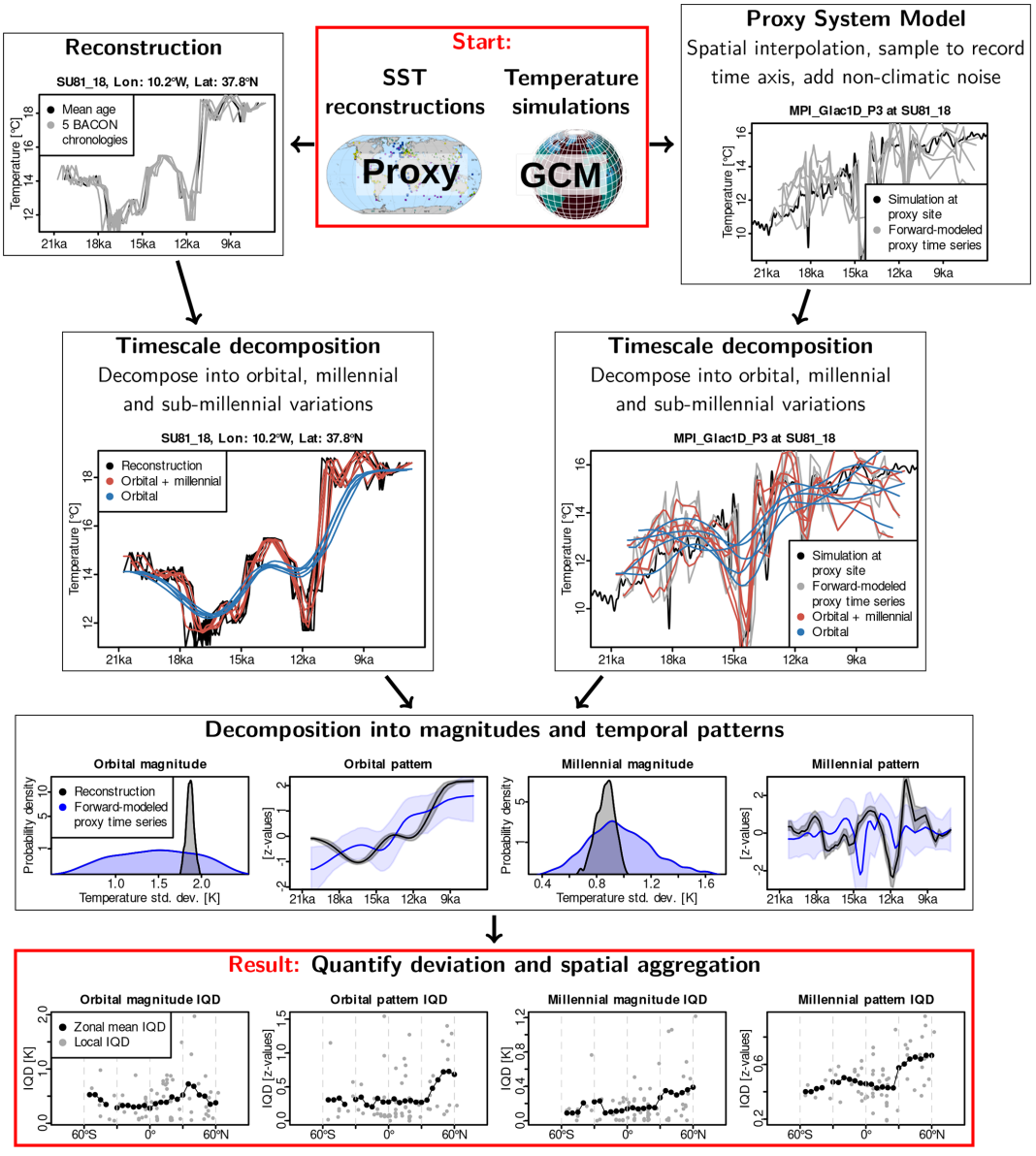

Figure 2. Flow chart describing the algorithm presented in this study (see Sect. 3 for details). We start at the top with two sets of data, reconstructed and simulated SSTs. Age uncertainties of the proxy records are quantified using multiple iterations from the age–depth model (top row, left). We apply a proxy system model (PSM) to the simulated SST fields to first obtain simulated time series interpolated to the proxy locations and then Monte Carlo realizations of forward-modeled proxy time series (top row, right). For each Monte Carlo realization, a timescale decomposition is performed to separate orbital- and millennial-scale variations using Gaussian smoothers (second row, left for reconstructions, right for forward-modeled proxy time series). Differences between the Monte Carlo realizations of reconstructions are due to chronological uncertainties, whereas differences in the Monte Carlo realizations of forward-modeled proxy time series result from the stochastic PSM. The orbital- and millennial-scale time series are decomposed into the magnitude and temporal pattern of the variations. This leads to probability distributions for reconstructions and forward-modeled proxy time series (third row). Finally, the integrated quadratic distance (IQD) between the probability distributions of reconstructions and forward-modeled proxy time series is computed for each of the four components (dots in the bottom row), and IQDs are averaged spatially. As an exemplary partition into regions, we show zonal mean IQDs in the bottom row for all latitudinal bands containing at least five proxy records (see Sect. 3.1 for a definition of the zonal mean averaging procedure).

Table 2. Information on the 74 proxy records selected for the deglacial model–data comparison.

ID	Core name	Long [°E]	Lat [°N]	Ocean basin	Proxy	Reference
1	108_658C	−18.6	20.7	Atlantic	U^k_{37}	Zhao et al. (1995)
2	323_U1340A	−179.5	53.4	Pacific	U^k_{37}	Schlung et al. (2013)
3	BOFS31_1K	−20.2	19.0	Atlantic	Plankt. foram. assembl.	Chapman et al. (1996)
4	BOFS31_1K	−20.2	19.0	Atlantic	U^k_{37}	Zhao et al. (1995)
5	BOFS31_1K	−20.2	19.0	Atlantic	MgCa (G. bulloides)	Elderfield and Ganssen (2000)
6	BOFS31_1K	−20.2	19.0	Atlantic	MgCa (G. inflata)	Elderfield and Ganssen (2000)
7	BOFS31_1K	−20.2	19.0	Atlantic	MgCa (G. ruber pink)	Elderfield and Ganssen (2000)
8	BOFS31_1K	−20.2	19.0	Atlantic	MgCa (N. incompta)	Elderfield and Ganssen (2000)
9	BOFS5K	−21.9	50.7	Atlantic	Plankt. foram. assembl.	Maslin et al. (1995) Vogelsang et al. (2001)
10	GeoB12615_4	39.8	−7.1	Indian	MgCa (G. ruber white)	Romahn et al. (2014)
11	GeoB16224_1	−52.1	6.7	Atlantic	MgCa (G. ruber white)	Crivellari et al. (2019)
12	GeoB16224_1	−52.1	6.7	Atlantic	Plankt. foram. assembl.	Crivellari et al. (2019)
13	GeoB16224_1	−52.1	6.7	Atlantic	U^k_{37}	Crivellari et al. (2019)
14	GeoB16224_1	−52.1	6.7	Atlantic	TEX86	Crivellari et al. (2019)
15	GeoB16602	113.7	19.0	Pacific	U^k_{37}	Huang et al. (2018)
16	GeoB16602	113.7	19.0	Pacific	MgCa (G. ruber white)	Cheng et al. (2018)
17	GeoB1711_4	12.4	−23.3	Atlantic	U^k_{37}	Kirst et al. (1999)
18	GeoB5844_2	34.7	27.7	Indian	U^k_{37}	Arz et al. (2003)
19	GeoB6211_2	−50.2	−32.5	Atlantic	MgCa (G. inflata)	Chiessi et al. (2008)
20	GeoB6211_2	−50.2	−32.5	Atlantic	MgCa (G. ruber white)	Chiessi et al. (2014, 2015)
21	GeoB9508_5	-17.9	15.5	Atlantic	U^k_{37}	Niedermeyer et al. (2009)
22	GeoB9508_5	−17.9	15.5	Atlantic	MgCa (G. ruber pink)	Zarriess et al. (2011)
23	GeoB9508_5	−17.9	15.5	Atlantic	MgCa (G. inflata)	Bouimetarhan et al. (2013)
24	GeoB9508_5	−17.9	15.5	Atlantic	MgCa (G. bulloides)	Bouimetarhan et al. (2013)
25	GeoB9526_5	−18.1	12.4	Atlantic	MgCa (G. ruber pink)	Zarriess et al. (2011)
26	GIK15612_2	−26.5	44.4	Atlantic	Plankt. foram. assembl.	Kiefer (1998)
27	GIK15637_1	−19.0	27.0	Atlantic	Plankt. foram. assembl.	Kiefer (1998)
28	GIK17286_1	89.9	19.74	Indian	U^k_{37}	Lauterbach et al. (2020)
29	GIK17940_2	117.4	20.1	Pacific	U^k_{37}	Pelejero et al. (1999)
30	GiK18515_3	119.4	−3.6	Pacific	MgCa (G. ruber white)	Schröder et al. (2016)
31	GIK18519_2	118.1	−0.6	Pacific	MgCa (G. ruber white)	Schröder et al. (2018)
32	GIK18522_3	119.1	1.4	Pacific	MgCa (G. ruber white)	Schröder et al. (2018)
33	GIK18526_3	118.2	−3.6	Pacific	MgCa (G. ruber white)	Schröder et al. (2018)
34	GIK18540_3	119.6	−6.9	Pacific	MgCa (G. ruber white)	Schröder et al. (2018)
35	GIK23415_9	−19.1	53.1	Atlantic	Plankt. foram. assembl.	Weinelt et al. (2003)
36	GL1090	−42.5	−24.9	Atlantic	MgCa (G. ruber white)	Santos et al. (2017)
37	H214	177.4	−36.9	Pacific	Plankt. foram. assembl.	Samson et al. (2005)
38	JR244_GC528	−58.0	−53.0	Atlantic	U^k_{37}	Roberts et al. (2016, 2017)
39	KNR159_5_36	−46.5	−27.5	Atlantic	MgCa (G. ruber white)	Carlson et al. (2008)
40	LV29_114_3	152.9	49.4	Pacific	MgCa (N. pachyderma)	Riethdorf et al. (2013)
41	M35003_4	−61.2	12.1	Atlantic	U^k_{37}	Rühlemann et al. (1999)
42	M35003_4	−61.2	12.1	Atlantic	Plankt. foram. assembl.	Hüls and Zahn (2000)
43	M77_2_059_1	−81.3	−4.0	Pacific	MgCa (G. ruber white)	Nürnberg et al. (2015)
44	M77_2_059_1	−81.3	−4.0	Pacific	MgCa (N. dutertrei)	Nürnberg et al. (2015)
45	M77_2_059_1	−81.3	−4.0	Pacific	U^k_{37}	Nürnberg et al. (2015)
46	MD01_2378	121.8	−13.1	Indian	MgCa (P. obliquiloculata)	Xu et al. (2006, 2008)
47	MD01_2378	121.8	−13.1	Indian	MgCa (G. ruber)	Xu et al. (2006, 2008)
48	MD01_2416	167.7	51.3	Pacific	Plankt. foram. assembl.	Gebhardt et al. (2008)
49	MD01_2416	167.7	51.3	Pacific	MgCa (N. pachyderma)	Gray et al. (2018)
50	MD02_2489	−148.9	54.4	Pacific	Plankt. foram. assembl.	Gebhardt et al. (2008)
51	MD02_2575	−87.1	29.0	Atlantic	MgCa (G. ruber white)	Ziegler et al. (2008)
52	MD06_3067	126.5	6.5	Pacific	MgCa (G. ruber)	Bolliet et al. (2011)
53	MD06_3067	126.5	6.5	Pacific	MgCa (P. obliquiloculata)	Bolliet et al. (2011)

Table 2. Continued.

ID	Core name	Long [°E]	Lat [°N]	Ocean basin	Proxy	Reference
54	MD88_770	96.5	−46.0	Indian	Plankt. foram. assembl.	Labeyrie et al. (1996)
55	MD95_2039	−10.3	40.6	Atlantic	Plankt. foram. assembl.	Salgueiro et al. (2014)
56	MD95_2042	−10.2	37.8	Atlantic	U^k_{37}	Pailler and Bard (2002)
57	MD95_2043	−2.6	36.1	Atlantic	U^k_{37}	Cacho et al. (1999)
58	MD98_2181	125.8	6.3	Pacific	MgCa (G. ruber)	Stott et al. (2002, 2007)
59	MD98_2181	125.8	6.3	Pacific	MgCa (T. sacculifer)	Stott et al. (2002)
60	NA87_22	−14.6	55.5	Atlantic	Plankt. foram. assembl.	Vogelsang et al. (2001)
61	PS75_056_1	−114.8	−55.2	Pacific	diatom assemblages	Benz et al. (2016)
62	RAPiD_15_4P	−17.1	62.3	Atlantic	MgCa (N. pachyderma)	Thornalley et al. (2011)
63	RS147_GC07	146.3	−45.2	Indian	U^k_{37}	Sikes et al. (2009)
64	RS147_GC07	146.3	−45.2	Indian	Plankt. foram. assembl.	Sikes et al. (2009)
65	SO201_2_12KL	162.4	54.0	Pacific	MgCa (N. pachyderma)	Riethdorf et al. (2013)
66	SO201_2_85	170.4	57.5	Pacific	MgCa (N. pachyderma)	Riethdorf et al. (2013)
67	SO42_74KL	57.3	14.3	Indian	Plankt. foram. assembl.	Schulz (1995)
68	SU81_18	−10.2	37.8	Atlantic	U^k_{37}	Bard et al. (2000)
69	SU81_18	−10.2	37.8	Atlantic	Plankt. foram. assembl.	Vogelsang et al. (2001)
70	TR163_22	−92.4	0.5	Pacific	MgCa (G. ruber)	Lea et al. (2006)
71	V25_59	−33.5	1.4	Atlantic	Plankt. foram. assembl.	Waelbroeck et al. (1998)
72	WIND_28K	51.0	−10.2	Indian	MgCa (G. ruber white)	Johnstone et al. (2014) Kiefer et al. (2006)
73	WIND_28K	51.0	−10.2	Indian	MgCa (T. sacculifer)	Johnstone et al. (2014)
74	WIND_28K	51.0	−10.2	Indian	MgCa (N. dutertrei)	Johnstone et al. (2014) Kiefer et al. (2006)

time series and the temporal pattern as the normalized, i.e., centered and standardized, time series (Fig. 2, third row).

Magnitude components quantify the strength of timescale-dependent variations, independent of their specific timing. Therefore, they are valuable for assessing the strength of the response to forcing, of spontaneous fluctuations, and of variations forced by time-uncertain boundary conditions. In contrast, temporal pattern components assess the direction, timing, and succession of timescale-dependent variations. They are particularly meaningful if variations are externally forced and if there are sufficiently tight constraints on the boundary condition reconstructions such that models can be expected to reproduce the timing of the observed pattern of variations. Since orbital- and millennial-scale variations are likely driven by different forcings and internal processes, we separate between deviations on these two timescales. We assess the deviations between forward-modeled proxy time series and proxy records for each component separately because computing a single score for the deviation between simulations and reconstructions is prone to conceal sources of discrepancies. For example, a simulation could reproduce the reconstructed spatio-temporal temperature pattern accurately but receive a poor score due to an underestimation of the LGM-to-Holocene temperature change.

3.1.3 Quantify deviations between reconstructions and forward-modeled proxy time series for individual proxy records.

The decompositions in step 2 result in probability distributions of forward-modeled proxy time series and the corresponding reconstructed SST records because we account for chronological uncertainties and include a noise process in the PSM. We quantify the deviations between these probability distributions with a distance function that takes into account the full, potentially multivariate probability distributions and not just summary statistics like the mean or standard deviation. We choose the integrated quadratic distance (IQD), which is a proper divergence function that has desirable mathematical properties for model selection as it penalizes overly confident or conservative uncertainty estimates compared to the unknown true uncertainties (Thorarinsdottir et al., 2013). Applying the distance function to the respective probability distributions results in a single number for the deviation between forward-modeled proxy time series and reconstructions for each of the four components in which we decompose the time series in step 2.

For two probability distributions $\mathbb{P}$ and $\mathbb{Q}$, the IQD takes positive values ($\mathrm{IQD}(\mathbb{P},\mathbb{Q}) \geq 0$). It is only zero when $\mathbb{P}$ and $\mathbb{Q}$ are equal ($\mathrm{IQD}(\mathbb{P},\mathbb{P}) = 0$). Smaller IQD values imply a smaller deviation and thus a better agreement of forward-modeled proxy time series and reconstructions. In the absence of age and proxy uncertainties, the IQD reduces to the

mean absolute difference between numbers (magnitudes) or time series (patterns). The IQD can be applied to quantities of arbitrary units. In our case, the units are temperature [K] for the comparison of magnitudes, and standard deviations [z] for patterns.

3.1.4 Average deviations in space.

Deviations between forward-modeled proxy time series and reconstructions can depend strongly on the unknown manifestation of non-climatic influences in the measured proxies and uncertainties in the PSM structure. Assuming that most non-climatic processes and PSM uncertainties are uncorrelated between proxy records, the influence of these processes can be reduced by spatially averaging deviations computed for individual proxy records. Computing averages in this last step instead of averaging temperature time series in the beginning avoids interpolating proxy records with irregular time axes to a common resolution. We analyze IQDs averaged on four spatial scales: locally, regionally (see color-coding of dots in Fig. 1b for the assignment of proxy records to the regions considered in this study), zonally, and globally. For local IQDs, we treat each proxy record individually, i.e., without averaging proxy records from the same core or nearby locations. Zonal IQDs are obtained by averaging over proxy records within overlapping bands of 20° width that move in 5° steps (Fig. 2, bottom row). We only consider latitudinal bands containing at least five proxy records to only incorporate spatial averages where we can assume that a substantial amount of non-climatic influences is averaged out.

3.2 Estimation of proxy system model parameters

The PSM described in Sect. 3.1 requires a SNR parameter quantifying the ratio between climatic and non-climatic variations and the specification of a temporal autocorrelation structure of the additive Gaussian noise process. Previous studies only estimated SNRs and autocorrelations for a subset of our proxy types (U^{k}_{37}, Mg / Ca) on sub-orbital timescales (Laepple and Huybers, 2014; Reschke et al., 2019). Therefore, we estimate the PSM parameters using the SST reconstruction database (see Sect. 2.2).

To obtain these estimates, we decompose the SST records into a similar structure as Eq. (1), i.e., the sum of a local mean SST signal $P_{\text{space}}(T)$ and a realization of a Gaussian noise process ε, which aggregates all deviations from the local mean SST signal. The decomposition starts by constructing clusters of SST records centered around each of the 74 SST records selected from the database. The clusters contain the records within a radius of $l \in \{100, 200, \ldots, 1000\}$ km around the central record (see Fig. 3 for an example cluster with $n = 3$ records centered around record SO201_2_12KL). For each cluster, we compute a local mean signal by averaging over the records in the cluster (red line in Fig. 3a). More specifically, we interpolate nearby records to a regular temporal resolution of 100 years, center the records, and average over the resulting time series. We use the mean age model of each record and not the age ensemble members since we account for chronological uncertainties at a different step of the PSM. Using the age ensembles instead of the mean ages strongly reduces the estimated SNR and likely biases it low (not shown). Note that we average records of different temporal resolutions, which tends to underestimate high-frequency contributions to ε. However, all records have at least a millennial resolution such that the relevant millennial and orbital timescales should be less affected by the interpolation and subsequent averaging.

For the record in the center of the cluster, we compute the residual from the local mean signal (green line in Fig. 3b), which is treated as a realization of the Gaussian noise process (ε in Eq. 1). We compute the variance ratio between the local mean signal and the residual which provides an estimate of the SNR. Due to the short time series length, the structure of the temporal autocorrelation cannot be determined from the residuals. We choose to describe ε as an autoregressive process of order one (AR1) because it is determined by only two parameters and as a compromise between a white noise process without temporal autocorrelation and power law processes with long-range autocorrelations. This AR1 process is specified by the SNR and a decorrelation length, which we estimate from the residual. We iterate this process for all 74 records if the clusters around the respective records contain at least a specified number of records. We then take the medians of the SNRs and the decorrelation lengths in all clusters to reduce the noise in the parameter estimates, which results from the predominantly small cluster sizes (most clusters contain less than five records). As the estimates can be sensitive to the construction of the clusters, we apply this procedure for cluster radii of $l \in \{100, 200, \ldots, 1000\}$ km and for the minimum required number of records in a cluster of $n \in \{2, 3\}$.

The median SNR over all sensitivity experiments is 1.6 ± 0.3 (1σ) and the median decorrelation length is 1289 ± 212 years. When we decompose the SST variability of each proxy record into a signal and a noise component according to $\text{SNR} = 1.6$, the mean noise level across all records is 0.9 ± 0.6 K. This estimate is consistent with an estimate of 0.6–1.3 K by Tierney et al. (2020) in a data assimilation framework characterizing LGM-to-Holocene anomalies. Our estimate is slightly higher than the SNR of 1.0 employed in the LGM climate field reconstruction by Paul et al. (2021).

3.3 Pseudo-proxy experiments

We use PPEs for the following three purposes: (i) to demonstrate the main features in the simulations that are captured by the model–data comparison algorithm; (ii) to diagnose how much model–data comparison results depend on limited temporal resolution, chronological uncertainties, and the magnitude and temporal autocorrelation structure of non-climatic noise; and (iii) to investigate how sensitive re-

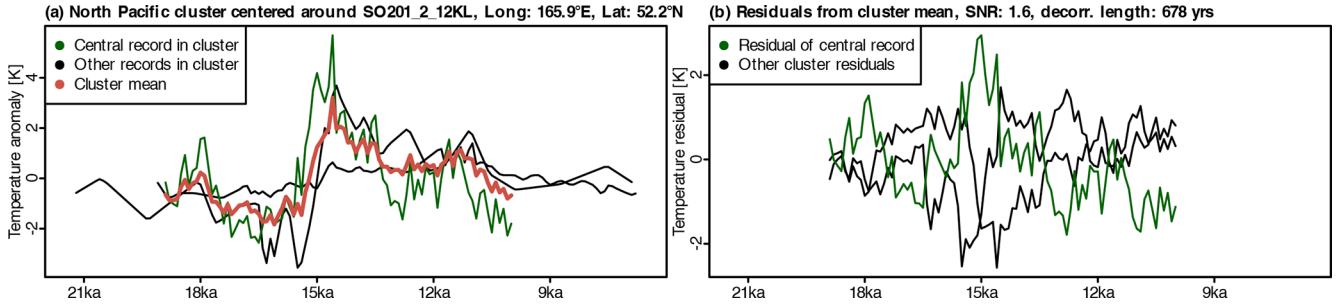

Figure 3. Visualization of the PSM parameter estimation as described in Sect. 3.2 for a cluster with 500 km radius and $n = 3$ records in the North Pacific centered around the proxy record SO201_2_12KL. **(a)** All SST records in the cluster and the corresponding local mean SST reconstruction (red line) with the central record of the cluster in green. **(b)** Residual deviations from the local mean reconstruction with the central record in green. The SNR and decorrelation length for the central record (green) are given in the caption. SNRs are estimated by comparing the variance of the mean reconstruction (signal) against the variance of the residuals (noise). The decorrelation length of the noise process is estimated from the residual time series.

sults are when noise magnitude and temporal autocorrelation structure in the PSM are different from their optimal values. Note that the difference between (ii) and (iii) is that (ii) is motivated by quantifiable limitations and uncertainties of reconstructions, while (iii) specifically targets the fact that the employed PSM is just an approximation of reality and its optimal parameters are unknown.

In PPEs, the underlying climate evolution is given by a reference simulation. The temperature time series of the reference simulation at each proxy location serves as the ground truth in the PPE. For each proxy record, the PSM from Sect. 3.1 is applied to the reference simulation to generate a single realization of forward-modeled proxy time series with a randomly selected iteration of the age–depth model and one realization of the non-climatic noise process. As this realization mimics the properties of the SST reconstructions, we call it a pseudo-proxy. We simulate pseudo-proxies at the locations and with the time axes and chronological uncertainties of the 74 selected proxy records from Sect. 2.2. Following this, the algorithm from Sect. 3.1 is employed to compute the deviations between $N = 100$ realizations of forward-modeled proxy time series derived from each simulation and the pseudo-proxies.

For (i), we use an example PPE with a subset of simulations to illustrate how simulation characteristics such as parameter configurations and the implementation of boundary conditions influence their ranking by our algorithm. We use MPI_Glac1D_P3 as reference simulation and PSM parameters given by the estimates from Sect. 3.2 (SNR = 1.6, decorrelation length = 1289 years). For the PPE, we select simulations that differ from the reference simulations in boundary conditions (MPI_Ice6G_P2_noMWF, TraCE-ALL), parameter configuration (MPI_Glac1D_PTK), and employed climate model (TraCE-ALL). Additionally, two idealized modifications of MPI_Glac1D_P3, which are shifted in time by 2 kyr in either direction (MPI_Glac1D_P3-2k, MPI_Glac1D_P3+2k), show the effects of a timing mismatch in the deglacial temperature evolution on the model–data comparison results (Fig. 4a).

For (ii) and (iii), we perform two sets of PPEs (Table 3). In the first set we assume that the noise magnitude and type in the PSM are known but we systematically vary the noise level of the records from very high (SNR = 1/4) to very low (SNR = 16) and include PPEs without additive noise process (SNR = Inf). We further vary the noise type between white noise (no autocorrelation), an AR1 process with a decorrelation length of 1 kyr, and a self-similar process following a power law distribution with exponent one (red noise). Using all 10 transient simulations as reference simulations to avoid spurious results from selecting a specific reference simulation, we perform in total 240 PPEs (8 SNRs, 3 noise types, 10 reference simulations).

In the second set, the PSM structure used for generating the forward-modeled proxy time series employed in the model–data comparison algorithm deviates from the one selected to simulate the pseudo-proxies, thus imitating the case where the PSM structure is uncertain. For each of the 10 reference simulations, we draw a realization of pseudo-proxies with AR1 noise (SNR = 2, decorrelation length = 1 kyr). For each pseudo-proxy realization, we first apply the model–data comparison algorithm with varying noise levels in the PSM (SNR = 1/4 to SNR = 16 and SNR = Inf) but the same autocorrelation structure as in the construction of the pseudo-proxies. Following this we apply the model–data algorithm with varying autocorrelation structure (white, AR1, and power-law noise) but the same noise level as in the construction of the pseudo-proxies.

Whether a certain IQD corresponds to an acceptable agreement between a simulation and a reconstruction is a subjective choice. Moreover, because the IQD uses the probability distribution of the forward-modeled proxy time series, its absolute value depends on the specification of the PSM. For example, a higher noise level results in a larger spread of the forward-modeled proxy time series created from the same

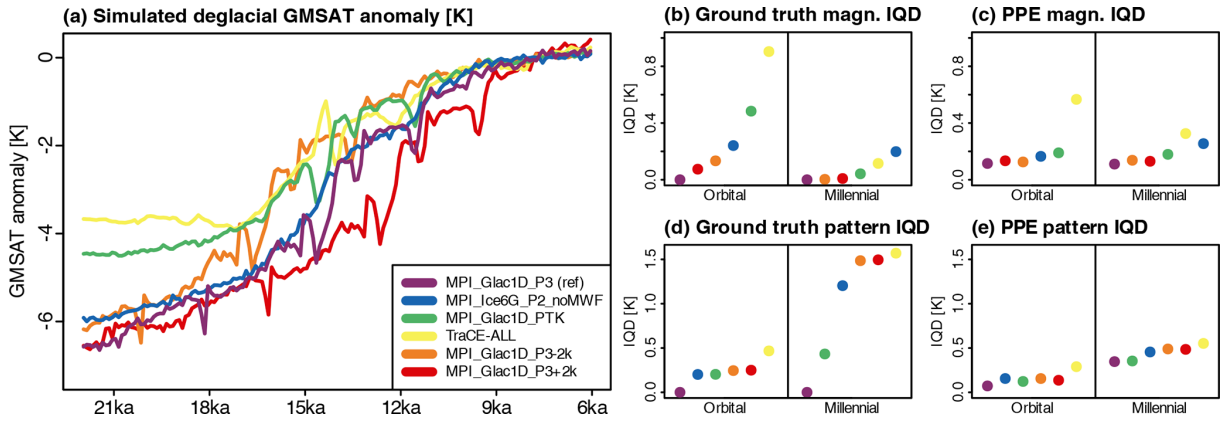

Figure 4. Visualization of the results for a PPE with SNR = 1.6, an AR1 noise process with a decorrelation length of 1289 years, and MPI_Glac1D_P3 as reference simulation. **(a)** GMSAT anomalies of the four simulations and the two time-shifted versions of MPI_Glac1D_P3 (anomalies with respect to the mean in the window 9 to 6 ka). Panels **(b)** and **(d)** show the ground-truth magnitude and pattern IQDs (see Sect. 3.3 for details). Panels **(c)** and **(e)** are the corresponding deviations between forward-modeled proxy time series and pseudo-proxies constructed from the reference simulation. Note that by definition the ground truth deviations in **(b)** and **(d)** of the reference simulation MPI_Glac1D_P3 from itself are zero.

Table 3. Characteristics of the example PPE and the two sets of PPEs described in Sect. 3.3. For set 1, all combinations of reference simulations, pseudo-proxy SNRs, and pseudo-proxy noise types are employed with the same settings for pseudo-proxies and forward-modeled proxy time series. For set 2, the 12 combinations of reference simulations, pseudo-proxy SNRs, and pseudo-proxy noise types are employed with all combinations of forward-modeled proxy time series SNRs and noise types.

Name	Reference simulations	Pseudo-proxy SNRs	Pseudo-proxy noise types	Forward-modeled proxy SNRs	Forward-modeled proxy noise type
Example	MPI_Glac1D_P3	1.6	AR1 (1289 years)	As pseudo-proxies	As pseudo-proxies
Set 1	All ensemble members	1/4, 1/2, 1, 2, 4, 8, 16, Inf	White AR1 (1000 years) power-law	As pseudo-proxies	As pseudo-proxies
Set 2	All ensemble members	2	AR1 (1000 years)	1/4, 1/2, 1, 2, 4, 8, 16, Inf	White, AR1 (1000 years), power-law

simulation, such that the IQD for a high noise level will differ from the IQD for a low noise level, even if the simulated and reconstructed SST time series are the same. Therefore, we focus on the ability of the algorithm to reliably discriminate between simulations, i.e., determining whether simulation A is closer to reality than simulation B. In PPEs, we can compute the "ground truth deviation" between a simulation and the reference climate history that was used to construct the pseudo-proxies. We choose the mean absolute deviation from the reference simulation at the locations of the proxy records as ground truth deviation because the IQD reduces to the mean absolute difference in the absence of uncertainties. We then compute a reference ranking by sorting the simulations according to their ground truth deviations. Similarly, we can rank the simulations according to the IQDs between the forward-modeled proxy time series and the pseudo-proxies, which is the ranking that would be obtained in a real-world model–data comparison situation in which only the pseudo-proxies are known but not the underlying reference climate history. We call this the pseudo-proxy ranking.

Finally, we compare the reference ranking with the pseudo-proxy ranking. If the model–data comparison algorithm discriminated perfectly between simulations, the reference ranking and pseudo-proxy ranking would be identical. However, due to reconstruction uncertainties and limitations, this will not always be the case. To quantify the similarity of the two rankings, we introduce a measure called the "fraction of pairwise reversed rankings" (FPRR). This measure is based on pairwise comparisons of the rankings of simulations: if simulation A ranks higher than simulation B in the reference ranking but ranks lower in the pseudo-proxy ranking, we say that the ranking of the two simulations is reversed in the pseudo-proxy ranking, i.e., the two simulations are erroneously ranked by the model–data comparison algo-

rithm. We assign 1 to the pairwise comparison if the ranking is reversed and 0 if it is not reversed. We compare the rankings for all pairs of simulations and define the FPRR as the mean of all pairwise comparisons. The FPRR is 0 when the pseudo-proxy and reference rankings are equal and is 1 if the two rankings are exactly reversed. The expected value for a random ranking of simulations is 0.5, which means that an FPRR below 0.5 indicates a better-than-random ranking. We focus on two aspects of the simulations' rankings: (i) the reliability of rankings, i.e., the expected probability of erroneously ranking simulations which we define as the median IQD in a set of PPEs with the same PSM parameters, and (ii) the robustness of rankings, i.e., how much the probability of an erroneous ranking depends on the reference climate history and the realization of non-climatic processes in the pseudo-proxies. Robustness is quantified by the spread of the IQD in a set of PPEs with the same PSM parameters and can be interpreted as a measure for the predictability of the reliability of model–data comparison results.

4 Results

We start this section with an example PPE that demonstrates the characteristics of the model–data comparison algorithm. We then use the PPE framework to systematically assess the dependency of model–data comparison results on uncertainties and limitations of SST reconstructions. Finally, we demonstrate our algorithm in a real-world setting by quantifying the deviations between deglacial simulations and SST reconstructions.

4.1 Pseudo-proxy experiments

4.1.1 Exemplifying pseudo-proxy experiment

As described in Sect. 3.3, we use an example PPE with MPI_Glac1D_P3 as reference simulation to demonstrate how a simulation's characteristics influence their ranking by our algorithm. The globally averaged ground truth deviations, i.e., IQDs between simulations and the reference simulation at the proxy locations with a regular temporal resolution, no chronological uncertainties, and no non-climatic noise are shown in Fig. 4b and d, and the IQDs from the comparison between forward-modeled proxy time series and pseudo-proxies are shown in Fig. 4c and e. For all four components of the deglacial temperature evolution (orbital magnitudes, millennial magnitudes, orbital patterns, and millennial patterns), the spread between IQDs corresponding to different simulations are smaller in the PPE (Fig. 4c, e) than in the ground truth deviations (Fig. 4b, d). This shows that in the presence of uncertainties, the forward-modeled proxy time series constructed from different simulations are harder to distinguish than the simulations in the uncertainty-free ground truth. However, the pseudo-proxy ranking mostly preserves the reference ranking (see Sect. 3.3 for a definition), which demonstrates the ability of the algorithm to still discriminate correctly between simulations in the presence of reconstruction limitations and uncertainties.

Comparing the IQDs with simulated global mean temperatures (Fig. 4a), we see that the orbital magnitude IQD rankings follow the differences in the magnitude of deglacial warming compared to the reference simulation. Meltwater fluxes have a strong influence on millennial magnitude rankings. MPI_Ice6G_P2_noMWF, in which no meltwater flux is applied, deviates substantially from the reference simulation. The varying spatial structure of millennial magnitudes due to the different meltwater history between TraCE-ALL and MPI_Glac1D_P3 seems to be exaggerated in the PPE. This leads to TraCE-ALL having a higher millennial magnitude IQD than MPI_Ice6G_P2_noMWF in the PPE but not in the ground truth.

The orbital pattern IQDs do not vary strongly between the MPI-ESM simulations, which all feature similar warming trends. In contrast, deglacial warming starts later and is more abrupt in TraCE-ALL, which results in a higher orbital pattern IQD. The difference in the meltwater histories is reflected in the millennial pattern component: MPI_Glac1D_P3 and MPI_Glac1D_PTK feature smaller IQDs than MPI_Ice6G_P2_noMWF, which does not exhibit pronounced millennial-scale fluctuations. The millennial pattern IQD is highest in TraCE-ALL, where a strong fluctuation around 14.5 ka is of opposite sign to MPI_Glac1D_P3.

In the reference rankings, as well as the PPE, the time-shifted versions of MPI_Glac1D_P3 are very similar to the reference simulation in the magnitude components (Fig. 4b, c). This is because the magnitude of orbital and millennial variations changes little under time shifts. In contrast, time-shifted versions deviate substantially from the reference simulation in the temporal patterns (Fig. 4d, e) because the timing of the start and end of the deglacial warming, as well as the millennial-scale fluctuations, differs from the reference simulation. This shows that the magnitude IQDs are insensitive to differences in the timing of events, whereas timing differences appear pronounced in the pattern IQDs.

4.1.2 Reliability and robustness of simulation rankings

We analyze the first set of 240 PPEs (see Sect. 3.3, set 1 in Table 3) by aggregating them according to the employed noise level and compare the respective FPRRs for three averaging scales: globally, zonally, and locally (Fig. 5). For all averaging scales, FPRRs increase for lower SNRs, i.e., pseudo-proxy rankings deviate more from the reference ranking for higher noise levels. However, even for the highest considered noise levels, the FPRRs are rarely above 0.5. Thus, there is almost always enough information of the underlying signal preserved to obtain a better than random ranking. There is no threshold behavior, but a steady FPRR increase for higher noise levels. This increase is expected since higher

non-climatic noise levels make it harder to distinguish simulations.

On average, rankings of orbital magnitudes differ least from the reference rankings, followed by orbital patterns, and millennial patterns. Millennial magnitude rankings are the least reliable under non-climatic noise. More reliable orbital than millennial rankings are expected because temperature variations are larger on orbital than millennial timescales whereas the noise level does not increase by the same rate on longer timescales. Median FPRRs mostly increase for decreasing spatial averaging scales, i.e., the reliability of rankings decreases from globally to locally averaged IQDs. The spread of FPRRs over the PPEs with the same noise level tends to increase with higher noise level and smaller spatial averaging scale, too. Thus, model–data comparison results are not just less reliable but also less robust for higher noise levels and smaller averaging scales (see also Sect. 3.3). For our SNR estimates from Sect. 3.2, the PPE results suggest below 10 % expected erroneous simulation rankings for orbital magnitudes and patterns and 10 %–20 % for millennial patterns and magnitudes.

In reality, the magnitude and temporal structure of non-climatic processes is uncertain. Therefore, we test how robust model–data comparison results are when either the noise level or the temporal autocorrelation structure in the forward-modeled proxy time series differs from the values selected to construct the pseudo-proxies (see set 2 in Table 3). Figure S10 in the Supplement shows the FPRR for overestimated or underestimated SNRs and for overestimated (power-law) or underestimated (white noise) temporal persistence of non-climatic processes. We find small influences from moderately (factor 2 to 4) overestimating or underestimating the noise level. Substantial differences from the results for the true noise level only occur for strong deviations (larger than factor 4) from the true level or when non-climatic processes are neglected entirely (SNR = Inf), especially for millennial magnitudes. For the latter, the reliability tends to decrease when the noise level is overestimated, whereas the robustness decreases when the noise level is underestimated. Neglecting non-climatic noise entirely for millennial magnitudes reduces the reliability more for global averages than on smaller spatial scales (see also Sect. 5.1). For all averaging scales and all four components, the effects of misspecified temporal autocorrelation structures are negligible. This supports the decision to choose an AR1 process in Sect. 3.2 instead of trying to estimate the structure of the temporal autocorrelation function.

4.2 Comparison of simulations against SST reconstructions

Next, we quantify the deviations between forward-modeled proxy time series derived from the 10 deglacial simulations (Sect. 2.1) and the 74 selected SST records (Sect. 2.2). We employ a PSM with an AR1 non-climatic noise process and vary the SNR between 1.1 and 2.2 and the decorrelation length between 865 and 1712 years (Sect. 3.2). We study globally and regionally averaged IQDs for the Southern Hemisphere extratropics ($n = 10$ proxy records), the tropics ($n = 44$), the extratropical North Atlantic ($n = 13$), and the extratropical North Pacific ($n = 7$) (Fig. 1). We select these regions based on detected inter-regional dissimilarities of the deglacial temperature evolution in an initial visual inspection of reconstructions and simulations. The averaged temporal evolution of the reconstructed temperatures and forward-modeled proxy time series at the proxy record locations is depicted for each of the four disjunct regions in Fig. 6. All regions contain more than five records and thus we expect the results to benefit from the spatial averaging effect found in the PPEs. Figure 7 shows the IQDs for all four components of the deglacial temperature evolution, simulations, and regions. An alternative visualization of the deviations, which combines magnitude and pattern deviations for a given timescale, is provided in the Supplement (Figs. S11, S12). In the next two subsections, we assess orbital- and millennial-scale variations of our main set of simulations. Finally, we analyze the model–proxy agreement of the three sensitivity experiments.

4.2.1 Orbital-scale variations

For orbital magnitudes, MPI_Glac1D_P3, MPI_Glac1D_PTK, and TraCE-ALL feature the smallest deviations between forward-modeled proxy time series and reconstructions in the global average (Fig. 7a). Among these three simulations, MPI_Glac1D_PTK and TraCE-ALL warm by ~ 4 K during the deglaciation (see Fig. 1) and deviate less from the reconstructions than other simulations in the Southern Hemisphere and tropics. Meanwhile, MPI_Glac1D_P3 has the strongest deglacial warming among the simulations and deviates significantly less from the reconstruction in the North Atlantic than all other simulations. In the global average, these regionally varying agreements compensate each other, which shows that global mean temperature alone is insufficient to explain the rankings. In the tropics and Southern Hemisphere, forward-modeled proxy time series with median orbital magnitudes around 1 K tend to deviate least from the reconstructions (Figs. 7a, 8a). In the North Atlantic, no simulation matches the high orbital magnitudes of the reconstructions (Fig. 8a). Here, the simulation with the highest magnitude (MPI_Glac1D_P3) features the lowest IQDs. In the North Pacific, orbital magnitudes are much smaller than in the North Atlantic in reconstructions and all simulations, and IQDs are relatively similar for all simulations.

Turning to orbital patterns, the globally averaged IQD differences between simulations are relatively small (Fig. 7b). In the North Atlantic, two distinct regional clusters appear in the reconstructions (Fig. 9a, c): along the Iberian Margin and in the Mediterranean Sea (denoted Mediterranean

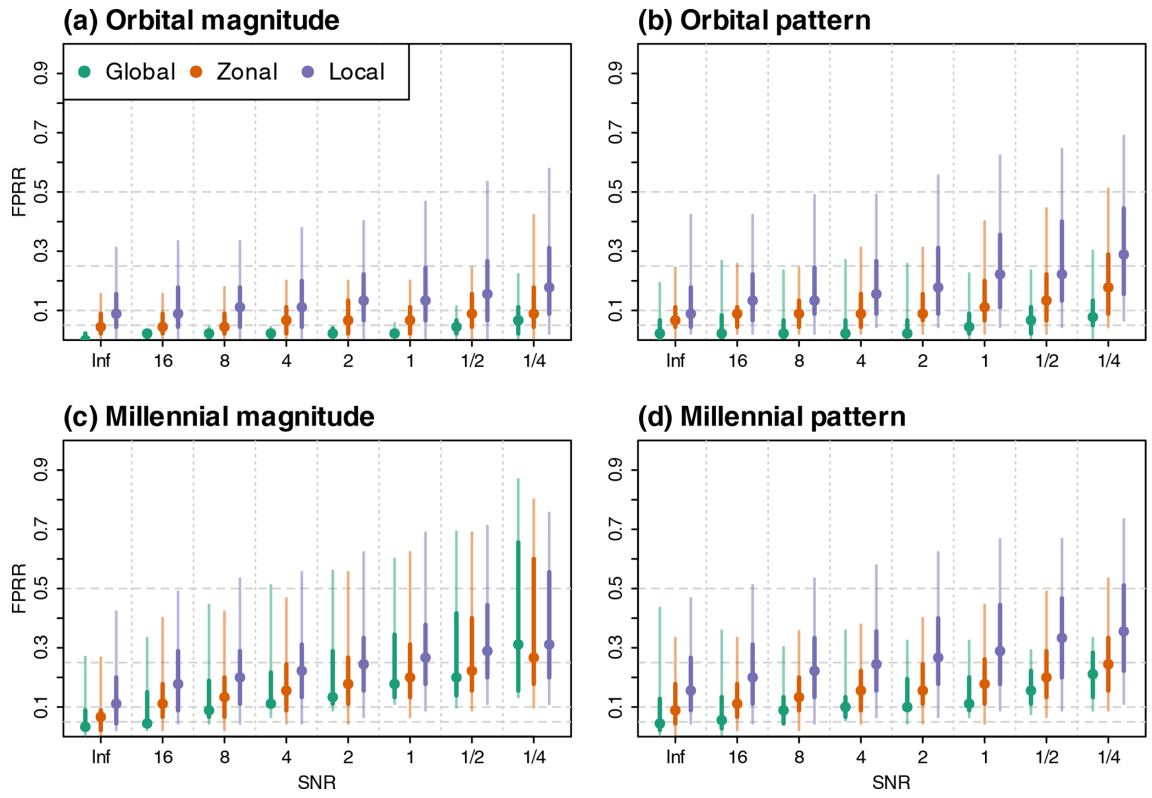

Figure 5. Fraction of pairwise reversed rankings (FPRR; see Sect. 3.3 for a definition) of simulations for globally averaged IQDs, zonally averaged IQDs, and IQDs of individual pseudo-proxy records. Shown are FPRRs for **(a)** orbital-scale magnitudes, **(b)** millennial-scale magnitudes, **(c)** orbital-scale temporal patterns, and **(d)** millennial-scale temporal patterns. Dots depict the medians across all PPEs with a given SNR ($n = 30$ for each SNR). Bars show the spread across PPEs. Darker colors depict the 25th to 75th percentiles, whereas lighter colors depict the 5th to 95th percentiles. SNR = Inf corresponds to PPEs without additive noise process. Dashed horizontal lines indicate FPRRs of 0.05, 0.1, 0.25, and 0.5. FPRRs above 0.5 are worse than expected for a randomized ranking.

North Atlantic; see Fig. 1b), the lowest SSTs occur during Heinrich Stadial 1 (∼ 17 ka), followed by two strong warming phases, which are interrupted by a warming hiatus during the Younger Dryas (∼ 12 ka). Meanwhile, warming is more monotonic in the subpolar North Atlantic (see Fig. 1b for a definition of the region). In contrast to the reconstructions, the orbital patterns are very similar between those two subregions of the North Atlantic in all of the simulations (Fig. 9a, c). Due to the differences between Subpolar and Mediterranean North Atlantic in the reconstructions, the lowest orbital pattern IQDs in the North Atlantic occur in MPI_Ice6G_P2_noMW and MPI_Glac1D_P3, which feature a smoother orbital pattern with weaker interruptions of the warming trend than other simulations. Among all examined regions, the highest orbital pattern IQDs occur in the North Pacific, where inter-model differences in orbital patterns are also the largest (Fig. 9e). Here, TraCE-ALL has the lowest IQD as it is the only simulation that somewhat resembles the pattern in the reconstructions with a temperature increase until ∼ 14 ka and subsequent cooling into the Holocene.

4.2.2 Millennial-scale variations

Millennial magnitude IQDs exhibit small differences between the simulations containing meltwater-induced abrupt events when averaged globally as well as in the Southern Hemisphere extratropics and in the Tropics (Fig. 7c). The highest millennial magnitudes in reconstructions and simulations occur in the North Atlantic (Fig. 8b). Here, two simulations with medium millennial magnitudes, TraCE-ALL and MPI_Glac1D_PTK, have the smallest IQDs, whereas the largest deviations from the reconstructions occur for the simulation without meltwater input, MPI_Ice6G_P2_noMWF. Compared to the North Atlantic, millennial-scale variations are weaker in the North Pacific in reconstructions and simulations and IQDs are more similar between simulations.

Turning to millennial patterns, MPI_Ice6G_P2_noMWF, a simulation without distinct millennial-scale variations, features the lowest globally averaged IQD (Fig. 7d). This is because no single simulation with distinct millennial-scale variations reproduces the reconstructed millennial patterns effectively in all regions. The agreement between simulations and reconstructions even differs within the North Atlantic and be-

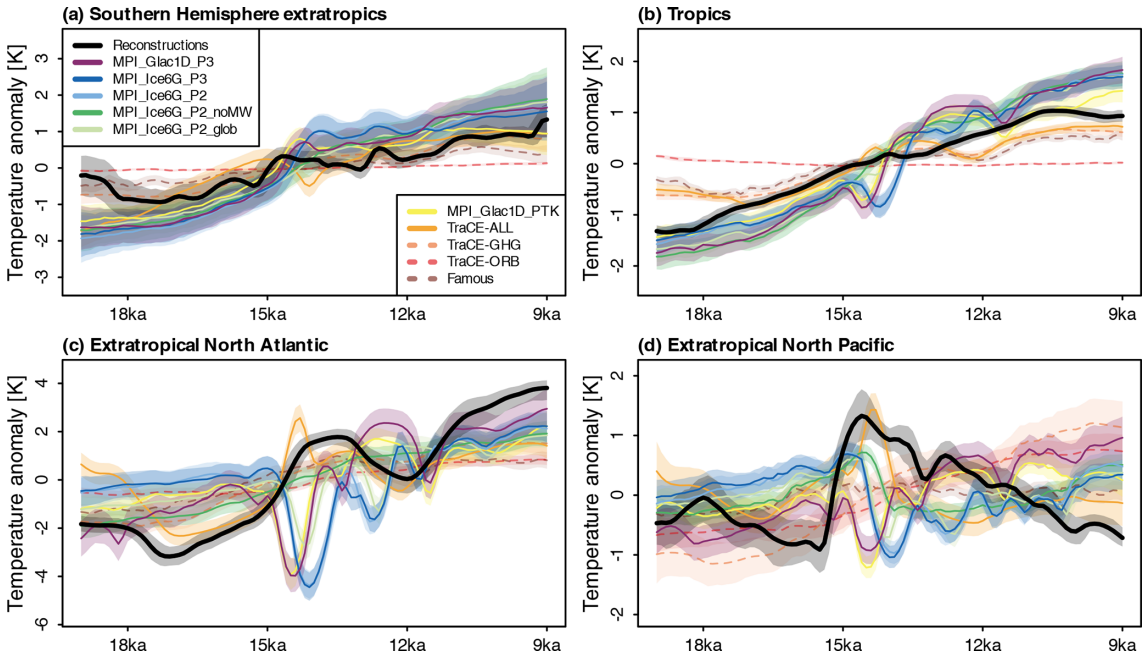

Figure 6. Regionally stacked SST variations for records in **(a)** the Southern Hemisphere extratropics ($n = 10$ proxy records), **(b)** the tropics ($n = 44$), **(c)** the extratropical North Atlantic ($n = 13$), and **(d)** the extratropical North Pacific ($n = 7$). Black lines denote the stacked reconstructions, whereas colored lines depict the stacked forward-modeled proxy time series derived from the 10 transient simulations. Shaded areas show uncertainties from chronologies and the PSM. Note that the stacks are not used in the model–data comparison algorithm but provide a visual impression of the reconstructed and simulated regional temporal evolution. The methodology to construct the stacks is described in the Supplement (Sect. S5).

tween North Atlantic and North Pacific (Fig. 9). Here, the meltwater fluxes extracted from the ice sheet reconstructions through dynamic river routing in the MPI-ESM simulations lead to abrupt millennial-scale temperature variations that do not align with the reconstructions. TraCE-ALL matches the millennial-scale variability pattern in the Mediterranean North Atlantic and therefore features the smallest IQDs in this area (Fig. 9b). However, it deviates strongly from the reconstructions in the Subpolar North Atlantic (Fig. 9d) and North Pacific (Fig. 9f).

4.2.3 Comparison of sensitivity experiments

Finally, we assess the model–proxy agreement of the three sensitivity experiment simulations, TraCE-GHG, TraCE-ORB, and FAMOUS. TraCE-GHG forward-modeled proxy time series have mostly comparable IQDs to the main set of simulations (Fig. 7). Only for millennial magnitudes, the TraCE-GHG IQDs are substantially higher than for the main set of simulations, in particular in the Southern Hemisphere. In the North Atlantic, all simulations with freshwater input have lower millennial magnitude IQDs than TraCE-GHG. In the global average, TraCE-ORB has the highest IQDs for orbital magnitudes, orbital patterns, and millennial magnitudes (Fig. 7). This is the result of lower orbital and millennial magnitudes than the other simulations (Fig. 8) and the absence of a deglacial warming trend in the Southern Hemisphere (Fig. 6). TraCE-ORB does not deviate substantially more from the reconstructions than the other simulations only for millennial pattern IQDs. FAMOUS features higher magnitude IQDs than the main set of simulations in the global average and in most regions (Fig. 7). For the pattern components, FAMOUS IQDs are in the range of the main set of simulations in the global average and in all regions other than the Southern Hemisphere, where it has higher IQDs for orbital and millennial patterns.

5 Discussion

Our study is a first step towards quantitative spatio-temporal model–data comparison for transient simulations of past climate transitions, as demonstrated here for the last deglaciation. In this section, we explore reasons for the PPE results and their implications. We then discuss the agreement between transient simulations of the last deglaciation and SST reconstructions, provide ideas for testing potential reasons for disagreements, and suggest improvements for future applications.

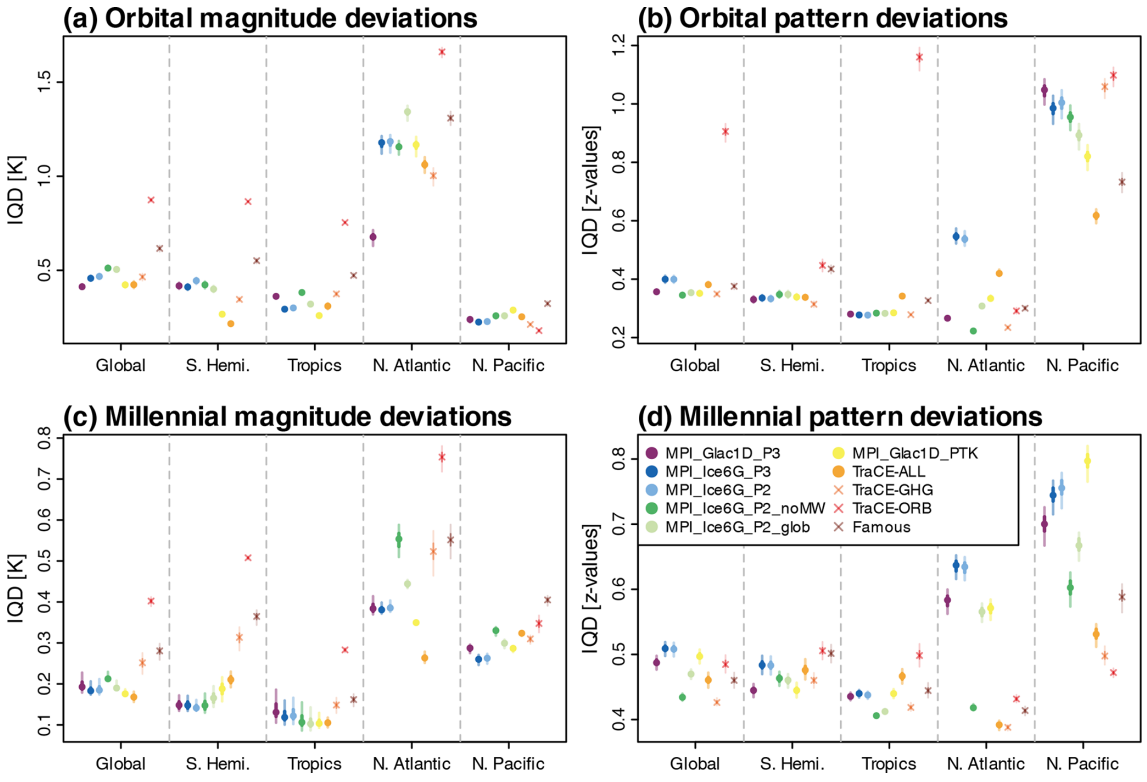

Figure 7. Global and regional mean IQDs of the 10 transient deglacial simulations from the 74 SST reconstruction records. Colored dots show median IQDs for **(a)** orbital magnitudes, **(b)** millennial magnitudes, **(c)** orbital temporal patterns, and **(d)** millennial temporal patterns. Darker colors depict the 25th to 75th percentiles resulting from varying the uncertain PSM parameters, whereas lighter colors depict the full range of uncertainties from varying the PSM parameters as described in Sect. 4.2. Note that the ranges of the y axes are different between the panels.

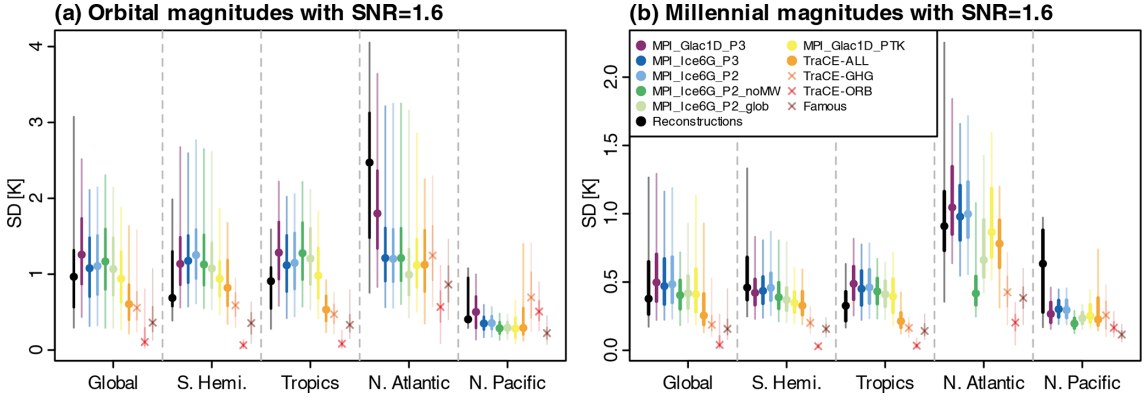

Figure 8. Mean absolute magnitudes of timescale-dependent variations of SST reconstructions (black) and forward-modeled proxy time series with the median PSM parameter estimates from Sect. 3.2 (color-coded). Depicted are globally and regionally averaged magnitudes of **(a)** orbital-scale and **(b)** millennial-scale variations. Points denote median magnitudes within a region. Darker color bars depict the 25th to 75th percentiles across all records within the respective region, whereas lighter colors depict the 5th and 95th percentile.

5.1 Reliability and robustness of the model–data comparison algorithm

The systematic PPEs show that the reliability and robustness of simulation rankings decrease with increasing noise levels. This result is not surprising as higher noise levels make it harder to identify the underlying temperature signal. The effect can be reduced by spatially averaging results from multiple records. As we assume the non-climatic noise to be independent between records, averaging over IQDs from multi-

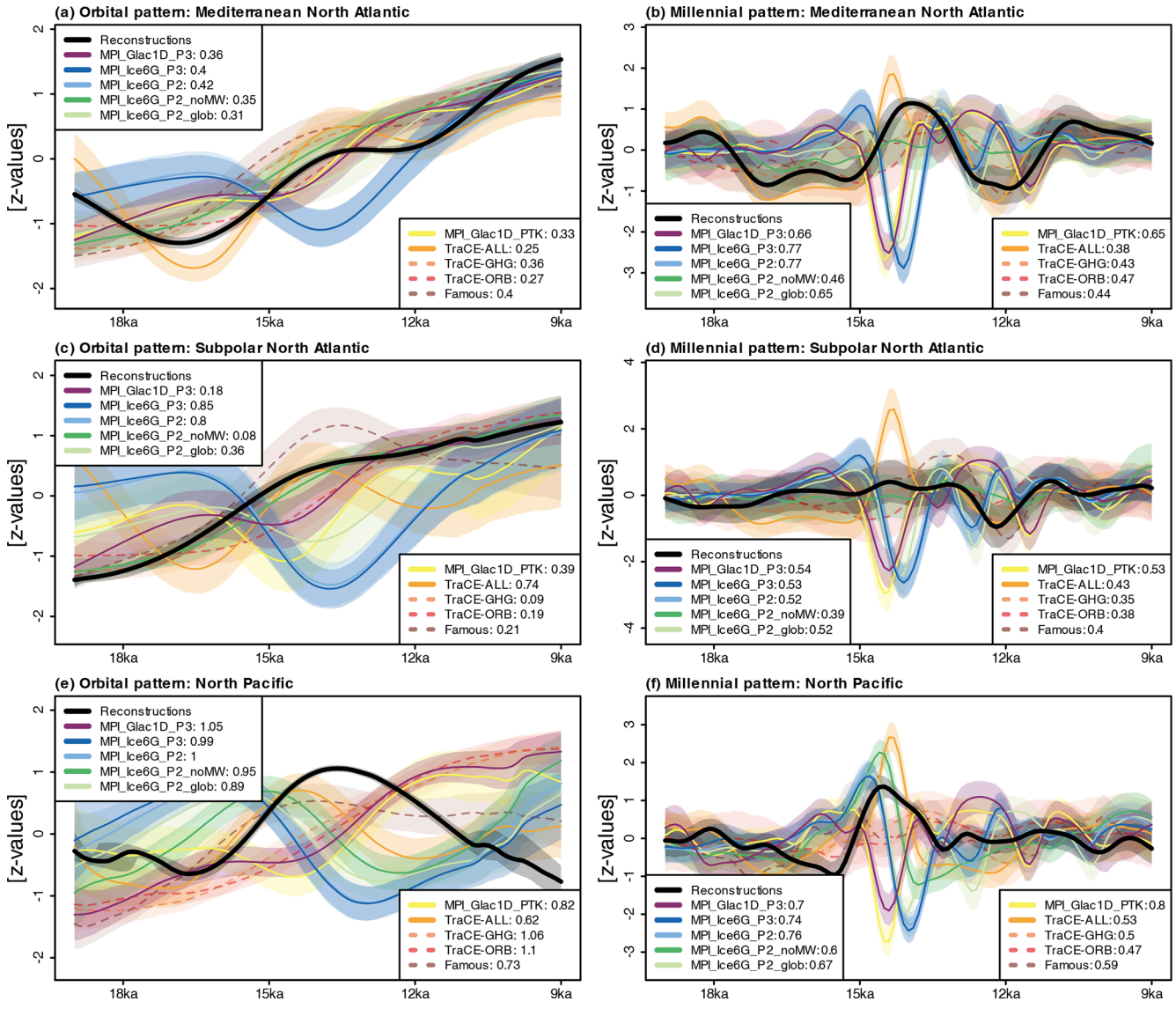

Figure 9. Regionally stacked temporal patterns of orbital-scale **(a, c, e)** and millennial-scale **(b, d, f)** variations for records in **(a, b)** the Mediterranean North Atlantic, **(c, d)** the Subpolar North Atlantic, and **(e, f)** the North Pacific (see Fig. 1 for the definition of the regions). Black lines denote the stacked reconstructions, whereas colored lines depict the stacked forward-modeled proxy time series derived from the 10 transient simulations. Shaded areas show uncertainties from chronologies and the PSM. The numbers in the legends next to each simulation are the averaged IQDs over all records in the respective regions. Note that the stacks are not used in the model–data comparison algorithm, but just facilitate the interpretation of the IQDs. The methodology to construct the stacks is described in the Supplement (Sect. S5).

ple records reduces the influence of the noise and thus effectively enhances the SNR. If modulations of the temperature signal were not independent between records in reality, the improvement from spatial averaging would be weakened.

Rankings for orbital-scale variations are more reliable and robust than for millennial-scale variations due to comparably smaller distortion by non-climatic noise. That orbital magnitude rankings tend to be more reliable and robust than orbital pattern rankings could be due to relatively subtle differences between simulations in the timing and shape of the deglacial warming trend compared to easier to identify differences in the magnitude of deglacial warming. On the other hand, we attribute more reliable and robust millennial pattern than magnitude rankings to the differing effects of non-climatic noise on these two components. Millennial patterns of simulations are often still distinguishable based on their most pronounced fluctuations that are comparatively less distorted by non-climatic noise. Meanwhile, non-climatic noise enhances the magnitude of reconstructed millennial-scale variations (in our PSM proportional to the variability of the simulation at a given location) and thus has a systematic effect on mil-

lennial magnitudes, which can further diminish the reliability of rankings.

If the assumed noise level in the model–data comparison is not strongly overestimated or underestimated (factor 4 and more), results remain reliable. Using explicitly conservative SNR values is not safeguarding from erroneous rankings as strongly overestimating noise levels reduces the reliability, whereas strongly underestimating noise levels reduces the robustness of rankings. Incorrect specifications of the temporal autocorrelation structure of non-climatic processes have a negligible effect in our PPEs. This rather unexpected result might be due to the relatively short time period of investigation (16 kyr) compared to the timescales we study. This hypothesis could be tested in future work by repeating the experiments for longer periods. Entirely neglecting existing non-climatic processes leads to less robust and reliable rankings for millennial-scale variations. On the one hand, this can be explained by non-climatic variations in reconstructions being interpreted as climate signals, such that rankings depend more on the unknown realization of non-climatic processes. On the other hand, underestimating millennial-scale variations by neglecting variability-enhancing processes can systematically distort millennial magnitude rankings. This effect is strongest for global averages.

Taken together, the PPE results suggest that the reliability and robustness of model–data comparison results can be improved the most by increasing the SNR. In contrast, reducing the uncertainty of SNR estimates or improving the specification of the temporal autocorrelation structures will barely improve rankings. A doubling of the SNR typically reduces erroneous rankings by 1–3 percentage points. Thus, incremental improvements, for example through process-based modeling of modulations of the recorded climate signal, will only have a small effect on the reliability of rankings. PPEs without non-climatic noise typically still have 5 %–10 % erroneous rankings for regionally averaged IQDs. This percentage could be reduced by more precise chronologies and higher temporal resolutions of records. Comparing global, zonal, and local estimates suggests that significantly improved reliability can also be achieved by increasing the number of proxy records and thus averaging over more records in regional averages, as long as non-climatic contributions are not strongly correlated between records.

5.2 Agreement of SST reconstructions and deglacial simulations

The diversity of the simulations in terms of employed climate models and experiment protocols makes interpreting the results challenging. Comparing TraCE-ALL and the six MPI-ESM simulations, we find that none of the simulations ranks among the simulations with the smallest deviation from the reconstructions across all four components and considered regions. We confirm this visual impression from Fig. 7 by computing rank histograms among the main set of simulations. Rankings are computed for each proxy record and each of the four components. Averaged over all records and components, the ranks of the simulations are between 3.8 (for MPI_Glac1D_PTK and MPI_Ice6G_P2_glob) and 4.2 (for MPI_Glac1D_P3) with TraCE-ALL at an average rank of 4.0 (Fig. S13 in the Supplement). However, the ranks of TraCE-ALL concentrate strongly at 1 (highest agreement) and 7 (lowest agreement). In contrast, the rank histograms of the MPI-ESM simulations are flatter; i.e., they feature more similar occurrence rates across ranks. Thus, TraCE-ALL IQDs are more often outside than inside the range of the MPI-ESM simulations, even though it does not feature a consistently higher or lower rank. The concentration of TraCE-ALL at extreme ranks tends to hold for all four components (Figs. S14–S17). At the moment, we cannot attribute the difference in the rank score histograms to differences between either the used climate models or the employed experiment protocols. This is due to the differences in the experiment protocol between TraCE-ALL and the MPI-ESM simulations, particularly regarding the location, timing, and magnitude of freshwater injections. Nevertheless, the flatter rank histograms of the MPI-ESM simulations, despite substantial experiment protocol and parameter configuration differences among them, hint at a substantial influence from climate model differences.

Examples of regionally varying mismatches between simulations, which compensate in global averages, are found for all four components of the deglacial temperature evolution (see Sect. 4.2). These compensations occur because simulations with higher variability than others have higher variability in almost all regions (Fig. 8). Additionally, simulations tend to have similar temporal patterns at least within each hemisphere (Figs. 6, 9). In contrast, the reconstructed variability magnitudes are the most similar to the simulations with the highest variability in some regions, but closer to those with low variability in others. Similarly, the reconstructed variability patterns vary more between and within ocean basins than in the simulations. Therefore, we attribute the absence of a simulation with consistently high agreement relative to the others to more regionally confined variability magnitudes and patterns in reconstructions than in simulations. In other words, the reconstructed spatial variability of the deglacial temperature evolution is higher than in all considered simulations. For the North Atlantic, the differences in the reconstructed deglacial temperature evolution between the Mediterranean and the Subpolar North Atlantic found in this study are consistent with a recent synthesis by Pedro et al. (2022).

This mismatch in the spatio-temporal variability structure could be caused by uncertainties in ice sheet reconstructions, shortcomings of the employed models, or temperature reconstruction characteristics that vary between regions. One can assess the role of systematic reconstruction deviations from mean annual SST by integrating process-based PSMs (e.g., Dolman and Laepple, 2018; Kretschmer et al., 2018; Osman

et al., 2021) into our algorithm in future work. This could disentangle the importance of different processes occurring during the recording, archiving, and measuring of the proxy, e.g., recording season and depth preferences, confounding environmental variables, and bioturbation. Moreover, our procedure to estimate the PSM parameters requires interpolating the proxy records to a common time axis which is otherwise avoided in the model–data comparison algorithm. Developing a more sophisticated method for the parameter estimation would be beneficial for future applications of our algorithm.

The locations of proxy records are biased towards coastal regions, and, for some regions, our results rely on records clustered in small areas. This could reduce the model–data agreement if the resolution of models was insufficient for an accurate simulation of zonal temperature heterogeneity, e.g., due to coastal upwelling or deficiencies in the simulation of gyre circulations and air–sea interactions (Judd et al., 2020; Kwon et al., 2010; Ma et al., 2016; Paul et al., 2021; Seager et al., 2003). As higher-resolution simulations of the deglaciation are currently precluded by computational limitations, including more proxy data and physically motivated downscaling of simulation output could help test this explanation. Finally, the reconstructed meltwater peaks could be too high or the models' responses to them too strong, leading to a spatially too homogeneous SST response (He and Clark, 2022). Insights into this potential explanation could be gained from coupled atmosphere–ocean–ice sheet simulations (Ziemen et al., 2019) or replacing local meltwater input with freshwater fingerprints obtained from eddy-resolving ocean models (Love et al., 2021).

The simulation with transient changes of orbital parameters only (TraCE-ORB) deviates significantly more from the reconstructions than all other simulations for orbital magnitudes, orbital patterns, and millennial magnitudes. This is due to too small magnitudes of variability in most regions and the absence of a deglacial warming trend in the Southern Hemisphere when GHG and ice sheet changes are neglected. We also find a systematically larger orbital magnitude mismatch between FAMOUS and the reconstructions compared to the main set of simulations because of weaker deglacial warming in FAMOUS. This could be explained by the acceleration in the forcing, which can delay global warming, but more simulations are needed to confirm this hypothesis.

In contrast, the neglected orbital and ice sheet forcing in TraCE-GHG does not lead to clearly higher disagreements for orbital-scale variability and millennial patterns. For millennial magnitudes, however, the absence of ice sheet forcing degrades results strongly. In particular, in the global average, all simulations with meltwater input show a better agreement with reconstructions for millennial magnitudes than those without meltwater input. The improved agreement originates mainly from a higher millennial-scale variability in the North Atlantic, where the meltwater-induced variability is the strongest. Moreover, the MPI-ESM simulation without meltwater input and TraCE-GHG have the smallest millennial pattern disagreement in the global average, which suggests that none of the employed meltwater schemes leads to a temporal pattern of millennial-scale variability that is globally consistent with the reconstructions. The uncertainties in ice sheet reconstructions (Abe-Ouchi et al., 2015; Ivanovic et al., 2016; Stokes et al., 2015) currently prevent determining the reason for the millennial pattern disagreements. The contrast between higher model–proxy agreement in simulating millennial magnitudes but no improvement for millennial patterns in the fully forced simulations hints at limitations in our current understanding of the spatio-temporal structure of millennial-scale variability during the deglaciation. Addressing these challenges with designated protocols in the context of inter-model comparison projects could be a promising way forward.

Our results suggest that reproducing the patterns of a small set of proxies might be an insufficient strategy to capture the spatial structure of millennial-scale temperature patterns. For example, reproducing the patterns of a specific Atlantic meridional overturning circulation (AMOC) proxy (e.g., Pa / Th ratios at Bermuda rise), as TraCE-ALL does (Liu et al., 2009), will not necessarily lead to a good model–proxy agreement for millennial-scale temperature patterns across different regions. Instead, other factors, such as the magnitude of the AMOC response or the background climatic state, could have a large influence on the regional manifestations of temperature variability. Alternatively, uncertainty regarding the origins of millennial-scale variability could lead to an adequate reproduction of the pattern of AMOC variability with an incorrect mechanism, which could result in a spatially varying degree of model–proxy agreement.

A single metric is likely insufficient for fully capturing the deviations between simulations and reconstructions in an interpretable way. When combining magnitude and pattern metrics in biplots (see Figs. S11, S12), simulations with local freshwater injection perform the best in the North Atlantic for either timescale: MPI_Glac1D_P3 for orbital timescales and TraCE-ALL for millennial timescales. While strong freshwater water-induced perturbations can have an imprint on the orbital-scale signal, when the perturbations are large enough to substantially influence time averages on orbital timescales, a good model–proxy agreement for orbital timescales does not imply a good agreement for millennial timescales and vice versa in our results. Instead, we argue that a varying importance of forcings and internal feedback processes on different temporal and spatial scales substantially affects the model–proxy agreements for different components.

As the PPEs and the real-world application have shown, the pattern IQDs are sensitive to the timing of timescale-dependent temperature fluctuations. Therefore, they are only meaningful if the goal of a simulation is to reproduce a specific succession of variations observed in reconstructions. Temporal alignment cannot be expected for internally driven

variations such as spontaneous millennial-scale fluctuations (Obase and Abe-Ouchi, 2019; Vettoretti et al., 2022) and in the presence of boundary conditions with large spatio-temporal uncertainties like deglacial meltwater fluxes. In these cases, the magnitude IQDs, which are insensitive to the timing of fluctuations, could be combined with a more insightful measure for temporal patterns, e.g., based on the similarity of spatial relationships in reconstructed and forward-modeled proxy time series (e.g., Adam et al., 2021).

Applications of our model–data algorithm are not restricted to SST reconstructions during the last deglaciation. With new syntheses becoming available (Herzschuh et al., 2023), an extension to terrestrial temperature records can be attempted. Moreover, other periods with climate transitions and changing background conditions can be assessed as long as a sufficient number of proxy records with absolute chronologies are available. Targets could, for example, be the penultimate deglaciation, the glacial inception, or the last glacial cycle. Finally, it is straightforward to adapt our algorithm for model–proxy comparison of other continuous variables such as oxygen isotopes, in particular if PSMs already exist that link the proxies to one or multiple simulated variables.

6 Conclusions

We present a new approach for the spatio-temporal comparison of reconstructed and simulated deglacial temperature evolutions. The algorithm applies proxy system models to simulation output and quantifies the deviation between the resulting forward-modeled proxy time series and temperature reconstructions. Thus, it can account for non-climatic processes that affect the temperature reconstructions and avoids the reconstruction of gridded fields or regional mean temperature time series from sparse and uncertain proxy data. We assess the reliability and robustness of the algorithm in pseudo-proxy experiments. For signal-to-noise ratios as estimated from a database of sea surface temperature reconstructions, the expected rate of simulation pairs that are ranked erroneously compared to the underlying ground truth is less than 10 % for magnitudes and temporal patterns of orbital-scale variations and 10 %–20 % for millennial-scale magnitudes and patterns, when deviations are regionally averaged. The quality of rankings is barely influenced by uncertainties in proxy system model parameters. The reliability and robustness of rankings could be improved most by including more data and increasing the signal-to-noise ratio.

Comparing 10 transient simulations of the last deglaciation with a global compilation of sea surface temperature reconstructions, we demonstrate that the algorithm provides insights into the importance of model differences and boundary conditions for explaining mismatches between simulations and reconstructions. The ranking of the simulations differs substantially between the considered regions and timescales, and no simulation features a consistently high agreement with the reconstructions. This suggests that optimizing for agreement with the temporal patterns of a specific proxy or reconstructions from a small region might be an inadequate strategy for capturing the spatial structure of millennial-scale temperature patterns during the deglaciation. We attribute these results to greater differences between and within ocean basins in reconstructions than in simulations. The mismatch could originate from uncertainties in boundary conditions, shortcomings of the employed climate models, or reconstruction characteristics that vary between regions. Further analyses are required to disentangle these potential explanations. In addition to assessing the temperature evolution during the last deglaciation, the proposed method can be applied to other continuous variables, e.g., oxygen isotopes, and other periods with climate transitions such as the penultimate deglaciation and the last glacial inception. Beyond quantifying disagreements between a given simulation and a database of reconstructions, our algorithm can be used for model tuning, testing the influence of uncertain boundary conditions, and understanding influences of non-climatic processes on model–data mismatches.

Code and data availability. R code to reproduce the results and plots of this study is available at https://doi.org/10.5281/zenodo.10497834 (Weitzel, 2024). The PalMod 130k marine paleoclimate data synthesis v1.1.1 is available at https://doi.org/10.5281/zenodo.7785766 (Jonkers et al., 2023). MPI-ESM simulation data were processed and provided by Marie Kapsch, Uwe Mikolajewicz, and Thomas Kleinen. Output from the MPI_Glac1D_P3, MPI_Ice6G_P3, MPI_Ice6G_P2, and MPI_Glac1D_PTK simulations is also available at https://doi.org/10.26050/WDCC/PMMXMCRTDGP132 (Mikolajewicz et al., 2023a), https://doi.org/10.26050/WDCC/PMMXMCRTDIP132 (Mikolajewicz et al., 2023b), https://doi.org/10.26050/WDCC/PMMXMCRTDIP122 (Mikolajewicz et al., 2023c), and https://doi.org/10.26050/WDCC/PMMXMCHTD (Kleinen et al., 2023b). TraCE data were obtained from https://www.earthsystemgrid.org/project/trace.html (Climate Data at the NSF National Center for Atmospheric Research, 2023), and FAMOUS data were obtained from https://catalogue.ceda.ac.uk/uuid/a43dcfaccfae4824ab9ab2b572703e72 (Lenton, 2008). More information on access to simulation output is available in the respective original publications (Kapsch et al., 2022; Kleinen et al., 2023a; Liu et al., 2009; Smith and Gregory, 2012).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/cp-20-865-2024-supplement.

Author contributions. NW, KR, and HA designed the study with input from OB, LJ, and AP. JPB, LJ, MK, TK, UM, NW, and

EZ processed the data. NW implemented and ran the model–data comparison algorithm. All authors discussed the results. NW wrote the manuscript with input from KR and HA. All authors commented on earlier versions of the manuscript and approved the final manuscript.

Competing interests. At least one of the (co-)authors is a member of the editorial board of *Climate of the Past*. The peer-review process was guided by an independent editor, and the authors also have no other competing interests to declare.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. This work originated from a workshop organized by Oliver Bothe and funded by Helmholtz-Zentrum Hereon and PalMod. All Max Planck Institute for Meteorology Earth System Model simulations were performed at the German Climate Computing Center (DKRZ). We thank Andrew Dolman for his helpful comments on a previous version of the manuscript. We thank the two anonymous reviewers and the editor, Marisa Montoya, for constructive feedback that improved the quality of the manuscript.

Financial support. Nils Weitzel, Elisa Ziegler, and Kira Rehfeld have been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; project no. 395588486). Heather Andres, Jean-Philippe Baudouin, Oliver Bothe, Lukas Jonkers, Marie-Luise Kapsch, Thomas Kleinen, Uwe Mikolajewicz, André Paul, and Nils Weitzel received funding from the German Federal Ministry of Education and Research (BMBF) within the Research for Sustainability initiative (FONA; https://www.fona.de/, last access: 10 May 2023) through the PalMod project, grant nos. (FKZ) 01LP1926C (Jean-Philippe Baudouin, Nils Weitzel), 01LP1509A (Oliver Bothe), 01LP1926B (Oliver Bothe), 01LP1922A (Lukas Jonkers), 01LP1504C (Marie-Luise Kapsch), 01LP1917B (Marie-Luise Kapsch), 01LP1921A (Thomas Kleinen), 01LP1915C (Uwe Mikolajewicz), and 01LP1511D (André Paul).

This open-access publication was funded
by the University of Tübingen.

Review statement. This paper was edited by Marisa Montoya and reviewed by two anonymous referees.

References

Abe-Ouchi, A., Saito, F., Kageyama, M., Braconnot, P., Harrison, S. P., Lambeck, K., Otto-Bliesner, B. L., Peltier, W. R., Tarasov, L., Peterschmitt, J.-Y., and Takahashi, K.: Ice-sheet configuration in the CMIP5/PMIP3 Last Glacial Maximum experiments, Geosci. Model Dev., 8, 3621–3637, https://doi.org/10.5194/gmd-8-3621-2015, 2015.

Adam, M., Weitzel, N., and Rehfeld, K.: Identifying Global-Scale Patterns of Vegetation Change During the Last Deglaciation From Paleoclimate Networks, Paleoceanogr. Paleocl., 36, e2021PA004265, https://doi.org/10.1029/2021PA004265, 2021.

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

Arz, H. W., Pätzold, J., Müller, P. J., and Moammar, M. O.: Influence of Northern Hemisphere climate and global sea level rise on the restricted Red Sea marine environment during termination I, Paleoceanography, 18, 1053, https://doi.org/10.1029/2002PA000864, 2003.

Bard, E., Rostek, F., Turon, J.-L., and Gendreau, S.: Hydrological Impact of Heinrich Events in the Subtropical Northeast Atlantic, Science, 289, 1321–1324, https://doi.org/10.1126/science.289.5483.1321, 2000.

Batchelor, C. L., Margold, M., Krapp, M., Murton, D. K., Dalton, A. S., Gibbard, P. L., Stokes, C. R., Murton, J. B., and Manica, A.: The configuration of Northern Hemisphere ice sheets through the Quaternary, Nat. Commun., 10, 3713, https://doi.org/10.1038/s41467-019-11601-2, 2019.

Benz, V., Esper, O., Gersonde, R., Lamy, F., and Tiedemann, R.: Last Glacial Maximum sea surface temperature and sea-ice extent in the Pacific sector of the Southern Ocean, Quaternary Sci. Rev., 146, 216–237, https://doi.org/10.1016/j.quascirev.2016.06.006, 2016.

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

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

Bolliet, T., Holbourn, A., Kuhnt, W., Laj, C., Kissel, C., Beaufort, L., Kienast, M., Andersen, N., and Garbe-Schönberg, D.: Mindanao Dome variability over the last 160 kyr: Episodic glacial cooling of the West Pacific Warm Pool, Paleoceanography, 26, PA1208, https://doi.org/10.1029/2010PA001966, 2011.

Bouimetarhan, I., Groeneveld, J., Dupont, L., and Zonneveld, K.: Low- to high-productivity pattern within Heinrich Stadial 1: Inferences from dinoflagellate cyst records off Senegal, Global Planet. Change, 106, 64–76, https://doi.org/10.1016/j.gloplacha.2013.03.007, 2013.

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, Nat. Clim. Change, 2, 417–424, https://doi.org/10.1038/nclimate1456, 2012.

Bühler, J. C., Roesch, C., Kirschner, M., Sime, L., Holloway, M. D., and Rehfeld, K.: Comparison of the oxygen isotope signatures in speleothem records and iHadCM3 model simulations for the last millennium, Clim. Past, 17, 985–1004, https://doi.org/10.5194/cp-17-985-2021, 2021.

Cacho, I., Grimalt, J. O., Pelejero, C., Canals, M., Sierro, F. J., Flores, J. A., and Shackleton, N.: Dansgaard-Oeschger and Heinrich event imprints in Alboran Sea paleotemperatures, Paleoceanography, 14, 698–705, https://doi.org/10.1029/1999PA900044, 1999.

Carlson, A. E., Oppo, D. W., Came, R. E., LeGrande, A. N., Keigwin, L. D., and Curry, W. B.: Subtropical Atlantic salinity variability and Atlantic meridional circulation during the last deglaciation, Geology, 36, 991, https://doi.org/10.1130/G25080A.1, 2008.

Chapman, M. R., Shackleton, N. J., Zhao, M., and Eglinton, G.: Faunal and alkenone reconstructions of subtropical North Atlantic surface hydrography and paleotemperature over the last 28 kyr, Paleoceanography, 11, 343–357, https://doi.org/10.1029/96PA00041, 1996.

Cheng, Z., Weng, C., Steinke, S., and Mohtadi, M.: Anthropogenic modification of vegetated landscapes in southern China from 6,000 years ago, Nat. Geosci., 11, 939–943, https://doi.org/10.1038/s41561-018-0250-1, 2018.

Chiessi, C. M., Mulitza, S., Paul, A., Pätzold, J., Groeneveld, J., and Wefer, G.: South Atlantic interocean exchange as the trigger for the Bølling warm event, Geology, 36, 919, https://doi.org/10.1130/G24979A.1, 2008.

Chiessi, C. M., Mulitza, S., Groeneveld, J., Silva, J. B., Campos, M. C., and Gurgel, M. H.: Variability of the Brazil Current during the late Holocene, Palaeogeography, Palaeoclimatology, Palaeoecology, 415, 28–36, https://doi.org/10.1016/j.palaeo.2013.12.005, 2014.

Chiessi, C. M., Mulitza, S., Mollenhauer, G., Silva, J. B., Groeneveld, J., and Prange, M.: Thermal evolution of the western South Atlantic and the adjacent continent during Termination 1, Clim. Past, 11, 915–929, https://doi.org/10.5194/cp-11-915-2015, 2015.

Clark, P. U., Shakun, J. D., Baker, P. A., Bartlein, P. J., Brewer, S., Brook, E., Carlson, A. E., Cheng, H., Kaufman, D. S., Liu, Z., Marchitto, T. M., Mix, A. C., Morrill, C., Otto-Bliesner, B. L., Pahnke, K., Russell, J. M., Whitlock, C., Adkins, J. F., Blois, J. L., Clark, J., Colman, S. M., Curry, W. B., Flower, B. P., He, F., Johnson, T. C., Lynch-Stieglitz, J., Markgraf, V., McManus, J., Mitrovica, J. X., Moreno, P. I., and Williams, J. W.: Global climate evolution during the last deglaciation, P. Natl. Acad. Sci. USA, 109, E1134–E1142, https://doi.org/10.1073/pnas.1116619109, 2012.

Cleator, S. F., Harrison, S. P., Nichols, N. K., Prentice, I. C., and Roulstone, I.: A new multivariable benchmark for Last Glacial Maximum climate simulations, Clim. Past, 16, 699–712, https://doi.org/10.5194/cp-16-699-2020, 2020.

Climate Data at the NSF National Center for Atmospheric Research: Simulation of the Transient Climate of the Last 21,000 Years (TraCE-21ka), NCAR Climate Data Gateway [data set], https://www.earthsystemgrid.org/project/trace.html, last access: 28 February 2023.

Crivellari, S., Chiessi, C. M., Kuhnert, H., Häggi, C., Mollenhauer, G., Hefter, J., Portilho-Ramos, R., Schefuß, E., and Mulitza, S.: Thermal response of the western tropical Atlantic to slowdown of the Atlantic Meridional Overturning Circulation, Earth Planet. Sc. Lett., 519, 120–129, https://doi.org/10.1016/j.epsl.2019.05.006, 2019.

Dallmeyer, A., Kleinen, T., Claussen, M., Weitzel, N., Cao, X., and Herzschuh, U.: The deglacial forest conundrum, Nat. Commun., 13, 6035, https://doi.org/10.1038/s41467-022-33646-6, 2022.

Dee, S., Parsons, L., Loope, G., Overpeck, J., Ault, T., and Emile-Geay, J.: Improved spectral comparisons of paleoclimate models and observations via proxy system modeling: Implications for multi-decadal variability, Earth Planet. Sc. Lett., 476, 34–46, https://doi.org/10.1016/j.epsl.2017.07.036, 2017.

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

Elderfield, H. and Ganssen, G.: Past temperature and $\delta^{18}O$ of surface ocean waters inferred from foraminiferal Mg / Ca ratios, Nature, 405, 442–445, https://doi.org/10.1038/35013033, 2000.

Evans, M., Tolwinski-Ward, S., Thompson, D., and Anchukaitis, K.: Applications of proxy system modeling in high resolution paleoclimatology, Quaternary Sci. Rev., 76, 16–28, https://doi.org/10.1016/j.quascirev.2013.05.024, 2013.

Gebhardt, H., Sarnthein, M., Grootes, P. M., Kiefer, T., Kuehn, H., Schmieder, F., and Röhl, U.: Paleonutrient and productivity records from the subarctic North Pacific for Pleistocene glacial terminations I to V, Paleoceanography, 23, PA4212, https://doi.org/10.1029/2007PA001513, 2008.

Gray, W. R., Rae, J. W. B., Wills, R. C. J., Shevenell, A. E., Taylor, B., Burke, A., Foster, G. L., and Lear, C. H.: Deglacial upwelling, productivity and CO_2 outgassing in the North Pacific Ocean, Nat. Geosci., 11, 340–344, https://doi.org/10.1038/s41561-018-0108-6, 2018.

Hargreaves, J. C., Annan, J. D., Ohgaito, R., Paul, A., and Abe-Ouchi, A.: Skill and reliability of climate model ensembles at the Last Glacial Maximum and mid-Holocene, Clim. Past, 9, 811–823, https://doi.org/10.5194/cp-9-811-2013, 2013.

Harrison, S. P., Bartlein, P. J., Brewer, S., Prentice, I. C., Boyd, M., Hessler, I., Holmgren, K., Izumi, K., and Willis, K.: Climate model benchmarking with glacial and mid-Holocene climates, Clim. Dynam., 43, 671–688, https://doi.org/10.1007/s00382-013-1922-6, 2014.

He, C., Liu, Z., Otto-Bliesner, B. L., Brady, E., Zhu, C., Tomas, R., Clark, P., Zhu, J., Jahn, A., Gu, S., Zhang, J., Nusbaumer, J., Noone, D., Cheng, H., Wang, Y., Yan, M., and Bao, Y.: Hydroclimate footprint of pan-Asian monsoon water isotope during the last deglaciation, Sci. Adv., 7, eabe2611, https://doi.org/10.1126/sciadv.abe2611, 2021.

He, F. and Clark, P. U.: Freshwater forcing of the Atlantic Meridional Overturning Circulation revisited, Nat. Clim. Change, 12, 449–454, https://doi.org/10.1038/s41558-022-01328-2, 2022.

Herzschuh, U., Böhmer, T., Li, C., Chevalier, M., Hébert, R., Dallmeyer, A., Cao, X., Bigelow, N. H., Nazarova, L., Novenko, E. Y., Park, J., Peyron, O., Rudaya, N. A., Schlütz, F., Shumilovskikh, L. S., Tarasov, P. E., Wang, Y., Wen, R., Xu, Q., and Zheng, Z.: LegacyClimate 1.0: a dataset of pollen-based climate reconstructions from 2594 Northern Hemisphere sites covering the last 30 kyr and beyond, Earth Syst. Sci. Data, 15, 2235–2258, https://doi.org/10.5194/essd-15-2235-2023, 2023.

Huang, E., Chen, Y., Schefuß, E., Steinke, S., Liu, J., Tian, J., Martínez-Méndez, G., and Mohtadi, M.: Precession and glacial-cycle controls of monsoon precipitation isotope changes over East Asia during the Pleistocene, Earth Planet. Sc. Lett., 494, 1–11, https://doi.org/10.1016/j.epsl.2018.04.046, 2018.

Hüls, M. and Zahn, R.: Millennial-scale sea surface temperature variability in the western tropical North Atlantic from planktonic foraminiferal census counts, Paleoceanography, 15, 659–678, https://doi.org/10.1029/1999PA000462, 2000.

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, Geosci. Model Dev., 9, 2563–2587, https://doi.org/10.5194/gmd-9-2563-2016, 2016.

Johnstone, H. J. H., Kiefer, T., Elderfield, H., and Schulz, M.: Calcite saturation, foraminiferal test mass, and Mg / Ca-based temperatures dissolution corrected using XDX-A 150 ka record from the western Indian Ocean, Geochem. Geophy. Geosy., 15, 781–797, https://doi.org/10.1002/2013GC004994, 2014.

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

Jonkers, L. and Kučera, M.: Sensitivity to species selection indicates the effect of nuisance variables on marine microfossil transfer functions, Clim. Past, 15, 881–891, https://doi.org/10.5194/cp-15-881-2019, 2019.

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 Syst. Sci. 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 version 1.1.1, Zenodo [data set], https://doi.org/10.5281/zenodo.7785766, 2023.

Judd, E. J., Bhattacharya, T., and Ivany, L. C.: A Dynamical Framework for Interpreting Ancient Sea Surface Temperatures, Geophys. Res. Lett., 47, e2020GL089044, https://doi.org/10.1029/2020GL089044, 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, Clim. Past, 17, 1065–1089, https://doi.org/10.5194/cp-17-1065-2021, 2021.

Kapsch, M., 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, Geophys. Res. Lett., 49, e2021GL096767, https://doi.org/10.1029/2021GL096767, 2022.

Kiefer, T.: Produktivität und Temperaturen im subtropischen Nordatlantik: zyklische und abrupte Veränderungen im späten Quartär, Tech. rep., Geologisch-Paläontologisches Institut und Museum, Christian-Albrechts-Universität, Kiel, https://doi.org/10.2312/REPORTS-GPI.1998.90, 1998.

Kiefer, T., McCave, I. N., and Elderfield, H.: Antarctic control on tropical Indian Ocean sea surface temperature and hydrography, Geophys. Res. Lett., 33, L24612, https://doi.org/10.1029/2006GL027097, 2006.

Kirst, G. J., Schneider, R. R., Müller, P. J., von Storch, I., and Wefer, G.: Late Quaternary Temperature Variability in the Benguela Current System Derived from Alkenones, Quaternary Res., 52, 92–103, https://doi.org/10.1006/qres.1999.2040, 1999.

Kleinen, T., Gromov, S., Steil, B., and Brovkin, V.: Atmospheric methane since the last glacial maximum was driven by wetland sources, Clim. Past, 19, 1081–1099, https://doi.org/10.5194/cp-19-1081-2023, 2023a.

Kleinen, T., Gromov, S., Steil, B., and Brovkin, V.: PalMod2 MPI-M MPI-ESM1-2-CR-CH4 transient-deglaciation-prescribed-glac1d-methane, World Data Center for Climate (WDCC) at DKRZ [data set], https://doi.org/10.26050/WDCC/PMMXMCHTD, 2023b.

Kretschmer, K., Jonkers, L., Kucera, M., and Schulz, M.: Modeling seasonal and vertical habitats of planktonic foraminifera on a global scale, Biogeosciences, 15, 4405–4429, https://doi.org/10.5194/bg-15-4405-2018, 2018.

Kucera, M., Weinelt, M., Kiefer, T., Pflaumann, U., Hayes, A., Weinelt, M., Chen, M.-T., Mix, A. C., Barrows, T. T., Cortijo, E., Duprat, J., Juggins, S., and Waelbroeck, C.: Reconstruction of sea-surface temperatures from assemblages of planktonic foraminifera: multi-technique approach based on geographically constrained calibration data sets and its application to glacial Atlantic and Pacific Oceans, Quaternary Sci. Rev., 24, 951–998, https://doi.org/10.1016/j.quascirev.2004.07.014, 2005.

Kwon, Y.-O., Alexander, M. A., Bond, N. A., Frankignoul, C., Nakamura, H., Qiu, B., and Thompson, L. A.: Role of the Gulf Stream and Kuroshio–Oyashio Systems in Large-Scale Atmosphere–Ocean Interaction: A Review, J. Climate, 23, 3249–3281, https://doi.org/10.1175/2010JCLI3343.1, 2010.

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 Syst. Sci. Data, 9, 363–387, https://doi.org/10.5194/essd-9-363-2017, 2017.

Labeyrie, L., Labracherie, M., Gorfti, N., Pichon, J. J., Vautravers, M., Arnold, M., Duplessy, J.-C., Paterne, M., Michel, E., Duprat, J., Caralp, M., and Turon, J.-L.: Hydrographic changes of the Southern Ocean (southeast Indian Sector) Over the last 230 kyr, Paleoceanography, 11, 57–76, https://doi.org/10.1029/95PA02255, 1996.

Laepple, T. and Huybers, P.: Ocean surface temperature variability: Large model–data differences at decadal and longer periods, P. Natl. Acad. Sci. USA, 111, 16682–16687, https://doi.org/10.1073/pnas.1412077111, 2014.

Lambeck, K., Rouby, H., Purcell, A., Sun, Y., and Sambridge, M.: Sea level and global ice volumes from the Last Glacial Maximum to the Holocene, P. Natl. Acad. Sci. USA, 111, 15296–15303, https://doi.org/10.1073/pnas.1411762111, 2014.

Lauterbach, S., Andersen, N., Wang, Y. V., Blanz, T., Larsen, T., and Schneider, R. R.: An $\sim$ 130 kyr Record of Surface Water Temperature and $\delta^{18}O$ From the Northern Bay of Bengal: Investigating the Linkage Between Heinrich Events and Weak Monsoon Intervals in Asia, Paleoceanography and Paleoclimatology, 35, e2019PA003646, https://doi.org/10.1029/2019PA003646, 2020.

Lea, D. W., Pak, D. K., Belanger, C. L., Spero, H. J., Hall, M. A., and Shackleton, N. J.: Paleoclimate history of Galápagos surface

waters over the last 135,000 yr, Quaternary Sci. Rev., 25, 1152–1167, https://doi.org/10.1016/j.quascirev.2005.11.010, 2006.

Lenton, T.: QUEST Quaternary: FAMOUS glacial cycle model data, NCAS British Atmospheric Data Centre [data set], https://catalogue.ceda.ac.uk/uuid/a43dcfaccfae4824ab9ab2b572703e72 (last access: 28 February 2023), 2008.

Liu, Z., Otto-Bliesner, B. L., He, F., Brady, E. C., Tomas, R., Clark, P. U., Carlson, A. E., Lynch-Stieglitz, J., Curry, W., Brook, E., Erickson, D., Jacob, R., Kutzbach, J., and Cheng, J.: Transient Simulation of Last Deglaciation with a New Mechanism for Bølling-Allerød Warming, Science, 325, 310–314, https://doi.org/10.1126/science.1171041, 2009.

Love, R., Andres, H. J., Condron, A., and Tarasov, L.: Freshwater routing in eddy-permitting simulations of the last deglacial: the impact of realistic freshwater discharge, Clim. Past, 17, 2327–2341, https://doi.org/10.5194/cp-17-2327-2021, 2021.

Ma, X., Jing, Z., Chang, P., Liu, X., Montuoro, R., Small, R. J., Bryan, F. O., Greatbatch, R. J., Brandt, P., Wu, D., Lin, X., and Wu, L.: Western boundary currents regulated by interaction between ocean eddies and the atmosphere, Nature, 535, 533–537, https://doi.org/10.1038/nature18640, 2016.

MARGO Project Members: Constraints on the magnitude and patterns of ocean cooling at the Last Glacial Maximum, Nat. Geosci., 2, 127–132, https://doi.org/10.1038/ngeo411, 2009.

Maslin, M. A., Shackleton, N. J., and Pflaumann, U.: Surface water temperature, salinity, and density changes in the northeast Atlantic during the last 45,000 years: Heinrich events, deep water formation, and climatic rebounds, Paleoceanography, 10, 527–544, https://doi.org/10.1029/94PA03040, 1995.

Menviel, L., Timmermann, A., Timm, O. E., and Mouchet, A.: Deconstructing the Last Glacial termination: the role of millennial and orbital-scale forcings, Quaternary Sci. Rev., 30, 1155–1172, https://doi.org/10.1016/j.quascirev.2011.02.005, 2011.

Mikolajewicz, U., Kapsch, M.-L., Gayler, V., Meccia, V. L., Riddick, T., Ziemen, F. A., and Schannwell, C.: PalMod2 MPI-M MPI-ESM1-2-CR Transient Simulations of the Last Deglaciation with prescribed ice sheets from GLAC-1D reconstructions (r1i1p3f2), World Data Center for Climate (WDCC) at DKRZ [data set], https://doi.org/10.26050/WDCC/PMMXMCRTDGP132, 2023a.

Mikolajewicz, U., Kapsch, M.-L., Gayler, V., Meccia, V. L., Riddick, T., Ziemen, F. A., and Schannwell, C.: PalMod2 MPI-M MPI-ESM1-2-CR Transient Simulations of the Last Deglaciation with prescribed ice sheets from ICE-6G reconstructions (r1i1p3f2), World Data Center for Climate (WDCC) at DKRZ [data set], https://doi.org/10.26050/WDCC/PMMXMCRTDIP132, 2023b.

Mikolajewicz, U., Kapsch, M.-L., Gayler, V., Meccia, V. L., Riddick, T., Ziemen, F. A., and Schannwell, C.: PalMod2 MPI-M MPI-ESM1-2-CR Transient Simulations of the Last Deglaciation with prescribed ice sheets from ICE-6G reconstructions (r1i1p2f2), World Data Center for Climate (WDCC) at DKRZ [data set], https://doi.org/10.26050/WDCC/PMMXMCRTDIP122, 2023c.

Niedermeyer, E. M., Prange, M., Mulitza, S., Mollenhauer, G., Schefuß, E., and Schulz, M.: Extratropical forcing of Sahel aridity during Heinrich stadials, Geophys. Res. Lett., 36, L20707, https://doi.org/10.1029/2009GL039687, 2009.

Nürnberg, D., Böschen, T., Doering, K., Mollier-Vogel, E., Raddatz, J., and Schneider, R.: Sea surface and subsurface circulation dynamics off equatorial Peru during the last ∼ 17 kyr, Paleoceanography, 30, 984–999, https://doi.org/10.1002/2014PA002706, 2015.

Obase, T. and Abe-Ouchi, A.: Abrupt Bølling-Allerød Warming Simulated under Gradual Forcing of the Last Deglaciation, Geophys. Res. Lett., 46, 11397–11405, https://doi.org/10.1029/2019GL084675, 2019.

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, Nat. Geosci., 12, 643–649, https://doi.org/10.1038/s41561-019-0400-0, 2019.

PAGES 2k-PMIP3 group: Continental-scale temperature variability in PMIP3 simulations and PAGES 2k regional temperature reconstructions over the past millennium, Clim. Past, 11, 1673–1699, https://doi.org/10.5194/cp-11-1673-2015, 2015.

Pailler, D. and Bard, E.: High frequency palaeoceanographic changes during the past 140 000 yr recorded by the organic matter in sediments of the Iberian Margin, Palaeogeography, Palaeoclimatology, Palaeoecology, 181, 431–452, https://doi.org/10.1016/S0031-0182(01)00444-8, 2002.

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

Pedro, J., Andersson, C., Vettoretti, G., Voelker, A., Waelbroeck, C., Dokken, T., Jensen, M., Rasmussen, S., Sessford, E., Jochum, M., and Nisancioglu, K.: Dansgaard-Oeschger and Heinrich event temperature anomalies in the North Atlantic set by sea ice, frontal position and thermocline structure, Quaternary Sci. Rev., 289, 107599, https://doi.org/10.1016/j.quascirev.2022.107599, 2022.

Pelejero, C., Grimalt, J. O., Heilig, S., Kienast, M., and Wang, L.: High-resolution U^{K}_{37} temperature reconstructions in the South China Sea over the past 220 kyr, Paleoceanography, 14, 224–231, https://doi.org/10.1029/1998PA900015, 1999.

Peltier, W. R., Argus, D. F., and Drummond, R.: Space geodesy constrains ice age terminal deglaciation: The global ICE-6G_C (VM5a) model: Global Glacial Isostatic Adjustment, J. Geophys. Res.-Sol. Ea., 120, 450–487, https://doi.org/10.1002/2014JB011176, 2015.

Rebotim, A., Voelker, A. H. L., Jonkers, L., Waniek, J. J., Meggers, H., Schiebel, R., Fraile, I., Schulz, M., and Kucera, M.: Factors controlling the depth habitat of planktonic foraminifera in the subtropical eastern North Atlantic, Biogeosciences, 14, 827–859, https://doi.org/10.5194/bg-14-827-2017, 2017.

Rehfeld, K., Marwan, N., Heitzig, J., and Kurths, J.: Comparison of correlation analysis techniques for irregularly sampled time series, Nonlin. Processes Geophys., 18, 389–404, https://doi.org/10.5194/npg-18-389-2011, 2011.

Reschke, M., Rehfeld, K., and Laepple, T.: Empirical estimate of the signal content of Holocene temperature proxy records, Clim. Past, 15, 521–537, https://doi.org/10.5194/cp-15-521-2019, 2019.

Riddick, T., Brovkin, V., Hagemann, S., and Mikolajewicz, U.: Dynamic hydrological discharge modelling for coupled climate model simulations of the last glacial cycle: the MPI-DynamicHD model version 3.0, Geosci. Model Dev., 11, 4291–4316, https://doi.org/10.5194/gmd-11-4291-2018, 2018.

Riethdorf, J.-R., Max, L., Nürnberg, D., Lembke-Jene, L., and Tiedemann, R.: Deglacial development of (sub) sea surface temperature and salinity in the subarctic northwest Pacific: Implications for upper-ocean stratification, Paleoceanography, 28, 91–104, https://doi.org/10.1002/palo.20014, 2013.

Roberts, J., Gottschalk, J., Skinner, L. C., Peck, V. L., Kender, S., Elderfield, H., Waelbroeck, C., Vázquez Riveiros, N., and Hodell, D. A.: Evolution of South Atlantic density and chemical stratification across the last deglaciation, P. Natl. Acad. Sci. USA, 113, 514–519, https://doi.org/10.1073/pnas.1511252113, 2016.

Roberts, J., McCave, I., McClymont, E., Kender, S., Hillenbrand, C.-D., Matano, R., Hodell, D., and Peck, V.: Deglacial changes in flow and frontal structure through the Drake Passage, Earth Planet. Sc. Lett., 474, 397–408, https://doi.org/10.1016/j.epsl.2017.07.004, 2017.

Romahn, S., Mackensen, A., Groeneveld, J., and Pätzold, J.: Deglacial intermediate water reorganization: new evidence from the Indian Ocean, Clim. Past, 10, 293–303, https://doi.org/10.5194/cp-10-293-2014, 2014.

Rühlemann, C., Mulitza, S., Müller, P. J., Wefer, G., and Zahn, R.: Warming of the tropical Atlantic Ocean and slowdown of thermohaline circulation during the last deglaciation, Nature, 402, 511–514, https://doi.org/10.1038/990069, 1999.

Salgueiro, E., Naughton, F., Voelker, A., de Abreu, L., Alberto, A., Rossignol, L., Duprat, J., Magalhães, V., Vaqueiro, S., Turon, J.-L., and Abrantes, F.: Past circulation along the western Iberian margin: a time slice vision from the Last Glacial to the Holocene, Quaternary Sci. Rev., 106, 316–329, https://doi.org/10.1016/j.quascirev.2014.09.001, 2014.

Samson, C. R., Sikes, E. L., and Howard, W. R.: Deglacial paleoceanographic history of the Bay of Plenty, New Zealand, Paleoceanography, 20, PA4017, https://doi.org/10.1029/2004PA001088, 2005.

Santos, T. P., Lessa, D. O., Venancio, I. M., Chiessi, C. M., Mulitza, S., Kuhnert, H., Govin, A., Machado, T., Costa, K. B., Toledo, F., Dias, B. B., and Albuquerque, A. L. S.: Prolonged warming of the Brazil Current precedes deglaciations, Earth Planet. Sc. Lett., 463, 1–12, https://doi.org/10.1016/j.epsl.2017.01.014, 2017.

Schlung, S. A., Christina Ravelo, A., Aiello, I. W., Andreasen, D. H., Cook, M. S., Drake, M., Dyez, K. A., Guilderson, T. P., LaRiviere, J. P., Stroynowski, Z., and Takahashi, K.: Millennial-scale climate change and intermediate water circulation in the Bering Sea from 90 ka: A high-resolution record from IODP Site U1340, Paleoceanography, 28, 54–67, https://doi.org/10.1029/2012PA002365, 2013.

Schröder, J. F., Holbourn, A., Kuhnt, W., and Küssner, K.: Variations in sea surface hydrology in the southern Makassar Strait over the past 26 kyr, Quaternary Sci. Rev., 154, 143–156, https://doi.org/10.1016/j.quascirev.2016.10.018, 2016.

Schröder, J. F., Kuhnt, W., Holbourn, A., Beil, S., Zhang, P., Hendrizan, M., and Xu, J.: Deglacial Warming and Hydroclimate Variability in the Central Indonesian Archipelago, Paleoceanography and Paleoclimatology, 33, 974–993, https://doi.org/10.1029/2018PA003323, 2018.

Schulz, H.: Meeresoberflächentemperaturen vor 10.000 Jahren – Auswirkungen des frühholozänen Insolationsmaximums, Tech. rep., Geologisch-Paläontologisches Institut und Museum, Christian-Albrechts-Universität, Kiel, https://doi.org/10.2312/REPORTS-GPI.1995.73, 1995.

Seager, R., Murtugudde, R., Naik, N., Clement, A., Gordon, N., and Miller, J.: Air–Sea Interaction and the Seasonal Cycle of the Subtropical Anticyclones, J. Climate, 16, 1948–1966, https://doi.org/10.1175/1520-0442(2003)016<1948:AIATSC>2.0.CO;2, 2003.

Sikes, E. L., Howard, W. R., Samson, C. R., Mahan, T. S., Robertson, L. G., and Volkman, J. K.: Southern Ocean seasonal temperature and Subtropical Front movement on the South Tasman Rise in the late Quaternary, Paleoceanography, 24, PA2201, https://doi.org/10.1029/2008PA001659, 2009.

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

Stokes, C. R., Tarasov, L., Blomdin, R., Cronin, T. M., Fisher, T. G., Gyllencreutz, R., Hättestrand, C., Heyman, J., Hindmarsh, R. C., Hughes, A. L., Jakobsson, M., Kirchner, N., Livingstone, S. J., Margold, M., Murton, J. B., Noormets, R., Peltier, W. R., Peteet, D. M., Piper, D. J., Preusser, F., Renssen, H., Roberts, D. H., Roche, D. M., Saint-Ange, F., Stroeven, A. P., and Teller, J. T.: On the reconstruction of palaeo-ice sheets: Recent advances and future challenges, Quaternary Sci. Rev., 125, 15–49, https://doi.org/10.1016/j.quascirev.2015.07.016, 2015.

Stott, L., Poulsen, C., Lund, S., and Thunell, R.: Super ENSO and Global Climate Oscillations at Millennial Time Scales, Science, 297, 222–226, https://doi.org/10.1126/science.1071627, 2002.

Stott, L., Timmermann, A., and Thunell, R.: Southern Hemisphere and Deep-Sea Warming Led Deglacial Atmospheric CO_2 Rise and Tropical Warming, Science, 318, 435–438, https://doi.org/10.1126/science.1143791, 2007.

Thorarinsdottir, T. L., Gneiting, T., and Gissibl, N.: Using Proper Divergence Functions to Evaluate Climate Models, SIAM/ASA J. Uncertainty Quantification, 1, 522–534, https://doi.org/10.1137/130907550, 2013.

Thornalley, D. J., Elderfield, H., and McCave, I. N.: Reconstructing North Atlantic deglacial surface hydrography and its link to the Atlantic overturning circulation, Global Planet. Change, 79, 163–175, https://doi.org/10.1016/j.gloplacha.2010.06.003, 2011.

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.

Tingley, M. P., Craigmile, P. F., Haran, M., Li, B., Mannshardt, E., and Rajaratnam, B.: Piecing together the past: statistical insights into paleoclimatic reconstructions, Quaternary Sci. Rev., 35, 1–22, https://doi.org/10.1016/j.quascirev.2012.01.012, 2012.

Vettoretti, G., Ditlevsen, P., Jochum, M., and Rasmussen, S. O.: Atmospheric CO_2 control of spontaneous millennial-scale ice age climate oscillations, Nat. Geosci., 15, 300–306, https://doi.org/10.1038/s41561-022-00920-7, 2022.

Vogelsang, E., Sarnthein, M., and Pflaumann, U.: d18O Stratigraphy, chronology, and sea surface temperatures of Atlantic sediment records (GLAMAP-2000 Kiel), Tech. rep., Institut für Geowissenschaften, Christian-Albrechts-Universität, Kiel, https://doi.org/10.2312/REPORTS-IFG.2001.13, 2001.

von Storch, H., Zorita, E., Jones, J. M., Dimitriev, Y., González-Rouco, F., and Tett, S. F. B.: Reconstructing Past Climate from Noisy Data, Science, 306, 679–682, https://doi.org/10.1126/science.1096109, 2004.

Waelbroeck, C., Labeyrie, L., Duplessy, J.-C., Guiot, J., Labracherie, M., Leclaire, H., and Duprat, J.: Improving past sea surface temperature estimates based on planktonic fossil faunas, Paleoceanography, 13, 272–283, https://doi.org/10.1029/98PA00071, 1998.

Weinelt, M., Rosell-Melé, A., Pflaumann, U., Sarnthein, M., and Kiefer, T.: The role of productivity in the Northeast Atlantic on abrupt climate change over the last 80,000 years, zdgg_alt, Zeitschrift der Deutschen Geologischen Gesellschaft, 154, 47–66, https://doi.org/10.1127/zdgg/154/2003/47, 2003.

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.: Code in support of "Towards spatio-temporal comparison of simulated and reconstructed sea surface temperatures for the last deglaciation", Zenodo [code], https://doi.org/10.5281/zenodo.10497834, 2024.

Xu, J., Kuhnt, W., Holbourn, A., Andersen, N., and Bartoli, G.: Changes in the vertical profile of the Indonesian Throughflow during Termination II: Evidence from the Timor Sea, Paleoceanography, 21, PA4202, https://doi.org/10.1029/2006PA001278, 2006.

Xu, J., Holbourn, A., Kuhnt, W., Jian, Z., and Kawamura, H.: Changes in the thermocline structure of the Indonesian outflow during Terminations I and II, Earth Planet. Sc. Lett., 273, 152–162, https://doi.org/10.1016/j.epsl.2008.06.029, 2008.

Zarriess, M., Johnstone, H., Prange, M., Steph, S., Groeneveld, J., Mulitza, S., and Mackensen, A.: Bipolar seesaw in the northeastern tropical Atlantic during Heinrich stadials, Geophys. Res. Lett., 38, L04706, https://doi.org/10.1029/2010GL046070, 2011.

Zhao, M., Beveridge, N. A. S., Shackleton, N. J., Sarnthein, M., and Eglinton, G.: Molecular stratigraphy of cores off northwest Africa: Sea surface temperature history over the last 80 Ka, Paleoceanography, 10, 661–675, https://doi.org/10.1029/94PA03354, 1995.

Ziegler, M., Nürnberg, D., Karas, C., Tiedemann, R., and Lourens, L. J.: Persistent summer expansion of the Atlantic Warm Pool during glacial abrupt cold events, Nat. Geosci., 1, 601–605, https://doi.org/10.1038/ngeo277, 2008.

Ziemen, F. A., Kapsch, M.-L., Klockmann, M., and Mikolajewicz, U.: Heinrich events show two-stage climate response in transient glacial simulations, Clim. Past, 15, 153–168, https://doi.org/10.5194/cp-15-153-2019, 2019.