BONNER METEOROLOGISCHE ABHANDLUNGEN

Heft 89 (2020) (ISSN 0006-7156)
Herausgeber: Andreas Hense

Nils Weitzel

CLIMATE FIELD RECONSTRUCTIONS FROM POLLEN AND MACROFOSSIL SYNTHESES USING BAYESIAN HIERARCHICAL MODELS

BONNER METEOROLOGISCHE ABHANDLUNGEN

Heft 89 (2020) (ISSN 0006-7156)
Herausgeber: Andreas Hense

Nils Weitzel

CLIMATE FIELD RECONSTRUCTIONS FROM POLLEN AND MACROFOSSIL SYNTHESES USING BAYESIAN HIERARCHICAL MODELS

Climate field reconstructions from pollen and macrofossil syntheses using Bayesian hierarchical models

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
M.Sc.
Nils Weitzel
aus
Kassel

Bonn, Mai, 2019

Diese Arbeit ist die ungekürzte Fassung einer der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn im Jahr 2019 vorgelegten Dissertation von Nils Weitzel aus Kassel.

This paper is the unabridged version of a dissertation thesis submitted by Nils Weitzel born in Kassel to the Faculty of Mathematical and Natural Sciences of the Rheinische Friedrich-Wilhelms-Universität Bonn in 2019.

Anschrift des Verfassers: Address of the author:

Nils Weitzel
Institut für Geowissenschaften, Abt. Meteorologie
Universität Bonn
Auf dem Hügel 20
D-53121 Bonn

1. Gutachter: Prof. Dr. Andreas Hense,
Rheinische Friedrich-Wilhelms-Universität Bonn
2. Gutachter: Prof. Dr. Thomas Litt,
Rheinische Friedrich-Wilhelms-Universität Bonn

Tag der Promotion: 28. Oktober 2019

Abstract

Studying past climate states is important to understand climate changes and climate variability on centennial to orbital timescales, to analyze the reaction of the Earth system to large-scale changes in external forcings, and to identify physical and biogeochemical processes that drive these changes. It improves not just the understanding of the climate system but can also lead to more accurate projections of future climate conditions. As the instrumental record is restricted to approximately the last 150 years, paleoclimatology has to rely on indirect observations of climate variables, so-called proxy data. Examples are isotope compositions in ice cores, pollen counts from lake sediment records, and geochemical indices measured in marine sediment cores. In addition, numerical Earth system models can run simulations with adjusted boundary conditions to test their ability to reproduce past climate states and to study mechanisms which control them.

Probabilistic spatial or spatio-temporal reconstructions of past climate states, known as climate field reconstructions, are an important tool to quantitatively study the climate system under different forcing conditions because they combine the information contained in a proxy synthesis in a comprehensible product. Unfortunately, they are subject to a complex uncertainty structure due to complicated proxy-climate relations and sparse data, which makes interpolation between samples difficult. Therefore, advanced statistical methods are required which are robust under sparse and noisy data. In this thesis, Bayesian hierarchical models are developed for spatial and spatio-temporal reconstructions from terrestrial proxy networks. The focus is on pollen and macrofossil records, which provide information on the past vegetation composition and indirectly on past temperature and precipitation distributions. Bayesian hierarchical models feature promising properties for climate field reconstructions like the possibility to include multiple sources of information and to quantify uncertainties in a statistically rigorous way. To interpolate between the proxy samples, this study combines geostatistical and data assimilation methods. While data assimilation techniques facilitate the use of physically consistent estimates of past climate states on regional scales provided by Earth system models, geostatistical methods are required to account for the small number of available state-of-the-art simulations and potentially large biases in the produced climate states. Bayesian inference is performed using Markov chain Monte Carlo methods following a Metropolis-within-Gibbs strategy. The Bayesian frameworks produce spatially or spatio-temporally distributed probability distributions that facilitate quantitative analyses which account for uncertainties.

The first application of this study is a reconstruction of European summer and winter temperature during the mid-Holocene using a published pollen and macrofossil synthesis in combination

with a multi-model climate simulation ensemble from the Paleoclimate Modelling Intercomparison Project Phase III. To transfer the pollen and macrofossil data into climate information, a forward version of the probabilistic indicator taxa method is applied. Different ways to incorporate the climate simulations in the Bayesian reconstruction framework are compared using identical twin and cross-validation experiments. The spatial reconstruction features dipole structures with warming in Northern Europe and cooling in Southern Europe in concordance with previous results from the literature. The reconstruction performs well in cross-validation experiments and exhibits a reasonable degree of spatial smoothing.

In a second application of the spatial reconstruction framework, summer temperature and mean annual precipitation during the Last Glacial Maximum in Siberia are reconstructed. A compilation of local reconstructions from pollen samples, provided by the Polar Terrestrial Environmental Systems Division at the Alfred-Wegener-Institute, is combined with the Last Glacial Maximum multi-model ensemble from the Paleoclimate Modelling Intercomparison Project Phase III. The reconstruction features a strong summer cooling in the mid latitudes but only moderate cooling in high latitudes perhaps due to the impact of lower sea levels. Our findings provide new insights into explanations for the absence of a Siberian ice sheet during the last Glacial.

To understand the climate evolution from the Last Glacial Maximum to the mid-Holocene, this study develops a data-driven Bayesian hierarchical model for reconstructing the spatio-temporal temperature evolution during the Last Deglaciation on continental scales. Eurasia is chosen as reconstruction domain because more proxy records are available compared to other continents. The Last Deglaciation features millennial-scale trends with a shift from Glacial to Interglacial climate conditions and additional abrupt climate changes. Therefore, a statistical model is required which can recover temporal and spatial non-stationarities. The Bayesian hierarchical model is tested in a controlled environment using pseudo-proxy experiments with a reference climate simulation. These experiments show that the model is recovering non-linear millennial-scale trends with high accuracy and abrupt climate changes are detected even though the magnitude of events tends to be slightly underestimated. Thus, the Bayesian hierarchical model is well-suited for future applications with new and existing proxy syntheses.

Contents

List of Figures

List of Tables

List of Abbreviations

AMOC Atlantic meridional overturning circulation

BA Bølling-Allerød

BHM Bayesian hierarchical model

BSS Brier skill score

cdf Cumulative distribution function

CFR Climate field reconstruction

CMIP Coupled Model Intercomparison Project

CRPS Continuous ranked probability score

CRPSS Continuous ranked probability skill score

CRU University of East Anglia Climatic Research Unit

CVE Cross-validation experiment

DA Data assimilation

DAG Directed acyclic graph

DKRZ German Climate Computing Center

ESGF Earth System Grid Federation

ESM Earth system model

GM Gaussian model

GMRF Gaussian Markov random field

iid independent and identically distributed

INLA Integrated nested Laplace approximation

ITE Identical twin experiment

ITM Indicator taxa method

KM Kernel model

LD Last Deglaciation

LGM Last Glacial Maximum

MAT Modern analogue technique

MCMC Markov chain Monte Carlo

MH Mid-Holocene

MTCO Mean temperature of the coldest month

MTWA Mean temperature of the warmest month

P_{ANN} Annual precipitation

PFT Plant functional type

PITM Probabilistic indicator taxa method

PMIP Paleoclimate Modelling Intercomparison Project

PPE Pseudo-proxy experiment

RM Regression model

RMSE Root mean square error

RMSEP Root mean square error of prediction

RW-MH Random walk Metropolis-Hastings

SPDE Stochastic partial differential equation

SST Sea surface temperature

TS Mean annual surface temperature

WA Weighted averaging

WAPLS Weighted averaging partial least squares

YD Younger Dryas

1 Introduction

The study of the climate system from instrumental records is restricted to observations from approximately the last 150 years. However, significantly larger changes of the climate state occurred prior to the instrumental period, typically defined as starting in 1880 CE. These changes were driven by variations in external forcings like the distribution of incoming solar radiation (e.g. Berger, 1978) and volcanic eruptions (e.g. Gao et al., 2008), as well as internal variability and feedback in the Earth system. Effects range from atmospheric circulation variations (e.g. Cheng et al., 2012), over changes in ocean circulation (e.g. Barker et al., 2010) and variations in the carbon cycle leading to changed greenhouse gas concentrations (e.g. Köhler et al., 2017), to the buildup and melting of ice sheets which lead to sea level and topography changes (e.g. Hughes et al., 2015).

In this study, the focus is on three time periods: The mid-Holocene (MH) around 6,000 years before present, the Last Glacial Maximum (LGM) around 21,000 years before present, and the transient evolution between those two time slices which we loosely refer to as the Last Deglaciation (LD), even though strictly speaking the deglaciation is just a subinterval of that period. In convention with the paleoclimatology literature, 'present' is defined as the year 1950 CE, and paleoclimate dates are denoted by ka (thousands of years before present), such that the MH was around 6ka and the LGM around 21ka.

The MH is part of the current interglacial period, the Holocene, but due to changes in orbital parameter configurations the seasonal and latitudinal insolation distribution was different from today (Otto-Bliesner et al., 2017). For example, the summer insolation and seasonality on the Northern Hemisphere was stronger than today (Fig. 1.1c). While proxy-based reconstructions indicate a global thermal maximum around the MH (Fig. 1.1d, Marcott et al., 2013) with subsequent cooling during the late Holocene, state-of-the-art Earth system model (ESM) simulations do not reproduce this maximum (Liu et al., 2014). In addition, greenhouse gas concentrations have been lower than today (Fig. 1.1e). These differences led to changes in the energy and moisture budget, prominently featured for example by a green Sahara (Jolly et al., 1998). Additional changes in large-scale atmospheric circulation are currently a major area of research (Mauri et al., 2014). Studying the MH provides important insights into internal climate feedback in response to changes in radiative forcing.

As part of the last Glacial period (around 115ka to 11.7ka, Fig. 1.1a), the LGM is defined as the last period of maximum global ice volume and near minimal eustatic sea level (approximately 115 to 130 m below the present level, Lambeck et al., 2014). While the orbital configuration was similar

to today (Fig. 1.1c), greenhouse gas concentrations were lower (Fig. 1.1e), atmospheric aerosol concentrations were higher, and large ice sheets over many continents led to a lower sea level and substantial changes in the land-sea mask (Kageyama et al., 2017). For example, the Bering Strait was closed, the Hudson Bay and the Baltic Sea were covered by ice sheets, the North Sea was a steppe-like area, and the nowadays shallow sea areas north of Siberia were above sea level. The global mean climate was cooler and dryer than today (Fig. 1.1b,d, Shakun et al., 2012), but the magnitude of changes varied in space (Bartlein et al., 2011; MARGO Project Members, 2009). In addition, Rehfeld et al. (2018) showed that centennial scale climate variability was higher than during the Holocene by comparing periods of equal length during the Holocene and the LGM, and proxy series covering both periods with similar sampling intervals (Fig. 1.1b). These substantial changes in multiple components of the Earth system make the LGM an interesting period for paleoclimate research to understand the mechanisms behind regional change patterns and the state dependency of climate variability as well as its governing processes. The difference in radiative forcing between the LGM and today is used in many studies to estimate climate sensitivity which is a popular quantity to summarize global mean temperature rise under increased CO_2 concentrations (e.g. Schmittner et al., 2011; Annan and Hargreaves, 2013; Harrison et al., 2015).

Between the LGM and the MH, a several millennia long climate change occurred in the form of the LD from approximately 19ka to 11ka. This period is characterized by a warming trend (Fig. 1.1d) accompanied by an increase of CO_2 concentrations (Fig. 1.1f), melting of ice sheets, and a subsequent sea level rise (Clark et al., 2012). Two abrupt events interrupt the relatively steady increase of global mean temperature, the Bølling-Allerød (BA) warming (approximately 14.7ka to 12.9ka) and the Younger Dryas (YD) cooling (approximately 12.9ka to 11.7ka), which are prominently seen in Greenland ice cores (Fig. 1.1b, North Greenland Ice Core Project members, 2004, Fig. 1.1b) and pollen records from the Eifel maars (Brauer et al., 1999). The mechanisms, that govern the deglacial evolution and especially regional change patterns, are an active research area (Ivanovic et al., 2016). Much effort is put for example into understanding the temperature - CO_2 relationship (Shakun et al., 2012) and the influence of freshwater fluxes on the Atlantic meridional overturning circulation (AMOC) (Clark et al., 2012; Liu et al., 2009b). Improved data-driven reconstructions of the spatio-temporal climate evolution will help to test statistical hypotheses on the influence of these processes.

To study paleoclimate phenomena, scientists rely on indirect observations, so-called paleodata, as indicators of past climate states. This data is often referred to as proxy data. Examples of popular proxy types are isotope compositions in ice cores from Greenland and Antarctica, pollen assemblages in lake sediments and peat bogs, and geochemical indices as well as Foraminifera assemblages in marine sediments. The cost of collecting these types of data is high, as sediment or ice cores need to be drilled and the processing of core samples in laboratories is labor intensive. In addition, the areas, from which paleodata can be extracted, is limited. For example, ice cores from the last Glacial are available just in Greenland and Antarctica, and pollen samples are only preserved in terrestrial regions which were not covered by ice sheets. Therefore, the spatio-

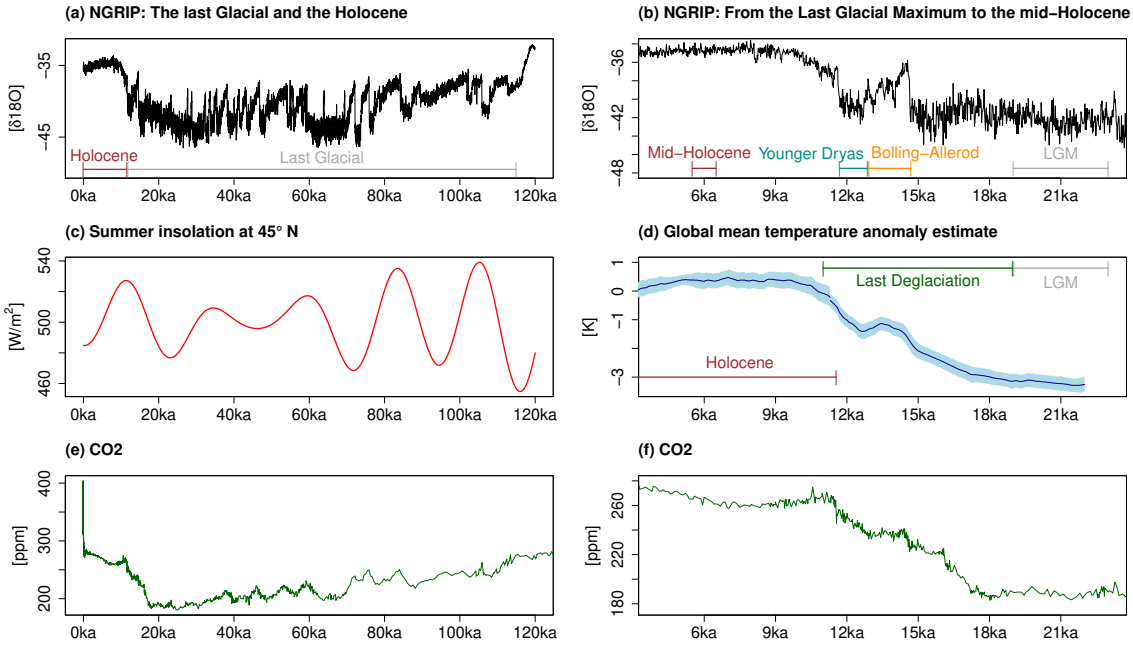

Figure 1.1: Temporal evolution of selected proxy records and forcings during the last Glacial and the Holocene. Left column: from the last Interglacial to today, (a) oxygen isotope evolution (δ18O) in the Greenland ice core NGRIP (North Greenland Ice Core Project members, 2004), (c) summer insolation at 45° N (Berger, 1978), and (e) CO_2 concentration (Köhler et al., 2017). Right column: from the Last Glacial Maximum to the mid-Holocene, (b) NGRIP δ18O evolution, (d) estimates of the global mean temperature anomaly during the Last Deglaciation (Shakun et al., 2012) and the Holocene (Marcott et al., 2013), and (f) CO_2 concentration.

temporal paleodata coverage is sparse compared to observations from the instrumental period. This demands sophisticated statistical techniques to extract robust signals of past climate changes from paleodata. Additionally, the relationship of paleodata and climate variables is often complex and stochastic. For example, pollen samples are an indicator of past vegetation, which in turn is controlled by seasonal temperatures and moisture availability, but also by other climatic and non-climatic factors like availability of light, soil properties, and plant competition. Furthermore, the relative importance of the climate factors, which influence vegetation composition, varies in space and time. Because of the stochasticity in the proxy-climate relation and simplifications made in the formulation of transfer functions, which describe the relationship of paleodata and climate variables, quantification of uncertainties is an important part of paleoclimate research to avoid overconfident statements on the state and evolution of past climate. Many international efforts (e.g., MARGO, MARGO Project Members (2009); INTIMATE, Blockley et al. (2012); PAGES2k, PAGES2k Consortium (2017); NEOTOMA, Williams et al. (2018)) aim at the combination of multiple proxy records and types in joint data products to increase the information

content in paleodata compilations.
The sparse coverage and low signal-to-noise ratio of paleodata as well as the complexity of the Earth system makes the identification of mechanisms, which cause inferred past climate changes, difficult. One strategy is the use of simulations with ESMs of varying degree of complexity. These so-called paleosimulations rely on reconstructions of boundary conditions like radiative forcing from orbital parameters and greenhouse gas concentrations, land-sea-masks, and ice sheet topography. ESMs are run with either fixed or time-dependent boundary conditions to produce equilibrium or transient simulations. Two issues of paleosimulations are the typically large deviation between simulated climate states and proxy reconstructions (Braconnot et al., 2012), and the complicated separation of physically reasonable mechanisms from model artifacts due to the complexity of state-of-the-art ESMs.

One way to combine information from proxy networks in a comprehensible and statistically consistent way are so-called climate field reconstructions (CFRs). CFRs of past near surface climate states combine information from proxy samples, which are mostly localized, with a model for interpolation between those samples on spatial or spatio-temporal domains. Thereby, CFRs are more suitable for many quantitative analyses of past climate than individual proxy records. For example, they can be used for statistical hypotheses testing, to identify common signals in proxy samples, for comparisons with paleosimulations, and as input data to numerical models of other compartments of the Earth system like vegetation or ice sheet models.
Unfortunately, the stochastic proxy-climate relation and the sparseness of available proxy data lead to complex CFR uncertainty structures. A meaningful reconstruction has to account for these uncertainties (Tingley et al., 2012). A natural way to represent uncertainties in the proxy-climate relation is to formulate transfer functions probabilistically, i.e. in terms of probability distributions (Ohlwein and Wahl, 2012). To account for uncertainties due to sparseness of proxy data, this study suggests the use of stochastic interpolation techniques. Most standard geostatistical methods like kriging or Gaussian process modeling are designed for interpolation in data rich situations, while in paleoclimatology we deal with sparse data. Therefore, their direct application to paleo situations is not suitable. Instead, this study proposes to use interpolation schemes that contain additional physical knowledge, such that the resulting product combines the information from a proxy network in a physically reasonable way (Gebhardt et al., 2008). This approach can in addition be used for structural extrapolation.

This work uses Bayesian statistical frameworks to combine the (local) proxy-climate relation and spatial or spatio-temporal interpolation. Bayesian statistics facilitates the combination of multiple data types and the accounting for several sources of uncertainty in a statistically rigorous way by estimating and inferring a multivariate probability distribution, the so-called posterior distribution (Gelman et al., 2013). In recent years, Bayesian hierarchical models (BHMs) and data assimilation (DA) frameworks formulated in a Bayesian way have been used for CFRs by several authors (e.g.,

Tingley and Huybers, 2010a; Li et al., 2010; Tingley and Huybers, 2013; Dee et al., 2016; Hakim et al., 2016), but all of these studies are restricted to the last two millennia. Beyond the Common Era, BHMs were developed for reconstructions of individual proxy records (e.g., Parnell et al., 2016) or to combine a small number of nearby records (Holmström et al., 2015), but only Gebhardt et al. (2008) and Simonis et al. (2012) incorporated physically motivated structures to reconstruct spatial climate fields in a variational framework. Extending CFRs with BHMs and DA methods beyond the Common Era is an important task of research which is tackled in this study.

Several new challenges arise in this extension. The proxy types change from mostly annually resolved proxies like tree rings or corals (PAGES2k Consortium, 2017) to proxies which record centennial- to millennial-scale climate changes such as pollen (Bartlein et al., 2011) or Foraminifera assemblages (MARGO Project Members, 2009). While using these proxies leads to reconstructions which are more uncertain on an absolute scale, the signal-to-noise ratio might be higher due to larger signals for example during glacial-interglacial transitions (Weitzel et al., 2019). An additional challenge is the lower spatial coverage of proxy data compared to the Common Era, especially on timescales beyond the Holocene. For reconstructions of the Common Era, the standard procedure calibrates parameters controlling the proxy-climate relation and the latent climate process against time series of instrumental records (e.g. Tingley and Huybers, 2013). The lower temporal resolution of proxy records beyond the Common Era makes this approach not feasible such that either a spatial calibration against modern climatologies or additional sources of information like physical mechanisms and paleosimulations need to be incorporated in the reconstructions. The statistical methods to study climate evolution beyond the Common Era have to be adapted, because non-stationarities become more important and a higher focus is on the detection and quantification of millennial-scale trends and abrupt changes like the BA warming or the YD cooling (see Fig. 1.1a,b). In contrast, most CFR methods for the Common Era assume a stationary climate evolution (Tingley et al., 2012).

This study is organized around the following research questions:

1. *Are skillful CFRs, which combine proxy syntheses and paleosimulations using Bayesian filtering methods, possible?*

   The sparse spatial coverage of paleodata and multiple sources of uncertainty make the reconstruction of physically reasonable climate fields, which are consistent with a given proxy synthesis, difficult. DA methods, which combine sparse observations and physical models, have emerged in many areas of geosciences (Carrassi et al., 2018). Past model-data comparison efforts demonstrated systematic mismatches between paleosimulations and proxy syntheses (Hargreaves et al., 2013; Braconnot et al., 2012) such that the applicability of standard DA methods in paleoclimatology is an open question. In this work, we study the ability of Bayesian filtering methods, designed with regard to the specific peculiarities of paleoclimatology, to produce skillful CFRs. The methods are evaluated using identical twin experiments (ITEs) with an ensemble of paleosimulations and cross-validation experiments

(CVEs) with a synthesis of pollen and macrofossil samples.

2. *Does a reconstruction with Bayesian filtering methods support the finding of dipole structures in the European MH climate by Mauri et al. (2014) despite the absence of such structures in most MH paleosimulations?*
   We apply the methods developed to tackle the previous research question to reconstruct summer and winter temperature in Europe during the MH. Using a European pollen synthesis, Mauri et al. (2014) found a dipole structure with warming in Northern and cooling in Southern Europe. Less pronounced, this dipole structure is found in Bartlein et al. (2011), too. In contrast, most paleosimulations exhibit a spatially persistent summer warming which is consistent with a stronger summer insolation on the Northern Hemisphere during the MH. Mauri et al. (2014) explain the deviation of models and paleodata with circulation changes that state-of-the-art ESMs do not reproduce. In this work, we test whether the dipole structure in pollen-based reconstructions is strong enough to persist even if the proxy synthesis is combined with spatial structures from paleosimulations, which do not exhibit distinct north-south temperature dipoles.

3. *Can pollen-based, spatial reconstructions of Siberian LGM climate improve the understanding of the absence of a Siberian ice sheet during the last Glacial?*
   Unlike other high latitude regions on the Northern Hemisphere, namely North America with the Laurentide ice sheet and Europe with the Fennoscandian ice sheet, Siberia was not covered by a large ice sheet during the last Glacial. Temperature and precipitation are two important controls of ice sheet surface mass balances (Reeh, 1989). This study produces the first continental-scale reconstruction of Siberian LGM summer temperature and annual precipitation amount, which can provide new insights into the respective roles of those two variables for the absence of a Siberian ice sheet.

4. *Can gradual trends and abrupt changes during the LD be identified from sparse and noisy proxy samples?*
   This work develops a new BHM for reconstructions of the climate evolution during the LD on continental scales based on existing efforts for the Common Era. The model is evaluated in a controlled test environment where the assumed truth shows a distinct non-stationary and multi-scale behavior, with a gradual warming on millennial timescales interrupted by abrupt climate changes that appear within a few centuries. The assumed spatio-temporal proxy coverage and the noise level of the proxy data are idealized but deduced from existing proxy databases and synthesis efforts like Bartlein et al. (2011), Tian et al. (2018), and Shakun et al. (2012). We evaluate whether non-stationarities in the first moments like millennial-scale trends and abrupt shifts can be reconstructed within this framework.

5. *Can (large-scale) dynamical properties of the LD climate evolution be extracted from data-driven CFRs?*

In addition to the reconstruction of trends and shifts, it is tested whether dynamical properties of the LD can be extracted from the CFRs described in the previous research question. We focus on the ability to separate between large- and small-scale changes, to reconstruct time series of large-scale changes, and to estimate parameters that describe second moment properties.

The structure of this thesis is as follows: In the next section, relevant concepts and literature, which are important for this work, are briefly summarized. This includes pollen and macrofossil data, paleosimulations, Bayesian statistics, statistical transfer functions for biological proxies, and CFRs.

In Chapter 3, methods for spatial reconstructions from pollen and macrofossil syntheses are tested and applied. These methods use Bayesian filtering techniques to combine proxy records, that constrain the local climate, with a multi-model ensemble of climate simulations, which produce physically consistent spatial fields for a given set of large-scale external forcings. We compare different possibilities to incorporate the information from the paleosimulations in a Bayesian reconstruction framework. Then, the most promising framework is applied to reconstruct MH summer and winter temperature in Europe as well as Siberian summer temperature and annual precipitation (P_{ANN}) amount during the LGM.

Chapter 4 extends the spatial reconstructions to spatio-temporal domains. A data-driven approach is chosen, which is less dependent on paleosimulations. This means, that a stochastic process controls the spatio-temporal interpolation between proxy samples. Paleosimulations can be added as probabilistic constraints for selected time slices through a spatial reconstruction module. The different elements of the framework, the proxy module, the spatial reconstruction module, and the spatio-temporal interpolation module are integrated in a BHM. After a detailed description of the BHM, the framework is tested in a controlled test environment given by a reference simulation, the transient TraCE-21ka simulation of the LD (Liu et al., 2009b).

Finally, the results with regard to the above formulated research questions are summarized in Chapter 5. In addition, suggestions and guidelines for future research directions are provided.

2 Important concepts and literature review

In this chapter, important concepts and existing literature for the research questions of this study are summarized. First, pollen and macrofossil data as well as paleosimulations, the two sources of information in this study, are discussed. Then, Bayesian statistics, statistical transfer functions for biological proxy data, and CFR methods are introduced, which are the main techniques applied in this work.

2.1 Pollen and macrofossil data

Pollen are the terrestrial proxy with the highest spatial coverage and they are routinely used to reconstruct vegetation and climate states (Bradley, 2015). Pollen samples are extracted from lake sediment cores, and in some cases from marine sediments or peat bogs. The cores are divided into intervals and each sample belongs to one interval.

There are few laminated cores published that are dated precisely with a sub-centennial resolution by varve counting. Examples are the Meerfelder Maar in Germany (Litt et al., 2009), Lago di Monticchio in Italy (Allen et al., 1999), and Sihailongwan Lake in China (Mingram et al., 2018). The more common dating approach is radiocarbon dating. The age of some samples is determined up to an uncertainty estimate by radiocarbon measurements (Bronk Ramsey, 2009), and the age of the remaining samples is inferred from interpolation. This means that the sedimentation rate of the core is estimated based on prior information and the radiocarbon dates (Parnell et al., 2011). Due to cost constraints, the number of samples, that are radiocarbon dated, is typically an order of magnitude smaller than the overall number of samples in a core.

The basic assumption of pollen as climate proxy is that vegetation is largely driven by climatic conditions (Bradley, 2015), such that a two-stage forward process structure can be formulated. The climate constraints the vegetation composition, which in turn determines the pollen production of each taxon. Then, a subset of the generated pollen is transported into the lake, from which the sediment core is extracted, to form a pollen assemblage. Here, a taxon is a taxonomic unit that can be distinguished in biological proxy data.

Pollen are an integrated proxy in the sense that they contain information about the climate in an area that is constrained by the influx domain via horizontal pollen transport. It is also integrated in time as each sample contains pollen from a time interval, which depends on the sedimentation rate and sampling interval depth in the core. In addition, the response of vegetation composition to climate changes is delayed such that the minimal timescale, for which climate changes can be

resolved, is on the order of decades.
The spatio-temporal coverage of pollen samples is very irregular. In general, the number of samples is lower the further back one looks in time. The most proxy records are available for Europe and North America (Williams et al., 2018), and a growing number is compiled for Asia (Tian et al., 2018).
Macrofossils like leaves, seeds, and fruits, are another biological proxy type, which can be found in sediment cores. They are transported only over very small distances such that disturbances via long distance transport, that can emerge for pollen, can be precluded (Hannon and Gaillard, 1997). Moreover, they are often identifiable on the species level, whereas pollen can mostly be just separated on family or genus level (Bradley, 2015). Therefore, the constraints from an individual macrofossil for climate reconstructions is stronger compared to a single pollen grain since the climatic niche of a species is often smaller than that of a family or genus. The main disadvantage of macrofossils is that they appear in a much lower quantity than pollen, which makes it impossible to draw any conclusions from their absence or the number of macrofossils from a given species in a sample. In addition, much fewer sediment records with classified macrofossils exist. Therefore, reconstructing climate or vegetation fields just from macrofossils is practically impossible, but combining information from pollen and macrofossils in vegetation and climate reconstructions can add value compared to reconstructions which rely solely on pollen (Birks and Birks, 2000).

2.2 Paleosimulations

As described above, the sparse coverage and indirect nature of proxy data makes the identification of mechanisms that govern detected climate changes challenging. A common strategy in paleoclimatology is to test proposed mechanisms with ESMs of various complexity and compare the simulation output with proxy data, although this comparison is highly challenging, particularly on glacial-interglacial timescales (Weitzel et al., 2019). Many parameterized processes in ESMs are tuned with modern observations as target. Therefore, the diagnosis of biases in paleosimulations can help to identify parameter settings that reproduce today's climate but do deviate significantly from reconstructions of past climate.
Two types of paleosimulations are common in the literature: equilibrium and transient simulations (Kageyama et al., 2018). Given a fixed set of forcing conditions, equilibrium simulations with ESMs produce a physically consistent estimate of the atmospheric and oceanic circulation, and the regional energy balance. Depending on the ESM, additional components like dynamical vegetation, sea ice, or dynamic ice sheets can be coupled to atmosphere and ocean modules. Starting from initial conditions that are either taken from a previous model run or a reference simulation, the ESM is integrated over a long time, typically around 1000 years, until atmosphere and ocean have reached an equilibrium state. This so called spin-up is necessary for coupled models as the ocean responds much slower to changes in forcings than the atmosphere. Subsequently, the model is integrated for typically 100 to 1000 years to analyze the climate state including inter-annual to

decadal-scale internal variability given a set of boundary conditions. In transient simulations, the forcings change over time, which is important to study transitions between climate states. The changes can include varying orbital configurations, greenhouse gas concentrations, volcanic forcing, and changing ice sheets and topographies. In addition, variable forcings can increase the variability of climate simulations, which is underestimated compared to proxy records when run with fixed forcings (Laepple and Huybers, 2014).

Another application of paleosimulations is to use them as physically consistent downscaling of large-scale forcing information. Multi-model ensembles of paleosimulations can provide estimates of climate states and spatio-temporal structures, which are physically reasonable under prescribed forcings. This interpretation of paleosimulations facilitates the use of multi-model ensembles as additional constraints in climate reconstructions (Hakim et al., 2016).

An important program to streamline paleosimulations is the Paleoclimate Modelling Intercomparison Project (PMIP) which is currently in its fourth phase (Kageyama et al., 2018), and is a subproject of the Coupled Model Intercomparison Project (CMIP). Within PMIP4, several key periods of interest have been defined. These are the MH, the LGM, the Last Millennium, the Last Interglacial (around 127ka), and the mid-Pliocene Warm Period (around 3.2ma) (Kageyama et al., 2018). For those periods, the participating modeling groups have agreed on a set of forcings, and simulations are run with many different ESMs. Common output variables and formats have been defined and all the data is archived through the Earth System Grid Federation (ESGF). In this work, we use simulations from PMIP3 as the PMIP4 simulations are still ongoing. The simulation output has been downloaded from the German ESGF node at the German Climate Computing Center (DKRZ).

2.3 Bayesian statistics

Bayesian statistics are a branch of statistics that is based on Bayes' theorem on conditional probabilities (Bayes, 1763). It has gained a lot of popularity in the past few decades due to advances of computational methods, and its ability to quantify uncertainties and deal with complicated test designs (Carlin and Louis, 2008). The general goal of statistics is to infer a quantity A given data B. Bayesian statistics expresses this goal in terms of conditional probability distributions, by estimating the probability distribution of A given B. This distribution is called posterior distribution and can be rewritten as

$$\overbrace{\mathbb{P}(A \mid B)}^{\text{Posterior}} = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A)\, \mathbb{P}(A)}{\mathbb{P}(B)} \tag{2.1}$$

$$\propto \underbrace{\mathbb{P}(B \mid A)}_{\text{Likelihood}} \underbrace{\mathbb{P}(A)}_{\text{Prior}} \tag{2.2}$$

Here, the first two equations result from applying Bayes' theorem twice. That the posterior is proportional to the product of the likelihood and the prior distribution follows because B is a fixed

quantity, such that $\mathbb{P}(B)$ is a normalizing constant in Eq. (2.1). The likelihood is also called observation operator in DA and transfer function in the paleoclimatology community. The strategy in Bayesian statistics is to model the likelihood and the prior probability distribution independent of each other and then obtain the posterior distribution (up to a constant) by multiplying those two. The likelihood describes the probability of getting the data B given any possible value of A. The prior distribution contains independent knowledge of A which is quantified prior to knowing the data B. There is a large literature on defining prior distributions (Gelman et al., 2013). Popular choices are Jeffreys priors (Jeffreys, 1946), which try to minimize the information content of the prior distribution while preserving desirable mathematical properties, and conjugate priors, which are chosen to result in an easy to evaluate closed form for the posterior (Gelman et al., 2013). Many popular probability distributions like Gaussian, Multinomial, or Poisson distributions possess conjugated priors.
Advantages of Bayesian methods, particularly in the context of paleoclimate reconstructions, are:

- The likelihood is formulated as forward model, which means that a model for B given A is sought. This corresponds to the logical structure of paleoclimate problems, where the climate should be inferred from proxy data, but the assumption is that the climate controls the specification of the proxies. Bayes' theorem provides a natural way to invert this forward structure.

- As each quantity in the Bayesian model Eq. (2.1) is modeled as a probability distribution, the uncertainty of each object in Eq. (2.1) is naturally included in the inference process. Therefore, Bayesian statistics are well-suited for the quantification of uncertainties, especially when multiple sources of uncertainty have to be combined.

- Additional quantities can be added to the basic Bayesian framework described above. Via multiple applications of Bayes' theorem, these quantities can be incorporated in a hierarchical way that leads to a number of independently specified conditional distributions. Then, the posterior distribution is recovered by multiplying all the conditional distributions. Such models are called BHM because the conditional distributions can be grouped into several hierarchical levels (Banerjee et al., 2014). An advantage of the hierarchical structure is that each BHM can be identified with a directed acyclic graph (DAG). In the DAG, each node represents a variable and the arrows indicate dependencies of variables (see e.g. Fig. 2.1).

- Bayesian models are well-suited to incorporate multiple sources of information. For example, in paleoclimate reconstructions multiple types of proxies can be combined in a BHM (Li et al., 2010), or information on radiative forcing can be included in the reconstruction (Gebhardt et al., 2008).

- Latent variables, that describe underlying dependencies, can be marginalized to focus on the target variables of a statistical inference within a complex Bayesian framework.

A problem of BHMs is that the posterior distribution can usually not be evaluated analytically or via direct numerical integration. This has hampered the spread of Bayesian methods for a long time, but the technical progress of the last decades has paved the way for new computational methods. Markov chain Monte Carlo (MCMC) methods of various complexity have become the standard tool to evaluate analytically intractable posterior distributions by sampling from a Markov chain which converges to the correct posterior distribution (Robert and Casella, 2004). Popular methods are Gibbs sampling, the Metropolis-Hastings algorithm, Hamiltonian Monte Carlo, and the Metropolis adjusted Langevin algorithm (Betancourt, 2018). An alternative are approximate Bayesian inference techniques like the integrated nested Laplace approximation (INLA, Rue et al., 2009), which have proven to be useful particularly for high dimensional posteriors.
These advances have led to the development of Bayesian frameworks for paleoclimate reconstruction problems in the last two decades (Tingley et al., 2012). Applications include the inversion of pollen-climate forward models (e.g. Ohlwein and Wahl, 2012; Salter-Townshend and Haslett, 2012; Holmström et al., 2015), data driven CFRs (e.g. Tingley and Huybers, 2010a; Li et al., 2010), and DA techniques (e.g. Steiger et al., 2014; Dee et al., 2016). Typically, Bayes' theorem is used in paleoclimatology to combine forward models for proxy data with interpolation methods for climate variables, which leads to BHMs that consist of three stages:

$$\underbrace{\mathbb{P}(C, \Theta \mid \mathrm{P})}_{\text{Posterior}} \quad \propto \quad \underbrace{\mathbb{P}(P \mid C, \Theta)}_{\text{Data Stage}} \quad \underbrace{\mathbb{P}(C \mid \Theta)}_{\text{Process stage}} \quad \underbrace{\mathbb{P}(\Theta).}_{\text{Prior stage}} \tag{2.3}$$

Here, C is the climate to be reconstructed, P denotes the proxy data, and Θ are model parameters that control the interpolation of climate variables and the proxy-climate relation. The data stage contains the transfer function formulated in a forward way, the process stage models the interpolation, and the prior stage defines prior distributions of the model parameters (Tingley et al., 2012).

2.4 Statistical transfer functions for biological proxy data

There is a long tradition of using biological proxies to infer past climate by applying statistical transfer functions (Birks et al., 2010; Ohlwein and Wahl, 2012). These transfer functions consist of a calibration step, where parameters (denoted by θ) are fitted to a modern calibration set, that consists of modern climate data (denoted by C_m) and a training set of modern pollen samples or vegetation distributions (denoted by P_m). Then, this statistical link is applied to a fossil proxy sample (denoted by P_p) to reconstruct the climate (denoted by C_p) at the estimated date and location of the sample using the fitted parameters θ. Fig. 2.1 shows a Bayesian formulation of this basic structure. Note, that the involved quantities are ordered as DAG, which facilitates the formulation of a posterior distribution for $\mathbb{P}\left(C_p, \theta \,|\, P_p, C_m, P_m\right)$. Transfer functions group the data of a sample into taxa, which are treated as homogeneous group.
Statistical transfer functions are based on several assumptions, summarized for example in Birks et al. (2010). The most important assumption is that the taxa, which are used for the reconstruction,

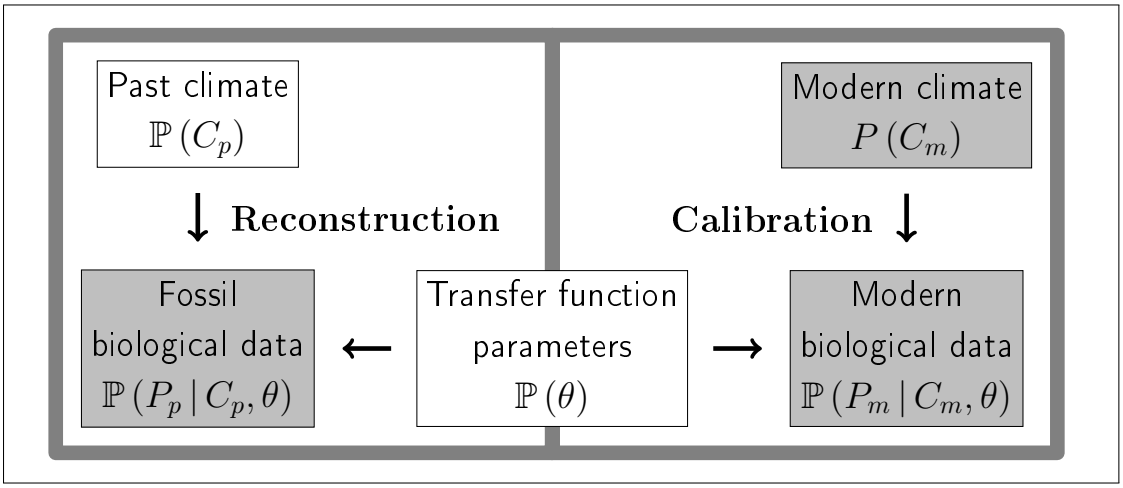

Figure 2.1: Structure of statistical transfer functions for reconstructions from biological proxies involving a calibration step (right) and a reconstruction step (left), which are connected through transfer function parameters. The gray nodes are given data, the white ones are quantities to be inferred. The arrows indicate the direction of dependencies of a corresponding Bayesian model as described in Eq. (2.3), in which both steps are combined.

are systematically related to the climate variables to be reconstructed, and that this relationship, as diagnosed from the training data, has not changed between the age of the fossil sample and the training data set. In addition, it is assumed that either other environmental variables have negligible influence on the fossil sample, or that their joint distribution with the reconstruction targets was the same in the past than today.

Two types of reconstruction methods can be distinguished: Assemblage methods and indicator taxa methods (Birks et al., 2010). Assemblage methods compare the counts or relative frequencies of the taxa in the fossil sample with the counts or frequencies in the training data. They rely on the assumption that the similarity of the climate, corresponding to two samples, is proportional to the deviation between the samples in a predefined metric. The distance can be evaluated directly as in the so-called modern analogue technique (MAT), or with a smoothing step in-between through the estimation of response surfaces. Alternatively, weighted averaging (WA) methods are used, where first the climate optimum for each taxon is determined from the training data and then these optima are weighted according to the abundance of the taxa in the fossil sample. An extension of the standard WA method is the use of partial least square calibration (WAPLS). WA is less sensitive to spatial autocorrelation in the training set and influences of other environmental variables on the reconstruction than MAT (Birks et al., 2010). In the past years, Bayesian versions of assemblage methods have been developed to facilitate an improved quantification of uncertainties. Examples are BUMMER by Korhola et al. (2002), BUMPER by Holden et al. (2017), and BClim, which is described in Salter-Townshend and Haslett (2012) and Parnell et al. (2016). While BUMMER and

BUMPER implement WA in a Bayesian way, BClim is a Bayesian version of response surface methods. Taxa can be grouped in biomes or plant functional types (PFT) to reduce the noise content in the samples (see e.g. Litt et al., 2012, for a Bayesian implementation). Assemblage methods are best suited for reconstructions from taxa that occur over a wide range of climates but whose abundance follows a sharp climatic gradient.

Indicator taxa methods (ITMs) reconstruct the joint climatic niche of selected taxa in a fossil sample. The assumptions behind ITMs are weaker than for assemblage methods, as it is only assumed that the climatic niches, where the taxa occur, do not change over time. The taxa are chosen based on expert elicitation or with statistical selection methods (Kühl et al., 2002). For each taxon, the climatic niche is estimated from the training data, and then the niches of each taxon are combined in the reconstruction step. The traditional ITM, which goes back to Iversen (1944), was reformulated in a probabilistic way by Kühl et al. (2002), Gebhardt et al. (2008), and Stolzenberger (2017) to formalize the reconstruction procedure mathematically and improve the quantification of uncertainty. In the literature, this probabilistic indicator taxa method (PITM) is often referred to as pdf-method (Kühl et al., 2002). ITMs are best suited for reconstructions from taxa with very small climatic niches. This can be the case for rare taxa in pollen samples or for macrofossils identified on species level.

2.5 Climate field reconstructions

CFR methods are statistical models that reconstruct spatial, temporal, or spatio-temporal near surface climate fields from a network of proxy records. They compute gridded fields based on the input proxy data and an interpolation model. The progress and challenges of CFRs have been reviewed by Christiansen and Ljungqvist (2017) for reconstructions of the past two millennia, and by Tingley et al. (2012) from a Bayesian point of view.

Reconstructing large-scale fields from sparse and noisy data is challenging from a paleoclimatological as well as from a statistical perspective. The main climatological issues are the formulation of appropriate models that link the target climate variables and the proxy data (Evans et al., 2013; Ohlwein and Wahl, 2012), the determination of spatio-temporal scales for which the proxy data provides information, and the definition of interpolation schemes that preserve as much physical consistency as possible. This challenge interferes with the identification of the right amount of smoothing in the interpolation scheme, possibly on multiple spatio-temporal scales, which is a topic in geostatistics. Additional statistical challenges are the quantification of uncertainties, and the treatment of the change of support problem (Tingley et al., 2012). This means that the proxy data is either point data or a weighted average over an irregular spatio-temporal domain, whereas the CFR contains grid box averages.

CFRs became popular in the late 1990s, especially in response to the regression-based reconstruction for the last millennium of Mann et al. (1998). In the following two decades a lot of controversies and scientific discussion pivoted around the merits and disadvantages of various re-

gression models, biases resulting from calibration specifications, the quantification of uncertainties in CFR methods, the selection of proxy records, and the preservation of low-frequency variability in the interpolation step (e.g., Masson-Delmotte et al., 2013; Christiansen and Ljungqvist, 2017). The tested methods include composite-plus-scaling (Jones et al., 1998; Mann and Jones, 2003; Moberg et al., 2005), principal component regression (Mann et al., 1998; Ammann and Wahl, 2007), point-to-point regression (Cook et al., 1999), univariate linear regression (D'Arrigo et al., 2006), total least squares (Hegerl et al., 2007), regularized expectation maximization (Mann et al., 2008), the LOC method of Christiansen and Ljungqvist (2011), which is based on local regression, and expectation maximization with sparse precision matrices derived from graphical models (Guillot et al., 2015). All those methods have in common that intervals of the last two millennia are reconstructed, and that the reconstructions are based on regression against the instrumental period.

As an alternative to these frequentist regression approaches, BHMs have been developed by Tingley and Huybers (2010a) and Li et al. (2010) to facilitate a better quantification of uncertainties and to clarify involved modeling assumptions. In addition to proxy data, Li et al. (2010) included information on solar, greenhouse gas, and volcanic radiative forcing as linear covariates. Besides those covariates, both models are stationary, linear, and the process stage follows an isotropic Gaussian structure. Especially the stationarity assumption is incompatible with the climate evolution during the LD. Therefore, these two models are not directly applicable to the research questions 4 and 5 from Chapter 1.

As a way to reconstruct physically more consistent CFRs and due to the increased number of available paleosimulations, so-called off-line DA methods have gained popularity in the last years (Goosse et al., 2016). The most common approach is the application of ensemble Kalman filters, where the observations are integrated quantities (typically annual averages), and the prior covariance is estimated from a database of ESM simulations (Steiger et al., 2014). This framework was extended by Dee et al. (2016) to include process-based forward models for the paleodata and applied to data from the PAGES2k project (PAGES2k Network, 2013) in the last millennium climate reanalysis project (Hakim et al., 2016). Similar to the other CFR approaches described before, these DA frameworks have only been applied to the Common Era.

CFRs have rarely been used for reconstructions beyond the last two millennia, and several challenges arise for the extension of previously used methods as pointed out in Chapter 1. Gebhardt et al. (2008) developed a variational approach that combined an advection-diffusion model with pollen and macrofossil samples for a reconstruction of European climate during the early Eemian (around 125ka). The advection-diffusion model was forced by insolation changes between today's climate and the Eemian. The method was expanded by Simonis et al. (2012) to improve the quantification of reconstruction uncertainties via importance sampling. They used pollen and macrofossil samples from four time slices, 6ka, 8ka, 12ka, and 13ka, for spatial reconstructions of European climate. Pattern matching, a form of multi-linear regression that is also used in climate change detection and attribution (Hegerl and Zwiers, 2011), was applied by Annan and Hargreaves

(2013) to reconstruct sea surface temperature (SST) from the MARGO LGM synthesis (MARGO Project Members, 2009) and surface air temperature from the Bartlein et al. (2011) pollen synthesis for the LGM. The patterns were extracted from the PMIP2 LGM simulations.

The evaluation of CFR methods is challenging, because only indirect paleodata is available. Therefore, it is not possible to distinguish biases in the observation model from biases in the interpolation model. Controlled experiments with reference datasets are an alternative to evaluations against paleodata. As the instrumental period is too short to serve as reference on the temporal scales that this study is interested in, paleosimulations are chosen instead. From these simulations, so-called pseudo-proxies can be simulated which imitate true proxy data. Then, CFR techniques can be applied to reconstruct the reference climate, and the fields can be evaluated against the reference simulation. These pseudo-proxy experiments (PPEs) are a frequently used instrument in the CFR community (e.g., von Storch et al., 2004; Mann et al., 2007; Christiansen et al., 2009; Werner et al., 2013; Gomez-Navarro et al., 2015; Steiger and Smerdon, 2017). In DA, similar types of experiments are called ITEs, when the reference simulation is taken from an ensemble of simulations and is inferred from the remaining ensemble members, and observation system study experiments, when the impact of a certain observational network shall be evaluated.

3 Combining proxy data and Earth system model simulations for spatial climate reconstructions using Bayesian filtering

In this chapter, we study the ability of Bayesian filtering methods, which combine pollen and macrofossil syntheses with multi-model ensembles of paleosimulations, to reconstruct spatial climate fields for time slices. During the time slices, the climate state is assumed to be approximately in equilibrium with external forcings. In the Bayesian filtering approach, a spatial climate prior distribution is updated by proxy data compilations. This distribution is computed from a climate simulation ensemble in combination with geostatistical techniques, which adjust for climate model inadequacy and the small number of available simulations. Multiple statistical models to define the climate prior can be envisaged. We test three different formulations of the process stage and two versions of the involved spatial covariance matrix, which results in six statistical models. To compare these models, ITEs with the PMIP3 MH ensemble and CVEs with the proxy data synthesis for the MH are designed and evaluated. Afterwards, results from applying the statistical model, which performs best in the comparison experiments, to the MH and LGM proxy syntheses are presented.

This chapter is structured as follows. First, the European pollen and macrofossil compilation for the MH, the local reconstructions from pollen samples of the Siberian LGM climate, and the PMIP3 ensembles for the MH and LGM are introduced. Then, the Bayesian filtering methods are presented. The subsequent results section starts with a comparison of the six statistical models using ITEs and CVEs, which is followed by spatial reconstruction results for the European MH and the Siberian LGM climate. The spatial reconstructions are compared with previous reconstruction efforts, and the robustness of the reconstructions as well as possible extensions are discussed.

3.1 Data

3.1.1 European pollen and macrofossil synthesis for the mid-Holocene

This study uses a pollen and macrofossil synthesis, which stems from Simonis et al. (2012) as part of the European Science Foundation project DECVeg (Dynamic European Climate-Vegetation impacts and interactions). Out of the four time slices (6ka, 8ka, 12ka, 13ka), which were compiled in Simonis et al. (2012), we only use the 6ka dataset because there is no ensemble of climate

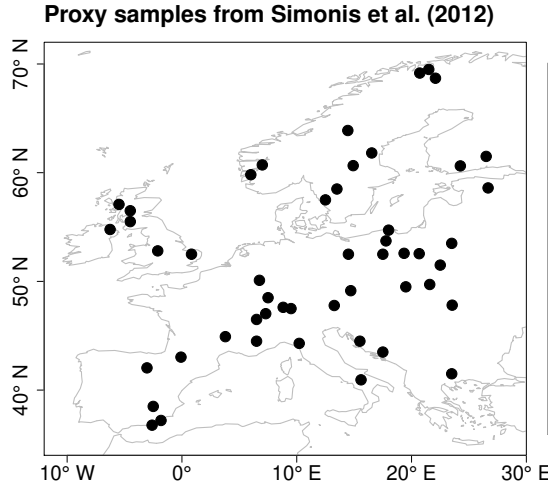

Figure 3.1: Locations of pollen and macrofossil samples from Simonis et al. (2012), marked by black dots.

simulations available for the other three time slices. For 50 paleosites, information on the occurrence of taxa is provided. 59 taxa occur at least at one site. For some sites, information from very nearby records are combined into a joint sample. 15 of the sites combine macrofossil and pollen data, three samples contain just macrofossil information, and for 32 sites only pollen data is available. In general, the macrofossil data provides more detailed taxonomic information than pollen, but sometimes only macrofossils from very few taxa are found. Therefore, additional pollen data at those sites and records with only pollen information are included to provide a more complete spatial picture of the European vegetation during the MH.

The 50 paleosites are sparsely but relatively uniformly distributed over Europe. Their locations are delimited by 6.5° W, 26.5° E, 37.5° N and 69.5° N. Compared with other recent syntheses like Bartlein et al. (2011), less records are included due to high quality control criteria. The raw pollen data and radiocarbon measurements, from which at least one was supposed to be close to 6ka, had to be available to recalculate age-depth models and ensure the use of calibrated radiocarbon dates as common timescale. Each site is assigned to the corresponding cell of a 2° by 2° grid, as a compromise between the resolutions of the paleodata and the paleosimulations. The locations of the proxy samples are depicted by black dots in Fig. 3.1. The full list of sites included in the synthesis can be found in Simonis et al. (2012). The list of taxa, which occur at the sites, is published in Simonis (2009).

Modern climate and vegetation data is used for the calibration of transfer functions. The climate data is computed from the University of East Anglia Climatic Research Unit (CRU) 1961 to 1990 reference climatology (CRU TS v.4.01; Harris et al., 2014; Harris and Jones, 2017). The vegetation data stems from digitized vegetation maps (Schölzel et al., 2002). The regions that have been used for the transfer function calibration were determined by pollen experts for each taxon (Kühl et al., 2007). The number of calibration samples varies between 14.543 and 28.844, depending on the

taxon.

For large parts of Europe, it was shown by Simonis (2009) that the pollen and macrofossil synthesis is well-suited for joint reconstructions of summer and winter temperature as measures for warmth of the growing season and cold of the winter, because at least one of these two variables is a limiting factor for most taxa growing in the mid and high latitudes of Eurasia during the Holocene. In contrast, the test of various climate variables, which indicate moisture availability, was less promising since moisture availability is rarely a limiting factor for European taxa (Simonis, 2009). Hence, in this study, mean temperature of the warmest month (MTWA) and mean temperature of the coldest month (MTCO) are chosen as target variables for climate reconstructions. To calculate MTWA and MTCO from time series of monthly averages, the data is first interpolated to the desired spatial grid. Then, for each hydrological year (October to September), the warmest and coldest month are extracted. The hydrological year is chosen instead of the calendar year to ensure that the months are taken from connected seasons. Finally, the climatological mean is calculated by averaging over the values for each year.

3.1.2 Local climate reconstructions of Siberian Last Glacial Maximum climate

For reconstructing the Siberian LGM climate, we use a compilation of local climate reconstructions from pollen records provided by the Polar Terrestrial Environmental Systems Division at the Alfred-Wegener-Institute (Ulrike Herzschuh, Xianyong Cao, and Thomas Böhmer, personal communication). The complete pollen synthesis contains 171 Siberian pollen records that cover parts of the last 40ka (Tian et al., 2018), which are many more Siberian records than in previous pollen syntheses like Bartlein et al. (2011). For each core, the raw radiocarbon dates as well as pollen percentages had to be available to be included in the synthesis. Therefore, it was possible to recalculate the age model with the Bayesian age model Bacon (Blaauw and Christen, 2011) as described in Cao et al. (2013). In this study, all samples with calibrated mean age between 19ka and 23ka are used. This leads to a total of 170 samples from 37 lake sediment cores, which are located between 69° E and 178° W and 42° N and 75° N. The locations of the cores are depicted by black dots in Fig. 3.4a. The taxonomy of the pollen counts is homogenized as described in Cao et al. (2013), Cao et al. (2014), and Tian et al. (2018). To apply statistical pollen-climate transfer functions, a modern pollen and climate calibration dataset was compiled, which is an expansion of the previously presented dataset for China and Mongolia (Cao et al., 2014). The data collection and processing of the samples is described in detail in Cao et al. (2014).

Several temperature and precipitation related variables have been tested for their ability to explain the observed pollen compositions similar to the analysis presented in Cao et al. (2014), using criteria such as explained variance (r^2), root mean square error of prediction (RMSEP) for fitted statistical transfer functions, canonical correspondence analysis, and species response models. These tests revealed that MTWA and P_{ANN} are the most promising variables for pollen-based climate reconstructions in Northern and Eastern Continental Asia.

Subsequently, reconstructions for each sample were performed at the Alfred-Wegener-Institute with WAPLS transfer functions following the methodology from Cao et al. (2014). Uncertainty estimates were calculated using the bootstrapped RMSEP from cross-validation of the modern calibration dataset, since the transfer function does not contain a natural error model for fossil samples. The WAPLS reconstructions as well as the error estimates were provided by Thomas Böhmer, Xianyong Cao, and Ulrike Herzschuh (personal communication). Recalculating the reconstructions from each sample with a Bayesian transfer function model was not feasible within this study. Anomalies of the reconstructions are calculated with respect to the reference climatology used for the calibration of transfer functions at the locations of the records. Since MTWA is calculated in the reference climatology by first computing monthly averages and then selecting the warmest month, this procedure is also applied in the corresponding spatial reconstructions.

The local reconstructions are summarized in Fig. 3.2. They show a cooler and drier LGM climate than today, with regional heterogeneity. A MTWA cooling of up to 4K is reached in large parts of the domain, but less cooling and in some grid boxes even positive anomalies are present near the coastlines of the Arctic and Pacific Oceans. MTWA values near the present temperature are found in the Southern Kolyma region (approximately 150° E, 60° N). The finding of relatively warm temperatures in this region is consistent with results of Meyer et al. (2017) who reconstructed LGM temperature on the Kamchatka Peninsula with geochemical proxies. The temperature uncertainties of the samples are relatively homogeneous in space, but multiple cores within one grid box and multiple samples in the cores can reduce the uncertainty of grid box averages compared to the individual samples. For most samples, a precipitation reduction of up to 50% compared to present day values is reconstructed. Exceptions are the coastlines of the Arctic Ocean and the Sea of Japan, where an increase of precipitation is inferred in several grid boxes. The standard deviations of the grid box averages range from 5 percentage points to 50 percentage points.

3.1.3 PMIP3 ensembles of climate simulations for the mid-Holocene and Last Glacial Maximum

To estimate a prior distribution which controls the spatial smoothing in the Bayesian framework, this study uses multi-model ensembles of climate simulations that were run within PMIP3, as part of CMIP5. Forcings are adjusted to the MH and the LGM. This includes changed orbital configurations and greenhouse gas concentrations for the MH and additional changes of ice sheets and topography, including a changed land-sea mask, for the LGM simulations (Braconnot et al., 2011). The ensemble contains all available MH and LGM simulations in the CMIP5 archive (downloaded from the DKRZ long term archive, available under https://cera-www.dkrz.de), which have a grid spacing of at least 2°. This constraint is chosen to better match the resolutions of pollen samples and simulations. The condition results in using seven MH model runs from the CCSM4, CNRM-CM5, CSIRO-Mk2-6-0, EC-Earth-2-2, HadGEM2-CC, MPI-ESM-P, and MRI-CGCM3, and four LGM model runs from the CCSM4, CNRM-CM5, MPI-ESM-P, and MRI-

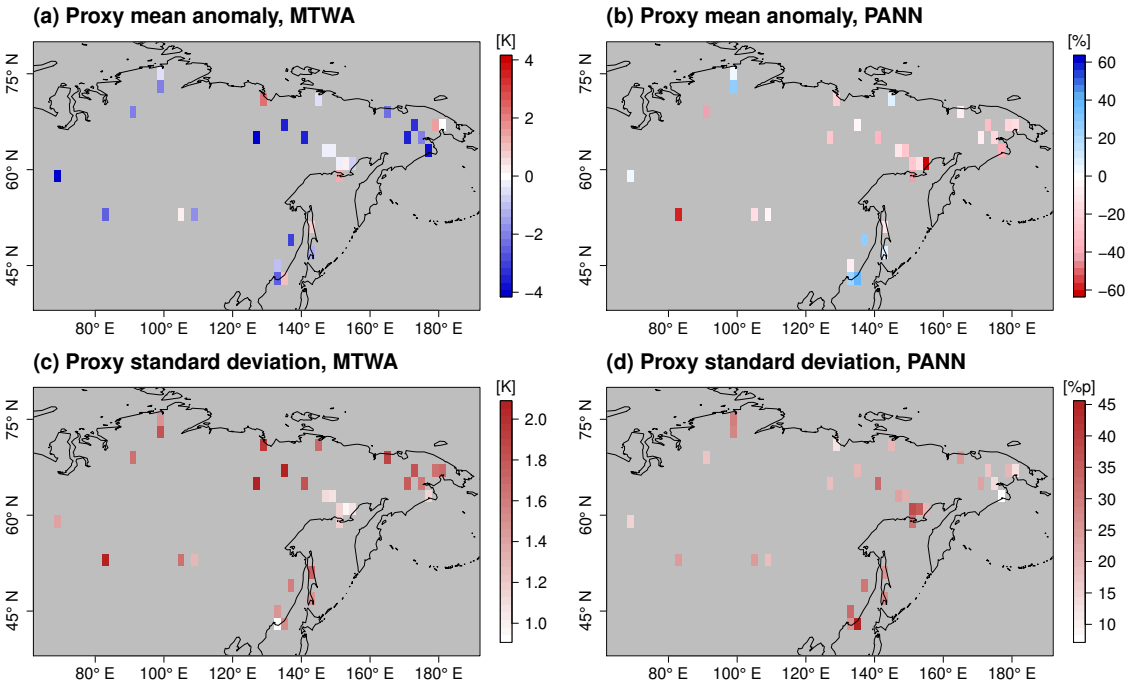

Figure 3.2: Summary statistics of local reconstructions with WAPLS transfer functions. Grid box values are computed by averaging over all samples from a grid box and assuming correlations among samples from the same record as estimated in the spatial reconstructions below. Top row: mean anomaly (left: MTWA, right: P_{ANN}), bottom row: standard deviation (left: MTWA, right: P_{ANN}).

CGCM3. Properties of the included simulations are given in Table 3.1. The simulations form a multi-model ensemble with common boundary conditions, i.e. they follow the same experiment protocol. The models are run to an equilibrium state (spin-up), followed by typically around 100 simulated years. Therefore, the differences within the ensemble represent modeling uncertainties (epistemic uncertainty).

All simulations are projected to a common 2° by 2° grid using bilinear interpolation. Then, the climatological means for the respective reconstruction variables, MTWA and MTCO for the European MH reconstruction, and MTWA and P_{ANN} for the Siberian LGM reconstruction, are extracted following the descriptions in Sect. 3.1.1 and 3.1.2. This minimizes inter-annual to decadal variability which is not resolved by the proxy data (Annan and Hargreaves, 2013). For the LGM simulations, anomalies with respect to the 1961 to 1990 climatologies of the respective models from the CMIP5 historical runs are computed. This is necessary to reduce model biases especially for precipitation as large differences between the models occurred due to spatial misalignment of precipitation patterns which are persistent between the modern climatologies and the LGM climatologies. This effect is small for the European MH climatologies such that the absolute values of the simulations are used. For P_{ANN}, we use the relative anomalies from the modern values, since

Table 3.1: Basic information on the PMIP3 climate simulations used to construct the process stage in the Bayesian framework (from https://pmip3.lsce.ipsl.fr).

Model	Institute	Atmospheric grid	Ocean grid	Model years MH	Model years LGM
CCSM4	NCAR	288x192xL26	320x384xL60	301	101
CNRM-CM5	CNRM-CERFACS	256x128xL31	362x292xL42	200	200
CSIRO-Mk3-6-0	CSIRO-QCCCE	192x96xL18	192x195xL31	100	—
EC-Earth-2-2	ICHEC	320x160xL62	362x292xL42	40	—
HadGEM2-CC	MOHC	192x144xL60	360x216xL40	35	—
MPI-ESM-P	MPI-M	192x96xL47	256x220xL40	100	100
MRI-CGCM3	MRI	320x160xL48	364x368xL51	100	100

they are more comparable across space.

The European mean summer climate, expressed as MTWA (Fig. 3.3a), in the MH ensemble is warmer than the CRU reference climatology (CRU TS v.4.01 over land, Harris et al. (2014), Harris and Jones (2017), and HadCRUT absolute over sea, Jones et al. (1999)) in large parts of Europe, especially Eastern Europe and the Norwegian Sea. These areas coincide predominantly with areas of large ensemble spreads, expressed as the grid box-wise empirical standard deviations in Fig. 3.3c. The standard deviations rise up to 4 K in some areas of Southern and Eastern Europe, which might originate from deviating change patterns of the general circulation over Europe in the models. In contrast to the ensemble mean MTWA climatology, the MH mean winter climate measured by MTCO in Fig. 3.3b shows a more dispersed structure with cooling in Fennoscandia, warming in the Mediterranean and Balkan peninsula, and mixed patterns in Western and Central Europe. The ensemble spread is predominantly small (Fig. 3.3d), but increases towards Northern Europe with very large inter-model differences in the Norwegian Sea and Eastern Fennoscandia.

The spatial average of the Siberian LGM MTWA ensemble mean anomaly is -2.8 K. The anomalies are most negative in the southern part of the domain, the magnitude of cooling decreases towards the north, and positive anomalies occur along the coastline of the Arctic Ocean and in areas that are below sea-level today but were above sea-level during the LGM (Fig. 3.4a). The ensemble spread has a clear north-south gradient with grid box wise standard deviations below 5 K in the southern parts of the domain and up to 20 K in the northernmost parts (Fig. 3.4c). The P_{ANN} ensemble mean is dryer than the modern climate across most of the domain (Fig. 3.4b). The strongest relative reduction of precipitation occurs in Western Siberia, whereas the smallest relative reduction is found in Mongolia, and Northeastern China. The ensemble spread exhibits a smooth spatial structure, with values up to 50 percentage points (Fig. 3.4d).

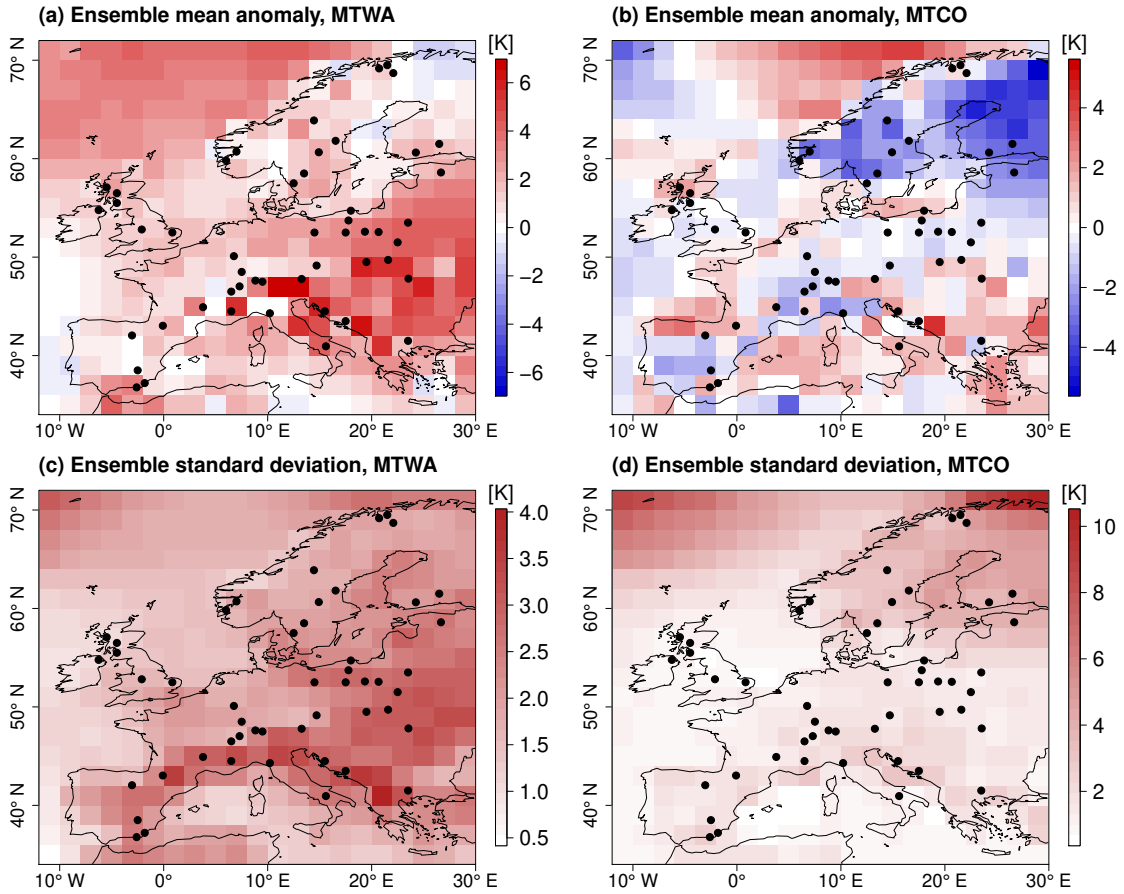

Figure 3.3: PMIP3 MH ensemble mean anomaly from CRU reference climatology, (a) MTWA, (b) MTCO, and empirical standard deviation, (c) MTWA, (d) MTCO. Black dots depict proxy samples from Simonis et al. (2012).

3.2 Bayesian filtering framework

To combine the networks of proxy samples and the ensembles of PMIP3 simulations described in Sect. 3.1, the Bayesian framework from Eq. (2.3) is used, which facilitates the inclusion of two types of uncertainty in the inference, the uncertainty from the proxy-climate relation and the uncertainty in the climate simulation ensemble. In this section, all quantities in the Bayesian framework are specified and the inference algorithms are described.

In the following, proxy data, which is either pollen data, macrofossil data, or local climate reconstructions, is denoted by P_p, past climate by C_p, modern vegetation and climate data for the calibration of transfer functions by P_m and C_m, respectively, and additional model parameters by Θ. These model parameters split into data stage and process stage parameters. The data stage parameters θ specify the proxy-climate relation and are defined in Sect. 3.2.1. Process stage parameters ϑ control the way how the climate simulations are incorporated in the Bayesian framework and are specified in Sect. 3.2.2.

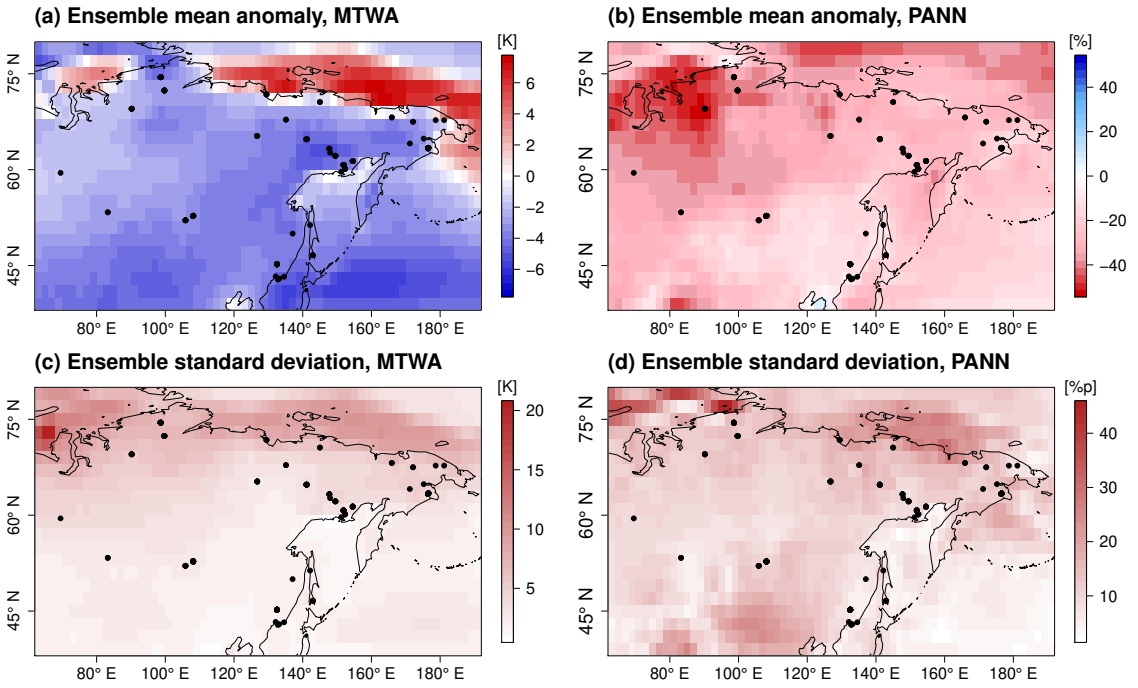

Figure 3.4: PMIP3 LGM ensemble mean anomaly, (a) MTWA, (b) P_{ANN}, and empirical standard deviation, (c) MTWA, (d) P_{ANN}. Black dots depict locations of proxy records.

After describing the data and process stage models in Sect. 3.2.1 and 3.2.2, the finalized Bayesian frameworks for the European MH and the Siberian LGM reconstructions and their respective inference procedures are presented in Sect. 3.2.3 and 3.2.4.

3.2.1 Data stage

In this section, the data stage in the Bayesian framework Eq. (2.3) is specified. The different types of proxy data require different methods to model the proxy-climate relation. For the pollen and macrofossil synthesis, a probabilistic transfer function is proposed which relates the occurrence of selected taxa to climate. For the local reconstructions of Siberian LGM climate, a Gaussian data stage is implemented, where the inferred local reconstructions and uncertainty estimates are interpreted as mean values and standard deviations of a Gaussian distribution. In both cases, it is assumed that the data stage acts locally. This means that given the climate at location x, a proxy sample at x is conditionally independent of the climate and proxy data at all other locations.

Probabilistic indicator taxa method for pollen and macrofossil data

To reconstruct the European MH climate from the Simonis et al. (2012) synthesis, we use a probabilistic transfer function to model proxy data, in our case occurrence information on taxa, given climate and transfer function parameters for which prior distributions have to be defined. To cali-

brate the parameters, which are denoted by θ, modern vegetation data P_m and climate data C_m are used. The past climate C_p is then reconstructed from fossil pollen and macrofossil data P_p. It is assumed, that the calibration procedure and the reconstruction are only linked through the transfer function parameters. Hence the data stage in Eq. (2.3) has the form

$$\mathbb{P}(P \mid C, \theta) = \underbrace{\mathbb{P}(P_p \mid C_p, \theta)}_{\text{Calibration stage}} \underbrace{\mathbb{P}(P_m \mid C_m, \theta)}_{\text{Observation stage}}, \tag{3.1}$$

where the data is $P = (P_p, P_m)$ and the climate C combines C_p and C_m. Following Parnell et al. (2015), the terms on the right hand side of Eq. (3.1) are called calibration and observation stage. As described above, our main reconstruction target is the bivariate climate $C = (C_1, C_2)$, where C_1 is MTWA and C_2 is MTCO.

We use the PITM model, which is a well established transfer function to quantitatively constrain past climate states from occurrence information on taxa, that are determined from pollen and macrofossil samples. It uses taxa, which are sensitive to MTWA and MTCO, and determines the climatic niche where they occur by fitting response functions. The classical ITM (Iversen, 1944) estimates binary limits, e.g. a taxon occurs above a certain temperature but not below it. PITM, also named pdf method in the literature (Kühl et al., 2002), is an extension of this method where probability distributions are fitted to acknowledge that most taxa have a preferred climate space. However, the transitions between climates where they usually occur and those where they do not grow is soft. Initially, Gaussian distributions were used for calibration (Kühl et al., 2002) against vegetation maps (Schölzel et al., 2002). Later, the model was extended to mixtures of Gaussians (Gebhardt et al., 2008) and quadratic logistic regression (Stolzenberger, 2011, 2017).

We integrate the forward formulation of PITM from Stolzenberger (2017) in our Bayesian framework. For each taxon, we fit a quadratic logistic regression model (response function) describing the probability of taxon occurrence for a given value of C. The idea of using quadratic logistic regression stems from the BIOMOD (BIOdiversity MODelling) software which is a model to predict species distributions (Thuiller, 2003). The regression for taxon T contains linear and quadratic terms for each of the climate variables as well as an interaction term:

$$\mathbb{P}(T = 1 \mid C = (C_1, C_2)) = \operatorname{logit}\left(\beta_1^T + \beta_2^T C_1 + \beta_3^T C_2 + \beta_4^T C_1 C_2 + \beta_5^T C_1^2 + \beta_6^T C_2^2\right). \tag{3.2}$$

Here, logit denotes the logistic function and $\beta_1^T, ..., \beta_6^T$ are regression coefficients. This regression leads to a unimodal response function which is anisotropic but has two symmetry axes, as can be seen for dwarf birch (Betula nana) and European ivy (Hedera helix) in Fig. 3.5.

To fit response functions, vegetation data is used instead of modern pollen or macrofossil data, because it contains more accurate information on the presence or absence of a taxon on the spatial scales of interest compared to modern pollen samples. Moreover, the nature of macrofossils makes it impossible to create a modern calibration dataset that contains pollen and macrofossils. The disadvantage of using vegetation data for the calibration is that the probability of a taxon being present is only valid in vegetation space on the spatial scale taken for the training data. However,

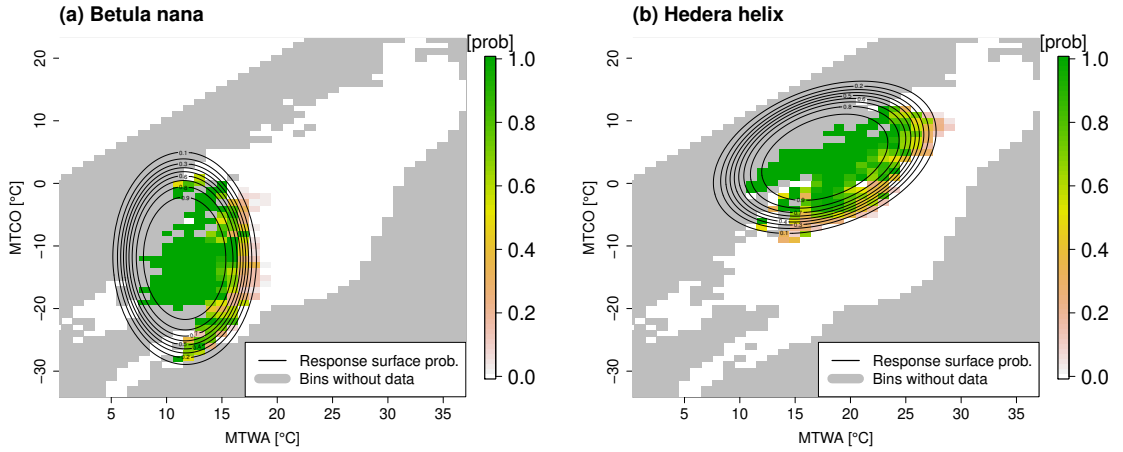

Figure 3.5: Response functions for Betula nana (a) and Hedera helix (b). The relative frequency of occurrence in 1K bins is shown in colors, and the contours depict the probability of presence as estimated by the logistic response function. Gray boxes denote bins without calibration data. In the climate space, combinations of MTWA and MTCO with MTWA < MTCO cannot occur by definition. White bins in the upper left depict artificial absence information added to account for this constraint.

it is not valid in the pollen or macrofossil space. Here, an absence of a taxon in a pollen or macrofossil sample can have multiple non-climatic reasons like local plant competition or pollen transport effects, as well as local climatic effects below the resolution of our reconstruction such that the taxon did not grow in the immediate surrounding of the core location.

For the calibration against the modern dataset, we use presence (T=1) as well as absence (T=0) information on the taxa which can be justified by assuming that the vegetation maps contain accurate information on presence as well as absence of taxa. From the definition given in Sect. 3.1.1, it follows that at any location MTWA is larger or equal than MTCO. Formally incorporating this constraint in the inference leads to a non-linear condition on the regression parameters, which is very hard to implement. Therefore, we choose the more practical way of adding artificial absence information for combinations of MTWA and MTCO such that MTCO > MTWA. While this leads to transfer functions, which do not preclude reconstructions of MTCO values larger than MTWA, it is at least very improbable. To apply the response functions for individual taxa to a set of proxy data, we assume that proxy samples $P(s)$, where $s = 1, ..., S$ subscripts the proxy samples, are conditionally independent given a climate field. In addition, we assume that $P(s)$ conditioned on $C(x_s)$, where x_s is the location of the s-th sample, is independent of the climate at all other locations. This leads to the following probabilistic model for the set of modern vegetation samples

$$\mathbb{P}(P_m \,|\, C_m, \theta) \;=\; \prod_{s=1}^{S_m} \prod_{T \in T(P)} \mathbb{P}\left(P_m^T(s) \,\middle|\, C_m(x_s), \beta_1^T, ..., \beta_6^T\right). \tag{3.3}$$

Here, $P_m^T(s)$ is the presence or absence of taxon T in the s-th calibration sample, $T(P)$ is the set of

all taxa occurring in the fossil pollen and macrofossil synthesis, and $\theta := (\beta_i^T,\ i = 1, ..., 6,\ T \in T(P))$.

As described above, the absence of a taxon in a pollen or macrofossil sample can have reasons that are not included in the absence probability estimated from Eq. (3.3), since this calibration is only valid in the vegetation space. As information on the absence of a taxon in the vegetation space (i.e. absence of the taxon in the grid box of interest at the respective time slice) is not available from pollen and macrofossil data, the only reliable occurrence information of a taxon in the respective grid box in the past is the presence of the taxon in a pollen or macrofossil sample (Gebhardt et al., 2003). Hence, only occurring taxa are included in the reconstruction step.

Violations of the assumption that taxa are treated as conditionally independent given climate, i.e. due to co-occurrence of taxa (Kühl et al., 2002), can lead to over-fitting and subsequently underestimation of uncertainty in the transfer functions. Therefore, a statistical pre-selection of taxa, which are present in a sample, is applied. This procedure uses the Mahalanobis distance (Mahalanobis, 1936) between the fitted distributions and is described in detail in Kühl et al. (2002) and Gebhardt et al. (2008). For the pollen and macrofossil synthesis used in this study, the pre-selection was carried out by Simonis (2009) and we follow his results.

Following these considerations and applying the same transfer function to the fossil samples that is fitted to the modern calibration data, $\mathbb{P}(P_p \,|\, C_p, \theta)$ is given by

$$\mathbb{P}(P_p \,|\, C_p, \theta) := \prod_{s=1}^{S_p} \prod_{T \in T(s)} \mathbb{P}\left(P_p^T(s) \,\middle|\, C_p(x_s), \beta_1^T, ..., \beta_6^T\right), \tag{3.4}$$

where $T(s)$ are the taxa, which occur in sample s and are picked by the pre-selection procedure of Simonis (2009).

Finally, we define a prior distribution for θ. We use a Gaussian distribution centered at 0 and a marginal variance of 10 for each parameter β_i^T. Due to the absence of prior information on the correlation structure, we assume independence between the taxa as well as within a taxon. Hence, the prior is given by

$$\mathbb{P}(\theta = (\beta_i^T,\ i = 1, ..., 6,\ T \in T(P))) = \prod_{T \in T(P)} \prod_{i=1}^{6} \mathcal{N}\left(\beta_i^T \,|\, 0, 10\right). \tag{3.5}$$

Here, $\mathcal{N}$ denotes a Gaussian distribution with mean and variance parameters. Due to the high information content in the calibration data set, the influence of the prior on the parameter estimates is negligible. Using a flat prior for $C_p(x_s)$ and removing spatial correlations, local climate reconstructions at the locations of the proxy samples can be calculated. These reconstructions depend only on the proxy data in grid box x_s. Results of local MH reconstructions for each grid box with proxy data are shown in Fig. 3.6, where the local reconstruction mean values and the marginal 90% credible intervals (CIs) are plotted for MTWA and MTCO.

Local reconstructions can also be used to evaluate the ability of the transfer functions to reconstruct modern climate which provides a reference for possible regional biases. For the PITM model such

evaluations have been performed by Gebhardt (2003), Gebhardt et al. (2008) and Stolzenberger (2011). All three evaluations show that the model tends to underestimate north-south gradients leading to positive biases in Fennoscandia, and slightly negative biases in the Mediterranean. The biases as well as the uncertainties are larger for winter temperature than for summer. Therefore, results for MTCO in Northern Fennoscandia should be treated with caution, while for all other regions biases of the reconstruction means are within reconstruction uncertainties.

A disadvantage of the PITM version used in this study is the inconsistent use of calibration and fossil data by using presence and absence information on taxa for the calibration but only occurring taxa in the reconstruction. Despite this inconsistency, the reconstructions in this study are in agreement with previous versions of PITM, where this inconsistency with the calibration did not appear as previously only occurrence information were used to fit the probability density functions. However, there is no simple solution for the problem that the calibration takes place in vegetation space whereas the absence of taxa in the fossil samples is an information in the pollen or macrofossil space. A promising idea might be to model the absence due to non-climatic reasons as a zero-inflated model by adding a latent variable to estimate the detection probability of a taxon (MacKenzie et al., 2002). But the estimation of detection probabilities is a very challenging task because it depends on many factors like pollen influx area of the fossil sample, local topography, soil properties, and plant competition which might change over time. It is a priori unclear which of these factors can be marginalized and which have to be included as covariates. In addition, the Simonis et al. (2012) synthesis combines macrofossils with pollen data. The processes that influence the detection probability of macrofossils are very different than for pollen. Therefore, a different detection probability has to be estimated for pollen than for macrofossils. Resolving the described issues is an interesting direction for future research, which requires extensive cooperation of (paleo)climatologists, (paleo)botanists, and statisticians, but is beyond the scope of this study.

Gaussian data stage

In the LGM reconstructions for Siberia, the data stage in Eq. (2.3) models the integration of local climate reconstructions from pollen into the spatial reconstruction framework. As described in Sect. 3.1.2, the WAPLS transfer functions infer best estimates C_s^l and corresponding standard errors $\sigma_p(s)$ for proxy samples $s = 1, ..., S$. These values can be interpreted naturally as a Gaussian observation model. Formulated in a forward way, i.e. as the probability of C_s^l given the climate field C_p, the model can be written as

$$\mathbb{P}\left(C_s^l \,\middle|\, C_p(x_s), \sigma_p(s)\right) \;=\; \mathcal{N}\left(C_s^l \,\middle|\, C_p(x_s), \sigma_p(s)^2\right). \tag{3.6}$$

Here, x_s denotes the grid box in which the s-th sample is located. Parts of the uncertainty contained in $\sigma_p(s)$ is shared among proxy samples, e.g. uncertainties that are related to the modern calibration and contribute in the same way to all samples which possess the same calibration dataset.

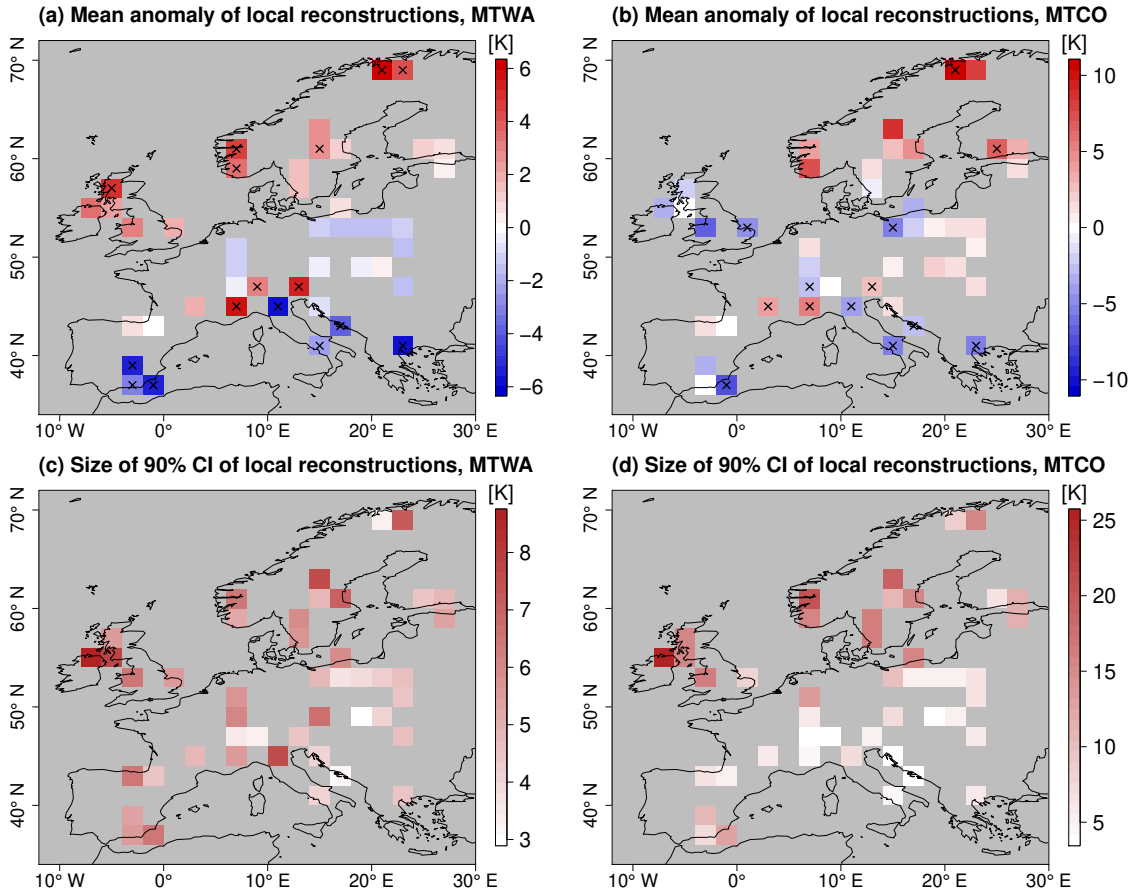

Figure 3.6: Summary statistics of local MH reconstructions using the PITM forward model. Top row: mean anomaly from CRU reference climatology (left: MTWA, right: MTCO), bottom row: uncertainty measured by the size of marginal 90% CIs (left: MTWA, right: MTCO). In the top row, significant anomalies (5% level) are marked by black crosses.

There is currently no quantitative estimate of the fraction of correlated (shared) uncertainty versus independent uncertainty available in the literature. Therefore, a proxy correlation parameter ϱ is introduced to estimate the fraction of correlated uncertainty. As each sample from the same pollen record shares the same calibration dataset, it is assumed that the correlation between samples from the same record is ϱ, while samples from different records are modeled as conditionally independent.

Summarizing, the proxy data P is given by the local reconstructions $P = \left(C_1^l, ..., C_S^l\right)$, and the data stage parameters are given by $\theta = (\sigma_p(1), ..., \sigma_p(S), \varrho)$, where $\sigma_p(1), ..., \sigma_p(S)$ are fixed and the prior distribution of ϱ is chosen as a weakly informative Beta distribution

$$\mathbb{P}(\varrho) = \text{Beta}\left(\tfrac{1}{2}, \tfrac{1}{2}\right). \tag{3.7}$$

A natural way to formulate the complete data stage model is by treating the local reconstructions

$(C_1^l, ..., C_S^l)$ and the spatial climate field $(C_p^1, ..., C_p^N)$ as vectors, and defining an $S \times N$ dimensional observation matrix O which projects the spatial climate field to the observation space. According to Eq. (3.6), this means that O is given by

$$O(s,n) \;=\; \begin{cases} 1 & \text{if } x_s = n, \\ 0 & \text{otherwise,} \end{cases} \tag{3.8}$$

where $x_s = n$ means that the s-th proxy sample is located in the n-th grid box. Next, an $S \times S$ dimensional proxy correlation matrix Ψ_{proxy} is defined following the considerations for ϱ:

$$\Psi_{\text{proxy}}(r,s) \;=\; \begin{cases} 1 & \text{if } r = s, \\ \varrho & \text{if } C_r^l \text{ and } C_s^l \text{ are from the same pollen record,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.9}$$

Here, r and s index the proxy samples. The corresponding proxy covariance matrix Σ_{proxy} is given by

$$\Sigma_{\text{proxy}} \;=\; \text{Diag}\left(\sigma_p(1), ..., \sigma_p(S)\right) \cdot \Psi_{\text{proxy}} \cdot \text{Diag}\left(\sigma_p(1), ..., \sigma_p(S)\right), \tag{3.10}$$

where $\text{Diag}\left(\sigma_p(1), ..., \sigma_p(S)\right)$ is a diagonal matrix with the standard errors of the reconstructions on the diagonal. Using these matrices, the data stage in Eq. (2.3) is given by

$$\mathbb{P}\left(P \,\middle|\, C_p, \theta\right) \;=\; \mathbb{P}\left(C_1^l, ..., C_S^l \,\middle|\, C_p, \varrho\right) \;=\; \mathcal{N}\left(\left(C_1^l, ..., C_S^l\right) \,\middle|\, O \cdot \left(C_p^1, ..., C_p^N\right), \Sigma_{\text{proxy}}\right). \tag{3.11}$$

3.2.2 Process stage

Similar to DA approaches in numerical weather and climate prediction, the ensemble of climate simulations is used to control the spatial structures of the reconstruction and to constrain the range of physically possible climate states for a given external forcing. This is achieved by computing a spatial prior distribution from the ensemble members. This distribution is combined with interpolation parameters ϑ to facilitate a more flexible adjustment to the proxy data in the process stage of Eq. (2.3). The estimation of the prior distribution is hampered by the small number of ensemble members with $K = 7$ for the MH and $K = 4$ for the LGM.

It is not obvious which method for estimating the prior distribution is best suited for the problem on hand and which additional model parameters are appropriate to preserve as much physical consistency contained in the climate simulations as possible but to correct for climate model inadequacies. Therefore, we perform a comparison study of six process stage models, that are composed of three techniques to formulate the process stage and two choices for the involved spatial covariance matrix. All those models are inspired by methods used in DA, postprocessing of forecasts, and climate change detection and attribution. The use of climate simulations in the process stage allows not just interpolation between proxy samples but also structural extrapolation through the eigenvectors of the spatial covariance matrix and the process stage parameters.

In the following, we first present the three techniques to formulate the process stage, which we call Gaussian, regression, and kernel models. Each of these techniques uses a spatial covariance matrix. The two methods to estimate this covariance matrix are introduced, thereafter. They are named glasso-based and shrinkage-based covariance matrices according to the involved regularization technique. Each process stage model combines one technique to formulate the process stage with one method to estimate the spatial covariance matrix.

Gaussian model

The most common approach in the DA literature is to assume that the ensemble members are independent and identically distributed (iid) samples from an unknown Gaussian distribution of possible climate states (Carrassi et al., 2018). In the following, the climatological means of the K ensemble members are denoted by μ_k. Subsequently the spatial prior distribution is given by $\mathcal{N}(\bar{\mu}, \Sigma_{\text{prior}})$, where $\bar{\mu}$ is the sample mean $\frac{1}{K}\sum_{k=1}^{K}\mu_k$, and Σ_{prior} is a spatial covariance matrix, which is given by a regularized version of the empirical covariance

$$\Sigma_{\text{emp}} \;=\; \frac{1}{K-1}\sum_{k=1}^{K}(\mu_k-\bar{\mu})\;(\mu_k-\bar{\mu})^t\,, \tag{3.12}$$

The superscript t denotes the matrix transpose. The covariance matrix Eq. (3.12) is based on the inter-model differences, which is an estimate of epistemic uncertainties. The regularization techniques of Σ_{emp} are specified below. The Gaussian distribution is multivariate and the state vector has dimension N, where N is the number of grid boxes times the number of jointly reconstructed variables.

The main advantage of this Gaussian model (GM) is that inference becomes simpler than in more complex probability density estimation techniques because the prior distribution is unimodal and Gaussian. The disadvantage is that it relies on the strong assumption that the samples μ_k are iid samples from an unknown Gaussian distribution. This assumption tends to be more realistic for samples from just one ESM, whereas statistics of multi-model ensembles are often not well described by purely Gaussian distributions (Knutti et al., 2010). A second disadvantage of this model is that the absence of additional parameters limits the possibilities to adjust the posterior distribution to the proxy data and correct for climate simulation inadequacies. The third disadvantage of the GM is that spatial structures of the individual models, which are directly derived from the physical equations solved in the ESMs, get lost by averaging over all ensemble members. Nevertheless, in many climate prediction applications multi-model averages outperformed each individual ESM (Krishnamurti et al., 1999, e.g.).

Regression model

A relaxation of the assumptions of the GM is the second model, that we call the regression model (RM) because it is inspired by regression-based models popular in postprocessing and climate

change detection and attribution (Hegerl and Zwiers, 2011). In the RM, the iid assumption is dropped for the first moments of the process stage by introducing weighted averages of the ensemble members with variable weights λ_k, $k = 1, ..., K$. As a result, samples, which fit better to the proxy data, are weighted higher in the posterior. The sum of the weights is set to one such that unrealistically warm or cold states are prevented. This leads to the process stage model

$$\mathbb{P}(C_p|\lambda_1, ..., \lambda_K) = \mathcal{N}\left(C_p \,\middle|\, \sum_{k=1}^{K} \lambda_k \mu_k, \Sigma_{\text{prior}}\right), \tag{3.13}$$

and an additional prior distribution for the model weights

$$\mathbb{P}(\lambda) = \text{Dir}\left(\lambda_1, ..., \lambda_K \,\middle|\, \tfrac{1}{2}, ..., \tfrac{1}{2}\right). \tag{3.14}$$

Dir denotes a Dirichlet distribution, which is chosen as prior distribution because it guarantees that the weights take values between 0 and 1 and sum up to 1. Conditioned on the ensemble member weights, the process stage distribution is Gaussian, but non-Gaussianity is permitted through the variable weights.

In addition, the RM has the advantage of possessing more degrees of freedom compared to the GM. The inference process becomes a little more involving than for the GM because the ensemble member weights have to be estimated, too, but the conditional Gaussian distribution of C_p helps designing efficient inference algorithms. Similar to the GM, spatial structures of the individual models can get lost, as we average over different ESM climatologies. However, the improved predictive power of weighted averages of multi-model ensembles compared to each individual ensemble member, diagnosed in many prediction studies (e.g. Krishnamurti et al., 1999), could justify this reduction of physical consistency.

Kernel model

The third model has been introduced in the DA literature by Anderson and Anderson (1999) to combine particle and Gaussian filtering approaches (Carrassi et al., 2018), where the particle part helps capturing non-Gaussian and non-linear features but the efficiency of Gaussian approximations in high dimensional filtering situations is still exploited. This kernel model (KM) assumes that each ensemble member is a sample from an unknown distribution of possible climate states given a set of forcings, but it does not assume that this unknown distribution is Gaussian. Instead, non-parametric kernel density estimation techniques (Silverman, 1986), where the probability distribution is given by a mixture of multivariate Gaussian kernels, are used. Each ensemble member climatology corresponds to the mean of a kernel.

Ideally, the covariance matrix of each kernel would correspond to the respective ESM, such that the spatial autocorrelation of that ESM is preserved when we sample from its kernel. Unfortunately, there is only one MH run available for each ESM and the internal variability in those runs is much smaller than the inter-model differences. Using the internal variability of those runs would thus lead to very distinct kernels and allow only a small range of climate states. Therefore, the

covariance of each kernel is estimated from the inter-model differences as a compromise that allows to sample from a much broader range of states even though autocorrelation of the individual models is lost. This compromise is a very common choice in kernel-based probability density approximations (Liu et al., 2016; Silverman, 1986, Chapter 3 and 4) if there is no good estimate available for the covariance corresponding to each kernel.

Compared to the GM, the empirical covariance matrix Σ_{emp} is scaled by the Silverman factor (Silverman, 1986)

$$f = \left(\frac{4}{K\cdot(N+2)}\right)^{\frac{2}{N+4}}, \tag{3.15}$$

which optimizes the variances of the kernels by empirically approximating the minimizer of the mean integrated square deviation from an unknown true Gaussian density. Hence in the KM, the scaled empirical covariance matrix $\tilde{\Sigma}_{\text{emp}}$ given by $f \cdot \Sigma_{\text{emp}}$ is regularized leading to the spatial covariance matrix $\tilde{\Sigma}_{\text{prior}}$. Note that the small number of ensemble members compared to the dimension of the probability distribution leads to a standard deviation reduction of only around 2% in our applications. This can be seen by inserting the dimension the climate vector and the number of ensemble members into Eq. (3.15).

Each kernel gets an assigned weight ω_k, $k = 1, ..., K$, which is inferred in the Bayesian framework. The weights sum up to one. The resulting process stage is a mixture distribution

$$\mathbb{P}\left(C_p \,|\, \omega_1, ..., \omega_K\right) = \sum_{k=1}^{K} \omega_k \, \mathcal{N}\left(C_p \,|\, \mu_k, \tilde{\Sigma}_{\text{prior}}\right). \tag{3.16}$$

A Dirichlet distributed prior is used for ω with parameter $\frac{1}{2}$ for each of the K components.

A computational disadvantage of the KM is that the process stage is multi-modal and non-Gaussian. We augment the model by an additional parameter z, which follows a categorical distribution, denoted by Cat. z selects a kernel k according to its weight ω_k, i.e. z is defined such that

$$\mathbb{P}(\omega) = \text{Dir}\left(\omega_1, ..., \omega_K \,|\, \tfrac{1}{2}, ..., \tfrac{1}{2}\right) \tag{3.17}$$

$$\mathbb{P}(z \,|\, \omega) = \text{Cat}\left(z_1, ..., z_K \,|\, \omega_1, ..., \omega_K\right) \tag{3.18}$$

$$\mathbb{P}(C_p \,|\, z) = \prod_{k=1}^{K} \left(\mathcal{N}\left(C_p \,|\, \mu_k, \Sigma_{\text{prior}}\right)\right)^{z_k}. \tag{3.19}$$

This restores that C_p is Gaussian conditioned on ω and z. Integrating out z yields the mixture distribution Eq. (3.16).

Two advantages of the KM are that it is not assumed that the unknown prior distribution is Gaussian and that the kernels do not rely on an iid assumption for their first moment properties. On the other hand, the KM relies on an iid assumption for the second moment properties, which is the compromise described above due to the absence of a suitable estimate of the uncertainty structures corresponding to each ensemble member. The KM preserves the spatial structures of each ESM in the first moments of the kernels and when sampling from Eq. (3.19), the mean of each sample belongs to one ESM and is not a weighted average over all ensemble members. This preservation

of physical consistency reduces the degrees of freedom compared to the RM. For example, if the true climate state lies exactly between μ_1 and μ_2 and far away from all other ensemble members, the weights of μ_1 and μ_2 can be increased compared to the other members in the KM, but the mode cannot be changed to $\frac{1}{2}(\mu_1 + \mu_2)$, which is possible in the RM. Another disadvantage of the KM is the more challenging design of efficient inference algorithms due to its multi-modality.

Glasso-based covariance matrices

The first technique to regularize the empirical covariance matrix (the scaled empirical covariance in the KM), which is applied in this study, is the graphical lasso algorithm (glasso, Friedman et al., 2008, implemented in the R-package glasso). This algorithm approximates the precision matrix (inverse covariance) by a positive definite, symmetric, and sparse matrix $\Sigma_{\text{prior}}^{-1}$. Therefore, Σ_{prior} is a valid N-dimensional covariance matrix. Glasso maximizes the penalized log-likelihood

$$\log\det\Sigma_{\text{prior}}^{-1} - \text{trace}(\Sigma_{\text{emp}}\,\Sigma_{\text{prior}}^{-1}) - \xi\,\|\Sigma_{\text{prior}}^{-1}\|_1, \tag{3.20}$$

where ξ is the penalty parameter, $\|\cdot\|_1$ is the vector L_1-norm, and the first two terms are the Gaussian log-likelihood. Because applying the glasso algorithm is computationally expensive, it is not feasible to formally include ξ in the Bayesian framework. Instead a suitable value of ξ has to be determined prior to the inference. In this study, ξ is chosen such that Σ_{prior} is a numerically stable covariance matrix and the performance in CVEs is optimized. Technical details of the determination of ξ are described in Appendix B.1.

The advantage of the glasso approach is that the empirical covariance matrix can be approximated very closely and the sparseness of the precision matrix facilitates the use of efficient Gaussian Markov random field (GMRF) techniques (Rue and Held, 2005) in the inference algorithm. On the other hand, the regularization does not add new spatial structures to Σ_{emp}. The effective number of spatial modes is much smaller than the dimension of the climate vector, which can lead to a collapse onto a very small subspace of the N-dimensional state space and subsequently biases and under-dispersion.

Shrinkage-based covariance matrices

To overcome the deficiencies of the glasso approach, we propose an alternative covariance regularization technique, which combines the empirical correlation matrix of the climate simulation ensemble with a regular correlation matrix. A so-called shrinkage approach (Hannart and Naveau, 2014) is used, which is a weighted average of the empirical correlation matrix and a reference matrix, which in our case contains additional spatial modes computed from a geostatistical model. Thereby, the effective number of spatial modes in the covariance matrix is increased. This allows deviations from the spatial structures prescribed by the climate simulation ensemble. Thus, it is a strategy to account for climate model inadequacies.

Let Ψ_{emp} be the empirical correlation matrix of the climate simulation ensemble, which is related to Σ_{emp} by

$$\Sigma_{\text{emp}} = \text{Diag}(\Sigma_{\text{emp}})^{\frac{1}{2}} \, \Psi_{\text{emp}} \, \text{Diag}(\Sigma_{\text{emp}})^{\frac{1}{2}}, \tag{3.21}$$

where $\text{Diag}(\Sigma_{\text{emp}})$ denotes a diagonal matrix with the same diagonal entries as Σ_{emp}, and the exponent $\frac{1}{2}$ means that the square root of each diagonal entry is taken. Replacing Ψ_{emp} by a weighted average of Ψ_{emp} and a shrinkage target Φ leads to the shrinkage covariance matrix

$$\Sigma_{\text{prior}} = \text{Diag}(\Sigma_{\text{emp}})^{\frac{1}{2}} \left(\alpha \, \Psi_{\text{emp}} + (1 - \alpha) \, \Phi\right) \text{Diag}(\Sigma_{\text{emp}})^{\frac{1}{2}}. \tag{3.22}$$

α is the weighting parameter, which takes values between zero and one. We compute Φ from a numerically efficient GMRF approximation of a stationary Matérn correlation matrix on a regular lat-lon grid (Lindgren et al., 2011). Matérn correlation functions correspond to diffusive transport of spatial white noise forcing (Whittle, 1954, 1963). The Matérn correlation matrix is controlled by three parameters, the smoothness, the range ρ, and the anisotropy ν. We fix the smoothness at a value which corresponds to the application of a standard Laplace operator. ρ controls the decorrelation length, and ν parameterizes the length of the meridional compared to the zonal decorrelation length. The implementation of the GMRF approximation for a single climate variable is described in detail in Appendix A. For joint reconstructions of multiple climate variables, independent correlation matrices for each variable are combined in a block structure, but with different parameters ρ and ν for each climate variable.

Ideally, the parameters α, ρ, and ν are estimated from the proxy data. But initial tests with weakly informative prior distributions, i.e. prior distributions such that the posterior distribution is mostly influenced by the proxy data, showed that the parameters could not be constrained by the available proxy data, because the signal in the proxy data is not strong enough to infer second moment properties of the process stage. Therefore, an ensemble of parameter combinations is created from fitting the shrinkage model to each of the climate simulation ensemble members given all other members. This results in seven consistent sets of α, ρ, and ν for the MH and four for the LGM. Those are passed to the reconstruction framework, such that each parameter set is chosen with a probability inferred from the proxy data. Thereby, each parameter combination is based on a fit against physically consistent structures, and uncertainty in the parameters is included in the inference, but the problem of non-identifiability of the parameters from proxy data alone is reduced. The parameter estimates are given in Table 3.2. They strongly depend on the chosen ESM, which leads to combinations that cover a wide range of possible values.

The main advantage of the shrinkage approach over the glasso-based matrices is the inclusion of more spatial modes in the covariance matrix. Thereby, the collapsing of the reconstruction onto a very low-dimensional subspace is mitigated and stronger deviations of the reconstruction from the climate simulation ensemble are facilitated. A disadvantage compared to the glasso estimated covariance matrices is that neither the shrinkage covariance matrix Σ_{prior} nor its inverse are sparse matrices such that the numerical inference algorithms are more costly.

Table 3.2: Parameter values for the shrinkage weight α, the anisotropy ν, and the range ρ from fitting to the MH and LGM ensemble members. α can take values between 0 and 1, ν quantifies the stretching in zonal versus meridional length (taking values between 0 and 90 degree), and ρ is approximately the 0.1 decorrelation length of the isotropic model. For the MH, α is fitted jointly for MTWA and MTCO.

MH Model	α	MTWA		MTCO	
		ν [deg]	ρ [°]	ν [deg]	ρ [°]
CCSM4	0.059	39.27	43.12	25.51	8.81
CNRM-CM5	0.596	29.43	15.81	10.60	11.60
CSIRO-Mk3-6-0	0.037	21.90	19.70	30.36	12.40
EC-Earth-2-2	0.063	40.98	22.96	17.47	7.69
HadGEM2-CC	0.157	30.18	24.85	21.27	10.33
MPI-ESM-P	0.626	18.51	23.79	22.69	9.54
MRI-CGCM3	0.346	21.14	31.54	21.88	12.20

LGM Model	MTWA			P_{ANN}		
	α	ν [deg]	ρ [°]	α	ν [deg]	ρ [°]
CCSM4	0.081	11.57	16.94	0.015	33.88	5.57
CNRM-CM5	0.065	18.78	6.07	0.006	28.42	3.25
MPI-ESM-P	0.801	22.12	12.23	0.250	22.95	22.78
MRI-CGCM3	0.858	9.25	16.99	0.029	28.55	6.99

3.2.3 Final framework and inference strategy for the mid-Holocene reconstructions

Putting together the PITM data stage from Sect. 3.2.1 and the process stage models from Sect. 3.2.2, the posterior distribution Eq. (2.3) of C_p and Θ given P_p, P_m, and C_m is given by

$$\underbrace{\mathbb{P}(C_p, \Theta \mid P_p, C_m, P_m)}_{\text{Posterior}} \propto \mathbb{P}(P_m \mid C_m, \theta) \quad \mathbb{P}(P_p \mid C_p, \theta) \quad \mathbb{P}(C_p \mid \vartheta) \quad \mathbb{P}(\vartheta) \quad \mathbb{P}(\theta). \tag{3.23}$$

Here, $\mathbb{P}(P_m \mid C_m, \theta)$ is defined in Eq. (3.2, 3.3), $\mathbb{P}(P_p \mid C_p, \theta)$ is given by Eq. (3.2, 3.4), $\mathbb{P}(C_p \mid \vartheta)$ and $\mathbb{P}(\vartheta)$ are defined according to the GM, RM, or KM, and $\mathbb{P}(\theta)$ is defined in Eq. (3.5). The MH framework Eq. (3.23) is visualized as DAG in Fig. 3.7.

Because PITM is non-Gaussian and non-linear, the posterior climate does not belong to a standard probability distribution. Therefore, MCMC techniques are used to asymptotically sample from the correct posterior distribution. These samples allow analyses beyond summary statistics like means and standard deviations. A Metropolis-within-Gibbs strategy is used. This means that we sample sequentially from the full conditional distributions (i.e. the distribution of the respective variable given all other variables) of θ, ϑ, and C_p in each update of the Markov chain. This strategy is chosen because for many variables the full conditional distributions follow probability distributions

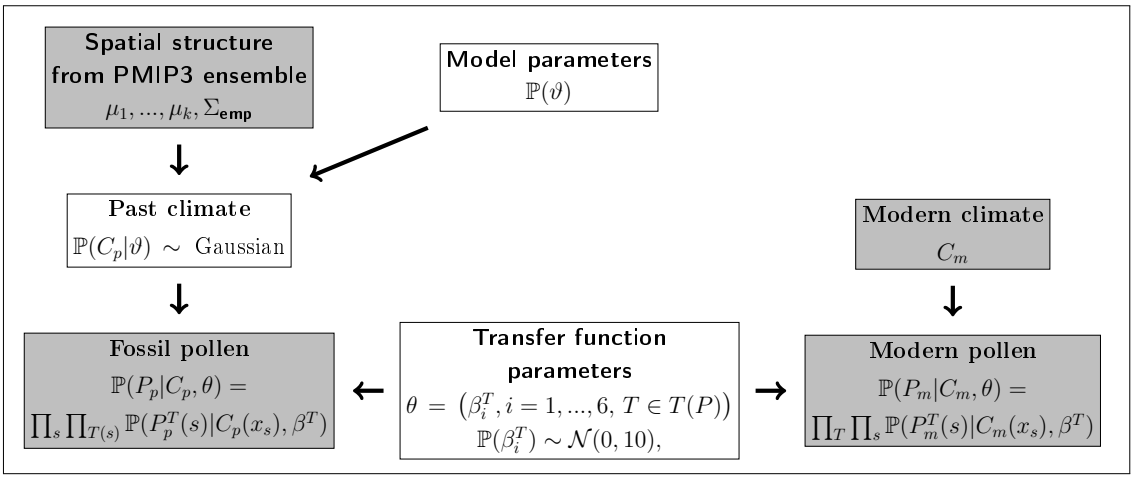

Figure 3.7: DAG corresponding to the Bayesian framework Eq. (3.23). Involved quantities are given by nodes and arrows indicate dependencies of variables. Details of the formulas are explained in Sect. 3.2.2 and 3.2.1. White: inferred quantities; gray: input data.

for which efficient sampling algorithms exist, and for the remaining variables Metropolis-Hastings updates are used for sampling (Robert and Casella, 2004). Metropolis-Hastings updates asymptotically sample from the correct full conditional distribution (Robert and Casella, 2004).

To sample the regression parameters θ in Eq. (3.2) to (3.4) efficiently, the data augmentation scheme of Polson et al. (2013) is used. For taxon T, the full conditional is only depending on C_p, C_m, P_p^T, and P_m^T, but not on other taxa. Therefore, we can sample $\beta_1^T, ..., \beta_6^T$ independently from the other taxa. Polson et al. (2013) introduce auxiliary variables $\gamma_l^T, l = 1, ..., L$, where L is the number of observations of taxon T, such that $\mathbb{P}(\gamma_l^T | \beta_1^T, ..., \beta_6^T, C_m, C_p)$ is Pólya-Gamma (PG) distributed, and $\mathbb{P}(\beta_1^T, ..., \beta_6^T | P_m^T, P_p^T, \gamma_1^T, ..., \gamma_L^T)$ is Gaussian. Therefore, the MCMC algorithm samples alternately from a PG distribution using the sampler of Windle et al. (2014) and from a Gaussian distribution. The PG sampler is implemented in the R package BayesLogit (Windle et al., 2013).

To sample from the full conditional of C_p, we separate the grid boxes x_P with at least one proxy record from those without any proxy records denoted by x_Q. There is no closed form available for the full conditionals of $C_p(x_P)$. Therefore, we use a random walk Metropolis-Hastings (RW-MH) algorithm to update $C_p(x_P)$ sequentially for all members of x_P. As the transfer functions act locally, $C_p(x_Q)$ is conditionally independent of P_p given $C_p(x_P)$ and ϑ. Therefore, we subsequently update $C_p(x_Q)$ by sampling from $\mathbb{P}\left(C_p(x_Q) | C_p(x_P), \vartheta\right)$ which is Gaussian.

Sampling from ϑ depends on the particular process stage model. In models with shrinkage covariance matrix Σ_{prior} given by Eq. (3.22), α, ρ, and ν are sampled from the K parameter sets in a Metropolis-Hastings step. The weights λ in the RM are sampled from a RW-MH update. In the KM, Eq. (3.17) to (3.19) lead to full conditionals for ω and z, which are again Dirichlet and categorically distributed but with updated parameters. Therefore, Gibbs sampling can be used to update ω and z.

The multi-modality of the KM makes inference for this model a lot more challenging than for the GM and RM. The problem of efficient MCMC algorithms for multi-modal posterior distributions is a widely acknowledged issue in the literature (Tawn and Roberts, 2018) and in this study the Metropolis coupled Markov chain Monte Carlo algorithm (MC^3; Geyer, 1991), which is also known as parallel tempering, is used to overcome this issue. Details of this procedure are provided in Appendix B.3.

To speed up the inference, grid boxes with proxy data and those without proxy data are treated sequentially. Because of the conditional Gaussian structure of the process stage and because the grid boxes without proxy data are not influencing the posterior of Θ, $C_p(x_Q)$ is integrated out to get an estimate of the joint distribution of Θ and $C_p(x_P)$. In a second step, we sample from $C_p(x_Q)$ conditioned on $C_p(x_P)$ and Θ, which leads to joint samples of C_p from the asymptotically correct posterior distribution. This strategy reduces the computation time of each MCMC update due to faster matrix operations.

The remaining bottleneck in computation time is the estimation of the transfer function parameters due to the large modern calibration set. While in theory the observation stage influences the updates of θ, in practice the influence of Eq. (3.4) on the posterior of θ is negligible. Therefore, a modularization approach (Liu et al., 2009a) is used similar to Parnell et al. (2015) in CVEs, where a sequence of reconstructions with subsets of the proxy samples is computed. This means that feedback between Eq. (3.3) and Eq. (3.4) is cut by first drawing as many MCMC samples as necessary from θ using only Eq. (3.3). Thereafter, C_p is reconstructed using these samples instead of sampling θ from its full conditional.

Detailed formulas for the full conditional distributions are given in Appendix B.2. Pseudo-code for the MCMC and MC^3 algorithms is provided in the Supplement. For a 798 dimensional climate posterior as it is the case in joint reconstructions of MTWA and MTCO, and 45 grid boxes that contain at least one proxy record, 75,000 MCMC samples are created. The first 25,000 samples are discarded as burn-in. To reduce the autocorrelation of subsequent samples, every fifth sample is extracted to create a set of 10,000 posterior samples which is used for further analyses. On a standard desktop computer, reconstructions with the modularized model can be computed in approximately 30 minutes. The convergence of MCMC variables is checked using the Gelman-Rubin-Brooks criterion (Brooks and Gelman, 1998) implemented in the R package coda (Plummer et al., 2006).

3.2.4 Final framework and inference strategy for the Last Glacial Maximum reconstructions

For the LGM reconstructions, only the RM with shrinkage covariance matrix is used due to properties of the LGM PMIP3 ensemble members. The anomalies of the ensemble member climatologies are spatially relatively homogeneous but the magnitude of MTWA cooling exhibits large differences between the simulations. Additionally, the ensemble is very small with only four members.

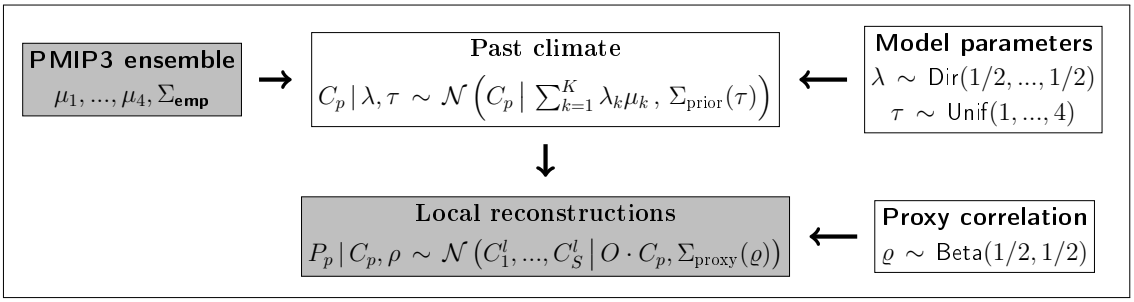

Figure 3.8: DAG corresponding to the Bayesian framework Eq. (3.24). Involved quantities are given by nodes and arrows indicate dependencies of variables. Details of the formulas are explained in Sect. 3.2.2 and 3.2.1. White: inferred quantities; gray: input data.

The RM is the most flexible process stage formulation such that the magnitude of MTWA cooling and P_{ANN} drying can be adapted best to the proxy data. The empirical covariance matrix Σ_{emp} is dominated by a very small number of spatial modes. Hence, the shrinkage covariance matrix approach is chosen, which adds effective degrees of freedom to the empirical covariance matrix and facilitates a better reconstruction of regional heterogeneity. Subsequently, the posterior distribution respects the proxy data better than in the glasso approach. Combining the shrinkage-based RM with a Gaussian process stage, leads to the following posterior distribution:

$$\mathbb{P}\left(C_p, \Theta \,\middle|\, P = \left(C_1^l, ..., C_S^l\right)\right) \quad \propto \quad \mathbb{P}\left(C_1^l, ..., C_S^l \,\middle|\, C_p, \varrho\right) \quad \mathbb{P}\left(C_p \,|\, \lambda, \tau\right) \quad \mathbb{P}(\lambda, \tau) \;\; \mathbb{P}(\varrho)\,. \tag{3.24}$$

Here, τ parameterizes the four shrinkage matrix parameter sets, $\mathbb{P}\left(C_1^l, ..., C_S^l \,|\, C_p, \varrho\right)$ is given by Eq. (3.11), $\mathbb{P}\left(C_p \,|\, \lambda, \tau\right)$ is defined in Eq. (3.13), and the prior distributions $\mathbb{P}(\lambda, \tau)$ and $\mathbb{P}(\varrho)$ follow Eq. (3.14) and (3.7). For the P_{ANN} reconstruction, the posterior probability mass of C_p below -100% is accumulated at -100%, since otherwise the anomalies would lead to negative precipitation values. A DAG, that visualizes the LGM framework Eq. (3.24), is provided in Fig. 3.8.

The inference strategy is similar to the reconstructions for the MH, but the Gaussian data stage facilitates simpler and more efficient MCMC sampling. We sample sequentially from the full conditional distributions of ϱ, λ, τ, and C_p in each step of the Markov chain. The full conditionals of ϱ, λ, and τ do not follow probability distributions for which efficient samplers exist. This is why Metropolis-Hastings steps are used to update those parameters. Conditioned on Θ, C_p is Gaussian distributed, such that Gibbs sampling can be used to update C_p. Similar to the MH framework, the climate vector is split into a portion $C_p(x_P)$ of grid boxes which contain at least one proxy sample, and a portion $C_p(x_Q)$ of grid boxes without proxy data. First, MCMC samples for $C_p(x_P)$ and Θ are drawn, and afterwards we sample $C_p(x_Q)$ conditioned on $C_p(x_P)$ and Θ. This does not change the posterior as described above and makes the MCMC algorithm a lot faster. The truncation of C_p in P_{ANN} reconstructions is executed by setting all values below -100%, i.e. negative precipitation amounts, in the MCMC samples to -100% which creates a point mass at $C_p(\cdot) = -100\%$. The

full conditional distributions of C_p, λ, τ, and ϱ as well as pseudo-code of the MCMC algorithm are given in Appendix B.2.
Convergence of the MCMC chains is checked using the Gelman-Rubin-Brooks criterion (Brooks and Gelman, 1998). The Gibbs sampling of C_p leads to sample autocorrelations of C_p near zero. 75.000 MCMC samples are created, from which the first 25.000 are discarded as burn-in, and every fifth is used to reduce the autocorrelation further. This leads to 10.000 samples of the posterior distribution. Drawing 75.000 samples takes approximately two hours on a standard desktop computer.

3.3 Comparison of process stage models

In this section, each combination of the three process stage formulations (GM, RM, KM) and the two covariance regularization approaches (glasso, shrinkage) is compared using the MH reconstruction framework (Sect. 3.2.3) and the MH proxy data (Sect. 3.1.1) and climate simulation ensemble (Sect. 3.1.3). Two different types of experiments are used to evaluate and compare the models. ITEs use the multi-model ensemble by simulating pseudo-proxy data from one ensemble member and trying to reconstruct the reference climatology from the simulated proxies and the remaining ensemble members. These experiments facilitate the understanding of different modeling approaches for the process stage in a controlled environment. In particular, the ability of the Bayesian frameworks to reconstruct the climate can be evaluated without having to rely on indirect observations as it is the case in real paleoclimate applications where the true climate state is unknown. The second type of experiments are CVEs, where spatial reconstructions with the Simonis et al. (2012) synthesis are performed but the samples from one grid box are left out. Then, the reconstructions for this grid box are evaluated against the left-out sample in the vegetation space by applying the PITM forward model to the reconstruction. The advantage of these experiments is that the models are compared in a real-world setting. The disadvantage of CVEs is that the target of a reconstruction is the climate state but there are no direct observations of paleoclimate available such that evaluations against data have to be indirect. Evaluating prediction models against indirect observations in the observation space through forward modeling is also a recent way of model skill evaluation in weather forecasting (e.g. Di Michele et al., 2012, for evaluation against indirect satellite observations).

3.3.1 Identical twin experiments

Experiment design

The ITEs make use of the fact that the PMIP3 ensemble is a multi-model ensemble such that the models can produce fairly different spatial structures. Therefore, trying to reconstruct the climate state of one ESM given the others is a more realistic test environment than doing the same with single-model ensembles, i.e. when all ensemble members are run with the same ESM but varied

parameterizations, initial conditions, or boundary conditions. The first step in an ITE is to choose a reference climate model with climate state C_p^{true}. Then, for each grid box with proxy samples from the Simonis et al. (2012) synthesis (denoted by x_P in accordance with the notation from Sect. 3.2.3), pseudo-proxies are simulated. The proxies follow a Gaussian approximation of the uncertainty structure of the local reconstructions depicted in Fig. 3.6. This means that a bivariate error covariance matrix $\Sigma_p^{x_s}$ is estimated for each grid box x_s with proxy data. The pseudo-proxies are assumed to be unbiased such that they are given by

$$P_{x_s} \sim \mathcal{N}\left(C_p^{\text{true}}(x_s), \Sigma_p^{x_s}\right), \quad x_s \in x_P. \tag{3.25}$$

Using unbiased Gaussian pseudo-proxies is a common strategy to test CFR techniques (e.g. Gomez-Navarro et al., 2015). An alternative would be to simulate pseudo-proxies with PITM, but the problem is that applying the forward model and then inverting it directly is only valid in the vegetation space whereas the data of this study are pollen and macrofossils, such that the statistical taxa pre-selection procedure would require extensive retuning. In addition, using unbiased Gaussian proxies facilitates a direct comparison of the process stage models without having to decompose biases into transfer functions effects and inadequacies of the interpolation method.
With the simulated proxies, a probabilistic spatial reconstruction is computed with the Bayesian framework Eq. (3.23), but the reference model climatology is removed from the climate simulation ensemble. This means that only $K - 1$ ensemble members are used to calculate the empirical covariance matrix Σ_{emp} and the ensemble mean in the GM, the weighted ensemble mean in the RM, or the kernel means in the KM, and to estimate the shrinkage matrix parameters. Biases, probabilistic reconstruction skill, and dispersiveness of the reconstructions are evaluated. ITEs are performed with each of the seven PMIP3 ensemble members as reference climate state and the six different process stage models. For each combination, five randomized ITEs are computed to separate random from systematic issues.

Metrics for ITE evaluation

To study the bias, the mean deviation of the reconstruction from the reference climatology is evaluated, either spatially averaged or its spatial distribution. The dispersiveness is studied by calculating coverage frequencies for 50% CIs and 90% CIs. This means that the frequency of the reference climate state to be included in the respective CIs is computed. Again, the overall coverage frequency for each of the six process stage models is analyzed as well as the spatial structures. The target frequencies are 50% for the 50% CIs and 90% for the 90% CIs. For the 50% CIs, coverage frequencies below 50% indicate under-dispersiveness, whereas values above 50% indicate over-dispersion. The same holds for 90% CIs.
A summary metric for probabilistic predictions, in the case of CFRs the ability of a reconstruction method to probabilistically predict past climate from sparse and noisy proxy data, is the continuous ranked probability score (CRPS). It is a strictly proper scoring function that generalizes the

absolute error to probabilistic forecasts (Matheson and Winkler, 1976). Given a reference climate state C_p^{true}, it is defined for each grid box x by

$$\text{CRPS}\left(F, C_p^{\text{true}}(x)\right) \;=\; \int_{-\infty}^{\infty} \left(F(y) - \mathbb{1}\{y \geq C_p^{\text{true}}(x)\}\right)^2 \, dy, \tag{3.26}$$

where F is the cumulative distribution function (cdf) of the probabilistic reconstruction C_p at grid box x. Defined in that way, the CRPS has a unique minimum at 0 and is positive unless F is a perfect prediction. Rewriting Eq. (3.26) as

$$\text{CRPS}\left(F, C_p^{\text{true}}(x)\right) \;=\; \mathbb{E}_F \left|X - C_p^{\text{true}}(x)\right| - \tfrac{1}{2}\, \mathbb{E}_F \left|X - Y\right|, \tag{3.27}$$

where $\mathbb{E}_F$ is the expected value with respect to F and X and Y are independent copies of random variables with cdf F, one can see that the CRPS is the mean absolute error of the prediction adjusted by the prediction uncertainty (Baringhaus and Franz, 2004; Gneiting and Raftery, 2007). Eq. (3.27) offers a natural way to estimate the CRPS from a set of MCMC samples by replacing the expectations with averages over all samples:

$$\overline{\text{CRPS}}\left(C_p(x), C_p^{\text{true}}(x)\right) \;:=\; \frac{1}{I} \sum_{i=1}^{I} \left|C_p^i(x) - C_p^{\text{true}}(x)\right| - \tfrac{1}{I(I-1)} \sum_{i=1}^{I} \sum_{j=i+1}^{I} \left|C_p^i(x) - C_p^j(x)\right|, \tag{3.28}$$

Here, $C_p(x)$ denotes the probabilistic reconstruction at x, and C_p^i, $i = 1, ..., I$, are the corresponding MCMC samples.

Results

Table 3.3: Summary measures for ITEs and CVEs with the six process stage models described in Sect. 3.2.2: For ITEs, mean deviation from reference climatology, mean size of grid box-wise 50% and 90% CIs, coverage frequency of 50% and 90% CIs, and mean CRPS are stated (see explanations in Sect. 3.3.1). In the last column, the mean BS of CVEs is provided (see explanations in Sect. 3.3.2).

Configuration	Mean deviation [K]	50% CI size [K]	90% CI size [K]	50% cov. freq. [%]	90% cov. freq. [%]	CRPS [K]	BS
GM glasso	0.00	1.19	2.91	29.2	64.1	1.03	0.186
RM glasso	0.00	1.19	2.92	29.6	64.4	1.03	0.186
KM glasso	-0.03	1.24	3.03	30.5	66.1	1.01	0.187
GM shrinkage	0.03	1.58	3.88	79.9	99.6	0.40	0.165
RM shrinkage	0.03	1.59	3.90	78.8	99.3	0.41	0.163
KM shrinkage	-0.14	1.76	4.36	76.5	99.5	0.47	0.161

Averaged over all ITEs with the same process stage model, the spatially averaged mean deviation between reference climate and posterior is close to 0 K for all process stage models with values between -0.14 K for the shrinkage KM and +0.03 K for the shrinkage RM (Table 3.3). However, the

variation across ITEs is larger for the glasso covariance models (standard deviation around 0.31 K) than for the shrinkage covariance models (standard deviation around 0.24 K), which shows that the additional spatial modes in the shrinkage covariances make the reconstructions less sensitive to biases originating from random errors in the proxy data (Fig. 3.9a). Separating into MTWA and MTCO, no pronounced patterns are found for the mean across all ITEs with the same process stage model, but the standard deviations for MTCO are generally larger than for MTWA. This follows from the larger noise level in the local MTCO reconstructions that makes the spatial MTCO reconstructions more susceptible to biases.

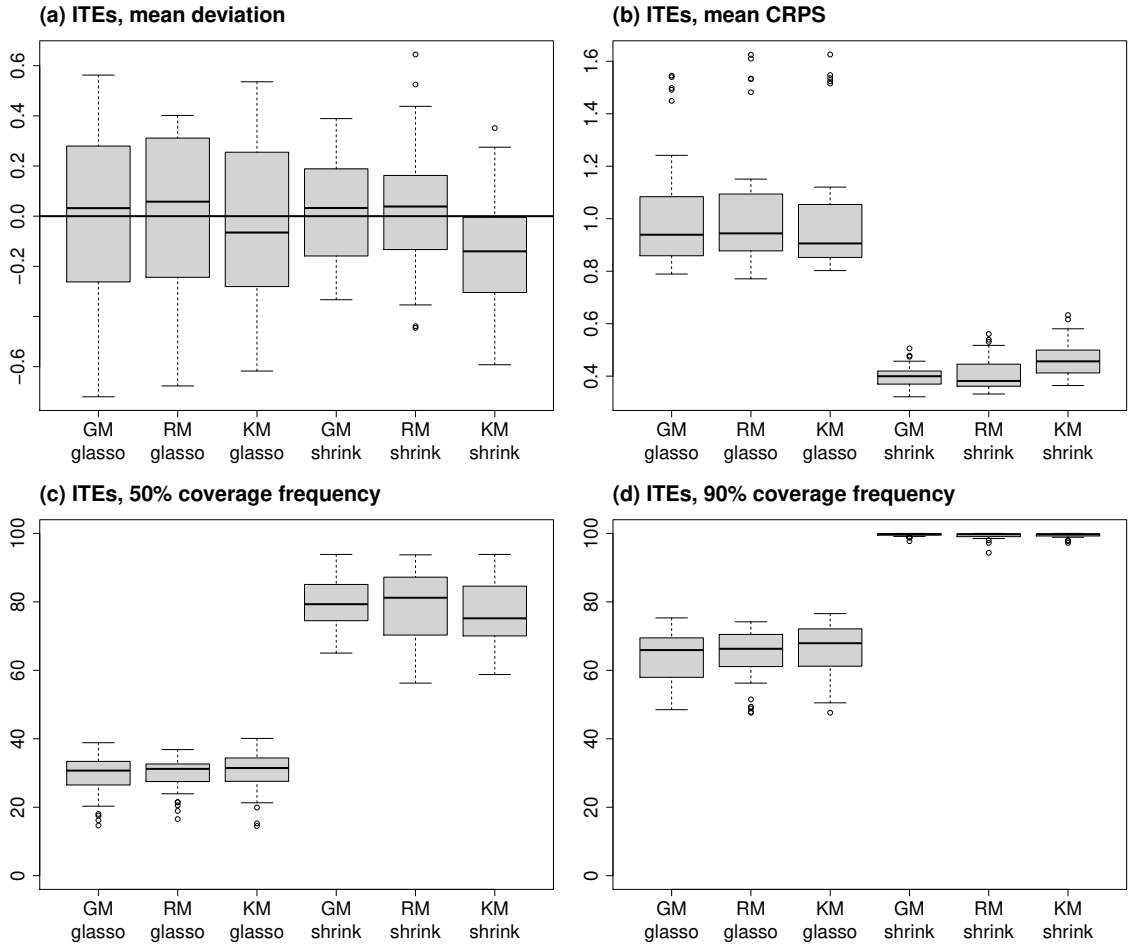

Figure 3.9: Results from ITEs. The boxplots depict the distribution of experiments with the same process stage model. (a) Mean deviation from reference climate, (b) mean CRPS, (c) coverage frequency of 50% CIs, (d) coverage frequency of 90% CIs.

A more diverse picture is found in the spatial patterns of the mean deviations. All ITEs with glasso covariance matrices have larger local biases than those with shrinkage covariances (3.10). While the magnitude of biases for the GM and RM models with shrinkage covariance matrices is much smaller, the shrinkage KM exhibits just slightly smaller local deviations than the glasso covariance

ITEs. This shows that the models with shrinkage matrix can reconstruct spatial patterns better than the models which use glasso, due to more effective degrees of freedom in the covariance matrix, and that models, which directly average over ensemble members, perform better than the KM.

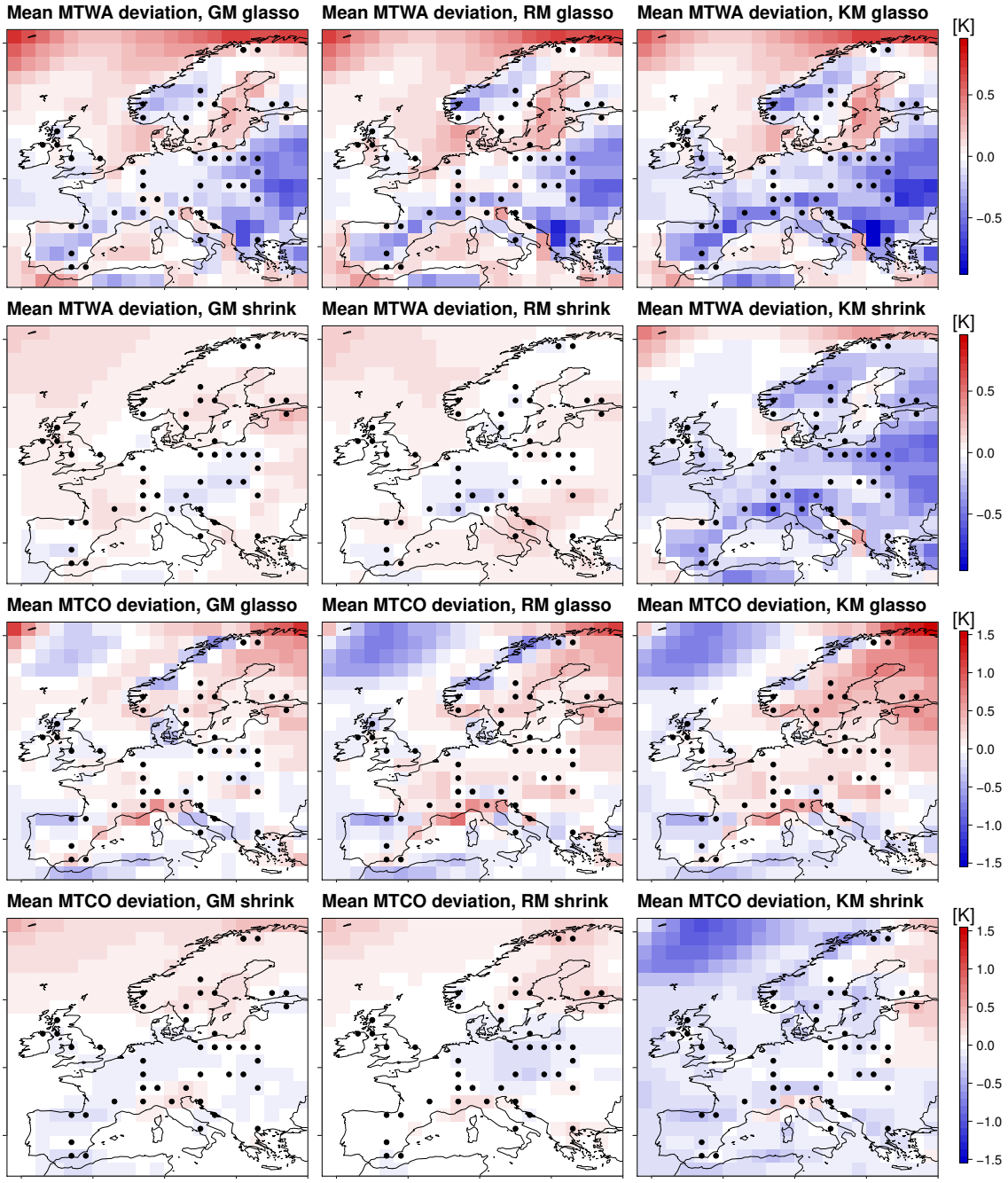

Figure 3.10: Mean deviation in ITEs for GM (left column), RM (middle column), and KM (right column). Top row: Models with glasso covariance matrix, MTWA. Second row: Models with shrinkage covariance matrix, MTWA. Third row: Models with glasso covariance matrix, MTCO, Bottom row: Models with shrinkage covariance matrix, MTCO. Grid boxes with simulated proxy data are depicted by black dots.

The higher number of spatial modes in the shrinkage covariance models leads to larger uncertainty estimates in the posterior distribution, because the limited information contained in the proxy data can constrain only a small number of spatial modes. For example, the grid box-wise 90% CIs in the ITEs with shrinkage matrix are on average 4.05 K wide, but only 2.95 K in the ITEs with glasso matrices (Table 3.3). This leads to diverging results for the coverage frequencies. The glasso models are under-dispersive and the shrinkage models are over-dispersive (Table 3.3). The coverage frequency for 50% CIs is below 41% in all ITEs with glasso covariance matrix and above 56% in all with shrinkage covariance matrix (Fig. 3.9c). Similarly, the coverage frequencies for 90% CIs are below 77% in all ITEs with glasso and above 94% in all ITEs with shrinkage matrix (Fig. 3.9d).

The coverage frequencies are below the target values at most grid boxes of the ITEs with glasso covariance matrix, whereas they are above the desired values at almost all grid boxes in the ITEs with shrinkage matrix (Fig. 3.11). The values are closest to the target at or near grid boxes with proxy data in all ITEs. This means that they tend to be higher than the spatial average in the ITEs with glasso matrix and lower than the spatial average in the ITEs with shrinkage matrix. This effect is more pronounced for coverage frequencies of 50% CIs than of 90% CIs, which indicates that the reconstruction identifies the centers of the probability distributions better than their tails.

The combined effect of mean deviation and dispersiveness can be seen in the CRPS values. With a spatially averaged mean of 1.02 K for all three models with glasso covariance matrix, the ITEs with glasso covariance matrix have a higher CRPS than the ITEs with shrinkage matrices that have a spatially averaged mean of 0.43 K (Table 3.3, Fig. 3.9b). This is a result of larger biases on the grid box level and under-dispersiveness of the posterior distribution. In addition, the variability between ITEs with the same process stage model is higher for models with glasso covariance, which shows that glasso-based covariances are less robust due to the small number of degrees of freedom in the spatial covariance matrix. The MTWA CRPS is slightly higher than the MTCO CRPS since the local reconstructions constrain MTWA more than MTCO. Among the process stage models with shrinkage matrix, the KM performs slightly worse than the GM and the RM which is a result of the larger biases on the grid box level described above. The spatial CRPS structures reflect the mean deviation and uncertainty patterns, with highest values at locations with highest mean deviations and largest reconstruction uncertainty (Fig. 3.12).

3.3.2 Cross-validation experiments

Experiment design

CVEs are a way to understand the ability of a CFR method to produce consistent reconstructions. In paleoclimatology, a challenge of indirect observations is that poor evaluation scores can result from errors both in the process or data stage and it is hard to disentangle those two error sources. The assumption behind our CVEs is that the data stage is unbiased or consistently biased among different proxy samples such that negative results can be attributed to the process stage. Cross-

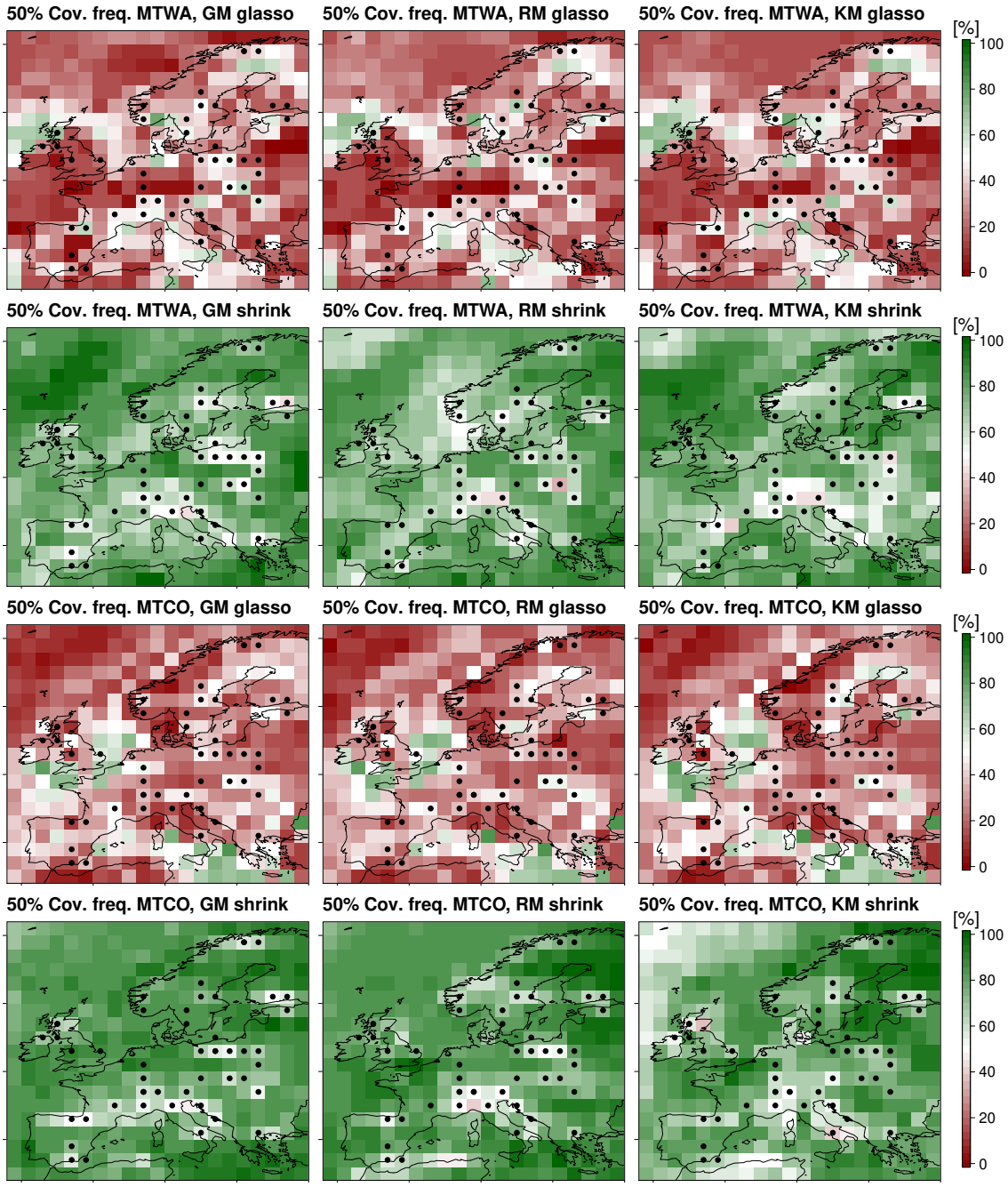

Figure 3.11: Same as Fig. 3.10, but for mean coverage frequency of 50% CIs.

validations are evaluated in the observation space, which is the vegetation space in this study, i.e. the occurrence of taxa in a grid box. As the only reliable information, which is available from the pollen and macrofossil synthesis, is the presence of certain taxa, this is the only data that is used for evaluation. Due to the sparseness of the proxy network, leave-one-out CVEs are performed and not further data is removed in each experiment.

In each CVE, a reconstruction with the Bayesian framework from Sect. 3.2.3 is computed with

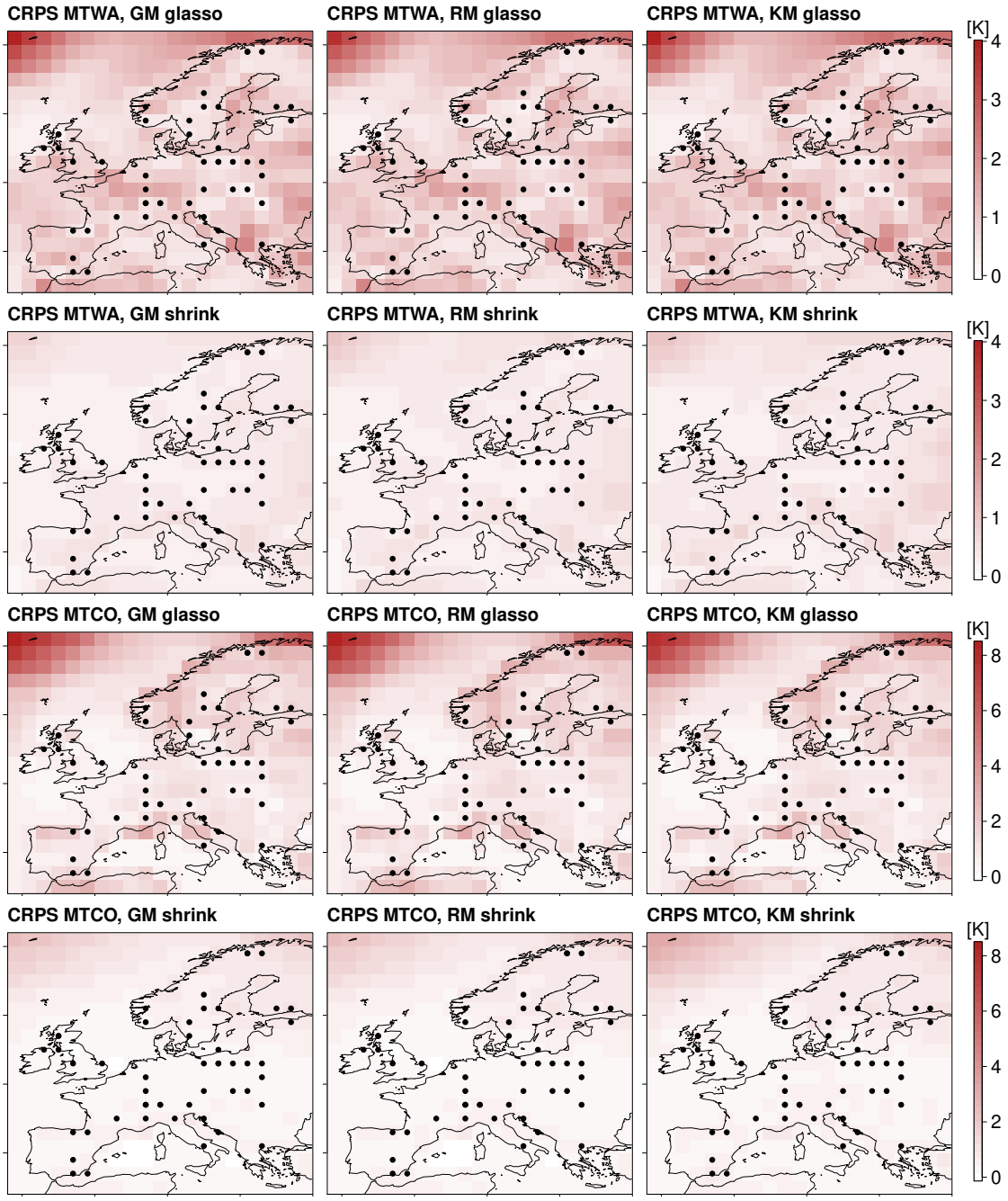

Figure 3.12: Same as Fig. 3.10, but for mean CRPS.

all proxy samples except for those in grid box x. Then, the reconstruction $C_p(x)$ at grid box x is extracted and treated as probabilistic prediction of the climate at x. Next, the PITM forward model from Eq. (3.4) is applied to $C_p(x)$ for each proxy sample located in grid box x to produce probabilistic predictions of the occurrence of the taxa which are found in those samples. Finally, a metric for probabilistic predictions of binary data is calculated. This procedure is repeated for each grid box that contains proxy samples.

Metric for CVE evaluations

In the vegetation space, the probabilistic prediction of the occurrence of a taxon T is represented by the probability of presence $p \in [0, 1]$. A common proper scoring function for binary variables is the Brier score (BS; Brier, 1950) given by

$$BS(T) := \frac{1}{2}\left((\delta_T(1) - p)^2 + (\delta_T(0) - (1 - p))^2\right) = \begin{cases} 1 + p^2 - 2p & \text{if } T = 1 \\ p^2 & \text{if } T = 0 \end{cases} \tag{3.29}$$

where δ_T denotes the indicator function of taxon T. The BS takes values between zero and one, where zero corresponds to a perfect prediction and one to the worst possible prediction. To calculate the BS for a probabilistic climate reconstruction, the PITM forward model is applied to each MCMC sample, which leads to a set of probabilistic predictions $p_i(T)$, $i = 1, ..., I$ for taxon T. Predictions are calculated for each taxa which occurs in sample $P(s)$. The joint score of $P(s)$ is then calculated by averaging the BS of each taxa and MCMC sample:

$$BS(P(s)) := \frac{1}{|T(s)|\, I} \sum_{T \in T(s)} \sum_{i=1}^{I} \left(1 + p_i(T)^2 - 2\, p_i(T)\right). \tag{3.30}$$

If multiple samples are assigned to one grid box, the mean score of those samples is taken.
A problematic step in the methodology described above is that the BS is only evaluated for occurring taxa and not for those which are absent in the sample. This can make the BS improper when comparing models which predict presence and absence of taxa. For example, one could 'cheat' by always predicting the presence of all taxa, since taxa, which do not occur, are removed from the evaluation. However, the goal of the methodology described above is an indirect evaluation of predictions of past climate via transfer functions. It would lead to inconsistencies between the local reconstructions and the BS evaluations if absent taxa were included for calculating the BS but not in the local reconstructions. Future research on estimating detection probabilities of taxa or groupings of taxa is needed to adequately estimate presence and absence probabilities in pollen and macrofossil samples to overcome this issue. As described in Sect. 3.2.1, this is beyond the scope of this study. For each taxon, the BS is minimal for a unique climate state. In addition, they are convex functions of the climate state near the minimizers. These two properties make the methodology described above suitable for the comparison of climate reconstructions and the indirect evaluation of CFR methods as long as the response surfaces are accurate.

Results

With mean BS values of 0.186 (GM, RM) or 0.187 (KM), the process stage models with glasso covariance matrices perform slightly worse than those with shrinkage covariances, which take values between 0.161 for the KM and 0.165 for the GM (Table 3.3). Similar to the ITEs, the differences between models with different covariance matrix type are larger than those with just different process stage formulations. Likewise, the largest differences at individual grid boxes

between models with different covariance matrix types are on the order of 10^{-1}, which is an order of magnitude larger than the difference between models with the same covariance type (Fig. 3.13). The models with shrinkage covariance matrices perform better in Western Europe and Fennoscandia, whereas in Central and Eastern Europe, the magnitude of differences is very small and in the majority of grid boxes, the models with glasso covariance matrices perform slightly better.

The overall better performance of the process stage models with shrinkage covariance matrix might be a result of the under-dispersive and less robust behavior of the models with glasso covariance matrices that was diagnosed in Sect. 3.3.1. This explanation is underlined by the spatial patterns of the performance differences. In Western Europe and Fennoscandia, the local reconstructions tend to be less informative than in Central and Eastern Europe. Therefore, the reconstructions at the grid boxes in those regions are less constrained by nearby proxy samples. This effect is enhanced in the glasso-based models, as the shrinkage target matrix has correlations which decay with the distance between two grid points. Therefore, reconstructions with glasso matrices can be more biased if information are inadequately transferred over large distances. In Central and Eastern Europe this effect is less problematic, as nearby locations are the main contributors to the reconstructions at the grid boxes of left out proxy samples.

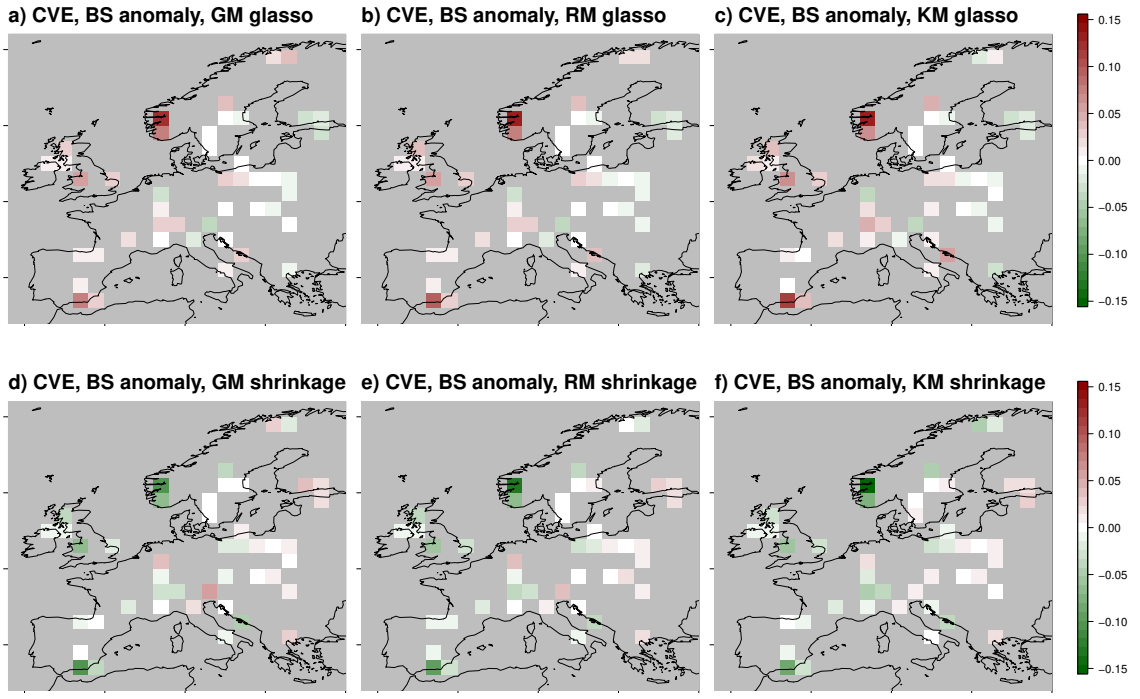

Figure 3.13: BS anomaly from the average BS over all six process stage models in CVEs. Top row: Models with glasso covariance matrix. Second row: Models with shrinkage covariance matrix.

3.3.3 Discussion and conclusions

The ITEs show that process stage models with shrinkage covariances are more dispersive, less biased, and more robust than those with glasso covariances. These properties transfer positively to CVEs where the models with shrinkage covariance matrix perform better, too. The results from process stage models with the same covariance matrix are very similar except that the KM with shrinkage covariance matrix is on average more biased than the respective GM and RM. This shows that the covariance matrix choice determines the reconstruction skill more than the general formulation of the process stage as Gaussian, regression, or kernel model.

The better performance of shrinkage covariance models indicates that the low number of spatial modes as a result of the small ensemble size is the main reason for the under-dispersiveness of the glasso-based models. On the other hand, the strong over-dispersiveness of the shrinkage-based models in ITEs is an indicator that this model is not under-dispersed even in real world applications which face additional challenges from potentially biased or under-dispersed transfer functions and from more sophisticated spatial structures in the climate state. Additionally, this over-dispersiveness shows that the ensemble spread is wide enough in most regions to lead to reconstructions without too narrow posterior distributions as long as enough spatial degrees of freedom are incorporated in the second-moment properties.

The larger biases of the KM with shrinkage covariance matrix compared to the GM and RM might be an effect of degeneracy of the ensemble member weights. It is a well-known issue of Bayesian model selection problems (Yang and Zhu, 2018), and therefore as well of particle filter methods (Carrassi et al., 2018), that ensemble member weights tend to degenerate towards the member which deviates least from the data. To mitigate this issue, the particle filter part of the KM is combined with a more robust Gaussian structure of each mixture component that leads to an adjustment of each ensemble member according to the signal in the proxy data. The ITEs show that this adjustment is strong enough to avoid under-dispersiveness compared to the GM and RM, but it cannot mitigate larger biases compared to process stage models that average over ensemble members directly.

Based on the results of ITEs and CVEs, process stage models with shrinkage matrix should be preferred over those with glasso covariances. In addition, the smaller biases and more robust nature of the GM and RM compared to the KM, makes them superior choices. Because the RM allows more flexible adjustments to the proxy data than the GM, this model is presumably better suited to deal with additional caveats of real world applications compared to the controlled test environment of ITEs. Therefore, this study advocates for the use of this model in paleoclimate reconstructions with Bayesian filtering methods.

Table 3.4: Summary measures for the joint MTWA and MTCO reconstructions (rows 1 and 2) and the separated reconstructions of MTWA (row 3) and MTCO (row 4). Numbers in brackets are minima and maxima of the corresponding 90% CIs.

Reconstruction name	Spatial mean anomaly	Spatially averaged 90% CI size	Grid box-wise uncertainty reduction	Spatial homogeneity	Median BSS	PMIP3 model with highest weight
Joint reconstruction (MTWA)	(0.99 K) 0.51 K (0.03 K)	4.15 K	38.1%	(1.90 K^2) 2.30 K^2 (2.69 K^2)	0.28	MPI-ESM-P
Joint reconstruction (MTCO)	(1.32 K) 0.69 K (0.10 K)	5.84 K	19.6%	(6.47 K^2) 8.18 K^2 (9.92 K^2)		
Separate MTWA reconstruction	(1.51 K) 1.04 K (0.55 K)	4.14 K	38.1%	(1.85 K^2) 2.32 K^2 (2.62 K^2)	0.27	MPI-ESM-P
Separate MTCO reconstruction	(0.60 K) -0.14 K (-0.82 K)	5.72 K	20.6%	(7.57 K^2) 9.61 K^2 (11.74 K^2)	0.05	HadGem2-CC

3.4 Results and discussion of the European mid-Holocene climate reconstructions

In this section, results are presented from the MH climate reconstructions for Europe with the Simonis et al. (2012) synthesis and the PMIP3 MH ensemble. We study the mean and uncertainty structure of the posterior distribution, the added value of the reconstruction compared to the unconstrained PMIP3 ensemble, the results of joint compared to separate MTWA and MTCO reconstructions, and the robustness of the reconstruction. Finally, our results are compared with previous reconstructions of the European MH climate. All results are from reconstructions with the RM using shrinkage covariance matrices as this model was found to be the most appropriate among all tested process stage models in Sect. 3.3. The results are summarized in Table 3.4. The sensitivity of the results with respect to the process stage model choice is studied in Appendix C.1.

3.4.1 Posterior mean and uncertainty structure

The spatially averaged mean temperature of the reconstruction (posterior mean) is 18.27 °C (90% CI: (17.79 °C, 18.75 °C)) for MTWA and 1.81 °C (90% CI: (1.22 °C, 2.45 °C)) for MTCO, which is in both cases warmer than the CRU reference climatology (+0.51 K for MTWA and +0.69 K for MTCO). Larger anomalies are found for subregions (Fig. 3.14a,b). For MTWA as well as MTCO, the temperature was cooler than today in many Southern European areas, while in Northern Europe the temperature was predominantly higher than today. More specifically, MTWA was warmer over

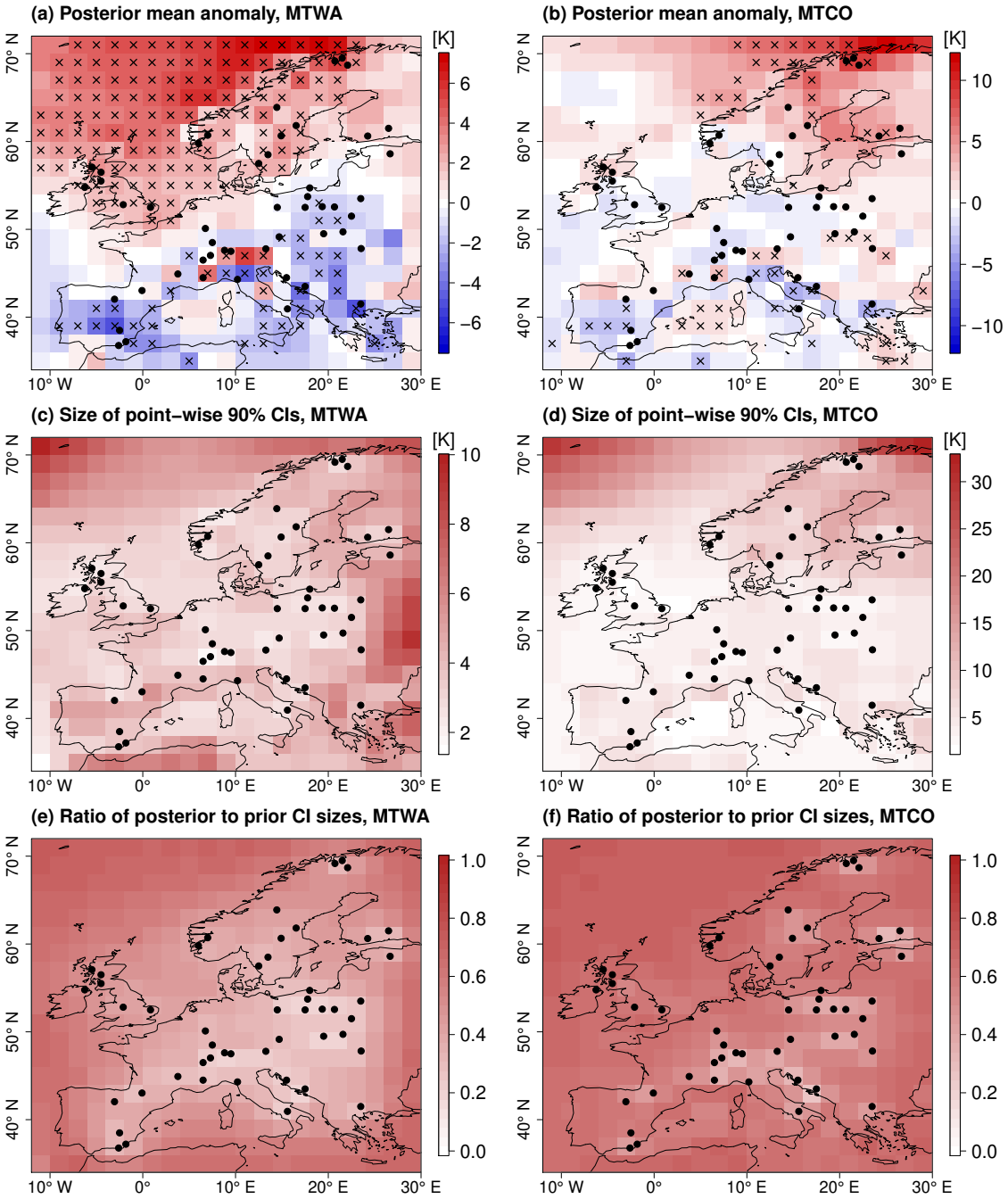

Figure 3.14: Spatial reconstruction for MH. Top row: Posterior mean anomaly from CRU reference climatology (left: MTWA, right: MTCO), middle row: reconstruction uncertainty plotted as size of grid box-wise 90% CIs (left: MTWA, right: MTCO), bottom row: reduction of uncertainty from posterior to prior measured by ratio of posterior to prior grid box-wise 90% CI sizes (left: MTWA, right: MTCO). Black dots depict proxy samples. In the top row, grid box-wise significant anomalies (5% level) are marked by black crosses.

Fennoscandia, the British Islands, and the Norwegian Sea. Most of these anomalies are significant on a 5% level. Here, a positive anomaly is called significant if the probability to exceed the reference climatology is at least 0.95. Significant negative anomalies are defined accordingly. The significance estimates are calculated grid box-wise. Negative MTWA anomalies are found in large parts of the Mediterranean and Eastern Europe, but fewer anomalies are significant on a 5% level than in Northwestern Europe. The largest positive MTCO anomalies are found in Fennoscandia and off the Norwegian coast. In the other parts of the domain, the majority of MTCO anomalies are negative, but the spatial pattern is more heterogeneous than for MTWA. A lot fewer MTCO anomalies are significant on a 5% level compared to MTWA anomalies. Using the joint information contained in the proxy synthesis and combining it with the spatial structure of the PMIP3 ensemble leads to more significant signals than in the individual data products.

Most of the taxa, which are used in the reconstruction, are stronger confined for MTWA than for MTCO because the occurrence of most European plants is more sensitive to environmental conditions during the growing season. This results in more constrained local MTWA reconstructions (Fig. 3.6c), which is in concordance with findings from Gebhardt et al. (2008). Hence, the uncertainty in the MTWA reconstruction is smaller than in the MTCO reconstruction with spatially averaged grid box-wise 90% CI sizes of 4.15 K and 5.84 K, respectively (Fig. 3.14c,d). The uncertainty is smallest at grid boxes with proxy records, and highest in the north-eastern and north-western parts of the domain where the PMIP3 ensemble spread is large and the constraint from proxy data is weak. For MTWA, additional regions with large uncertainties are found at the eastern and southern boundaries of the domain due to weak proxy data constraints. Besides, the reconstruction uncertainty has small spatial variations. The ratio of the CI size of spatially averaged temperature and the spatially averaged grid box-wise CIs can be interpreted as a measure for the spatial degrees of freedom in the reconstruction.

The highest reduction of uncertainty due to the inclusion of proxy data compared to the unconstrained PMIP3 ensemble is found at grid boxes with proxy data. At these grid boxes, the spatially averaged reduction of grid box-wise CI sizes from unconstrained process stage to posterior distribution is 50.1%, compared to 26.0% for grid boxes without proxy data (Fig. 3.14e,f). The uncertainty reduction for MTWA is higher for terrestrial grid boxes than marine ones, but the smaller PMIP3 ensemble spread over the British Islands, the North Sea, and the Bay of Biscay leads to similar posterior CI sizes in these areas. For MTCO, the reduction of uncertainty is generally smaller than for MTWA due to the weaker proxy data constraint.

To study whether the degree of spatial smoothing of the reconstruction is reasonable, we calculate a discrete approximation H of the energy norm given by

$$\frac{1}{2\,|\text{Domain}|}\int_{\text{Domain}} |\nabla C(x)|^2\; dx, \tag{3.31}$$

where the integral is taken over the reconstruction domain and normalized by the size of the domain. This is a measure of the spatial homogeneity as it measures the local fluctuations of the climate field C. A reconstruction with a good degree of smoothing is expected to have a spatial

homogeneity similar to the PMIP3 ensemble and the reference climatology, as the energy norm depends mainly on local features like orography or land-sea contrasts. We expect these features to affect the local climate of the MH and today's climate similarly, since reconstructions suggest only small changes in topography between the MH and today. To approximate Eq. (3.31), we replace $\nabla C(x)$ by the finite differences of C in zonal and meridional direction. We use the mean of the absolute values of the differences to both sides as approximation of $|\nabla C(x)|$. Finally, the spatial average over all grid boxes is calculated. Then, this homogeneity measure H of the posterior, the climatologies of the PMIP3 ensemble members, and the reference climatology are compared. For MTWA, the posterior mean value is 2.30 K^2 (90% CI: (1.91 K^2, 2.69 K^2)), which is comparable to 2.77 K^2 for the reference climatology and values between 1.28 K^2 and 2.64 K^2 for the PMIP3 climatologies. The heterogeneity of MTCO is higher than of MTWA, but the mean posterior value of 8.18 K^2 (90% CI: (6.47 K^2, 9.92 K^2)) is at least of comparable magnitude as the reference climatology (4.86 K^2) and the PMIP3 climatologies (between 4.24 K^2 and 7.74 K^2). From these results, it is deduced that the posterior has a reasonable degree of spatial smoothing.

3.4.2 Comparison of unconstrained PMIP3 ensemble and posterior distribution

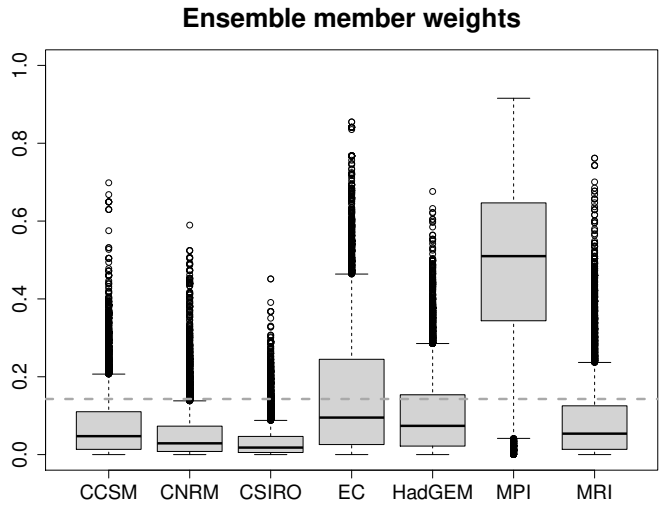

Figure 3.15: Posterior ensemble member weights λ of the spatial MH reconstruction. The mean of the prior weights is denoted by the dashed line.

By comparing the posterior with the PMIP3 ensemble and the local reconstructions, it can be seen that for most areas with nearby proxy records the posterior mean resembles the local reconstructions more than the PMIP3 ensemble mean. This shows that the uncertainty in the prior distribution is large enough to lead to a reconstruction which is mostly determined by proxy data, where available. The posterior MTWA mean is warmer than the prior mean in Northern Europe and cooler in Southern and Eastern Europe. For MTCO, the posterior mean is much warmer than the prior mean in Fennoscandia and slightly cooler in Southern Europe.

The posterior weights λ of the PMIP3 ensemble members are a combination of the prior distribution of λ and the likelihood of C_p for each combination of ensemble member weights (see Appendix B.2 for details). λ provides information about which combination of ensemble members fits best to the proxy data. In our reconstruction, the MPI-ESM-P climatology has the highest posterior weights (mean of 0.485) (see Fig. 3.15), followed by the EC-Earth-2-2 climatology (posterior mean of 0.154) and the Had-GEM2-CC climatology (posterior mean of 0.104). Note, that the weights of the MPI-ESM-P and the EC-Earth-2-2 are the only ones that are on average higher than the prior mean of 1/7. The large differences of the weights are a result of the large differences between the ensemble member climatologies. Because there is less uncertainty in the local MTWA reconstructions, it is the major variable for determining the posterior weights. Among all included models, the MPI-ESM-P simulation is closest to the dipole structure with MTWA warming in Northern and cooling in Southern Europe, which explains the high model weight.

3.4.3 Added value of the reconstruction

CVEs provide insight into the value that is added to the unconstrained PMIP3 ensemble, represented by the process stage Eq. (3.13), by constraining it with the Simonis et al. (2012) synthesis. To quantify the added value, the BS from Eq. (3.30) is calculated for the unconstrained process stage, which is called BS(Prior), and compared to the BS of the posterior, BS(Posterior), calculated from leave-one-out CVEs. Then, the Brier skill score (BSS)

$$\text{BSS} := \frac{\text{BS(Prior)} - \text{BS(Posterior)}}{\text{BS(Prior)}} \tag{3.32}$$

is computed, which is a measure of the added value of the spatial reconstruction. For positive BSS values, the posterior distribution is superior to the prior. On the other hand, the posterior distribution is inferior to the prior for negative values. This would indicate inconsistencies in the local proxy reconstructions, a scaling mismatch of simulations and data, or the existence of spurious correlations in the spatial covariance matrix.

For most left-out proxy samples, the BSS is positive (68.9% of grid boxes) with a median of 0.28 (Fig. 3.16). The BSS values are predominantly positive for all regions but the British Islands and the Alps. This indicates a high consistency of the reconstruction in large parts of the domain. In particular, consistent MTWA cooling in Southern and Eastern Europe in the local reconstructions compared to the prior distribution leads to cooling and reduction of uncertainty in the posterior compared to the unconstrained PMIP3 ensemble. Similarly, the consistent MTWA and MTCO warming of the local reconstructions in the north-eastern part of the domain lead to positive BSS values.

The persistent negative BSS values for the British Islands cannot be explained by data outliers but warrant a systematic issue. For this region, the uncertainty in the local reconstructions is larger than for other areas, such that the local proxy records constrain the posterior less than the posterior ensemble member weights and some of the more distant proxy records. This leads to a reduction of

the posterior uncertainty compared to the unconstrained PMIP3 ensemble, but without improving the concordance of the mean state with the local reconstructions, which in turn results in negative BSS values. In and near the Alps, negative BSS might be a result of not accounting enough for orographic effects in the different sources of information. Both of those features might be a result of the limited resolution of PMIP3 simulations, as the location of the British Islands between two oceans and the complex topography of the Alps is not accurately represented in the current generation of PMIP3 global simulations. This could lead to an erroneous relationship between the British Islands, the Alps, and other European regions in the simulations. The use of downscaled regional simulations might resolve these issues in the future.

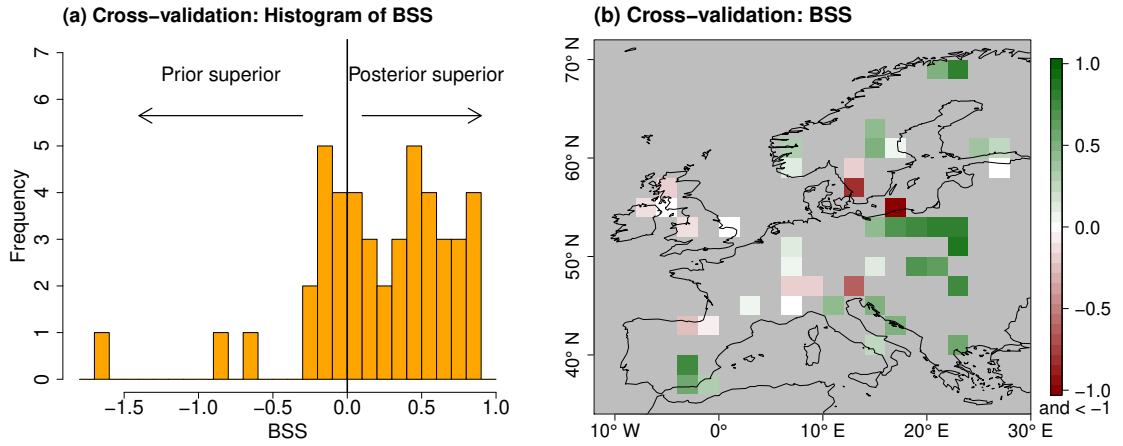

Figure 3.16: BSS from leave-one-out cross-validation. For positive values, the posterior is superior to the prior distribution from the unconstrained PMIP3 ensemble, while for negative values the posterior is inferior to the prior. (a) histogram, (b) spatial distribution

3.4.4 Joint versus separate MTWA and MTCO reconstructions

To study the effect of reconstructing MTWA and MTCO jointly compared to separately, additional reconstructions with only one climate variable are computed. Note that the interactions of MTWA and MTCO are twofold in the joint reconstruction: (a) the response functions have an interaction term in the logistic regression Eq. (3.2), and (b) the process stage contains joint ensemble member weights for MTWA and MTCO as well as inter-variable correlations in the empirical correlation matrix.

The separate MTWA reconstruction is on average around 0.5 K warmer than the joint reconstruction, whereas the spatially averaged posterior mean of the separate MTCO reconstruction is 0.83 K cooler (Table 3.4). Hence, the seasonal difference is smaller in the joint reconstruction, due to smoothing from the PMIP3 ensemble and slightly positive correlations between MTWA and MTCO in most of the joint local reconstructions. The MTWA only reconstruction is warmer in most land areas, with largest differences in Southern and Eastern Europe, but the differences are almost never significant on a 5% level (Fig. 3.17a). As this part of the domain is best con-

strained by proxy data, and because the posterior ensemble member weights are similar to the joint reconstruction (mean λ_k of 0.419 for MPI-ESM-P), it is likely that the additional warming is due to the missing interaction in the transfer function. On the other hand, the posterior ensemble member weights change a lot for the separate MTCO reconstruction, with HadGEM2-CC and MRI-CGCM3 being the models with the highest weights (mean λ_k of 0.426 and 0.199, respectively). Together with the less constrained transfer functions for MTCO than MTWA and the missing interaction term, this leads to a cooler reconstruction for most areas except for some parts of the Mediterranean (Fig. 3.17b). The cooling is strongest in Scandinavia, the British Islands, the Norwegian Sea, and the Iberian peninsula, but almost never significant on a 5% level. As these are the regions which are least constrained by proxy data, choosing different PMIP3 ensemble members affects the reconstruction more than in Central Europe, where MTCO is best constrained by proxy data. The reconstruction uncertainties are of similar magnitude in the joint and the separate reconstructions (Table 3.4), because the interactions in the transfer functions do not reduce the marginal uncertainties and the shrinkage target Φ does not contain correlations of MTWA and MTCO.

The BSS pattern in the MTWA only reconstruction is mostly the same than in the joint reconstruction except for slightly positive skill in the British Islands (Table 3.4, Fig. 3.17e). This shows that the added value of the joint reconstruction compared to the unconstrained PMIP3 ensemble is mainly determined by the MTWA reconstruction. On the other hand, the added value of the MTCO only reconstruction is much smaller (Table 3.4, Fig. 3.17f) due to larger uncertainties in the local MTCO reconstructions.

The results show that the more constrained local MTWA reconstructions lead to a higher influence on the joint reconstruction than the local MTCO reconstructions. Therefore, the MTCO only reconstruction differs more from the MTCO estimate in the joint reconstruction. Reconstructing MTWA and MTCO jointly should in theory lead to a physically more reasonable reconstruction by creating samples drawn from the same combination of ensemble members. In the example of this study, this effect can be seen in the smaller seasonal differences in the joint than in the separate reconstructions. On the other hand, Rehfeld et al. (2016) show that multi-variable reconstructions from pollen assemblages can be biased when signals from a dominant variable are transferred to a minor variable. While the PITM model might be less sensitive to this issue than the WA transfer function used in Rehfeld et al. (2016) because it better respects the larger MTCO uncertainties and because it is unclear whether taxa occurrence is similarly susceptible to the issue than pollen ratios, it will be subject to future work to study whether joint or separate reconstructions lead to more reliable results.

3.4.5 Robustness of the reconstruction

Our approach is designed with the goal of being more suitable for sparse data situations than standard geostatistical models. To understand the robustness of the Bayesian framework with

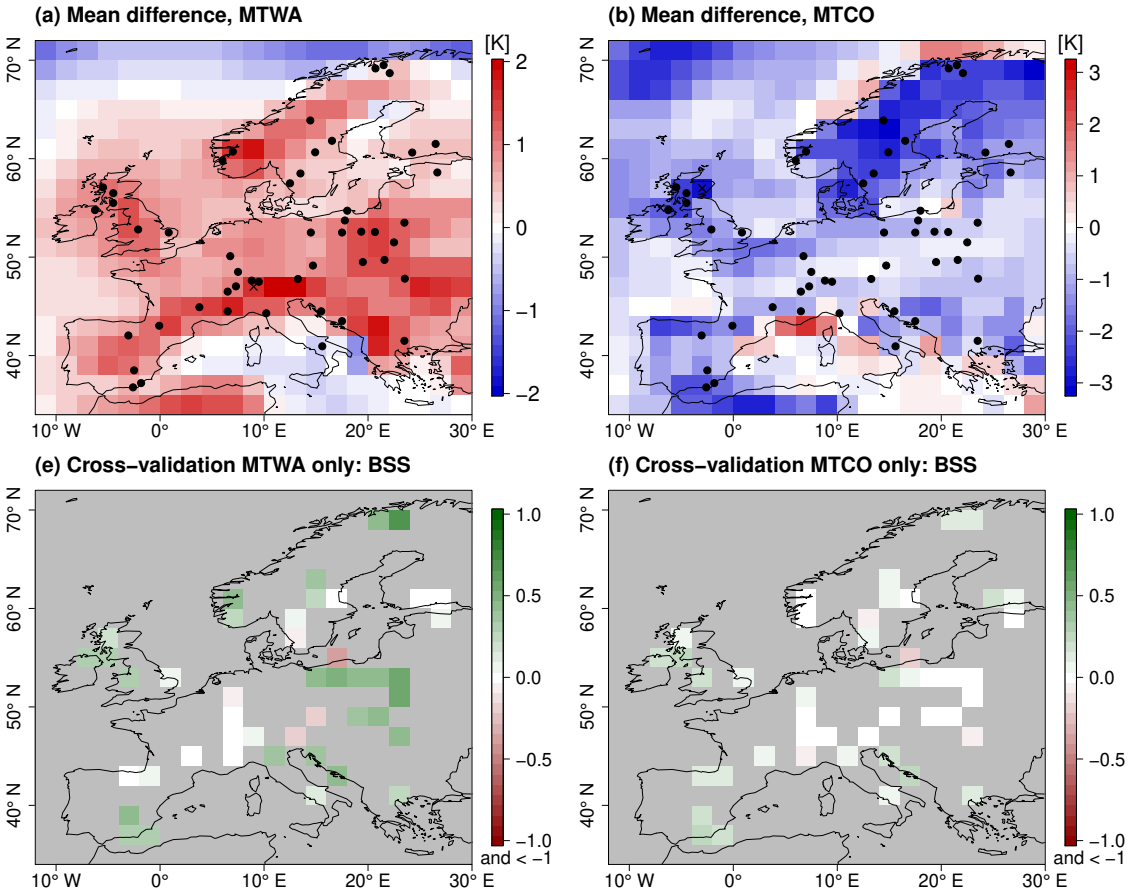

Figure 3.17: Differences of joint and separate reconstructions of MTWA and MTCO. Top row: posterior mean difference (left: MTWA, right: MTCO); bottom row: BSS of the separate reconstructions (left: MTWA, right: MTCO). Black dots depict proxy samples. In the top row, grid box-wise significant differences (5% level) between the separate and the joint reconstructions are marked by black crosses.

respect to the amount of data included in a proxy synthesis, five experiments with only half of the samples are performed, which are either selected to retain the spatial distribution of proxy samples or chosen randomly. In all of the tests, the general spatial structure of the posterior distribution, including the anomaly patterns, is preserved. However, depending on the chosen proxy samples, the local anomalies and magnitude of changes vary which should be expected when such a large portion of the already sparse data is left out. Only the Norwegian Sea in the MTCO reconstruction changes substantially in some experiments. Plots from the experiments with reduced proxy samples are provided in Appendix C.3.

The mean spatial averages differ up to 0.6 K for MTWA as well as MTCO, but none of the changes is significant on a 5% level. In contrast, the uncertainty estimates are consistent across all experiments with reduced proxy samples. The averaged grid box-wise 90% CIs grow by up to 0.4 K compared to the reconstruction with the full proxy synthesis. Considering that half the proxy

samples are left out, this number is low. In all experiments, the spatial homogeneity H is not significantly different from the values reported in Table 3.4, which shows that the spatial homogeneity as a local feature is more controlled by the process stage than the proxy data. In all but one experiment, the MPI-ESM-P remains the ensemble member with the largest weight λ_k, and the three ESMs which are favored neither in the MTWA nor in the MTCO reconstruction retain very low weights in all experiments. Depending on whether proxy samples in which MTCO is much less constrained than MTWA are removed or not, the weights of the four models with the highest values in the joint and separate reconstructions can vary such that in one case the average weight of the EC-Earth-2-2 is 0.1 higher than the weight of MPI-ESM-P. These changes of the weights explain the MTCO changes in the Norwegian Sea as this is the region which is most influenced by the ensemble member weights. The experiments show that the reconstruction is robust with respect to the number of proxy samples as long as the remaining samples are informative and relatively uniformly distributed across space. In our example, this is not the case for the Norwegian Sea. Combining pollen records with SST proxies could potentially overcome this issue.

The large PMIP3 ensemble spread for most grid boxes shows that the prior distribution, which is calculated from the ensemble, contains a wide range of possible states. In areas which are well constrained by proxy data, this large total uncertainty leads to a reconstruction which depends little on the climatologies of the ensemble members. Hence, in these areas, the reconstruction is not sensitive to the particular formulation of the process stage (see Appendix C.1). This shows that our method is applicable despite well-known model-data mismatches for the MH (Mauri et al., 2014). On the other hand, the spatial correlation structure controls the spread of local information into space. Different formulations of the spatial correlation matrix can lead to substantially different reconstructions in regions that are not well constrained by proxy samples and in particular a spatial covariance with too large spatial correlations can lead to overly optimistic uncertainty estimates.

3.4.6 Comparison with previous reconstructions

Several reconstructions of European climate during the MH have been compiled previously. Here, we compare our reconstructions to those of Mauri et al. (2015), Simonis et al. (2012), and Bartlein et al. (2011).

Mauri et al. (2015) use a PFT MAT transfer function and a thin plate spline interpolation for pollen samples stemming mostly from the European pollen database. The simpler interpolation method allows the treatment of the samples as point data and the interpolation of the local reconstructions to an arbitrary grid. Among other variables, summer and winter temperature are reconstructed. We find a dipole anomaly structure similar to Mauri et al. (2015) in our reconstructions, with mostly positive anomalies in Northern Europe and negative anomalies in Southern Europe. In Mauri et al. (2015) as well as in our reconstruction, the Alps are the only region with significant warming in Central and Southern Europe for summer temperature. Generally, the amplitude of summer anomalies in the two reconstructions is similar, although locally there are differences with cooler

anomalies over Southwestern Fennoscandia in our reconstruction as well as warmer anomalies in Finland. For winter temperature, the cooling in the Mediterranean and the British Islands is less pronounced and spatially less consistent in our reconstruction compared to Mauri et al. (2015). As for summer temperature, we reconstruct smaller anomalies in Southern Fennoscandia. In contrast, our reconstruction shows higher anomalies in Northern Scandinavia.

The same pollen dataset and another version of PITM are used in Simonis et al. (2012) to reconstruct July and January temperature, such that differences between the two reconstructions are mostly related to the different smoothing technique. Simonis et al. (2012) minimize a cost function which combines pollen samples with an advection-diffusion model that is driven by insolation changes between the MH and today. In Simonis et al. (2012), the dipole structure is not found in the same way than in our reconstruction, which might be due to the different way how regions, that are not well constrained by proxy data, are treated. Both reconstructions share positive summer temperature anomalies in Northern Europe as well as negative anomalies in Central Europe and the Iberian peninsula. Unlike our reconstruction, Simonis et al. (2012) find positive anomalies in Southeastern Europe. For winter temperature, the reconstruction of Simonis et al. (2012) exhibits an east-west dipole in contrast to the north-east to south-west dipole in our reconstruction. This different structure might be due to the smaller proxy data control of the winter reconstructions, which leads to a higher importance of the interpolation schemes.

A reconstruction designed to evaluate the PMIP3 simulations was provided by Bartlein et al. (2011). They combine a large number of pollen-based local reconstructions from the literature to produce a gridded product of six climate variables including MTWA and MTCO. In contrast to our reconstruction, the used local reconstructions are not smoothed across space but only within a grid box. Their results show a dipole structure but less pronounced than in our reconstruction and in Mauri et al. (2015). In particular, they find a cooling for Eastern Fennoscandia in summer, a much smaller warming of Northern Fennoscandia than our reconstruction, and a warming in Germany and France. On the other hand, the reported anomalies in Bartlein et al. (2011) for the Mediterranean and Eastern Europe are similar to our results.

The comparisons show that patterns like the dipole type anomaly structure, which are not present in the PMIP3 ensemble, seem to be consistent across reconstructions with different pollen transfer functions. However, it cannot be ruled out that the dipole structures are a result of the tendency of most pollen transfer functions to be biased towards the mean values of the calibration dataset. While some of the differences between the existing literature and our results can be explained by the used transfer functions and proxy syntheses, the choice of an appropriate interpolation method plays an important role, too, especially in areas with very sparse and weakly informative proxy data and to determine the degree of smoothing.

Table 3.5: Summary measures for the Siberian LGM reconstructions of MTWA and P_{ANN}. Numbers in brackets are minima and maxima of the corresponding 90% CIs.

Reconstruction variable	Spatial mean anomaly	Spatially averaged 90% CI size	Grid box-wise uncertainty reduction	Median CRPSS	PMIP3 model with highest weight	Mean ϱ
MTWA	(-0.63 K) -1.40 K (-2.13 K)	4.87 K	59.5%	0.45	MRI-CGCM3	0.91
P_{ANN}	(-15.8%) -21.2% (-26.6%)	34.8 %points	8.4%	0.01	MPI-ESM-P	0.80

3.5 Results and discussion of the Siberian Last Glacial Maximum climate reconstructions

This section presents results from the Siberian Last Glacial Maximum climate reconstructions. The reconstructions of MTWA and P_{ANN} combine the local reconstructions described in Sect. 3.1.2 and the PMIP3 LGM ensemble (Sect. 3.1.3). First, the mean and uncertainty structure of the MTWA reconstruction will be studied, before looking at insights that the reconstruction provides on the different PMIP3 simulations and its added value compared to the unconstrained PMIP3 ensemble. Then, the corresponding results for the P_{ANN} reconstruction are presented. Finally, we discuss the results with a focus on reasons for the non-existence of a Siberian ice sheet during the last Glacial.

3.5.1 Spatial MTWA reconstruction

The spatially averaged mean anomaly of the MTWA reconstruction is -1.40 K with a 90% CI of (-2.13 K, -0.63 K). The spatial average is the result of a meridional gradient with largest cooling of up to 5 K in the southern part of the domain, only very small cooling in Northern Siberia, and MTWA warming of up to 10 K in Arctic Ocean regions that are below sea level today but were above sea level during the LGM (Fig. 3.18a). Very weak cooling and temperature similar to the modern climate are also found in the Southern Kolyma region and the Northern Sea of Okhotsk (around 150° E, 60° N). The maximal warming in the spatial reconstruction is a lot stronger than in the local reconstructions (Fig. 3.2) due to structural extrapolation via the process stage. In the local reconstructions, temperatures are near the modern values or slightly warmer along the modern coastline of the Arctic Ocean. PMIP3 simulations with similar values along the modern coastline produce pronounced summer warming north of the modern coastline. The negative anomalies in the southern and central part of the domain are mostly significant on a 5% level. While many positive anomalies in the north-eastern part of the domain are also significant on a 5% level, those

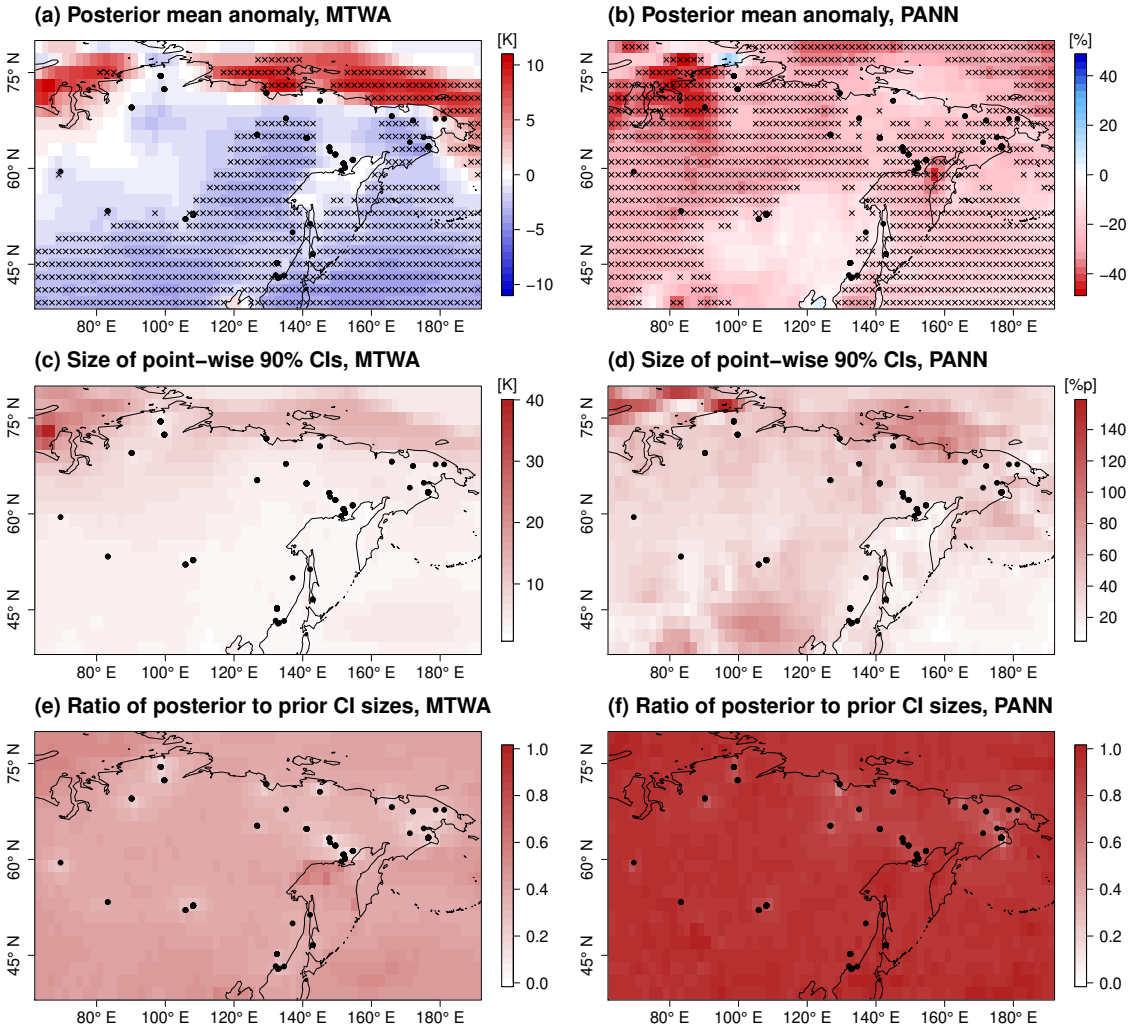

Figure 3.18: Spatial reconstruction for LGM. Top row: Posterior mean anomaly (left: MTWA, right: P_{ANN}), middle row: reconstruction uncertainty plotted as size of grid box-wise 90% CIs (left: MTWA, right: P_{ANN}), bottom row: reduction of uncertainty from posterior to prior measured by ratio of posterior to prior grid box-wise 90% CI sizes (left: MTWA, right: P_{ANN}). Black dots depict proxy samples. In the top row, grid box-wise significant anomalies (5% level) are marked by black crosses.

in the north-western part are not because this is the region with largest reconstruction uncertainties. Grid box-wise 90% CIs feature a strong north-south gradient with CI sizes down to 1 K in the southern part of the domain and up to 40 K at the north-western coastline of the Arctic Ocean (Fig. 3.18c). This pattern is similar to the ensemble spread in the PMIP3 simulations, which differ most in northern areas. The uncertainty in the unconstrained PMIP3 ensemble is reduced most near proxy records, but also large-scale features are more constrained (Fig. 3.18e). Therefore, the spatially averaged grid box-wise reduction of uncertainty from the unconstrained process stage to the posterior distribution is 59.5%.

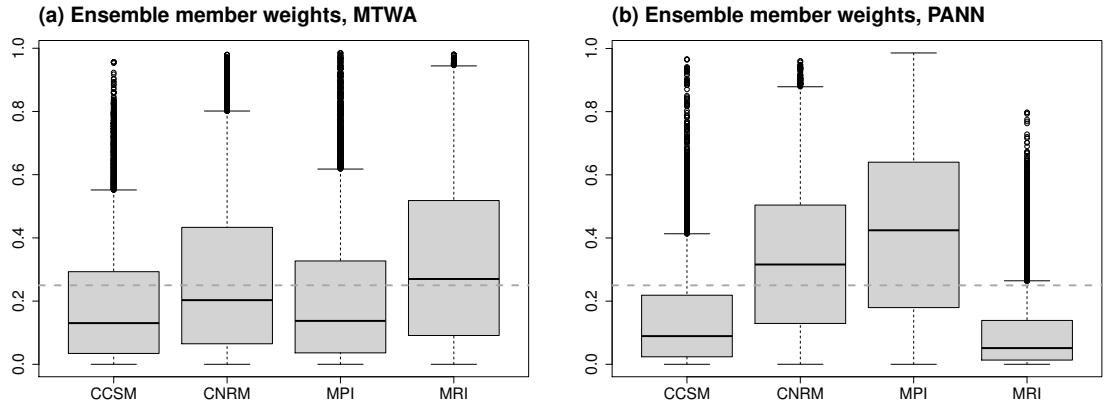

Figure 3.19: Posterior ensemble member weights λ of the reconstructions, (a) MTWA, (b) P_{ANN}. The mean of the prior weights is depicted by the dashed line.

The overall moderate cooling of the local reconstructions leads to a preference of warmer simulations in the spatial reconstruction, in particular those with similar north-south gradients. As a result, the MRI-CGCM3 climatology has the largest mean weight λ_k with 0.323, followed by the CNRM-CM5 with a mean weight of 0.272, the MPI-ESM-P with 0.210, and the CCSM4 with 0.195 (Fig. 3.19a). The latter is the coldest of the four models. While the large-scale north-south gradient in the first three models is in concordance with the local reconstructions, more regional heterogeneity is featured in the local reconstructions than in all ensemble members. These regional patterns are contained in the spatial reconstruction, which indicates the added value from constraining the PMIP3 ensemble by proxy data. Meanwhile, local features of the local reconstructions, which are not reflected by nearby proxy samples, tend to be smoothed out in the spatial reconstruction. The posterior mean is warmer than the PMIP3 ensemble mean across almost the whole domain. On average, this leads to an 1.38 K warmer posterior distribution.

The proxy correlation parameter ϱ estimates the correlation of the transfer function uncertainty between samples from the same sediment core. It is mostly determined by discrepancies between local reconstructions from nearby pollen cores. The posterior distribution of ϱ is concentrated between 0.86 and 0.94, which means that among samples from one pollen record only around 10% of the observation uncertainty is independent. This indicates that the majority of the uncertainty is related to shared error sources like the modern calibration.

The added value of the LGM reconstructions can be quantified similar to the MH reconstruction. Since the local reconstructions are already in the climate space, a different score function is required. We use the CRPS, which was introduced in Sect. 3.3.1 for evaluations of probabilistic predictions in temperature space. We create CVEs where all local reconstructions C_s^l from one grid box x are left out. The spatial reconstruction C_p at x is used as probabilistic prediction of MTWA at that location. To account for transfer function uncertainty, the local reconstruction uncertainty $\sigma_p(s)$ is added to $C_p(x)$. The resulting probability distributions are compared to the local reconstructions C_s^l by calculating the CRPS, which we denote by CRPS(Posterior). Similarly, the

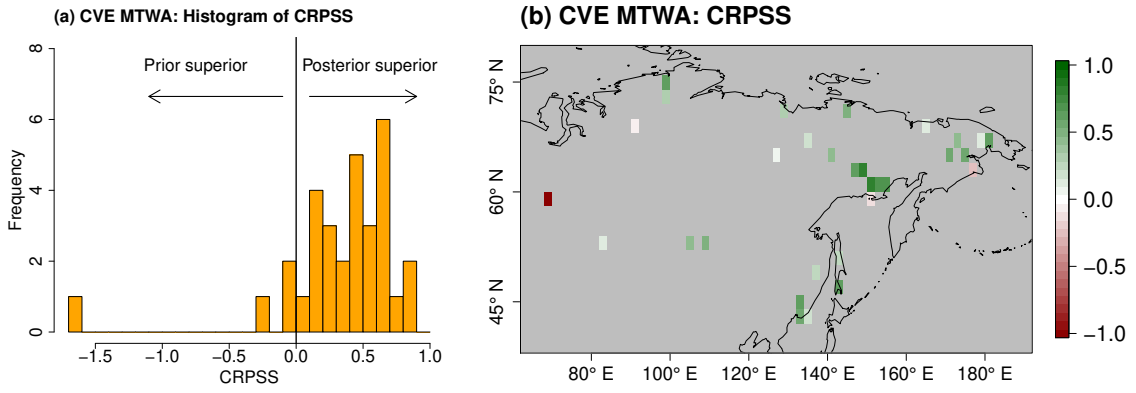

Figure 3.20: Same as Fig. 3.16 but for CRPSS from leave-one-out cross-validation in the MTWA reconstruction.

unconstrained process stage is evaluated by sampling from it, adding the transfer function uncertainty, and calculating the CRPS, which is called CRPS(Prior). Finally, the mean over all left out samples is taken. To compare the two scores, the Continuous Ranked Probability Skill Score (CRPSS), defined by

$$\text{CRPSS} = \frac{\text{CRPS(Prior)} - \text{CRPS(Posterior)}}{\text{CRPS(Prior)}}, \tag{3.33}$$

is calculated which takes values between $-\infty$ and $+1$. Positive values correspond to a gain of reconstruction skill and negative values indicate that the unconstrained process stage is superior to the posterior.

The CRPSS is predominantly positive, with values above 0 at 87% of grid boxes and a median of 0.45. This shows that constraining the PMIP3 ensemble by pollen-based local reconstructions adds value in most parts of the domain (Fig. 3.20). The positive CRPSS values are an effect of less uncertainty in the posterior distribution and regionally consistent local reconstructions, which lead to a reduction of the deviation between PMIP3 ensemble and local reconstructions. The most-western grid box with proxy samples exhibits a pronounced negative CRPSS. This grid box does not possess nearby proxy records such that leaving out the samples in that grid box leads to a warmer posterior distribution in that region compared to the unconstrained PMIP3 ensemble due to higher posterior weights for warmer ensemble members. This is not in agreement with the local reconstructions in that grid box and leads to the negative CRPSS value.

3.5.2 Spatial P_{ANN} reconstruction

Spatially averaged, the P_{ANN} posterior mean is drier than the modern climate, with 21.1% less P_{ANN}. The 90% CI of that estimate is (-26.6%, -15.8%). The posterior mean features the strongest negative anomalies with up to 50% less precipitation in the north-western part of the domain. Precipitation amounts similar to today's climate are found in Manchuria. In all other regions, the precipitation amounts were between 10% and 30% lower than today. The anomalies are signifi-

cant (5% level) in most regions except for Manchuria, and parts of the northernmost and north-easternmost areas of the reconstruction domain (Fig. 3.18b).

The spatial average of the grid box-wise 90% CIs is 34.8 percentage points, which is just an 8.5% reduction of uncertainty compared to the unconstrained process stage. This is less than in all other spatial reconstructions of this study. At grid boxes with proxy data, the reduction of uncertainty is largest with 20.1%, but the proxy signals are not spread out in space a lot. The smallest uncertainties are found in the same regions than for the unconstrained PMIP3 ensemble. The grid box-wise uncertainties range from 10 percentage points in the south-eastern part of the domain to 150 percentage points in the north-western part (Fig. 3.18d,f).

In general, the large-scale patterns of the spatial reconstruction reflect the structures of the PMIP3 ensemble. The posterior mean is on average 4.2 percentage points wetter than the ensemble mean. This is reflected in the posterior model weights λ (Fig. 3.19b), which are highest for the two wettest simulations, the MPI-ESM-P climatology (mean λ_k: 0.425) and the CNRM-CM5 climatology (mean λ_k: 0.332). The reason for the small reduction of uncertainty and the similarity of the spatial structures of posterior and PMIP3 ensemble mean is the smaller ensemble spread compared to the uncertainty of the local reconstructions. Therefore, the proxy data constrains the PMIP3 ensemble less than in all other applications of this study, in which the proxy uncertainty is mostly smaller than the ensemble spread.

This result is also reflected in CVEs. The median CRPSS is 0.01, which means that on average no value is added to the PMIP3 ensemble from constraining it by proxy data. The values are small for most left out grid boxes with CRPSS values ranging from -0.4 to +0.5 and no pronounced spatial CRPSS structures (Fig. 3.21).

The proxy correlation parameter ϱ is concentrated between 0.75 and 0.85, which is slightly lower than for MTWA, but still only 20% of the reconstruction uncertainty of samples within one record is independent. This shows that for P_{ANN}, similar to MTWA, most of the transfer function uncertainty is originating from sources that are shared among samples from the same core.

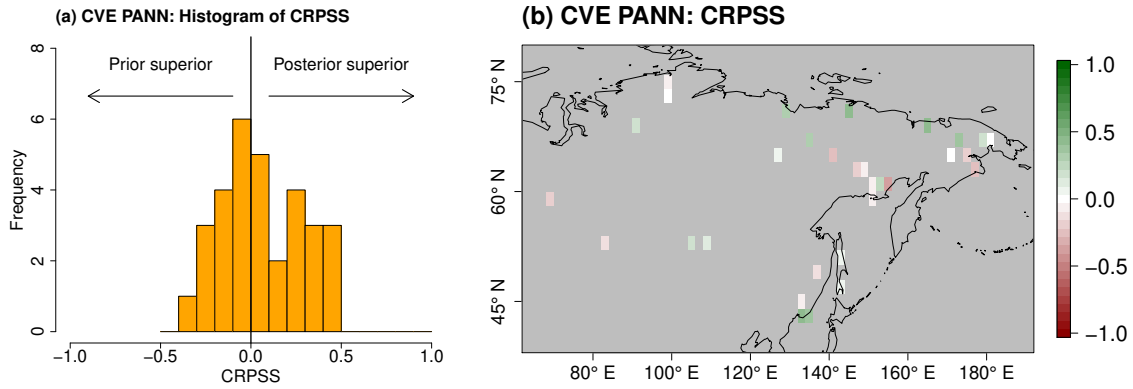

Figure 3.21: Same as Fig. 3.16 but for CRPSS from leave-one-out cross-validation in the P_{ANN} reconstruction.

3.5.3 Reconstruction properties in a controlled test environment

To understand the statistical properties of the LGM reconstructions, we perform ITEs similar to the ones described in Sect. 3.3.1. As the results presented above are produced with the shrinkage RM, only this process stage model is studied. The four PMIP3 LGM models are used as test environment.

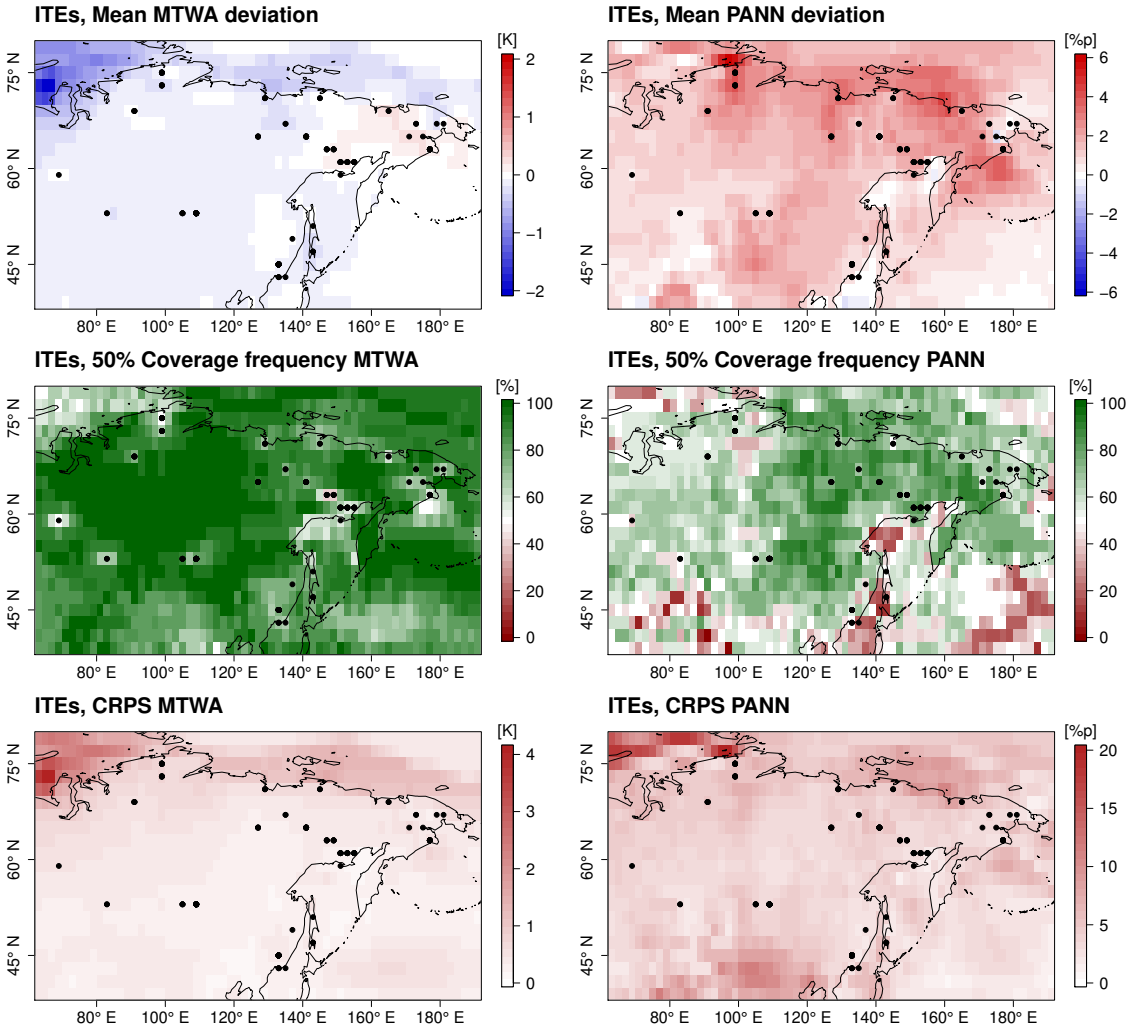

Figure 3.22: ITE results for MTWA (left column) and P_{ANN} (right column). Depicted are mean deviation (top row), coverage frequencies for 50% CIs (middle row), and CRPS (bottom row).

For the MTWA ITEs, the mean deviation from the reference climatology is -0.11 K. The reconstructions show similar robustness as the MH ITEs with non of the ITEs deviating on average more than 0.6 K from the reference climatology. The biases are small across space, with the north-western part of the domain being the only area that exhibits mean deviation magnitudes of more than 1 K. The ITEs are over-dispersive with a coverage frequency above the target value for all experiments, and mean coverage frequencies of 89.9% for 50% CIs, and 99.7% for 90%

CIs. The mean CRPS values are below 1 K in all ITEs which shows the overall good predictive properties of the MTWA reconstructions. CRPS values are highest in the north-western parts of the domain due to biases in that area. Besides, the values tend to be worse in areas with large uncertainties, because the reconstructions are not very sharp in these regions. Summarizing, the ITEs for MTWA show similar statistical properties as the MH ITEs with the shrinkage RM. This includes mostly small biases, over-dispersiveness, and good probabilistic prediction results. This leads to high confidence in the MTWA reconstruction results presented above for all areas except for the north-western part of the domain.

For the P_{ANN} ITEs, the interpretation is more difficult due to the absence of a comparison from the MH. The mean deviation, averaged over all ITEs, is 1.3 percentage points with a standard deviation of 4.1 percentage points. The mean deviations are below 6 percentage points in all grid boxes. Most grid boxes are on average slightly too dry with largest biases along the Arctic Ocean coastline. The experiments are less over-dispersed than the MTWA ITEs, with a mean coverage frequency of 61.5% for the 50% CIs and 98.5% for 90% CIs. Under-dispersion is found particularly in southern and south-eastern parts of the domain, while most other areas are over-dispersed. Similar to the MTWA experiments, the CRPS structures follow mostly the uncertainty patterns, as the systematic biases of the reconstructions are rather small. Spatially averaged, the CRPS is below 6 percentage points in all ITEs, but it exhibits values up to 20 percentage points in the north-western part of the domain. Summarizing, the ITEs show that the confidence in the P_{ANN} reconstruction, in particular the spatial patterns, should be lower than in the MTWA reconstruction, because there are areas with systematic under-dispersive behavior which is not the case in all other applications of this study.

A second issue, that hampers the confidence in the P_{ANN} reconstructions, is that the spatial structures of the posterior are very similar to the unconstrained PMIP3 ensemble (see Sect. 3.5.2). This means, that the proxy data has a substantial influence on the spatially averaged precipitation amount, but small influence on the regional distribution. This is a result of the smaller PMIP3 ensemble spread compared to the uncertainty in the local reconstructions. This issue should be investigated further in future research by increasing the ensemble size or including additional parameters in the BHM, which control the interpolation uncertainty. Doing this in a way which preserves as much physical consistency as possible is challenging and beyond the scope of this study.

3.5.4 Discussion of reconstruction results

Our reconstructions show that Siberia was colder and drier than today during the LGM but warmer and wetter than the PMIP3 ensemble mean. MTWA and P_{ANN} are two important factors for controlling the buildup of ice sheets (Reeh, 1989). Therefore, reconstructions of these two variables can improve the understanding of the reasons for the absence of a Siberian ice sheet during the last Glacial. While LGM reconstructions are just a short snapshot into the whole Glacial cycle and

particularly the Glacial inception needs to be studied to understand why no ice sheet was formed at the onset of the last Glacial, the LGM is widely used as a representative sample of glacial climate conditions (Masson-Delmotte et al., 2013).

The finding of relatively warm summer temperature in Northern Siberia is supported by studies of individual records and subregions (e.g. Meyer et al., 2017; Kienast et al., 2005; Alfimov and Berman, 2001). For Western Siberia, the increased continentality was proposed as the reason for not much colder summer temperature than today (Kienast et al., 2005). In our reconstructions, the regions with warmest MTWA anomalies coincide with regions that are most affected by higher sea level during the LGM, which indicates that the changed land-sea-mask is an important factor for the relatively warm summers in Northern Siberia. Meyer and Barr (2017) inferred from indirect glacier extent modeling that precipitation amount must have been similar to modern values in Pacific Siberia. Our reconstruction does not exhibit much smaller precipitation amounts in Eastern Siberia, which supports their findings of summer temperature being the limiting factor for ice sheet growth in Eastern Siberia. Many studies considered aridity the main reason for the absence of large ice sheets (e.g. Stauch and Gualtieri, 2008), which was supported by the finding of very arid conditions in interior Siberia by Kienast et al. (2005). For Western Siberia, we reconstruct the strongest reduction of P_{ANN}, such that an important role of aridity for the absence of ice sheets in that part of the domain cannot be ruled out.

In our reconstruction framework, higher ensemble member weights are inferred for warmer and wetter PMIP3 models. However, this study does not provide direct insights into the mechanisms that control the climate state of Siberia during the LGM. For example, all PMIP3 simulations are run with the same land-sea-mask and some ESMs produce very cold summers across the whole domain. This shows that increased continentality originating in the lower sea level might not be the only cause of the meridional temperature anomaly gradient. To understand the physical mechanisms behind the spatial patterns, it is necessary to study the differences of the PMIP3 models more extensively. In this comparison, the divergent fit of the ensemble members to the proxy data can provide guidance. New ESM simulations with interactive ice sheets will additionally improve the understanding of the interactions between ice sheet buildup and regional climate conditions during the last Glacial.

3.6 Shortcomings and possible extensions

In this section, shortcomings of the spatial reconstructions, potential solutions, and extensions, which are envisaged for future research, are discussed. We cover issues regarding the inclusion of additional proxy data and climate simulations, the joint reconstruction of multiple climate variables, and the introduction of alternative process stage structures. Common themes are methods to account for climate model inadequacies, to improve the robustness of the reconstruction framework, to raise the physical consistency of the estimates, and to increase the comparability of CFR techniques.

3.6.1 Additional proxy data

In this study, only pollen and macrofossil samples are used as proxy data. Moreover, for the MH reconstructions, a common transfer function is used for both data types by merging them to a common data type, the occurrence of taxa. Thereby, additional information, namely the pollen ratios between different taxa, are discarded. A combination of the PITM framework with a transfer function for pollen ratios on a coarser taxonomic level like biomes or plant functional types could lead to sharper local reconstructions (Thoma, 2017). However, additional non-climatic factors and more complex dependence structures among taxa have to be respected to avoid reconstruction biases. The inclusion of alternative pollen-climate transfer functions in the BHM can be managed by changing the data stage accordingly. This will provide new application areas for the Bayesian framework and new possibilities to compare different transfer functions for example via CVEs.
The incorporation of additional types of paleodata can strengthen the proxy signal and make the BHM applicable in regions where only few pollen samples are available. Two promising extensions are speleothems and SST proxies like Foraminifera assemblages. The accessibility of speleothem data has improved a lot in recent years with community efforts like the PAGES Sisal Working Group (Atsawawaranunt et al., 2018). Speleothems are often more precisely dated than sediment cores and provide an independent climate proxy that can constrain particularly hydroclimate conditions (Bradley, 2015). The combination of pollen and speleothem data is a promising idea to produce joint temperature and precipitation reconstructions. With SST proxies, the BHM can be used to reconstruct marine environments. The combination of pollen and Foraminifera assemblages can produce joint reconstructions of SST over the ocean and near surface temperature over land. For example, the North Atlantic, Pacific, and Arctic Oceans are the least constrained areas of the European and Siberian climate reconstructions. Thus, the inclusion of SST proxies would fill a gap in the reconstructions presented above.

3.6.2 Extended climate simulation ensembles

The ensemble size of the applications presented in this study is very small with $K = 7$ for the MH and $K = 4$ for the LGM, because to date very few simulations with a high enough resolution are publicly available. New and higher resolution simulations should be available in the next years from ongoing community efforts like PMIP4 (Kageyama et al., 2018), the PalMod project (Latif et al., 2016), or the PAGES PALEOLINK Working Group (Ludwig et al., 2019). These simulations will provide new opportunities for reconstructions with Bayesian filtering methods but also new challenges as the set-up and output of the different projects will be less structured than using only PMIP3 simulations.
The empirical correlation structure becomes more robust for larger ensemble sizes which makes the Bayesian reconstruction framework more robust. Therefore, it would be desirable to have larger ensembles available for future applications of our method. The current ensemble covers only modeling uncertainty, while internal variability and uncertainties from forcing reconstructions are

not explicitly included. An extension of our method would use ensembles which account for all these types of uncertainty in a structured way. Uncertainty hierarchies could be build for the different types of uncertainty and climate model specific uncertainty modules would provide a way to better preserve physical consistency in the reconstruction.

Hopefully, larger ensembles can also constrain the range of physically reasonable climate states for a given set of external forcings better. In some regions, the empirical marginal variances from the inter-model differences are currently very large due to large differences between the ensemble members and the small ensemble size. On the other hand, the uncertainty in the precipitation simulations of Siberia is smaller than the proxy uncertainty which warrants an underestimation of the prior uncertainty (see Sect. 3.5.2). This can lead to problems for the spatial reconstructions as described above.

3.6.3 Joint reconstructions of multiple climate variables

In Sect. 3.4.4, we studied effects of joint versus separate reconstructions of climate variables. In ITEs joint reconstructions perform equally or better than separated reconstructions. Good ITE results can be an indicator of a good statistical model but they can also result from unrealistically small inter-model differences. For the Siberian LGM, we reconstructed MTWA and P_{ANN} separately because the local reconstructions have been computed independently for MTWA and P_{ANN} such that there is no proxy constraint on the correlation of the two variables.

The different ensemble member weights λ in joint versus separate MH reconstructions show that different climate models fit the proxy data better for different climate variables. A potential problem arises if the model weights in joint reconstructions are dominated by one variable, as it is the case in the MH reconstructions with MTWA being more constrained by proxy data than MTCO. This can transfer a signal inaccurately from one climate variable to another. Therefore, it might be preferable to have separate ensemble member weights even in joint reconstructions for different climate variables. On the other hand, this reduces the physical consistency of the state estimates as samples are not drawn from the same combination of ensemble members anymore. Balancing physical consistency of state estimates and their fit with proxy data is a challenging but very important task for future research.

The inclusion of more and more proxy types and climate variables in a common BHM is a way to get closer to reconstructions that have the character of climate 'reanalyses', i.e. of physically consistent estimates of past climate states which are adjusted to large compilations of proxy data. Such a 'reanalysis' would facilitate a more direct analysis of physical mechanisms behind past climate changes and provide a more complete picture of past climate states. But it also requires careful modeling of inter-variable dependence structures in the process as well as the data stage.

3.6.4 Process stage structure

To account for inadequacies of climate models in simulating past climate states, we introduced flexible ensemble member weights λ and the shrinkage matrix approach which combines the empirical covariance matrix of the climate simulation ensemble with a matrix that is derived from an independent physically motivated model. Combining ensemble filtering methods with additional techniques to correct model biases in a physically consistent way is an important but also challenging direction of future work on CFRs as a balance has to be found between under-dispersion of the posterior distribution by inducing physical structure and over-fitting to noisy proxy data by enhancing the degrees of freedom. Beyond the strategies implemented in this work, some directions that can be envisaged are the increase of permitted spatial structures in the prior mean by adding patterns calculated from alternative physically motivated models, and the introduction of multiple shrinkage targets in the spatial covariance matrix (Gray et al., 2018).

In this study, the process stage model for precipitation reconstructions does account for non-Gaussian features, that are typically present in precipitation data, only in a very simple way. The reason for this is that the fit of Gaussian approximations of the anomaly fields produced by the climate simulations in the Siberian LGM framework was found acceptable as long as the probability distribution is concentrated significantly away from -100%, which is the case at almost all grid boxes. A rewriting of the current framework to better model non-Gaussianity is challenging but important for future work to avoid the approximations made in this study even though we do not think that this would change the results of our P_{ANN} LGM reconstruction significantly. One possibility is the use of non-linear transformations with Copulas as in Schölzel (2005) and Thoma (2017).

In addition to those extensions of the current filtering framework, different types of process stages, which are independent of simulations with ESMs but still computed from physical principles, can be envisaged. One example are stochastic energy balance models, which are simple stochastic climate models that simulate the energy fluxes between external forcing, atmosphere, and surface. These types of models have become important tools in studies of climate variability (Rypdal et al., 2015) and could be reformulated as process stage in BHMs. The advantage of such a process stage is that it can be fully integrated in the BHM instead of using an offline run with an ESM which does not learn from the proxy data. In addition, spatial reconstructions that are independent of ESM output are more suitable for comparisons of proxy data and climate simulations than the Bayesian filtering approach.

Another valuable extension of our approach would be the computation of reconstructions on hemispheric to global scales. In this case, a more flexible structure for the ensemble member weights is desirable to account for the varying skill of climate models in different regions. On the other hand, the weights should not change too much within small domains to avoid unreasonable spatial heterogeneity. An additional problem is that the stationary shrinkage target matrix would no longer be a good approximation as different regions require different target matrices. This means

that non-stationarities have to be integrated in the shrinkage target. A straightforward approach to accomplish this way is to introduce non-constant parameters in the stochastic partial differential equation (SPDE) behind the Matérn model (Lindgren et al., 2011).
The implementation of these different types of process stages would facilitate more quantitative comparisons of reconstructions and allow a fair testing of modeling assumptions by using ITEs and CVEs. In particular, the reformulation of existing CFR techniques as BHM (Tingley et al., 2012) offers many model comparison techniques that are currently not available as the borders between very different statistical techniques, especially to estimate reconstruction uncertainties, have to be crossed.

3.7 Conclusions

Using ITEs and CVEs, we showed that robust spatial reconstructions with Bayesian filtering methods, which exhibit small biases and are not under-dispersed, are possible as long as the BHM is flexible enough to account for deficiencies of climate simulations and to avoid filter degeneracy. However, both properties can be lost when the BHM does not adequately account for those two issues. The resulting Bayesian model, that is used for spatial reconstructions of European MH and Siberian LGM climate, combines a weighted average of the involved ensemble member climatologies with a shrinkage matrix approach for spatial interpolation and structural extrapolation of the proxy data. It performs well in ITEs except for a strong over-dispersion which is less harmful than under-dispersion because it does not lead to overconfident conclusions. Additionally, the over-dispersion in the controlled test environment of ITEs leaves a margin of error for a potential reduction of the model performance in real world applications. Constraining the PMIP3 ensemble by proxy data adds value as long as the spatial covariance matrix Σ_{prior} contains enough spatial degrees of freedom and a wide enough range of potential climate states. This is not the case in the reconstruction of P_{ANN} in Siberia during the LGM and the European MH MTCO reconstruction, where the process stage uncertainty is smaller than the proxy uncertainty.
Our spatial reconstruction of the European MH climate exhibits dipole structures with warming in Northern and cooling in Southern Europe for MTWA as well as MTCO. Such a structure is also present in the reconstruction of Mauri et al. (2015) and less pronounced in Bartlein et al. (2011), but not in most PMIP3 simulations. The dipole structures occur in our reconstruction because of the overall stronger influence of the proxy data, which supports north-south gradients, on the spatial reconstruction compared to the climate simulations. This shows that spatial reconstructions, which combine the strength of paleodata and climate simulations, and which are in good agreement with local reconstructions from proxy samples not included in the reconstruction, are possible despite existing model data mismatches as long as the chosen CFR method accounts for these mismatches.
Applying the Bayesian filtering framework to pollen-based local reconstructions of Siberian LGM climate leads to a reconstruction with cooler LGM summer temperature and a smaller precipitation

amount than the modern climate which is in agreement with the literature. However, the reconstruction mean is warmer and wetter than that of the PMIP3 ensemble, which is accompanied by higher posterior ensemble member weights for comparably warm and wet simulations. This suggests that some ESMs simulate too cold summer temperature and too small precipitation amounts when run with LGM forcings. Systematic probabilistic model-data evaluations like Stolzenberger (2017) based on our results can assess this hypothesis and should be part of future research. Our results indicate that a North Siberian summer just slightly colder than today is a major factor for the absence of a Siberian ice sheet during the last Glacial. To understand the physical mechanisms behind these findings is an important goal of future research which needs extensive collaboration of paleodata experts and climate modelers.

4 Developing a Bayesian hierarchical model for reconstructing the temperature evolution during the Last Deglaciation

In this chapter, we develop and test a BHM for reconstructing the spatio-temporal temperature evolution during the LD on continental-scale domains. We use a Bayesian framework, which is structured into hierarchical modules. These modules are well-suited to deal with the challenges named in Chapter 1 and 2. In particular, we use the ability to include multiple sources of uncertainty and quantify uncertainties. Our model can be used to better understand spatio-temporal climate variability during the LD, and to analyze the spatial homogeneity of gradual as well as abrupt climate changes. The framework is modular to facilitate flexible modifications of modeling choices and future model augmentations, which account for current shortcomings discussed in Sect. 4.5.

Many proxy data records are available for parts of the LD, but so far only few studies have combined multiple proxy records in a quantitative way. In these studies, the indices, that have been calculated, are mostly regional or global averages (e.g., Clark et al., 2012; Shakun et al., 2012). Going beyond the study of individual records and large-scale averages by computing spatio-temporally distributed CFRs is an important next step to facilitate the quantitative study of the LD. Several ideas are taken from Tingley and Huybers (2010a) and adapted to the non-stationary case. We choose a simpler temporal parameter structure than Parnell et al. (2016) to facilitate an easier extension to gridded spatial domains but still maintain enough temporal flexibility to recover non-stationary structures. Our model's structure shares similarities with Holmström et al. (2015), even though we use different parameterizations to model the temporal update steps.

We develop our model for mean annual temperature and time steps of 200 and 500 years, but an application to other temperature variables like seasonal means is possible. Because of the low temporal resolution of many available proxy records, the focus is on centennial- to millennial-scale climate variations. We acknowledge that there are some very high resolution records available (e.g. Sihailongwan Lake, Mingram et al., 2018), but as the goal of this study is to combine as much of the available records as possible, we choose a lower temporal resolution. We restrict our study to a continental domain. Eurasia is chosen as it features a high paleodata coverage compared to other continents via pollen databases like Neotoma (Williams et al., 2018) and compiled syntheses like Tian et al. (2018). As first and last time step of the reconstruction, we choose the LGM and

MH because of the amount of proxy data and paleosimulations, which are available to additionally constrain those two time slices through a spatial reconstruction module. We test the BHM with PPEs using the transient LD simulation TraCE-21ka with the climate model CCSM3 (Liu et al., 2009b). Thereby, the skill of the BHM to reconstruct point-wise as well as aggregated quantities in idealized experiments can be studied.
In the next section, the transient LD reference simulation for testing spatio-temporal variations of model parameters and for performing PPEs is described. Then, we present the proposed BHM for spatio-temporal reconstructions. In Sect. 4.3 and 4.4, the PPE design and results are shown. Finally, we discuss shortcomings and potential extensions of our framework in Sect. 4.5 and conclude.

4.1 Reference simulation

We use the TraCE-21ka simulation for two purposes: To test the identifiability of BHM parameters under perfect observations (see Sect. 4.2.4), and as reference simulation in PPEs (see Sect. 4.3 and 4.4).
The simulation was run with the coupled atmosphere-ocean model CCSM3 in T31 horizontal grid spacing (around 3.75°) and with 26 vertical levels (Yeager et al., 2006). CCSM3 consists of the atmospheric model CAM3, the ocean model POP, the dynamical vegetation model, which is based on LPJ-DGVM, and the sea ice model CSIM. The simulation started at 22ka from an equilibrium LGM simulation (Otto-Bliesner et al., 2006) and was integrated until 1990 CE. Forcings in the simulation were changed transiently with orbital parameters following Berger (1978), atmospheric greenhouse gases following Joos and Spahni (2008), and the ice sheet configuration as well as the corresponding land-sea-mask and topography is updated every 500 to 1000 years according to the ICE-5G reconstruction by Peltier (2004). Additionally, meltwater discharge into the ocean due to ice sheet retreat was adjusted manually to reproduce reconstructions of sea level rise and geological records on the speed of ice sheet melting (He, 2011).
Our reference fields are computed from the published decadal averages of mean annual surface temperature (TS). We center the first and last time steps of our reconstructions around 21ka and 6ka, and the temperature values are averaged over the interval corresponding to each time step. The spatial domain of interest covers Europe, large parts of Asia, and Northern Africa. The domain ranges from 13.13° W to 166.88° W, encompassing 55 grid boxes in zonal direction, and from 18.56° N to 77.92° N, containing 16 grid boxes in meridional direction. Open ocean areas of the Pacific, Atlantic, Indian, and Arctic Ocean have been excluded from the reconstruction as we only simulate terrestrial proxies and do not account for non-stationarities due to land-sea contrasts. In addition, the Eurasian ice sheet, as inferred by the ICE-6G reconstruction (Argus et al., 2014; Peltier et al., 2015), is masked out. In the late Glacial and early Holocene, additional proxy records are available in northern European regions formerly covered by ice sheets. These records contain valuable information on the European climate evolution and we want our model to be able to

incorporate those records. Therefore, the ice sheet mask is applied transiently, i.e. the domain of interest changes over time. Technically, this is implemented by running the BHM on the whole 55×16 grid box lat-lon grid without constraints from proxy data in the masked out areas. However, PPEs are only evaluated over the domain of interest.
Restricted to the parts of the domain, which are included during the whole reconstruction period, the TraCE-21ka simulation exhibits a spatially averaged warming of around 5 K from the LGM to the MH, and approximately 6 K from the coolest time step (17ka) to the warmest (6ka). The spatially averaged temperature is nearly constant from 21ka to 19ka, decreases gradually from 19ka to 17ka, and increases first gradually and then abruptly into the BA warming around 14.5ka, as shown by the red line in Fig. 4.6 (plotted as anomaly from the MH state). The YD cooling is very weak in the simulation (He, 2011), and followed by another abrupt warming from 12ka to 11.5ka. Finally, a gradual warming until 6ka takes place.
Fig. 4.2 plots the difference of LGM and MH temperature for each grid box of the domain which is never covered by an ice sheet. The spatial structure is characterized by a north-south gradient with warming below 5 K in most areas in the southern half of the domain and up to 20 K at the Taimyr Peninsula and nearby parts of Siberia. An exception to the north-south gradient is the Pacific Ocean coast which exhibits smaller warming than more continental areas of the same latitude.

4.2 Bayesian hierarchical model

4.2.1 Bayesian framework

The goal of the reconstruction is to estimate the temperature field $C(t, x)$, where $t = 1, ..., T$ indexes the time steps and x indexes the grid boxes in the domain of interest, which are identified by the coordinates of their centers. $t = 1$ corresponds to 21ka, and $t = T$ corresponds to 6ka. The estimation of $C(t, x)$ is relying on proxy data $P(t, s)$, where $s = 1, ..., S_t$ are the samples corresponding to time step t. In real world applications, the dating of the samples is uncertain. Including age uncertainties in the model will be a goal of future research, which is discussed in Sect. 4.5. In addition, we include prior information in the form of spatial reconstructions C^{LGM} and C^{MH} for the LGM and MH, which are itself uncertain and contain parameters λ^{LGM} and λ^{MH} to optimize the fit with the proxy data and the dynamical constraints on the spatio-temporal temperature evolution. Finally, the dynamical evolution of C is controlled by a parameter vector ϑ. The involved quantities are structured in the following three modules:

- The proxy data module, which models proxy data conditioned on climate fields. As the focus of this study is on understanding the ability of our model to recover spatio-temporal patterns, a simple Gaussian observation model without correlated errors is implemented. In principal, the module is able to incorporate more complicated proxy-climate structures.

- A spatial reconstruction module, which additionally constrains the first and last time step by knowledge from a precomputed spatial reconstruction on the same spatial grid as the

spatio-temporal reconstruction. These reconstructions can contain information from paleosimulations and independent proxy data like SST reconstructions. The module supports the existence of parameters which adjust the spatial reconstructions to the proxy data. Mathematically, we insert the spatial reconstruction as prior information which introduce physical limits on the first and last time slices of the spatio-temporal reconstruction.

- The spatio-temporal climate evolution module interpolates between observations in space and time using geostatistical methods. It models the temporal trend as a random walk with variable step size, and spatial deviations from this trend by a Matérn-type anisotropic covariance matrix. Since many studies suggest that deglacial warming was stronger in high latitudes compared to lower latitudes, a meridional dependence of the random walk step sizes is introduced.

To combine the three modules in one statistical model, we use Bayesian statistics. Mathematically speaking, the conditional probability distribution of $C(t, x)$, ϑ, λ^{LGM}, and λ^{MH} given $P(t, s)$, C^{LGM}, and C^{MH} is estimated. Inserting the quantities into the Bayesian framework Eq. (2.3), we get the the posterior distribution

$$\begin{aligned}\mathbb{P}\left(C, \vartheta, \lambda^{\mathrm{LGM}}, \lambda^{\mathrm{MH}} \mid P, C^{\mathrm{LGM}}, C^{\mathrm{MH}}\right) \;\propto\; & \mathbb{P}\left(P \mid C\right)\; \mathbb{P}\left(C \mid \vartheta, \lambda^{\mathrm{LGM}}, \lambda^{\mathrm{MH}}, C^{\mathrm{LGM}}, C^{\mathrm{MH}}\right) && (4.1)\\ & \mathbb{P}\left(\vartheta, \lambda^{\mathrm{LGM}}, \lambda^{\mathrm{MH}}\right). && (4.2)\end{aligned}$$

Here, we assume that the proxy data, given the climate evolution, is independent of the spatial reconstruction module and the climate evolution parameter vector ϑ. Next, we separate the spatial reconstruction module for the first and last time step from the climate evolution module, by assuming that $C(2, \cdot)$, ... , $C(T-1, \cdot)$ are conditionally independent of C^{LGM} and C^{MH} given $C(1, \cdot)$ and $C(T, \cdot)$, and that ϑ is controlling the climate evolution but not the spatial reconstruction modules:

$$\begin{aligned}\mathbb{P}\left(C \mid \vartheta, \lambda^{\mathrm{LGM}}, \lambda^{\mathrm{MH}}, C^{\mathrm{LGM}}, C^{\mathrm{MH}}\right) \;=\; & \mathbb{P}\left(C(2, \cdot), ..., C(T-1, \cdot) \mid C(1, \cdot), C(T, \cdot), \vartheta\right) && (4.3)\\ & \mathbb{P}\left(C(1, \cdot), C(T, \cdot) \mid \lambda^{\mathrm{LGM}}, \lambda^{\mathrm{MH}}, C^{\mathrm{LGM}}, C^{\mathrm{MH}}\right) && (4.4)\\ & \mathbb{P}\left(\vartheta\right)\; \mathbb{P}\left(\lambda^{\mathrm{LGM}}, \lambda^{\mathrm{MH}}\right). && (4.5)\end{aligned}$$

Finally, we assume that the spatial MH and LGM reconstructions are independent of each other, and reorganize the terms of Eq. (4.2) - (4.5) to identify the contributions to the BHM from the

three different modules:

$$\underbrace{\mathbb{P}\left(C, \vartheta, \lambda^{\mathrm{LGM}}, \lambda^{\mathrm{MH}} \mid P, C^{\mathrm{LGM}}, C^{\mathrm{MH}}\right)}_{\text{Posterior distribution}} \tag{4.6}$$

$$\propto \quad \underbrace{\mathbb{P}(P \mid C)}_{\text{Proxy module}} \quad \underbrace{\mathbb{P}(C(2,\cdot), ..., C(T-1,\cdot) \mid C(1,\cdot), C(T,\cdot), \vartheta)\ \mathbb{P}(\vartheta)}_{\text{Climate evolution module}} \tag{4.7}$$

$$\underbrace{\mathbb{P}\left(C(1,\cdot) \mid C^{\mathrm{LGM}}, \lambda^{\mathrm{LGM}}\right) \mathbb{P}\left(\lambda^{\mathrm{LGM}}\right)}_{\text{Spatial reconstruction module LGM}} \tag{4.8}$$

$$\underbrace{\mathbb{P}\left(C(T,\cdot) \mid C^{\mathrm{MH}}, \lambda^{\mathrm{MH}}\right) \mathbb{P}\left(\lambda^{\mathrm{MH}}\right)}_{\text{Spatial reconstruction module MH}}. \tag{4.9}$$

In the next three subsections, the proxy data module, the spatial reconstruction modules, and the climate evolution module are worked out in detail. Afterwards, we present the MCMC algorithm that is used for inference. A summarizing DAG of the full BHM is depicted in Fig. 4.1.

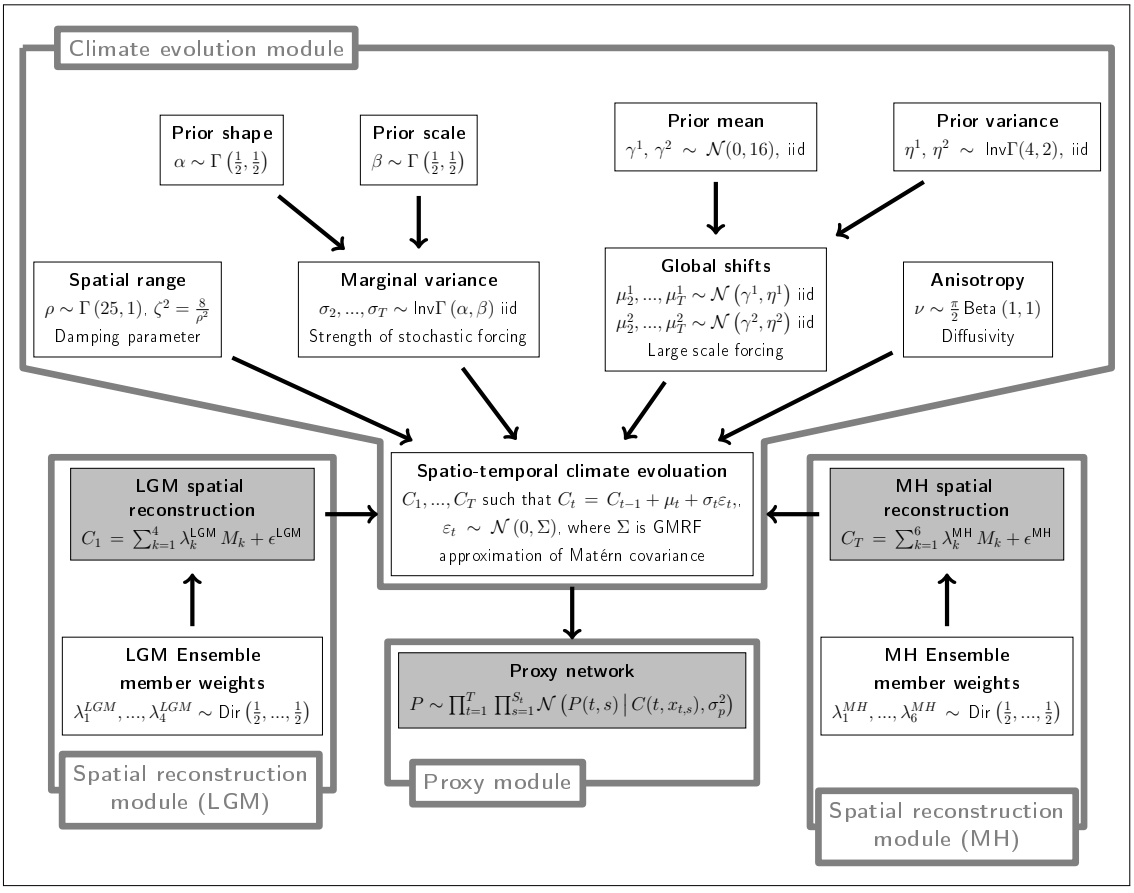

Figure 4.1: DAG corresponding to the spatio-temporal BHM. Known quantities are denoted by gray nodes, while inferred quantities are denoted by white nodes. The proxy module, the two spatial reconstruction modules, and the climate evolution module are indicated by gray boxes. Arrows denote dependencies of the model components.

4.2.2 Proxy data module

In this study, a simple Gaussian proxy module is used, comparable to other PPE studies (e.g. Werner et al., 2013). We assume a Gaussian uncertainty structure with iid observation noise. In Sect. 4.5.1, we discuss limitations of this proxy module and possible extensions. The assumption of iid observation uncertainties leads to a product likelihood for the proxy data module. In addition, we assume that the proxies act locally, i.e. given $C(t, x_{t,s})$, where $x_{t,s}$ is the grid box in which proxy sample $P(t, s)$ is located, $P(t, s)$ is conditionally independent of the climate at all other time steps and locations. Incorporating these assumptions, the proxy data module is given by

$$\mathbb{P}(P \mid C) \;=\; \prod_{t=1}^{T} \prod_{s=1}^{S_t} \mathbb{P}\left(P(t,s) \mid C(t, x_{t,s})\right). \tag{4.10}$$

It remains to specify $\mathbb{P}\left(P(t,s) \mid C(t, x_{t,s})\right)$. We introduce a global proxy uncertainty parameter σ_p quantifying the standard deviation of the proxy samples, and assume that the proxy data $P(t, s)$ is an unbiased estimator of the climate in grid box $x_{t,s}$:

$$\mathbb{P}\left(P(t,s) \mid C(t, x_{t,s})\right) \;=\; \mathcal{N}\left(P(t,s) \mid C(t, x_{t,s}), \sigma_p^2\right). \tag{4.11}$$

4.2.3 Spatial reconstruction module

The spatial reconstruction module is another way of controlling the reconstruction besides the proxy data module. It contains independent constraints for individual time steps, in our case the MH and the LGM. In addition, it serves two purposes:

1. In the Bayesian framework a prior distribution for the climate process has to be defined. The climate evolution module contains dynamical constraints on how to get from one time slice to another. But to provide a closed Bayesian model, at least one time step needs a spatial prior distribution (Tingley and Huybers, 2010a). This corresponds to specifying initial conditions in a climate simulation whereas the evolution module corresponds to the partial differential equation solver in a climate model. Because of the sparse proxy data, having an informative prior distribution for $C(1, \cdot)$ is desirable. Ideally, this distribution is based on physical constraints like large-scale forcings for the respective time slice.

2. The dynamical evolution module is non-stationary and could run beyond physical limits if the paleodata does not provide enough control due to sparsity or high noise levels. To avoid such a behavior, the spatial reconstruction induces limits for the magnitude of climate changes and the spatial heterogeneity. Using spatial reconstructions for the LGM and the MH leads to a LD reconstruction, which is a bridge between those two time slices.

In the PPEs presented in Sect. 4.3 and Sect. 4.4, the spatial reconstruction module has a simple form which is motivated by the RM for Bayesian filtering from Chapter 3. Similar to the climate simulation ensemble, we use samples $M_1, ..., M_K$ of the climate state. The climate during the respective time slice is then modeled as weighted average of the samples (with weights $\lambda_1, ..., \lambda_K$)

and an additive Gaussian noise component ϵ, which represents uncertainties in the spatial reconstructions:

$$C = \sum_{k=1}^{K} \lambda_k M_k + \epsilon, \tag{4.12}$$

$$\epsilon \sim \mathcal{N}(0, \Sigma), \tag{4.13}$$

where Σ is a spatial covariance matrix. The weights $\lambda = (\lambda_1, ..., \lambda_K)$ are fitted in the BHM as described in Eq. (4.9). We use $K = 4$ for the LGM ($t = 1$) and $K = 6$ for the MH ($t = T$) to account for the fact that the number of available paleosimulations is small and that there are more simulations for the MH than for the LGM. For the covariance matrix, we employ the Matérn-type model, which is used as shrinkage target in Sect. 3.2.2, but with spatially homogeneous variances and fixed parameters ρ and ν. We choose $\rho = 75°$ and $\nu = 27°$, which means a longer decorrelation length in zonal direction (see Appendix A for details). To respect the larger uncertainty of LGM reconstructions, we use marginal standard deviations of 5 K for the LGM and 2.5 K for the MH. These values are designed for demonstrative purposes. Finally, a prior distribution for λ has to be specified. A Dirichlet distribution with parameter $\frac{1}{2}$ for each of the K components is used, which gives equal prior weight to each sample and leads to weak constraints on the posterior weights.

Summarizing, we get the following expressions for the two spatial reconstructions that enter Eq. (4.9):

$$\mathbb{P}\left(C(1,\cdot)\,\middle|\, C^{\mathrm{LGM}}, \lambda^{\mathrm{LGM}}\right)\mathbb{P}\left(\lambda^{\mathrm{LGM}}\right) = \mathcal{N}\left(C(1,\cdot)\,\middle|\, \sum_{k=1}^{4} \lambda_k^{\mathrm{LGM}} M_k^{\mathrm{LGM}},\ \Sigma^{\mathrm{LGM}}\right) \tag{4.14}$$

$$\mathrm{Dir}\left(\lambda^{\mathrm{LGM}}\,\middle|\, \left(\tfrac{1}{2}, ..., \tfrac{1}{2}\right)\right), \tag{4.15}$$

$$\mathbb{P}\left(C(T,\cdot)\,\middle|\, C^{\mathrm{MH}}, \lambda^{\mathrm{MH}}\right)\mathbb{P}\left(\lambda^{\mathrm{MH}}\right) = \mathcal{N}\left(C(T,\cdot)\,\middle|\, \sum_{k=1}^{6} \lambda_k^{\mathrm{MH}} M_k^{\mathrm{MH}},\ \Sigma^{\mathrm{MH}}\right) \tag{4.16}$$

$$\mathrm{Dir}\left(\lambda^{\mathrm{MH}}\,\middle|\, \left(\tfrac{1}{2}, ..., \tfrac{1}{2}\right)\right). \tag{4.17}$$

Simulated samples from the TraCE-21ka simulation for the respective time slice are used for $M_1, ..., M_K$. We tried to use simulations with ESMs from the PMIP3 project as well, but found that there where systematic differences between the reference simulation and the PMIP3 ensemble that could not be offset by adjusting λ^{LGM} and λ^{MH}. Since using more sophisticated spatial reconstruction methods is beyond the scope of this study, we chose to use unbiased samples from the reference simulation perturbed by Σ^{LGM} and Σ^{MH} instead. This problem is further discussed in Sect. 4.5.2.

4.2.4 Climate evolution module

Temporal update equation

The climate evolution module has to recover temporal non-stationarities, in the form of millennial-scale trends as well as potential abrupt changes. As can be seen in Fig. 4.6, the spatially averaged trend in the TraCE-21ka is highly nonlinear, which is in agreement with previous proxy syntheses (Clark et al., 2012; Shakun et al., 2012). Therefore we choose a model that does not prescribe a specific parametric form if the trend but let the data inform the trend.
We choose a Markov time series model to facilitate efficient computational structures. The update equation combines the climate state from the previous time step with a global shift and a spatially correlated zero-mean noise process:

$$C(t,\cdot) = C(t-1,\cdot) + \mu_t + \sigma_t \varepsilon_t, \tag{4.18}$$

$$\varepsilon_t \sim \mathcal{N}(0,\Sigma). \tag{4.19}$$

Here, Σ is a 2D spatial covariance matrix. The global shift μ_t recovers large-scale changes from one time step to the next one, and therefore contains most of the warming signal during the deglaciation. As this signal is spatially non-homogeneous, μ_t depends not only on t but also on x as discussed below. The noise process ε_t models regional and local climate changes. The magnitude of small-scale changes is determined by σ_t. A physical interpretation of μ_t, σ_t, and ε_t is that signals from large-scale changes in forcings or atmosphere-ocean circulation are aggregated in μ_t, while small-scale changes, which resemble stochastic forcing, determine the realization of the noise process ε_t.

The spatial covariance matrix Σ is a computationally efficient GMRF approximation of the 2D Matérn matrix (Lindgren et al., 2011), which is a widely used covariance function in geostatistics. The matrix, that we construct, is non-isotropic but stationary in space and constant over time. The justifiability of these modeling choices is discussed below. The matrix is constructed on a regular lat-lon grid using the distance in degrees as metric. A priori, this is an unintuitive measure, because the distance between cell centers of neighboring grid boxes depends on latitude. On the other hand, we find that zonal decorrelation lengths in the TraCE-21ka simulation are shorter in higher latitudes and that the scaling of that shortening is comparable to the decrease in distance between zonally neighboring grid boxes.
Matérn covariance functions are characterized by three parameters: the smoothness, the anisotropy ν, and the range parameter ρ, which controls the spatial decorrelation length. Matérn matrices correspond to stationary solutions of the stochastic Laplace equation (cf. Appendix A), which facilitates a physical interpretation of ε_t and its parameters. We fix the smoothness such that it corresponds to the application of a standard Laplace operator in diffusive transport of spatial white noise $\mathcal{W}$ (Whittle, 1954, 1963). The range parameter ρ models the damping of the dynamic system, and the anisotropy ν corresponds to different strengths of diffusivity in zonal and meridional

direction. Details of the construction of Σ are described in Appendix A.

For ρ and ν, prior distributions have to be specified. For ρ, a Gamma distribution with shape parameter 25 and rate parameter 1 is used, which means that the mean and the variance of the prior distribution both correspond to a decorrelation length of 93.75 degree (25 grid boxes). To parameterize ν, we use the transformation $\nu = \frac{\pi}{2}\chi$, such that χ takes values between 0 and 1. We use a Beta distribution with parameters (1,1) as prior distribution for χ, which means that all values have equal weight. Empirical testing revealed that the influence of these prior distributions on the posterior distributions of ρ and ν is negligible.

Space-time dependence of parameters

We study the space-time dependence of the parameters μ_t, σ_t, ν, and ρ by fitting them to different sub-regions and time periods of the TraCE-21ka simulation. This might artificially increase the skill of the PPEs presented below by a small factor, because we specify parts of the model along the same reference simulation that is used in the PPEs. But since many of the findings from this subsection are supported by paleodata, using a different dataset to perform the tests in this section should lead to similar modeling decisions. Unfortunately, there is currently no comparable gridded dataset available to test this hypothesis. We first introduce the spatio-temporal parameter dependencies that are included in the model, and then present evidence for those choices from experiments with the reference simulation.

From Fig. 4.6 (red line) and 4.2, it can be seen that the large-scale trends of the deglacial evolution are variable in space and time. Many of the spatially averaged changes are very small, but abrupt events can reach up to 2 K differences within 500yr time slices. Moreover, on the timescales that we look at, there is little temporal structure in the global shifts, with small changes followed by abrupt jumps and vice versa. Therefore, the large-scale shifts are modeled as independent random variables for each time step, which are equipped with a common prior distribution such that μ_t makes large jumps only if there is strong data support for them. In addition, proxy records suggest that deglacial warming was much stronger in high latitudes than in low latitudes. Hence, we introduce a linear dependence of shifts on latitude which is parameterized by a weighted average of the two extreme values. In other words, $\mu_t(x)$ is given by

$$\mu_t(x) \;=\; w(x_{\text{lat}})\,\mu_t^1 \;+\; (1 - w(x_{\text{lat}}))\,\mu_t^2, \tag{4.20}$$

where $w(x_{\text{lat}})$ are weights, which depend linearly on latitude.

We use one prior distribution for μ_t^1 and one for μ_t^2, where $t = 2, ..., T$. Both prior distributions are Gaussians, i.e. they are given by $\mathcal{N}(\gamma^1, \eta^1)$ and $\mathcal{N}(\gamma^2, \eta^2)$. Additionally, we have to define hyperpriors γ^1, γ^2, η^1, and η^2, which we choose such that very large jumps (more than approximately 8 K) are unlikely. The hyperpriors are conjugate to μ_t^1 and μ_t^2 which facilitates efficient inference

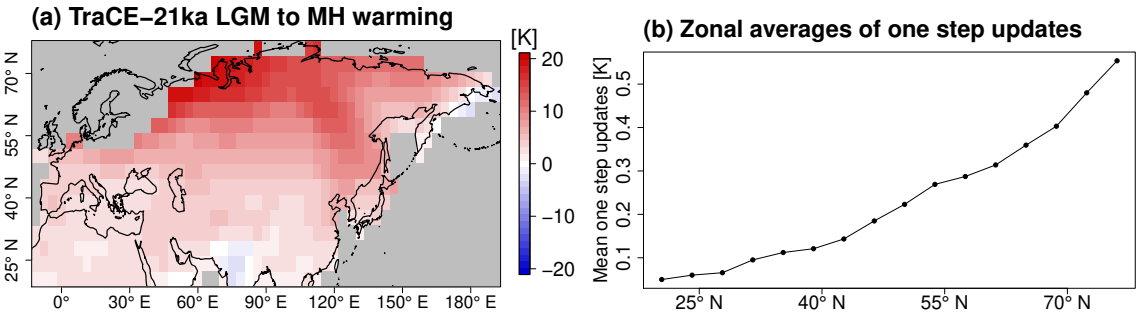

Figure 4.2: Spatial dependence of deglacial warming: Spatial distribution on the left and zonal averages on the right.

with MCMC methods and are given by

$$\gamma^1, \gamma^2 \sim \mathcal{N}(0, 16), \text{ iid} \tag{4.21}$$

$$\eta^1, \eta^2 \sim \text{Inv}\Gamma(4, 2), \text{ iid.} \tag{4.22}$$

Similar to the global shifts, the magnitude of small-scale changes varies strongly between time steps without much temporal structure. Therefore, we model σ_t^2, where $t = 2, ..., T$, as independent samples from a common prior distribution. An inverse gamma distribution is used because it is conjugate to the Gaussian structure of the climate fields, i.e.

$$\sigma_t^2 \sim \text{Inv}\Gamma(\alpha, \beta), \text{ iid,} \tag{4.23}$$

where the shape parameter α and the rate parameter β are hyperpriors. For α, there is no conjugate prior distribution for which a sampler is available. Therefore, we choose $\alpha \sim \Gamma(0.5, 0.5)$. For β, the Gamma distribution is a conjugate prior. Therefore, we set the prior distribution to $\beta \sim \Gamma(0.5, 0.5)$. Both, the prior distributions for α as well as β, are weakly informative and the influence on the posterior distributions was negligible in empirical tests. ρ and ν are fixed in space and time with weakly informative prior distributions.

To demonstrate the adequacy of our modeling choices, we fit the parameters μ_t, γ^1, γ^2, η^1, η^2, σ_t, α, β, ρ, and ν to 500yr averages of the TraCE-21ka simulation for nine different spatial domains: three meridional stripes (40 grid boxes in meridional direction, 5 grid boxes in zonal direction), three zonal stripes (40 grid boxes in zonal direction, 5 grid boxes in meridional direction), a western and an eastern squared domain over Eurasia (24 grid boxes in zonal direction, 24 grid boxes in meridional direction), and an extended Eurasian domain (61 grid boxes in zonal direction, 22 grid boxes in meridional direction). In addition, we fit parameters for three shorter time periods (6ka to 10.5ka, 11.5ka to 16ka, 16.5ka to 21ka) using the extended Eurasian domain. These are very rough analyzes because we do not respect surface changes between subregions like varying land-ice-sea proportions. But they provide indications for spatio-temporal non-stationarity of model parameters. A map of the regions is provided in Fig. 4.3.

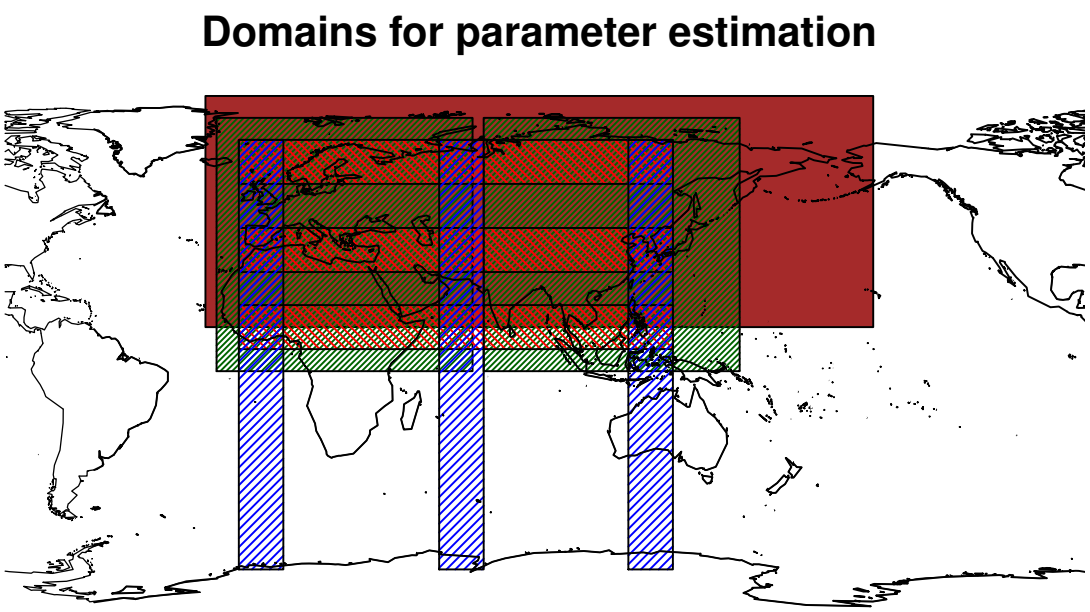

Figure 4.3: Subregions for which the parameters of the BHM are fitted to study spatio-temporal dependencies and identifiability of climate evolution module parameters.

In general, μ_t, σ_t, ν, and ρ are identified well, while there is more uncertainty in some of the hyperpriors. This indicates that it might be possible to improve the assumption of common hyperpriors for μ_t^1, μ_t^2, and σ_t. Since this seems to not hinder the estimation of μ_t and σ_t, it is not problematic in our context.

The latitudinal dependence of μ_t is shown in Fig. 4.4. The total warming is much stronger in higher latitudes and there is more temporal variance as can be seen from higher η values in higher latitudes. But this does not lead to a significant difference in γ, because the estimation uncertainty is higher for the northern zonal stripe. Comparing the other spatial domains, no structural differences can be found, except that the western meridional stripe behaves different than all other domains with higher σ_t and lower μ_t. This means that more small and less large-scale changes occur than in the other domains, which might be a result of the comparably large ocean and small land areas, as well as the retreating British-Irish ice sheet. We find high temporal variations of μ_t between the three Eurasian experiments with shorter time periods, which is in line with different deglacial regimes during different time periods of the LD and early Holocene. Fig. 4.5a and Fig. 4.5b compare μ_t^1 and μ_t^2 for the Eurasian domain. This is another evidence of the larger magnitude of changes in higher latitudes. Fig. 4.2a,b show that linear dependence is a reasonable first approximation of the true latitudinal structure of μ_t. Additional covariates, which might be considered in future work, are the land-ice-sea mask, large-scale circulation, or a different parameterization of meridional dependence.

Fig. 4.5c shows that σ_t has high temporal variability, which justifies modeling it as temporally non-stationary. Large magnitudes of σ_t tend to coincide with large global shifts (large absolute

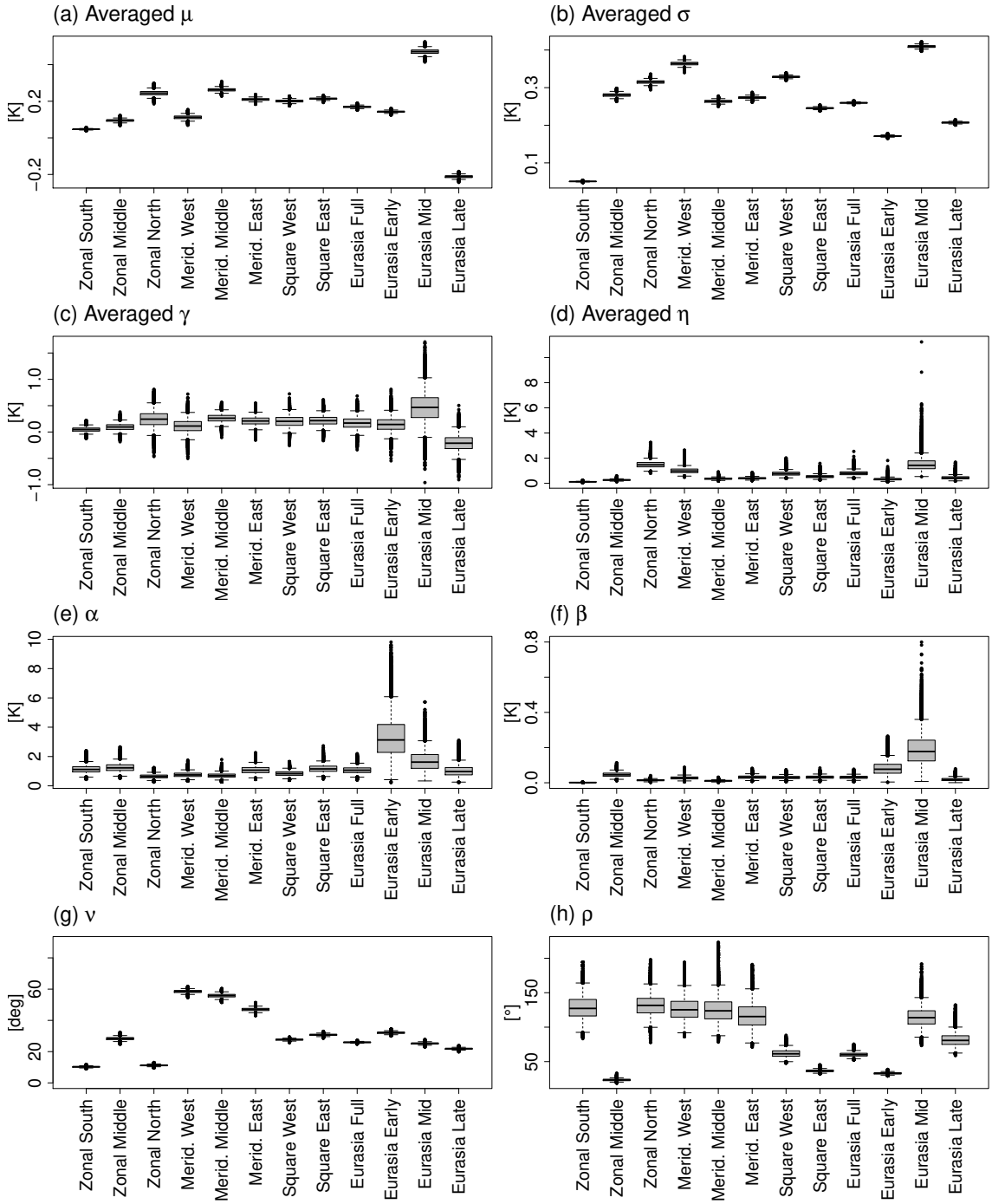

Figure 4.4: Parameter fits for subregions. Plotted are (a) mean of μ_t, (b) mean of σ_t, (c) mean of γ, (d) mean of η, (e) α, (f) β, (g) ν, and (h) ρ.

values of μ_t), which indicates that in the reference simulation strong large-scale changes often come together with large small-scale variability. σ_t exhibits latitudinal dependence (Fig. 4.4b), but less pronounced than μ_t in the mid to high latitudes. Therefore, σ_t is modeled constant in space. α and β show some variations between the spatial domains, but the values are mostly on

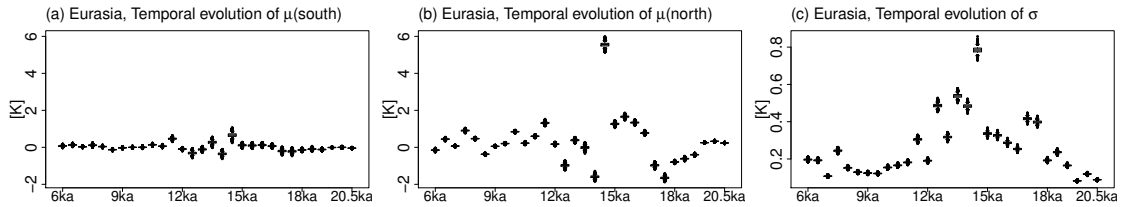

Figure 4.5: Temporal evolution of (a) μ_t^1, (b) μ_t^2, and (c) σ_t for the extended Eurasian domain.

the same order. Therefore, they are kept constant in space.
ρ depends weakly on the geometry of the domain with higher values for the stripes than the more balanced domains, and features one outlier (the middle zonal stripe). Besides that outlier, values are mostly on the same order and therefore ρ is kept constant in space and time. The dependence of ν on the geometry of the domain is much stronger than for ρ with large differences between the zonal and meridional stripes. This can be expected because the manifestation of spatial correlations in the narrow direction is hard to estimate. For the domains with a more balanced geometry, values are relatively similar. That is why we model ν constant in space and time.

4.2.5 Inference algorithm

As there is no closed formula for sampling from the posterior distribution Eq. (4.9), we use a MCMC algorithm that follows a Metropolis-within-Gibbs strategy similar to the MCMC algorithms of Chapter 3. For most variables, the full conditional distributions belong to a family of random variables for which efficient samplers exist. Metropolis-Hastings update steps are used for the remaining variables. The full conditional distributions of all variables are listed in Appendix D.

We update the variables in the following blocks:

1. λ^{MH} is updated using a RW-MH step with Dirichlet distributed proposals;
2. λ^{LGM} is updated using a RW-MH step with Dirichlet distributed proposals;
3. γ^1 and γ^2 are updated using Gibbs sampling from the Gaussian distributed full conditional distributions;
4. η^1 and η^2 are updated using Gibbs sampling from the inverse Gamma distributed full conditional distributions;
5. (μ_t^1, μ_t^2), $t = 2, ..., T$, are updated using Gibbs sampling from the Gaussian distributed full conditional distributions;
6. α is updated using a RW-MH step with Gamma distributed proposals;
7. β is updated using Gibbs sampling from the Gamma distributed full conditional distribution;

8. σ_t^2, $t = 2, ..., T$, is updated using Gibbs sampling from the inverse Gamma distributed full conditional distributions;

9. ν and ρ are updated using a RW-MH step with Beta distributed proposal for ν and Gamma distributed proposal for ρ;

10. $C(1, \cdot)$, ... , $C(T, \cdot)$ are updated sequentially using Gibbs sampling from the Gaussian distributed full conditionals for each time step.

A problem of the algorithm is high autocorrelation of MCMC samples due to the sequential updates of the high dimensional posterior distribution (the number of variables is on the order of 10^5 to 10^6). Therefore, in the PPEs presented below, we simulate 250,000 MCMC samples, discard the first 50,000 samples as burn-in, and retain only every 10th sample (thinning) to reduce the autocorrelation of the samples. The remaining 20,000 samples are used to study the posterior distribution.

4.3 Design of pseudo-proxy experiments

The idea of PPEs is to test CFR methods in an idealized setting, where the true fields are known. Many previous PPEs used replications of existing proxy datasets by taking dates and locations of proxy samples from published proxy compilations. PPEs with artificially created proxy locations and dates were mostly used to study very specific properties of CFR algorithms like the ability to reconstruct long range memory in Nilsen et al. (2018). Here, we choose to not replicate the dates and locations of existing proxy compilations for the LD, because the existing global syntheses (e.g., Clark et al., 2012; Shakun et al., 2012; Sánchez Goñi et al., 2017) do not have a spatial coverage which is high enough to fit our model. Existing compilations with high spatial coverage are only available for smaller regions. One of the goals of this study is to test whether combining several of those compilations for parts of Eurasia (e.g. Bartlein et al., 2011; Cao et al., 2013; Moreno et al., 2014; Tian et al., 2018) creates a sufficient dataset for continental-scale reconstructions of the LD temperature evolution. Hence, the spatio-temporal coverage of proxy data is based on simple estimates of the amount of available pollen samples. To study the influence of specific characteristics of the used proxy synthesis on the reconstruction skill we perform randomized experiments, where we draw a new spatial structure of proxy records and new proxy values in each experiment.

4.3.1 Specification of pseudo-proxy experiments

From the TraCE-21ka reference simulation as described in Sect. 4.1, four different PPE configurations are constructed: low temporal resolution (500yr averages) vs. high temporal resolution (200yr averages), and small proxy noise level (standard deviation $\sigma_p = 1$ K) vs. large proxy noise level ($\sigma_p = 2$ K). 500yr time slices is a lower temporal resolution than that of many available

proxy records. Therefore, we expect that many proxy records have samples in subsequent time slices. For the 200yr case, the spatial proxy sample coverage is lower in each time slice. In addition, the reference simulation features a less homogeneous evolution in space and the BA warming is more pronounced in the 200yr case. The low proxy noise level is an optimistic assumption if for each space-time point only one proxy sample is available but it might be realistic if multiple samples within one grid box can be aggregated. $\sigma_p = 2$ K is a more conservative estimate. With 1.5 K, the uncertainty estimates of Shakun et al. (2012) are between the two cases.

For both 500yr configurations, ten experiments are run, where the proxy network as well as the spatial reconstruction samples $M_1^{\mathrm{LGM}}, ..., M_4^{\mathrm{LGM}}$, and $M_1^{\mathrm{MH}}, ..., M_6^{\mathrm{MH}}$ are randomized. Only five experiments are run for each of the 200yr cases, because of the high computational cost of the MCMC algorithm. A larger number of experiments would be desirable but requires larger computational resources or a better optimized MCMC algorithm. As described in Sect. 4.2.3, $M_1^{\mathrm{LGM}}, ..., M_4^{\mathrm{LGM}}$ are iid samples from $\mathcal{N}\left(C^{\mathrm{ref}}(1, \cdot), \Sigma^{\mathrm{LGM}}\right)$, and $M_1^{\mathrm{MH}}, ..., M_6^{\mathrm{MH}}$ are iid samples from $\mathcal{N}\left(C^{\mathrm{ref}}(T, \cdot), \Sigma^{\mathrm{MH}}\right)$, where C^{ref} denotes the reference simulation.

To specify the proxy network, we first extract proxy locations for the MH and LGM from Bartlein et al. (2011) and Tian et al. (2018), and assign them to their respective grid boxes. This yields 97 grid boxes with proxy samples for the LGM and 279 grid boxes with proxy samples for the MH (see Fig. 4.13 for all proxy locations). The time slices, from which proxy samples were collected for the two syntheses, are 1000yr for the MH (5.5ka to 6.5ka) and 4000yr for the LGM (19ka to 23ka). This is longer than our time steps. Therefore, we use a lower number of records for the LGM and the MH in each PPE. For the 500yr time steps, we remove 25% of the samples, while for the 200yr time steps, we remove 50% of the samples. Between 21ka and 6ka, the number or records is increased linearly. This is a simplification of the progression of available proxy samples, which tends to increase slower during the late glacial and then increases faster as ice sheets retreat and new areas are vegetated such that pollen records are build. The procedure leads to proxy sample numbers ranging from 77 for the LGM to 209 for the MH in the 500yr case, and from 50 (LGM) to 140 (MH) in the 200yr case (see Fig. 4.13 for an exemplifying proxy network). For each time step, we sample a random subset from the originally extracted proxy locations.

Finally, pseudo-proxy values are simulated by corrupting the reference simulation with white noise as specified in Eq. (4.11). Here, we use either $\sigma_p = 1$ or $\sigma_p = 2$. We assume that the noise level of the proxies is known during the inference, because for many proxy types the uncertainty estimates of statistical transfer functions are determined either using calibration structures that do not fit into the BHM structure Eq. (4.9) or the estimation is computationally too demanding for the BHM. In this case, the local reconstructions are computed prior to the spatio-temporal interpolation (Parnell et al., 2015).

4.3.2 Evaluation measures

We evaluate the PPEs by analyzing spatially or temporally averaged summary statistics as well as the spatio-temporal evolution of one exemplifying experiment. The evaluation is centered around four questions:

1. Do the reconstructions exhibit spatio-temporal characteristics which are similar to the reference simulation?

2. Are the uncertainty estimates statistically meaningful and how do they depend on the PPE configuration?

3. Are the reconstructions a good probabilistic prediction of past climate given proxy data?

4. Are the climate evolution parameters ϑ identifiable from sparse and noisy proxy data?

For all questions, the robustness of the answers with respect to the length of time steps, the proxy noise level, and the sampling variability of proxy locations and values is analyzed.
To study the first question, we compute the root mean square error (RMSE) of the posterior mean, and the mean difference of posterior distribution and reference simulation. The RMSE is a common measure for the similarity of the posterior mean and the underlying truth, while the mean difference can identify persistent warming or cooling biases. We focus on the ability to identify slow (millennial-scale) trends, and the recognition and magnitude of the BA warming as an example for abrupt changes. In addition, we compare the sequence of spatially averaged μ_t with the spatially averaged temperature evolution of the reference simulation. Ideally, μ_t should recover the domain-wide trends. The uncertainty estimates are analyzed by looking at coverage frequencies of 50% and 90% CIs, as defined in Sect. 3.3.1. Moreover, we compare the sizes of 50% and 90% CIs in the different experiments to study the dependence of the reconstruction uncertainty on the PPE specification. To summarize the predictive properties of the posterior distribution at a space-time point, we calculate the CRPS, which was defined in Sect. 3.3.1. Finally, we look at the identifiability of the climate evolution parameters ϑ by comparing the posterior mean and spread of parameters, and by studying the sensitivity of parameter estimates with respect to randomization of PPEs. As an indication of the effect of sparse and noisy proxy data on the estimates, we compare the posterior distributions of the 500yr time step PPEs with the parameter estimates for the extended Eurasian domain from Sect. 4.2.4.
Previous PPEs have used additional measures that are based on the correlation of time series extracted from the reconstruction and the reference simulation (e.g., Tingley and Huybers, 2010b; Werner et al., 2013). These are not calculated in this study as the correlation is only a meaningful measure for stationary time series. Since a major goal of our BHM is to estimate non-stationarities, it is not interesting to detrend the reconstructions and calculate correlation-based scores afterwards.

Table 4.1: Statistical properties of PPEs for Eurasia with 500yr and 200yr time steps as well as $\sigma_p = 1$ and $\sigma_p = 2$. We report the RMSE of the posterior mean, the mean deviation from the reference simulation, the mean size of 50% and 90% CIs, the coverage frequency of 50% and 90% CIs, and the mean CRPS. For each configuration, we report the minimum, mean and maximum of the summary statistics from the randomized PPEs.

Experiment	RMSE of posterior mean	Mean deviation	Mean size of 50% CIs	Mean size of 90% CIs	Coverage frequency 50% CIs	Coverage frequency 90% CIs	Mean CRPS
$\triangle t = 500yr$, $\sigma_p = 1$	0.566	-0.160	0.966	2.354	57.5%	93.3%	0.315
	0.615	-0.053	0.984	2.397	59.5%	93.6%	0.329
	0.641	0.005	1.019	2.477	61.3%	94.3%	0.346
$\triangle t = 500yr$, $\sigma_p = 2$	0.716	-0.143	1.115	2.733	53.5%	89.6%	0.397
	0.778	-0.025	1.152	2.811	55.3%	91.8%	0.421
	0.809	0.134	1.183	2.868	58.4%	92.7%	0.435
$\triangle t = 200yr$, $\sigma_p = 1$	0.687	-0.126	0.806	1.941	45.0%	85.3%	0.351
	0.702	-0.072	0.825	1.993	48.2%	87.5%	0.358
	0.733	0.016	0.868	2.121	51.1%	90.1%	0.373
$\triangle t = 200yr$, $\sigma_p = 2$	0.798	-0.173	0.990	2.377	47.2%	87.1%	0.422
	0.817	-0.063	1.019	2.461	48.8%	88.2%	0.433
	0.838	0.038	1.052	2.524	50.5%	88.8%	0.440

4.4 Pseudo-proxy experiment results

We evaluate the PPEs along the research questions formulated in Sect. 4.3.2. Summary statistics of the experiments, aggregated according to the four configurations described in Sect. 4.3.1, are provided in Table 4.1.

4.4.1 Evaluation of spatio-temporal reconstruction patterns

Fig. 4.6 shows the temporal evolution of the spatial mean for all PPEs derived from the respective posterior distribution. We see that all PPEs recognize the major features present in the reference simulation (red line in Fig. 4.6). All PPEs exhibit small variations from around 21ka to 19ka, followed by a cooling till 17ka, and a strong warming until 14.5ka, which ends in the abrupt BA warming from around 15ka to 14.5ka. Thereafter, the 200yr time step PPEs show more pronounced variations than the 500yr time step PPEs until around 11ka. Finally, a slow warming until 6ka is found in all PPEs even though some PPEs show divergent signals for individual time steps.

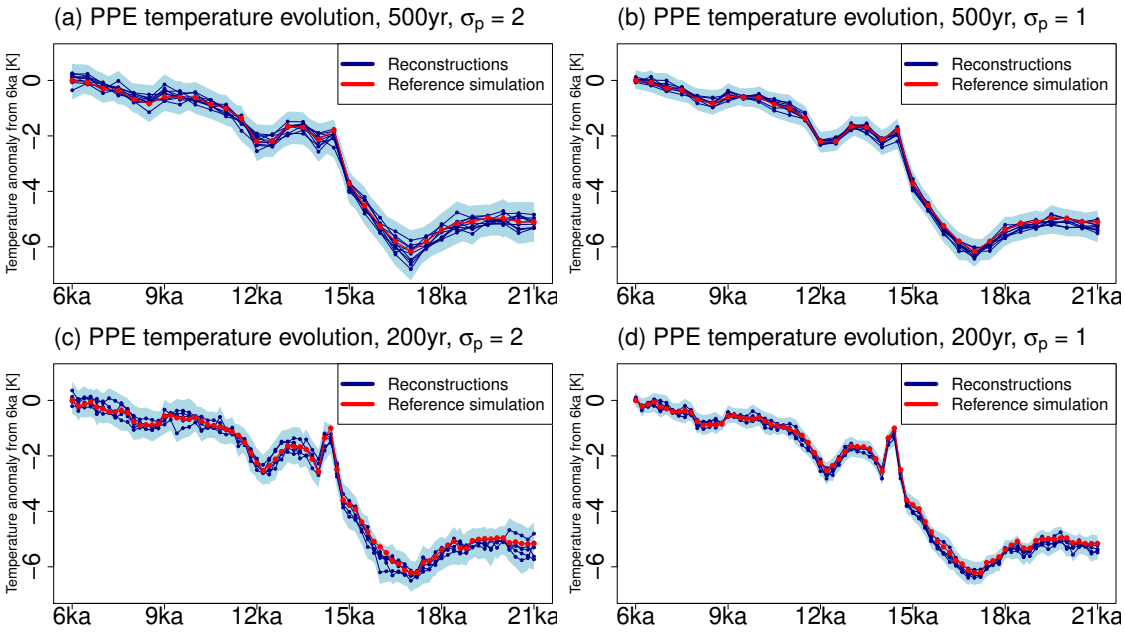

Figure 4.6: Temporal evolution of the PPEs, for 500yr with $\sigma_p = 2$ (a) and with $\sigma_p = 1$ (b), 200yr with $\sigma_p = 2$ (c) and with $\sigma_p = 1$ (d). Blue lines with dots depict the posterior means, of PPEs, shaded areas are 90% CIs, and the red lines depict the reference simulation.

As expected, the uncertainty of PPEs with $\sigma_p = 2$, i.e. higher proxy noise level, is larger than in those with $\sigma_p = 1$ (light blue areas in Fig. 4.6). In general, the PPEs with lower proxy noise level follow the reference simulation more closely. Several 500yr time step experiments with high proxy noise level are persistently too cold between 19ka and 16ka, while others are too warm from 10ka to 6ka. In the high proxy noise experiments, the short range evolution is less precise since many one time step changes are diverging from the reference simulation. Moreover, the low proxy noise level experiments are less sensitive to randomization of the proxy network.

On average, higher noise levels and less proxy data for each time step lead to higher RMSEs of the posterior mean (Table 4.1). The temporally and spatially averaged RMSE is relatively insensitive to randomization of the proxy network with differences of less than 0.1 K among PPEs with the same configuration. All experiments have RMSEs of less than 1 K, which shows that the major features are well recovered considering that the reference simulation exhibits a global warming from LGM to MH of around 5 K and it contains global jumps of up to 2 K from one time step to another. The RMSE decreases from earlier to later times mostly as a result of less proxy coverage during the late Glacial compared to the Holocene (Fig. 4.7a). The highest values are present from 16ka to 17ka, i.e. around the coldest time slice, which is not always well represented in the randomized proxy network. The RMSE variability is higher among PPEs with $\sigma_p = 2$. The most extreme outliers in the 200yr experiments are found around 14.5ka (corresponding to the BA warming) and 16ka (1.3 K RMSE for one PPE), while for the 500yr experiments the largest

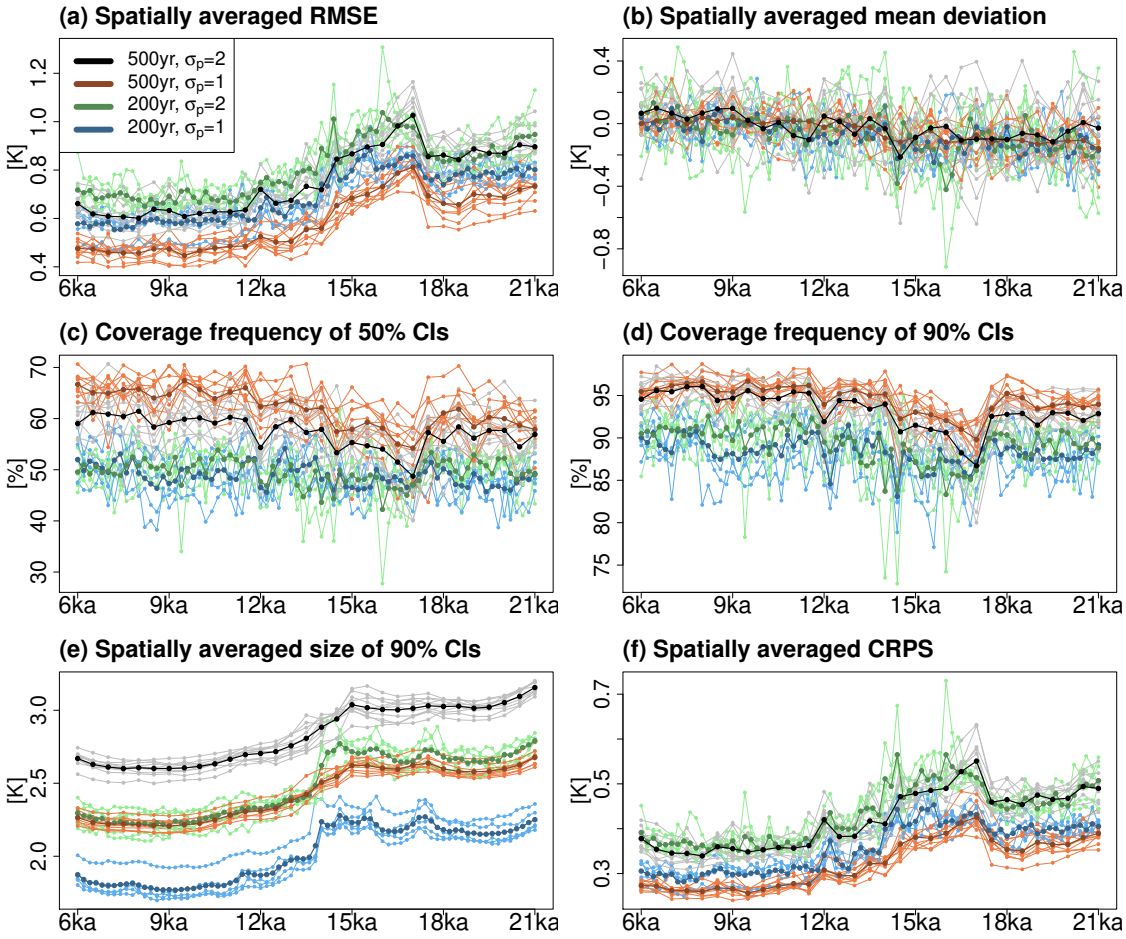

Figure 4.7: Temporal evolution of score functions, (a) RMSE of posterior mean, (b) mean deviation, (c,d) Coverage frequency of 50% CIs, (e) size of 90% CIs, and (f) mean CRPS. PPEs with 500yr time step and $\sigma_p = 2$ are depicted black, 500yr time step and $\sigma_p = 1$ brown, 200yr time step and $\sigma_p = 2$ green, and 200yr time step and $\sigma_p = 1$ blue. Light colors depict individual PPEs and thick lines depict means over all PPEs with the same configuration.

RMSE is exhibited at 17ka.

Averaging over all randomized experiments with the same configuration, we see similar spatial RMSE patterns for all configurations (Fig. 4.8). The spatial distribution is mostly a result of the proxy coverage, with lowest values in Central Europe, Northern Africa, Central Asia, and China, where the proxy density is highest. The RMSE is higher in Siberia, where less proxies are available, and in India, the Middle East, and the Arabian Peninsula, from where no proxy samples are included. In addition, higher RMSEs are present along the domain boundaries, which are often less constrained by proxy data, in particular in the easternmost parts of Siberia. Despite the low proxy coverage, North Africa features low RMSEs due to a combination of less variability in that region compared to other areas and helpful constraints from the numerous European proxy sam-

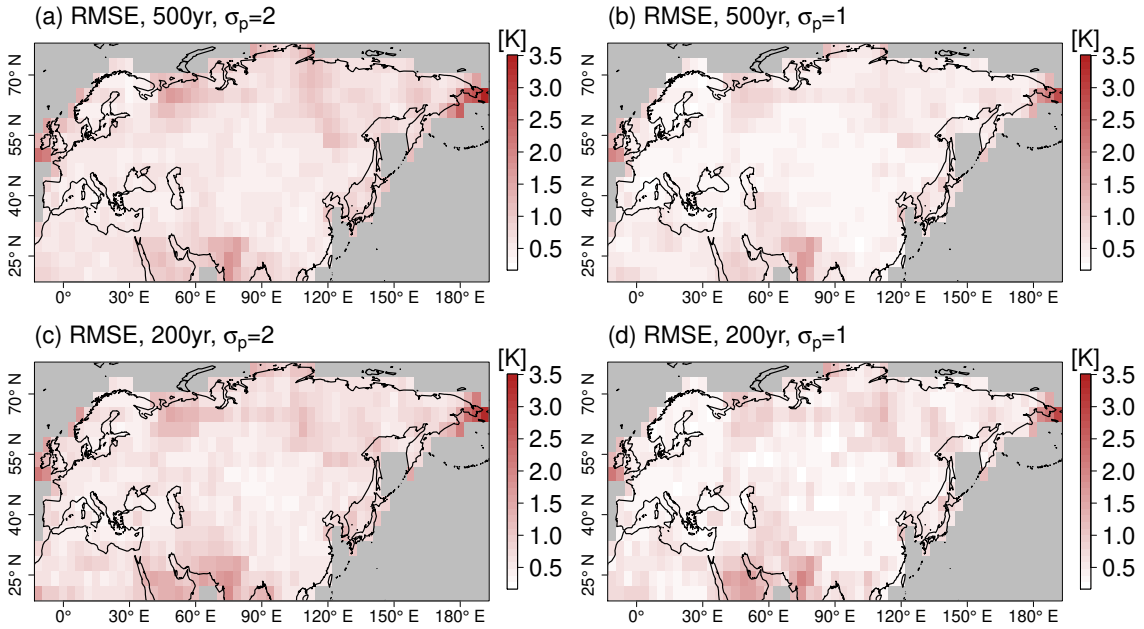

Figure 4.8: Spatial distribution of RMSE, averaged over PPEs with (a) 500yr time step and $\sigma_p = 2$, (b) 500yr time step and $\sigma_p = 1$, (c) 200yr time step and $\sigma_p = 2$, (d) 200yr time step and $\sigma_p = 1$.

ples.

To further understand whether the RMSEs are originating from systematic (large-scale) or random (small-scale) differences, we look at the mean differences of the posterior distribution and the reference simulation. Averaged over time and space, this deviation is mostly small with absolute values below 0.2 K in all PPEs (see Table 4.1), and spatially averaged deviations of less than 1 K for all time steps and PPEs. For the majority of PPEs, we exhibit a small negative mean bias. This can be traced back to mostly negative deviations during the late Glacial, whereas less structure is found in the biases during the Holocene (Fig. 4.7b). The only exception are the predominantly positive deviations in the 500yr time step experiments with $\sigma_p = 2$ from around 10ka to 6ka. The 200yr time step and the high proxy noise level PPEs are more sensitive to changes in the proxy network.

The most consistent negative deviations are found during the BA warming at 14.5ka, which is the time slice with the largest global temperature jump. The BA warming is recognized in all PPEs with high warming rates compared to all other periods, but the magnitude of the warming is still underestimated in most PPEs. This can be explained by the BHM structure. Large jumps are only recovered if they are strongly supported by the proxy data. The BA warming is strongest in the northernmost areas of the domain, but these areas are underrepresented in the proxy network. Therefore, more proxy data in high latitudes would be necessary to better reconstruct the warming signal.

The parameters μ_t are designed to account for large-scale climate changes. Therefore, we compare the spatially averaged evolution of the reference simulation with the evolution of μ_t, by extracting the one time step changes in the reference simulation (Fig. 4.9a-d) and by integrating μ_t backwards from the MH (Fig. 4.9e-h). By looking at the one step changes, we can see that μ_t generally detects the different phases between 21ka and 6ka as described above. However, the uncertainty of the reconstructed climate evolution C is smaller than the uncertainty of μ_t which is expected because C is directly related to the observations whereas μ_t is located in a lower hierarchy level of the BHM. This also shows that the BHM cannot always distinguish between large-scale (μ_t) and small-scale (ε_t) changes, i.e. ε_t contains a portion of the large-scale changes (see also Sect. 4.4.4). Additionally, Fig. 4.9 shows that the abrupt climate changes between 13.5ka and 15ka related to the BA warming are underestimated by μ_t in most PPEs.

Looking at the integrated evolution of μ_t, we find that the overall warming from the LGM to the MH is underestimated by 1 K to 1.5 K. As the overall warming of C is not underestimated (Fig. 4.6), the remaining warming must have been attributed to small-scale changes by the BHM. The large-scale warming from 17ka to 14.5ka is underestimated by μ_t, whereas the integrated warming from 12ka to 6ka is slightly overestimated.

4.4.2 Evaluation of uncertainty estimates

The coverage frequencies of all PPEs are at most five percentage points below the target frequencies of 50% and 90% respectively, and on average 1.6 percentage points higher than the target frequency (Table 4.1). In general, too conservative uncertainty estimates are more desirable than too narrow ones, because too small uncertainty estimates can lead to erroneous conclusions. The coverage frequencies are relatively insensitive to the randomization of the proxy network as can be seen by the differences of at most six percentage points between the highest and lowest values within each configuration (Table 4.1). For 500yr time step PPEs, the coverage frequencies are higher than for 200yr time step PPEs, particularly for the 50% CIs coverage frequencies. This might stem from the higher number of proxy samples per time step or from the existence of more unrecognized small-scale features in the reference simulation when averaged over 200yr time slices. For 500yr time step PPEs, we find higher coverage frequencies for the higher noise level experiments compared to the lower noise level ones, whereas no difference is detected for 200yr time step PPEs.

As expected, the coverage frequencies are lowest between 17.5ka and 14ka, which is the period when the largest changes occur in the reference simulation and when the largest reconstruction biases have been detected. However, even in this period the coverages frequencies are not much below the target frequencies (Fig. 4.7c,d). Low coverage frequencies are present near the reconstruction domain boundaries due to weak proxy constraints in these areas. In general, the spatial structure of coverage frequencies is more heterogeneous than the spatial RMSE patterns (Fig.

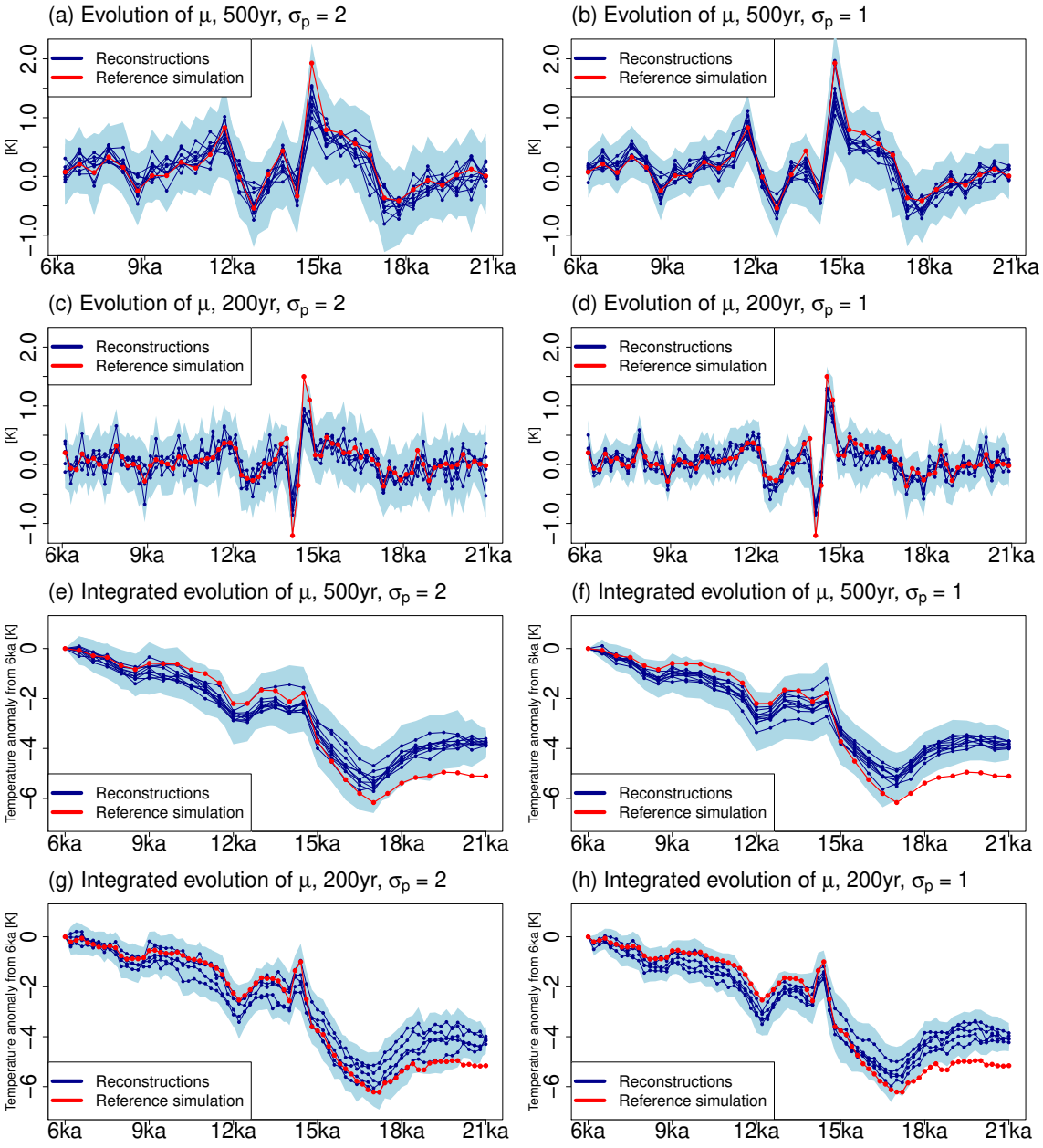

Figure 4.9: Temporal evolution of μ_t in the two top rows and integrated evolution of μ_t backwards from the MH in the two bottom rows. PPEs with (a,e) 500yr time step and $\sigma_p = 2$, (b,f) 500yr time step and $\sigma_p = 1$, (c,g) 200yr time step and $\sigma_p = 2$, (d,h) 200yr time step and $\sigma_p = 1$ are shown. Blue lines with dots depict the posterior means of PPEs, shaded areas are 90% CIs, and the red lines depict the spatially averaged changes in the reference simulation.

4.10).

The CI sizes are larger for the higher proxy noise PPEs as the signal from the proxy data is weaker.

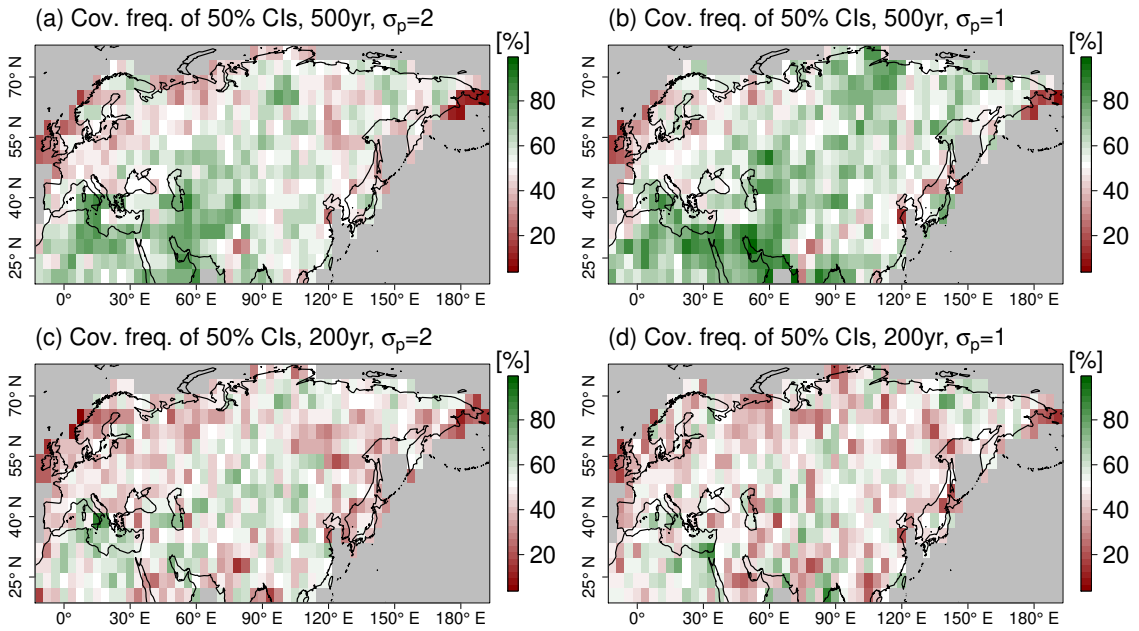

Figure 4.10: Spatial distribution of 50% CIs coverage frequencies, averaged over PPEs with (a) 500yr time step and $\sigma_p = 2$, (b) 500yr time step and $\sigma_p = 1$, (c) 200yr time step and $\sigma_p = 2$, (d) 200yr time step and $\sigma_p = 1$.

In addition, the point-wise CIs are larger for the 500yr than the 200yr time step experiments which might be an effect of simulating a larger total number of proxy samples for the 200yr time step PPEs without increasing the proxy noise level. Averaged over time and space, the CI sizes are insensitive to randomization of the proxy network with differences of less than 0.2 K between PPEs with the same configuration.

The reconstruction uncertainty is small compared with local climate changes, such that a meaningful detection of change signals is possible not just for continental scales but also for local ones. The CIs are larger for the late Glacial than the Holocene (Fig. 4.7e) due to more simulated proxy samples in later times and less uncertainty in the included spatial MH than LGM reconstruction. The spatial distribution of the CI sizes is strongly related to the spatial proxy coverage, with much larger values for areas with low proxy coverage (Fig. 4.11). The smallest reconstruction uncertainties are found in Europe and China, which contain the most proxy samples, followed by the less covered Siberia and North Africa, and largest reconstruction uncertainties in the Arabian Peninsula, India, and the Middle East, from where no proxy records are included.

4.4.3 Probabilistic evaluation using CRPS

The CRPS is a measure to evaluate probabilistic reconstructions. It combines effects from biases with the quality of uncertainty estimates. In that sense, the CRPS is summarizing the last two sections. As for the RMSE, the low proxy noise PPEs exhibit lower CRPS averages. In addition, the

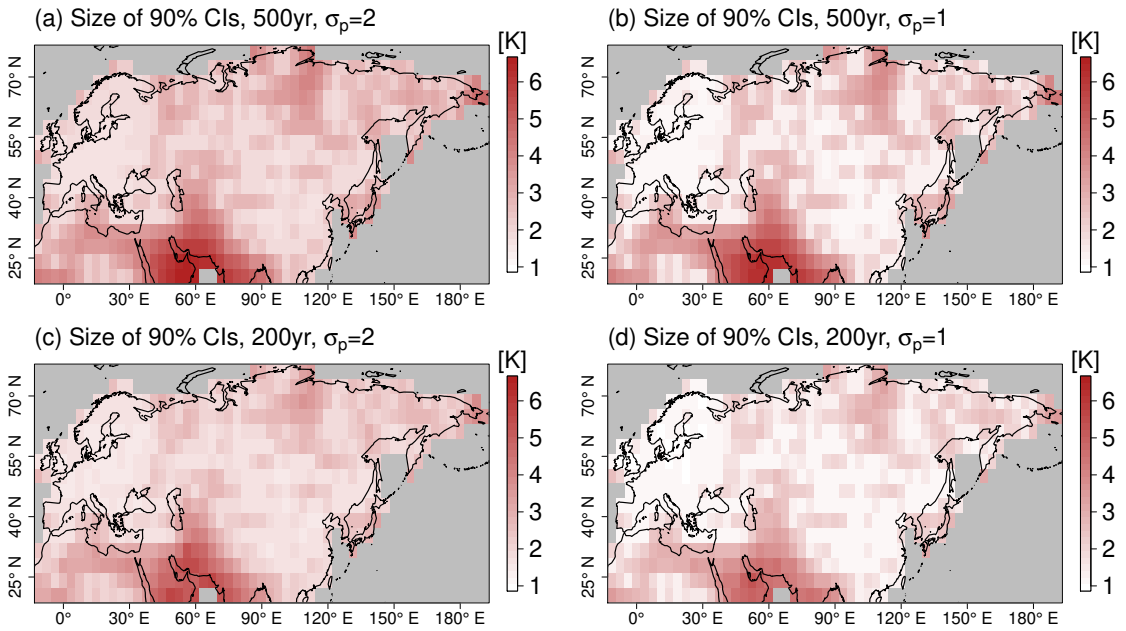

Figure 4.11: Spatial distribution of size of 90% CIs, averaged over PPEs with (a) 500yr time step and $\sigma_p = 2$, (b) 500yr time step and $\sigma_p = 1$, (c) 200yr time step and $\sigma_p = 2$, (d) 200yr time step and $\sigma_p = 1$.

500yr time step PPEs have a slightly lower CRPS than the 200yr time step PPEs due to featuring more proxy samples per time slice (Table 4.1). The reconstructions are insensitive to randomization of the proxy network with mean CRPS differences of less than 0.04 K for PPEs with the same configuration (Table 4.1). These values are even lower than the RMSE sensitivity because some of the mean biases are offset by higher uncertainty estimates in regions with weak proxy signals. The spatial and temporal CRPS structure follows mostly the RMSE patterns but the variations are smaller due to the concurrence of large biases and high posterior uncertainty. This shows that the uncertainty estimates are meaningful in the sense that larger uncertainty estimates occur at time slices and regions where weak proxy signals make a reconstruction of the reference simulation difficult. The CRPS increases towards earlier times due to the higher reconstruction uncertainty in the late Glacial than the Holocene and is highest between 14ka and 17.5ka, where we find the most disagreements between reconstructions and reference simulation (Fig. 4.7f). The highest CRPS values occur in the easternmost parts of Siberia (Fig. 4.12), where the reconstruction uncertainty is underestimated such that biases are not offset by larger uncertainty estimates. The other regions with relatively high CRPS are the previously detected regions with low proxy coverage: along some boundaries of the domain and in Siberia, India, the Arabian Peninsula, and the Middle East.

Finally, we examine one PPE with 500yr time steps and $\sigma_p = 2$ in detail to further investigate the spatio-temporal CRPS patterns (Fig. 4.13). The spatially and temporally averaged CRPS in this PPE is very close to the mean of all experiments with the same configuration. On average,

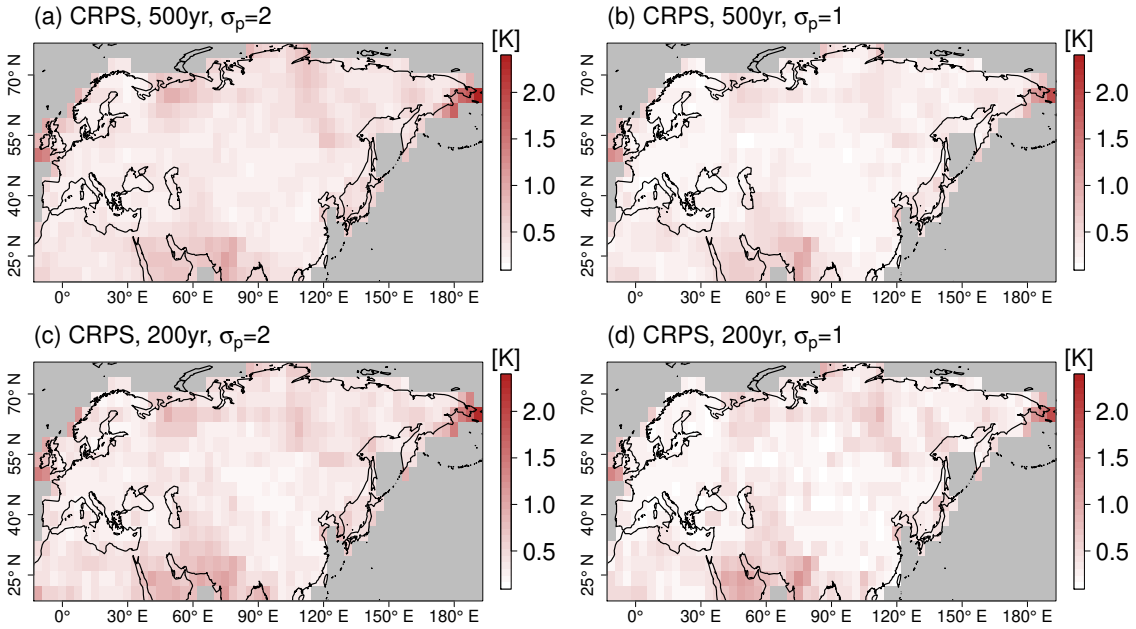

Figure 4.12: Spatial distribution of CRPS, averaged over PPEs with (a) 500yr time step and σ_p = 2, (b) 500yr time step and σ_p = 1, (c) 200yr time step and σ_p = 2, (d) 200yr time step and σ_p = 1.

the highest CRPS values are found in the easternmost parts of Siberia, followed by India, and the northern and north-western boundaries of the domain. Compared with other PPEs, the CRPS is relatively low at the Arabian Peninsula and the Middle East.

During the Holocene, we find mostly low CRPS values except for some regions near the domain boundaries. Starting from 12ka, higher CRPS values occur, in particular near the boundaries of the Fennoscandian ice sheet. From 15ka on backwards, consistently high values are featured in India leading to the high average CRPS in India. Apparently, local features of the late Glacial evolution are not captured well by the reconstruction because there are no proxy records from India included in the pseudo-proxy network and the signal from far away proxy samples does not suffice to recover the local evolution. This is in agreement with India featuring a different late Glacial evolution than China in the reference simulation, where the closest pseudo-proxy samples are located. Several regions with worse than average CRPS values occur for time periods of up to 3k years, particularly near the Caspian and Japanese Sea, and in parts of Siberia and Central Asia.

4.4.4 Parameter identifiability

Fig. 4.14 shows the posterior distribution of the climate evolution parameters ϑ, where the mean of μ_t^1 and μ_t^2 is taken, and μ_t and σ_t are averaged in time. In general, we find that the estimates are mostly consistent across PPEs with the same configuration, which shows that they are insensitive to randomization of the proxy network. The parameters, that define the local change structure (σ_t,

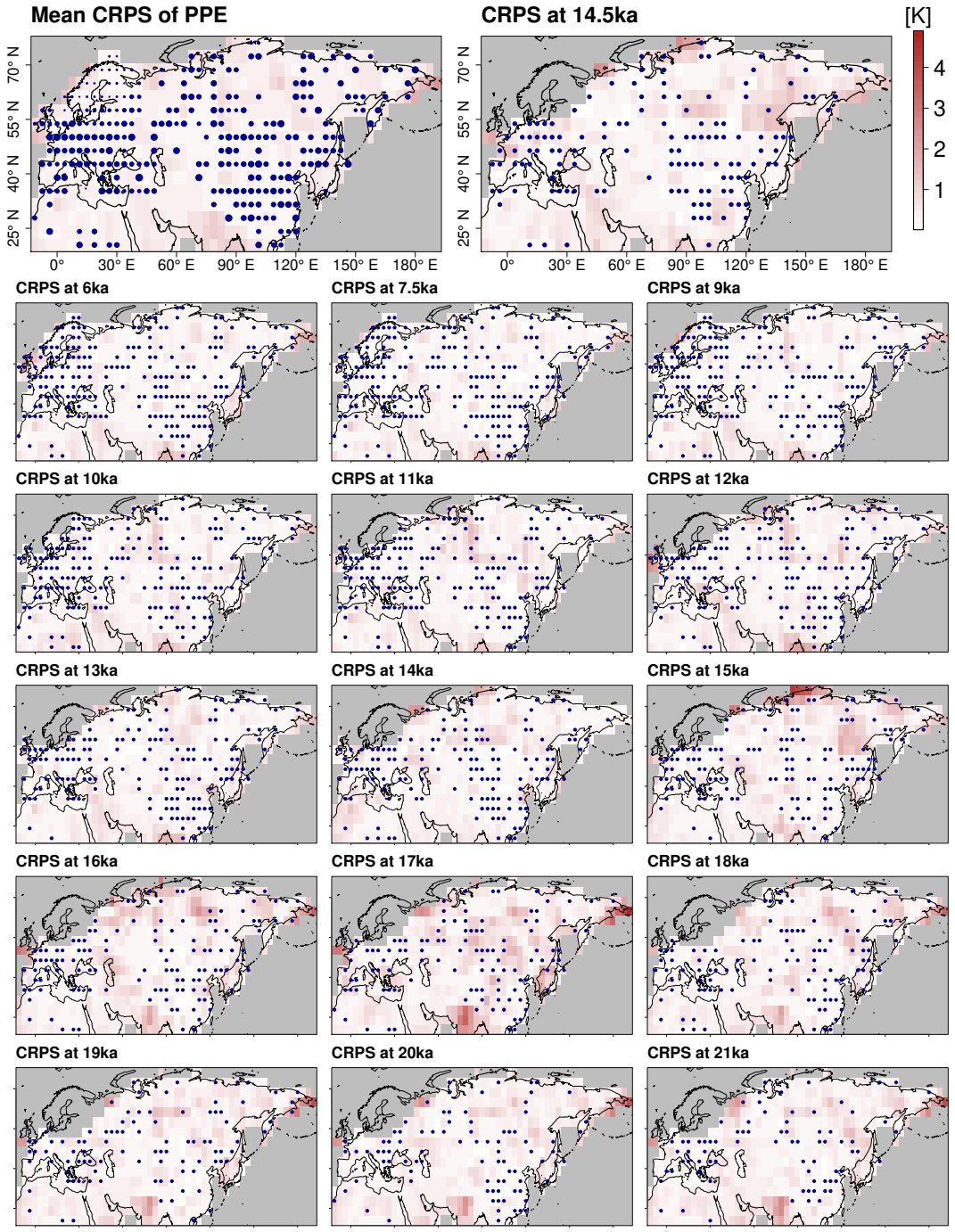

Figure 4.13: Spatio-temporal evolution of CRPS in an exemplifying PPE with 500yr time steps and $\sigma_p = 2$. In the top left, the temporal average of CRPS is plotted with the pseudo-proxy samples given by blue dots. The size of the dots represents the frequency of proxy samples at the grid boxes. On the top right, the CRPS at the BA warming is provided. The other plots show the spatial CRPS distribution of selected time steps. Pseudo-proxy samples at the given time steps are depicted by blue dots.

α, β, ρ, ν), are less consistent than the parameters which characterize the global shifts (μ_t, γ, η). Most parameters are slightly more uncertain in the high proxy noise experiments. When examining these results, it should be kept in mind that global climate models at T31 horizontal grid spacing simulate large scale features more reliably than small scale ones. Therefore, the reconstruction results for global structures should translate better to reconstructions with actual paleodata. On the one hand, the low climate model resolution should lead to improved PPE reconstruction skill of local changes with our simplified statistical model as processes are coarser resolved than in reality. On the other hand, the abrupt change of boundary conditions like topography in the TraCE-21ka (He, 2011) can lead to more abrupt local changes in the reference simulation than in reality which would deteriorate the reconstruction skill.

The identifiability of parameters, measured by the spread of the posterior distribution compared to the posterior mean, is weaker the lower in the BHM hierarchy the parameters are located and the fewer data is available to estimate the parameters. For example, the hyperparameters of μ_t and σ_t are less identifiable than the time averaged μ_t and σ_t, and the signals of μ_t and σ_t at individual time steps are less consistent across PPEs and exhibit more uncertainty than the averages over multiple time steps. Very few of the one step global shifts μ_t differ significantly from 0 K (5% significance level), despite the time average being significantly above 0 K in all PPEs. For example, the only significant shift from one time step to the next one in all 500yr time step experiments is the high latitude BA warming (μ_t^2) from 15ka to 14.5ka. In addition, a few other μ_t^2 values are significantly positive in individual PPEs.

When we compare the parameter estimates from the fully known reference simulation in the extended Eurasian domain (Fig. 4.4) with the PPE estimates, we find two major differences. While the global features μ_t, γ, and η, are mostly on the same order, the estimates from the known reference simulation recognize more local variability with higher values of σ_t, lower values of ρ, and less anisotropy ν. While this might also be a result of having more spatial non-stationarity in the extended Eurasian domain due to the combination of surfaces covered by land, water and ice, it shows that the reconstructions smooth out several local features, in particular for parameters on lower levels of the BHM hierarchy, due to the sparse and noisy proxy samples. In addition, the estimates of μ_t for individual t are a lot more uncertain in the PPEs than in the direct fits from the reference simulation. On average, the estimates of μ_t, when inferred directly from the reference simulation, are a little higher such that integrating μ_t in time leads to a better estimate of the total large-scale warming from the LGM to the MH than in the PPEs. This also indicates that the partition into global changes recovered by μ_t and local variations parameterized by σ_t, ρ and ν is more strict in the estimates from the reference simulation than in the PPEs.

Summarizing, drawing conclusions on the dynamics of the LD from the estimation of model parameters ϑ should be carried out with great caution because of the large uncertainty in many parameters, in particular those attributed to small-scale features. In addition, some parameters depend substantially on the characteristics of the randomized proxy networks, which means that they are prone to biases originating from proxy network specifications. When averaging over long

enough timescales, estimates of μ_t become less uncertain and more consistent across PPEs such that meaningful statistical hypothesis testing is possible for continental-scale patterns on long timescales. Extracting features from the posterior distribution of C might be a more promising strategy towards testing hypotheses on the dynamics of the LD than using the model parameters ϑ on lower BHM levels.

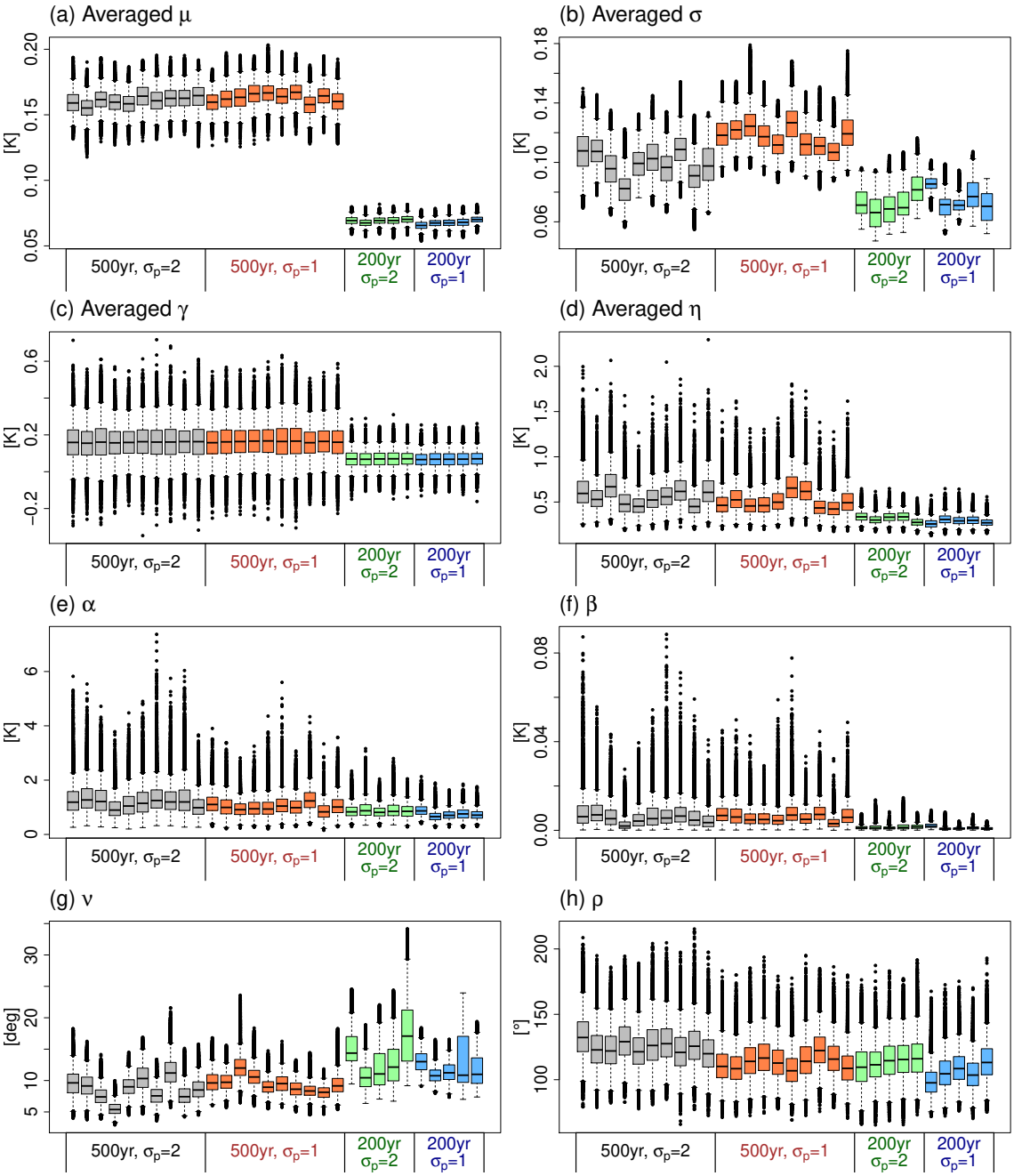

Figure 4.14: Posterior parameter distribution in all PPEs. Each bar corresponds to one PPE. Plotted are (a) mean of μ_t, (b) mean of σ_t, (c) mean of γ, (d) mean of η, (e) α, (f) β, (g) ν, and (h) ρ.

4.5 Potential shortcomings and extensions

This study is a first step towards data-driven spatio-temporal CFRs of the LD. Our BHM is in many regards simplified compared to ESMs used for paleosimulations of the LD, and the PPEs, which we performed, are idealized. While the PPE results are promising, some issues arise from the experiments. In this section, we discuss these issues as well as potential extensions of the BHM. We structure them by first looking at differences between our simulated proxies and real-world compilations of terrestrial proxies. Then we comment on the current requirements of the spatial reconstruction module and on possibilities to improve the ability of the BHM to reconstruct the dynamical structure of the LD. Finally, we discuss strategies to improve the computational efficiency of the MCMC algorithm.

4.5.1 Towards realistic proxy networks

As mentioned above, the pseudo-proxies that we simulate are highly idealized. The reason is that the main goal of this study is to test the ability of the climate evolution module to reconstruct climate fields from sparse and noisy proxy data. Accounting for additional issues of terrestrial proxies is necessary for the application of our BHM to existing and ongoing proxy synthesis efforts. The next step for testing our model in more realistic PPEs is the use of a proxy network, which better resembles the proxy records available in the literature, instead of a randomized proxy network. Such a synthesis is ongoing in the German PalMod project (Tian et al., 2018) and an application of our model to this synthesis is envisaged.

Uncertainties from dating paleodata are an important factor that hampers the extraction of temporal patterns, in particular for proxy data beyond the late Holocene. Including age uncertainties in our BHM will be an important next step, even though 500yr time slices are longer than the random age uncertainty of many Eurasian proxy records during the LD (Binney et al., 2017). However, dating errors might still lead to wrong assignments of proxy samples to time slices and systematic age estimation errors can lead to systematic biases in reconstructions. One way to include age uncertainties in the BHM would be to introduce a probability for assigning proxy samples to time slices and then simulate the corresponding time slice of each proxy sample in every MCMC step. The effect of this strategy would be that less small-scale signals are recovered by the BHM as the proxy signals are smoothed. The probabilities of different time slices could be extracted from dating uncertainties estimated with Bayesian age models. An extension of this strategy is to simulate an age model for each proxy record and then use one MCMC sample from a Bayesian age model in each MCMC step of our BHM to assign the time slices of the proxy samples (e.g. Parnell et al., 2015). Thereby, the temporal structure of dating uncertainty is better respected than just simulating from the marginal dating uncertainty of each sample. The most sophisticated way of integrating dating uncertainties in the BHM is to include a Bayesian age model directly in the BHM similarly to Werner and Tingley (2015). However, their climate evolution module as well as

age model is less sophisticated than our BHM and state-of-the-art algorithms for age-depth modeling of sediment records. Hence, an integration of their approach is not straightforward.

A second issue for the use of sediment cores is that the time steps of proxy samples are irregular because of the non-constant sedimentation rate and the length of time steps do not align between proxy records. In contrast, our BHM currently uses regular time steps. An extension to irregular time steps would be possible but even then uniform time steps across the whole domain would be required. Just refining the temporal discretization to respect the time structure of each proxy record is not feasible with the implemented MCMC scheme, because of computational complexity issues. Therefore, proxy samples need to be assigned to the time slice that fits best. One way to handle this problem is to discard proxy samples which are too far away from the center of time slices to minimize wrong assignments. However, this also means that the proxy network becomes even sparser than it already is.
An alternative strategy is to preprocess proxy records with a temporal BHM for individual proxy records like the BClim model (Parnell et al., 2015), which allows the interpolation of the resulting time series to arbitrary time steps including uncertainty estimates. In addition, BClim already contains the Bayesian age model BChron, such that the estimates from BClim combine uncertainties from pollen-climate transfer functions, age-depth models, and temporal interpolation (Parnell et al., 2016). This means that the preprocessed time series' can be used directly as input in our BHM. A disadvantage of this strategy is that the temporal interpolation in BClim smooths the raw proxy data, such that the combined smoothing of BClim and our BHM could lead to an underestimation of temporal variability.

Currently, our BHM uses not just a regular temporal discretization but also a regular spatial grid, and similar to the assigning of proxy samples to time slices, the assigning to grid boxes can lead to biases if the proxy record is not representative for the climate in the respective grid box. The current parameterization of the spatial covariance matrix Σ uses a triangulation with linear finite elements (Lindgren et al., 2011). To account for the irregular spatial coverage of proxy records, software for the creation of irregular triangulations on the sphere can be used. An issues with this approach is that the parameterization of the anisotropy ν and the latitudinal dependence of μ_t rely on the regular grid. Adaptations of these parameterizations are necessary if an irregular grid is used.

Pollen assemblages record signals from multiple climate variables, including temperature during the growing season, winter temperature, and moisture availability (Huntley, 2012). Our BHM is currently univariate and extending it to multiple climate variables is challenging. While it is reasonable to assume that the structure of the climate evolution module can be adjusted to alternative temperature-related variables like MTWA, precipitation follows a very different dynamical structure and therefore the climate evolution module is probably not well-suited for moisture-related

variables. Finding pollen samples in continental-scale domains, which are sensible to one common variable like MTWA, is a challenging task in itself and reconstructing a variable of minor importance can lead to reconstruction biases (Rehfeld et al., 2016). Therefore, extensive screening of proxy syntheses is necessary to identify which variables are most promising for continental-scale reconstructions.

4.5.2 Requirements of the spatial reconstruction module

As mentioned in Sect. 4.2.3, the current form of the spatial reconstruction module does not account for large-scale biases between the spatial reconstructions and the proxy data in the corresponding time slice. We tried to use PMIP3 simulations for LGM and MH instead of the random samples of the reference simulation, but the included PMIP3 simulations are all warmer than the TraCE-21ka simulation, for the LGM as well as the MH, perhaps due to systematic mismatches between the T31 TraCE-21ka simulation and the finer grid spacing in PMIP3 ESMs or differences between the PMIP2 and PMIP3 experiment protocols. This led to large biases in the reconstructions (not shown), because the MH and LGM estimates were too warm, and this bias was extended to nearby time slices. Therefore, either improved spatial reconstructions have to be included by using not just ESM simulations but also additional independent proxy data to correct large-scales biases of the simulation, or the spatial reconstruction module has to be extended to correct for large-scale biases. The most simple form would be a global shift of the simulations which is estimated from the proxy data, but also more advanced techniques like adjustments along physically motivated spatial structures are conceivable.

4.5.3 Improving the physical consistency of the climate evolution module

The dynamical evolution module is a very simple stochastic climate model, which only captures dynamical features occurring in the proxy data. Therefore, whether dynamical features can be extracted from the reconstruction in future applications relies strongly on the quality of the input proxy compilation. There are possibilities to increase the physical consistency of the module by testing different parameterizations of spatio-temporal non-stationarities, by using an update equation instead of Eq. (4.18), which is motivated by physical considerations, or by including additional information from reconstructions of large-scale forcings. Testing the model with other LD simulations will show whether the model is able to reconstruct the spatio-temporal climate evolution independently of the chosen reference simulation.

The structure of non-stationarities in the BHM is currently restricted to the random walk type temporal updates, the latitudinal dependence of μ_t, and the temporal variability of μ_t and σ_t. There are other covariates that can be imagined for spatial non-stationarities, for example topography or insolation variations. Moreover, to include proxy data from different archive types, in particular marine sediment cores or ice cores, we have to account for the effect of different land surface types

and the effects that occur along land-ice-sea boundaries. In addition, the assumption of spatially uniform ρ and ν could be oversimplifying when too many different circulation patterns are present in the domain, even though the long time steps average out many circulation-related features. This can be seen for example in the different deglacial evolution patterns in the reference simulation for China and India despite their geographical proximity (see Sect. 4.4.3). This problem induces biases in our PPEs as discussed in Sect. 4.4.1.
A more sophisticated temporal structure of μ_t and σ_t could improve the temporal consistency of reconstructions. Possibilities would be time series models with more memory like autoregressive moving-average processes or stochastic volatility models as used in Parnell et al. (2015). The stochastic volatility model leads to Normal-inverse-Gaussian marginal distributions of the climate process C. A computationally efficient implementation of random fields with spatial Normal-inverse-Gaussian noise was developed by Bolin and Wallin (2018). Their model extends the SPDE approach of Lindgren et al. (2011), which we use to implement our Matérn-type covariance model (cf. Appendix A). One issue in our PPEs is the detection of the magnitude of abrupt changes like the BA warming due to the limited flexibility of the BHM to deviate very far from the mean state of μ_t and σ_t (see Sect. 4.4.1). Time series models, which contain more memory in the updates like the ones described above, facilitate easier deviations from the mean shifts.

In addition to improving the spatio-temporal structure of the model parameters, it is also possible to modify the update equation Eq. (4.18) itself with the goal of sharpening the reconstruction by giving less probability to paths which are physically less consistent. This can be achieved by using equations which resemble those used in climate models more closely. For temperature related variables, these are (stochastic) energy balance models (Rypdal et al., 2015). A simple spatio-temporal stochastic energy balance model contains diffusive transport, deterministic large-scale and stochastic small-scale forcings and a linear damping term:

$$d_t C(t,x) - \nabla \cdot (\nu(t,x)\,\nabla)\, C(t,x)\, dt \;=\; \mu(t,x)\, dt + \sigma(t,x)\, d\mathcal{W}(t,x) - \zeta^2(t,x)\, C(t,x)\, dt. \qquad (4.24)$$

Here, the notation is chosen in concordance with Sect. 4.2.4. Discretizing this model leads to a non-separable spatio-temporal covariance model which increases the computational cost of sampling from the full conditional distributions. In contrast to Eq. (4.18), the solution of Eq. (4.24) is a stationary stochastic process for constant parameters ν, μ_t, σ_t, and ζ^2. Hence, to preserve the flexibility of the current update equation, effort has to be put into the spatio-temporal modeling of the parameters. Eq. (4.24) can be extended further by including advective transport in addition to the diffusive transport (Gebhardt et al., 2008). Moreover, the right hand side of Eq. (4.24) can be extended by additional non-linear terms to better resemble radiative processes. In a deterministic context, such an energy balance model has been described and implemented for applications in paleoclimatology by Fanning and Weaver (1996). The main issues with adding advective transport and non-linear terms is that the full conditional distributions of C can no longer be calculated analytically, such that efficient MCMC sampling is a lot more challenging. To improve the iden-

tifiability of parameters in Eq. (4.24) and to allow changes of the equilibrium temperature in space and time, informative priors for the model parameters can be implemented using additional sources of information, for example reconstructions of large-scale forcings. This would push the mostly geostatistical reconstruction problem considered in this work more into the direction of DA problems which combine simplified dynamical models with sparse and noisy data.

4.5.4 Improving the computational efficiency of the MCMC scheme

Due to the high autocorrelation in the MCMC samples resulting from the Metropolis-within-Gibbs strategy, much longer Markov chains have to be simulated to reach convergence of the chains compared with previous BHMs in paleoclimatology. Having to simulate long Markov chains increases the computational cost. Hence, improving the computational efficiency of the MCMC scheme would be desirable to facilitate extended testing of the model.

Two strategies are possible. The first one is to transfer parts of the code to faster programming languages than R. While the use of R makes it possible to use many packages to sample the various random variables involved in the model, many operations are slower in R than in higher level programming languages like C++ or Fortran. At the current stage, where we are still testing and developing the model, we decided to use R to focus our resources on the BHM development. Optimizing code by moving it to faster languages was postponed to future work.

A second strategy would be to use an MCMC algorithm that updates more variables at once to speed up mixing of Markov chains. This would reduce sample autocorrelation and subsequently shorter Markov chains would be required. The problem with updating more variables at once is that the full conditional distributions are no longer available for direct sampling. Instead, approximated proposal densities have to be used. The creation of efficient proposals for high dimensional problems is a very challenging task (Betancourt, 2018). In general, gradient-based methods seem to be most promising. Examples of such techniques are Hamiltonian Monte Carlo or the Metropolis adjusted Langevin algorithm (Apte et al., 2007). Implementing our model using one of those algorithms and compare the performance to the Metropolis-within-Gibbs strategy is desirable for future work.

4.6 Conclusions

We presented a data-driven BHM for reconstructions of the spatio-temporal climate evolution during the LD. The BHM is designed for continental-scale reconstructions using sparse and noisy proxy samples from terrestrial records. Nevertheless, the BHM was developed for comparably proxy rich regions like Eurasia and North America. The model builds on previous spatio-temporal BHMs like Tingley and Huybers (2010a), but modifications were necessary to account for the non-linear trend in temperature between the LGM and the MH as well as other non-stationarities. We use PPEs with the TraCE-21ka simulation as reference simulation to evaluate the reconstruc-

tion skill of the BHM. We find that all PPEs recover the large-scale warming trend between the LGM and the MH well and they detect the abrupt BA warming, even though its magnitude is underestimated. The uncertainty estimates of the reconstruction are reasonable with larger reconstruction spread in areas and time periods with larger biases due to low proxy coverage, unresolved small-scale changes, or underestimated abrupt climate changes. Hence, studying the full posterior probability distribution adds value compared to just analyzing the posterior mean.

Except for abrupt changes, the reconstruction uncertainty is too large to detect statistically significant small-scale signals. However, when averaged over time periods long enough and subdomains large enough, climate changes become statistically significant. While in total, the large-scale temperature evolution is reconstructed well, the partition into large-scale changes (measured by μ_t) and small-scale changes (measured by σ_t and ε_t) is not optimal and parameters lower in the model hierarchy exhibit larger uncertainties. Therefore, drawing conclusions about the dynamics of the LD from the posterior distribution of the model parameters should be treated with caution.

The biggest challenge before our BHM can be applied to a real-world synthesis is to modify the proxy data module because in the current study we do not account for dating uncertainties, irregular temporal spacing of proxy records, and correlated proxy errors. These simplifications are made to focus on the ability of the climate evolution module to reconstruct the deglacial climate evolution from sparse and noisy data. As the PPE results are very promising, future work will focus on overcoming these simplifications. In addition, improving the spatial reconstruction module to mitigate issues from large-scale differences between input spatial reconstructions and proxy data will be important.

5 Conclusions and outlook

In this study, new BHMs were developed for spatial and spatio-temporal reconstruction of past climate states during the Holocene and the last Glacial. The models aim at spatially distributed reconstructions on continental scales which quantify uncertainties from using sparse and noisy proxy data in a statistically rigorous way. Common themes of both BHMs are the use of multiple sources of information, the combination of geostatistical and DA methods, and the application of MCMC methods that utilize Metropolis-within-Gibbs strategies. Pollen and macrofossil samples are the main source for our reconstructions as they are the most abundant terrestrial proxy types. In addition, multi-model ensembles of ESM simulations are included in the spatial reconstruction framework to increase the physical consistency of the reconstructed fields compared to using standard geostatistical methods. The BHMs facilitate quantitative testing of hypotheses in paleoclimatology and the assessment of the consistency of a given proxy synthesis. This includes the fit with large-scale structures, the spatial homogeneity, and the quantitative quality control of the proxy data by identification of potential outliers.

In Chapter 3, it is shown that Bayesian filtering methods possess desirable statistical properties like small biases, over-dispersiveness, and robustness with respect to random errors and sparse data, as long as the BHM accounts appropriately for the small number of ensemble members and biases in the ESM simulations. In our framework, this is accomplished by allowing weighted averages over the ensemble members and by imposing additional spatial structure in the second moments. We test different filtering frameworks in the setting of the European MH climate reconstruction using ITEs with the PMIP3 MH ensemble as controlled test environment, and CVEs with the Simonis et al. (2012) pollen and macrofossil synthesis. PITM is used as transfer function to convert occurrence data for taxa into temperature information. The shrinkage RM and GM perform best in the experiments. For subsequent applications, we use the shrinkage RM because the more flexible structure of this model makes it better suited for additional challenges in real world applications.

In Sect. 3.4, we presented results from reconstructing MTWA and MTCO in Europe during the MH with the Bayesian filtering framework. The reconstructions combine the proxy synthesis of Simonis et al. (2012) with the multi-model PMIP3 MH ensemble. Our reconstruction exhibits a dipole structure with positive anomalies in Northern Europe and negative anomalies in Southern Europe for summer temperature and in a weaker form for winter temperature. These large-scale spatial patterns are in agreement with previous work (Mauri et al., 2015; Bartlein et al., 2011). As the posterior mean is more similar to the local proxy reconstructions than to the prior mean in most regions, we see that the main role of the simulation ensemble is to provide spatial patterns for

interpolation. In addition, this shows that a reconstruction with Bayesian filtering methods, which is in line with reconstructions that do not include ESM simulation output, is possible despite well-known model-data mismatches (Mauri et al., 2014). CVEs reveal that the spatial reconstruction predominantly adds value to the unconstrained PMIP3 ensemble, and analyses of the spatial homogeneity of the posterior distribution indicate a reasonable degree of smoothing through the induced correlation structure.

Using a new synthesis of Siberian pollen records described in Tian et al. (2018), we produced the first spatial reconstruction of Siberian LGM climate. The local reconstructions from the pollen data were computed in the Polar Terrestrial Environmental Systems Division at the Alfred-Wegener-Institute. We reconstructed MTWA and P_{ANN} by combining the local reconstructions with the multi-model PMIP3 LGM ensemble. Our reconstructions show that MTWA was cooler during the LGM than today, and that the amount of P_{ANN} was lower than today. But the spatially-distributed posterior distribution exhibits also a latitudinal gradient with strongest cooling in the mid latitudes and weak cooling or even LGM warming in the highest latitudes of the reconstruction domain. This might be an effect of changes in the land-sea mask, as some of the northernmost areas of the reconstruction domain were above sea level during the LGM but below sea level today. Additional contributions might stem from circulation changes, originating from the different ice sheet topography and land-sea mask. Our reconstructions are warmer and wetter than the PMIP3 ensemble mean. Further examination of the ensemble is necessary to identify mechanisms behind the different climate states that the ESMs produce. For the MTWA reconstructions, CVEs show mostly added value compared to the unconstrained PMIP3 ensemble and ITEs show similar statistical properties than in the MH reconstructions. Meanwhile, the added value of the P_{ANN} reconstructions is very small, and they are less dispersed in ITEs due to a smaller ensemble spread compared to the proxy uncertainty. Our results indicate that a relatively warm summer temperature was a major factor for the absence of a Siberian ice sheet during the last Glacial and changes in precipitation were less important. This result is in agreement with recent findings of Meyer and Barr (2017).

Using PPEs with the TraCE-21ka simulation of the LD, we showed that spatial reconstruction methods can be extended to spatio-temporal domains using geostatistical methods optimized for sparse and noisy proxy networks. These interpolation methods allow data-driven bridges between spatial reconstructions at time slices for which additional information, for example in the form of climate simulation ensembles, are available. The BHM, that we developed, is non-stationary in space and time such that millennial-scale trends are recovered and abrupt events like the BA warming are detected with high accuracy. The BHM provides state estimates and large-scale dynamical structures which exhibit desirable statistical properties, like small biases of the posterior mean and coverage frequencies near the target values. On the other hand, the magnitude of abrupt events tends to be underestimated due to the sparseness of proxy data and local dynamical properties are often not well recovered because of limited signal in the proxy samples and a simplified representation of local spatio-temporal structures.

Possible extensions of the BHMs in future work were presented in Sect. 3.6 and 4.5. In addition to these ideas, some focuses and directions of current research in paleoclimatology will stimulate the use of BHMs for CFRs.

Community efforts like PAGES, PMIP, and Neotoma aim at common data formats for paleodata and paleosimulations which improves the accessibility of available proxy records and ESM simulations. Therefore, more data will be available for the inclusion in BHMs in the next years. Moreover, a focus on providing raw and metadata for the quantification of age-depth uncertainties and uncertainties in the proxy-climate relationship will improve transfer functions and subsequently the quantification of reconstruction uncertainties.

Another major topic of current research is a better scientific understanding of paleodata. This includes new methods to test which climate variables are recorded by proxies and advances in the understanding of spatio-temporal dependencies of proxy uncertainties. While much progress has been made in understanding the error structures of marine proxies and ice cores via process modeling (Dolman and Laepple, 2018; Münch and Laepple, 2018), similar advances for terrestrial proxies, particularly pollen-climate transfer functions, are still in progress.

The increase of computing capacities leads to a growing number of transient paleosimulations and first community efforts for transient simulations beyond the Holocene (Ivanovic et al., 2016). These reconstructions can provide additional constraints for the temporal structure of CFRs similar to the incorporation of equilibrium simulations in the spatial reconstructions of this study.

In addition to equilibrium simulations, high resolution regional climate simulations of past conditions are a growing branch of paleoclimatology (Ludwig et al., 2019). Because the number of available simulations is still very small and multi-model ensembles are not yet available, a replacement of the global simulations used in this study by regional ones will not be possible in the near future. But regional simulations can provide new ways for the downscaling of global simulations to bridge the gap in spatial scales that are resolved by ESMs and proxy data.

Technical advances have facilitated the application of Bayesian methods in the last three decades. But standard MCMC strategies like the Metropolis-within-Gibbs algorithms used in this study reach their limits for spatio-temporal reconstructions with grid box numbers on the order of 10^4. To increase reconstruction domains, spatio-temporal resolution, and the number of jointly reconstructed variables, new methods need to be introduced. On the one hand, the parallelization of algorithms and the use of graphical processing units for faster matrix operations can speed-up currently used MCMC algorithms. On the other hand, more sophisticated MCMC techniques like gradient-based techniques, and particle MCMC methods produce samples with less autocorrelation such that shorter Markov chains are required. In addition, approximate Bayesian methods like INLA provide less precise approximations of the posterior but need much less computation time. All these techniques can help to overcome the computational bottlenecks in the MCMC algorithms used in this study.

Summarizing, current research efforts aim at making more paleodata and paleosimulations available to the community, at improving our understanding of the climate signals contained in proxy

records, and at providing more sophisticated Bayesian inference techniques to overcome current computational limitations. This study is a starting point for developing CFR methods using Bayesian statistical frameworks beyond the Common Era. If successful, all these scientific advances will facilitate improved spatial and spatio-temporal climate reconstructions with BHMs and lead to a better understanding of past climate changes and the mechanisms behind these changes.

Appendix

A Gaussian Markov random field approximation of Matérn covariance matrices

In this appendix, we describe technical details of the GMRF approximation of Matérn covariance matrices, which is used in the shrinkage approach of Chapter 3 and as model for local climate changes in Chapter 4.

Matérn covariance functions are characterized by three parameters: the smoothness, the anisotropy ν, and the scale ζ. ζ is proportional to the inverse of the range ρ. We parameterize Σ by ρ instead of ζ because ρ can be directly interpreted as a decorrelation length. Following Lindgren et al. (2011), we use the empirical relation $\zeta^2 = \frac{8}{\rho^2}$ such that ρ corresponds approximately to the radius of the 0.1 correlation isocline in the isotropic Matérn model. For simplicity, we fix the smoothness of the Matérn covariance function at 1, which means that the resulting Gaussian fields are differentiable. Moreover, we let the anisotropy parameter ν parameterize a diagonal diffusion matrix $\upsilon := \text{Diag}\,(\sin(\nu), \cos(\nu))$. Thus, ν models the ratio of zonal versus meridional decorrelation length.

A continuous mean zero Gaussian process $\hat{\varepsilon}$ with this Matérn covariance is the stationary solution of the stochastic Laplace equation

$$\left(\zeta^2 - \nabla \cdot (\upsilon\, \nabla)\right) \hat{\varepsilon}(x) \;=\; \mathcal{W}(x) \tag{A.1}$$

on $\mathbb{R}^2$ (Whittle, 1954, 1963),which is a linear SPDE. Here, $\mathcal{W}$ is spatial Gaussian white noise with unit variance, ∇ denotes the gradient, and $\nabla\cdot$ the divergence operator. Based on Eq. (A.1), Lindgren et al. (2011) compute a GMRF approximation of Matérn fields using a linear finite element discretization. This GMRF approximation is given by a sparse precision matrix (inverse covariance). More precisely, every row and column has at most 13 non-zero entries.

The discretization of Lindgren et al. (2011) uses linear basis functions ϕ_n, $n = 1, ..., N$, where N is the number of grid boxes, and ϕ_n is equal to one at the center of the n-th grid box and zero at the center of all other grid boxes. By standard finite element theory, the Galerkin discretization of Eq. (A.1) leads to the following linear system

$$\left(\zeta^2 M + G\right)\varepsilon \;=\; F, \tag{A.2}$$

where $M_{i,j} = \langle \phi_i, \phi_j \rangle$ is the mass matrix, $G_{i,j} = \langle \nabla\phi_i, \nabla\phi_j \rangle_\upsilon$ is the stiffness matrix, and $F \sim \mathcal{N}(0, M)$ is the stochastic forcing (Lindgren et al., 2011). Lindgren et al. (2011) show that replacing M by the diagonal matrix $\tilde{M}$ with entries $\tilde{M}_{i,i} := \langle \phi_i, 1 \rangle$ in Eq. (A.2) yields a good GMRF approximation of the stationary solution of Eq. (A.1). This means that we letε be the solution of the linear system

$$\left(\zeta^2 \tilde{M} + G\right)\varepsilon \;=\; \tilde{F}. \tag{A.3}$$

ε is a multivariate zero mean Gaussian random variable with precision matrix

$$\Sigma^{-1}(\nu, \rho) \;:=\; K\,\tilde{M}^{-1}\,K, \tag{A.4}$$

$$\text{where } K := \left(\zeta^2 \tilde{M} + G\right). \tag{A.5}$$

For the regular 2D grid, that is used in this study, it can be shown that Σ^{-1} is assembled by sequentially adding the contribution of each basis function to Σ^{-1}. This contribution is given by the 5×5 matrix

$$\begin{array}{ccccc}
0 & 0 & \upsilon[1,1]^2 & 0 & 0 \\
0 & 2\upsilon[1,1]\upsilon[2,2] & -2\upsilon[1,1]a & 2\upsilon[1,1]\upsilon[2,2] & 0 \\
\upsilon[2,2]^2 & -2\upsilon[2,2]a & a^2 + 2(\upsilon[1,1]^2 + \upsilon[2,2]^2) & -2\upsilon[2,2]a & \upsilon[2,2]^2 \\
0 & 2\upsilon[1,1]\upsilon[2,2] & -2\upsilon[1,1]a & 2\upsilon[1,1]\upsilon[2,2] & 0 \\
0 & 0 & \upsilon[1,1]^2 & 0 & 0
\end{array} \tag{A.6}$$

Here, $a = 2*(\upsilon[1,1]+\upsilon[2,2])+\zeta^2$. In Eq. (A.6), the middle element is the grid box corresponding to the respective basis function, while columns belong to meridional and rows to zonal neighbors. Thus, the $N \times N$ matrix Σ^{-1} can be assembled by adding up N matrices of size 5×5, which is computationally efficient.

Because our reconstruction domain is a polygon, boundary conditions for Eq. (A.1) need to be specified. We use zero Neumann boundary conditions, such that no additional terms have to be added besides Eq. (A.6). Instead, we just have to delete the respective rows and columns of Eq. (A.6) for grid boxes near the boundary. A disadvantage of using Neumann boundary conditions is that the marginal variances near the boundary are reduced compared to the marginal variance of the stationary solution on $\mathbb{R}^2$. Based on empirical testing, we add three rows and columns of grid boxes on all four sides of the domain. These are discarded after the inference.

B Additional technical information for spatial climate reconstructions

B.1 Determination of glasso penalty parameter

To determine the glasso penalty parameter ξ, we first recognize that for values smaller than $\xi = 0.3$ the resulting matrices become numerically unstable in our application due to the small ensemble size. Five values for ξ were tested: 0.3, 0.5, 0.7, 1.0, and 2.0. Larger values lead to sparser precision matrices and therefore to weaker spatial correlations. This in turn means that the local information from the proxy data is spread less into space. For each of the five parameters, we perform CVEs with the RM and compare the resulting BS (see Sect. 3.3.2). While the smallest penalty parameters have the best mean BS, the differences are generally small (see Appendix C.1). On the other hand, the influence of the penalty term in Eq. (3.20) on the overall regression increases from 79.5% for $\xi = 0.3$ to 98.5% for $\xi = 2.0$. Based on these diagnostics, we choose $\xi = 0.3$ for the reconstructions in Sect. 3.4 to get a numerically stable covariance, which performs at least as good as other choices of ξ in CVEs, and is comparably little influenced by the penalty term.

B.2 Full conditional distributions

The Metropolis-within-Gibbs approach samples (asymptotically) from the full conditional distribution of each variable, i.e. the distribution of the variable given all other variables. Some variables are treated block-wise to account for correlations between them, while others are updated sequentially if they are independent from each other or the joint distribution is too complicated for efficient sampling. In this appendix, we detail the conditional distributions that are used for sampling.

To sample the transfer function parameters, we introduce PG distributed augmented variables γ_l^T, where $T \in T(P)$ and $l = 1, ..., L(T)$ are the number of observations for taxon T (Polson et al., 2013). γ_l^T is PG distributed given $\beta^T = (\beta_1^T, ..., \beta_6^T)$ and climate data $C(l) = (C_1(l), C_2(l))$, where

C_1 and C_2 denote MTWA and MTCO. More precisely, the full conditional distribution is given by

$$\gamma_l^T \mid \beta^T,\, C(l) \quad \sim \quad \mathrm{PG}(n = 1, X_{C(l)} \cdot \beta^T), \tag{B.1}$$

$$\text{where } X_{C(l)} := \left(1, C_1(l), C_2(l), C_1(l)C_2(l), C_1(l)^2, C_2(l)^2\right). \tag{B.2}$$

Including the Gaussian prior defined in Sect. 3.2.1, the full conditional of β^T is Gaussian distributed:

$$\beta^T \mid P_m^T,\, P_p^T,\, \gamma_1^T, ..., \gamma_{L(T)}^T \quad \sim \quad \mathcal{N}\left(V_\gamma X^t \kappa^T, V_\gamma\right), \tag{B.3}$$

$$\text{where } V_\gamma := \left(X^t \Gamma X + B^{-1}\right)^{-1}. \tag{B.4}$$

Here, X is a matrix with rows $X_{C(l)}$ for $l = 1, ..., L(T)$, Γ is a diagonal matrix with entries γ_l^T, B is the 6×6 prior covariance matrix of β^T, and κ^T is a vector with entries $\left(P^T(l) - \frac{1}{2}\right)$, where $P^T(l)$ is the presence or absence of taxon T in observation l. In our case, B is a diagonal matrix with all values equal to 10. Details on the definition of PG variables and the augmented Gibbs sampler can be found in Polson et al. (2013).

The full conditional distribution of the proxy correlation parameter ϱ in the LGM reconstructions with the Gaussian data stage Eq. (3.11) is given by

$$\varrho \mid P, C_p \quad \sim \quad \mathcal{N}\left(\left(C_1^l, ..., C_S^l\right) \,\middle|\, O \cdot C_p, \Sigma_{\text{proxy}}\right) \,\mathrm{Beta}\left(\tfrac{1}{2}, \tfrac{1}{2}\right). \tag{B.5}$$

This distribution does not follow a standard probability distribution. Therefore, we use RW-MH to update ϱ.

Sampling from ϑ depends on the specific version of the process stage which is used. In the RM, $\lambda = (\lambda_1, ..., \lambda_K)$ is influenced by its prior and by the Gaussian distribution of C_p given λ:

$$\lambda \mid C_p \quad \sim \quad \mathrm{Dirichlet}\left(\tfrac{1}{2}, ..., \tfrac{1}{2}\right) \,\mathcal{N}\left(C_p \,\middle|\, \sum_{k=1} \lambda_k^K \mu_k, \Sigma_{\text{prior}}\right). \tag{B.6}$$

This full conditional does not follow a probability distribution from which we can sample directly. Therefore, a RW-MH update is used for updating λ.

In the KM, the full conditional of $\omega = (\omega_1, ..., \omega_K)$ is Dirichlet distributed given $z = (z_1, ...z_K)$ and its Dirichlet prior:

$$\omega \mid z \quad \sim \quad \mathrm{Dirichlet}\left(\tfrac{1}{2} + z_1, ..., \tfrac{1}{2} + z_K\right). \tag{B.7}$$

Given ω and C_p, z is categorically distributed:

$$z \mid \omega,\, C_p \quad \sim \quad \mathrm{Cat}\left(\alpha_1, ..., \alpha_K\right), \tag{B.8}$$

$$\text{where } \alpha_k := \frac{\omega_k \cdot \exp\left(-\frac{1}{2}\,(C_p - \mu_k)^t\, \Sigma_{\text{prior}}^{-1}\,(C_p - \mu_k)\right)}{\sum_{i=1}^K \left(\omega_i \cdot \exp\left(-\frac{1}{2}\,(C_p - \mu_i)^t\, \Sigma_{\text{prior}}^{-1}\,(C_p - \mu_i)\right)\right)} \tag{B.9}$$

If shrinkage covariance matrices are used, the parameters (α,ρ,ν) have to be chosen in each MCMC step from one of the K predefined parameter sets. We use a uniform prior, i.e. all seven parameter sets have the same prior probability. Then, the full conditional of τ, which indexes the parameter set, is given by

$$\tau \,\Big|\, C_p, \hat{\mu} \quad \sim \quad \mathcal{N}\left(C_p \,\Big|\, \sum_{k=1} \hat{\mu}, \Sigma_{\text{prior}}(\alpha_\tau, \rho_\tau, \nu_\tau)\right), \tag{B.10}$$

where $\hat{\mu}$ is given according to conditioning the process stage on the process stage parameters ϑ of GM, RM, or KM. We update τ using an Metropolis-Hastings step with uniformly distributed proposals, which choose $\tau = k$ with probability $\frac{1}{K}$.

In the MH reconstruction framework with PITM data stage, we update $C_p(x)$ for $x \in x_P$ sequentially using RW-MH sampling. The set of all grid boxes besides x is denoted by Y_x, and let $\Sigma^{-1}_{\text{prior}}(a, b)$ be the block of the inverse covariance matrix containing the rows a and columns b. The full conditional distribution of $C_p(x)$ depends on the pollen samples $P_p(s)$ with location $x_s = x$, the climate $C_p(Y_x)$ at the other locations, and the process stage parameters ϑ. It does not follow a standard distribution:

$$C_p(x) \,\Big|\, P_p,\, C_p(Y_x),\, \vartheta \quad \sim \quad \mathcal{N}\left(\tilde{\mu}_k(x), \left(\Sigma^{-1}_{\text{prior}}(x, x)\right)^{-1}\right) \prod_{\substack{s \\ \text{with} \\ x_s = x}} \prod_{T \in T(s)} \text{logit}\left(X_{C_p(x)} \cdot \beta^T\right), \tag{B.11}$$

$$\text{where} \;\; \tilde{\mu}(x) \;=\; \hat{\mu}(x) - \left(\Sigma^{-1}_{\text{prior}}(x, x)\right)^{-1} \Sigma^{-1}_{\text{prior}}(x, Y_x)\left(C_p(Y_x) - \hat{\mu}(Y_x)\right). \tag{B.12}$$

The step size of the random walk proposals is controlled by the conditional variance $\left(\Sigma^{-1}_{\text{prior}}(x, x)\right)^{-1}$.

In the LGM reconstruction framework with Gaussian data stage, the full conditional distribution of $C_p(x_P)$ follows a Gaussian distribution:

$$C_p(x_P) \,\Big|\, \left(C^l_1, ..., C^l_S\right),\, \varrho,\, \vartheta \quad \sim \quad \mathcal{N}\left(\tilde{\mu}(x_P), \left(\Sigma^{-1}_{\text{prior}} + O^t\, \Sigma^{-1}_{\text{proxy}}\, O\right)^{-1}\right), \tag{B.13}$$

$$\text{where} \;\; \tilde{\mu}(x_P) \;=\; \left(\Sigma^{-1}_{\text{prior}} + O^t\, \Sigma^{-1}_{\text{proxy}}\, O\right)^{-1} \left(\Sigma^{-1}_{\text{prior}}\, \hat{\mu}(x_P) + O^t\, \Sigma^{-1}_{\text{proxy}} \left(C^l_1, ..., C^l_S\right)\right). \tag{B.14}$$

Conditioned on $C_p(x_P)$ and ϑ, $C_p(x_Q)$ follows a Gaussian distribution:

$$C_p(x_Q) \,\Big|\, C_p(x_P),\, \vartheta \quad \sim \quad \mathcal{N}\left(\tilde{\mu}(x_Q), \left(\Sigma^{-1}_{\text{prior}}(x_Q, x_Q)\right)^{-1}\right), \tag{B.15}$$

$$\text{where} \;\; \tilde{\mu}(x_Q) \;=\; \hat{\mu}(x_Q) - \left(\Sigma^{-1}_{\text{prior}}(x_Q, x_Q)\right)^{-1} \Sigma^{-1}_{\text{prior}}(x_Q, x_P)\left(C_p(x_P) - \hat{\mu}(x_P)\right). \tag{B.16}$$

Pseudo-code for the MCMC algorithm is given in Algorithm B.1. In the algorithm, M denotes the number of MCMC steps.

B.3 Metropolis coupled Markov chain Monte Carlo algorithm

As described in Sect. 3.2.3, the multi-modality of the KM in combination with the high dimension of the posterior makes the standard MCMC algorithm very inefficient. In our specific formulation,

Algorithm B.1 Pseudo-code for MCMC algorithm

```
Initialize MCMC chain
for j = 1, ..., M do
  if Data Stage = PITM then
    if TF Modus = 'FULL' then
      for T ∈ T(P) do
        for l = 1, ..., L(T) do
          Sample from full conditional of γ_l^T
        end for
        Sample from full conditional of β^T
      end for
    else if TF Modus = 'MODULARIZED' then
      Load j-th sample of β^T
    end if
  end if
  if Data Stage = Gaussian then
    Update ϱ with RW-MH sampler
  end if
  if Process stage formulation = RM then
    Update λ with RW-MH sampler
  end if
  if Process stage formulation = KM then
    Sample from full conditional of ω
    Sample from full conditional of z
  end if
  if Covariance matrix = Shrinkage then
    Update τ with independent Metropolis-Hasting sampler
    Set Σ_prior according to τ
  end if
  if Data Stage = PITM then
    for x ∈ x_P do
      Update C_p(x) with RW-MH sampler
    end for
  end if
  if Data Stage = Gaussian then
    Sample from full conditional of C_p(x_P)
  end if
end for
for i = 1, ..., M do
  Sample from full conditional of C_p(x_Q)
end for
```

the inefficiency is manifested in a very slow mixing of z, because conditioned on C_p the likelihood of choosing a new model z_k from one MCMC step to the next one is very small. This problem could be shifted to other variables by integrating out z, but than the conditional Gaussian structure of C_p would be lost which would lead to new challenges for generating efficient MCMC strategies. Therefore, we apply a MC^3 or parallel tempering strategy. The original strategy from Geyer (1991) was adapted to parallel computer architectures by Altekar et al. (2004) and applied in a paleoclimate reconstruction problem by Werner and Tingley (2015).

We run A MCMC chains in parallel, and after every B steps, we use an additional Metropolis-Hastings step to swap the states of the Markov chains a_1 and a_2 with probability $0 \leq p_{a_1,a_2} \leq 1$, where p_{a_1,a_2} is calculated from the Metropolis-Hastings odds ratio. The Markov chains are created by exponentiating the process stage and the data stage by constants $g_1 = 1 > ... > g_A > 0$. The first Markov chain ($g_1 = 1$) asymptotically retains the original posterior distribution for all variables, whereas the subsequent chains sample from a flatter posterior distribution, in which it is easier to jump from one kernel to another. Following empirical testing, we run the European MH reconstructions with $A = 8$ parallel chains, levels $g_1 = 1, g_2 = 1.25^{-1}, ..., g_8 = 1.25^{-7}$, and swaps after every $B = 30$ steps. The number of outer loops J is chosen such that the number of MCMC samples M equals $M = JB$. Pseudo-code for the MC^3 algorithm is given in the Algorithm B.2.

Algorithm B.2 Pseudo-code for MC^3 algorithm

```
Initialize A MCMC chains
for j = 1, ..., J do
  for a = 1, ..., (A − 1) do
    Run B steps of chain a following the MCMC Algorithm 1 to update γ, β, ω, z, τ, and
    C_p(x_P)
  end for
  for a = 1, ..., (A − 1) do
    Calculate Metropolis-Hastings ratio r of chains a and a + 1
    Swap chains a and a + 1 with probability p = max(1, r)
  end for
end for
for i = 1, ..., M do
  Sample from full conditional of C_p(x_Q)
end for
```

C Additional results from spatial reconstructions of the European mid-Holocene climate

C.1 Result from spatial reconstructions with alternative process stage models

In Sect. 3.4 of the manuscript, we only presented results with the RM using the shrinkage covariance matrix approach. In this Section, we provide results from spatial reconstructions of European MH climate with the other five process stage models, which where described in Sect. 3.2.2 and compared in Sect. 3.3 using ITEs and CVEs. Summary measures are given in Table C.1. Figures C.1 to C.6 show results of the different experiments. Similar to the ITEs, the choice of covariance matrix impacts the reconstruction more than using a different process stage formulation, i.e. GM, RM or KM.

The three reconstructions with shrinkage matrix have very similar estimates of the spatially averaged mean MTWA anomaly and the spatial structures are in very good agreement. For most of the domain, the posterior mean structure is also very similar for MTCO, but differences are found in the north-western part of the domain, that is least constrained by proxy data. Therefore, this is the region that is most affected by the different process stage formulations. Here, the KM is warmer than the RM, while the GM is cooler than the RM. This can be explained by the RM being the most flexible of the three such that the GM is pulled more towards the ensemble mean while the KM is pulled towards the MPI-ESM-P climatology, which is the favored model. On the other hand, the RM departs from the ensemble mean towards the MPI-ESM-P climatology but not as strong as in the KM. The difference in the northern-western part of the domain is carried over to the spatially averaged mean MTCO anomaly which is lower in the GM reconstruction than in the RM one, but higher in the KM reconstruction than the RM one. The spatially averaged CI sizes as well as the spatial uncertainty pattern are in very good agreement between the three reconstructions. The degree of spatial smoothing of the posteriors, measured by the spatial homogeneity H (see Sect. 3.4.1), is very similar between the three reconstructions.

Likewise, the three reconstructions with glasso covariance matrices are very similar, but differ in some regards from the shrinkage-based reconstructions. The MTWA reconstruction is cooler in all three models, while the MTCO reconstruction is warmer in all models. This is an effect

Table C.1: Summary measures for the joint MTWA and MTCO reconstructions with the five alternative process stage models. Numbers in brackets are minima and maxima of the corresponding 90% CIs. Provided are the spatially averaged posterior mean anomaly from the CRU reference climatology for MTWA and MTCO, the spatially averaged 90% CI sizes for MTWA and MTCO, the spatial homogeneity for MTWA and MTCO, and the PMIP3 ensemble member with the highest posterior weight, where applicable.

Reconstruction name	Mean anomaly MTWA	Mean anomaly MTCO	Mean 90% CI size MTWA	Mean 90% CI size MTCO	Spatial homogeneity MTWA	Spatial homogeneity MTCO	PMIP3 model with highest weight
GM shrinkage	(0.04 K) 0.53 K (1.02 K)	(-0.05 K) 0.55 K (1.14 K)	4.10 K	5.64 K	(1.87 K^2) 2.32 K^2 (2.73 K^2)	(6.29 K^2) 8.15 K^2 (9.99 K^2)	—
KM shrinkage	(0.02 K) 0.48 K (0.97 K)	(0.37 K) 0.92 K (1.51 K)	4.08 K	5.75 K	(1.92 K^2) 2.31 K^2 (2.68 K^2)	(6.14 K^2) 7.79 K^2 (9.42 K^2)	MPI-ESM-P
GM glasso	(-0.29 K) 0.05 K (0.36 K)	(0.66 K) 1.01 K (1.38 K)	2.89 K	3.15 K	(2.19 K^2) 2.47 K^2 (2.73 K^2)	(5.17 K^2) 5.78 K^2 (6.35 K^2)	—
RM glasso	(-0.31 K) 0.03 K (0.34 K)	(0.63 K) 1.02 K (1.37 K)	2.91 K	3.17 K	(2.22 K^2) 2.49 K^2 (2.77 K^2)	(5.18 K^2) 5.77 K^2 (6.37 K^2)	MPI-ESM-P
KM glasso	(-0.32 K) 0.01 K (0.33 K)	(0.74 K) 1.11 K (1.50 K)	2.90 K	3.18 K	(2.27 K^2) 2.57 K^2 (2.88 K^2)	(5.20 K^2) 5.77 K^2 (6.36 K^2)	MPI-ESM-P

of the stronger interaction between MTWA and MTCO in the glasso covariance matrices, as the shrinkage target Φ does not contain correlations between MTWA and MTCO. The spatial mean patterns tend towards more extreme values in the glasso than in the shrinkage-based reconstructions. This can be explained by the fewer spatial modes in the glasso covariances, that lead to less robust reconstructions. The uncertainty estimates are significantly smaller than in the reconstructions with shrinkage covariances, due to the stronger spatial correlations in the covariance matrix. This fits with the diagnosed under-dispersiveness of the glasso models in ITEs (see Sect. 3.3.1). The difference in the uncertainty estimates is larger for MTCO than MTWA, because the stronger inter-variable interaction in the glasso covariance matrices leads to a stronger uncertainty reduction compared to the unconstrained process stages in those three reconstructions. This also leads to a higher degree of spatial smoothing in the glasso-based reconstructions for MTCO, while the estimates for MTWA are very similar to the reconstructions with shrinkage covariances.

All reconstructions with the RM as well as the KM favor the MPI-ESM-P climatology, which has the highest posterior weights in all those reconstructions. In the KM reconstruction, particle filter degeneracy, which is discussed in the manuscript as a major issue of the KM, can be diagnosed as the MPI-ESM-P climatology is chosen in more than 98% of MCMC samples.

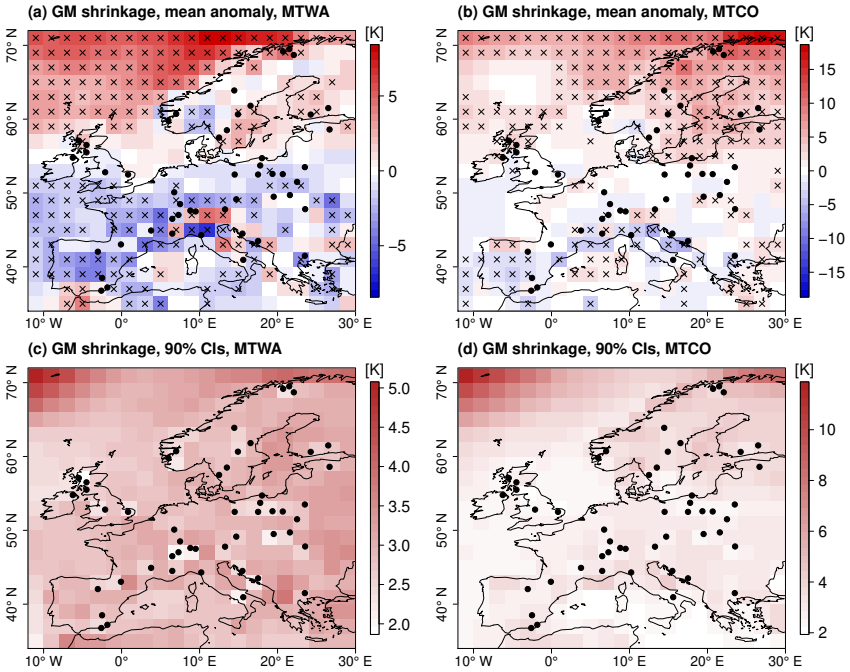

Figure C.1: Spatial reconstruction for MH with the GM using the shrinkage covariance matrix. Top row: Posterior mean anomaly from CRU reference climatology (left: MTWA, right: MTCO), bottom row: reconstruction uncertainty plotted as size of grid box-wise 90% CIs (left: MTWA, right: MTCO). Black dots depict proxy samples. In the top row, point-wise significant anomalies (5% level) are marked by black crosses.

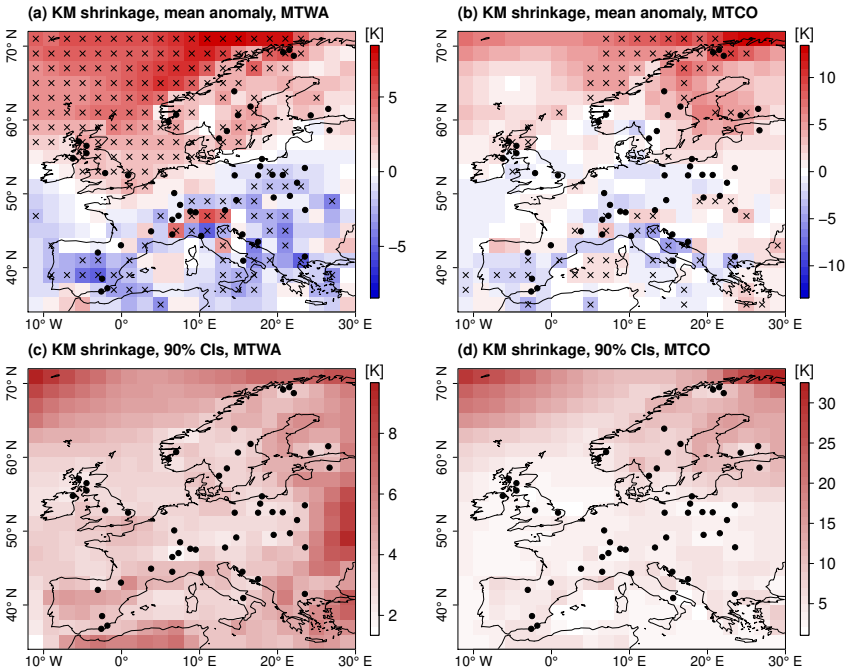

Figure C.2: Same as Fig. C.1, but for KM with shrinkage covariance matrix.

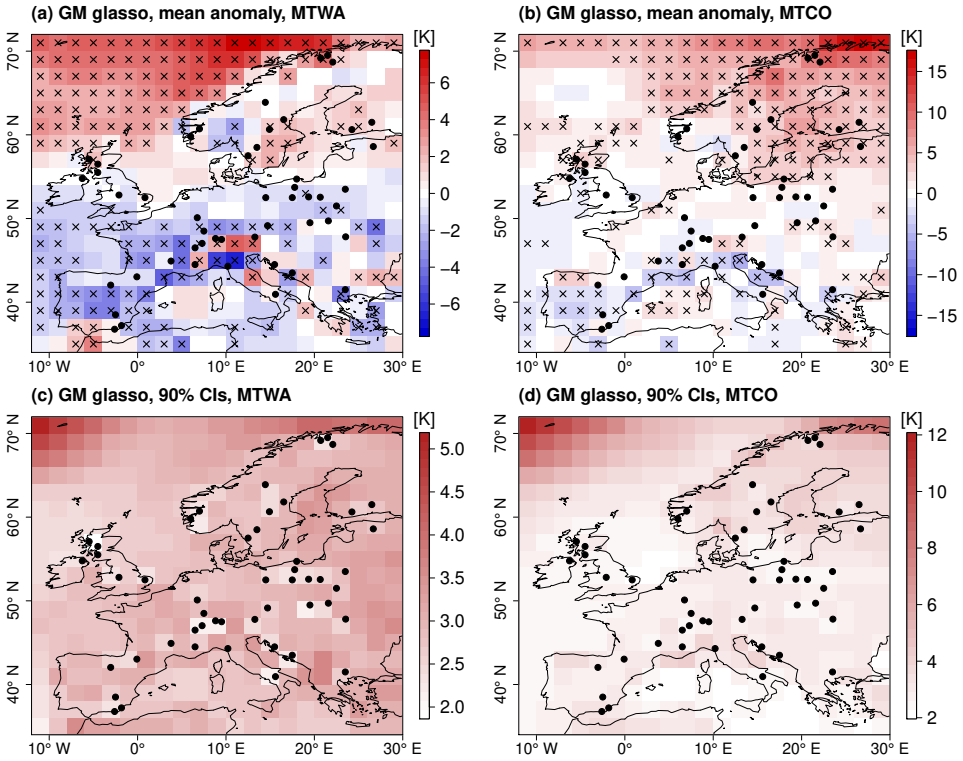

Figure C.3: Same as Fig. C.1, but for GM with glasso covariance matrix.

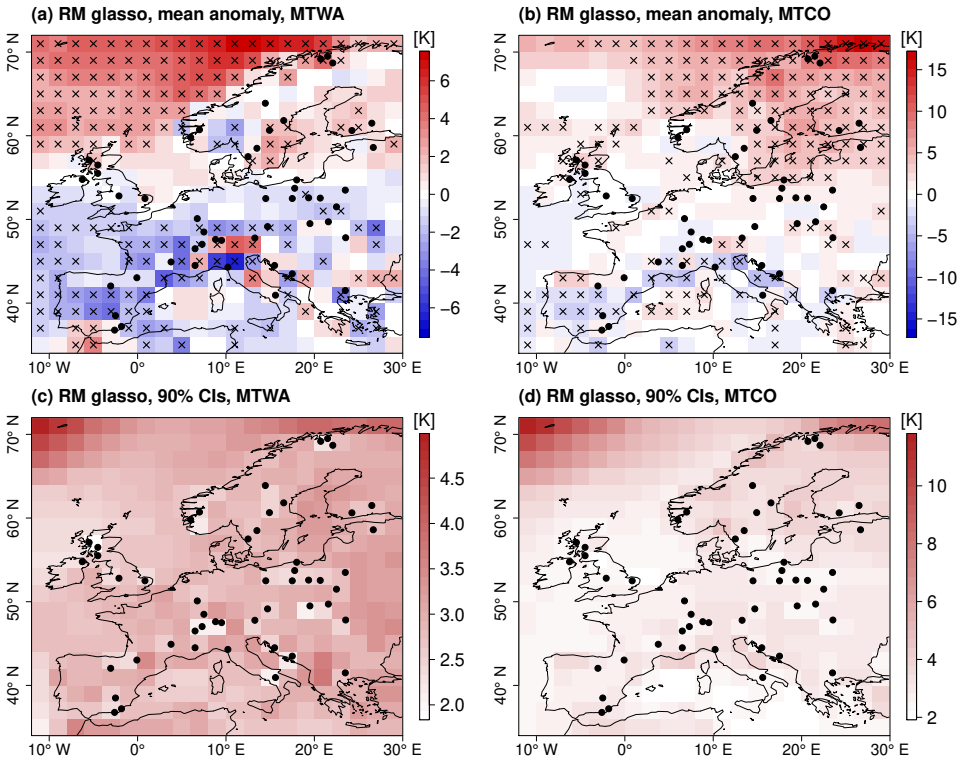

Figure C.4: Same as Fig. C.1, but for RM with glasso covariance matrix.

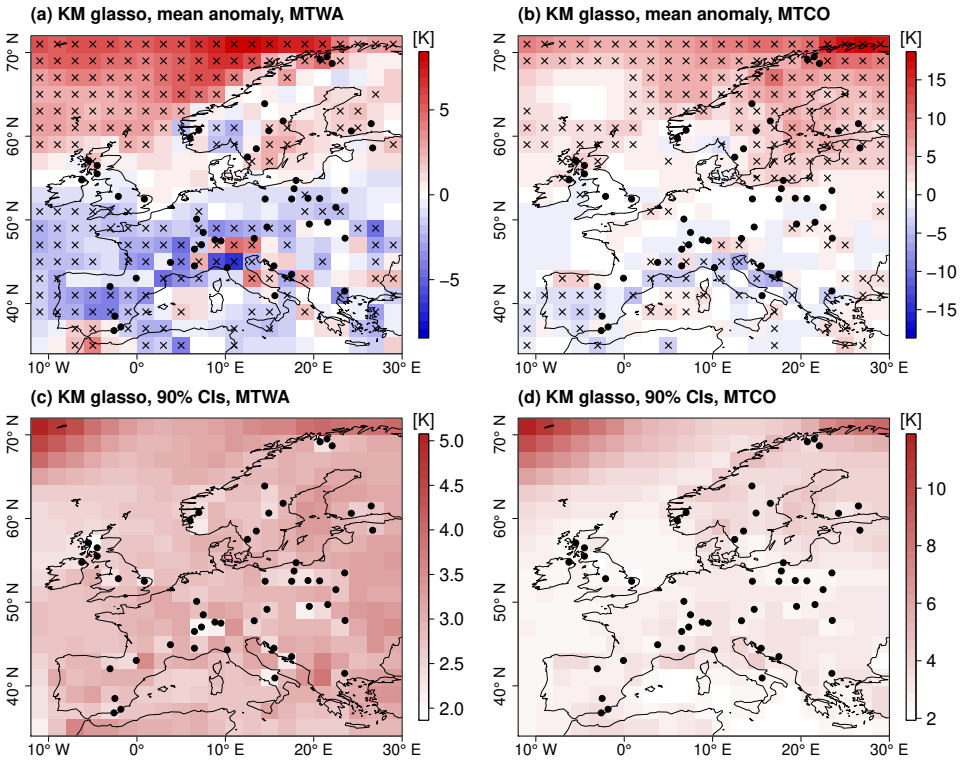

Figure C.5: Same as Fig. C.1, but for KM with glasso covariance matrix.

C.2 Sensitivity with respect to the glasso penalty parameter

As described in Sect. 3.2.2, the glasso algorithm, which regularizes the covariance matrix estimated from the PMIP3 ensemble, requires the specification of a penalty parameter ξ. Values of at least 0.3 produce numerically stable matrices. Larger values lead to smaller correlations, and therefore higher posterior uncertainties due to less spatial transfer of proxy information. To test the influence of ξ on the reconstruction results, we compute reconstructions and cross-validations with the RM for $\xi = 0.3, 0.5, 0.7, 1.0, 2.0$.

The fraction of the penalty term on the total cost given by Eq. (14) increases from 79.5% for $\xi = 0.3$ to 98.5% for $\xi = 2.0$ (Fig. C.7) showing the generally high influence of the penalty term due to the small ensemble size and the increasing importance for larger ξ. While the mean BS is lowest for $\xi = 0.3$ with 0.19 and increases with ξ, the differences are small in total with 0.22 being the highest value for $\xi = 2.0$ (Fig. C.7). The mean value of the posterior climate and the posterior ensemble member weights are insensitive to changes in ξ with differences in the spatial average of the posterior mean of less than 0.2 K (Fig. C.7) as well as no substantial regional differences, and mean λ values which differ less than 0.06 for each of the ESM climatologies. Meanwhile, the posterior uncertainty grows substantially for larger ξ values. While the spatially averaged size of grid box-wise 90% CIs is 3.04 K for $\xi = 0.3$, the value increases to 6.41 K for $\xi = 2.0$ (Fig. C.7).

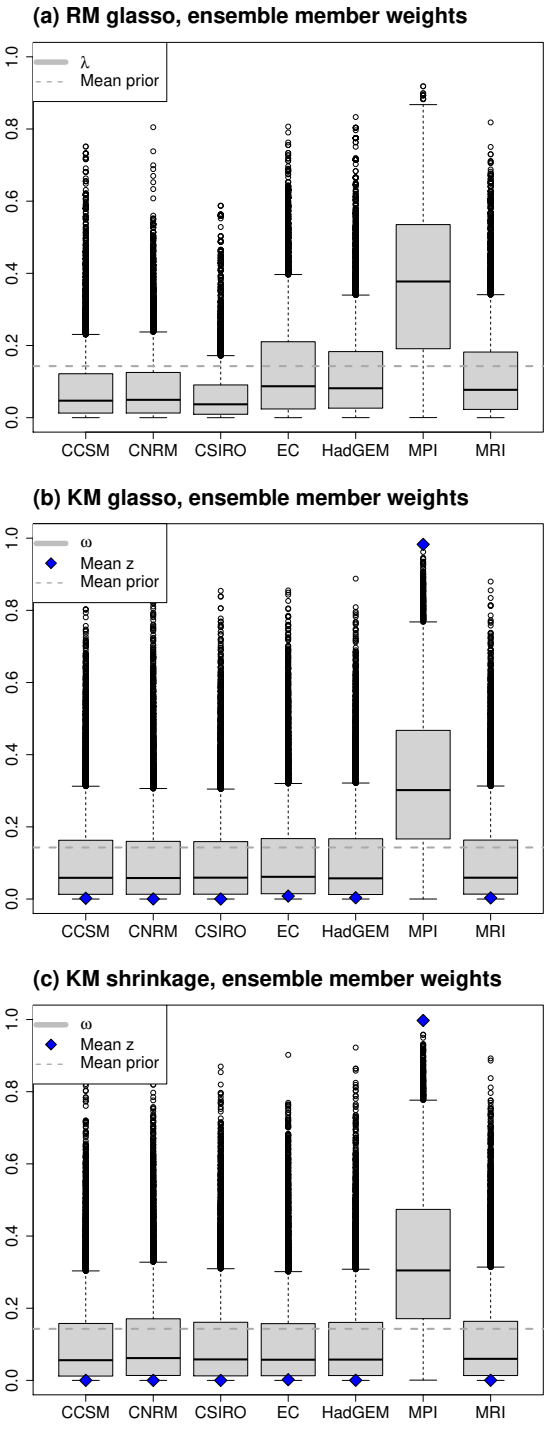

Figure C.6: Posterior ensemble member weights of the reconstructions with RM glasso, KM glasso, and KM shrinkage. The mean of prior weights is denoted by the dashed line.

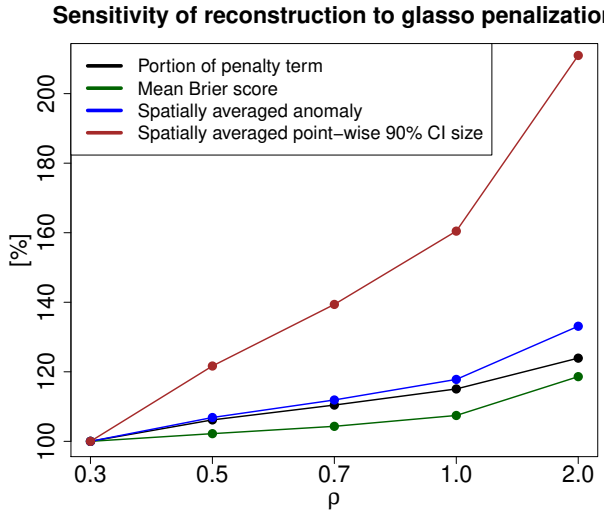

Figure C.7: Sensitivity of reconstructions with the glasso RM to changes in the glasso penalty parameter ξ. Shown are the portion of the penalty term on the total cost in Eq. (14), the mean BS of cross-validations, the spatially averaged posterior mean anomaly from CRU reference climatology, and the spatially averaged size of grid box-wise 90% CIs. All quantities are plotted as percentage of the respective values for $\xi = 0.3$. Note the non-linear scaling of the x-axis.

C.3 Results from experiments with reduced proxy network

In Sect. 3.4.5, results from five reconstructions with a reduced proxy dataset were discussed. In each of those five experiments, half of the proxy samples are removed from the synthesis, and a reconstruction is computed with the remaining records. Here, we provide plots corresponding to each of the reconstructions, numbered by Experiment 1 to Experiment 5. For each experiment, we plot the posterior mean anomaly for MTWA and MTCO, the size of 90% CIs for MTWA and MTCO, and the posterior ensemble member weights λ.

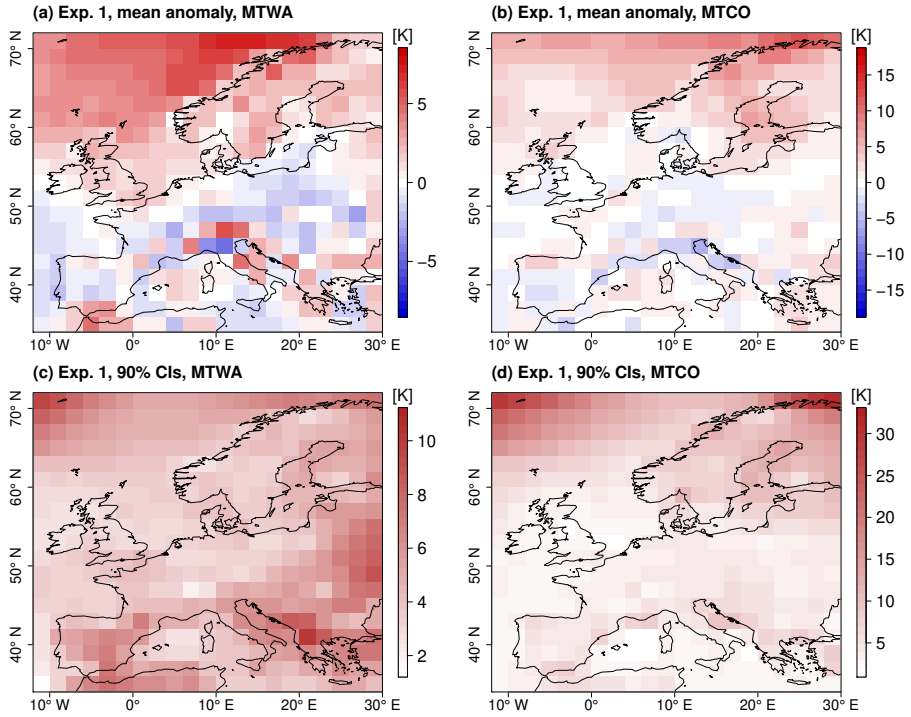

Figure C.8: Reduced data experiment 1. Top row: Posterior mean anomaly from CRU reference climatology (left: MTWA, right: MTCO), bottom row: size of grid box-wise 90% CIs (left: MTWA, right: MTCO).

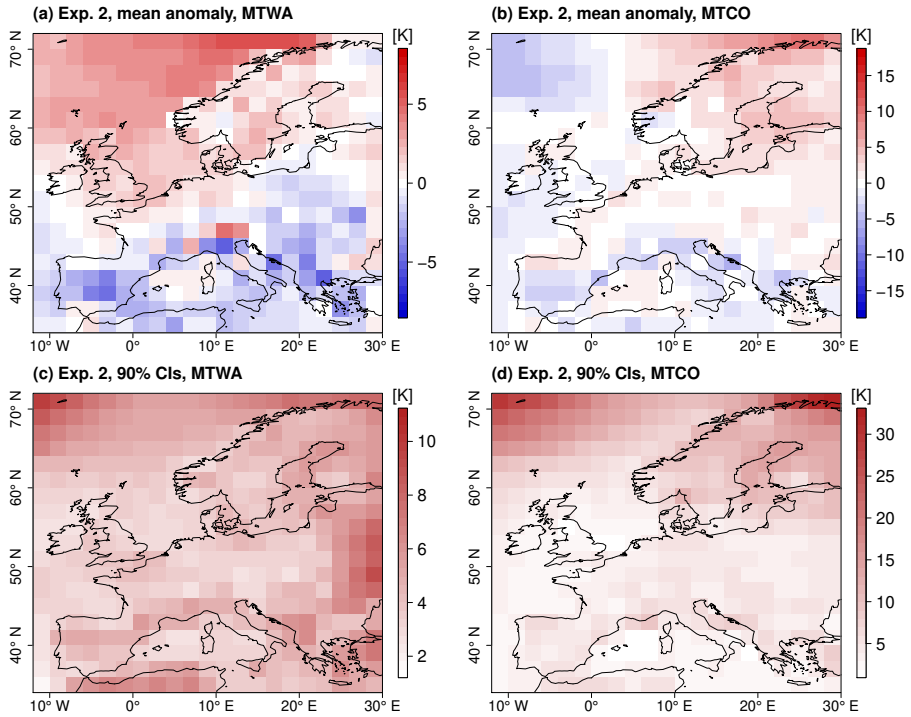

Figure C.9: Same as Fig. C.8, but for experiment 2.

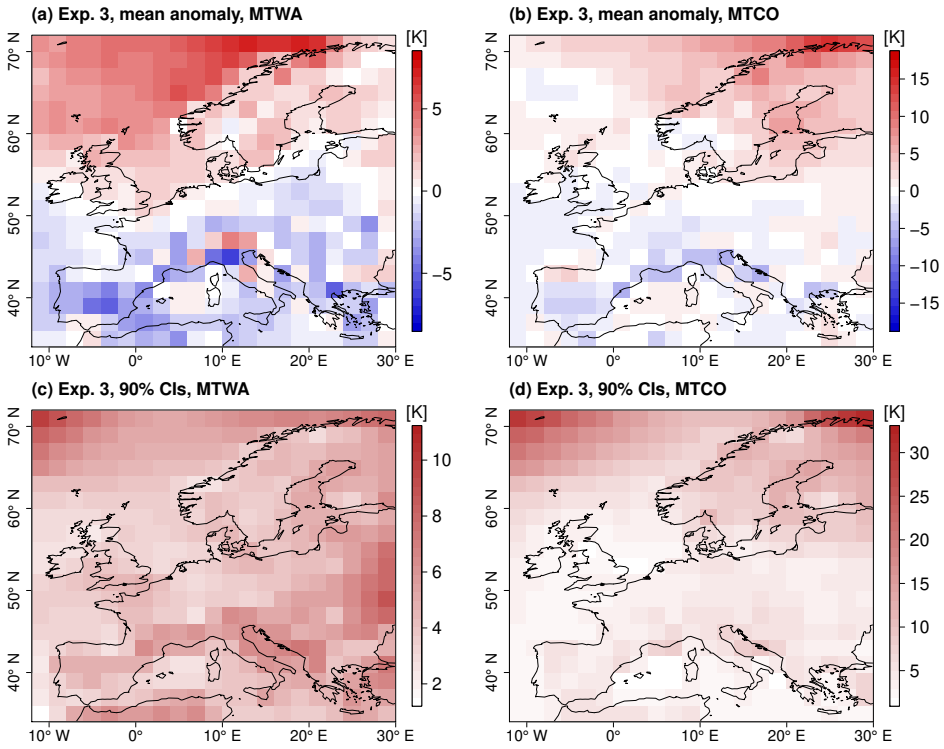

Figure C.10: Same as Fig. C.8, but for experiment 3.

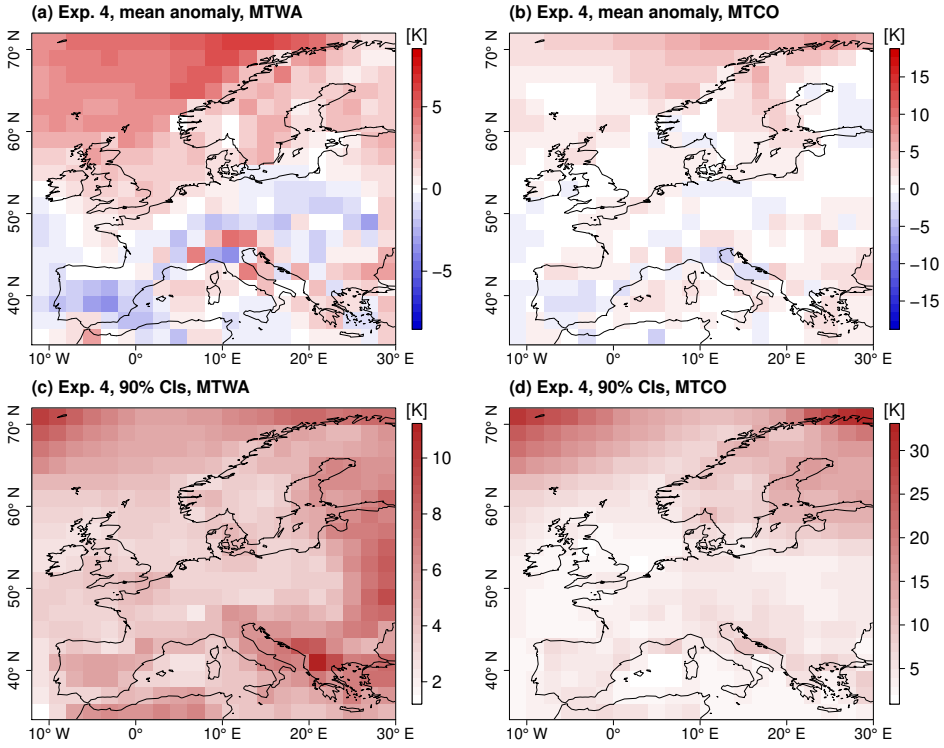

Figure C.11: Same as Fig. C.8, but for experiment 4.

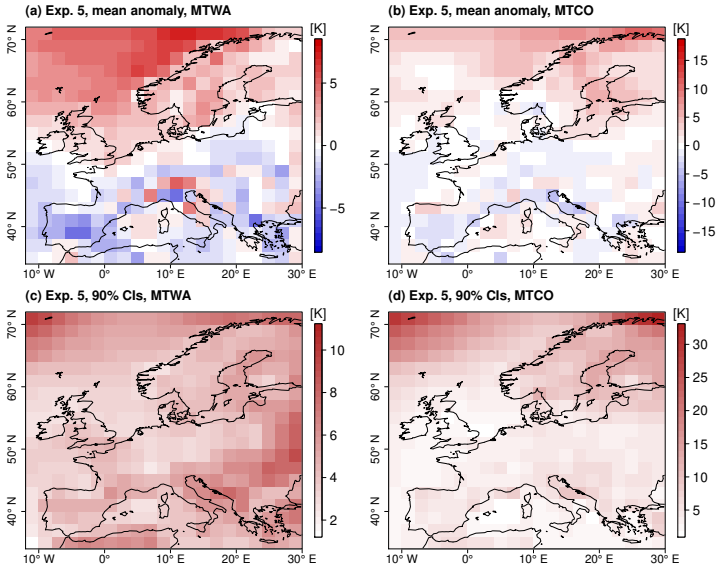

Figure C.12: Same as Fig. C.8, but for experiment 5.

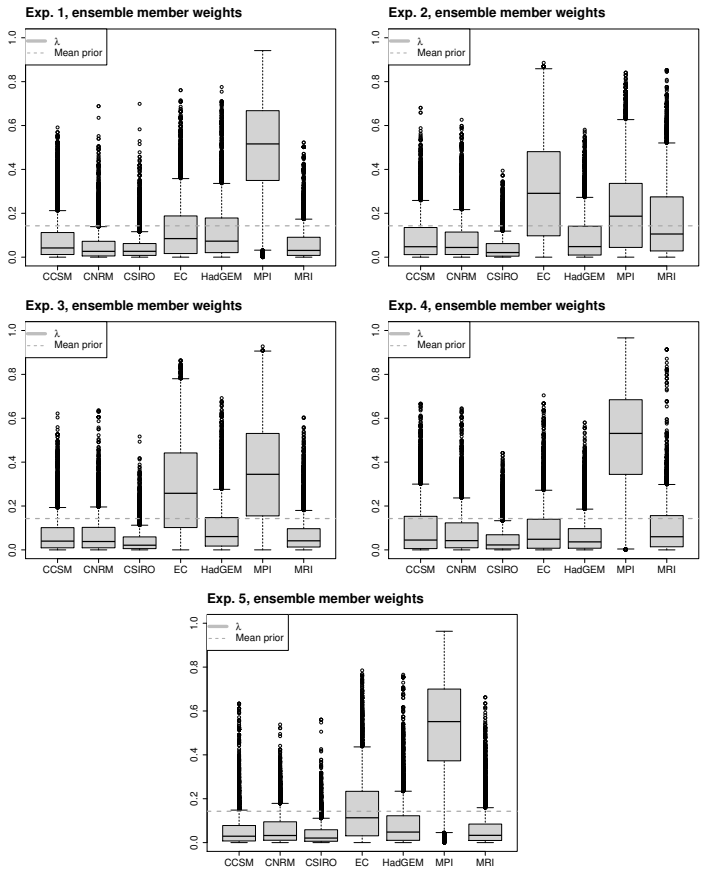

Figure C.13: Posterior ensemble member weights of the five experiments with reduced proxy syntheses. The mean of prior weights is denoted by the dashed line.

D Full conditional distributions in the spatio-temporal BHM

In this appendix, the full conditional distributions of the spatio-temporal BHM are provided. Sampling from the parameters is described in Sect. 4.2.5.
In the spatial reconstruction module, the parameters λ^{LGM} and λ^{MH} are influenced by their priors and by the states $C(1,\cdot)$ and $C(T,\cdot)$. The full conditional distributions are given by

$$\lambda^{\mathrm{LGM}} \,\Big|\, C(1,\cdot) \quad \sim \quad \mathrm{Dirichlet}\left(\tfrac{1}{2},...,\tfrac{1}{2}\right) \mathcal{N}\left(C(1,\cdot)\,\Big|\, \sum_{k=1}^{4} \lambda_k M_k^{\mathrm{LGM}}, \Sigma^{\mathrm{LGM}}\right), \tag{D.1}$$

$$\lambda^{\mathrm{MH}} \,\Big|\, C(T,\cdot) \quad \sim \quad \mathrm{Dirichlet}\left(\tfrac{1}{2},...,\tfrac{1}{2}\right) \mathcal{N}\left(C(T,\cdot)\,\Big|\, \sum_{k=1}^{6} \lambda_k M_k^{\mathrm{MH}}, \Sigma^{\mathrm{MH}}\right). \tag{D.2}$$

These full conditionals do not follow a probability distribution from which we can sample directly. Therefore, a RW-MH update is used for updating λ^{LGM} and λ^{MH}.

The full conditionals of γ^1 and γ^2 are Gaussian due to the Gaussian prior distribution and likelihood of μ. They are given by

$$\gamma^i \,\Big|\, \mu_2^i,...,\mu_T^i,\eta^i \quad \sim \quad \mathcal{N}\left(\gamma^i \,\Big|\, \tilde{\sigma}^2\left(\sum_{t=2}^{T} \frac{\mu_t^i}{\eta^i}\right), \tilde{\sigma}^2\right), \quad i=1,2, \tag{D.3}$$

$$\text{where} \quad \tilde{\sigma}^2 \;=\; \left(\tfrac{1}{16} + \tfrac{T-1}{\eta^i}\right)^{-1}. \tag{D.4}$$

Therefore, we can sample directly from the full conditionals.

The full conditional of η^1 and η^2 follow an inverse Gamma distribution from which we can sample directly. They are determined by the inverse Gamma prior distribution and the Gaussian likelihood of μ. They are given by:

$$\eta^i \,\Big|\, \mu_2^i,...,\mu_T^i,\gamma^i \quad \sim \quad \mathrm{Inv}\Gamma\left(\eta^i \,\Big|\, 4+\tfrac{T-1}{2},\; 2+\tfrac{1}{2}\sum_{t=2}^{T}\left(\mu_t^i-\gamma^i\right)^2\right), \quad i=1,2 \tag{D.5}$$

The full conditional distributions of $(\mu_2^i,...,\mu_T^i)$, $i=1,2$, are determined by the Gaussian prior distribution with parameters γ^i and η^i, and the Gaussian likelihood. They follow a Gaussian dis-

tribution given by

$$\left(\mu_t^1, \mu_t^2\right) \mid C, \sigma_t, \Sigma, \gamma^1, \gamma^2, \eta^1, \eta^2 \quad \sim \quad \mathcal{N}\left(\left(\mu_t^1, \mu_t^2\right) \,\middle|\, \tilde{\Sigma}\,\tilde{\mu}_t,\, \tilde{\Sigma}\right), \quad t = 2, ..., T \tag{D.6}$$

$$\text{where } \tilde{\Sigma} = \left(\text{Diag}\left(\tfrac{1}{\eta^1}, \tfrac{1}{\eta^2}\right) + \tfrac{1}{\sigma_t^2}\, O\, \Sigma^{-1}\, O^t\right)^{-1} \tag{D.7}$$

$$\text{and } \tilde{\mu}_t \;=\; \text{Diag}\left(\tfrac{1}{\eta^1}, \tfrac{1}{\eta^2}\right)\left(\gamma^1, \gamma^2\right) + \tfrac{1}{\sigma_t^2}\, O\, \Sigma^{-1}\, (C(t, \cdot) - C(t-1, \cdot))\,. \tag{D.8}$$

Here, O is a projection matrix that maps the climate field to the two parameters $\left(\mu_t^1, \mu_t^2\right)$ which control the global shifts and depend linearly on latitude. We can sample directly from this Gaussian distribution.

The full conditional distribution of α depends on it prior and the likelihood of σ^2. It is given by

$$\alpha \,\middle|\, \sigma_t, \beta \quad \sim \quad \Gamma\left(\tfrac{1}{2}, \tfrac{1}{2}\right) \prod_{t=2}^{T} \text{InvΓ}\left(\sigma_t^2 \,\middle|\, \alpha, \beta\right). \tag{D.9}$$

No direct sampler is available for this full conditional distribution. Therefore, we use a RW-MH step with Gamma distributed proposals for updating α.

The full conditional distribution of β is Gamma distributed and influenced by the Gamma distributed prior and the inverse Gamma distributed likelihood of σ^2. The full conditional is given by

$$\beta \,\middle|\, \sigma_t, \alpha \quad \sim \quad \Gamma\left(\beta \,\middle|\, \tfrac{1}{2} + (T-2)\,\alpha,\; \tfrac{1}{2} + \sum_{t=2}^{T} \tfrac{1}{\sigma_t^2}\right). \tag{D.10}$$

We can update β by sampling directly from this full conditional distribution.

The full conditional distribution of σ_t^2 is influenced by its prior and the likelihood of C. It is again inverse Gamma distributed with parameters given by

$$\sigma_t^2 \,\middle|\, C, \Sigma, \left(\mu_t^1, \mu_t^2\right), \alpha, \beta \quad \sim \quad \text{InvΓ}\left(\sigma_t^2 \,\middle|\, \alpha + \tfrac{N}{2}, \beta + \tfrac{1}{2}\tilde{\beta}\right), \tag{D.11}$$

$$\text{where } \tilde{\beta} \;=\; \left(C(t, \cdot) - C(t-1, \cdot) - O^t\left(\mu_t^1, \mu_t^2\right)\right)^t \Sigma^{-1} \left(C(t, \cdot) - C(t-1, \cdot) - O^t\left(\mu_t^1, \mu_t^2\right)\right). \tag{D.12}$$

Here, O is the projection operator as defined above and N is the number of grid boxes. We update σ_t^2, $t = 2, ..., T$, sequentially by sampling from their full conditional distributions.

The full conditionals of $\nu = \frac{\pi}{2}\chi$ and ρ are determined by the Gaussian likelihood of C and its prior distributions. They are given by

$$\chi, \rho \,\middle|\, C, \sigma, \mu \quad \sim \quad \text{Beta}(1,1)\,\Gamma(25,1) \prod_{t=2}^{T} \mathcal{N}\left(C(t, \cdot) - C(t-1, \cdot) \,\middle|\, O^t\left(\mu_t^1, \mu_t^2\right),\, \sigma_t^2\, \Sigma\left(\tfrac{\pi}{2}\chi, \rho\right)\right). \tag{D.13}$$

We update ν and ρ jointly using a RW-MH step with Beta distributed proposals for χ and Gamma distributed proposals for ρ.

$C(1,\cdot), ..., C(T,\cdot)$ are updated sequentially using Gibbs sampling from the Gaussian distributed full conditionals for each time step. The full conditional distributions are influenced by the climate evolution parameters, the spatial reconstructions for LGM and MH, and the proxy data P.
For the first time step (LGM), the full conditional of $C(1,\cdot)$ is given by

$$C(1,\cdot) \,\Big|\, C(2,\cdot), P(1,\cdot), \lambda^{\text{LGM}}, M^{\text{LGM}}, \mu_2, \sigma_2, \Sigma, \Sigma^{\text{LGM}} \quad \sim \quad \mathcal{N}\left(C(1,\cdot) \,\middle|\, \tilde{\Sigma}\tilde{\mu},\, \tilde{\Sigma}\right), \tag{D.14}$$

$$\text{where } \tilde{\Sigma} \;=\; \left(\left(\Sigma^{\text{LGM}}\right)^{-1} + \tfrac{1}{\sigma_2^2}\Sigma^{-1} + \Sigma_p^{-1}\right)^{-1}, \tag{D.15}$$

$$\text{and } \tilde{\mu} \;=\; \tfrac{1}{\sigma_2^2}\Sigma^{-1}\left(C(2,\cdot) - O^t\left(\mu_2^1, \mu_2^2\right)\right) + \left(\Sigma^{\text{LGM}}\right)^{-1} \sum_{k=1}^{4} \lambda_k^{\text{LGM}} M_k^{\text{LGM}} + \Sigma_p^{-1} P(1,\cdot). \tag{D.16}$$

Here, $P(1,\cdot)$ is the vector of proxy samples corresponding to time step 1, and Σ_p is a diagonal matrix with entries σ_p^2 that represents the joint proxy uncertainty of $P(1,\cdot)$.
For time steps $t = 2$ to $t = T-1$, the full conditional of $C(t,\cdot)$ is given by

$$C(t,\cdot) \,\Big|\, C(t-1,\cdot), C(t+1,\cdot), P(t,\cdot), \mu_t, \mu_{t+1}, \sigma_t, \sigma_{t+1}, \Sigma \quad \sim \quad \mathcal{N}\left(C(t,\cdot) \,\middle|\, \tilde{\Sigma}\tilde{\mu},\, \tilde{\Sigma}\right), \tag{D.17}$$

$$\text{where } \tilde{\Sigma} \;=\; \left(\left(\tfrac{1}{\sigma_t^2} + \tfrac{1}{\sigma_{t+1}^2}\right)\Sigma^{-1} + \Sigma_p^{-1}\right)^{-1}, \tag{D.18}$$

$$\text{and } \tilde{\mu} \;=\; \tfrac{1}{\sigma_t^2}\Sigma^{-1}\left(C(t-1,\cdot) + O^t\left(\mu_t^1, \mu_t^2\right)\right) + \tfrac{1}{\sigma_{t+1}^2}\Sigma^{-1}\left(C(t+1,\cdot) - O^t\left(\mu_{t+1}^1, \mu_{t+1}^2\right)\right) + \Sigma_p^{-1} P(t,\cdot). \tag{D.19}$$

For the last time step (MH), the full conditional $C(T,\cdot)$ is given by

$$C(T,\cdot) \,\Big|\, C(T-1,\cdot), P(T,\cdot), \lambda^{\text{MH}}, M^{\text{MH}}, \mu_T, \sigma_T, \Sigma, \Sigma^{\text{MH}} \quad \sim \quad \mathcal{N}\left(C(T,\cdot) \,\middle|\, \tilde{\Sigma}\tilde{\mu},\, \tilde{\Sigma}\right), \tag{D.20}$$

$$\text{where } \tilde{\Sigma} \;=\; \left(\left(\Sigma^{\text{MH}}\right)^{-1} + \tfrac{1}{\sigma_T^2}\Sigma^{-1} + \Sigma_p^{-1}\right)^{-1}, \tag{D.21}$$

$$\text{and } \tilde{\mu} \;=\; \tfrac{1}{\sigma_T^2}\Sigma^{-1}\left(C(T-1,\cdot) + O^t\left(\mu_T^1, \mu_T^2\right)\right) + \left(\Sigma^{\text{MH}}\right)^{-1} \sum_{k=1}^{6} \lambda_k^{\text{MH}} M_k^{\text{MH}} + \Sigma_p^{-1} P(T,\cdot). \tag{D.22}$$

E Data and code availability

All analyses are performed in R version 3.3.1 (R Core Team, 2016). The preprocessing of data uses the packages ncdf4 (Pierce, 2017), Matrix (Bates and Maechler, 2017), mvtnorm (Genz et al., 2017), glasso (Friedman et al., 2014), and MASS (Venables and Ripley, 2002). The MCMC algorithms use the packages MCMCpack (Martin et al., 2011), BayesLogit (Polson et al., 2013), mvtnorm, spam (Furrer and Sain, 2010), and parallel (R Core Team, 2016) for sampling. For post-processing and plotting, the following packages are used: fields (Nychka et al., 2017), maptools (Bivand and Lewin-Koh, 2017), maps (Becker et al., 2017), coda (Plummer et al., 2006).
R code for computing the spatial reconstructions of European MH climate is provided in a Bitbucket repository available under https://bitbucket.org/nils_weitzel/spatial_reconstr_repo. Code for the Siberian LGM climate reconstructions and for running the spatio-temporal BHM are available on request from the author (email: nils.weitzel@uni-bonn.de).
The PMIP3 simulations are available in the CMIP5 archives. In this study, they were downloaded from the DKRZ long term archive CERA (https://cera-www.dkrz.de). The pollen and macrofossil synthesis for European MH climate reconstructions is published in Simonis (2009). It is available as an R list object in the Bitbucket repository. The modern climate data was downloaded from CRU, available at http://www.cru.uea.ac.uk/data/. The vegetation data for the calibration of PITM was provided by Thomas Litt and Norbert Kühl.
The Siberian LGM reconstructions with WAPLS transfer functions and the modern climatology for calculating anomalies of the local reconstructions were provided by Ulrike Herzschuh, Xianyong Cao, and Thomas Böhmer from the Polar Terrestrial Environmental Systems Division at the Alfred-Wegener-Institute. The locations of the proxy records and corresponding biome reconstructions are given in Tian et al. (2018). The locations of MH and LGM climate reconstructions in Bartlein et al. (2011) are available in the online supplement of Bartlein et al. (2011).
The TraCE-21ka simulations are described at http://www.cgd.ucar.edu/ccr/TraCE/. The simulation output is available on the 'Climate Data Gateway at NCAR' and was downloaded from https://www.earthsystemgrid.org/project/trace.html.

Bibliography

Alfimov, A. V. and Berman, D. I.: Beringian climate during the Late Pleistocene and Holocene, Quaternary Science Reviews, 20, 127–134, 2001.

Allen, J. R. M., Brandt, U., Brauer, A., Hubberten, H.-W., Huntley, B., Keller, J., Kraml, M., Mackensen, A., Mingram, J., Negendank, J. F. W., Nowaczyk, N. R., Oberhänsli, H., Watts, W. A., Wulf, S., and Zolitschka, B.: Rapid environmental changes in southern Europe during the last glacial period, Nature, 400, 740–743, 1999.

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F.: Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference, Bioinformatics, 20/3, 407–415, 2004.

Ammann, C. M. and Wahl, E. R.: The importance of the geophysical context in statistical evaluations of climate reconstruction procedures, Climatic Change, 85, 71–88, 2007.

Anderson, J. L. and Anderson, S. L.: A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts, Monthly Weather Review, 127, 2741–2758, 1999.

Annan, J. and Hargreaves, J.: A new global reconstruction of temperature changes at the Last Glacial Maximum, Climate of the Past, 9, 367–376, 2013.

Apte, A., Hairer, M., Stuart, A., and Voss, J.: Sampling the posterior: An approach to non-Gaussian data assimilation, Physica D, 230, 50–64, 2007.

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

Atsawawaranunt, K., Comas-Bru, L., Amirnezhad Mozhdehi, S., Deininger, M., Harrison, S. P., Baker, A., Boyd, M., Kaushal, N., Ahmad, S. M., Ait Brahim, Y., Arienzo, M., Bajo, P., Braun, K., Burstyn, Y., Chawchai, S., Duan, W., Hatvani, I. G., Hu, J., Kern, Z., Labuhn, I., Lachniet, M., Lechleitner, F. A., Lorrey, A., Pérez-Mejías, C., Pickering, R., Scroxton, N., and Members, S. W. G.: The SISAL database: a global resource to document oxygen and carbon isotope records from speleothems, Earth System Science Data, 10, 1687–1713, 2018.

Banerjee, S., Carlin, B. P., and Gelfand, A. E.: Hierarchical Modeling and Analysis for Spatial Data, vol. 2, Chapman & Hall, CRC Press, New York, 2014.

Baringhaus, L. and Franz, C.: On a new multivariate two-sample test, Journal of multivariate analysis, 88, 190–206, 2004.

Barker, S., Knorr, G., Vautravers, M. J., Diz, P., and Skinner, L. C.: Extreme deepening of the Atlantic overturning circulation during deglaciation, Nature Geoscience, 3, 567–571, 2010.

Bartlein, P., Harrison, S., Brewer, S., Connor, S., Davis, B., Gajewski, K., Guiot, J., Harrison-Prentice, T., Henderson, A., Peyron, O., Prentice, I., Scholze, M., Seppä, H., Shuman, B., Sugita, S., Thompson, R., Viau, A., Williams, J., and Wu, H.: Pollen-based continental climate reconstructions at 6 and 21 ka: a global synthesis, Climate Dynamics, 37, 775–802, 2011.

Bates, D. and Maechler, M.: Matrix: Sparse and Dense Matrix Classes and Methods, URL `https://CRAN.R-project.org/package=Matrix`, r package version 1.2-12, 2017.

Bayes, T.: LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S, Philosophical transactions of the Royal Society of London, 53, 370–418, 1763.

Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P., and Deckmyn, A.: maps: Draw Geographical Maps, URL `https://CRAN.R-project.org/package=maps`, r package version 3.2.0, 2017.

Berger, A. L.: Long-Term Variations of Daily Insolation and Quaternary Climatic Changes, Journal of the Atmospheric Sciences, 35, 2362–2367, 1978.

Betancourt, M.: The Convergence of Markov Chain Monte Carlo Methods: From the Metropolis Method to Hamiltonian Monte Carlo, Annalen der Physik, https://doi.org/10.1002/andp.201700214, 2018.

Binney, H., Edwards, M., Macias-Fauria, M., Lozhkin, A., Anderson, P., Kaplan, J. O., Andreev, A., Bezrukova, E., Blyakharchuk, T., Jankovska, V., Khazina, I., Krivonogov, S., Kremenetski, K., Nield, J., Novenko, E., Ryabogina, N., Solovieva, N., Willis, K., and Zernitskaya, V.: Vegetation of Eurasia from the last glacial maximum to present: Key biogeographic patterns, Quaternary Science Reviews, 157, 80–97, 2017.

Birks, H. H. and Birks, H. J. B.: Future uses of pollen analysis must include plant macrofossils, Journal of Biogeography, 18, 31–35, 2000.

Birks, H. J. B., Heiri, O., Seppä, H., and Bjune, A. E.: Strengths and Weaknesses of Quantitative Climate Reconstructions Based on Late-Quaternary Biological Proxies, The Open Ecology Journal, 3, 68–110, 2010.

Bivand, R. and Lewin-Koh, N.: maptools: Tools for Reading and Handling Spatial Objects, URL `https://CRAN.R-project.org/package=maptools`, r package version 0.9-2, 2017.

Blaauw, M. and Christen, J. A.: Flexible Paleoclimate Age-Depth Models Using an Autoregressive Gamma Process, Bayesian Analysis, 6/4, 457–474, 2011.

Blockley, S., Lane, C., Turney, C., and Bronk Ramsey, C.: The INTegration of Ice core, MArine and TErrestrial records of the last termination (INTIMATE) 60,000 to 8000 BP, Quaternary Science Review, 36, 1, 2012.

Bolin, D. and Wallin, J.: Multivariate Type-G Matern fields, arXiv:1606.08298v2, pp. 1–27, 2018.

Braconnot, P., Harrison, S. P., Otto-Bliesner, B., Abe-Ouchi, A., Jungclaus, J., and Peterschmitt, J.-Y.: The Paleoclimate Modeling Intercomparison Project contribution to CMIP5, CLIVAR Exchanges, 56/16/2, 15–19, 2011.

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

Bradley, R. S.: Paleoclimatology - Reconstructing Climates of the Quaternary, Academic Press, Oxford, 3 edn., 2015.

Brauer, A., Endres, C., Günter, C., Litt, T., Stebich, M., and Negendank, J. F.: High resolution sediment and vegetation responses to Younger Dryas climate change in varved lake sediments from Meerfelder Maar, Germany, Quaternary Science Reviews, 18, 321–329, 1999.

Brier, G.: Verification of Forecasts Expressed in Terms of Probability, Monthly Weather Review, 78, 1–3, 1950.

Bronk Ramsey, C.: Bayesian Analysis of Radiocarbon Dates, Radiocarbon, 51, 337–360, 2009.

Brooks, S. P. and Gelman, A.: General Methods for Monitoring Convergence of Iterative Simulations, Journal of Computational and Graphical Statistics, 7/4, 434–455, 1998.

Cao, X., Ni, J., Herzschuh, U., Wang, Y., and Zhao, Y.: A late Quaternary pollen dataset from eastern continental Asia for vegetation and climate reconstructions: Set up and evaluation, Review of Palaeobotany and Palynology, 194, 21–37, 2013.

Cao, X., Herzschuh, U., Telford, R., and Ni, J.: A modern pollen-climate dataset from China and Mongolia: Assessing its potential for climate reconstructions, Review of Palaeobotany and Palynology, 211, 87–96, 2014.

Carlin, B. P. and Louis, T. A.: Bayesian Methods for Data Analysis, vol. 3, Chapman & Hall, CRC Press, New York, 2008.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data assimilation in the geosciences. An overview on methods, issues and perspectives, WIREs Climate Change, Available at: https://arxiv.org/abs/1709.02798, 2018.

Cheng, H., Sinha, A., Wang, X., Cruz, F. W., and Edwards, R. L.: The Global Paleomonsoon as seen through speleothem records from Asia and the Americas, Climate Dynamics, 39, 1045–1062, 2012.

Christiansen, B. and Ljungqvist, F. C.: Reconstruction of the Extratropical NH Mean Temperature over the Last Millennium with a Method that Preserves Low-Frequency Variability, Journal of Climate, 24, 6013–6034, 2011.

Christiansen, B. and Ljungqvist, F. C.: Challenges and perspectives for large-scale temperature reconstructions of the past two millennia, Reviews of Geophysics, 55, 40–96, 2017.

Christiansen, B., Schmith, T., and Thejll, P.: A Surrogate Ensemble Study of Climate Reconstruction Methods: Stochasticity and Robustness, Journal of Climate, 22, 951–976, 2009.

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, Proceedings of the National Academy of Sciences, 109/19, 1134–1142, 2012.

Cook, E. R., Meko, D. M., Stahle, D. W., and Cleaveland, M. K.: Drought Reconstructions for the Continental United States, Journal of Climate, 12, 1145–1162, 1999.

D'Arrigo, R., Wilson, R., and Jacoby, G.: On the long-term context for late twentieth century warming, Journal of Geophysical Research, 111, 2006.

Dee, S., Steiger, N. J., Hakim, G. J., and Emile-Geay, J.: On the utility of proxy system models for estimating climate states over the common era, Journal of Advances in Modeling Earth Systems, 8, 1164–1179, 2016.

Di Michele, S., Ahlgrimm, M., Forbes, R., Kulie, M., Bennartz, R., Janiskova, M., and Bauer, P.: Interpreting an evaluation of the ECMWF global model with CloudSat observations: ambiguities due to radar reflectivity forward operator uncertainties, Quarterly Journal of the Royal Meteorological Society, 138, 2047–2065, 2012.

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

Evans, M., Tolwinski-Ward, S., Thompson, D., and Anchukaitis, K.: Application of proxy system modeling in high resolution paleoclimatology, Quaternary Science Reviews, 76, 16–28, 2013.

Fanning, A. F. and Weaver, A. J.: An Atmospheric energy-moisture balance model: climatology, interpentadal climate change, and coupling to an ocean general circulation model, Journal of Geophysical Research, 101, 15 111–15 128, 1996.

Friedman, J., Hastie, T., and Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso, Biostatistics, 9/3, 432–441, 2008.

Friedman, J., Hastie, T., and Tibshirani, R.: glasso: Graphical lasso- estimation of Gaussian graphical models, URL `https://CRAN.R-project.org/package=glasso`, r package version 1.8, 2014.

Furrer, R. and Sain, S. R.: spam: A Sparse Matrix R Package with Emphasis on MCMC Methods for Gaussian Markov Random Fields, Journal of Statistical Software, 36, 1–25, 2010.

Gao, C., Robock, A., and Ammann, C.: Volcanic forcing of climate over the past 1500 years: An improved ice core-based index for climate models, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2008JD010239, 2008.

Gebhardt, C.: Varational Reconstruction of Quaternary temperature fields using mixture models as botanical-climatological transfer functions, Dissertation, Meteorologisches Institut der Rheinischen Friedrich-Wilhelms-Universität Bonn, 2003.

Gebhardt, C., Kühl, N., Hense, A., and Litt, T.: Multi-Scale Processes and the Reconstruction of Palaeoclimate, pp. 325–336, In: Neugebauer, Horst J. and Simmer, Clemens, Dynamics of Multiscale Earth Systems, Springer, Berlin, 2003.

Gebhardt, C., Kühl, N., Hense, A., and Litt, T.: Reconstruction of Quaternary temperature fields by dynamically consistent smoothing, Climate Dynamics, 30, 421–437, 2008.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D.: Bayesian data analysis, vol. 3, Chapman & Hall, CRC Press, Boca Raton, 2013.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T.: mvtnorm: Multivariate Normal and t Distributions, URL `http://CRAN.R-project.org/package=mvtnorm`, r package version 1.0-6, 2017.

Geyer, C.: Markov chain Monte Carlo maximum likelihood, pp. 156–163, In: Keramidas (ed.), Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, Interface Foundation, Fairfax Station, 1991.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association, 102/477, 359–378, 2007.

Gomez-Navarro, J. J., Werner, J., Wagner, S., Luterbacher, J., and Zorita, E.: Establishing the skill of climate field reconstruction techniques for precipitation with pseudoproxy experiments, Climate Dynamics, 45, 1395–1413, 2015.

Goosse, H., Evans, M., and Khatiwala, S.: DAPS - Paleoclimate Reanalyses, Data Assimilation and Proxy System modeling, PAGES Magazine, 24, 74, 2016.

Gray, H., Leday, G. G., Vallejos, C. A., and Richardson, S.: Shrinkage estimation of large covariance matrices using multiple shrinkage targets, arXiv:1809.08024v1, pp. 1–32, 2018.

Guillot, D., Rajaratnam, B., and Emile-Geay, J.: Statistical paleoclimate reconstructions via Markov random fields, The Annals of Applied Statistics, 9, 324–352, 2015.

Hakim, G. J., Emile-Geay, J., Steig, E. J., Noone, D., Anderson, D. M., Tardif, R., Steiger, N., and Perkins, W. A.: The last millennium climate reanalysis project: Framework and first results, Journal of Geophysical Research: Atmospheres, 121, 6745–6764, 2016.

Hannart, A. and Naveau, P.: Estimating high dimensional covariance matrices: A new look at the Gaussian conjugate framework, Journal of Multivariate Analysis, 131, 149–162, 2014.

Hannon, G. E. and Gaillard, M.-J.: The plant-macrofossil record of past lake-level changes, Journal of Paleolimnology, 18, 15–28, 1997.

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, Climate of the Past, 9, 811–823, 2013.

Harris, I. and Jones, P.: CRU TS4.01: Climatic Research Unit (CRU) Time-Series (TS) version 4.01 of high-resolution gridded data of month-by-month variation in climate (Jan. 1901- Dec. 2016), https://doi.org/doi:10.5285/58a8802721c94c66ae45c3baa4d814d0, 2017.

Harris, I., Jones, P., Osborn, T., and Lister, D.: Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset, International Journal of Climatology, 34/3, 623–642, 2014.

Harrison, S. P., Bartlein, P. J., Izumi, K., Li, G., Annan, J., Hargreaves, J., Braconnot, P., and Kageyama, M.: Evaluation of CMIP5 palaeo-simulations to improve climate projections, Nature Climate Change, 5, 735–743, 2015.

He, F.: Simulating Transient Climate Evolution of the Last Deglaciation with CCSM3, Dissertation, University of Wisconsin-Madison, 2011.

Hegerl, G. and Zwiers, F.: Use of models in detection and attribution of climate change, WIREs Climate Change, 2, 570–591, 2011.

Hegerl, G. C., Crowley, T. J., Allen, M., Hyde, W. T., and Pollack, H. N.: Detection of human influence on a new, validated 1500-year temperature reconstruction, Journal of Climate, 20, 650–666, 2007.

Holden, P. B., Birks, H. J. B., Brooks, S. J., Bush, M. B., Hwang, G. M., Matthews-Bird, F., Valencia, B. G., and van Woesik, R.: BUMPER v1.0: a Bayesian user-friendly model for palaeoenvironmental reconstruction, Geoscientific Model Development, 10, 483–498, 2017.

Holmström, L., Ilvonen, L., Seppä, H., and Veski, S.: A Bayesian Spatiotemporal Model for Reconstructing Climate from Multiple Pollen Records, The Annals of Applied Statistics, 9/3, 1194–1225, 2015.

Hughes, A. L. C., Gyllencreutz, R., Lohne, O. S., Mangerud, J., and Svendsen, J. I.: The last Eurasian ice sheets - a chronological database and time-slice reconstruction, DATED-1, Boreas, 45/1, 1–45, 2015.

Huntley, B.: Reconstructing palaeoclimates from biological proxies: Some often overlooked sources of uncertainty, Quarternary Science Reviews, 31, 1–16, 2012.

Ivanovic, R. F., Gregoire, L. J., Kageyama, M., Roche, D. M., Valdes, P. J., Burke, A., Drummond, R., Peltier, W. R., and Tarasov, L.: Transient climate simulations of the deglaciation 21-9 thousand years before present (version 1) - PMIP4 Core experiment design and boundary conditions, Geoscientific Model Development, 8, 9045–9102, 2016.

Iversen, J.: Viscum, Hedera and Ilex as climate indicators, Geologiska Foereningens i Stockholm foerhandlingar, 66, 463–483, 1944.

Jeffreys, H.: An invariant form for the prior probability in estimation problems, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 186, 453–461, 1946.

Jolly, D., Prentice, I. C., Bonnefille, R., Ballouche, A., Bengo, M., Brenac, P., Buchet, G., Burney, D., Cazet, J.-P., Cheddadi, R., Edorh, T., Elenga, H., Elmoutaki, S., Guiot, J., Laarif, F., Lamb, H., Lezine, A.-M., Maley, J., Mbenza, M., Peyron, O., Reille, M., Reynaud-Farrera, I., Riollet, G., Ritchie, J. C., Roche, E., Scott, L., Ssemmanda, I., Straka, H., Umer, M., Van Campo, E., Vilimumbalo, S., Vincens, A., and Waller, M.: Biome reconstruction from pollen and plant macrofossil data for Africa and the Arabian peninsula at 0 and 6000 years, Journal of Biogeography, 25, 1007–1027, 1998.

Jones, P., New, M., Parker, D., Martin, S., and Rigor, I.: Surface air temperature and its variations over the last 150 years, Reviews of Geophysics, 37, 173–199, 1999.

Jones, P. D., Briffa, K. R., Barnett, T. P., and Tett, S. F. B.: High-resolution palaeoclimatic records for the last millennium: interpretation, integration and comparison with General Circulation Model control-run temperatures, The Holocene, 8, 455–471, 1998.

Joos, F. and Spahni, R.: Rates of change in natural and anthropogenic radiative forcing over the past 20,000 years, Proceedings of the National Academy of Sciences, 105, 1425–1430, 2008.

Kageyama, M., Albani, S., Braconnot, P., Harrison, S. P., Hopcroft, P. O., Ivanovic, R. F., Lambert, F., Marti, O., Peltier, W. R., Peterschmitt, J.-Y., Roche, D. M., Tarasov, L., Zhang, X., Brady, E. C., Haywood, A. M., LeGrande, A. N., Lunt, D. J., Mahowald, N. M., Mikolajewicz, U., Nisancioglu, K. H., Otto-Bliesner, B. L., Renssen, H., Tomas, R. A., Zhang, Q., Abe-Ouchi, A., Bartlein, P. J., Cao, J., Li, Q., Lohmann, G., Ohgaito, R., Shi, X., Volodin, E., Yoshida, K., Zhang, X., and Zheng, W.: The PMIP4 contribution to CMIP6 – Part 4: Scientific objectives and experimental design of the PMIP4-CMIP6 Last Glacial Maximum experiments and PMIP4 sensitivity experiments, Geoscientific Model Development, 10, 4035–4055, 2017.

Kageyama, M., Braconnot, P., Harrison, S. P., Haywood, A. M., Jungclaus, J., Otto-Bliesner, B. L., Peterschmitt, J.-Y., Abe-Ouchi, A., Albani, S., Bartlein, P. J., Brierley, C., Crucifix, M., Dolan, A., Fernandez-Donado, L., Fischer, H., Hopcroft, P. O., Ivanovic, R. F., Lambert, F., Lunt, D. J., Mahowald, N. M., Peltier, W. R., Phipps, S. J., Roche, D. M., Schmidt, G. A., Tarasov, L., Valdes, P. J., Zhang, Q., and Zhou, T.: The PMIP4 contribution to CMIP6 – Part1: Overview and over-arching analysis plan, Geoscientific Model Development, 11, 1033–1057, 2018.

Kienast, F., Schirrmeister, L., Siegert, C., and Tarasov, P.: Palaeobotanical evidence for warm summers in the East Siberian Arctic during the last cold stage, Quaternary Research, 63, 283–300, 2005.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in combining projections from multiple climate models, Journal of Climate, 23, 2739–2758, 2010.

Köhler, P., Nehrbass-Ahles, C., Schmitt, J., Stocker, T. F., and Fischer, H.: A 156 kyr smoothed history of the atmospheric greenhouse gases CO2, CH4, and N2O and their radiative forcing, Earth System Science Data, 9, 363–387, 2017.

Korhola, A., Vasko, K., Toivonen, H. T., and Olander, H.: Holocene temperature changes in northern Fennoscandia reconstructed from chironomids using Bayesian modelling, Quarternary Science Reviews, 21, 1841–1860, 2002.

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., Gadgil, S., and Surendran, S.: Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble, Science, 285, 1548–1550, 1999.

Kühl, N., Gebhardt, C., Litt, T., and Hense, A.: Probability Density Functions as Botanical-Climatological Transfer Functions for Climate Reconstruction, Quarternary Research, 58, 381–392, 2002.

Kühl, N., Litt, T., Schölzel, C., and Hense, A.: Eemian and Early Weichselian temperature and precipitation variability in northern Germany, Quarternary Science Reviews, 26, 3311–3317, 2007.

Laepple, T. and Huybers, P.: Ocean surface temperature variability: Large model-data differences at decadal and longer periods, Proceedings of the National Academy of Sciences, 111/47, 16 682–16 687, 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, Proceedings of the National Academy of Sciences, 111, 15 296–15 303, 2014.

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

Li, B., Nychka, D. W., and Ammann, C. M.: The Value of Multiproxy Reconstruction of Past Climate, Journal of the American Statistical Association, 105/491, 883–895, 2010.

Lindgren, F., Rue, H., and Lindström, J.: An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach, Journal of the Royal Statistical Society, Statistical Methodology, Series B, 73/4, 423–498, 2011.

Litt, T., Schölzel, C., Kühl, N., and Brauer, A.: Vegetation and climate history in the Westeifel Volcanic Field (Germany) during the past 11 000 years based on annually laminated lacustrine maar sediments, Boreas, 38, 679–690, 2009.

Litt, T., Ohlwein, C., Neumann, F. H., Hense, A., and Stein, M.: Holocene climate variability in the Levant from the Dead Sea pollen record, Quarternary Science Reviews, 49, 95–105, 2012.

Liu, B., Ait-El-Fquih, B., and Hoteit, I.: Efficient Kernel-Based Ensemble Gaussian Mixture Filtering, Monthly Weather Review, 144, 781–800, 2016.

Liu, F., Bayarri, M., and Berger, J.: Modularization in Bayesian analysis, with emphasis on analysis of computer models, Bayesian Analysis, 4, 119–150, 2009a.

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, 2009b.

Liu, Z., Zhu, J., Rosenthal, Y., Zhang, X., Otto-Bliesner, B. L., Timmermann, A., Smith, R. S., Lohmann, G., Zheng, W., and Elison Timm, O.: The Holocene temperature conundrum, Proceedings of the National Academy of Sciences, 111, 3501–3505, 2014.

Ludwig, P., Gómez-Navarro, J. J., Pinto, J. G., Raible, C. C., Wagner, S., and Zorita, E.: Perspectives of regional paleoclimate modeling, Annals of the New York Academy of Sciences, 1436, 54–69, 2019.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., and Langtimm, C. A.: Estimating site occupancy rates when detection probabilities are less than one, Ecology, 83, 2248–2255, 2002.

Mahalanobis, P.: On the generalized distance in statistics, Proceedings of the National Institute of Sciences of India, 12, 49–55, 1936.

Mann, M. E. and Jones, P. D.: Global surface temperatures over the past two millennia, Geophysical Research Letters, 30, 2003.

Mann, M. E., Bradley, R. S., and Hughes, M. K.: Global-scale temperature patterns and climate forcing over the past six centuries, Nature, 392, 779–787, 1998.

Mann, M. E., Rutherford, S., Wahl, E., and Ammann, C.: Robustness of proxy-based climate field reconstruction methods, Journal of Geophysical Research, 112, 2007.

Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S., and Ni, F.: Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia, Proceedings of the National Academy of Sciences, 105, 13 252–13 257, 2008.

Marcott, S. A., Shakun, J. D., Clark, P. U., and Mix, A. C.: A Reconstruction of Regional and Global Temperature for the Past 11,300 Years, Science, 339, 1198–1201, 2013.

MARGO Project Members: Constraints on the magnitude and patterns of ocean cooling at the Last Glacial Maximum, Nature Geoscience, 2, 127–132, 2009.

Martin, A. D., Quinn, K. M., and Park, J. H.: MCMCpack: Markov Chain Monte Carlo in R, Journal of Statistical Software, 42, 2011.

Masson-Delmotte, V., Schulz, M., Abe-Ouchi, A., Beer, J., Ganopolski, A., Gonzalez Rouco, J., Jansen, E., Lambeck, K., Luterbacher, J., Naish, T., Osborn, T., Otto-Bliesner, B., Quinn, T., Ramesh, R., Rojas, M., X., S., and Timmermann, A.: Information from Paleoclimate Archives, In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Matheson, J. and Winkler, R.: Scoring Rules for Continuous Probability Distributions, Management Science, 22, 1087–1096, 1976.

Mauri, A., Davis, B., Collins, P., and Kaplan, J.: The influence of atmospheric circulation on the mid-Holocene climate of Europe: a data-model comparison, Climate of the Past, 10, 1925–1938, 2014.

Mauri, A., Davis, B., Collins, P., and Kaplan, J.: The climate of Europe during the Holocene: a gridded pollen-based reconstruction and its multi-proxy evaluation, Quaternary Science Reviews, 112, 109–127, 2015.

Meyer, V. and Barr, I.: Linking glacier extent and summer temperature in NE Russia - Implications for precipitation during the global Last Glacial Maximum, Palaeogeography, Palaeoclimatology, Palaeoecology, 470, 72–80, 2017.

Meyer, V. D., Hefter, J., Lohmann, G., Max, L., Tiedemann, R., and Mollenhauer, G.: Summer temperature evolution on the Kamchatka Peninsula, Russian Far East, during the past 20 000 years, Climate of the Past, 13, 359–377, 2017.

Mingram, J., Stebich, M., Schettler, G., Hu, Y., Rioual, P., Nowaczyk, N., Dulski, P., You, H., Opitz, S., Liu, Q., and Liu, J.: Millennial-scale East Asian monsoon variability of the last glacial deduced from annually laminated sediments from Lake Sihailongwan, N.E. China, Quaternary Science Reviews, 201, 57–76, 2018.

Moberg, A., Sonechkin, D. M., Holmgren, K., Datsenko, N. M., and Karlén, W.: Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data, Nature, 433, 614–617, 2005.

Moreno, A., Svensson, A., Brooks, S. J., Connor, S., Engels, S., Fletcher, W., Genty, D., Heiri, O., Labuhn, I., Persoiu, A., Peyron, O., Sadori, L., Valero-Garces, B., Wulf, S., Zanchetta, G., and data contributors: A compilation of Western European terrestrial records 60-8 ka BP: towards an understanding of latitudinal climatic gradients, Quaternary Science Reviews, 106, 167–185, 2014.

Münch, T. and Laepple, T.: What climate signal is contained in decadal- to centennial-scale isotope variations from Antarctic ice cores?, Climate of the Past, 14, 2053–2070, 2018.

Nilsen, T., Werner, J. P., Divine, D. V., and Rypdal, M.: Assessing the performance of the BARCAST climate field reconstruction technique for a climate with long-range memory, Climate of the Past, 14, 947–967, 2018.

North Greenland Ice Core Project members: High-resolution record of Northern Hemisphere climate extending into the last interglacial period, Nature, 431, 147–151, 2004.

Nychka, D. W., Furrer, R., Paige, J., and Sain, S.: fields: Tools for spatial data, R package version 9.6, https://doi.org/10.5065/D6W957CT, 2017.

Ohlwein, C. and Wahl, E. R.: Review of probabilistic pollen-climate transfer methods, Quaternary Science Reviews, 31, 17–29, 2012.

Otto-Bliesner, B., Brady, E. C., Clauzet, G., Tomas, R., Levis, S., and Kothavala, Z.: Last Glacial Maximum and Holocene Climate in CCSM3, Journal of Climate, 19, 2526–2544, 2006.

Otto-Bliesner, B. L., Braconnot, P., Harrison, S. P., Lunt, D. J., Abe-Ouchi, A., Albani, S., Bartlein, P. J., Capron, E., Carlson, A. E., Dutton, A., Fischer, H., Goelzer, H., Govin, A., Haywood, A., Joos, F., LeGrande, A. N., Lipscomb, W. H., Lohmann, G., Mahowald, N., Nehrbass-Ahles, C., Pausata, F. S. R., Peterschmitt, J.-Y., Phipps, S. J., Renssen, H., and Zhang, Q.: The PMIP4 contribution to CMIP6 – Part 2: Two interglacials, scientific objective and experimental design for Holocene and Last Interglacial simulations, Geoscientific Model Development, 10, 3979–4003, 2017.

PAGES2k Consortium: A global multiproxy database for temperature reconstructions of the Common Era, Scientific Data, 4, https://doi.org/10.1038/sdata.2017.88, 2017.

PAGES2k Network: Continental-scale temperature variability during the past two millennia, Nature Geoscience, 6, 339–346, 2013.

Parnell, A. C., Buck, C. E., and Doan, T. K.: A review of statistical chronology models for high-resolution, proxy-based Holocene palaeoenvironmental reconstruction, Quarternary Science Reviews, 30, 2948–2960, 2011.

Parnell, A. C., Sweeney, J., Doan, T. K., Salter-Townshend, M., Allen, J. R., Huntley, B., and Haslett, J.: Bayesian inference for palaeoclimate with time uncertainty and stochastic volatility, Journal of the Royal Statistical Society, Applied Statistics, Series C, 64/1, 115–138, 2015.

Parnell, A. C., Haslett, J., Sweeney, J., Doan, T. K., Allen, J. R., and Huntley, B.: Joint Palaeoclimate reconstruction from pollen data via forward models and climate histories, Quarternary Science Reviews, 151, 111–126, 2016.

Peltier, W. R.: Global glacial isostasy and the surface of the ice-age Earth - The ICE-5G(VM 2) model and GRACE, Annual Review of Earth and Planetary Sciences, 32, 111–149, 2004.

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

Pierce, D.: ncdf4: Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files, URL `https://CRAN.R-project.org/package=ncdf4`, r package version 1.16, 2017.

Plummer, M., Best, N., Cowles, K., and Vines, K.: CODA: Convergence Diagnosis and Output Analysis for MCMC, R News, 6, 7–11, URL `https://www.r-project.org/doc/Rnews/Rnews_2006-1.pdf`, 2006.

Polson, N. G., Scott, J. G., and Windle, J.: Bayesian inference for logistic models using Polya-Gamma latent variables, arXiv:1205.0310v3, pp. 1–42, 2013.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`, 2016.

Reeh, N.: Parameterization of Melt Rate and Surface Temperature on the Greenland lce Sheet, Polarforschung, 59, 113–128, 1989.

Rehfeld, K., Trachsel, M., Telford, R., and Laepple, T.: Assessing performance and seasonal bias of pollen-based climate reconstructions in a perfect model world, Climate of the Past, 12, 2255–2270, 2016.

Rehfeld, K., Münch, T., Ho, S. L., and Laepple, T.: Global patterns of declining temperature variability from the Last Glacial Maximum to the Holocene, Nature, 554, 356–359, 2018.

Robert, C. and Casella, G.: Monte Carlo Statistical Methods, vol. 2, Springer, New York, 2004.

Rue, H. and Held, L.: Gaussian Markov random fields : theory and applications, Chapman & Hall / CRC (Taylor & Francis Group), Boca Raton, 2005.

Rue, H., Martino, S., and Chopin, N.: Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations, Journal of the Royal Statistical Society, Statistical Methodology, Series B, 71, 319–392, 2009.

Rypdal, K., Rypdal, M., and Fredriksen, H.-B.: Spatiotemporal Long-Range Persistence in Earth's Temperature Field: Analysis of Stochastic-Diffusive Energy Balance Models, Journal of Climate, 28, 8379–8395, 2015.

Salter-Townshend, M. and Haslett, J.: Fast inversion of a flexible regression model for multivariate pollen counts data, Environmetrics, 23, 595–605, 2012.

Sánchez Goñi, M. F., Desprat, S., Daniau, A.-L., Bassinot, F. C., Polanco-Martínez, J. M., Harrison, S. P., Allen, J. R. M., Anderson, R. S., Behling, H., Bonnefille, R., Burjachs, F., Carrión, J. S., Cheddadi, R., Clark, J. S., Combourieu-Nebout, N., Mustaphi, Debusk, G. H., Dupont, L. M., Finch, J. M., Fletcher, W. J., Giardini, M., González, C., Gosling, W. D., Grigg, L. D., Grimm, E. C., Hayashi, R., Helmens, K., Heusser, L. E., Hill, T., Hope, G., Huntley, B., Igarashi, Y., Irino, T., Jacobs, B., Jiménez-Moreno, G., Kawai, S., Kershaw, A. P., Kumon, F., Lawson, I. T., Ledru, M.-P., Lézine, A.-M., Liew, P. M., Magri, D., Marchant, R., Margari, V., Mayle, F. E., McKenzie, G. M., Moss, P., Müller, S., Müller, U. C., Naughton, F., Newnham, R. M., Oba, T., Pérez-Obiol, R., Pini, R., Ravazzi, C., Roucoux, K. H., Rucina, S. M., Scott, L., Takahara, H., Tzedakis, P. C., Urrego, D. H., van Geel, B., Valencia, B. G., Vandergoes, M. J., Vincens, A., Whitlock, C. L., Willard, D. A., and Yamamoto, M.: The ACER pollen and

charcoal database: a global resource to document vegetation and fire response to abrupt climate changes during the last glacial period, Earth System Science Data, 9, 679–695, 2017.

Schmittner, A., Urban, N. M., Shakun, J. D., Mahowald, N. M., Clark, P. U., Bartlein, P. J., Mix, A. C., and Rosell-Melé, A.: Climate Sensitivity Estimated from Temperature Reconstructions of the Last Glacial Maximum, Science, 334, 1385–1388, 2011.

Schölzel, C.: Palaeoenvironmental Transfer Functions in a Bayesian Framework with Application to Holocene Climate Variability in the Near East, Dissertation, Meteorologisches Institut der Rheinischen Friedrich-Wilhelms-Universität Bonn, 2005.

Schölzel, C., Hense, A., Hübl, P., Kühl, N., and Litt, T.: Digitization and geo-referencing of botanical distribution maps, Journal of Biogeography, 29, 851–856, 2002.

Shakun, J. D., Clark, P. U., He, F., Marcott, S. A., Mix, A. C., Liu, Z., Otto-Bliesner, B., Schmittner, A., and Bard, E.: Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation, Nature, 484, 49–55, 2012.

Silverman, B.: Density Estimation for Statistics and Data Analysis, vol. 26 of *Monographs on Statistics and Applied Probability*, Chapman & Hall / CRC, Boca Raton, 1986.

Simonis, D.: Reconstruction of possible realizations of the Late Glacial and Holocene near surface climate in Central Europe, Dissertation, Meteorologisches Institut der Rheinischen Friedrich-Wilhelms-Universität Bonn, 2009.

Simonis, D., Hense, A., and Litt, T.: Reconstruction of late Glacial and Early Holocene near surface temperature anomalies in Europe and their statistical interpretation, Quaternary International, 274, 233–250, 2012.

Stauch, G. and Gualtieri, L.: Late Quaternary glaciations in northeastern Russia, Journal of Quaternary Science, 23, 545–558, 2008.

Steiger, N. J. and Smerdon, J. E.: A pseudoproxy assessment of data assimilation for reconstructing the atmosphere-ocean dynamics of hydroclimate extremes, Climate of the Past, 13, 1435–1449, 2017.

Steiger, N. J., Hakim, G. J., Steig, E. J., Battisti, D. S., and Roe, G. H.: Assimilation of Time-Averaged Pseudoproxies for Climate Reconstruction, Journal of Climate, 27, 426–441, 2014.

Stolzenberger, S.: Untersuchungen zu botanischen Paläoklimatransferfunktionen, Diploma thesis, Meteorologisches Institut der Rheinischen Friedrich-Wilhelms-Universität Bonn, 2011.

Stolzenberger, S.: On the probabilistic evaluation of decadal and paleoclimate model predictions, Dissertation, Meteorologisches Institut der Rheinischen Friedrich-Wilhelms-Universität Bonn, 2017.

Tawn, N. G. and Roberts, G. O.: Accelerating Parallel Tempering: Quantile Tempering Algorithm (QuanTA), arXiv:1808.10415v1, pp. 1–39, 2018.

Thoma, B. M.: Palaeoclimate reconstruction in the Levant and on the Balkans, Dissertation, Meteorologisches Institut der Rheinischen Friedrich-Wilhelms-Universität Bonn, 2017.

Thuiller, W.: BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change, Global Change Biology, 9, 1353–1362, 2003.

Tian, F., Cao, X., Dallmeyer, A., Lohmannn, G., Zhang, X., Ni, J., Andreev, A., Anderson, P. M., Lozhkin, A. V., Bezrukova, E., Rudaya, N., Xu, Q., and Herzschuh, U.: Biome changes and their inferred climatic drivers in northern and eastern continental Asia at selected times since 40 cal ka BP, Vegetation History and Archaeobotany, 27, 365–379, 2018.

Tingley, M. P. and Huybers, P.: A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part I: Development and Applications to Paleoclimate Reconstruction Problems, Journal of Climate, 23, 2759–2781, 2010a.

Tingley, M. P. and Huybers, P.: A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part II: Comparison with the Regularized Expectation-Maximization Algorithm, Journal of Climate, 23, 2782–2800, 2010b.

Tingley, M. P. and Huybers, P.: Recent temperature extremes at high northern latitudes unprecedented in the past 600 years, Nature, 496, 201–205, 2013.

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 Science Reviews, 35, 1–22, 2012.

Venables, W. N. and Ripley, B. D.: Modern Applied Statistics with S, Springer, New York, 4 edn., 2002.

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, 2004.

Weitzel, N., Wagner, S., Sjolte, J., Klockmann, M., Bothe, O., Andres, H., Tarasov, L., Rehfeld, K., Zorita, E., Widmann, M., Sommer, P., Schädler, G., Ludwig, P., Kapp, F., Jonkers, L., Garcia-Pintado, J., Fuhrmann, F., Dolman, A., Dallmeyer, A., and Brücher, T.: Diving into the Past: A Paleo Data-Model Comparison Workshop on the Late Glacial and Holocene, Bulletin of the American Meteorological Society, 100, 1–4, 2019.

Werner, J. P. and Tingley, M. P.: Technical Note: Probabilistically constraining proxy age-depth models within a Bayesian hierarchical reconstruction model, Climate of the Past, 11, 533–545, 2015.

Werner, J. P., Luterbacher, J., and Smerdon, J. E.: A Pseudoproxy Evaluation of Bayesian Hierarchical Modeling and Canonical Correlation Analysis for Climate Field Reconstructions over Europe, Journal of Climate, 26, 851–867, 2013.

Whittle, P.: On stationary processes in the plane, Biometrika, 41, 434–449, 1954.

Whittle, P.: Stochastic processes in several dimensions, Bulletin of the International Statistical Institute, 40, 974–994, 1963.

Williams, J. W., Grimm, E. C., Blois, J. L., Charles, D. F., Davis, E. B., Goringa, S. J., Graham, R. W., Smith, A. J., Anderson, M., Arroyo-Cabrales, J., Ashworth, A. C., Betancourt, J. L., Bills, B. W., Booth, R. K., Buckland, P. I., Curry, B. B., Giesecke, T., Jackson, S. T., Latorres, C., Nichols, J., Purdum, T., Roth, R. E., Stryker, M., and Takahara, H.: The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource, Quaternary Research, 89, 156–177, 2018.

Windle, J., Polson, N. G., and Scott, J. G.: BayesLogit: Bayesian logistic regression, URL http://cran.r-project.org/web/packages/BayesLogit/index.html, 2013.

Windle, J., Polson, N. G., and Scott, J. G.: Sampling Pólya-Gamma random variates: alternate and approximate techniques, arXiv:1405.0506v1, pp. 1–26, 2014.

Yang, Z. and Zhu, T.: The good, the bad, and the ugly: Bayesian model selection produces spurious posterior probabilities for phylogenetic trees, arXiv:1810.05398v1, pp. 1–6, 2018.

Yeager, S. G., Shields, C. A., Large, W. G., and Hack, J. J.: The Low-Resolution CCSM3, Journal of Climate, 19, 2545–2566, 2006.

Acknowledgments

I thank my supervisor, Andreas Hense, who guided my transition from mathematics to paleoclimatology. I am grateful for his help and support, and the freedom he gave to pursue my research interests. He made this work possible.
In summer 2016, I experienced a five-month research stay at the National Center for Atmospheric Research in Boulder (Colorado). I thank Doug Nychka for providing me with this opportunity, for his support, and for inspiring discussions during my visit.
I thank Ulrike Herzschuh, Xianyong Cao, and Thomas Bömer from the Polar Terrestrial Environmental Systems Division at the Alfred-Wegener-Institute for providing the Siberian climate reconstructions and for discussions which helped to improve the spatial climate reconstructions presented in this study.
I thank Kira Rehfeld for giving me the time and a little bit of necessary pressure to finish my thesis after starting to work in her group.
Christian Ohlwein, Florian Kapp, Sophie Stolzenberger, and Benno Thoma helped me to arrive in paleoclimatology, gave me information on paleodata and reconstruction methods, and provided me programming code and data. I thank them for their support.
I thank Florian, Maike, Moritz, Janica, and Benni for reading earlier versions of the manuscript and providing valuable feedback.
I thank Maike, Volker, and Astrid for always interesting discussions in our office and everything I learned about numerical weather prediction while sharing the office with them. I thank Rüdiger for interesting discussions about the use of statistical methods in climatology. I thank all members of the climate dynamics working group in Bonn and the STACY group in Heidelberg for their feedback and all the encouraging comments.
My work was supported by the German Federal Ministry of Education and Research (BMBF) as Research for Sustainability initiative (FONA, www.fona.de) through the Palmod project (FKZ: 01LP1509D). My research stay in Boulder (Colorado) was additionally supported by NCAR. NCAR is funded by the National Science Foundation. Since January 2019, I am supported by the German Research Foundation (DFG) (code RE3994-2/1). I want to thank BMBF, DFG, and NCAR for their financial support which made this work possible.
Finally, I acknowledge all groups involved in producing and making available the PMIP3 multi-model ensemble. I acknowledge the World Climate Research Programme's Working Group on Coupled Modeling, which is responsible for CMIP.

BONNER METEOROLOGISCHE ABHANDLUNGEN

Herausgegeben vom Institut für Geowissenschaften der Universität Bonn, Abteilung Meteorologie, durch Prof. Dr. H. FLOHN (Hefte 1-25), Prof. Dr. M. HANTEL (Hefte 26-35), Prof. Dr. H.-D. SCHILLING (Hefte 36-39), Prof. Dr. H. KRAUS (Hefte 40-49), ab Heft 50 durch Prof. Dr. A. HENSE.

Heft 1-69: siehe `http://www.meteo.uni-bonn.de/bibliothek/bma`

70-89: open access, verfügbar unter `https://uni-bn.de/kpSDaQffe1`

Heft 70: ***A S M Mostaquimur Rahman***: Influence of subsurface hydrodynamics on the lower atmosphere at the catchment scale, 2015, 98 S. + XVI.

Heft 71: ***Sabrina Wahl***: Uncertainty in mesoscale numerical weather prediction: probabilistic forecasting of precipitation, 2015, 108 S.

Heft 72: ***Markus Übel***: Simulation of mesoscale patterns and diurnal variations of atmospheric CO_2 mixing ratios with the model system TerrSysMP-CO_2, 2015, [erschienen] 2016, 158 S. + II

Heft 73: ***Christian Bernardus Maria Weijenborg***: Characteristics of Potential Vorticity anomalies associated with mesoscale extremes in the extratropical troposphere, 2015, [erschienen] 2016, 151 S. + XI

Heft 74: ***Muhammad Kaleem***: A sensitivity study of decadal climate prediction to aerosol variability using ECHAM6-HAM (GCM), 2016, 98 S. + XII

Heft 75: ***Theresa Bick***: 3D Radar reflectivity assimilation with an ensemble Kalman filter on the convective scale, 2016, [erschienen] 2017, 96 S. + IX

Heft 76: ***Zied Ben Bouallegue***: Verification and post-processing of ensemble weather forecasts for renewable energy applications, 2017, 119 S.

Heft 77: ***Julia Lutz***: Improvements and application of the STatistical Analogue Resampling Scheme STARS, 2016, [erschienen] 2017, 103 S.

Heft 78: ***Benno Michael Thoma***: Palaeoclimate Reconstruction in the Levant and on the Balkans, 2016, [erschienen] 2017, XVI, 266 S.

Heft 79: ***Ieda Pscheidt***: Generating high resolution precipitation conditional on rainfall observations and satellite data, 2017, V, 173 S.

Heft 80: ***Tanja Zerenner***: Atmospheric downscaling using multi-objective genetic programming, 2016, [erschienen] 2017, X, 191 S.

Heft 81: ***Sophie Stolzenberger***: On the probabilistic evaluation of decadal and paleoclimate model predictions, 2017, IV, 122 S.

Heft 82: ***Insa Thiele-Eich***: Flooding in Dhaka, Bangladesh, and the challenge of climate change, 2017, V, 158 S.

Heft 83: ***Liselotte Bach***: Towards a probabilistic regional reanalysis for Europe, 2017 [erschienen] 2018, VI, 114 S.

Heft 84: ***Yen-Sen Lu***: Propagation of land surface model uncertainties in terrestrial system states, 2017, [erschienen] 2018, X, 120 S.

Heft 85: ***Rüdiger Hewer***: Stochastic physical models for wind fields and precipitation extremes, 2018, 99 S.

Heft 86: ***Sebastian Knist***: Land-atmosphere interactions in multiscale regional climate change simulations over Europe, 2018, 147 S.

Heft 87: ***Jessica Keune***: Integrated terrestrial simulations at the continental scale: Impact of groundwater dynamics and human water use on groundwater-to-atmosphere feedbacks during the European heatwave in 2003, 2019, IX, 172 S.

Heft 88: ***Christoph Beekmans***: 3-D Cloud Morphology and Evolution Derived from Hemispheric Stereo Cameras, 2019, [erschienen] 2020, VIII., 118 S.

Heft 89: ***Nils Weitzel***: Climate field reconstructions from pollen and macrofossil syntheses using Bayesian hierarchical models, 2019, [erschienen] 2020, XII., 153 S.

Institut für Geowissenschaften
Abteilung Meteorologie
Mathematisch Naturwissenschaftliche Fakultät
Universität Bonn

institut für geowissenschaften
METEOROLOGIE