BONNER METEOROLOGISCHE ABHANDLUNGEN

Heft 81 (2017) (ISSN 0006-7156)
Herausgeber: Andreas Hense

Sophie Stolzenberger

On the Probabilistic Evaluation of Decadal and Paleoclimate Model Predictions

BONNER METEOROLOGISCHE ABHANDLUNGEN

Heft 81 (2017) (ISSN 0006-7156)
Herausgeber: Andreas Hense

Sophie Stolzenberger

On the Probabilistic Evaluation of Decadal and Paleoclimate Model Predictions

On the Probabilistic Evaluation of Decadal and Paleoclimate Model Predictions

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

vorgelegt von
Sophie Stolzenberger
aus
Deggendorf

Bonn, 2017

Diese Arbeit ist die ungekürzte Fassung einer der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn im Jahr 2017 vorgelegten Dissertation von Sophie Stolzenberger aus Deggendorf.

This paper is the unabridged version of a dissertation thesis submitted by Sophie Stolzenberger born in Deggendorf to the Faculty of Mathematical and Natural Sciences of the Rheinische Friedrich-Wilhelms-Universität Bonn in 2017.

Anschrift des Verfassers:

Address of the author:

Sophie Stolzenberger
Meteorologisches Institut der
Universität Bonn
Auf dem Hügel 20
D-53121 Bonn

1. Gutachter: Prof. Dr. Andreas Hense, Universität Bonn
2. Gutachter: PD Dr. Martin Schultz, Universität Bonn

Tag der Promotion: 12. Juli 2017
Erscheinungsjahr: 2017

Abstract

The development of climate prediction systems for years to decades is an area of current research, as these time scales are important e.g. for the planning horizon of decision-makers. Those prediction systems can be improved by including the knowledge of past climate states. To get a better understanding of climate variations, paleoclimate models simulate climate for certain periods in the past, often key periods such as the Last Glacial Maximum (21,000 years before present), the Mid-Holocene (6,000 years before present), etc. For both decadal and paleoclimate applications, ensembles of climate predictions (a set of predictions instead of the most likely one) are evaluated to quantify the uncertainty of the predictions. The verification of such ensemble climate predictions is an ongoing field in climate research. In this thesis, the quality of decadal and paleoclimate ensemble predictions is assessed by a probabilistic evaluation that comprises different attributes such as reliability/calibration and skill.

Creating decadal climate predictions is challenging and still in an experimental stage due to little experiences (e.g. with the initialization of the model components) compared to weather forecasting. We consider three experiments (b1-LR, pr-GECCO, pr-ORA) of the MiKlip (Mittelfristige Klimaprognosen) decadal prediction system. These experiments differ in the way the atmospheric and oceanic model components are initialized, the number of ensemble members, etc. Each ensemble experiment is validated using one observational dataset, i.e. we assume no observational uncertainties. The three-dimensional evaluation in the atmosphere and in the ocean shows skillful and reliable areas, especially in the subtropics and mid-latitudes. However, in all experiments, we detect deficiencies in the tropical Pacific region at higher altitudes, which may result from falsely generated dynamics in the model physics. For the ocean, we see clear differences between the experiments mostly caused by differences in the initialization data. In the Pacific and the subtropical belt around the equator, pr-GECCO outperforms b1-LR and pr-ORA also in deeper layers of the ocean whereas in the North Atlantic, b1-LR and pr-ORA are more reliable compared to pr-GECCO.

Pollen and macrofossils in sediment cores provide the basis for the local reconstruction of vegetation and, thus, climate for a state in the past. We determine probabilistic information of the observed pollen by estimating botanical climate transfer functions using the generalized linear model. This probabilistic information is used to optimize a multi-model ensemble created from members of PMIP3 (Paleoclimate Modelling Intercomparison Project Phase 3). For the Mid-Holocene, summer temperatures change clearly (up to 0.4 K over land) when assimilating the PMIP3 multi-model ensemble to the observed pollen data. The added value is evidenced by the predominantly positive Brier skill scores (improvement of ca. 20% on average).

Another approach to estimate climate transfer functions is the quadratic discriminant analysis as used in the Bayesian biome model. To apply the Bayesian biome model, the environmental vegetation needs to fulfill similar conditions as in the Dead Sea basin, where three vegetation zones (Mediterranean, Irano-Turanian, and Saharo-Arabian territory) are considered at the transition from arid to sub-humid climate. We apply the Bayesian biome model to a sediment core drilled at the Dead Sea, which encompasses the last ca. 220,000 years. For the Eemian warming phase (approx. 130,000 to 115,000 years before present), we find similar winter temperatures and annual precipitation as for today. For the Last Glacial, the reconstructed values show generally higher precipitation rates and lower winter temperatures compared to today's climate.

Contents

Chapter 1

Introduction

A prominent question in climate science is how reliable and accurate is a prediction for a certain time frame, especially in the future. As we do not know the future evolution of climate, we are not able to assess the quality of a single prediction. However, we can measure the reliability and accuracy of the prediction system itself and find out if we can trust it by analyzing a time period in the past. The quality of a climate prediction system can be improved by including the knowledge of paleoclimatology. Understanding the climate of the past can help us to deal with climate fluctuations e.g. ice ages and warm phases. However, climate does not only change because of natural reasons: as Alley (2016), amongst others, points out, it is also the anthropogenic influence, e.g. the carbon dioxide emissions, which should not be underestimated as the emissions affect the climate system with long-lasting consequences depending on the amount of emissions.

Paleoanthropologists found that climate variability over the last glacial cycles had effects on human evolution. Based on oxygen measurements in skeletons of *foraminifera* (microorganisms with calcareous shells living on the sea floor), oxygen stable isotope curves can be obtained indicating climate information. Comparing these curves with e.g. brain enlargement in human evolution, there is a correlation especially between 800,000-200,000 years before present.[1] During this period, climate strongly fluctuates while the brain size relative to body size increased. With larger and more complex brains, it is possible to increase social skills, deal with abstract problems, survive etc. This is one evolutionary characteristic of the *homo sapiens* compared to ancestors, which made him able to move and resettle.[2] The *homo sapiens* first hunted and gathered, later cultivated crop, herbs, fruits, etc. (which depend on the climate conditions), where the living conditions were most suitable.

One key question in paleoclimatology is how exact can we determine climatic conditions together with the corresponding uncertainty information for certain time slices, e.g. when the *homo sapiens* came to Europe, during the Last Glacial or in the Holocene. Highly resolved spatial and temporal information of climate variables for the Late Quaternary are not available as meteorological records were started to be kept in the 19th century. The knowledge about former climate states can only be obtained by analyzing indirect climate indices, so called proxies. Proxy information can be taken from ice cores, tree rings, speleothems, etc. Pollen and macro fossils are suitable proxy data and represent well the vegetation history as they can be assigned to the species or at least to the genus. Pollen remain on

[1] In the following, 1,000 years before present is equal to 1 ka BP.

[2] http://humanorigins.si.edu/research/climate-and-human-evolution/climate-effects-human-evolution

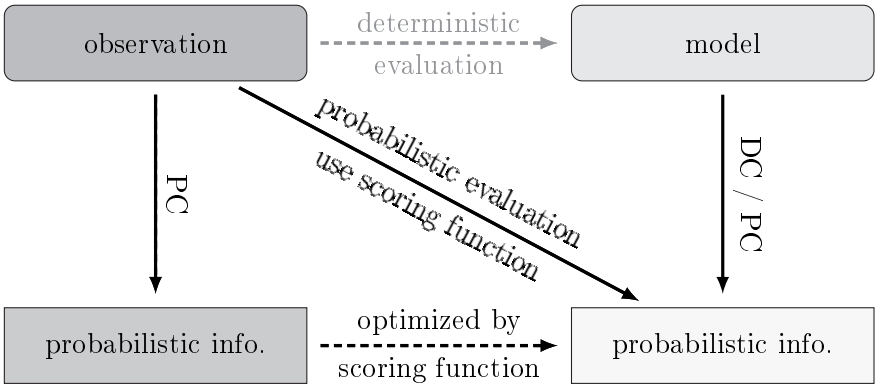

Figure 1.1: *Thesis concept and overview for evaluating model predictions with observations for both considered fields: decadal climate predictions (DC) and paleoclimate predictions (PC). Further information can be found in the text.*

the ground of waterbodies for millenia and can be well preserved in the sediment layers found when drilling e.g. in lakes. Based on these biological proxy data from terrestrial archives, we can receive probabilistic information of climate conditions, under which these plants have existed. In contrast to those observational based reconstructions, paleoclimate model simulations are done by using numerical climate models with adjusted boundary conditions for a certain time step in the past. Finding accurate boundary conditions is ambitious e.g. due to proxy heterogeneities in time and space.

Looking at shorter time scales such as seasonal to decadal scales, there are phenomena induced by climate, such as El Niño Southern Oscillation (ENSO), North Atlantic Oscillation (NAO), etc. Another prominent and ongoing example is the global surface warming hiatus during the first 15 years in the 21st century (Karl et al., 2015; Meehl et al., 2013). However, during this hiatus decade, the net energy imbalance at the top of the atmosphere is up to 1 Wm^{-2}, which is associated with increases of the deep ocean heat content below 750 m. This phenomenon can be connected with processes such as the Pacific decadal oscillation (PDO), which is related to internal decadal variations of the climate system. Creating decadal climate predictions is a challenging task: in the last Assessment Report (AR5) of the Intergovernmental Panel on Climate Change (IPCC), the essay "If you cannot predict the weather next month, how can you predict for the coming decade?"[3] presents the differences and requirements of weather forecasting and seasonal to decadal predicting. Both systems use mathematical equations to describe the atmosphere, need initial conditions to start with, etc. but the time scales are different. The chaotic, non-linear signals of the high-dimensional climate system impose natural limits for single realizations or the deterministic view but (hopefully) not for the probabilistic representation. In weather forecasting, e.g. forecasting the occurrence and development of a single convective rain cell at a specific place and time beyond the next 12 hours is completely impossible. However, forecasting the probability of occurrence of a convective rain cell seems possible under specific circumstances. For decadal climate predictions, the role of external forcing, e.g. the increase of greenhouse gases, is another field of ongoing

[3] https://www.ipcc.ch/pdf/assessment-report/ar5/wg1/WG1AR5_Chapter11_FINAL.pdf, FAQ 11.1, p.964

research, which has to be explicitly studied. Decadal climate predictions find vast interest in the public: potential users can be found in agriculture, economy, politics and society, who plan and decide with a horizon up to ten years.

How can we evaluate these climate models? Probabilistic forecasts, as mentioned above, should be evaluated in a probabilistic way. To evaluate these forecasts, we determine the predictive probability/cumulative density functions (pdfs/cdfs). Compared to probabilistic forecasts, deterministic forecasts are only one-sample forecasts, which do not provide any uncertainty information. For evaluating the pdfs, we use the only available observation, which is assumed to be true. The comparison of model data with observations is called verification. Murphy (1991), Jolliffe and Stephenson (2012) and others point out the importance of comparability and ranking, when two or more forecasting systems exist. A detailed review of proper scoring rules, which form the basis of verifying predictive pdfs/cdfs with the corresponding observation, is given by Gneiting and Raftery (2007). However, including observational uncertainties is important e.g. due to measurement errors or extrapolation discrepancies (Röpnack et al., 2013).

In this study, we will present a comprehensive evaluation of two products, which are (originally) connected to the Coupled Model Intercomparison Project Phase 5 (CMIP5): i) the MiKlip (Mittelfristige Klimaprognosen; mid-term climate forecasts) decadal prediction system and ii) paleoclimate simulations for the Mid-Holocene (6 ka BP), which are part of the Paleo Modelling Intercomparison Project Phase 3 (PMIP3). Figure 1.1 gives an overview of the strategy we follow within this work: for both fields, the decadal climate (DC) and the paleoclimate (PC) field, we work with model ensembles. Comparing observations with the ensemble mean is a deterministic way to evaluate a prediction system. As already mentioned, we can determine probabilistic information by creating pdfs. In DC, the ensembles are based on one model, where single realizations are generated compared to PC, where a multi-model ensemble is created. In DC, we use one observation at one time step / at one grid point and treat it as true, whereas in PC, we determine probabilistic information of the occurring proxies. In case there is one true observation, an ensemble can be verified by using a score function. The result can be the basis for further comparison between different ensembles. In case probabilistic information of model and observations are available, the probabilistic information of the observations can be used for verifying and optimizing the predictions.

The major goals for DC are i) revealing a three-dimensional and process-oriented probabilistic evaluation for atmosphere and ocean and ii) comparing the different baseline experiments within the MiKlip prediction system to each other. For PC, the major aims are i) verifying and optimizing a PMIP3 multi-model ensemble for the Mid-Holocene and ii) providing two different ways to estimate the probabilistic information of proxies.

This work has been carried out within the project MiKlip (Mittelfristige Klimaprognosen) funded by the Federal Ministry for Education and Research (Bundesministerium für Bildung und Forschung) and

started in autumn 2011. The major aim of MiKlip is to create a probabilistic prediction system for climate variability on a decadal time scale. MiKlip rises to four challenges: firstly, the initialization of climate models for decadal predictions including the creation of effective ensembles of prediction runs; secondly, the incorporation of those processes in climate models that are important for realistic representation of decadal climate variability and the understanding of important processes in the numerical prediction system; thirdly, the exploration of predictive skill on the regional scale; fourthly, the systematic evaluation of the decadal prediction system. In the subproject VeCAP (Verification, Calibration and Assessment of Predictability of medium-range climate predictions using satellite data), we have worked on a software tool for evaluating ensemble predictions in general, which is used for the analyses in this work.

The paleoclimatic studies have been done in the framework of the second phase of the Collaborative Research Centre 806 (CRC 806) "Our Way to Europe - Culture-Environment Interaction and Human Mobility in the Late Quaternary" funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) and started in summer 2013. Understanding the history of modern humans and their way from East Africa to Europe by using geoscientific and archaeological methods is the focal point of interest in the CRC 806. Two trajectories are considered: the eastern corridor via the Middle East, Anatolia and the Balkans or the western corridor via North East Africa and the Iberian Peninsula. In the subproject B3 (Environmental Response on Climate Impact in the Levant during the last 200 ka based on a Long Continental Record from the Dead Sea), we focus, amongst others, on the reconstruction of vegetation and climate particularly in the Levant.

This thesis is arranged as follows: in Chapter 2, background information concerning the IPCC structure and the CMIP5 experiments are given including a brief literature review for decadal climate predictions and paleoclimate model simulations. The methods we use to reveal the quality of the predictions are presented in Chapter 3. After this, the thesis is divided in two parts. The first part deals with the probabilistic evaluation of the MiKlip prediction system. The model data and verifying observational data are described in Chapter 4. Chapter 5 shows the results for the three-dimensional evaluation of geopotential height and temperature. Moreover, in the process-oriented evaluation, we examine the MiKlip predictions for certain regions (North Atlantic, Pacific, subtropical regions along the equator). Additionally, we present a concept for the joint probabilistic evaluation of zonal and meridional wind components. The second part of this thesis looks at statistical paleoclimate reconstructions. Chapter 6 presents a new approach for spatial paleoclimate reconstructions including probabilistic information of both model and proxy data. This three-step method is described in detail and applied to the Mid-Holocene (6 ka BP). In Chapter 7, we show another way to determine probabilistic information of proxy data in the Dead Sea region by applying the Bayesian biome model. Moreover, we present climate reconstructions in the Dead Sea regions for the last approx. 150 ka. Finally, the Conclusion is given in Chapter 8. All computations are done with the freely available R statistical language (R Core Team, 2016) and the freely available climate data operators (CDO, 2015).

Chapter 2

Predictions on different time scales

The role of the IPCC and its Assessment Reports are important for the climate community but also for decision-maker far from science, who need basic and actual knowledge for deciding. This chapter includes some important points about the IPCC Assessment Reports and the underlying data basis CMIP5. Furthermore, a brief literature review is given for the field of decadal climate predictions and paleoclimate model simulations.

2.1 Background

2.1.1 IPCC Assessment Reports

The Intergovernmental Panel on Climate Change (IPCC) was established in 1988 by the United Nations Environment Programme (UNEP) and the World Meteorological Organisation (WMO). The main task is to provide the current and comprehensive state of knowledge concerning climate change, which should inform, and a "Summary for Policymakers" (SPM), which gives recommendations to e.g. politics. These so-called Assessment Reports consist of three parts: contribution on physical science basis (by Working Group I), on impacts, adaption and vulnerability (by Working Group II), and on mitigation of climate change (by Working Group III). In the following paragraphs, a brief summary is given based on the findings of Working Group I.

In the first Assessment Report (Houghton et al., 1990), it is clearly stated at the beginning that there is a natural and an anthropogenic global climate change. Further definitions and concepts are specified e.g. the greenhouse effect and the most important gases, which lead to the increasing greenhouse gas concentration. Furthermore, the climate system itself is described, the natural factors that determine climate, and direct and indirect aerosol effects, etc.

Dealing with modeled climate evolution, we have to distinguish between the terms prediction and projection (Stocker et al., 2013):

Climate prediction "A climate prediction or climate forecast is the result of an attempt to produce (...) an estimate of the actual evolution of the climate in the future (...). Because the future evolution of the climate system may be highly sensitive to initial conditions, such predictions are usually probabilistic in nature."

Projection/Climate projection "A projection is a potential future evolution of a quantity or a set of quantities, often computed with the aid of a model. Climate projections are distinguished from climate predictions by their dependence on the emission/concentration/radiative forcing scenario used, which is in turn based on assumptions concerning, for example, future socioeconomic and technological developments that may or may not be realized."

In the fifth Assessment Report (AR5), former scenarios are replaced by the so-called RCPs (Representative Concentration Pathways), which are characterized by the stabilization or peak of radiative forcing at the end of the 21st century: the lowest case RCP2.6 will increase the radiative forcing during the 21st century up to 3 W/m^2 and decrease it to 2.6 W/m^2 by 2100; the midrange scenarios RCP4.5 and RCP6 aim to stabilize the radiative forcing to 4.5 W/m^2 and 6 W/m^2; the high-emission scenario RCP8.5 implies a rising of 8.5 W/m^2 until 2100 with a potential increase of radiative forcing after that. The RCPs, however, cannot fully represent the range of emissions, e.g. looking at aerosols.

A prominent topic within all Assessment Reports is the discussion about sea level changes. Questions such as how did the global sea level change during the past and what can we expect for the future are addressed. Strongly connected to this topic is paleoclimate modeling, which has been firstly presented in an own chapter, together with proxy data as observational basis, in the fourth Assessment Report (Solomon et al., 2007). In the SPM of AR5, it is stated that "the rate of sea level rise since the mid-19th century has been larger than the mean rate during the previous two millenia (high confidence). Over the period 1901 to 2010, global mean sea level rose by 0.19 m (...)". The projections for global mean sea level change are between 0.3 and 0.6 m for RCP2.6 and between 0.55 and 1 m for RCP8.5 by the end of the 21st century. Decadal ("Near-term") predictions and projections have been firstly analyzed in AR5. Looking at decadal climate predictions, there are areas where positive skill is exhibited, mainly discussed for surface temperature and precipitation.

2.1.2 CMIP5 framework

Data basis for the analyses presented in AR5 are the CMIP5 (Coupled Model Intercomparison Project Phase 5) experiments. CMIP5 is a model comparison platform, which was arranged by the Working Group on Coupled Modelling (WGCM) of the World Climate Research Programme (WCRP) together with the International Geosphere-Biosphere Programme (IGBP) and Analysis, Integration and Modeling of the Earth System (AIMES) project in 2008. Taylor et al. (2012) give a comprehensive overview of CMIP5 and list the innovations compared to earlier CMIP experiments, which will be briefly summarized here.

The main goals of CMIP5 are to provide i) a quality check for the models, ii) projections until the end of the 21 century and iii) an understanding how differences in the multi model output can clarify uncertainties such as e.g. coupling effects due to clouds and the carbon cycle (Taylor et al., 2009). As there are pre-defined standards for the experiments, the participating model output can be compared

at its best. In CMIP5, two kinds of experiments are executed: i) near-term experiments (also called decadal predictions) for time scales on 10 to 30 years and ii) long-term experiments for a century or longer time intervals. Both experiments are generated by using atmosphere-ocean global climate models (AOGCMs). For the long-term experiments, some AOGCMs are coupled with a carbon cycle model, which are then called Earth system models (ESMs), trying to ensure a closed carbon cycle. The use of Earth system models of intermediate complexity (EMICs) is restricted to long-term simulations only.

Both experimental types consist of a "core", which includes the basic simulations, and one or two "tiers" around the core, which concentrate on more specific simulations. The core simulations of the near-term experiment comprise a 10 years 3 member ensemble initialized every five years from 1960 using observed forcing factors until 2005 and the RCP4.5 scenario afterwards. In the second core experiment, the time frame is extended to a 30 years 3 member ensemble initialized in 1960, 1980 and 2005. The longer timescale takes into account the importance of the external forcing from the increasing greenhouse gases. The tier 1 experiment includes the increasing of the number of ensemble members to at least 10, testing alternative initialization methods e.g. initializing every year, leaving out the volcanoes, etc. For the long-term simulation, the core experiments contain a coupled control run, AMIP (Atmospheric Model Intercomparison Project) runs and at least one 20th century experiment with all forcings. As projection, two scenarios - RCP8.5 and RCP4.5 - are proposed. For ESMs, control runs and 20th century simulations, the high scenario is used as core experiment. Tier 1 considers e.g. the other two scenarios (RCP 2.6 and RCP6) or the temporal extension of RCP4.5 to the end of the 23th century. Paleoclimate simulations for the Mid-Holocene (6 ka BP) and Last Glacial Maximum (LGM, 21 ka BP) are also included in tier 1. In tier 2, one experiment explores simulations over the last 1,000 years (850-1850). These paleoclimate experiments are also part of PMIP3 (Paleoclimate Modelling Intercomparison Project Phase 3, see Section 2.2.2 for further information). The diagnostic core and proceeding experiments are not further described here.

One advantage of CMIP5 compared to earlier CMIP releases is the number of participating institutions and models. In CMIP3, 17 institutions contributed with 24 models (Meehl et al., 2007), whereas in CMIP5 more than 20 institutions contributed with more than 50 models. Another advantage is the higher spatial resolution of the models, which ranges from 0.5 to 4° for the atmospheric and from 0.2 to 2° for the oceanic model component. Some of the 35 experiments, which are explained in detail in Taylor et al. (2009), are even generated with a higher resolution of the atmospheric model component. Furthermore, the model documentation and the description for the single experiment conditions are improved.

2.2 Related work

2.2.1 Decadal climate predictions

The IPCC has firstly mentioned decadal climate predictions in a particular chapter in its last Assessment Report ("Near-term Climate Change: Projections and Predictability", Chapter 11). Initializing

models in order to produce predictions on a time scale of 10 to 30 years is challenging. Meehl et al. (2009) highlight the characteristics of decadal climate predictions concerning its initialization, which is schematically presented in Figure 2.1. For climate projections, which mainly try to describe climate trends on the evolution of greenhouse gases and aerosols, the boundary condition problem is considered. For shorter time scales, such as weather forecasting or seasonal to interannual climate predictions, the initial value problem is used as the current observations are essential for the starting conditions. Comparing the impact of initial conditions and external forcings in decadal climate predictions, Corti et al. (2015) find that predictability for sea surface temperatures (SSTs) on a global scale arise from external forcing for time scales longer than one year. For selected regions, however, the impact of initialization is longer for SSTs and the 0 to 700 m depth oceanic heat content. Analyzing the Atlantic meridional overturning circulation (AMOC) shows that the impact of initial conditions last up to 5 years or longer, but the degree of predictability tends to be more model dependent.

Predictability and skill for decadal predictions are affected by the initialization of the corresponding model components (Smith et al., 2007; Keenlyside et al., 2008; Pohlmann et al., 2009; Polkova et al., 2014; Romanova and Hense, 2015). There are different initialization techniques: we mainly differ between full-field and anomaly initialization. Full-field initialization uses 3 dimensional fields of atmospheric and/or oceanic variables whereas anomaly initialization uses anomalies of such 3 dimensional fields. Furthermore, the ensemble size has an impact of the skill of decadal predictions (Scaife et al., 2014; Sienz et al., 2016). Due to computational reasons the number of ensemble members is restricted but increasing the number of members increases the prediction skill.

Experiences with decadal climate predictions are still rare compared to weather forecasts or climate projections. Goddard et al. (2012) demonstrate how the experience with seasonal predictions can benefit decadal climate predictions. Looking for differences, the predicting time scales are obvious: seasonal predictions cover upcoming months, whereas decadal predictions cover up to 10 years. Moreover, there are different processes on different time scales: the El Niño-Southern Oscillation (ENSO) phenomenon is a prominent example for seasonal scales, whereas the Pacific decadal variability (PDV) or the Atlantic multi-decadal variability (AMV), which can influence SSTs, are observed on decadal time scales. First tests show that initializing a coupled model only with SST anomalies lead to positive skill in the North Atlantic region (Keenlyside et al., 2008). This implies skill in initializing the Atlantic Meridional Overturning Circulation (AMOC). A similar initializing technique is used by Matei et al. (2012) showing that the monthly mean AMOC at 26.5°N is predictable up to 4 years in advance. This skill arises from mid-ocean transport.

Goddard et al. (2013) give recommendations for verifying decadal predictions. Metrics analyzing the following two issues are chosen: firstly, the benefit of initializing decadal climate predictions compared to uninitialized predictions and secondly, the comparison of the ensemble spread and prediction uncertainty on average. The basic evaluation tools include bias adjustments, correlations, temporal and spatial aggregations. Additionally, they include the mean squared skill score (MSSS), which is based

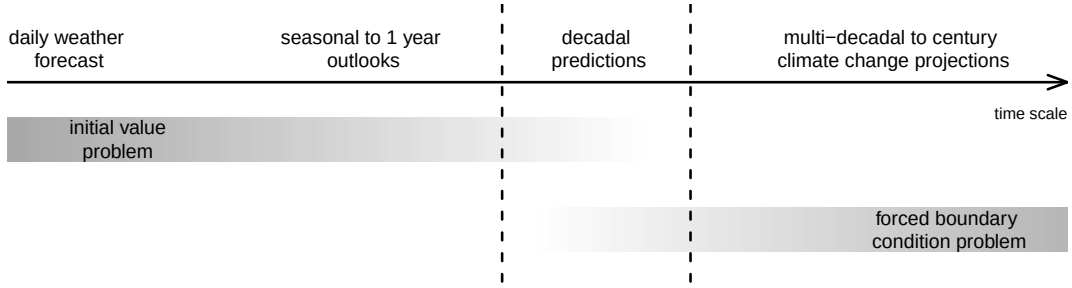

Figure 2.1: *The influence of initial conditions for weather forecasts and climate predictions on different time scales redrawn after Meehl et al. (2009).*

on the mean squared error (MSE) of the ensemble mean following Murphy (1988) and the continuous ranked probability (skill) score after Gneiting et al. (2007). Goddard et al. (2013) suggest to average the temporal information at different scales: lead year 1, 2-5, 6-9 and 2-9. Analyzing lead year 1 is of central interest as it is the transition of seasonal to multi-year prediction, where a rapid decrease of predictability is expected (Stolzenberger et al., 2016).

For near surface temperature and precipitation of the Decadal Climate Prediction System (DePreSys) by the Hadley Centre (Smith et al., 2010), Goddard et al. (2013) find only some regions, e.g. the North Atlantic, with improved skill compared to the uninitialized predictions for lead year 2-9. Smith et al. (2007) analyze the impact of initial conditions also for near surface temperatures of DePreSys and find skill improvement for lead year 1 and 1-9. Their latest seasonal prediction version of DePreSys3 is set up for 16 months starting each 1 November with 40 ensemble members. Dunstone et al. (2016) show that not only the first winter North Atlantic Oscillation (NAO) has skill but also for the second winter. The key drivers for this second winter NAO skill are ENSO and the stratospheric polar vortex strength, which is significantly correlated to the total solar irradiance forcing.

Within the MiKlip[1] project (Marotzke et al., 2016), decadal climate predictions based on the MPI-ESM model are evaluated for certain processes (more information on the prediction system is given in Section 4.2). The accuracy of surface temperature and precipitation is improved by initializing atmosphere and ocean (compared to uninitialized predictions), especially for lead year 1 (Kadow et al., 2016). Stolzenberger et al. (2016) show by using the example of the freshwater flux (evaporation minus precipitation) that potential predictability and skill only exist for lead year 1 and only for the tropical Pacific region. These signals disappear almost completely in lead year 2. This is one evidence that verifying near surface variables is a difficult task as long as the model physics and dynamical processes are in an experimental stage. Therefore, they recommend a three-dimensional evaluation of prognostic variables, such as temperature and geopotential height. Comparing atmosphere-ocean initialization to ocean-only initialization shows that the predictability in the inner tropics increases from 1 to 2 years

[1] http://www.fona-miklip.de

for e.g. geopotential height at higher altitudes (Stolzenberger et al., 2016).

Another process-oriented application is done, amongst others, by Kruschke et al. (2016), who look at the skill of extra-tropical cyclones and find positive skill for lead year 2-5 and 2-9 in winter. The two additionally applied initializing strategies (anomaly- versus full-field-initialization) show no significant differences. Spangehl et al. (2016) evaluate cloud parameters by implementing a satellite simulator in the MPI-ESM model. Analyzing different cloud types reveals the challenging task of evaluating cirrus clouds. Predictability for total cloud cover can be found e.g. in parts of the North Atlantic.

Besides MiKlip, the projects SPECS[2] (Seasonal-to-decadal climate Prediction for the improvement of European Climate Services), EUPORIAS[3] (European Provision Of Regional Impacts Assessments on Seasonal and Decadal Timescales), and NACLIM[4] (North Atlantic Climate), all funded by the European Commission, deal with seasonal to decadal predictions under certain aspects. The major aim of SPECS is to improve European forecasting models, also on regional scales to produce quasi-operational climate information. EUPORIAS is more societal oriented: one objective is to communicate/collaborate with stakeholders and provide them with tools e.g. for calibrating or downscaling climate information for their needs. NACLIM aims to optimize the observations in the North Atlantic and Arctic Ocean in order to improve the initial conditions for these regions. All of these projects want to achieve a better understanding of predictability on seasonal to decadal timescales (Hewitt et al., 2013).

2.2.2 Paleoclimate model simulations

The Paleoclimate Modelling Intercomparison Project (PMIP) is a prominent example focusing on climate reconstructions for certain time slices and periods in the past in order to get a better understanding of climatic changes. Its activity started in 1991. In Braconnot et al. (2011), a detailed overview of PMIP, its third phase and the connection to CMIP5 is presented. One major goal of PMIP is to get a better understanding about past climate changes and the efficiency of models. Bothe et al. (2013) study the consistency of the PMIP3 (PMIP Phase 3) multi-model ensemble including eight members for the past 1,000 years. Evaluating these simulations with global temperature reconstructions, they find regionally limited consistency e.g. the western tropical Pacific.

A global comparison for annual and seasonal surface temperatures in the Eemian interglacial (approx. 127 to 116 ka BP) is done by Otto-Bliesner et al. (2013) using the CCSM3 (Community Climate System Model, Version 3, also part of PMIP) model and different kinds of proxy data of more than 300 sites (over land and ocean). Discrepancies between observations and model data are explained by the chosen boundary conditions (present-day vegetation and polar ice sheets, etc.). A comparison between simulations based on the ECHO-G model and statistical climate reconstructions based on pollen proxies is done by Kaspar et al. (2005) for one time slice within the Eemian. For pollen-

[2] http://www.specs-fp7.eu
[3] http://www.euporias.eu
[4] http://www.naclim.eu

based reconstructions, they use the so-called pdf-method (Kühl et al., 2002), where the distribution for each occurring taxon is represented as pdf. Combining the individual pdfs lead to the most probable reconstructed climate state. For Europe, Kaspar et al. (2005) find that modeled and reconstructed January and July temperatures fit quantitatively well together and that orbitally induced changes in insolation is the main indicator to explain reconstructed temperature pattern. Lohmann et al. (2013) compare a model ensemble to derived SSTs from proxies taken from marine cores for the Mid-Holocene and discuss occurring differences in magnitudes for both fields: the underestimation of SST trends in the model can appear when e.g. the model is not sensitive enough regarding the insolation or cannot fully capture the natural range of climate variability. Uncertainties for reconstructing SSTs based on proxies can be due to assumptions, which are being made and the shifts in seasonality and habitat depth.

A local comparison at Lake Van (Litt et al., 2012a) between proxy data and model simulations is done by Stockhecke et al. (2016), who examine mainly the last 360 ka. Stockhecke et al. (2016) find a weak AMOC signal in the model simulations during Dansgaard Oeschger variability and extended droughts in the eastern Mediterranean region, which fit to e.g. low deciduous *Quercus* pollen percentages. Furthermore, spring/early summer precipitation and winter precipitation as measured by winter storm tracks are increased due to increased precession, which is observed during marine isotope stages (MIS) 5e, 5c and 5a in both model simulations and proxies. MIS represent climate change signals, which are based on benthic oxygen isotopes of several, globally distributed sediment cores (Lisiecki and Raymo, 2005). MIS 5 encompasses the time period between 130 and 80 ka BP, roughly the Last Interglacial and the beginning of the Last Glacial. Sub-stage MIS 5e corresponds to the Eemian warming phase (ca. 130 to 115 ka BP).

Harrison et al. (2015) give an overview of how future climate projections can be improved including the knowledge of past climate changes as recent observations encompass only a limited range of climate variability. By evaluating the CMIP5/PMIP3 Mid-Holocene and LGM simulations, Harrison et al. (2015) find only modest signals for global and regional skill: for different variables, either the magnitudes are under- or overestimated or the signal of the models is wrong. However, they find also model improvements compared to earlier model versions, which are not necessarily associated with the increased model complexity. von der Heydt et al. (2016) address further the difficulties of estimating radiative forcing although the proxy reconstructions become more accurate (e.g. CO_2 levels or ice sheet extends).

The climate modeling group of the CRC 806 looks, amongst others, at precipitation changes during the LGM over Europe (Ludwig et al., 2016). Therefore they use a weather typing approach, which characterizes four different regions dependent on the prevailing present-day circulation forms and apply it to four models, which are part of PMIP3. For western Europe, they find an increase of precipitation, compared to the control run, which can be associated with stronger evaporation over the North Atlantic. Comparing these model simulations to proxy-based reconstructions, the model simulations show higher precipitation values, especially over western Europe.

Within the platform Past Earth Network (PEN) funded by the British Engineering and Physical Sciences Research Council (EPSRC), it is studied what can be learned of past climate for future climatic changes. Mainly paleoclimate scientists and statisticians work together in four working groups looking at uncertainty quantification of observational data, uncertainty quantification of model data, data-model comparison and forecasting/future projections.[5]

Another framework for paleoclimate research is the project PalMod (Paleo Modeling: A national paleoclimate modeling initiative) funded by the German Ministry of Education and Research, which aims to model the climate of the last glacial cycle using comprehensive Earth system models instead of Earth system models of intermediate complexity (Latif et al., 2016). Another goal is to rate future climate projections with those adjusted models.

[5] http://www.pastearth.net

Chapter 3

Quality assessment

To make a clear statement about certain forecasts - not only "good" or "bad" (Murphy, 1993) - we analyze forecasts in different ways. Probabilistic forecasts need to be evaluated with probabilistic methods. Although it is challenging to compare e.g. predictive distributions with point observations, attributes were introduced to rank the predictions qualitatively. Murphy (1993) defines essential attributes, which are partially listed in the following:

Accuracy	Average correspondence between individual pairs of forecasts and observations
Skill	Accuracy of forecasts of interest relative to accuracy of forecasts produced by standard of reference
Calibration/Reliability	Correspondence between conditional mean observation and conditioning forecast
Sharpness	Variability of forecasts as described by distribution of forecasts

These aspects and corresponding metrics will be explained in the following sections, where essential parts have already been published in Stolzenberger et al. (2016).

3.1 Accuracy and Skill

Proper Scoring Rules

Scoring rules provide a framework to measure the accuracy of a forecast. To rank the forecast, a quantitative score is chosen based on the forecast probability and the underlying observation. For dichotomous variables (e.g. the occurrence (1) / non-occurrence (0) of precipitation), probabilistic forecasts produce a predictive probability $p \in [0, 1]$, which is compared to the observed 0 or 1. For continuous variables (e.g. temperature), probabilistic forecasts produce a predictive probability/cumulative density function (pdf/cdf), which is compared to an observed temperature value.

A scoring rule is proper if the forecast is assessed honestly (Gneiting et al., 2005) and treated without cheating, e.g. including only the best forecast performance. Let $S(P, y)$ be a scoring rule with the predictive distribution P and the verifying observation y. The expected score with respect to the

observational distribution G can be expressed by

$$S(P, G) = \int S(P, y) \mathrm{d}G(y). \tag{3.1}$$

The forecaster's aim is to optimize $S(P, G)$ in order to predict the truth ($P = G$). A negatively oriented score ("the smaller the score the better the prediction") is called proper if

$$S(G, G) \leq S(P, G) \tag{3.2}$$

and even strictly proper if $P = G$ (Gneiting and Raftery, 2007). A proper scoring rule can be understood as minimized expected score when the predictive distribution agrees with the "true" distribution (Thorarinsdottir et al., 2013). As propriety is an essential property for evaluating forecasts, the scores which are used in this work are mostly proved to be proper.

Skill Scores

Applying scores to real data, the score can be estimated by its sample value averaged over a fixed set of forecast/observation situations (Gneiting and Raftery, 2007),

$$\mathcal{S}_n = \frac{1}{n} \sum_{i=1}^{n} S(P_i, y_i), \tag{3.3}$$

where n is the number of observations in time and/or space and y_n is an independent and identically distributed (iid) sample from the observational pdf G. This iid assumption is often based on the climatological distribution as the actual distribution is completely unknown.

To classify the improvement of a forecast, it is useful to relate the forecast to a reference forecast. It is common to choose the observational climatology for the reference forecast although other forecasting systems (e.g. earlier experiment releases) can be used instead (Wilks, 2011). The skill score is given by

$$\mathcal{S}_n^{\mathrm{skill}} = \frac{\mathcal{S}_n^{\mathrm{fcst}} - \mathcal{S}_n^{\mathrm{ref}}}{\mathcal{S}_n^{\mathrm{perf}} - \mathcal{S}_n^{\mathrm{ref}}}, \tag{3.4}$$

where $\mathcal{S}_n^{\mathrm{perf}}$ is the score of a perfect forecast, and $\mathcal{S}_n^{\mathrm{ref}}$ is the score of a reference forecast. Assuming that for negatively orientated scores the perfect score $\mathcal{S}_n^{\mathrm{perf}}$ is zero, the skill score can be expressed as

$$\mathcal{S}_n^{\mathrm{skill}} = 1 - \frac{\mathcal{S}_n^{\mathrm{fcst}}}{\mathcal{S}_n^{\mathrm{ref}}}. \tag{3.5}$$

If the skill score $\mathcal{S}_n^{\mathrm{skill}}$ is one, the forecast is perfect (relative to the reference forecast). If the skill score is zero, the forecast has no improvement compared to the reference forecast. If the skill score is negative, the reference forecast performs better than the forecast. The scores in the following section can be expressed as skill scores (Equation 3.5). Note that, in general, the skill score is not proper anymore.

In the following section we will present common tools to assess the accuracy of forecasts without uncertainty information, i.e. only using the ensemble mean, and probabilistic forecasts.

3.1.1 Brier Score (BS)

The Brier Score (Brier, 1950; Wilks, 2011) can be used for the verification of dichotomous events, e.g. rain/no rain. It is the squared difference between forecast probability and the occurrence of an observed event, and the scoring rule can be written as

$$S(F,y)_{\mathrm{BS}} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2, \tag{3.6}$$

where p is the forecast probability and k the occurring event ($o = 1$ if the event occurs and $o = 0$ if the event does not occur) for N forecasts. Equation 3.6 differs to the original Brier score (Brier, 1950), which summed over both occurring and non-occurring events.

3.1.2 Continuous ranked probability score (CRPS)

The CRPS is a generalization of the Brier Score to continuous threshold values (Hersbach, 2000). The underlying scoring rule,

$$S(P,y)_{\mathrm{CRPS}} = \int_{-\infty}^{\infty}(P(x) - \mathcal{H}(x-y))^2\mathrm{d}x, \tag{3.7}$$

quantifies the difference between the predictive cdf $P(x)$ and the cdf of the observation $\mathcal{H}(x-y)$. The latter expression is known as Heaviside function, which is 0 if $y < x$ and 1 otherwise, assuming perfect observations y. For a deterministic forecast, the CRPS is identical to the mean absolute error (MAE) (Hersbach, 2000).

We use the analytic solution after Gneiting and Raftery (2007) for the standard Gaussian distribution $\mathcal{N}(\mu,\sigma^2)$,

$$\mathrm{CRPS}(\mathcal{N}(\mu,\sigma^2),y) = \sigma\left[\frac{1}{\sqrt{\pi}} - 2\phi\left(\frac{y-\mu}{\sigma}\right) - \frac{y-\mu}{\sigma}\left(2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1\right)\right], \tag{3.8}$$

where ϕ and Φ indicate the pdf and cdf. The CRPS is (as the Brier score) a negatively oriented score, which means the smaller the value the better the model skill. It is a very common score fulfilling the condition of a proper score (Gneiting and Raftery, 2007).

In case of forecast distributions, which include mixture Gaussian distributions, Grimit et al. (2006) give an analytical solution for a Gaussian mixture distribution, which returns

$$\begin{aligned}\mathrm{CRPS}\left(\sum_{m=1}^{M} w_m\mathcal{N}(\mu_m,\sigma_m^2),y\right) = \\ \sum_{m=1}^{M} w_m A(y-\mu_m,\sigma_m^2) - \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M} w_m w_n A(\mu_m-\mu_n,\sigma_m^2-\sigma_n^2),\end{aligned} \tag{3.9}$$

where

$$A(\mu,\sigma^2) = 2\sigma\phi\left(\frac{\mu}{\sigma}\right) + \mu\left(2\Phi\left(\frac{\mu}{\sigma}\right) - 1\right). \tag{3.10}$$

The expectation value and the standard deviation for each kernel density are represented by μ_m and σ_m. The number of realizations is denoted by M and the weights are set to be nonnegative and summing to one, $\sum_i^M w_i = 1$. Note that within this study all ensemble members are treated equally, i.e. $w_i = \frac{1}{M}$.

3.1.3 Energy score (ES)

The energy score (ES) is a multivariate generalization of the CRPS, which can be expressed as

$$S(P, y)_{\mathrm{ES}} = E_P||x - y|| - \frac{1}{2}E_P||x - x'||, \tag{3.11}$$

where E_P denotes the expectation value, $||\cdot||$ denotes the Euclidean norm, and $\vec{x}$ and $\vec{x}'$ are independent random values with the distribution P (Gneiting and Raftery, 2007). The first term of Equation 3.11 can be interpreted as the divergence between the predictive distribution and the observation. The second term is called entropy of the forecasts and is closely related to sharpness (see Section 3.3). Gneiting et al. (2008) introduce a computationally efficient Monte Carlo approximation for the ES (Equation 3.11),

$$\widehat{\mathrm{ES}}(P, y) = \frac{1}{k}\sum_{i=1}^{k}||x_i - y|| - \frac{1}{2(k-1)}\sum_{i=1}^{k-1}||x_i - x_{i+1}||, \tag{3.12}$$

where the vector x_i is a sample of size $k \geq 1,000$ from the predictive distribution P.

3.1.4 Mean square error skill score (MSESS)

In Stolzenberger et al. (2016), the MSESS introduced in Goddard et al. (2013) and Murphy and Epstein (1989) is presented in a version comparable to the ANOVA basics (see Section 3.3.1). Here, the mean squared differences between ensemble mean and observations at time j_l are compared with the mean squared differences between the observations at time j_l and all ensemble means except for j_l, the so-called model climate. The mean square error (MSE) between the model data and the observations at time step j_l, $\mathrm{MSE}(m, j_l)$, can be expressed as

$$\mathrm{MSE}(m, j_l) = \sigma(j_l)^2 + (\bar{x}(j_l) - y(j_l))^2, \tag{3.13}$$

where $\sigma^2(j_l)$ denotes the ensemble variance, $\bar{x}(j_l)$ the ensemble mean and $y(j_l)$ the observation, all at forecast time j_l. For the reference MSE, $\mathrm{MSE}(c, j_l)$, the model data for the whole time frame are used without verifying the observations at time j_l,

$$\mathrm{MSE}(c, j_l) = \frac{1}{N-1}\sum_{k=1,k\neq l}^{N}\sigma(j_k)^2 + \frac{1}{N-1}\sum_{k=1,k\neq l}^{N}(\bar{x}(j_k) - y(j_l))^2. \tag{3.14}$$

The MSESS measures the treatment variance compared to the overall variance (Stolzenberger et al., 2016),

$$\mathrm{MSESS} = 1 - \frac{\frac{1}{N}\sum_{j_l=1}^{N}\mathrm{MSE}(m, j_l)}{\frac{1}{N}\sum_{i=1}^{N}\mathrm{MSE}(c, j_l)}. \tag{3.15}$$

The MSESS is positive if the differences between the model predictions and the verifying observations are closer compared to the differences between the observations and all other predictions.

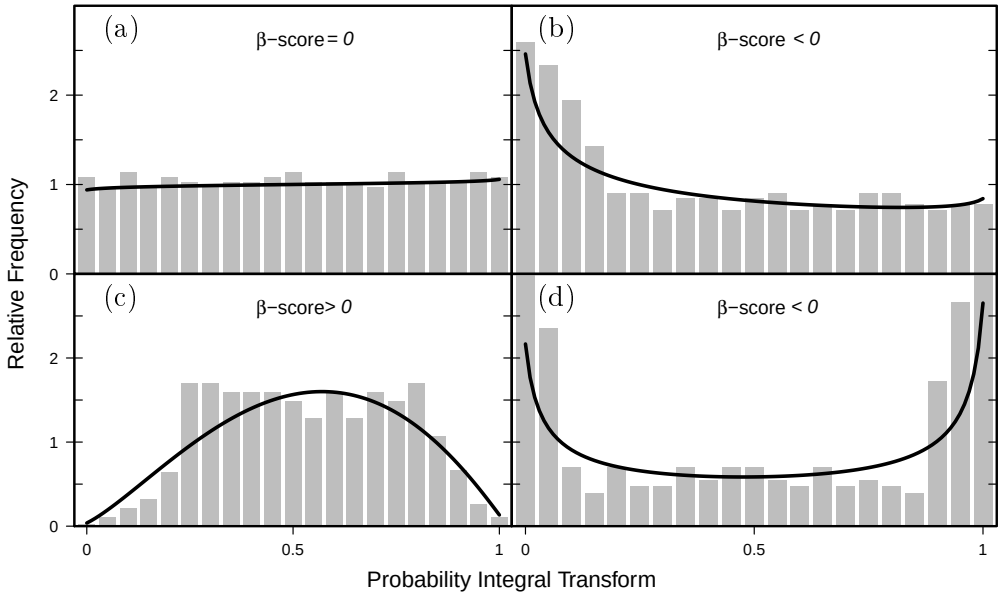

Figure 3.1: *Four cases of probability integral transform (PIT) histograms with the corresponding β-scores. These examples are based on 20 ensemble members representing (a) a perfectly calibrated ensemble, (b) a negatively biased ensemble, (c) an overdispersive ensemble and (d) an underdispersive ensemble.*

3.2 Calibration

3.2.1 Probability integral transform (PIT) histograms & β-scores

One method to analyze calibration is by creating probability integral transform (PIT) histograms. Assuming P as the predictive cdf of the prediction variable x, PIT is the value of the predictive cdf at the observation y. If y is a realization of P, PIT is uniformly distributed, which indicates a perfectly calibrated ensemble (Gneiting et al., 2008). A PIT histogram is obtained by plotting the histogram of the PIT values. Another well-known method is the analysis rank histogram, ARH (Anderson, 1996; Candille and Talagrand, 2005; Keller and Hense, 2011), which is the non-parametric analog to the PIT histogram (Keune et al., 2014).

In Keller and Hense (2011), the β-score, which summarize the graphical shape of the PIT histogram (or the ARH) in one single number, is introduced. The pdf of a β-distribution, $B(\alpha, \beta)$, is fitted to the PIT histogram. Based on the estimated shape and rate parameters α and β, the so-called β-score

$$\beta_S = 1 - \sqrt{\frac{1}{\alpha\beta}} \tag{3.16}$$

can be determined. Figure 3.1 shows different PIT histograms with the corresponding β-score. A U-shaped or biased PIT histogram is denoted by a negative β-score (too small ensemble spread) and an inverse U-shaped PIT histogram is denoted by a positive β-score (too large ensemble spread). The β-score is zero if the PIT histogram is flat (perfectly calibrated ensemble).

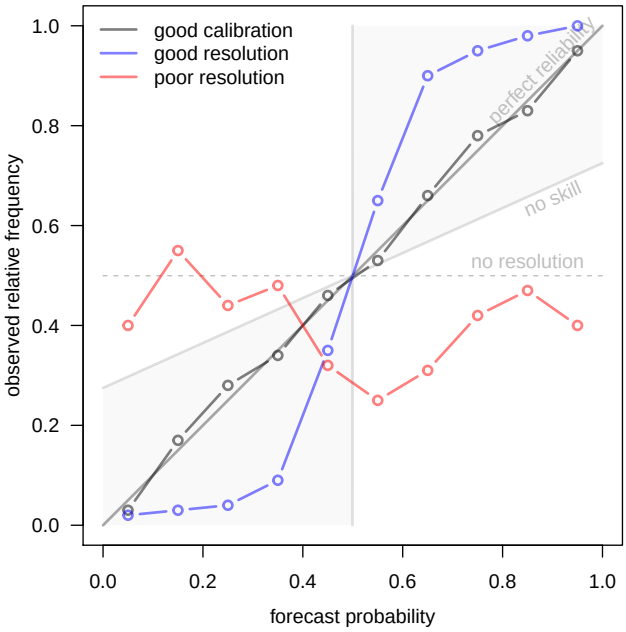

Figure 3.2: *Reliability diagram for three cases: perfectly calibrated ensemble (black line), the underconfident case (blue line) and the overconfident case (red line). Points on the diagonal "perfectly calibration"-line denote perfect calibration. Points on the dashed "no resolution"-line indicate forecasts which are unable to resolve occasions when the event probability is similar to the overall climatological probability. Points on the "no-skill"-line and the white area below indicate a negative forecast skill according to the Brier Score with the climatology as reference (Wilks, 2011).*

3.2.2 Reliability classifications

Another way to determine calibration of a forecast ensemble is a further analysis of reliability diagrams, which is, again, a graphical method to compare forecast probabilities with observed relative frequencies. The concept of reliability classifications has already been published in Stolzenberger et al. (2016), and we follow their description.

Before calculating reliability diagrams, model and observational data have to be prepared. First, we define the climatological median as threshold value at every gridpoint. Forecast probabilities are obtained by counting how many ensemble members are above this threshold. The binary observational-based field is obtained by exceeding or dipping below the threshold. Wilks (2011), Bröcker and Smith (2007) and others, explicitly explain how to calculate reliability diagrams and based on their description, we will briefly summarize this in two steps: Firstly, we partition the forecast probabilities into a fixed number of bins - in our case there are five bins assuming that the forecast probabilities are uniformly distributed. Secondly, we count how many observed events fall into each of the predefined bins. These rates are called observed relative frequencies, which are the maximum likelihood estimators of a binomial distributed random variable with the probability of occurrence π. The reliability diagram is a plot of forecast probabilities p versus the estimated values of π. In Figure 3.2, three examples of reliability

diagrams are shown. A perfect reliable binary forecast would give $\pi = p$ or a slope $m = 1$ (black line in Figure 3.2). The blue line indicates overforecasting biases for smaller p and underforecasting biases for larger p and the red line the reverse case (Wilks, 2011).

Using maximum likelihood estimation, we can determine the uncertainty ranges of the estimated π. As the variance of a binomial random variable equals $n_p\pi(1-\pi)$, where n_p is the number of forecasts predicting the occurrence of the binary events, the uncertainty of the estimated π is given by $\sqrt{\frac{\pi(1-\pi)}{n_p}}$. The inverse of this standard deviation is used for a weighted least squared regression fit. By calculating the slope m and intercept of this regression fit, the reliability diagram is classified into three categories (Weisheimer and Palmer, 2014). Model data are well calibrated and reliable if the slope $m = 1$. The model data are potentially useful if the slope is $0.5 < m < 1.5$, and the model data are unreliable if $m < 0.5 \vee m > 1.5$ (Stolzenberger et al., 2016).

These analyses are applied at every gridpoint aggregated over ± 1,000 km in a great circle distance if horizontal fields are analyzed and ± 50 hPa in the vertical if applied to zonally averaged variables at all available pressure levels. For zonally averaged oceanic variables the gridpoints are aggregated over ± 50 m in the vertical.

3.3 Sharpness

3.3.1 Analysis of variance (ANOVA)

The ANOVA measures the sharpness, or the potential predictability, i.e. the co-variability between different data series (von Storch and Zwiers, 2001). As this analysis looks not at the verification of forecast predictions, observations are not used. Based on the ensemble realizations, it is determined if the information content of the initialized forecasts compared to the internal variability (climate noise) is large enough to generate a predictive value.

We assume the following statistical model for each realization i at treatment j,

$$y_{ij} = \mu + a_j + \epsilon_{ij}, \tag{3.17}$$

where μ is the overall mean and a_j the so-called treatment effects, $a_j = \mu_j - \mu$. The errors ϵ_{ij} fulfill the iid assumption and are normally distributed.

Partitioning of the variance into the treatment and error components can be obtained by building the total sum of squares

$$\mathcal{SST} = \sum_{i=1}^{M}\sum_{j=1}^{N}(y_{ij} - \bar{y})^2 \tag{3.18}$$

$$= \mathcal{SSA} + \mathcal{SSE}, \tag{3.19}$$

where M denotes the number of ensemble realizations and N the number of treatments. The overall mean is represented by $\bar{y} = \frac{1}{NM}\sum_{i=1}^{M}\sum_{j=1}^{N} y_{ij}$ and is, thus, an unbiased estimator of μ. The ensemble mean at treatment j, $\bar{y}_j = \frac{1}{M}\sum_{i=1}^{M} y_{ij}$, is an unbiased estimator of $\mu + a_j$. $\mathcal{SSA}$ represents the treatment components

$$\mathcal{SSA} = M\sum_{j=1}^{N}(\bar{y}_j - \bar{y})^2 \tag{3.20}$$

and $\mathcal{SSE}$ represents the error components

$$\mathcal{SSE} = \sum_{i=1}^{M}\sum_{j=1}^{N}(y_{ij} - \bar{y}_j)^2. \tag{3.21}$$

The proportion of variance due to the treatments is given by R^2, which is also known as coefficient of multiple determination,

$$R^2 = \frac{1}{\mathcal{SST}}\left(\mathcal{SSA} - \frac{(N-1)}{N(M-1)}\mathcal{SSE}\right). \tag{3.22}$$

The treatment effect can be analyzed by accepting/rejecting the null hypothesis $\mathbf{H_0} : \sum_{j=1}^{N} a_j^2 = 0$, which means that there are no treatments. To test $\mathbf{H_0}$, the F-ratio

$$\mathbf{F} = \frac{\mathcal{SSA}/(N-1)}{\mathcal{SSE}/(N(M-1))} \tag{3.23}$$

is used. By comparing Equation 3.23 with the p-quantile of the F-distribution $F(1-\alpha, N-1, N(M-1))$, $\mathbf{H_0}$ is tested at a significance level α.

As treatments, we use the prediction years as in Stolzenberger et al. (2016). The ANOVA, which is described here, is known as one-way ANOVA. The two-way ANOVA is extended to account for the effects of two treatments including the interaction effects (the two-way ANOVA is not used within this work).

Part I

Probabilistic evaluation of decadal climate predictions

Chapter 4

Verifying observations and predictions

This chapter provides an overview of the decadal prediction system MiKlip including the different baseline experiments. Additionally, we introduce the observational datasets for atmosphere and ocean with which the decadal MiKlip predictions are compared.

4.1 Atmospheric and oceanic reanalyses

The observational atmospheric and oceanic data come from reanalyses provided by the ECMWF (Uppala et al., 2005; Dee et al., 2011; Balmaseda et al., 2013) and the University of Hamburg (Köhl and Stammer, 2008). From now on, we call these datasets "observations" although they are not in its strict sense. Dynamical and statistical models are used to aggregate the data from the different observing systems (Hense, 2005).

4.1.1 Atmospheric data

ERA-40 & ERA-Interim

The reanalysis ERA-40 was produced in 2003 and is explained explicitly in Uppala et al. (2005). It covers 45 years starting from September 1957 until August 2002. ERA-40's original horizontal resolution is T155 before interpolating it to the ECHAM-T63 Gaussian grid (approx. 1.9° in longitude and latitude). The ERA-Interim reanalysis starts in 1979 and is available until today (Dee et al., 2011). The increased spatial resolution of T255 is due to the advanced data assimilation system with which also the temporal resolution is increased. This dataset is also interpolated to the coarser ECHAM-T63 grid. For both reanalysis products, there are 60 vertical levels up to 0.1 hPa.

In order to evaluate the whole available time frame given by the predictions (see Section 4.2), we use the ERA-40 data for the period 1960-1989 and ERA-Interim for the period 1990-2014 as it is done for the prediction system (Pohlmann et al., 2013).

4.1.2 Oceanic data

In contrast to observed atmospheric data, the long term observation of oceanic data with an appropriate horizontal resolution is problematic. To get a better understanding of the dynamics and processes

in this climatic subsystem, measurements in the ocean are steadily extended.

One measuring component to observe the ice-free ocean is the use of ARGO floats.[1] Battery-powered floats measure temperature and salinity profiles to a depth of 2,000 m. These data are transmitted to satellites when they surface again (the battery lasts for about 150 cycles). The positions of more than 3,900 ARGO floats are determined by satellites or GPS signals. Another method to observe oceanic data is exploiting the shipping routes. The Ocean Climate Observation Program of NOAA (National Oceanic and Atmospheric Administration) developed a world-wide network of so-called XBTs (expendable Bathythermographs), which are measurements for temperature and salinity. They are dropped into the ocean from a ship and measure along certain transects in the upper ocean. By repeating these measurements several times a year, it is possible to get a in-depth knowledge about upper layer circulations. Several ships taking part in this network do additional oceanic measurements, e.g. temperature and salinity depth-profiles, CO_2 partial pressure, etc.[2]

The following two oceanic reanalyses use the data that are obtained by ARGO floats and XBTs, amongst other measurements, for assimilation. Although it is a challenging task to extend the observations in the ocean with higher spatial and temporal resolution, these measurements become more important when we look at oceanic variabilities on e.g. decadal time scales.

ORAS4

ORAS4 (Ocean ReAnalysis System 4) is an ocean reanalysis provided by the ECMWF for the time period from 1958 until present (Balmaseda et al., 2013). Compared to earlier releases, ORAS4 uses a different ocean model: NEMO (Nucleus for European Modelling of the Ocean) Version 3 by Madec (2008). As data assimilation system, a newly developed 3D variational assimilation system NEMOVAR (Mogensen et al., 2012) has been applied. The horizontal resolution is 1° with a refined meridional resolution of 0.3° at the equator. It has 42 vertical levels with the highest resolution in the upper 200 m with 10-15 m level thickness.

GECCO2

Köhl and Stammer (2008) present the second version of GECCO (German contribution to Estimating the Circulation and Climate of the Ocean), which is provided by the University of Hamburg, Germany. It covers the time frame from 1948 to today and is, thus, longer than the previous version GECCO, which is available from 1952 to 2001. The zonal resolution is 1° and the meridional resolution is 0.3° at the equator, whereas the horizontal resolution of GECCO is 1°. The number of vertical levels has been increased (from 23 levels in GECCO to 50 levels in GECCO2). The configuration is based on the MITgcm model (Massachusetts Institute of Technology general circulation model, Marshall et al. (1997)) and GECCO2 uses a 4D-VAR adjoint method for assimilating data.

[1] ww.argo.ucsd.edu

[2] www.oco.noaa.gov/xBTsOOPS.html

Table 4.1: *Overview of the experiments within the MiKlip prediction system (baseline 0, baseline 1 and prototype) including major differences. Basic variables are temperature (T), salinity (S), vorticity (V), divergence (D), and surface pressure (P). ERA-Interim is abbreviated as ERA-I.*

	baseline 0	baseline 1	prototype
model	MPI-ESM-LR/MR	MPI-ESM-LR/MR	MPI-ESM-LR
initialization atmosphere	no initialization	full field V, D, T & log(P) from ERA-40/ERA-I	full field V, D, T & log(P) from ERA-40/ERA-I
initialization ocean	anomalous T&S from NCEP forced MPIOM	anomalous T&S from ORAS4	full field T&S from ORAS4+GECCO2
ensemble size	3	10/5	15 + 15

4.2 MiKlip prediction system

The quality of a prediction system can be obtained by using retrospective forecasts, so-called hindcasts or reforecasts. Hindcasts are forecasts initialized with a numerical model (Hamill et al., 2006) for a certain time step in the past, which are used to perform the forecasts for the subsequent time period. Forecast skill can be determined by probabilistic evaluating these hindcasts with the available observations.

The MiKlip experiments are based on the model output of the MPI-ESM (Max Planck Institute Earth System Model) in a low (LR) and mixed (MR) resolution. For LR, the atmospheric resolution is T63 with 47 vertical levels up to 0.1 hPa. The oceanic component is based on a bi-polar grid with a resolution of approx. 1.5° at the equator. For MR, the horizontal atmospheric resolution is equal to the LR version but the number of vertical levels is increased (95 levels up to 0.1 hPa). For the ocean, MR uses a tri-polar grid with a resolution of approx. 0.4° at the equator. Both LR and MR have 40 vertical levels in the ocean. Detailed information of the MPI-ESM model are given in Giorgetta et al. (2013).

There are three development stages (baseline 0, baseline 1, prototype) of the MiKlip prediction system, which are briefly described in the following. Major differences are summarized in Table 4.1.

Baseline 0

Baseline 0 (b0) is the CMIP5 version of the MPI-ESM-LR model. After forcing the MPIOM (Max Planck Institute Ocean Model) with atmospheric energy, water and momentum fluxes from the NCEP / NCAR reanalysis (Kalnay et al., 1996), three-dimensional fields of ocean temperature (T) and salinity (S) can be obtained. From this ocean-only simulation, T and S anomalies are nudged into the coupled model (anomaly initialization). This coupled model is initialized yearly (for the time period between 1960-1990) and runs free for 10 years forced with observed greenhouse gases, solar and volcanic variations, and anthropogenic aerosols. It includes 3 ensemble members. For the same time frame, every

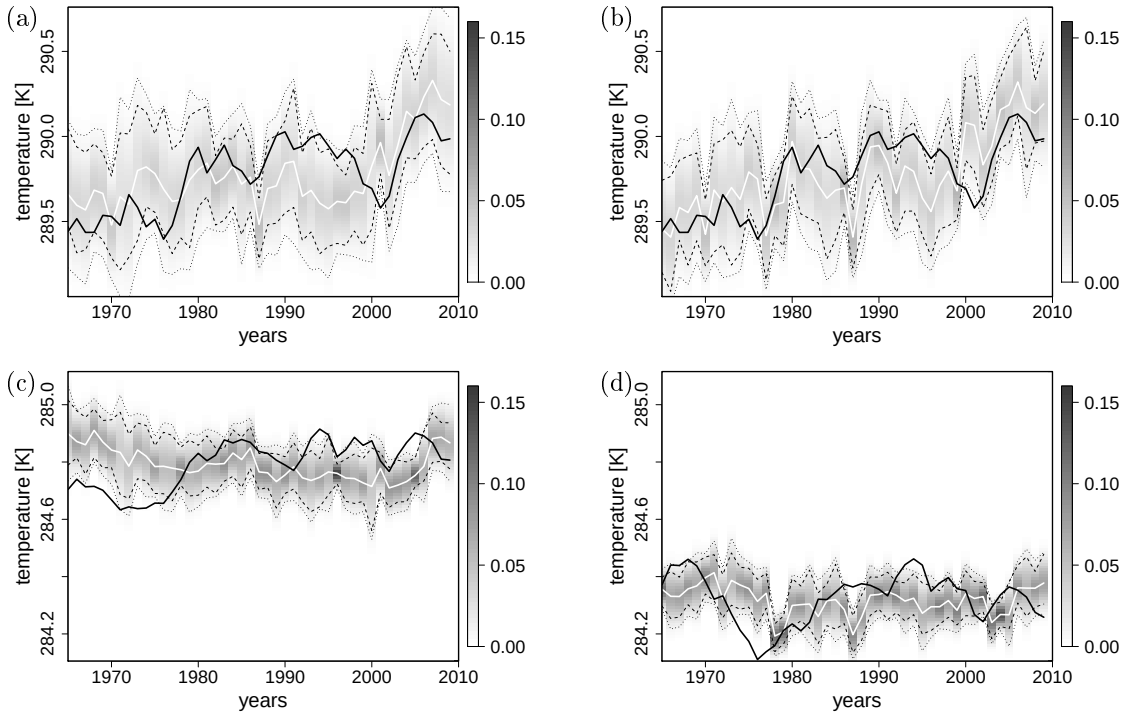

Figure 4.1: *Predictive pdfs (gray shadings) and ensemble mean (white solid line) for prediction year 2-5 and corresponding observations (solid black line) for T as a function of time. T is spatially averaged over the region North-East Pacific; (a) pr-GECCO and ERA (combination of ERA-40 and ERA-Interim) at 850 hPa; (b) pr-ORA and ERA at 850 hPa; (c) pr-GECCO and GECCO2 averaged over 100 and 500 m depth; (d) pr-ORA and ORAS4 averaged over 100 and 500 m depth.*

5 years the coupled model is initialized for 10 ensemble members. From 2000, b0-LR is a ten-member ten year forecast ensemble that is yearly initialized.

The b0 experiment is also set up for the MPI-ESM-MR version. The model is initialized every 5 years from 1961-1999 and yearly from 2000-2012 for a three-member ensemble, respectively. The baseline experiments b0-LR and b0-MR will not be analyzed within this work due to the small number of ensemble members and the small time frame of yearly initialized simulations.

Baseline 1

For baseline 1 (b1), the ocean is initialized in the same way as in b0 (see Section 4.2), but with three-dimensional T and S anomalies of the ORAS4 reanalysis provided by ECMWF (Balmaseda et al., 2013). The atmosphere is initialized with full three-dimensional fields (so-called full-field initialization) of vorticity (V), divergence (D), T and surface pressure (P) of ERA-40 (Uppala et al., 2005) and ERA-Interim reanalyses (Dee et al., 2011) from ECMWF. After Pohlmann et al. (2013), these two ERA products are merged in 1989/1990. Both model resolutions (LR and MR) are used, both are yearly initialized (1961-2014) but differ in the ensemble size (10 members for b1-LR, 5 members for b1-MR).

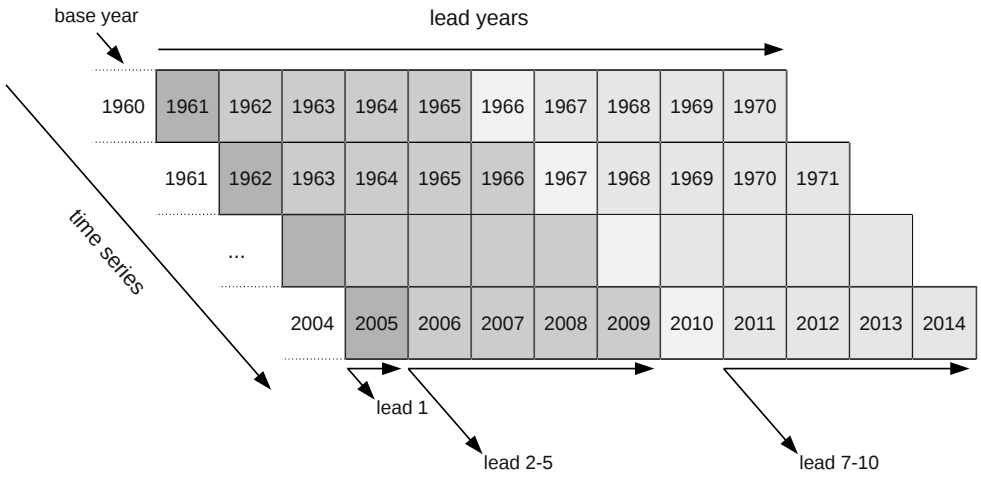

Figure 4.2: *Flow chart for temporal averaging of lead years for the entire time period in dependence on the base year. Details can be found in Section 4.2.1.*

Prototype

The prototype (pr) system only uses the MPI-ESM-LR model output. The atmosphere is initialized the same way as in b1 with a full three-dimensional field initialization. For the ocean initialization, also full fields of T and S are taken from the ORAS4 reanalysis (Balmaseda et al., 2013) and the GECCO2 reanalysis (Köhl, 2015). For each pr-set (based on GECCO2 and ORAS4), there are 15 ensemble members available. From now on, we rename both sets as pr-GECCO and pr-ORA.

Combining both pr-sets to a 30 member ensemble is not recommended, especially for the oceanic variables such as e.g. T. Figure 4.1 presents the differences between the ensemble uncertainty and the observations. For the atmosphere at 850 hPa, the averages over the North-East Pacific show similar characteristics for both pr-sets. The temporal evolution of the ensemble spread is almost analog for both cases as they include exactly the same initializing dataset. For the ocean (Figures 4.1c and d), however, the differences are clearly visible. Averaging over the North-East Pacific area and the depth from 100 to 500 m shows that the pdf-shapes of the pr-GECCO and pr-ORA vary up to 0.4 K as well as the respective observations. Moreover, the corresponding observations are only partly within the ensemble spread.

4.2.1 Temporal average of the MiKlip predictions

To verify the MiKlip predictions, we analyze single years and multi-year averages as it has been already shown e.g. in Goddard et al. (2013), Müller et al. (2012) and others. Lead years 2-5 or 7-10 are prominent examples on the decadal scale as they reveal the differences between the early and late

phase of the predictions (Stolzenberger et al., 2016). Figure 4.2 exemplarily shows how the predictions are sorted with respect to the lead time. For each base year and each ensemble member, there is one file provided that includes monthly predictions with a horizon up to 10 years. For e.g. lead year 2-5, we take the single lead years from 2 to 5 for each starting date and average it over the time frame for each ensemble member.

Chapter 5

Verification of the MiKlip hindcasts

For an overall quality check of the MiKlip prediction system, we apply a variety of scores and other metrics, which are described in Chapter 3. We firstly highlight the three-dimensional evaluation of atmosphere and ocean by looking at reliability, skill and potential sharpness of geopotential height and temperature. Moreover, we concentrate on the evaluation of temperature in the North Atlantic, tropical Pacific and the structure along the equator. Finally, we present a method to jointly verify the zonal and meridional wind components.

5.1 Three-dimensional evaluation of atmosphere and ocean

In this section, we will concentrate on annual means of temperature (T) and geopotential height (Z) at those 17 different pressure levels in the atmosphere and 35 depth layers in the ocean, which are directly available in both model and reanalysis. Global climate models predict T and Z as prognostic variables and extrapolate near surface variables e.g. near surface T afterwards. As decadal climate predictions are still experimental, information of dynamical structures and processes in upper layers of the atmosphere and deeper layers of the ocean can be provided to the modeling community. This section contains parts of Stolzenberger et al. (2016).

As mentioned in Section 4.2.1, we use multi-year averages (Goddard et al., 2013; Müller et al., 2012). Prominent examples are lead year 2-5 and 7-10 as differences between the early and late phase can be figured out. Lead year 1 plays a special role as it is the transition between seasonal and multi-year predictions. The model simulations are bias corrected for each lead horizon separately, which means that climate trends are still preserved but model drifts are excluded (Stolzenberger et al., 2016).

Probabilistic verification of Z

Reliability for the baseline experiments b1-LR, b1-MR, pr-GECCO, and pr-ORA is analyzed by applying the β-scores for Z at 850 hPa and 200 hPa as a function of latitude (see Figure 5.1). The uncertainty area is calculated by sampling over the parameters α and β (see Section 3.2.1) and taking the 5%- and 95%-quantiles when calculating the β-score.

At 850 hPa, the structure of the β-scores is similar for the four different baseline experiments and the

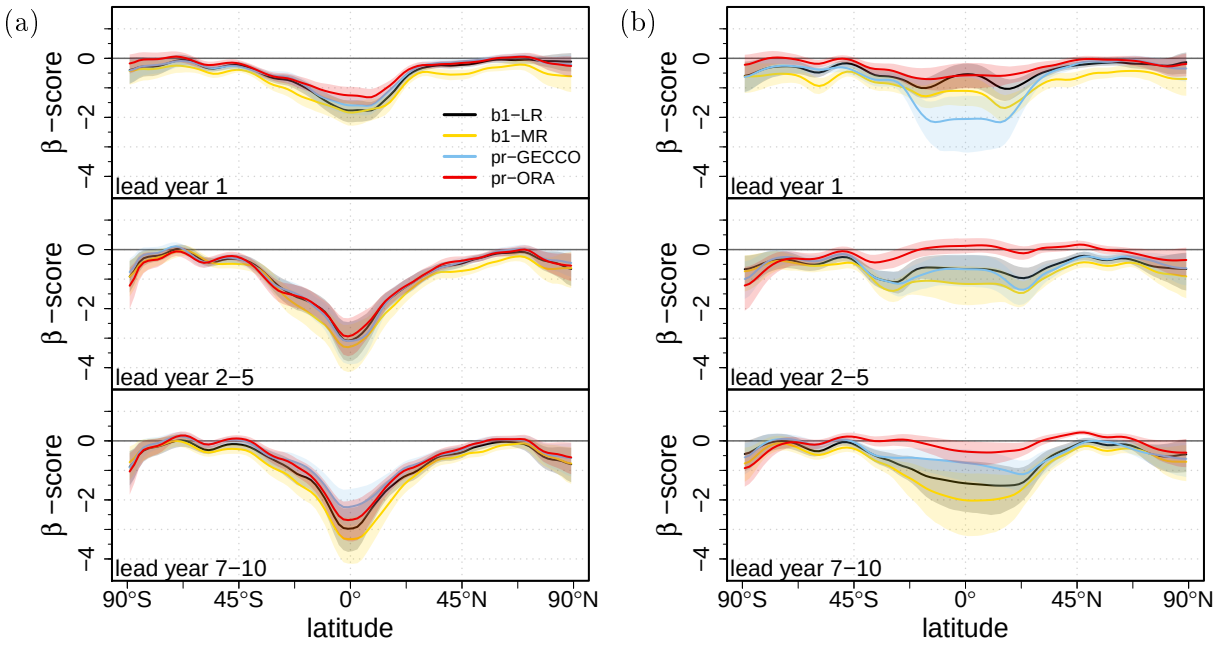

Figure 5.1: *β-scores in dependence of latitude for Z at 850 hPa (a) and 200 hPa (b). For both levels, the β-scores for b1-LR (black), b1-MR (yellow), pr-GECCO (blue) and pr-ORA (red) are shown for lead year 1 (top), lead year 2-5 (middle) and lead year 7-10 (bottom). The lightly colored areas correspond to the 5%- and 95%-quantile. The ensemble is perfectly calibrated if the β-score is zero.*

different time spans. β-scores close to zero are found in the mid-latitudes, whereas strongly negative values appear in the tropical region approx. between 30°S and 30°N. This minimum in the tropics intensifies from lead year 1 to lead year 2-5 with β-scores between -3 and -4. For lead year 7-10, this structure is still similar but in the tropical region, the baseline experiments differ: b1-MR shows the lowest β-scores followed by b1-LR, pr-ORA and pr-GECCO.

At 200 hPa, the β-scores show a different structure in the tropics compared to 850 hPa. At the equator, the β-scores are generally closer to zero but some experiments have local minima at approx. 20°S and 20°N. The β-scores for pr-ORA are almost perfect for all latitudes and lead times. Although the initialization in the atmosphere is identical for the considered prediction systems, the β-scores differ especially for lead year 1 in the tropics. Pr-GECCO clearly fails compared to pr-ORA and the b1 experiments with β-scores of -2. For lead year 2-5, differences between pr-GECCO and b1-LR are marginal. However, for lead year 7-10, pr-GECCO shows better results compared to the b1-experiments between 45°S and 45°N. Despite of the increased vertical resolution, b1-MR shows no benefits compared to its counterpart b1-LR for any lead year and altitude.

In Figure 5.2, the CRPSS for pr-ORA at 850 hPa, 500 hPa and 200 hPa is shown with observational climate as reference CRPS for lead year 2-5. The experiment has skill if the values are positive (red

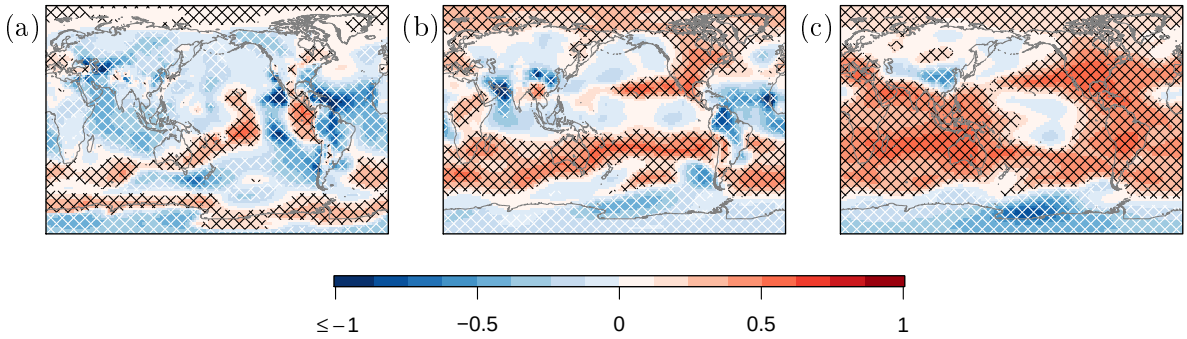

Figure 5.2: *CRPSS for pr-ORA with climatology as reference forecast for Z at 850 hPa (a), 500 hPa (b) and 200 hPa (c). The skill of pr-ORA is better than climate if the CRPSS is positive (red shadings). White/black hatching means that the CRPSS is negatively/positively robust.*

shadings). If the skill score is zero or negative, the reference forecast is as good as the baseline experiment or even better. The skill robustness is tested by sampling uncertainty, which is assessed by bootstrapping. This is indicated by the white (for negative skill robustness) and black (for positive skill robustness) hatching. The bootstrap sample is generated by randomly drawing n forecast/observation pairs of fields with replacement from the original n pairs and calculating the CRPSS for each sample. The bootstrapping is done on the complete fields and not on each gridpoint separately in order to preserve the spatial dependencies. From the bootstrap sample (sample size is 1,000), the p-value for a CRPSS of zero is estimated. If this p-value is smaller than 0.05 (larger than 0.95), we call the CRPSS to be robust in terms of its positive (negative) skill (Stolzenberger et al., 2016).

At 850 hPa, there are only a few areas with positive robust skill, e.g. over south-east and south-west of South Africa, Australia and parts of the tropical Pacific, and the western coast of South and Middle America. For most parts, however, the skill scores are negative and robust, i.e. the climatology predicts significantly better than pr-ORA. The areas of positive skill enlarge at higher altitudes. At 500 hPa, a southerly belt parallel to the equator is visible, where the skill scores are positive and robust. Also the American and European North Atlantic coastal area is skillful but in the central North Atlantic, the skill scores are negative. The CRPSS in the central North Atlantic get positive at 200 hPa, but these skill scores are partly not robust. In general, most parts of the globe represent positive CRPSS-values except for the tropical Pacific, east Europe and continental Asia. The results for b1-LR and pr-GECCO are similar (see Appendix A.1). For lead year 7-10, the skill is preserved as for lead year 2-5 but the CRPSS-values are partly decreasing, even though the robustness is still given.

For Z at 500 hPa, we compare the baseline experiments to study which experiment performs best. In Figure 5.3, six CRPSS combinations are shown for lead year 2-5. Comparing b1-MR with b1-LR as reference, the CRPSS is negative for most parts. There are some areas of slightly positive CRPSS, e.g. in the tropical region of Africa, but these results are not robust. This means that the increased vertical model resolution, which is used for b1-MR, is less beneficial compared to the lower vertically

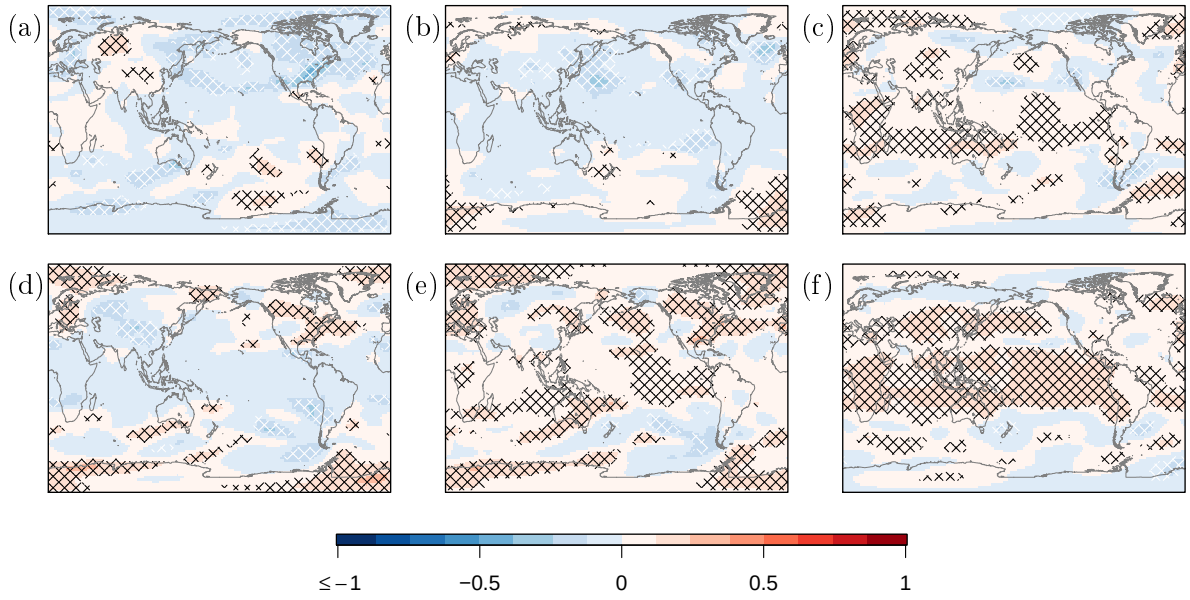

Figure 5.3: *CRPSS for b1-LR, b1-MR, pr-GECCO and pr-ORA with different reference forecasts for Z at 500 hPa: (a) CRPSS for b1-MR with b1-LR as reference CRPS; (b) CRPSS for pr-GECCO with b1-LR as reference CRPS; (c) CRPSS for pr-ORA with b1LR as reference CRPS; (d) CRPSS for pr-GECCO with b1-MR as reference CRPS; (e) CRPSS for pr-ORA with b1-MR as reference CRPS; (f) CRPSS for pr-ORA with pr-GECCO as reference CRPS. These results are based on lead year 2-5. The skill of the different baseline experiments is better than its reference if the CRPSS is positive (red shadings). White/black hatching means that the CRPSS is negatively/positively robust.*

resolved b1-LR experiment. For pr-GECCO versus b1-LR there are, again, no significant skill scores indicating that pr-GECCO performs as good as b1-LR. Examining the CRPSS for pr-ORA with b1-LR as reference forecast, we determine a profitable signal for pr-ORA especially along the equator and the tropical Pacific although in this region, climate predicts better than pr-ORA (see Figure 5.2b). Comparing the prototype predictions with b1-MR, the differences between pr-GECCO and b1-MR are smaller compared to the differences between pr-ORA to b1-MR. For the latter case, pr-ORA is conclusive, again, in the tropical Pacific, southern Indian Ocean, and also in Europe and parts of the North Atlantic. The last case examines the differences between the two prototype sets, which are assumed to predict similarly, i.e. the CRPSS should be zero. However, there are positive and robust skill scores in the tropics and in parts of the northern mid-latitudes indicating a more skillful behavior for pr-ORA. Only small areas can be found, where pr-GECCO predicts better than pr-ORA, e.g. over New Zealand and the western South Atlantic. At 850 hPa (see Appendix A.1), we can also see a predominance of pr-ORA, especially over sea, whereas pr-GECCO has more skillful areas over land. At 200 hPa (see Appendix A.1), the CRPSS-values for pr-ORA increase in the tropics but pr-GECCO preserve more skill in the southern mid-latitudes.

As b1-MR does not show any benefits concerning reliability and skill and only contains 5 members,

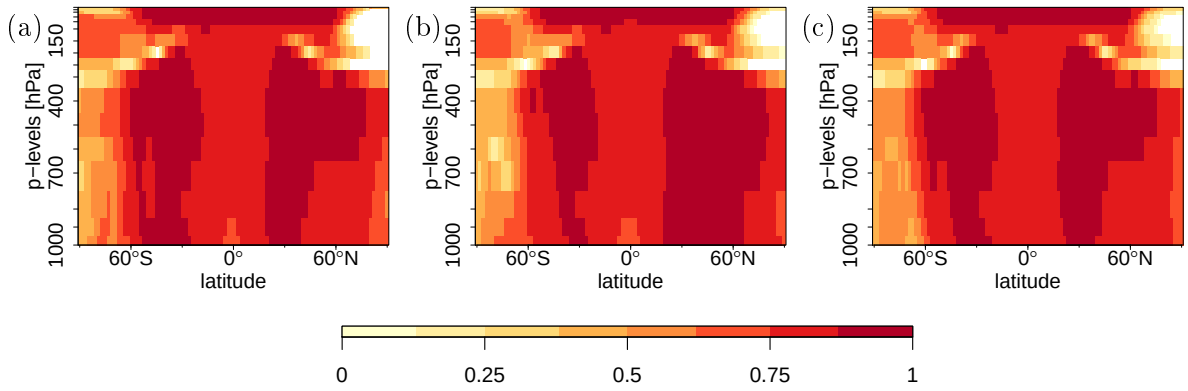

Figure 5.4: *Anova (R^2) for b1-LR (a), pr-GECCO (b) and pr-ORA (c) for zonally averaged T and lead year 2-5. The pressure levels (p-levels) are plotted on the ordinate.*[1]

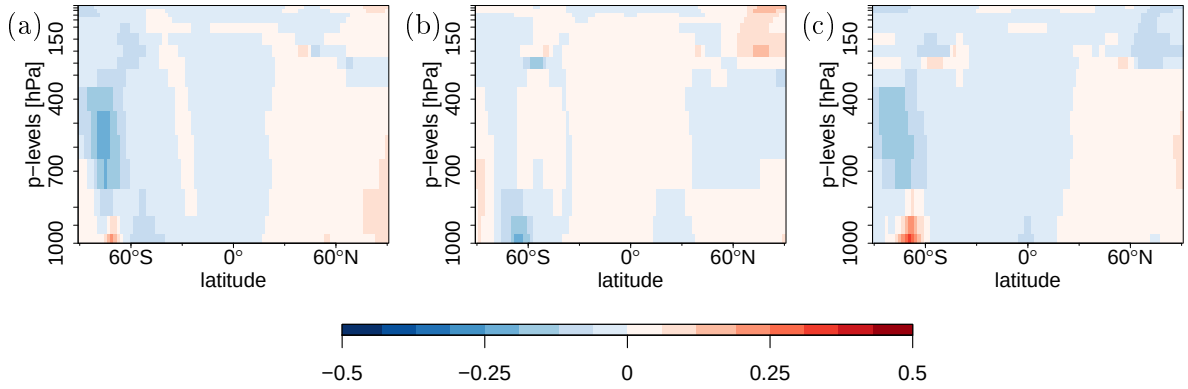

Figure 5.5: *ANOVA differences for zonally averaged T and lead year 2-5: (a) pr-GECCO minus b1-LR; (b) pr-ORA minus b1-LR; (c) pr-GECCO minus pr-ORA.*

which is little for statistical studies based on the ensemble member size, we exclude it from further analyses. CRPSS for other pressure-levels and other lead years can be found in the Appendix A.1.

Probabilistic verification of T

Analyzing the sharpness of the three remaining baseline experiments, we look at the ANOVA for zonally averaged T as latitude-height cross section for lead year 2-5 (see Figure 5.4). For all experiments, we find high values (more than 85%) in the subtropics and mid-latitudes throughout the troposphere and lower stratosphere. The ANOVA reaches values above 75% for the tropical region although we detect a strong underdispersive signal (negative β-scores in Figure 5.1). Stolzenberger et al. (2016) find high ANOVA values for the uninitialized ensemble in the low resolution but not as pronounced as the other baseline experiments in the subtropics. They state that potential sharpness arises from the external forcing and not from the initialization. Looking at the ANOVA differences between the experiments, we detect only small variations (see Figure 5.5). The ANOVA values for pr-GECCO are slightly higher for the northern subtropics and mid-latitudes compared to b1-LR and pr-ORA, negligible more than

[1] In the following, pressure-level is abbreviated as p-level.

10%, whereas in the southern hemisphere, the other two experiments outperform respectively. For the tropical area, pr-ORA has slightly positive values compared to b1-LR and pr-GECCO.

For zonally averaged T in the atmosphere and in the ocean, we look at the reliability classifications (see Figure 5.6). The shaded areas represent the three classes: reliable, potentially useful and not useful. Furthermore, we add the MSESS to the reliable classifications.

In the atmosphere, b1-LR shows a few reliable and potentially useful areas, e.g. in the northern subtropics and mid-latitudes and in the lower stratosphere below 150 hPa. This structure is similar for the two prototype sets, where reliable areas between 30°N and 60°N are increased. The highest positive MSESS values appear between 1,000 and 600 hPa and 60°S and 60°N. For all experiments, the MSESS turns strongly negative between 500 and 200 hPa and between 30°S and 30°N but are positive again in the lower stratosphere. In the southern mid-latitudes between 1,000 and 700 hPa the MSESS has slightly negative values, which fits to the unreliable areas.

As MSESS and ANOVA have the same statistical basics, we are able to directly compare both results with each other. The negative MSESS e.g. in the subtropical mid-troposphere or in the mid-latitudes on the southern hemisphere indicate that the climatology predicts better than the baseline experiments. For these regions, the ANOVA shows values of partly more than 75%. This implies that ensemble members agree on the predictions but the predictions are false (Stolzenberger et al., 2016). This signal is unimproved during the development stages in the MiKlip prediction system.

In the ocean, we find in all experiments predominantly potentially useful areas concerning the reliability (see Figure 5.6). At the surface layer, unreliable areas in b1-LR turn into partly reliable areas in pr-GECCO and pr-ORA. Also in the deeper oceanic layers, reliable areas increase in the pr-sets compared to b1-LR. Due to plotting artefacts of the contour lines and shadings, variances can appear, i.e. high MSESS values and not reliable areas. One exception is presented by pr-GECCO between 500 and 2,500 m depth and 30°S and 30°N. In this region, pr-GECCO shows indeed an improvement up to 80% compared to the model climate reference but these predictions are neither reliable nor potentially useful. High MSESS values do not necessarily need to be associated with reliable or potentially useful areas as both measures are based on different calculations. The reliability diagrams are, thus, created that observations and model predictions exceed the climatic median. The MSESS, however, improves compared to the reference if large amplitudes are better predicted (Stolzenberger et al., 2016).

We examine more detailed the unreliable area despite high MSESS values in pr-GECCO. Slope and y-intercept (see Figure 5.7) play the essential role for the reliability classification (see Section 3.2.2). The slope is negative and the y-intercept is above 0.5 indicating the overconfident case where underforecasting biases are associated with small forecast probabilities and overforecasting biases are associated with large forecast probabilities (Wilks, 2011). When using lower or upper terciles to receive the observational relative frequencies (Weisheimer and Palmer, 2014), the lack of reliability in pr-GECCO

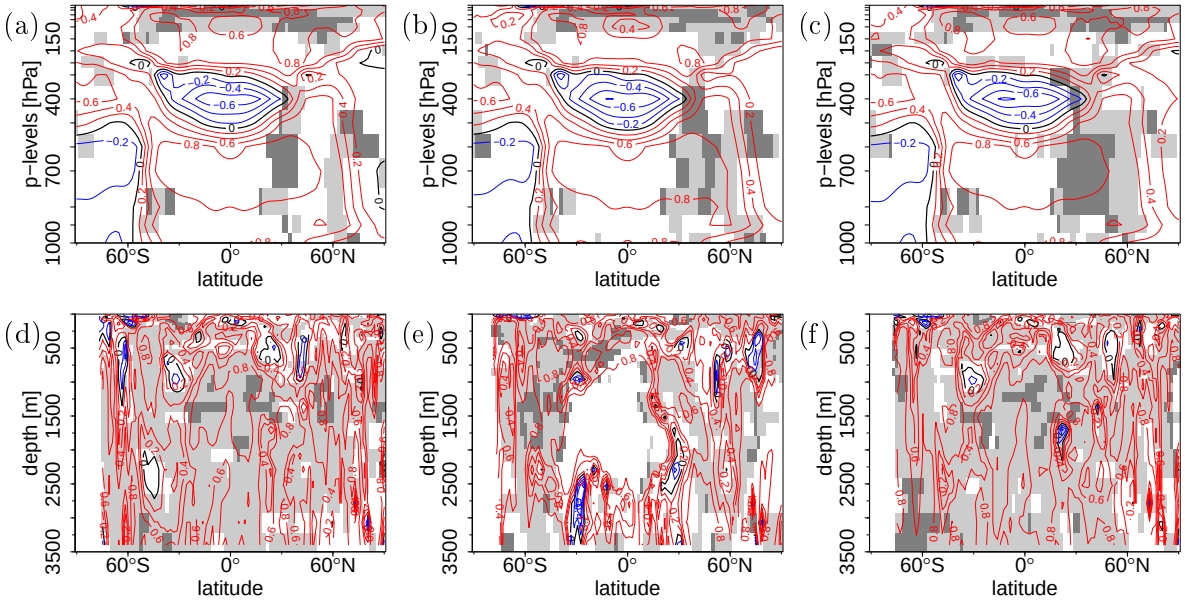

Figure 5.6: *Reliability classifications and MSESS for zonally averaged T in atmosphere and ocean. The results are shown for b1-LR in atmosphere (a) and ocean (d), for pr-GECCO in atmosphere (b) and ocean (e), and for pr-ORA in atmosphere (c) and ocean (f). The results are based on lead year 2-5. The contour lines display positive (red), negative (blue) and zero (black) MSESS values. Dark gray shadings correspond to reliable areas, light gray shadings to potentially useful areas and white areas to not useful areas.*

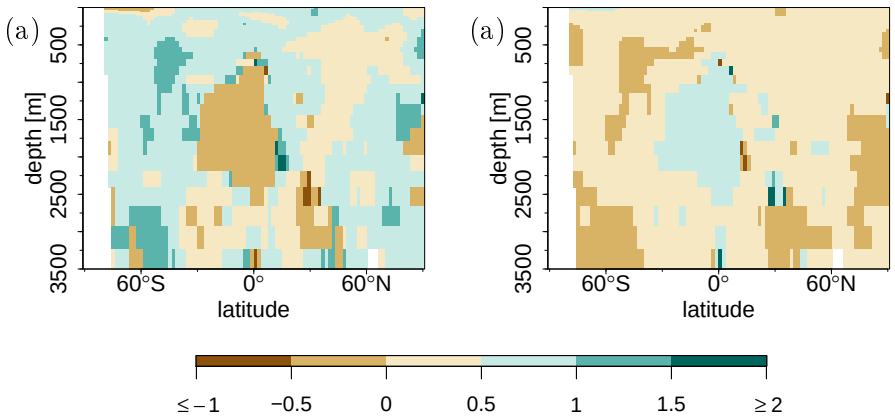

Figure 5.7: *Slope (a) and y-intercept (b) based on the reliability diagram for pr-GECCO for zonally averaged T in the ocean. The results are based on lead year 2-5. The ensemble is perfectly calibrated if the slope is one and the y-intercept is zero.*

persists (see Figure A13 in the Appendix). The reliable areas can change though when aggregating over a larger vertical area (larger than ± 50 m depth).

For the atmosphere, lower or upper terciles as thresholds show similar reliability structures compared to the median except for the stratosphere and the tropical mid-troposphere. For the ocean, we detect more reliable and potentially useful areas for negative than for positive anomalies (see Appendix A.2).

5.2 Process-oriented evaluation

5.2.1 Reliability and skill for certain regions

We examine reliability and skill for the North Atlantic (NA, 55°E-0°, 30°N-50°N) and the tropical Pacific region (PAC, 120°E-120°W, 15°S-15°N). As shown in previous studies by Müller et al. (2012) and others, the North Atlantic region is of central interest when looking at the skill and predictability of decadal climate predictions. Quantities such as the AMOC (Matei et al., 2012) or the North Atlantic heat content (Pohlmann et al., 2009) influence the European climate on decadal time scales. In the Pacific ocean, climate variations on decadal time scales can be observed (Pacific decadal oscillation, PDO), which are connected with winter climate for North America and Asian monsoons (Mochizuki et al., 2010). We concentrate on the variable T in the atmosphere and the ocean for lead year 2-5 and analyze b1-LR, pr-GECCO and pr-ORA.

In the atmosphere, the β-scores evolve similarly for the three baseline experiments and both considered regions (see Figure 5.8). For NA and between 1,000 and 400 hPa, the β-scores for pr-ORA and b1-LR are very close, whereas the β-scores for pr-GECCO are systematically smaller. Pr-ORA and b1-LR use the same data for initializing the ocean although the initialization technique differs. Coupling effects can be the reason why pr-ORA and b1-LR are more similar in the lower troposphere. At higher altitudes, the β-scores get smaller with values up to -3. For PAC, there are generally negative β-scores, which is in line with the results for Z (see Figure 5.1) and there are no clear differences between the experiments. From 1,000 hPa to approx. 300 hPa, there is a slight decrease from -1.5 to 0.5, but at higher p-levels the β-scores decrease.

In the ocean, the β-scores are negative as in the atmosphere indicating that the ensembles are underdispersive, but we can detect clear differences between the baseline experiments. While the β-scores at the surface of NA and PAC are similar and close to zero, the characteristics change in deeper layers. In the tropical Pacific, the β-scores for pr-GECCO are closer to zero compared to the other two experiment which is clearly different for the North Atlantic, where the β-scores reach values up to -30 below 2,000 m. The β-scores for b1-LR and pr-ORA show a similar evolution especially in the upper layers until 1,000 m depth for both regions.

The CRPSS is analyzed between 15°S and 15°N along the equator in the atmosphere (see Figure 5.9).

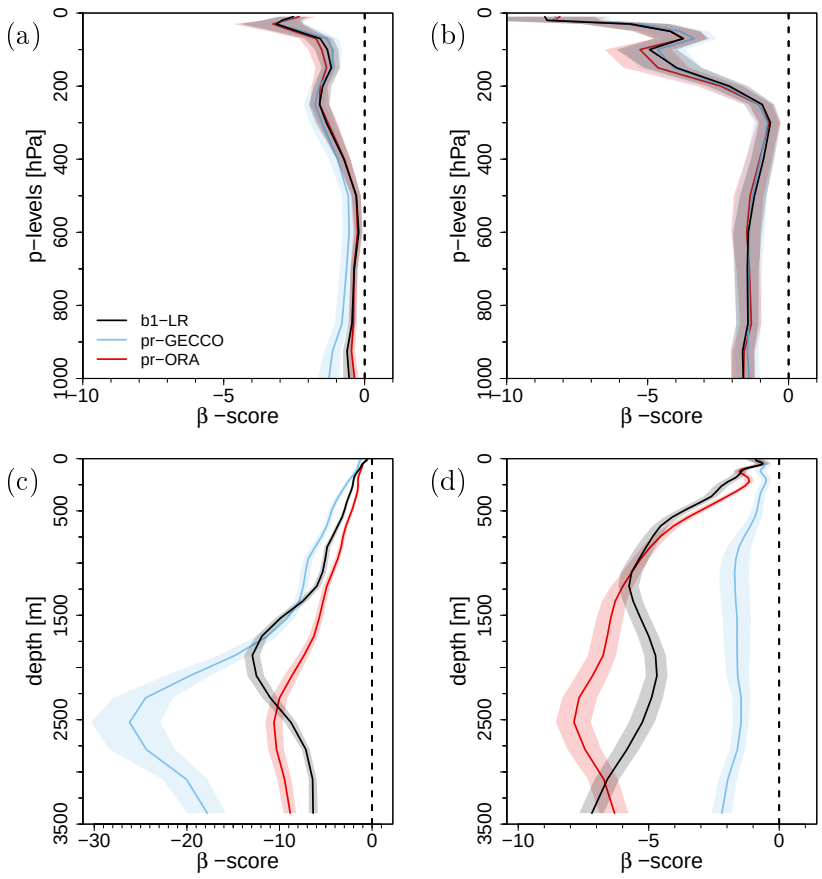

Figure 5.8: *β-scores for b1-LR, pr-GECCO and pr-ORA in the North Atlantic and the tropical pacific region for T in atmosphere and ocean. (a) β-scores in the North Atlantic (atmosphere); (b) β-scores in the tropical pacific region (atmosphere); (c) β-scores in the North Atlantic (ocean); (d) β-scores in the tropical pacific region (ocean). For all graphics, the β-scores for b1-LR (black), pr-GECCO (blue) and pr-ORA (red) are shown for lead year 2-5. The lightly coloured areas correspond to the 5%- and 95%-quantile. The ensemble is perfectly calibrated if the β-score is zero. Here, we define the North Atlantic region from 55°E to 0° and 30°N to 50°N and the tropical equator region from 120°E to 120°W and 15°S to 15°N. Note that for (c) the abscissa ranges from -30 to 1 compared to the other cases.*

We detect only slight differences between the baseline experiments when comparing the CRPS with climate as reference CRPS. Between 400 and 250 hPa, the skill scores are positive and robust except for the tropical Pacific region. For this region, the negative skill scores are noticeable throughout lower pressure levels to the surface. For the western part of South America and the Atlantic, the skill scores are positive between 900 and 700 hPa. At the same levels, the CRPSS are also positive for the area around Indonesia and Malaysia. Above 250 hPa, there is a layer of negative CRPSS followed by positive skill scores in the stratospheric levels. Comparing the baseline experiments with each other, only marginal differences can be determined. Pr-ORA shows slight but robust improvements compared to b1-LR for the tropical Pacific and some small areas at lower pressure levels. For pr-GECCO versus

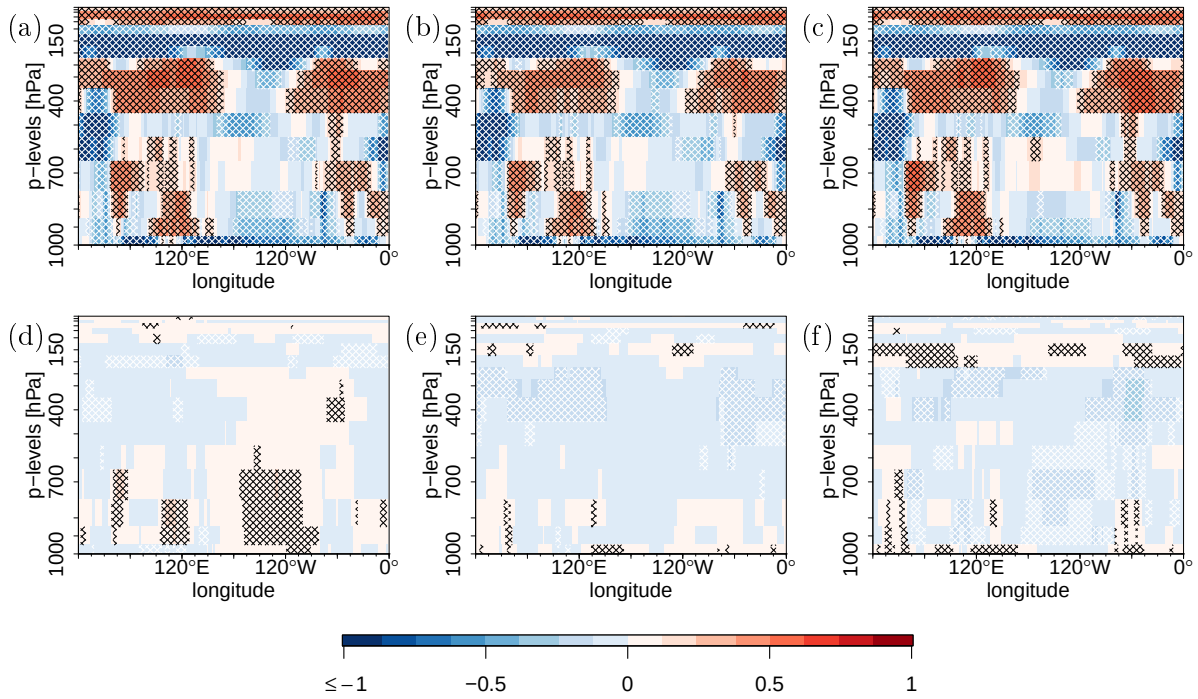

Figure 5.9: *CRPSS for b1-LR, pr-GECCO and pr-ORA with different reference forecasts for T in the atmosphere along the equator averaged between 15°S - 15°N: (a) CRPSS for b1-LR with climate as reference CRPS; (b) CRPSS for pr-GECCO with climate as reference CRPS; (c) CRPSS for pr-ORA with climate as reference CRPS; (d) CRPSS for pr-ORA with b1-LR as reference CRPS; (e) CRPSS for pr-GECCO with b1-LR as reference CRPS; (f) CRPSS for pr-GECCO with pr-ORA s reference CRPS. These results are based on lead year 2-5. The skill of the different baseline experiments is better than its reference if the CRPSS is positive (red shadings). White/black hatching means that the CRPSS is negatively/positively robust.*

b1-LR, we cannot see clear improvements, but robust benefits for b1-LR at 400 to 250 hPa. At higher altitudes, positive CRPSS-values for pr-GECCO with pr-ORA as reference forecast can be seen. For the tropical Pacific and between 900 to 700 hPa, the CRPSS for pr-GECCO is negative and robust indicating that pr-ORA predicts better compared to pr-GECCO although the magnitudes are only small. In general, the differences between the baseline experiments are slight especially between the two pr-sets.

For the ocean, we detect clear changes in the CRPSS structure (Figure 5.10). Comparing the CRPS for the baseline experiments with the corresponding climate as references CRPS (here, the observational climate bases on the data which are used for initialization), it is noticeable that especially the CRPSS for pr-GECCO includes positive and robust values for most parts of the considered area. Whereas in the upper layers (surface to 500 m depth) the values are mostly negative or zero, the amplitudes even increase in deeper layers, e.g. the tropical Pacific region between 2,000 and 3,000 m depth. The CRPSS for b1-LR and pr-ORA, both with climate as reference, show only a few areas where positive skill exists, e.g. at the surface. Comparing pr-ORA with b1-LR, we determine clear improvements

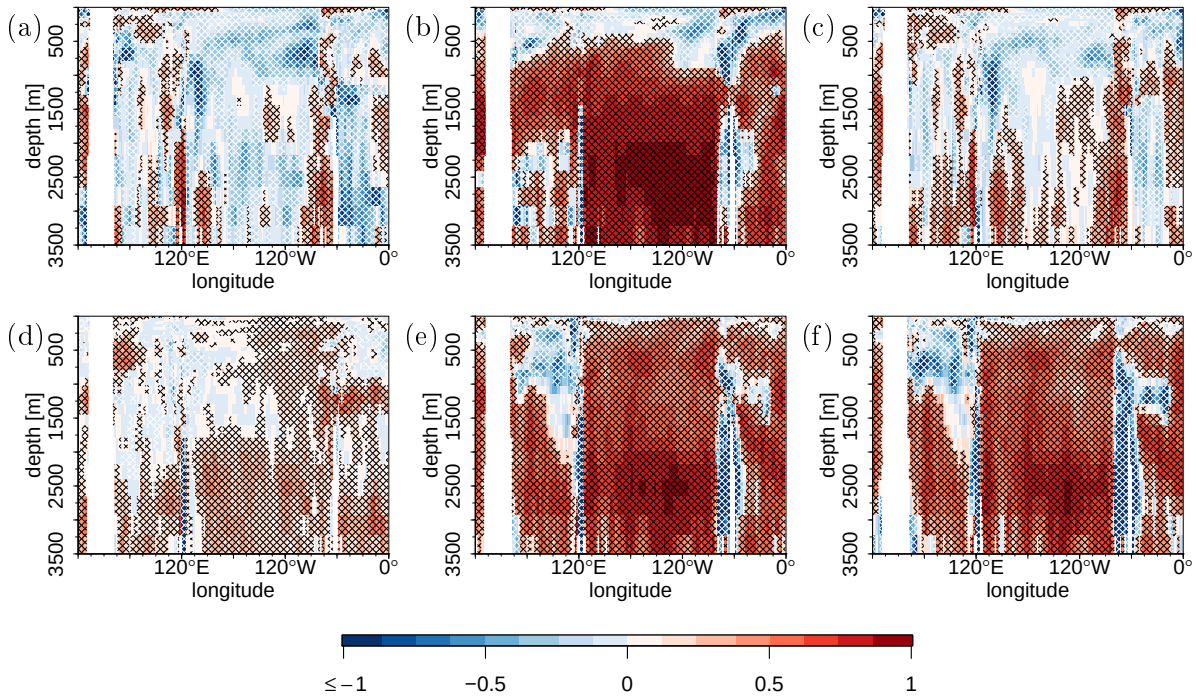

Figure 5.10: *Same as in Figure 5.9 but for T in the ocean.*

for pr-ORA especially in the deeper layers and the Pacific area throughout the surface. This signal increases when looking at the CRPSS for pr-GECCO with b1-LR as reference CRPS, where only small areas are negative and robust. Large discrepancies appear when analyzing the CRPSS for pr-GECCO with pr-ORA as reference CRPS, which is against our expectations as the two pr-sets are similarly initialized. Especially between 120°E and 120°W, the skill scores are positive and robust. Also in the deeper layers of the Indian ocean and the Atlantic, pr-GECCO predicts better compared to pr-ORA, or to be more precise, pr-GECCO fits better to the underlying observational dataset GECCO2 than pr-ORA to ORAS4.

The skill scores for the ocean are in concordance to the reliability results achieved by applying the β-scores. Whereas the β-scores for b1-LR and pr-ORA show similar structures in the tropical Pacific with strongly negative values, the β-scores for pr-GECCO are also negative but closer to zero through all depth layers. This indicates that the choice of the dataset which is used for the ocean initialization has more impact on reliability and skill than the initialization technique, at least for the considered region.

Merge the prototype ensembles

As a similar predicting behavior of the pr-sets in the atmosphere is provided, we merge each 15 member ensemble to one prototype ensemble including 30 members. The CRPS is calculated as above with observational climate as reference forecast (see Figure 5.11). The results look similar as in Figure 5.9a-c.

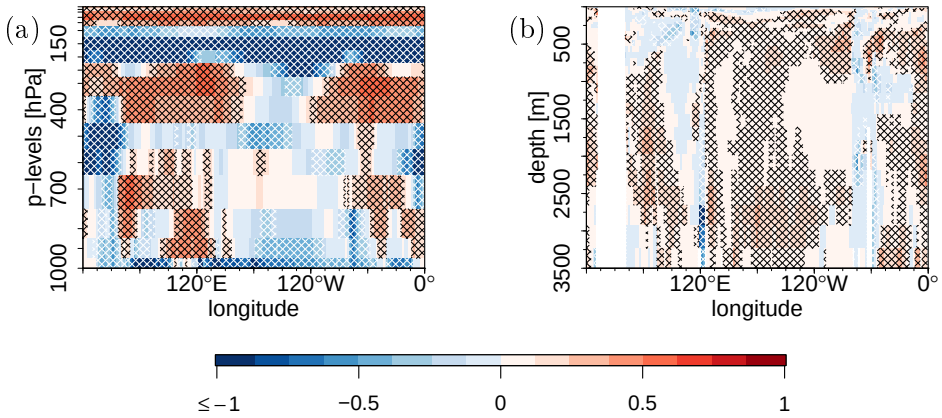

Figure 5.11: *CRPSS for merged pr-sets in atmosphere and ocean: (a) CRPSS for the merged pr-ensemble (30 ensemble members from pr-GECCO and pr-ORA) with ERA-climate as reference CRPS in the atmosphere; (b) CRPSS for the pr-ensemble verified with GECCO2 and as reference CRPS the pr-ensemble verified with ORAS4. Note that for (b) the CRPS for Gaussian mixture distributions is used. The analyses are done for T averaged between 15°S and 15°N and lead year 2-5.*

However, enlarging the ensemble by merging both pr-sets brings together the benefits of each pr-set (e.g. the positive and robust area at 175°W and 600 hPa) and diminish the weak characteristics due to averaging over all ensemble member.

For the ocean, the discrepancies between pr-GECCO and pr-ORA are large (see Figure 5.10f) and thus not recommended to merge as for the atmosphere. One method to jointly verify the two pr-sets for oceanic T is to apply a Gaussian mixture model. The mixture model density is a sum of weighted Gaussian distributions. The sum of the weights has to be one and in our case, all ensemble members are treated equally. For these skill analyses, we use the analytical solution of the CRPS for Gaussian mixture distributions (Grimit et al., 2006), which has already been introduced in Chapter 3 (Equation 3.9). Firstly, we determine the CRPS for pr-GECCO and pr-ORA with GECCO2 as verifying observation and secondly, we calculate it with ORAS4 as verifying observation. In Figure 5.11b the relation of these two CRPS is shown. There are areas of positive skill, i.e. verifying the 30-member ensemble with GECCO2 has higher skill values in comparison of ORAS4, e.g. in the deeper layers of the Indian or Pacific Ocean. However, there are areas where the CRPSS is zero, which means that GECCO2 is as good as ORAS4 for verifying the ensemble. As both pr-sets and the corresponding reanalyses show large discrepancies, the CRPS approach for Gaussian mixture distributions is a useful tool to analyze the skill of the joint ensemble.

5.2.2 Joint evaluation of zonal and meridional wind field

We present a concept to verify bivariate variables. As example we choose the zonal and meridional wind component at 10 m height for January and July. The model simulations are taken from the b1-LR

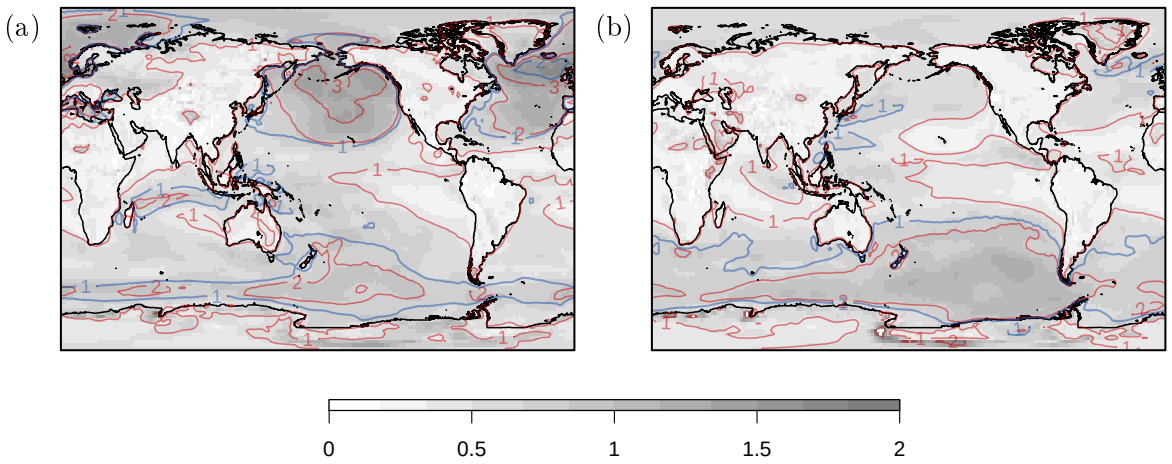

Figure 5.12: *Bivariate CRPS (ES) for the u- and v-wind component for January (a) and July (b). Divergence (red contour line) and entropy (blue contour line) are shown with the resulting energyscore (gray shading). Model basis is b1-LR for lead year 2-5. ES, divergence and entropy are displayed in m/s.*

experiment for lead year 2-5. Firstly, the joint covariance matrix for both wind components is estimated at each grid point for each time step. Secondly, we generate 1,000 simulations by random sampling from the predictive density. Here we assume a multivariate kernel dressing. The energy score (ES) is the difference between divergence and entropy and a negatively oriented score (see Section 3.1.3). The best case is, thus, to obtain low values for the divergence indicating small differences between prediction and observation and high values for the entropy indicating potential predictability.

The results for the ES with its components divergence and entropy are shown in Figure 5.12 for January and July. For both months, the ES has lower values over land compared to over sea, as the wind velocities in 10 m height are generally higher over the ocean. In January, there are two prominent areas, the North Atlantic and the northern Pacific, where the ES reaches highest values up to 2 m/s. The divergence shows values up to 3 m/s, whereas the entropy has moderate values (1 m/s). For July, we detect higher ES values in the southern Pacific whereas in the northern Pacific the ES is partly zero.

The ES components for the North Atlantic (NA) and the northern Pacific (PAC) regions are examined as time series (see Figure 5.13). The North Atlantic region is restricted between 52.5°W to 7.5°E and 40.3°N to 62.3°N and the Pacific region between 136.9°E to 110.6°W and 15.9°N to 55.3°N. Again, the smaller the divergence values and the higher the entropy values, the smaller the resulting ES. Especially for the North Atlantic, the divergence shows a slightly negative trend, whereas the entropy fluctuates constantly around 1.5 m/s (1.2 m/s in the PAC). The divergence of the climatic component evolves similarly as for b1-LR in both regions but for the PAC region, the climatic part has generally lower values compared to the prediction. Looking at the entropy, there are several years, where the climate shows higher values for NA but not for PAC. In contrast to January, the results for July show

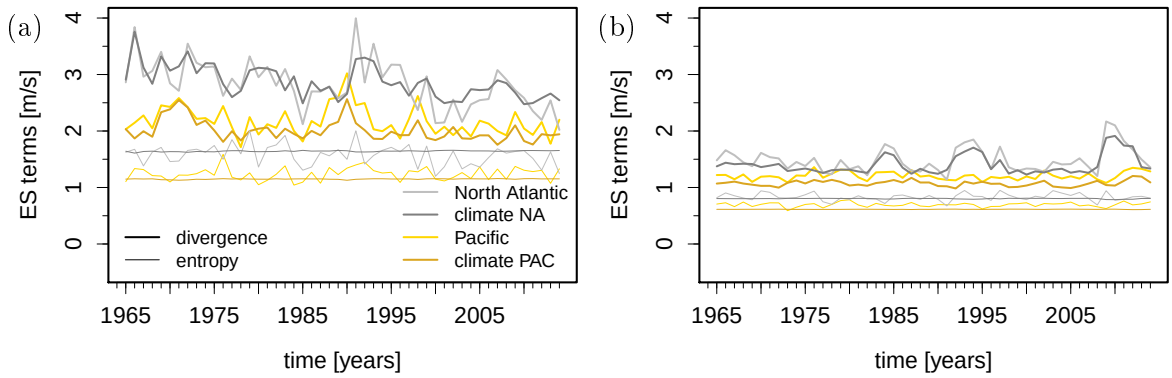

Figure 5.13: *Time series for divergence (thick line) and entropy (thin line) of joint u- and v-wind component for North Atlantic (NA, gray line) and Pacific (PAC, brown line), again for January (a) and July (b). Model basis is b1-LR for lead year 2-5.*

generally lower values for divergence and entropy and smaller differences between the two considered regions. Moreover the amplitudes of the fluctuations are lower especially for the Pacific region. For the NA, we detect three prominent maxima in the divergence, where the last maximum has its dip at about 2010. This indicates not only discrepancies between prediction and observation, but also within the observational climatology. The climatic components have generally lower values compared to the predictions for both divergence and entropy and for both regions.

The energy skill score (ESS) is shown in Figure 5.14. The robustness has been estimated in the same way as for the CRPSS uncertainty. For both months, we see in general negative ESS values indicating that the reference forecast, here the observational climate, predicts better and robust compared to b1-LR for lead year 2-5. Over land, we find areas e.g. Europe/Eurasia, parts of Africa and North America, where the values are closer to zero or even positive and robust. In January, there are positively robust signals in the south and south-east of South Africa. Especially in the Atlantic, we can determine a structure parallel to the equator, where the northern subtropics show positive ESS values and the southern subtropics show negative ESS values. This structure can be associated with the trade winds and a slightly shifted position of the Intertropical Convergence Zone (ITCZ), which might not be represented correctly e.g. by the convection scheme.

Moemken et al. (2016) downscale wind speed and wind energy output for Central Europe and Germany and find negative MSESS (calculated after Goddard et al. (2013), which differ from the MSESS presented in Section 3.1.4) values, where the reference MSE is based on the uninitialized ensemble for winter and summer months. They state that for annual variables the skill increases but lasts only for short lead times (lead year 1-3). These findings are in concordance to our suggestions that verifying near surface variables such as the wind velocity in the boundary layer is difficult as the prediction system is still in an experimental stage (Stolzenberger et al., 2016). Additionally, it depends on the quality of the model, how well e.g. the cumulus convection scheme or other related variables are parametrized.

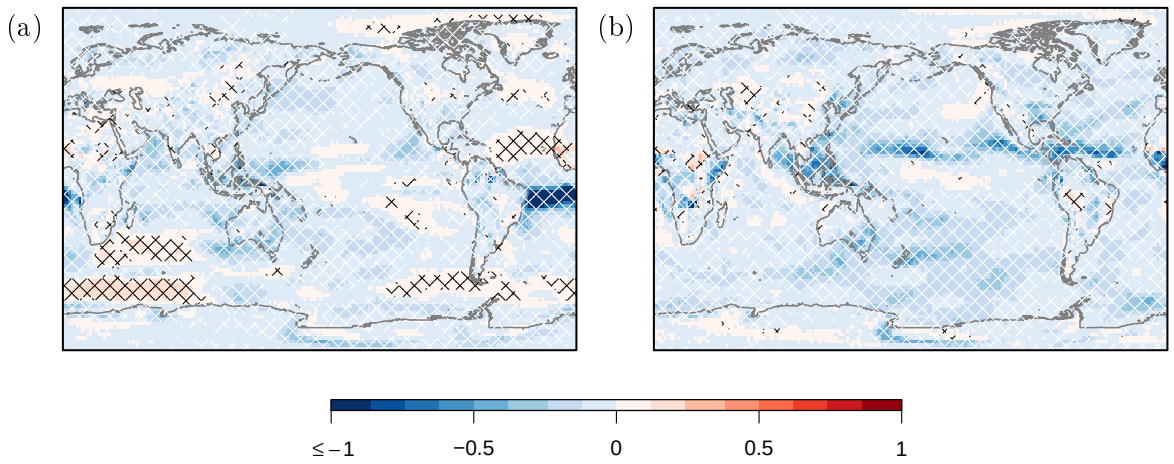

Figure 5.14: *Energy skill score for the u- and v-wind component for January (a) and July (b). Model basis is b1-LR for lead year 2-5 with climate as reference ES. Divergence (red contour line) and entropy (blue contour line) are shown with the resulting energyscore (gray shading). The skill of the b1-LR experiment is better than its reference if the ESS is positive (red shadings). White/black hatching means that the CRPSS is negatively/positively robust.*

However, this approach can be applied to wind velocities at higher altitudes. With this method, two-dimensional vector fields can be jointly evaluated keeping the physical basis including wind velocity and direction.

This approach can also be applied to e.g. down-welling and up-welling shortwave radiation at the surface or at the top of the atmosphere, in addition to look at budgets. Furthermore, the approach can be extended to combined thermodynamic variables such as dry static energy $s_d = c_pT + gz$ and moist static energy $s_m = c_pT + gz + L_vr$, where c_p is the specific heat capacity of air at constant pressure, T the absolute temperature, g the gravitational acceleration, z the height above some reference level, L_v the latent heat of vaporization and r the water vapor mixing ratio in the air[1].

5.3 Summary and Discussion

The MiKlip system is a decadal prediction system based on the MPI-ESM model output. Here, we analyze four baseline experiments, which mainly differ in the model resolution, the ocean initialization technique, and the number of ensemble members. For b1-LR/MR (10/5 members), T and S anomalies are used to initialize the ocean, whereas for the pr-sets (15 members each) full fields of T and S are used.

Evaluating T and Z in three dimensions (horizontal fields and latitude/longitude-height cross-sections), we find skillful and reliable areas especially at higher altitudes for all experiments. However, there is no skill and reliability in the subtropical mid-troposphere. This signal is visible for the b1 experiments as

[1] http://glossary.ametsoc.org/

well as for the two prototype sets. Stolzenberger et al. (2016) argue that processes such as the Walker or the Hadley circulation is not represented properly in the model dynamics. Moreover, we cannot determine big differences between lead year 2-5 and 7-10 when comparing the predictions to climate as reference (see Appendix A.1). The β-scores for lead year 1 at 200 hPa, however, show that pr-GECCO is not well calibrated in the subtropics. As b1-MR shows no additional benefits concerning the quality assessment and only includes 5 ensemble members, it is not used for further statistical analyses.

Potential sharpness exists (indicated by large ANOVA values) especially in the subtropics and mid-latitudes. The high ANOVA values in combination with the negative skill scores and the lack of reliability indicate that the ensemble members predict similarly but falsely. As the uninitialized runs show also high ANOVA values, the predictability arises rather from external forcing than from initializing the model (Stolzenberger et al., 2016).

For T in the ocean, we find reliable areas, which are presented as latitude-height cross section (by reliability classification) and for selected regions (by β-scores) such as the North Atlantic and the Pacific. Most areas are potentially useful with some small reliable clusters, which are supported by predominantly positive MSESS values. Pr-GECCO clearly outperforms pr-ORA and b1-LR in the tropical Pacific, whereas pr-GECCO fails in the deeper layers of the North Atlantic compared to pr-ORA and b1-LR.

Skill analyses for the atmosphere between 15°S and 15°N along the equator show that the baseline experiments hold strong similarities as the atmosphere is initialized exactly in the same way with the same datasets. For the ocean, however, we can determine strong differences especially between the pr-sets as the underlying observations for initializing the ocean are different. We find that pr-GECCO is more consistent with the observation GECCO2 in contrast to pr-ORA with ORAS4. This is one reason why merging the pr-sets to a 30-member ensemble in the ocean is not recommended. Therefore we use the CRPS for Gaussian mixture distributions. Especially in the tropical Pacific and deeper oceanic layers, positive and robust skill scores are found indicating that GECCO2 is better suitable as verifying observation than ORAS4. Thus, the choice of the dataset, which is used for initialization and for verification, is of central importance when interpreting the results.

Marotzke et al. (2016) find missing skill for the North Atlantic SSTs in the early lead years and increasing skill in the late lead years for the pr-sets, especially for pr-ORA. They state that this skill behavior may be connected with a model drift, which is problematic when initializing the model components. We cannot find skill for T in the North Atlantic (from surface to 1,500 m depth) for b1-LR and for the pr-sets (not shown here), which fit to the results of Marotzke et al. (2016). However, recent analyses of the latest seasonal version of the Met Office Decadal Prediction System (DePreSys3) show skill in the North Atlantic amongst other regions for the first two lead years. DePreSys3 uses full-field initialization for ocean and atmosphere but their hindcasts start in contrast to the MiKlip experiments each 1 November (Roberts et al., 2016; Dunstone et al., 2016), which is a known and proven initial-

ization date in seasonal climate predicting. This date is set for the latest MiKlip experiment Pre-Op (Pre-Operational), which is developed during MiKlip phase II.

Moreover, the number of ensemble members increased from the b1-LR experiment to the pr-sets, but it is still comparatively small (e.g. DePreSys3 includes 40 ensemble members). In our study, we cannot detect that improved skill in the pr-sets comes from 5 additional members. However, Scaife et al. (2014) and Sienz et al. (2016) state that the prediction skill increases if the number of ensemble members increase regardless of the computational time.

Finally, we have presented a way to jointly verify the zonal and meridional wind component by applying a bivariate CRPS (ES). For the 10 m wind and lead year 2-5, the ESS shows negative and robust skill scores for most parts. The temporal evolution of divergence and entropy in the Pacific and North Atlantic shows high variability especially for January. This application illustrates a test case for b1-LR and can be applied to other baseline experiments and upper pressure levels, which is likely to be more promising in terms of skill.

Part II

Statistical paleoclimate reconstructions

Chapter 6

Spatial reconstructions over Europe

In this chapter, a statistical concept for spatial paleoclimate reconstructions is presented. This concept includes probabilistic information of both a multi-model ensemble and observations. We apply the method to estimate winter and summer temperatures in Europe for the Mid-Holocene (6 ka BP).

6.1 Data basis

6.1.1 Paleoclimate simulations

PMIP3 has been described in Chapter 2. For this study, we use the experiments for the Mid-Holocene (6 ka BP) and select those models which firstly, are available at the CERA database[1], secondly, are CMIP5 member and thirdly, do not show any problems such as longterm drifts (Bothe et al., 2013). In all models, the atmospheric and oceanic components are coupled and two of them include the carbon cycle (see Table 6.1).

We extract the global model data to Europe (7.5°W-27.5°E, 37.5°N-70°N) and interpolate (bilinear interpolation) to a regular grid with a horizontal resolution of 2.5°. This coarse grid resolution is aimed at a balance between the different model resolutions. We obtain climatologies for winter and summer temperatures at 2 m by averaging the monthly values over the total simulation time.

6.1.2 Observational data

Climate Research Unit (CRU) data

Since the 1990s, the climate research unit (CRU) of the University of East Anglia in Norwich, UK, publishes global, gridded observational datasets at a high resolution. We use the time series CRU TS 3.22 for the period from 1901 to 2013 (Harris et al., 2014). This land-only dataset is created by interpolating station anomalies (1961-1990) to a regular grid ($0.5° \times 0.5°$ longitude/latitude). Variables such as mean temperature, precipitation and vapor pressure and also secondary variables such as potential evapotranspiration, which are derived from the primary ones, are available.

In Figure 6.1 the differences between the CRU climatic mean and the PMIP3 multi-model mean are shown. In Scandinavia, the winter temperatures of the 20th century are up to 5 K higher than for

[1] https://www.dkrz.de/daten-en/cera

Table 6.1: *Considered PMIP3 members[2] for the multi-model ensemble. First column represents the institution including the nation in brackets, second column represents the model name, third column represents the original atmospheric resolution (gridpoints in longitude times gridpoints in latitude times vertical levels), fourth column represents the original oceanic resolution and fifth column represents whether the model includes the carbon cycle.*

Participant		model	atmospheric grid	oceanic grid	carbon cycle
NCAR	(US)	CCSM4	288x192 x L26	320x384 x L60	no
CNRM/CERFACS	(FR)	CNRM-CM5	256x128 x L31	362x292 x L42	no
QCCCE/CSIRO	(AU)	CSIRO-Mk3-6-0	192x96 x L18	192x192 x L31	no
KNMI	(NL)	EC-Earth 2.2	320x160 x L62	362x292 x L42	no
Hadley Center	(UK)	HadGEM2-CC	192x145 x L38	360x216 x L40	yes
IPSL	(FR)	IPSL CM5ALR	96x96 x L39	182x149 x L31	yes
MPI	(DE)	ECHAM6/MPIOM	192x96 x L47	256x220 x L40	no
MRI	(JP)	MRI-CGCM3	320x160 x L48	364x368 x L51	no

6 ka BP whereas for south-east Europe, UK and the north of Spain, the winter temperature in the Mid-Holocene is up to 4 K above today's climate. For Europe especially in east and south-east Europe, today's summer temperatures are generally higher compared to 6 ka BP.

Recent vegetation data

To get a relation between botanical and climatic data, we need recent distribution maps of occurring taxa. Schölzel et al. (2002) developed a software to digitize maps from atlantes, which were created in the 20th century for Europe and Eurasia. These digitized maps have a regular horizontal resolution of 0.5°, which is equal to the resolution of the CRU-dataset. To date, a pool of more than 300 taxa distributions are available for Eurasia.

Paleobotanical data

Simonis et al. (2012) provide a list of pollen data from coring sites all over Europe, which were collected for different time slices (6, 8, 12 and 13 ka BP). The distribution of the coring sites for 6 ka BP is shown in Figure 6.2 and a detailed list is given in the Appendix B.2. The statistical concept presented in the following requires the occurring taxa to be statistically independent from each other at each coring site. Taxa with similar distributions and hence similar pdfs will be excluded. For statistical climate reconstructions, the pdfs are multiplied. In case that the pdfs are too similar, there will be no additional information only a reduction of the variance. The taxa selection is obtained by calculating the Mahalanobis distance, which measures the distance between two Gaussian distributions. If the

[2] https://wiki.lsce.ipsl.fr/pmip3/doku.php/pmip3:database:expected

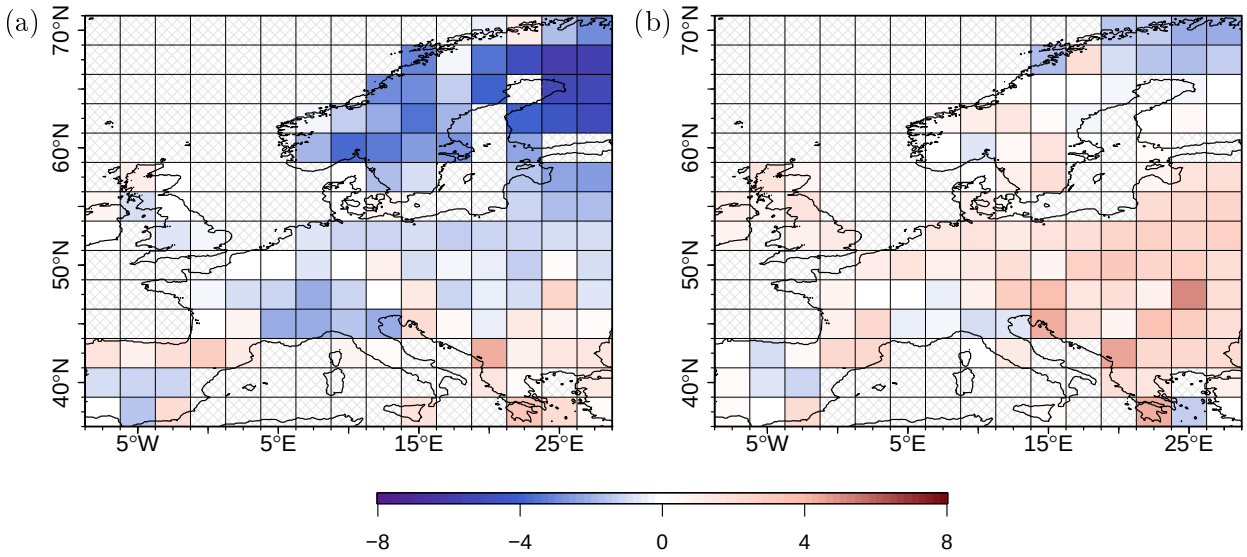

Figure 6.1: *Difference between CRU climatic mean (1901-1913) and PMIP3 multi-model mean (6 ka BP) for (a) winter and (b) summer 2 m temperatures in K.*

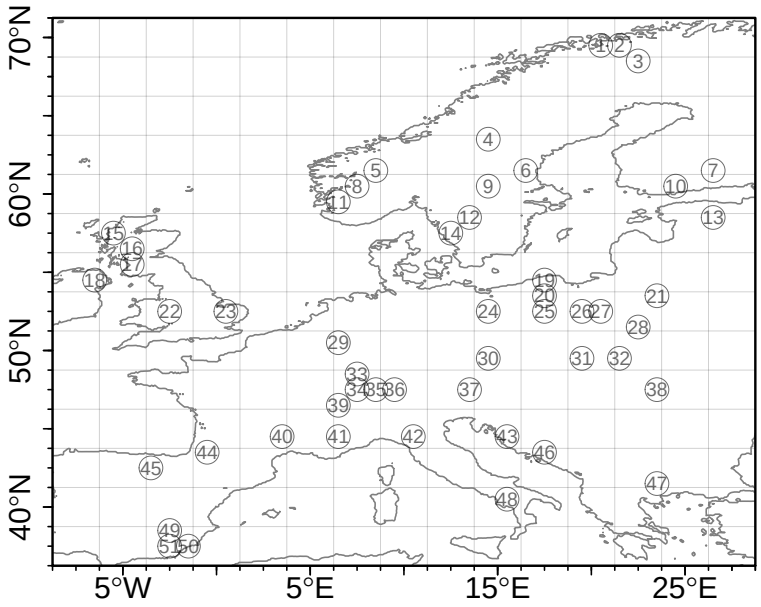

Figure 6.2: *Coring sites at 6 ka BP taken from Simonis et al. (2012). A detailed list is given in the Appendix B.2.*

Mahalanobis distance is smaller than 0.2, the taxon is excluded. Table B.1 contains the statistically chosen 59 taxa for the 51 coring sites.

6.2 Statistical approach

We compare statistical climate reconstructions based on pollen data with the PMIP3 multi-model ensemble in order to optimize these models. This method consists of three steps (see Figure 6.3).

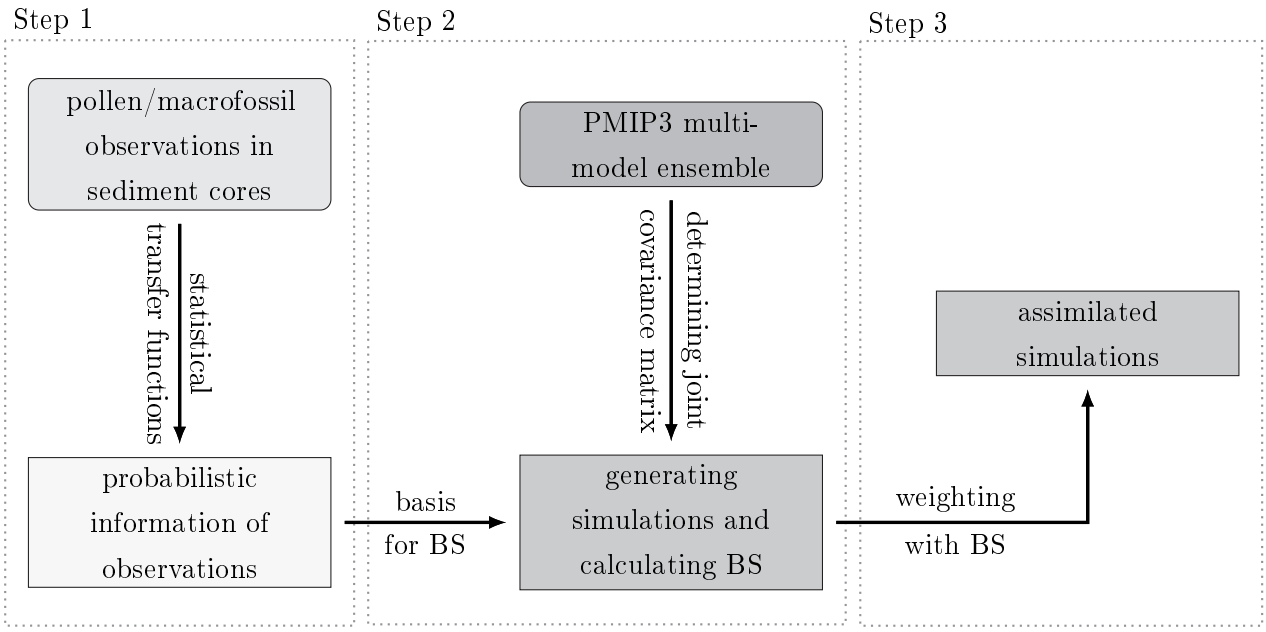

Figure 6.3: *Overview of reconstruction concept using probabilistic information of both proxies and model data. A detailed description can be found in the text.*

In the first step, the uncertainties of the observed pollen data are assessed by calculating statistical botanical climate transfer functions. In the second step, we build a PMIP3 multi-model ensemble and generate simulations from the multivariate normal distribution. In each generation step, we calculate the Brier score (and skill score) based on the probabilistic information of the observations. In the third step, the generated simulations are weighted with the corresponding Brier score to obtain simulations that are assimilated to observed pollen data.

6.2.1 Step 1: Calculating transfer functions

Botanical climate transfer functions describe the statistical relation between recent botanical and recent climatic data. For calculating the transfer functions, we use the pdf-method (Kühl et al., 2002) based on generalized linear models (GLMs). An overview of the GLM and the advantages of the GLM based pdf-method (Stolzenberger, 2011) is given in the following section.

Pdf-method The pdf-method (Kühl et al., 2002) is based on the mutual climatic range method by Grichuk (1969). The climate space (e.g. winter and summer temperatures) in which a certain taxon occurs is determined and illustrated as a closed area, which is called mutual climatic range. The overlapping area of the mutual climatic ranges, which occur at one location, represents the reconstructed climate state. As this graphical method has some disadvantages (overfitting due to sharp borders, equal probability for the occurrence of a taxon within the distributional area, etc.), the climate ranges are replaced with pdfs of e.g. Gaussian or Gamma distributions.

For the statistical method two assumptions are made: firstly, we assume that the vegetation only de-

pends on the climate parameters that are to be reconstructed. Interactions or competitions between the plants, which can lead to relocations, are not taken into account. Secondly, we assume that the climate conditions under which the taxon can grow nowadays did not change over the last millennia and hence the genetics of the plants are still the same.

Generalized linear model (GLM)

GLMs were firstly introduced by Nelder and Baker (1972). We briefly describe the GLM formalism and the parameter estimation after Fahrmeir and Tutz (1994).

The GLM is a generalization of the classical linear model which is, for ungrouped normal responses and deterministic covariates, defined by

$$\vec{y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}, \tag{6.1}$$

where $\vec{\beta}$ is a vector of unknown parameters of dimension p. The errors $\vec{\epsilon}$ are assumed to be independent and normally distributed, $\vec{\epsilon} \sim \mathcal{N}(0, \sigma^2)$. The design matrix $\mathbf{X}$ includes n observations of the p-dimensional vectors of covariates,

$$\mathbf{X} = (\vec{x_1}, ..., \vec{x_n})^{\mathrm{t}} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}. \tag{6.2}$$

We can transform the linear model to a GLM in a natural way. The observations $\vec{y}$ are independent and normally distributed, $\vec{y} \sim \mathcal{N}(\vec{\mu}, \sigma^2)$, with the expectation value $\vec{\mu} = E(\vec{y})$ given by the linear combination $\vec{\mu} = \mathbf{X}\vec{\beta}$. If the pairs $(y_i, \vec{x_i})$ are independent and identically distributed, the observations y_i given the covariates $\vec{x_i}$ are conditionally independent.

The GLM is defined by the following two assumptions:

1. The observations y_i are (conditionally) independent for given covariates x_i, and the distribution of y_i is part of a basic exponential family defined by

$$f(y_i|\theta_i, \phi, \omega_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi}\omega_i + c(y_i, \phi, \omega_i)\right), \tag{6.3}$$

   where

θ_i	are the so-called natural parameters,
$b(), c()$	are specific functions dependent on the type of the exponential family,
ϕ	is the scale or dispersion parameter, which is not dependent on i. (For simplification, we exclude overdispersion, i.e. $\phi = 1$.)
ω_i	are known weights with $\omega_i = 1$ for ungrouped data.

The expectation value and variance of the exponential family is given by

$$E(y_i) = \mu_i(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta_i} \tag{6.4}$$

$$var(y_i) = \sigma^2(\mu_i) = \frac{\phi}{\omega_i}\frac{\partial^2 b(\theta_i)}{\partial^2 \theta_i}. \tag{6.5}$$

2. The relation between the expectation $\vec{\mu}$ and the linear predictor $\vec{\eta} = \mathbf{X}\vec{\beta}$ is given by

$$\vec{\mu} = h(\vec{\eta}) = h(\mathbf{X}\vec{\beta}) \qquad \text{and} \qquad \vec{\eta} = g(\vec{\mu}), \tag{6.6}$$

where

$$h \quad \text{is the response function and} \tag{6.7}$$

$$g \quad \text{is the link function (inverse of } h). \tag{6.8}$$

The Bernoulli distribution

$$\mathcal{B}(1, \pi_i) = \pi_i^{k_i}(1-\pi_i)^{(1-k_i)}, \qquad \text{with} \quad k_i \in [0, 1], \tag{6.9}$$

where π_i is the probability for the occurrence of an event ($k_i = 1$), can be rewritten in terms of the exponential family formalism (see Equation 6.3) as,

$$f_{\mathcal{B}}(y_i|\theta_i, \phi, \omega_i) = \exp\left(\log\left(\pi_i^{k_i}(1-\pi_i)^{(1-k_i)}\right)\right) \tag{6.10}$$

$$= \exp\left(k_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)\right) \tag{6.11}$$

and the corresponding components of the exponential family can be derived: the natural parameter is defined by

$$\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) \tag{6.12}$$

and the specific functions are defined by

$$b(\theta_i) = \log(1-\pi_i) = -\log(1-\exp(\theta_i)) \tag{6.13}$$

$$c(y_i, \phi, \omega_i) = 0. \tag{6.14}$$

The scale parameter and weights equal one, $\phi = 1$ and $\omega_i = 1$ for all $i = 1, ..., n$.

Estimating unknown parameters The unknown parameter vector $\vec{\beta}$ can be obtained by using maximum likelihood estimation. The log-likelihood of the observation y_i can be derived from Equation 6.3 and is given by

$$l = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi}\omega_i = \sum_{i=1}^{n} l_i, \tag{6.15}$$

where $c(.)$ is not considered as it does not depend on θ_i.

The first derivative of l is the score function

$$s(\vec{\beta}) = \sum_{i=1}^{n} s_i(\vec{\beta}) = \frac{\partial l}{\partial \vec{\beta}} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\eta_i}{\vec{\beta}}. \tag{6.16}$$

The single partial derivatives are:

$$\frac{\partial l_i}{\partial \theta_i} = \frac{\omega_i}{\phi} \left(y_i - \frac{\partial b(\theta_i)}{\partial \theta_i} \right); \tag{6.17}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i} \right)^{-1} = \frac{\partial^2 b(\theta_i)}{\partial^2 \theta_i} = \frac{1}{\upsilon(\mu_i)}; \tag{6.18}$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^{-1} = \frac{\partial g(\mu_i)}{\partial \mu_i} = \frac{\partial h(\eta_i)}{\partial \eta_i}; \tag{6.19}$$

$$\frac{\partial \eta_i}{\partial \vec{\beta}} = \mathbf{X_i}^{\mathrm{t}}; \tag{6.20}$$

The individual score function is then defined by

$$s_i(\vec{\beta}) = (y_i - \mu_i)\sigma^{-2} D_i(\vec{\beta}) \mathbf{X_i}^{\mathrm{t}} \tag{6.21}$$

$$= (y_i - \mu_i) w_i(\vec{\beta}) \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^{-1} \mathbf{X_i}^{\mathrm{t}}, \tag{6.22}$$

where $\sigma_i^2(\vec{\beta}) = \upsilon(\eta_i)\frac{\phi}{\omega_i}$. The first derivative of the response function $h(\eta_i)$ evaluated at η_i is denoted by $D_i(\vec{\beta}) = \frac{\partial h(\eta_i)}{\partial \eta_i}$. Additionally, the weight functions $w_i(\vec{\beta}) = D_i^2(\vec{\beta})\sigma_i^{-2}(\vec{\beta})$ are introduced.

The expected Fisher information matrix is determined by

$$F(\vec{\beta}) = cov\left(s(\vec{\beta})\right) = \sum_{i=1}^{n} F_i(\vec{\beta}), \tag{6.23}$$

where $F_i(\vec{\beta}) = \mathbf{X_i}^{\mathrm{t}} \mathbf{X_i} w_i(\vec{\beta})$.
In matrix notation and with the diagonal elements of

$$\Sigma(\vec{\beta}) = diag(\sigma_i^2(\vec{\beta})), \tag{6.24}$$

$$W(\vec{\beta}) = diag(w_i(\vec{\beta})) \qquad \text{and} \tag{6.25}$$

$$D(\vec{\beta}) = diag(D_i(\vec{\beta})), \tag{6.26}$$

the score function and the Fisher information matrix are defined by

$$s(\vec{\beta}) = \mathbf{X_i}^{\mathrm{t}} D(\vec{\beta}) \Sigma^{-1}(\vec{\beta})(\vec{y} - \vec{\mu}(\vec{\beta})), \tag{6.27}$$

$$F(\vec{\beta}) = \mathbf{X_i}^{\mathrm{t}} W(\vec{\beta}) \mathbf{X_i}. \tag{6.28}$$

The maximum likelihood estimators for $\vec{\beta}$ minimize the score function

$$s(\hat{\vec{\beta}}) = 0. \tag{6.29}$$

The log-likelihood $l(\vec{\beta})$ is concave for many models which means that local and global maxima coincide. As the likelihood equations are generally nonlinear, they have to be solved iteratively, e.g. by using the Fisher scoring scheme

$$\hat{\vec{\beta}}_{k+1} = \hat{\vec{\beta}}_k + F^{-1}(\hat{\vec{\beta}}_k)s(\hat{\vec{\beta}}_k), \qquad k = 0, 1, 2, ..., \tag{6.30}$$

with $\hat{\vec{\beta}}_0$ as initial estimate. The iterations continue until a certain threshold ($\epsilon > 0$) is reached

$$\frac{\parallel \hat{\vec{\beta}}_{k+1} - \hat{\vec{\beta}}_k \parallel}{\parallel \hat{\vec{\beta}}_k \parallel} \leq \epsilon. \tag{6.31}$$

If the observation vector is given by

$$\tilde{\vec{y}}(\vec{\beta}) = (\tilde{y_1}(\vec{\beta}), ..., \tilde{y_n}(\vec{\beta}))^{\mathrm{t}}, \tag{6.32}$$

$$\tilde{y_i}(\vec{\beta}) = \mathbf{X_i}\vec{\beta} + D_i^{-1}(\beta)\left(y_i - \mu_i(\vec{\beta})\right), \tag{6.33}$$

the Fisher scoring iterations can be expressed by

$$\hat{\vec{\beta}}_{k+1} = \hat{\vec{\beta}}_k + (\mathbf{X}^{\mathrm{t}}W(\hat{\vec{\beta}}_k)\mathbf{X})^{-1}s(\hat{\vec{\beta}}_k) \tag{6.34}$$

$$= (\mathbf{X}^{\mathrm{t}}W(\hat{\vec{\beta}}_k)\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}W(\hat{\vec{\beta}}_k)\tilde{\vec{y}}(\hat{\vec{\beta}}_k), \tag{6.35}$$

which can be interpreted as iteratively weighted least squares.

GLM-probability With the estimated parameters $\hat{\vec{\beta}}$, we can derive the GLM probability from Equation 6.12 given the covariates $\mathbf{X}$

$$\pi = \frac{\exp\left(\mathbf{X}\hat{\vec{\beta}}\right)}{1 + \exp\left(\mathbf{X}\hat{\vec{\beta}}\right)}. \tag{6.36}$$

In our application, the GLM probability π denotes the probability for the occurrence of a taxon given the winter or summer temperature. This relation is based on recent data. The covariates $\mathbf{X}$ represent the winter or summer temperatures, which are normalized concerning the climatic mean. The parameters $\vec{\beta}$ are estimated for the GLM-probability and used to rank the simulations, which is described in Step 2 (see Section 6.2.2).

Advantage of the GLM based pdf-method By using the GLM (Equation 6.36), it is possible to directly determine the probability for the occurrence of a taxon (V) given a certain climate (C). In general, we can express this probability by using the Bayes theorem,

$$p(V = 1|C) = \frac{p(C|V = 1)p(V = 1)}{p(C)} \tag{6.37}$$

where $p(C)$ is the marginal distribution, $p(C) = p(C|V = 1)p(V = 1) + p(C|V = 0)p(V = 0)$, the sum over the probabilities of a climate state for all possible conditions. It is not trivial to calculate $p(C)$ as for determining $p(C|V = 0)$ the absence information of a taxon is missing (Stolzenberger, 2011). This can be avoided by using the GLM as it includes the presence/absence information by fitting a Bernoulli distribution. Additionally, the outcome of the GLM is a likelihood and not a density, which is beneficial for further processing.

6.2.2 Step 2: Generating simulations from joint covariance matrix

The PMIP3 multi-model ensemble consists of eight members. From this multivariate Gaussian distribution we sample simulations. Using parametric sampling includes interpolating the eight existing realizations (ensemble members) according to the covariance structure and, thus, generating new simulations. Each ensemble member is treated equally. The method we use here is the Gaussian ensemble kernel dressing (EKD, see Sections 3.1.2 and 5.2.1). The sum of the (equally weighted) Gaussian distributions produce the mixture model density,

$$f_{\mathrm{EKD}}(\vec{x}|\vec{x}_1,...,\vec{x}_m) = \frac{1}{m}\sum_{i=1}^{m} f_{\vec{x}_i}(\vec{x}) \tag{6.38}$$

$$= \frac{1}{m}\sum_{i=1}^{m} \exp\left(-\frac{1}{2}(\vec{x}-\vec{x}_i)^{\mathrm{t}}\Sigma_\epsilon^{-1}(\vec{x}-\vec{x}_i)\right), \tag{6.39}$$

where m is the number of ensemble members, $\vec{x}_i$ the model realization of dimension q and Σ_ϵ is the dressing covariance matrix. The estimation of the joint covariance matrix for f_{EKD} is described in the following subsection.

The Brier Score (see Section 3.1.1) is calculated for each generation step by using the GLM-probability (given the generated temperature) for each coring site and for each occurring taxon. To rank the Brier score, we look at the Brier skill score. As reference Brier score, we use the GLM-probability without the influence of the covariates, i.e. the intercept: $\vec{x}\vec{\beta} = \vec{\beta}_0$. If the coefficients $\vec{\beta}_{1,2}$ are zero, the GLM-probability π (Equation 6.36) is constant, which means that the probability for the occurrence of a taxon does not depend on the given temperatures. If the Brier skill score is positive/negative, the inclusion of the probabilistic observational information has a qualitatively positive/negative effect.

Estimating the dressing covariance matrix

A detailed description of multivariate kernel dressing and estimating the covariance matrix is given in Schölzel and Hense (2011).

The unknown covariance matrix Σ_ϵ (Equation 6.39) is estimated from the multi-model ensemble $\vec{x}_i$. The estimator for Σ_ϵ is expressed by Σ_D.

The estimator of the raw covariance matrix is given by

$$\hat{\Sigma}_{\mathrm{raw}} = \frac{1}{2N_{tot}}\sum_{i=1}^{m}\sum_{i=1}^{m}(\vec{x}_i-\vec{x}_j)(\vec{x}_i-\vec{x}_j)^{\mathrm{t}}, \tag{6.40}$$

where $N_{tot} = \frac{1}{2}m(m-1)$ is the number of possible combinations of the m ensemble members. As $\vec{x} = \vec{x}_i + \vec{\epsilon}_i$, where $\vec{x}$ is the "true" but unknown state vector and $\vec{\epsilon}_i$ is the internal noise, the differences between the single model realizations can be understood as a prewhitening filter to remove the true signal $\vec{x}$ (Röpnack et al., 2013). If we assume that the ensemble is unbiased and the ensemble members

are indistinguishable, the estimators for the dressing covariance matrix and for the raw covariance matrix are connected through

$$\Sigma_D = h_{opt} \cdot \hat{\Sigma}_{\text{raw}}, \tag{6.41}$$

where $h_{opt} = \left(\frac{4}{m(q+2)}\right)^{\frac{1}{q+4}}$ is the so-called Silverman's factor with q representing the vector dimension. Silverman (1986) introduce the factor h_{opt} in order to find the optimal bandwidth.

For equally treated ensemble realizations, the pdf multivariate Gaussian ensemble kernel dressing can be expressed by

$$f_{\text{EKD}} = \frac{1}{m\sqrt{(2\pi)^q \det \Sigma_D}} \sum_{i=1}^{m} \exp\left(-\frac{1}{2\pi}(x - x_i)^{\mathrm{T}} \Sigma_D^{-1} (x - x_i)\right), \tag{6.42}$$

which can be multi-modal if the spread between the ensemble members is larger than the spread of the noise.

Graphical lasso

Estimating the covariance and inverse covariance matrix is challenging when the number of ensemble members is clearly smaller than the dimension, $m \ll q$, as the ordinary maximum likelihood estimate does not exist. Even if the number of ensemble members is equal or larger than the dimension, $m \geq q$, the maximum likelihood estimate can be distorted (Mazumder and Hastie, 2012). The graphical lasso (glasso) algorithm is used to estimate a covariance matrix assuming that its inverse covariance matrix is sparse (Friedman et al., 2008). The approach is to maximize the penalized log-likelihood

$$\log\det\Theta - tr(\Sigma_D\Theta) - \rho_{gl}||\Theta||_1, \tag{6.43}$$

where Θ is the precision or concentration matrix, i.e. the inverse covariance matrix $\Theta = \Sigma_{gl}^{-1}$. The empirical covariance matrix is represented by Σ_D an tr denotes the trace of a matrix. The last term (penalty term) includes the non-negative so-called regularization parameter ρ_{gl} and the L_1 norm of the precision matrix $||\Theta||_1 = \Sigma_{ij}\theta_{ij}$, which is the sum of the absolute values of the off-diagonal coefficients θ_{ij}. The inverse covariance matrix does not change if $\rho_{gl} = 0$, but it will get more sparse with increasing ρ_{gl}. A detailed description of the glasso algorithm including the used block-coordinate method, is given e.g. in Friedman et al. (2008), Mazumder and Hastie (2012) amongst others. One advantage of glasso is the fast and easy computation with the R-package `glasso`.[3]

To assess the influence of ρ_{gl}, Weinert (2015) introduce a threshold value α_{gl}, which measures how much of the covariance remains after modification with glasso,

$$||\Sigma_{gl}|| = \alpha_{gl}||\Sigma_D||, \tag{6.44}$$

with $||\Sigma_*|| = \sum_{i\neq j} |\Sigma_*|$. This implies the larger ρ_{gl} the smaller α_{gl} or if $\alpha_{gl} = 1$ then $\rho_{gl} = 0$.

[3] https://cran.r-project.org/web/packages/glasso/index.html

Marginal and partial correlations[4]

Marginal correlations can be derived directly from the covariance matrix,

$$\rho_{marg} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}, \tag{6.45}$$

where σ_{ij} are the elements of the covariance matrix Σ. The marginal correlations describe the linear relation between x_i and x_j without including any other vector elements.

Partial correlations can be derived directly from the precision matrix Θ with its components θ_{ij}. The diagonal elements of Θ are equal to the inverted partial variances

$$\theta_{ii} = \frac{1}{var(x_i|x_{\setminus i})}, \tag{6.46}$$

where $x_{\setminus i}$ is the vector x without the component x_i. The diagonal element θ_{ii} is, hence, the partial variance of x_i considering the linear effect of the vector elements $x_{\setminus i}$. The non-diagonal element θ_{ij} is the partial variance between x_i and x_j after eliminating the linear relation of the other vector components and can be expressed by

$$\theta_{ij} = -\sqrt{\theta_{ii}\theta_{jj}}\rho(x_i, x_j|x_{\setminus ij}) \quad \text{for} \quad i \neq j, \tag{6.47}$$

where θ_{ii} and θ_{jj} are the diagonal elements of the precision matrix, and $\rho(x_i, x_j|x_{\setminus ij})$ is the partial correlation between the vectors x_i and x_j without including the components x_i and x_j, denoted by $x_{\setminus ij}$. The partial correlations

$$\rho_{part} = \rho(x_i, x_j|x_{\setminus ij}) = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}, \tag{6.48}$$

can thus be obtained from the precision matrix. If x_i and x_j are partially uncorrelated, the corresponding entry θ_{ij} in the precision matrix equals zero.

6.2.3 Step 3: Weighting simulations

There are several methods to obtain reconstructed fields of temperatures including the probabilistic information of the PMIP3 multi-model ensemble and the pollen observations. One option is to weight each generated simulation X_i with the corresponding Brier score and then, sum it up over all N simulations,

$$x_{\text{assim}} = \sum_{i=1}^{N} w_i \cdot X_i, \tag{6.49}$$

where w_i with $\sum_i^N w_i = 1$ is the inverse Brier score (as the Brier score is negatively oriented). For each gridpoint, we obtain the value x_{assim}, which is assimilated to the observed pollen.

6.3 Application to the Mid-Holocene

In this section, we apply the three-step reconstruction concept from Section 6.2 to winter and summer temperatures for the Mid-Holocene.

[4] https://www.elab.moodle.elearning.lmu.de/course/view.php?id=820

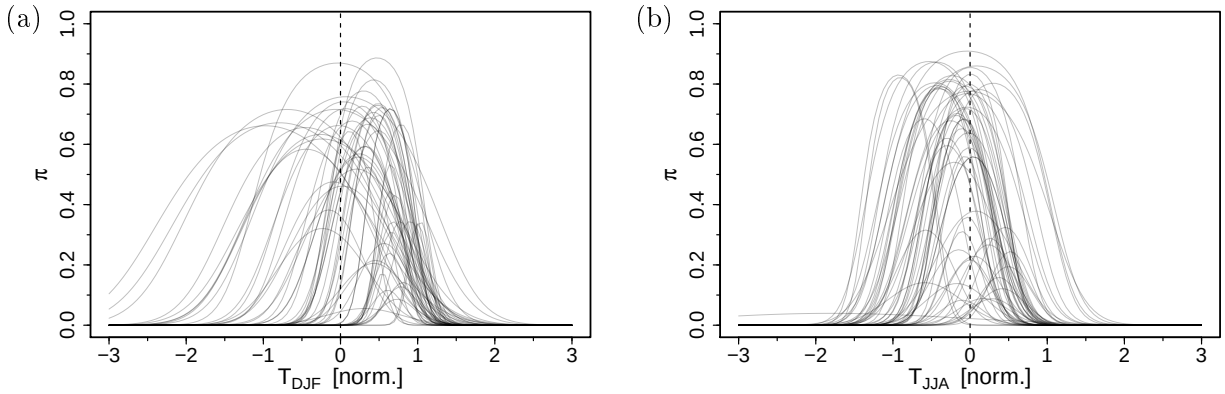

Figure 6.4: *GLM-probabilities for 59 occurring taxa for normalized winter (a) and summer (b) temperatures.*[5]

Step 1: Calculating transfer functions

We first calculate the transfer functions for each occurring taxon dependent on the normalized winter and summer temperatures (see Figure 6.4). For winter, the transfer functions are wider for some taxa. Moreover, the GLM-probabilities for winter are not as centered around zero as for summer, e.g. for *Salix pentandra, Rubus idaeus* and *Juniperus communis*, which means that these taxa occur when it is colder than the average. These taxa mainly appear in Central and northern Europe and Siberia mainly at higher altitudes. They are robust to cold winter temperatures but their growth is restricted when it is too warm in summer.

For one occurring taxon *Hippophae rhamnoides* (sea buckthorn), the GLM-probability, especially in summer, is very flat with its maximum at around -2 (Figure 6.4b) indicating that i) it is robust to cold summer temperatures and ii) it has only a small distribution area. *Hippophae rhamnoides* appears mainly in mountainous regions such as the Alps or the Himalayas. Especially in mountainous areas, the resolution of 0.5° (approx. 50 km) is too coarse to represent microclimatic conditions within a gridbox and under which a plant can occur. Strongly varying elevation can influence insolation, precipitation and wind at the surface. Thus, for estimating transfer functions, gridboxes are excluded, where the difference between mean and minimal height is larger than 400 m (Kühl et al., 2002). These settings are applied by using the topographical dataset from the National Geophysical Data Centre (NGDC), which is part of NOAA (National Oceanic and Atmospheric Administration).

At this point, we prepare the basis for the reference Brier score, which is needed for Step 2 and 3. Therefore, we take the GLMs without temperature influence, $\pi(\beta_0)$, i.e. only β_0 is used in the exponential terms (see Equation 6.36). In Figure 6.5, these probabilities are shown for all occurring taxa. There are some taxa where $\pi(\beta_0)$ equals zero for winter temperatures. In these cases, the estimated β_1 and β_2 show large amplitudes compared to other taxa. Such taxa occur predominantly in

[5]The abbreviations DJF (December, January, February) and JJA (June, July, August) are used in the following graphics instead of winter and summer.

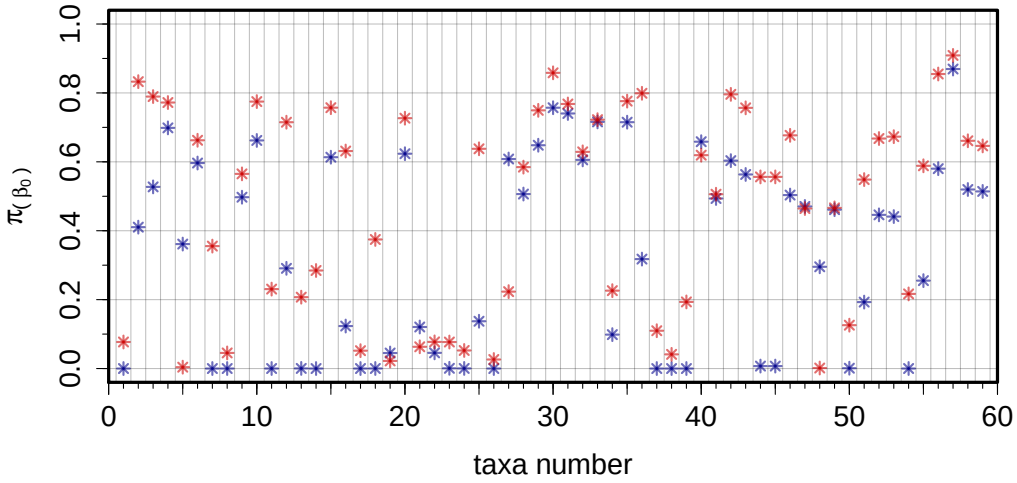

Figure 6.5: *GLM-probability without temperature influence,* $\pi(\beta_0) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$*, for winter (blue dots) and summer (red dots) for 59 occurring taxa.*

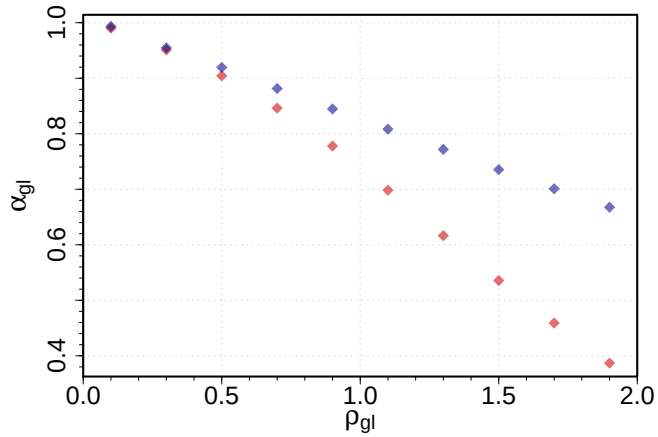

Figure 6.6: *Relation between* ρ_{gl} *and* α_{gl} *for winter (blue squares) and summer (red squares).*

southern Europe and are sensitive to cold winter temperatures, e.g. *Najas marina* or *Quercus pubescens*, which appear only at one coring site.

Step 2: Generating simulations

We estimate the covariance matrix for the PMIP3 multi-model ensemble using glasso (see Section 6.2.2) for winter and summer temperatures. Figure 6.6 shows the relation between the regularization parameter ρ_{gl} and α_{gl} for winter and summer temperatures. With increasing ρ_{gl}, the covariance matrix for summer looses its structure earlier than for winter. The relation for winter is nearly linear whereas for summer, it follows a parabolic curve. The following results are based on a small regularization parameter $\rho_{gl} = 0.1$ as firstly, we want as much information as possible to remain in the covariance matrix and secondly, α_{gl} is equal for both seasons.

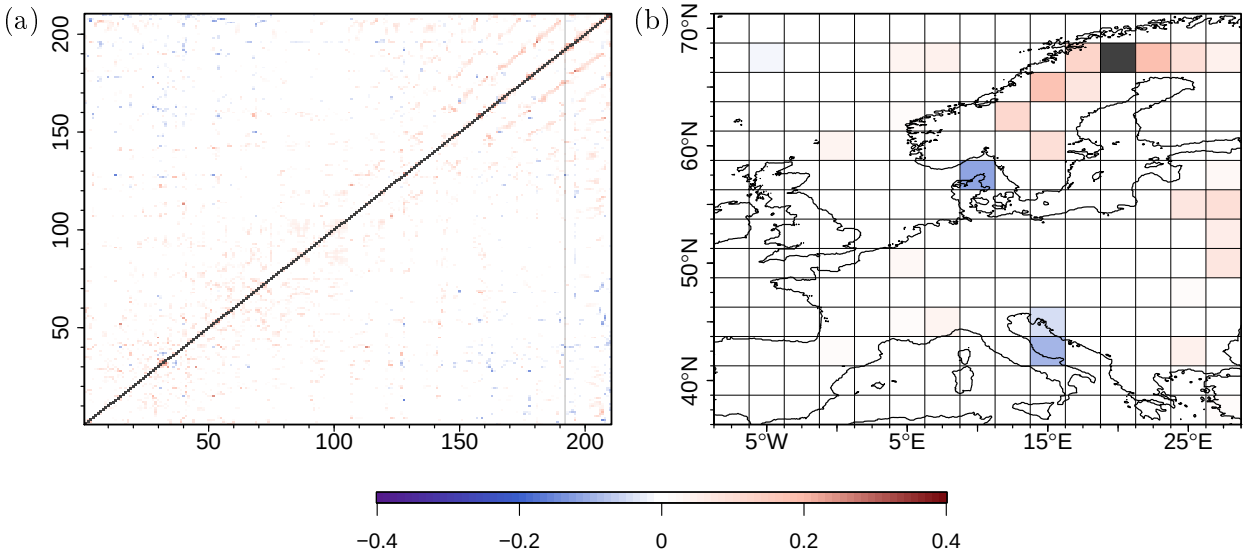

Figure 6.7: *Partial correlations (ρ_{part}) for $\rho_{gl} = 0.1$. (a) shows ρ_{part} for a $q \times q$ matrix and (b) shows the map of the cross section denoted by the gray line in (a) at 192 both for winter. Black colour on the diagonal (a) and at one grid point (b) represent the value 1 (initial point).*

The partial correlations (see Figure 6.7) exemplarily give an impression of the covariance structure for a low regularization parameter ($\rho_{gl} = 0.1$). The dimension of the partial correlation matrix equals the number of gridpoints ($n = 210$) in the examined area. Around some main diagonal elements we find small blocks (three to five gridpoints) of positive partial correlations. Positive partial correlations are also noticed on secondary diagonals. The values increase for the northern European area when $n \geq 160$. Figure 6.7b shows the spatial map of partial correlations with gridpoint 192 (black square). There are higher correlation values (approx. 0.3) for gridpoints surrounding the initial point especially in zonal direction. There are also positive correlations in the east of the considered area (especially over land) but there also a few gridpoints where the correlations are negative (highest amplitudes over sea). With increasing ρ_{gl}, the structure of the partial correlations remains but the correlation values decrease (not shown). Guillot et al. (2015) find similar structures when applying glasso to the precision matrix of a near-surface temperature field. Depending on the initial point, geophysical structures are detected such as anisotropic climate features when the initial point is located in the ocean (e.g. Atlantic structures are related to the subtropical gyre circulation).

With the estimated precision matrix, we can generate temperature simulations by varying the expectation value depending on each ensemble member. Figure 6.8 presents boxplots of the generated normalized winter temperatures ($n = 10,000$) and the corresponding Brier scores and Brier skill scores at 51 coring sites. In the upper panel, the range of the temperatures shows similarities except for the northern stations (blue boxplots), where the interquartile range and the upper and lower inner fence are larger. This means that the differences between the predictions of the PMIP3 ensemble members are generally small except for the northern coring sites. The middle panel shows the Brier scores, which are determined from the GLM-probabilities given the generated temperatures together with the

presence/absence information of all taxa at each coring site. The red points denote the reference Brier scores, which are independent of the generated temperatures. The Brier skill scores are presented in the lower panel. Some coring sites, e.g. 19-21 and 24-26, have predominantly negative Brier skill scores. The Spanish sites (49-51) show the best Brier skill scores compared to the other sites. The differences between the Spanish stations are minimal as similar pollen were found (without taking into account the occurring frequencies). Averaging over all coring sites, the median and the lower quartile are positive. For winter temperatures, there is a slight benefit when including the probabilistic information of the proxies.

The results for the generated summer temperatures are shown in the upper panel of Figure 6.9. The ranges of the generated summer temperatures are generally wider compared to the winter temperatures. The Brier scores in the middle panel, thus, have also wider ranges. The medians of the Brier skill scores (lower panel) are positive for all coring sites. Again, the Spanish coring sites in the south (49-51) show positive Brier skill scores with a few negative outliers. The average over all coring sites shows a Brier skill score with a median slightly above 0.2. In summer, it can be clearly seen that the fully calculated GLM-probabilities predict better than the GLM-probabilities excluding the temperature influence.

Similar to the verification of the MiKlip decadal prediction system (see Chapter 5), we can directly verify the PMIP3 models. However, for the paleoclimate application, the uncertainty information of the observations is additionally taken into account by using the GLM.

How consistent is the input of each ensemble member? In Figure 6.10 the Brier scores are presented as function of the ensemble members for both seasons. The eight boxplots represent the BS averaged over all coring sites for each ensemble member. They are randomly sorted as this study will not point out which model outperforms. The dashed line denotes the reference BS. For winter, the ranges of the boxplots are similar, some include outliers but, judging by eye, the differences between the boxplots are small. Boxplots, which have the lowest BS in winter, do not necessarily show low BS for summer and vice versa. The interquartile distances of each member do not show big differences. For both seasons, there are indeed no model particular extremes in both directions which means that we neither need to exclude one ensemble member nor include a weighting function for post-processing.

Step 3: Weighting the simulations

We obtain assimilated temperature fields by using the Brier scores as weights for the generated temperatures and fulfilling the assumption that summing up over all weights equals one. "Assimilated" in this context means that we use the BS, which include the GLM-probability and thus the proxy uncertainty, to optimize the PMIP3 model data. For winter (Figure 6.11a), the assimilated temperatures are lower in the north-eastern part of Europe and higher in the southern part (up to 0.4 K) compared to the PMIP3 ensemble mean. Deviations from the multi-model mean are smallest over land at latitudes

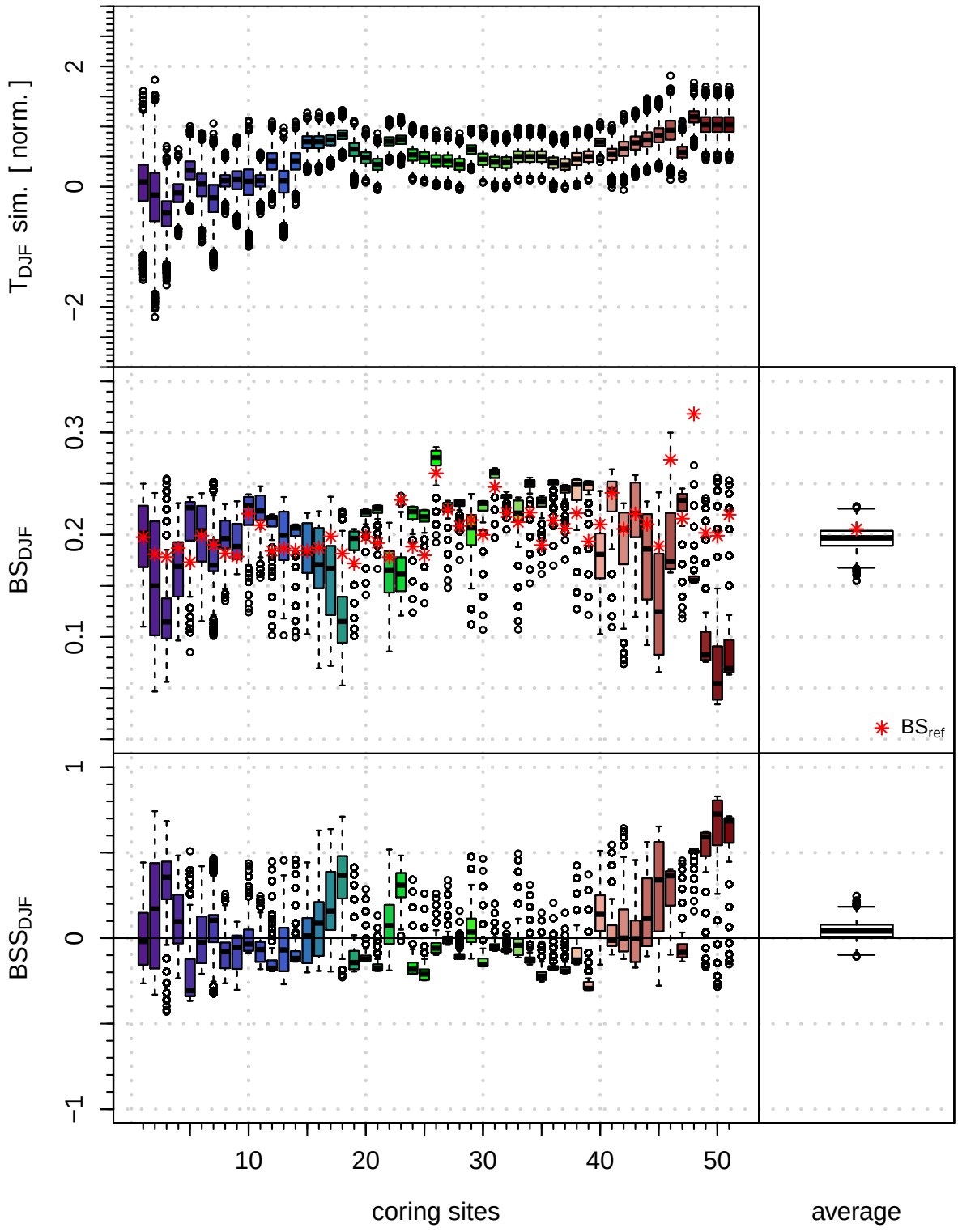

Figure 6.8: *Boxplots of the generated winter temperatures (upper panel), corresponding Brier scores (BS) with the reference BS denoted by red stars (middle panel) and Brier skill scores (BSS, lower panel) dependent on the 51 coring sites. The averages over all coring sites are presented next to BS and BSS.*

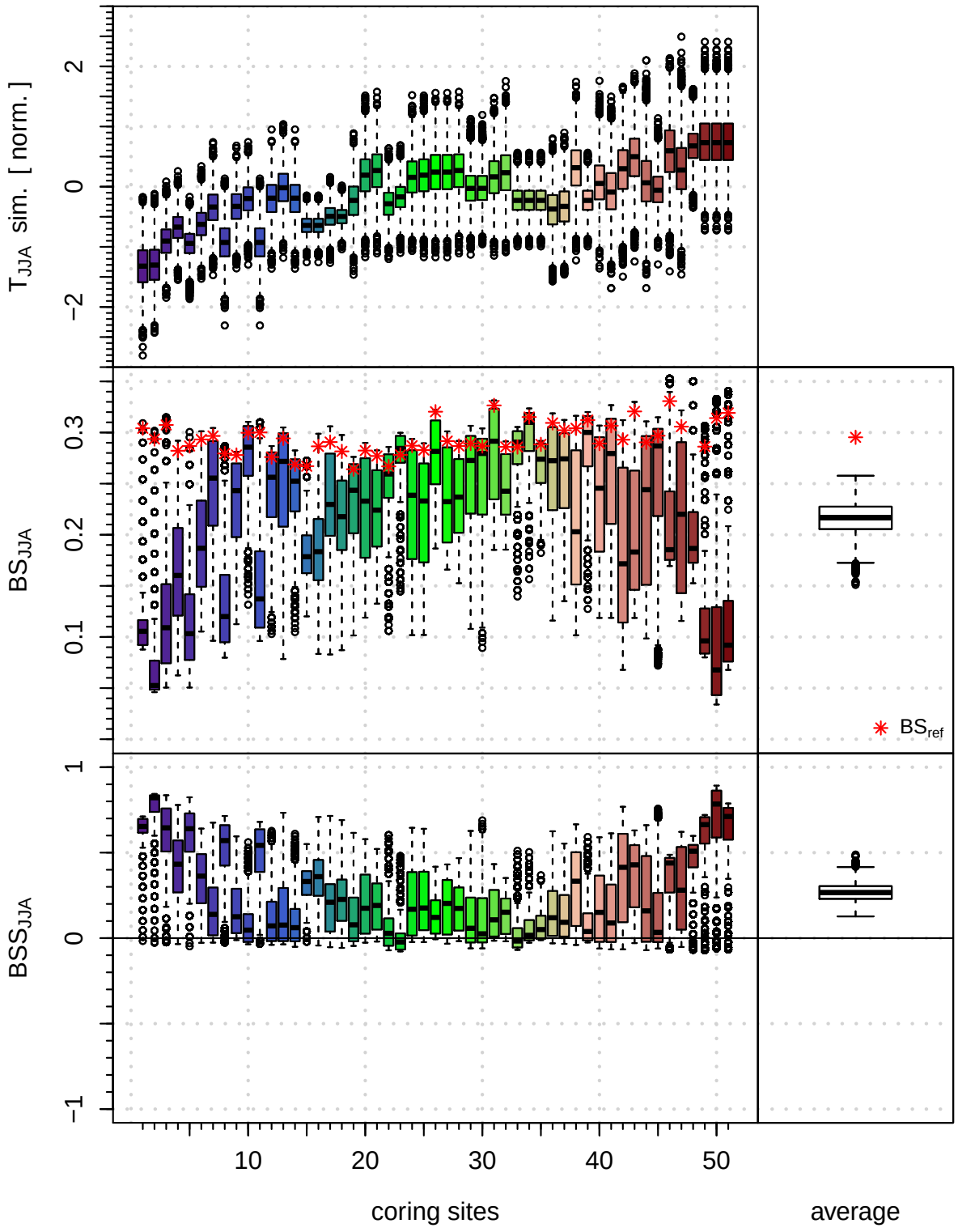

Figure 6.9: *Same as in Figure 6.8, but for summer temperatures.*

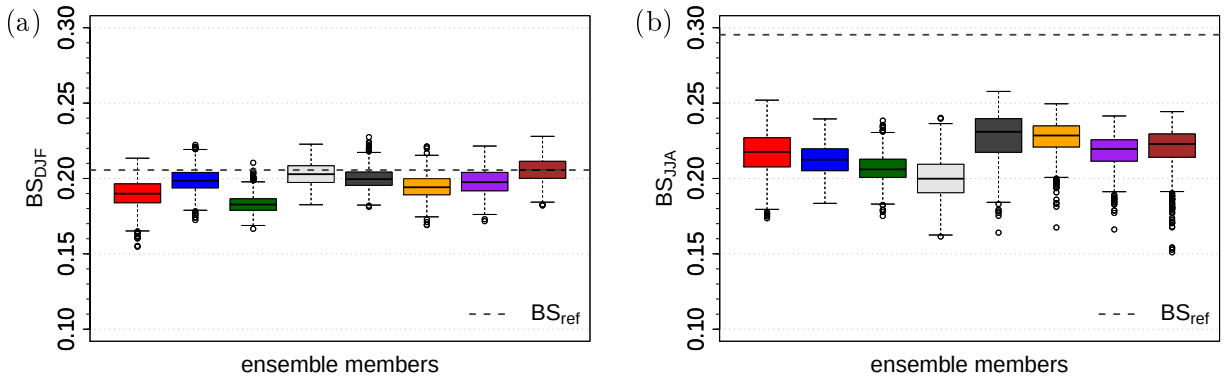

Figure 6.10: *Brier score (BS) in dependence on the ensemble members, averaged over all coring sites, for winter (a) and summer (b). The boxplots are represented in a random order. The dashed line denotes the reference BS, which is independent of the models.*

between 50° and 52.5°N.

As the BSS medians of more than 50% of the coring sites are clearly negative (see Figure 6.8), the assimilated winter temperatures have to be treated with care, especially when cold winter temperatures are expected (e.g. north and north-east Europe). If there is more than one coring site in one gridbox, it is likely that these coring sites have negative BSS (e.g. 33-35), but this is not generally valid. For instance, the BSS of the Spanish sites (49-51) are positive although they are closely spaced. Another example is coring site 47 (south-west Bulgaria), which is located more than 2,000 m above sea level and where cold temperatures in winter likely occur. Another explanation for the negative BSS at coring site 47 can be due to the pollen composition or uncertainties in the pollen measurements: pollen transports from lower altitudes to mountainous regions can lead to biases in the reconstructions (Simonis et al., 2012). For more than 10% of the coring sites, the medians of the BSS are zero, indicating that the reference BS predicts as good as the BS based on the fully estimated GLM. Positive BSS occur at coring sites where mild winter temperatures are expected, e.g. south England and the Mediterranean area. At these coring sites, positive anomalies (assimilated field compared to the multi-model mean) can be observed.

For summer (Figure 6.11b), the assimilated temperatures are generally higher compared to the PMIP3 model mean, especially over land. Over sea, the differences are still positive except for the Baltic Sea and the German North Sea coast. For the north-western part of the considered region, the differences between assimilated fields and the PMIP3 ensemble mean increase again.

Gebhardt et al. (2008) state that in their test case proxy data can apparently detect warm monthly temperatures better than cold monthly temperatures, which is in concordance with our findings. However, for building the transfer functions, Gebhardt et al. (2008) use Gaussian distributions, which only take into account the presence information of a taxon. In contrast, the GLM used in this study can

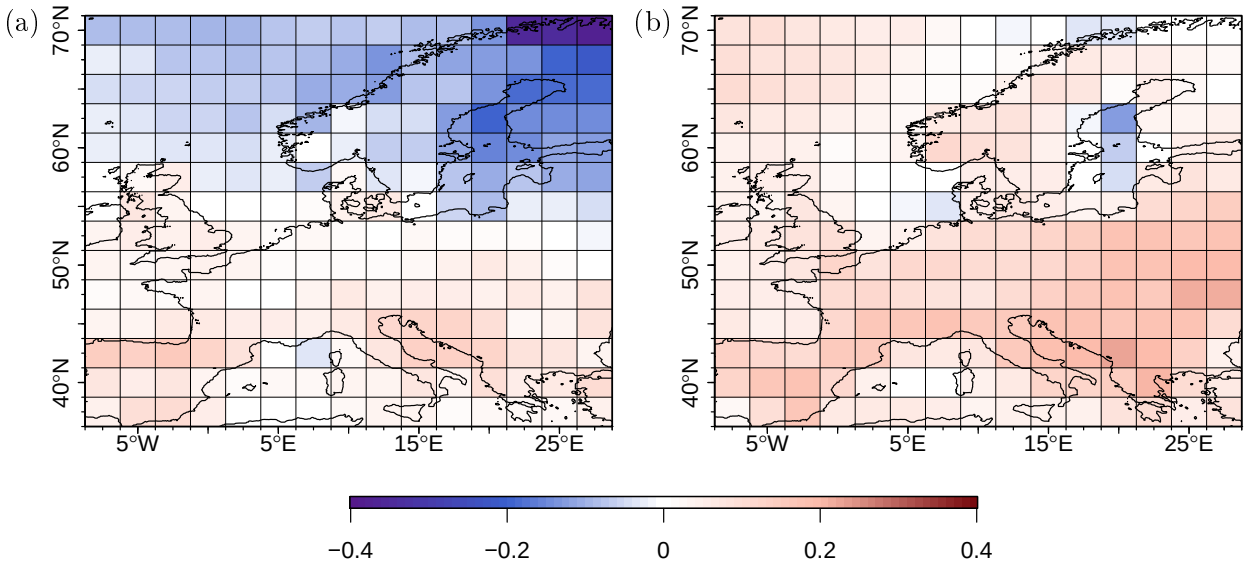

Figure 6.11: *Differences between assimilated 2 m temperatures and PMIP3 multi-model mean for winter (a) and summer (b) temperatures in K.*

process both presence and absence information.

6.4 Summary and Outlook

A statistical approach is presented in order to merge proxy (pollen) data and model data. Firstly, probabilistic information of pollen data is estimated by calculating botanical climate transfer functions using the GLM. Secondly, a multi-model ensemble of eight PMIP3 models is created by applying ensemble kernel dressing, and simulations are generated by random sampling through estimating the joint covariance matrix. In each generation step, the Brier (skill) score is determined. Thirdly, the generated simulations are weighted with the corresponding Brier scores.

Including the probabilistic information of the observed pollen data change the PMIP3 multi-model mean by up to 0.5 K. For winter, the assimilated temperatures are lower in Scandinavia and over the North Sea, whereas the assimilated temperatures are higher in the southern part compared to the original PMIP3 multi-model mean. Although there are no observations available e.g. over the North Sea, information outside the coring sites can be extrapolated through the covariance modeling, which is based on the coring sites. In summer, the assimilated temperatures are generally higher especially over land. Particularly for summer, we can detect an added value (20% improvement on average), as the Brier skill scores are predominantly positive for nearly all coring sites. This can be associated with the growth phase of plants, which has its maximum in summer (at mid-latitudes).

Is there an overfitting problem due to the ensemble kernel distribution approach compared to a Gaussian distribution fit? Generating simulations from a Gaussian covariance matrix indicate that the ensemble members are independently and identically distributed (iid), which is a sharp assumption as

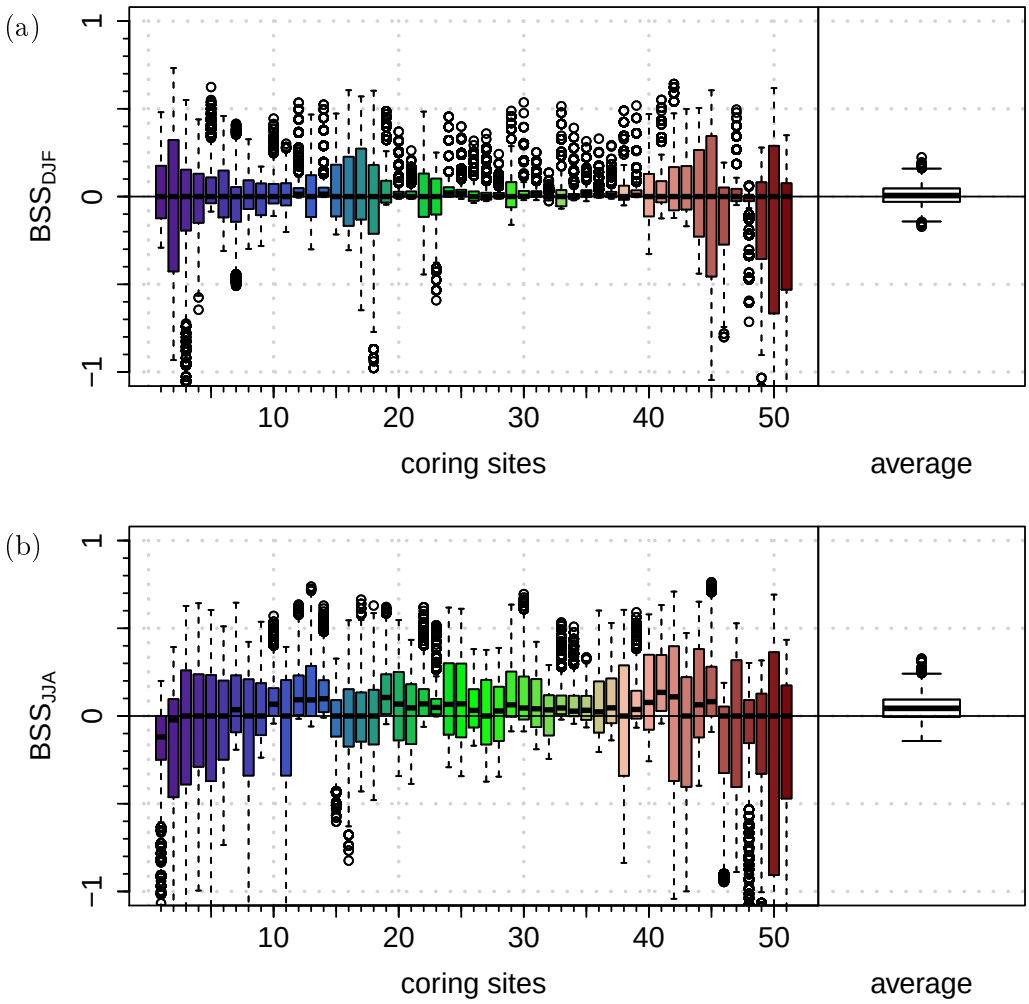

Figure 6.12: *BSS for BS based on EKD with BS based on the Gaussian distribution fit as reference. The BSS is shown for winter (a) and summer temperatures (b) dependent on the 51 coring sites. Next to the BSS, the averages over all coring sites are presented.*

the ensemble members are not necessarily independent from each other. Figure 6.12 illustrates the Brier skill scores for the BS from ensemble kernel dressing with the BS from a Gaussian distribution fit as reference BS. Averaging over all stations, the median of the skill scores for winter is zero, i.e. the BS of both approaches do not differ. For summer, the averaged Brier skill score is slightly positive, however, we cannot detect clear overfitting due to the ensemble kernel dressing.

Also the results for spring and autumn temperatures are promising (see Appendix B.1.2). The Brier skill scores are predominantly positive for the different coring sites and clearly positive when averaging over all coring sites. Looking at the differences between assimilated temperatures and the PMIP3 multi-model mean, the transition periods show what we expect: we see a similar structure for spring/winter and autumn/summer, however with smaller amplitudes in spring/autumn. This underlines that for cold temperatures especially above 45°N, the inclusion of probabilistic information of proxies has a

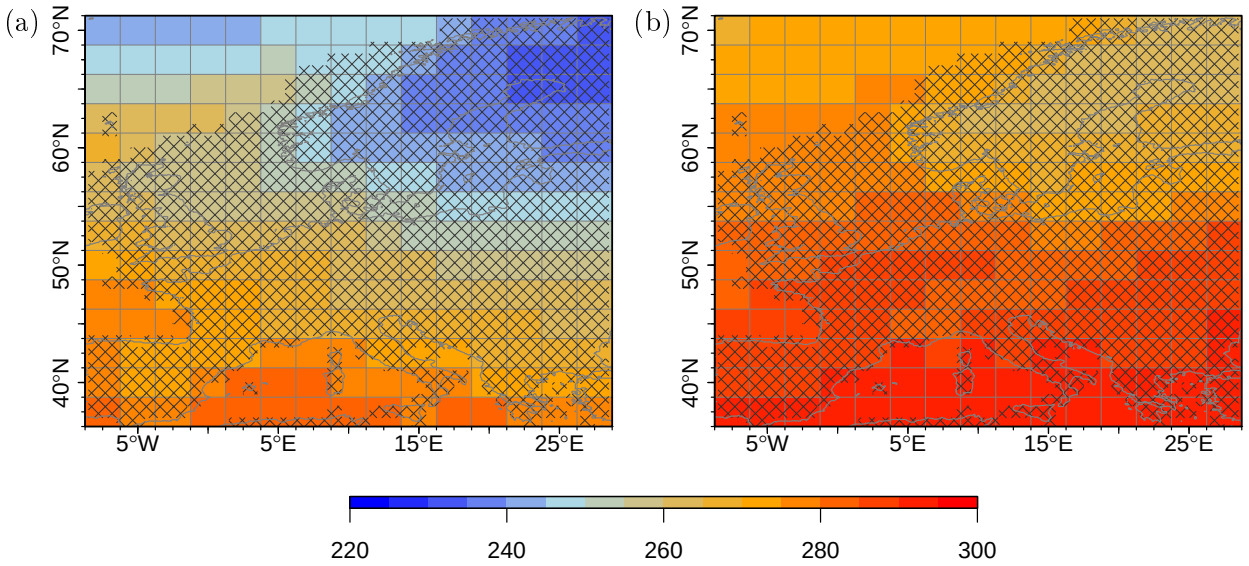

Figure 6.13: *PMIP3 multi-model average during the LGM for winter (a) and summer (b) 2 m temperatures in K. Black hatching denotes the land-sea mask (including the ice-sheet extends) which is used for generating the model experiments.*

smaller effect compared to warmer temperatures.

As the *homo sapiens* started to cultivate e.g. crop in the Holocene, the presented concept can be applied to growing degree days (GDDs) e.g. for summer wheat, emmer, einkorn, etc. As the growing phase strongly depends on moisture, it would be worth to study also e.g. precipitation. For non-Gaussian distributed variables, the estimation of the covariance matrix and generation of the corresponding simulations have to be refined.

For further analyses, the considered area can be enlarged to the west (Portugal) and east (Greece, Turkey, etc.) by including available pollen data for the needed time slice. Also filling gaps e.g. in northern France and considering north African coring sites can be beneficial. Including proxy data from marine sediment cores would be another interesting option if the corresponding recent vegetation data are available.

As the PMIP3 experiments have also been generated for the Last Glacial Maximum (LGM, 21 ka BP), applying the reconstruction concept to this time slice could be beneficial. In Figure 6.13, the multi-model mean for winter and summer temperatures is shown based on the following six models: CCSM4 (by NCAR), CNRM-CM5 (by CNRM/CERFACS), COSMOS-ASO (by FUB), IPSL-CM5A-LR (by IPSL), MPI-ESM (by MPI) and MRI-CGCM3 (by MRI). The hatched area displays the land-sea mask during the LGM, which was used for the model initialization. When estimating the transfer functions for the LGM, one has to consider that i) the land-sea distribution has changed over the years and ii) the climate conditions under which a taxon can exist might have changed.

Chapter 7

Local reconstruction at the Dead Sea

In this chapter, we show a different, well-known method to obtain probabilistic information of pollen proxies. The Bayesian biome model can be applied in regions that offer certain vegetation conditions, e.g. at the Dead Sea area. These probabilistic information can be used instead of the GLM-probabilities for the statistical approach to merge proxy and model data (see Section 6.2). Firstly, we describe the Dead Sea basin and necessary datasets as basis for the reconstructions. After a brief summary of the Bayesian biome model, we analyze the climate conditions for the last approx. 150 ka in the Dead Sea region.

7.1 Dead Sea area and data basis

The local situation of the Dead Sea basin is special and unique. It is located on the lowest continental depression with a level of 427 m below mean sea level in 2013 (Neugebauer et al., 2014). Since the beginning of the Holocene, the Dead Sea surface is ca. 76 km long and between 15 and 17 km wide (Litt et al., 2012b). It is a terminal lake, i.e. water that comes in has no outlet, receiving fresh water from the Jordan River. The lake is hypersaline with a water density of 1.234 g/l and is mainly inhabited by certain green algae and archaeobacteria (Niemi et al., 1997). The Dead Sea consists of two basins separated by the Lisan Peninsula in the eastern part (Bentor, 1961). The basin structure is mainly a result of the strike-slip faults along the Jericho and Arava valley (Niemi et al., 1997). Its drainage area is about 40,000 km^2 large including the Mount Hermon massif and Lake Kinneret in the north, the northern part of Arava in the south, and the area between the Judean Mountains in the west and the Jordan Plateau in the east (Bentor, 1961).

Within this area, a variety of vegetation zones due to the transition between arid and sub-humid climate, so-called biomes, is present. A biome can be defined as a generalization of plant functional types (PFTs). A PFT groups plants according to their functions in the ecosystem (Smith et al., 1993) such as morphological characteristics (plant size, leave shape, size, etc.). For the Dead Sea basin, four different biomes can be identified: the Mediterranean territory (including e.g. *Quercus ithaburensis, Quercus calliprinos, Olea europaea*), the Irano-Turanian territory (steppe vegetation including e.g. *Ephedra, Artemisia*), the Saharo-Arabian territory (desert vegetation including e.g. *Tamarix*, Chenopodiaceae), and the Sudanian penetration territory (subtropical vegetation including e.g. *Ziziphus spina-christi*). The latter biome will not be considered for further statistical analyses as it mainly depends on ground

water and, thus, is supposed to be a weak climate indicator (Litt et al., 2012b).

We use the biome distributions after Thoma (2016). Thoma (2016) slightly changes the biome distribution areas, which originally come from Meusel et al. (1965), and finds qualitatively improvements in the reconstructions for the Levant (setup 3 in Thoma (2016) is shown in Figure 7.1). The modifications involve enlarging the Mediterranean biome to the north and north-west, including Cyprus and the southern coastal area of Turkey, and negligibly relocating the Irano-Turanian biome to the south.

Climate data are taken from the Climate Research Unit (CRU), which have already been described in Section 6.1.2. Figures 7.1 shows the climatic means (1901-2013) of winter temperatures and annual precipitation (annual sum of monthly precipitation rates). There are winter temperatures of up to 25°C and annual precipitation below 200 mm in the southern part of the Levant. The winter temperatures decrease northwards (values below 0°C), whereas precipitation values increase up to 800-1,000 mm/year in the Golan heights. Averaging over the Dead Sea basin, the 20th century winter temperature is ca. 10°C and annual precipitation is ca. 200 mm. At the drilling location, the climatic mean for the winter temperature is approx. 12°C and for annual precipitation approx. 200 mm. The data are available in a high resolution (with 0.5° in longitude/latitude), and the area is restricted to 30°E-50°E and 20°N-40°N.

Within the ICDP (International Continental Scientific Drilling Program) deep drilling project at the Dead Sea in 2010/2011, a 460 m long sediment core could be obtained encompassing the last 220 ka (Stein et al., 2011; Neugebauer et al., 2014; Torfstein et al., 2015). The main aim of this project is to receive a dataset in a highly temporal resolution in order to reconstruct the paleoenvironment and paleoclimate but also the paleoseismicity and paleomagnetism (Stein et al., 2011). Palynological analyses, which serve as an essential basis for statistical climate reconstructions, have been executed by Andrea Miebach and Chunzhu Chen at the Steinmann Institute of the University of Bonn. They have divided the palynological assemblages in two phases based on the sedimentology: phase 1 looks at the Samra Formation and the upper Amora Formation spanning the period between 146 and 90 ka BP (including the Last Interglacial), and phase 2 looks at the Lisan Formation and the upper Zeelim Formation spanning the period between 88 and 9 ka BP (including the Last Glacial).[1]

7.2 Bayesian biome model (BBM)

The Bayesian biome model (BBM) is used here as the vegetational situation of the Dead Sea basin is unique (three vegetation zones within a few kilometers, see Section 7.1) and characterized by a large catchment area (Litt et al., 2012b) compared to other regions, e.g Birkat Ram in the Golan Heights (Neumann et al., 2007). The BBM has already been introduced and used for statistical climate reconstructions, e.g. for the Holocene Dead Sea (Litt et al., 2012b; Schölzel, 2006).

[1] The calibrated ages are taken from Miebach (2016) and Chunzhu Chen (PhD-thesis in preparation).

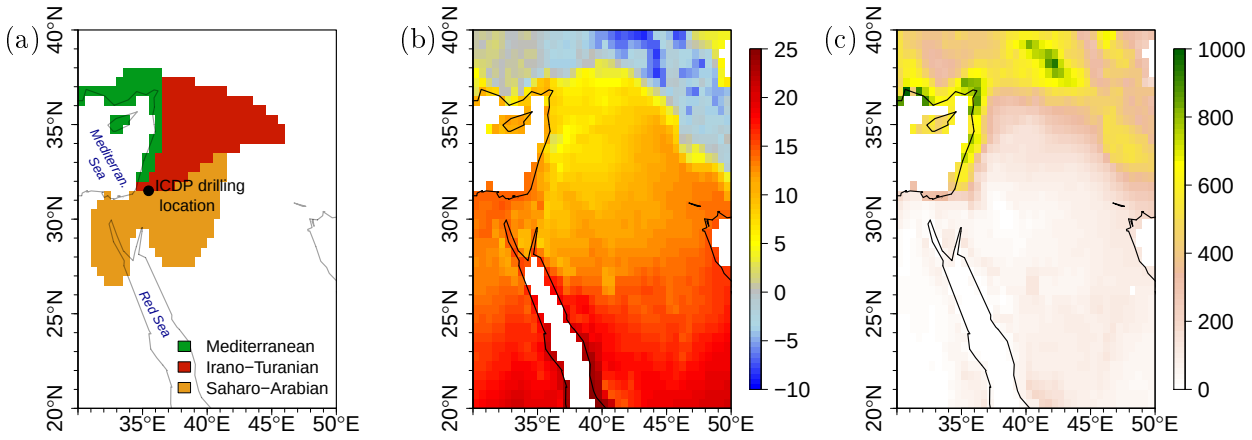

Figure 7.1: *Distribution areas (setup 3 taken from Thoma (2016)) of the Mediterranean, the Irano-Turanian, and the Sahara-Arabian biomes (a), climatic mean of winter temperatures (b), and annual precipitation (c).*

The BBM is based on Bayesian hierarchical models (BHM), with which a complicated model can be separated into several levels (Ohlwein and Wahl, 2012; Li et al., 2007). The BBM mainly consists of estimating the following probabilities: firstly, the probability for the occurrence of each biome given the corresponding pollen spectra; secondly, the biome-climate transfer function; thirdly, the prior distribution.

As explained in Ohlwein and Wahl (2012), the reconstruction problem in a Bayesian framework can be generally expressed as a joint probability density function,

$$P(PO, C, \Theta) = P(PO|C, \Theta) \cdot P(C|\Theta) \cdot P(\Theta), \tag{7.1}$$

where PO denotes the raw pollen counts. The climatic state is summarized in the multivariate variable C, and Θ includes a set of statistical parameters. The first probability on the right-hand side of Equation 7.1 describes the data stage, i.e. the statistical relation between pollen and climate data. The second term describes the process stage including the climate process based on space and time, and the third term describes the prior distribution.

For the biome approach, a further vegetation step V is inserted in Equation 7.1, resulting

$$P(PO, C, \Theta) = \sum_i P(PO|V_i, \Theta) \cdot P(V_i|C, \Theta) \cdot P(C|\Theta) \cdot P(\Theta), \tag{7.2}$$

where the relation between pollen counts and climate is split into $P(PO|V_i, \Theta)$ and $P(V_i|C, \Theta)$. The probability $P(PO|V_i, \Theta)$ describes the relation between pollen counts and vegetation (e.g. biomes) and assumes that the pollen production does not depend on the climate when a vegetation step is inserted. $P(V_i|C, \Theta)$ is the probability for the occurrence of a biome in dependence on climate. Applying the

Bayes theorem to Equation 7.2, we get the following relation for the posterior distribution $P(C, \Theta|PO)$:

$$\begin{aligned} P(C, \Theta|PO) &= \frac{\sum_i P(PO, V_i, C, \Theta)}{P(PO)} \\ &\propto \sum_i P(PO|V_i, \Theta) \cdot P(V_i|C, \Theta) \cdot P(\Theta) \end{aligned} \tag{7.3}$$

Here, we omit the process stage $P(C|\Theta)$ as it describes the evolution of climate fields (Tingley and Huybers, 2010), which is not necessary for local reconstructions. Furthermore, it is assumed that Equation 7.3 is valid for both recent and past climate. In the following subsections, we give a brief overview of the steps (terms) within the BBM after Litt et al. (2012b), Schölzel (2006), Ohlwein and Wahl (2012), and others.

7.2.1 Biome ratios

Firstly, a statistical relation between the three biome compositions is done, which is represented by $P(PO|V_i, \Theta)$ in Equation 7.3. Applying the Bayes theorem, we estimate the conditional probability for the occurrence of a biome given the pollen spectra, as these are the essential proxy data. This probability can be estimated by the so-called affinity score A_{ik} (Prentice et al., 1996),

$$A_i = \sum_j \delta_{ij} \sqrt{\max(0, p_j - \theta_j)}, \tag{7.4}$$

i.e. the affinity of pollen samples to a biome. The biome is denoted by i and the occurring taxon by j. Here, δ_{ij} is the taxa-biome affiliation, p_j the pollen percentage, and θ_j a threshold percentage for the occurrence of a taxon. This threshold is introduced to reduce the background noise. Therefore, we analyze the empirical cdf (ECDF) for a considered taxon and set the threshold usually at the first striking step. This procedure is done for each taxon individually and, thus, not objectively decided. For each sediment layer, the biome ratios are normalized to one,

$$prob(V_i|ps) = \frac{A_i}{\sum_i A_i} \tag{7.5}$$

where ps denote the pollen spectrum and V_i the biome i with $i = 1, 2, 3$.

7.2.2 Biome climate transfer functions

The statistical relation between one biome i and climate, $P(V_i|C, \Theta)$, can be expressed by using e.g. a generalized linear model (GLM) or a quadratic discriminant analysis (QDA). The latter concept is used for further analyses.

Probabilistic classification based on the QDA

The discriminant analysis is a process of estimating a function to correspond a variable to one of several possible groups. The QDA uses, in contrast to the linear discriminant analysis (LDA), a separately

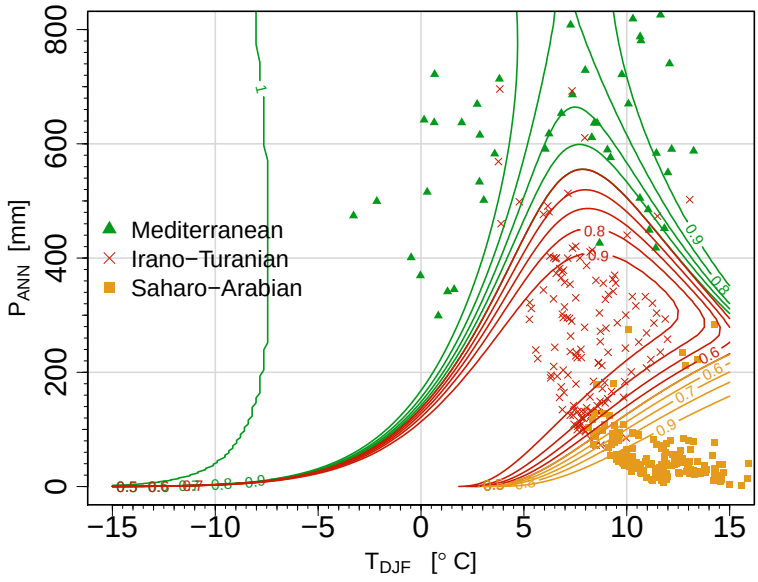

Figure 7.2: *Bivariate distribution of the two dimensional climate state vector (triangles, crosses, squares) for the Mediterranean biome (green), the Irano-Turanian biome (red), and the Saharo-Arabian biome (orange). The contour lines represent the biome transfer functions based on the QDA.*

estimated covariance matrix for each group.

The probabilistic classification (Wilks, 2011), i.e. the probability for a certain climate state vector x belonging to one biome V_i with $i = 1, 2, 3$, can be obtained via the Bayes theorem resulting in

$$pdf(V_i|x) = \frac{p_i^* f_i(x)}{\sum_i p_i^* f_i(x)} \tag{7.6}$$

with

$$f_i(x) \propto (\det \hat{\Sigma}_i)^{-1/2} \exp\left(-\frac{1}{2}(x - \overline{x}_i)^T (\hat{\Sigma}_i)^{-1}(x - \overline{x}_i)\right), \tag{7.7}$$

where $\hat{\Sigma}_i$ is the estimated covariance matrix and $\overline{x}_i$ the climate mean both within biome i. In each biome, the data are assumed to be normally distributed. Thus, precipitation data are modified before calculating the transfer functions: a gamma cdf and Gaussian quantile function are used and the data are transformed to a normally distributed random variable. The prior distribution for the biome membership is denoted by p_i^*, where $p_i^* = \frac{n_i}{N}$ with $N = \sum_i n_i$, the size of the training dataset. The probability for the occurrences of the Mediterranean biome, the Irano-Turanian biome, and the Saharo-Arabian biome given winter temperature and annual precipitation is shown in Figure 7.2.

One advantage of using the QDA compared to the GLM is that the probability for each biome is dependent on the other two biomes. Furthermore, the probabilities are normalized to one. Using the GLM, the probability for the occurrence of each biome given climate is calculated independently from

the other biomes, whereby overlapping areas can appear. However, the QDA uses only the presence information of each biome, which is different for the GLM where also the absence information is used.

7.2.3 Selecting prior distribution

If there is any knowledge about the climate in the Near East, one can incorporate this information in the prior distribution $P(\Theta)$ (see Equation 7.3). As in Litt et al. (2012b), we choose an almost non-informative prior distribution, i.e. a bivariate distribution with a normal distribution for winter temperature and a gamma distribution for annual precipitation both as marginal distributions. Selecting an almost non-informative prior has the advantage of not affecting the posterior relating its ability to learn from the data and not from the prior (Ohlwein and Wahl, 2012).

7.2.4 Assumptions for the BBM

Again, the same assumptions are valid as for the transfer functions calculated with the GLM (see GLM paragraph in Section 6.2.1): firstly, the occurrence of a biome only depends on those climate variables that are to be reconstructed; secondly, the climatic dependencies under which a taxon appears did not change over the years; thirdly, the composition of the biomes did not change over the years but caused a relocation.

7.3 Application to Dead Sea sediment core

In this section, we apply the BBM to the Dead Sea core. As already mentioned, we split the reconstructions in two phases: phase 1 (199.2-340.6 m depth; ca. 89.1-147.3 ka BP) and phase 2 (63.2-199.1 m depth; ca. 8.7-87.6 ka BP). In Figure 7.3 the biome probabilities are shown for phase 1 and phase 2. In the upper panel, the probability for the Mediterranean biome is presented, in the middle for the Irano-Turanian biome, and in the lower for the Saharo-Arabian biome. A list of considered taxa and pollen percentages including thresholds, which are necessary for the biome probabilities, are given in Appendix B.2. Additionally, the boundaries for the pollen assemblage zones (PAZs), which are defined by using a cluster analysis, and the corresponding marine isotope stages (MIS) are marked.

During late MIS 6, the Mediterranean and the Irano-Turanian biomes are the dominant zones, while the Saharo-Arabian biome only appears sporadically. Changing to early MIS 5e, the Eemian-Interglacial, the Saharo-Arabian biome probability massively increase up to 80%, while the probabilities for the Irano-Turanian biome decrease to less than 20%, and the Mediterranean biome decrease to less than 10%. Between 290 and 240 m depth, the Mediterranean biome probability slowly increases, while the Irano-Turanian biome probability decreases from 80% to 20% with some fluctuations. In PAZ II5, we can observe a negative trend for the occurrence of the Mediterranean biome, while the Irano-Turanian biome probability reaches values up to 60% in the middle of this PAZ, and while the Saharo-Arabian biome probability fluctuates around 40%.

Looking at phase 2, the probability for the Irano-Turanian biome is nearly constant (approx. 30%). The main drivers are thus those taxa that are part of the Saharo-Arabian biome, which only consists of two taxa (Chenopodiaceae and *Tamarix*) and the Mediterranean biome. For late MIS 4 and MIS 3, the Saharo-Arabian biome probability decelerates but tends to increase again during MIS 2 until 10 ka BP, whereas the Mediterranean biome probability reaches values above 50% during MIS 3 and decreases to 10% until ca. 12 ka BP. The probability for the Irano-Turanian biome proceeds constantly with values around 30% and a peak at early MIS 2 (60%), which is mainly formed by the raised occurrence of *Artemisa* (see Appendix B.2).

At the beginning of the Holocene (10-8.7 ka BP), the probability for the Mediterranean biome decreases. This is not necessarily a signal for climate change and/or a decreasing lake level but more likely a signal for the anthropogenic influence including cultivating crops and clearing forests. As the sum of biome probabilities is designed to be one for each time slice, the Saharo-Arabian biome increases automatically at the beginning of the Holocene. Compared to the biome ratios in Litt et al. (2012b), our results show smaller values for the Mediterranean (ca. -0.1) and Irano-Turanian biome (ca. -0.3) and higher values for the Saharo-Arabian biome (ca. +0.4) in the early Holocene.

The boundaries of the PAZs fit well to the evolution of the biome probabilities for both phases. This implies that our taxa selection and the estimated thresholds, which serve as basis for the biome probabilities, are representative for the overall biome reconstruction.

Finally, we estimate the posterior probability density functions for the climate state vector conditional on the biome composition for each layer in the sediment core. In Figure 7.4, the reconstructed winter temperature and annual precipitation are shown for phase 1. Furthermore, annual and winter insolation at 30°N (Laskar et al., 2004) are presented and interpolated appropriate to the depth. Between 340 and 320 m depth, the expected winter temperatures vary around 3°C with an interdecile range between -5 and 9 °C. The expected annual precipitation for this depth has an interdecile range between 180 and 600 mm. The prominent increase of the Saharo-Arabian biome in Figure 7.3 is also visible in the reconstructions: the expected increase of winter temperatures up to 7°C and the decrease of precipitation to 220 mm, which is less uncertain (increased pdf-values compared to other time slices). This expected climate state is similar to climate conditions of the 20th century in the Dead Sea basin (as already mentioned, 10°C for winter temperature and 200 mm for annual precipitation). After 122 ka BP, winter temperatures vary around 2°C with an interquartile range between -2 and 5°C. The expected annual precipitation raises again up to 340 mm and declines to 240 mm between 100 and 85 ka BP while the expected winter temperatures increase from 2 to 6°C.

Both insolation curves show a similar evolution with slightly shifted maxima. Winter insolation with its two maxima can be quantitatively associated with higher reconstructed winter temperatures between 315 and 290 m and at around 220 m with insolation values up to 258 W/m^2.

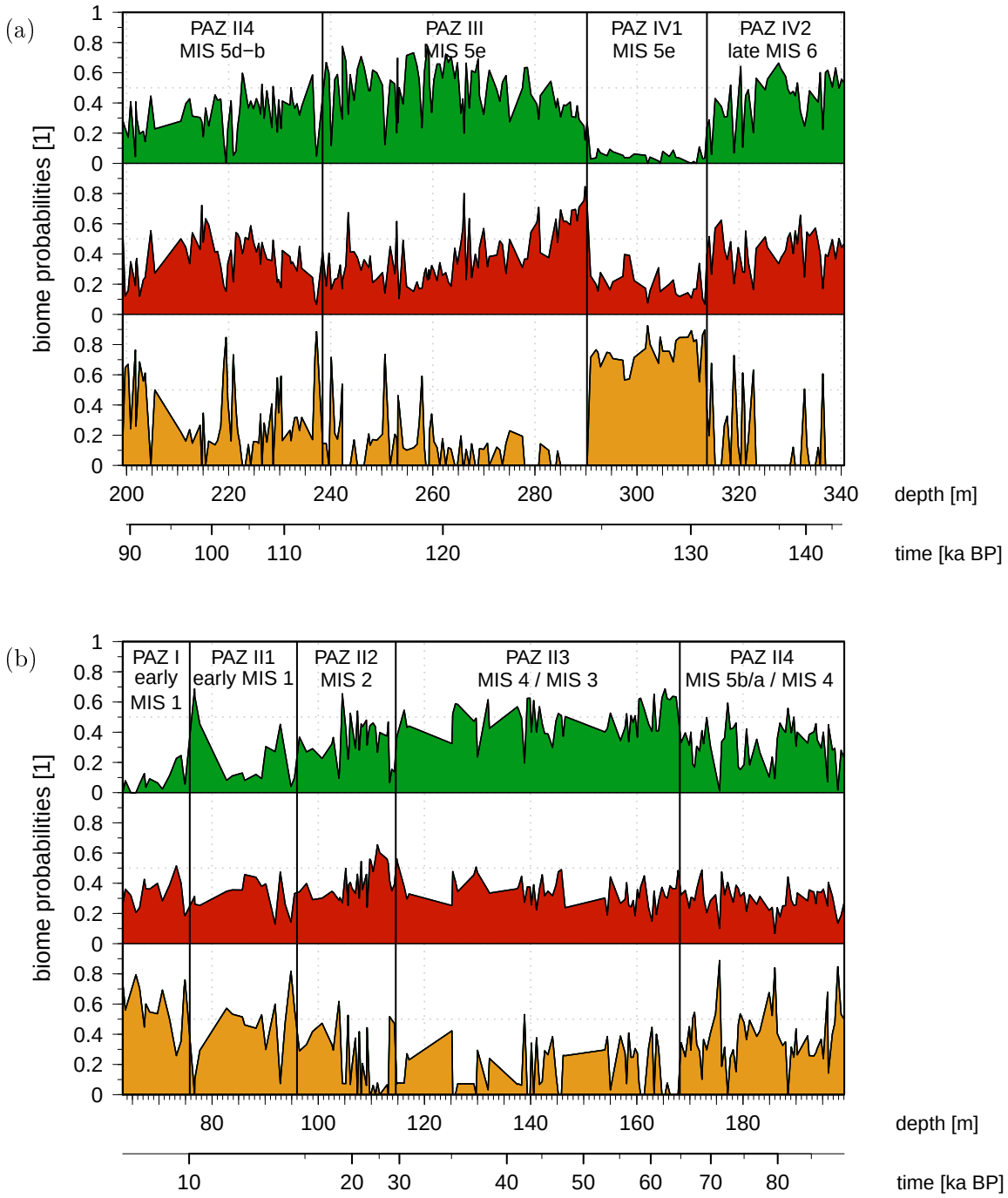

Figure 7.3: *Biome probabilities for phase 1 (a) and phase 2 (b). The probability for the Mediterranean biome is shown in the top panel (green), for the Irano-Turanian biome in the middle panel (red) and for the Saharo-Arabian biome in the lower panel (orange). Additionally, the pollen assemblage zones (PAZs) are marked with the corresponding MIS (marine isotope stage). The unlabeled ticks on the time axis denote 5-year intervals.*

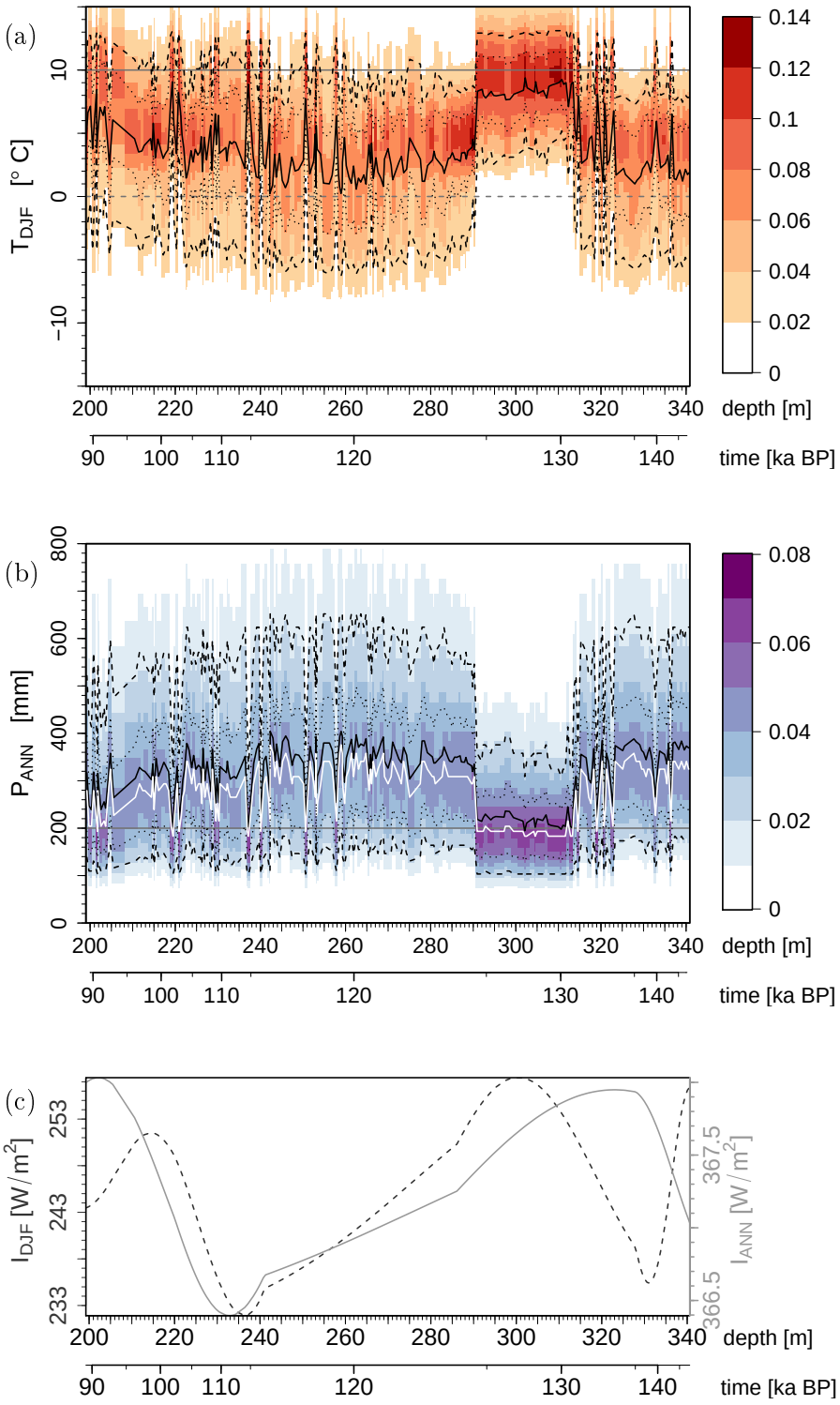

Figure 7.4: *Posterior probability density function for winter temperature (a) and annual precipitation (b) for phase 1. The solid black line denote the mean, the white solid line in (b) the median, the doted lines the 25%- and 75%-quantiles and the dashed lines the 10%- and 90%-quantiles. The horizontal lines at 10°C and 200 mm represent the climate of the last century for the Dead Sea basin. The winter insolation (black dotted line) and annual insolation (gray line) for 30°N (Laskar et al., 2004) is interpolated to the depth values and shown in (c). The unlabeled ticks on the time axis denote 5-year intervals.*

Before 65 ka BP, the expected winter temperatures fluctuate between 8 and 1°C and precipitation rates vary around 280 mm (see Figure 7.5). During late MIS 4 and MIS 3, temperatures decrease while precipitation tends to increase. In this stage, the pdf-values especially for winter temperatures are lower compared to the other time intervals with interquartile ranges between -2 and 8°C. Interstadial phases (so-called Dansgaard-Oeschger events), which are short warming phases during the Glacials, e.g. during MIS 4 and MIS 3, cannot be observed in our reconstructions. One problem is the appearance of slumps, which can distort the age model and through which sediment gaps can arise. For the late Glacial, the expected winter temperatures increase again up to 7°C until the early Holocene including an abrupt cold snap at a depth of ca. 75 m. In this time frame, precipitation decreases to 200 mm which is similar to today's conditions.

Looking at winter insolation, we can observe three local maxima. The first winter insolation maximum at 185 m depth corresponds to raised expected winter temperatures. The second peak at approx. 150 m depth cannot be clearly seen in the expected winter temperatures as slumps occur. However, the pdf-values have comparatively low values during this time period. At the winter insolation maximum around 105 m depth and subsequent increased winter and annual insolation values compared to other depths, the winter temperatures tend to increase as well. As the insolation curve is presented as function of depth and not of time (as it is received from the original data), sharp bends occur e.g. at 241 or 286 m due to the non-linear behavior between depth and age.

For both phases, the expectation values for winter temperatures are often slightly above zero. We emphasize that the temperature and precipitation reconstructions represent a probabilistic climate state for the Dead Sea basin and not for a specific drilling point at the Dead Sea. As already mentioned, the Dead Sea area encompasses a large catchment area including mountainous regions, e.g. Mount Hermon in the north, where winter temperatures below 0°C are possible during the Last Glacial (Ayalon et al., 2013). Negative winter temperatures at the Dead Sea are not likely to appear, even during cold stages, which is due to the orography and the knowledge about existing frost-sensitive plants. For precipitation, there is the same effect: our reconstructions for the Dead Sea basin show an interdecile range between 100 and 600 mm during the last Glacial. Directly at the Dead Sea, values below the expected reconstructed values, which are around 200 mm, are likely to appear. However, for the Soreq Cave, which is located in the Judean Mountains in the west of the Dead Sea (400 m above sea level), precipitation rates up to 500 mm/year are expected between 85 and 80 ka BP (Bar-Matthews et al., 2003).

7.4 Summary and Outlook

The Dead Sea catchment is special as vegetation zones change within a few kilometers. We consider three different biomes: the Mediterranean, the Irano-Turanian, and the Saharo-Arabian biome. A wide region is covered by collected pollen. These characteristics form the basis to apply the BBM for

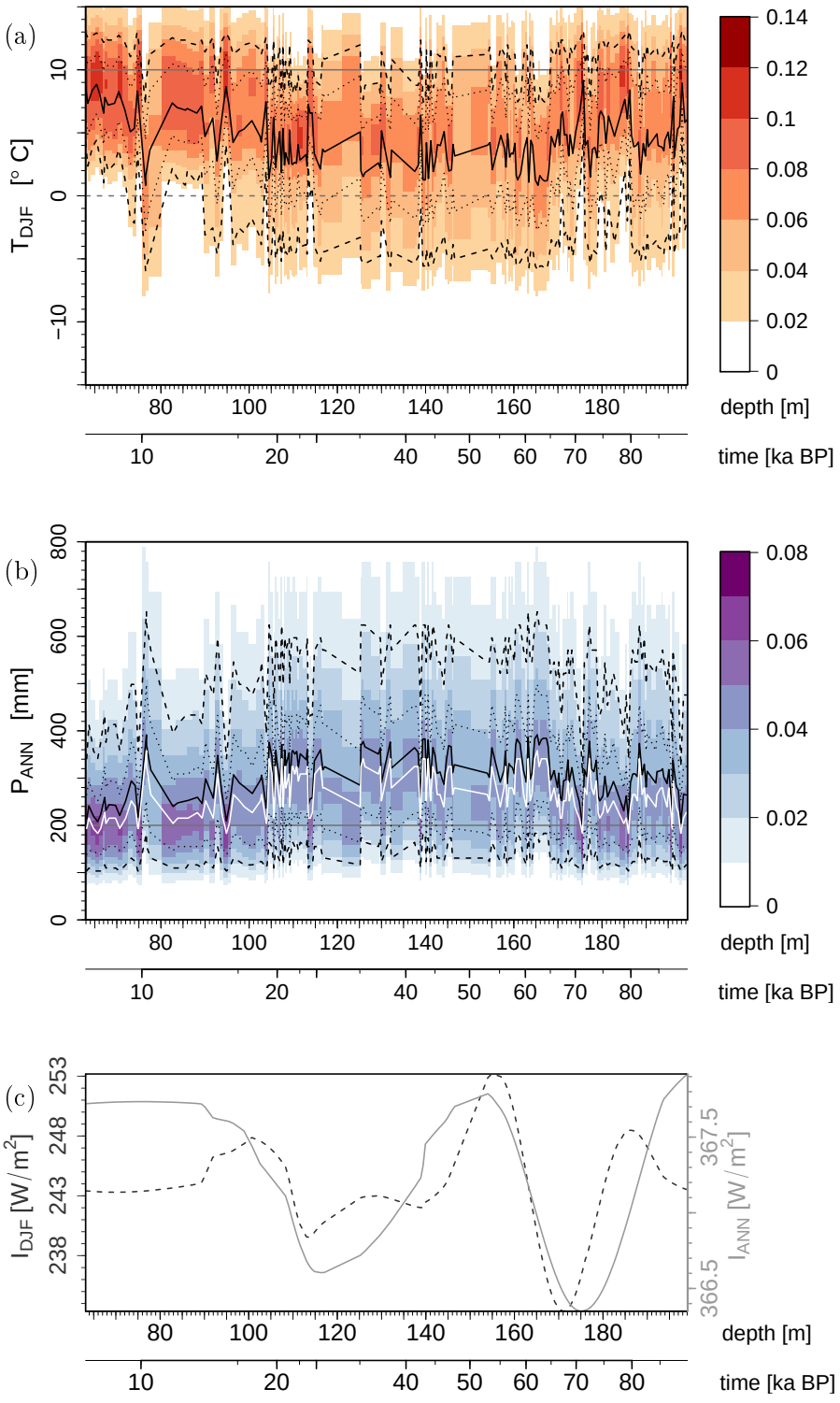

Figure 7.5: *Same as in Figure 7.4 but for phase 2.*

the pollen counts that are taken from a sediment core at the Dead Sea encompassing the time period between 146 and 9 ka BP.

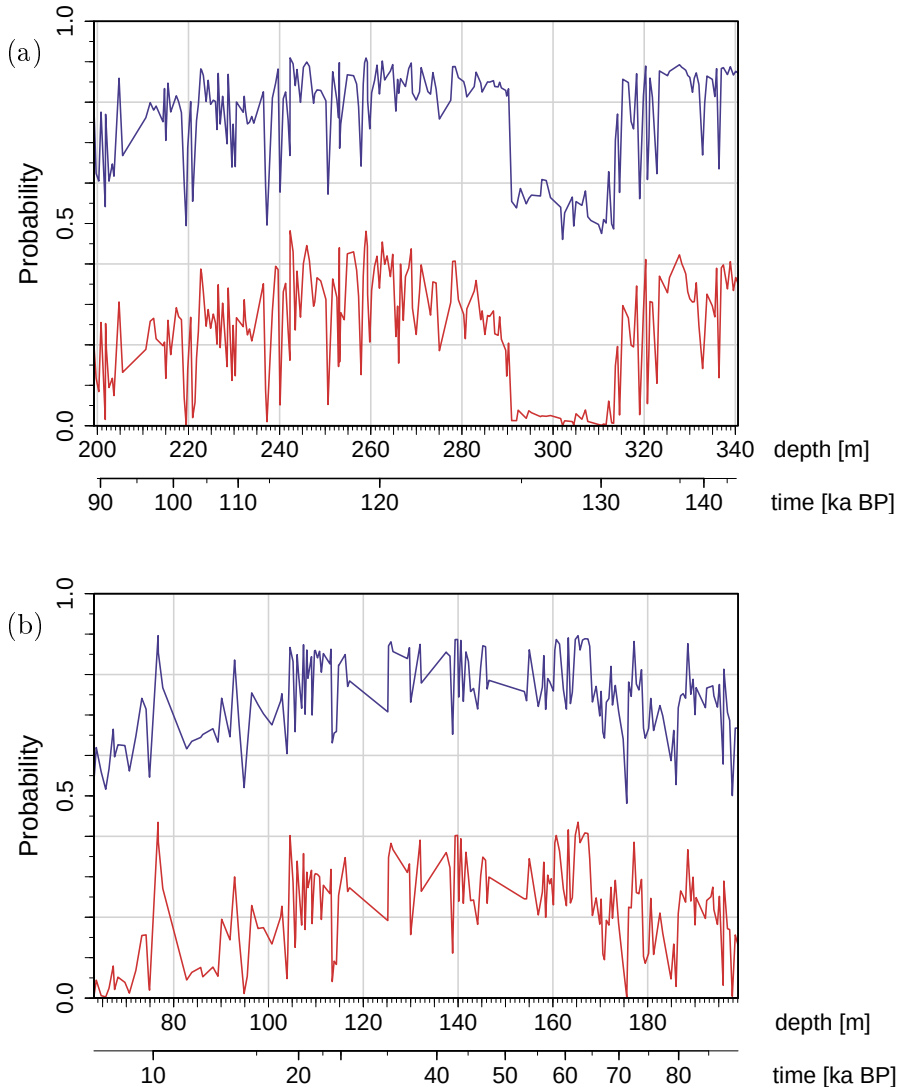

Figure 7.6: *Summarizing representation of the statistical climate reconstruction for phase 1 (a) and phase 2 (b). The blue line displays the probability for annual precipitation being above 200 mm/year, and the red line displays the probability for the reconstructed winter temperature laying below 0°C. The unlabeled ticks on the time axis denote 5-year intervals.*

The BBM can mainly be described in three steps: firstly, we determine the probability for each biome given the counted pollen spectra; secondly, the biome climate transfer functions are estimated, which are the conditional pdfs for the occurrence of a biome given climate (this relation is determined by applying the QDA); thirdly, an almost non-informative prior distribution is selected.

In general, higher biome probabilities are achieved for phase 1 (including the Last Interglacial) with values to some extent higher than 80%. Especially during MIS 5e including the Eemian, the Saharo-Arabian biome probability as well as the Mediterranean biome probability is up to 80%, which is not reachable during phase 2 (including the Last Glacial). The explicit biome composition for both time periods is listed in Appendix B.2. A detailed description of the reconstructed vegetation history can

be found in Miebach (2016) and Chunzhu Chen (PhD-thesis in preparation).

The expected values of the probabilistic reconstructions are valid for the Dead Sea basin and not for the specific drilling position. The general evolution of winter temperatures quantitatively fit to the insolation rate. Figure 7.6 summarizes the information of the climate reconstructions above. The red lines show the probabilities for winter temperatures below 0°C and the blue lines show the probabilities for annual precipitation above 200 mm. The progression of both curves correspond well to each other: the more probable that it was wet, the more probable that there were temperatures below 0°C and vice versa. For phase 1 and phase 2, the probability for winter temperatures below 0°C is mostly between 20% and 40% with some fluctuations and warming phases, e.g. the transition to the Eemian. At the beginning of the Holocene, these probability values decrease to ca. 0% which should not be over-interpreted due to anthropogenic influence.

Moisture in form of precipitation and evaporation is an important factor for the plant growth. Thoma (2016) also reconstructs the annual climatic water deficit (difference between potential evapotranspiration and precipitation), which turns out to be robust and beneficial. The growing phase also depends on the seasonality, which is assumed in the southern Levant (Miebach, 2016) and for Lake Van (Pickarski et al., 2015; Stockhecke et al., 2016). Thus, as an outlook, precipitation or other moisture variables could be reconstructed on seasonal scales.

Chapter 8

Conclusions

In this study, we have investigated a probabilistic quality assessment for model predictions in two application areas: decadal climate predictions and paleoclimate model simulations for the Mid-Holocene (6 ka BP). In both fields, we derive probabilistic information from the model predictions as they are formed as ensembles. Only for the paleoclimate application, we use probabilistic information of the observations. The verification of the ensemble mean with the corresponding observation without taking into account the variance of the ensemble (so-called deterministic evaluation) is not done within this work.

To rank the model predictions, we have presented four attributes to measure the forecast quality: accuracy and skill, calibration/reliability and sharpness (see Chapter 3). To examine accuracy, we use proper scoring rules based on score functions for the Brier score, the continuous ranked probability score and the energy score. These scores can be expressed as skill score when an appropriate reference forecast is available. Additionally, we present the mean squared error skill score (MSESS) after Stolzenberger et al. (2016). For calibration (also known as reliability) analyses, we use PIT (potential integral transform) histograms and the corresponding β-scores, which summarize the graphical character of the PIT histogram in one number. Additionally, we have presented a method to classify reliability into three categories (reliable, potentially useful and not useful). This method is based on the shape of reliability diagrams. Sharpness or potential predictability is determined by the ANOVA (analysis of variance), where no observations are taken into account.

As the decadal climate predictions are available as ensemble predictions, we derive probabilistic information out of these forecasts as probability/cumulative density functions (pdf/cdf). Basis for the decadal climate predictions is the MiKlip (Mittelfristige Klimaprognosen) prediction system, which includes three experiments (baseline 0, baseline 1, prototype) and which has been refined during the project phase of MiKlip (see Chapter 4). The major differences between the considered experiments baseline 1 (b1-LR, b1-MR) and the two prototype sets pr-GECCO and pr-ORA are the initialization techniques in the ocean (from anomaly initialization in b1-LR/MR to full-field initialization in the pr-sets) and the number of ensemble members (10/5 members for b1-LR/MR and 15 members each for pr-GECCO and pr-ORA). To verify the ensembles, we use one observation at one gridpoint / time step coming from atmospheric (ERA-40, ERA-Interim) and oceanic (GECCO2, ORAS4) reanalyses.

The three-dimensional evaluation of geopotential height (Z) and temperature (T) in the atmosphere shows skillful and reliable clusters (see Chapter 5). In the tropical and subtropical mid-troposphere, a lack of skill and reliability is determined, which can be connected to a deficient representation of dynamics in the model physics. In this tropical area, there are high values for potential predictability (the ensemble members agree on the predictions up to 75%) but they predict falsely. The differences between the baseline experiments and between the early and late lead years are only marginal but for the tropical area at 500 and 200 hPa, pr-ORA outperforms pr-GECCO and b1-LR at least for lead year 2-5.

In the ocean, there are skillful areas, which are also reliable. Through all layers in the Pacific ocean, pr-GECCO shows the best skill scores and β-scores but for the North Atlantic, the reliability signal is lost for all experiments especially for pr-GECCO in the deeper layers. Along the equator, pr-GECCO is more consistent with GECCO2 compared to pr-ORA with ORAS4. Pr-GECCO and pr-ORA differ clearly as the initialization data (GECCO2, ORAS4) show large discrepancies. For oceanic variables, we, thus, do not recommend merging the pr-sets to a 30-member ensemble but we suggest to apply e.g. the continuous ranked probability score for Gaussian mixture distributions.

Moreover, we have presented a method based on the energy score to jointly verify the u- and v-wind component by keeping the physical basis of wind velocity and direction. The energy skill score for lead year 2-5 in 10 m height shows predominantly negative and robust values indicating that the observational climate predicts better compared to b1-LR. Predicting near surface variables such as wind velocity in the boundary layer is difficult as the MiKlip decadal prediction system is still in an experimental stage and, thus, dynamical processes in the model physics need to be optimized (Stolzenberger et al., 2016). This verification method can be applied to wind velocities at higher altitudes, though, or combined thermodynamic variables such as dry/moist static energy.

To verify paleoclimate models, we use probabilistic information from pollen data (see Chapter 6). This is done by calculating botanical climate transfer functions using generalized linear models. For the PMIP3 (Paleoclimate Modelling Intercomparison Project Phase 3) multi-model ensemble, a joint covariance matrix is estimated from which simulations are generated (by random sampling). In each generation step, the observational probability is used to determine the Brier scores, with which the simulations are weighted. The PMIP3 multi-model ensemble is, thus, optimized by including the pollen observations if the Brier skill scores are positive. Especially for European summer temperatures, we detect predominantly positive Brier skill scores indicating an improvement of ca. 20% on average when including the probabilistic information of the observed pollen. The assimilated temperatures are higher compared to the PMIP3 ensemble mean (up to 0.4 K over land). This assimilation technique shows also clear benefits for spring and autumn temperatures but for winter temperatures, the inclusion of pollen data is only useful at a few coring sites.

Another way to estimate probabilistic information of the pollen data is done within the Bayesian biome model (BBM, see Chapter 7). The transfer functions are estimated by using the quadratic discriminant

analysis (QDA). The BBM has been developed for the surroundings of the Dead Sea. Its location is characterized by a unique vegetation as within a few kilometers the vegetation zones (Mediterranean, Irano-Turanian and Saharo-Arabian territory) change. The BBM is applied to pollen founds of a sediment core, which was drilled at the Dead Sea and encompasses approx. the last 220 ka. Generally, the climate reconstructions at the Dead Sea basin show lower winter temperatures and higher precipitation rates compared to the climate of the 20th century. We detect the transition to the Eemian warming phase and the subsequent increasing of precipitation and decreasing of winter temperatures. The climate conditions during the Eemian are similar to today's climate. For the Last Glacial, the probability for precipitation above 200 mm/year is at 70% on average and for winter temperatures below 0°C at 40% until the Holocene begins. The BBM can also be adjusted to other regions if the environmental vegetation is similar.

When verifying decadal climate predictions, we assume the observations to be true. The large differences between e.g. ORAS4 and GECCO2 show that one has to pay attention, which dataset is selected. As an outlook, an ensemble of observations can be used to evaluate the decadal climate predictions as it is done for the PMIP3 multi-model ensemble. If enough observational datasets are available, which is challenging particularly for the ocean, this extension would be beneficial (as for the paleoclimate application) since uncertainties in the observations could be included.

Appendix A

Part I

A.1 Additional figures for geopotential height (Z)

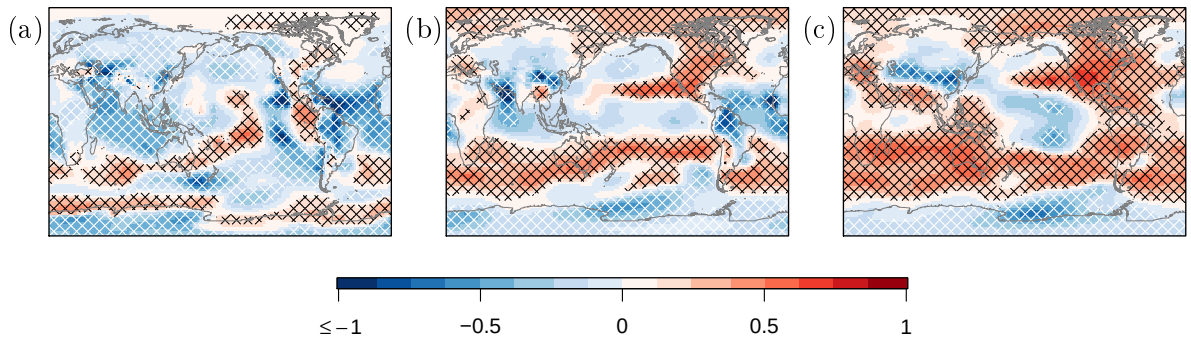

Figure A1: *CRPSS for b1-LR with climatology as reference forecast for Z at 850 hPa (a), 500 hPa (b) and 200 hPa (c) for lead year 2-5. The skill of b1-LR is better than climate if the CRPSS is positive (red shadings). White/black hatching means that the CRPSS is negatively/positively robust.*

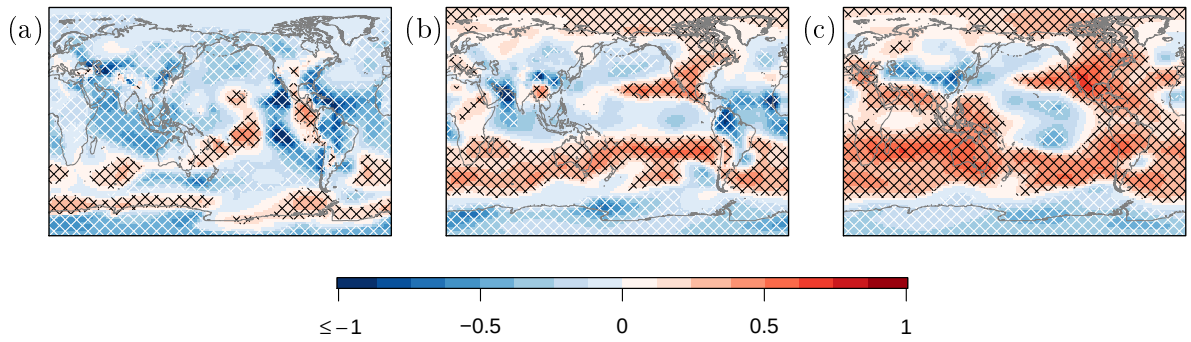

Figure A2: *Same as in A1 but for b1-MR.*

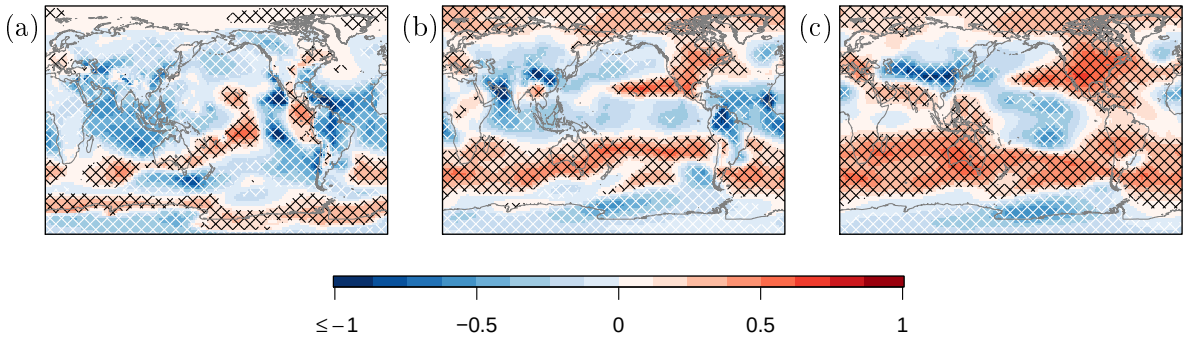

Figure A3: *Same as in A1 but for pr-GECCO.*

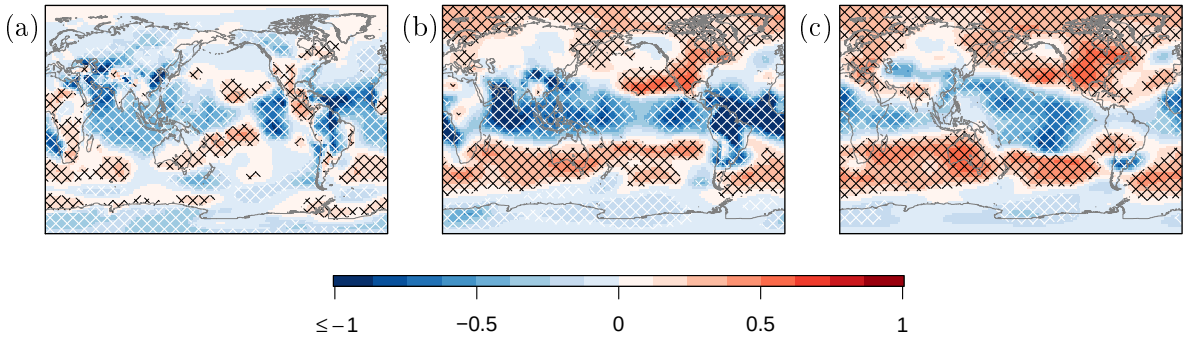

Figure A4: *Same as in A1 but for b1-LR and lead year 7-10.*

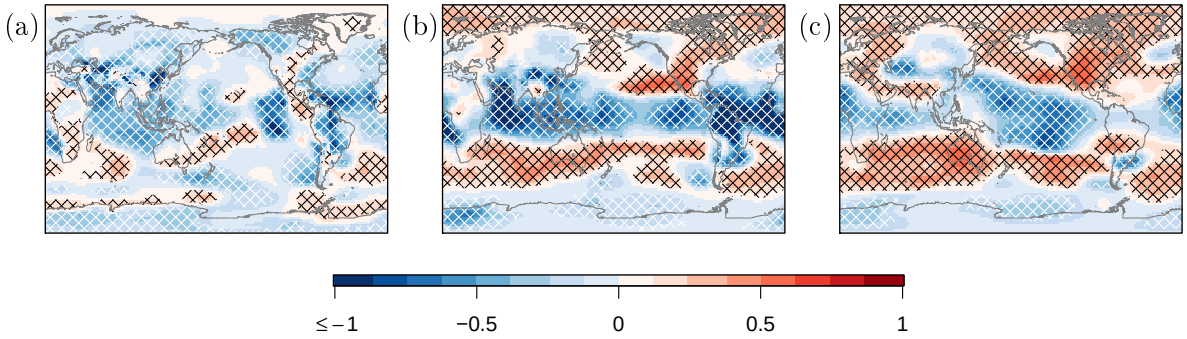

Figure A5: *Same as in A1 but for b1-MR and lead year 7-10.*

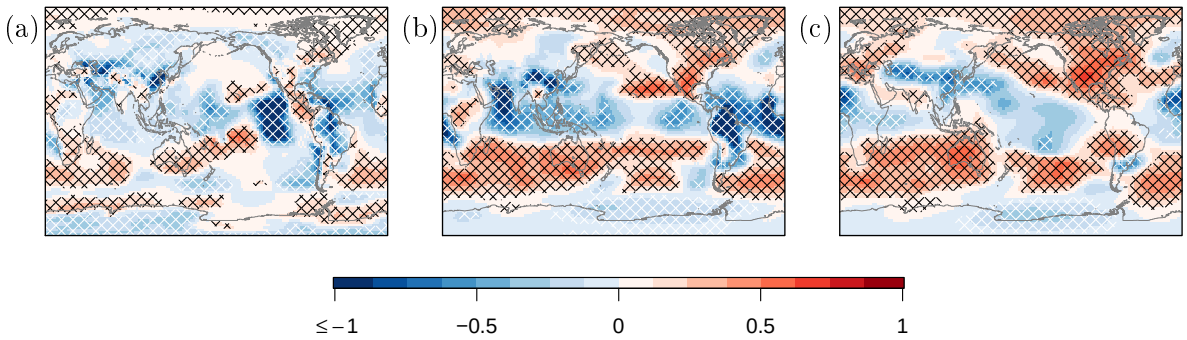

Figure A6: *Same as in A1 but for pr-GECCO and lead year 7-10.*

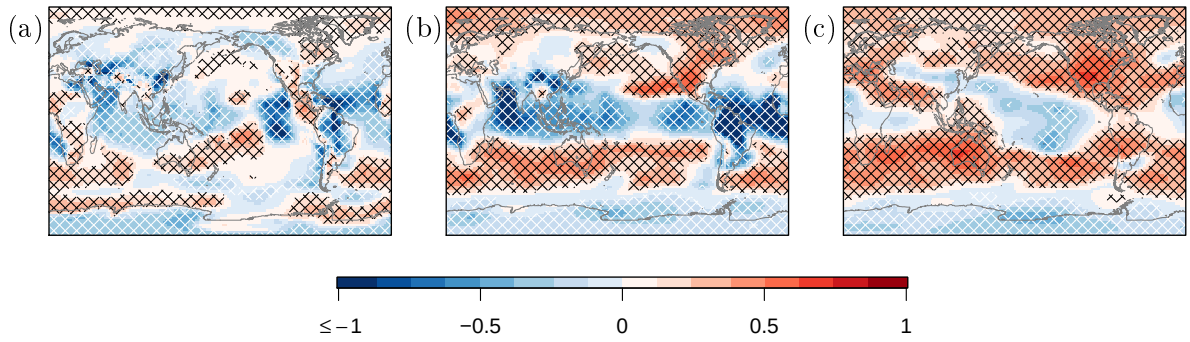

Figure A7: *Same as in A1 but for pr-ORA and lead year 7-10.*

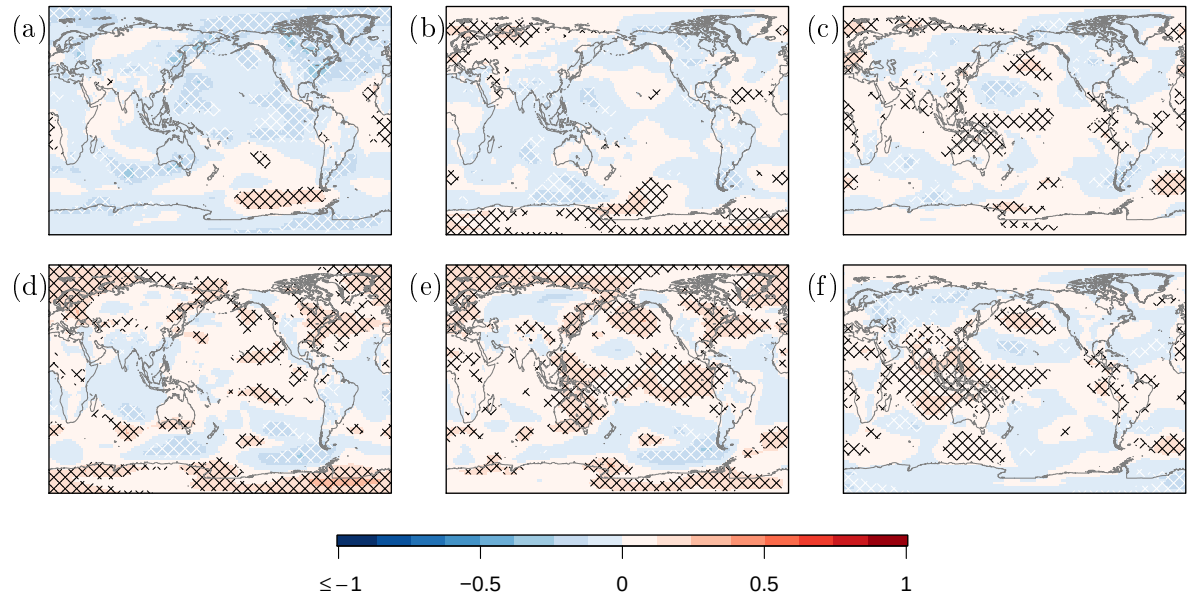

Figure A8: *CRPSS for b1-LR, b1-MR, pr-GECCO and pr-ORA with different reference forecasts for Z at 850 hPa: (a) CRPSS for b1-MR with b1-LR as reference CRPS; (b) CRPSS for pr-GECCO with b1-LR as reference CRPS; (c) CRPSS for pr-ORA with b1LR as reference CRPS; (d) CRPSS for pr-GECCO with b1-MR as reference CRPS; (e) CRPSS for pr-ORA with b1-MR as reference CRPS; (f) CRPSS for pr-ORA with pr-GECCO as reference CRPS. These results are based on lead year 2-5. The skill of the different baseline experiments is better than its reference if the CRPSS is positive (red shadings). White/black hatching means that the CRPSS is negatively/positively robust.*

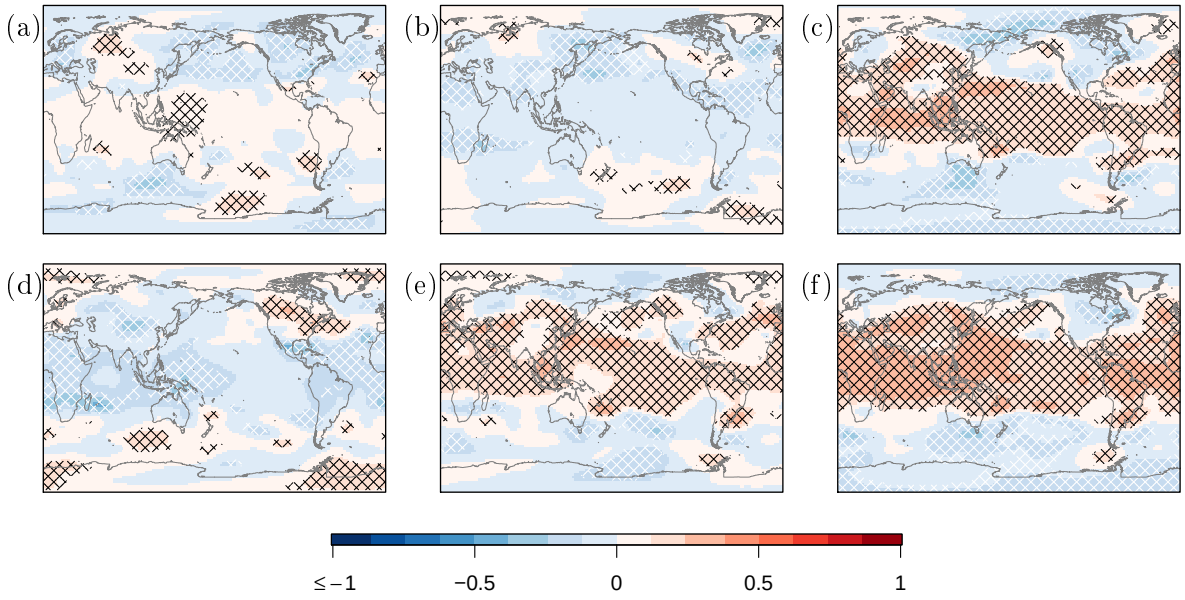

Figure A9: *Same as in A8 but for Z at 200 hPa*

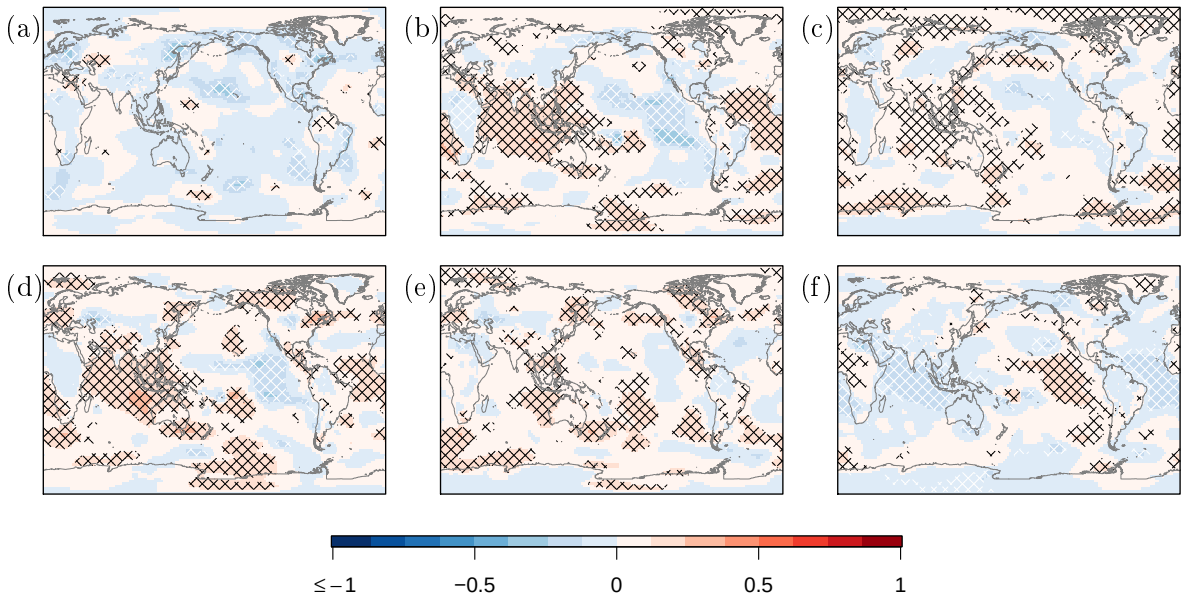

Figure A10: *Same as in A8 but for Z at 850 hPa for lead year 7-10.*

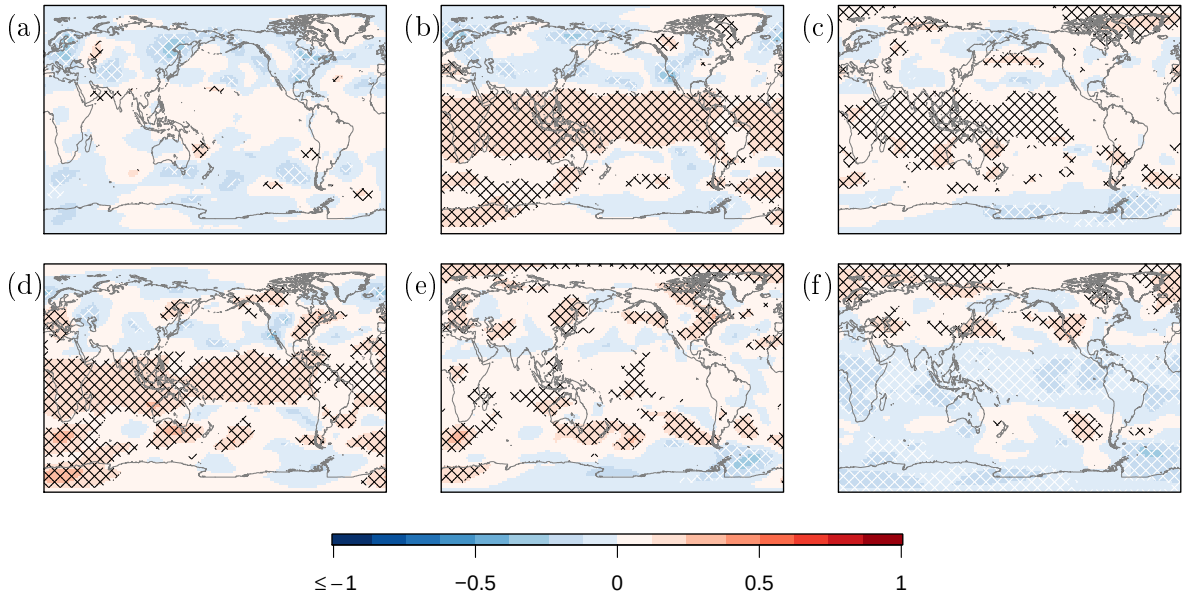

Figure A11: *Same as in A8 but for Z at 500 hPa for lead year 7-10.*

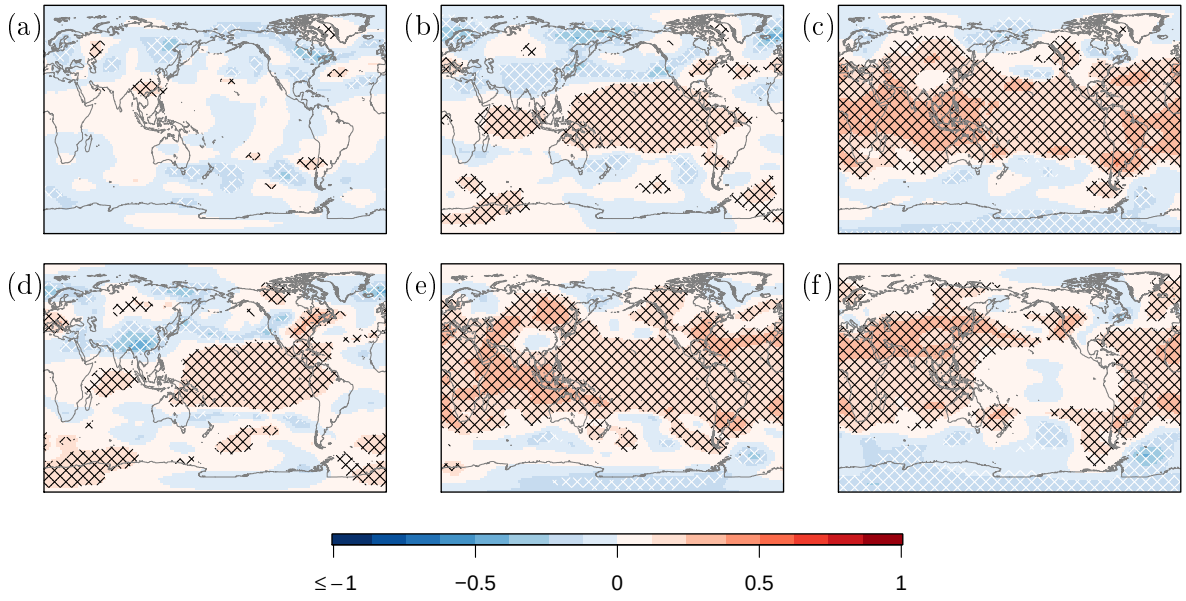

Figure A12: *Same as in A8 but for Z at 200 hPa for lead year 7-10.*

A.2 Additional figures for temperature (T)

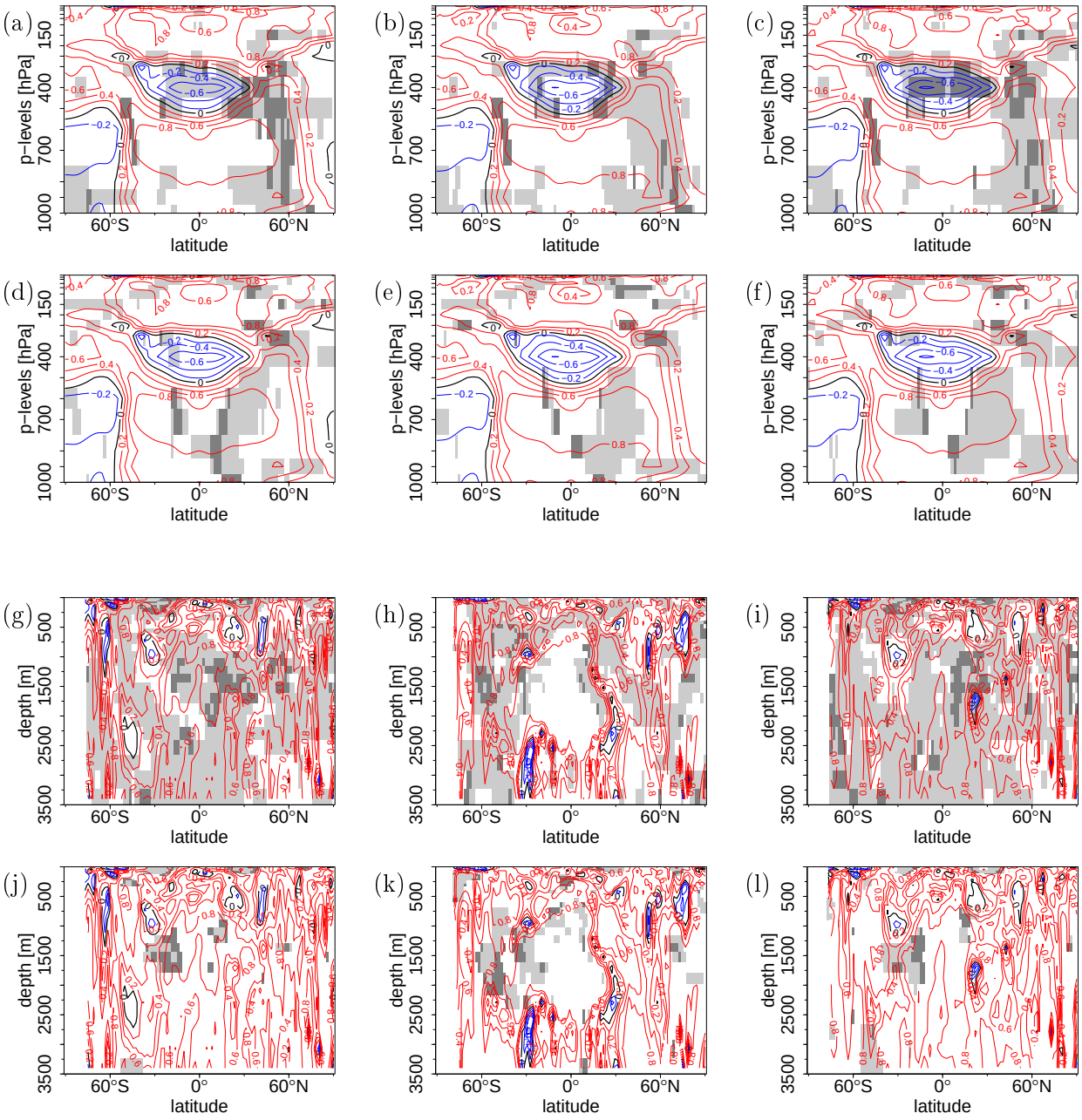

Figure A13: *Reliability classifications and MSESS for zonally averaged T in atmosphere and ocean with terciles as thresholds. The results are shown for atmosphere (a-f) and ocean (g-l), where lower terciles (a-c,g-i) and upper terciles (d-f, j-l) are used as thresholds. b1-LR is basis in subfigures a, d, g, j, pr-GECCO is basis in subfigures b, e, h, k, and pr-ORA is basis in c, g, i, l. The results are based on lead year 2-5. The contour lines display positive (red), negative (blue) and zero (black) MSESS values. Dark gray shadings correspond to reliable areas, light gray shadings to potentially useful areas and white areas to not useful areas.*

Appendix B

Part II

B.1 Additional information for spatial reconstructions

B.1.1 List of coring sites

Table B.1: *Selected taxa for 6 ka BP, which are used for spatial reconstructions (Simonis et al., 2012): second column represents the name of the coring site; third and fourth column represent the longitude and latitude information in [°]; fifth column represents the altitude in [m]; sixth column represents the selected taxa. The longitude of coring site 7 (Laihalampi) is corrected.*

nr	site	lon	lat	alt	selected taxa
1	Dalmutladdo	20.5	69.5	355	Alnus, Betula, Fraxinus excelsior, Juniperus communis, Thalictrum aquilegifolium, Selaginella selaginoides, Picea abies
2	Toskaljavri	21.5	69.5	704	Juniperus communis, Pinus sylvestris, Calluna vulgaris
3	Lake Tsuolbmajavri	22.5	68.5	526	Betula, Pinus sylvestris, Picea abies
4	Abbortjärnen	14.5	63.5	250	Alnus, Fraxinus excelsior, Juniperus communis, Pinus sylvestris
5	Brurskardtjørni	8.5	61.5	1,310	Betula, Ranunculus acetosella, Rumex acetosella
6	Klotjärnen	16.5	61.5	160	Betula, Fraxinus excelsior, Hippophae rhamnoides, Juniperus communis, Calluna vulgaris, Potamogeton, Rumex acetosa
7	Laihalampi	26.5	61.5	137	Juniperus communis, Pinus sylvestris, Populus tremula, Tilia, Ulmus, Picea abies
8	Trettetjørn	7.5	60.5	819	Juniperus communis, Pinus sylvestris, Quercus deciduous, Calluna vulgaris Ranunculus acetosella

9	Holtjärnen	14.5	60.5	110	Alnus, Juniperus communis, Pinus sylvestris, Quercus deciduous, Tilia cordata, Calluna vulgaris, Picea abies
10	Hirvilampi	24.5	60.5	114	Cornus mas, Corylus avellana, Sambucus, Tilia, Calluna vulgaris, Filipendula, Picea abies
11	Vestre	6.5	59.5	570	Alnus, Juniperus communis, Myrica gale, Pinus sylvestris, Populus tremula, Quercus deciduous
12	Flarken	13.5	58.5	110	Betula, Corylus avellana, Tilia, Calluna vulgaris, Potamogeton, Rumex acetosa
13	Raigastvere	26.5	58.5	52	Betula, Quercus deciduous, Tilia, Picea abies
14	Sämbosjön	12.5	57.5	35	Betula, Quercus deciduous, Tilia, Viburnum, Calluna vulgaris, Plantago lanceolata
15	Loch Maree	-5.5	57.5	107	Alnus glutinosa Fraxinus excelsior, Juniperus communis, Pinus sylvestris. Calluna vulgaris, Plantago major, Rumex acetosa
16	Dubh Lochan	-4.5	56.5	75	Betula, Calluna vulgaris, Nymphaea alba, Potamogeton natans
17	Machrie Moor	-4.5	55.5	50	Cladium mariscus, Alnus, Betula, Quercus deciduous
18	Sluggan Moss	-6.5	54.5	52	Alnus, Betula, Quercus deciduous
19	Zarnowiec	17.5	54.5	5	Alnus, Quercus deciduous, Tilia, Potamogeton gramineus, Sparganium minimum, Uritca dioica
20	Maly Suszek	17.5	53.5	115	Acer, Alnus, Betula, Carpinus betulus, Quercus deciduous, Calluna vulgaris, Picea abies
21	Stare Biele	23.5	53.5	143	Alnus, Carpinus betulus, Juniperus communis, Picea abies, Quercus deciduous, Tilia cordata, Calluna vulgaris, Filipendula
22	King's Pool	-2.5	52.5	100	Betula, Quercus deciduous, Tilia, Plantago lanceolata, Potamogeton, Rumex acetosa
23	Hockam Mere	0.5	52.5	33	Alnus glutinosa, Betula, Fraxinus excelsior, Hedera helix, Sambucus nigra, Taxus baccata, Tilia cordata, Plantago lanceolata, Potamogeton

24	Treppelsee	14.5	52.5	52	Alnus, Betula, Quercus deciduous, Calluna vulgaris, Thalictrum aquilegifolium, Picea abies
25	Skrzetuszewskie	17.5	52.5	109	Alnus, Betula, Quercus deciduous, Tilia, Calluna vulgaris, Picea abies
26	Lake Gosciaz	19.5	52.5	64	Betula, Fagus sylvatica, Fraxinus excelsior, Hedera helix, Picea abies, Salix pentandra, Taxus baccata, Tilia cordata, Viburnum opulus, Calluna vulgaris, Cladium mariscus
27	Bledowo	20.5	52.5	78	Corylus avellana, Fagus sylvatica, Hedera helix, Picea abies, Tilia, Calluna vulgaris, Filipendula, Plantago lanceolata, Rumex acetosa, Typha latifolia
28	Lukcze	22.5	51.5	163	Corylus avellana, Fagus sylvatica, Juniperus communis, Picea abies, Tilia, Calluna vulgaris, Rumex acetosa, Rumex acetosella, Typha latifolia
29	Meerfelder Maar, Hitsche	6.5	50.5	335	Acer, Alnus, Betula, Hedera helix, Quercus deciduous, Myriophyllum alterniflorum
30	Svarcenberk, Rezabinec	14.5	49.5	400	Alnus, Abies alba, Quercus deciduous, Tilia, Filipendula
31	Malopolskie	19.5	49.5	656	Abies alba, Betula nana, Fagus sylvatica, Fraxinus excelsior, Hedera helix, Picea abies, Pinus sylvestris, Calluna vulgaris, Plantago lanceolata, Plantago major, Rumex acetosa, Thalictrum aquilegifolium
32	Tarnowiec, Besko, Rostoki	21.5	49.5	240	Abies alba, Acer, Alnus, Fraxinus excelsior, Picea abies, Rubus idaeus, Tilia platyphyllos, Calluna vulgaris, Filipendula, Potamogeton natans, Thalictrum aquilegifolium
33	Feigne d'Artimont	7.5	48.5	1,100	Abies alba, Alnus, Betula, Hedera helix, Quercus deciduous, Tilia, Plantago lanceolata
34	Lobsigensee, Loermoss	7.5	47.5	550	Abies alba, Betula, Carpinus betulus, Fraxinus excelsior, Tilia, Ranunculus acetosella, Picea abies

35	Nussbaumerseen, Rotsee, Breitnau, Soppensee, Steerenmoss	8.5	47.5	610	Abies alba, Pinus sylvestris, Quercus deciduous, Tilia, Calluna vulgaris, Ranunculus acetosella, Scheuchzeria palustris, Plantago lanceolata, Potamogeton, Picea abies
36	Ried bei Oberschan	9.5	47.5	640	Abies alba, Betula, Corylus avellana, Fagus sylvatica, Tilia, Potamogeton, Picea abies
37	Fuschlsee	13.5	47.5	663	Abies alba, Betula, Quercus deciduous, Tilia cordata, Tilia platyphyllos, Plantago major, Ranunculus acetosella, Picea abies
38	Steregoiu	23.5	47.5	1,300	Corylus avellana, Fagus sylvatica, Hedera helix, Tilia, Rumex acetosella, Picea abies
39	Le Tronchet, Marais du Rosey	6.5	46.5	600	Abies alba, Alnus, Pinus sylvestris, Tilia, Picea abies
40	Lac du Bouchet	3.5	44.5	1,200	Alnus, Betula, Corylus avellana, Hedera helix, Quercus ilex, Tilia, Polygonum aviculare
41	Lac du Saint	6.5	44.5	1,308	Abies alba, Acer, Alnus, Hedera helix, Quercus ilex, Nymphaea alba, Filipendula
42	Lago Padula	10.5	44.5	1,187	Abies alba, Acer, Alnus, Corylus avellana
43	Lake Vrana	15.5	44.5	13	Abies alba, Alnus, Betula, Fraxinus excelsior, Quercus ilex, Tilia, Picea abies
44	Biscaye	-0.5	43.5	410	Corylus avellana, Quercus ilex, Viburnum, Filipendula, Plantago lanceolata, Thalictrum aquilegifolium,
45	Las Pardillas	-3.5	42.5	1,850	Betula, Quercus ilex, Ulmus, Potamogeton natans
46	Malo Jezero	17.5	43.5	0	Acer, Alnus, Carpinus betulus, Carpinus orientalis, Fraxinus excelsior, Fraxinus ornus, Pinus sylvestris, Quercus ilex, Quercus pubescens, Chenopodium glaucum
47	Lake Dalgoto, Lake Ribno	23.5	41.5	2,310	Carpinus betulus, Carpinus orientalis, Corylus avellana, Tilia, Filipendula, Plantago lanceolata, Picea abies
48	Lago di Monticchio	15.5	40.5	1,326	Abies alba, Carpinus betulus, Carpinus orientalis, Fraxinus ornus, Hedera helix, Tilia, Najas marina, Najas minor, Nymphaea alba

49	Canada de la Cruz	-2.5	38.5	1,350	Acer, Betula, Corylus avellana, Quercus evergreen, Potamogeton
50	Antas	-1.5	37.5	0	Betula, Quercus evergreen
51	San Rafael	-2.5	37.5	0	Alnus, Corylus avellana, Phillyrea, Quercus evergreen

B.1.2 Additional figures for spring and autumn temperatures

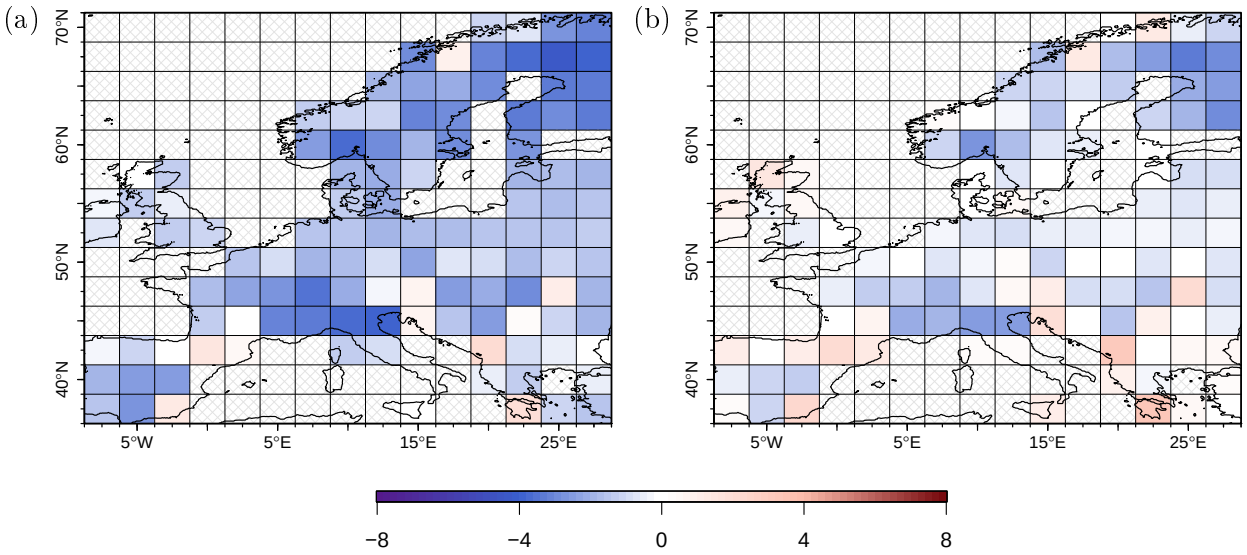

Figure B1: *Difference between CRU climatic mean (1901-1913) and PMIP3 multi-model mean (6 ka BP) for (a) spring and (b) autumn 2 m temperatures in K.*

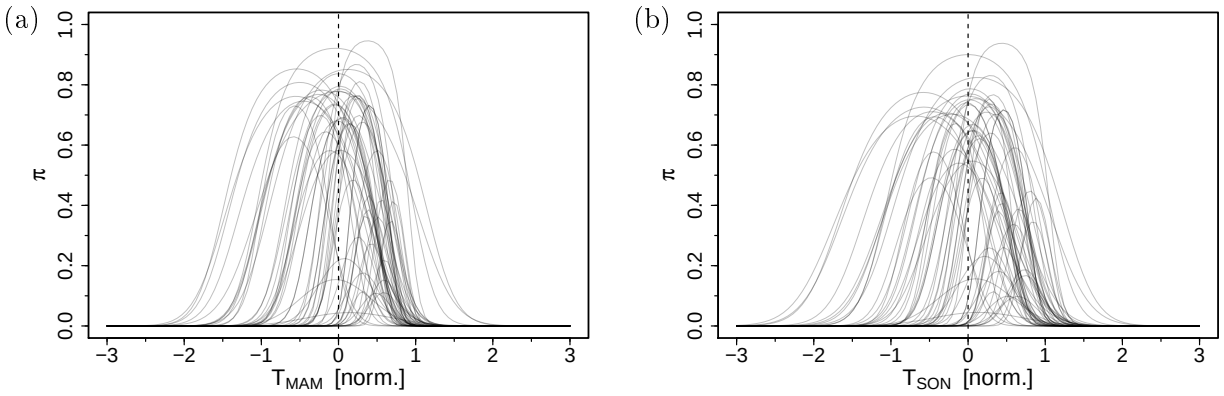

Figure B2: *GLM-probabilities for 59 occuring taxa for spring (a) and autumn (b) temperatures.*

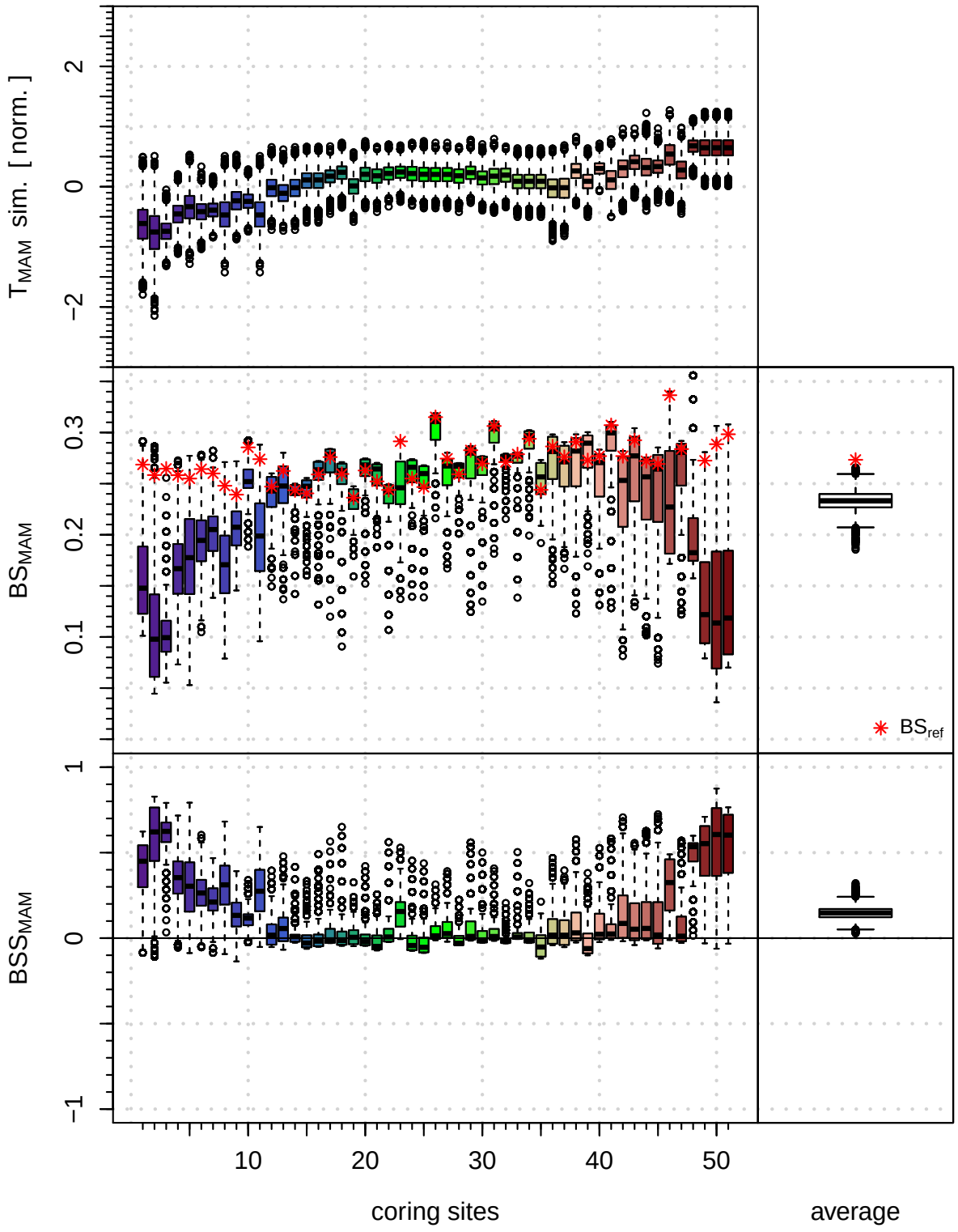

Figure B3: *Boxplots of the generated spring temperatures (upper panel), corresponding Brier scores (BS) with the reference BS denoted by red stars (middle panel) and Brier skill scores (BSS, lower panel) dependent on the 51 coring sites. The averages over all coring sites are presented next to BS and BSS.*

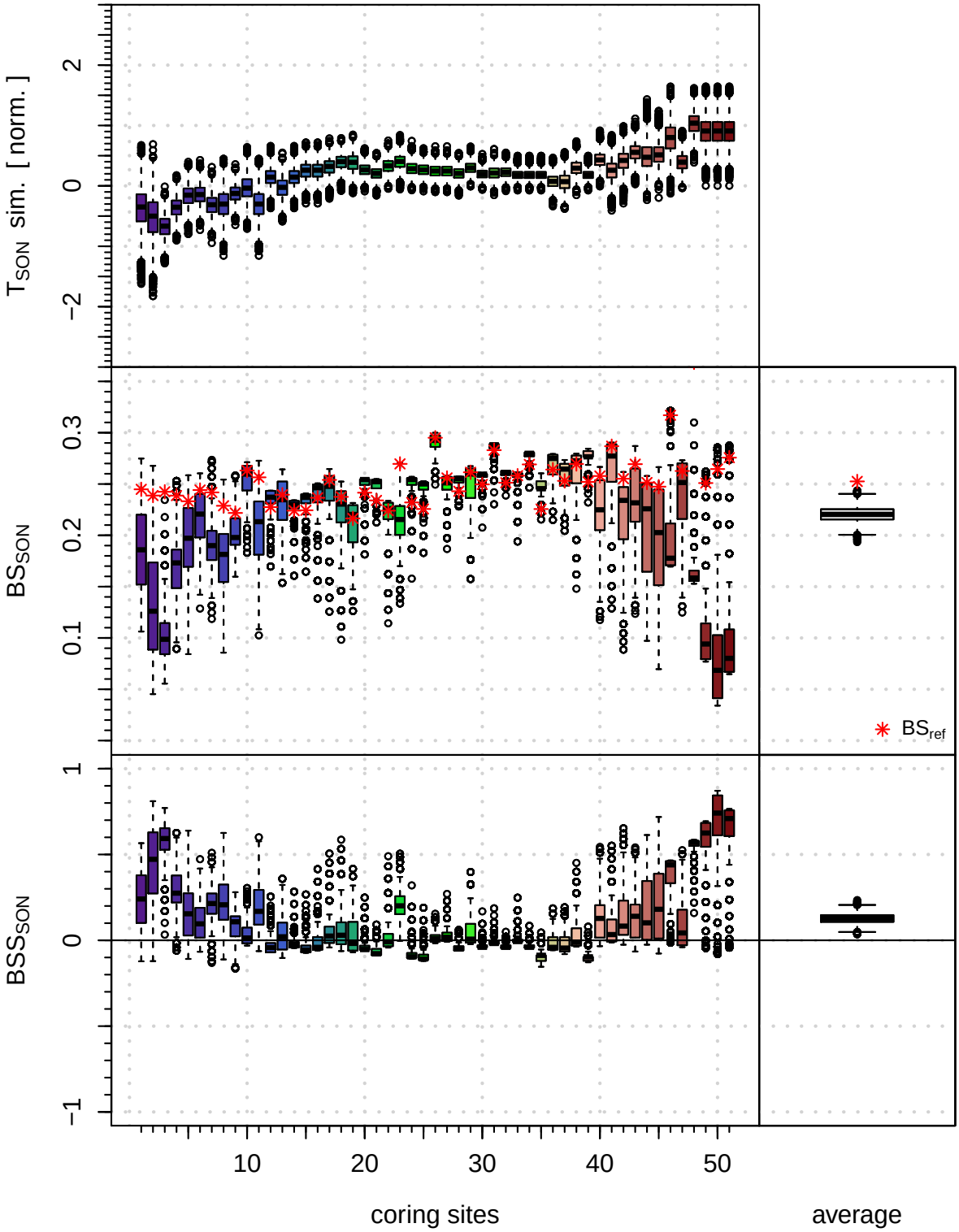

Figure B4: *Same graphic as in B3, but for autumn temperatures.*

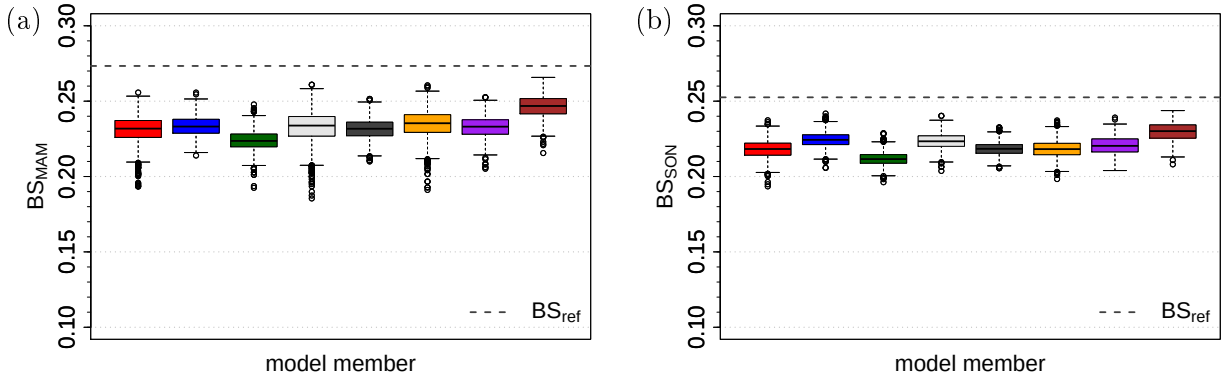

Figure B5: *Brier score (BS) in dependence on the ensemble members, averaged over all coring sites, for spring (a) and autumn (b). The boxplots are represented in a random order. The dashed line denotes the reference BS, which is independent of the models.*

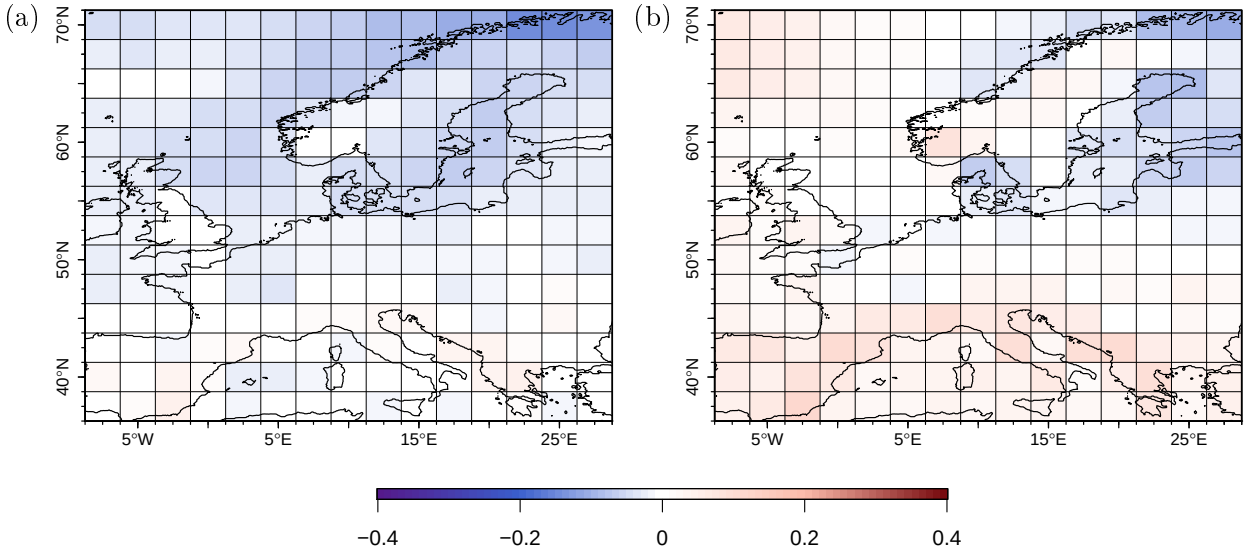

Figure B6: *Differences between assimilated 2 m temperatures and PMIP3 multi-model mean for spring (a) and autumn (b) temperatures in K.*

B.2 Biome compositions and additional figures

Table B.2: *Biome occurrences for phase 1 (147-89 ka BP) and phase 2 (89-9 ka BP). First column represents the name of the taxon, second column represents the taxon abbreviation, third column represents the biome (M: Mediterranean biome; S: Saharo-Arabian biome; I: Irano-Turanian biome), and fourth/fifth column represent the occurrence of the taxon for phase 1/phase 2.*

taxon	taxon acronym	biome	phase 1	phase 2
Quercus ithaburensis type	Qu.ith	M	✓	✓
Quercus calliprinos type	Qu.call	M	✓	✓
Pistacia	Pista	M	✓	✓
Olea europaea	Olea	M	✓	✓
Phillyrea	Phillyre	M	✓	✓
Pinus	Pinus	M	✓	✓
Cedrus	Cedrus	M	✓	✓
Cupressaceae	Cupressa	M	✓	✓
Phoenix	Phoenix	S	✓	
Tamarix	Tamarix	S	✓	✓
Ziziphus Spina-christi	Ziziphus	S	✓	
Acacia	Acacia	S	✓	
Cerealea type	Cereal	I	✓	✓
Poaceae	Poa	I	✓	✓
Artemisia	Artem	I	✓	✓
Ephedra	Ephedra.f	I	✓	✓
Chenopodiaceae	Chenop	S	✓	✓
Tubuliflorae	Tubul	I	✓	✓
Sarcopoterium Spinosum	Sarcopot	M		✓
Cistus type	Cistus	M	✓	✓
Helianthemum	Helianth	M	✓	✓
Centaurea	Centaure.j	I	✓	✓
Zygophyllum	Zygophyl	S	✓	
Liguliflorae	Ligulif	I	✓	✓

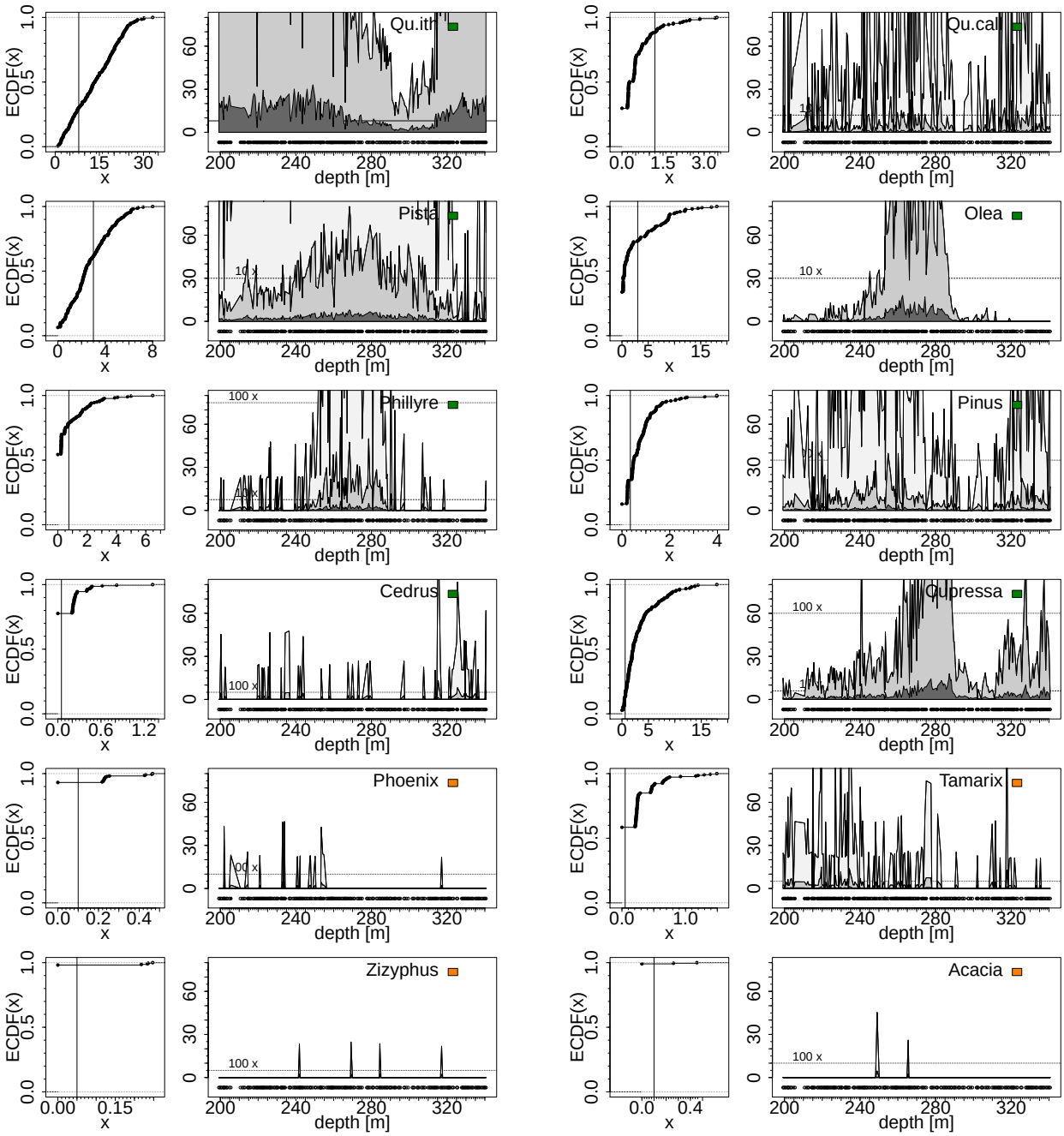

Figure continues on next page.

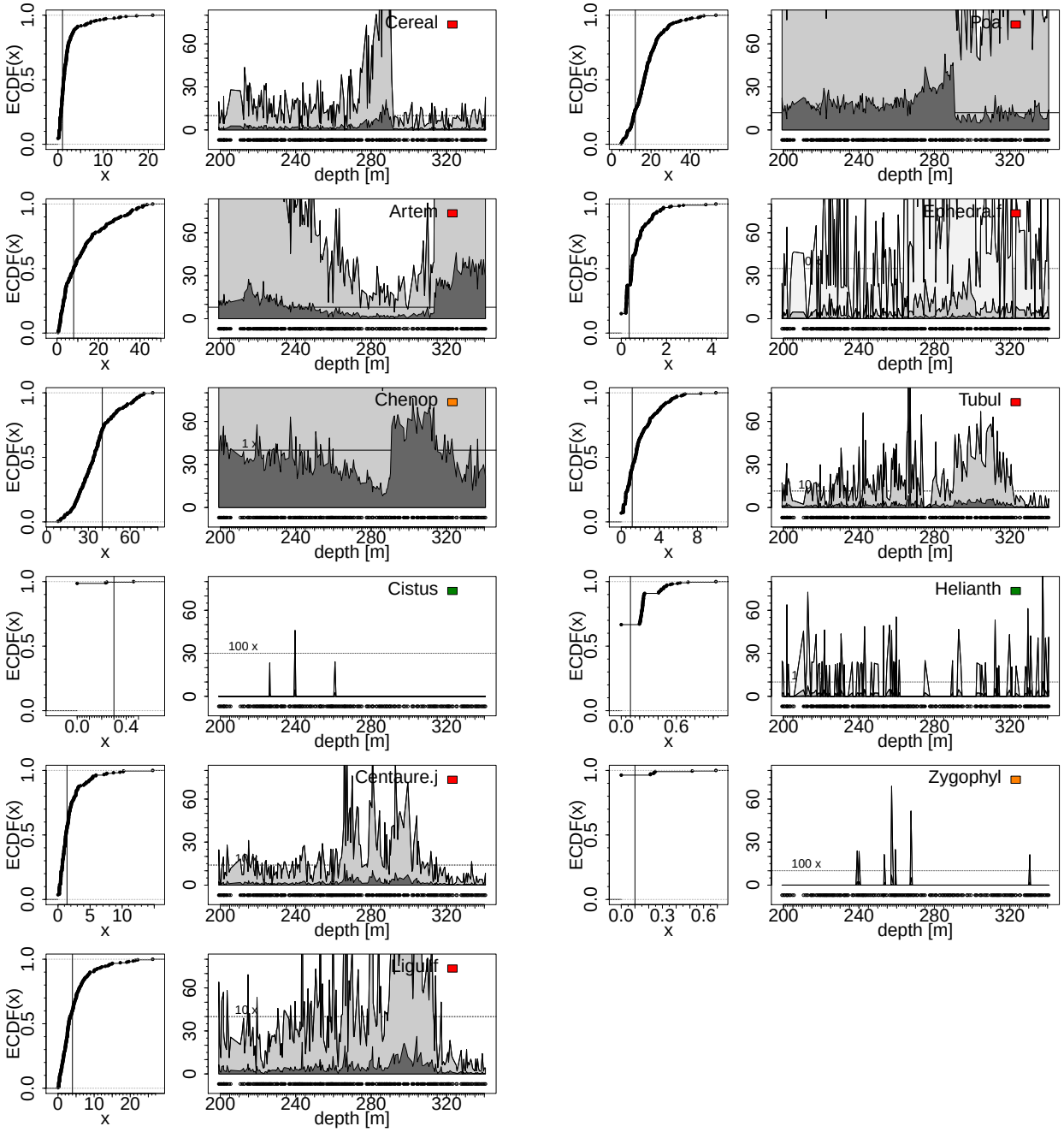

Figure B7: *Empirical cdfs (ECDFs) including individual threshold values (left) and depth profiles of pollen spectra for occurring pollen (right) during phase 1. The acronyms are listed in Table B.2.*

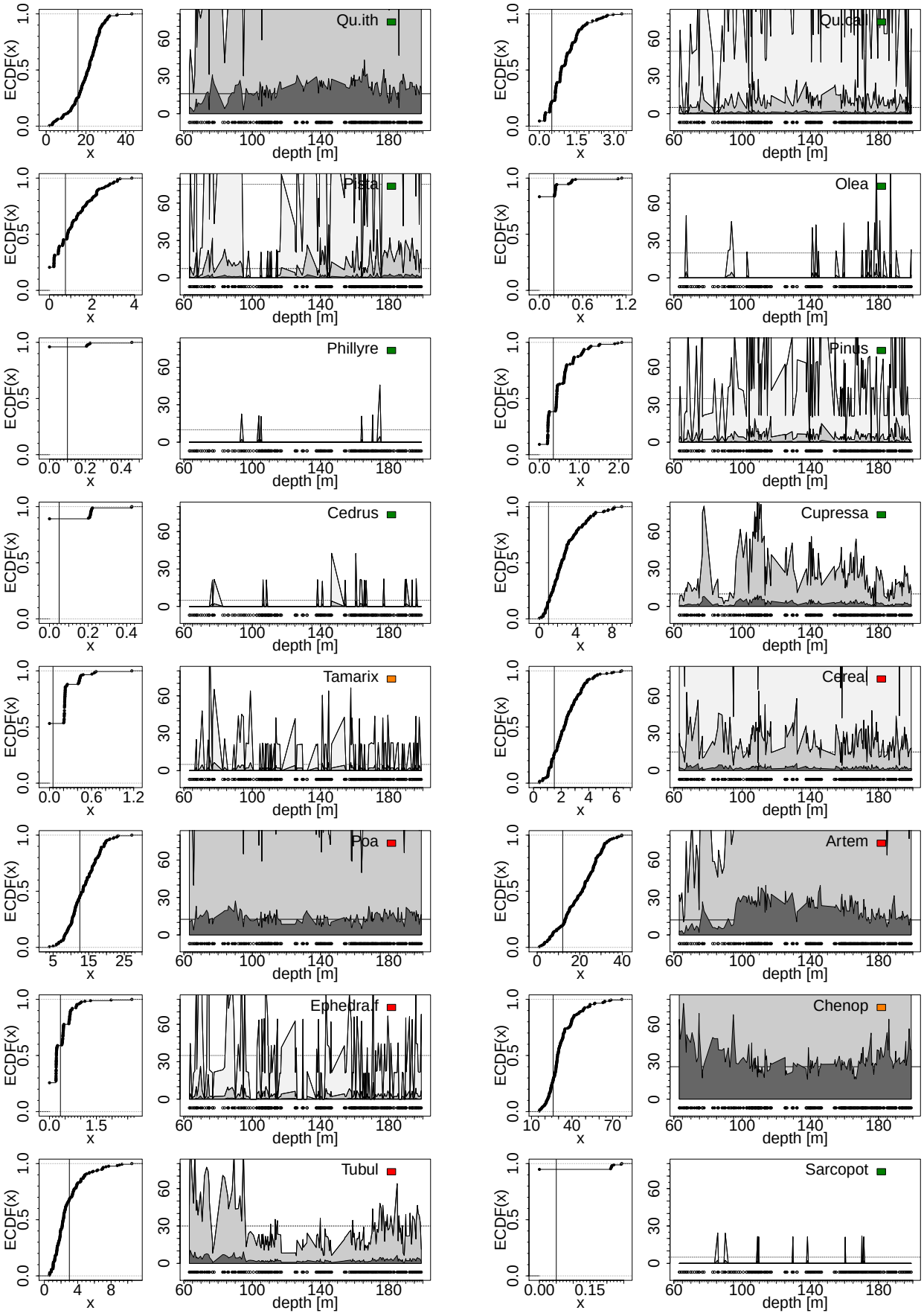

Figure continues on next page.

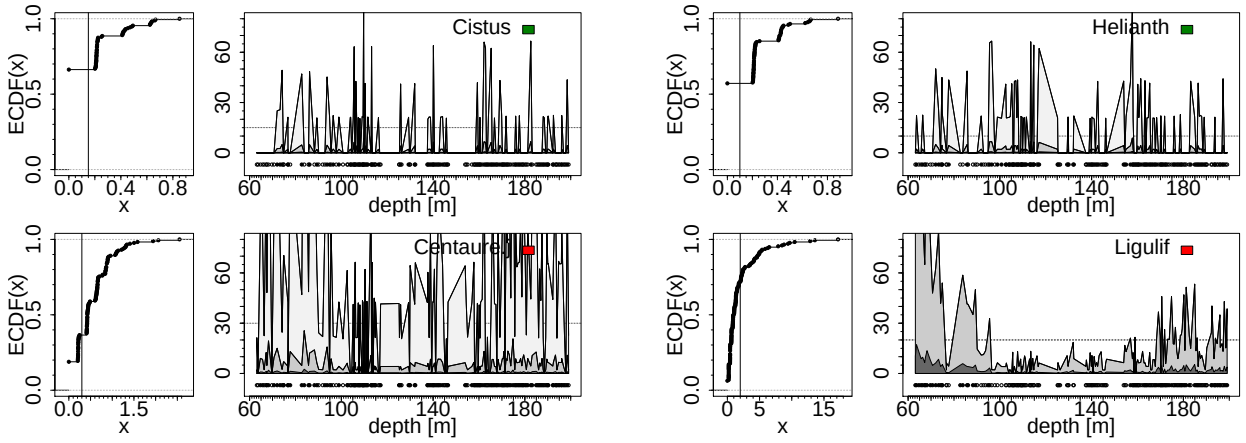

Figure B8: *Empirical cdfs (ECDFs) including individual threshold values (left) and depth profiles of pollen spectra for occurring pollen (right) during phase 2. The acronyms are listed in Table B.2.*

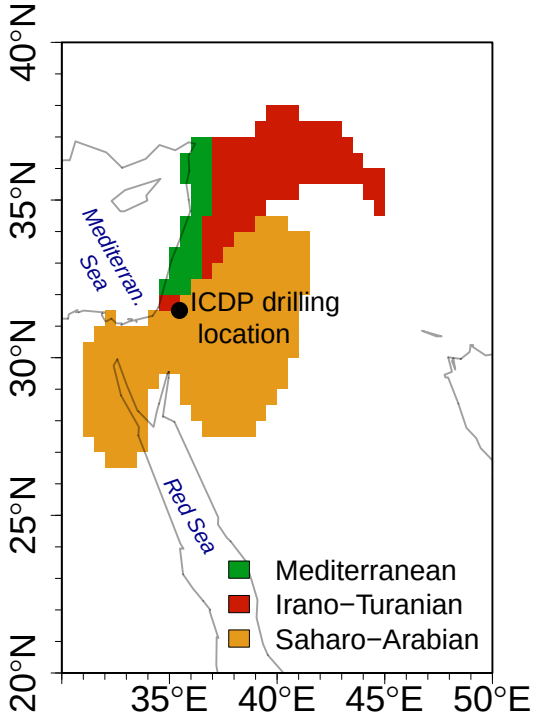

Figure B9: *Distributions of the Mediterranean, the Irano-Turanian and Sahara-Arabian biomes after Meusel et al. (1965).*

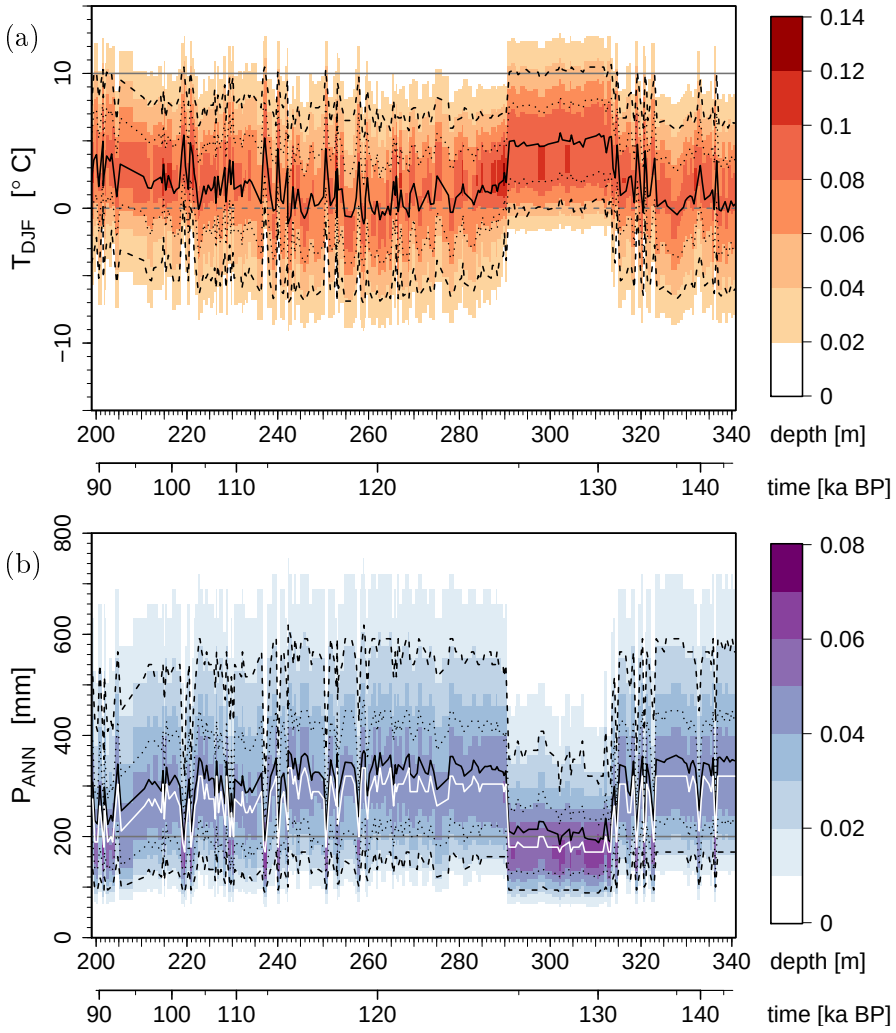

Figure B10: *Posterior probability density function for winter temperature (a) and annual precipitation (b) for phase 1. The solid black line denote the mean, the white solid line in (b) the median, the doted lines the 25%- and 75%-quantiles and the dashed lines the 10%- and 90%-quantiles. The reconstructions are based on the biome distribution after Meusel et al. (1965).*

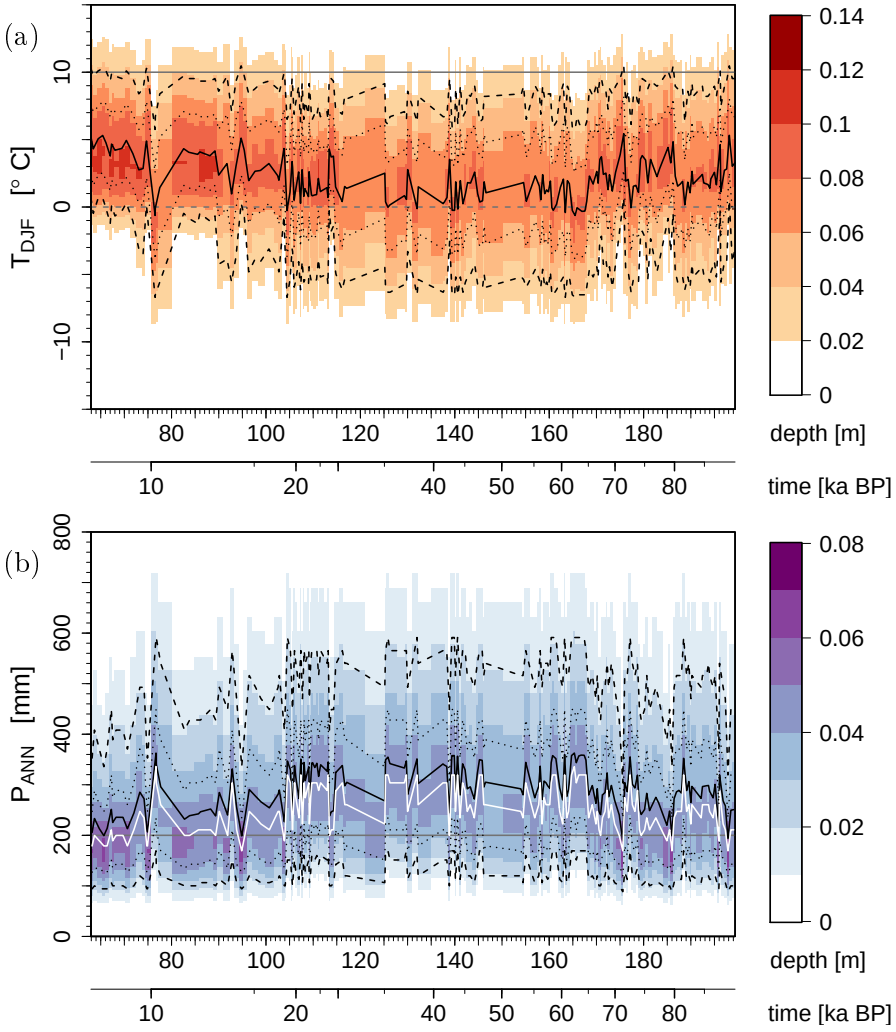

Figure B11: *Same as in Figure B10 but for phase 2.*

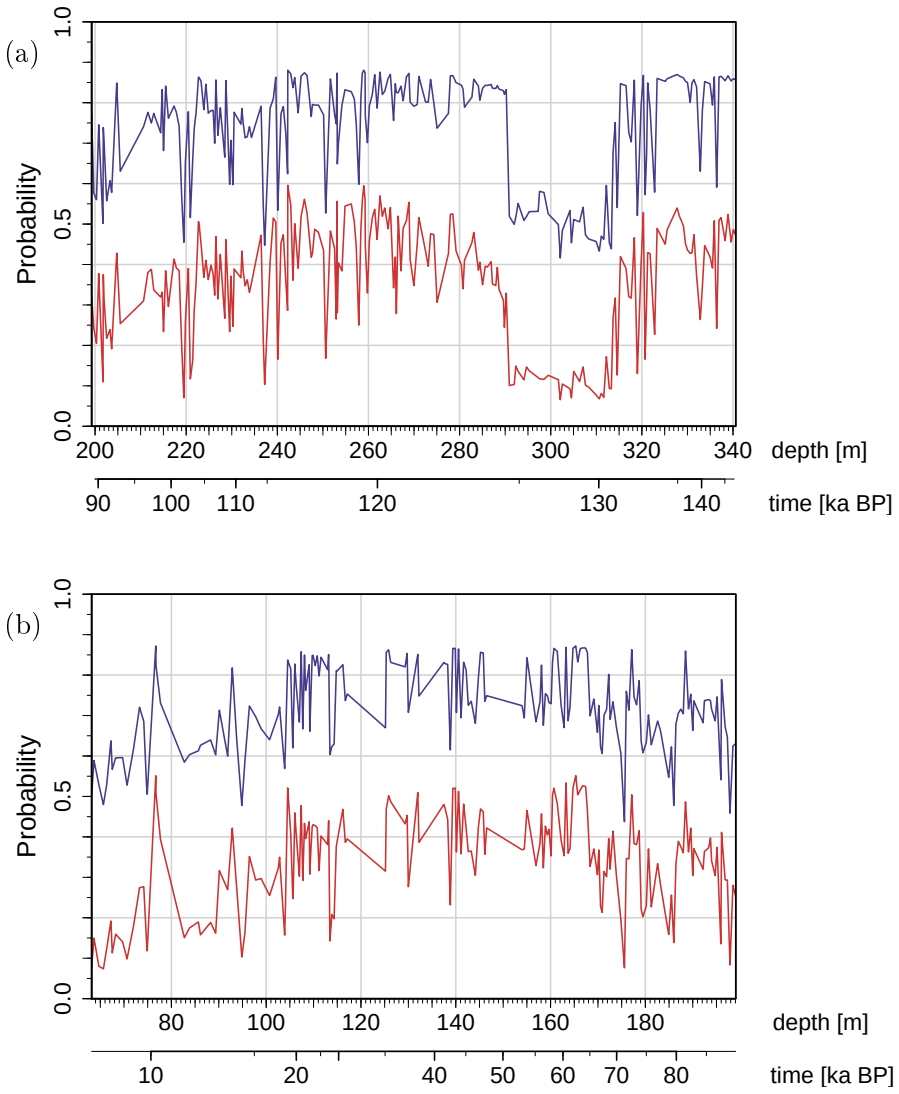

Figure B12: *Summarizing representation of the statistical climate reconstruction for phase 1 (a) and phase 2 (b) based on the original biome distribution after Meusel et al. (1965). The blue line displays the probability for annual precipitation being above 200 mm/year, and the red line displays the probability for the reconstructed winter temperature laying below 0°C. The unlabeled ticks on the time axis denote 5-year intervals.*

Bibliography

Alley, R. (2016). A heated mirror for future climate. *Science*, 352(6282):151–152.

Anderson, J. L. (1996). A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations. *Journal of Climate*, 9(7):1518–1530.

Ayalon, A., Bar-Matthews, M., Frumkin, A., and Matthews, A. (2013). Last Glacial warm events on Mount Hermon: the southern extension of the Alpine karst range of the east Mediterranean. *Quaternary Science Reviews*, 59:43–56.

Balmaseda, M. A., Mogensen, K., and Weaver, A. T. (2013). Evaluation of the ECMWF ocean reanalysis system ORAS4. *Quarterly Journal of the Royal Meteorological Society*, 139(674):1132–1161.

Bar-Matthews, M., Ayalon, A., Gilmour, M., Matthews, A., and Hawkesworth, C. (2003). Sea-land oxygen isotopic relationships from planktonic foraminifera and speleothems in the Eastern Mediterranean region and their implication for paleorainfall during interglacial intervals. *Geochimica et Cosmochimica Acta*, 67(17):3181–3199.

Bentor, Y. K. (1961). Some geochemical aspects of the Dead Sea and the question of its age. *Geochimica et Cosmochimica Acta*, 25(4):239–260.

Bothe, O., Jungclaus, J., and Zanchettin, D. (2013). Consistency of the multi-model CMIP5/PMIP3-past100 ensemble. *Climate of the Past*, 9:2471–2487.

Braconnot, P., Harrison, S., Otto-Bliesner, B., Abe-Ouchi, A., Jungclaus, J., and Peterschmitt, J. (2011). The Paleoclimate Modeling Intercomparison Project Contribution to CMIP5. *CLIVAR Exchanges No. 56*, 16(2):15–19.

Brier, G. (1950). Verification of forecasts expressed in terms of reliabilty. *Monthly Weather Review*, 78(1):1–3.

Bröcker, J. and Smith, L. A. (2007). Increasing the reliability of reliability diagrams. *Weather Forecasting*, 22(3):651–661.

Candille, G. and Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, 131:2131–2150.

CDO (2015). *Climate Data Operators*. Availabe at: http://www.mpimet.mpg.de/cdo.

Corti, S., Palmer, T., Balmaseda, M., Weisheimer, A., Drijfhout, S., Dunstone, N., Hazeleger, W., Kröger, J., Pohlmann, H., Smith, D., von Storch, J.-S., and Wouters, B. (2015). Impact of Initial Conditions versus External Forcing in Decadal Climate Predictions: A Sensitivity Experiment. *Journal of Climate*, 28:4454–4470.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., Berg, L. v. d., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., Rosnay, P. d., Tavolato, C., Thépaut, J.-N., and Vitart, F. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597.

Dunstone, N., Smith, D., Scaife, A., Hermanson, L., Eade, R., Robinson, N., Andrews, M., and Knight, J. (2016). Skilfull predictions of the winter North Atlantic Oscillation one year ahead. *Nature Geoscience*, 9:809–814.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models*. Springer New York.

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

Gebhardt, C., Kühl, N., Hense, A., and Litt, T. (2008). Reconstruction of Quaternary temperature fields by dynamically consistent smoothing. *Climate dynamics*, 30(4):421–437.

Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Böttinger, M., Brovkin, V., Crueger, T., Esch, M., Fieg, K., Glushak, K., Gayler, V., Haak, H., Hollweg, H.-D., Ilyina, T., Kinne, S., Kornblueh, L., Matei, D., Mauritsen, T., Mikolajewicz, U., Müller, W., Notz, D., Pithan, F., Raddatz, T., Rast, S., Redler, R., Roeckner, E., Schmidt, H., Schnur, R., Segschneider, J., Six, K. D., Stockhause, M., Timmreck, C., Wegner, J., Widmann, H., Wieners, K.-H., Claussen, M., Marotzke, J., and Stevens, B. (2013). Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. *Journal of Advances in Modeling Earth Systems*, 5(3):572–597.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17(2):211–235.

Goddard, L., Hurrell, J. W., Kirtman, B. P., Murphy, J., Stockdale, T., and Vera, C. (2012). Two time scales for the price of one (almost). *Bulletin of the American Meteorological Society*, 93(5):621–629.

Goddard, L., Kumar, A., Solomon, A., Smith, D., Boer, G., Gonzalez, P., Kharin, V., Merryfield, W., Deser, C., Mason, S., Kirtman, B., Msadek, R., Sutton, R., Hawkins, E., Fricker, T., Hegerl, G., Ferro, C., Stephenson, D., Meehl, G., Stockdale, T., Burgman, R., Greene, A., Kushnir, Y., Newman, M., Carton, J., Fukumori, I., and Delworth, T. (2013). A verification framework for interannual-to-decadal predictions experiments. *Climate Dynamics*, 40(1-2):245–272.

Grichuk, V. (1969). An attempt to reconstruct certain elements of the climate of the northern hemisphere in the Atlantic period of the Holocene. *Golotsen. Izd-vo Nauka, Moscow*, pages 41–57.

Grimit, E., Gneiting, T., Berrocal, V., and Johnson, N. (2006). The Continuous Ranked Probability Score for Circular Variables and its application to Mesoscale Forecast Ensemble Verification. *Quarterly Journal of the Royal Meteorological Society*, 132:2925–2942.

Guillot, D., Rajaratnam, B., and Emile-Geay, J. (2015). Statistical paleoclimate reconstructions via Markov random fields. *The Annals of Applied Statistics*, 9(1):324–352.

Hamill, T., Whitaker, J., and Mullen, S. (2006). Reforecasts: An important dataset for improving weather predictions. *Bulletin of the American Meteorological Society*, 87(1):33–46.

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

Harrison, S., Bartlein, P., Izumi, K., Li, G., Annan, J., Hargreaves, J., Braconnot, P., and Kageyama, M. (2015). Evaluation of CMIP5 palaeo-simulations to improve climate projections. *Nature Climate Change*, 7:735–742.

Hense, A. (2005). Processing of observational data and its implication for climate analysis. In Hantel, M., editor, *Observed Global Climate*, volume 6 of *Landolt-Börnstein - Group V Geophysics*, pages 1–11. Springer Berlin Heidelberg.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecasting*, 15(5):559–570.

Hewitt, C., Buontempo, C., and Newton, P. (2013). Using climate predictions to better serve society's needs. *EOS*, 94(11):105–107.

Houghton, J., Jenkins, G., and Ephraums, J., editors (1990). *Climate Change: The IPCC Scientific Assessment*. Cambridge University Press, Cambridge.

Jolliffe, I. and Stephenson, D. (2012). *Forecast Verification: A Practitioner's Guide in Atmospheric Science, 2nd Ed.* Wiley, Oxford.

Kadow, C., Illing, S., Kunst, O., Rust, H. W., Pohlmann, H., Müeller, W. A., and Cubasch, U. (2016). Evaluation of Forecasts by Accuracy and Spread in the MiKlip Decadal Climate Prediction System. *Meteorologische Zeitschrift*, 25(6):631–643.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., W.Higgins, Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Jenne, R., and Joseph, D. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–471.

Karl, T., Arguez, A., Huang, B., Lawrimore, J., McMahon, J., Menne, M., Peterson, T., Vose, R., and Zhang, H.-M. (2015). Possible artefacts of data biases in the recent global surface warming hiatus. *Science*, 348(6242).

Kaspar, F., Kühl, N., Cubasch, U., and Litt, T. (2005). A model-data comparison of European temperatures in the Eemian interglacial. *Geophysical Research Letters*, 32:L11703.

Keenlyside, N. S., Latif, M., Jungclaus, J., Kornblueh, L., and Roeckner, E. (2008). Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature*, 453(7191):84–88.

Keller, J. D. and Hense, A. (2011). A new non-Gaussian evaluation method for ensemble forecasts based on analysis rank histograms. *Meteorologische Zeitschrift*, 20(2):107–117.

Keune, J., Ohlwein, C., and Hense, A. (2014). Multivariate probabilistic analysis and predictability of medium-range ensemble weather forecasts. *Monthly Weather Review*, 142(11):4074–4090.

Köhl, A. (2015). Evaluation of the GECCO2 ocean synthesis: transports of volume, heat and freshwater in the Atlantic. *Quarterly Journal of the Royal Meteorological Society*, 141:166–181.

Köhl, A. and Stammer, D. (2008). Variability of the meridional overturning in the North Atlantic from the 50-year GECCO state estimation. *Journal of Physical Oceanography*, 38(9):1913–1930.

Kruschke, T., Rust, H. W., Kadow, C., Müller, W. A., Pohlmann, H., Leckebusch, G. C., and Ulbrich, U. (2016). Probabilistic evaluation of decadal predictions of Northern Hemisphere winter storms. *Meteorologische Zeitschrift*, 25(6):721–738.

Kühl, N., Gebhardt, C., Litt, T., and Hense, A. (2002). Probability density functions as botanical-climatological transfer functions for climate reconstruction. *Quaternary Research*, 58(3):381–392.

Laskar, J., Robutel, P., Joutel, F., Gastineau, M abd Correia, A., and Levrard, B. (2004). A long-term numerical solution for the insolation quantities of the Earth. *Astronomy & Astrophysics*, 428:261–285.

Latif, M., Claussen, M., Schulz, M., and Brücher, T. (2016). Comprehensive Earth system models of the last glacial cycle. *EOS*, 97.

Li, B., Nychka, D., and Ammann, C. (2007). The value of multi-proxy reconstruction of past climate (with discussions and rejoinder).

Lisiecki, L. and Raymo, M. (2005). A Pliocene-Plistocene stack of 57 globally distributed benthic δ^{18}O records. *Paleoceanography*, 20. PA1003, doi:10.1029/2004PA001071.

Litt, T., Anselmetti, F. S., Baumgarten, H., Beer, J., Cagatay, N., Cukur, D., Damci, E., Glombitza, C., Haug, G., Heumann, G., Kallmeyer, H., Kipfer, R., Krastel, S., Kwiecien, O., Meydan, A. F., Orcen, S., Pickarski, N., Randlett, M.-E., Schmincke, H.-U., Schubert, C. J., Sturm, M., Sumita, M., Stockhecke, M., Tomonaga, Y., Vigliotti, L., Wonik, T., and the PALEOVAN Scientific Team (2012a). 500,000 Years of Environmental History in Eastern Anatolia: The PALEOVAN Drilling Project. *Scientific Drilling*, 14:18–29.

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

Lohmann, G., Pfeiffer, M., Laepple, T., Leduc, G., and Kim, J.-H. (2013). A model-data comparison of the Holocene global sea surface temperature evolution. *Climate of the Past*, 9:1807–1839.

Ludwig, P., Schaffernicht, E., Shao, Y., and Pinto, J. (2016). Regional atmospheric circulation over Europe during the Last Glacial Maximum and its links to precipitation. *Journal of Geophysical Research: Atmospheres*, 121(5):2130–2145.

Madec, G. (2008). NEMO reference manual, ocean dynamics component: NEMO-OPA. Preliminary Version. *Note du Pôle de modélisation, Institute Pierre-Simon Laplace (IPSL), France*, (27).

Marotzke, J., Müller, W. A., Vamborg, F. S. E., Becker, P., Cubasch, U., Feldmann, H., Kaspar, F., Kottmeier, C., Marini, C., Polkova, I., Prömmel, K., Rust, H. W., Stammer, D., Ulbrich, U., Kadow, C., Köhl, A., Kröger, J., Kruschke, T., Pinto, J. G., Pohlmann, H., Reyers, M., Schröder, M., Sienz, F., Timmreck, C., and Ziese, M. (2016). MiKlip - a National Research Project on Decadal Climate Prediction. *Bulletin of the American Meteorological Society*, 97:2379–2394.

Marshall, J., Hill, C., Perelman, L., and Adcroft, A. (1997). *Geophysical Research*, 102:5733–5752.

Matei, D., Baehr, J., Jungclaus, J. H., Haak, H., Müller, W. A., and Marotzke, J. (2012). Multiyear prediction of monthly mean atlantic meridional overturning circulation at 26.5°N. *Science*, 335:76–79.

Mazumder, R. and Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6:2125–2149.

Meehl, G., Hu, A., Arblaster, J., Fasullo, J., and Trenberth, K. (2013). Externally Forced and Internally Generated Decadal Climate Variability Associated with the Interdecadal Pacific Oscillation. *Journal of Climate*, 26:7298–7310.

Meehl, G. A., Covey, C., Taylor, K. E., Delworth, T., Stouffer, R. J., Latif, M., McAvaney, B., and Mitchell, J. F. (2007). The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bulletin of the American Meteorological Society*, 88(9):1383–1394.

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., Dixon, K., Giorgetta, M. A., Greene, A. M., Hawkins, E., Hegerl, G., Karoly, D., Keenlyside, N., Masahide, K., Kirtman, B., Navarra, A., Pulwarty, R., Smith, D., Stammer, D., and Stockdale, T. (2009). Decadal prediction: can it be skillful? *Bulletin of the American Meteorological Society*, 90:1467–1485.

Meusel, H., Jäger, E., and Weinert, E. (1965). *Vergleichende Chorologie der zentraleuropäischen Flora*. VEB Fischer, Jena.

Miebach, A. (2016). *Climate- and Human-Induced Vegetation Changes in Northwestern Turkey and in the Southern Levant since the Last Glacial*. PhD thesis, University of Bonn.

Mochizuki, T., Ishii, M., Kimoto, M., Chikamoto, Y., Watanabe, M., Nozawa, T., Sakamoto, T. T., Shiogama, H., Awaji, T., Sugiura, N., Toyoda, T., Yasunaka, S., Tatebe, H., and Mori, M. (2010). Pacific decadal oscillation hindcasts relevant to near-term climate prediction. *Proceedings of the National Academy of Sciences of the United States of America*, 107(5):1833–1837.

Moemken, J., Reyers, M., Buldmann, B., and Pinto, J. (2016). Decadal predictability of regional scale wind speed and wind energy potentials over Central Europe. *Tellus A*, 68:29199.

Mogensen, K., Balmaseda, M. A., and Weaver, A. (2012). The NEMOVAR ocean data assimilation system as implemented in the ECMWF ocean analysis for System 4. Technical report, European Centre for Medium-Range Weather Forecasts.

Müller, W., Baehr, J., Haak, H., Jungclaus, J., Kröger, J., Matei, D., Notz, D., Pohlmann, H., von Storch, J., and Marotzke, J. (2012). Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology. *Geophysical Research Letters*, 39(22):L22707.

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116(12):2417–2424.

Murphy, A. H. (1991). Forecast verification: Its complexity and dimensionality. *Monthly Weather Review*, 119:1590–1601.

Murphy, A. H. (1993). What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Monthly Weather Review*, 8:281–293.

Murphy, A. H. and Epstein, E. S. (1989). Skill Scores and Correlation Coefficients in Model Verification. *Monthly Weather Review*, 117(3):572–581.

Nelder, J. A. and Baker, R. J. (1972). Generalized linear models. *Encyclopedia of Statistical Sciences*.

Neugebauer, I., Brauer, A., Schwab, M., Waldmann, N., Enzel, Y., Kitagawa, H., Torfstein, A., Frank, U., Dulski, P., Agnon, A., Ariztegui, D., Ben-Avraham, Z., Goldstein, S., and Stein, M. (2014). DSDDP Scientific Party: Lithology of the long sediment record recovered by the ICDP Dead Sea Deep Drilling (DSDDP). *Quaternary Science Reviews*, 102:149–165.

Neumann, F., Schölzel, C., Litt, T., Hense, A., and Stein, M. (2007). Holocene vegetation anad climate history of the northern Golan heights (Near East). *Vegetation History and Archaeobotany*, 16:329–346.

Niemi, T., Ben-Avraham, Z., and Gat, J. (1997). *The Dead Sea: The Lake and Its Setting*. Number 36. Oxford University Press, USA.

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

Otto-Bliesner, B., Rosenbloom, N., Stone, E., McKay, N., Lunt, D., Brady, E., and Overpeck, J. (2013). How warm was the last interglacial? New model-data comparisons. *Philosophical Transactions of the Royal Society A*, 371(2001):20130097.

Pickarski, N., Kwiecien, O., Langgut, D., and Litt, T. (2015). Abrupt climate and vegetation variability of eastern Anatolia during the last glacial. *Climate of the Past*, 11:1491–1505.

Pohlmann, H., Jungclaus, J. H., Köhl, A., Stammer, D., and Marotzke, J. (2009). Initialized decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic. *Journal of Climate*, 22:3926–3938.

Pohlmann, H., Müller, W., Kulkarni, K., Kameswarrao, M., Matei, D., Vamborg, F., Kadow, C., Illing, S., and Marotzke, J. (2013). Improved forecast skill in the tropics in the new MiKlip decadal climate predictions. *Geophysical Research Letters*, 40(21):5798–5802.

Polkova, I., Köhl, A., and Stammer, D. (2014). Impact of initialization procedures on the predictive skill of a coupled ocean-atmosphere model. *Climate Dynamics*, 42:3151–3169.

Prentice, I., Guiot, J., Huntley, B., Jolly, D., and Cheddadi, R. (1996). Reconstructing biomes from paleoecological data: a general method and its application to European pollen data at 0 and 6 ka. *Climate Dynamics*, 12:185–194.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/.

Roberts, C., Calvert, D., Dunstone, N., Hermanson, L., Palmer, M., and Smith, D. (2016). On the Drivers and Predictability of Seasonal-to-Interannual Variations in Regional Sea Level. *Journal of Climate*, 29:7565–7585.

Romanova, V. and Hense, A. (2015). Anomaly transform methods based on total energy and ocean heat content norms for generating ocean dynamic disturbances for ensemble climate forecasts. *Climate Dynamics*. doi:10.1007/s00382-015-2567-4.

Röpnack, A., Hense, A., Gebhardt, C., and Majewski, D. (2013). Bayesian model verification of nwp ensemble forecasts. *Monthly Weather Review*, 141(1):375–387.

Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R. T., Dunstone, N., Eade, R., Fereday, D., Folland, C. K., Gordon, M., Hermanson, L., Knight, J. R., Lea, D. J., MacLachlan, C., Maidens, A., Martin, M., Peterson, A. K., Smith, D., Vellinga, M., Wallace, E., Waters, J., and Williams, A. (2014). Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, 41(7):2514–2519. 2014GL059637.

Schölzel, C. (2006). Paleoenvironmental Transfer Functions in a Bayesian Framework with Application to Holocene Climate Variability in the Near East. *Bonner Meteorologische Abhandlungen 62*.

Schölzel, C. and Hense, A. (2011). Probabilistic assessment of regional climate change in Southwest Germany by ensemble dressing. *Climate Dynamics*, 36:2003–2014.

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

Sienz, F., Müller, W., and Pohlmann, H. (2016). Ensemble size impact on the decadal predictive skill assessment. *Meteorologische Zeitschrift*, 25(6):645–655.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall/CRC, London New York.

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

Smith, D. M., Cusack, S., Colman, A. W., Folland, C. K., Harris, G. R., and Murphy, J. M. (2007). Improved Surface Temperature Prediction for the Coming Decade from a Global Climate Model. *Science*, 317:796–799.

Smith, D. M., Eade, R., Dunstone, N. J., Fereday, D., Murphy, J. M., Pohlmann, H., and Scaife, A. A. (2010). Skilfull multti-year predictions of Atlantic hurricane frequency. *Nature Geoscience*, 3:846–849.

Smith, T., Shugart, H., Woodward, F., and Burton, P. (1993). Plant functional types. In *Vegetation Dynamics & Global Change*, pages 272–292. Springer.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K., Tignor, M., and Miller, H., editors (2007). *IPCC, 2007: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Spangehl, T., Schröder, M., Stolzenberger, S., Glowienka-Hense, R., Mazurkiewicz, A., and Hense, A. (2016). Evaluation of the MiKlip decadal prediction system using satellite based cloud products. *Meteorologische Zeitschrift*, 25(6):695–707.

Stein, M., Ben-Avraham, Z., and Goldstein, S. (2011). Dead Sea deep cores: a window into the past climate seismicity. *EOS*, 92(49):453–454.

Stocker, T. F., Qin, D., Plattner, G., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., editors (2013). *IPCC, 2013: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Stockhecke, M., Timmermann, A., Kipfer, R., Haug, G., Kwiecien, O., Friedrich, T., Menviel, L., Litt, T., Pickarski, N., and Asnelmetti, F. (2016). Millenial to orbital-scale variations of drought intensity in the Eastern Mediterranean. *Quaternary Science Reviews*, 33:77–95.

Stolzenberger, S. (2011). Untersuchungen zu botanischen Paläoklimatransferfunktionen. Diplomarbeit, Meteorologisches Institut der Rhein. Friedr.-Wilh.-Universität, X.

Stolzenberger, S., Glowienka-Hense, R., Spangehl, T., Schröder, M., Mazurkiewicz, A., and Hense, A. (2016). Revealing skill of the MiKlip decadal prediction system by three-dimensional probabilistic evaluation. *Meteorologische Zeitschrift*, 25(6):657–671.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. (2009). A summary of the CMIP5 experiment design. *PCDMI Report*, page 33 pp.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4):485–498.

Thoma, B. (2016). *Paleoclimate Reconstructions in the Levant and the Balkans*. PhD thesis, University of Bonn, submitted.

Thorarinsdottir, T. L., Gneiting, T., and Gissibl, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1):522–534.

Tingley, M. and Huybers, P. (2010). 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.

Torfstein, A., Goldstein, S., Kushnir, Y., Enzel, Y., Haug, G., and Stein, M. (2015). Dead sea drawdown and monsoonal impacts in the levant during the last interglacial. *Earth Planet. Sc. Lett*, 412:235–244.

Uppala, S. M., Kållberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Van De Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth,

K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J. (2005). The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, 131(612):2961–3012.

von der Heydt, A., Dijkstra, H., van de Wal, R., Caballero, R., Crucifix, M., Foster, G., Huber, M., Köhler, P., Rohling, E., Valdes, P., Ashwin, P., Bathiany, S., Berends, T., van Bree, L., Ditlevsen, P., Ghil, M., Haywood, A., Katzav, J., Lohmann, G., Lohmann, J., Lucarini, V., Marzocchi, A., Pälike, H., Baroni, I., Simon, D., Sluijs, A., Stap, L., Tantet, A., Viebahn, J., and Ziegler, M. (2016). Lessons on climate sensitivity from past climate changes. *Curr Clim Change Rep*, 2:148–158.

von Storch, H. and Zwiers, F. (2001). *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge, New York.

Weinert, H. (2015). Multivariate verification of decadal climate predictions based on surface air temperature. Masterarbeit, Meteorologisches Institut der Rhein. Friedr.-Wilh.-Universität, XXVI.

Weisheimer, A. and Palmer, T. (2014). On the Reliability of Seasonal Climate Forecasts. *J. R. Soc. Interface*, 11(96):20131162.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences, 3rd Ed.* Academic press, Oxford, Amsterdam.

BONNER METEOROLOGISCHE ABHANDLUNGEN

Herausgegeben vom Meteorologischen Institut der Universität Bonn 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 1-63: siehe `http://www.meteo.uni-bonn.de/bibliothek/bma`

64-80: open access, verfügbar unter `https://uni-bn.de/kpSDaQffe1`

Heft 64: ***Michael Weniger***: Stochastic parameterization: a rigorous approach to stochastic three-dimensional primitive equations, 2014, 148 S. + XV.

Heft 65: ***Andreas Röpnack***: Bayesian model verification: predictability of convective conditions based on EPS forecasts and observations, 2014, 152 S. + VI.

Heft 66: ***Thorsten Simon***: Statistical and Dynamical Downscaling of Numerical Climate Simulations: Enhancement and Evaluation for East Asia, 2014, 48 S. + VII. + Anhänge

Heft 67: ***Elham Rahmani***: The Effect of Climate Change on Wheat in Iran, 2014, [erschienen] 2015, 96 S. + XIII.

Heft 68: ***Pablo A. Saavedra Garfias***: Retrieval of Cloud and Rainwater from Ground-Based Passive Microwave Observations with the Multi-frequency Dual-polarized Radiometer ADMIRARI, 2014, [erschienen] 2015, 168 S. + XIII.

Heft 69: ***Christoph Bollmeyer***: A high-resolution regional reanalysis for Europe and Germany - Creation and Verification with a special focus on the moisture budget, 2015, 103 S. + IX.

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.

METEOROLOGISCHES INSTITUT
MATHEMATISCH NATURWISSENSCHAFTLICHE FAKULTÄT
UNIVERSITÄT BONN

Meteorologisches Institut
Universität Bonn