UNIVERSITÉ DE LAUSANNE

DOCTORAL THESIS

Software and Numerical Tools for Paleoclimate Analysis

Author:
Philipp S. SOMMER

Supervisor:
Dr. Basil A. S. Davis

UNIL | Université de Lausanne

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Science

in the

Davis Group
Institute of Earth Surface Dynamics (IDYST)
Faculty of Geosciences and Environment (GSE)

Jury members:
Prof. Dr. Christian Kull (*president*), Prof. Dr. Grégoire Mariéthoz (*director*),
Basil A. S. Davis (*co-director*),Prof. Dr. John Robert Tipton (*examiner*),
Prof. Dr. Christoph Cornelius Raible (*examiner*)

February 26, 2020

"The purpose of computing is insight, not numbers."

Richard Wesley Hamming

UNIVERSITÉ DE LAUSANNE

Summary

Faculty of Geosciences and Environment (GSE)
Institute of Earth Surface Dynamics (IDYST)

Doctor of Science

Software and Numerical Tools for Paleoclimate Analysis

by Philipp S. SOMMER

Data-model comparisons of Holocene (11,700 years ago to present) climate provide an ideal basis for evaluating climate model performance outside the range of modern climate variability. The Holocene is recent enough so that boundary conditions of the underlying physics and forcing are well known, while paleoenvironmental archives are abundant and dated with enough precision to comprehensively reconstruct climate. To date, efforts to reconstruct the spatial patterns of Holocene climate change have been mainly focused on the mid-Holocene (about 6'000 years ago), but significant discrepancies have already been identified in data-model comparisons.

These data-model discrepancies can be investigated using instrumental datasets covering continental or hemispheric scales which allow us to reconstruct large-scale climatic features, such as atmospheric dynamics or latitudinal temperature gradients. The generation of these datasets for times prior to the 19th century however faces considerable challenge because there are very few direct measurements of climatic variables. We rely on climate proxies as indirect measurements of the paleo climate. The most abundant one is fossil pollen data, i.e. pollen that are produced by vegetation and can be preserved over thousands of years in terrestrial (or coastal) archives (e.g. lake sediments). This proxy is available from all non-glaciated continents over the world in many different climate regimes, and the primary data is becoming increasingly accessible through large publicly available and community-driven relational databases. Our ability to use this proxy for continental-scale climate reconstructions, however, depends on our ability to analyze, explore and find patterns in these rich and heterogeneous databases. In particular, this requires a proper understanding of the uncertainties that are related to the indirect measurement of climate.

In the first part of this thesis, I present three new software tools that tackle the challenge to make this large amount of data accessible, and to build and develop a continental-scale pollen database. These tools cover a wide range of possible applications to leverage our work with site-based proxy data to a continental scale. The first tool I present is a web framework that is built around a map-based interactive database viewer, developed primarily for the Eurasian Modern Pollen Database, EMPD. This new tool makes the database accessible to other researchers and to the general public, and it allows a continuous and stable development of the community-driven database. In addition to the EMPD, I present an extension of this viewer that makes a large northern-hemispheric fossil pollen database accessible and allows its visual exploration.

The second tool tackles the challenge to fill the gaps in certain geographic areas in the pollen database. *straditize* is a digitization software for stratigraphic diagrams, and pollen diagrams in particular. It can be used to generate new data for the pollen

database from publications of the pre-digital era, i.e. from publications where the primary pollen data is not accessible anymore but through the visualization in form of a pollen diagram in a peer-reviewed publication.

Finally, I present the generic python visualization framework *psyplot*, that bridges the gap between visualization, computation and publication in the day-to-day work of scientists, and that has been used in multiple parts of the thesis. This flexible software can be integrated and enhanced by a variety of applications and already contains multiple convenient visualization methods useful for climate science, particularly the visualization of geo-referenced data and it handles data that is too large to fit into memory or lives on different structured or unstructured grids.

The second part of my thesis contains two new statistical methods to estimate large-scale paleo climatic environments based on modern day relationships. The first one, *pyleogrid*, uses a large pollen database and turns it into a gridded climate reconstruction that can cover continental, hemispheric or even global scales. This software focuses a lot on the integration of the intrinsic uncertainties in the proxy data. The outcome of this gridding procedure allows a comparison of computational climate models with an independent observational database that comes with reliable estimates of uncertainty.

The last chapter of this thesis applies a converse strategy and uses modern statistical relations within climate variables to inform a computational model. The global weather generator (*GWGEN*) has been parameterized with thousands of global weather stations and provides a statistical tool that downscales monthly to daily climatology on a global scale. This tool can be embedded in a global paleo vegetation model where it efficiently simulates the necessary daily meteorology.

UNIVERSITÉ DE LAUSANNE

Résumé

Faculté des géosciences et de l'environnement
Institut des dynamiques de la surface terrestre

Docteur ès Sciences

Software and Numerical Tools for Paleoclimate Analysis

by Philipp S. SOMMER

Les comparaisons données-modèles du climat de l'Holocène (d'il y a 11 700 ans à aujourd'hui) fournissent une base idéale pour évaluer la performance des modèles climatiques en dehors de la plage moderne de variabilité climatique. L'Holocène est assez récent pour que les conditions aux limites et les différents forçages soient bien connus, tandis que les archives paléoenvironnementales sont abondantes et datées avec suffisamment de précision pour reconstruire complètement le climat. Jusqu'à présent, les efforts pour reconstruire les changements climatiques de l'Holocène spatialement se sont principalement concentrés sur l'Holocène moyen (il y a environ 6 000 ans), mais des divergences significatives ont déjà été identifiées lors des comparaisons données-modèles.

Ces écarts entre les modèles et les données peuvent être étudiés à l'aide de collection de données d'observation couvrant des échelles continentales ou hémisphériques qui nous permettent de reconstruire des caractéristiques climatiques à grande échelle, comme la dynamique atmosphérique ou les gradients de température latitudinaux. La génération de ces ensembles de données pour des périodes antérieures au XIXe siècle est cependant confrontée à un défi considérable, car il existe très peu de mesures directes de ces variables climatiques. Nous nous appuyons sur des mesures indirectes du paléoclimat à l'aide d'indicateurs climatiques. Le plus abondant est le pollen fossile, c'est-à-dire le pollen produit par la végétation et qui peut être conservé pendant des milliers d'années dans des archives terrestres ou côtières (e.g. les sédiments lacustres). Ce « proxy » est disponible sur tous les continents non couverts par les glaces à travers de nombreux régimes climatiques différents à travers le monde, et les données primaires sont de plus en plus accessibles par le biais de grandes bases de données relationnelles accessibles au public et gérées par la communauté. Notre capacité à utiliser ce proxy pour les reconstructions climatiques à l'échelle continentale, cependant, dépend de notre capacité d'analyser, d'explorer et de trouver des modèles dans ces bases de données riches et hétérogènes. En particulier, cela exige une bonne compréhension des incertitudes liées à la mesure indirecte du climat.

Dans la première partie de cette thèse, je présente trois nouveaux outils logiciels qui relèvent le défi de rendre accessible cette grande quantité de données et de construire et développer une base de données de pollen à l'échelle continentale. Ces outils couvrent un large éventail d'applications possibles pour tirer parti de notre travail avec des données proxy basées sur des sites à l'échelle continentale.

Le premier outil que je présente est une application Web qui s'articule autour d'un visualiseur de base de données interactif basé sur des cartes, développé principalement pour la base de données eurasienne moderne sur le pollen (EMPD, Eurasian Modern Pollen Database). Ce nouvel outil rend la base de données accessible à d'autres chercheurs et au grand public, et il permet un développement continu et

stable. En plus des données de l'EMPD, je présente une extension de ce visualiseur qui rend accessible une vaste base de données sur les pollens fossiles de l'hémisphère Nord et permet son exploration visuelle.

Le deuxième outil s'attaque au défi de combler les lacunes dans certaines zones géographiques de la base de données de pollen. straditize est un logiciel de numérisation pour les diagrammes stratigraphiques, et les diagrammes de pollen en particulier. Il peut être utilisé pour générer de nouvelles données pour la base de données sur le pollen à partir de publications de l'ère pré-numérique, c'est-à-dire de publications dont les données primaires sur le pollen ne sont plus accessibles, mais dont la visualisation sous forme de diagramme de pollen est possible dans une publication.

Enfin, je présente le package python de visualisation générique psyplot, qui comble l'écart entre la visualisation, le calcul et la publication dans le travail quotidien des scientifiques, et qui a été utilisé dans plusieurs parties de la thèse. Ce logiciel flexible peut être intégré et amélioré par une variété d'applications et contient déjà de multiples méthodes de visualisation pratiques utiles pour la climatologie, en particulier la visualisation de données géoréférencées et il traite des données qui sont trop grandes pour tenir en mémoire ou qui vivent sur différentes grilles, structurées ou non.

La deuxième partie de ma thèse contient deux nouvelles méthodes statistiques pour estimer les environnements paléoclimatiques à grande échelle basées sur les relations modernes. La première, pyleogrid, utilise une grande base de données de pollen et la transforme en une reconstruction climatique maillée qui peut couvrir des échelles continentales, hémisphériques ou même globales. Ce logiciel se concentre sur l'intégration des incertitudes intrinsèques des données proxy. Le résultat de cette méthode de maillage permet de comparer des modèles climatiques computationnels avec une base de données d'observation indépendante qui fournit des estimations fiables de l'incertitude.

Le dernier chapitre de cette thèse applique une stratégie inverse et utilise les relations statistiques des variables climatiques modernes pour informer un modèle. Le générateur météorologique global (GWGEN, global weather generator) a été paramétré avec des données provenant de milliers de stations météorologiques mondiales et fournit un outil qui permet de passer de l'échelle mensuelle à l'échelle quotidienne (« downscaling ») pour le monde entier. Cet outil peut être intégré dans un modèle global de paléovégétation où il simule efficacement la météorologie quotidienne nécessaire.

Acknowledgements

This work and my entire PhD would not have been possible without the lot of support I received from numerous people. Most of all I have to thank my supervisor Basil Davis for becoming my mentor after the first year of my PhD and introducing me into the great world of paleo-climatic research. I especially want to thank him for his support and understanding during my critical third year, which would have been so much worse without it. I am deeply grateful for his advice, ideas and experiences that constantly improved my understanding of the big climate picture, I am thankful for his support during the rush in the last weeks of my PhD. I further want to thank (my unofficial second supervisor) Manuel Chevalier for his great experience, his open mind and his impressive skill to quickly familiarize with new topics and discuss them. Thank you both that I could always come to you and discuss any issue with my research. All the software packages that I present in this thesis would not have been useful at all without your opinions and your input.

I also want thank Jed O. Kaplan for the collaboration that we had during the development of the weather generator and for being my supervisor during the first year of my PhD. And I want to thank John Tipton for his inspiring methodological experience and his great advice during our collaboration, particularly for his input in the *pyleogrid* chapter. Special thanks goes to Grégoire Mariéthoz for his methodological advice with several statistical issues that I had during the last four years.

It has been a great pleasure to work in the Institute of Earth Surface Dynamics at the University of Lausanne and I want to thank the Swiss National Science Foundation, as well as the Faculty of Geosciences and Environment at UNIL for the support of our HORNET (200021_169598) and the ACACIA (CR10I2_146314) projects. I was very lucky to do my research in such a wonderful environment, particularly with Mathieu Gravey, Inigo Irarrazaval, Luiz Gustavo Rasera, Harsh Beria, Anthony Michelon (for all the little teasing), and especially with Dilan Rech, Lucien Goldenschue, Leanne Phelps, Ryan Hughes, John Shekeine and Andrea Kay. I thank you all for the nice discussions that we had, for your advice in the preparation of multiple talks, and for being good friends during the past four years.

Finally, this entire work would not have been possible without the skills I developed during my Master at the University of Hamburg and the Max-Planck-Institute for Meteorology. I am very thankful for the possibilities I had that allowed me to learn programming during this time, and I am thankful for the ongoing friendship with my former colleagues, Stefan Hagemann, Tobias Stacke and Philipp de Vrese.

To conclude this long list here, which is still not finished, I want to thank my family, particularly my parents and brothers for the ongoing wonderful relationship that we have, and I want to thank my new family, Sylvia Weiß-Fiebig and Gernot Fiebig for their support of my new little family that I got during this PhD. This little family is the best thing in my life and I am extremely thankful to the most important part of it, Bianca, for staying with me after all the difficulties we experienced, being such a good friend, a wonderful partner and a loving mom.

Dedicated to Bianca and Leo, my personal key to happiness…

Contents

Summary iii

Résumé v

Acknowledgements vii

1 Introduction 1
1.1 Motivation . . . 1
1.2 Learning from the past — Why we study paleo-climates . . . 2
1.2.1 Pollen as a climate proxy . . . 2
1.3 Software for Paleoclimatology . . . 3
1.3.1 Sofware for Proxy Data Analysis, Visualization and Distribution 4
1.3.2 Methods and Workflows in Open-Source Software Development 5
Version Control . . . 5
Automated Tests, Test Coverage and Continuous Integration . 6
Automated Documentation . . . 7
Distribution through package managers and virtual environments . . . 7
1.4 Challenges tackled in this thesis . . . 8
References . . . 9

Part I New Software Tools for Paleoclimate Analysis 19

2 The EMPD and POLNET web-interfaces 21
2.1 Summary . . . 21
2.2 The EMPD web framework . . . 22
2.2.1 The EMPD viewer . . . 23
The Web Interface . . . 23
Implementation details . . . 24
2.2.2 The EMPD2 data repository . . . 24
2.2.3 The EMPD-admin . . . 27
Implementation details . . . 27
2.2.4 Distribution of the tools . . . 27
2.3 The POLNET viewer . . . 28
References . . . 30

3 Straditize: *A digitization software for pollen diagrams* 33
3.1 Introduction . . . 33
3.2 Methods: Treatment of stratigraphic diagram features . . . 34
3.2.1 Structure of a stratigraphic diagram . . . 34
3.2.1.1 Stratigraphic columns . . . 34
3.2.1.2 Diagram types . . . 36

3.2.1.3 Informative features . . . 36
3.2.2 Digitization procedure . . . 37
3.2.2.1 Defining the data part of the diagram . . . 37
3.2.2.2 Separating the columns . . . 38
3.2.2.3 Cleaning up the diagram . . . 38
3.2.2.4 Handling low taxon values . . . 39
3.2.2.5 Digitizing the diagram . . . 40
3.2.2.6 Finding the samples . . . 41
3.3 Discussion . . . 42
3.4 Conclusions . . . 43
References . . . 44

4 Psyplot: *A flexible framework for interactive data analysis* 45
4.1 Summary . . . 45
4.2 Introduction . . . 45
4.3 The psyplot framework . . . 46
4.3.1 Data model . . . 46
Psyplot and xarray . . . 46
Psyplot core structure . . . 47
4.3.2 Psyplot plugins . . . 49
psy-simple: The psyplot plugin for simple visualizations . . . 49
psy-maps: The psyplot plugin for visualizations on a map . . . 49
psy-reg: The psyplot plugin for visualizing and calculating regression plots . . . 50
psy-strat: A psyplot plugin for stratigraphic plots . . . 50
4.3.3 The psyplot Graphical User Interface . . . 50
Console . . . 52
Help explorer . . . 52
Plot creator . . . 52
Project content . . . 52
Formatoptions . . . 53
Figures and plots . . . 53
4.4 Conclusions . . . 53
4.5 Outlook . . . 54
4.A Example call of a plot method . . . 55
4.B psy-simple plot methods . . . 56
4.C psy-maps plot methods . . . 57
4.D psy-reg plot methods . . . 57
4.E psy-strat plot methods . . . 58
References . . . 58

Part II Computational Models 63

5 pyleogrid: *A Probabilistic Approach for Gridding Paleo Climate Data* 65
5.1 Introduction . . . 65
5.2 Data . . . 66
5.2.1 Pollen database . . . 67
5.2.2 Sample site: Tigalmamine . . . 67
5.2.3 Site-based holocene temperature estimates . . . 69
5.2.4 Age uncertainties . . . 70

5.3 Method . . . 71
5.3.1 Constrained age sampling . . . 74
The intuitive approach . . . 74
The random sorting approach . . . 75
The Gibbs sampling approach . . . 76
5.3.2 Temperature sampling . . . 78
5.3.3 Gridding . . . 80
5.3.4 Implementation . . . 82
5.4 Results . . . 83
5.4.1 Site-based realized climate reconstruction: a use-case . . . 83
5.4.2 Gridded summer temperature . . . 83
Determinstic vs. ensemble approach . . . 83
Temperature sampling parameters . . . 87
Uncertainties for spatial extrapolation . . . 89
5.5 Discussion . . . 89
5.6 Conclusions . . . 90
5.A Estimated age uncertainties . . . 92
5.B Example of generated age distributions . . . 92
5.C Maps of uncertainties . . . 93
References . . . 95

6 GWGEN: *A global weather generator for daily meteorology* 101
6.1 Introduction . . . 101
6.2 Model description . . . 103
6.3 Model development . . . 105
6.3.1 Development of a global weather station database . . . 105
6.3.2 Parameterization . . . 108
Precipitation occurrence . . . 108
Precipitation amount . . . 110
Temperature . . . 112
Cloud fraction . . . 115
Wind speed . . . 117
Cross correlation . . . 119
6.3.3 Model Evaluation . . . 120
6.3.4 Bias correction . . . 120
6.3.5 Sensitivity analysis . . . 123
6.4 Limitations . . . 124
6.5 Discussion and Outlook . . . 125
6.6 Conclusions . . . 126
6.7 Code availability . . . 126
6.A Supplementary material . . . 127
6.A.1 Sensitivity analysis . . . 127
References . . . 127

7 Conclusions 133
References . . . 135

Appendices 139

List of Figures 139

List of Abbreviations 141

A Publications and Conference contributions **143**
A.0.1 Peer-reviewed 143
A.0.2 Conference contributions 143

Chapter 1

Introduction

1.1 Motivation

Our understanding of the climate system is based on computational models that operate on large spatial scales and simulate a complex system of closely related environmental parameters. The evaluation of these models poses a considerable challenge because we need data on continental scale to reconstruct large-scale climatic features, such as atmospheric dynamics or latitudinal temperature gradients. This instrumental data, however, is limited to not even the past two centuries and overlaps highly with the *comfort zone* of the models, i.e. the period where they have both been developed and tested.

A comparison of models with paleo-environmental proxy records, i.e. records from past climates prior to the systematic measurement of meteorology and climatology, provides therefore the only possibility to evaluate the predictive skill of our models for climates that are very different from today. The Holocene, ranging from 11'700 years ago to present, provides an ideal basis for it because (1) it is recent enough so that boundary conditions and forcings are well known, and (2) paleoenvironmental archives are abundant and dated with enough precision to comprehensively reconstruct climate. Each of these archives represent the regional climate condition in the surrounding environment and, when grouped together into large databases; they allow an informed estimate of the climate state over a large period of time and vast geographic areas.

A direct comparison of climate model output and proxies is however still challenging because even the climate proxy record, as an indirect measurement of climate, relies on an inverse modelling approach with associated uncertainties that are not always easy to quantify.

Key challenges for large-scale data-model comparisons on past climates are therefore (1) to gather enough climate proxy information from a spatially large area and a variety of climates, and (2) to provide reliable estimates of the uncertainties associated to the indirect measurements.

These challenges will be addressed in this thesis via the development of new software tools that cover flexible data analysis (chapters 2 and 4), a tool for data gathering (chapter 3), as well as new predictive methods for large-scale paleoenvironmental modelling (chapters 5 and 6).

All these tools are open-source with a strong emphasis on a proper software development that includes documentation and reproducibility.

In the following section 1.2, I will describe the paleo-climate of the current epoch and why this is of interest for future climate predictions. In the subsequent section 1.3, I introduce the influence of software development in this paleo-climatic research, and some of the common open-source software development contents. I conclude this chapter by providing an overview on the contents of this thesis in section 1.3.2.

1.2 Learning from the past — Why we study paleo-climates

Mankind is facing large infrastructural challenges during this century, such as the loss of biodiversity, an exponentially growing world population and an acceleration of growth and globalization of markets (e.g. Ceballos et al., 2015; United Nations, 2019; World Bank, 2002). They all interact with a global climate change that may lead to a new environment none of us ever experienced (Collins et al., 2013). Any future global planning has to account highly diverse responses that range from regional to continental scales (Christensen et al., 2013). As such the complex climate system will enter a state that is significantly different from everything we had since the beginning of the satellite era, i.e. the beginning of global meteorological data acquisition, and even different from what has been experienced within the last 2'000 years (Neukom et al., 2019a,b).

Our knowledge about this new climate is therefore mainly based on computational Earth System Models (ESMs). They face the challenge of simulating a new climate based on our present knowledge of the interactions between the different compartments Ocean, Land and Atmosphere. The validation of it becomes conceptually difficult because of the aforementioned transition into a warmer world during the next century. We are entering a new state and it is questionable how well our models perform (Hargreaves et al., 2013; Mauri et al., 2014).

To evaluate the predictive skill, we rely on our knowledge of paleo-climates, i.e. climates before the systematic measurement of temperature, precipitation, etc.. They provide the only opportunity for a large-scale evaluation of ESMs under conditions very different than today. Paleo-climatic research has therefore been an integral part for climate sciences since the 80s (COHMAP Members, 1988; Joussaume and Taylor, 1995), particularly in the Paleoclimate Modelling Intercomparison Project (PMIP) (Braconnot et al., 2012, 2007a,b; Jungclaus et al., 2017; Kageyama et al., 2016; Otto-Bliesner et al., 2017).

The Holocene interglacial period (11'700 years ago to present) (Walker et al., 2009) is particularly important because it is sufficiently close in time to provide paleo-climate archives and the forcings and boundary conditions are well known (Wanner et al., 2008). With the end of the Younger Dryas around 11'700 years ago, the Earth experienced a climate warming due to changes in orbital precession and obliquity of the Earth, as well as the disappearing residual ice sheets of the Last Glacial Maximum (LGM) (Berger and Loutre, 1991; Peltier, 2004). This results in a multitude of large-scale effects in the atmospheric circulation, such as an increasing amplitude and frequency of the El Niño–Southern Oscillation (ENSO) (Donders et al., 2008), stronger westerly circulation in winter indicating a more positive AO/NAO over mid-latitudes and the arctic (Funder et al., 2011; Mauri et al., 2014) and changes in the polar amplification and a weakening of the latitudinal temperature gradient (Davis and Brewer, 2009).

Hence, this epoch is of particular interest because the continental setup is comparable to nowadays while still having a climate that is significantly different from present day.

1.2.1 Pollen as a climate proxy

Before 1850, there is almost no instrumental measurement of temperature. Instead we rely on archives such as lake sediments, glaciers, peat bogs, or speleothems that

preserve climate proxies. The latter is a set of variables that are influenced by climate conditions and therefore allow an indirect measurement of climate(-related) parameters at ancient times, e.g. temperature, precipitation or sea-level.

The most abundant climate proxy, that I will also focus on in the next chapters, are pollen assemblages. It is the geographically most spread paleo-climate proxy (Birks and Birks, 1980) and has a long history in quantitative paleo-climatologic reconstructions (e.g. Bradley, 1985; Iversen, 1944; Nichols, 1967, 1969).

The chemically stable polymer sporopollenin allows the pollen grain to be preserved over very long periods of time, in various terrestrial archives such as lakes, wetlands or ocean sediments (Fægri et al., 1989; Havinga, 1967). Pollen are produced by seed-bearing plants (spermatophytes, Wodehouse, 1935) and as such have a high spatial continuity and prevalence (Chevalier et al., in prep). Their compositions are strongly influenced by the surrounding climate, although other factors, such as soil compositions or inter-species competition also play an important role. This dependency allows to reconstruct the driving factor, i.e. climate parameters such as winter and summer temperature, or precipitation from the observed pollen data (Brewer et al., 2007; Chevalier et al., in prep; Juggins, 2013; Juggins and Birks, 2012).

This high abundance of pollen led to multiple regional efforts to combine and homogenize fossil pollen data. This makes pollen particularly useful for large-scale data-model intercomparisons. The earliest examples are the European Pollen Database (EPD) and North American Pollen Database (NAPD) that both started around 1990 and developed a similar structure in order to be compatible (Fyfe et al., 2009; Grimm, 2008) . This let to the development of other regional pollen databases, such as the Latin American Pollen Database (LAPD) (Flantua et al., 2015; Marchant et al., 2002) in 1994 or the African Pollen Database (APD) (Vincens et al., 2007) in 1996, and others (see Grimm, 2008). These attempts finally let to the development of the Neotoma database (Williams et al., 2018), a global multiproxy database that incorporates many of the regional pollen databases.

The use of pollen for paleo-climate reconstruction has a long academic tradition in geology (Bradley, 1985) and provides the source of large-scale paleo-climatic reconstructions in number of different studies (Davis et al., 2003; Fischer and Jungclaus, 2011; Marsicek et al., 2018; Mauri et al., 2015, and more). Such a reconstruction, however, has multiple uncertainties that are often difficult to quantify and to consider (see chapter 5). Key challenges for a data-model comparison are dating uncertainties, influences of seasonality on the proxy (e.g. whether it represents summer, winter or annual temperature) and quality and temporal resolutions of the record. Another challenge is the proper handling of uncertainties related to the inverse modelling approach (e.g. Guiot and Vernal, 2011; Telford and Birks, 2005, 2009), and the spatial coverage of the proxy (see chapter 3).

1.3 Software for Paleoclimatology

The usage of software is crucial for the quantitative reconstruction of Earth's Climate. Paleoclimate research is facing an information overload problem and requires innovative methods in the realm of visual analytics, i.e. the interplay between automated analysis techniques and interactive visualization (Keim et al., 2008; Nocke, 2014). As such, a visual representation of the paleoclimate reconstruction has been essential for both, proxies (Bradley, 1985; Grimm, 1988; Nichols, 1967) and models

(Böttinger and Röber, 2019; Nocke, 2014; Nocke et al., 2008; Phillips, 1956; Rautenhaus et al., 2018), although the visualization methods significantly differ due to the differences in data size and data heterogeneity.

The second important aspect for software and paleoclimate is the distribution of data to make it accessible to other researchers, the community and policy makers, which is commonly established through online accessible data archives and recently also through map-based web interfaces (Bolliet et al., 2016; Williams et al., 2018).

The following sections provide an overview on the different techniques used by palynologists to visualize and distribute their data and concludes with an introduction into Open-Source Software Development, which forms the basis of all the software solutions that are presented later in this thesis.

1.3.1 Sofware for Proxy Data Analysis, Visualization and Distribution

Due to the nature of stratigraphic data, proxies, especially pollen assemblages, are often treated as a collection of multiple time-series (one-dimensional arrays). The size of one dataset is generally small (in the range of kB) and can be treated as plain text files. Traditionally, numerical and statistical analysis are separated from the visualization.

In palynology, standard analytical tools are Microsoft Excel[1] and the R software for statistical computing (R Core Team, 2019). The latter also involves multiple packages for paleoclimatic reconstruction, such as `rioja` (Juggins, 2017) and `analogue` (Simpson, 2007; Simpson and Oksanen, 2019), and bayesian methods exists in a variety of programming languages (e.g. Haslett et al., 2006; Holmström et al., 2015; Nolan et al., 2019; Parnell et al., 2014; Tipton, 2019). Alternatively, desktop applications exist, such as Polygon[2] by Nakagawa et al., 2002 or the CREST software presented by Chevalier et al., 2014 and Chevalier, 2019.

It is a long-standing tradition to visualize stratigraphic data, and especially pollen data, in form of a stratigraphic (pollen) diagram (Bradley, 1985; Grimm, 1988). Especially during the 20th century, when it was not yet common to distribute data alongside a peer-reviewed publication, pollen diagrams have been the only possibility to publish the entire dataset (see also chapter 3). The generation of these diagrams is usually based on desktop applications such as C2 (Juggins, 2007) or Tilia[3] (Grimm, 1988, 1991). A more recent implementation into the psyplot framework (Sommer, 2017, chapter 4) is also provided with the psy-strat plugin[4] (Sommer, 2019).

Raw pollen data is at present made available through web archives, such as PANGAEA[5] or the National Climatic Data Center (NCDC) by the National Oceanic and Atmospheric Administration (NOAA)[6]. Collections of data, such as regional pollen databases or project specific collections (e.g. Davis et al., 2013; Whitmore et al., 2005) are usually published in one of the above-mentioned archives or associated with a publication. A different approach has been taken by Bolliet et al., 2016 who developed a small web application as an interface into the data collection, the *ClimateProxiesFinder* (Brockmann, 2016, chapter 2).

Outstanding compared to the previous data interfaces is the new infrastructure for the Neotoma database (Williams et al., 2018). It consists of the map-based web

[1] https://products.office.com/en/excel
[2] http://polsystems.rits-paleo.com
[3] https://www.tiliait.com/
[4] https://psy-strat.readthedocs.io
[5] https://pangaea.de/
[6] https://www.ncdc.noaa.gov/data-access/paleoclimatology-data

interface, the Neotoma Explorer[7], a RESTful api[8] that allows an interaction with other web services, the neotoma R package (Goring et al., 2015) and an interface into the Tilia software for stratigraphic and map-based visualizations (Williams et al., 2018). This rich functionality is, however, bound to the structure of Neotoma and as such, different from the Javascript-based approach developed in chapter 2 because it cannot easily be transferred to other projects.

1.3.2 Methods and Workflows in Open-Source Software Development

The importance and necessity of software for visualization and data analysis led to the development of the software packages I present in this thesis. Most of them are written in the programming language Python (Perez et al., 2011), on the one hand due to my personal preference, but mainly due to the recent developments in out-of-core computing with the establishment of xarray and dask (Dask Development Team, 2016; Hoyer and Hamman, 2017; Rocklin, 2015). Another important reason, especially for psyplot (chapter 4) and straditize (chapter 3) was the availability of a highly flexible and stable package for graphical user interfaces, PyQt [9], and the comparably simple possibility to implement an in-process python console into the PyQ5 application[10] that allows to handle the software functionalities both, from the command line and from the GUI.

The tools that I present in the following chapter are all available as open-source software packages. But modern Free and Open-Source Software (FOSS) development is not only about making the source code available, but rather about providing a sustainable and maintainable package that allows continuous and transparent development under the aspect of rapidly evolving environment. In the following sections, I will introduce the most important FOSS development concepts (e.g. Shaw, 2018; Stodden and Miguez, 2014) and the necessary vocabulary. These concepts are used by many of the well-established software packages, such as matplotlib (Hunter, 2007), numpy (Oliphant, 2006), and scipy (Jones et al., 2001).

Version Control

Version control systems record changes to a file and enables the user to roll-back to previous versions of it. The usage of a such a system is inevitable for sustainable FOSS packages. It enables contributions by other FOSS developers and the usage through external packages.

The packages I present in the following chapters are hosted on Github[11], a freely available web platform for hosting projects that are managed with git (Chacon et al., 2019).

Version control with git has a specific terminology (see also chapter 2). Central aspects are *repositories* (project folders), *commits* (change of the project files), *issues* (bug reports), *branches* and *forks* (copies of the (main) project), and *pull requests* (contributions to a project). The following list explains this vocabulary in a bit more detail because the terminology is used in several parts of this thesis, particular in chapter 2 and 4. A more complete list is provided in Github, Inc., 2019.

[7] https://apps.neotomadb.org/Explorer

[8] https://api.neotomadb.org

[9] http://pyqt.sourceforge.net/Docs/PyQt5/installation.html

[10] https://qtconsole.readthedocs.org

[11] The packages are available at https://github.com/Chilipp. Other potential platforms for version control are sourceforge (https://sourceforge.net) and Bitbucket (https://bitbucket.org)

Repositories are the most basic elements of git and Github. It can be compared to a folder that contains all the necessary files associated with a project (e.g. the source code and documentation of a software package). It also contains all the different versions (revisions) of the project files.

Commits or revisions track the changes in the repository. Each commit is a change to a specific file (or a set of files) that is associated with a unique ID and a message of the author that describes the changes.

Issues are suggested improvements, bug reports or any other question to the repository. Every issue has an associated discussion page for the communication between repository owners and the users.

Branches are parallel versions of a repository. Often one incorporates new developments into a separate branch that does not affect the main version of the repository (the *master* branch) and merge the two versions when the new developments are fully implemented.

Forks are copies of repositories. When someone wants to contribute to a software package (repository) that does not belong to him, he can *fork* (copy) it, implement the changes, and then create a *pull request* to contribute to the official version. Different from a branch, that is a (modified) copy of another branch, forks are copies of the entire repository, i.e. all existing branches.

Pull request are the proposed changes to a repository. One can create a fork of the repository from someone else, implement changes in this fork and then create a pull request to merge it into the original repository. Every pull request has an associated discussion page that allows the repository owner to moderate and discuss the suggested changes.

Webhooks are general methods for web development. Github can trigger a hook to inform a different web service (such as a Continuous Integration (CI) service, see next section)) that a repository has changed or that someone contributed in a discussion. In chapter 2 we use Github webhooks for an automated administration of a repository.

Automated Tests, Test Coverage and Continuous Integration

The most important aspect for FOSS development, especially considering the rapid evolution of this area, is the existence of automated tests. One distinguishes unit tests (tests designed to cover one specific routine) and integration tests (tests of one or more routines within the framework) (Shaw, 2018). The boundary between the two tests is rather vague and the decision about what is used highly depends on the structure of the software that is supposed to be tested. For complex frameworks (such as psyplot or straditize), integration tests are needed to ensure the operability within the framework. Other more simple software packages, (such as docrep or model-organization (Sommer, 2018a,b)) go well with unit tests only.

Another good standard for such a test suite is to use an automated test discovery tool (e.g. the Python unittest package (Python Software Foundation, 2019) or pytest (Krekel et al., 2004)) that also reports the test coverage (i.e. the fraction of the code that is tested by the test suite). These functionalities are then implemented on

a CI service, such as Travis CI[12], Appveyor[13] or CircleCi[14] that are integrated into the Github repository (section 1.3.2). Every commit to the Github repository, or any new pull requests then triggers the tests. This transparently allows to ensure the operability of the software, and the test coverage report ensures that the newly implemented functionality is properly tested. A software development concept that is entirely built on this idea is the test-driven development. Within this framework, new features are implemented by starting with the test that should be fulfilled by the new feature and then improving the software until this test pass (Beck, 2002).

Automated Documentation

Documentation is the key aspect of a sustainable software and much of the geoscientific code has a lack of proper documentation (based on personal experience). For the software in this thesis, four different levels of the documentation play an important role:

The Application programming interface (API) documentation is meant to document the major parts of the software code that is subject to be used by external scripts or packages. It is usually implemented in the code and documents the essential subroutines and methods of the software.

The graphical user interface (GUI) documentation provides help for the most high-level functionality for the software. The GUI is a user interface into the software through graphical elements (such as buttons, checkboxes, etc.). Unlike the API documentation, it should not require knowledge about programming.

The contributing and/or developers guide is targeting other software developers that might want to contribute to the software package. This document states how other software developers should contribute to the software and introduces the central structural aspects and frameworks of the software.

The manual (or also commonly referred to as *the* documentation) is the document that contains all necessary information for the software, such as installation instructions, tutorials, examples, etc.. It often includes some (or multiple) of the above parts.

The documentations for the software in this thesis have been automatically generated with Sphinx, a Python tool to generate documentations in various different formats (such as HTML, PDF, etc.) (Hasecke, 2019; Perez et al., 2011). It is also implemented as a webhook into the Github repository (see section 1.3.2) to automatically generate an up-to-date documentation of the software for each commit to the Github repository. This provides an additional automated test for the software, and especially its high-level-interface, in addition to the automated test suite described in the previous section. Most of the manuals for the software packages in this thesis are hosted and build online with the free services offered by readthedocs.org.

Distribution through package managers and virtual environments

FOSS software is meant to be extensible and to build upon other FOSS packages. This requires an accurate and transparent handling of its dependencies and requirements which is usually provided through the so-called packaging of the software

[12] https://travis-ci.org/
[13] https://appveyor.com
[14] https://circleci.com/

(e.g. Torborg, 2016). There exists a variety of package managers and the choice most often depends on the framework of the software.

The software in this thesis is mainly distributed via two systems. The first one is python's own package manager *pip* which is based on the packages uploaded to pypi.org. The second one, which got increasing importance during the recent past, is the open-source Anaconda Distribution[15]. Both work on multiple operating systems (Windows, Linux and Mac OS), but the Anaconda Distribution contains also non-python packages (e.g. written in C or C++) that multiple Python packages rely on; and it contains a rich suite of R packages.

One step further, compared to package managers, are the distribution of virtual environments. These systems do not only provide the software, but also a full operating system and the installed dependencies. A popular platform (used also for the Eurasian Modern Pollen Database (EMPD) database) is provided through so-called Docker containers[16]. Compared to package managers, this system has the advantage of simplifying the installation procedure for the user because he only has to download the corresponding docker image. The docker image itself then runs independent of the local file system in a separate isolated mode.

1.4 Challenges tackled in this thesis

In part I of this thesis I present several new software tools that tackle the data analysis, data gathering and data distribution aspects described in the previous section 1.3.

Chapter 2 in this first part describes new tools for the data analysis and distribution of pollen data on a large continental scale. In this chapter I present the new infrastructural tools I developed for the sustainable management of the community-driven Eurasian Modern Pollen Database (EMPD). These tools consist of a flexible and lightweight map-based web interface into the data, the EMPD-viewer, and a webserver for an automated administration of the database. Within this chapter, I also present another use case for the map-viewer that is adapted to a large northern-hemispheric database of fossil and modern pollen records.

The second chapter in this part, chapter 3, describes the new *straditize* software that addresses the problem of gathering proxy data that has been collected during the pre-digital area. This software is a semi-automated digitization package for stratigraphic diagrams, and particularly pollen diagrams that we use to fill gaps in our database in data-poor regions.

I conclude the first part with the presentation of the generic visualization framework psyplot in chapter 4. It is a suite of python packages that are designed for an interactive visual analysis of data, both from a GUI and the command line. This software is the base infrastructure for many of the tools described in the other chapters. It has a very general scope is not limited to paleoclimate analysis.

In the second part I present two new models that leverage site-based observations (or paleo climate reconstruction) onto a continental, or even global scale. The first model in chapter 5 presents the very recent *pyleogrid* package that extends the methodology of (Mauri et al., 2015). The ensemble method I present in this study provides a spatio-temporal gridding of side-based proxy-climate estimates under

[15] https://www.anaconda.com
[16] https://www.docker.com

the consideration of both, their dating and reconstruction uncertainties. The outcome of this model can be conveniently used for data-model intercomparisons because it contains reliable estimates of the methodological uncertainties.

Finally, I describe the weather generator *GWGEN* in chapter 6, a statistical model that uses modern relationships in observational meteorological data to inform large-scale paleo vegetation models with temporally downscaled temperature, precipitation, cloud cover and wind speed records. This weather generator has been parameterized with more than 50 million daily weather observation to be applicable on the entire globe.

This thesis finishes with the conclusions in chapter 7 where I summarize the new tools from this thesis and provide an outlook for the further development of the methods. Three of them are already published in peer-reviewed journals by the time that this thesis has been submitted. These software packages are psyplot (Sommer, 2017), GWGEN (Sommer and Kaplan, 2017) and straditize (Sommer et al., 2019). In appendix A I provide a list of all the peer-reviewed publications I have been involved as main or co-author during my thesis and that have been accepted by the time that this thesis has been submitted.

References

Beck, Kent (Dec. 1, 2002). *Test Driven Development. By Example*. Addison Wesley. 192 pp. ISBN: 978-0-321-14653-3. URL: https://www.ebook.de/de/product/3253611/kent_beck_test_driven_development_by_example.html.

Berger, A. and M. F. Loutre (1991). "Insolation values for the climate of the last 10 million years". In: *Quat. Sci. Rev.* 10.4, pp. 297–317. ISSN: 02773791. DOI: 10.1016/0277-3791(91)90033-q.

Birks, Harry John Betteley and Hilary H Birks (1980). *Quaternary palaeoecology*. Edward Arnold London.

Bolliet, Timothé, Patrick Brockmann, Valérie Masson-Delmotte, Franck Bassinot, Valérie Daux, Dominique Genty, Amaelle Landais, Marlène Lavrieux, Elisabeth Michel, Pablo Ortega, Camille Risi, Didier M. Roche, Françoise Vimeux, and Claire Waelbroeck (Aug. 2016). "Water and carbon stable isotope records from natural archives: a new database and interactive online platform for data browsing, visualizing and downloading". In: *Climate of the Past* 12.8, pp. 1693–1719. DOI: 10.5194/cp-12-1693-2016.

Böttinger, Michael and Niklas Röber (2019). "Visualization in Climate Modelling". In: *International Climate Protection*. Ed. by Michael Palocz-Andresen, Dòra Szalay, Andràs Gosztom, Làszlò Sìpos, and Tìmea Taligàs. Cham: Springer International Publishing, pp. 313–321. ISBN: 978-3-030-03816-8. DOI: 10.1007/978-3-030-03816-8_39. URL: https://doi.org/10.1007/978-3-030-03816-8_39.

Braconnot, P., Sandy P Harrison, Masa Kageyama, Patrick J Bartlein, Valerie Masson-Delmotte, Ayako Abe-Ouchi, Bette Otto-Bliesner, and Yan Zhao (2012). "Evaluation of climate models using palaeoclimatic data". In: *Nature Climate Change* 2.6, p. 417. DOI: 10.1038/nclimate1456. URL: https://www.nature.com/articles/nclimate1456.

Braconnot, P., B. Otto-Bliesner, S. Harrison, S. Joussaume, J.-Y. Peterchmitt, A. Abe-Ouchi, M. Crucifix, E. Driesschaert, Th. Fichefet, C. D. Hewitt, M. Kageyama, A. Kitoh, M.-F. Loutre, O. Marti, U. Merkel, G. Ramstein, P. Valdes, L. Weber, Y. Yu, and Y. Zhao (June 2007a). "Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum – Part 2: feedbacks with emphasis on the

location of the ITCZ and mid- and high latitudes heat budget". In: *Climate of the Past* 3.2, pp. 279–296. DOI: 10.5194/cp-3-279-2007. URL: https://www.clim-past.net/3/279/2007/.

Braconnot, P., Otto-Bliesner, S. P. Harrison, S. Joussaume, J.-Y. Peterchmitt, A. Abe-Ouchi, M. Crucifix, E. Driesschaert, Th. Fichefet, C. D. Hewitt, M. Kageyama, A. Kitoh, A. Laîné, M.-F. Loutre, O. Marti, U. Merkel, G. Ramstein, P. Valdes, S. L. Weber, Y. Yu, and Y. Zhao (June 2007b). "Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum – Part 1: experiments and large-scale features". In: *Climate of the Past* 3.2, pp. 261–277. DOI: 10.5194/cp-3-261-2007. URL: https://www.clim-past.net/3/261/2007/.

Bradley, Raymond S (1985). *Quaternary paleoclimatology : methods of paleoclimatic reconstruction*. eng. Boston ; London [etc.]: Allen and Unwin. ISBN: 0045510679.

Brewer, S., Joel Guiot, and Doris Barboni (2007). "Pollen data as climate proxies". In: *Encyclopedia of Quaternary Science*. Elsevier, pp. 2497–2508. URL: https://hal.archives-ouvertes.fr/hal-00995404.

Brockmann, Patrick (2016). *ClimateProxiesFinder: dc.js + leaflet application to discover climate proxies*. Last accessed: 2019-08-30. URL: https://github.com/PBrockmann/ClimateProxiesFinder (visited on 10/06/2016).

Ceballos, Gerardo, Paul R. Ehrlich, Anthony D. Barnosky, Andrès Garcìa, Robert M. Pringle, and Todd M. Palmer (June 2015). "Accelerated modern human–induced species losses: Entering the sixth mass extinction". In: *Science Advances* 1.5. DOI: 10.1126/sciadv.1400253. eprint: https://advances.sciencemag.org/content/1/5/e1400253.full.pdf. URL: https://advances.sciencemag.org/content/1/5/e1400253.

Chacon, Scott, Ben Straub, and Pro Git Contributors (2019). *Pro Git*. 2nd ed. Last accessed: 2019-08-31. URL: https://github.com/progit/progit2 (visited on 08/31/2019).

Chevalier, M., R. Cheddadi, and B. M. Chase (Nov. 2014). "CREST (Climate REconstruction SofTware): a probability density function (PDF)-based quantitative climate reconstruction method". In: *Climate of the Past* 10.6, pp. 2081–2098. DOI: 10.5194/cp-10-2081-2014.

Chevalier, M., B.A.S. Davis, K. Gajewski, H. Seppä, O. Heiri, J. Guiot, J. Marcisek, B.M. Chase, N. Kühl, J. Tipton, A. Dawson, L. Holmström, K. Izumi, T. Lacourse, W. Finsinger, R.J. Telford, L.N. Phelps, S.Y. Maezumi, V. Carter, M. Zanon, A. Mauri, F. Vallè, A. de Vernal, S. Goring, M. Chaput, P. S. Sommer, D. Kuprianov, and C. Nolan (in prep). "A review of statistical methods to quantify past climates from fossil pollen data". In:

Chevalier, Manuel (Apr. 2019). "Enabling possibilities to quantify past climate from fossil assemblages at a global scale". In: *Global and Planetary Change* 175, pp. 27–35. DOI: 10.1016/j.gloplacha.2019.01.016.

Christensen, J.H., K. Krishna Kumar, E. Aldrian, S.-I. An, I.F.A. Cavalcanti, M. de Castro, W. Dong, P. Goswami, A. Hall, J.K. Kanyanga, A. Kitoh, J. Kossin, N.-C. Lau, J. Renwick, D.B. Stephenson, S.-P. Xie, and T. Zhou (2013). "Climate Phenomena and their Relevance for Future Regional Climate Change". In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P.M. Midgley, T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P.M. Midgley, T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P.M. Midgley, T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J.

Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. Chap. 14, pp. 1217–1308. ISBN: ISBN 978-1-107-66182-0. DOI: 10.1017/CBO9781107415324.028. URL: www.climatechange2013.org.

COHMAP Members (1988). "Climatic Changes of the Last 18,000 Years: Observations and Model Simulations". In: *Science* 241.4869, pp. 1043–1052. ISSN: 00368075, 10959203. URL: http://www.jstor.org/stable/1702404.

Collins, M., R. Knutti, J. Arblaster, J.-L. Dufresne, T. Fichefet, P. Friedlingstein, X. Gao, W.J. Gutowski, T. Johns, G. Krinner, M. Shongwe, C. Tebaldi, A.J. Weaver, and M. Wehner (2013). "Long-term Climate Change: Projections, Commitments and Irreversibility". In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. Chap. 12, pp. 1029–1136. ISBN: ISBN 978-1-107-66182-0. DOI: 10.1017/CBO9781107415324.024. URL: www.climatechange2013.org.

Dask Development Team (2016). *Dask: Library for dynamic task scheduling*. URL: https://dask.org.

Davis, B. A. S., S. Brewer, A. C. Stevenson, and J. Guiot (2003). "The temperature of Europe during the Holocene reconstructed from pollen data". In: *Quat. Sci. Rev.* 22.15-17, pp. 1701–1716. ISSN: 02773791. DOI: 10.1016/s0277-3791(03)00173-2.

Davis, B. A. S., M. Zanon, P. Collins, A. Mauri, J. Bakker, D. Barboni, A. Barthelmes, C. Beaudouin, A. E. Bjune, E. Bozilova, R. H. W. Bradshaw, B. A. Brayshay, S. Brewer, E. Brugiapaglia, J. Bunting, S. E. Connor, J. L. de Beaulieu, K. Edwards, A. Ejarque, P. Fall, A. Florenzano, R. Fyfe, D. Galop, M. Giardini, T. Giesecke, M. J. Grant, J. Guiot, S. Jahns, V. Jankovska, S. Juggins, M. Kahrmann, M. Karpinska-Kolaczek, P. Kolaczek, N. Kuhl, P. Kunes, E. G. Lapteva, S. A. G. Leroy, M. Leydet, J. Guiot, S. Jahns, V. Jankovska, S. Juggins, M. Kahrmann, M. Karpinska-Kolaczek, P. Kolaczek, N. Kuehl, P. Kunes, E. G. Lapteva, S. A. G. Leroy, M. Leydet, J. A. L. Saez, A. Masi, I. Matthias, F. Mazier, V. Meltsov, A. M. Mercuri, Y. Miras, F. J. G. Mitchell, J. L. Morris, F. Naughton, A. B. Nielsen, E. Novenko, B. Odgaard, E. Ortu, M. V. Overballe-Petersen, H. S. Pardoe, S. M. Peglar, I. A. Pidek, L. Sadori, H. Seppa, E. Severova, H. Shaw, J. Swieta-Musznicka, M. Theuerkauf, S. Tonkov, S. Veski, W. O. van der Knaap, J. F. N. van Leeuwen, J. Woodbridge, M. Zimny, and J. O. Kaplan (2013). "The European Modern Pollen Database (EMPD) project". In: *Vegetation History and Archaeobotany* 22.6, pp. 521–530. ISSN: 0939-6314. DOI: 10.1007/s00334-012-0388-5. URL: http://link.springer.com/article/10.1007%2Fs00334-012-0388-5.

Davis, Basil A. S. and Simon Brewer (Feb. 2009). "Orbital forcing and role of the latitudinal insolation/temperature gradient". In: *Climate Dynamics* 32.2, pp. 143–165. ISSN: 1432-0894. DOI: 10.1007/s00382-008-0480-9. URL: https://doi.org/10.1007/s00382-008-0480-9.

Donders, Timme H., Friederike Wagner-Cremer, and Henk Visscher (Mar. 2008). "Integration of proxy data and model scenarios for the mid-Holocene onset of modern ENSO variability". In: *Quaternary Science Reviews* 27.5-6, pp. 571–579. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2007.11.010. URL: http://www.sciencedirect.com/science/article/pii/S0277379107003186.

Fægri, K., P. E. Kaland, and K. Krzywinski (1989). *Textbook of pollen analysis*. Ed. 4. Chichester, UK: John Wiley & Sons Ltd.

Fischer, N. and J. H. Jungclaus (Nov. 2011). "Evolution of the seasonal temperature cycle in a transient Holocene simulation: orbital forcing and sea-ice". In: *Climate of the Past* 7.4, pp. 1139–1148. DOI: 10.5194/cp-7-1139-2011. URL: https://www.clim-past.net/7/1139/2011/.

Flantua, Suzette G.A., Henry Hooghiemstra, Eric C. Grimm, Hermann Behling, Mark B. Bush, Catalina González-Arango, William D. Gosling, Marie-Pierre Ledru, Socorro Lozano-Garcìa, Antonio Maldonado, Aldo R. Prieto, Valentì Rull, and John H. Van Boxel (Dec. 2015). "Updated site compilation of the Latin American Pollen Database". In: *Review of Palaeobotany and Palynology* 223, pp. 104–115. DOI: 10.1016/j.revpalbo.2015.09.008.

Funder, Svend, Hugues Goosse, Hans Jepsen, Eigil Kaas, Kurt H. Kjær, Niels J. Korsgaard, Nicolaj K. Larsen, Hans Linderson, Astrid Lyså, Per Möller, Jesper Olsen, and Eske Willerslev (Aug. 2011). "A 10,000-Year Record of Arctic Ocean Sea-Ice Variability–View from the Beach". In: *Science* 333.6043, pp. 747–750. ISSN: 0036-8075. DOI: 10.1126/science.1202760. eprint: https://science.sciencemag.org/content/333/6043/747.full.pdf. URL: https://science.sciencemag.org/content/333/6043/747.

Fyfe, Ralph M., Jacques-Louis de Beaulieu, Heather Binney, Richard H. W. Bradshaw, Simon Brewer, Anne Le Flao, Walter Finsinger, Marie-Josè Gaillard, Thomas Giesecke, Graciela Gil-Romera, Eric C. Grimm, Brian Huntley, Petr Kunes, Norbert Kühl, Michelle Leydet, Andrè F. Lotter, Pavel E. Tarasov, and Spassimir Ton-kov (Mar. 2009). "The European Pollen Database: past efforts and current activities". In: *Vegetation History and Archaeobotany* 18.5, pp. 417–424. DOI: 10.1007/s00334-009-0215-9.

Github, Inc. (2019). "GitHub glossary". In: Last accessed: 2019-08-31. URL: https://help.github.com/en/articles/github-glossary (visited on 08/31/2019).

Goring, Simon, Andria Dawson, Gavin L Simpson, Karthik Ram, Russell W Graham, Eric C Grimm, and Jack W. Williams (2015). "neotoma: A Programmatic Interface to the Neotoma Paleoecological Database". In: *Open Quaternary* 1.1, p. 2. URL: http://doi.org/10.5334/oq.ab.

Grimm, Eric C. (1988). "Data analysis and display". In: *Vegetation history*. Ed. by B. Huntley, T. Webb, B. Huntley, and T. Webb. Dordrecht: Springer Netherlands, pp. 43–76. ISBN: 978-94-009-3081-0. DOI: 10.1007/978-94-009-3081-0_3. URL: https://doi.org/10.1007/978-94-009-3081-0_3.

— (1991). "Tilia and Tiliagraph". In: *Illinois State Museum, Springfield* 101.

— (2008). "Neotoma: an ecosystem database for the Pliocene, Pleistocene, and Holocene". In: *Illinois State Museum Scientific Papers E Series* 1. URL: https://www.neotomadb.org/uploads/NeotomaManual.pdf.

Guiot, Joel and Anne de Vernal (Oct. 2011). "QSR Correspondence "Is spatial autocorrelation introducing biases in the apparent accuracy of palaeoclimatic reconstructions?" Reply to Telford and Birks". In: *Quaternary Science Reviews* 30.21, pp. 3214–3216. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2011.07.023. URL: http://www.sciencedirect.com/science/article/pii/S0277379111002344.

Hargreaves, J. C., J. D. Annan, R. Ohgaito, A. Paul, and A. Abe-Ouchi (2013). "Skill and reliability of climate model ensembles at the Last Glacial Maximum and mid-Holocene". In: *Clim. Past* 9.2, pp. 811–823. ISSN: 1814-9332. DOI: 10.5194/cp-9-811-2013.

Hasecke, Jan Ulrich (2019). *Software-Dokumentation mit Sphinx: Zweite überarbeitete Auflage (Sphinx 2.0) (German Edition)*. Independently published. ISBN: 1793008779. URL: https://www.amazon.com/Software-Dokumentation-mit-Sphinx-%C3%BCberarbeitete-Auflage/dp/1793008779?SubscriptionId=AKIAIOBINVZYXZQZ

2U3A%5C&tag=chimbori05-20%5C&linkCode=xm2%5C&camp=2025%5C&creative=165953%5C&creativeASIN=1793008779.

Haslett, J., M. Whiley, S. Bhattacharya, M. Salter-Townshend, Simon P. Wilson, J. R. M. Allen, B. Huntley, and F. J. G. Mitchell (July 2006). "Bayesian palaeoclimate reconstruction". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169.3, pp. 395–438. DOI: 10.1111/j.1467-985x.2006.00429.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-985X.2006.00429.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-985X.2006.00429.x.

Havinga, A.J. (June 1967). "Palynology and pollen preservation". In: *Review of Palaeobotany and Palynology* 2.1-4, pp. 81–98. DOI: 10.1016/0034-6667(67)90138-8.

Holmström, Lasse, Liisa Ilvonen, Heikki Seppä, and Siim Veski (Sept. 2015). "A Bayesian spatiotemporal model for reconstructing climate from multiple pollen records". In: *Ann. Appl. Stat.* 9.3, pp. 1194–1225. DOI: 10.1214/15-AOAS832. URL: https://doi.org/10.1214/15-AOAS832.

Hoyer, S. and J. Hamman (2017). "xarray: N-D labeled arrays and datasets in Python". In: *Journal of Open Research Software* 5.1. DOI: 10.5334/jors.148. URL: http://doi.org/10.5334/jors.148.

Hunter, J. D. (May 2007). "Matplotlib: A 2D Graphics Environment". In: *Computing in Science Engineering* 9.3, pp. 90–95. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.55.

Iversen, Johs (Sept. 1944). "Viscum, Hedera and Ilex as Climate Indicators". In: *Geologiska Föreningen i Stockholm Förhandlingar* 66.3, pp. 463–483. DOI: 10.1080/11035894409445689. eprint: https://doi.org/10.1080/11035894409445689. URL: https://doi.org/10.1080/11035894409445689.

Jones, Eric, Travis Oliphant, Pearu Peterson, et al. (2001). *SciPy: Open source scientific tools for Python*. [Online; accessed 2017-02-18]. URL: http://www.scipy.org/.

Joussaume, S and KE Taylor (1995). "Status of the paleoclimate modeling intercomparison project (PMIP)". In: *World Meteorological Organization-Publications-WMO TD*, pp. 425–430.

Juggins, Steve (2007). "C2: Software for ecological and palaeoecological data analysis and visualisation (user guide version 1.5)". In: *Newcastle upon Tyne: Newcastle University* 77. URL: https://www.staff.ncl.ac.uk/stephen.juggins/software/C2Home.htm.

— (Mar. 2013). "Quantitative reconstructions in palaeolimnology: new paradigm or sick science?" In: *Quaternary Science Reviews* 64, pp. 20–32. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2012.12.014. URL: http://www.sciencedirect.com/science/article/pii/S0277379112005422.

— (2017). *rioja: Analysis of Quaternary Science Data*. R package version 0.9-21. URL: http://www.staff.ncl.ac.uk/stephen.juggins/.

Juggins, Steve and H. John B. Birks (2012). "Quantitative Environmental Reconstructions from Biological Data". In: *Tracking Environmental Change Using Lake Sediments: Data Handling and Numerical Techniques*. Ed. by H. John B. Birks, André F. Lotter, Steve Juggins, and John P. Smol. Dordrecht: Springer Netherlands, pp. 431–494. ISBN: 978-94-007-2745-8. DOI: 10.1007/978-94-007-2745-8_14. URL: https://doi.org/10.1007/978-94-007-2745-8_14.

Jungclaus, J. H., E. Bard, M. Baroni, P. Braconnot, J. Cao, L. P. Chini, T. Egorova, M. Evans, J. F. Gonzàlez-Rouco, H. Goosse, G. C. Hurtt, F. Joos, J. O. Kaplan, M. Khodri, K. Klein Goldewijk, N. Krivova, A. N. LeGrande, S. J. Lorenz, J. Luterbacher, W. Man, A. C. Maycock, M. Meinshausen, A. Moberg, R. Muscheler, C. Nehrbass-Ahles, B. I. Otto-Bliesner, S. J. Phipps, J. Pongratz, E. Rozanov, G. A.

Schmidt, H. Schmidt, W. Schmutz, A. Schurer, A. I. Shapiro, M. Sigl, J. E. Smerdon, S. K. Solanki, C. Timmreck, M. Toohey, I. G. Usoskin, S. Wagner, C.-J. Wu, K. L. Yeo, D. Zanchettin, Q. Zhang, and E. Zorita (2017). "The PMIP4 contribution to CMIP6 – Part 3: The last millennium, scientific objective, and experimental design for the PMIP4 *past1000* simulations". In: *Geosci. Model Dev.* 10.11, pp. 4005–4033. DOI: 10.5194/gmd-10-4005-2017. URL: https://www.geosci-model-dev.net/10/4005/2017/.

Kageyama, M., P. Braconnot, S. P. Harrison, A. M. Haywood, J. Jungclaus, B. L. Otto-Bliesner, J.-Y. Peterschmitt, A. Abe-Ouchi, S. Albani, P. J. Bartlein, C. Brierley, M. Crucifix, A. Dolan, L. Fernandez-Donado, H. Fischer, P. O. Hopcroft, R. F. Ivanovic, F. Lambert, D. J. Lunt, N. M. Mahowald, W. R. Peltier, S. J. Phipps, D. M. Roche, G. A. Schmidt, L. Tarasov, P. J. Valdes, Q. Zhang, and T. Zhou (2016). "PMIP4-CMIP6: the contribution of the Paleoclimate Modelling Intercomparison Project to CMIP6". In: *Geosci. Model Dev. Discuss.* 2016, pp. 1–46. DOI: 10.5194/gmd-2016-106. URL: https://www.geosci-model-dev.net/11/1033/2018/gmd-11-1033-2018.html.

Keim, Daniel, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon (2008). "Visual Analytics: Definition, Process, and Challenges". In: *Information Visualization: Human-Centered Issues and Perspectives*. Ed. by Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, Chris North, Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 154–175. ISBN: 978-3-540-70956-5. DOI: 10.1007/978-3-540-70956-5_7. URL: https://doi.org/10.1007/978-3-540-70956-5_7.

Krekel, Holger, Bruno Oliveira, Ronny Pfannschmidt, Floris Bruynooghe, Brianna Laugher, and Florian Bruhin (2004). *pytest 5.1*. URL: https://github.com/pytest-dev/pytest.

Marchant, Robert, Lucia Almeida, Hermann Behling, Juan Carlos Berrio, Mark Bush, Antoine Cleef, Joost Duivenvoorden, Maarten Kappelle, Paulo De Oliveira, Ary Teixeira de Oliveira-Filho, Socorro Lozano-Garìia, Henry Hooghiemstra, Marie-Pierre Ledru, Beatriz Ludlow-Wiechers, Vera Markgraf, Virginia Mancini, Marta Paez, Aldo Prieto, Olando Rangel, and Maria Lea Salgado-Labouriau (Aug. 2002). "Distribution and ecology of parent taxa of pollen lodged within the Latin American Pollen Database". In: *Review of Palaeobotany and Palynology* 121.1, pp. 1–75. DOI: 10.1016/s0034-6667(02)00082-9.

Marsicek, Jeremiah, Bryan N. Shuman, Patrick J. Bartlein, Sarah L. Shafer, and Simon Brewer (Feb. 2018). "Reconciling divergent trends and millennial variations in Holocene temperatures". In: *Nature* 554.7690, pp. 92–96. DOI: 10.1038/nature25464.

Mauri, A., B. A. S. Davis, P. M. Collins, and J. O. Kaplan (2014). "The influence of atmospheric circulation on the mid-Holocene climate of Europe: a data-model comparison". In: *Clim. Past* 10.5, pp. 1925–1938. ISSN: 1814-9324. DOI: 10.5194/cp-10-1925-2014. URL: http://www.clim-past.net/10/1925/2014/cp-10-1925-2014.pdf.

— (2015). "The climate of Europe during the Holocene: a gridded pollen-based reconstruction and its multi-proxy evaluation". In: *Quat. Sci. Rev.* 112, pp. 109–127. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2015.01.013. URL: http://www.sciencedirect.com/science/article/pii/S0277379115000372.

Nakagawa, Takeshi, Pavel E. Tarasov, Kotoba Nishida, Katsuya Gotanda, and Yoshinori Yasuda (Oct. 2002). "Quantitative pollen-based climate reconstruction in central Japan: application to surface and Late Quaternary spectra". In: *Quaternary Science Reviews* 21.18-19, pp. 2099–2113. DOI: 10.1016/s0277-3791(02)00014-8.

Neukom, Raphael, Luis A. Barboza, Michael P. Erb, Feng Shi, Julien Emile-Geay, Michael N. Evans, Jörg Franke, Darrell S. Kaufman, Lucie Lücke, Kira Rehfeld, Andrew Schurer, Feng Zhu, Stefan Brönnimann, Gregory J. Hakim, Benjamin J. Henley, Fredrik Charpentier Ljungqvist, Nicholas McKay, Veronika Valler, Lucien von Gunten, and P. A. G. E. S. 2k Consortium (2019a). "Consistent multidecadal variability in global temperature reconstructions and simulations over the Common Era". In: *Nature Geoscience* 12.8, pp. 643–649. ISSN: 1752-0908. URL: https://doi.org/10.1038/s41561-019-0400-0.

Neukom, Raphael, Nathan Steiger, Juan Josè Gòmez-Navarro, Jianghao Wang, and Johannes P. Werner (July 2019b). "No evidence for globally coherent warm and cold periods over the preindustrial Common Era". In: *Nature* 571.7766, pp. 550–554. DOI: 10.1038/s41586-019-1401-2.

Nichols, Harvey (1967). "The Post-glacial history of vegetation and climate at Ennadai Lake, Keewatin, and Lynn Lake, Manitoba (Canada)". In: E&G – Quaternary Science Journal 18.1. DOI: 10.23689/fidgeo-1124.

— (1969). "The Late Quaternary History of Vegetation and Climate at Porcupine Mountain and Clearwater Bog, Manitoba". In: *Arctic and Alpine Research* 1.3, p. 155. ISSN: 00040851. DOI: 10.2307/1550287. URL: http://www.jstor.org/stable/1550287.

Nocke, Thomas (2014). "Images for Data Analysis: The Role of Visualization in Climate Research Processes". In: *IMAGE POLITICS OF CLIMATE CHANGE: VISUALIZATIONS, IMAGINATIONS, DOCUMENTATIONS*. Ed. by Schneider, B and Nocke, T. Vol. 55. Image-Series, 55–77. ISBN: 978-3-8394-2610-4; 978-3-8376-2610-0.

Nocke, Thomas, Till Sterzel, Michael Böttinger, Markus Wrobel, et al. (2008). "Visualization of climate and climate change data: An overview". In: *Digital earth summit on geoinformatics*, pp. 226–232.

Nolan, Connor, John Tipton, Robert K Booth, Mevin B Hooten, and Stephen T Jackson (May 2019). "Comparing and improving methods for reconstructing peatland water-table depth from testate amoebae". In: *The Holocene* 29.8, pp. 1350–1361. DOI: 10.1177/0959683619846969.

Oliphant, Travis E (2006). *A guide to NumPy*. Vol. 1. Trelgol Publishing USA. URL: http://www.numpy.org/.

Otto-Bliesner, B. L., P. Braconnot, S. P. Harrison, D. J. Lunt, A. Abe-Ouchi, S. Albani, P. J. Bartlein, E. Capron, A. E. Carlson, A. Dutton, H. Fischer, H. Goelzer, A. Govin, A. Haywood, F. Joos, A. N. LeGrande, W. H. Lipscomb, G. Lohmann, N. Mahowald, C. Nehrbass-Ahles, F. S. R. Pausata, J.-Y. Peterschmitt, S. J. Phipps, H. Renssen, and Q. Zhang (2017). "The PMIP4 contribution to CMIP6 – Part 2: Two interglacials, scientific objective and experimental design for Holocene and Last Interglacial simulations". In: *Geosci. Model Dev.* 10.11, pp. 3979–4003. DOI: 10.5194/gmd-10-3979-2017. URL: https://www.geosci-model-dev.net/10/3979/2017/.

Parnell, Andrew C., James Sweeney, Thinh K. Doan, Michael Salter-Townshend, Judy R. M. Allen, Brian Huntley, and John Haslett (July 2014). "Bayesian inference for palaeoclimate with time uncertainty and stochastic volatility". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 64.1, pp. 115–138. DOI: 10.1111/rssc.12065. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssc.12065. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12065.

Peltier, W. R. (2004). "Global glacial isostasy and the surface of the Ice-Age earth: The ICE-5G (VM2) model and GRACE". In: *Annu. Rev. Earth Planet. Sci.* 32.1,

pp. 111–149. ISSN: 0084-6597 1545-4495. DOI: 10.1146/annurev.earth.32.082503.144359.

Perez, Fernando, Brian E. Granger, and John D. Hunter (Mar. 2011). "Python: An Ecosystem for Scientific Computing". In: *Computing in Science & Engineering* 13.2, pp. 13–21. DOI: 10.1109/mcse.2010.119.

Phillips, Norman A. (1956). "The general circulation of the atmosphere: a numerical experiment". In: *Quarterly Journal of the Royal Meteorological Society* 82.352, pp. 123–164.

Python Software Foundation (2019). *unittest – Unit testing framework*. URL: https://docs.python.org/3.7/library/unittest.html (visited on 09/02/2019).

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rautenhaus, Marc, Michael Böttinger, Stephan Siemen, Robert Hoffman, Robert M. Kirby, Mahsa Mirzargar, Niklas Rober, and Rudiger Westermann (Dec. 2018). "Visualization in Meteorology—A Survey of Techniques and Tools for Data Analysis Tasks". In: *IEEE Transactions on Visualization and Computer Graphics* 24.12, pp. 3268–3296. DOI: 10.1109/tvcg.2017.2779501.

Rocklin, Matthew (2015). "Dask: Parallel Computation with Blocked algorithms and Task Scheduling". In: *Proceedings of the 14th Python in Science Conference*. Ed. by Kathryn Huff and James Bergstra, pp. 130–136.

Shaw, Anthony (2018). *Getting Started With Testing in Python*. Last accessed: 2019-09-02. URL: https://realpython.com/python-testing/ (visited on 10/22/2018).

Simpson, G. L. (2007). "Analogue Methods in Palaeoecology: Using the analogue Package". In: *Journal of Statistical Software* 22.2, pp. 1–29.

Simpson, G. L. and J. Oksanen (2019). *analogue: Analogue and weighted averaging methods for palaeoecology*. R package version 0.17-3. URL: https://cran.r-project.org/package=analogue.

Sommer, Philipp S. (Aug. 2017). "The psyplot interactive visualization framework". In: *The Journal of Open Source Software* 2.16. DOI: 10.21105/joss.00363. URL: https://doi.org/10.21105/joss.00363.

— (2018a). *docrep: A Python Module for intelligent reuse of docstrings*. Last accessed: 2018-02-03. URL: https://github.com/Chilipp/docrep (visited on 02/03/2018).

— (2018b). *model-organization: Organize your computational models transparently*. Last accessed: 2018-02-03. URL: https://github.com/Chilipp/model-organization (visited on 02/03/2018).

— (Aug. 2019). *psy-strat v0.1.0: A Python package for creating stratigraphic diagrams*. DOI: 10.5281/zenodo.3381753. URL: https://doi.org/10.5281/zenodo.3381753.

Sommer, Philipp S. and Jed O. Kaplan (Oct. 2017). "A globally calibrated scheme for generating daily meteorology from monthly statistics: Global-WGEN (GWGEN) v1.0". In: *Geosci. Model Dev.* 10.10, pp. 3771–3791. DOI: 10.5194/gmd-10-3771-2017.

Sommer, Philipp S., Dilan Rech, Manuel Chevalier, and Basil A. S. Davis (Feb. 2019). "straditize: Digitizing stratigraphic diagrams". In: *Journal of Open Source Software* 4.34, p. 1216. DOI: 10.21105/joss.01216. URL: https://doi.org/10.21105/joss.01216.

Stodden, Victoria and Sheila Miguez (July 2014). "Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research". In: *Journal of Open Research Software* 2.1. DOI: 10.5334/jors.ay.

Telford, R.J. and H.J.B. Birks (Nov. 2005). "The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance". In: *Quaternary Science Reviews* 24.20, pp. 2173–2179. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2005.05.001. URL: http://www.sciencedirect.com/science/article/pii/S027737910500168X.

— (June 2009). "Evaluation of transfer functions in spatially structured environments". In: *Quaternary Science Reviews* 28.13, pp. 1309–1316. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2008.12.020. URL: http://www.sciencedirect.com/science/article/pii/S0277379108003806.

Tipton, John (2019). *BayesComposition: Fit forward and inverse prediction Bayesian functional models for compositional data*. R package version 1.0. URL: https://github.com/jtipton25/BayesComposition.

Torborg, Scott (2016). *python-packaging: Tutorial on how to structure Python packages*. Revision 35daf993. URL: https://python-packaging.readthedocs.io (visited on 09/02/2019).

United Nations (2019). *World Population Prospects 2019: Highlights*. Department of Economic and Social Affairs, Population Division.

Vincens, Annie, Anne-Marie Lézine, Guillaume Buchet, Dorothée Lewden, and Annick Le Thomas (2007). "African pollen database inventory of tree and shrub pollen types". In: *Rev. Palaeobot. Palynol.* 145.1-2, pp. 135–141. ISSN: 00346667. DOI: 10.1016/j.revpalbo.2006.09.004.

Walker, Mike, Sigfus Johnsen, Sune Olander Rasmussen, Trevor Popp, Jørgen-Peder Steffensen, Phil Gibbard, Wim Hoek, John Lowe, John Andrews, Svante Björck, Les C. Cwynar, Konrad Hughen, Peter Kershaw, Bernd Kromer, Thomas Litt, David J. Lowe, Takeshi Nakagawa, Rewi Newnham, and Jakob Schwander (2009). "Formal definition and dating of the GSSP (Global Stratotype Section and Point) for the base of the Holocene using the Greenland NGRIP ice core, and selected auxiliary records". In: *J. Quat. Sci.* 24.1, pp. 3–17. ISSN: 02678179 10991417. DOI: 10.1002/jqs.1227.

Wanner, Heinz, Jürg Beer, Jonathan Bütikofer, Thomas J. Crowley, Ulrich Cubasch, Jacqueline Flückiger, Hugues Goosse, Martin Grosjean, Fortunat Joos, Jed O. Kaplan, Marcel Küttel, Simon A. Müller, I. Colin Prentice, Olga Solomina, Thomas F. Stocker, Pavel Tarasov, Mayke Wagner, and Martin Widmann (Oct. 2008). "Mid- to Late Holocene climate change: an overview". In: *Quaternary Science Reviews* 27.19-20, pp. 1791–1828. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2008.06.013. URL: http://www.sciencedirect.com/science/article/pii/S0277379108001479.

Whitmore, J., K. Gajewski, M. Sawada, J.W. Williams, B. Shuman, P.J. Bartlein, T. Minckley, A.E. Viau, T. Webb, S. Shafer, P. Anderson, and L. Brubaker (Sept. 2005). "Modern pollen data from North America and Greenland for multi-scale paleoenvironmental applications". In: *Quaternary Science Reviews* 24.16-17. DOI: 10.1016/j.quascirev.2005.03.005.

Williams, John W., Eric C. Grimm, Jessica L. Blois, Donald F. Charles, Edward B. Davis, Simon J. Goring, Russell W. Graham, Alison J. Smith, Michael Anderson, Joaquin Arroyo-Cabrales, Allan C. Ashworth, Julio L. Betancourt, Brian W. Bills, Robert K. Booth, Philip I. Buckland, B. Brandon Curry, Thomas Giesecke, Stephen T. Jackson, Claudio Latorre, Jonathan Nichols, Timshel Purdum, Robert E. Roth, Michael Stryker, and Hikaru Takahara (Jan. 2018). "The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource". In: *Quaternary Research* 89.1, pp. 156–177. DOI: 10.1017/qua.2017.105.

Wodehouse, Roger Philip (1935). *Pollen grains: Their structure, identification and significance in science and medicine*. McGraw-Hill Book Co.

World Bank (2002). *Globalization, growth, and poverty : building an inclusive world economy*. The World Bank. URL: http://documents.worldbank.org/curated/en/954071468778196576/Globalization-growth-and-poverty-building-an-inclusive-world-economy.

Part I

New Software Tools for Paleoclimate Analysis

Chapter 2

The EMPD and POLNET web-interfaces

2.1 Summary

The Eurasian (née European) Modern Pollen Database (EMPD) was established in 2013 as a public database of quality controlled and standardized modern pollen surface sample data to compliment the European Pollen Database (EPD) for fossil pollen (Davis et al., 2013). The first version of the EMPD (referenced herein as the EMPD1) contained almost 5000 samples, submitted by over 40 individuals and research groups from all over Europe. Over the last 6 years more data has continued to be submitted, and more efforts have been made to incorporate more data held in open data repositories such as PANGAEA, and as supplementary information in published studies. This data is now released as the Eurasian Modern Pollen Database, version 2 (Davis et al., in prep) with an increase of 80 percent to 8663 samples (see figure 2.1).

The EMPD remains the only public and open access database of modern pollen samples covering the Eurasian continent and is entirely driven by the community of its data contributors. This effort of creating an open and accessible database led to the development of new open source data management tools that we present in this chapter. The EMPD2 is now hosted on the version control platform Github, with a dedicated web viewer at EMPD2.github.io and an automated administration app, the EMPD-admin (see table 2.1 for a list of the web resources). The new web-viewer provides an intuitive interface into the database and displays the essential meta information for every sample, as well as the pollen and climate data in a comprehensive bar plot. The integration with the EMPD-admin provides a simplified and transparent administration of multiple contributions from different sources

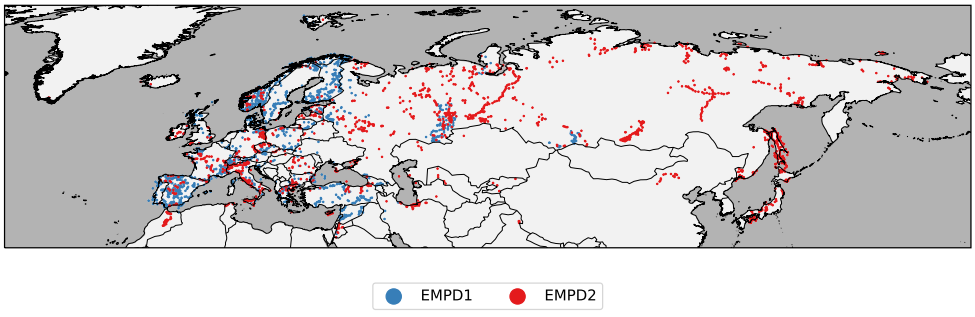


FIGURE 2.1: Modern calibration samples in the Eurasian Modern Pollen Database (EMPD).

TABLE 2.1: EMPD Web resources

	Description	Online Access
EMPD2	Github Organization	github.com/EMPD2
EMPD-Viewer	Map-based web interface to the EMPD database	github.com/EMPD2/EMPD-Viewer empd2.github.io
EMPD-Data	Version controlled data repository of the EMPD	github.com/EMPD2/EMPD-data
EMPD-Admin	Automated administration web app for the EMPD	github.com/EMPD2/EMPD-admin empd-admin.herokuapp.com EMPD2.github.io/EMPD-admin

and people to the database. All web components are hosted without any additional costs. The integration for the EMPD, that we present here, is only one example of a regional database. This framework can be extended to make other community-based (regional) pollen databases accessible, for instance the Latin American Pollen Database (LAPD) (Flantua et al., 2015) or African Pollen Database (APD) (Vincens et al., 2007). Especially the light-weight EMPD-viewer web interface can be ported to other database (as shown in section 2.3) to make heterogeneous data accessible to the broad public.

2.2 The EMPD web framework

The EMPD web framework is built on very common open source software development tools that have been adopted for a transparent data management, in favor of open science. The EMPD is now hosted on the web platform Github at github.com/EMPD2. This web platform, free of charge, hosts the source code for many popular open source software packages but can also be used to host a diverse, but small database (in terms of megabytes), such as the EMPD. Github builts upon the version control system *git* that transparently manages changes to documents by providing a full history of their revisions. The web platform is intrinsically designed for community-based projects that focus on collaboration and contains many features for a transparent communication between users, maintainers and contributors of a project. Besides others, the platform provides repository (i.e. project) specific discussion pages, so-called issues, where users can provide feedback, report bugs, or discuss any other aspect of the project. These issues are often linked to so-called pull requests, where each pull request is a proposal for a change in the source files of the project. This is then discussed between project maintainer and contributor in a dedicated discussion/review page.

Another common feature for Github repositories are integrations with so-called Continuous Integration (CI) services, e.g. for automated testing and/or packaging the software. These services run predefined scripts (for example test scripts) every time someone contributes to the repository, or creates a pull request.

The following sections describe how these software development tools are implemented in the three components of the EMPD web framework, the EMPD-viewer (section 2.2.1), the EMPD data repository (section 2.2.2) and the EMPD-admin (section 2.2.3).

2.2.1 The EMPD viewer

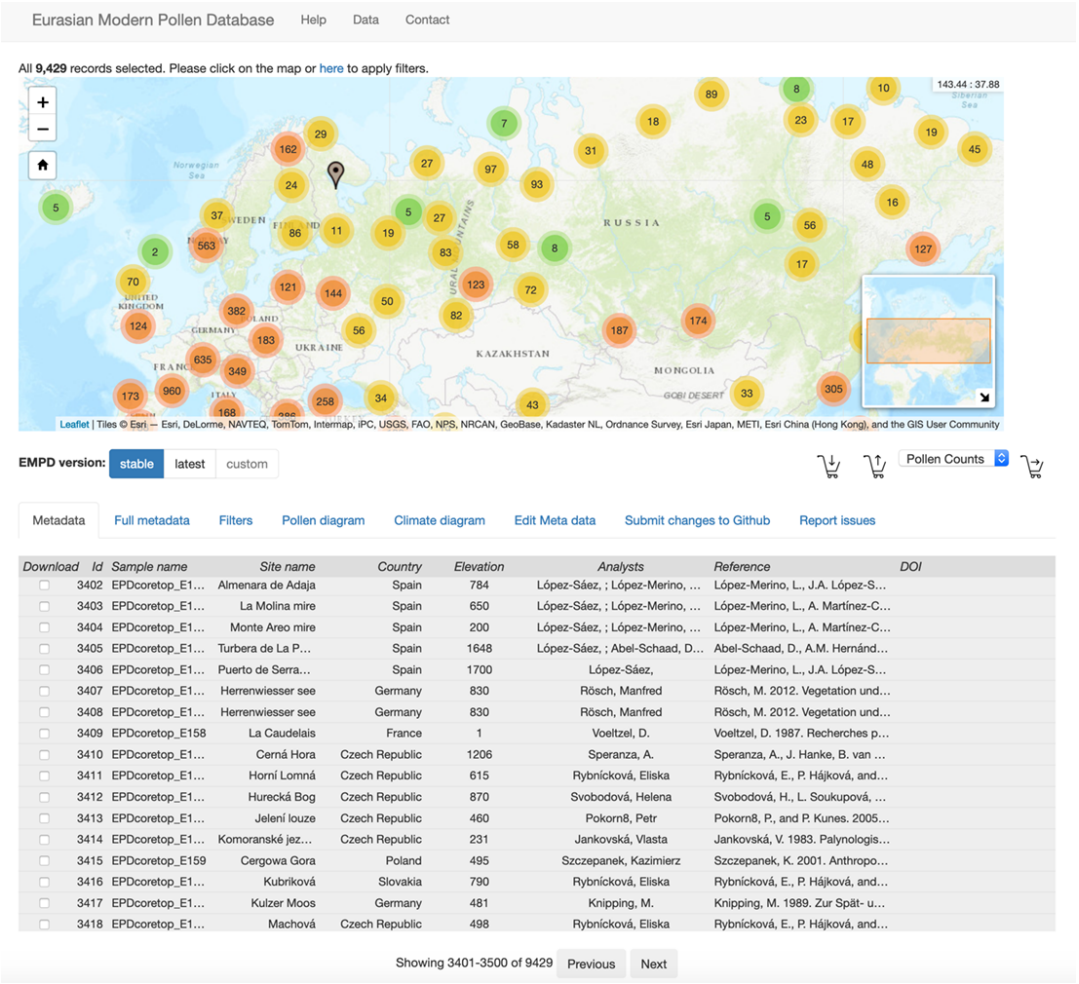


FIGURE 2.2: Screenshot of the EMPD viewer

The main public interface into the EMPD is an interactive web viewer accessible from EMPD2.github.io. This JavaScript-based application (see figure 2.2 for a screenshot) provides an intuitive interface into the database without requiring any particular computer expertise. It enables the user to view the data on a map and to select and download subsets of the database. The webpage involves no server-side processing and such it can be hosted for free using the service provided by Github Pages (pages.github.com). This provides a stable access to the database, independent of funding availabilities.

The Web Interface

The EMPD-Viewer has been initially based on the climate proxies finder (Bolliet et al., 2016; Brockmann, 2016) which can still be seen in it the layout and design of its graphical interface (i.e. its front-end). The code base, however, has been changed entirely, updated to the latest available versions of the underlying JavaScript dependencies and extended with multiple additional tools, shown in table 2.2. The central element of the viewer is a map to show the sample locations. It also allows to intuitive access to the essential meta data of every sample through the popup of the corresponding marker on the map. The detailed meta data can also be seen in

the meta data table, together with all the other samples. Another key element of the viewer are the meta data filters, that subset the data using efficient and intuitive filtering tools. This allows to search the database, or to select specific countries, climatic regimes, sample types, samples of a specific data contributor/analyst, and more.

Additional information on the sample is revealed through a bar diagram of the associated pollen data, which is dynamically created when the user clicks on the sample. The viewer also displays monthly, seasonal and annual precipitation and temperature values at the side, based on the WorldClim dataset, version 2 (Fick and Hijmans, 2017).

Finally, the viewer contains elements that allow scientists to contribute to the database, even without dedicated knowledge about the Github framework. The meta data editor allows to edit a sample and then submit it via the data submission form. The request is handled by the EMPD-admin webapp (see section 2.2.3) that pushes the data to the corresponding pull request on Github that is then reviewed by the core database maintainers. Another implemented element is an issue report form that allows the user to highlight erroneous sample information which is then, again through the EMPD-admin, submitted as a Github issue to the data repository.

The web app is fully integrated into the Github framework of the EMPD and loads the displayed data from the online repository. As such, it also provides a further quality control check and allows the data contributors/maintainers to review and edit new contributions before they are merged into the database.

Implementation details

The viewer itself is very light-weight and can be flexibly adapted to other database systems (see for example section 2.3). As the climate proxies finder (Bolliet et al., 2016; Brockmann, 2016), the EMPD-viewers main viewing/filtering functionality it is built upon the *dc* (Zhu and the dc.js Developers, 2019), *crossfilter* (Square, Inc. and crossfilter contributors, 2019) and *leaflet* (Agafonkin and leaflet contributors, 2019) open source JavaScript libraries. We ported the app to the *npm* package manager (npmjs.com) which enables a better and more secure monitoring of the app dependencies. This package manager is also used for an automated testing of the viewer on a Continuous Integration (CI) service, prior to deployment on the official web page. Due to time constrains, the viewer is not yet fully adapted to mobile devices.

2.2.2 The EMPD2 data repository

The raw data of the EMPD2 is accessible as plain text files in the *EMPD-data* Github repository (see table 2.1). The software development framework of Github (see introductory part of section 2.2) is adopted such that issues in the data repository can highlight errors in the database, or provide room for the discussion of potential new efforts that should be considered within the community-database. Pull requests into the repository are new data contributions that can be reviewed by the maintainers before being merged into the official database.

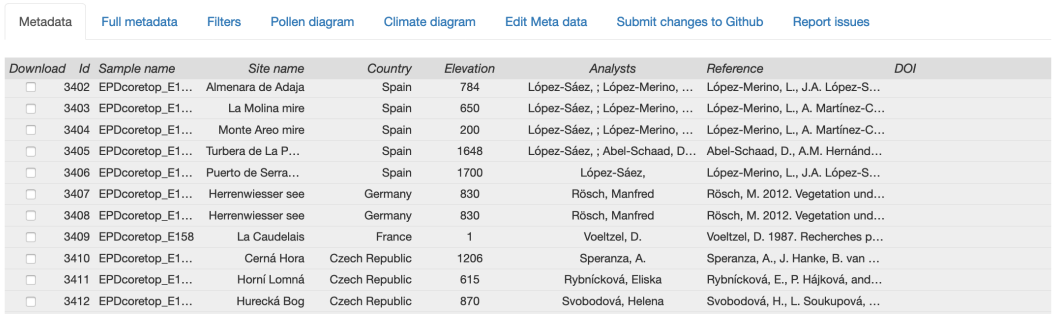
This method allows a fully transparent traceback of changes made to the EMPD through version control. The online access to the raw data files through Github also allows the EMPD viewer to interface with different versions of the database (see previous section).

The EMPD-data repository additionally uses the CI services from Travis CI (travis-ci.org) for automated tests of the meta data in each sample.

TABLE 2.2: Tools in the EMPD viewer

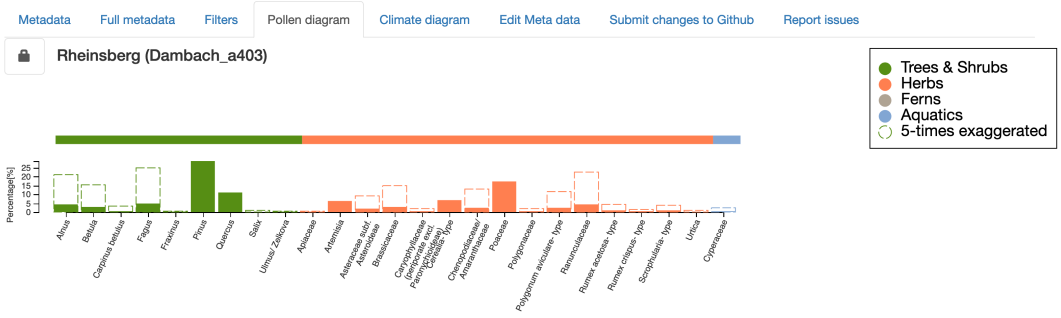
Map interface

Meta data table



Download	Id	Sample name	Site name	Country	Elevation	Analysts	Reference	DOI
	3402	EPDcoretop_E1…	Almenara de Adaja	Spain	784	López-Sáez, ; López-Merino, …	López-Merino, L., J.A. López-S…	
	3403	EPDcoretop_E1…	La Molina mire	Spain	650	López-Sáez, ; López-Merino, …	López-Merino, L., A. Martínez-C…	
	3404	EPDcoretop_E1…	Monte Areo mire	Spain	200	López-Sáez, ; López-Merino, …	López-Merino, L., A. Martínez-C…	
	3405	EPDcoretop_E1…	Turbera de La P…	Spain	1648	López-Sáez, ; Abel-Schaad, D…	Abel-Schaad, D., A.M. Hernánd…	
	3406	EPDcoretop_E1…	Puerto de Serra…	Spain	1700	López-Sáez,	López-Merino, L., J.A. López-S…	
	3407	EPDcoretop_E1…	Herrenwiesser see	Germany	830	Rösch, Manfred	Rösch, M. 2012. Vegetation und…	
	3408	EPDcoretop_E1…	Herrenwiesser see	Germany	830	Rösch, Manfred	Rösch, M. 2012. Vegetation und…	
	3409	EPDcoretop_E158	La Caudelais	France	1	Voeltzel, D.	Voeltzel, D. 1987. Recherches p…	
	3410	EPDcoretop_E1…	Cerná Hora	Czech Republic	1206	Speranza, A.	Speranza, A., J. Hanke, B. van …	
	3411	EPDcoretop_E1…	Horní Lomná	Czech Republic	615	Rybnícková, Eliska	Rybnícková, E., P. Hájková, and…	
	3412	EPDcoretop_E1…	Hurecká Bog	Czech Republic	870	Svobodová, Helena	Svobodová, H., L. Soukupová, …	

Pollen Data



Climate Data

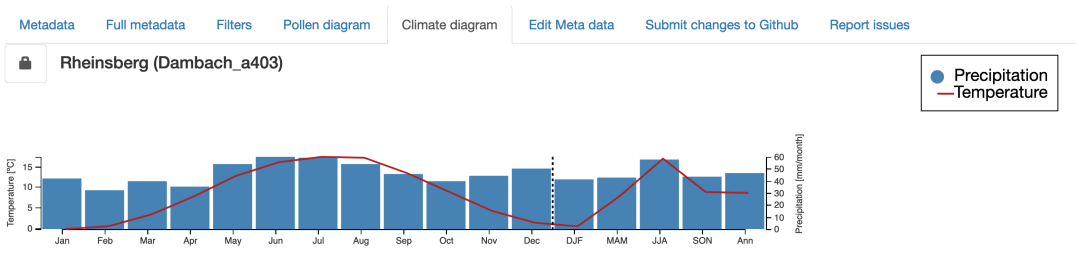
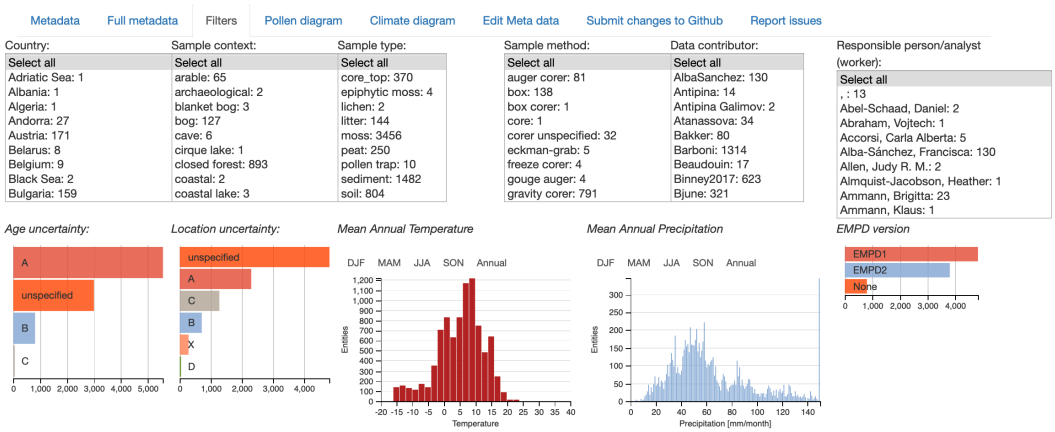
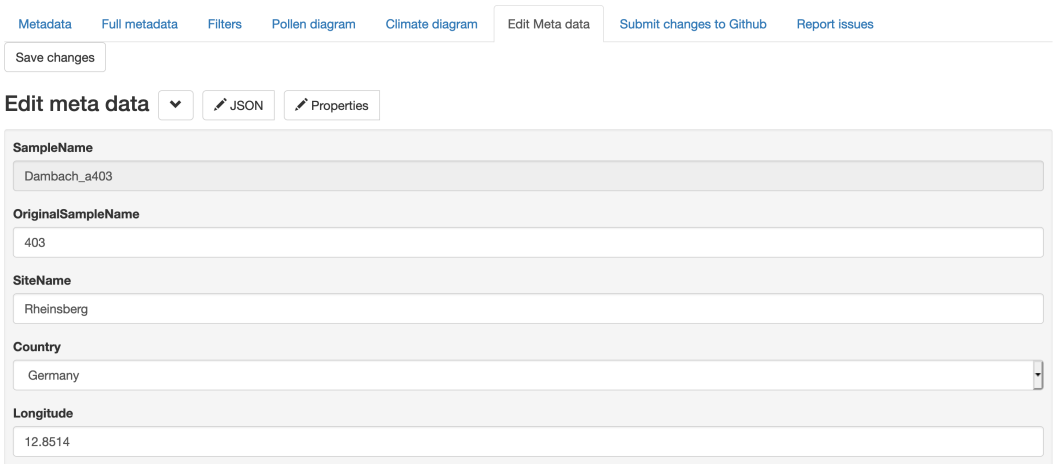


TABLE 2.2: Tools in the EMPD viewer (continued)

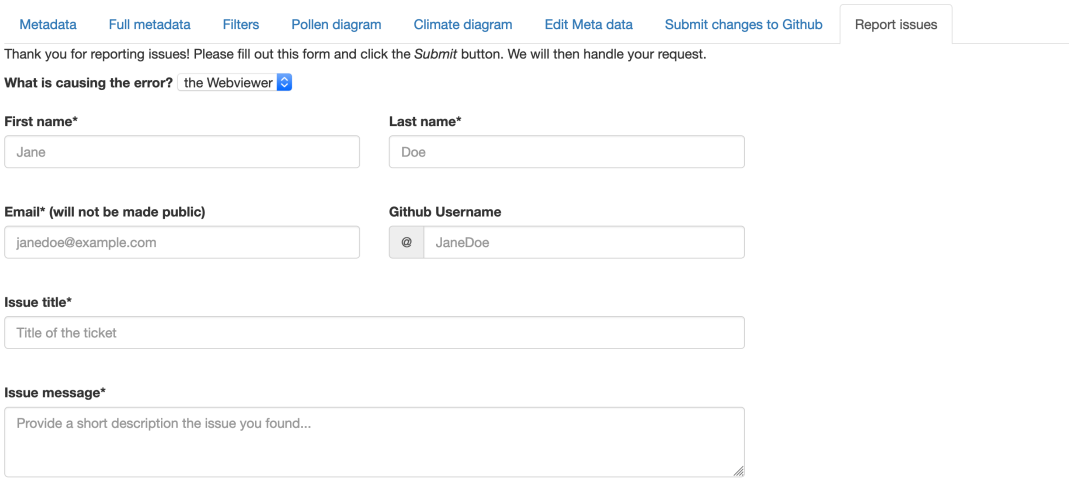
Meta data filter



Meta Data Editor



Issue submission form



2.2.3 The EMPD-admin

In addition to the standard CI services, we developed the EMPD-admin webapp. Inspired by the web management tools of the conda-forge community[1], this tool provides an automated handling of data contributions from within Github Pull Requests. It behaves like a standard CI service and runs tests on the data contribution, every time changes have been made to the pull request.

But the main purpose of the EMPD-admin is to provide a web tool for an automated administration of the database, which is helpful for a community-project with changing maintainers. Hence, the EMPD-admin web app acts like a bot that reacts on comments from within a pull request (i.e. a data contribution). Maintainers and contributors can use this functionality and directly contact the bot, for instance, to subset the data, run specific tests on subsets of the data, or automatically fix certain meta data issues, such as wrong countries or missing elevation.

The bot is also integrated in the EMPD-viewer (see previous section 2.2.1). Bug reports or edited data are processed by the EMPD-admin and put online as an issue in the github repository, or it updates the corresponding data contribution.

As such, the administration of the database can be done entirely remotely, without having to install dedicated software on a local computer.

Implementation details

The EMPD-admin webapp is hosted for free at Heroku (https://www.heroku.com) at empd-admin.herokuapp.com with a software package documentation hosted at EMPD2.github.io/EMPD-admin. This, again, allows stability independent on the availability of funding. The package can, also be installed locally and used from the command-line, independent of Github and Heroku, which is sometimes helpful for very large data contributions..

The Python library is based on the tornado web framework[2], as well as pandas (McKinney, 2010), a tabular data analysis library for Python, and sqlalchemy (Bayer, 2012), a Python SQL toolkit.

2.2.4 Distribution of the tools

The EMPD is hosted within the EMPD2 Github organization (github.com/EMPD2) in the EMPD-data repository. The source files of the viewer are accessible in the EMPD-viewer, and for the EMPD-admin in the EMPD-admin repository (see also table 2.1).

The EMPD-data and the EMPD-admin are additionally both available as so-called Docker container image at `https://hub.docker.com/u/empd2`. These containers are lightweight, standalone, executable packages of software that include everything needed to run an application: code, runtime, system tools, system libraries and settings. As such, they extend standard software packaging systems by providing an entire operating system that contains the target application. This makes it especially useful for web applications (such as the EMPD-admin) that can, as such, operate in a well-defined and portable environment.

The EMPD-admin can, however, also be installed through the standard python package manager `pip`.

[1] conda-forge.org
[2] www.tornadoweb.org

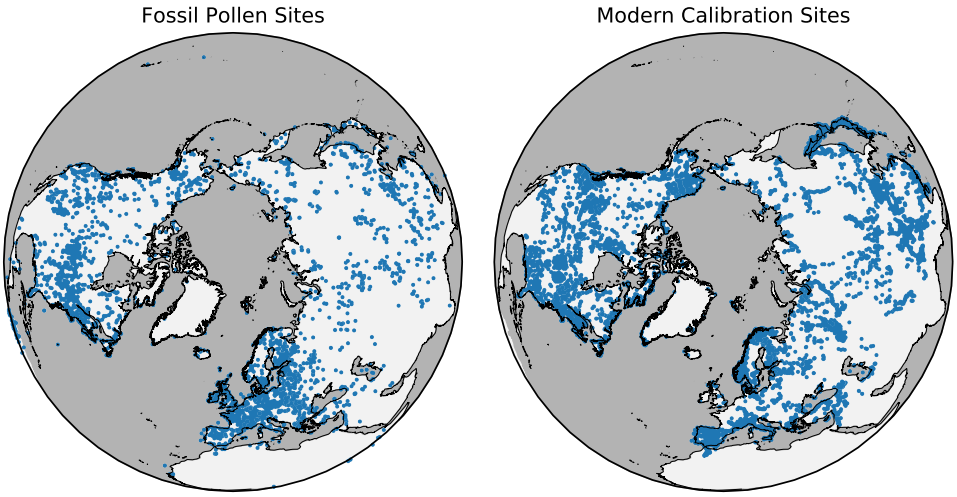


FIGURE 2.3: Maps of (left) fossil and (right) modern pollen sites in the POLNET database.

2.3 The POLNET viewer

The adaptability of the EMPD-viewer gave the motivation for an application with the POLNET database. This database, currently in development status, is a northern hemispheric, extra-tropical collection of modern and fossil pollen assemblages (Davis and Kaplan, 2017; Sommer et al., 2019). The purpose of this database is to generate the source for large-scale climate reconstruction during the Holocene (past 12'000 years) that can be used for model-data comparisons. It contains about 3'300 fossil pollen sites and about 13'200 modern surface samples (see figure 2.3) and is at present the largest existing collection of fossil and modern pollen samples. The database will soon be made publicly available through a dedicated web interface, the POLNET viewer. We present it here as a sample application of the EMPD-viewer to demonstrate how this web interface can be extended and applied to other datasets, in order to make them more accessible.

Like its core application, the EMPD-viewer, the POLNET-viewer is a map-based interface with implemented meta data filters. As it is a data exploration and distribution tool only, we did not include the functionalities to edit the meta data or to submit issues. Instead we implemented new features to visualize the essential aspects of this database: fossil pollen data and climate reconstructions.

The fossil pollen data is loaded upon request from the dedicated Github repository. It is afterwards visualized in form of a stratigraphic pollen diagram, with the age of the samples on the vertical y-axis, and the pollen taxa organized as vertically aligned diagram columns (see figure 2.4).

Climate reconstructions are displayed in two different manners: The site-based reconstructions are visualized as line plots in a separate diagram, together with their associated uncertainties. The gridded temperature reconstruction, i.e. the final product of the database (see also chapter 5) is visualized as an overlay on the map of the web application. This results in a combined visualizations of site-based and gridded reconstructions (see figure 2.5) which enables an intuitive regional analysis of the reconstruction method.

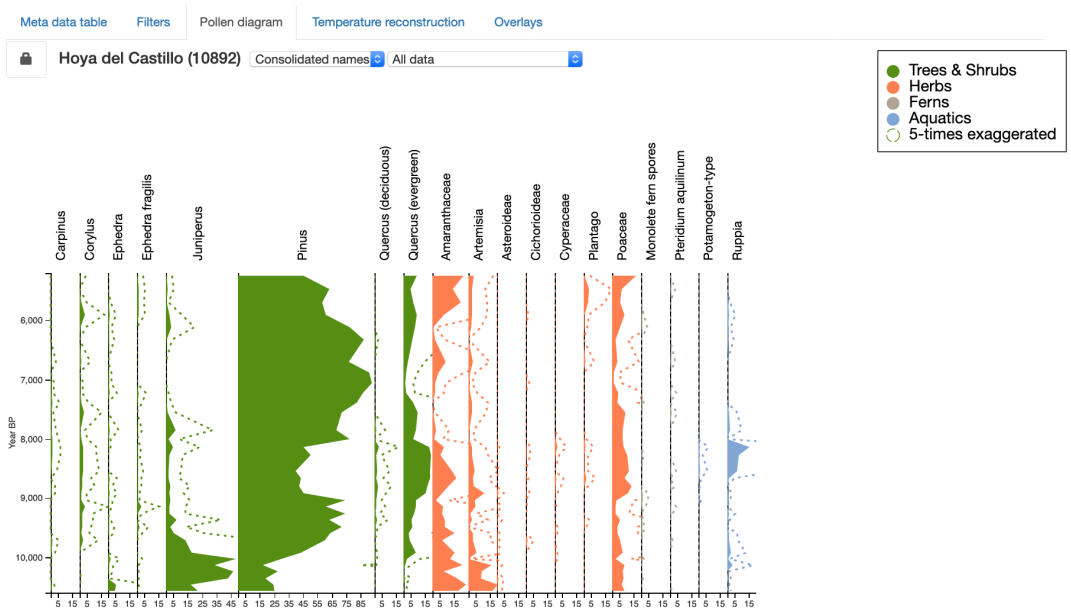


FIGURE 2.4: Screenshot of an automatically generated pollen diagram in the POLNET database viewer. The left dropdown menu above the pollen diagram allows to select the different naming schemes (here consolidated names that were used for the pollen-climate reconstruction). The right dropdown menu selects either the entire data or specific samples that are then displayed as a bar diagram (see the pollen data in table 2.2).

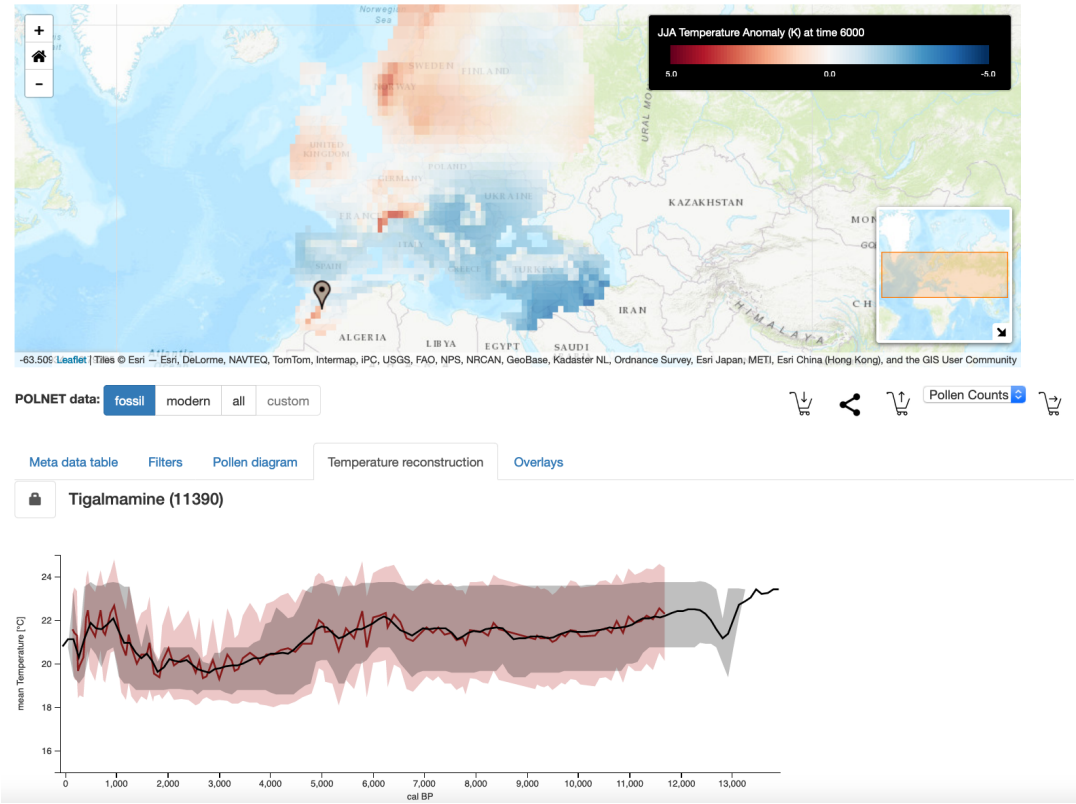


FIGURE 2.5: Exemplary screenshot how the climate reconstruction is visualized in the POLNET viewer. The map at the top figure shows the gridded temperature reconstruction (here 6k BP after Mauri et al., 2015). The lower plot shows the single site-based reconstruction (here Tigalmamine (Cheddadi et al., 1998)) for different reconstruction methods.

References

Agafonkin, Vladimir and leaflet contributors (2019). *Leaflet - an open-source JavaScript library for mobile-friendly interactive maps*. URL: http://crossfilter.github.io/crossfilter/ (visited on 05/14/2019).

Bayer, Michael (2012). "SQLAlchemy". In: *The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks*. Ed. by Amy Brown and Greg Wilson. aosabook.org. URL: http://aosabook.org/en/sqlalchemy.html.

Bolliet, Timothé, Patrick Brockmann, Valérie Masson-Delmotte, Franck Bassinot, Valérie Daux, Dominique Genty, Amaelle Landais, Marlène Lavrieux, Elisabeth Michel, Pablo Ortega, Camille Risi, Didier M. Roche, Françoise Vimeux, and Claire Waelbroeck (Aug. 2016). "Water and carbon stable isotope records from natural archives: a new database and interactive online platform for data browsing, visualizing and downloading". In: *Climate of the Past* 12.8, pp. 1693–1719. DOI: 10.5194/cp-12-1693-2016.

Brockmann, Patrick (2016). *ClimateProxiesFinder: dc.js + leaflet application to discover climate proxies*. Last accessed: 2019-08-30. URL: https://github.com/PBrockmann/ClimateProxiesFinder (visited on 10/06/2016).

Cheddadi, R., H. F. Lamb, J. Guiot, and S. van der Kaars (Oct. 1998). "Holocene climatic change in Morocco: a quantitative reconstruction from pollen data". In: *Climate Dynamics* 14.12, pp. 883–890. DOI: 10.1007/s003820050262.

Davis, B. A. S., M. Zanon, P. Collins, A. Mauri, J. Bakker, D. Barboni, A. Barthelmes, C. Beaudouin, A. E. Bjune, E. Bozilova, R. H. W. Bradshaw, B. A. Brayshay, S. Brewer, E. Brugiapaglia, J. Bunting, S. E. Connor, J. L. de Beaulieu, K. Edwards, A. Ejarque, P. Fall, A. Florenzano, R. Fyfe, D. Galop, M. Giardini, T. Giesecke, M. J. Grant, J. Guiot, S. Jahns, V. Jankovska, S. Juggins, M. Kahrmann, M. Karpinska-Kolaczek, P. Kolaczek, N. Kuhl, P. Kunes, E. G. Lapteva, S. A. G. Leroy, M. Leydet, J. Guiot, S. Jahns, V. Jankovska, S. Juggins, M. Kahrmann, M. Karpinska-Kolaczek, P. Kolaczek, N. Kuehl, P. Kunes, E. G. Lapteva, S. A. G. Leroy, M. Leydet, J. A. L. Saez, A. Masi, I. Matthias, F. Mazier, V. Meltsov, A. M. Mercuri, Y. Miras, F. J. G. Mitchell, J. L. Morris, F. Naughton, A. B. Nielsen, E. Novenko, B. Odgaard, E. Ortu, M. V. Overballe-Petersen, H. S. Pardoe, S. M. Peglar, I. A. Pidek, L. Sadori, H. Seppa, E. Severova, H. Shaw, J. Swieta-Musznicka, M. Theuerkauf, S. Tonkov, S. Veski, W. O. van der Knaap, J. F. N. van Leeuwen, J. Woodbridge, M. Zimny, and J. O. Kaplan (2013). "The European Modern Pollen Database (EMPD) project". In: *Vegetation History and Archaeobotany* 22.6, pp. 521–530. ISSN: 0939-6314. DOI: 10.1007/s00334-012-0388-5. URL: http://link.springer.com/article/10.1007%2Fs00334-012-0388-5.

Davis, Basil A. S., Manuel Chevalier, Philipp S. Sommer, et al. (in prep). "The Eurasian Modern Pollen Database (EMPD), Version 2". In: *Earth System Science Data ESSD*.

Davis, Basil A. S. and Jed O. Kaplan (Feb. 2017). *HORNET Holocene Climate Reconstruction for the Northern Hemisphere Extra-tropics*. SNF-Research-Plan. last accessed Jan, 30th, 2018. URL: http://p3.snf.ch/project-169598#.

Fick, Stephen E. and Robert J. Hijmans (May 2017). "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas". In: *International Journal of Climatology* 37.12, pp. 4302–4315. DOI: 10.1002/joc.5086. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.5086. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5086.

Flantua, Suzette G.A., Henry Hooghiemstra, Eric C. Grimm, Hermann Behling, Mark B. Bush, Catalina González-Arango, William D. Gosling, Marie-Pierre Ledru, Socorro Lozano-Garcìa, Antonio Maldonado, Aldo R. Prieto, Valentì Rull, and

John H. Van Boxel (Dec. 2015). "Updated site compilation of the Latin American Pollen Database". In: *Review of Palaeobotany and Palynology* 223, pp. 104–115. DOI: 10.1016/j.revpalbo.2015.09.008.

Mauri, A., B. A. S. Davis, P. M. Collins, and J. O. Kaplan (2015). "The climate of Europe during the Holocene: a gridded pollen-based reconstruction and its multi-proxy evaluation". In: *Quat. Sci. Rev.* 112, pp. 109–127. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2015.01.013. URL: http://www.sciencedirect.com/science/article/pii/S0277379115000372.

McKinney, Wes (2010). "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman, pp. 51–56.

Sommer, Philipp S., Basil A. S. Davis, Manuel Chevalier, Jian Ni, and John Tipton (2019). "The HORNET project: applying 'big data' to reconstruct the climate of the Northern Hemisphere during the Holocene". In: *20th Congress of the International Union for Quaternary Research (INQUA)*. International Union for Quaternary Research. URL: https://app.oxfordabstracts.com/events/574/program-app/submission/94623.

Square, Inc. and crossfilter contributors (2019). *Crossfilter - Fast Multidimensional Filtering for Coordinated Views*. URL: http://crossfilter.github.io/crossfilter/ (visited on 05/14/2019).

Vincens, Annie, Anne-Marie Lézine, Guillaume Buchet, Dorothée Lewden, and Annick Le Thomas (2007). "African pollen database inventory of tree and shrub pollen types". In: *Rev. Palaeobot. Palynol.* 145.1-2, pp. 135–141. ISSN: 00346667. DOI: 10.1016/j.revpalbo.2006.09.004.

Zhu, Nick and the dc.js Developers (2019). *dc.js - Dimensional Charting Javascript Library*. URL: https://dc-js.github.io/dc.js/ (visited on 05/14/2019).

Chapter 3

Straditize

A digitization software for pollen diagrams

Straditize *is published in the Journal of Open Source Software:*

Sommer, Philipp S., Dilan Rech, Manuel Chevalier, and Basil A. S. Davis (Feb. 2019). "straditize: Digitizing stratigraphic diagrams". In: *Journal of Open Source Software* 4.34, p. 1216. DOI: 10.21105/joss.01216. URL: https://doi.org/10.21105/joss.01216.

Abstract. The conversion of printed diagrams or figures into numerical data has become extremely important in ensuring that scientific work, especially from the pre or early digital age, is not lost to science. One of the most common figures used in the paleo-sciences is the stratigraphic diagram, where the results of the analysis of samples are plotted against a common y-axis, usually representing age or depth. Currently this type of diagram is laborious and error prone to digitize using current software designed for simple x/y graphs. Here we present a new open source software written in python that is specifically designed to quickly and accurately digitize stratigraphic diagrams based on a user controlled semi-automatic process. The software is optimized for use with pollen diagrams, but will work well with many other types of similar diagram. The software is fully documented and includes integrated help and tutorials.

3.1 Introduction

As with almost all areas of science, the digital capture, storage, manipulation and sharing of data has almost completely transformed the way that paleo-science is now undertaken compared to just 20-30 years ago. This digital transformation has created entirely new types of datasets, analysis, collaborations and visualizations, but it has also created a profound divide between the science that is available in digitized form, and that which is only available in analogue or paper form. This is a particular problem for science that was undertaken and published before this digital revolution, or where the original digital version of a dataset is unavailable, perhaps through retirement or other personnel changes, accidental damage to data storage devices, incompatible or out of date hardware storage or data file formats.

This data however may still be available as a published or printed diagram, which can be turned into numerical data by digitization, either manually or often using graph digitizing software such as *Graphclick*, *Engage Digitizer*, *Plot Digitizer*, *g3data*, *Digitizelt* and *WebPlotDigitizer* (Rohatgi, 2019). While this approach may be optimal for simple graphs with a single x and y axis, it can rapidly become extremely

time-consuming and error prone for stratigraphic diagrams with a shared y-axis and multiple x-axis. In a stratigraphic diagram the y-axis commonly takes the form of a depth or age scale (or both) reflecting sampling down a sediment core or open sediment section, which is then accompanied by a series of x-axis that plot the results of the analysis on each of these samples. Each sample may have been analyzed for a variety biotic or abiotic paleo-environmental indicators, and plotted on a variety of x-axis scales.

Here we present an open-source software *straditize* (Sommer et al., 2019) that has been specifically developed for digitizing stratigraphic diagrams. The software assumes that the figure follows the standard format associated with stratigraphic figures with a depth or age scale on the y-axis, and then a series of columns that represent the results of the analysis on common samples at specific depths or ages. The design is optimized for pollen diagrams (figure 3.1), but can be used without modification with any similar diagram design irrespective of the type of data being presented. The software first interprets the structure of the stratigraphic diagram and then reconstructs the data associated with each sample. This is done using a semi-automated process whereby many aspects of the software are automated but still editable by the user. The software allows the user to make continual visual checks on the digitization process, and provides the functionality to export the entire project in a data format that is independent of the platform and software[1]. The software is open-source and written in the programming language python (Perez et al., 2011) which makes it very flexible and easy to adapt. It is also equipped with an extensively documented graphical user interface, interactive visualizations and tutorials that allow the user to discover and to use the semi-automated methodologies. Additionally, *straditize* comes with an extensive test suite for a sustainable workflow with automated checks that also ensure the basic functionality of the various features in the software.

3.2 Methods: Treatment of stratigraphic diagram features

In this section we describe the common features of a stratigraphic pollen diagram and their handling within *straditize*. The emphasis is on pollen diagrams A pollen diagram highlighting the features in a stratigraphic diagram is provided in figure 3.1.

3.2.1 Structure of a stratigraphic diagram

3.2.1.1 Stratigraphic columns

A stratigraphic diagram consists of multiple sub diagrams, each visualizing one or more different variables, for example the percentages of different pollen taxa, the concentration of different chemical elements, or the various percentages of different grain sizes. These sub diagrams share one common axis which is usually the age or depth of the core (see fig. 3.1a). The diagram is then divided into multiple sub diagrams which we refer to as the columns of the stratigraphic diagram (fig. 3.1b). Each column visualizes the data of one variable, such as a pollen taxon, or multiple variables where these are plotted within the same column (same x-axis), such as winter and summer precipitation shown in the left most column of figure 3.1. The current version of *straditize* requires that the columns do not overlap and that

[1] *straditize* projects are exported as NetCDF file (Rew et al., 1989) that allows a platform and programming language independent access and sharing of the data

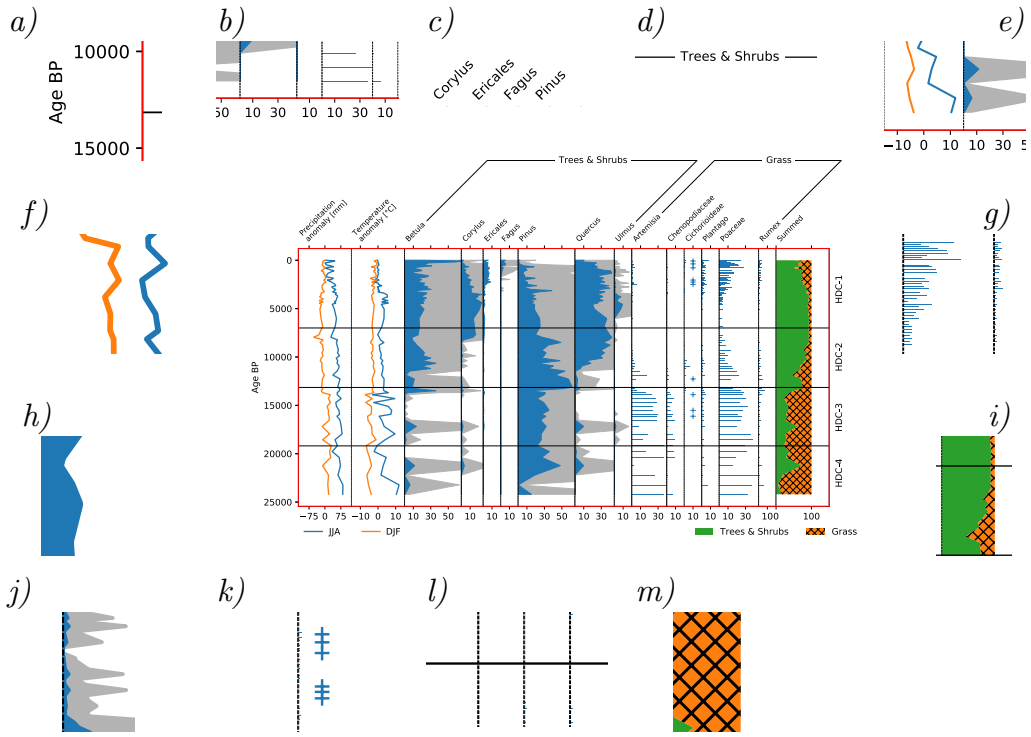


FIGURE 3.1: Common features in a pollen diagram. The center of the image is a pollen diagram with the data part being highlighted in red, the surrounding subfigures show some of it's special aspects. Subfigures a) – e) show common features in the diagram structure, subfigures f) – i) the different plotting styles and j) – m) some of the special features.

Each variable shares the same vertical axis a), is plotted as one column (sub diagram) marked by a separate vertical line b) and has a rotated title with the name of the variable c). The variables are grouped together d) and potentially have differing units on the horizontal axis e). Common plotting styles are line diagrams f), bar diagrams g), stacked diagrams h), or most commonly for pollen, as filled area diagram i). They may also be enhanced through exaggerations j), the visualization of taxon occurences k), horizontal lines as subdivisions of the core and vertical lines as y-axes for the columns l) and with hatches on the area plots for visual distinction m).

multiple variables plotted in the same column do not overlap either. The software can process multiple variables in a column so long as these are stacked or additive, and therefore they never overlap. An example in a pollen diagram could be where multiple species of, for instance Betula, are plotted as a stacked diagram in a single column so that the sum of the species also shows the total *Betula*.

At the top of each column it is usual to find the name of the variable plotted in the column. This may be shown at a variety of angles, but usually either vertically, or at an angle or rotated to make it easier to read and fit within the diagram (fig. 3.1c). This label can be automatically read by *straditize* and the name assigned to the respective column data, although care should be taken as the label is sometimes offset from the column it represents.

Variables are also often grouped together and the group labelled, for instance in pollen diagrams into *trees & shrubs*, and *herbs* (fig. 3.1d). For pollen data these groups usually share the same x-axis units/scaling (as in fig. 3.1b). In *straditize* the units/scaling can be set and applied to a whole group of columns/variables, or set and assigned for each individual column/variable (see fig. 3.1e).

3.2.1.2 Diagram types

A number of different diagram types are shown in figure 3.1 that are commonly used in pollen diagrams, as well as other stratigraphic diagrams. These can all be identified and read by *straditize*. One of the most commonly found diagrams in pollen diagrams are line diagrams (fig. 3.1f), or line diagrams where the area underneath of the line is filled to make an area diagram (fig. 3.1h). Data is also often commonly presented as bar diagrams that make it clearer where the individual samples are located (fig. 3.1g). Both bar and line/area diagrams may also be stacked, where (as we have already mentioned) columns may contain multiple variables stacked one upon the other (fig. 3.1i). These various diagram types require different digitization strategies, which we discuss in the digitization section below (see section 3.2.2).

3.2.1.3 Informative features

Other more specialized features can also be found in pollen diagrams that provide additional information for the reader, but are more difficult to interpret for the software. For instance the taxa or variables in a pollen diagram are usually all plotted on the same scale even if they are in different columns, so that direct visual comparison can be made between them. However, whilst this works well for large percentage values, it can often be difficult to see changes in low percentage values, which may still be ecologically important. To help the reader see these changes in low percentages, pollen diagrams often include a vertical exaggeration. This means that the percentages for a pollen taxa in a column will be plotted with two lines, one showing the percentage value shown on the scale, and the other showing the percentage value multiplied or exaggerated by a certain factor (usually 5 or 10) (fig. 3.1j). A different approach to the same problem is to mark the low percentages with a symbol or marker. For instance, a common method is to mark all samples with less than 0.5% or 1.0% with a "+" symbol (fig. 3.1k).

Other features that are often added to pollen diagrams and other stratigraphic diagrams are vertical and horizontal lines. Vertical lines often denote the start of a column, representing the baseline of the y-axes. These are often a continuous or discontinuous dashed line, and when it is quite a thick line it can be difficult to define

its position relative to the x-axis. Horizontal lines usually run across columns and are often used to denote different sections of the diagram. For instance, in pollen diagrams they are often used to denote zones or sections of the diagram that have a similar ecological assemblage. These horizontal lines do not usually provide useful numerical data and their intersection with column lines can make reading the column lines more difficult for the software. Another difficulty are hatch patterns which are especially common in old monochrome diagrams that predate the use of shading (fig. 3.1m).

3.2.2 Digitization procedure

Straditize digitizes the multiple columns or curves within a diagram in a single but editable action. This is different from other digitization software that usually requires the user to digitize each curve individually. This makes the digitization of a diagram with many columns much faster, and at the same time it enables the software to use all of the information in all of the different columns to help infer knowledge common to all columns, such as sample depths and percentage values (see section 3.2.2.6). This all-in-one digitization strategy however requires that *straditize* is able to understand and capture the structure of the diagram without encountering too many interpretation problems. Hence, instead of selecting the features that should be digitized, *straditize* first requires the user to remove all of the features that should not be digitized.

In summary, the digitization procedure for a stratigraphic diagram using *straditize* follows the following steps:

1. Define the data part of the diagram
2. Identify the columns representing the different variables
3. Clean-up the diagram by removing any unnecessary informative features (see section 3.2.2.3)
4. Decide how to handling exaggerations and rare occurrences
5. Digitize the diagram
6. Identify the samples
7. Translate the data into the correct x- and y-units
8. Read in the variable names

All of these steps are semi-automated and the results can (and should) be checked and edited by the user. Each step is fully reversible and the digitization process can be interrupted, saved and reopened at any time. In the following subsections we describe the algorithms of the different steps.

(1) Defining the data part of the diagram

The data part (see the red rectangle in fig. 3.1), displays the data of the diagram. Defining this part properly in the diagram image helps *straditize* to identify the part of the diagram from which data is to be extracted. Ideally it should not contain any labels for the horizontal or vertical axes, or any column headers or titles. If this cannot be avoided, these parts will have to be removed afterwards (see section 3.2.2.3).

Straditize uses a simple procedure to automatically detect the data part of the diagram. It looks for the two outer most pixel rows/columns that cover a certain fraction of the entire image (by default 70%). This algorithm works if the data part of the diagram is enclosed by a rectangle, as is shown in figure 3.1. If there is no rectangle enclosing the data then the user has to define the rectangle.

The data part is then transformed into a binary (black-and-white) version of the diagram image, and all of the informative features need to be removed. In the end, only data pixels, i.e. meaningful pixels that represent the numerical data behind the diagram, should be left (see the following section 3.2.2.3 and figure 3.2).

(2) Separating the columns

The next important aspect of the diagram structure that needs to be defined are the start of each of the columns (see section 3.2.1.1). This information is particularly important because a small error in each column can quickly sum up when digitizing a diagram with multiple columns. For instance, an error of one pixel in defining the start of every column in a pollen diagram that is about 1400 pixels wide and has 27 taxa is equivalent to an error of about 0.5 percent per taxon. In total this can easily introduce a summed error of up to 12% per sample. *Straditize* uses several criteria to detect the various columns. The start of each column is detected using the following procedure. Let $D(i)$ be the number of data pixels in a pixel column i (see black areas in 3.2. Then we assume a column start at pixel column i if

1. the previous pixel column $i-1$ did not contain any data ($D(i-1) = 0$)
2. the amount of data points doubled compared to $i-1$ ($D(i) \geq 2 \cdot D(i-1)$)
3. the amount of data points steadily increases within the next few columns to a value twice as large as the previous column ($D(i+n) \geq 2 \cdot D(i-1)$ with $n > 0$ and $D(i+j) \geq D(i)$ for all $0 < j \geq n$)

Additionally, the start of each potential column has to contain a user defined number of data pixels, which is by default ten percent of the height of the area containing the data within the diagram (the red rectangle in figure 3.1).

(3) Cleaning up the diagram

In order for the automated digitization algorithms to work effectively, *straditize* has to know which pixels contain data (i.e. is part of a line, or a bar), and which pixels are purely informative (such as y-axes or horizontal lines, see fig. 3.1l). It is necessary for the user to remove informative features from the data part of the diagram to get a clean version that will not confuse the algorithm. Figure 3.2 shows what this looks like for the sample diagram in figure 3.1. *Straditize* has multiple tools to facilitate the removing of informative features, documented in the software and manual, but their applicability depends on the diagram that is subject to digitization.

The most important step of cleaning the diagram is the recognition of the vertical axes (y- axes) that are usually at the start of every column. This is important because it defines the start of the x-axis, and therefore the value assigned to each of the data points, but also because in some cases the line can obscure part of the data itself. For instance, it is possible that the lowest values on the x-axis (see section 3.2.2.4) are obscured by the vertical line marking the y-axis, making accurate digitization difficult.

FIGURE 3.2: Cleaned binary image of the data part of figure 3.1. Informative features (Y-axes and horizontal lines) have been removed. Exaggerations and occurrences are still in the binary image and are considered separately in section 3.2.2.4.

Straditize therefore has an automatic algorithm to detect vertical axes that tries to minimize the removing of real data pixels. This detection is described by the following algorithm:

Let $C(i)$ be the most frequent color in the pixels of a pixel column i, and $D(i)$ the amount of data pixels in this column. A pixel column that is covered by data with more than a user-defined threshold (by default 30 percent of the data part height, see section 3.2.2.1) is considered as part of a y-axis if

- it is either the first pixel column in the subdiagram with data (i.e. $D(j) = 0$ for $j < i$ and j being larger or equal than the column start of the diagram),
- or the dominant color of the pixel column is the same as for the previous pixel column (i.e. $C(i) = C(i-1)$) and the number of data points is approximately the same (i.e. $D(i) \approx D(i-1)$).

This procedure results in one vertical line per column and works independent of whether it is a dotted, dashed or solid line. However, the line width is critical and may vary a lot. If a diagram column contains a filled line (fig. 3.1h) that merges with the vertical line defining the y-axis, then the algorithm could potentially overestimate the width of the y-axis line. Therefore, we use the median of all of the estimated lines from the various columns as the width of the y-axis line, and reduce the width of each of the lines to this amount. The algorithm then looks for informative features or other features that appear to be part of the data columns but are not part of the data behind the image. These include small features such as axis tick marks for example, and lighter pixels (i.e. close to white) that are usually a result from the rasterization of the diagram image. *straditize* then moves the start of the column because it assumes that the starting point for the x-axis (for pollen taxa it would be the 0% line) is in the middle of the vertical line marking the y-axis.

(4) Handling low taxon values

One particular feature often associated with pollen diagrams is the use of vertical exaggeration to help visualize changes in low percentages. Ordinarily, low percentages

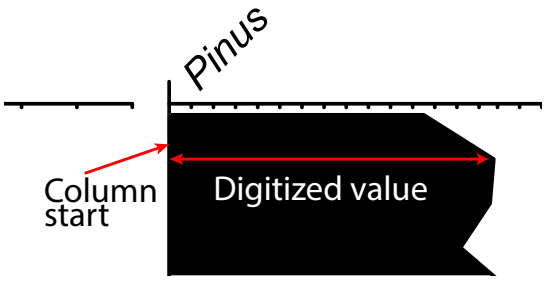


FIGURE 3.3: Illustration of the basic digitization strategy of *straditize*. For each diagram column in the binary image (fig. 3.2), we use the pixel that is located furthest to the right on the curve and take the distance to the column start as the digitized value. This is then transformed from the pixel scale into the data units based on the user input.

could be viewed better by changing the x-axis scale for the taxa with low percentages, but with pollen data it is also important to be able to make a visual comparison across all of the taxa listed, including those with high pollen percentages. Therefore, a common scale for all taxa is important. The exaggeration (fig. 3.1j) usually takes the form of a second line outside of the first, usually representing x5 or x10 exaggeration of the scale presented on the x-axis. This line could be expected to have greater precision than the first for any given pixel resolution, but problems emerge when the value of the exaggerated value exceeds the width of the scale on the x-axis, so that the line marking the exaggerated values is truncated at high values above a certain threshold. Another common way to help the reader identify low percentages in pollen diagrams is to use a marker (often a "+" symbol, fig. 3.1k) for values below a certain threshold. This can be particularly useful for very low counts where the author wants the reader to be aware that pollen of a certain taxa was found, even if the pollen counted was very low. This is often used for instance with "important" taxa such as *Cereals* that can indicate human agricultural activity, and taxa like *Larix* (Larch) that have notoriously low pollen productivity and where <1.0% in a pollen diagram may actually represent 20% of *Larix* trees in the surrounding landscape. For the purpose of digitization, the *straditize* user can either remove these exaggerations, or use some of the functions available in *straditize* to consider both the non-exaggerated and the exaggerated information in the diagram. In the case of the use of symbols to represent values below a threshold, it needs to be decided what value the symbol will represent once it is digitized and turned into numerical data. In any case, exaggerations and occurences are automatically removed from the image before the diagram is digitized (see next step 3.2.2.5).

(5) Digitizing the diagram

After removing the informative features (see section 3.2.2.3) and exaggerations (see section 3.2.2.4), *straditize* can automatically digitize the various columns on a pixel basis. In general *straditize* treats every column of the diagram separately and uses different algorithms for the various plotting types:

Area and line diagrams such as those shown in the fig. 3.1f and fig. 3.1h are digitized based on the pixel located furthest to the right on the curve in any particular column (illustrated in figure 3.3).

Bar diagrams as in figure 3.1g are also digitized based on the the pixel located furthest to the right on the bar in any particular column. Additionally, *straditize*

distinguishes between two adjacent bars by using a user defined threshold (by default two pixels). Additionally, it identifies bars that are significantly wider than the others (which would indicate two or more overlapping bars) and then the user can split them manually.

Stacked diagrams as in figure 3.1i have to be digitized manually. The user has to manually distinguish the different areas using the selection tools in *straditize*.

These procedures each result in one value per pixel row in each variable column in the data part. The next step is then to extract only the rows that are necessary to regenerate the diagram, that is, the pixel rows that correspond to the samples.

(6) Finding the samples

A key function of *straditize* is its ability to identify sample levels in the data, so that measurements of x-axis values in each column for each variable are assigned to the appropriate y-axis sample depth or age across all columns and variables. The search and assignment of sample levels can be done either automatically or manually, and if done automatically, this can also be later edited following manual checking.

The algorithm is thereby split into two steps:

1. For each column: Identify the intervals that contain exactly one sample (the rough locations, i.e. certain consecutive pixel rows in every column where we know that there is a sample, but we do not know where exactly)

2. Align the overlapping intervals between the columns to estimate the exact location

The implementation of step 1 necessarily differs between bar and area/stacked or line diagrams. With bar diagrams *straditize* uses the bars identified in the previous digitization step (section 3.2.2.5) to define these rough sample locations, while for the other diagram types the algorithm looks for local extrema in the graph line, i.e. intervals that are lower or higher than the surrounding areas. This implies that each sample is associated with a local minimum or maximum in at least one of the diagram columns. This holds well for pollen diagrams that usually sum up to 100% across all of the columns or variables in the diagram, but it is however not generally true for all stratigraphic diagrams.

Step 2 then aligns these rough intervals and uses the overlapping information from the different columns to estimate the exact location of the sample. This is described by the following procedure, where we focus on a simple case of only two diagram columns. Assume that the two columns i and j have a sample in the corresponding overlapping intervals I_i and I_j (i.e. $I_i = \{r_{i,1}, r_{i,2}, \ldots\}$, and $I_j = \{r_{j,1}, r_{j,2}, \ldots\}$ with $I_i \cap I_j \neq \emptyset$). To find the exact location, *straditize* distinguishes the following cases:

If one of the intervals contains only one pixel row (i.e. $|I_i| = 1$ or $|I_j| = 1$), *straditize* sets the sample at exactly this location

If each of the intervals contains multiple rows, *straditize* uses the mean of all the row indices in each of the intervals (i.e. $y = \overline{I_i \cup I_j}$). This then weighs overlapping areas in the intervals above non-overlapping areas.

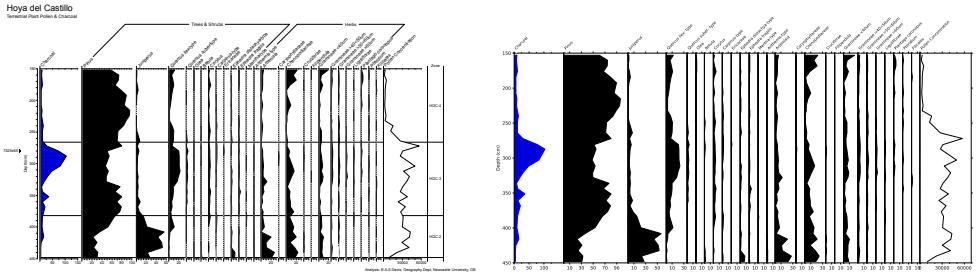


FIGURE 3.4: Pollen diagram for Hoya del Castillo after Davis and Stevenson, 2007. Left, the original diagram, right, the digitized version obtained (and plotted) using *straditize*.

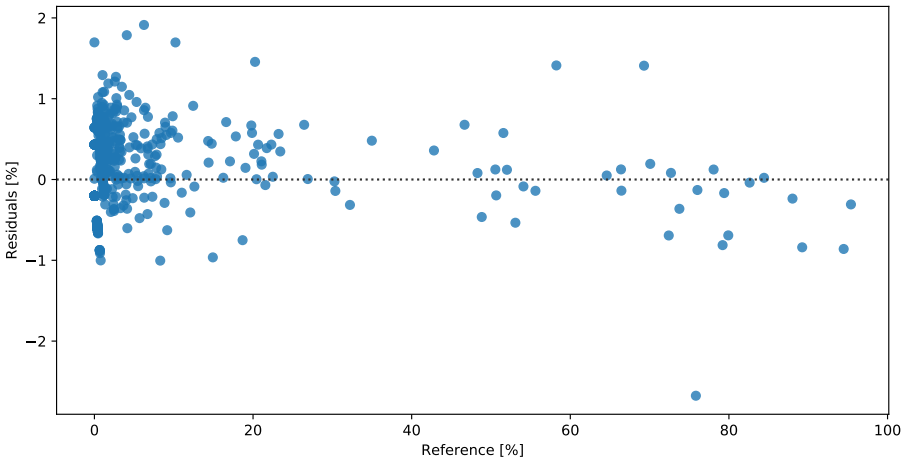


FIGURE 3.5: A plot of the residuals based on a comparison between the digitized Hoya del Castillo pollen diagram, and the original reference data that was used to generate the diagram (see fig. 3.4). The y-axis shows the residuals (digitized data percentage minus original reference data percentage) and the x-axis shows the original reference data percentage for Hoya del Castillo. Each dot represents a single pollen sample. The dotted line denotes the one-to-one line where digitization result and reference are the same.

Finally, samples that are close to each other (by default, closer than 5 pixel rows) are merged together. This is necessary because it may happen that, due to the quality of the diagram image, two rough locations do not exactly overlap although they belong to the same sample.

3.3 Discussion

As an example of the application of *straditize*, we digitized the Holocene pollen diagram for Hoya del Castillo (Spain) a site in Los Monegros, NE Spain, published by Davis and Stevenson, 2007 (see figure 3.4). The diagram was generated using the popular pollen diagram plotting software package Tilia[2] (Grimm, 1988, 1991) and exported as an image with a resolution of 450 dots per inch (dpi). We then followed the strategy described in section 3.2.2 to digitize the diagram. First, we selected the

[2] https://www.tiliait.com/

data part, then the columns and cleaned the image by removing the y-axes and horizontal lines. Then we digitized the diagram and used the *straditize* sample finding algorithm to extract the sample locations.

Comparing the digitized data with the original data, we find that *straditize* was able to successfully identify all 34 samples in the diagram. The root mean square error (RMSE) for the depth of each sample digitized from the y-axis and normalized by the range of the vertical axes, is very low and corresponds to an error of 0.2%. Additionally *straditize* gives a good measure of the individual taxon percentage with the RMSE being only 0.5%. Nevertheless, *straditize* shows a tendency to overestimate the real percentage. About 43% of the samples have percentages that are higher than the orginal percentages, whereas only about 10% have lower percentages (fig. 3.5), the remainder are exact (and most of the time zero percentage). This is probably due to a systematic bias in the positioning of the exact column start, i.e. 0 start point of the x-axis relative to y-axis baseline, and is something that could be systematically corrected. Overall this error is in the range of less than one percent per sample/taxon.

Irrespective of the performance of the software, other factors will also influence the accuracy of the digitization process. The quality and type of diagram image is very important, especially if the diagram is not of a high pixel resolution, or has poorly aligned and marked axis, or has many closely located samples. But also the skill and experience of the user will have some impact, especially if they are also unfamiliar with the type of data being analyzed and the way that it is commonly presented in a diagram. To help the user evaluate how accurate the digitization process has been, *straditize* also allows the user to plot the digitized data in a way that allows a direct visual comparison with the source diagram (fig. 3.4) using the stratigraphic visualization features of *psy-strat* (Sommer, 2019).

Keeping in mind these many caveats, we generally estimate that *straditize* will allow an experienced user to reliably obtain a numerical estimate through diagram digitization in the order of 1% of the original data for each sample/variable. In the case of pollen diagrams, this uncertainty should be viewed from the perspective of the inherent uncertainty associated with the counting of each pollen sample. Any pollen count is an estimate of the pollen sample that is displayed on a pollen slide. Although pollen samples are usually displayed as percentages, the reason for this is that the size of the pollen count varies from sample to sample, and percentages allow different samples to be directly compared on a common scale. Each count is a sample of the total pollen assemblage on a slide, and therefore each count represents an estimate of the composition of the total pollen assemblage represented on the slide. The more of that pollen assemblage that is counted, the closer that estimate will be to the actual pollen assemblage. This mean that each pollen sample plotted on a pollen diagram has an inherent uncertainty that is related to the total number of pollen grains counted. The bigger the count, the lower the uncertainty. Typical 0.95 confidence intervals for individual taxa based on a typical pollen count of around 300-500 pollen grains are easily in the order of 2-5% (Maher, 1972).

3.4 Conclusions

In this paper we present a new open-source software that is capable of greatly reducing the time required to accurately digitize stratigraphic diagrams. These diagrams are characterized by a series of horizontally or vertically aligned diagrams that plot

various variables representing the results of the analysis of a series of common samples that are aligned on the same y-axis representing age or depth. The software is currently optimized for use with pollen diagrams, but should work well with any similar type of data plotted in a similar style. The x-axis values can be percentages or absolute values of any kind, and the y-axis could also represent distance down a river of any other linear scale. The program is freely available, well documented with integrated help and training, is written in python, and is also open for adaptation for other uses.

References

Davis, Basil A. S. and A. C. Stevenson (2007). "The 8.2ka event and Early–Mid Holocene forests, fires and flooding in the Central Ebro Desert, NE Spain". In: *Quat. Sci. Rev.* 26.13-14, pp. 1695–1712. ISSN: 02773791. DOI: 10.1016/j.quascirev.2007.04.007. URL: https://dx.doi.org/10.1016/j.quascirev.2007.04.007.

Grimm, Eric C. (1988). "Data analysis and display". In: *Vegetation history*. Ed. by B. Huntley, T. Webb, B. Huntley, and T. Webb. Dordrecht: Springer Netherlands, pp. 43–76. ISBN: 978-94-009-3081-0. DOI: 10.1007/978-94-009-3081-0_3. URL: https://doi.org/10.1007/978-94-009-3081-0_3.

— (1991). "Tilia and Tiliagraph". In: *Illinois State Museum, Springfield* 101.

Maher, Louis J. (Apr. 1972). "Nomograms for computing 0.95 confidence limits of pollen data". In: *Review of Palaeobotany and Palynology* 13.2, pp. 85–93. ISSN: 0034-6667. DOI: 10.1016/0034-6667(72)90038-3. URL: http://www.sciencedirect.com/science/article/pii/0034666772900383.

Perez, Fernando, Brian E. Granger, and John D. Hunter (Mar. 2011). "Python: An Ecosystem for Scientific Computing". In: *Computing in Science & Engineering* 13.2, pp. 13–21. DOI: 10.1109/mcse.2010.119.

Rew, Russ, Glenn Davis, Steve Emmerson, Cathy Cormack, John Caron, Robert Pincus, Ed Hartnett, Dennis Heimbigner, Lynton Appel, and Ward Fisher (1989). *Unidata NetCDF*. en. DOI: 10.5065/D6H70CW6.

Rohatgi, Ankit (Apr. 2019). *WebPlotDigitizer, version 4.2*. URL: https://automeris.io/WebPlotDigitizer.

Sommer, Philipp S. (Aug. 2019). *psy-strat v0.1.0: A Python package for creating stratigraphic diagrams*. DOI: 10.5281/zenodo.3381753. URL: https://doi.org/10.5281/zenodo.3381753.

Sommer, Philipp S., Dilan Rech, Manuel Chevalier, and Basil A. S. Davis (Feb. 2019). "straditize: Digitizing stratigraphic diagrams". In: *Journal of Open Source Software* 4.34, p. 1216. DOI: 10.21105/joss.01216. URL: https://doi.org/10.21105/joss.01216.

Chapter 4

Psyplot

A flexible framework for interactive data analysis

4.1 Summary

From

Sommer, Philipp S. (Aug. 2017e). "The psyplot interactive visualization framework". In: *The Journal of Open Source Software* 2.16. DOI: 10.21105/joss.00363. URL: https://doi.org/10.21105/joss.00363.

psyplot (Sommer, 2017e) is a cross-platform open source python project that mainly combines the plotting utilities of matplotlib (Hunter, 2007) and the data management of the xarray (Hoyer and Hamman, 2017) package and integrates them into a software that can be used via command-line and via a graphical user interface (GUI).

The main purpose is to have a framework that allows a fast, attractive, flexible, easily applicable, easily reproducible and especially an interactive visualization of data.

The ultimate goal is to help scientists in their daily work by providing a flexible visualization tool that can be enhanced by their own visualization scripts.

The framework is extended by multiple plugins: psy-simple (Sommer, 2017c) for simple visualization tasks, psy-maps (Sommer, 2017a) for georeferenced data visualization and psy-reg (Sommer, 2017b) for the visualization of fits. It is furthermore extended by the optional graphical user interface psyplot-gui (Sommer, 2017d).

4.2 Introduction

The mathematical and statistical processing of climate data is closely related to its visualization and analysis. But in traditional visual analytics literature, these two aspects are commonly treated in separate manners. Keim et al., 2008 for instance, (following Wijk, 2005) distinguish two steps of visual analytics, the initial data processing with statistical or mathematical techniques, and a *sense-making loop* of visualization, exploration and the gain of new knowledge. Böttinger and Röber, 2019 distinguish the *filtering* step (data processing), and *mapping/rendering* step that describes the visualization. Also in the literature there is a clear division between the climate visualization (or visual analytic) papers and the standard statistical or climate literature that describes new methods for data processing. Visualization research focuses mainly on advanced visualization tools such as ParaView (Ayachit,

2015), VAPOR (Clyne et al., 2007) or Avizo[1] (e.g. Böttinger and Röber, 2019; Nocke et al., 2015; Rautenhaus et al., 2018; Wong et al., 2014) whereas statistical or climate literature commonly uses R (R Core Team, 2019), Python (Oliphant, 2006; Perez et al., 2011), Climate Data Operators (CDOs) (Schulzweida, 2019) or other command-line tools.

This separation, however, devalues the interplay between the new knowledge from the visualization step, that commonly raises the need for more statistical and mathematical processing of the initial data. This calls for integrated and flexible tools that tackle both steps: the data processing and the visualization, a requirement that is currently not fulfilled by the visualization tools described above. An example software that integrates data processing and data visualization is provided with the Earth System Model Evaluation Tool (ESMValTool) (Eyring et al., 2016). This framework provides common diagnostics for Earth System Models (ESMs) to enable model intercomparisons. The tool, however, has limited interactivity and a slow learning curve for the implementation of new diagnostics.

This lack leads to large efforts of climate scientists to develop scripts for the data processing and visualization. They usually do not follow a systematic framework and as such need to be adapted every time a new project starts which also make them difficult to share with other researchers. The new *psyplot* framework wants to generalize this data processing and visualization by providing a framework that is highly flexible, interoperates with standard computational data processing tools and enables flexible visualizations and adaptations. The software is written in the programming language Python (Perez et al., 2011) and builds upon the visualization package *matplotlib* (Hunter, 2007) and the N-dimensional array processing package *xarray* (Hoyer and Hamman, 2017), that closely interoperates with the numeric packages numpy and scipy (Jones et al., 2001; Oliphant, 2006) and the parallel computing library dask (Dask Development Team, 2016). Due to the flexibility of Python, it can be used from the command-line, a graphical user interface (GUI) (section 4.3.3) or jupyter notebooks[2] (Kluyver et al., 2016). As such, it supports out-of-core computation (i.e. the processing of data too large to fit into memory), a rich set of visualization methods from matplotlib, and can be extended to other visualization packages, such as the 3D-visualization framework VTK (Sommer, 2019b).

The next section 4.3 provide an overview of the framework with its data model, plugins and GUI. Sections 4.4 and 4.5 finally discuss further usage and extensions to the software. For more information, usage and implementation examples I also refer to the online documentation `https://psyplot.readthedocs.io`.

4.3 The psyplot framework

The psyplot framework consists of three parts: The core structure that is built upon *xarray* and provides the general infrastructure (section 4.3.1), the plugins that use the plotting functionalities of *matplotlib* (section 4.3.2), and the GUI (section 4.3.3).

4.3.1 Data model

Psyplot and xarray

psyplot acts as a high-level interface into the packages *xarray* and *matplotlib*. The first one is a recent package for N-dimensional labeled arrays that adopts Unidata's

[1] `https://www.fei.com/software/avizo3d/`
[2] `https://jupyter.org/`

self-describing Common Data Model on which the network Common Data Form (netCDF) is built (Brown et al., 1993; Hoyer and Hamman, 2017; Rew and Davis, 1990). The package integrates with standard python from the python environment, such as the computing and analysis packages numpy (Oliphant, 2006), scipy (Jones et al., 2001; Oliphant, 2007), pandas (McKinney, 2010) and statsmodels (Seabold and Perktold, 2010), but also offers intuitive interfaces for other packages, such as a package for empirical orthogonal functions (EOFs, Dawson, 2016), CDOs (Müller, 2019), fourier transforms (Uchida et al., 2019) and many more[3]. This large potential for extension distinguishes psyplot from other high-level visualization software, such as ParaView or Vapor, as such python packages can be implemented as a so-called *formatoption* (without hyphen, see below) or used in a pre-processing step.

Psyplot core structure

The core structure of psyplot consists of five base classes that interact with each other, the visualization objects *Plotter* and its *Formatoptions*, the data objects *DataArray*, an *InteractiveList* of them, and a collection of all of them, the psyplot *Project*. It is schematically visualized in figure 4.1.

The most high-level Application programming interface (API) object is the psyplot project that consists of multiple data objects that are (or are not) visualized. The main purpose is a parallel handling of multiple plots/arrays that may also interact with each other (e.g. through the sharing of *formatoptions*). It mainly spreads update commands to its contained objects, but also serves as a filter for the data objects. Furthermore, one project may be split up into sub projects which then only control a specific part of the main project, e.g. for a specific formatting of only a small part of the data.

The next level is the *DataArray* from the xarray package (or more explicitly, its accessor, the *InteractiveArray*[3]), that holds the data of one (or more) variables (e.g. temperature) and its corresponding coordinates (e.g. time, latitude, longitude, etc.). It may be one or multidimensional depending on the chosen visualization method. psyplot offers several methods to provide the coordinates for the plotting of different grids to make the visualization easier. The software can interpret CF Conventions[4] and UGRID conventions for unstructured grids (Jagers et al., 2018).

Multiple of these arrays can also be grouped together into an *InteractiveList* that shall be visualized by the same plot method (e.g. multiple lines or a scalar field with overlying vector field).

The visualization part in the framework is managed by the *Plotter* class, a collection of multiple *Formatoptions*. Each plotter subclass is designed to visualize the data in a specific manner (e.g. via line plots, violin plots, or map plots) and is completely defined through it's *formatoptions*.

Formatoptions are the core of the psyplot structure. The standard functionality of a formatoption is to control the visual appearance of one aspect of the plot (e.g. through the colormap, figure title, etc.). It is, however, completely unlimited and can also do data manipulations or calculations. The psy-reg plugin for example (see section 4.3.2) implements a formatoption that performs a regression through the data

[3] several packages related to xarray are listed in the docs at `http://xarray.pydata.org/en/stable/related-projects.html` and psyplots integration (accessors) in particular is shown at `https://psyplot.readthedocs.io/en/latest/accessors.html`.

[4] `http://cfconventions.org`

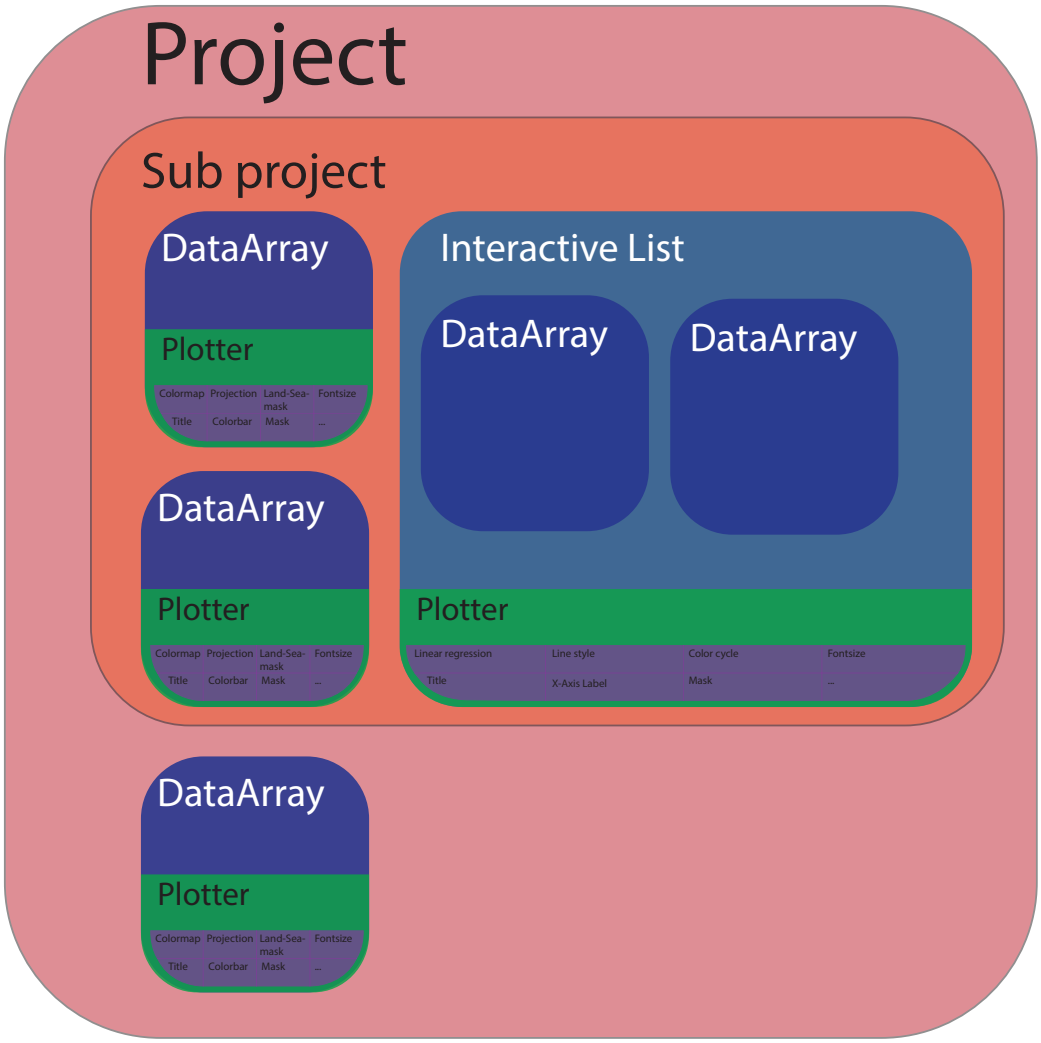


FIGURE 4.1: The psyplot core framework. A (sub) project consists of n-dimensional data arrays or a list of these that are each visualized by a plotter. Each plotter consists of a set of *formatoptions* that control the appearance of the plot or performs data manipulation.

that is then visualized. As mentioned earlier, each plotter is set up through its *formatoptions* where each formatoption has a unique formatoption key inside the plotter. This formatoption key (e.g. *title* or *cmap*) is what is used for updating the plot, manipulating the data, etc.. *Formatoptions* might also interact with other *formatoptions* inside the plotter or from other plotters. This concept of *formatoptions* allows to use the same formatoption with all different kinds of plotters and the interaction of multiple plots with each other. Common plot features, such as the figure title, colormap, etc., therefore do not have to be implemented explicitly for every plotter but can be used from existing implementations. This framework also allows a very easy integration and development of own *formatoptions* with a low or high level of complexity.

4.3.2 Psyplot plugins

The psyplot package provides the core of the data management described in the previous section 4.3.1. The real visualization is implemented in external plugins. The advantage of this approach is an increased flexibility of the entire framework (collaborations can evolve through dedicated plugins) and of managing the various dependencies of the packages. As such, the dependencies of psyplot are rather week (only xarray is needed), but the dependencies of the plugins can be more extensive (e.g. for geo-referencing or advanced statistics).

Each plugin defines new *Plotters* and *Formatoptions* that are specific to the purpose of the visualization/analysis task. The plotters can also be implemented as a plot method (see supplements 4.B to 4.E) and accessed through the psyplot core API (see supplements 4.A for an example).

The current lists of plugins include *psy-simple* for rather simple and standard visualization tasks, *psy-maps* for geo-referenced plots, *psy-reg* for statistical analysis visualization, and *psy-strat* for stratigraphic diagrams.

psy-simple: The psyplot plugin for simple visualizations

Much of the functionality that is used by other plugins is developed in the psy-simple plugin. This package targets simple visualizations and currently includes plot methods for one-dimensional data: line plots, bar plots and violin plots; for two-dimensional data: scalar plots, vector plots and combined scalar and vector plots; and plots that do not require complex data manipulation: a density plot and a plot of the weighted geographic mean. A full list of examples is provided in the supplementary material, section 4.B.

This package also implements most of the functionality to handle unstructured grids in 2D visualizations and defines most of the commonly used *formatoptions*. The latter include text manipulation (such as plot title, figure title, x- and y-axis labels, etc.), data masking, x- and y-axis tick labeling and positioning, as well as color coding for 2D plots (colormap, colormap sections, etc.).

psy-maps: The psyplot plugin for visualizations on a map

psy-maps builds on top of the psy-simple plugin and extends its functionality for visualizations on a map using the functionalities of the cartopy package (Met Office, 2010 - 2015) (see supplements 4.C for examples). It simplifies as such the automated generation of maps for climate model data through the flexibility of the psyplot framework.

psy-maps currently implements additional *formatoptions* for choosing the projection of the map, selecting the geographic region, drawing the contintens or shaded reliefs of land and ocean, and more. One feature that distinguishes psy-maps from other visualization software, even from pure cartopy, is the ability to visualize unstructured geo-referenced grids on the map. For this purpose, triangles are projected in a pre-processing step to the target projection, prior to the visualization with matplotlib. This drastically increases the performance and makes it possible to visualize even very large data sets. As such, psy-maps visualizes a global scalar field on a hexagonal grid of roughly 4.4 million grid cells ($\approx$ 13 km resolution) in roughly 3.5 minutes. The interactive usage of such a large dataset is however limited by the functionalities of matplotlib to handle such an immense amount of data.

psy-reg: The psyplot plugin for visualizing and calculating regression plots

psy-reg performs regression analysis on 1D variables using the methods of the statsmodels (Seabold and Perktold, 2010) and scipy (Jones et al., 2001; Oliphant, 2007) packages, and visualizes the results with the functionalities of the psy-simple plugin. As such, it implements *formatoptions* for univariate regressions, confidence intervals via bootstrapping, and combined plots of the data density and the fitted model (see also supplements 4.D). The necessity for this package arose from the need to visualize a regression model, compare it (visually) with the original data and to use it afterwards. Other python packages either focus only on the generation of the regressions (such as statsmodels or scipy), or on their visualization (such as seaborn (Waskom et al., 2018)). The psyplot plugin makes it possible to generate the visualization and to access the underlying regression model parameters and uncertainties.

psy-reg has been heavily used for the parameterization of the weather generator in chapter 6 which also gave the initial motivation for the package.

psy-strat: A psyplot plugin for stratigraphic plots

psy-strat (Sommer, 2019a) is the latest plugin for psyplot that has been developed for stratigraphic diagram visualization. It is particularly designed for the straditize software (Sommer et al., 2019, chapter 3) and was motivated by the need for an automated creation of pollen diagrams. One example of such a diagram is provided in the supplementary material, section 4.E.

As the psy-reg and psy-maps plugins, psy-strat uses the functionalities of the psy-simple plugin for a visualization of multiple variables in separate diagrams that share one common vertical axis (usually age or depth)[5]. Additionally, besides the integration that is common for every psyplot plugin (see next section 4.3.3), psy-strat contains additional functionalities for the psyplot GUI. This implementation allows the user to select and reorder the variables (pollen taxa) that are shown in the stratigraphic diagram.

4.3.3 The psyplot Graphical User Interface

Psyplots objective of providing a platform for flexible and convenient data analysis is further approached with the *psyplot-gui* package. This extension to the framwork provides a GUI for simplified access to the plotting features in psyplot.

A strong focus of this interface is, again, the flexibility. psyplot-gui is based on the cross-platform PyQt5 library[6], a very flexible and frequently used package for

[5] See psy-strat.readthedocs.io for an example of psy-strat.

[6] PyQt5 can be accessed via `https://riverbankcomputing.com/software/pyqt/intro`.

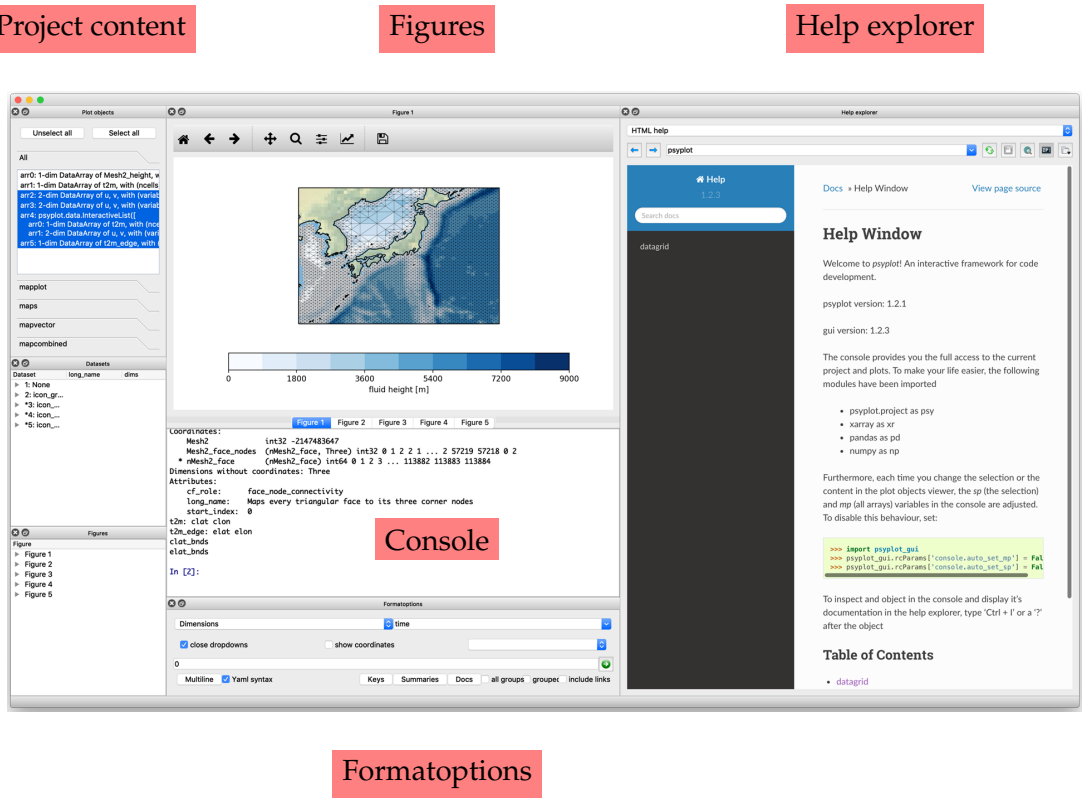


FIGURE 4.2: Screenshot of the psyplot GUI. The left part shows the content of the psyplot project, the upper center the plots, and the right part contains the help explorer. Below the plots, there is also the IPython console for the usage from the command line and a widget to update the *formatoptions* of the current project.

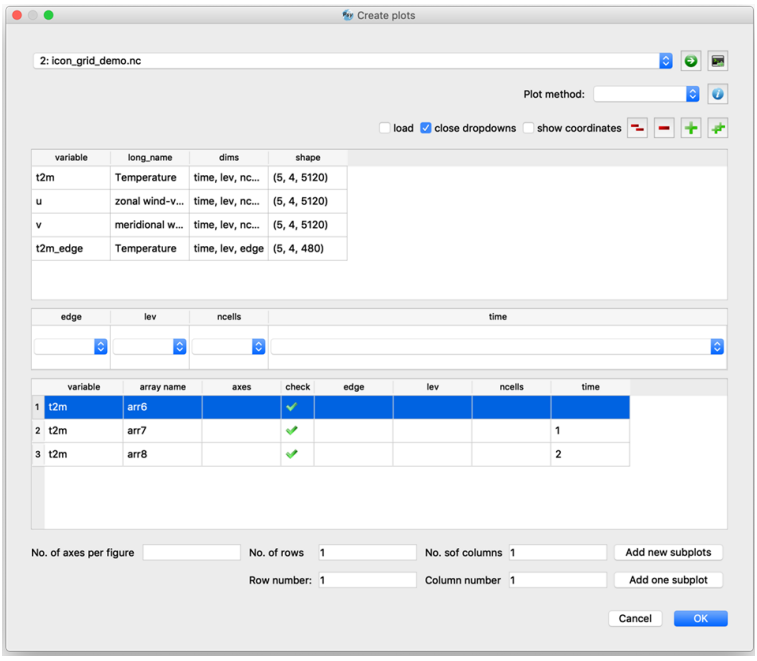


FIGURE 4.3: Plot creation dialog to generate new figures from an xarray dataset.

graphical user interfaces. This enables other software to develop additional features for the package (see psy-strat in the previous section 4.3.2, for instance, or straditize in chapter 3) and to flexibly change the layout of the application. The GUI is complemented with an interactive console to provide a fully integrated python environment for data analysis.

The next paragraphs provide an overview on the various widgets, that are also displayed in figure 4.2 and 4.3.

Console

The central aspects to guarantee flexibility of the application is an in-process IPython console, based on the qtconsole package[7] that provides the possibility to communicate with the psyplot package via the command line and to load any other module or to run any other script or notebook, or even to run commands in different programming languages, such as R (R Core Team, 2019) or Julia (Bezanson et al., 2017). The console is fully integrated both ways into the GUI. The documentation of every python object in the terminal, for instance, can be viewed in the help explorer of the GUI. And vice versa: a change of the current project through the project content widgets, also changes the corresponding python variable in the shell.

Help explorer

As a complement to the console, the GUI contains a help explorer to provide immediate and dynamic access to the documentation of python objects in the console, rendered as an HTML webpage[8]. Furthermore, the help explorer is connected to multiple other widgets of the GUI in order to provide a dynamically generated documentation. The documentation of available *formatoptions* in the psyplot project, for instance, are rendered as HTML upon request, in order to make the various plot methods more accessible. The same principle works for the plot methods that are accessible in the plot creator.

Plot creator

The plot creator (figure 4.3) is the starting point of the GUI into the psyplot framework (at least, if one does not use the console or a script to generate the plots). It loads data from the disk or the in-process console, and essentially provides a wrapper around the psyplot plotting call (see suppl. section 4.A). It additionally displays the documentation of the method and its associated *formatoptions*. This widget creates new plots, that are appended to the psyplot project and are accessible through the console and the project content widgets.

Project content

The psyplot project is the most high-level API element in the psyplot framework (see section 4.3.1) and is displayed in the project content widgets of the GUI. All other elements, such as the *formatoptions* widget or the plot creator, are interfering with the project, and it is accessible as a variable in the console. The project content widget can be used to see the various items in the project, but it is also used to select

[7] https://github.com/jupyter/qtconsole

[8] The help explorer widget has been originally motivated by the *Help* widget of the Scientific PYthon Development EnviRonment, Spyder (https://www.spyder-ide.org/) and uses the sphinx package (Hasecke, 2019) to convert restructured Text into HTML.

the specific items for the so-called *current* sub-project. The latter is dynamically set in the console through the `sp` variable and it is used by the *formatoptions* widget to update the plotting parameters of the selected items.

Formatoptions

As mentioned in section 4.3.1, *formatoptions* are the core elements in psyplot that control the figure aesthetics of the plots and/or perform data manipulations. The generic *formatoptions* widget provides access to these parameters, in order to update them for the selected items in the current project. The formatoption itself (i.e. the python object) can in turn generate a widget that is implemented in the *formatoptions* widget, to make the available options more accessible. The *title* formatoption, for instance, generates a drop-down menu to select variable attributes (e.g. variable name, variable units, etc.) which is then embedded in the *formatoptions* widget. The modifications of the *formatoptions* via this widgets, updates the figures of the selected items.

Figures and plots

The plots generated by the plotting methods are displayed in dedicated widgets inside the GUI and can be dynamically adjusted using the *formatoptions* widget or the console. The underlying library of the current implemented psyplot plugins, matplotlib, implements multiple backends to display the data interactively, or to export them as PDF, PNG, etc. The psyplot GUI has implemented a backend on top of the PyQt5 backend of matplotlib, which embeds the figures in the GUI. psyplot can, however, work with any backend of matplotlib and does not depend on the specific implementation.

4.4 Conclusions

psyplot (Sommer, 2017e) is a new data visualization framework that integrates rich computational and mathematical software into a flexible framework for visualization. It differs from most of the visual analytic software such that it focuses on extensibility in order to flexibly tackle the different types of analysis questions that arise in pioneering research. The design of the high-level API of the framework enables a simple and standardized usage from the command-line, python scripts or jupyter notebooks. A modular plugin framework enables a flexible development of the framework that can potentially go into many different directions. The additional enhancement with a flexible GUI makes it the only visualization framework that can be handled from the conveniently command-line, and via point-click handling. It also allows to build further desktop applications on top of the existing framework.

The plugins of psyplot currently provide visualization methods that range from simple line plots, to density plots, regression analysis and geo-referenced visualization in two dimensions. The software is currently entirely based on the visualization methods of matplotlib (Hunter, 2007), the most established visualization package in the scientific python community. However, the framework itself is agnostic to the underlying visualization method and can, as such, leverage a variety of existing analytical software.

4.5 Outlook

The possibilities for further development of the psyplot framework are numerous, due to its intrinsic generality. The core of the psyplot framework will, in the future, be extended with a standardized algorithm for the generation of animations. Psyplot projects already have the functionality of being saved to a file and reloaded, but they will also be exportable as python scripts for a more flexible reusability and adaptability. The update process within a psyplot project (currently every item in the project is updated in parallel) also has potential for improvement by using a single-threaded scheduler approach that better reflects if one formatoption depends on the *formatoptions* of another plotter.

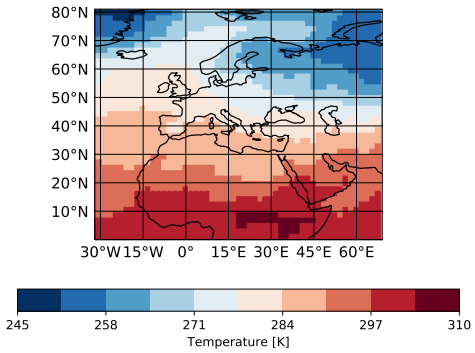
The GUI has especially high potential for further development, as it still lacks widgets to quickly and intuitively modify the visual appearance of the plots. The only possibility inside the GUI (besides the console) is to use the *formatoptions* widget whose main focus however is on flexibility, rather than usability and has, as such, limited possibilities for adaptation to specific use cases.

Another focus will be the development of new plot methods inside the psyplot framework. The major aspect will be on the development of 3D visualization methods of geo-referenced data, using recently published software that builds on top of the visualization toolkit VTK (Sullivan and Kaszynski, 2019; Sullivan and Trainor-Guitton, 2019), see Sommer, 2019b. psyplot has the unique potential to generate 3D visualizations conveniently from the command line, a distinguishing feature, compared to other visualization software packages, such as ParaView or Vapor. Further potential enhancements for visualizations can involve standard interactive visual analytic tools, e.g. such that the interactive selection of features in one plot affects the visualization in another plot (so-called brushing and linking).

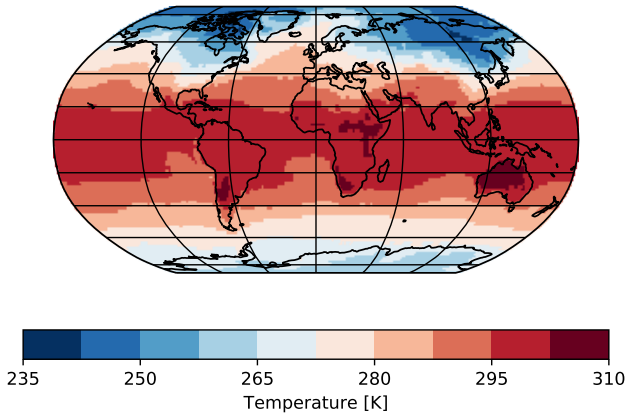
Supplementary material

4.A Example call of a plot method

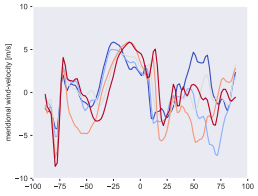
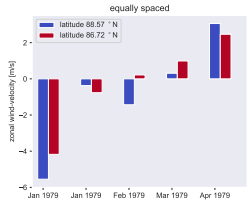
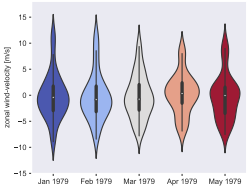
```
# example call for generating a map
import psyplot.project as psy

maps = psy.plot.mapplot(
    'psy-maps-demo.nc',  # input file name, can also be data in memory
    name='t2m',          # variable to plot (can also be multiples
    ### formatoptions
    # colorbar label uses meta attributes of netCDF variable
    clabel='%(long_name)s [%(units)s]',
    # select colormap
    cmap='RdBu_r',
    # focus on a specific lonlatbox given by [lonmin, lonmax, latmin, latmax]
    lonlatbox=['Europe', 'Europe', 0, 'Europe'])

maps.show()
```

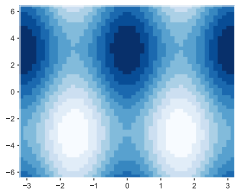
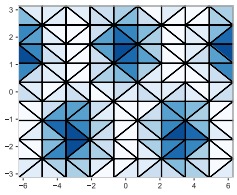
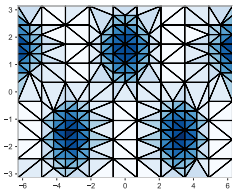


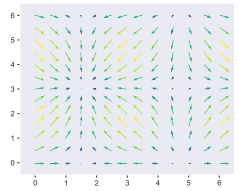
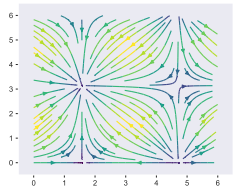
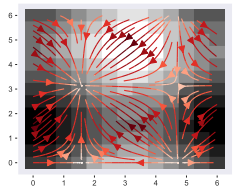

```
# Update the plot, e.g. change projection, plot global

maps.update(projection='robin', lonlatbox=None)
```

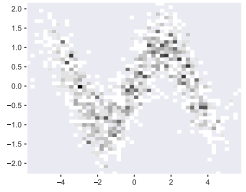
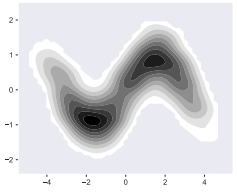
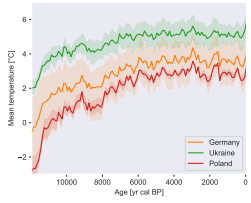


4.B psy-simple plot methods

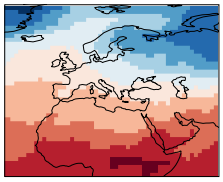
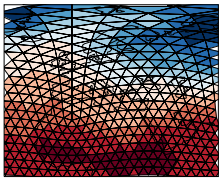
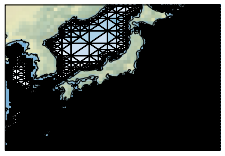
Plot method	lineplot	barplot	violinplot
Example			

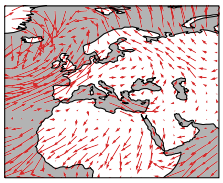
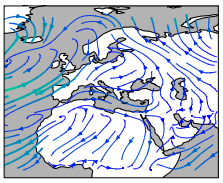
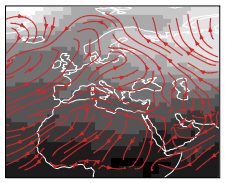
Plot method	plot2d	
Grid type	rectilinear	unstructured
Example		 

Plot method	vector	combined
Example	 	

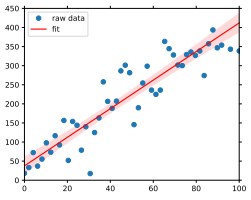
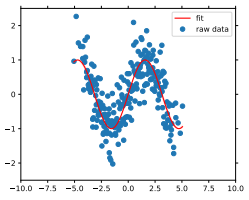
Plot method	density	fldmean
Example	 	

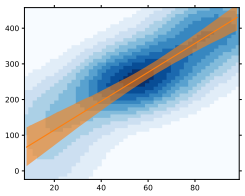
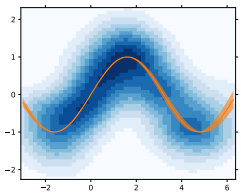
4.C psy-maps plot methods

Plot method	mapplot		
Grid type	rectilinear	unstructured	
Example			

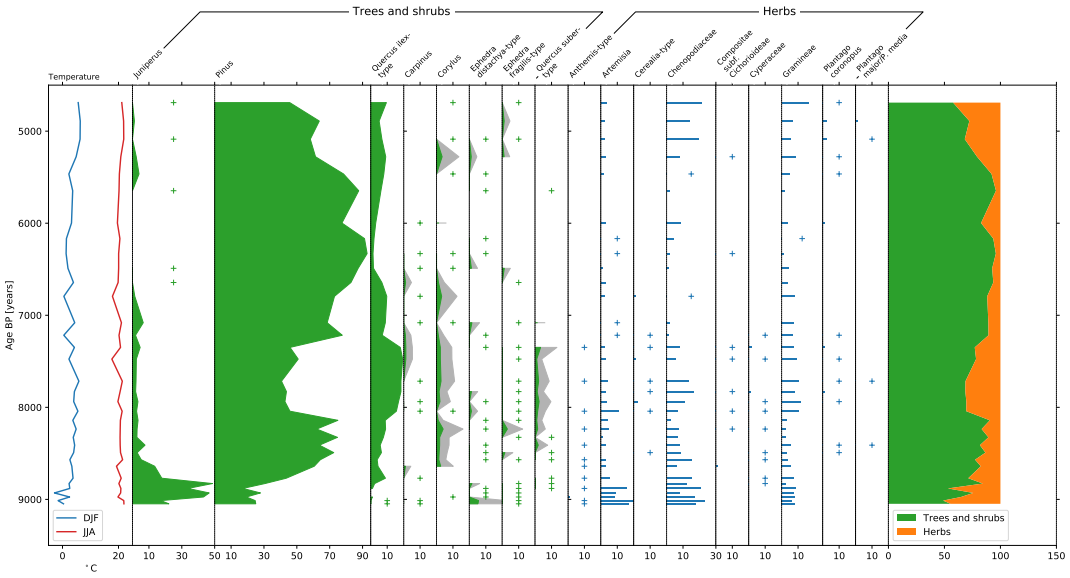
Plot method	mapvector		combined
Example			

4.D psy-reg plot methods

Plot method	linreg	
Example		

Plot method	densityreg	
Example		

4.E psy-strat plot methods



References

Ayachit, Utkarsh (2015). *The paraview guide: a parallel visualization application*. Kitware, Inc.

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B. Shah (Jan. 2017). "Julia: A Fresh Approach to Numerical Computing". In: *SIAM Review* 59.1, pp. 65–98. DOI: 10.1137/141000671. eprint: https://doi.org/10.1137/141000671. URL: https://doi.org/10.1137/141000671.

Böttinger, Michael and Niklas Röber (2019). "Visualization in Climate Modelling". In: *International Climate Protection*. Ed. by Michael Palocz-Andresen, Dòra Szalay, Andràs Gosztom, Làszlò Sìpos, and Tìmea Taligàs. Cham: Springer International Publishing, pp. 313–321. ISBN: 978-3-030-03816-8. DOI: 10.1007/978-3-030-03816-8_39. URL: https://doi.org/10.1007/978-3-030-03816-8_39.

Brown, Stewart A., Mike Folk, Gregory Goucher, Russ Rew, and Paul F. Dubois (1993). "Software for Portable Scientific Data Management". In: *Computers in Physics* 7.3, p. 304. DOI: 10.1063/1.4823180.

Clyne, John, Pablo Mininni, Alan Norton, and Mark Rast (Aug. 2007). "Interactive desktop analysis of high resolution simulations: application to turbulent plume dynamics and current sheet formation". In: *New Journal of Physics* 9.8, pp. 301–301. DOI: 10.1088/1367-2630/9/8/301.

Dask Development Team (2016). *Dask: Library for dynamic task scheduling*. URL: https://dask.org.

Dawson, Andrew (Apr. 2016). "eofs: A Library for EOF Analysis of Meteorological, Oceanographic, and Climate Data". In: *Journal of Open Research Software* 4. DOI: 10.5334/jors.122.

Eyring, Veronika, Mattia Righi, Axel Lauer, Martin Evaldsson, Sabrina Wenzel, Colin Jones, Alessandro Anav, Oliver Andrews, Irene Cionni, Edouard L. Davin, Clara Deser, Carsten Ehbrecht, Pierre Friedlingstein, Peter Gleckler, Klaus-Dirk Gottschaldt, Stefan Hagemann, Martin Juckes, Stephan Kindermann, John Krasting, Dominik Kunert, Richard Levine, Alexander Loew, Jarmo Mäkelä, Gill Martin,

Erik Mason, Adam S. Phillips, Simon Read, Catherine Rio, Romain Roehrig, Daniel Senftleben, Andreas Sterl, Lambertus H. van Ulft, Jeremy Walton, Shiyu Wang, and Keith D. Williams (May 2016). "ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP". In: *Geoscientific Model Development* 9.5, pp. 1747–1802. DOI: 10.5194/gmd-9-1747-2016. URL: https://www.geosci-model-dev.net/9/1747/2016/.

Hasecke, Jan Ulrich (2019). *Software-Dokumentation mit Sphinx: Zweite überarbeitete Auflage (Sphinx 2.0) (German Edition)*. Independently published. ISBN: 1793008779. URL: https://www.amazon.com/Software-Dokumentation-mit-Sphinx-%C3%BCberarbeitete-Auflage/dp/1793008779?SubscriptionId=AKIAIOBINVZYXZQZ2U3A%5C&tag=chimbori05-20%5C&linkCode=xm2%5C&camp=2025%5C&creative=165953%5C&creativeASIN=1793008779.

Hoyer, S. and J. Hamman (2017). "xarray: N-D labeled arrays and datasets in Python". In: *Journal of Open Research Software* 5.1. DOI: 10.5334/jors.148. URL: http://doi.org/10.5334/jors.148.

Hunter, J. D. (May 2007). "Matplotlib: A 2D Graphics Environment". In: *Computing in Science Engineering* 9.3, pp. 90–95. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.55.

Jagers, Bert, David Stuebe, Tom Gross, Chris Barker, Brian Zelenke, Rich Signell, Bob Oehmke, Alex Crosby, Karen Schuchardt, David Ham, Brian Blanton, Cristina Forbes, Charles Seaton, Dave Forrest, Bill Howe, Geoff Cowles, and Phil Elson (Aug. 2018). *UGRID Conventions (v1.0)*. URL: http://ugrid-conventions.github.io/ugrid-conventions (visited on 09/05/2019).

Jones, Eric, Travis Oliphant, Pearu Peterson, et al. (2001). *SciPy: Open source scientific tools for Python*. [Online; accessed 2017-02-18]. URL: http://www.scipy.org/.

Keim, Daniel, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon (2008). "Visual Analytics: Definition, Process, and Challenges". In: *Information Visualization: Human-Centered Issues and Perspectives*. Ed. by Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, Chris North, Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 154–175. ISBN: 978-3-540-70956-5. DOI: 10.1007/978-3-540-70956-5_7. URL: https://doi.org/10.1007/978-3-540-70956-5_7.

Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing (2016). "Jupyter Notebooks – a publishing format for reproducible computational workflows". In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press, pp. 87–90. DOI: 10.3233/978-1-61499-649-1-87.

McKinney, Wes (2010). "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman, pp. 51–56.

Met Office (2010 - 2015). *Cartopy: a cartographic python library with a matplotlib interface*. Exeter, Devon. URL: http://scitools.org.uk/cartopy.

Müller, Ralf (2019). *cdo-bindings: Ruby/Python bindings for CDO*. URL: https://github.com/Try2Code/cdo-bindings (visited on 09/05/2019).

Nocke, T., S. Buschmann, J. F. Donges, N. Marwan, H.-J. Schulz, and C. Tominski (Sept. 2015). "Review: visual analytics of climate networks". In: *Nonlinear Processes in Geophysics* 22.5, pp. 545–570. DOI: 10.5194/npg-22-545-2015. URL: https://www.nonlin-processes-geophys.net/22/545/2015/.

Oliphant, Travis E (2006). *A guide to NumPy*. Vol. 1. Trelgol Publishing USA. URL: http://www.numpy.org/.

— (2007). "Python for Scientific Computing". In: *Computing in Science & Engineering* 9.3, pp. 10–20. DOI: 10.1109/mcse.2007.58.

Perez, Fernando, Brian E. Granger, and John D. Hunter (Mar. 2011). "Python: An Ecosystem for Scientific Computing". In: *Computing in Science & Engineering* 13.2, pp. 13–21. DOI: 10.1109/mcse.2010.119.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rautenhaus, Marc, Michael Böttinger, Stephan Siemen, Robert Hoffman, Robert M. Kirby, Mahsa Mirzargar, Niklas Rober, and Rudiger Westermann (Dec. 2018). "Visualization in Meteorology—A Survey of Techniques and Tools for Data Analysis Tasks". In: *IEEE Transactions on Visualization and Computer Graphics* 24.12, pp. 3268–3296. DOI: 10.1109/tvcg.2017.2779501.

Rew, R. and G. Davis (July 1990). "NetCDF: an interface for scientific data access". In: *IEEE Computer Graphics and Applications* 10.4, pp. 76–82. DOI: 10.1109/38.56302.

Schulzweida, Uwe (Feb. 2019). *CDO User Guide*. DOI: 10.5281/zenodo.2558193. URL: https://doi.org/10.5281/zenodo.2558193.

Seabold, Skipper and Josef Perktold (2010). *Statsmodels: Econometric and statistical modeling with python*.

Sommer, Philipp S. (Aug. 2017a). "Chilipp/psy-maps: v1.0.0: First official and stable release". In: DOI: 10.5281/zenodo.845712. URL: https://doi.org/10.5281/zenodo.845712.

— (Aug. 2017b). "Chilipp/psy-reg: v1.0.0: First official and stable release". In: DOI: 10.5281/zenodo.845717. URL: https://doi.org/10.5281/zenodo.845717.

— (Aug. 2017c). "Chilipp/psy-simple: v1.0.0: First official and stable release". In: DOI: 10.5281/zenodo.845705. URL: https://doi.org/10.5281/zenodo.845705.

— (Aug. 2017d). "Chilipp/psyplot-gui: v1.0.1: Graphical User Interface for the psyplot package". In: DOI: 10.5281/zenodo.845726. URL: https://doi.org/10.5281/zenodo.845726.

— (Aug. 2017e). "The psyplot interactive visualization framework". In: *The Journal of Open Source Software* 2.16. DOI: 10.21105/joss.00363. URL: https://doi.org/10.21105/joss.00363.

— (Aug. 2019a). *psy-strat v0.1.0: A Python package for creating stratigraphic diagrams*. DOI: 10.5281/zenodo.3381753. URL: https://doi.org/10.5281/zenodo.3381753.

— (2019b). *psy-vtk: A VTK plugin for psyplot*. Last accessed: 2019-05-27. URL: https://github.com/Chilipp/psy-vtk (visited on 05/27/2019).

Sommer, Philipp S., Dilan Rech, Manuel Chevalier, and Basil A. S. Davis (Feb. 2019). "straditize: Digitizing stratigraphic diagrams". In: *Journal of Open Source Software* 4.34, p. 1216. DOI: 10.21105/joss.01216. URL: https://doi.org/10.21105/joss.01216.

Sullivan, C. and Alexander Kaszynski (May 2019). "PyVista: 3D plotting and mesh analysis through a streamlined interface for the Visualization Toolkit (VTK)". In: *Journal of Open Source Software* 4.37, p. 1450. DOI: 10.21105/joss.01450.

Sullivan, C. and Whitney Trainor-Guitton (June 2019). "PVGeo: an open-source Python package for geoscientific visualization in VTK and ParaView". In: *Journal of Open Source Software* 4.38, p. 1451. DOI: 10.21105/joss.01451.

Uchida, Takaya, Ariel Rokem, Tom Nicholas, Ryan Abernathey, Jake Vanderplas, Yaroslav Halchenko, Andreas Mayer, Greg Wilson, Kai Pak, and Aurélien Ponte (2019). *xgcm/xrft v0.2.0*. DOI: 10.5281/zenodo.1402635.

Waskom, Michael, Olga Botvinnik, Drew O'Kane, Paul Hobson, Joel Ostblom, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian De Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Thomas Brunner, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, , Brian, and Adel Qalieh (2018). *mwaskom/seaborn: v0.9.0 (July 2018)*. DOI: 10.5281/zenodo.592845.

Wijk, J. J. van (Oct. 2005). "The Value of Visualization". In: *VIS 05. IEEE Visualization, 2005.* IEEE, pp. 79–86. DOI: 10.1109/visual.2005.1532781.

Wong, Pak Chung, Han-Wei Shen, Ruby Leung, Samson Hagos, Teng-Yok Lee, Xin Tong, and Kewei Lu (Nov. 2014). "Visual analytics of large-scale climate model data". In: *2014 IEEE 4th Symposium on Large Data Analysis and Visualization (LDAV)*. IEEE, pp. 85–92. DOI: 10.1109/ldav.2014.7013208.

Part II

Computational Models

Chapter 5

pyleogrid

A Probabilistic Approach for Gridding Paleo Climate Data

5.1 Introduction

Paleo-climate reconstructions are most often undertaken on a site by site basis to provide a record of climate change at a specific place through time. The integration of data obtained from multiple sites however provides the basis for investigating spatially explicit reconstructions of climate through time. This spatio-temporal perspective can provide powerful insights into the climate system that are not easily discernible from the typical 1-dimensional approach associated with single site records. Spatially explicit data allows us to see how spatial patterns in climate variables change through time, providing a way of identifying the underlying causes of climate change. It also allows us to match the spatial scale of Earth-system models, which are based on grid-boxes that often reflect climatic changes at a very different spatial resolution than that experienced at the scale of a single site.

Here, we describe a computationally efficient methodology for integrating multiple paleo-climate records from different sites into a single spatio-temporal record that simultaneously takes into account the associated uncertainties. This method also involves projecting the data onto a uniform spatial grid and regular time-step. This approach is different from the conventional approach to gridding, often called *pseudo-gridding*, in which records that fall within a grid box are simply combined in some way to represent the grid box value (e.g. Bartlein et al., 2010; Marcott et al., 2013; Marsicek et al., 2018; Waelbroeck et al., 2009). Similarly, samples from the records within a grid box are also combined or binned into time-windows to create a regular time-step. Our method instead does not aggregate the reconstruction spatially or temporally but rather interpolates the data to a user-defined 3-dimensional spatial grid and regular time-step. This approach has been used in previous studies (Davis et al., 2003; Mauri et al., 2014, 2015), but here we integrate chronological and reconstruction uncertainties into the gridding process, allowing us to propagate these uncertainties through time and space onto our grid network.

Gridding has many advantages over other simple mapping approaches, including *pseudo-gridding*. It allows us to calculate more accurately area-averages and energy balances, as well as to make direct comparisons with Earth system models at a comparable grid box size and regular time-step. Gridding also allows a changing paleo-climate site/sample network through time to be stabilized, making it easier to compare one time period with another. We are also able to create a more complete record of climate in time and space, using the entire sampling network to infer climate in places and at times where we may not have a site/sample.

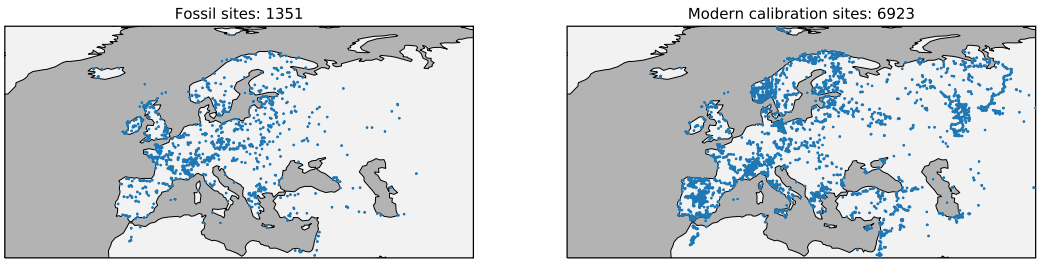


FIGURE 5.1: Site locations of the (left) fossil and (right) modern pollen database.

Our new method applies a probability approach to data integration, whereby the full uncertainty of each sample is considered in the gridding process, rather than just the sample mean used in previous methods (Mauri et al., 2015). From this, we can calculate the uncertainties associated with the temporal and spatial distances between our reconstruction samples/sites and the points on the grid network. We do this through an ensemble bootstrapping approach in which we repeatedly grid the data, each time using a different set of samples that are randomly selected to reflect the reconstruction and chronological uncertainties of the site network. In this study, we use pollen-data, which provides the most accessible and spatially distributed paleo-climate data available for the late-Quaternary period.

The strength of this new method is that it treats all of the paleo-climate samples and sites as a single integrated paleo-climate record from which it is possible to extract a single regionally coherent climatic reconstruction, complete with uncertainties. Pollen-based reconstructions often have high sample to sample uncertainties, and the vegetation at individual sites can be influenced by non-climatic factors such as soils, disease, fire, migration lag and human impact. Our method allows us to fully utilize the large quantity of pollen-data that is available, to extract the regional background climate signal from what may be locally quite noisy data. It also has a significant advantage over other methods that integrate records using Bayesian approaches (e.g. Holmström et al., 2015), in that it is much more computationally efficient, making it possible to undertake analysis at continental scales with hundreds and thousands of sites and samples.

5.2 Data

The ensemble based gridding method is adapted to paleo-climates. In this study, we describe the method using a large set of western Eurasian fossil pollen assemblages that have been transformed to summer (June, July and August) (JJA) temperatures. We focus on pollen data because it is the spatially most widely available proxy during the Holocene, but it is important to mention that the reconstruction method is agnostic to the climate proxy, because it does not explicitly use the pollen assemblages but rather alters the standard climate reconstruction method under the aspect of its methodological uncertainties. As such, the following sections describe the fossil and modern pollen database for this use case (section 5.2.1) and the associated uncertainties of the temperature reconstruction method (section 5.2.3) and the dating of the fossil pollen samples (section 5.2.4).

5.2.1 Pollen database

The source data for this study is a subset of the latest development version of the POLNET database, a northern hemispheric, extra-tropical collection of pollen assemblages (Davis and Kaplan, 2017; Sommer et al., 2019). The purpose of this database is to generate the source for large-scale climate reconstruction during the Holocene (past 12'000 years) that can be used for model-data comparisons. The subset that we use in this study to describe and develop the gridding method contains fossil and modern pollen assemblages of western Eurasia, a region that has already been under investigation in the previous study by Mauri et al., 2015.

The fossil database contains raw pollen counts with in total about 1350 datasets that consists of 80500 fossil samples. The majority of the fossil pollen data (left part of figure 5.1) comes from the European Pollen Database (EPD) (Fyfe et al., 2009, ca. 94%) and other publicly available databases. The presented dataset extends the database used by Mauri et al., 2015 especially with a few sites towards the eastern part of the map.

The modern calibration dataset (6900 sames, see right map in figure 5.1) is mainly based on the version 2 of the Eurasian Modern Pollen Database (EMPD) (Davis et al., 2013, ca. 87%, see also chapter 2) and core tops of EPD (10%) that were younger than 250 years cal BP.

5.2.2 Sample site: Tigalmamine

We chose the pollen record of Tigalmamine in Morocco (32.9N, 5.34W, 1626m) to evaluate our method. The site was first studied by Lamb and Kaars, 1995, and the pollen data was downloaded from the European Pollen Database. The chronology and choice of control points used here is that from Giesecke et al (Giesecke et al., 2013). The site is well dated with 11 radiocarbon dates, and spans the entire Holocene with 110 samples. The data has been used for a previously published pollen reconstruction based on the modern analogue method (Cheddadi et al., 1998), although this study used a calibration dataset that included modern pollen samples from Morocco that have subsequently been found to have geolocation errors (Davis et al., 2013). None of these problematic modern samples have been used in our analysis.

The site (red cross in figure 5.2) is located on the southern edge of our study region in an area with a montane Mediterranean vegetation and climate. The Mediterranean has traditionally been considered to be a particularly challenging environment for pollen-based reconstructions because of the effects of long term human impact, and the interplay of precipitation and temperature on vegetation distribution. The fossil pollen record of Tigalmamine shows a mainly forested montane Mediterranean assemblage throughout the Holocene, dominated by evergreen oak, but with an important transition between the early Holocene and late Holocene marked by a change from deciduous Oak to Cedrus (see the pollen diagram in Cheddadi et al., 1998). The occurrence of Cedrus represents an interesting challenge for any pollen climate transfer function, since this particular taxa is limited in its distribution (and in our calibration dataset) to Morocco and the Lebanon region, while all of the other taxa in the assemblage are widely distributed across the Mediterranean. The strong presence of evergreen Oak also makes the site interesting, because although this taxa is mainly associated with the Mediterranean region, its distribution extends all the way up the west coast of France to Brittany.

Figure 5.2 illustrates these challenges for the particular site. The climate analogues (see next section 5.2.3) span a large summer temperature regime of about 10

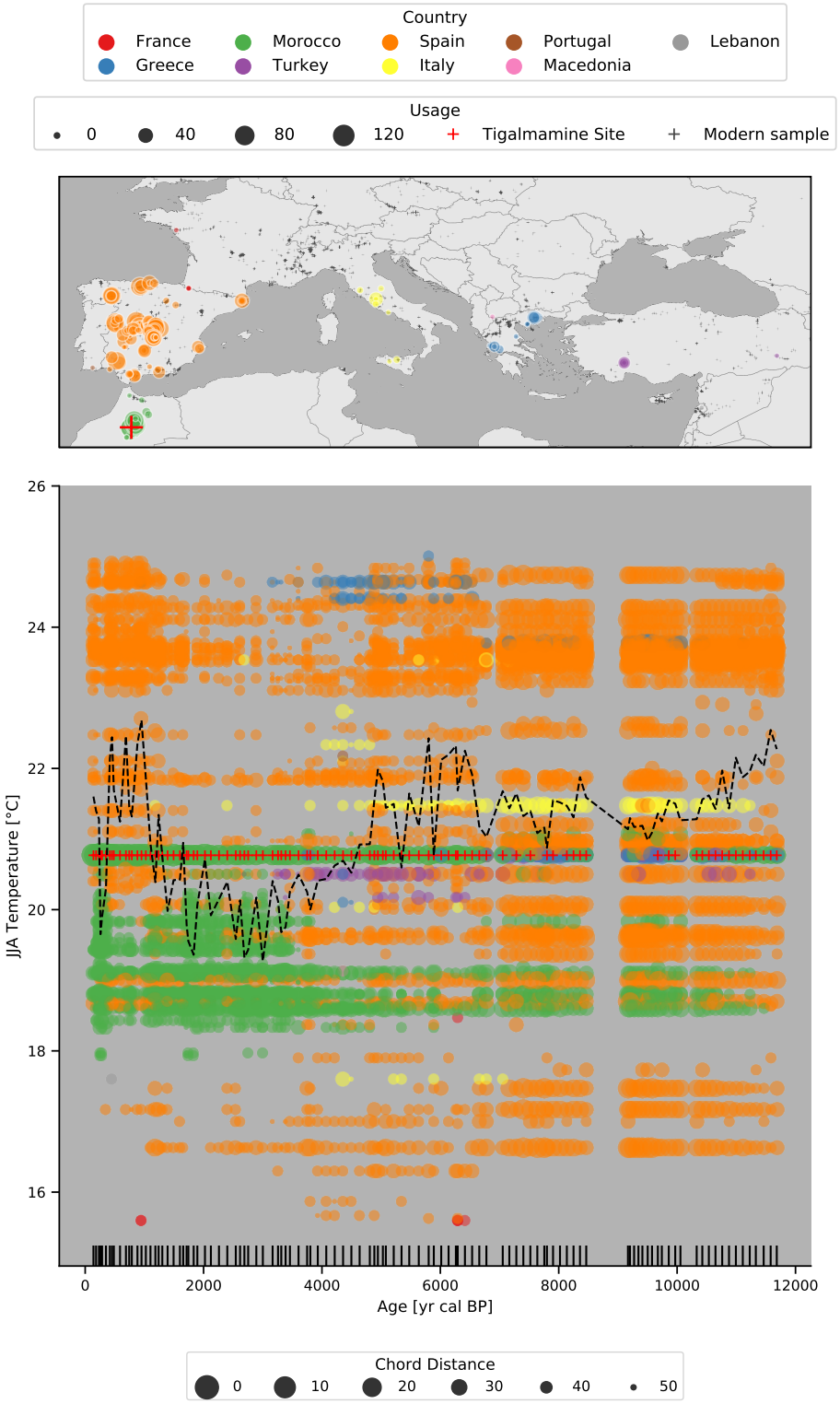


FIGURE 5.2: Climate analogues of the Tigalmamine site (red cross). Every circle corresponds to one modern analogue that was one of the fifties closest analogues in at least one sample within the Tigalmamine dataset. The color-coding of each circle is based its corresponding country (see legend at the top). The marker size in the top plot depends on the usage of the sample as modern analogue. The larger the marker, the more samples in the Tigalmamine dataset use it as modern analogue. Tiny crosses in the map show the locations of the rest of the modern calibration data. The lower plot shows the summer temperature for the analogue (y-axis) at the age of the Tigalmamine sample (x-axis). The marker size in this plot corresponds to the chord distance between modern and fossil pollen assemblage. I.e., the larger the dot, the closer (and more important) the analogue. The dashed line shows the weighted average of all the climate analogues per sample. The age of each Tigalmamine sample is shown with the vertical lines at the bottom of the plot. Red crosses in the lower plot show the Tigalmamine core top sample that has been used as an analogue in 58 out of the 110 samples.

degrees, from 15 to 25 °C. The upper temperature range is dominated by analogue climates from Spain (orange) which in general shows the highest number analogue matches. The early Holocene (12k to 8k BP) is dominated by modern samples from Spain, with a wider and more uniform temperature regime, when compared to the later periods. During the transition in the mid-Holocene (8k to 4k BP), analogues from across the Mediterranean Sea play a more important role, in particular from Greece, Italy and Turkey. The lower temperature regime is then dominated by Moroccan samples (green) that are of particular importance during the late Holocene (4k BP to present) due to the above mentioned presence of Cedrus.

The weighted average of the analogues (black dashed line in figure 5.2) is in general about one to two degrees lower than the one in Cheddadi et al., 1998 (very likely due to the above-mentioned erroneous calibration data they used). The trends are however similar: Higher temperatures in early Holocene (the *spanish analogues* dominate) with a drop around 6k BP (Moroccan climates). Our weighted average however also shows a clear increase during the past 2000 years, again driven by spanish analogues.

The climatic and geographic space that is covered by the analogues is further discussed in section 5.3.2.

5.2.3 Site-based holocene temperature estimates

A standard approach for site-based climate reconstruction from fossil pollen assemblages is the modern analogue technique (MAT) (also called *k*-nearest neighbors). This technique estimates the climate of the fossil sample as the (weighted) climate average of the most similar modern samples (i.e. the closest modern analogues). It has the major advantage that it requires little parameterization efforts and can be applied over a large spatial area that covers many different climate regimes (Mauri et al., 2015).

For this purpose, we follow the standard approach and assign JJA temperature values for each modern calibration sample (figure 5.1), taken from the corresponding grid cell in the WorldClim dataset, version 2 at 30 seconds (Fick and Hijmans, 2017).

Every pollen assemblage is then transformed from raw counts to percentages, based on the total sum of terrestrial pollen counts per sample that we deem useful for the reconstruction. This excludes low samples with low counts and taxa with low occurrences. We use squared-chord distance from the R package *rioja* (Juggins, 2017) as a measure of similarity. For a given transformed fossil pollen assemblage $\{f_t\}$ and a modern pollen assemblage $\{m_t\}$, where $t = \{1, \ldots, N\}$ denotes one of the N individual taxa in the assemblages, this distance measure is defined as

$$d = \sum_{t=1,\ldots,N} \left(\sqrt{f_t} - \sqrt{m_t} \right)^2$$

This distance is calculated between every modern and every fossil sample in the entire database (section 5.2.1). The standard, non-probabilistic setup would now compute the climate of the fossil sample as the mean climate of the k closest analogues (e.g. $k = 6$), eventually weighted by their corresponding distance d. There are many variations of this technique (see for example Birks et al., 2010, including various measures of similarity, choices about k, the maximum allowed distance d between modern and fossil assemblage, subsampling of the calibration dataset to avoid spatial autocorrelation (Guiot and Vernal, 2011; Telford and Birks, 2005, 2009), and by grouping pollen taxa into so-called plant-functional types (PFTs) (Davis et

al., 2003; Mauri et al., 2015, e.g.). They all, however, have in common that the categorical, multi-modal distribution of the climate of the modern analogues is simplified into a unimodal distribution, represented by the mean of the analogue climates. Therefore, in our ensemble approach, we do not take the mean but sample the climate of the analogues directly. This is further discussed in the methods section 5.3.2 and 5.4.1.

5.2.4 Age uncertainties

In addition to the methodological uncertainties of the climate reconstruction method (previous section 5.2.3), we provide a framework to handle dating uncertainties. During the gridding step (see next section 5.3.3), every sample is weighted by the age difference to the target reconstruction age. The previous studies by Davis et al., 2003 and Mauri et al., 2015 do not take this uncertainty, that can be as high as multiple centuries, into account although they influence the gridded temperature reconstruction.

The reason is a systematic problem of pollen samples that we overcome here with the recent developments in the pollen community. In palynology, each sample in a sediment core is is dated using a so-called age-depth model, a function that maps each depth of the sediment core to an age. This function is based on a few chronological control points where the age has been determined instrumentally (for lake sediments in the Northern Hemisphere, these are commonly radiocarbon (^{14}C dates) and interpolates/extrapolates to the depths of the sample locations. Various methodologies exist to define these age-depth models, ranging from simple linear interpolation methods (Bennett, 1994) to the more recently developed bayesian techniques of the Bchron (Haslett and Parnell, 2008) and BACON (Blaauw and Christen, 2011) models.

The early approaches have been proven to provide unreliable uncertainty estimates (Telford et al., 2004) and there has been no standardized way to report the uncertainties, if they are reported at all. For this reason we (and previous studies) cannot rely on the age uncertainties reported in the pollen database. An alternative approach is to recalculate the chronology for every dataset in the database (see Goring, 2019, for instance), but this also requires parameterization for reliable uncertainties and goes beyond the scope of this study.

Instead, we follow an approach that is based on two aspects: age uncertainties are higher for older samples, and samples that are farther away from the radiocarbon dates (i.e. chronological control points). Additionally, samples behave differently if the sample is surrounded by two chronological points (i.e. the sampe age is interpolated) or not (sample age is extrapolated). To quantify these relationship, we perform a study based on all datasets (ca. 30'000 samples) from the Neotoma paleoecology database (Williams et al., 2018) that have age-depth models estimated with BACON, a model that has been proven to provide more reliable age uncertainty estimates (Trachsel and Telford, 2016). For the sake of implementation (the age sampling in section 5.3.1 assumes a normal distribution), we apply several assumptions and approximations to the Neotoma samples, in particular:

1. We assume that every dataset with a BACON chronology in Neotoma keeps the defaults and reports the limits of the 95% confidence interval (CI)

2. We keep only the maximal distance of the CI limits from the reported age (i.e. we assume a symmetric distribution)

3. We assume that the distribution is normal (i.e. the 95% CI corresponds to the 2σ interval, where σ^2 denotes the scale parameter) and a division in half of the maximal distance (see previous assumption) gives the standard deviation σ (which is what we call the age uncertainty)

The resulting data is illustrated in figure 5.3. The grayscale density plots in the background shows the high dispersal of the data and the number of samples decreases strongly with higher distance to the control point or older samples (red lines). Nonetheless, the mean of the data (blue lines) reveals the increasing nature of both relationships, as mentioned before.

Figure 5.3 also shows two models that have been fitted to the data. The first one is a standard simple univariate linear model $y = a + b \cdot x$ (orange line). This model simulates the increasing trend of both variables although it does not capture the non-linear relationship between age and age-uncertainty. A reason for this non-linearity arises from the time-dependency of the radiocarbon calibration curve and its associated errors. This non-linear behavior gives the motivation to use a constrained linear Generalized Additive Model (GAM), a smooth semi-parametric model of the form

$$\mathbb{E}[y|X] = \beta_0 + f_1(X_1)$$

in the univariate case, or

$$\mathbb{E}[y|X] = \beta_0 + f_1(X_1) + f_2(X_2)$$

in the bivariate case. The feature functions f_1 and f_2 are based on penalized B splines with a constraint for monotonic increasing, the expected value $\mathbb{E}[y|X]$ is based on a normal distribution. The GAM model has been fitted with the *pyGAM* software package (Servén et al., 2018). This model enables to better simulate the non-linear features as can be seen with the green lines in figure 5.3.

These results approve the initial hypotheses and justify the choice of a bivariate GAM for predicting age uncertainties based on the distance to the chronological control point, and the age of the sample. The two models, together with a bivariate simple linear regression model, and again for interpolated and extrapolated samples, are shown in the central column of figure 5.4. Both model classes (simple linear and GAM) are able to reproduce the general shape of the observed data, although the GAM better resolves the non-linear relationship between the three variables.

The final uncertainties, predicted for the data set presented in the previous section 5.2.1, are shown in the supplementary figure 5.17.

5.3 Method

With the intrinsic methodological uncertainties of climate and dating in mind, we present a new ensemble-based approach on gridding the reconstructions from the individual sites. Each ensemble member is generated with randomized sample ages and climate, derived from the corresponding uncertainty measures (see previous sections 5.2.3 and 5.2.4), with additional constraints arising from the integrity of the individual dataset (sediment core). We explain these in more details in sections 5.3.1 and 5.3.2. The final gridding step for each ensemble member is based on a modified setup of Mauri et al., 2015, but can also be extended with other interpolation algorithms, as described in section 5.3.3). We implemented the method as the python package *pyleogrid* that efficiently scales to large datasets and ensemble sizes, and shortly describe it in section 5.3.4.

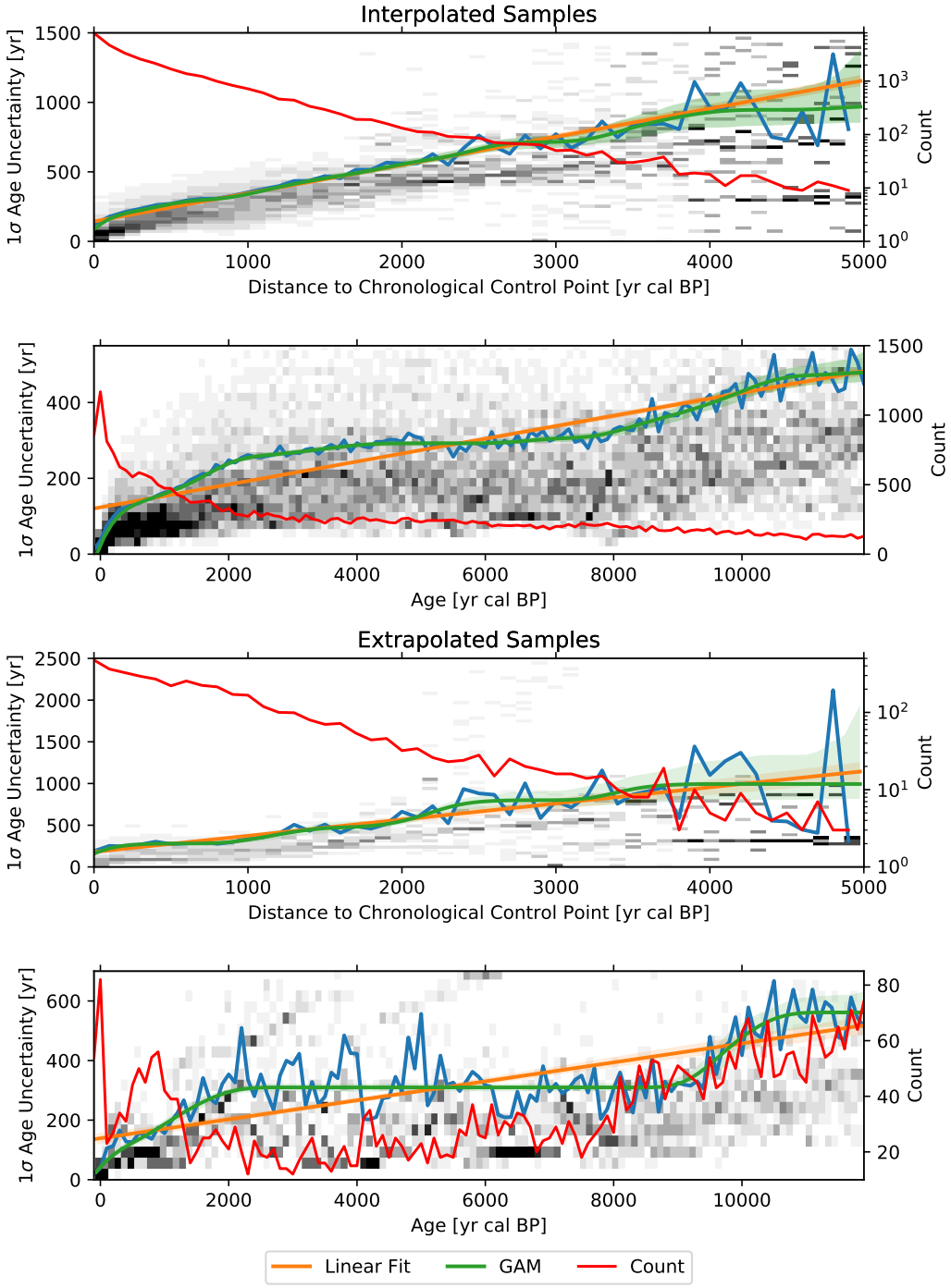


FIGURE 5.3: Univariate regression plots of (first and third) distance to chronological points, and (second and fourth) age to the one sigma dating uncertainty of the sample. The upper two plots contain only interpolated samples (i.e. samples that lie between two chronological control points), the lower extrapolated samples. Blue lines show the mean age uncertainty for the given distance (age). Orange and green lines show the linear and GAM fits of distance (age) to age uncertainty, and red lines show the number of samples for a given distance (age). The grayscale plot in the background shows a two-dimensional histogram (density plot) to illustrate the underlying data of the fits. For the purpose of a better visualization, each vertical bin of this histogram has been normalized to one. Means, counts and histogram are all based on 100 year bins in distance (age). The fits are estimated based on the unbinned data, the source data are all Neotoma datasets with BACON-based age-depth models. Note the logarithmic scale of the right count axis on the first and third plot.

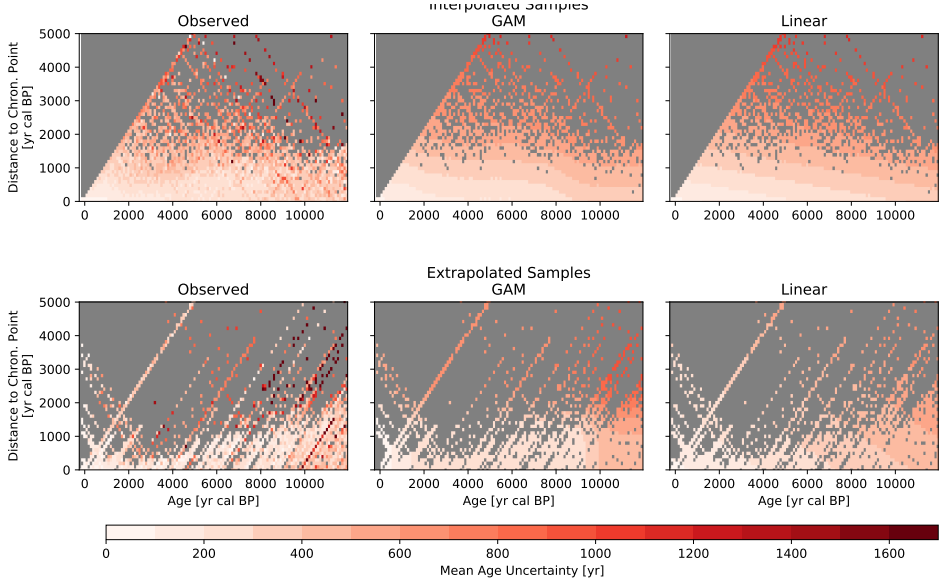


FIGURE 5.4: Bivariate models of age uncertainty. Shown are the mean 1σ age uncertainties of ca. 30'000 samples from the Neotoma database sites with BACON-based age-depth models. Each sample has an age uncertainty that depends on the age of the sample (x-axis) and the distance to the closest chronological control point (y-axis). For the purpose of visualization, we grouped the samples into categories of 100 years in x- (age) and 100 years in y- (control point distance) direction, and calculated the average age uncertainty of the groups. These averages are shown with the color coding in the plots, and the gray area represents the space without any observation. The top row shows interpolated samples (i.e. samples that lie between two chronological control points), the bottom row extrapolated samples. Plots in the left column show the observed mean age uncertainties, central and right columns show the means of the predicted age uncertainties from the bivariate linear GAMs and bivariate linear regression models respectively.

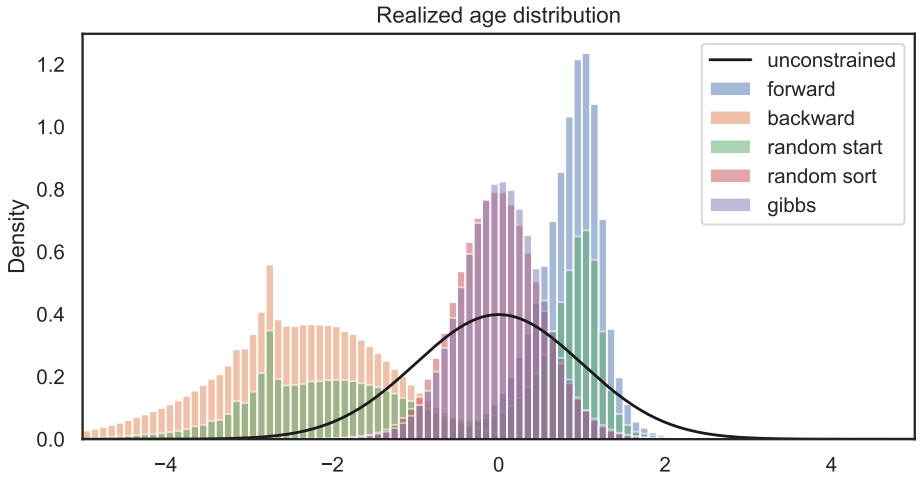


FIGURE 5.5: Histograms of age sampling methods for the site in section 5.2.2 with an ensemble size of 10'000. Every sampled age has been centered at the reported age of the corresponding sample and scaled by its age uncertainty. The black line shows the unconstrained distribution (a standard normal with a standard deviation of 1), the other histograms show the realized distributions for each of the age sampling methods (section 5.3.1). Note that *random sort* and *Gibbs* histograms highly overlap.

5.3.1 Constrained age sampling

Every dataset has an intrinsic monotonicity constraint that the sample deeper down the core has an older age. An inversion of this constraint is very rare and is usually visible in the stratigraphy of the core, such that affected samples are ruled-out before. As such, a classic unconstrained sampling of ages[1] using a normal distribution centered at reported sample age and a scale corresponding to the estimated age uncertainty (section 5.2.4) violates this constraint. Samples are inverted in such a case when their uncertainty intervals overlap and as such the individual ensemble member would not maintain the integrity of the individual core. We illustrate an example for such a core in section 5.4.1.

pyleogrid therefore implements different variants of this constraint with the Gibbs sampling being the one that is finally used.

The intuitive approach

The most intuitive approach is to randomly draw a sample age and constrain the age of the neighboring sample with it. This can be done in a *forward* manner, such that every older sample has to be older than the previous younger sample, or in a backward manner, i.e. the younger sample has to be younger than the neighboring older sample. We will show in the paragraphs below that this method is biased, nevertheless we mention it here because of the intuitivity of the approach and because the reason for the failure is non-trivial.

As such, we demonstrate three different algorithms:

[1] We call it the unconstrained distribution for convenience, but keeping in mind that every sampled age has to be older than -70 yr cal BP.

forward Starting with an unconstrained age distribution for the youngest sample in the core, every consecutive sample has to be older than the previous (i.e. the method works forward in age, but backward in time

backward Starting with an unconstrained age distribution for the oldest sample in the core, every consecutive sample has to be younger than the previous (i.e. the method works backward in age, but forward in time)

random start Starting with an unconstrained age distribution of a random sample in the core, we apply the *backward* algorithm for younger and *forward* algorithm for younger samples.

As such, *forward* and *backward* algorithms always start with an unconstrained age distribution of the youngest (oldest) sample for every ensemble member. Within the *random start* algorithm, every sample gets the chance to start with an unconstrained age distribution, because the starting point is random for every ensemble member. The constrained age distributions for the consecutive samples are implemented as truncated normal distributions.

The resulting age distributions from the three algorithms are shown in figure 5.5, together with another method, that is described later in this section. The figure shows the sampled age distributions by the various above-mentioned sampling methods for the site described in section 5.2.2. To make these age distributions comparable, we transformed them to a standard normal distribution (visualized as the unconstrained distribution in figure 5.5) prior to visualization, by subtracting the reported age and dividing by the estimated age uncertainty of the corresponding sample. It is obvious from this figure that all of the above-mentioned algorithms produce an artificial bias to the age distribution. The *forward* approach pushes the samples to the upper tail of the distribution, the *backward* approach pushes everything to the lower tail. The *random start* method produces a bimodal distribution with peaks at the upper and lower tail.

This is also shown with three exemplary samples from the site in the supplementary figure 5.18. The forward method works well for the young sample but pushes all older samples to the upper tail of their distribution, The backward method does the opposite and the random sort method creates a bimodal distribution for the sample in the center of the core, and backward behaves like the forward (backward) algorithm at the older (younger) part of the core.

We explain this initially unexpected results with the overlapping age uncertainties in the core. The site that we describe here has 110 samples. As such, the probability that one sample draws a random age at the lower or upper tail of the distribution is very high. Now, most of the dating uncertainty intervals overlap and this forces all the consecutive samples to the tail of their age distributions. Another problem, that is not shown here, arises from the differing sizes of the age uncertainties which highly depends on the distance to the chronological control point (see section 5.2.4). This can also lead to unsatisfiable requirements, if one sample is close to a control point (and as such has a lower age uncertainty) and the previous sample has been pushed far outside of the 95% confidence interval.

The random sorting approach

These strong biases of the intuitive approach led to another method, that we also show in red in figure 5.5 and supplementary figure 5.18, the *random sort* method. This method consists of two steps: in the first step we draw random age for each

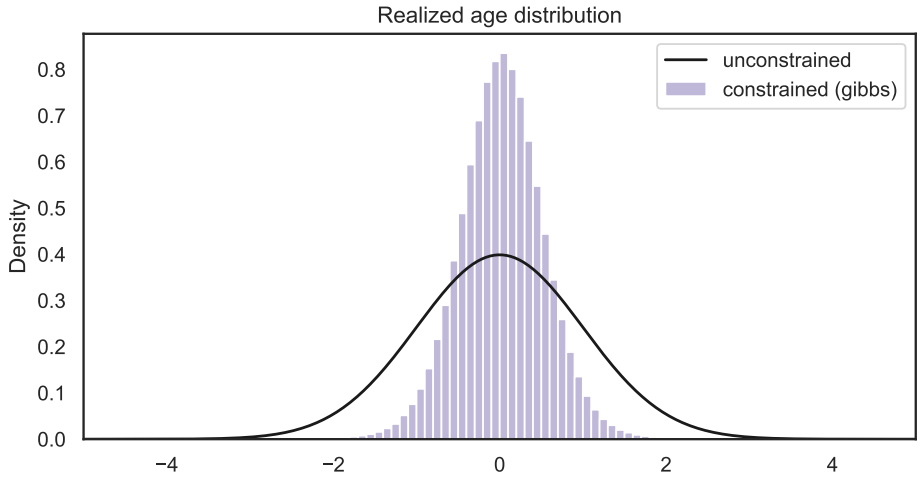


FIGURE 5.6: Realized age distribution for the entire dataset (section 5.2.1) with the *random* method (section 5.3.1). The individual sample distributions have been centered and scaled as in figure 5.5.

sample based on its unconstrained distribution[1]. In the second step, we order these random ages while maintaining the order of samples in each dataset. As such, we assign an age to each sample that is not necessarily drawn from its own distribution, but rather from the one of a neighboring sample. When samples overlap, this then truncates the tails of realized distribution and effectively decreases the reported age uncertainty, as can be seen in the figures 5.5, 5.18. This approach is mathematically difficult to justify because it violates the common methodology that each sample has a unique confidence interval that it needs to explore. Therefore the method might introduce some hidden biases in the sampled distributions that are difficult to quantify. Nevertheless, the algorithm is very fast and much closer to the desired joint distribution, than the previous *intuitive* approach. But in order to avoid any biases and to guarantee a mathematically correct result, we chose to implement a Gibbs sampling algorithm to sample from the desired constrained distribution.

The Gibbs sampling approach

Algorithm 1 Accept/Reject algorithm. $\mathcal{N}(\mu, \sigma)$ denotes the normal distribution with location parameter μ and shape parameter σ.

1: Set $i = 0$
2: Set $\boldsymbol{\mu}$ as vector of the reported ages in *dataset*
3: Set $\boldsymbol{\sigma}$ as vector of estimated age uncertainties
4: Set $\mathbf{a}$ (the target age vector) to be of length $\boldsymbol{\mu}$
5: **while** $i < 1$ **or not** *is_monotonic*$(\mathbf{a})$ **do**
6: $\quad \mathbf{a} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$
7: $\quad$ Set $i = i + 1$
8: **end while**

The biases of the above-mentioned algorithms led to the development of a Markov chain Monte Carlo (MCMC) sampling algorithm. An accept/reject algorithm, which draws a set of random ages for all unconstrained sample distributions in a core at

once and accepts the draw if the monotonicity condition, is satisfied and rejects the sample if the monotonicity condition was initially explored. For one realization of the ages **a** in a given dataset, this is described with the pseudo-code in algorithm 1. This standard approach however did not find a monotonic solution within ten million iterations for a high-resolution site such as it has been used in the previous section.

Therefore we decided to implement a Gibbs sampler, an algorithm that is commonly used in Bayesian inference to obtain a sequence of samples from conditional probability distributions, which generate samples from a multivariate joint distribution when this distribution is unknown and/or cannot be sampled directly. In our case, this distribution if the distribution of all sample ages in one dataset, where each sample age is conditioned by its younger and older neighbor. Let $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots \mu_N)$ be the reported ages of the N pollen samples in one individual dataset with estimated age uncertainties $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \dots \sigma_N)$. The reported ages fulfill the monotonicity constrain, i.e. $\mu_j \leq \mu_k$ for all j, k with $1 \leq j \leq k \leq N$. The objective of our sampling approach is to generate M random realizations of $\boldsymbol{\mu}$, denoted by $\mathbf{X}^{(m)} = \left(X_1^{(m)}, X_2^{(m)}, \dots X_N^{(m)}\right)$ with $m = 1, \dots, M$, that all fulfill the monotonicity constrain. In other words, the realizations $\mathbf{X}^{(m)}$ are constrained to fulfill

$$X_j^{(m)} \leq X_k^{(m)}, \text{ for all } j, k \text{ with } 1 \leq j \leq k \leq N \text{ and } 1 \leq m \leq M. \tag{5.1}$$

We set the intial value to the reported ages ($\mathbf{X}^{(1)} = \boldsymbol{\mu}$) where we know that the constrain is fullfilled. For the following realizations $\mathbf{X}^{(m+1)}$ with $1 < m \leq M$ we sample each component $X_j^{(m)}$ with $1 \leq j \leq N$ conditioned by its previous sample $X_{j-1}^{(m)}$ and, most importantly, conditioned by the next sample, but from the previous realization, i.e. $X_{j+1}^{(m-1)}$. As such, we define the sampled age of $X_j^{(m)}$ with

$$X_j^m = \mathcal{N}(X_{j-1}^{(m)}; X_{j+1}^{(m-1)}; \mu_j, \sigma_j^2) \tag{5.2}$$

where $\mathcal{N}(a; b; \cdot, \cdot)$ denotes a random variate of the truncated normal distribution with lower limit a and upper limit b. Although this algorithm always starts with the youngest sample in the dataset for every realization, such as the *forward* method, it does not push every sample to the lower tail of the distribution because every sampled age is conditioned by the age of the next pollen sample from the previous realization. It is mathematically proven that the combined realizations $\left\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots \mathbf{X}^{(M)}\right\}$ of this algorithm approximates the joint distribution of the sample ages in the dataset under the given constraint, and that each marginal distribution of a the age of a particular pollen sample $1 \leq j \leq N$ is approximated by $\left\{X_j^{(1)}, X_j^{(2)}, \dots, X_j^{(M)}\right\}$.

As it is common for a MCMC algorithm, each realization is correlated with nearby realizations. The first samples are particularly correlated with the initial value $\boldsymbol{\mu}$ and it is therefore common practice to discard the first 1000 realizations, the so-called *burn-in* period.

To avoid an autocorrelation between the successive realizations, we *thin* our set of sampled ages and keep only every tenth realization until we have the desired amount of M realizations. The value of 10 has been shown to be sufficient using an autocorrelation analysis of the different samples in the Tigalmamine record.

As can be seen in figure 5.5 and 5.18, the outcomes are very close to the above mentioned *random sort* approach. However, a look into the realized distribution of

the last sample in figure 5.18 reveals a negative bias of the distribution sampled with the *random sort* approach.

The realized (and standardized) distribution of the entire database presented in section 5.2.1 is finally shown in figure 5.6. The comparison with the unconstrained distribution in this figure highlights the need for a constrained sampling because the latter significantly reduces the width of the distribution.

5.3.2 Temperature sampling

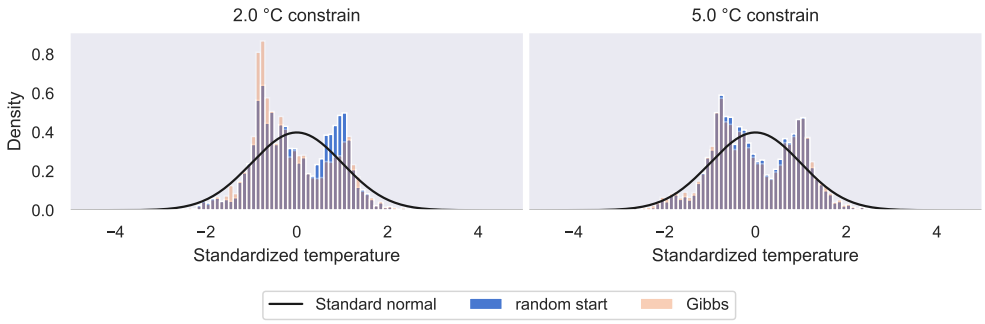


FIGURE 5.7: Histograms of temperature sampling methods for the site in section 5.2.2 with an ensemble size of 10'000 for a climatic constraint of (left) 2 °C, and (right) 5 °C. As in figure 5.5, every sampled temperature has been centered at the weighted average of the corresponding modern analogues and scaled by the corresponding weighted standard deviation. The black line shows the unconstrained distribution (a standard normal with a standard deviation of 1), the other histograms show the realized distributions for *random start* and *Gibbs* temperature sampling method (section 5.3.1).

As already mentioned in 5.2.3, our sampling approach does not use the temperature and uncertainty reported for every single variable. Instead, it samples the underlying distribution. As such, our method can be adapted to multiple site-specific reconstruction methods, such as weighted averaging (WA), weighted-averaging partial least squares (WAPLS) (Birks et al., 1990; Braak and Juggins, 1993) or other approaches (e.g. Birks et al., 2010; Brewer et al., 2007; Juggins, 2013). In this study, we use a modern analogue technique (MAT) approach (see section 5.2.3) and sample the discrete set of climate analogues for each sample. The probability to select an analogue (i.e. its weight) is thereby determined by the chord distance between the fossil and modern pollen assemblages. The closer the assemblages (relative to the other potential analogues), the higher the weight.

This methodology is substantially different from the standard approach, such that it takes the multimodality of the analogues into account, whereas the standard approach (weighted average of the k closest analogues) estimates a unimodal distribution. It additionally better represents the discrete nature of the analogue approach whereas the standard method intrinsically assumes a continuity in the distribution. In fact, only a small part of the available climate space is actually represented by the modern analogues. The 5'500 modern analogues for Tigalmamine, for instance (110 samples with 50 analogues each), are represented by only 240 distinct modern pollen samples with only 131 distinct JJA temperatures and they eventually span a large climate space (see figure 5.2).

The latter gives the motivation for a climatic constraint that ensures the integrity of the individual dataset. It is, for instance, impossible that two samples from the

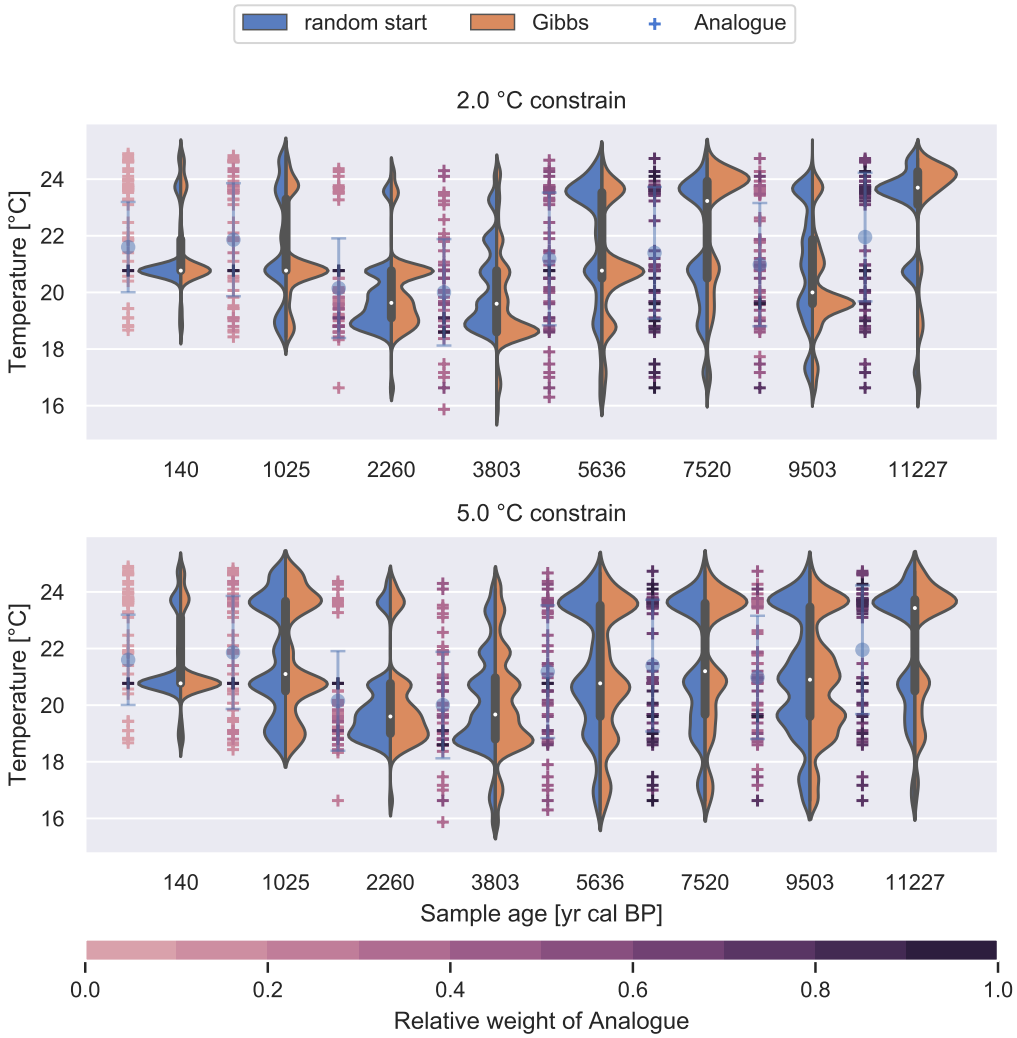


FIGURE 5.8: Violin plots for sampled temperature distributions of some samples in the Tigalmamine record with a climatic threshold of (top) 2 and (bottom) 5 °C. Blue (left) distributions are realized with the *random start* method, ocher (right) areas with Gibbs sampling. The crosses to the left of each violin shows the locations of the climate analogues. Each cross is color coded by its chord distance relative to the chord distance of the closest analogue (i.e. the closest analogue has always a weight of 1).

same dataset but 200 years apart experience a temperature difference of five degrees or more between each other. This is, however, a possible combination, considering the underlying set of analogues (see figure 5.2 for instance). We therefore perform a constrained sampling, as in section 5.3.1, and implement a fixed temperature threshold T. Every sampled analogue in each dataset (i.e. every choice of the discrete distribution for each pollen sample) is constraint to not differ by more than T degree Celsius from its temporally neighboring samples. The exact choice of T is a critical assumption and has a major impact on the realized temperature distribution for each sample. We decided for a very conservative estimate of five degree Celsius, which is only applied if the samples do not differ by more than 1000 years. These choices are further discussed in section 5.4.

In the remainder of this section, we focus on the implementation of this conditional sampling, because of its substantial impact on the realized distribution, as

already shown for the sampled ages in section 5.3.1. We briefly discuss the same approaches as in section 5.3.1 (without the *random sort* algorithm because we do not enforce monotonicity here). The core of the method is the same for all approaches: If a climate analogue differs by more than T degrees from the temperature of the conditioning sample, its probability is set to zero. The choice about the *conditioning sample* is dependent on sampling algorithm. Here, we discuss following methods:

forward The temperature of every older sample must not differ by more than T from its younger sample

backward The temperature of every younger sample must not differ by more than T from its older sample

random start Starting with a random sample in a data, we apply the forward method to older and the backward method to younger samples

Gibbs The choice for each sample is constrained by the younger sample and the older sample from the previous realization of the dataset.

We described these algorithms already in detail in section 5.3.1 and therefore focus only on the comparison of results. The only difference is that now, without the monotonicity constraint, *forward* and *backward* methods give the same result as the *random start* method. Therefore we will only focus on the last two methods.

Figure 5.7 shows the realized distributions for the two methods. As in the corresponding figure for the age sampling (figure 5.5), we subtracted the weighted average of the climate analogues of the corresponding pollen sample by each of the randomly sampled temperature values, and afterwards divided by the weighted standard deviation, in order to make the drawn temperature values at the different ages comparable. Both methods realize a bimodal distribution (a feature that is also visible in the spread of the climate analogues in figure 5.2) and result in the similar distributions when considering a climate constraint of 5 °C. This does not hold for the stronger 2 °C constraint, where the Gibbs method gives more weight to samples below the weighted average. This is caused by the additional constraint of the Gibbs sampling approach where each sample is constrained by the sample of the previous realization. The algorithm always starts with the youngest sample which has a higher probability in the Moroccan regime (green area, figure 5.2). Due to the constraint of the sample in the previous realization, it then is more likely that we stay in this regime. As such, the distribution tends to get more unimodal compared to the other method, where each realization is entirely independent of the other.

This difference between the two methods and the two climatic constraints is further illustrated in figure 5.8, which shows the violin plots for a selection of samples in the Tigalmamine record, separated by method and climatic constraint. As already shown with the standardized temperatures in figure 5.7, the two methods are approximately equal under the 5 °C constraint and significantly reduce the realized temperature regime with a more multimodal distribution under the 2 °C constraint. The Gibbs method tends to sharpen the distribution at the locations of the analogue climates stronger than the *random start* method, which implies that the latter is again prone to (potentially) large and unknown biases.

5.3.3 Gridding

The underlying gridding algorithm is the same as in Mauri et al., 2015, the thin plate spline regression method (*Tps*) of the *fields* R-package (Nychka et al., 2017; R Core

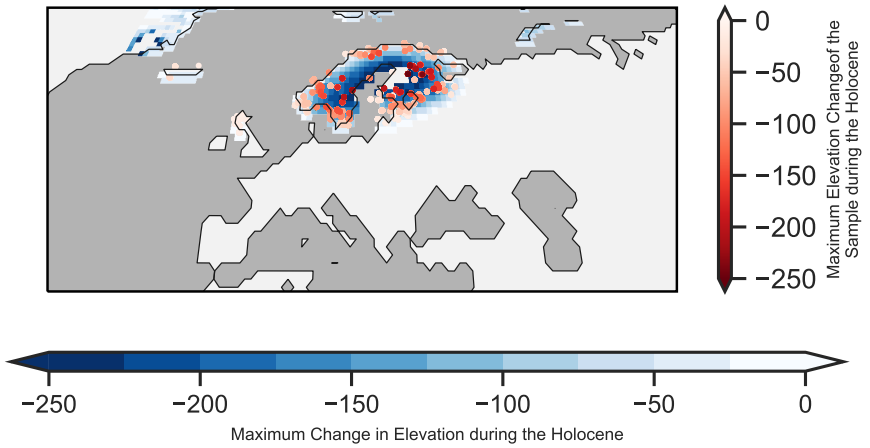


FIGURE 5.9: Maximum change of elevation during the Holocene per grid cell (blue background), based on the data from the ICE-6G model. The dots show the applied (maximal) corrections to the samples in these locations.

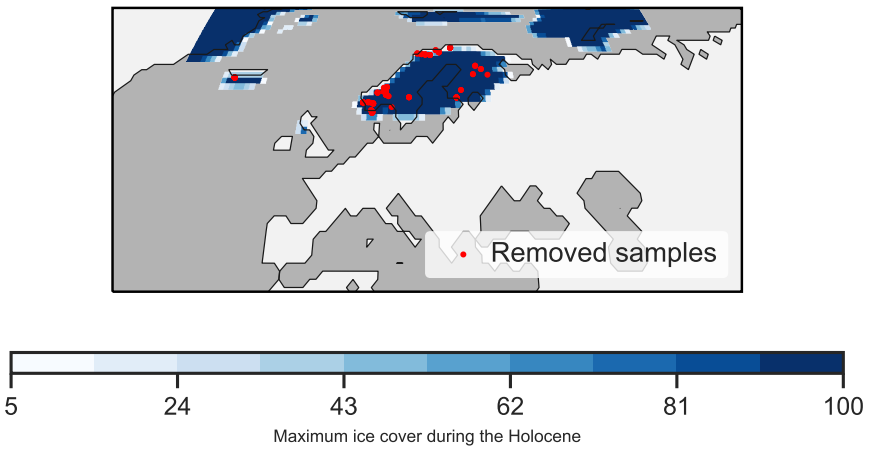


FIGURE 5.10: Locations of samples (red dots) that have been removed from the input data because they are covered by ice (according to the ICE-6G model, blue background).

Team, 2019). This method interpolates the scalar variable (temperature) from the irregularly spaced four dimensional sample data onto a two dimensional surface (defined by latitude, longitude and elevation) in 3D space. The time component is considered by giving a higher weight to samples that are temporally closer to the target time of the interpolation.

The major difference compared to Mauri et al., 2015 is the ensemble approach, which can also be interpreted as a bootstrapping approach. We apply the gridding many times to different realizations of the input data. The major advantage of this ensemble approach is the sophisticated distribution of temperatures per grid cell. This allows a better estimate of the uncertainty of the gridded climate reconstruction, compared to Mauri et al., 2015. But it should be noted that this ensemble approach is structurally independent of the underlying regression algorithm. Hence, it can potentially also be extended to other gridding methods.

Another difference between our study and Mauri et al., 2015 is the type of climatic variable. Mauri et al., 2015 calculated interpolated anomalies, whereas we interpolate the absolute climate climate variable as it is derived from the pollen-climate reconstruction method. The advantage is that we can directly interpolate to the elevation at a given timestep, whereas Mauri et al., 2015 applied an a posteriori isostatic correction.

For the reconstruction presented in section 5.4, we use the data of the ICE-6G-C model (Argus et al., 2014; Peltier et al., 2015), that is also used in the PMIP4 experiments (Ivanovic et al., 2016; Kageyama et al., 2018; Otto-Bliesner et al., 2017). To account for the change in elevation, *pyleogrid* also implements a method to correct the elevation of the samples based on the elevation difference in the given input raster for the different time steps. The results of this correction can be seen in figure 5.9. In addition to this elevation correction, we removed samples from the input data where the ICE-6G model reports an ice coverage of more than 50% in the grid cell (figure 5.10). Affected samples are all during the early Holocene in northern Europe.

5.3.4 Implementation

The ensemble method presented in this thesis is available as the python package *pyleogrid* from github.com/Chilipp/pyleogrid . The documentation of the package is available at pyleogrid.readthedocs.io. This module also contains the models for predicting age uncertainties (section 5.2.4), that are based on the *pyGAM* software (Servén et al., 2018). The github repository and the documentation contains the notebooks that are used to run the analysis of this study.

The sampling methods of *pyleogrid* that generate entirely independent realizations of the input data (forward, backward, random start and random sort methods) are built using the functionalities of the numerical numpy and scipy packages (Jones et al., 2001; Oliphant, 2006) to efficiently generate thousands of constrained realizations of the input data. The computationally more expensive Gibbs sampling algorithms (section 5.3.1 and 5.3.2) were every realization is constrained by the previous realizations, is implemented in Cython (Behnel et al., 2011), an optimising static compiler for Python, and uses the corresponding Cython Application programming interfaces (APIs) of numpy and scipy.

As already mentioned in section 5.3.3, *pyleogrid* uses the *Tps* method of the *fields* package (Nychka et al., 2017) and as such interfaces into an R environment for the gridding.

pyleogrid is designed to scale to large amounts of input data (e.g. for a global reconstruction or hemispheric reconstruction) on a local computer or a large parallelized cluster, using the xarray and dask packages for parallel computing and out-of-core computation (Dask Development Team, 2016; Hoyer and Hamman, 2017).

5.4 Results

In this section we briefly present the results of the final temperature reconstruction, both for the site-based reconstruction at Tigalmamine (section 5.4.1), and the western Eurasian gridded temperature record for selected time slices (section 5.4.2). The purpose of this section is present the results of the ensemble approach with a special focus on the choices that have been made in the methods section, particularly the number of analogues and the climatic constraint. We focus on the ensemble mean, but it is important to keep in mind that our method approximates the joint spatio-temporal distribution of the climate.

5.4.1 Site-based realized climate reconstruction: a use-case

The final reconstruction of the Tigalmamine site for an ensemble size of 10'000 realizations is displayed in figure 5.11. It shows a density plot of the realized temperature and age regime, as well as the weighted average of the *standard approach*. For comparison we also show the mean of the marginal temperature distribution record from our method (we call this the *realized mean* from now). The plot shows how our method realizes the site within the ensemble. It shows the *Moroccan* regime during the late Holocene (lower left dark green area between 0 and 4'000 years cal BP, see also figure 5.2) and the *Spanish* regime during the early to mid Holocene (upper right dark green area in figure 5.11). The figure also displays the differences between the two temperature constraints. As already mentioned in section 5.3.2, the 2 °C constraint results in a more multimodal distribution, especially towards the end of the record. The the realized mean is therefore significantly different from the weighted average of the standard approach.

Figure 5.12 shows the same figure but for a varying number of analogues. The trends of the three images behave the same and show all a decrease in temperature during the early holocene and a small plateau during the last millenium. Also the temperature regimes conform with a the above-mentioned dominance of the *Moroccan regime* in the late Holocene the *Spanish regime* during the early Holocene. With a lower number of analogues, however, (top plot) the temperature changes in the means are much smaller over a short period of time. The distribution for this scenario (green 2D histgram in the plot) shows a strong multimodality with a few temperature values that have a very high probability. The scenarios with higher analogues (20 or 50) result in a much smoother mean because the underlying distribution covers more of the available age-temperature space.

5.4.2 Gridded summer temperature

Determinstic vs. ensemble approach

The gridded reconstruction for the entire dataset is shown in figure 5.13 for 3k, 6k, 9k and 12k BP. The figure shows the temperature anomaly to the modern time (0k BP), both for the ensemble method (10'000 realizations, 5 °C climate constraint) and the deterministic case, i.e. without climate and temperature sampling. For the deterministic case, we followed the approach by Mauri et al., 2015 and mask cells that are

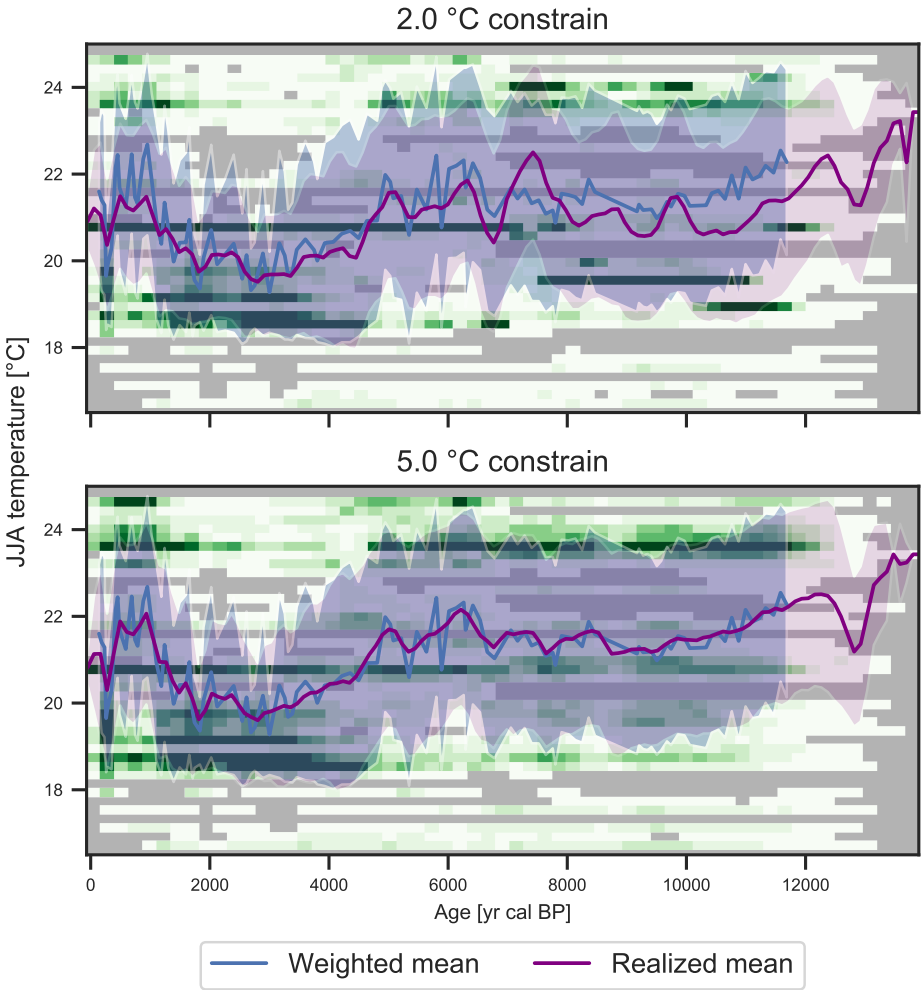


FIGURE 5.11: Realized summer temperature reconstruction of Tigalmamine from the Gibbs sampling with a temperature constraint of (top) 2 °C and (bottom) 5 °C. The background shows a density plot of the sampled age-temperature pairs within the ensemble of 10'000 realizations. The purple line is the mean of all sampled temperatures within 100-year bins, i.e. it represents the mean of the marginal temperature distribution at a given age. The blue curve is the weighted average of the standard MAT approach (black dashed line in figure 5.2). The shaded areas correspond to the standard deviations of the corresponding line.

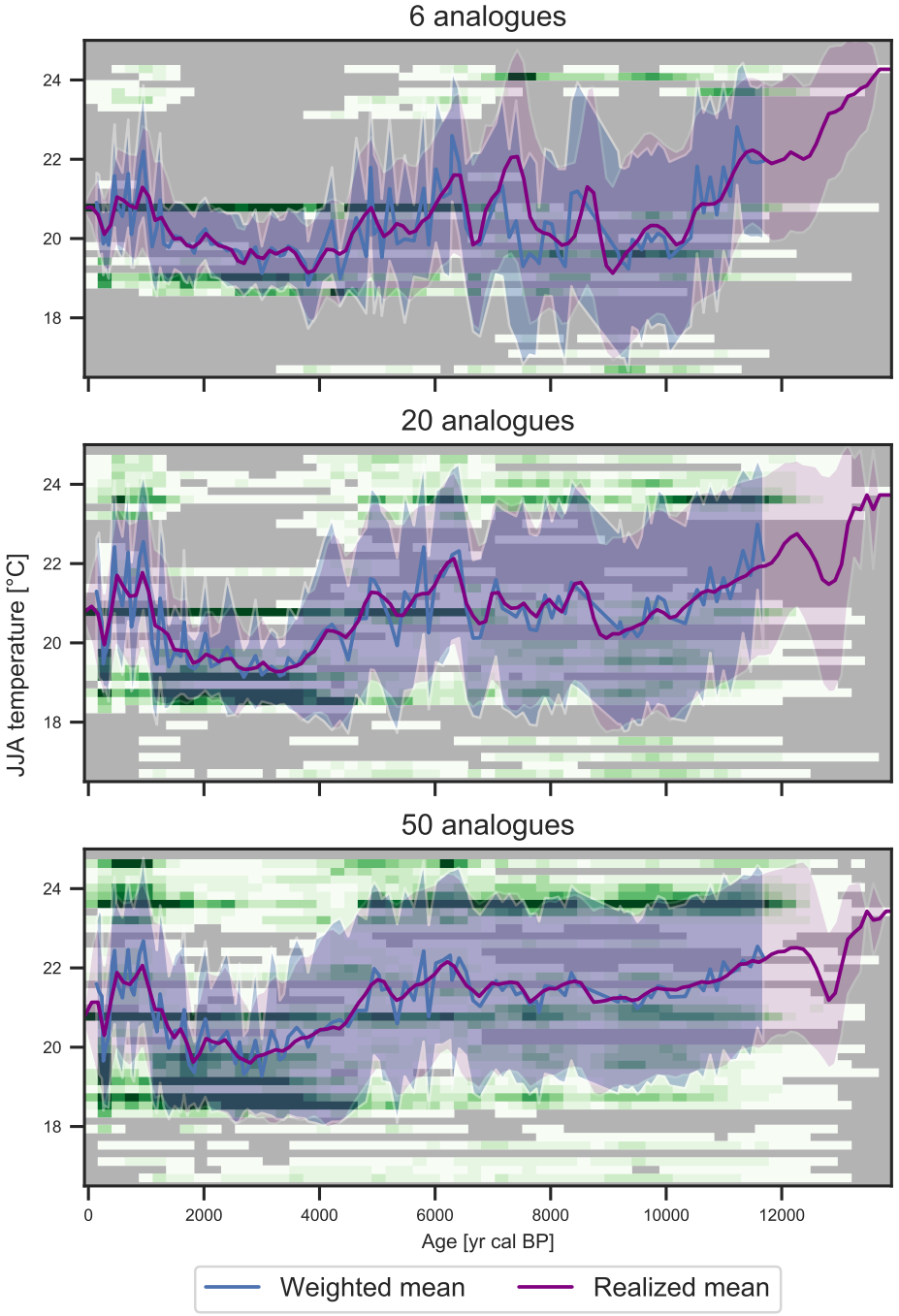


FIGURE 5.12: Realized summer temperature reconstruction of Tigalmamine from the Gibbs sampling with 5 °C constraint for (top) 6 analogues, (middle) 20 analogues and (bottom) the default 50 analogues. The weighted means (blue curves) have been calculated using only the 6, 20 or 50 analogues. See figure 5.11 for a description of the elements in the plot.

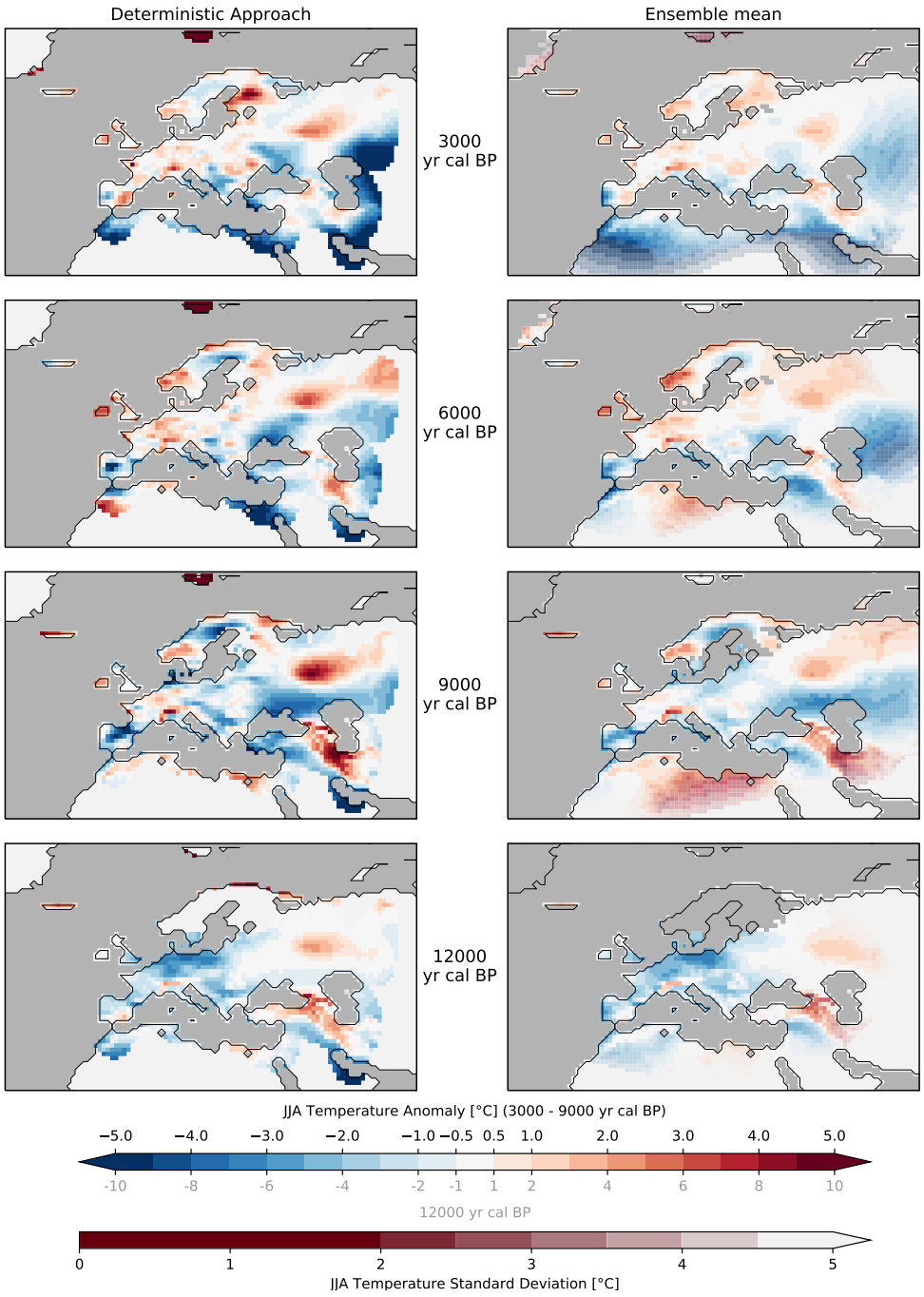


FIGURE 5.13: Deterministic approach vs. ensemble mean (10'000 realizations) at 3k, 6k, 9k and 12k BP (note the different color coding for 12k BP). Each map shows the anomaly with respect to the gridded reconstruction at 0k BP. The deterministic approach (left column) is the input data with temperatures as weighted averages of the 50 closest analogues and without age sampling. The gridded reconstruction has been masked when more than 500km away from the closest sample. The ensemble mean (right column) is overlayed by the corresponding ensemble standard deviation of the anomaly (suppl. figure 5.19). This overlay is transparent for standard deviations smaller than 2.5 °C and afterwards gets more opaque until it reaches 5 °C. Grid cells in the maps that were covered by more than 50 percent with ice, according to the ICE-6G data, have been masked.

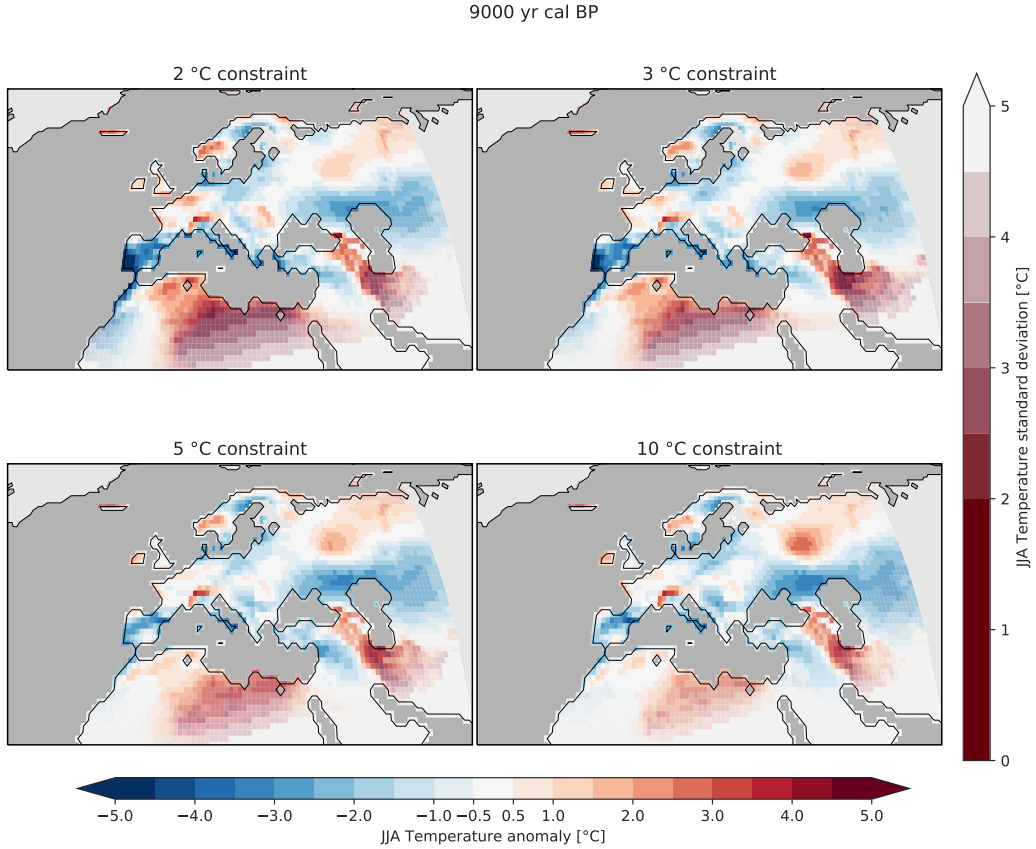


FIGURE 5.14: Ensemble mean summer temperature anomaly at 9k BP (1'000 realizations) for different temperature constraints. The standard deviation is visualized on top of the mean with an increasing opaqueness (same as in figure 5.13).

further away than 500km from any of the pollen sites. For the ensemble approach on the other hand we used the ensemble standard deviation as a gray overlay that is getting more opaque for higher standard deviations.

Both methods show similar patterns and span the same temperature range at the onset of the Holocene. The temperature ranges in the other maps, however, are different with the deterministic approach being more extreme in certain locations, particularly in Spain, eastern Russia and the western Mediterranean region. The method predicts absolute temperature anomalies of more than four degrees, whereas the anomalies of the ensemble approach commonly range between -2.5 and 2.5 °C are more coherent. The ensemble standard deviation of the anomaly (suppl. figure 5.19) in Europe ranges between 0.5 and 1 °C for the mid- to late Holocene timesteps at 6k and 3k BP, and between 1 and 2 °C at 12k and 9k BP. The standard deviation is particularly high at the map boundaries of the 12k reconstruction, very likely due to the smaller availability of fossil samples in this period.

Temperature sampling parameters

Figure 5.14 shows the influence of the choice for the climatic constraint (see section 5.3.2) on the gridded reconstruction for a selected timestep at 9k BP. The comparison of the 5 °C and the very relaxed 10 °C constraint shows only minor differences. Stronger constraints (2 °C and 3 °C) result in a strengthening of the anomaly, particularly in Spain and towards the southern and south-eastern borders of the map.

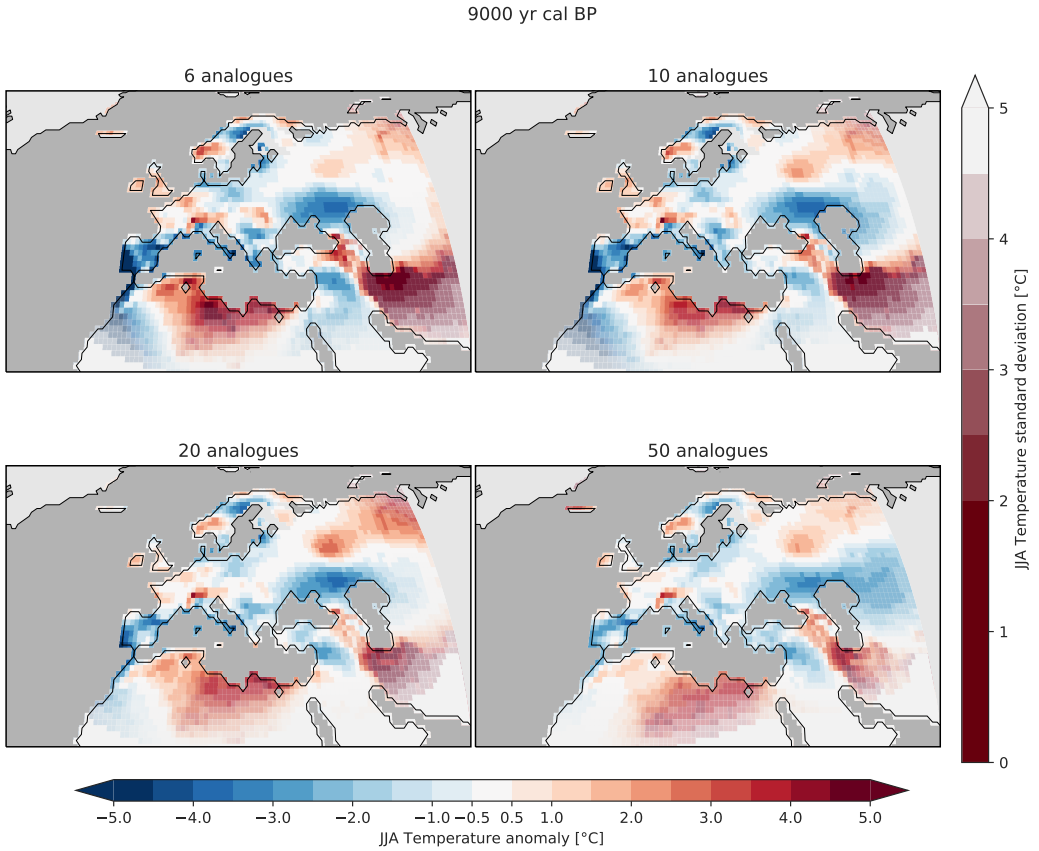


FIGURE 5.15: Ensemble mean summer temperature anomaly at 9k BP for numbers of analogues. The standard deviation is visualized on top of the mean with an increasing opaqueness (same as in figure 5.13).

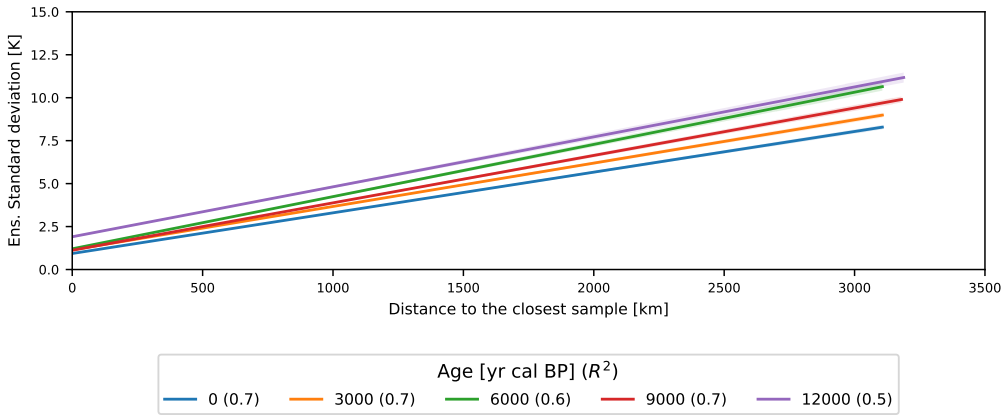


FIGURE 5.16: Relationship between distance to proxy sites and the ensemble standard deviation. Each line corresponds to a linear fit where the distance to the closest site in the database for a given grid cell is regressed against the ensemble standard deviation in this cell (figure 5.19). The different lines represent the different time steps from figure 5.13, the value in brackets after each legend label shows the coefficient of determination (R^2) for the given regression fit.

These impacts are coherent with a reduced number of analogues, as shown in figure 5.15. Lower numbers of analogues, e.g. 6 or 10, result in a similar strengthening of the anomaly in these geographic regions.

Uncertainties for spatial extrapolation

One important issue of gridding is the question about how far we can extrapolate spatially outside of the geographic domain of our proxy database. Mauri et al., 2015 for instance chose a fixed threshold of 500km to the closest pollen site (also shown in the deterministic plots of figure 5.13). Grid cells that are further away than this, are masked.

The ensemble method we present here does not require such a formal definition and should rather be interpreted with respect to the calculated uncertainties. Figure 5.16 shows a clear linear relationship between this distance and the ensemble uncertainty. It also reveals the caveats of using fixed threshold because the relationship between standard deviation and distance to the closest site is time dependent, likely because of the smaller amount of available samples during the early Holocene.

5.5 Discussion

In this section, we will discuss the methodological uncertainties of the method that are not necessarily easy to quantify. They arise from first, the underlying site-based proxy-climate reconstruction method (in this case MAT), second the constrained climate sampling to approximate the joint distribution, and third, the gridding algorithm (*Tps*).

The results from the side-based reconstruction in section 5.4.1, and the gridded reconstruction in section 5.4.2, show that the first two aspects (site-based reconstruction method uncertainty and climate sampling) are closely linked, and potentially

have a strong influence on the outcome. The effects of a reduced number of analogues (figures 5.12 and 5.15), as well as a stronger climatic constraint (figures 5.11 and 5.14) resulted in more extreme temperature anomalies over the same geographic regions. This result is coherent with the sampled temperatures presented in the methods section (figure 5.8) where we show that a higher climatic constraint particularly strengthens the closer climate analogues and as such effectively reduces the number of analogues that is used in the reconstruction. There is no clear and definitive answer to the question how many analogues one should use and how strong the climatic constraint should be. Based on the results however, we favor a less strong climatic constraint of 5 °C and a high number of analogues (50). The philosophy behind is to let the method choose which analogue it uses. If there are inconsistencies within the set of modern analogues for a particular dataset, then this should be reflected by the ensemble standard deviation, and can eventually be compensated by the spatially neighboring samples in the gridding process.

These are all problems that arise from the underlying discrete distribution of modern analogues. An alternativity would be to use a PDF based method (Chevalier et al., 2014; Chevalier, 2019, for instance) that can potentially overcome these weeknesses by providing a more continuous distribution for sampling the climate.

The third aspect of the uncertainty is related to the gridding algorithm itself which can be added on top of the uncertainty of our ensemble method. For the *Tps* method, this uncertainty can be estimated conveniently using the *predictSE* function of the *fields* R-package that approximates the covariances of the prediction based on a linear combination of the observed data under the assumption of fixed covariance parameters (see Nychka et al., 2017 for details). A calculation of this standard error for 20 ensemble members revealed that it is rather independent of the individual realization. As such, we present the averaged standard error of these 20 members in the supplementary figure 5.20. We can see that this uncertainty estimate is high towards boundaries of the interpolation domain, but smaller than the ensemble standard deviations in between.

5.6 Conclusions

With *pyleogrid* we present a new methodological framework that transforms multiple site-based proxy-climate reconstructions into a joint spatio-temporal probabilistic climate reconstruction. Our method exploits the climatic and temporal space that is spanned by the intrinsic uncertainties related to the proxy-climate reconstruction method and the dating of the samples, in order to approximate the distribution of potential climate states in the geographic area of interest. Our approach requires little parameterization, is computationally efficient and can be scaled to large hemispheric or even global areas. The generic ensemble approach we present is in principle agnostic to the underlying proxy-climate reconstruction method and to the gridding method and can therefore be extended to a wide range of potential applications.

Compared to previous approaches of a large-scale gridded reconstruction, our method therefore provides a more reliable uncertainty estimate based on our constrained sampling approaches, which is essential for the comparison with climate model output.

The methodology comes with two side-products that are essential for the spatio-temporal ensemble approach. The first one is a methodology to estimate dating uncertainties based on a bivariate model of the age of the sample and its temporal distance to the closest chronological control point in the dataset. This model is

based on all samples with BACON-based chronologies from the Neotoma database under the assumption that they all report the age uncertainties as the 95th percent confidence interval. This strong assumption can be relaxed for future improvements of this method by using the very recent peer-reviewed dataset by Wang et al., 2019 which contains standardized chronologies for more than 500 datasets.

The other side-product, is a probabilistic variant of the site-based modern analogue technique (MAT) that uses a Gibbs sampling algorithm for the ages and analogue climates in order to approximate the joint distribution within a single dataset. This sampling algorithm is constrained for the individual dataset through first, the monotonicity of the sample ages, and second, a climatic threshold that must not be overcome between too temporally neighboring samples. We compare this sampling algorithm to computationally faster algorithms that involve a forward sampling (climate/age of a sample is constrained by its younger predecessor), its inversion, the backward sampling, as well as a combined approach that uses forward and backward sampling. For age sampling we also test an algorithm that starts with an unconstrained sampling of the ages and then applies an a posteriori sorting in order to maintain the correct distribution of samples in the core (sections 5.3.1 and 5.3.2). All of these approaches, however reveal biases when compared to the computationally more demanding, but statistically correct distribution of the Gibbs sampler. The software package *pyleogrid* therefore implements a computationally efficient version of this Gibbs sampling algorithm that efficiently scales from one single dataset to tenth of thousands of realizations of more than a thousand individual datasets, as it has been used in this study.

This sampling successfully reconstructs a probabilistic version of the individual dataset and provides reliable uncertainty estimates. An evaluation of this realization with respect to the number of modern analogues, and the climatic constraint of the sampling algorithm reveals a close linkage of the two parameters. Further developments might therefore explore the usage of other proxy-climate reconstruction methods that provide a more continuous distribution to sample from.

Supplementary material

5.A Estimated age uncertainties

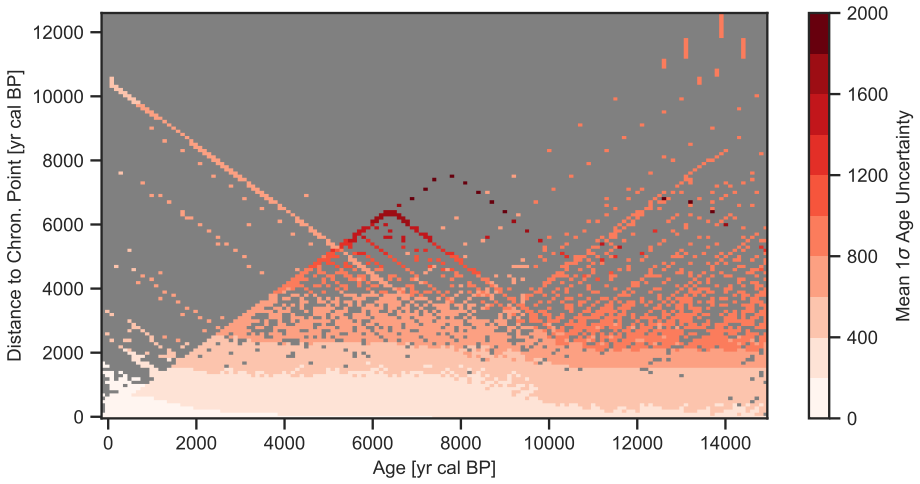


FIGURE 5.17: Estimated age uncertainties for the Eurasian dataset from section 5.2.1 with the same formatting as in figure 5.4.

5.B Example of generated age distributions

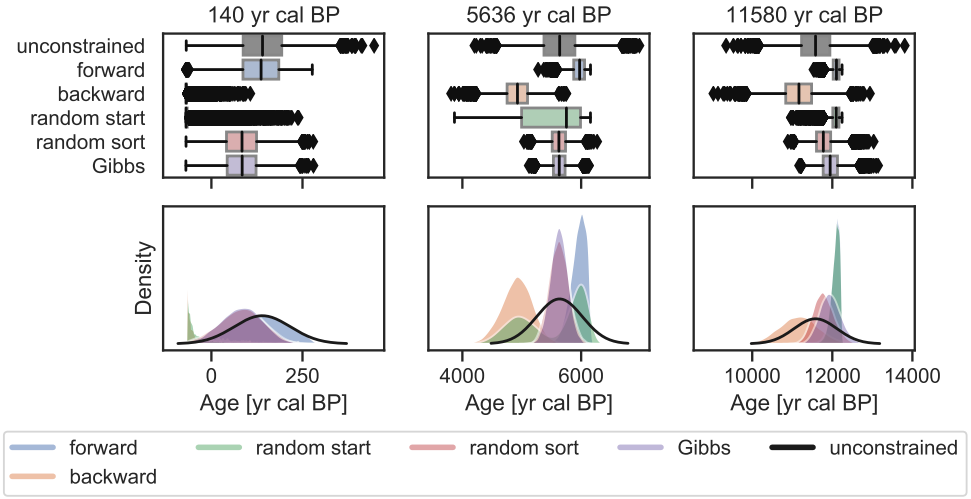


FIGURE 5.18: Example of three samples from the site in section 5.4.1 and their realized distributions. Sampling algorithms are explained in section 5.3.1. Top plots show the box plots of the realized distribution that are visualized with a kernel density estimate in the lower row.

5.C Maps of uncertainties

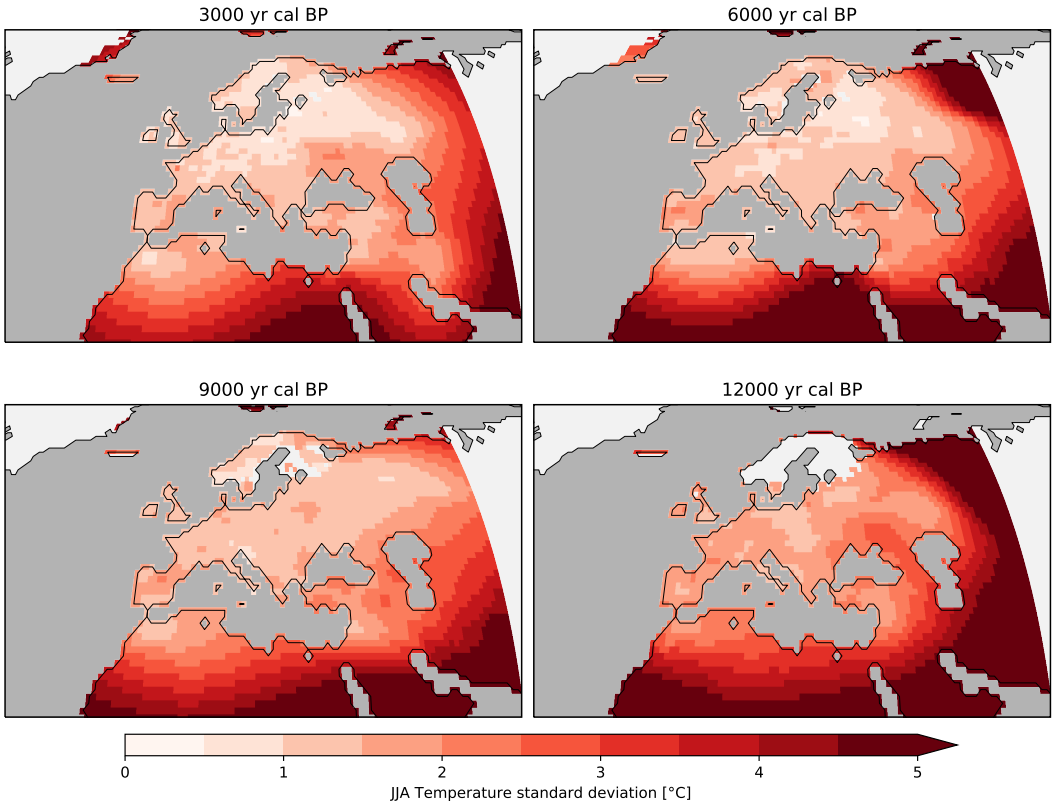


FIGURE 5.19: Standard deviation of the ensemble mean summer temperature anomaly (10'000 realizations) at 3k, 6k, 9k and 12k BP.

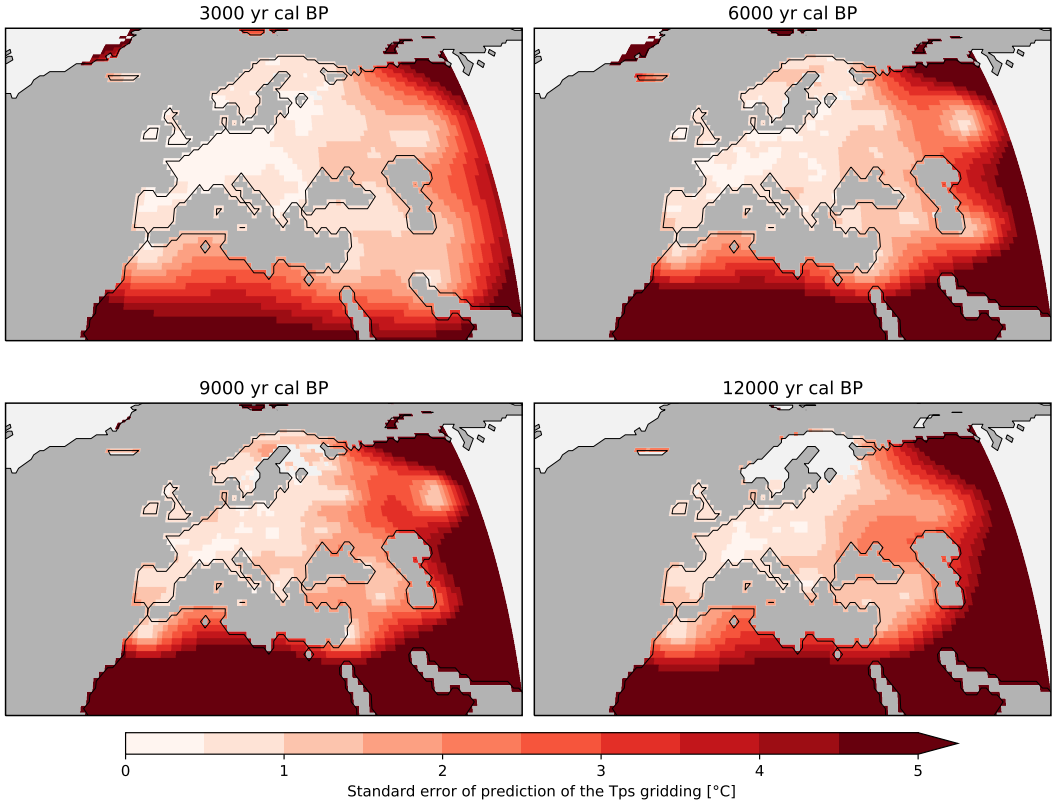


FIGURE 5.20: Standard error of the gridding algorithm, predicted with the *predictSE* method of the *fields* package (Nychka et al., 2017) for the different timeslices. The standard errors are an average of 20 ensemble members, although there is only very little difference between the individual maps.

References

Argus, Donald F., W. R. Peltier, R. Drummond, and Angelyn W. Moore (May 2014). "The Antarctica component of postglacial rebound model ICE-6G_C (VM5a) based on GPS positioning, exposure age dating of ice thicknesses, and relative sea level histories". In: *Geophysical Journal International* 198.1, pp. 537–563. ISSN: 0956-540X. DOI: 10.1093/gji/ggu140. eprint: http://oup.prod.sis.lan/gji/article-pdf/198/1/537/17366620/ggu140.pdf. URL: https://doi.org/10.1093/gji/ggu140.

Bartlein, P. J., S. P. Harrison, S. Brewer, S. Connor, B. A. S. Davis, K. Gajewski, J. Guiot, T. I. Harrison-Prentice, A. Henderson, O. Peyron, I. C. Prentice, M. Scholze, H. Seppä, B. Shuman, S. Sugita, R. S. Thompson, A. E. Viau, J. Williams, and H. Wu (Sept. 2010). "Pollen-based continental climate reconstructions at 6 and 21 ka: a global synthesis". In: *Climate Dynamics* 37.3-4, pp. 775–802. ISSN: 0930-7575 1432-0894. DOI: 10.1007/s00382-010-0904-1.

Behnel, Stefan, Robert Bradshaw, Craig Citro, Lisandro Dalcin, Dag Sverre Seljebotn, and Kurt Smith (Mar. 2011). "Cython: The Best of Both Worlds". In: *Computing in Science & Engineering* 13.2, pp. 31–39. ISSN: 1521-9615. DOI: 10.1109/mcse.2010.118.

Bennett, K.D. (Dec. 1994). "Confidence intervals for age estimates and deposition times in late-Quaternary sediment sequences". In: *The Holocene* 4.4, pp. 337–348. DOI: 10.1177/095968369400400401. eprint: https://doi.org/10.1177/095968369400400401. URL: https://doi.org/10.1177/095968369400400401.

Birks, H. J. B., J. M. Line, S. Juggins, A. C. Stevenson, and C. J. F. T. Braak (Mar. 1990). "Diatoms and pH Reconstruction". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 327.1240, pp. 263–278. DOI: 10.1098/rstb.1990.0062. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.1990.0062. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rstb.1990.0062.

Birks, H. John B., Oliver Heiri, Heikki Seppä, and Anne E. Bjune (Mar. 2010). "Strengths and Weaknesses of Quantitative Climate Reconstructions Based on Late-Quaternary Biological Proxies". In: *The Open Ecology Journal* 3.1, pp. 68–110. DOI: 10.2174/1874213001003020068.

Blaauw, Maarten and J. Andrés Christen (Sept. 2011). "Flexible paleoclimate age-depth models using an autoregressive gamma process". In: *Bayesian Analysis* 6.3, pp. 457–474. ISSN: 1931-6690. DOI: 10.1214/11-ba618.

Braak, Cajo J. F. ter and Steve Juggins (Oct. 1993). "Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages". In: *Hydrobiologia* 269.1, pp. 485–502. ISSN: 1573-5117. DOI: 10.1007/bf00028046. URL: https://doi.org/10.1007/BF00028046.

Brewer, S., Joel Guiot, and Doris Barboni (2007). "Pollen data as climate proxies". In: *Encyclopedia of Quaternary Science*. Elsevier, pp. 2497–2508. URL: https://hal.archives-ouvertes.fr/hal-00995404.

Cheddadi, R., H. F. Lamb, J. Guiot, and S. van der Kaars (Oct. 1998). "Holocene climatic change in Morocco: a quantitative reconstruction from pollen data". In: *Climate Dynamics* 14.12, pp. 883–890. DOI: 10.1007/s003820050262.

Chevalier, M., R. Cheddadi, and B. M. Chase (Nov. 2014). "CREST (Climate REconstruction SofTware): a probability density function (PDF)-based quantitative climate reconstruction method". In: *Climate of the Past* 10.6, pp. 2081–2098. DOI: 10.5194/cp-10-2081-2014.

Chevalier, Manuel (Apr. 2019). "Enabling possibilities to quantify past climate from fossil assemblages at a global scale". In: *Global and Planetary Change* 175, pp. 27–35. DOI: 10.1016/j.gloplacha.2019.01.016.

Dask Development Team (2016). *Dask: Library for dynamic task scheduling*. URL: https://dask.org.

Davis, B. A. S., S. Brewer, A. C. Stevenson, and J. Guiot (2003). "The temperature of Europe during the Holocene reconstructed from pollen data". In: *Quat. Sci. Rev.* 22.15-17, pp. 1701–1716. ISSN: 02773791. DOI: 10.1016/s0277-3791(03)00173-2.

Davis, B. A. S., M. Zanon, P. Collins, A. Mauri, J. Bakker, D. Barboni, A. Barthelmes, C. Beaudouin, A. E. Bjune, E. Bozilova, R. H. W. Bradshaw, B. A. Brayshay, S. Brewer, E. Brugiapaglia, J. Bunting, S. E. Connor, J. L. de Beaulieu, K. Edwards, A. Ejarque, P. Fall, A. Florenzano, R. Fyfe, D. Galop, M. Giardini, T. Giesecke, M. J. Grant, J. Guiot, S. Jahns, V. Jankovska, S. Juggins, M. Kahrmann, M. Karpinska-Kolaczek, P. Kolaczek, N. Kuhl, P. Kunes, E. G. Lapteva, S. A. G. Leroy, M. Leydet, J. Guiot, S. Jahns, V. Jankovska, S. Juggins, M. Kahrmann, M. Karpinska-Kolaczek, P. Kolaczek, N. Kuehl, P. Kunes, E. G. Lapteva, S. A. G. Leroy, M. Leydet, J. A. L. Saez, A. Masi, I. Matthias, F. Mazier, V. Meltsov, A. M. Mercuri, Y. Miras, F. J. G. Mitchell, J. L. Morris, F. Naughton, A. B. Nielsen, E. Novenko, B. Odgaard, E. Ortu, M. V. Overballe-Petersen, H. S. Pardoe, S. M. Peglar, I. A. Pidek, L. Sadori, H. Seppa, E. Severova, H. Shaw, J. Swieta-Musznicka, M. Theuerkauf, S. Tonkov, S. Veski, W. O. van der Knaap, J. F. N. van Leeuwen, J. Woodbridge, M. Zimny, and J. O. Kaplan (2013). "The European Modern Pollen Database (EMPD) project". In: *Vegetation History and Archaeobotany* 22.6, pp. 521–530. ISSN: 0939-6314. DOI: 10.1007/s00334-012-0388-5. URL: http://link.springer.com/article/10.1007%2Fs00334-012-0388-5.

Davis, Basil A. S. and Jed O. Kaplan (Feb. 2017). *HORNET Holocene Climate Reconstruction for the Northern Hemisphere Extra-tropics*. SNF-Research-Plan. last accessed Jan, 30th, 2018. URL: http://p3.snf.ch/project-169598#.

Fick, Stephen E. and Robert J. Hijmans (May 2017). "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas". In: *International Journal of Climatology* 37.12, pp. 4302–4315. DOI: 10.1002/joc.5086. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.5086. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5086.

Fyfe, Ralph M., Jacques-Louis de Beaulieu, Heather Binney, Richard H. W. Bradshaw, Simon Brewer, Anne Le Flao, Walter Finsinger, Marie-Josè Gaillard, Thomas Giesecke, Graciela Gil-Romera, Eric C. Grimm, Brian Huntley, Petr Kunes, Norbert Kühl, Michelle Leydet, Andrè F. Lotter, Pavel E. Tarasov, and Spassimir Ton-kov (Mar. 2009). "The European Pollen Database: past efforts and current activities". In: *Vegetation History and Archaeobotany* 18.5, pp. 417–424. DOI: 10.1007/s00334-009-0215-9.

Giesecke, Thomas, Basil Davis, Simon Brewer, Walter Finsinger, Steffen Wolters, Maarten Blaauw, Jacques-Louis de Beaulieu, Heather Binney, Ralph M. Fyfe, Marie-José Gaillard, Graciela Gil-Romera, W. O. van der Knaap, Petr Kuneš, Norbert Kühl, Jacqueline F. N. van Leeuwen, Michelle Leydet, André F. Lotter, Elena Ortu, Malte Semmler, and Richard H. W. Bradshaw (Mar. 2013). "Towards mapping the late Quaternary vegetation change of Europe". In: *Vegetation History and Archaeobotany* 23.1, pp. 75–86. DOI: 10.1007/s00334-012-0390-y.

Goring, S.J. (2019). *Bulk Baconizing*. https://github.com/NeotomaDB/bulk-baconizing. DOI: 10.5281/zenodo.2545891.

Guiot, Joel and Anne de Vernal (Oct. 2011). "QSR Correspondence "Is spatial autocorrelation introducing biases in the apparent accuracy of palaeoclimatic reconstructions?" Reply to Telford and Birks". In: *Quaternary Science Reviews* 30.21, pp. 3214–3216. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2011.07.023. URL: http://www.sciencedirect.com/science/article/pii/S0277379111002344.

Haslett, John and Andrew Parnell (Sept. 2008). "A simple monotone process with application to radiocarbon-dated depth chronologies". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 57.4, pp. 399–418. DOI: 10.1111/j.1467-9876.2008.00623.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9876.2008.00623.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2008.00623.x.

Holmström, Lasse, Liisa Ilvonen, Heikki Seppä, and Siim Veski (Sept. 2015). "A Bayesian spatiotemporal model for reconstructing climate from multiple pollen records". In: *Ann. Appl. Stat.* 9.3, pp. 1194–1225. DOI: 10.1214/15-AOAS832. URL: https://doi.org/10.1214/15-AOAS832.

Hoyer, S. and J. Hamman (2017). "xarray: N-D labeled arrays and datasets in Python". In: *Journal of Open Research Software* 5.1. DOI: 10.5334/jors.148. URL: http://doi.org/10.5334/jors.148.

Ivanovic, R. F., L. J. Gregoire, M. Kageyama, D. M. Roche, P. J. Valdes, A. Burke, R. Drummond, W. R. Peltier, and L. Tarasov (2016). "Transient climate simulations of the deglaciation 21–9 thousand years before present (version 1) – PMIP4 Core experiment design and boundary conditions". In: *Geosci. Model Dev.* 9.7, pp. 2563–2587. DOI: 10.5194/gmd-9-2563-2016. URL: https://www.geosci-model-dev.net/9/2563/2016/.

Jones, Eric, Travis Oliphant, Pearu Peterson, et al. (2001). *SciPy: Open source scientific tools for Python*. [Online; accessed 2017-02-18]. URL: http://www.scipy.org/.

Juggins, Steve (Mar. 2013). "Quantitative reconstructions in palaeolimnology: new paradigm or sick science?" In: *Quaternary Science Reviews* 64, pp. 20–32. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2012.12.014. URL: http://www.sciencedirect.com/science/article/pii/S0277379112005422.

— (2017). *rioja: Analysis of Quaternary Science Data*. R package version 0.9-21. URL: http://www.staff.ncl.ac.uk/stephen.juggins/.

Kageyama, Masa, Pascale Braconnot, Sandy P. Harrison, Alan M. Haywood, Johann H. Jungclaus, Bette L. Otto-Bliesner, Jean-Yves Peterschmitt, Ayako Abe-Ouchi, Samuel Albani, Patrick J. Bartlein, Chris Brierley, Michel Crucifix, Aisling Dolan, Laura Fernandez-Donado, Hubertus Fischer, Peter O. Hopcroft, Ruza F. Ivanovic, Fabrice Lambert, Daniel J. Lunt, Natalie M. Mahowald, W. Richard Peltier, Steven J. Phipps, Didier M. Roche, Gavin A. Schmidt, Lev Tarasov, Paul J. Valdes, Qiong Zhang, and Tianjun Zhou (Mar. 2018). "The PMIP4 contribution to CMIP6 – Part 1: Overview and over-arching analysis plan". In: *Geoscientific Model Development* 11.3, pp. 1033–1057. DOI: 10.5194/gmd-11-1033-2018. URL: https://www.geosci-model-dev.net/11/1033/2018/.

Lamb, Henry F. and Sander van der Kaars (Dec. 1995). "Vegetational response to Holocene climatic change: pollen and palaeolimnological data from the Middle Atlas, Morocco". In: *The Holocene* 5.4, pp. 400–408. DOI: 10.1177/095968369500500402. eprint: https://doi.org/10.1177/095968369500500402. URL: https://doi.org/10.1177/095968369500500402.

Marcott, S. A., J. D. Shakun, P. U. Clark, and A. C. Mix (Mar. 2013). "A Reconstruction of Regional and Global Temperature for the Past 11,300 Years". In: *Science* 339.6124, pp. 1198–1201. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking). DOI:

10.1126/science.1228026. URL: http://www.ncbi.nlm.nih.gov/pubmed/23471405.

Marsicek, Jeremiah, Bryan N. Shuman, Patrick J. Bartlein, Sarah L. Shafer, and Simon Brewer (Feb. 2018). "Reconciling divergent trends and millennial variations in Holocene temperatures". In: *Nature* 554.7690, pp. 92–96. DOI: 10.1038/nature25464.

Mauri, A., B. A. S. Davis, P. M. Collins, and J. O. Kaplan (2014). "The influence of atmospheric circulation on the mid-Holocene climate of Europe: a data-model comparison". In: *Clim. Past* 10.5, pp. 1925–1938. ISSN: 1814-9324. DOI: 10.5194/cp-10-1925-2014. URL: http://www.clim-past.net/10/1925/2014/cp-10-1925-2014.pdf.

— (2015). "The climate of Europe during the Holocene: a gridded pollen-based reconstruction and its multi-proxy evaluation". In: *Quat. Sci. Rev.* 112, pp. 109–127. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2015.01.013. URL: http://www.sciencedirect.com/science/article/pii/S0277379115000372.

Nychka, Douglas, Reinhard Furrer, John Paige, and Stephan Sain (2017). *fields: Tools for Spatial Data*. R package version 9.8-3. Boulder, CO, USA: University Corporation for Atmospheric Research. DOI: 10.5065/d6w957ct. URL: https://github.com/NCAR/Fields.

Oliphant, Travis E (2006). *A guide to NumPy*. Vol. 1. Trelgol Publishing USA. URL: http://www.numpy.org/.

Otto-Bliesner, B. L., P. Braconnot, S. P. Harrison, D. J. Lunt, A. Abe-Ouchi, S. Albani, P. J. Bartlein, E. Capron, A. E. Carlson, A. Dutton, H. Fischer, H. Goelzer, A. Govin, A. Haywood, F. Joos, A. N. LeGrande, W. H. Lipscomb, G. Lohmann, N. Mahowald, C. Nehrbass-Ahles, F. S. R. Pausata, J.-Y. Peterschmitt, S. J. Phipps, H. Renssen, and Q. Zhang (2017). "The PMIP4 contribution to CMIP6 – Part 2: Two interglacials, scientific objective and experimental design for Holocene and Last Interglacial simulations". In: *Geosci. Model Dev.* 10.11, pp. 3979–4003. DOI: 10.5194/gmd-10-3979-2017. URL: https://www.geosci-model-dev.net/10/3979/2017/.

Peltier, W. R., D. F. Argus, and R. Drummond (Jan. 2015). "Space geodesy constrains ice age terminal deglaciation: The global ICE-6G_C (VM5a) model". In: *Journal of Geophysical Research: Solid Earth* 120.1, pp. 450–487. DOI: 10.1002/2014jb011176. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2014JB011176. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014JB011176.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Servén, Daniel, Charlie Brummitt, and Hassan Abedi (2018). *Dswah/Pygam: V0.8.0*. DOI: 10.5281/zenodo.1476122.

Sommer, Philipp S., Basil A. S. Davis, Manuel Chevalier, Jian Ni, and John Tipton (2019). "The HORNET project: applying 'big data' to reconstruct the climate of the Northern Hemisphere during the Holocene". In: *20th Congress of the International Union for Quaternary Research (INQUA)*. International Union for Quaternary Research. URL: https://app.oxfordabstracts.com/events/574/program-app/submission/94623.

Telford, R. J., E. Heegaard, and H. J. B. Birks (Jan. 2004). "All age–depth models are wrong: but how badly?" In: *Quaternary Science Reviews* 23.1, pp. 1–5. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2003.11.003. URL: http://www.sciencedirect.com/science/article/pii/S0277379103003160.

Telford, R.J. and H.J.B. Birks (Nov. 2005). "The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance". In: *Quaternary Science Reviews* 24.20, pp. 2173–2179. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2005.05.001. URL: http://www.sciencedirect.com/science/article/pii/S027737910500168X.

— (June 2009). "Evaluation of transfer functions in spatially structured environments". In: *Quaternary Science Reviews* 28.13, pp. 1309–1316. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2008.12.020. URL: http://www.sciencedirect.com/science/article/pii/S0277379108003806.

Trachsel, Mathias and Richard J Telford (Nov. 2016). "All age–depth models are wrong, but are getting better". In: *The Holocene* 27.6, pp. 860–869. DOI: 10.1177/0959683616675939. eprint: https://doi.org/10.1177/0959683616675939. URL: https://doi.org/10.1177/0959683616675939.

Waelbroeck, C., A. Paul, M. Kucera, A. Rosell-Melé, M. Weinelt, R. Schneider, A. C. Mix, A. Abelmann, L. Armand, E. Bard, S. Barker, T. T. Barrows, H. Benway, I. Cacho, M.-T. Chen, E. Cortijo, X. Crosta, A. de Vernal, T. Dokken, J. Duprat, H. Elderfield, F. Eynaud, R. Gersonde, A. Hayes, M. Henry, C. Hillaire-Marcel, C.-C. Huang, E. Jansen, S. Juggins, N. Kallel, T. Kiefer, M. Kienast, L. Labeyrie, H. Leclaire, L. Londeix, S. Mangin, J. Matthiessen, F. Marret, M. Meland, A. E. Morey, S. Mulitza, U. Pflaumann, N. G. Pisias, T. Radi, A. Rochon, E. J. Rohling, L. Sbaffi, C. Schäfer-Neth, S. Solignac, H. Spero, K. Tachikawa, J.-L. Turon, and M. A. R. G. O. Project Members (Jan. 2009). "Constraints on the magnitude and patterns of ocean cooling at the Last Glacial Maximum". In: *Nature Geoscience* 2.2, pp. 127–132. ISSN: 1752-0908. DOI: 10.1038/ngeo411. URL: https://doi.org/10.1038/ngeo411.

Wang, Yue, Simon J. Goring, and Jenny L. McGuire (2019). "Bayesian ages for pollen records since the last glaciation in North America". In: *Scientific Data* 6.1, p. 176. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0182-7. URL: https://doi.org/10.1038/s41597-019-0182-7.

Williams, John W., Eric C. Grimm, Jessica L. Blois, Donald F. Charles, Edward B. Davis, Simon J. Goring, Russell W. Graham, Alison J. Smith, Michael Anderson, Joaquin Arroyo-Cabrales, Allan C. Ashworth, Julio L. Betancourt, Brian W. Bills, Robert K. Booth, Philip I. Buckland, B. Brandon Curry, Thomas Giesecke, Stephen T. Jackson, Claudio Latorre, Jonathan Nichols, Timshel Purdum, Robert E. Roth, Michael Stryker, and Hikaru Takahara (Jan. 2018). "The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource". In: *Quaternary Research* 89.1, pp. 156–177. DOI: 10.1017/qua.2017.105.

Chapter 6

GWGEN v1.0

A globally calibrated scheme for generating daily meteorology from monthly statistics

From

Sommer, Philipp S. and Jed O. Kaplan (Oct. 2017a). "A globally calibrated scheme for generating daily meteorology from monthly statistics: Global-WGEN (GWGEN) v1.0". In: *Geosci. Model Dev.* 10.10, pp. 3771–3791. DOI: 10.5194/gmd-10-3771-2017.

Abstract. While a wide range of earth system processes occur at daily and even sub-daily timescales, many global vegetation and other terrestrial dynamics models historically used monthly meteorological forcing, both to reduce computational demand and because global datasets were lacking. Recently, dynamic land surface modeling has moved towards resolving daily and subdaily processes, and global datasets containg daily and sub-daily meteorology have become available. These meteorological datasets, however, cover only the instrumental era of the last ca. 120 years at best, are subject to considerable uncertainty, and represent extremely large data files with associated computational costs of data input/output and file transfer. For periods before the recent past or into the future, global meteorological forcing can be provided by climate model output, but the quality of these data at high temporal resolution is low, particularly for daily precipitation frequency and amount. Here we present GWGEN, a globally applicable statistical weather generator for the temporal downscaling of monthly climatology to daily meteorology. Our weather generator is parameterized using a global meteorological database and simulates daily values of five common variables: minimum and maximum temperature, precipitation, cloud cover, and wind speed. GWGEN is lightweight, modular, and requires a minimal set of monthly mean variables as input. The weather generator may be used in a range of applications, for example, in global vegetation, crop, soil erosion, or hydrological models. While GWGEN does not currently perform spatially autocorrelated multi-point downscaling of daily weather, this additional functionality could be implemented in future versions.

6.1 Introduction

The development of the first global vegetation models in the 1970's (e.g., Lieth, 1975) brought about the demand for meteorological forcing datasets with global extent and relatively high spatial resolution, e.g., $1^\circ \times 1^\circ$. While a global weather

station-based monthly climate dataset was available at this time (Walter and Lieth, 1967), limitations in computers and storage allowed only the simplest treatment of these data. The first global simulations of the net primary productivity of the terrestrial biosphere (Lieth, 1975), thus used rasterized polygons of annual meteorological variables that had been crudely interpolated from the station-based climatology. A decade later saw the development of better computers and more sophisticated global vegetation models (Prentice et al., 1992; Prentice, 1989) that recognized the need for forcing at a sub-annual timestep and development of these models was done in parallel with the first global, gridded high resolution (0.5°) monthly climatology (Leemans and Cramer, 1991). At the time, monthly meteorological data was the only feasible global data that could be produced, in terms of the raw station data available to feed the interpolation process, the processing time required to produce gridded maps, and the data storage and transfer capabilities of contemporary computer systems and networks. Global gridded monthly climate data thus became the standard for not only large-extent vegetation modeling (Haxeltine and Prentice, 1996; Haxeltine et al., 1996; Kaplan et al., 2003; Kucharik et al., 2000; Woodward et al., 1995), but also for a wide range of studies on biodiversity and species distribution (e.g., Elith et al., 2006), vegetation trace gas emissions (e.g., Guenther et al., 1995), and even the geographic distribution of human diseases (e.g., Bhatt et al., 2013)

Over subsequent years, the global gridded monthly climate datasets were improved (New et al., 1999, 2002), developed with very high spatial resolution (Hijmans et al., 2005), and expanded beyond mean climate to cover continuous timeseries over decades (Harris et al., 2014; Mitchell and Jones, 2005; New et al., 2000). The latter was an essential requirement for forcing dynamic global vegetation models (DGVMs) (e.g., Sitch et al., 2003). However, despite increasing quality, spatial resolution, and temporal extent in these datasets, the basic time step remained monthly, partly for legacy reasons — models had been developed in an earlier era subject to computational limitations and therefore used a monthly timestep for efficiency even if this was no longer strictly a constraint — and partly because of the challenge in developing a global, high-resolution climate dataset with a daily or shorter timestep still presented a major data management challenge.

On the other hand, there was increasing awareness that accurate simulation of many earth surface processes required representation of processes at a shorter-than-monthly timestep. Global simulation of surface hydrology (Gerten et al., 2004), crop growth (Bondeau et al., 2007), or biogeophysical processes (Krinner et al., 2005) needed sub-monthly forcing to produce reliable results. To address this need for better forcing data, two main approaches were taken: either monthly climate data were downscaled online using a stochastic weather generator (e.g., Pfeiffer et al., 2013), or a sub-daily, high-resolution, gridded climate timeseries was generated directly by merging high-temporal-resolution reanalysis data (e.g., NCEP, 6h, 2.5°) with high-spatial-resolution monthly climate data (e.g., CRU, 0.5°). The latter process resulted in the CRUNCEP dataset (Viovy and Ciais, 2016; Wei et al., 2014), which, while global, is large even by modern standards (ca. 350 Gb), is not available at spatial resolution greater than 0.5°, and covers only the period 1901-2014.

Forcing data for global vegetation and other models with shorter than monthly resolution at higher spatial resolutions than 0.5°, or for any other period than the last ca. 120 years, e.g., for the future or the more distant past, may therefore only be available through downscaling techniques. One approach to overcome the limitations of currently available datasets could be to use GCM output directly, however, most GCM output currently available does not have greater than 0.5° spatial resolution,

with the current generation of GCMs typically approaching ca. $1^\circ \times 1^\circ$. Furthermore, there is a general observation that daily meteorology produced by GCMs is not realistic, particularly for precipitation (Dai, 2006; Stephens et al., 2010; Sun et al., 2006). An alternative approach is, therefore, to perform temporal downscaling on monthly meteorological data using a statistical weather generator.

Statistical weather generators were first developed primarily for crop and hydrological modeling at the field to catchment scale (Richardson, 1981; Woolhiser and Pegram, 1979; Woolhiser and Roldan, 1982). The weather generator was parameterized using daily meteorological observations at one or more weather stations close to the area of interest, although some attempts were made to generalize the parameterization over larger, sub-continental regions (e.g., Wilks, 1998, 1999b; Woolhiser and Roldán, 1986). Locally parameterized weather generators have been applied to a very wide range of studies (Wilks and Wilby, 1999; Wilks, 2010), and enhanced to include additional meteorological variables beyond the original precipitation, temperature, and solar radiation (e.g., Parlange and Katz, 2000). Applications of a weather generator at continental to global scales was still limited, however, because of the need to perform local parameterization.

The need to simulate daily meteorology in regions of the world with short, unreliable, or unavailable daily meteorological timeseries brought about the realization that certain features of weather generator parameterization might be generalized across a range of climates (Geng et al., 1986; Geng and Auburn, 1987). This ultimately led to the development of globally applicable weather generators (Friend, 1998), and their incorporation in DGVMs (Bondeau et al., 2007; Gerten et al., 2004; Pfeiffer et al., 2013). The original global parameterization (Geng et al., 1986) of these weather generators was, however, limited to seven weather stations, mostly in the temperate latitudes. Friend, 1998 does not publish the parameters used in his global weather generator, but we assume these were the same as the original Geng and Auburn, 1987 and Geng et al., 1986 models. Given the availability of 1) large datasets of daily meteorology, and 2) computers powerful enough to process these data, we therefore decided that it would be valuable to revisit these parameterizations, perform a systematic and quantitative evaluation of the resulting downscaled meteorology, and potentially improve our ability to perform monthly-to-daily downscaling of common meteorological variables with a single, globally applicable parameterization.

In the following sections we describe Global-WGEN (GWGEN, Sommer and Kaplan, 2017a), a weather generator parameterized using more than 50 million daily weather observations from all continents and latitudes. We demonstrate how updated schemes for simulating precipitation occurrence and amount, and for bias correcting wind speed, further improve the quality of the model simulations. We perform an extensive model evaluation and parameter uncertainty analysis in order to settle on a parameter set that provides the most accurate, globally applicable results. We comment on the limitations of the model and priorities for future research. GWGEN is an open-source, stand-alone model that may be incorporated into any number of models designed to work at global scale, including, e.g., vegetation, hydrology, climatology, and animal distribution models.

6.2 Model description

GWGEN requires the following six monthly summary values as input: 1) total monthly precipitation, 2) the number of days in the month with measurable precipitation (i.e.,

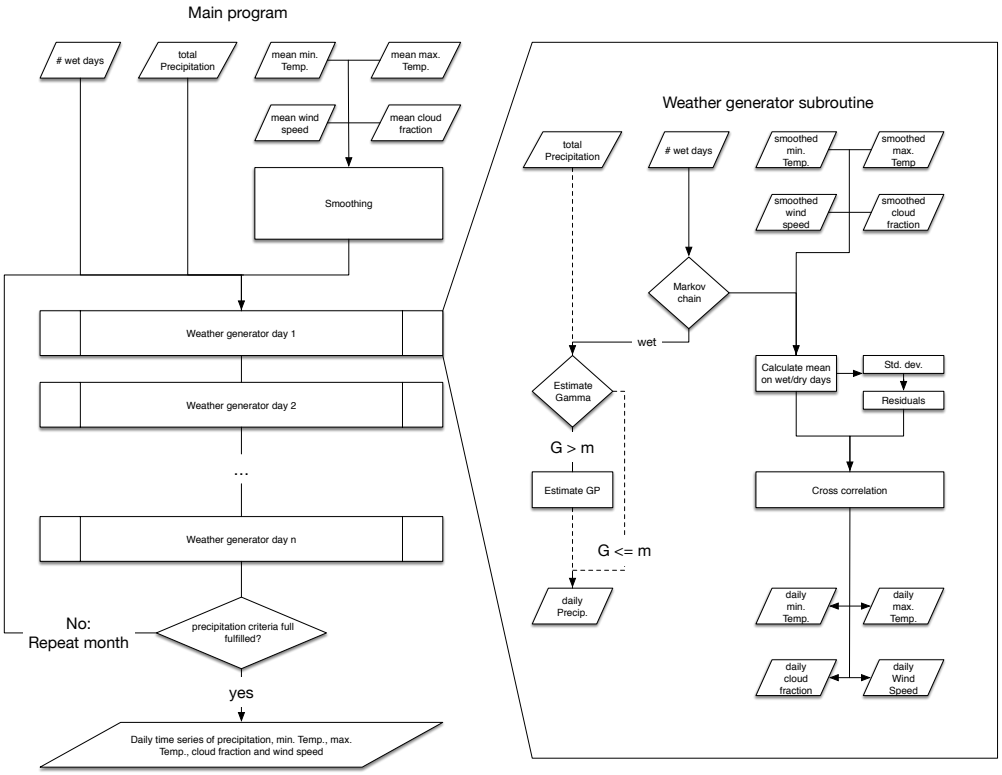


FIGURE 6.1: Schematic workflow of GWGEN. After smoothing the monthly input, the Markov Chain is used to decide, whether it is a dry or a wet day. If it is a wet day, we draw a random number from the Gamma-GP distribution. Furthermore, the other means of the variables ($\bar{T}_{\min/\max}, \bar{c}, \bar{w}$) are adjusted and their daily values are calculated using the estimated standard deviations and residuals. The wind speed furthermore undergoes a square root transformation before applying the cross correlation and in the end is corrected using the bias correction. A quality check in the end restricts our model to be within a 5% range of the observed total precipitation and to replicate the number of wet days from the input.

wet days), 3-4) monthly mean daily minimum and maximum temperature, 5) mean cloud fraction, and 6) wind speed. The model outputs are the same variables at daily resolution. This section summarizes the basic workflow in the model which is also shown schematically in Figure 6.1 and Algorithm 2.

The first approximation of the daily variables comes from smoothing the monthly time series using a mean-preserving algorithm (Rymes and Myers, 2001).

For precipitation we then first use the Markov Chain approach (section 6.3.2) to decide the wet/dry state of the day. If it is a wet day, we calculate the gamma parameters using the equations (6.7) and (6.8). The resulting distribution allows us to draw a random number, the precipitation amount of the currently simulated day. If we are above the threshold μ, we draw a second random number from the Generalized Pareto (GP) distribution parameterized via equation (6.9) and the chosen GP shape.

The next step modifies the means of temperature, wind speed and cloud fraction depending on the wet/dry state of the day (lines 11 and 15 in algorithm 2). After that, we use the cross-correlation approach described in Richardson, 1981 (lines 18 - 20 and Equation 6.3.2) and calculate the daily values of these variables. Finally we use the quantile-based bias correction described in section 6.3.4 to correct the simulated wind speed.

We restrict the weather generator to reproduce the exact number of wet days (± 1) as the input and to be within a 5% range of the total monthly precipitation (with a maximum allowed deviation of 0.5 mm). If the program cannot produce these results, the procedure described above is repeated (see line 4).

6.3 Model development

GWGEN is based on the WGEN weather generator (Richardson, 1981), using the method of defining the model parameters based on monthly summaries described by Geng et al., 1986 and Geng and Auburn, 1987. GWGEN diverges from the original WGEN by using a hybrid-order Markov chain to simulate precipitation occurrence (Wilks, 1999a), and a hybrid Gamma-GP distribution (Furrer and Katz, 2008; Neykov et al., 2014) to estimate precipitation amount. Temperature, cloud cover, and wind speed are calculated following (Richardson, 1981), using cross correlation and depending on the wet/dry-state of the day. We further add a quantile-based bias correction for wind speed and minimum temperature, which improves the simulation results significantly.

In the following subsections, we first describe the global weather station database used to develop and evaluate the model, then describe the underlying relationships that we use to define GWGEN's parameters.

6.3.1 Development of a global weather station database

To parameterize GWGEN, we assembled a global dataset of daily meteorological observations. Precipitation and minimum and maximum daily temperature come from the daily Global Historical Climatology Network (GHCN-Daily) database (Menne et al., 2012a,b). The GHCN-Daily consists of observations collected at ca. 100'000 weather stations on all continents and many oceanic islands. As the GHCN-Daily stations are highly concentrated in some parts of the world, particularly in the conterminous United States, we selected stations for our study using a geographic anti-aliasing filter to avoid an especially strong geographic bias in the generation of the

Algorithm 2 Basic workflow of GWGEN

Require: monthly precipitation $P_{\text{in}}\,[\text{mm}]$, cloud cover fraction c_{in}, minimum ($T_{\text{min,in}}\,[^\circ\text{C}]$) and maximum ($T_{\text{max,in}}\,[^\circ\text{C}]$) temperature, wind speed $w_{\text{in}}\,[\text{m/s}]$, number of wet days n_{in}

Output: daily $P_i\,[\text{mm/d}], c_i, T_i\,[^\circ\text{C}], w_i\,[\text{m/s}]$ and the wet/dry state $s_i \in \{0, 1\}$

for month m in $input$ **do**
  smooth the monthly data using Rymes and Myers, 2001
  Set $j = 0, \chi = 0$
  while $j \equiv 0$ or $\left|\sum_{d_i \in m} P_i - P_{\text{in}}\right| > \min\left(5\% \cdot P_{\text{in}}, 0.5mm\right)$ or $|n_{\text{sim}} - n_{\text{in}}| > 1$ **do**
    for day d_i in m **do**
      Calculate p_{11}, p_{101}, p_{001} after equations (6.1) - (6.3) using n {Precipitation occurence after Wilks, 1999a}
      Use the Markov chain to determine whether d_i is wet ($s_i = 1$) or dry ($s_i = 0$)
      if $s_i = 1$ **then**
        Calculate θ, α and σ via eq. (6.7)-(6.9) {Precipitation amount after Neykov et al., 2014}
        Draw a random number P_i from the Gamma-GP distribution, eq. (6.6)
        Set $T_{\text{min},i} = T_{\text{min,wet}}, T_{\text{max},i} = T_{\text{max,wet}}, c_i = c_{\text{wet}}, w_i = w_{\text{wet}}$ from eq. (6.10) and (6.12) and tables 6.1, 6.3
        Set $\sigma_{T_{\text{min}},i} = \sigma_{T_{\text{min}},\text{wet}}, \sigma_{T_{\text{max}},i} = \sigma_{T_{\text{max}},\text{wet}}, \sigma_{w,i} = \sigma_{w,\text{wet}}, \sigma_{c,i} = \sigma_{c,\text{wet}}$ from eq. (6.11), (6.13) and (6.14) and tables 6.1, 6.2, 6.3
      else
        Set $P_i = 0\,\text{mm/d}$
        Set $T_{\text{min},i} = T_{\text{min,dry}}, T_{\text{max},i} = T_{\text{max,dry}}, c_i = c_{\text{dry}}, w_i = w_{\text{dry}}$ from eq. (6.10) and (6.12) and tables 6.1, 6.3
        Set $\sigma_{T_{\text{min}},i} = \sigma_{T_{\text{min}},\text{dry}}, \sigma_{T_{\text{max}},i} = \sigma_{T_{\text{max}},\text{dry}}, \sigma_{w,i} = \sigma_{w,\text{dry}}, \sigma_{c,i} = \sigma_{c,\text{dry}}$ from eq. (6.11), (6.13) and (6.14) and tables 6.1, 6.2, 6.3
      end if
      Draw 4 normally distributed random numbers $\epsilon \in \mathbb{R}^4$ {Cross correlation after Richardson, 1981}
      Set the residuals $\chi_i = \begin{pmatrix}\chi_{T_{\text{min}}} & \chi_{T_{\text{max}}} & \chi_c & \chi_w\end{pmatrix} = A\chi_{i-1} + B\epsilon \in \mathbb{R}^4$ with A and B from eq. (6.17)
      Calculate daily variables via
      $T_{\text{min},i} = \chi_{T_{\text{min}}} \cdot \sigma_{T_{\text{min}},i} + T_{\text{min},i} \qquad c_i = \chi_c \cdot \sigma_{c,i} + c_i$
      $T_{\text{max},i} = \chi_{T_{\text{max}}} \cdot \sigma_{T_{\text{max}},i} + T_{\text{max},i} \qquad w_i = \left(\chi_w \cdot \sqrt{\sigma_{w,i}} + \sqrt{w_i}\right)^2$
      Apply bias correction w (eq. (6.23))
      $j = j + 1$
    end for
  end while
end for

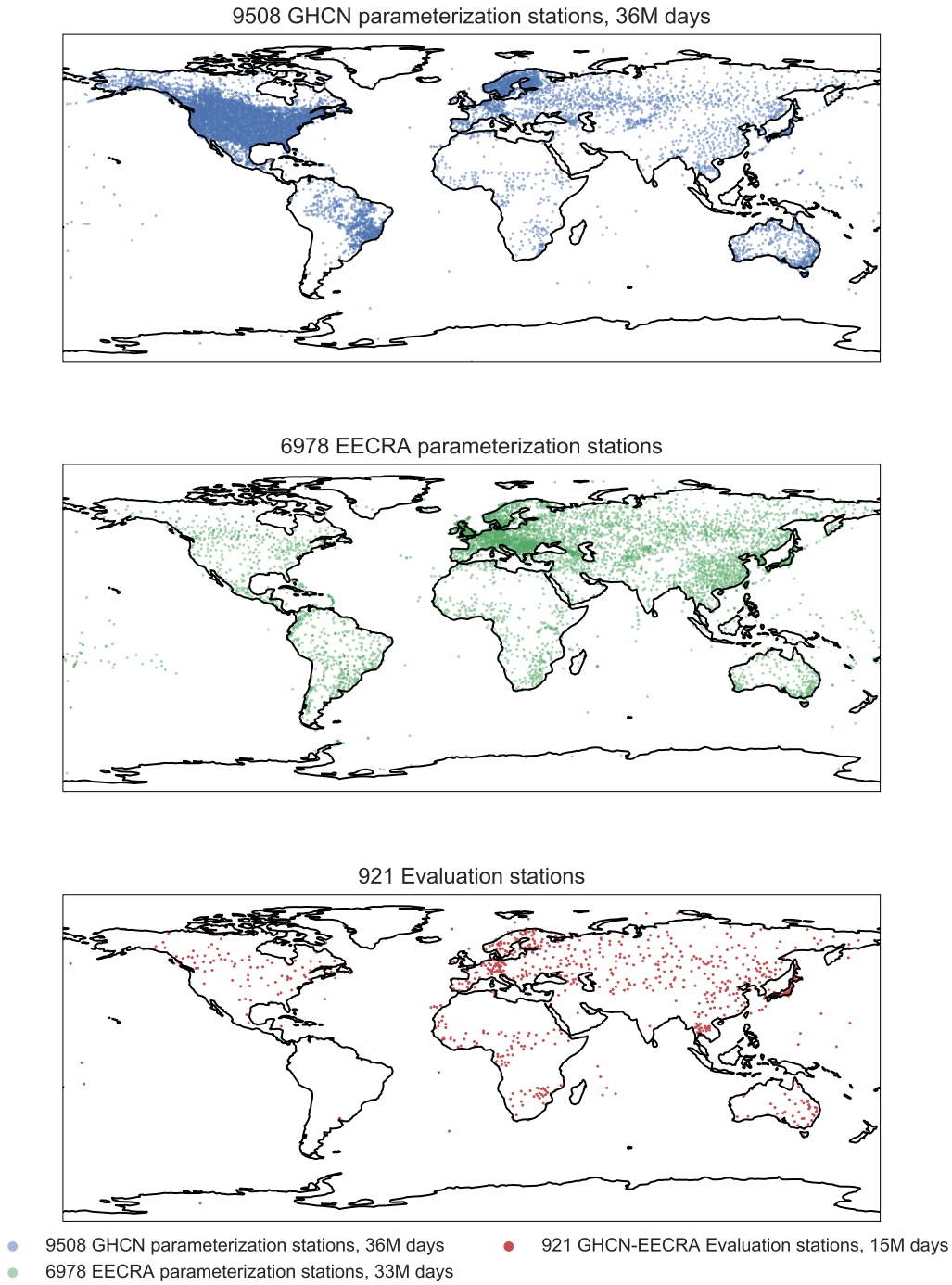


FIGURE 6.2: Weather stations used for parameterization and evaluation of the weather generator. The uppermost panel shows the locations of the stations used for parameterizing precipitation and temperature, the middle panel shows the stations for cloud fraction and wind speed, as well as for calculating the cross correlations between temperature, cloud fraction, and wind speed. The lower plot shows the location of the stations used to evaluate the model, which were excluded from the parameterization stations.

model parameters. Dividing the world up into a 0.5° grid, we selected the single station with the longest record in each cell, if one was present. While the GHCN-Daily units for precipitation have a nominal precision of 0.1 mm, several of the stations in the United States reported precipitation in fractions of an inch, which were later converted to mm. To ensure uniform precision across all of our calibration stations — this was particularly important when generating the probability density functions for precipitation amount — we selected only those GHCN-Daily stations where all precipitation amounts between 0.1 and 1.0 mm d^{-1} were reported in the record. This resulted in 9508 stations covering all continents, although the distribution is strongly heterogenous, with the majority of the stations in North America, despite our geographic filter (Figure 6.2, top panel). For cloud cover, windspeed, and to calculate cross-correlations between temperature, cloud cover, and windspeed, we used the Extended Edited Cloud Report Archive (EECRA) database (Hahn and Warren, 1999). The geographic distribution of the 6978 EECRA stations we selected is different than the GHCN-Daily, with more stations in Europe (Figure 6.2, middle panel), but overall a relatively similar number of stations were used from both datasets. For the observations from both GHCN-Daily and EECRA, we made one additional filtering step, selecting only complete months, i.e., months with no days having missing observations, for further processing. In total, our database of daily meteorological observations used in the model parameterization contains ca. 69 million individual records.

Finally, we reserved some weather station records for model evaluation that were not used for model parameterization. These were individual stations, or two stations separated by a maximum distance of 1 km, where all of the daily meteorological variables that GWGEN simulates (P, T_{min}, T_{max}, c, w) were recorded on the same dates in the EECRA database. This merged selection from EECRA and GHCN resulted in a set of 921 stations representing ca. 15 million daily records, with observations on all continents, although the geographic distribution is once again highly heterogenous, with a particularly high density of stations in Japan and Germany (Figure 6.2, bottom panel).

6.3.2 Parameterization

Precipitation occurrence

Following Geng et al., 1986, we expect to find a good relationship between the fraction of days in a month with measurable precipitation and the probability that any given day will be wet. Following Wilks, 1999a we use a hybrid-order model that retains first-order Markov dependence for wet spells but allows second-order dependence for dry sequences; this hybrid-order scheme has been shown to be a good compromise between performance and simplicity. To parameterize the precipitation occurrence part of the model, we thus calculated transition probabilities for a wet day being followed by a wet day (p_{11}), for a wet day being followed by a dry day being followed by a wet day (p_{101}) and for two dry days being followed by a wet day (p_{001}). We perform this analysis on a station and month-wise basis, i.e., we first extract each of the (complete) Januaries, Februaries, etc. for a given station, and then merge all of the Januaries (Februaries, Marches, etc...) for this station into a single series representing each month. Merging months over several years is particularly important for stations that have relatively little precipitation in a given month; for example, it could take several years of observations to observe a single (p_{101}) event. The final transition probabilities were then regressed against the fraction of days

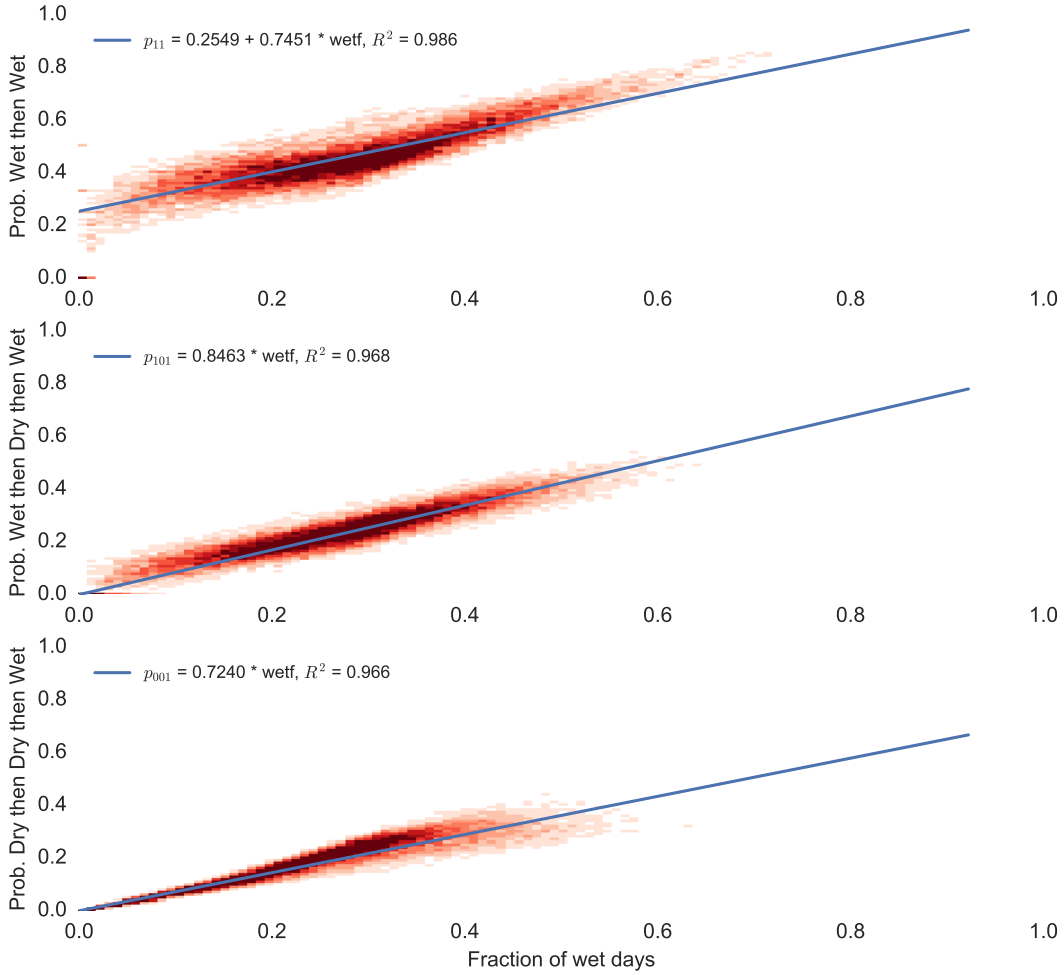


FIGURE 6.3: Transition probabilities vs. wet fraction. The red density plot in the background shows the density of the observations, and the blue lines are the linear regression line of the probability against the wet fraction. The fit for the p_{11} transition probability was forced to the point $(1, 1)$, the others were forced to $(0, 0)$. The underlying data for the fits correspond to the means of the the multi-year series for each month for each station.

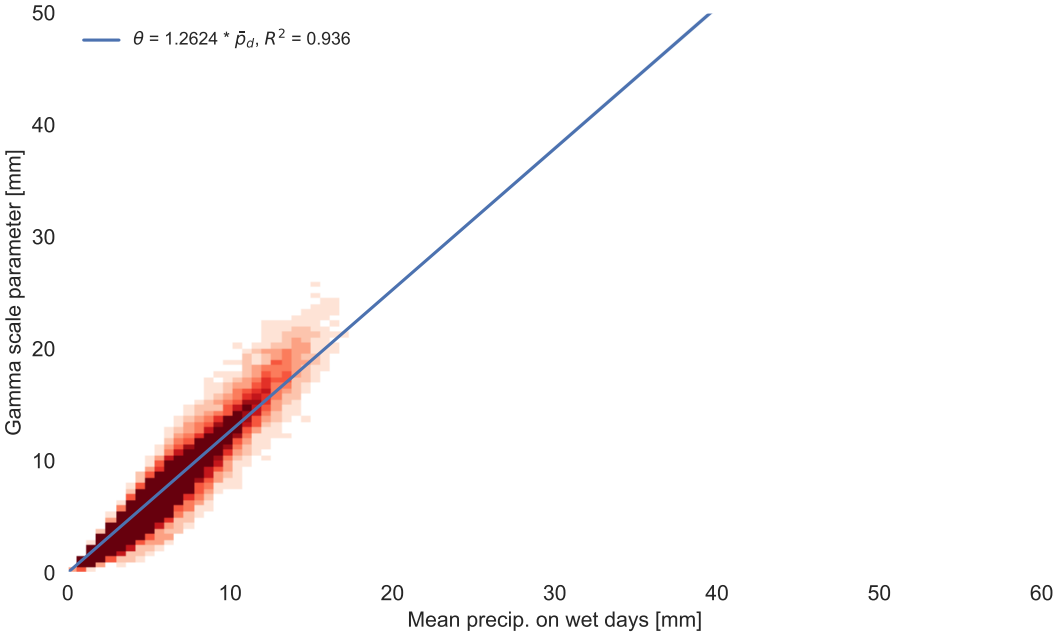


FIGURE 6.4: Mean precipitation - Gamma scale relationship. The blue line represents the best fit line of the mean precipitation on wet days to the estimated gamma scale parameter of the corresponding distribution. Each data point corresponds to one multi-year series of one month for one station.

in the month with precipitation, which show the characteristic linear relationship described by Geng et al., 1986 (Figure 6.3).

Because the transition probabilities (p_{001}) and (p_{101}) must be zero by definition when the fraction of wet days (f_{wet}) is zero, i.e., a completely dry month, we force the linear regression between these quantities to pass through the origin. Likewise, we require the regression line for (p_{11}) to equal 1 when f_{wet} is 1. One has to note, however, that this method artificially increases the R^2 coefficient for the fit because we fix the intercept (see for example Gordon, 1981).

The analysis results in the the following relationships:

$$p_{11} = 0.2549 + 0.7451 \cdot f_{\text{wet}} \tag{6.1}$$
$$p_{101} = 0.8463 \cdot f_{\text{wet}} \tag{6.2}$$
$$p_{001} = 0.7240 \cdot f_{\text{wet}}. \tag{6.3}$$

In the weather generator (see line 6 in algorithm 2) we determine if any given day will have precipitation by calculating the appropriate probability density function selected from equations (6.1)-(6.3) on the basis of the precipitation state of the previous day (or two). Comparing the calculated probability from the selected equation with a random number $u \in [0, 1]$, a precipitation day is simulated if u is greater than its corresponding probability.

Precipitation amount

Following the original WGEN (Richardson, 1981), GWGEN disaggregates precipitation amount using a statistical distribution. A number of different probability

density functions have been used to estimate precipitation amount in weather generators including, e.g., single exponential or mixed exponential, one or two parameter gamma, or Weibull distribution (Wilks and Wilby, 1999). The strong relationship between the gamma scale parameter and the mean precipitation on wet days noted by Geng et al., 1986 makes generation of precipitation amounts with only monthly input data feasible. It is based upon the fact that the expected value of a gamma random variable equals the product of its two parameters. i.e $E(\Gamma) = \alpha\theta$. The gamma distribution, however, shows poor performance in simulating high-precipitation events consistent with observations. Furrer and Katz, 2008 and Neykov et al., 2014 suggest that a hybrid probability density function, based on both gamma and the generalized pareto (GP) distribution, has superior accuracy in simulating extreme precipitation events when compared to gamma alone. Because of its superior accuracy and ease of implementation, we therefore adopt the hybrid gamma-GP distribution for simulating precipitation amount in GWGEN.

The probability density function (pdf) of the gamma distribution is defined as

$$f(x) = \begin{cases} \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^{\alpha}\Gamma(\alpha)} & \text{for } x > 0 \\ 0 & \text{for } x = 0 \end{cases} \tag{6.4}$$

where $\alpha > 0$ is the shape, and $\theta > 0$ the scale parameter. The pdf of the generalized pareto (GP) distribution is defined via

$$g(x) = \begin{cases} \frac{1}{\sigma}\left(1 + \frac{\xi\,(x-\mu)}{\sigma}\right)^{-\frac{1}{\xi}-1} & \text{for } \xi \neq 0 \\ \frac{1}{\sigma}e^{-\frac{x-\mu}{\sigma}} & \text{for } \xi = 0 \end{cases} \tag{6.5}$$

with $\sigma > 0$ being the scale parameter and $\xi \in \mathbb{R}$ the shape parameter. μ is the location parameter.

Following Furrer and Katz, 2008, we define the hybrid gamma-GP pdf as

$$h(x) = \begin{cases} f(x) & \text{for } x \leq \mu \\ (1 - F(\mu))\, g(x) & \text{for } x > \mu \end{cases}, \tag{6.6}$$

where $F(\mu)$ describes the cumulative gamma distribution function at the threshold μ. In our weather generator however, we first draw a random number from the gamma distribution and, if we are above the threshold, we draw another random number from the GP distribution. Thus, the frequency of precipitation events larger than μ is determined by the gamma distribution, but the actual amount of precipitation simulated when above the threshold μ is determined by the GP distribution (Furrer and Katz, 2008).

To determine the parameters of the hybrid distribution for precipitation, we started with the simple strategy by Geng et al., 1986. As above when calculating the Markov chain parameters, we created multi-year series for each of the parameterization stations for each month and extracted the days with precipitation. If a series contained more than 100 entries, we fit a gamma distribution using maximum likelihood to it in order to estimated the α and θ parameters.

Following Geng et al., 1986, we then fit a regression line of the gamma scale parameter against the mean precipitation on wet days $\bar{p}_d$ (see figure 6.4) and found the relationship

$$\theta = 1.262\,\bar{p}_d. \tag{6.7}$$

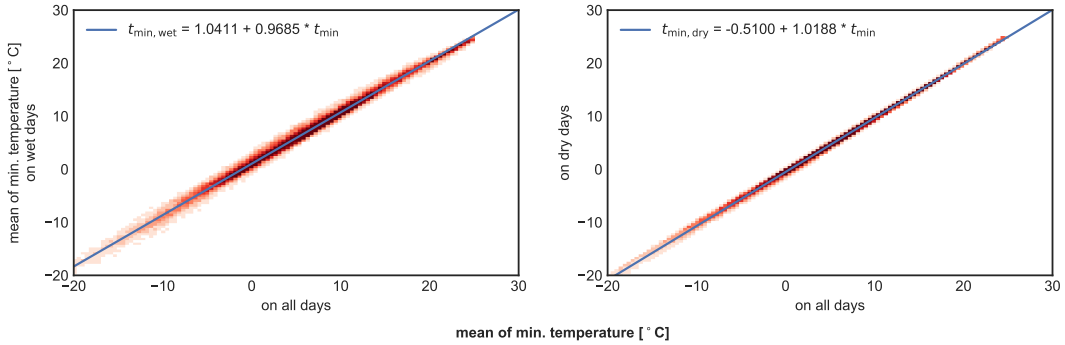


FIGURE 6.5: Correlation of minimum temperature on wet and dry days to the monthly mean. The y-axes show the mean minimum temperature on wet or dry days respectively, the blue line corresponds to the best fit line. Parameters of the fits are also shown in table 6.1.

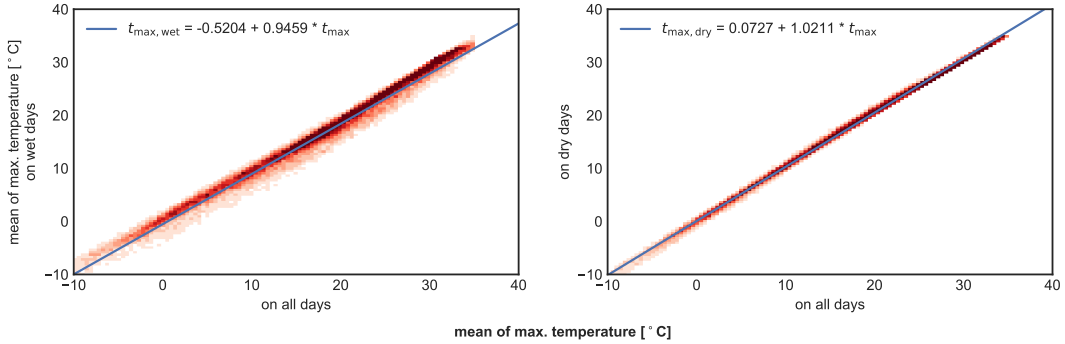


FIGURE 6.6: Correlation of maximum temperature on wet and dry days to the monthly mean. The y-axes show the mean maximum temperature on wet or dry days respectively, the blue line corresponds to the best fit line. Parameters of the fits are also shown in table 6.1.

As proposed by Geng et al., 1986, we use this relationship in our model to estimate the scale parameter of the distribution. Using this approach, the gamma shape parameter α is a constant, given via

$$\alpha = \frac{\bar{p}_d}{\theta} = \frac{1}{1.262}. \tag{6.8}$$

The GP scale parameter σ on the other hand is calculated during the simulation following Neykov et al., 2014 via

$$\sigma = \frac{1 - F(\mu)}{f(\mu)}. \tag{6.9}$$

The other parameters of the GP distribution are obtained through a sensitivity analysis described in section 6.3.5.

Temperature

Following the standard WGEN methodology (Richardson, 1981) and Geng et al., 1986, daily temperature is determined through 2 processes: First, the wet/dry state of the day, and second the cross correlation (section 6.3.2).

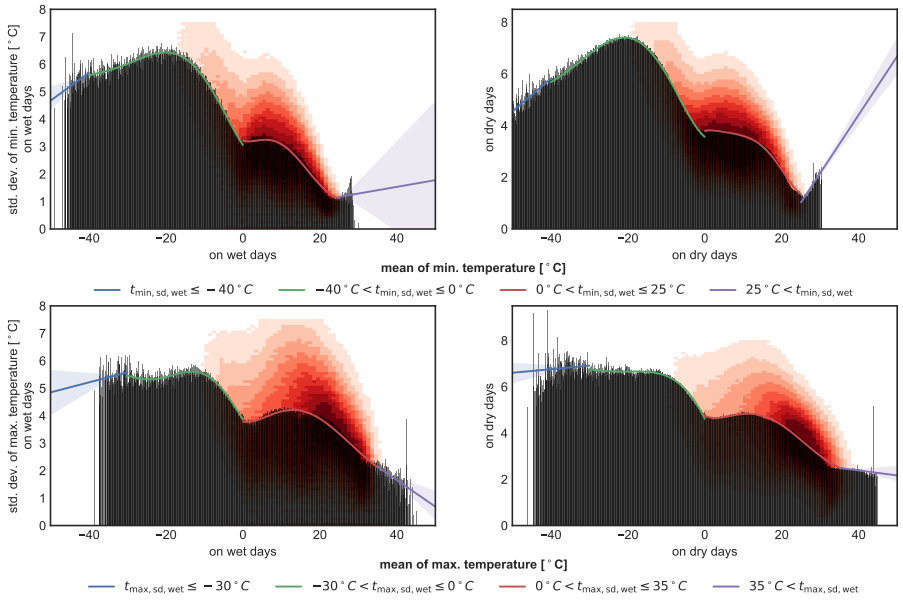


FIGURE 6.7: Correlation of standard deviation of the minimum and maximum temperature on wet and dry days to the monthly mean. The y-axes show the standard deviation, the x-axes the mean on wet or dry days respectively. The bars have a width of $0.1°C$ (the data accuracy) and indicate the mean standard deviation for a given mean minimum temperature in one month. The lines are fitted to these bars, where the green and red polynomials of order 5 are the use all the data below or above $0°C$ respectively and the blue and violet lines are a linear extrapolation of the data below $-40°C$ (or $-30°C$ for $T_{\max}$) or above $25°C$ (or $35°C$) respectively. The red density plot in the background indicates the spread of the data. The bars and the density plot are based on the single month for each station (i.e. not the multi-year monthly series as for, e.g. mean temperature (figure 6.5 and 6.6)). Parameters of the fits are also shown in table 6.1.

TABLE 6.1: Fit results of temperature correlation for wet and dry days for figures 6.5, 6.6, 6.10 and 6.11. The coefficients c_0 to c_3 correspond to the coefficients used in equations (6.10) and (6.14).

plot	variable	R^2	c_0	c_1	c_2	c_3
6.6	$T_{\text{max,dry}}$	0.9969	0.0727	1.0211	0	0
6.6	$T_{\text{max,wet}}$	0.9752	-0.5204	0.9459	0	0
6.5	$T_{\text{min,dry}}$	0.9972	-0.5100	1.0188	0	0
6.5	$T_{\text{min,wet}}$	0.9840	1.0411	0.9685	0	0
6.11	$w_{\text{sd,dry}}$	0.4243	0	1.0860	-0.2407	0.0222
6.11	$w_{\text{sd,wet}}$	0.5003	0	0.8184	-0.1263	0.0093
6.10	w_{dry}	0.9930	0	0.9437	0	0
6.10	w_{wet}	0.9723	0	1.0937	0	0

In the weather generator, we know from the Markov chain (section 6.3.2), whether the current simulated day is a wet or dry day. Based upon the simple linear relationships

$$\begin{aligned} \bar{x}_{\text{wet}} &= c_{0,x,\text{wet}} + c_{1,x,\text{wet}} \cdot \bar{x} \\ \bar{x}_{\text{dry}} &= c_{0,x,\text{dry}} + c_{1,x,\text{dry}} \cdot \bar{x} \end{aligned} \tag{6.10}$$

we adjust the monthly mean $\bar{x}$ of the variable $x \in \{T_{\text{min}}, T_{\text{max}}\}$.

To estimate the values of the parameters c_0 and c_1 in the above equations, we follow the same procedure as for the parameters of the Markov chain (section 6.3.2). We extracted the complete months for T_{min} and T_{max} from the GHCN-Daily dataset and created a multi-year series for each month and station. We then regressed the mean on wet and dry days separated against the overall mean of each month (Figures 6.5 and 6.6). Through this procedure, we estimate the parameters necessary for equations (6.10) (see table 6.1).

To estimate residual noise, we also need an estimate of the standard deviation of the variable (see Equation 6.3.2). Figure 6.7 shows the correlation between standard deviation on wet and dry days and the corresponding mean. The means of the standard deviations (black bars in figure 6.7) indicate a strong but non-linear relationship between the standard deviation and the corresponding mean. The correlation changes particularly at $0^\circ C$. We therefore use two different polynomials of order 5 for the values below and above the freezing point. Furthermore, to account for the sparse data below $-40^\circ C$ and above $25^\circ C$ for minimum temperature (or $-30^\circ C$ and $35^\circ C$ for maximum temperature), we use an extrapolation for the extremes as indicated by the blue and violet lines in figure 6.7. The formulae for the standard deviations σ of minimum and maximum temperature are therefore a combination of 4 polynomials:

TABLE 6.2: Fit results of the correlation of temperature standard deviation with the corresponding mean on wet/dry days for figure 6.7. The underlying equations are shown in equation (6.11).

variable	interval	R^2	c_0	c_1	c_2	c_3	c_4	c_5
$T_{\text{max,sd,dry}}$	(-∞, -30]	0.0125	7.3746	0.0154	0	0	0	0
	(-30, 0.0]	0.6721	4.6170	-0.3387	-0.0188	-0.0003	0.000003	0.0000001
	(0.0, 35]	0.9744	4.7455	-0.0761	0.0189	-0.0013	0.00003	-0.0000002
	(35, ∞)	0.0390	3.2554	-0.0218	0	0	0	0
$T_{\text{max,sd,wet}}$	(-∞, -30]	0.0366	6.6720	0.0364	0	0	0	0
	(-30, 0.0]	0.7362	3.8601	-0.2186	0.0039	0.0015	0.00006	0.0000007
	(0.0, 35]	0.9508	3.7919	-0.0313	0.0161	-0.0012	0.00003	-0.0000002
	(35, ∞)	0.2530	5.5529	-0.0973	0	0	0	0
$T_{\text{min,sd,dry}}$	(-∞, -40]	0.6006	10.8990	0.1271	0	0	0	0
	(-40, 0.0]	0.9509	3.5676	-0.1154	0.0282	0.0020	0.00004	0.0000003
	(0.0, 25]	0.9825	3.7941	0.0330	-0.0150	0.0019	-0.0001	0.000002
	(25, ∞)	0.7784	-4.6194	0.2261	0	0	0	0
$T_{\text{min,sd,wet}}$	(-∞, -40]	0.1661	9.7272	0.1011	0	0	0	0
	(-40, 0.0]	0.9285	3.0550	-0.2116	0.0137	0.0014	0.00004	0.0000003
	(0.0, 25]	0.9633	3.2187	-0.0451	0.0209	-0.0026	0.00010	-0.000001
	(25, ∞)	0.0089	0.5571	0.0244	0	0	0	0

$$\sigma_{T_{\text{min}},\text{wet/dry}} = \begin{cases} p_1(\bar{T}_{\text{min,wet/dry}}), & \text{for } \bar{T}_{\text{min,wet/dry}} \leq -40^\circ C \\ p_5(\bar{T}_{\text{min,wet/dry}}), & \text{for } -40^\circ C < \bar{T}_{\text{min,wet/dry}} \leq 0^\circ C \\ p_5(\bar{T}_{\text{min,wet/dry}}), & \text{for } 0^\circ C < \bar{T}_{\text{min,wet/dry}} \leq 25^\circ C \\ p_1(\bar{T}_{\text{min,wet/dry}}), & \text{for } 25^\circ C < \bar{T}_{\text{min,wet/dry}} \end{cases}$$

$$\sigma_{T_{\text{max}},\text{wet/dry}} = \begin{cases} p_1(\bar{T}_{\text{max,wet/dry}}), & \text{for } \bar{T}_{\text{max,wet/dry}} \leq -30^\circ C \\ p_5(\bar{T}_{\text{max,wet/dry}}), & \text{for } -30^\circ C < \bar{T}_{\text{max,wet/dry}} \leq 0^\circ C \\ p_5(\bar{T}_{\text{max,wet/dry}}), & \text{for } 0^\circ C < \bar{T}_{\text{max,wet/dry}} \leq 35^\circ C \\ p_1(\bar{T}_{\text{max,wet/dry}}), & \text{for } 35^\circ C < \bar{T}_{\text{max,wet/dry}} \end{cases}. \quad (6.11)$$

p_1 in eq. (6.11) denotes a polynomial of order 1, p_5 a polynomial of order 5. The coefficients of the different polynomials are shown in table 6.2.

These coefficients are based on the means of the standard deviation (black bars in figure 6.7). We chose this procedure to give the same weight to all temperatures. Otherwise the fit would be dominated by the temperature values around the freezing points.

Cloud fraction

Monthly mean cloud fraction is disaggregated, as for temperature, using the standard WGEN procedure of adding statistical noise to a wet- or dry-day mean and accounting for cross-correlation among the different weather variables. For the parameterization of the cloud fraction equations, we used the EECRA dataset. The original dataset contains eight measurements per day of the total cloud cover in

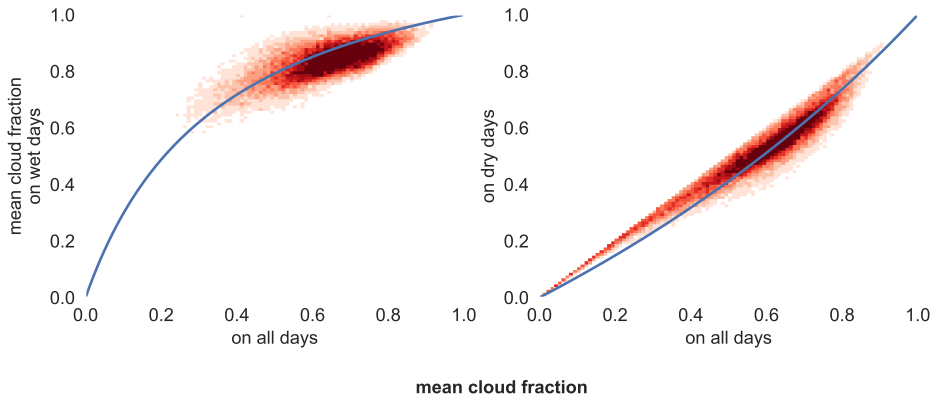


FIGURE 6.8: Correlation of cloud fraction on wet and dry days to the monthly mean.The y-axes show the mean cloud fraction on wet or dry days respectively, the blue line corresponds to the best fit line. Parameters of the fits are also shown in table 6.3.

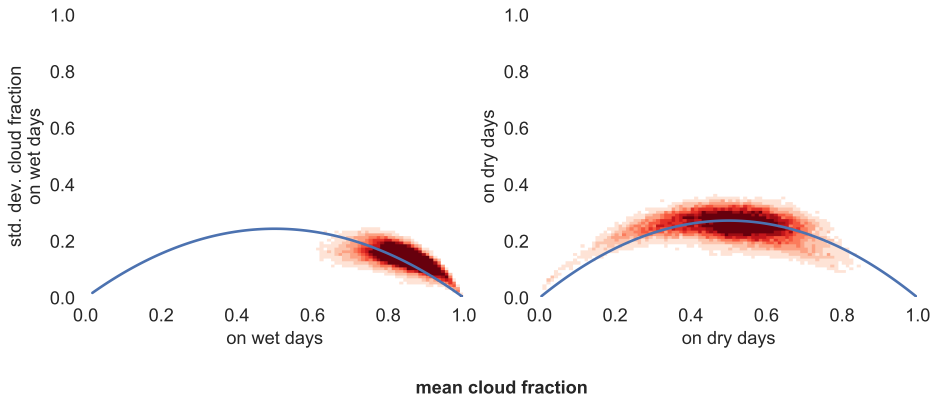


FIGURE 6.9: Correlation of standard deviation of the cloud fraction on wet and dry days to the corresponding monthly mean. The y-axes show the standard deviation, the x-axes the mean on wet or dry days respectively. The blue line corresponds to the best fit line. Parameters of the fits are also shown in table 6.3.

TABLE 6.3: Fit results of cloud correlation for wet and dry days for figure 6.8

plot	variable	a	std. err. of a	R^2
6.8	c_{dry}	0.4302	0.0013	0.8745
6.8	c_{wet}	-0.7376	0.0006	0.3881
6.9	$c_{\text{sd,dry}}$	1.0448	0.0004	0.2803
6.9	$c_{\text{sd,wet}}$	0.9881	0.0006	0.0802

units of octas, i.e., values ranging from 0 (clear sky) to 8 (overcast). Hence, to calculate the daily cloud fraction, those values were averaged and divided by 8 to produce a daily mean.

To adjust the monthly mean depending on the wet/dry state of the day, we could not use a simple linear relationship as we used for temperature because cloud fraction is bounded by a lower limit 0 and an upper limit of 1. Furthermore, we observed that cloud cover on wet days is usually greater or equal to the monthly mean cloud cover, whereas the cloud cover on dry days is usually less or equal to the monthly mean cloud cover. This results in a concave curve for the wet case and a convex curve for dry days. We used a qualitative graphical analysis to develop "best guess" equations that had the desired shape and propose the following formulae for the regression linking cloud cover on wet or dry days to the overall mean:

$$\begin{aligned}\bar{c}_{\text{wet}} &= \frac{-a_{\text{c,wet}} - 1}{a_{\text{c,wet}}^2 \cdot \bar{c} - a_{\text{c,wet}}^2 - a_{\text{c,wet}}} - \frac{1}{a_{\text{c,wet}}} \\ \bar{c}_{\text{dry}} &= \frac{-a_{\text{c,dry}} - 1}{a_{\text{c,dry}}^2 \cdot \bar{c} - a_{\text{c,dry}}^2 - a_{\text{c,dry}}} - \frac{1}{a_{\text{c,dry}}}\end{aligned} \tag{6.12}$$

with $a_{\text{c,wet}} < 0$ and $a_{\text{c,dry}} > 0$.

The standard deviation of cloud cover fraction becomes 0 when the mean monthly cloud fraction reaches both the minimum or maximum limits of 0 and 1. Hence, for $c_{\text{sd,dry}}$ and $c_{\text{sd,wet}}$ we have an concave parabola with the formula

$$\begin{aligned}\sigma_{\text{c,wet}} &= a_{\text{c,wet}}^2 \cdot \bar{c}_{\text{wet}} \cdot (1 - \bar{c}_{\text{wet}}) \\ \sigma_{\text{c,dry}} &= a_{\text{c,dry}}^2 \cdot \bar{c}_{\text{dry}} \cdot (1 - \bar{c}_{\text{dry}})\end{aligned} \tag{6.13}$$

with $a_{\text{c,wet}}, a_{\text{c,dry}} \geq 0$. Results of the fits can be seen in figure 6.8, 6.9 and the parameters in table 6.3.

Wind speed

The parameterization of the mean wind speed is based upon the same linear equation (6.10) as temperature. For the standard deviation however, we use a third-order polynomial given that is forced through the origin, given via

$$\begin{aligned}\sigma_{w,\text{wet}}(\bar{w}_{\text{wet}}) &= c_{1,w,\text{wet}}\, \bar{w}_{\text{wet}} + c_{2,w,\text{wet}}\, \bar{w}_{\text{wet}}^2 + c_{3,w,\text{wet}}\, \bar{w}_{\text{wet}}^3 \\ \sigma_{w,\text{dry}}(\bar{w}_{\text{dry}}) &= c_{1,w,\text{dry}}\, \bar{w}_{\text{dry}} + c_{2,w,\text{dry}}\, \bar{w}_{\text{dry}}^2 + c_{3,w,\text{dry}}\, \bar{w}_{\text{dry}}^3.\end{aligned} \tag{6.14}$$

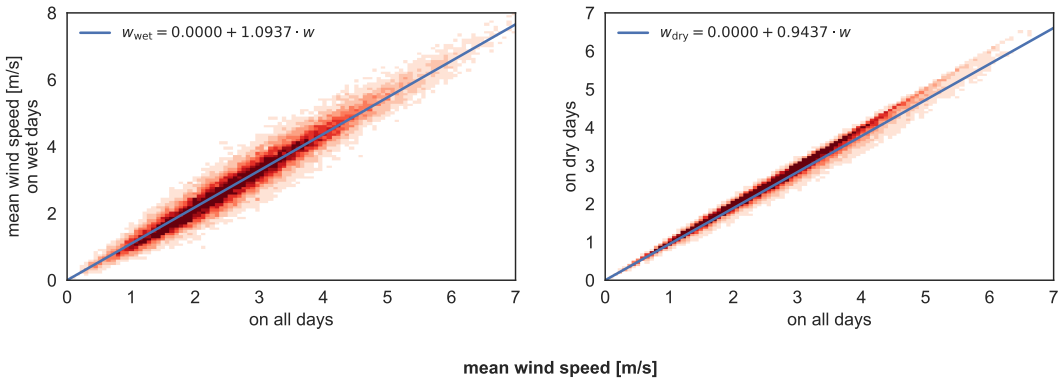


FIGURE 6.10: Correlation of wind speed on wet and dry days to the monthly mean.The y-axes show the mean cloud fraction on wet or dry days respectively, the blue line corresponds to the best fit line. Parameters of the fits are also shown in table 6.1.

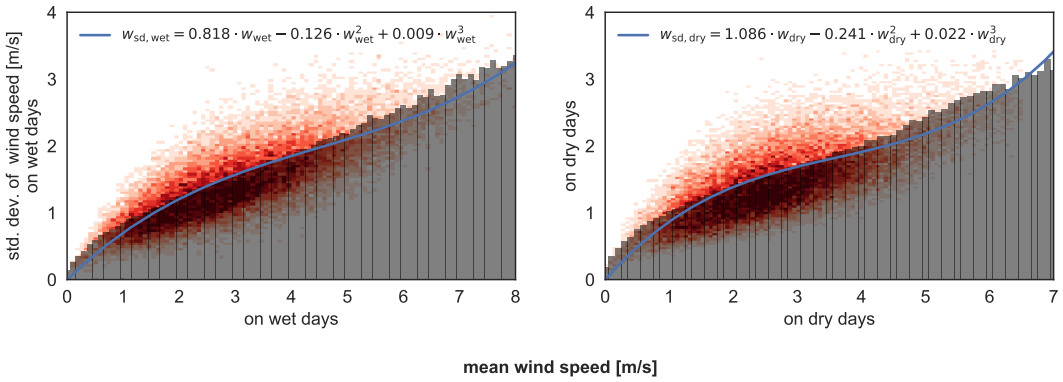


FIGURE 6.11: Correlation of standard deviation of wind speed on wet and dry days to the corresponding monthly mean. The y-axes show the standard deviation, the x-axes the mean on wet or dry days respectively. The blue line corresponds to the best fit line, a third order polynomial to the underlying red density plot. The black bars have a width of 0.1 $\mathrm{m\,s^{-1}}$, the accuracy of the input data, and indicate the mean standard deviations for the given interval range. Parameters of the fits are also shown in table 6.1.

This better resolves the complex behavior close to $0\,\mathrm{m\,s^{-1}}$ compared to a linear fit. The plots are shown in the figures 6.10 and 6.11 and the parameters for the fits are shown in table 6.1.

Cross correlation

Following Richardson, 1981 we use cross correlation to add additional residual noise to the simulated meteorological variables, which provides more realism in the daily weather result. This method, based on Matalas, 1967 preserves the serial and the cross correlation between the simulated variables. It implies that the serial correlation of each variable may be described by a first-order linear autoregressive model

Given the cross correlation matrix $M_0 \in \mathbb{R}^4 \times \mathbb{R}^4$ and the lag-1 correlation matrix $M_1 \in \mathbb{R}^4 \times \mathbb{R}^4$, we calculate

$$A = M_1 M_0^{-1} \qquad BB^T = M_0 - M_1 M_0^{-1} M_1^T. \tag{6.15}$$

The matrices A, B, M_0 and M_1 are calculated using the stations from the EECRA database in figure 6.2. The results are

$$M_0 = \begin{pmatrix} 1. & 0.565 & 0.041 & 0.035 \\ 0.565 & 1. & -0.089 & -0.043 \\ 0.041 & -0.089 & 1. & 0.114 \\ 0.035 & -0.043 & 0.114 & 1. \end{pmatrix}$$
$$M_1 = \begin{pmatrix} 0.933 & 0.55 & 0.016 & 0.03 \\ 0.557 & 0.417 & -0.066 & -0.043 \\ 0.004 & -0.095 & 0.599 & 0.093 \\ 0.011 & -0.063 & 0.061 & 0.672 \end{pmatrix}. \tag{6.16}$$

leading to

$$A = \begin{pmatrix} 0.916 & 0.031 & -0.018 & 0.001 \\ 0.485 & 0.135 & -0.069 & -0.047 \\ 0.004 & -0.043 & 0.592 & 0.023 \\ 0.012 & -0.043 & -0.02 & 0.672 \end{pmatrix} \qquad B = \begin{pmatrix} 0.358 & 0. & 0. & 0. \\ 0.112 & 0.809 & 0. & 0. \\ 0.142 & -0.06 & 0.785 & 0. \\ 0.077 & -0.016 & 0.061 & 0.733 \end{pmatrix}. \tag{6.17}$$

The columns and rows in the two matrices correspond to min. and max. temperature, cloud fraction and square root of wind speed, respectively.

In the weather generator, the variables $T_{\min}, T_{\max}, c$ and w are then calculated using a combination of residual noise χ_i (where i denotes the current simulated day) and the mean of the variables. χ_i is determined by the other variables and the previous day using A and B from above (Matalas, 1967; Richardson, 1981). Hence, χ_i is given via

$$\chi_i = \begin{pmatrix} \chi_{T_{\min}} & \chi_{T_{\max}} & \chi_c & \chi_w \end{pmatrix} = A\chi_{i-1} + B\epsilon \in \mathbb{R}^4. \tag{6.18}$$

The daily values for the variables are then calculated via

$$T_{\min,i} = \chi_{T_{\min}} \cdot \sigma_{T_{\min},\mathrm{wet/dry}} + \bar{T}_{\min,\mathrm{wet/dry}} \qquad c_i = \chi_c \cdot \sigma_{c,\mathrm{wet/dry}} + \bar{c}_{\mathrm{wet/dry}} \tag{6.19}$$

$$T_{\max,i} = \chi_{T_{\max}} \cdot \sigma_{T_{\max},\mathrm{wet/dry}} + \bar{T}_{\max,\mathrm{wet/dry}} \qquad w_i = \left(\chi_w \cdot \sqrt{\sigma_{w,\mathrm{wet/dry}}} + \sqrt{\bar{w}_{\mathrm{wet/dry}}}\right)^2 \tag{6.20}$$

with $\sigma_{T_{\min},\text{wet/dry}}, \sigma_{T_{\max},\text{wet/dry}}$ from eq. (6.11), $\sigma_{c,\text{wet/dry}}$ from eq. (6.13), $\sigma_{w,\text{wet/dry}}$ from eq. (6.14), $\bar{T}_{\min,\text{wet/dry}}, \bar{T}_{\max,\text{wet/dry}}, \bar{w}_{\text{wet/dry}}$ from eq. (6.10) and $\bar{c}_{\text{wet/dry}}$ from eq. (6.12).

Since this procedure always requires the residuals from the previous day, χ_{i-1}, we initialize χ_0 with 0, simulate the month and then simulate it again.

Note that, through the entire procedure, wind speed is subject to a square-root transformation (also when calculating M_0 and M_1) to account for the fact that it is not normally distributed.

6.3.3 Model Evaluation

To evaluate GWGEN, we started with the daily meteorology at the evaluation stations described above and calculated monthly summaries. We used this monthly data to drive the model and simulate daily meteorology. The resulting daily series now has the same length as the observed meteorology from the GHCN and EECRA database. Because we cannot expect the weather generator to reproduce the weather exactly as observed, for example the number of rainy days in a month may be the same as observed but they may not occur in precisely the same order, our evaluation is restricted to comparing the statistical properties of the input observed versus the output simulated daily meteorology.

Figure 6.12 shows the comparison of simulated versus observed values for each of the five meteorological variables handled by GWGEN. For temperature, wind, and cloud fraction, the model does an excellent job of downscaling monthly input to daily resolution[1]. The comparison between precipitation amounts looks good when considering all of the data, however a closer look into the results (Fig. 6.13) shows that while the higher precipitation percentiles are well captured using the hybrid Gamma-GP distribution, the lower percentiles show somewhat worse results. This observation of poor performance for very low values also holds true for wind speed (not shown here). The lower values of the two variables, however, are very close to the precision of the observation (0.1 mm for precipitation and $0.1\,\text{m}\,\text{s}^{-1}$ for wind speed). Very small precipitation amounts and low wind speeds are also less biophysically and ecologically important compared to the higher percentiles. We therefore consider the results of the evaluation largely acceptable.

In table 6.4 we also compare the simulated versus the observed frequencies. For very light rain (<=1mm), light rain (1-10mm), heavy rain (10-20mm) and very heavy rain (>20mm). As we can see, our model underestimates the occurrence of very light rain events (28.6% instead of 36.4%) and overestimates the light rain events (58.3% instead of 48.6%) but generally performs much better than GCMs (Dai, 2006; Sun et al., 2006), especially when it comes to the heavy rain events.

6.3.4 Bias correction

After evaluating the results of GWGEN for wind speed for the different quantiles (see previous subsection 6.3.3) we found a strong, systematic bias between the simulated and the observed values. This observation led us to adopt a further measure to improve the quality of the model output by implementing a quantile-based bias correction.

We use an empirical distribution correction approach (quantile-mapping) (Lafon et al., 2012) to a posteriori correct the simulated data. In the quantile evaluation (previous subsection 6.3.3) we saw that the simulated wind speed is a linear function of

[1]Note that the plot for wind speed has been bias corrected using the approach in subsection 6.3.4.

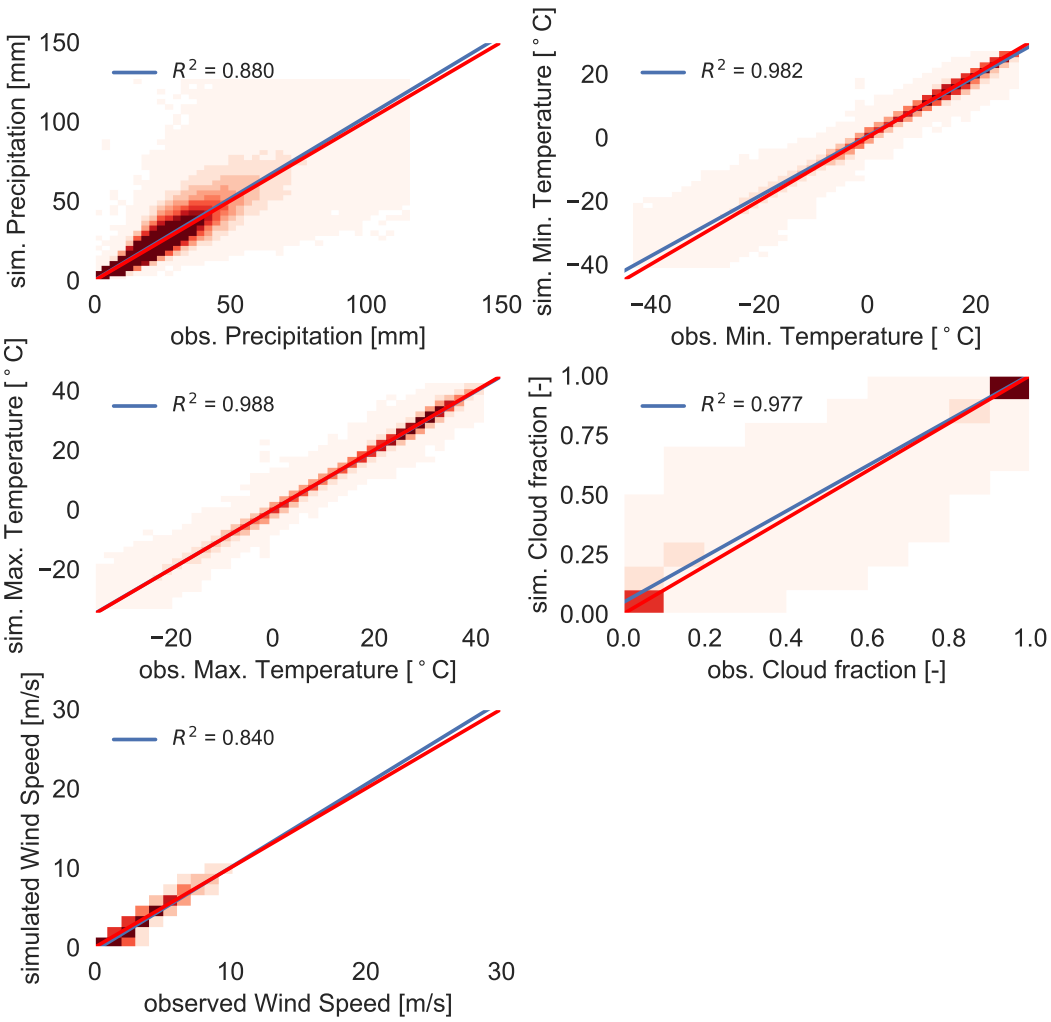


FIGURE 6.12: QQ-plots for all variables with all quantiles (1, 5, 10, 25, 50, 75, 90, 95 and 99) for $\mu = 5.0\,\text{mm}\,\text{mm}$, $\xi = 1.5$. The blue lines are linear regression from simulation to observation. The red line shows the ideal fit (the identity line). Blue shaded areas represent the 95% confidence interval. The plots compares the simulated quantile from the list above of one year of one station to the corresponding observed quantile of the same year and station. The plot for wind speed underwent used the bias correction from subsection 6.3.4.

TABLE 6.4: Simulated and observed precipitation frequencies for certain ranges. The frequency is defined as the number of precipitation occurences in the specified range, divided by the total number of precipitation occurences.

Precip. range [mm]	Simulated	Observed
(0, 1]	0.285688	0.364014
(1, 10]	0.583330	0.486415
(10, 20]	0.074063	0.090178
(20, ∞]	0.056920	0.059392

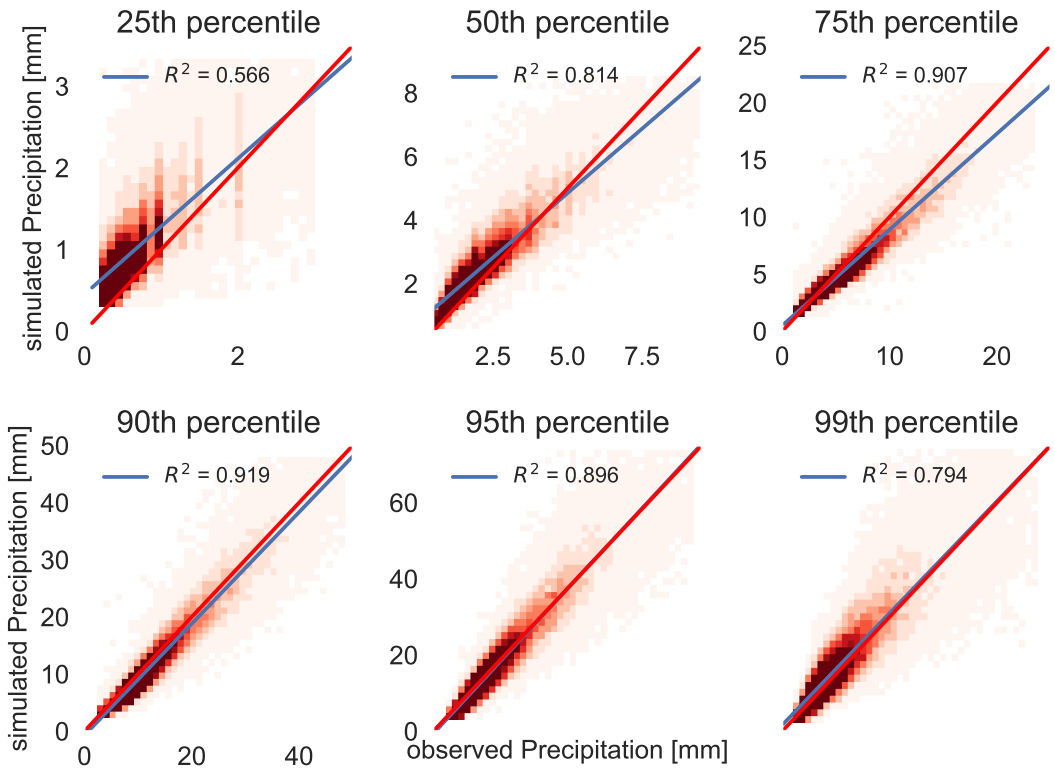


FIGURE 6.13: QQ-plot for different quantiles for precipitation for $\mu = 5.0$mm, $\xi = 1.5$. The blue lines are linear regression from simulation to observation. The red line shows the ideal fit (the identity line). Blue shaded areas represent the 95% confidence interval. The plots compare the simulated quantile of one year of one station to the corresponding observed quantile of the same year and station.

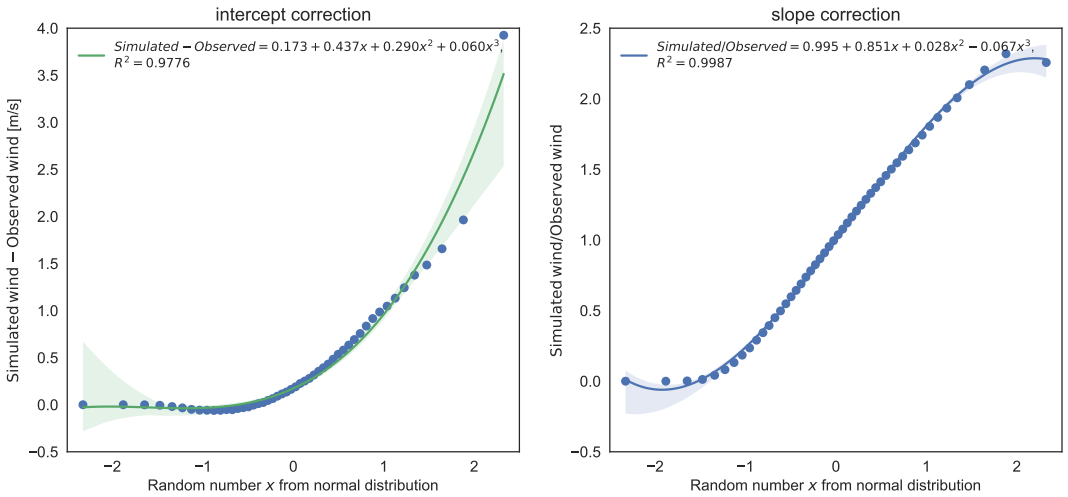


FIGURE 6.14: Basis for the wind bias correction. For the left plot, each data point corresponds to the difference of a simulated percentile to the observed percentile. For the right plot (wind speed), each data point corresponds to the fraction of simulated to the observed wind speed for a given percentile. The random number on the x-axis represents the residual value from a normal distribution centered at 0 with standard deviation of unity, as it is used in the cross correlation approach (Richardson, 1981).

the observed wind speed, i.e. $w_{sim} = \text{intercept} + \text{slope} \cdot w_{obs}$ (best fit line in figure 6.12). Therefore, we use two steps here, one is for the difference between simulation and observation (ideally 0), the other one is the fraction of observation and simulation (ideally 1). The first one corresponds to the intercept with the y-axis in figure 6.12, the second one to the slope of the best fit line. The analysis is based on every second percentile between 1 and 100 (i.e. $1, 3, 5, \ldots$) and mapped to its corresponding random number $u \in \mathbb{R}$ from a normal distribution as it is used for the cross correlation in the weather generator (section 6.3.2, x-axis in figure 6.14 and Richardson, 1981).

Regarding the intercept (fig. 6.14, left) we see that it strongly follows an exponential function given through

$$f_{exp}(u) = \mathrm{e}^{a\,u+b}, \qquad a, b, u \in \mathbb{R}. \tag{6.21}$$

The slope (fig. 6.14, right) on the other hand can be described by a simple third-order polynomial given by

$$p3(u) = c_0 + c_1\,u + c_2\,u^2 + c_3\,u^3, \qquad c_0, c_1, c_2, c3, u \in \mathbb{R} \tag{6.22}$$

Hence, given the best fit lines in figure 6.14, the simulated wind speed is corrected via

$$w'_{sim} = \frac{w_{sim} - f_{exp}(u)}{p3(u)} \tag{6.23}$$

with $a = 1.1582, b = -1.3359, c_0 = 0.9954, c_1 = 0.8508, c_2 = 0.0278, c_3 = -0.0671$.

6.3.5 Sensitivity analysis

The Generalized-Pareto part of the hybrid Gamma-GP distribution, which we used to simulate precipitation amount, has two parameters: the GP shape, and the threshold parameter. Unlike the gamma parameters, we were unable to relate these GP parameters to any of the monthly summary data we use as input to GWGEN. Hence, we decided to set fixed values for these parameters, and determine them through a sensitivity analysis.

To select the "best" values of the GP parameters, we compared simulated with observed precipitation amounts, running GWGEN with a wide range of realistic parameter values. To quantitatively assess the model performance, we used two metrics: 1) direct comparison of the quantiles (see previous section), and 2) a Kolmogorov-Smirnov (KS) test that evaluates whether two data samples come from significantly different distributions. Our criteria were

1. The R^2 correlation coefficient between simulated and observed quantiles
2. The fraction $\frac{\text{simulated precipitation}}{\text{observed precipitation}}$ from the slopes in figure 6.13 and its deviation from unity
3. the fraction of simulated (station specific) years that are significantly different (KS test) from the observation
4. The mean of the above values

We tried two different approaches to select the gamma-GP crossover threshold: first we tried a fixed crossover point, second we used a quantile-based crossover point. For the latter, the model chooses to use the GP distribution if the quantile of

the random number drawn from the gamma distribution is above a certain quantile threshold. This introduces a flexible crossover point in our hybrid distribution which, however, did not improve the results significantly. We therefore show here only the results using the fixed crossover point.

The values of the crossover point for our sensitivity analysis were 2, 2.5, 3, 4 and from 5 to 20 in steps of 2.5 and 20 to 100 in steps of 5. Furthermore we varied the GP shape parameter from 0.1 to 3 in steps of 0.1 (810 experiments in total). The results of this sensitivity analysis are shown in the supplementary material, figure 6.15.

In general we found that the three criteria 1, 2 and 3 could not be optimized all together at the same time. The R^2 is best for high thresholds and low GP shape parameters, the slope is best for low to intermediate thresholds and a low GP shape and the KS statistic is best for low threshold and intermediate GP shape parameters.

However, R^2 did not vary that much (from 0.68 to 0.74) and from a visual evaluation of the corresponding quantile plots we saw that the higher quantiles (>90) were much better represented for a better KS result. Hence, we chose to follow the KS test criteria, which is also the strictest of our evaluation methods but again compared the different quantile plots to get good results for the higher quantiles. Finally, we chose a threshold of 5 mm and a GP shape parameter of 1.5. For this setting, 81.7% of the simulated years do not show a significant difference compared to the observation, the mean R^2 of the plots in figure 6.13 is 0.81 and the mean deviation of the slope from unity is 0.10 and for the upper quantiles (90 to 100), 0.017.

Nevertheless, in total the results seem to be fairly independent of the two parameters since even the amount of years without significant differences vary from 73% to only 83%. It is however better than the gamma distribution alone which still has 78.6% of station years not differing significantly but with a slope deviation from unity for the upper quantiles of 0.16. Thus using the hybrid Gamma-GP distribution improves the simulation of high-amount precipitation events by roughly factor 10 compared to a standard Gamma approach.

6.4 Limitations

As demonstrated above, GWGEN successfully downscales monthly to daily meteorology with good correlation and low bias when compared to observations. However, there are a few limitations of the model as currently described that should be noted. Importantly, this version of GWGEN neither downscales all conceivable meteorological variables, nor does it provide a mechanism for generating daily meteorological timeseries across multiple points that are spatially autocorrelated. Concerning the former point, while GWGEN simulates daily precipitation, temperature, cloud cover, and windspeed, it does not currently handle other variables that might be important in land surface modeling, such as humidity or wind direction. On the latter point, the lack of explicit simulation of spatial autocorrelation may make GWGEN unsuitable for certain applications, e.g., regional high-resolution hydrological modeling in small catchments (< ca. 2500 km^2), where having the capability to simulate flood and other extremes is important. This is because the the weather generator could, e.g., simulate rainfall on different days in different parts of the catchment, where in reality storm events would be highly autocorrelated in space and controlled by mesoscale meteorological conditions.

6.5 Discussion and Outlook

GWGEN successfully downscales monthly to daily meteorology, for any point on the globe, in any climate, in any season, and in any time in recent earth history and into the near future (e.g., next century). It extends the original Richardson-type weather generators to simulate wind speed along with precipitation, temperature, and cloud cover. The model requires only monthly values of the meteorological variables to be downscaled, and does not rely on any other spatial information, e.g., whether or not the location is in the tropics.

In general, the results of our downscaled meteorology are excellent, with all simulated variables showing both very high correlation and limited bias when compared to observations. We improved the simulation of daily precipitation amount by replacing the Gamma distribution used in the original Richardson-type weather generators with a hybrid Gamma-GP distribution, which results in the improved simulation of heavy precipitation events. The GP distribution is based upon a globally fixed shape and location parameter, which may be an oversimplification, but is still ten times more accurate than traditional methods that used Gamma alone. Our extensive sensitivity analysis to determine the best coefficients for the shape and location parameters of the GP distribution suggest that further improvements might come through correlating the GP parameters to geographic region and/or seasonality (Maraun et al., 2009; Rust et al., 2009) or by introducing a dynamical location parameter (Frigessi et al., 2002). Finally, we introduced a step to correct for systematic bias in the downscaling of temperature and wind speed.

Despite the limitations noted above, GWGEN will be useful in a wide range of applications, from global vegetation and crop modeling, to large-scale hydrologic analyses, to understanding animal behavior, to forecasting of fire, insect outbreaks, and other ecosystem disturbances. GWGEN may even be envisaged as a potential replacement for very large and cumbersome gridded datasets of high-temporal resolution meteorology such as CRUNCEP (Viovy and Ciais, 2016), especially for models that use meteorological forcing at a daily timestep. The weather generator is particularly suited for the incorporation into models that run on a spatial grid, for example, GWGEN can readily be incorporated into existing DGVMs such as LPJ-LMfire (Pfeiffer et al., 2013) or LPJ-ML (Bondeau et al., 2007) that already rely on a weather generator to provide daily meteorology for certain processes.

While GWGEN does not handle spatial autocorrelation, in most DGVMs there is no lateral connection between gridcells, and therefore an explicit representation of spatial autocorrelation in the driving daily meteorological data would have no effect on the model output. We further note that if the monthly data used to drive the model are spatially autocorrelated — this would be the case when using gridded climatology for example — then the result of the weather generator will also preserve this autocorrelation, at least when the model results are analyzed on monthly or longer timescales.

The limitations present in this version of GWGEN could be addressed in future versions. Methods for simultaneous multisite weather generation exist (Wilks, 1998, 1999b,c) and could be adapted to GWGEN. However, even simpler methods to approximate spatial autocorrelation could be possible. Running GWGEN with gridded monthly meteorology — this is the primary application we foresee for the current version of the model — means that the input variables are already highly correlated in space, i.e., the monthly climate in one gridcell generally closely resembles neighboring cells, outside of complex terrain containing sharp, monotonic climate

gradients, e.g., rain shadows. Thus, one simple way of achieving a measure of spatial autocorrelation in GWGEN would be to impose a spatial autocorrelation field on the sequence of random numbers used to impose stochastic noise in the downscaling functions. If the random number sequence is similar between gridcells, then, e.g., rain is likely to fall on the same day, given that the transition probabilities will likely also be similar. Over moderate distances, e.g., <50's of km, it might even be sufficient to use the same random seed across all gridcells in a neighborhood. This would have the effect of producing strongly autocorrelated daily meteorology in space, with the only variations being imposed by the underlying input monthly climatology.

Furthermore, it would be straightforward to include additional meteorological variables in the model framework, handling, e.g., humidity in the same way that temperatures, cloud cover, and wind speed are disaggregated. Other variables, such as pressure and wind direction, might be more difficult using the basic GWGEN structure because of the importance of autocorrelation, particularly at high spatial resolution, and might benefit from a different approach towards weather generation. Finally, GWGEN only downscales meteorology from monthly to daily values; for models that require an even shorter timestep, e.g., 6-hourly, some extension of the model functionality would be required. For certain variables, e.g., temperatures, sub-daily downscaling could be easily implemented (Cesaraccio et al., 2001), for other variables, such as precipitation, a large literature on downscaling methods exists (e.g. Bennett et al., 2016), and global datasets of hourly meteorology for model calibration are available (e.g., the Integrated Surface Database, Smith et al., 2011).

6.6 Conclusions

Compiling a global database of daily precipitation, temperature, cloud cover, and wind speed measurements, we explored the relationship between daily meteorology and monthly summaries first described in the context of weather downscaling by Geng and Auburn, 1987. Our analysis of more than 50 million individual records showed that daily-to-monthly relationships are relatively stable in space and time, and constant across a very wide range of stations from all latitudes and climate zones. With the resulting relationships, we parameterized a WGEN/SIMMETEO-type weather generator, with the intention of creating a generic scheme that could be applied anywhere over the earth's land surface for the past, present, and (near) future.

6.7 Code availability

GWGEN, is open source software, and the code, utility programs for parameterization, evaluation and manipulating the raw weather station data, and complete documentation are available at (Sommer and Kaplan, 2017b). The original weather station database can be made available upon request to the authors or downloaded from Hahn and Warren, 1999 and Menne et al., 2012b. The weather generator module is programmed in FORTRAN, the parameterization, evaluation and other supplementary tools are written in Python mainly using the numerical python libraries numpy and scipy (Jones et al., 2001), statsmodels (Seabold and Perktold, 2010), as well as matplotlib (Hunter, 2007) and psyplot (Sommer, 2017) for the visualization. Detailed installation instructions can be found in the user manual: `https://arve-research.github.io/gwgen/`.

6.A Supplementary material

6.A.1 Sensitivity analysis

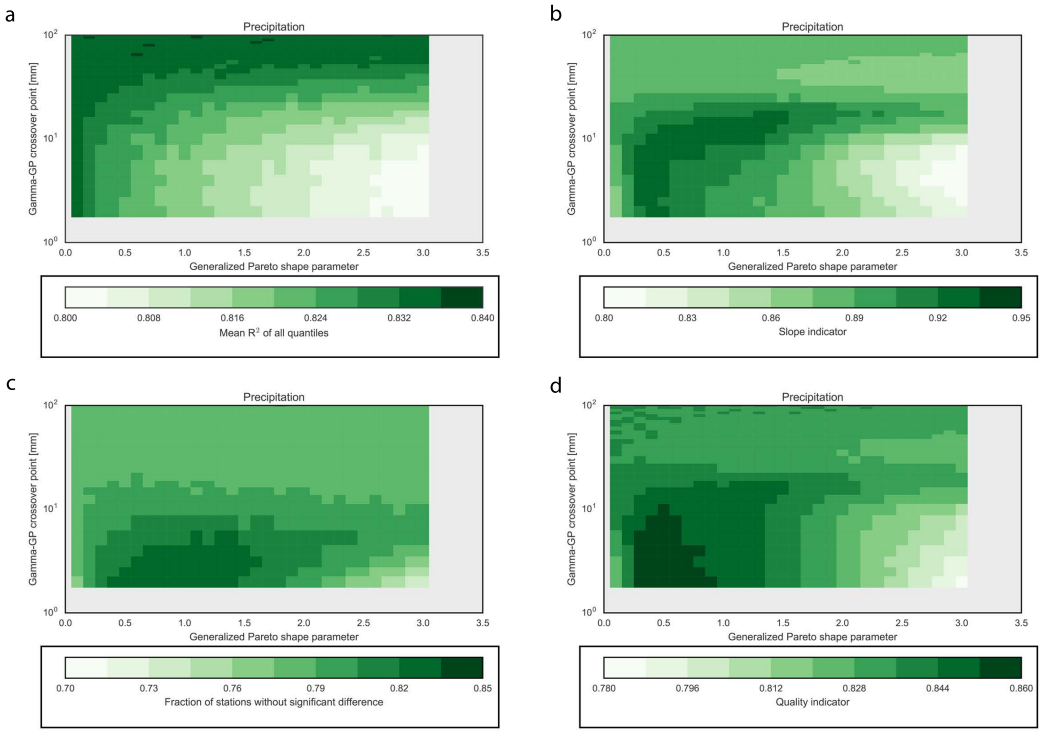


FIGURE 6.15: Results of the sensitivity analysis for the (a) correlation coefficient R^2, (b) deviation from a slope of unity, (c) the fraction of significant different station years, (d) the mean of (a) - (c). For the plots in (a) and (b) we used the means of the 25th, 50th, 75th, 90th, 95th and 99th percentiles. In general, 1 (dark green) is best, 0 (white) is worst. The dark red fields indicate experiments that failed because of a too low threshold and too high GP shape parameter. Note the logarithmic scale on the y-axis.

Author contributions. JOK conceived the model and analyses, wrote the prototype code and performed preliminary analyses, PS developed and documented the final version of the code (including parameterization and evaluation), performed all of the final analyses, and created the graphical output. Both authors contributed to the writing of the manuscript

Acknowledgements. This work was supported by the European Research Council (COEVOLVE, 313797) and the Swiss National Science Foundation (ACACIA, CR10I2_146314). We thank Shawn Koppenhoefer for assistance compiling and querying the weather databases and Alexis Berne and Grégoire Mariéthoz for helpful suggestions on the analyses. We are grateful to NOAA NCDC and the University of Washington for providing free of charge the GHCN-Daily and EECRA databases, respectively.

References

Bennett, James C., David E. Robertson, Phillip G.D. Ward, H.A. Prasantha Hapuarachchi, and Q.J. Wang (2016). "Calibrating hourly rainfall-runoff models with

daily forcings for streamflow forecasting applications in meso-scale catchments". In: *Environmental Modelling & Software* 76, pp. 20–36. ISSN: 1364-8152. DOI: http://dx.doi.org/10.1016/j.envsoft.2015.11.006. URL: http://www.sciencedirect.com/science/article/pii/S1364815215300979.

Bhatt, Samir, Peter W. Gething, Oliver J. Brady, Jane P. Messina, Andrew W. Farlow, Catherine L. Moyes, John M. Drake, John S. Brownstein, Anne G. Hoen, Osman Sankoh, Monica F. Myers, Dylan B. George, Thomas Jaenisch, G. R. William Wint, Cameron P. Simmons, Thomas W. Scott, Jeremy J. Farrar, and Simon I. Hay (Apr. 2013). "The global distribution and burden of dengue". In: *Nature* 496.7446, pp. 504–507. DOI: 10.1038/nature12060. URL: http://dx.doi.org/10.1038/nature12060.

Bondeau, Alberte, Pascalle C. Smith, Sönke Zaehle, Sibyll Schaphoff, Wolfgang Lucht, Wolfgang Cramer, Dieter Gerten, Hermann Lotze-Campen, Christoph Müller, Markus Reichstein, and Benjamin Smith (2007). "Modelling the role of agriculture for the 20th century global terrestrial carbon balance". In: *Global Change Biol.* 13.3, pp. 679–706. ISSN: 1354-1013 1365-2486. DOI: 10.1111/j.1365-2486.2006.01305.x.

Cesaraccio, C., D. Spano, P. Duce, and R. L. Snyder (2001). "An improved model for determining degree-day values from daily temperature data". In: *Int. J. Biometeorol.* 45.4, pp. 161–9. ISSN: 0020-7128 (Print) 0020-7128 (Linking). DOI: 10.1007/s004840100104. URL: http://www.ncbi.nlm.nih.gov/pubmed/11769315.

Dai, Aiguo (Sept. 2006). "Precipitation Characteristics in Eighteen Coupled Climate Models". In: *J. Climate* 19.18, pp. 4605–4630. DOI: 10.1175/jcli3884.1. URL: http://dx.doi.org/10.1175/JCLI3884.1.

Elith, Jane, Catherine H. Graham, Robert P. Anderson, Miroslav Dudík, Simon Ferrier, Antoine Guisan, Robert J. Hijmans, Falk Huettmann, John R. Leathwick, Anthony Lehmann, Jin Li, Lucia G. Lohmann, Bette A. Loiselle, Glenn Manion, Craig Moritz, Miguel Nakamura, Yoshinori Nakazawa, Jacob McC. M. Overton, A. Townsend Peterson, Steven J. Phillips, Karen Richardson, Ricardo Scachetti-Pereira, Robert E. Schapire, Jorge Soberón, Stephen Williams, Mary S. Wisz, and Niklaus E. Zimmermann (2006). "Novel methods improve prediction of species' distributions from occurrence data". In: *Ecography* 29.2, pp. 129–151. ISSN: 1600-0587. DOI: 10.1111/j.2006.0906-7590.04596.x. URL: http://dx.doi.org/10.1111/j.2006.0906-7590.04596.x.

Friend, A. D. (1998). "Parameterisation of a global daily weather generator for terrestrial ecosystem modelling". In: *Ecol. Modell.* 109.2, pp. 121–140. ISSN: 0304-3800. DOI: Doi10.1016/S0304-3800(98)00036-2.

Frigessi, Arnoldo, Ola Haug, and Håvard Rue (2002). "A Dynamic Mixture Model for Unsupervised Tail Estimation without Threshold Selection". In: *Extremes* 5.3, pp. 219–235. ISSN: 1572-915X. DOI: 10.1023/A:1024072610684. URL: http://dx.doi.org/10.1023/A:1024072610684.

Furrer, Eva M. and Richard W. Katz (2008). "Improving the simulation of extreme precipitation events by stochastic weather generators". In: *Water Resour. Res.* 44.12, n/a–n/a. ISSN: 00431397. DOI: 10.1029/2008wr007316.

Geng, S., F. W. T. P. Devries, and I. Supit (1986). "A Simple Method for Generating Daily Rainfall Data". In: *Agric. For. Meteorol.* 36.4, pp. 363–376. ISSN: 0168-1923. DOI: 10.1016/0168-1923(86)90014-6.

Geng, Shu and J. S. Auburn (1987). "Weather simulation models based on summaries of long-term data". In: *Weather and Rice: Proceedings of the international workshop on the Impact of Weather Parameters on Growth and Yield of Rice, 7-10 Apr 1986*. Ed. by

International Rice Research Institute. Los Baños, Philippines: International Rice Research Institute, pp. 237–254.

Gerten, Dieter, Sibyll Schaphoff, Uwe Haberlandt, Wolfgang Lucht, and Stephen Sitch (2004). "Terrestrial vegetation and water balance—hydrological evaluation of a dynamic global vegetation model". In: *J. Hydrol.* 286.1-4, pp. 249–270. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2003.09.029.

Gordon, H. A. (1981). "Errors in Computer Packages. Least Squares Regression Through the Origin". In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 30.1, pp. 23–29. ISSN: 00390526, 14679884. URL: http://www.jstor.org/stable/2987701.

Guenther, A., C. N. Hewitt, D. Erickson, R. Fall, C. Geron, T. Graedel, P. Harley, L. Klinger, M. Lerdau, W. A. Mckay, T. Pierce, B. Scholes, R. Steinbrecher, R. Tallamraju, J. Taylor, and P. Zimmerman (1995). "A Global-Model of Natural Volatile Organic-Compound Emissions". In: *Journal of Geophysical Research-Atmospheres* 100.D5, pp. 8873–8892. ISSN: 2169-897x. DOI: Doi10.1029/94jd02950.

Hahn, C.J. and S.G. Warren (1999). "Extended Edited Synoptic Cloud Reports from Ships and Land Stations Over the Globe, 1952-1996 (with Ship data updated through 2008)". In: DOI: 10.3334/CDIAC/cli.ndp026c. URL: http://dx.doi.org/10.3334/CDIAC/cli.ndp026c.

Harris, I., P. D. Jones, T. J. Osborn, and D. H. Lister (2014). "Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset". In: *Int. J. Climatol.* 34.3, pp. 623–642. ISSN: 08998418. DOI: 10.1002/joc.3711.

Haxeltine, A. and I. C. Prentice (1996). "BIOME3: An equilibrium terrestrial biosphere model based on ecophysiological constraints, resource availability, and competition among plant functional types". In: *Global Biogeochem. Cycles* 10.4, pp. 693–709. ISSN: 0886-6236. DOI: Doi10.1029/96gb02344.

Haxeltine, Alex, I. Colin Prentice, and Ian David Creswell (1996). "A coupled carbon and water flux model to predict vegetation structure". In: *J. Veg. Sci.* 7.5, pp. 651–666. ISSN: 1654-1103. DOI: 10.2307/3236377. URL: http://dx.doi.org/10.2307/3236377.

Hijmans, Robert J., Susan E. Cameron, Juan L. Parra, Peter G. Jones, and Andy Jarvis (2005). "Very high resolution interpolated climate surfaces for global land areas". In: *Int. J. Climatol.* 25.15, pp. 1965–1978. ISSN: 0899-8418 1097-0088. DOI: 10.1002/joc.1276.

Hunter, J. D. (May 2007). "Matplotlib: A 2D Graphics Environment". In: *Computing in Science Engineering* 9.3, pp. 90–95. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.55.

Jones, Eric, Travis Oliphant, Pearu Peterson, et al. (2001). *SciPy: Open source scientific tools for Python*. [Online; accessed 2017-02-18]. URL: http://www.scipy.org/.

Kaplan, J. O., N. H. Bigelow, I. C. Prentice, S. P. Harrison, P. J. Bartlein, T. R. Christensen, W. Cramer, N. V. Matveyeva, A. D. McGuire, D. F. Murray, V. Y. Razzhivin, B. Smith, D. A. Walker, P. M. Anderson, A. A. Andreev, L. B. Brubaker, M. E. Edwards, and A. V. Lozhkin (2003). "Climate change and Arctic ecosystems: 2. Modeling, paleodata-model comparisons, and future projections". In: *Journal of Geophysical Research-Atmospheres* 108.D19. ISSN: 2169-897x. DOI: Artn 817110.1029/2002jd002559.

Krinner, G., Nicolas Viovy, Nathalie de Noblet-Ducoudré, Jérôme Ogée, Jan Polcher, Pierre Friedlingstein, Philippe Ciais, Stephen Sitch, and I. Colin Prentice (2005). "A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system". In: *Global Biogeochem. Cycles* 19.1, n/a–n/a. ISSN: 08866236. DOI: 10.1029/2003gb002199.

Kucharik, Christopher J., Jonathan A. Foley, Christine Delire, Veronica A. Fisher, Michael T. Coe, John D. Lenters, Christine Young-Molling, Navin Ramankutty, John M. Norman, and Stith T. Gower (2000). "Testing the performance of a dynamic global ecosystem model: Water balance, carbon balance, and vegetation structure". In: *Global Biogeochem. Cycles* 14.3, pp. 795–825. ISSN: 1944-9224. DOI: 10.1029/1999GB001138. URL: http://dx.doi.org/10.1029/1999GB001138.

Lafon, Thomas, Simon Dadson, Gwen Buys, and Christel Prudhomme (May 2012). "Bias correction of daily precipitation simulated by a regional climate model: a comparison of methods". In: *Int. J. Climatol.* 33.6, pp. 1367–1381. DOI: 10.1002/joc.3518. URL: http://dx.doi.org/10.1002/joc.3518.

Leemans, Rik and Wolfgang P Cramer (1991). "The IIASA database for mean monthly values of temperature, precipitation, and cloudiness on a global terrestrial grid". In: *International Institute for Applied Systems Analysis, Laxenburg, Austria.*

Lieth, Helmut (1975). "Modeling the Primary Productivity of the World". In: *Primary Productivity of the Biosphere*. Ed. by Helmut Lieth and Robert H. Whittaker. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 237–263. ISBN: 978-3-642-80913-2. DOI: 10.1007/978-3-642-80913-2_12. URL: http://dx.doi.org/10.1007/978-3-642-80913-2_12.

Maraun, D., H. W. Rust, and T. J. Osborn (Oct. 2009). "The annual cycle of heavy precipitation across the United Kingdom: a model based on extreme value statistics". In: *Int. J. Climatol.* 29.12, pp. 1731–1744. DOI: 10.1002/joc.1811. URL: http://dx.doi.org/10.1002/joc.1811.

Matalas, N. C. (1967). "Mathematical assessment of synthetic hydrology". In: *Water Resour. Res.* 3.4, pp. 937–945. ISSN: 1944-7973. DOI: 10.1029/WR003i004p00937. URL: http://dx.doi.org/10.1029/WR003i004p00937.

Menne, Matthew J., Imke Durre, Bryant Korzeniewski, Shelley McNeill, Kristy Thomas, Xungang Yin, Steven Anthony, Ron Ray, Russell S. Vose, Byron E. Gleason, and Tamara G. Houston (2012a). "Global Historical Climatology Network - Daily (GHCN-Daily), Version 3.22". In: DOI: 10.7289/V5D21VHZ. URL: http://dx.doi.org/10.7289/V5D21VHZ.

Menne, Matthew J., Imke Durre, Russell S. Vose, Byron E. Gleason, and Tamara G. Houston (July 2012b). "An Overview of the Global Historical Climatology Network-Daily Database". In: *J. Atmos. Oceanic Technol.* 29.7, pp. 897–910. DOI: 10.1175/jtech-d-11-00103.1. URL: http://dx.doi.org/10.1175/JTECH-D-11-00103.1.

Mitchell, Timothy D. and Philip D. Jones (2005). "An improved method of constructing a database of monthly climate observations and associated high-resolution grids". In: *Int. J. Climatol.* 25.6, pp. 693–712. ISSN: 0899-8418 1097-0088. DOI: 10.1002/joc.1181.

New, M., M. Hulme, and P. Jones (1999). "Representing twentieth-century space-time climate variability. Part I: Development of a 1961-90 mean monthly terrestrial climatology". In: *J. Climate* 12.3, pp. 829–856. ISSN: 0894-8755. DOI: Doi10.1175/1520-0442(1999)012<0829:Rtcstc>2.0.Co;2.

— (2000). "Representing twentieth-century space-time climate variability. Part II: Development of 1901-96 monthly grids of terrestrial surface climate". In: *J. Climate* 13.13, pp. 2217–2238. ISSN: 0894-8755. DOI: Doi10.1175/1520-0442(2000)013<2217:Rtcstc>2.0.Co;2.

New, M., D. Lister, M. Hulme, and I. Makin (2002). "A high-resolution data set of surface climate over global land areas". In: *Climate Research* 21.1, pp. 1–25. ISSN: 0936-577x. DOI: DOI10.3354/cr021001.

Neykov, N. M., P. N. Neytchev, and W. Zucchini (2014). "Stochastic daily precipitation model with a heavy-tailed component". In: *Natural Hazards and Earth System Science* 14.9, pp. 2321–2335. ISSN: 1684-9981. DOI: 10.5194/nhess-14-2321-2014.

Parlange, Marc B. and Richard W. Katz (May 2000). "An Extended Version of the Richardson Model for Simulating Daily Weather Variables". In: *J. Appl. Meteorol.* 39.5, pp. 610–622. DOI: 10.1175/1520-0450-39.5.610. URL: http://dx.doi.org/10.1175/1520-0450-39.5.610.

Pfeiffer, M., A. Spessa, and J. O. Kaplan (2013). "A model for global biomass burning in preindustrial time: LPJ-LMfire (v1.0)". In: *Geosci. Model Dev.* 6.3, pp. 643–685. ISSN: 1991-959x. DOI: 10.5194/gmd-6-643-2013. URL: http://www.geosci-model-dev.net/6/643/2013/gmd-6-643-2013.pdf.

Prentice, I. C., W. Cramer, S. P. Harrison, R. Leemans, R. A. Monserud, and A. M. Solomon (1992). "A Global Biome Model Based on Plant Physiology and Dominance, Soil Properties and Climate". In: *J. Biogeogr.* 19.2, pp. 117–134. ISSN: 0305-0270. DOI: Doi10.2307/2845499.

Prentice, I.C. (Aug. 1989). *Developing a Global Vegetation Dynamics Model: Results of an IIASA Summer Workshop*. IIASA Research Report. IIASA, Laxenburg, Austria. URL: http://pure.iiasa.ac.at/3223/.

Richardson, C. W. (1981). "Stochastic simulation of daily precipitation, temperature, and solar radiation". In: *Water Resour. Res.* 17.1, pp. 182–190. ISSN: 00431397. DOI: 10.1029/WR017i001p00182.

Rust, H. W., D. Maraun, and T. J. Osborn (July 2009). "Modelling seasonality in extreme precipitation". In: *The European Physical Journal Special Topics* 174.1, pp. 99–111. DOI: 10.1140/epjst/e2009-01093-7. URL: http://dx.doi.org/10.1140/epjst/e2009-01093-7.

Rymes, M.D. and D.R. Myers (2001). "Mean preserving algorithm for smoothly interpolating averaged data". In: *Sol. Energy* 71.4, pp. 225–231. DOI: 10.1016/s0038-092x(01)00052-4. URL: http://dx.doi.org/10.1016/S0038-092X(01)00052-4.

Seabold, Skipper and Josef Perktold (2010). *Statsmodels: Econometric and statistical modeling with python*.

Sitch, S., B. Smith, I. C. Prentice, A. Arneth, A. Bondeau, W. Cramer, J. O. Kaplan, S. Levis, W. Lucht, M. T. Sykes, K. Thonicke, and S. Venevsky (2003). "Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model". In: *Global Change Biol.* 9.2, pp. 161–185. ISSN: 1354-1013. DOI: 10.1046/j.1365-2486.2003.00569.x. URL: http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2486.2003.00569.x/abstract.

Smith, Adam, Neal Lott, and Russ Vose (2011). "The Integrated Surface Database: Recent Developments and Partnerships". In: *Bull. Amer. Meteor. Soc.* 92.6, pp. 704–708. DOI: 10.1175/2011BAMS3015.1. eprint: https://doi.org/10.1175/2011BAMS3015.1. URL: https://doi.org/10.1175/2011BAMS3015.1.

Sommer, Philipp S. (Aug. 2017). "The psyplot interactive visualization framework". In: *The Journal of Open Source Software* 2.16. DOI: 10.21105/joss.00363. URL: https://doi.org/10.21105/joss.00363.

Sommer, Philipp S. and Jed O. Kaplan (Oct. 2017a). "A globally calibrated scheme for generating daily meteorology from monthly statistics: Global-WGEN (GWGEN) v1.0". In: *Geosci. Model Dev.* 10.10, pp. 3771–3791. DOI: 10.5194/gmd-10-3771-2017.

— (Sept. 2017b). "GWGEN v1.0.2: A global weather generator for daily data". In: DOI: 10.5281/zenodo.889213. URL: https://github.com/ARVE-Research/gwgen.

Stephens, Graeme L., Tristan L'Ecuyer, Richard Forbes, Andrew Gettelmen, Jean-Christophe Golaz, Alejandro Bodas-Salcedo, Kentaroh Suzuki, Philip Gabriel, and John Haynes (2010). "Dreary state of precipitation in global models". In: *Journal of Geophysical Research: Atmospheres* 115.D24, n/a–n/a. ISSN: 2156-2202. DOI: 10.1029/2010JD014532. URL: http://dx.doi.org/10.1029/2010JD014532.

Sun, Ying, Susan Solomon, Aiguo Dai, and Robert W. Portmann (Mar. 2006). "How Often Does It Rain?" In: *J. Climate* 19.6, pp. 916–934. DOI: 10.1175/jcli3672.1. URL: http://dx.doi.org/10.1175/JCLI3672.1.

Viovy, N. and P. Ciais (2016). Online Database. URL: http://dods.extra.cea.fr/data/p529viov/cruncep.

Walter, H and H Lieth (1967). "Climate diagram world atlas". In: *VEB Gustav Fischer Verlag Jena, Jena*.

Wei, Y., S. Liu, D. N. Huntzinger, A. M. Michalak, N. Viovy, W. M. Post, C. R. Schwalm, K. Schaefer, A. R. Jacobson, C. Lu, H. Tian, D. M. Ricciuto, R. B. Cook, J. Mao, and X. Shi (2014). "The North American Carbon Program Multi-scale Synthesis and Terrestrial Model Intercomparison Project – Part 2: Environmental driver data". In: *Geosci. Model Dev.* 7.6, pp. 2875–2893. ISSN: 1991-9603. DOI: 10.5194/gmd-7-2875-2014. URL: http://www.geosci-model-dev.net/7/2875/2014/.

Wilks, D. S. (1998). "Multisite generalization of a daily stochastic precipitation generation model". In: *J. Hydrol.* 210.1-4, pp. 178–191. ISSN: 0022-1694. DOI: 10.1016/S0022-1694(98)00186-3.

— (1999a). "Interannual variability and extreme-value characteristics of several stochastic daily precipitation models". In: *Agric. For. Meteorol.* 93.3, pp. 153–169. ISSN: 0168-1923. DOI: 10.1016/S0168-1923(98)00125-7.

— (1999b). "Multisite downscaling of daily precipitation with a stochastic weather generator". In: *Climate Research* 11.2, pp. 125–136. ISSN: 0936-577x. DOI: 10.3354/cr011125.

— (1999c). "Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain". In: *Agric. For. Meteorol.* 96.1-3, pp. 85–101. ISSN: 0168-1923. DOI: 10.1016/S0168-1923(99)00037-4.

Wilks, D. S. and R. L. Wilby (1999). "The weather generation game: a review of stochastic weather models". In: *Prog. Phys. Geog.* 23.3, pp. 329–357. ISSN: 0309-1333. DOI: 10.1177/030913339902300302.

Wilks, Daniel S. (2010). "Use of stochastic weathergenerators for precipitation downscaling". In: *Wiley Interdiscip. Rev. Clim. Change* 1.6, pp. 898–907. ISSN: 17577780. DOI: 10.1002/wcc.85.

Woodward, F. Ian, Thomas M. Smith, and William R. Emanuel (1995). "A global land primary productivity and phytogeography model". In: *Global Biogeochem. Cycles* 9.4, pp. 471–490. ISSN: 1944-9224. DOI: 10.1029/95GB02432. URL: http://dx.doi.org/10.1029/95GB02432.

Woolhiser, D. A. and G. G. S. Pegram (1979). "Maximum Likelihood Estimation of Fourier Coefficients to Describe Seasonal-Variations of Parameters in Stochastic Daily Precipitation Models". In: *J. Appl. Meteorol.* 18.1, pp. 34–42. ISSN: 0894-8763. DOI: Doi10.1175/1520-0450(1979)018<0034:Mleofc>2.0.Co;2.

Woolhiser, D. A. and J. Roldan (1982). "Stochastic Daily Precipitation Models: 2. A Comparison of Distributions of Amounts". In: *Water Resour. Res.* 18.5, pp. 1461–1468. ISSN: 0043-1397. DOI: DOI10.1029/WR018i005p01461.

Woolhiser, D. A. and José Roldán (1986). "Seasonal and Regional Variability of Parameters for Stochastic Daily Precipitation Models: South Dakota, U.S.A". In: *Water Resour. Res.* 22.6, pp. 965–978. ISSN: 00431397. DOI: 10.1029/WR022i006p00965.

Chapter 7

Conclusions

Paleoclimatological large-scale reconstructions allow an independent evaluation of global climate models skill to simulate climates outside the range of modern climate variability. In the previous chapters of this thesis I described several new open-source tools that can be used to leverage the single site-based proxy-climate reconstruction onto continental, or even global scale, by using a combination of thousands of different records.

For the Eurasian Modern Pollen Database (EMPD) (Davis et al., in prep) I developed a web-framework to communicate and manage the community-driven database in a transparent and sustainable way (chapter 2). The framework consists of an interactive web-based interface into the data and an automated administration webapp. The entire framework is based on the free webservices provided by the version control platform Github and as such allow to trace back every change to the database and provides a variety of tools to manage new contributions and/or changes to the database. This methodology assures stable and intuitive access to the database, independent of the available funding, and contributors or maintainers. One can think of many further potential applications for this framework that can be applied to any regional pollen (or in general, proxy) database. The EMPD is only one example, other potential use cases are the Latin American Pollen Database (LAPD) (Flantua et al., 2015), or the African Pollen Database (APD) (Vincens et al., 2007). The method can also be applied to communicate a study-specific collection of proxy sites and use already implemented analysis and visualization tools for the data, or add new methods specific to the scope of the study. The future plans with this project therefore include a further generalization of the methods, particularly the visualization methods of the EMPD- and POLNET-viewer (section 2.3), to make it widely applicable. The integration with Github allows an easy way to share the source, and to host the interactive interface on the same platform without any costs.

The next tool I presented is the stratigraphic digitization software straditize in chapter 3 (Sommer et al., 2019). This package transforms stratigraphic diagrams, i.e. diagrams where the analysis of samples are plotted against a common y-axis, usually representing age or depth. The potential applications for this software are numerous because of the existence of hundreds of pollen datasets (and more) that are only available as pollen diagrams in the publications. This software provides the unique possibility to make this data from the pre-digital era accessible in a reasonable amount of time. Further extensions to this package will involve the support of new diagram types (e.g. multiple lines in a single diagram column). A strong focus will lie on the documentation of the software in order to make it easier and accessible. This will involve video tutorials, more tutorials for the various diagrams directly embedded into the software, and there are still some parts of the software that are not yet sufficiently document.

In chapter 4, I further described the interactive visualization framework psyplot (Sommer, 2017), a cross-platform open source python project that combines plotting and data management into a single framework that can be used conveniently via command-line and a graphical user interface (GUI). The software differs from most of the visual analytic software such that it focuses on extensibility and flexibility and can therefore be used in a variety of research questions. It particularly provides the basis for the straditize software (chapter 3) and for multiple of the analysis in the chapters 5 and 6. psyplot currently provide visualization methods that range from simple line plots, to density plots, regression analysis and geo-referenced visualization in two dimensions. In the future, this will be enhanced with 3D visualization methods to provide the first visualization tool in climate research that can be used conveniently from the command line (Sommer, 2019).

In the second part of my thesis I described two new computational models that leverage site-based information into models for large-scale prediction of paleo environments. The first one is the pyleogrid package, a new method for a spatio-temporal gridded climate estimate from a database of proxy-climate reconstruction. It is a probabilistic extension of the method by Mauri et al., 2015 and Davis et al., 2003 that provides reliable uncertainties by incorporating the intrinsic dating and proxy-climate reconstruction uncertainties of the individual sites, in order to generate a product that can conveniently be used for data-model intercomparison. In addition, this framework contains two novel methods, (1) a model to predict dating uncertainties based on the age of the pollen sample and the time-difference to the closest chronological control point, and (2) a new probabilistic version of the modern analogue technique (MAT), based on constrained Gibbs sampling algorithms for the age of the samples and the climate reconstructions. The ensemble method is very scalable, both in terms of the size of the spatio-temporal domain, and the computational resources that are used for the prediction. This method will be used in the near future to generate a climate reconstruction for the entire northern hemisphere that is based on the POLNET database as described in Davis and Kaplan, 2017. Further developments will also concentrate on a revision of the age uncertainty estimate using the recent database by Wang et al., 2019 which contains standardized chronologies for more than 500 datasets.

Finally, the second model in chapter 6 describes the new global weather generature GWGEN (Sommer and Kaplan, 2017), a statistical model for a temporal downscaling of monthly to daily climatology. This model can be applied on the entire globe whilst being parameterized based on a large dataset of weather stations with more than 50 million records. The aim is to provide a tool that can be implemented in a global vegetational model for paleo environments. These models require daily meteorology as input which poses a considerable challenge considering the long simulation period they have.

I additionally developed another model in collaboration, that is not related to paleo and therefore not included in the main part of this thesis. This model, the Integrated Urban Complexity Model (IUCM) (Cremades and Sommer, 2019), presents a new method that simulates the transformation of urban areas while focusing on a low energy consumption from urban mobility. I mention this model here because it also incorporates the infrastructural methodologies that I developed for psyplot and GWGEN. This additional use-case highlights one of the important aspects of open-source software development, that is a flexible, extensible and sustainable modular framework, where the packages related to a specific product can be used in multiple other products. Other examples for it are the *docrep* and *sphinx-nbexamples* packages (Sommer, 2018a,b) that I primarily developed for psyplot but are used in a variety

of recently developed packages now (e.g. Abernathey et al., 2017; Banihirwe et al., 2019; Uchida, 2018).

References

Abernathey, Ryan, Takaya Uchida, Julius Busecke, Dhruv Balwada, Ci Zhang, and Anirban Sinha (July 2017). *xgcm/xgcm: v0.1.0*. Version v0.1.0. DOI: 10.5281/zenodo.826926. URL: https://doi.org/10.5281/zenodo.826926.

Banihirwe, Anderson, Matthew Long, Alper Altuntas, Riley Brady, Michael Levy, and Sudharsana K J L (Apr. 2019). *NCAR/esmlab: v2019.04.27*. Version v2019.04.27. DOI: 10.5281/zenodo.2652704. URL: https://doi.org/10.5281/zenodo.2652704.

Cremades, R. and Philipp S. Sommer (2019). "Computing climate-smart urban land use with the Integrated Urban Complexity model (IUCm 1.0)". In: *Geoscientific Model Development* 12.1, pp. 525–539. DOI: 10.5194/gmd-12-525-2019. URL: https://www.geosci-model-dev.net/12/525/2019/.

Davis, B. A. S., S. Brewer, A. C. Stevenson, and J. Guiot (2003). "The temperature of Europe during the Holocene reconstructed from pollen data". In: *Quat. Sci. Rev.* 22.15-17, pp. 1701–1716. ISSN: 02773791. DOI: 10.1016/s0277-3791(03)00173-2.

Davis, Basil A. S., Manuel Chevalier, Philipp S. Sommer, et al. (in prep). "The Eurasian Modern Pollen Database (EMPD), Version 2". In: *Earth System Science Data ESSD*.

Davis, Basil A. S. and Jed O. Kaplan (Feb. 2017). *HORNET Holocene Climate Reconstruction for the Northern Hemisphere Extra-tropics*. SNF-Research-Plan. last accessed Jan, 30th, 2018. URL: http://p3.snf.ch/project-169598#.

Flantua, Suzette G.A., Henry Hooghiemstra, Eric C. Grimm, Hermann Behling, Mark B. Bush, Catalina González-Arango, William D. Gosling, Marie-Pierre Ledru, Socorro Lozano-Garcìa, Antonio Maldonado, Aldo R. Prieto, Valentì Rull, and John H. Van Boxel (Dec. 2015). "Updated site compilation of the Latin American Pollen Database". In: *Review of Palaeobotany and Palynology* 223, pp. 104–115. DOI: 10.1016/j.revpalbo.2015.09.008.

Mauri, A., B. A. S. Davis, P. M. Collins, and J. O. Kaplan (2015). "The climate of Europe during the Holocene: a gridded pollen-based reconstruction and its multi-proxy evaluation". In: *Quat. Sci. Rev.* 112, pp. 109–127. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2015.01.013. URL: http://www.sciencedirect.com/science/article/pii/S0277379115000372.

Sommer, Philipp S. (Aug. 2017). "The psyplot interactive visualization framework". In: *The Journal of Open Source Software* 2.16. DOI: 10.21105/joss.00363. URL: https://doi.org/10.21105/joss.00363.

— (2018a). *docrep: A Python Module for intelligent reuse of docstrings*. Last accessed: 2018-02-03. URL: https://github.com/Chilipp/docrep (visited on 02/03/2018).

— (2018b). *sphinx-nbexamples: Create an examples gallery with sphinx from Jupyter Notebooks*. Last accessed: 2018-02-03. URL: https://github.com/Chilipp/sphinx-nbexamples (visited on 02/03/2018).

— (2019). *psy-vtk: A VTK plugin for psyplot*. Last accessed: 2019-05-27. URL: https://github.com/Chilipp/psy-vtk (visited on 05/27/2019).

Sommer, Philipp S. and Jed O. Kaplan (Oct. 2017). "A globally calibrated scheme for generating daily meteorology from monthly statistics: Global-WGEN (GWGEN) v1.0". In: *Geosci. Model Dev.* 10.10, pp. 3771–3791. DOI: 10.5194/gmd-10-3771-2017.

Sommer, Philipp S., Dilan Rech, Manuel Chevalier, and Basil A. S. Davis (Feb. 2019). "straditize: Digitizing stratigraphic diagrams". In: *Journal of Open Source Software* 4.34, p. 1216. DOI: 10.21105/joss.01216. URL: https://doi.org/10.21105/joss.01216.

Uchida, Takaya (Aug. 2018). *xgcm/xrft: First Official Release*. Version v0.1-beta. DOI: 10.5281/zenodo.1402636. URL: https://doi.org/10.5281/zenodo.1402636.

Vincens, Annie, Anne-Marie Lézine, Guillaume Buchet, Dorothée Lewden, and Annick Le Thomas (2007). "African pollen database inventory of tree and shrub pollen types". In: *Rev. Palaeobot. Palynol.* 145.1-2, pp. 135–141. ISSN: 00346667. DOI: 10.1016/j.revpalbo.2006.09.004.

Wang, Yue, Simon J. Goring, and Jenny L. McGuire (2019). "Bayesian ages for pollen records since the last glaciation in North America". In: *Scientific Data* 6.1, p. 176. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0182-7. URL: https://doi.org/10.1038/s41597-019-0182-7.

Appendices

List of Figures

2.1 Modern calibration samples in the EMPD 21
2.2 Screenshot of the EMPD viewer 23
2.3 Map of sites in the POLNET database 28
2.4 Screenshot of an automatically generated pollen diagram in the POLNET database viewer. The left dropdown menu above the pollen diagram allows to select the different naming schemes (here consolidated names that were used for the pollen-climate reconstruction). The right dropdown menu selects either the entire data or specific samples that are then displayed as a bar diagram (see the pollen data in table 2.2). . 29
2.5 Climate reconstructions visualized in the POLNET viewer 29

3.1 Common features in a pollen diagram 35
3.2 Cleaned binary image of the data part 39
3.3 Illustration of the basic digitization strategy of *straditize* 40
3.4 Pollen diagram for Hoya del Castillo 42
3.5 Residuals plot of the digitized Hoya del Castillo diagram 42

4.1 The psyplot core framework 48
4.2 Screenshot of the psyplot GUI 51
4.3 psyplot Gui plot creation dialog 51

5.1 Pollen Database 66
5.2 Climate analogues of the Tigalmamine site 68
5.3 Univariate age uncertainty models 72
5.4 Univariate age uncertainty models 73
5.5 Scaled histograms of age sampling methods 74
5.6 Realized age distribution for the entire dataset 76
5.7 Scaled histograms of temperature sampling methods 78
5.8 Realized temperature distributions for a few samples 79
5.9 Elevation difference and corrections 81
5.10 Ice sheet mask of the pollen samples 81
5.11 Realized summer temperature reconstruction of Tigalmamine with different constraints 84
5.12 Realized summer temperature reconstruction of Tigalmamine with different analogues 85
5.13 Deterministic approach vs. Ensemble mean 86
5.14 Impact of climate constraint on gridded results 87
5.15 Impact of the number of analogues on the gridded results 88
5.16 Relationship between distance to proxy sites and ensemble standard deviation 89
5.17 Estimated age uncertainties 92
5.18 Example of sampled distribution 92
5.19 Ensemble standard deviation 93

5.20 Standard error of the Tps gridding method 94

6.1 Schematic workflow of GWGEN 104
6.2 Weather stations used for parameterization and evaluation of the weather generator. 107
6.3 Transition probabilities vs. wet fraction 109
6.4 Mean precipitation - Gamma scale relationship 110
6.5 Correlation of minimum temperature on wet and dry days to the monthly mean 112
6.6 Correlation of maximum temperature on wet and dry days to the monthly mean 112
6.7 Correlation of standard deviation of min. and max. temperature to the monthly mean 113
6.8 Correlation of cloud fraction on wet and dry days to the monthly mean 116
6.9 Correlation of standard deviation of cloud fraction to the monthly mean 116
6.10 Correlation of wind speed on wet and dry days to the monthly mean . 118
6.11 Correlation of standard deviation of wind speed to the monthly mean 118
6.12 QQ-plots for all variables and all quantiles 121
6.13 QQ-plots for different quantiles for precipitation 122
6.14 wind bias correction 122
6.15 Results of the sensitivity analysis 127

List of Abbreviations

APD African Pollen Database. 3, 22, 133

API Application programming interface. 7, 47, 49, 52, 82

CDO Climate Data Operator. 46, 47

CI Continuous Integration. 6, 7, 22, 24, 27

EMPD Eurasian Modern Pollen Database. 8, 21, 22, 23, 24, 25, 27, 28, 67, 133

EPD European Pollen Database. 67

ESM Earth System Model. 2, 46

FOSS Free and Open-Source Software. 5, 6, 7

GAM Generalized Additive Model. 71, 72, 73

GUI graphical user interface. 7, 8, 45, 46, 50, 51, 52, 53, 54, 134

JJA summer (June, July and August). 66, 69, 78

LAPD Latin American Pollen Database. 3, 22, 133

LGM Last Glacial Maximum. 2

MAT modern analogue technique. 69, 78, 84, 89, 91, 134

MCMC Markov chain Monte Carlo. 76, 77

NAPD North American Pollen Database. 3

NCDC National Climatic Data Center. 4

NOAA National Oceanic and Atmospheric Administration. 4

PMIP Paleoclimate Modelling Intercomparison Project. 2

Appendix A

Publications and Conference contributions

A.0.1 Peer-reviewed

Cremades, R. and P. S. Sommer (2019). "Computing climate-smart urban land use with the Integrated Urban Complexity model (IUCm 1.0)". In: *Geoscientific Model Development* 12.1, pp. 525–539. DOI: 10.5194/gmd-12-525-2019. URL: https://www.geosci-model-dev.net/12/525/2019/.

Sommer, Philipp, Dilan Rech, Manuel Chevalier, and Basil Davis (Feb. 2019). "straditize: Digitizing stratigraphic diagrams". In: *Journal of Open Source Software* 4.34, p. 1216. DOI: 10.21105/joss.01216. URL: https://doi.org/10.21105/joss.01216.

Weitzel, Nils, Sebastian Wagner, Jesper Sjolte, Marlene Klockmann, Oliver Bothe, Heather Andres, Lev Tarasov, Kira Rehfeld, Eduardo Zorita, Martin Widmann, Philipp S. Sommer, Gerd Schädler, Patrick Ludwig, Florian Kapp, Lukas Jonkers, Javier García-Pintado, Florian Fuhrmann, Andrew Dolman, Anne Dallmeyer, and Tim Brücher (Sept. 2018). "Diving into the past – A paleo data-model comparison workshop on the Late Glacial and Holocene". In: *Bulletin of the American Meteorological Society*. DOI: 10.1175/bams-d-18-0169.1.

Sommer, Philipp S (Aug. 2017). "The psyplot interactive visualization framework". In: *The Journal of Open Source Software* 2.16. DOI: 10.21105/joss.00363. URL: https://doi.org/10.21105/joss.00363.

Sommer, Philipp S. and Jed O. Kaplan (Oct. 2017). "A globally calibrated scheme for generating daily meteorology from monthly statistics: Global-WGEN (GWGEN) v1.0". In: *Geosci. Model Dev.* 10.10, pp. 3771–3791. DOI: 10.5194/gmd-10-3771-2017.

A.0.2 Conference contributions

Sommer, P. S., B. A. S. Davis, and M. Chevalier (Apr. 2019a). "Github and Open Research Data; an example using the Eurasian Modern Pollen Database". In: *EGU General Assembly Conference Abstracts*. Vol. 21. EGU General Assembly Conference Abstracts, p. 5669. URL: https://meetingorganizer.copernicus.org/EGU2019/EGU2019-5669.pdf.

Sommer, Philipp S., Basil A. S. Davis, Manuel Chevalier, Jian Ni, and John Tipton (2019b). "The HORNET project: applying 'big data' to reconstruct the climate of the Northern Hemisphere during the Holocene". In: *20th Congress of the International Union for Quaternary Research (INQUA)*. International Union for Quaternary Research. URL: https://app.oxfordabstracts.com/events/574/program-app/submission/94623.

Sommer, P. S. (Apr. 2018). "Psyplot: Interactive data analysis and visualization with Python". In: *EGU General Assembly Conference Abstracts*. Vol. 20. EGU General Assembly Conference Abstracts. Provided by the SAO/NASA Astrophysics Data System, p. 4701. URL: http://adsabs.harvard.edu/abs/2018EGUGA..20.4701S.

Sommer, P. S., B. A. S. Davis, and M. Chevalier (Apr. 2018a). "STRADITIZE: An open-source program for digitizing pollen diagrams and other types of stratigraphic data". In: *EGU General Assembly Conference Abstracts*. Vol. 20. EGU General Assembly Conference Abstracts. Provided by the SAO/NASA Astrophysics Data System, p. 4433. URL: http://adsabs.harvard.edu/abs/2018EGUGA..20.4433S.

Sommer, Philipp S., Manuel Chevalier, and Basil A. S. Davis (July 2018b). "STRADITIZE: An open-source program for digitizing pollen diagrams and other types of stratigraphic data". In: *AFQUA - The African Quaternary*. Nairobi (Kenya): AFQUA. URL: https://afquacongress.wixsite.com/afqua2018.

Sommer, P. and J. Kaplan (Apr. 2017). "Quantitative Modeling of Human-Environment Interactions in Preindustrial Time". In: *PAGES OSM 2017, Abstract Book*, pp. 129–129.

Sommer, P. (Apr. 2016). "Psyplot: Visualizing rectangular and triangular Climate Model Data with Python". In: *EGU General Assembly Conference Abstracts*. Vol. 18. EGU General Assembly Conference Abstracts. Provided by the SAO/NASA Astrophysics Data System, p. 18185. URL: http://adsabs.harvard.edu/abs/2016EGUGA..1818185S.

Sommer, P. and J. Kaplan (Apr. 2016a). "Fundamental statistical relationships between monthly and daily meteorological variables: Temporal downscaling of weather based on a global observational dataset". In: *EGU General Assembly Conference Abstracts*. Vol. 18. EGU General Assembly Conference Abstracts. Provided by the SAO/NASA Astrophysics Data System, EPSC2016–18183. URL: http://adsabs.harvard.edu/abs/2016EGUGA..1818183S.

— (May 2016b). "Fundamental statistical relationships between monthly and daily meteorological variables: Temporal downscaling of weather based on a global observational dataset". In: *Workshop on Stochastic Weather Generators*. Vannes (France): University of Bretagne Sud. URL: https://www.lebesgue.fr/content/sem2016-climate-program.