

UCLA

UCLA Electronic Theses and Dissertations

Title

Uncertainty Quantification in Structured High Dimensional Regression Tasks with Missing Data

Permalink

<https://escholarship.org/uc/item/9x5374b7>

Author

Hwangbo, Nathan

Publication Date

2025

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Uncertainty Quantification in Structured High Dimensional Regression Tasks with Missing
Data

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Nathan Hwangbo

2025

© Copyright by
Nathan Hwangbo
2025

ABSTRACT OF THE DISSERTATION

Uncertainty Quantification in Structured High Dimensional Regression Tasks with Missing
Data

by

Nathan Hwangbo

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2025

Professor Karen McKinnon, Chair

Modern statistical analyses often involve inference in the setting where the number of parameters exceeds the number of datapoints. Valid inference in this high dimensional setting relies on modeling assumptions that constrain the dimensionality of the parameter space. The set of assumptions appropriate for a given task should ideally make use of domain knowledge about the underlying structure of the data. This thesis develops statistical models for high dimensional regression tasks in two scientific domains, with particular attention paid to quantifying uncertainty due to these structural assumptions. Chapters 1 and 2 discuss the field of paleoclimate reconstruction, whereby climate of the past is inferred through natural archives like tree-ring widths (TRW). The paleoclimate reconstruction task is best formulated as an inverse problem, where the unobserved climate of the past is viewed as missing data that must be estimated by inverting a forward model which maps climate to tree-ring widths. Without constraints, this problem is ill-posed, in the sense that it is impossible to disentangle the contribution of climatic and non-climatic influences on variation in the paleoclimate proxy. We consider the Bayesian approach to the inverse problem, which estimates the probability distribution of the inverse solution and constrains the solution space by imposing prior distributions on both past climate and the climate-proxy relationship. Chapter 2 proposes a Bayesian hierarchical model for TRW-based paleoclimate reconstructions at a

single spatial location, with emphasis placed on propagating uncertainties in the forward model. One source of uncertainty in existing paleoclimate reconstruction workflows is the temporal change of support problem, whereby TRWs produce a single measurement per year, but are typically not equally sensitive to climate in all seasons. The proposed model resolves the change-of-support problem by modeling climate at monthly temporal resolution, simultaneously estimating the season which the proxies are sensitive to alongside the reconstruction. Chapter 3 considers uncertainty in paleoclimate field reconstructions, where proxies at multiple spatial locations are used to generate a gridded spatial reconstruction of the past. The high dimensional spatiotemporal nature of this task make computational feasibility a limiting factor, so that popular existing methods for climate field reconstruction make use of simplifying assumptions for practical use. In Chapter 3, Bayesian hierarchical models for climate field reconstruction are developed which include these existing methods as special cases. This framework is used alongside pseudoproxy experiments to highlight the simplifying assumptions implicit in existing climate field reconstruction methods, and to assess the influence of these assumptions on reconstruction skill. Insights gained from these experiments are used to develop a simplified model for the common paleoclimate task of estimating a large-scale spatial average. Chapter 4 discusses the application of high dimensional regression to a different scientific domain, the identification of biomarkers in metabolomic profiles of cerebrospinal fluid (CSF). Recent advances in high-throughput biology have allowed for the measurement of concentrations of all small molecules in a biological sample (known as the metabolome). These metabolomic profiles are influenced by both the genome and the environment, and can be used to identify biomarkers of different phenotypes. We focus on the identification of biomarkers for Alzheimer’s and Parkinson’s disease using metabolomic profiles of CSF. Due to the invasive procedure required to extract CSF, very few samples are available for analysis, leading to a “small n , large p ” setting where the number of metabolites greatly exceeds the sample size. In addition, multiple technologies and processing techniques are available for annotating and identifying metabolites, so that the results of any

metabolomic analysis are sensitive to the particular data collection methods used. These metabolomic profiles are subject to measurement error and frequently feature missing data, so that the imputation of missing regressors (here, metabolite concentrations) is a common theme shared across the thesis. We compare biomarker identification across metabolomic profiles and multiple statistical analyses to assess the robustness of our findings to these data and methodological choices. Chapter 5 contains a discussion of possible extensions to the methods developed here to obtain more reliable inference in each of these scientific domains.

The dissertation of Nathan Hwangbo is approved.

Kevin J. Anchukaitis

Mark S. Handcock

Frederick Paik Schoenberg

Karen McKinnon, Committee Chair

University of California, Los Angeles

2025

Contents

Abstract	ii
List of Figures	viii
Acknowledgements	xii
1 Introduction	1
2 Simultaneous optimal target season estimation and local climate reconstruction using tree-ring widths	8
2.1 Introduction	9
2.2 Data and methods	11
2.2.1 Model output and tree ring chronologies	11
2.2.2 A hierarchical Bayesian model for monthly reconstructions	11
2.2.3 Pseudoproxy construction	14
2.3 Results	16
2.3.1 The impact of misspecified seasonality	16
2.3.2 Pseudoproxy Experiments for the monthly-resolved model	18
2.3.3 Case study: California blue oaks	21
2.4 Discussion	23
3 Towards Principled Uncertainty Quantification in Paleoclimate Spatial Field Reconstructions	26
3.1 Introduction	26
3.2 Data and Methods	29
3.3 Statistical Models	30
3.3.1 Offline Paleoclimate Data Assimilation	32
3.3.2 Uncertainty due to the climate-proxy relationships	35
3.3.3 Uncertainty due to spatial covariance model	40
3.3.4 Large-scale average reconstruction	49
3.4 Discussion	54
4 Predictive Modeling of Alzheimer’s and Parkinson’s Disease Using Metabolomic and Lipidomic Profiles from Cerebrospinal Fluid	58
4.1 Introduction	59
4.2 Materials and Methods	60

4.2.1	Metabolomics	61
4.2.2	Preprocessing	64
4.2.3	Regression modeling	65
4.2.4	Pathway analysis	66
4.3	Results	67
4.3.1	PCA	68
4.3.2	Prediction Results	68
4.3.3	Missing Data	72
4.3.4	Pathway and Set Enrichment Analysis	73
4.4	Discussion	75
4.4.1	Modeling limitations	78
5	Discussion	85
	Appendices	89
A	Chapter 2 Appendix	90
A.1	Details of the statistical model	90
A.2	Eliciting informative prior weights	93
B	Chapter 3 Appendix	98
B.1	Prior specification	99
C	Chapter 4 Appendix	101
C.1	Abbreviations	101
C.2	Summary of data by phenotype	103
C.3	Flowchart of analysis	103
C.4	Distribution in date of LP draws	104
C.5	Univariate untargeted patterns of missingness	105
C.6	Precision-Recall curves and threshold-dependent metrics	106
C.7	ROC curves of predictive models including age and sex	107
C.8	MSEA output tables	108
C.9	Removing batch effects for PD vs Control analysis	109
C.10	Lipids tables for discriminating AD from PD	114
C.11	Metabolites appearing in all leave one out models	114

List of Figures

2.1	The problem with reconstruction target season misspecification using idealized pseudoproxy experiments from a Southern California grid cell (35.1°N, 120°W) in a run from the CESM Last Millenimum Experiment. Baseline reconstructions of log precipitation (standardized to have zero mean and unit variance from 1950-2004) for the seasons (a) mean DJF (the correct pseudoproxy seasonal window), (b) the annual mean, (c) January, and (d) JJA. The posterior mean reconstruction is shown in purple with shading representing 90% pointwise credible intervals. Black lines display true (i.e. climate model output) standardized log precipitation.	18
2.2	Demonstration of the proposed model with linear pseudoproxies. (a) The reconstruction of weighted monthly log precipitation anomalies. The posterior mean reconstruction is shown in purple, with shading for 90% credible intervals. The black line represents the truth, which is obtained by applying the learned weights to monthly climate model output. (b) The posterior mean (bars) of the learned weights for each month in the water year (October - September), with error bars representing 90% credible intervals. A reference line is drawn at height 1/12, which represents the uniform prior weights across months. (c) The reconstruction of (purple) and true (black) DJF precipitation, which is transformed from the original reconstruction target (log precipitation anomalies). 21	21
2.3	Application of the proposed model to blue oaks. Similar to Figure 2.2, panel (a) shows the posterior mean and 90% credible interval of the targeted weighted log precipitation anomaly average in purple alongside instrumental weighted log precipitation anomalies in black, panel (b) shows the posterior mean weights with 90% credible intervals, and panel (c) shows the DJF reconstruction. . .	22
3.1	Subset of PAGES2K TRW database used for case study. Locations of each TRW chronology is represented by a red point, with the gray boxes representing the gridcells where temperature will be reconstructed. Each chronology is assumed to be sensitive to the temperature in the gridcell which contains it.	31
3.2	Pseudoproxy experiment setup. (a) The grid represents the spatial domain used for all pseudoproxy experiments, with yellow cells representing locations where pseudoproxies are generated. (b) An example of one such pseudoproxy used in a VS-Lite based PPE. Each point represents a different year, with the simulated TRW on the vertical axis and JJA mean temperature on the horizontal axis.	38

3.3	Out of sample (1600-1950) evaluation of an idealistic PPE ($\text{SNR} = 2$) looking at the impact of modeling the proxy-climate relationship. The top row shows the gridcell-by-gridcell correlation between the posterior mean reconstruction and the truth (i.e. climate model output), while the bottom row shows the frequentist 90% coverage over this period. The columns show the output of Bayesian models which learn the calibration alongside the reconstruction, including (right) and excluding (left) a model discrepancy term.	39
3.4	PPE learning covariance from proxies which experience different spatial covariance in the reconstruction and calibration periods. (a) The learned posterior mean spatial correlation matrix of temperature using a squared exponential parametrization. (b) The sample temperature correlation matrix in the reconstruction period (i.e., with shuffled location labels) (c) The sample temperature correlation matrix in the calibration period (i.e., with correct location labels). The subtitles report the Frobenius distances to the learned posterior mean correlation matrix in panel (a).	44
3.5	PPE learning covariance from proxies. (a) The sample spatial temperature correlation matrix from climate model output. (b) The learned posterior mean correlation matrix using an LKJ(0.5) prior, and (c) a squared exponential covariance function prior. The subtitles report the Frobenius distance between the posterior mean covariance matrix and the true covariance matrix.	46
3.6	PPE spatial mean reconstruction of JJA mean temperature using a Bayesian model which fixes covariance at the sample instrumental mean (top), and one that learns covariance from the data, using a modified BARCAST model (bottom). The posterior mean (colored lines) and 90% credible intervals are shown, alongside the true climate model run spatial mean.	48
3.7	North American Southwest reconstruction summary using PAGES2k tree ring width chronologies. Maps of per-gridcell (a) correlation and (b) 90% frequentist coverage between the posterior mean reconstruction and GISTEMP instrumental surface temperature across all out-of-sample years with instrumental surface temperature (1880-1949), alongside (c) the spatial mean reconstruction	50
3.8	Comparison of PPE spatial mean reconstruction, using (a) the joint normality assumption to fill in gridcells with no available data before taking the mean vs (b) ridge regression to directly estimate the spatial mean from the gridcells with available data	53
4.1	Untargeted data projected onto the first two Principal Components (PC). Each point represents a subject, colored by their phenotype. Percentages in the axis titles refer to the percentage of variation of the data explained by the respective PC. 95% confidence ellipses assuming the t -distribution are also plotted. The first two principal components do not clearly separate the disease phenotypes.	69

4.2	Receiver Operating Characteristic (ROC) Curves for binomial elastic net regressions classifying (a) controls against subjects with AD, (b) controls against subjects with PD, and (c) subjects with AD against subjects with PD. Solid lines represent models formed using each of the five missing data imputed datasets. The dotted $y = x$ line represents the ROC curve under a model which makes predictions at random. The average Area Under the Curve (AUC) across the five ROC curves is displayed in the bottom right.	70
4.3	Pathway Analysis of Mummichog on the positive mode untargeted metabolites from univariate logistic models classifying PD and AD against Controls, sorted by $-\log_{10}$ Benjamini Hochberg corrected p -values, with the vertical dashed line marking $p = 0.05$	74
A.1	Demonstration of the proposed model with VSLite pseudoproxies. (a) The reconstructed (red) and true (black, from the climate model) DJF precipitation. The reconstructed DJF precipitation is transformed from the original reconstruction target, which is in log-precipitation space. (b) The posterior mean (bars) of the learned weights for each month in the water year (October - September), with error bars representing 90% credible intervals. (c) Monthly correlations of the noiseless VSLite pseudoproxy and each month's precipitation	94
A.2	Demonstration of the proposed model with simplified VS-Lite pseudoproxies. As in Fig. 2, but using VS-Lite to produce the pseudoproxies with SNR of 0.75 and setting growth outside of DJF to 0.	95
A.3	Sensitivity of the weight inference to prior choice and SNR in VS-Lite pseudoproxies with growing season truncated to lie solely in DJF. The posterior mean and 90% credible intervals of monthly weights for pseudoproxies with varying SNR (0.25, 0.5, and 0.75), using (a) an uninformative prior with $w_{\text{input}} = 1$ in all months, (b) a prior with $w_{\text{input}} = 10$ in SONDJF and 1 in all other months, and (c) a prior with $w_{\text{input}} = 10$ in DJF and 1 in all other months.	96
A.4	Reference model for the blue oaks. The baseline model applied to the same blue oaks, choosing DJF as the reference season to facilitate comparisons with Figure 3c. Calibration and reconstruction time periods match that of Section 3.3.	97
B.1	Temporal availability of proxies at each gridcell in the subset of PAGES2K being used. Each row represents a gridcell containing proxy data, with the length of the bar representing the years with data in at least one chronology.	98
B.2	Ridge regression coefficients (i.e. the posterior mean coefficients) for estimating the spatial mean over all shown gridcells, given pseudoproxies in the colored gridcells. Each location's temperature time series is normalized to have mean 0 and variance one prior to model fitting, so that the coefficient is unitless. .	99
C.1	(a) Distribution of subject age, split by phenotype. (b) a comparison of missingness between profiles, split by phenotype	103
C.2	Flowchart outlining the analysis performed on each of the three profiles. . . .	103
C.3	Distribution of sample collection date, split by phenotype.	104

C.4	For metabolites with widely varying amounts of abundance missingness by phenotype (AD, PD, age-matched controls), PD tend to have the least missing data. Plotted is the percent missingness of untargeted metabolites for which univariate logistic regressions classifying its missingness across subjects using phenotype contained a $FDR < 0.05$. The controls used for this analysis were found by matching each member of the AD/PD cohort to the control of closest age, removing duplicates. Metabolites are listed by their retention time, neutral mass (DA), and mode, with each value separated by an underscore	105
C.5	Precision-Recall Curves for binomial elastic net regressions classifying (a) controls against subjects with AD, (b) controls against subjects with PD, and (c) subjects with AD against subjects with PD, using the same predictions as Figure 4.2. Each line represent models formed using one of the five missing data imputed datasets. The average Area Under the Curve (AUC) is displayed in the bottom right.	106
C.6	ROC displaying leave one out predictive accuracy of models including age and sex rather than detrending each metabolite. Age and sex are not penalized in the elastic net procedure to reflect the known associations between age and sex with AD and PD. As expected, we generally find that the predictive performance of these models is slightly better than those presented in the main analysis.	107
C.7	ROC curves displaying leave one out predictive accuracy of models fit on the targeted profile after the batch orthogonalization procedure. We find that these models are largely unable to distinguish AD from controls after this procedure, and are still able to distinguish PD from controls (but with less predictive accuracy compared to the results in main analysis).	112
C.8	Comparison of univariate p -values before and after orthogonalization, using univariate logistic regressions fit on the targeted profile to classify PD against controls. p -values are Benjamini-Hochberg corrected within each dataset, with a $-\log_{10}$ transformation applied. The $y = x$ line is also drawn, with values above the line representing metabolites which are more differentially abundant after orthogonalization. Points are colored according to false discovery rate cutoffs of 0.05, indicating whether a metabolite is considered differentially abundant in both analyses, only one analysis, or in neither analysis.	113

Acknowledgements

This dissertation would not be possible without the support of a large community full of fantastic people. To start, I could not have asked for a better advisor than Karen McKinnon. Thank you for sharing your expertise as a statistician and climate scientist, for providing direction and insightful feedback to my work, for fostering my skills as an independent researcher, and for helping me navigate the wild west that is the PhD experience. I would also like to thank the rest of my thesis committee: thank you to Kevin Anchukaitis for introducing me to the fascinating world of paleoclimatology and guiding me through it, and to Mark Handcock and Rick Schoenberg for the valuable statistical advice.

I've had the opportunity to learn from a great research group working at the intersection of statistics and climate science. Interactions with this group have greatly influenced the way I think about science. Thank you to Sam Baugh, Suqin Duan, Wenwen Kong, Kyle McEvoy, Mady Beckner, and Talia Anderson for the advice, conversation, and camaraderie.

I would like to thank the Department of Statistics and Data Science for their support. Thank you to Mahtash Esfandiari and Vivian Lew for the opportunity to work at the statistics consulting center, to Rob Gould for nominating me to serve as the TAC and for encouraging me to think critically about statistics pedagogy, to Michael Tsiang and Thomas Maierhofer for helping me figure out how to lecture an undergraduate course, and to Dale Kim for requesting me as a TA whenever our schedules lined up.

There are many friends and housemates that have supported me along the way (and I'll thank each of you in person), but a special thank you goes to James Kim for being an ideal housemate throughout my entire time at UCLA, and to Jonathon Ahn, Alan Dang, Steve Kim, Brian Lee, Raymond Nguyen, and Randy Yiv for their support from day 1. I was also very fortunate to be joined by a great cohort of graduate students, and I would especially like to thank Evan Becker, Navin Souda, and Luciano Vinas for being with me every step of the way.

Graduate school was not on my radar until I enrolled in classes taught by Alex Franks.

Thank you Alex for giving me my first opportunity to do research, for investing time and resources into these projects, and for your mentorship in setting me down this path. Thank you to Daniel Promislow for the encouragement and for the contributions to these first research projects, and Julien Brun for fostering my interest in environmental data science and developing my skills as an applied scientist.

Last but certainly not least, thank you Mom and Dad for providing a safe environment for me to learn and grow up in, and thank you to Tim, Hanna, and Nick for always being role models I can look up to.

Vita

Education

University Of California, Los Angeles	September 2020 - June 2025
Statistics Ph.D (Expected)	
University Of California, Santa Barbara	September 2016 - June 2020
Statistical Science B.S and Mathematical Sciences B.S	

Awards

Don Ylvisaker Award for the Best Practice of Statistics	June 2024
---	-----------

Chapter 1

Introduction

Modern advances in computational capability, statistical theory, and data collection practices have given way to the development of statistical techniques for inference when the number of unknown parameters exceeds the number of available data points. A common theme in this high dimensional setting is that structural assumptions on the data are often required to obtain reliable parameter estimates. The choice of structural assumption in an applied data analysis should rely on domain expertise and will vary depending on the task at hand. These assumptions are often made at the discretion of the researcher, and represent a “researcher degree of freedom”, a choice made by the analyst whose uncertainty might not be captured by the output of existing statistical models [1, 2]. This thesis contains methodological advancements towards more complete quantification of this type of uncertainty for two regression tasks: paleoclimate reconstruction and metabolomics-based biomarker identification.

Paleoclimate reconstruction aims to extend the climate record into the past, before the development of modern weather instruments. The routine measurement of surface air temperature, precipitation, and other climate variables provides a source of data which has proved to be crucial for improving our understanding of earth systems. However, very few high quality instrumental measurements exist prior to 1850, and this temporal length limits

statistical analysis of the climate system. For example, quantifying the natural variability in the climate system requires the characterization of multidecadal climate patterns (e.g. [3]). The statistical estimation of these patterns from the relatively short instrumental record is subject to large sampling variability, so that physical assumptions and understanding must be heavily relied upon to obtain estimates of these forms of natural variability. Furthermore, the quantity of interest in many climate studies is often extreme events, whose characterization is particularly important for climate adaptation. Statistical estimation of extremes requires much more data to constrain than measures of central tendency (e.g the mean). As a result, the length of the instrumental record also limits our ability to obtain reliable estimates of these distribution tails. One additional challenge inherent with using the instrumental record in climate studies is that the technological advances which led to modern weather instruments coincide with a period of industrialization where humans have undoubtedly influenced the climate system. High quality instrumental climate data is not available prior to industrialization, making it difficult to disentangle anthropogenic influence from natural climate variability from existing data. In short, the length of the instrumental record and its positioning in time present obstacles which obfuscate our ability to characterize natural climate variability and the climate response to anthropogenic forcings.

Paleoclimate reconstructions attempt to extend the instrumental record by inferring past climate from natural archives whose formation depends on climate. Tree-ring widths (TRW), the paleoclimate proxy most commonly referred to in this thesis, provide one record per year, with wider rings typically reflecting more favorable growing conditions. With this proxy type, statistical methods for paleoclimate reconstruction require the development of a forward model which maps climate to TRWs, and a method for inverting this relationship to obtain estimates of past climate from long-lived trees. Paleoclimate reconstruction studies requires many scientific and statistical considerations [4]: for example, the researcher must, for each paleoclimate proxy, determine what time period the proxy records, determine what climate variable(s), if any, the proxy is sensitive to, and identify non-climatic influences on growth.

Each of these factors requires thoughtful analysis from an expert on the particular proxy type. Once a biologically-informed statistical model describing proxy formation from climate is developed, it must then be inverted to obtain estimates of past climate given ancient proxy measurements. This so-called inverse problem is notoriously ill-conditioned, so that that small modifications to the data and the model can lead to large differences in the resulting paleoclimate reconstruction. This not only makes point estimation difficult, but also tends to lead to unreliable quantification of uncertainty. Modeling assumptions are necessary for reliable estimation, and the sensitivity to these choices must be assessed. This complex, multifaceted nature of paleoclimate studies makes reconstruction uncertainty difficult to fully account for. Chapters 2 and 3 propose Bayesian hierarchical models (BHMs) for paleoclimate reconstruction which provide principled uncertainty quantification at different stages of the reconstruction workflow, focusing on annual reconstructions of the Common Era.

One challenge with TRW-based paleoclimate reconstructions is that TRWs produce a single measurement each year, but growth is not uniformly influenced by climate throughout the year. This forms a many-to-one relationship between sub-annual instrumental climate data and annual tree ring widths. Dendrochronologists commonly resolve this temporal mismatch by first identifying seasons that the proxies are sensitive to, then reconstructing seasonal averages from the TRWs. The choice of seasonal target is often not obvious, given the complex biological mechanisms which drive proxy formation and the high autocorrelation typically present in climate variables at sub-annual resolution. Uncertainty in the choice of season does not get propagated to the final reconstruction, and it is unclear how impactful this choice is to the analysis. Chapter 2 quantifies the impact of this decision on reconstruction skill and proposes a statistical model which simultaneously learns which season the proxy is most sensitive to alongside the reconstruction. In addition to estimating proxy seasonality, the output from the proposed model can be used to create reconstructions of arbitrary monthly averages which incorporate uncertainty due to the choice of averaging period. More broadly, we demonstrate that uncertainty in misspecified forward models is not properly propagated

to the reconstruction, even when the forward model is learned simultaneously alongside the reconstruction. In particular, when the linearity assumption is a poor fit for the proxy–climate relationship, we show that the introduction of a mean-zero, time varying “model discrepancy” term improves uncertainty quantification in the reconstruction. The large number of latent parameters and short instrumental climate record make model evaluation with observations difficult, so emphasis is placed on the development of pseudoproxy experiments, which use monthly output of physics-based climate models alongside predetermined forward models to construct synthetic proxy data. Pseudoproxies based on climate model output are based on realistic spatiotemporal climate variability and can extend into the last millennium, providing a thorough testbed for the evaluation of statistical models. The contents of this chapter are currently in review for publication by *Geophysical Research Letters*.

In recent years, paleoclimatologists have compiled impressive collections of paleoclimate proxies across time and space, enabling the construction of estimates of past climate globally. Obtaining reconstructions of spatiotemporal climate fields from proxies is a monumental task requiring collaboration from researchers across many scientific domains [5]. On the statistical side, spatiotemporal reconstruction studies combines the aforementioned challenges with the inverse problem in local models, with the challenging of interpolating climate information between proxy sites to locations which do not yet have available proxy measurements. Chapter 3 develops Bayesian hierarchical models for spatial paleoclimate reconstructions, jointly modeling local proxy–climate relationships across different proxy sites, then interpolating inferred local climate at these sites to estimate a spatially gridded paleoclimate reconstruction surface. The proposed Bayesian hierarchical model links together existing spatially resolved reconstruction methods as special cases. Most notably, the popular method of paleoclimate data assimilation (PaleoDA) is shown to be an approximate sampler to a variant of the proposed model, and this connection is used to reveal different simplifying assumptions implicit in PaleoDA. Pseudoproxy experiments are constructed to better understand the implications of each of these assumptions on reconstruction skill, especially with regards to

uncertainty quantification. One simplifying assumption made by PaleoDA is that the forward models need not be learned simultaneously with the reconstruction, and can instead be learned first and approximated using point estimates of the parameters. As a result, PaleoDA reconstruction uncertainty estimates do not typically reflect uncertainty due to the learned forward model. As in the previous chapter, we find that simultaneously learning the forward model alongside the reconstruction can account for this uncertainty, and that the inclusion of a zero-mean model discrepancy term in the forward model improves reconstruction uncertainty quantification when the forward model is misspecified.

A defining feature of PaleoDA is the choice to estimate spatial covariance from the output of physics-based climate models, rather than using any observational instrumental or proxy data. This modeling decision is critical to the success of PaleoDA, as the estimation of a spatial covariance matrix from proxy data represents a computational bottleneck in Bayesian models for climate field reconstruction. However, choosing not to learn spatial covariance from the observations risks missing out on information about spatial covariance that can be learned from the proxies themselves. To evaluate the impact of this choice, we develop pseudoproxy experiments which quantify the extent to which proxies can inform spatial covariance in the Common Era. We find that PaleoDA and the proposed BHMs are complementary, with the simplifying assumptions of PaleoDA offering insight into where modeling assumptions can be made without sacrificing reconstruction skill.

Oftentimes, the quantity of interest in climate field reconstructions is the estimation of a large-scale spatial mean (e.g. a time series reconstruction of average Northern Hemisphere temperature). When the reconstruction of a spatial mean is the only estimand of interest, we propose a modification to our Bayesian model which treats the estimation of a large scale average as a high dimensional regression problem, whereby inferred local climate at gridcells with proxy data are used as regressors in a linear model to directly predict the large scale spatial average. This step of the analysis naturally fits into the modeling framework by adding this regression step as an addition layer to the Bayesian hierarchical

model. The advantage of this method is that it does not require the estimation of a full gridded climate field, instead only utilizing reconstructions at gridcells which contain proxy observations. Pseudoproxy experiments demonstrate that this method is computationally efficient without loss of reconstruction skill. This framing of spatiotemporal climate problems as high dimensional regression tasks where different spatial locations are treated as predictors has started to gain traction in recent years (e.g. [6]). This treatment shares quite a lot in common with the high dimensional regression tasks in high-throughput biology, in the sense that the predictors are highly correlated, and the correlation has structure which is partially known through field expertise (i.e. climate system spatiotemporal patterns and biological pathways). Chapter 4 contains an application of a high dimensional regression task in metabolomic profiles.

The metabolome, the collection of small molecules in an organism, acts as a bridge between genotype and phenotype, in the sense that the metabolome is influenced by both genetic and environmental factors. This makes the metabolome a unique and powerful tool for investigation into biological mechanisms for observed characteristics. Characterization of the metabolome includes the measurement of many different metabolite concentrations, so that jointly modeling the metabolome naturally requires high dimensional statistical models. Furthermore, processing metabolite data requires the use of expensive and sensitive scientific equipment, so that these studies also typically feature a small number of human subjects. Statistical analysis in this setting of small sample size, many metabolites requires additional modeling assumptions to be well-posed. One challenge associated with metabolomic studies comes from the many possible technologies for measuring metabolite concentrations, each producing datasets with varying degrees of metabolite annotation and identification. The goal of the analysis in Chapter 4 is to use metabolomic profiles of cerebrospinal fluid (CSF) from a small sample of patients to identify potential biomarkers for Alzheimer’s and Parkinson’s disease (AD and PD, respectively). To assess variability due to the choice of metabolomic profile, we conduct our analysis in parallel using data resulting from three different processing

techniques: (1) a targeted panel, which identifies concentrations from a small set of well-known, predetermined metabolites, (2) an untargeted panel, which records the full spectrum of measured metabolites in a sample, and (3) a lipidomic panel which exclusively measures fat-soluble metabolites. Elastic net logistic regression is used to classify subjects with Alzheimer’s and Parkinson’s disease from their metabolomic profiles and offer exploratory analysis for potential biomarkers that might be relevant to the predictions. Obtaining CSF samples requires lumbar punctures, an invasive medical procedure. As a result, a small number of samples are available for analysis relative to the number of metabolites, and particular care is paid to the robustness of the inference to different statistical techniques for this small- n , large p problem. Additionally, biomarker identification is assessed across untargeted, targeted, and lipidomic profiles to assess whether predictive drivers from each profile can be linked to the same biological pathways. The contents of this paper are published in [7].

Despite being from distinct scientific fields, paleoclimate reconstruction and metabolite biomarker identification share many statistical characteristics. Both regression tasks are high-dimensional in nature, either in the form of many metabolites or many latent years of climate data to be inferred, potentially at many different spatial locations. The data for both fields is highly structured, either through biological pathways in the metabolome, or through spatiotemporal relationships in the earth system. Both tasks feature missing data problems, either through missing metabolite concentrations resulting from the use of sensitive mass spectrometry equipment, or through “missing” climate values prior to the development of modern weather instruments. The work presented here represents domain-specific methods towards a more complete accounting of uncertainty in such applied analyses.

Chapter 2

Simultaneous optimal target season estimation and local climate reconstruction using tree-ring widths

Tree-ring widths provide a natural archive of past environmental variability and are a valuable resource for paleoclimate reconstructions of the Common Era. However, tree-rings widths only provide climate information during a portion of the year and uncertainty in the seasonality of climate influence on the proxy formation adds to the challenge of separating climatic signal from non-climatic noise. Here we propose a Bayesian hierarchical model for the simultaneous estimation of a target reconstruction season and reconstruction of local climate. We estimate the target reconstruction season through the introduction of latent monthly weights, whose priors can be adjusted to reflect expert knowledge. Model behavior is explored using pseudoproxy experiments and applications to tree-ring width chronologies with known seasonal biases. Our proposed model provides meaningful information about the true growth-determining season of the proxy and leads to improved uncertainty quantification

The contents of this chapter are in review at Geophysical Research Letters

while retaining overall model skill.

2.1 Introduction

Obtaining a record of past climate that extends before instrument-based measurements is desirable for a better understanding of how modern changes in climate compare to past climate variability [8, 9, 10]. To this end, paleoclimate proxies — natural systems that record past climatic conditions — can be used to reconstruct the climate of the past. In this paper, we consider local climate reconstructions using tree-ring widths (TRW), which are some of the most widely distributed and highly spatio-temporally resolved proxies of the Common Era [11, 12]. Wider tree rings indicate favorable growing conditions, which can include a variety of climactic influences (e.g. precipitation, temperature, soil moisture) depending on the location and tree species.

One common difficulty is that proxy records are seldom equally good recorders of climate in all seasons. TRWs typically reflect climate during only part of the year, when the climate is a dominant control on growth. While some studies – especially multiproxy studies – aim to reconstruct mean annual climate (e.g. [13, 14, 15]), the majority of tree-ring reconstructions use expert or prior knowledge to make an informed assumption about their archive’s seasonal window and limit their reconstructions to that season. When the seasonal window is correctly specified, statistical models that reconstruct seasonal climate have more skill than those reconstructing mean annual climate; however, this choice is not always clear, particularly when multiple tree-ring chronologies are used. For instance, a recent experiment [4] assessed methodology-based reconstruction uncertainty by tasking different research groups with creating Northern Hemisphere temperature reconstructions from the same data. Across 15 research groups, temperature was reconstructed for four different seasonal windows: June-July, June-August, May-September, and June-October. Given the different climatologies represented in these time periods, the choice of seasonal window can meaningfully change the

interpretation of the resulting reconstruction [16, 17, 18]. This choice is further complicated by the fact that the optimal reconstruction season need not be the growing season, as TRWs can integrate climate information from prior months through biological-induced autocorrelation (e.g. [19, 20]). These challenges motivate our development of a consistent, objective, and data-driven approach to address the paleoclimate challenge of seasonality.

Here we propose a Bayesian statistical model that simultaneously infers a target reconstruction season and reconstructs past climate from the proxy data. Previous work with other proxies has also attempted to infer a reconstruction target season, but as a separate pre-processing step prior to fitting the regression model used for reconstruction [21, 22, 23, 24]. These approaches either depend heavily on expert judgment or rely on screening procedures which perform many simultaneous tests on autocorrelated data, which is prone to producing spurious correlations without statistical corrections [25, 26]. In contrast, the statistical model proposed here explicitly includes estimation of the seasonal window as part of the calibration and reconstruction, which avoids the need for multiple tests.

We focus our attention on linear inverse Bayesian hierarchical models, a class of statistical model which explicitly combines domain knowledge with data to obtain reconstructions with uncertainty quantification (see [27] for an overview). Section 2 contains a discussion of the data and statistical model. In Section 2.3.1, pseudoproxy experiments (PPE), wherein synthetic tree-ring records are constructed as a response to specified climate variables [28], are developed to identify the potential problems with misspecified monthly sensitivity. Our statistical model is applied to pseudoproxy data in Section 2.3.2, showing the model’s skill in both the identification of a growth-dominant seasonal window as well as reconstruction. In Section 2.3.3, the model is applied to real tree-ring width data that are known to be sensitive to water year precipitation and we test the extent to which our model recovers the same seasonal window. Of particular interest is the model’s ability to accurately capture and reflect uncertainty in its final reconstruction, which is an ongoing challenge in paleoclimate studies [20, 29, 25].

2.2 Data and methods

2.2.1 Model output and tree ring chronologies

The pseudoproxies used to test the proposed methods are constructed from monthly-average temperature and precipitation. We rely on climate model output rather than actual observations so that we can straightforwardly evaluate the skill of the statistical models in reconstructing climate before the instrumental period. Model output is from the first ensemble member of the CESM1 Last Millennium Ensemble Project with all transient forcings [30] with a spatial resolution of 2° . The full simulations spans the years 850-2005, from which we use years 1600-2004 to evaluate our method.

The tree-ring records for the case study are from 165 individual blue oak (*Quercus douglasii*) increment cores from two southern California sites: Los Lobos Creek (34.92°N , 119.24°W) and Figueroa Mountain (34.74°N , 120°W), which span the years 1293 - 2004 [31, 32]. Annual tree-ring width measurements for this species at these sites have been shown to be sensitive to local cool-season and/or water year precipitation [33, 34, 35]. Observational monthly precipitation is taken from the Parameter-elevation Regressions on Independent Slopes Model (PRISM), a high resolution (4km) gridded precipitation dataset (e.g. [36]). We extract monthly precipitation from the nearest grid cell to the blue oaks between the years 1895-2004. In all analysis, the water year is used as our annual averaging period.

2.2.2 A hierarchical Bayesian model for monthly reconstructions

The statistical model we propose builds upon existing approaches to hierarchical Bayesian regression for paleoclimate calibration and reconstruction [27, 37, 38, 39, 40, 41, 42, 43]. Following previous work [44, 45], we model the log of tree-ring width as a linear function of an annualized climate variable of interest and tree age. In this paper, we use log-transformed precipitation as the climate variable, although the same methods could be used for another variable such as temperature (the log transform ensures that reconstructed precipitation

remains positive, and in many cases leads to a better linear fit with tree-ring widths [46, 47]). Tree age is a known non-climatic factor confounding the climate-proxy relationship, due to the fact that older trees have wider circumferences, so that the same volume of growth results in narrower rings [48]. The following annually-resolved model for TRW will be referred to as the “baseline” model:

$$\begin{aligned}
\log(\text{TRW})_{i,t} &= \beta_{0i} + \beta_{1i}\text{age}_{i,t} + \beta_c\text{climate}_t + \epsilon_{i,t} \\
\text{climate}_{t+1} - \mu &= \alpha(\text{climate}_t - \mu) + e_t \\
\epsilon_{i,t} &\sim N(0, \sigma_i^2) \\
e_t &\sim N(0, \sigma_e^2)
\end{aligned} \tag{2.1}$$

where i indexes trees, and t indexes time in years, beginning with the first year with tree ring data and ending with the last year with tree ring data. To capture climate signal that might be present in individual trees but not the mean response [49, 50], each tree is allowed its own intercept (β_{0i}), age-width regression coefficient (β_{1i}), and residual variance (σ_i), and these tree-specific parameters are assumed to be draws from distributions which are shared by all trees (see SI Text S1 for more detail). β_c captures the proxy-climate relationship, which is assumed to be linear and shared between trees. α is an autoregressive coefficient which captures year-to-year dependence in the climate time series, with innovations e_t . Given that the tree-ring records extend further back in time than the instrumental climate record, the reconstruction task is to estimate the values $(\text{climate}_1, \text{climate}_2, \dots, \text{climate}_{t_{\text{ins}}-1})$, where t_{ins} denotes the first year of instrumental data. The generic notation “climate” is used to emphasize, as noted above, that any climate variable can be modeled using this approach, although log-precipitation is the chosen reconstruction target throughout the paper. We treat each yearly value in the pre-instrumental climate time series as a model parameter to be estimated alongside the regression parameters. This baseline model combines characteristics

of previous approaches: tree-specific parameters are included to allow for flexibility in the tree-ring width-climate relationship (similar to [44, 45]), while also including an autoregressive term on the climate variable (which is included in [39, 40]) to better estimate reconstruction variability. This autoregressive structure is in contrast to models with an autoregressive term on tree-ring widths [45], which might have a different temporal structure than the climate series due to memory in the biological process of growth itself [51, 52, 53].

In the model described above, there is a single annual value of climate associated with the tree-ring widths, which is usually chosen to be the average climate over a predetermined subset of months. This motivates our proposed model, which learns monthly weights as part of the calibration. Let climate_{t_j} now refer to *monthly* climate anomalies in year t and month j , so that each month's time series has zero mean and unit standard deviation over the training period. A simplified version of the proposed model can be written as follows (full details are available in SI Text S1):

$$\log(\text{tree-ring width})_{i,t} = \beta_{0i} + \beta_{a,i}\text{age}_{i,t} + \beta_c\text{climate}_t + \eta_t + \epsilon_{i,t} \quad (2.2)$$

$$\epsilon_{i,t} \sim N(0, \sigma_i^2)$$

$$\text{climate}_t = \sum_{j=1}^{12} w_t \text{climate}_{t_j}$$

$$\text{climate}_{t_m} - \mu = \alpha_{ar}(\text{climate}_{t_{m-1}} - \mu) + \epsilon_{ar}$$

$$\eta_t \sim N(0, \sigma_{\eta_t}^2)$$

$$\theta_m \sim N_{12}(w_{\text{input}}, \Sigma)$$

$$w_m = \frac{\exp(\theta_m)}{\sum_{j=1}^{12} \exp(\theta_j)}$$

where m indexes months. The first two lines are similar to their analogue in model (2.1), where the logarithm of tree-ring widths is assumed to have a linear relationship with age and climate. However, rather than pre-specifying which months are averaged to produce a

yearly climate value, the yearly climate variable (climate_t) is now modeled as a weighted average of monthly values (climate_{t_j}). The prior distribution on the vector of monthly weights depends on hyperparameters w_{input} (also a vector of length 12) and Σ (a 12×12 covariance matrix), which can be specified by the user to encode prior knowledge about the target reconstruction season. The relative size of elements of w_{input} indicate which months are thought to be included in the target reconstruction season, while the absolute magnitudes indicate the strength of this prior knowledge (a non-informative uniform prior is used through this paper, and guidance on the selection of w_{input} is discussed in SI Text S2). The other hyperparameter, Σ , encodes prior belief about relationships between monthly weights. We allow Σ to learn the strength of these relationships, but enforce that months closer together in time are more similar to those further away (e.g. weights for December will look more similar to January than July). Weights are constrained to be nonnegative and sum to one by taking the unconstrained weights (θ) and normalizing them by their exponentiated sum. Another difference between the baseline model in Eqn. (2.1) and the proposed model is that the autoregressive term is now on the monthly process rather than the yearly one, with α_{ar} (constrained between 0 and 1) modeling the strength of this month-to-month relationship. Finally, we note the inclusion of η_t , a zero mean “model discrepancy” term with changing variance σ_{η_t} . This term accounts for approximation error in the assumed linear calibration relationship, and is needed to quantify uncertainty in the reconstruction [54]. The model produces a fully probabilistic reconstruction, which can then be evaluated in the validation period (e.g. as far back as the first year of available instrumental climate data)

2.2.3 Pseudoproxy construction

The proposed model will be evaluated using pseudoproxy experiments in Sec. 2.3.2. Pseudoproxies are constructed from the monthly climate model output with pre-specified growing seasons, which enables us to evaluate both reconstruction skill and seasonal window inference. Our pseudoproxies come in two general forms. The first is a “best case” scenario, where the

pseudoproxies are created from the same data-generating process as the proposed statistical model in Eqn. (2.2) (namely a linear combination of monthly log-precipitation anomalies), so that the statistical model perfectly captures the pseudoproxy-climate relationship. The second set of pseudoproxies are generated using the more realistic VS-lite forward model [55], where tree-ring widths are modeled as a potentially non-linear multivariate function of temperature and precipitation. In this model, growth is controlled by the limiting variable between temperature and soil moisture, the latter of which is a function of precipitation, and is modeled using the CPC leaky bucket model [56]. The primary parameters for the VS-lite model are thresholds for the minimum and maximum temperature and soil moisture needed for growth. Temperatures or soil moisture values above the maximum are assumed to have no additional impact on growth. In order to create pseudo tree rings that are primarily moisture sensitive, analogous to the blue oaks used in the case study (Sec. 2.3.3), we chose the minimum and 0.1 quantile for temperature, and the 0.1 and 0.95 quantiles for soil moisture, for the minimum and maximum parameters at each location. Depending on the climatology, growth is limited by moisture in 90% - 95% of months in our experiments. Defaults from the `VSLiteR` package are used for all other parameter choices [57]. Both methods of generating pseudoproxies yield a single yearly tree-ring width time series $W(y)$, which is standardized to have zero mean and unit standard deviation, then turned into multiple trees through scaling with noise (following [58]):

$$W(y)_i \sim N(\sqrt{1 - \sigma^2}W(y), \sigma^2)$$

where i indexes the trees as before. Values of σ can be chosen by specifying the signal-to-noise ratio, defined here as by $\sqrt{\frac{1 - \sigma^2}{\sigma^2}}$. In all pseudoproxy experiments, the validation period is chosen to be 1600-1949 with training period 1950-2004.

2.3 Results

2.3.1 The impact of misspecified seasonality

We first demonstrate the implications of misspecifying the seasonal sensitivity of the proxy when constructing climate reconstructions using pseudoproxy data. There are three broad categorizations of seasonal misspecification in annual-resolution climate reconstructions: the reconstruction target season being too large (e.g. ignoring seasonality and attempting to reconstruct mean annual climate, which has been warned against in [59, 20]), the target season being too small, and the target season having no overlap with the proxy’s true seasonal sensitivity. To illustrate the consequences of each of these misspecifications, we conduct a series of pseudoproxy experiments using precipitation data from one member of the CESM Last Millennium Ensemble for the gridcell closest to the location of the case study presented in Section 2.3.3 (35°N, 119°W). In each experiment, ten pseudoproxies are constructed to have a linear relationship with average December - February (DJF) log precipitation with a signal-to-noise ratio (SNR) of 0.75, which mimics relatively high quality paleoclimate proxies [28]. We then fit a non-seasonally resolving hierarchical Bayesian model (described in Section 2.2.2, Equation 2.1) which matches the pseudoproxy generating process, except that the target reconstruction seasons are chosen to be correctly specified (Fig 2.1a), longer than DJF (Fig 2.1b), shorter than DJF (Fig 2.1c), and completely misspecified (Fig 2.1d). Because the data generating process almost exactly matches the statistical model, the differences in reconstruction skill between the pseudoproxy experiments can be attributed solely to seasonal misspecification. The statistical model produces a probabilistic reconstruction, meaning that a posterior probability distribution for precipitation is estimated for each hindcast year. The reconstructions are evaluated in two ways: in the skill of the posterior mean, which is measured by the Pearson correlation coefficient R with the true log precipitation, and in uncertainty estimation, which is measured by 90% frequentist coverage, defined as the proportion of years in which the 90% credible interval contains the true log precipitation.

Figure 2.1a represents the best case scenario, in which the pseudoproxies are sensitive to DJF-average log precipitation, and the paleoclimatologist has chosen DJF-average log precipitation as the reconstruction target. The model follows the form of Equation (2.1) and is calibrated on the years 1950-2004, with reconstruction skill being evaluated on the years 1600-1949. In this case, the correlation between the posterior mean and the true log precipitation over the reconstruction period is high at 0.89, and the uncertainty quantification is close to the nominal level, with 91% of years being captured by the 90% credible interval. In contrast, Figure 2.1b represents the case when the chosen target season is too large, with mean annual log precipitation chosen as the target when the pseudoproxies are only sensitive to DJF. The model skill is reduced compared to the correctly specified case, and it noticeably underestimates the reconstruction uncertainty ($R = 0.75$, 90% coverage of 0.80). This misspecification error tends to be the least harmful seasonal misspecification, as the correct months are still entirely contained in the chosen reconstruction target. The addition of the rest of the year acts as noise in the regression model, which in this specific case is relatively small since DJF makes up a large proportion of yearly precipitation in the gridcell (the correlation between DJF and annual precipitation is $R = 0.91$). Figure 2.1c considers the case where the reconstruction target season is smaller than the growing season of DJF, targeting only January. Again, both mean reconstruction skill and coverage degrade relative to the experiment with correct seasonal specification, and more so than overestimating the reconstruction season ($R = 0.55$, 90% coverage of 0.70). In general, the consequence of this misspecification will depend on the strength of the climate-proxy relationship in the months that are left out and on the strength of the correlation between the growing and target seasons; the larger reduction in skill in this case of misspecification is linked to the relatively low correlation ($R=0.61$) between January and DJF precipitation. Finally, Figure 2.1d considers the case when the target season is completely misspecified, reconstructing JJA log precipitation instead of DJF. While this extreme scenario is unlikely to occur in practice, it illustrates the point that reconstruction skill (in both posterior mean correlation

and uncertainty quantification) suffers most when two seasons are nearly uncorrelated with one another ($R = 0.14$, 90% coverage = 0.66).

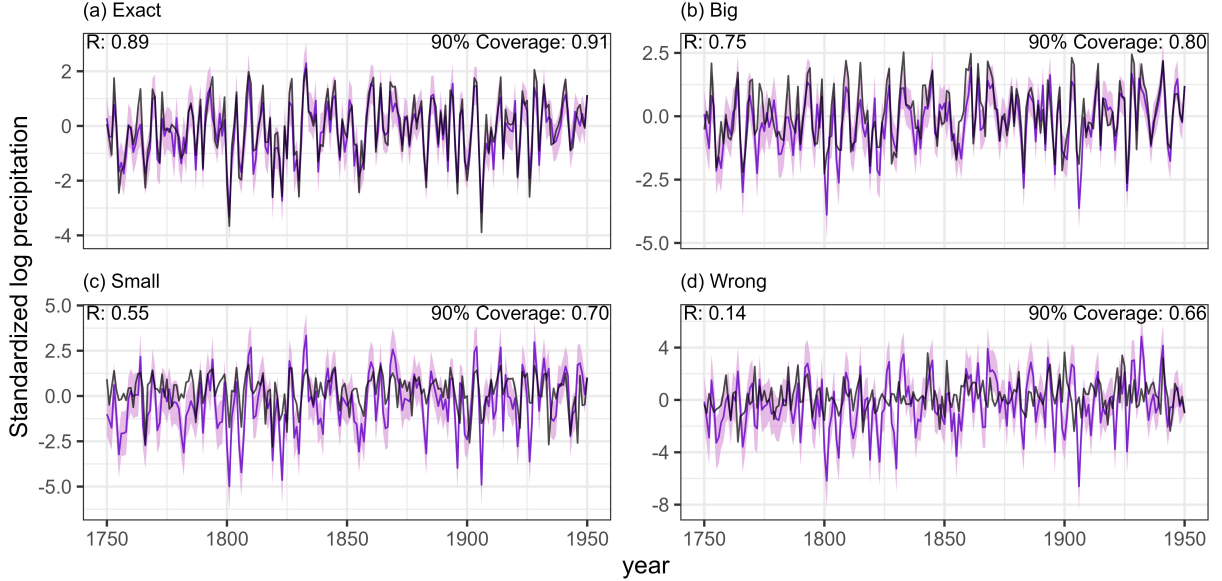


Figure 2.1: The problem with reconstruction target season misspecification using idealized pseudoproxy experiments from a Southern California grid cell (35.1°N , 120°W) in a run from the CESM Last Millenimum Experiment. Baseline reconstructions of log precipitation (standardized to have zero mean and unit variance from 1950-2004) for the seasons (a) mean DJF (the correct pseudoproxy seasonal window), (b) the annual mean, (c) January, and (d) JJA. The posterior mean reconstruction is shown in purple with shading representing 90% pointwise credible intervals. Black lines display true (i.e. climate model output) standardized log precipitation.

2.3.2 Pseudoproxy Experiments for the monthly-resolved model

Having demonstrated the negative implications of misspecifying the reconstruction season, we turn to evaluating the proposed statistical model (Eqn. (2.2)), which jointly estimates the contribution of each month’s climate to the growth of the annual tree ring and reconstructs past climate. The pseudoproxies used for the evaluation follow the same construction as in

Section 2.3.1, generating ten DJF-sensitive pseudoproxies with SNR of 0.75, calibrated in years 1950-2004 and evaluated in years 1600-1949.

The results of the first pseudoproxy experiment are shown in Figure 2.2. The model is largely successful at inferring from the data that year-to-year tree growth was controlled only by DJF precipitation (Fig. 2.2b); the weights in all other months are small, and their credible intervals do not exceed the prior of equal weight across all months. However, the weights do have uncertainty, and do not exactly align with the 1/3 weight on each month in DJF that would be identical to the DJF-average. As a result of this added uncertainty, the proposed model performance of the reconstruction target is noisier than that of the perfectly specified baseline model (Fig. 2.2a; $R = 0.86$ and 90% coverage of 0.91).

So far, all model results have been expressed in weighted averages of log precipitation, which is the specified reconstruction target. However, this metric is not typically of interest by itself; rather, in addition to learning which months the proxy is sensitive to, we would like to get estimates of the actual (not log-transformed) precipitation during a particular season. This is demonstrated in Figure 2.2c, where the learned monthly log precipitation anomalies in DJF are scaled by the monthly calibration standard deviations, added to the monthly calibration means, exponentiated, and averaged to obtain a reconstruction of DJF-mean precipitation in their original units (here, mm/month). While this process slightly degrades the quality of the reconstruction, as expected, it remains relatively good with a correlation of 0.78 and a 90% coverage of 0.84 for the reconstruction period. This ability to generate arbitrary annual-resolution averages can be used as an alternative to fitting a separate version of the baseline model to each season of interest, as was done in Figure 2.1.

The clear identification of DJF as the growth-determining season in Figure 2.2b is optimistic for real proxy data, because the pseudoproxies were constructed using the same statistical relationship as assumed in the reconstruction model, i.e. as a linear combination of log-precipitation plus noise. To assess model performance under misspecification, we use pseudoproxies generated by VS-lite (see Section 2.2.2). This forward model creates a

single noiseless tree-ring width series as a function of temperature and precipitation; we then add noise to obtain pseudoproxies with a specified SNR. There are two types of model misspecification at play when fitting the proposed statistical model to VS-lite pseudoproxies: the model assumes that tree-ring width is a function of log-transformed precipitation when it is a function of soil moisture in VS-lite, and there is a nonlinear relationship in VS-lite between soil moisture and tree ring width, whereas the statistical model assumes linearity. The analogous version of Figure 2.2 for the VS-Lite pseudoproxies with 0.75 SNR demonstrates slightly worse correlation in the posterior mean (0.78 compared to 0.68) and similar frequentist coverage (0.84 for a nominal level of 0.90) (SI Figure S1a compared to Figure 2.2c), which reflects the impact of both forms of model misspecification present in the VS-Lite experiment. Note that the optimal target season for VS-Lite pseudoproxies is not known *a priori*, but the inferred weights are similar to what we might expect from calculating the correlation of each month’s precipitation to tree-ring width (compare SI Figure S1b and S1c).

To better verify the extent to which the model can infer the target season under misspecification, we perform a number of sensitivity tests. For all tests, we use a truncated version of the VS-lite pseudoproxies where growth is set to zero outside of DJF, which allows for weight inference comparisons to the linear pseudoproxy experiment. First, we perform the same reconstruction task as above but using the truncated version. In this case, the posterior distribution of the inferred weights have more uncertainty than in the linear case (compare Figure 2.2b and SI Figure S2b), but both correctly identify DJF as the dominant season. Second, to test the sensitivity of the model to proxy noise, we run a series of VS-Lite experiments similar to above, but with more challenging conditions, varying the SNR (at 0.25, 0.50, 0.75). The posterior mean weights look similar at each of these SNR levels, but a higher SNR corresponds to less uncertainty in the posterior (e.g. SI Figure S3a). Finally, we assess sensitivity to the choice of prior weights, varying the prior to be uniform across the year (Fig. S3a), weakly specified to be sensitive to a half year (SONDJF, Fig. S3b), and correctly specified as DJF (Fig. S3c). A weakly informative prior helps reduce uncertainties

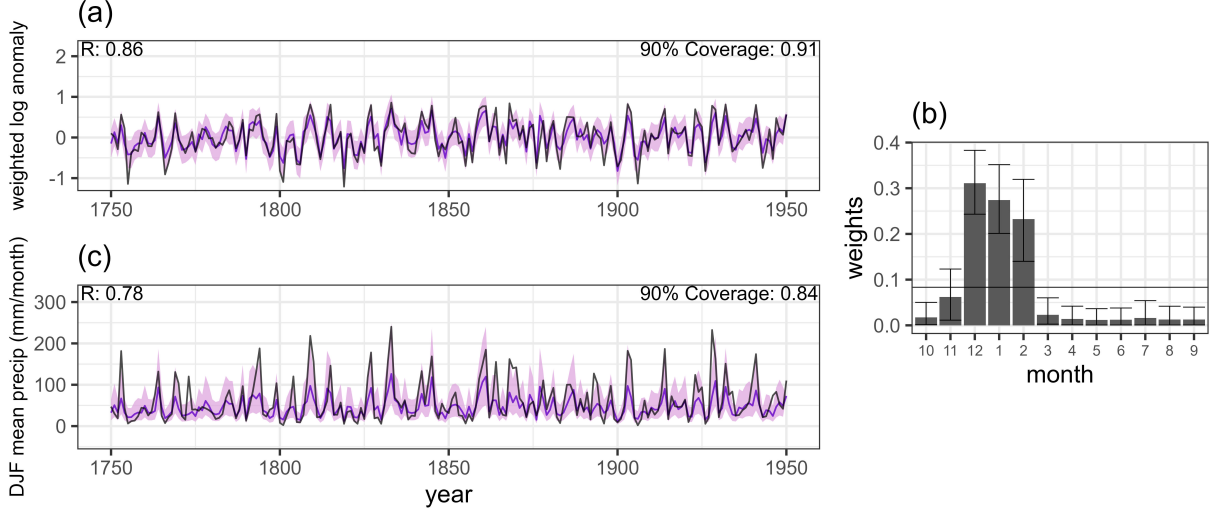


Figure 2.2: Demonstration of the proposed model with linear pseudoproxies. (a) The reconstruction of weighted monthly log precipitation anomalies. The posterior mean reconstruction is shown in purple, with shading for 90% credible intervals. The black line represents the truth, which is obtained by applying the learned weights to monthly climate model output. (b) The posterior mean (bars) of the learned weights for each month in the water year (October - September), with error bars representing 90% credible intervals. A reference line is drawn at height $1/12$, which represents the uniform prior weights across months. (c) The reconstruction of (purple) and true (black) DJF precipitation, which is transformed from the original reconstruction target (log precipitation anomalies).

and aids with model identification, especially for parameters for months with small weights (e.g. comparing Fig. S3a and Fig. S3c).

2.3.3 Case study: California blue oaks

Blue oak trees growing in Southern California have been successfully used to reconstruct cool season precipitation [31, 34, 33, 35]. This prior information allows us to evaluate our statistical model on a real proxy archive through the comparison of this expert knowledge to the model’s estimate of the growth-determining season. In practice, this information should be used to inform the prior distribution on the weights to improve the reconstruction, but we utilize the weak uninformative uniform prior ($w_{input} \equiv 1$) in this case study to test what the model learns from the data alone. For comparison, we also fit a baseline Bayesian

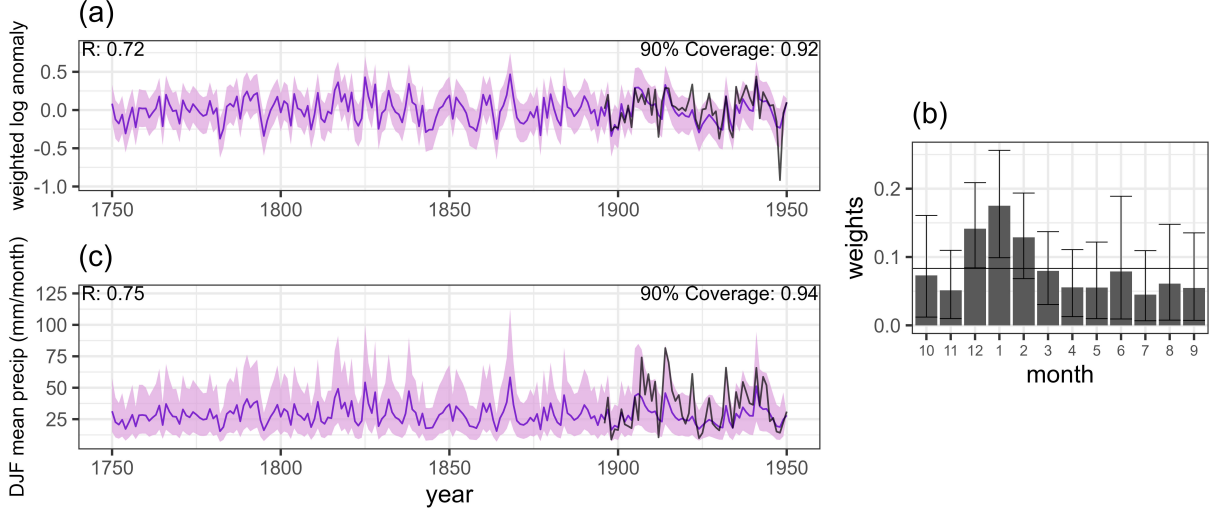


Figure 2.3: Application of the proposed model to blue oaks. Similar to Figure 2.2, panel (a) shows the posterior mean and 90% credible interval of the targeted weighted log precipitation anomaly average in purple alongside instrumental weighted log precipitation anomalies in black, panel (b) shows the posterior mean weights with 90% credible intervals, and panel (c) shows the DJF reconstruction.

hierarchical model in Eqn. (2.1), pre-specifying DJF as the target season and including a model discrepancy term. The model is trained using observed precipitation from 1950-2004, and validated on 1895-1949. The resulting reconstruction from the proposed model is shown in Figure 2.3 alongside instrumental precipitation. The DJF reconstruction skill matches what we would expect from previous work [33], and is similar between the two methods (posterior mean correlation of 0.75 and frequentist coverage of 0.94 for a nominal 0.9 level in Figure 2.3c compared to correlation of 0.71 and coverage of 0.87 in SI Figure S4). The estimation of the target season is noisy (all 90% credible intervals have widths larger than 0.1), but the cool season is still identified as the dominant growth-determining season, with remaining months in the water year still contributing to the target reconstruction. This is also consistent with expert findings.

2.4 Discussion

We propose a new climate reconstruction method which explicitly estimates the optimal reconstruction season on monthly timescales. The advantage of the proposed model is its ability to simultaneously (1) estimate the climate target for the reconstruction and (2) separate uncertainty in growth-dominant season from other sources of noise in the final reconstruction. Pseudoproxy experiments demonstrate that the impact of misspecified seasonality can manifest in both decreased reconstruction skill and unreliable uncertainty estimates (Figure 2.1), whereas our model can help avoid this class of misspecification error. The separation of uncertainty enables the reconstruction of any annual climate aggregate without having to fit a separate model to that particular season. These reconstructions may still have widely varying degrees of skill, depending on the strength of the paleoclimate signal in the proxies, but provide principled uncertainty quantification.

The model presented in this paper is relatively simple: we consider reconstruction of a single climate variable at a single grid cell from a single proxy type (TRW) and assume a linear climate-proxy relationship. The model’s simplicity allows for reconstruction and model parameter estimation to be done simultaneously; however, the model could be extended to multiproxy, spatially extensive, and nonlinear conditions. Multiproxy and climate field reconstructions would benefit the most from the use of seasonally inferring models [11, 20], especially when different proxy archives are sensitive to different seasonal windows, which is often the case in tree-ring based studies in the Northern Hemisphere [60, 61, 62, 63]. Due to the model’s ability to simultaneously learn a seasonal window and reconstruct annualized climate targets, information about the desired reconstruction target season would be efficiently shared between proxies depending on their seasonal SNRs. For instance, a tree-ring based reconstruction of average Northern Hemisphere annual temperature could involve fitting the proposed model at each proxy site: the learned weights would indicate the seasonal window a proxy is sensitive to, and the model would simultaneously produce a reconstruction of mean annual temperature that reflects the extent to which that seasonal window informs the

annual average (a source of uncertainty for Northern Hemisphere summer sensitive terrestrial proxies shown in Figure S6 of [11] and discussed in Section 3 of [20]).

Additionally, the reconstruction approach used here could be improved by using more complex forward models [64, 57]. Linear forward models are convenient approximations, but are unlikely to fully capture the climate-proxy relationship [65, 66]. Model discrepancy terms (η_t) are added to our model to account for this approximation error, but better representations of the proxy-climate relationship would decrease this model discrepancy and improve reconstruction skill. The main challenges associated with the use of more complex forward models are computational in nature; inverse models struggle with model identifiability, particularly if reconstruction is done simultaneously alongside the calibration. Various approaches to reducing the runtime of Bayesian models have been taken in the paleoclimate context (e.g, [58, 39, 43, 21]), but efficient Bayesian computation for inverse problems is an active area of research. Two strategies for improving runtime are the incorporation of prior knowledge, and the separation the calibration and reconstruction tasks: fitting our model on the calibration period alone produces a seasonal window estimate which can be used in any yearly reconstruction model, as is commonly done in data assimilation-based reconstructions [67, 23]. Adapting our model to work alongside existing reconstruction approaches would add uncertainty quantification to the calibration and estimation of a target season, which represents a single step towards fully accounting for the many sources of uncertainties in paleoclimate reconstructions needed to more accurately represent our understanding of climate in the past [20, 4].

Analysis was done in R while models were fit using the probabilistic programming language Stan. The R package `prism` was used to quickly access precipitation data from PRISM [68], and the `VSLiteR` package was used to generate the VS-lite pseudoproxies [57]. The CESM1 Last Millenium Ensemble run used for the pseudoproxy experiments is available via the Earth System Grid (<https://earthsystemgrid.org>), and the tree-ring width data is available via the NCEI Paleo Data Search (<https://www.ncei.noaa.gov/access/paleo-search/>

[study/8558](#) and <https://www.ncei.noaa.gov/access/paleo-search/study/8553>). All code necessary to reproduce these results (including more detailed data access information) are available at https://github.com/nathanhwangbo/trw_seasonal_window.

Chapter 3

Towards Principled Uncertainty

Quantification in Paleoclimate Spatial Field Reconstructions

3.1 Introduction

The climate system is highly complex and includes processes which occur on varying spatial and temporal scales. A perennial challenge in climate science is attempting to understand these processes (and our effects on them) given an instrumental climate record which exists only for the past century and a half. A longer record would allow for better quantification of natural climate variability and the contextualization of current climate events in history. One approach to this problem involves the use of natural archives which record climate (e.g. tree ring widths and ice cores). Given that these archives (known as paleoclimate proxies) have existed long before modern weather instruments, they can be used to extend the instrumental climate record further into the past. The task of inferring a spatially complete record of past climate from paleoclimate proxies is known as climate field reconstruction.

Paleoclimate data assimilation (PaleoDA) is an increasingly popular method of climate

field reconstruction which has been used with great success in recent years (see [69] for a review). Conceptually, this method uses the output of physics-based climate models to produce a prior estimate of past climate, and treats proxy values as imperfect measurements of the climate in a Kalman filter to update this background state. Offline PaleoDA is the most common variant of this method, which uses the same background state at each time point so that there is a single Kalman update (details in Section 3.3.1). Offline PaleoDA is computationally quick, straightforward to implement, and can be used to simultaneously estimate multiple climate fields over long timescales (e.g. [67, 70]).

Inferring climate from paleoclimate proxies comes with many challenges. These archives are noisy recorders of their local climate (i.e the climate that the proxy experienced), as proxy formation is also dependent on non-climatic sources of variability. The causal direction of the proxy-climate relationship is that climate influences proxy formation, so that inferring past climate requires the inversion of a forward model (that is, a model which estimates proxy formation from climate). Furthermore, the forward model is often unknown and can depend on multiple climate variables, so that special care must be taken to ensure that the simultaneous estimation of the forward model and the inverse problem is well posed. In addition, these archives are unevenly distributed over both time and space, so that climate field reconstruction also includes the interpolation of climate information to spatial locations which have no proxy measurements at varying points in time. Ideally, methods for climate field reconstruction should be able to account for each of these uncertainties. In this work, a series of Bayesian hierarchical models (BHM) are used to examine these sources of uncertainty in climate field reconstruction and understand how the assumptions PaleoDA makes about how these sources impact the reconstruction.

BHMs have been used previously for climate field reconstruction (e.g. [71, 39, 72, 73]), though not nearly as frequently as data assimilation. These models are “hierarchical” in the sense that they stack multiple estimation problems on top of one another, resulting in a reconstruction which integrates uncertainty from each layer in the hierarchy (see [27, 5]

for a review of the Bayesian statistical framing of paleoclimate reconstruction). As we will discuss, this approach shares much in common with PaleoDA, but is more flexible and more naturally quantifies uncertainty. These theoretical benefits come at the cost of increased computational burden and difficulty with model identifiability, especially when the number of learned parameters is very large relative to the amount of available information. These challenges are often prohibitive for large reconstruction projects and largely explain the preference in the literature towards PaleoDA for climate field reconstruction.

In this work, we bridge the gap between PaleoDA and BHM, using a series of BHMs to evaluate modeling assumptions and their tradeoffs with computational tractability. We demonstrate these techniques on reconstructions of surface temperature in the Common Era, but the ideas discussed here generalize to other reconstruction contexts. In Section 3.3.1, we write the statistical model underlying PaleoDA as an estimator of the posterior mean and variance of a probabilistic reconstruction model. We note that the primary computational gains of this approach come from the use of a fixed spatial covariance matrix, and show that similar gains can be accomplished in a BHM by similarly specifying the covariance matrix. In Section 3.3.2, we extend the underlying PaleoDA model to a hierarchical Bayesian model which simultaneously learns the calibration parameters alongside the reconstruction. This model is used to highlight the challenge of propagating calibration uncertainty to the reconstruction. In Section 3.3.3, we extend the statistical model further to learn spatial covariance in addition to the calibration parameters, and connect this model to BARCAST, a fully Bayesian model for climate field reconstruction [71, 39]. Pseudoproxy experiments are developed to determine the extent to which proxies can learn spatial covariance, and to compare covariance parameterizations. In Section 3.3.4, special attention is paid to the reconstruction of a spatial mean, which is often the quantity of interest in climate field reconstruction studies. We propose several model simplifications that can speed up a BHM when the large-scale mean is the only desired estimand. We demonstrate these techniques by producing reconstructions of temperature in the Common Era, using tree ring widths

(TRW) as the sole paleoclimate proxy type. Broadly, we argue that BHMs and PaleoDA are mutually beneficial: assumptions underlying PaleoDA can be used to simplify and speed up BHMs, and Bayesian inference can be used to test the sensitivity of reconstructions to PaleoDA assumptions when a BHM is impractically slow.

3.2 Data and Methods

The first member of the 2° spatial resolution CESM1 LAST Millennium Ensemble temperature and precipitation products are used for all pseudoproxy experiments [30]. NASA GISTEMP GHCNv4 Temperature 2.5° product is used for instrumental temperature measurements [74, 75], and is re-gridded to match the CESM1 coordinates using bilinear interpolation. Only the North American Southwest (defined as land-centered gridcells between 124-105E and 32-45N) between the years 1600-2003 are used in the experiments. In all experiments, the years 1950-2003 are used for calibration, so that instrumental (or climate model output) temperature are inputs to the statistical model, while the years 1600-1949 are the years targeted for reconstruction. To mimic the spatial sparsity of proxy data in context of the desired field reconstruction, pseudoproxies are generated for the full period 1600-2003 in 11 of the 52 randomly selected CESM gridcells in the North American Southwest, so that 20% of the modeled gridcells are informed by proxies (see Figure 3.2(a) for a map of which gridcells are used). Simple pseudoproxies (denoted Y) are constructed using a linear model of standardized temperature (i.e. mean zero and standard deviation 1 over the entire period 1600-2003, denoted X) as follows:

$$Y_t := \beta_0 + \beta_1 X_t + \epsilon_t$$

$$\epsilon_t \stackrel{iid}{\sim} N\left(0, \frac{1}{0.75}\right),$$

with location-specific β_0 and β_1 generated uniformly between 0 and 3, and with location-specific white noise ϵ_t , whose variance is chosen so that the proxies have a signal-to-noise

ratio of 0.75, which mimic relatively high quality proxies [28]. These pseudoproxies are used for all experiments outside of 3.3.2, which instead uses pseudoproxies based on VS-Lite, a simple biologically motivated model for tree ring growth [55]. The four primary inputs for VS-Lite are the minimum and maximum temperature / soil moisture values required for growth, which are chosen at each location to be equal to the 5th and 100th percentiles for temperature, and the 0 and 10th percentile for leaky bucket [56] derived soil moisture. These quantiles are chosen alongside a strong signal-to-noise ratio of 2 to exaggerate the nonlinear climate-proxy relationships that can arise from VS-Lite (Figure 3.2(b)).

PAGES2K TRW chronologies within the North American Southwest are used for the case studies [11]. This left 22 chronologies which inform 13 of the 52 GISTEMP gridcells (Figure 3.1). 16 sites were classified by PAGES2K as being sensitive to JJA mean temperature, while the other 6 sites were classified as being sensitive to annual mean temperature. All but one chronology captures the period 1600-1983; fewer are available during the remainder of the calibration period, with only 2 chronologies extending to 2003 (Appendix Figure B.1).

3.3 Statistical Models

PaleoDA and all Bayesian models considered here can be written as extensions of the following probabilistic model: Let $Y_{t,s,i} \in \mathbb{R}$ denote the i -th proxy record at year t and spatial location s , alongside climate variable $X_{t,s} \in \mathbb{R}$. Let $X_t := (X_{t,s_1}, \dots, X_{t,s_{n_s}}) \in \mathbb{R}^{n_s}$ denote the desired spatial grid of temperature at time t (which contains n_s gridcells). For simplicity, we will consider a single proxy type (tree ring widths), and a single climate variable (temperature) in this work. Note that climate field reconstructions typically model proxies at the site level, so that $i > 1$ only in the case when multiple proxy sites are located within the same gridcell. Then an initial probabilistic model can informally be written as:

$$Y_{t,s,i}|X_{t,s} \sim N(\beta_{0,i} + \beta_{1,i}X_{t,s}, \sigma_i^2), \text{ for all } (t, s) \text{ with proxy values} \quad (3.1)$$

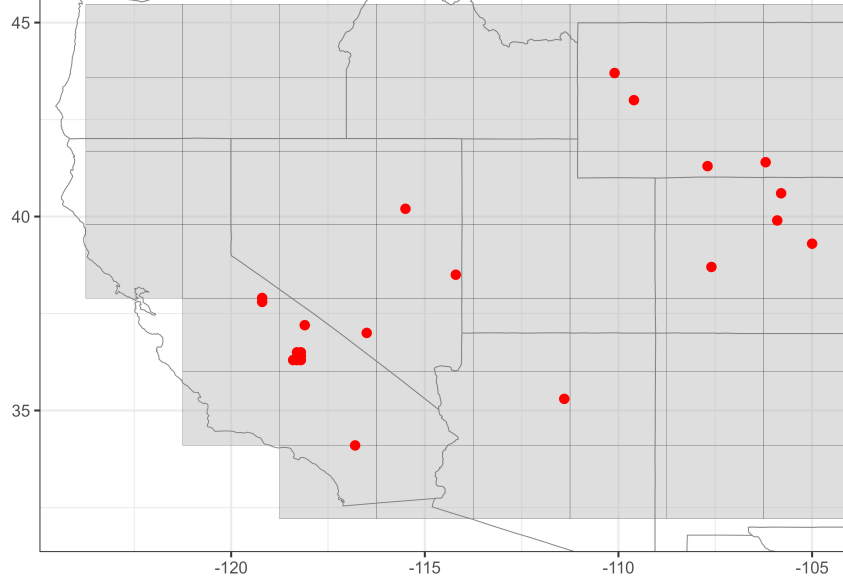


Figure 3.1: Subset of PAGES2K TRW database used for case study. Locations of each TRW chronology is represented by a red point, with the gray boxes representing the gridcells where temperature will be reconstructed. Each chronology is assumed to be sensitive to the temperature in the gridcell which contains it.

$$X_t \sim N_{n_s}(\mu, B), \text{ for all years } t$$

β_0, β_1 , and σ_y^2 are scalars which make up the linear forward model, mapping temperature to TRWs. $\mu \in \mathbb{R}^{n_s}$ is a spatial vector representing the mean state of temperature with corresponding covariance matrix $B \in \mathbb{R}^{n_s \times n_s}$. Note that this presentation of the model abstracts the spatiotemporal availability of each variable: for the experiments in this paper, t is chosen to span the years 1600–2003, but instrumental temperature measurements X_t are only available for the most recent set of observations (e.g. 1880–2017, in the case of the GISTEMP product). The reconstruction task, then, is to infer X_t for t ranging from 1600–2017. On the other hand, at least one TRW chronology is present at each year between 1600–2003, but not necessarily at each gridcell, as each chronology starts and ends at different points in time (Figure B.1). Conceptually, the model can be thought of in two steps: first, TRWs $Y_{t,s}$ which are temporally rich but spatially sparse are used to estimate their local

climate $X_{t,s}$ at all time points prior to the instrumental record (e.g. 1880). Second, this local estimation is spread out spatially to estimate a full spatial field X_t . All of the statistical models discussed below are connected to this initial model, with the differences being choices in which parameters are learned vs fixed, how the model is estimated, and modifications to the functional form of the climate-proxy relationship.

3.3.1 Offline Paleoclimate Data Assimilation

Of the methods discussed here, PaleoDA is distinct in that it heavily utilizes the assumptions of the model in Eqn. (3.1), using the fact that the conjugacy of the normal distribution implies that the posterior distribution $X|Y$ will be normally distributed, that a normal distribution is completely specified by its mean and covariance, and that the proxy-climate relationship is assumed to be linear. These assumptions allow us to analytically determine the posterior distribution $X|Y$, and estimate the distribution by only estimating the posterior mean and covariance.

To start, we can rewrite the first line of 3.1 following [76] with slight notational abuse as $Y_t|X_t \sim N_{n_{sp}(t)}(HX_{t,1:n_{sp}(t)}, R)$, where $n_{sp}(t)$ denotes the number of gridcells which are informed by proxy observations at time t , and $X_{t,1:n_{sp}(t)}$ represents the subset of X_t which matches the spatial availability of Y_t as well as a column of ones for the regression intercept. H contains the regression coefficients $(\beta_{0,i}, \beta_{1,i})$ mapping between proxy and climate, and R is the diagonal matrix containing σ_i^2 along the diagonal.

If μ , H , R , and B are fixed, then through Bayes' theorem, the posterior distribution $X_t|Y_t$ can be straightforwardly expressed as:

$$X_t|Y_t \sim N_{n_s}(\mu + K(Y_t - H\mu), (I - KH)B), \text{ where} \quad (3.2)$$

$$K = BH^\top(HBH^\top + R)^{-1}$$

Offline PaleoDA estimates the posterior mean and covariance of Eqn. (3.2) through what can be thought of as a single-step ensemble Kalman filter, which is sufficient to characterize the posterior distribution due to its normality. This single step involves sampling a series of background states (denoted $X^{\text{prior},j}$), and using proxy data at each time Y_t to obtain an estimate of the climate at time t (denoted $X_t^{a,j}$, which can be thought of as approximate samples from $X_t|Y_t, X^{\text{prior},1}, \dots, X^{\text{prior},m}$). These models are “offline” in the sense that the update equation can be computed in parallel through time, as there is no temporal dependence between updates built into the method – rather, all temporal persistence must propagate from the proxy data itself. PaleoDA fixes H and R , often using the least squares estimates over the time period where X_t and Y_t overlap (e.g. [23]). Offline PaleoDA most commonly uses an Ensemble Kalman Filter [77, 78], where the basic algorithm is as follows:

1. Obtain $X^{\text{prior},1}, \dots, X^{\text{prior},m} \in \mathbb{R}^{n_s}$ by randomly sampling m years from climate model output (CESM1 Last Millennium run in this example)
2. Generate “pseudoproxy” samples $\tilde{Y}^{\text{prior},1}, \dots, \tilde{Y}^{\text{prior},m}$ via the forward model, $\tilde{Y}^{\text{prior},j} := HX^{\text{prior},j} + \epsilon^{\text{prior},j}$, where $\epsilon^{\text{prior},j} \sim N(0, R)$ are independent draws using specified covariance matrix R ($j = 1, 2, \dots, m$)
3. Estimate K via $\hat{K} = \widehat{\text{Cov}}(X^{\text{prior}}, \tilde{Y}^{\text{prior}}) \widehat{\text{Cov}}(\tilde{Y}^{\text{prior}}, \tilde{Y}^{\text{prior}})^{-1}$, where $\widehat{\text{Cov}}(\cdot, \cdot)$ denotes the sample covariance matrix across the m samples. Note that $\widehat{\text{Cov}}(X^{\text{prior}}, \tilde{Y}^{\text{prior}})$ is a sample estimate of $BH^\top$ while $\widehat{\text{Cov}}(\tilde{Y}^{\text{prior}}, \tilde{Y}^{\text{prior}})$ is a sample estimate of $HBH^\top + R$, where B represents the spatial covariance matrix of the prior distribution $P(X^{\text{prior}})$.
4. For each year t : update each prior sample via $X_t^{a,j} := X^{\text{prior},j} + \hat{K}(Y_t - \tilde{Y}^{\text{prior},j})$, where Y_t represents the observed proxy at time t .
5. Treat $X_t^{a,1}, \dots, X_t^{a,m}$ as samples from the posterior distribution $X_t|Y_t$, i.e. using the sample mean as a point reconstruction and using the ensemble spread as a measure of uncertainty.

We note that the procedure most commonly used in PaleoDA is a variation of the above algorithm known as the ensemble square root Kalman filter [79]. The addition of noise $\epsilon^{\text{prior},j} \sim N(0, R)$ to the pseudoproxies in step 2 is used to obtain an unbiased estimator of the posterior covariance, but adds sampling noise to the analysis. The ensemble square root Kalman filter updates the ensemble mean separately from deviations from the mean. A modified Kalman gain matrix, $\tilde{K} := \hat{B}H^\top((\sqrt{H\hat{B}H^\top} + R)^{-1})^\top(\sqrt{H\hat{B}H^\top} + R + \sqrt{R})^{-1}$, is estimated for the deviations, which allows for unbiased covariance estimation without adding noise $\epsilon^{\text{prior},j}$, *while still converging in the large ensemble limit* [80].

Assuming that the climate model distribution is well described by a normal distribution with fixed mean and covariance, the ensemble of $X_t^{a,1}, \dots, X_t^{a,m}$ can be considered samples (albiet, biased samples) from the posterior distribution (3.1), which can be formalized through the connection to a sequential Monte Carlo sampler using equal weights. Adapting the presentation of [81] to the paleoclimate data assimilation algorithm reveals the following: if $X^{\text{prior},j}$ is a draw from the climate model temperature distribution $P(X^{\text{prior}})$, then a general importance sampler for the prior distribution can be approximated as $\hat{P}(X^{\text{prior}}) = \sum_{j=1}^m \delta(X^{\text{prior},j}) w^{\text{prior},j}$, where $\delta(\cdot)$ is the Dirac delta function. EnKF simplifies this by assuming that $w^{\text{prior},j} \equiv \frac{1}{m}$, so that $\hat{P}(X^{\text{prior}}) = \frac{1}{m} \sum_{j=1}^m \delta(X^{\text{prior},j})$. Then by the linearity of the forward model (H) and the normality of the likelihood $P(Y|X)$, it follows that pushing these prior samples through the forward model gives us an estimate of the likelihood, $\hat{P}(Y|X^{\text{prior},1}, \dots, X^{\text{prior},m}) = \frac{1}{m} \sum_{j=1}^m \delta(HX^{\text{prior},j} + \epsilon^{\text{prior},j}) \equiv \frac{1}{m} \sum_{j=1}^m \delta(\tilde{Y}^{\text{prior},j})$. Because this distribution is assumed to be normal, it can be fully characterized by its mean and variance. Therefore, through conjugacy of the normal distribution, the sample mean and covariance of $\tilde{y}$ can be pushed through the Kalman filter update to get samples from $X|Y$. This connection highlights two properties of PaleoDA: first, the assumptions of both normality and linearity are necessary for these procedures to produce approximate samples of Model (3.1) (also shown empirically in [82]), and second, that even if those assumptions are met, these procedures do not produce unbiased samples from the posterior distribution, as they use uniform $\frac{1}{m}$ weights.

Unbiased sampling would instead require the use of importance sampling weights, which in this case, should be proportional to the likelihood due to the use of the prior distribution to generate samples. As a result of this sampling bias, it is unclear how well the PaleoDA posterior ensemble spread captures reconstruction uncertainty.

We emphasize that in PaleoDA, instrumental climate observations are not used in the update equations 1-5, and are primarily used to help with the calibration (i.e. learning H and R). Instead, the output of climate models is used as the prior climate distribution, so that the spatial covariance of the reconstruction is driven by the dynamics of the climate model(s) used. In addition, the spread in the PaleoDA posterior ensemble depends only on the observed proxy information insofar as the observations are used to inform R (e.g. Theorem 7.2 of [83]). This suggests that understanding uncertainty in the estimation of H and R (which are considered fixed in the Kalman update) is crucial for the interpretation of the posterior spread of PaleoDA. In the following, we use pseudoproxy experiments to address these uncertainties. To this end, we will use a Markov Chain Monte Carlo (MCMC) sampler, a more flexible sampling technique than the ensemble Kalman filter which is able to obtain samples from the posterior distribution of extensions of Model (3.1) which jointly model the distribution of all random quantities.

3.3.2 Uncertainty due to the climate-proxy relationships

PaleoDA assumes that the relationship between proxy and climate is linear, an assumption fundamental to Kalman filters. When this relationship is learned from observational data, existing reconstruction methods either use this assumption directly by fitting a linear model inferring proxies from climate (using proxy records which coincide in time with instrumental climate observations) then using a point estimate of these parameters in the reconstruction model (e.g. [67, 23]), or by using fitting more complex biologically motivated models (known as proxy system models [84]), which are then linearized, with point estimates then used in the reconstruction model [70]. As such, the climate reconstruction produced are conditioned

on fixed values of calibration parameters. Paleoclimate proxies are generally noisy recorders of climate, and the calibration parameters H and R have substantial estimation error [29]. Particular care is given to the residual covariance R , which is either fixed at point estimates from the linear regression calibration values (e.g. [23]) or manually calibrated, using multiple choices for R to assess their impact on the reconstruction [85, 73, 86]. A comparison of offline DA results using a linear forward model and the correctly specified (non-linear) forward model has been done in pseudoproxies experiments [87], which finds that the extent to which using correctly specified non-linear forward models performs better than using a linear approximation (in terms of mean reconstruction skill) depends on the proxy type, identifying an increase in reconstruction skill for tree ring widths. This result is despite the fact that the theoretical underpinnings of DA rely on the linearity of the forward model, so that there are no strong theoretical guarantees on a DA based reconstruction using a nonlinear forward model.

While significant progress has been made on the front of developing better forward models (e.g. [55, 88, 21, 89]), progress on this problem is a monumental task due to the complexities involved with modeling each of the many proxy types (e.g. modeling proxy measurement uncertainty, aging uncertainty, on top of uncertainty in the climate-proxy relationship itself). Here, we develop pseudoproxy experiments to evaluate the impact of using a misspecified linear model on reconstruction uncertainty. To begin, we consider the initial probabilistic reformulation of PDA described in Eqn. (3.1) but treat $\beta_{0,i}$, $\beta_{1,i}$, and σ_i^2 as parameters to be estimated (prior distributions for these parameters are specified in Appendix Section B.1. To estimate this model, we replace the PaleoDA algorithm with an MCMC sampler which is able to jointly estimate these parameters alongside the reconstruction. To isolate the impact of calibration uncertainty in this model, we fix the climate prior parameters μ and B to be the sample mean and covariance of our climate model output (the first member of CESM Last Millennium Ensemble).

To assess the impact of this form of model misspecification, we fit the model on VS-Lite

based pseudoproxies [55] which are constructed to exhibit a nonlinear relationship between TRW and temperature 3.2(b). This setup guarantees that the model (3.1) uses an incorrectly specified forward model. One challenge associated with learning an incorrectly specified forward model jointly with the reconstruction is that with sufficient data, the model will learn the “best fitting line” with very little uncertainty, so that the posterior variance of the linear model parameters will not match the uncertainty in the calibration procedure [54]. The resulting reconstruction from a misspecified model will tend to be overconfident in its predictions. Fitting a correctly specified BHM which uses VS-Lite as the forward model [57] for comparison would not permit us to test the impacts of model misspecification, so we instead opt to test an extension of model (3.1) which adds a zero mean time-varying model discrepancy term which accounts for the bias due to model misspecification. This extension can be written as follows (full model specification in Appendix Section B.1):

$$\begin{aligned}
Y_{t,s,i}|X_{t,s} &\sim N(\beta_{0,i} + \beta_{1,i}X_{t,s} + \eta_t, \sigma_i^2), \text{ for all } (t, s) \text{ with proxy values} & (3.3) \\
X_t &\sim N_{n_s}(\mu, B), \text{ for all years } t \\
\eta_t &\sim N(0, \sigma_\eta^2), \text{ for all years } t \text{ with proxy values}
\end{aligned}$$

Without the model discrepancy term, the posterior variance of the reconstruction tends to be too small. On average (across gridcells), the 90% credible intervals of the reconstruction contained only 67% of the out of sample years. (Figure 3.3(c)). If the model were to have correct frequentist coverage, then we would expect the 90% credible interval to contain 90% of the out of sample years. Adding in the model discrepancy term improves this coverage to an average of 79% (Figure 3.3(d)). Even with the addition of this model discrepancy term, the reconstruction tends to be overconfident, which comes from two sources. First, the addition of the model discrepancy term led to improved coverage in 80% of the 52 gridcells, and improved coverage in all 11 of the gridcells which contain proxy data. This result is unsurprising, as the

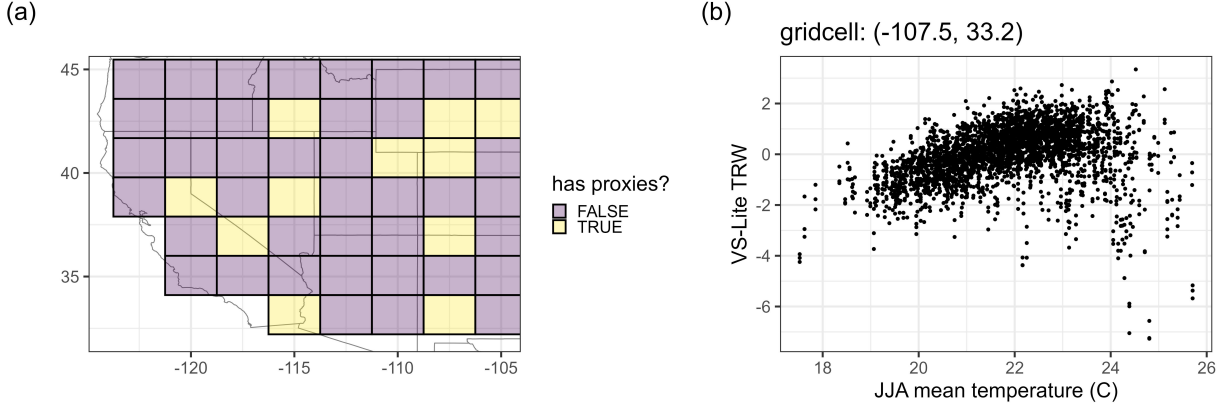


Figure 3.2: Pseudoproxy experiment setup. (a) The grid represents the spatial domain used for all pseudoproxy experiments, with yellow cells representing locations where pseudoproxies are generated. (b) An example of one such pseudoproxy used in a VS-Lite based PPE. Each point represents a different year, with the simulated TRW on the vertical axis and JJA mean temperature on the horizontal axis.

model discrepancy term should have the most immediate impact on gridcells which contain proxy data, rather than relying on the prior spatial covariance to obtain a reconstruction. This suggests that the sparse spatial availability of the proxies limits the influence of the model discrepancy term. The other source of undercoverage comes from the use of a single η_t shared across space, which does not allow for model discrepancy to be fully captured by the model. Ideally, each proxy location should have its own model discrepancy term to account for the varying degrees of misspecification. However, in order for a spatial η_t to be identifiable, there must be multiple proxies at each location. The experiments shown here mimic the common setup where chronologies (rather than individual proxy measurements) are used for the reconstruction, so that there are too few observations to learn a spatially varying η_t . However, reconstructions which model individual proxies instead of preprocessed chronologies would allow for such a term to be learned and for model discrepancy to be more

fully accounted for. The model discrepancy term generally decreases the skill of the posterior mean, measured here as the correlation between the posterior mean and the true climate model output temperature over the out-of-sample years 1600-1949 (comparing Figure 3.3(a) and (b)). The posterior mean reconstruction typically has compressed variance compared to the truth [71], and the model discrepancy term further acts to flatten the posterior mean in order to obtain the correct coverage. The extent to which this is true will depend on the degree of model misspecification (note that the experiment here constructs pseudoproxies which exaggerate model discrepancy).

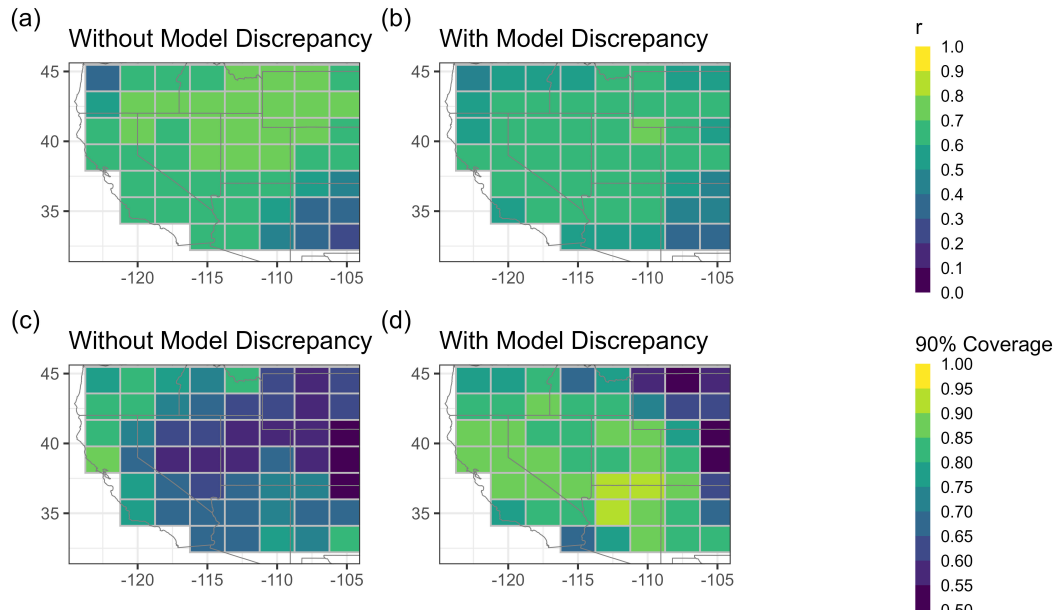


Figure 3.3: Out of sample (1600-1950) evaluation of an idealistic PPE ($\text{SNR} = 2$) looking at the impact of modeling the proxy-climate relationship. The top row shows the gridcell-by-gridcell correlation between the posterior mean reconstruction and the truth (i.e. climate model output), while the bottom row shows the frequentist 90% coverage over this period. The columns show the output of Bayesian models which learn the calibration alongside the reconstruction, including (right) and excluding (left) a model discrepancy term.

3.3.3 Uncertainty due to spatial covariance model

Another key modeling choice central to PaleoDA is the use of climate model output to obtain the "background" (i.e. prior) climate state. While there have been multiple methods proposed to obtain a background state from a climate model [90], we consider the simple case where temperature fields from randomly sampled years of the climate models are used to calculate a sample mean and covariance to estimate the Kalman gain (denoted $\hat{K}$ above), and the same samples are updated using the observed proxy data to obtain the reconstruction (denoted $Y_t - \tilde{Y}^{\text{prior},j}$ above). The use of climate model output to obtain a prior covariance is of particular interest, given that it does not make use of any instrumental or proxy observations. The decision to use a fixed covariance matrix is a rather strong assumption, but if reasonable, can considerably speed up common computational bottlenecks in BHMs. The natural statistical alternative to this approach given model (3.1) involves learning a covariance matrix using some combination of the instrumental and proxy observations, often specifying some functional form for the covariance. However, methods for learning covariance matrices from a normal model can be prohibitively slow on large domains, as the likelihood evaluation requires the inversion of the covariance matrix, an operation which scales cubically with the number of gridcells computationally and quadratically in memory. This operation must be done for each iteration of a MCMC sampler, making MCMC-based methods for paleoclimate reconstruction (like the BHMs shown here) orders of magnitude slower than approximate methods like PaleoDA. Furthermore, even if computation was not a problem, it is possible that using climate model sample covariance might illustrate better climate dynamics than a highly structured learned matrix (e.g. the squared exponential covariance). Timescale, the climate variable of interest, and the spatial domain are all important factors determining the importance of the covariance matrix. For reconstructions of the deep past, the true spatial covariance structure has undoubtedly changed in drastic ways such that any choice of a fixed, stationary covariance matrix is inappropriate. Different climate variables operate on widely varying spatial scales with varying degrees of temporal persistence, so that the

appropriateness of parametric assumptions on learned covariance matrices will be application specific. To begin to understand the implication of this modeling choice of PaleoDA, we consider the following questions in the context of small-scale temperature reconstructions of the Common Era:

1. To what extent can spatial covariance be learned from proxies?
2. If we choose to learn the spatial covariance from data, how much does the assumed parametric structure of the covariance matrix change the resulting reconstruction?

Learning spatial covariance from proxies

One defining characteristic of the PaleoDA approach is that spatial covariance is not learned by instrumental or proxy observations. Instead, information about how temperature varies over space is taken from the output of a climate model. In contrast, BHMs which learn a covariance matrix do so by assuming stationarity in time, which allows both the observed instrumental temperature data and the inferred temperature from the proxy values to inform spatial covariance. The extent to which these methods are preferable to the PaleoDA approach depend on the relative contributions of the proxy data to the learned covariance. If the proxy records are too noisy and spatially sparse to inform spatial covariance, then the computational savings associated with the PaleoDA approach are a reasonable simplification. On the other hand, if proxy records can meaningfully constrain spatial covariance, then BHMs should be developed which learn spatial covariance, with parameterizations which allow for changes through time. To disentangle the relative influences of instrumental record and proxy records on learned spatial covariance matrices, we construct the following pseudoproxy experiment:

1. At each grid cell from the first member of CESM1 LME, standardize the JJA average temperature such that each time series has mean zero and unit variance in both the reconstruction period (1600-1949) and the calibration period (1950-2003).

2. In the reconstruction period, “mislabel” the locations of each standardized temperature time series, randomly permuting this segment of the time series across spatial locations.
3. Construct linear pseudoproxies from these new time series, following the experimental procedure described in Section 3.2 (e.g. using 11 gridcells).

The spatial covariance of the temperature field underlying the pseudoproxy construction follows that of the original climate model output temperature data during the calibration period (where the statistical model is given access to “instrumental” temperatures), but temperature has a very different covariance structure in the reconstruction period (where the statistical model is not given access to “instrumental” temperature), as the locations are randomly permuted (Figure 3.4 (b) and (c)). The BHM with model discrepancy term in Eqn. (3.3) can be further extended to learn the spatial covariance matrix by adding another layer to the hierarchy:

$$Y_{t,s,i}|X_{t,s} \sim N(\beta_{0,i} + \beta_{1,i}X_{t,s} + \eta_t, \sigma_i^2), \text{ for all } (t, s) \text{ with proxy values} \quad (3.4)$$

$$X_t \sim N_{n_s}(\mu, B), \text{ for all years } t$$

$$\eta_t \sim N(0, \sigma_\eta^2), \text{ for all years } t \text{ with proxy values}$$

$$B_{s_1, s_2} = \begin{cases} \sigma_x^2 \exp\left(-\frac{\|s_1 - s_2\|^2}{2\rho^2}\right) & i \neq j \\ \sigma_x^2 + \tau^2 & i = j, \end{cases}$$

where $\|\cdot\|$ refers to geodesic distance. The scale σ_x , the lengthscale ρ (determining the smoothness of the spatial decay), and the nugget effect τ are all parameters learned in the model. This new layer specifies a squared exponential covariance matrix which is both stationary in space and time – that is, the spatial covariance of temperature is assumed to be the same at all time points, and the covariance between two points depends only on spatial distance between the two, regardless of their absolute positions. The assumption of

temporal stationary of the covariance structure is purposefully misspecified in this experiment, since the pseudoproxies were constructed to experience two distinctly different covariances between the instrumental and reconstruction periods. This magnitude of the difference in the covariance is unrealistic, but allows us to clearly test whether it is possible to learn covariance information from the proxies. This model is fit on this pseudoproxy data, and the posterior mean covariance matrix is then converted to a correlation matrix and compared to the spatial correlation matrix of the CESM LME temperature run before and after shuffling (Figure 3.4). Note that there are 53 years in the calibration compared to 349 years in the reconstruction, so that if proxies were perfect recorders of temperature, and if proxies were available at each gridcell, then we would expect the learned covariance matrix to more closely resemble the reconstruction covariance (Figure 3.4 panel (b)) than the calibration (panel (c)). However, the proxies are both noisy recorders of climate and sparsely spatially available (11 of the 52 gridcells). As a result, the learned posterior mean is more informed by the instrumental temperature observations than proxies (Frobenius distances between the posterior mean correlation matrix are 245 and 98, for the reconstruction and calibration periods, respectively). To test whether the proxies have any influence on the covariance matrix, we compare these results to another experiment where locations are not shuffled between the instrumental and reconstruction periods. In this experiment, the learned covariance matrix yields an even closer fit to the instrumental covariance, indicating that the pseudoproxies exert weak influence on the covariance matrix (comparing Figure 3.4(c) to Figure 3.5(c)).

Learned spatial covariance parametrization

In the above experiment, the pseudoproxies experience a dramatic structural shift in spatial covariance structure that is not present in the instrumental data. While useful for assessing the model’s ability to learn from the proxies alone, a structured parametrized covariance matrix like the squared exponential is unlikely to be able to learn such a change. The ability of a model to represent large shifts would be useful for paleoclimate reconstruction

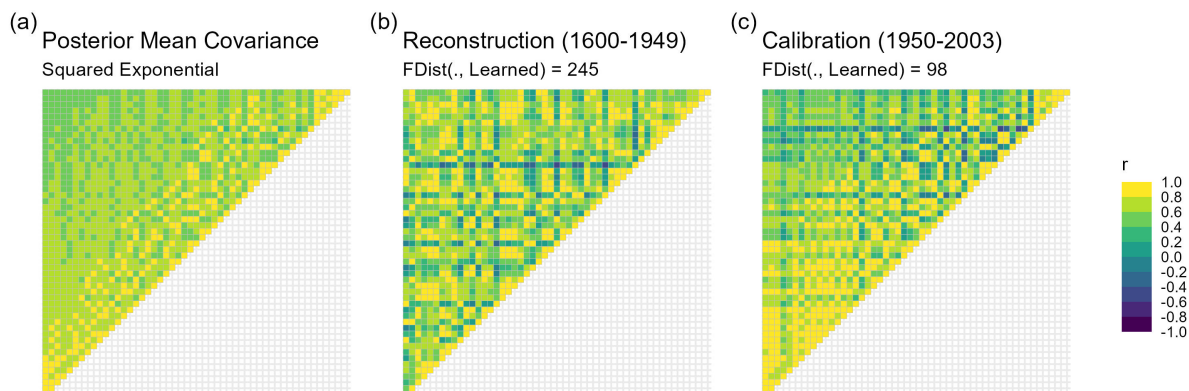


Figure 3.4: PPE learning covariance from proxies which experience different spatial covariance in the reconstruction and calibration periods. (a) The learned posterior mean spatial correlation matrix of temperature using a squared exponential parametrization. (b) The sample temperature correlation matrix in the reconstruction period (i.e., with shuffled location labels) (c) The sample temperature correlation matrix in the calibration period (i.e., with correct location labels). The subtitles report the Frobenius distances to the learned posterior mean correlation matrix in panel (a).

in deep time, which has experienced such changes in covariance. It is possible that a more flexible covariance parametrization would be more capable of inferring changes in spatial covariance from the proxies. Here, we consider a Lewandowski-Kurowicka-Joe (LKJ) prior on the correlation matrix [91] and independent diffuse half-normal distributions for the variances, a popular covariance model which decomposes covariance into correlations and scale [92].

The LKJ prior is parameterized with shape parameter set to 0.5, which weakly encodes the prior knowledge that temperatures are expected to be correlated in space. We assess the feasibility of this flexible parameterization in the standard pseudoproxy setup (i.e., without any spatial shuffling). Comparing this parameterization to the squared exponential kernel in the model described in Eqn. 3.4, we find that the more flexible parameterization is unable to learn the covariance matrix, instead staying near the prior of a uniform correlation matrix (Figure 3.5 (b)). For the use case presented here (mean JJA temperature fields in the North American Southwest), the squared exponential parameterization is a suitable model, adding considerable computational speedup on top of a better model fit (comparing Figure 3.5 (b) and (c)). The appropriateness of this structural assumption will depend heavily on the variable of interest, the spatial region, and the spatial resolution of the data. Experiments like the one shown here are in the Bayesian context are an important step of validating different covariance parametrizations, when the spatial covariance is chosen to be learned from data.

Linking PaleoDA to existing BHMs

Perhaps the most popular climate field reconstruction technique which properly samples from a posterior distribution is BARCAST [71, 39], which has been studied in pseudoproxies [93, 94, 41] as well as in observations [72, 95, 96]. BARCAST can be informally written as follows:

$$(X_{t+1} - \mu) | X_t \sim N(\alpha(X_t - \mu), B_t), \text{ for all years } t$$

$$Y_t | X_t \sim N(\beta_0 + \beta_1 X_t, \sigma_y^2), \text{ for years when proxy data } Y_t \text{ is available}$$

$$\tilde{X}_t | X_t \sim N(X_t, \tau^2), \text{ for years when instrumental temperature } \tilde{X}_t \text{ is available}$$

$$B_{s_1, s_2} = \sigma_x^2 \exp \left\{ \frac{\|s_1 - s_2\|^2}{-2\rho^2} \right\}$$

In comparison to the model (3.4), an autoregressive coefficient(1) term is added to the temperature field, instrumental measurement noise is added in place of the nugget effect τ , but the proxy-climate relationship is assumed to be the same for all proxies (i.e. $\beta_0, \beta_1, \sigma_y^2$ are

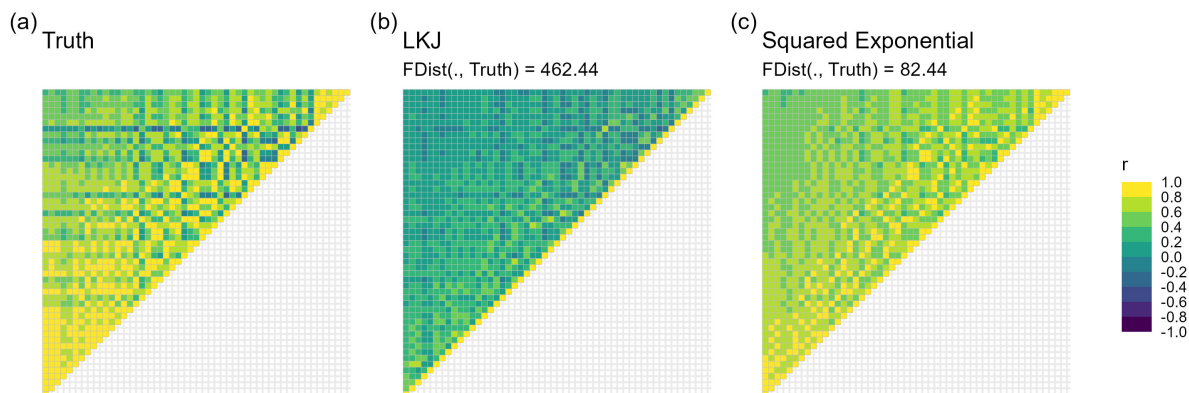


Figure 3.5: PPE learning covariance from proxies. (a) The sample spatial temperature correlation matrix from climate model output. (b) The learned posterior mean correlation matrix using an LKJ(0.5) prior, and (c) a squared exponential covariance function prior. The subtitles report the Frobenius distance between the posterior mean covariance matrix and the true covariance matrix.

the same for all proxies), and no model discrepancy term is included. Similar to the previous BHMs but in contrast to PaleoDA, all parameters (H, B, R, μ) are simultaneously estimated using instrumental temperatures and proxy data.

In our implementation of the model, we omit the instrumental measurement noise variance, instead fixing it at the value 0.015 for the pseudoproxy experiments (a value within the range of estimated measurement error in the original paper). Because we use a single temperature product (and therefore do not have repeated instrumental measurements of temperature at

any year t), the identification of the instrumental measurement error variance τ^2 and the latent time series X_t relies heavily on proxies with overlap with the instrumental period, requiring the inverse model to produce “repeated measurements” of X_t which are of sufficiently high quality. When the proxies are noisy and the forward model is uncertain (especially in the presence of model discrepancy), the model is unable to distinguish the measurement error τ^2 from variability in latent X_t due to uncertainty in inversion of the forward model. In general, instrumental measurement error will be best learned using multiple instrumental products in the analysis to obtain true repeated observations.

The addition of the autoregressive term α marks a conceptual change from the previously discussed models, in that BARCAST explicitly models a temporal process for temperature. BARCAST is therefore able to better propagate uncertainty through time, allowing the posterior reconstructions to better be thought of as trajectories [97]. Again, we note that in the usual context of the uncertain inversion of noisy proxies, the identifiability of α depends on the instrumental record rather than from temperature inferred from the proxies. We consider a BHM which combines BARCAST with model 3.4:

$$\begin{aligned}
Y_{t,s,i}|X_{t,s} &\sim N(\beta_{0,i} + \beta_{1,i}X_{t,s} + \eta_t, \sigma_i^2), \text{ for all } (t, s) \text{ with proxy values} & (3.5) \\
(X_{t+1} - \mu)|X_t &\sim N_{n_s}(\alpha(X_t - \mu), B_t), \text{ for all years } t \\
\tilde{X}_t|X_t &\sim N(X_t, \tau^2), \text{ for years when instrumental temperature } \tilde{X}_t \text{ is available} \\
\eta_t &\sim N(0, \sigma_\eta^2), \text{ for all years } t \text{ with proxy values} \\
B_{s_1, s_2} &= \sigma_x^2 \exp \left\{ \frac{\|s_1 - s_2\|^2}{-2\rho^2} \right\}
\end{aligned}$$

This model modifies BARCAST by making the calibration proxy-specific, adding a model discrepancy term, and using user-selected measurement error variance τ^2 (note that this measurement error term can be omitted and replaced with the nugget effect of model (3.4) to ensure that the resulting covariance matrix is positive definite). This model, BARCAST, and

all other models discussed in Section 3.3.3 learn spatial covariance from both the instrumental record and proxy data, making them much slower computationally than methods which use a fixed spatial covariance. Improving spatial covariance specification can be especially important for the reconstruction of a spatial mean, which is very often a target of paleoclimate reconstructions. In the simplest case, the spatial mean is estimated by averaging posterior estimates from the entire field. Given the sparsity of paleoclimate proxies in space, the spatial covariance specification is critical for the infilling of the reconstruction in gridcells without proxy data. In our pseudoproxy experiments, a reconstruction of the spatial mean from model (3.3) using covariance fixed at the sample instrumental record covariance tends to be overconfident compared to model (3.5) which learns a squared exponential spatial covariance matrix (90% frequentist coverage of 0.82 in Figure 3.6(a) compared to 0.91 in panel (b)).

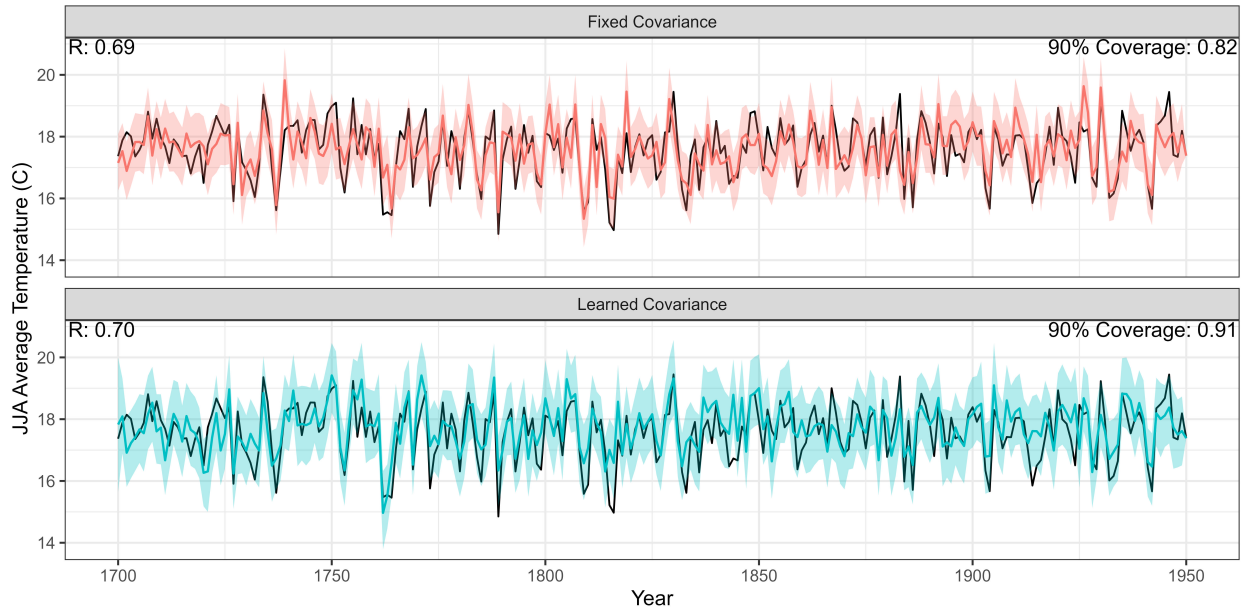


Figure 3.6: PPE spatial mean reconstruction of JJA mean temperature using a Bayesian model which fixes covariance at the sample instrumental mean (top), and one that learns covariance from the data, using a modified BARCAST model (bottom). The posterior mean (colored lines) and 90% credible intervals are shown, alongside the true climate model run spatial mean.

Case Study using PAGES2K TRW in the American Southwest

The model (3.5) is also assessed with real tree ring width chronologies from the PAGES2K proxy database [11], limiting the proxies to match the same study region as the pseudoproxy experiments (details in Section 3.2). The posterior mean reconstruction skill, measured by correlation over the out-of-sample period 1880-1949, varies between gridcells, ranging between 0.21 and 0.53 (Figure 3.7(a)). The gridcells with lowest correlation are in the northwest quadrant of the region, where there are no proxy observations (comparing Figure 3.1 and 3.7(a)). This result is unsurprising, demonstrating that reconstruction skill will tend to be best in gridcells which are directly informed by proxy observations, rather than relying on the learned spatial covariance to obtain a reconstruction. The 90% credible intervals cover between 67% and 0.95% of out of sample years depending on the gridcell, with an average of 82% (Figure 3.7(b)). This tendency for individual gridcells to have coverage lower than the nominal rate is likely due to forward models which are not sufficiently flexible. Most gridboxes are informed by only a single proxy chronology (Figure 3.1), so that the model discrepancy term is learned by assuming that the model discrepancy is the same at all gridcells. As a result, uncertainty in the forward model is not being able to fully propagate back to reconstruction. The posterior draws of the gridded reconstruction are also averaged to obtain a time series reconstruction of the spatial mean. The posterior mean has with a correlation of 0.50 with the true spatial average over the out-of-sample evaluation years (1880-1950), but has noticeably damped variance compared to the true spatial average (3.7(c)). The shrinkage of the posterior mean towards zero is unsurprising, and is a characteristic shared by BHMs for climate field reconstruction [71]. The credible interval has the correct nominal coverage over the out of sample period, containing 90% of the true out of sample years.

3.3.4 Large-scale average reconstruction

Large-scale spatial averages are often the quantity of interest in a climate field reconstruction and are worth special consideration as an estimand. In particular, global and hemispheric

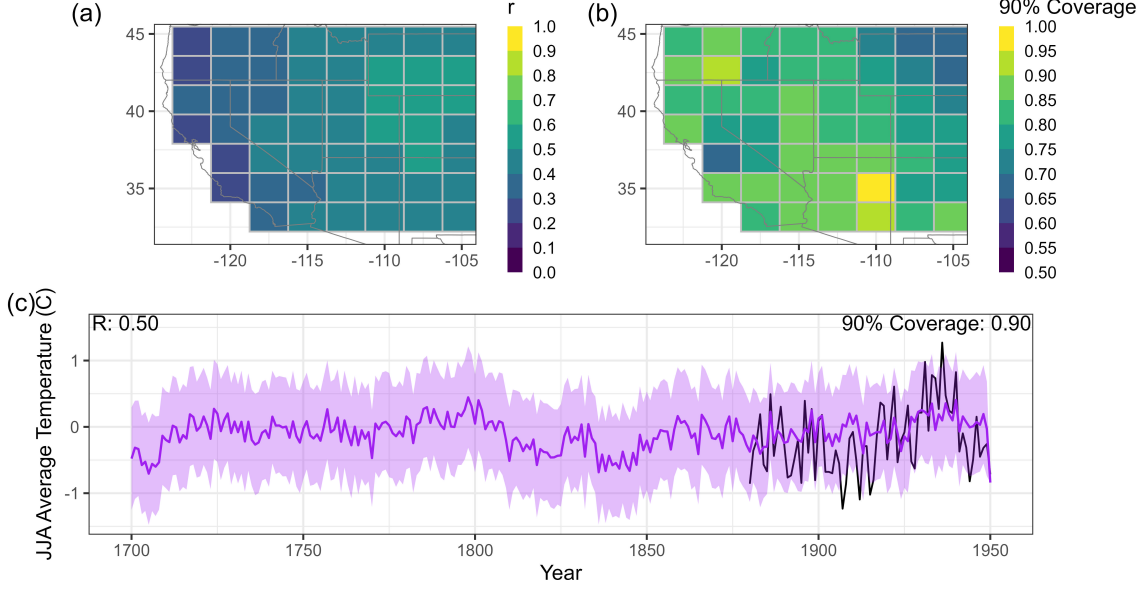


Figure 3.7: North American Southwest reconstruction summary using PAGES2k tree ring width chronologies. Maps of per-gridcell (a) correlation and (b) 90% frequentist coverage between the posterior mean reconstruction and GISTEMP instrumental surface temperature across all out-of-sample years with instrumental surface temperature (1880-1949), alongside (c) the spatial mean reconstruction

reconstructions of the Common Era have both been the subject of much research interest [20]. As previously discussed, the estimation of a spatial mean is complicated by the fact that proxy observations are typically only available in a small fraction of the available grid cells, requiring a spatial covariance matrix to spread out the local information learned from proxies to the rest of the spatial field. This problem of estimating spatial averages from spatially incomplete data is nontrivial even when instrumental observations are available, and is known to be sensitive to the choice of spatial covariance [98]. One common approach is to fill in missing values of the grid and then calculate an average on the spatially complete field. This approach is natural in the context of a climate field reconstruction: given a gridded probabilistic climate field reconstruction, the posterior distribution of the simple (i.e. ignoring area-weighting) spatial mean can be estimated by averaging each posterior draw of the climate field, i.e. $\text{spatial mean}_t^{(m)} = \frac{1}{n_s} \sum_{i=1}^{n_s} X_{t,i}^{(m)}$ for each posterior sample m .

Given that proxies provide limited information about spatial covariance (Section 3.3.3) and that learning covariance is the main computational bottleneck of BHMs for climate field reconstruction, it is worthwhile to consider modifications to the BHMs which avoid having to learn spatial covariance when the spatial mean is the sole estimand of interest. The most natural approach makes use of the normality assumption on the temperature field, using the well-known property that subsets of a multivariate normal vector are also normal after conditioning on the complement. That is, if the temperature field at time t is reorganized and decomposed as $X_t \equiv (X_{t,1:n_{s_1}}, X_{t,(n_{s_1}+1):n_s})$ with $X_{t,1:n_{s_1}}, X_{t,(n_{s_1}+1):n_s} \subset \mathbb{R}^{\kappa \sim}$ denoting temperature at the subset of gridcells with and without observed proxy data, then

$$X_{t,(n_{s_1}+1):n_s} | X_{t,1:n_{s_1}} \sim N(\mu_{(n_{s_1}+1):n_s} + B_{21}B_{11}^{-1}(X_{t,1:n_{s_1}} - \mu_{1:n_{s_1}}), B_{22} - B_{21}B_{11}^{-1}B_{12}), \quad (3.6)$$

where B_{11} and B_{22} represent the principle submatrices corresponding to $X_{t,1:n_{s_1}}$ and $X_{t,(n_{s_1}+1):n_s}$, B_{12} and B_{21} represent the corresponding off-diagonal blocks, and $\mu_{(n_{s_1}+1):n_s}$ and $\mu_{1:n_{s_1}}$ represent the corresponding subsets of μ . A simplified BHM can be used to reconstruct temperatures at locations for which proxy data is available ($X_{t,1:n_{s_1}}$ and $\mu_{1:n_{s_1}}$), and posterior draws of these parameters can be plugged into the above equation alongside fixed B and $\mu_{(n_{s_1}+1):n_s}$ to generate posterior samples of $X_{t,(n_{s_1}+1):n_s}$ from model (3.3).

An alternative approach, proposed for observational data [6], opts not to directly use a spatial covariance matrix at all, instead fitting a spatially unaware ridge regression model to predict the large-scale mean using individual gridcell values as predictors. Ridge regression is a natural choice, as it is adept at handling a large number of correlated predictors (gridcells). We can adapt this approach to the Bayesian paleoclimate context by adding an additional layer to model (3.3):

$$Y_{t,s,i} | X_{t,s} \sim N(\beta_{0,i} + \beta_{1,i}X_{t,s} + \eta_t, \sigma_i^2), \text{ for all } (t, s) \text{ with proxy values}$$

$$\begin{aligned}
X_{t,1:n_{s_1}} &\sim N_{n_{s_1}}(\mu, B), \text{ for all years } t \\
\eta_t &\sim N(0, \sigma_\eta^2), \text{ for all years } t \text{ with proxy values} \\
\frac{1}{n_s} \sum_{i=1}^{n_s} X_{t,i} &\sim N(\beta_r X_{t,1:n_{s_1}}, \sigma_r^2), \text{ for all years } t \\
\beta_r &\sim N_{n_{s_1}}(0, \sigma_b I),
\end{aligned} \tag{3.7}$$

$$\tag{3.8}$$

where β_r, σ_r^2 , and σ_b^2 are the parameters corresponding to the the added layer. This so-called “ridge” prior on β_r provides regularization to 0 so that the posterior mean (and mode) is equivalent to the ridge estimator with $\lambda := \frac{\sigma_s}{\sigma_b}$. The entries of β_r , shown in Figure B.2, describe the contribution of the temperature at each gridcell to the spatial mean. The results of these approaches on linear pseudoproxy data can be see in Figure 3.8, which show that the time series reconstruction of the spatial mean is almost identical to that of model (3.3) fit using the conditional normal distribution, with both posterior mean correlation and coverage being within 0.02 of one another. It is not too surprising that a linear model would work well to estimate the spatial mean, as the expectation of the conditional spatial mean is a linear function of the observed temperature values, with Eqn (3.6) implying that the conditional distribution of the spatial mean can be written

$$\begin{aligned}
\left(\frac{1}{n_s} \sum_{i=1}^{n_s} X_{t,i} \right) \Big| X_{t,(n_{s_1}+1):n_s} &\sim N(M, S), \text{ where} \\
M &:= \frac{1}{n_s} \left(\sum_{i=1}^{n_{s_1}} \mu_{1|2} + \sum_{i=n_{s_1}+1}^{n_s} X_{t,i} \right) \\
S &:= \frac{1}{n_s^2} \left(\sum_{i=1}^{n_{s_1}} \text{Var}(X_{t,i}) + 2 \sum_{1 \leq i < j \leq n_{s_1}} \text{Cov}(X_{t,i}, X_{t,j}) \right) \\
\mu_{2|1} &:= \mu_{(n_{s_1}+1):n_s} + B_{21} B_{11}^{-1} (X_{t,1:n_{s_1}} - \mu_{1:n_{s_1}})
\end{aligned}$$

The choice of which approximation to use will depend on the field being modeled and the available proxy data. If the climate field has a stable, well-behaved spatial covariance structure (like the US Southwest temperature used in this experiment), then assuming multivariate normality with a fixed covariance matrix will be able to infill the spatial grid well, which will lead to quality estimation even when the proxies are sparsely located. Ridge regression will suffer in comparison when there are few proxies, and when these proxies are in locations which are not representative of the spatial mean. On the other hand, the flexibility of the ridge regression approach might be preferable if the spatial covariance matrix is highly uncertain or if the assumption of normality is inappropriate. Both of these approaches offer considerable computational speedups for estimating a spatial mean when a gridded reconstruction is not of interest, as these methods only require gridcells informed by the proxies to be sampled at each MCMC iteration.

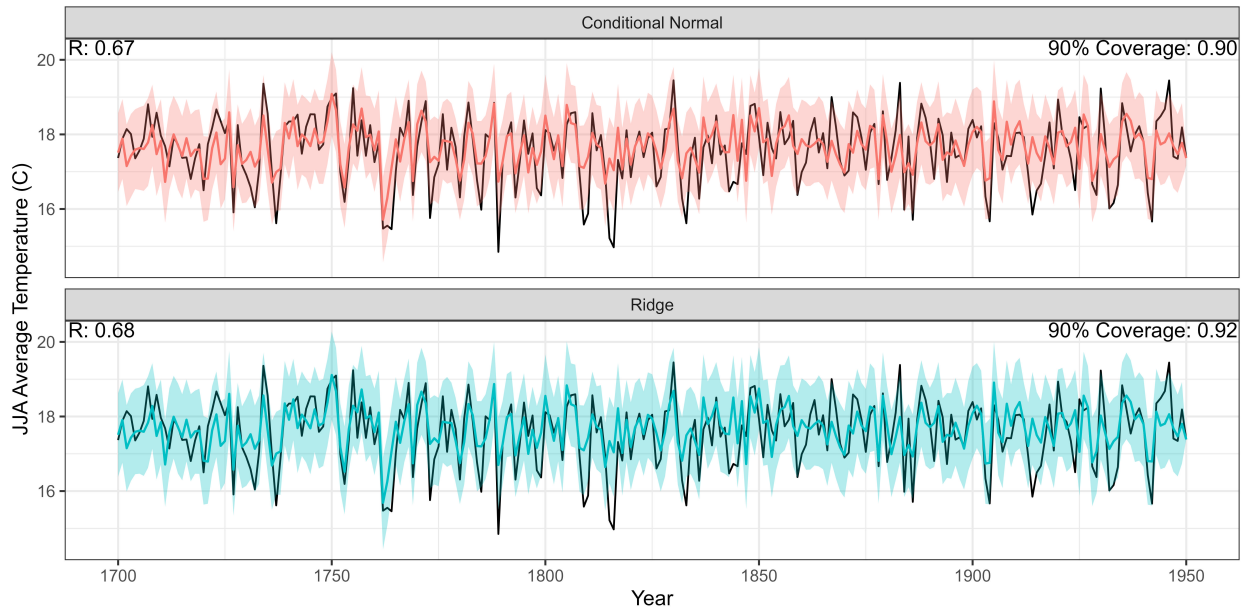


Figure 3.8: Comparison of PPE spatial mean reconstruction, using (a) the joint normality assumption to fill in gridcells with no available data before taking the mean vs (b) ridge regression to directly estimate the spatial mean from the gridcells with available data

3.4 Discussion

We connect offline paleoclimate data assimilation to its underlying probabilistic model, using MCMC samplers to better understand the implications that PaleoDA assumptions have on reconstruction uncertainty. A series of Bayesian hierarchical models are developed to bridge the gap between PaleoDA and BARCAST, two existing methods for climate field reconstruction.

In pseudoproxy experiments, we find that calibration uncertainty can lead to substantial changes in uncertainty quantification of the resulting reconstruction. Existing methods for climate field reconstruction commonly assume that the proxy–climate relationship is the same at all proxy sites, or that the proxy–climate can be modeled as a pre-processing step, with parameter point estimates used to obtain a reconstruction. Both of these approaches are useful for speeding up computation, but when the forward model is noisy or vary across sites, small changes in the forward model can lead to meaningful changes in the reconstruction. We demonstrate that this uncertainty can be better propagated to the reconstruction by simultaneously learning the forward model alongside the reconstruction, and that forward models can be made more flexible by using hierarchical parametrizations which allow for each proxy site to have its own regression parameters, which are assumed to be draws from a common distribution across sites.

In this setting, we demonstrate that the inclusion of a zero-mean model discrepancy term to a linear forward model can be helpful for proper uncertainty quantification when linearity is a poor fit for the proxy-climate relationship. However, our experiments suggest that the model discrepancy parameterization presented here is a useful first step but insufficient for fully capturing calibration uncertainty. In particular, we assume that model discrepancy is shared over all gridcells with proxy data, learning a single time series assumed to be shared across space. This assumption is needed to guarantee identifiability when only a single proxy time series is available at each gridcell, which is frequently the case in existing proxy networks. Ideally, a separate model discrepancy term should be included for each proxy series. One

possible solution in the BHM framework is to model individual proxy measurements rather than starting with a site-level chronology as input, so that the removal of non-climatic influences can also be included as a layer in the model (see [44] for an example of this approach at a single location). A location-specific model discrepancy term included at this stage of the model could be learned across individual proxy records at each site, which would allow for more accurate uncertainty quantification in the reconstruction. These experiments into model discrepancy add to the literature suggesting that the linearity assumption can be quite impactful in paleoclimate reconstruction [87, 65, 23], and highlights the importance of developing principled forward models for reconstruction.

We then consider the PaleoDA modeling decision to fix spatial covariance using sample covariance from climate model output. Pseudoproxy experiments with models learning spatial covariance show that this assumption, while strong, can be quite practical, as sparse and noisy proxies provide only limited information about spatial covariance. When learning a covariance matrix in a BHM, we find that prior information about the covariance structure must be included for inference when the spatial domain is large. Restricting the spatial covariance to the class of squared exponential covariance functions tends to be a reasonable assumption for Common Era temperature reconstructions in the North American Southwest, in both pseudoproxy experiments and in a case study using TRW chronologies from PAGES2k. Other covariance parametrizations should be studied for different climate fields. For instance, one component of common PaleoDA applications neglected here is the use of a localization matrix applied to the spatial covariance, which modifies a sample covariance by imposing that correlation decreases as a function of spatial distance. Localization has been shown to be impactful for DA applications in both numerical weather applications [99] and in PaleoDA [100], but the choice of localization radius is often ad-hoc in paleoclimate applications. Existing BHMs can be extended to include covariance localization with learned length-scale, and can also include a class of so-called “hybrid” estimators which assume that the spatial covariance is a convex combination of a climate model sample covariance and instrumental

period observed covariance [101].

When a large scale mean is the inferential target, BHM’s can be simplified by assuming a fixed spatial covariance and shrinking the parameter space to model only locations informed by proxy data. These simplifications should be further tested on other spatially aggregated quantities (e.g. the impact of nonlinear proxies on ENSO reconstruction [65]).

The models fit here all assume that the spatial covariance structure is stationary in time (i.e beyond an autoregressive term). While this assumption might be reasonable for reconstructions of the Common Era, stationarity is certainly violated on longer timescales, and the extent to which proxies can inform spatial covariance on these timescales should be examined. The experiments performed here demonstrate the difficulties of learning even a stationary covariance matrix, so exploration into highly parameterized spatiotemporal structures (e.g. spectral approaches) should be considered first. One benefit of using Bayesian models to assess modeling assumptions is that it highlights that these methodological choices are not (and should not) be independent of one another. For example, different climate-proxy models will change the impact of the prior covariance, so that the effect on the final reconstruction is different from what would be observed if each choice was assessed one at a time, holding the other parameters fixed. As such, PPEs using variations on a Bayesian model like the ones shown here offer a useful complement to the excellent existing experiments that operate solely using PaleoDA to assess the sensitivity of PaleoDA to the forward model [87], prior spatial covariance choice [90], and the quality of the input data [102].

The ideas discussed here also broadly apply to large reconstruction projects that span multiple proxy types, multiple climate fields, long timescales, and large spatial domains (e.g. [67, 23, 70]). Ideally, climate field reconstructions should simultaneously estimate well-tuned, biologically motivated forward models for each proxy alongside a statistical inverse model which is capable of producing a probabilistic reconstruction of multiple climate fields. Each of these extensions, however, adds to the computational challenge associated with fitting large high dimensional Bayesian models, and will rely on expert guidance to ensure that the

model is identifiable. BHMs which are able to jointly estimate all of these components are still likely too computationally expensive for practical use. This work considers additional modeling assumptions that can speed up these samplers, but are still noticeably slower than approximate methods like PaleoDA. Advances in more efficient Bayesian inference should be tested for their application to paleoclimate reconstruction. Covariance estimation is the largest computational bottleneck, so particular attention should be paid to the large and ever-growing body of literature on approximating Gaussian spatial random fields (see [103, 104] for reviews). In the meantime, we advocate the use and development of BHMs to assess the assumptions underlying modeling choices in PaleoDA, both in pseudoproxy experiments (e.g. using [105]) and in smaller scale observational experiments.

Chapter 4

Predictive Modeling of Alzheimer’s and Parkinson’s Disease Using Metabolomic and Lipidomic Profiles from Cerebrospinal Fluid

In recent years, metabolomics has been used as a powerful tool to better understand the physiology of neurodegenerative diseases and identify potential biomarkers for progression. We used targeted and untargeted aqueous, and lipidomic profiles of the metabolome from human cerebrospinal fluid to build multivariate predictive models distinguishing patients with Alzheimer’s disease (AD), Parkinson’s disease (PD), and healthy age-matched controls. We emphasize several statistical challenges associated with metabolomic studies where the number of measured metabolites far exceeds sample size. We found strong separation in the metabolome between PD and controls, as well as between PD and AD, with weaker separation between AD and controls. Consistent with existing literature, we found alanine, kynurenine,

The contents of this chapter appeared in paper [7]

tryptophan, and serine to be associated with PD classification against controls, while alanine, creatine, and long chain ceramides were associated with AD classification against controls. We conducted a univariate pathway analysis of untargeted and targeted metabolite profiles and find that vitamin E and urea cycle metabolism pathways are associated with PD, while the aspartate/asparagine and c21-steroid hormone biosynthesis pathways are associated with AD. We also found that the amount of metabolite missingness varied by phenotype, highlighting the importance of examining missing data in future metabolomic studies.

4.1 Introduction

Alzheimer’s (AD) and Parkinson’s (PD) disease are the two most common neurodegenerative disorders, and the fifth and tenth leading causes of death in the United States among individuals aged 65 or older [106]. Projections using the 2010 US census estimate that there will be approximately 1.2 million cases of PD (among individuals aged 45 or older) and 8.4 million cases of AD by 2030 [107, 108]. Currently, there are no therapies that can slow or halt progression of these debilitating diseases. Biomarkers for these diseases could improve our ability to diagnose these conditions at early stages, and track progression of neurodegeneration and the effect of therapies on slowing or halting disease progression.

Recent research has focused on the discovery of biomarkers to identify the presence of these diseases [109, 110]. Cerebrospinal fluid (CSF) surrounds the brain and is peripherally accessible, making it a promising medium to use for identifying biomarkers of neurodegeneration in AD and PD [111, 112]. The metabolome, consisting of the small molecules (< 2000 Da) circulating in an organism, has been shown to vary across many phenotypic traits, including (but not limited to) human longevity [113], cancer [114, 115], and type 2 diabetes [116]. Alterations in the metabolome have also been identified in AD versus control subjects [117, 118], and PD versus control subjects [119, 120, 121].

In the present study, we examined whether the CSF metabolome can discriminate between

healthy controls, AD, and PD subjects by fitting predictive models for each disease. We compared untargeted aqueous, targeted aqueous, and lipidomic approaches to profiling the metabolome in their ability to discriminate between AD, PD, and control CSF. Using a logistic elastic net regression to classify phenotypes and select relevant features [122], we were able to best discriminate PD versus control in the untargeted metabolite profiles (0.998 AUC). Our models were also able to discriminate AD versus control in untargeted and lipid profiles well (0.81 and 0.75 AUC, respectively). Our model’s ability to differentiate between AD and PD in the untargeted metabolomic profile was also excellent (0.976 AUC). We performed univariate analyses to identify potential metabolomic pathways associated with each disease, revealing metabolites and lipids that have been previously identified in association with AD and PD versus controls, as well as some new putative biomarkers. We address statistical methods for handling potential data quality issues related to metabolomics, including corrections for assay drift and an examination of statistically informative missingness. Our work should be validated in follow-up studies to determine if our findings are robust in larger cohorts, and in longitudinal studies to identify if any metabolites are biomarkers for disease progression. To our knowledge, this is the first paper to compare multivariate classification performance of AD and PD between metabolomic profiles.

4.2 Materials and Methods

Our analysis was based on cerebrospinal fluid (CSF) samples of 85 healthy subjects, 57 AD subjects, and 56 PD subjects, previously utilized and described in [123]. Sex at birth and APOE variant is known for all subjects, while Glucosylceramidase Beta variant (GBA) and Levodopa Equivalent Daily Dosage (LEDD) (calculated following [124]) are also known for the PD cohort. All samples were collected under University of Washington or VA Puget Sound Health Care System IRB-approved protocols. Written informed consent was given by all participants, or their legally authorized representative, before any study procedures took

place. The VA Northwest Mental Illness Research, Education, and Clinical Center (MIRECC) Sample and Data Repository provided the samples for AD and control subjects, while the Pacific Udall Center provided the PD samples used for this analysis.

Samples were collected while subjects were in the fasting state between 0900-1100 hours, after the subjects were in supine bedrest for at least 40 minutes. CSF was extracted via negative pressure using a Sprotte 24g atraumatic spinal needle, while subjects were placed in the lateral decubitus position. At the bedside, CSF was collected into sterile polypropylene syringes, placed into 0.5 ml aliquots in polypropylene cryotubes, and frozen immediately on dry ice. All samples were stored at -80°C prior to assay. Appendix C.3 displays the distribution of the date of LP draws. All participants with AD met the NINDS-ADRDA criteria for probable AD. A summary of the neuropsychological testing results is available in Table 1.

All PD patients were enrolled in the Pacific Udall Center Clinical Core [125]. All PD subjects underwent physical, neurological, and neuropsychological assessments, which are summarized in Table 1, and described in detail in [126]. Cognitive and motor diagnoses were made at consensus conferences, which included at least two movement disorders specialists, a neuropsychologist, and study support personnel. For more detailed information about sample collection for the PD cohort, see [127].

4.2.1 Metabolomics

Details of metabolite processing can be found in previous analyses [123, 129]. In brief, we used a LC-MS/MS platform to carry out both global and targeted metabolomics analysis. An Agilent 1200 LC system and a 6520 Q-TOF mass spectrometer were used to perform global metabolomic analysis. We prepared 200 μL CSF samples by dissolving in 1000 μL methanol before vortexing and incubating at 20°C . After centrifuging at 18000g for 20 min, we collected 750 μL of the supernatant, which was dried and reconstituted in a 100 μL solution of 40% H_2O : 60% acetonitrile. We then analyzed the samples in positive and

Table 4.1: Summary of subject data split by phenotype (top), with cognitive test results (middle), and with additional PD subject information (bottom). The cognitive status for PD was classified at three levels: no cognitive impairment, mild cognitive impairment (MCI), and dementia. Age of onset refers to the age of onset of motor symptoms. GBA refers to the carrier frequency for pathogenic GBA mutations and the E326K polymorphism. A chi-squared test of independence for sex and phenotype reports a p -value of 0.051. A one-way ANOVA of age at time of LP and phenotype reports a p -value of 2.099×10^{-9} [128]

* For PD subjects, Trail Making Test Part A was truncated at 150 seconds, and Part B was truncated at 300 seconds

	Control	AD	PD
n	85	57	56
Age at time of LP	53.7 ± 20.3	70.6 ± 9.8	65 ± 10
Duration of disease	N/A	4.3 ± 2.8	9 ± 5
ApoE genotype	2,3 (9%) 2,4 (1%) 3,3 (53%) 3,4 (33%) 4,4 (4%)	2,3 (3.5%) 2,4 (3.5%) 3,3 (44%) 3,4 (33%) 4,4 (16%)	2,2 (1.8%) 2,3 (14.3%) 3,3 (50%) 3,4 (26.8%) 4,4 (7.1%)
Race (% white)	91.7%	94.7%	94.6%
Sex	53% M (41 F)	49% M (29 F)	70% M (17 F)

	Control	AD	PD
MMSE total score (0-30)	29.4 ± 0.9	21.4 ± 5.6	N/A
Logical memory immediate recall (0-25)	13.5 ± 4.0	5.2 ± 13.0	12.6 ± 3.0
Category fluency (animals) (0-999)	23.3 ± 6.3	9.7 ± 5.1	20.6 ± 5.3
Trail Making Test Part A (sec)*	25.4 ± 2.0	107.3 ± 182.1	30.9 ± 12.2
Trail Making Test Part B (sec)*	68.1 ± 34.1	406.4 ± 357.3	76.9 ± 43.5
Logical memory delayed recall (0-25)	12.5 ± 4.1	3.4 ± 13.1	11.7 ± 3.6

	All	No cognitive impairment	MCI	Dementia
n	56	16	36	4
Sex	70% M (17 F)	56% M (7 F)	70% M (11 F)	100% M
Race (% white)	94.6%	100%	91.7%	100%
Age of onset of motor symptoms	56 ± 10	52 ± 11.1	57 ± 9	58 ± 16.6
Age at time of LP	65 ± 10	60 ± 9.1	67 ± 8.6	65 ± 21.4
Duration of disease	9 ± 5	8 ± 5.2	10 ± 4.9	7 ± 5.3
Levodopa equivalent dose	714 ± 536	441 ± 324.9	802 ± 592.3	1029 ± 221.7
MDS-UPDRS III	25 ± 12.8	21 ± 11.4	24 ± 12.4	38 ± 17.6
Hoehn & Yahr stage	2 ± 0.6	2 ± 0.4	2 ± 0.5	3 ± 1.3
MoCA	25 ± 3.8	27 ± 2.4	25 ± 2.1	16 ± 5.4

negative ESI modes using $5\mu\text{L}$ and $10\mu\text{L}$ injection volumes, respectively. We performed the chromatographic separation using a Waters XBridge BEH Amide column ($15\text{cm} \times 2.1\text{ mm}$, $2.5\mu\text{m}$) heated to 40°C , using a flow rate of $0.3\text{mL}/\text{min}$, $10\text{mM NH}_4\text{HCO}_3$ in $100\% \text{H}_2\text{O}$ (mobile phase A) and 100% acetonitrile (mobile phase B) with a gradient of 95% - 10% B from 0 to 5 min, 10% B from 5 - 40 min, 10% - 100% B from 40 - 45 min, and 100% from 40-70

min. We calibrated the Q-TOF/MS before each batch run, and achieved mass accuracy of <1 ppm using a G1969-85000 Agilent Technologies tuning mixture. We then used Agilent Mass Hunter and Mass Profiler Professional to process the MS data. We used an m/z scan range of 100-2000 and acquisition rate of 1.0 spectra/s.

For targeted analysis, a Sciex 6500+ LC-MS/MS system was used to target 203 known metabolites across more than 25 metabolomic pathways, along with 33 stable-isotope internal standards from Cambridge Isotope Laboratory, Tewksbury, MA. We injected 2 μ L of each sample for positive ESI ionization mode analysis, and 10 μ L for negative mode. In both cases, we conducted chromatographic separations using hydrophilic interaction chromatography, employing a Waters XBridge BEH Amide column as before, with flow rate of 0.3 mL per minute, a column compartment temperature of 40° C, and autosampler temperature of 4° C. We used 5mM ammonium acetate in H₂O with 0.5% acetic acid and 0.5% acetonitrile (Solvent A) and acetonitrile with 0.5% acetic acid and 0.5% water (Solvent B) to compose the mobile phase. For both negative and positive ionization modes, we used the following gradient: isocratic elution of 10% A for 0-1.5 min, a linear increase of A to 65% from 1.5-9 min, constant 65% A from 9-14 min, and 10% at 15 min. We used authentic commercially obtained compounds (Fisher Scientific, Pittsburg, PA or Sigma-Aldrich, Saint Louis, MO) to develop the assays, and MultiQuant 3.0.2 was used to obtain metabolite concentrations, as described previously [130].

Details of lipidomic analysis have been described previously [131, 132, 123]. For lipidomic analysis, we added 54 isotope labeled internal standards (spanning 13 lipid classes) purchased from Sciex (Framingham, MA). We thawed the frozen CSF samples at 25° C for 30 minutes, then vortexed and transferred 100 μ L of CSF to a borosilicate glass culture tube. We extracted lipids using dichloromethane/methanol, then concentrated the extracts under nitrogen and reconstituted the samples in 100 μ L of 10 mM ammonium acetate in dichloromethane:methanol (50:50). We analyzed the resulting lipids using the Sciex Lipidizer platform: specifically, we used a Shimadzu LC and AB Sciex QTRAP 5500 MS/MS system, where differential mobility

spectrometry (DMS) is conducted using SelexION. We then used multiple reaction monitoring to quantify lipids in both positive and negative ionization modes, with and without DMS, before processing the data using Sciex Analyst 1.6.3 and Lipidomics Workflow Manager 1.0.5.0. The Lipidizer platform outputs an abbreviated lipid annotation that does not exactly match the LIPID MAPS shorthand annotation [133]. Table S1A of [132] provides a table of translation between the two shorthand notations.

As described in [130], we monitored instrument performance over long periods and assisted with normalization by running both a pooled study Quality Control (QC) sample and a commercially pooled lab QC sample once every 10 study samples, and at the beginning and end of all sample batches. The pooled study QC sample is made by combining 10-20 μ L of each sample, which are prepared as usual alongside the study samples.

4.2.2 Preprocessing

Samples were prepared for assay in seven batches of 28 or 29 samples each. As described in [123], batch effects are reduced using the finite selection model from [134] to balance age, sex, APOE status, and disease status across the batches. An XGboost model (a nonparametric regression method implemented in the ‘xgboost’ R package), was fit separately on each metabolite using run index, age, and sex as covariates. Cross validation was used to select tuning parameters for these models [135]. The residuals of this procedure were used for the remainder of the analysis, which reduces the likelihood that the predictive performance of our models solely reflects metabolite drift over time, or the known associations of age and sex with both AD and PD [136, 137, 138]. The inclusion of run index in this process supplements the QC sample normalization done in the metabolite processing, and can be helpful especially when considering metabolite drift between QC runs. The benefit of using a flexible tree-based method like XGboost is that it is capable of handling and correcting non-linear and discontinuous metabolite drift. Examples of this drift correction procedure in our dataset are shown in Supplementary figure 2 of Hwangbo et al. After this correction, each

metabolite was centered and scaled to have unit variance. Within each feature, abundances further than 3 Median Absolute Deviations (MAD) were removed and treated as missing values to limit the influence of unreasonably small or large abundances, which are most likely the result of processing noise. To limit the impact of missing data imputation on the analysis, we excluded features with more than 50% missingness from this analysis. The remaining missing abundances were imputed using multiple imputation, whereby multiple (in our case, 5) versions of imputed data are created. Identical analysis is done on each of the five imputed datasets, which provides a measure of the variability induced by the missing data procedure. The Amelia package in R [139] was used to perform this procedure, with details of the implementation to accommodate large metabolomic datasets available in [123]. Missing data imputation was done within each cross validated loop (described in Section 2.3) to avoid the imputation depending on information from both the training and testing sets.

4.2.3 Regression modeling

To model the extent to which our data discriminate AD and PD against controls, we fit regularized logistic regression models, which extend the classic logistic regression model to allow for the number of features in our dataset to exceed the number of rows. In particular, we used elastic net regression, which fits

$$\min_{\beta_0, \beta_1 \in \mathbb{R}^{1+p}} - \left[\frac{1}{n} \sum_{i=1}^n y_i (\beta_0 + x_i^\top \beta) - \log \left(1 + e^{\beta_0 + x_i^\top \beta} \right) + \lambda \left(\frac{(1 - \alpha) \|\beta\|_2^2}{2} + \alpha \|\beta\|_1 \right) \right] \quad (4.1)$$

Here, α is a value between 0 and 1 which determines the type of regularization, while λ is a non-negative value that determines the total amount of regularization. We used the R package glmnet to fit these models, choosing $\alpha = 0.5$ and selecting the value of λ that resulted in the smallest deviance in leave one out cross validation [122]. This choice of α allowed for variable selection without imposing a strict upper limit on the number of selected variables. To mitigate the effects of class imbalance, observations were weighted by inverse proportion of

their class label. For interpretability, in Tables 2, 3, and 4 we show exponentiated coefficients for models fit on the full data (for the targeted and lipidomic profiles) to classify AD against controls, PD against controls, and AD against PD,, respectively. Because the covariates are standardized, these exponentiated coefficients correspond to the expected odds ratio (OR) resulting from a standard deviation increase in metabolite/lipid concentration.

To estimate out of sample predictive power, leave one out prediction was performed, i.e. missing data imputation and regression modeling was iteratively done on the training data after removing a single observation, and a prediction was formed on the left out observation. Rather than imputing missing values on a left out observation, missing features were removed from that iteration’s test set, using a process known as reduced feature modeling [140]. These left out predictions were used to create Receiver Operating Characteristic (ROC) curves and compute Area Under the ROC Curve (AUC). To summarize the performance of our models, we compute the average AUC value across all five missing data imputations for each leave-one-out iteration.

4.2.4 Pathway analysis

To perform pathway analysis and metabolite set enrichment analysis, we construct univariate classification models by separately regressing each metabolite against disease status, which allows for correlated and collinear metabolites to be included in their respective pathways. For univariate untargeted metabolomic analysis, we used logistic regression models to perform the classification, and then inputted mass:charge ratios, retention times, t -values and Benjamini Hochberg corrected p -values into Mummichog via a Python script [141] to perform pathway analysis. Positive and negative ionization modes were analyzed separately.

For univariate targeted analysis, we used the same logistic regression model technique, and inputted the names of all metabolites with false discovery rate < 0.05 into Metabolite Set Enrichment Analysis (MSEA), which performed hypergeometric tests of associations with known metabolite sets [142]. To perform MSEA on the targeted dataset, we used the open

source MetaboAnalystR package [143].

4.3 Results

CSF samples from AD, PD, and healthy control subjects obtained through two different longitudinal studies were analyzed in this study: 85 healthy subjects, 57 subjects with AD, and 56 subjects with PD. Age at time of CSF collection ranged from 20 to 88, with a median age of 65 (Table 4.1).

The average age of control subjects was 54 at the time of LP, ranging from 20 to 86 years old, with 48% female and 93% white. The average age of AD subjects at the time of LP was 71, ranging from 52 to 87 years old, with 51% female and 95% white. The average duration of disease of AD subjects was 4.3 ± 2.8 years. Consistent with the diagnosis of AD, the average MMSE score for AD subjects was 21.4 ± 5.6 , compared to 29.4 ± 0.9 for controls. Additional cognitive measures for AD and control subjects are included in Table 4.1.

The average age of the PD subjects was 65, with 9 years average duration of disease at the time of LP. The average levodopa equivalent daily dose (LEDD) was 714 mg/day, where only three PD subjects were not taking dopaminergic medication. The average MDS-UPDRS part III score for motor symptoms was 25, with average Hoehn and Yahr stage of 2, consistent with mid-stage PD patients on a moderate amount of dopaminergic medication. The average MOCA score for the PD cohort was 25.2 ± 3.8 . The majority of the cohort had the consensus diagnosis of mild cognitive impairment (64%), while 29% had no cognitive impairment and 7% had dementia at the time of LP. Six PD subjects were carriers of pathogenic GBA gene mutations and two subjects were carriers of the GBA E326K polymorphism. None of the most common pathogenic mutations of leucine rich repeat kinase 2 (LRRK2) G2019S or R1441C/G/H/S mutation carriers were identified in this cohort.

Metabolomic profiling of CSF was performed using untargeted and targeted metabolite methods, as well as lipid profiling. The untargeted, targeted, and lipid profiling yielded 6735,

108, and 1070 features respectively. These samples came with widely varying degrees of missingness, with 16%, 3%, and 81% of the data missing from each of the profiles, respectively. A summary of subject age and missing data is displayed in Appendix C.1, and a flowchart outlining the the analysis performed in this study is available in Appendix C.2

4.3.1 PCA

We first used the unsupervised method of principal component analysis (PCA) (see, e.g. [144]) to identify the combinations of metabolites responsible for the majority of the variation present in the untargeted metabolomic data. Twenty percent of the variation in the data was explained by the first two principal components (Figure 2). This variation did not associate strongly with disease (AD, PD, controls), as the projection of the subjects’ untargeted profile onto the first two principal components does not show significant separation between subject groups. In a previous analysis on the dataset of only control subjects, we found the first principal component to be associated with age [123].

4.3.2 Prediction Results

We next used supervised methods to seek explicit combinations of metabolites and lipids that can differentiate AD, PD, and healthy controls. For both its predictive power and interpretable results, we fit binary logistic elastic net regression models on each of these three pairwise phenotype combinations (AD, PD, control), and on three profiles (untargeted and targeted aqueous metabolites and untargeted lipids), using leave one out prediction to classify the disease status for each subject. The overall predictive performance of models fit on the untargeted profiles is displayed via ROC curves in Figure 4.2. Alternative characterizations of model performance are presented in Appendix C.6, including precision-recall curves and examples of threshold-dependent performance metrics. As stated in Section 4.2.2, age and sex are detrended from each metabolite prior to modeling fitting to better distinguish the signal from the metabolome from the known association of these covariates between AD, PD,

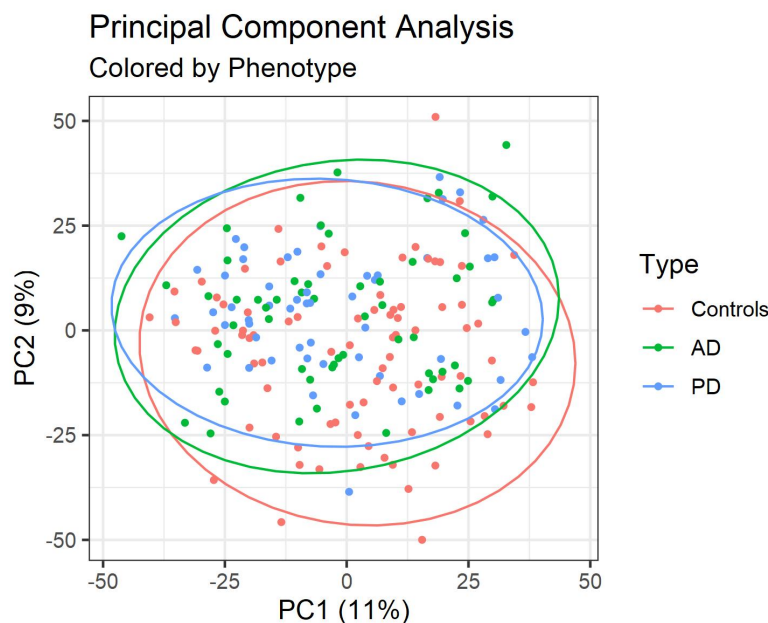


Figure 4.1: Untargeted data projected onto the first two Principal Components (PC). Each point represents a subject, colored by their phenotype. Percentages in the axis titles refer to the percentage of variation of the data explained by the respective PC. 95% confidence ellipses assuming the t -distribution are also plotted. The first two principal components do not clearly separate the disease phenotypes.

and controls. The predictive performance of models fit including age and sex can be found in Appendix C.7.

Classifying AD against controls

The predictive accuracy of models classifying AD against controls varied across the profiles, reporting Area Under the ROC Curves (AUC) between 0.5 and 0.7. The untargeted profile yielded the highest predictive accuracy. Recall that age and sex are detrended from each metabolite, so that these models mostly disregard the discriminatory power of age and sex. These models contain less discriminatory power than age and sex alone, as the leave one out performance of classical logistic regression models using only age and sex yields an AUC of 0.734. In addition to the leave one out modeling done to generate these ROC curves, a logistic elastic net model was fit on the full lipid and targeted profiles to get a sense of which metabolites separate AD from controls in our data. Exponentiated coefficients larger than 1.1

or smaller than 0.9 for these models are displayed in Table 4.2, and represent the odds ratio (OR) of classifying the subject as having AD resulting from a standard deviation increase in the metabolite/lipid.

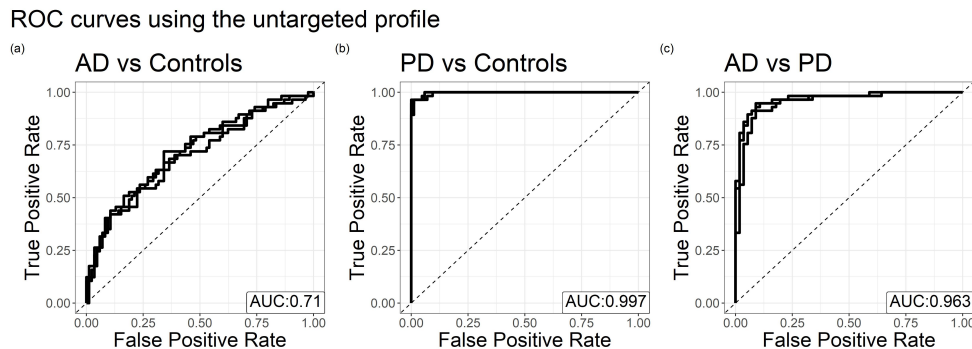


Figure 4.2: Receiver Operating Characteristic (ROC) Curves for binomial elastic net regressions classifying (a) controls against subjects with AD, (b) controls against subjects with PD, and (c) subjects with AD against subjects with PD. Solid lines represent models formed using each of the five missing data imputed datasets. The dotted $y = x$ line represents the ROC curve under a model which makes predictions at random. The average Area Under the Curve (AUC) across the five ROC curves is displayed in the bottom right.

Classifying PD against controls

The models correctly classify PD patients with near perfect accuracy in the untargeted metabolomic profile, and classify better than chance in the targeted metabolomic and lipidomic profiles. For reference, the leave one out predictions of logistic regression models fit using only age and sex yield AUC of 0.671. Table 4.3 displays the odds ratios associated with an elastic net model fit on the full targeted metabolite and lipid profiles, akin to Table 4.2. To test whether this near perfect prediction in the untargeted profile can be attributed to a small subset of metabolites, we proceeded to run a second elastic net regression, following the same procedure described in Section 4.2.3, but this time removing all metabolites selected by the first elastic net regression model from consideration (fit using all controls and PD subjects). This second elastic net model was still predictive of PD, reporting an average AUC of 0.83 over the five imputations and the leave one out analysis, indicating that the discriminatory power of our model is not limited to a small set of metabolites.

Table 4.2: Names and Odds Ratios (OR) for targeted metabolites and lipids retained in all five imputations of elastic net models fit on the full data to classify AD patients against controls. The tables are sorted by magnitude and split into positive and negative coefficient tables. Because the features were standardized prior to fitting the models, these ORs represent the expected odds ratio resulting from a standard deviation increase in concentration. Only ORs with magnitude greater than 1.1 or less than 0.9 are shown here. A two standard deviation interval is shown for the OR to quantify variability across the five missing data imputations.

Targeted Metabolites - Positive Coefficients

Metabolite	OR $\pm$ 2SD (> 1)
1-Methyladenosine	1.52 (1.43, 1.62)
Glycine	1.38 (1.3, 1.46)
Alanine	1.38 (1.32, 1.43)
Sarcosine	1.21 (1.16, 1.26)
Acetylcarnitine	1.19 (1.16, 1.21)
4-Methoxyphenylacetic acid	1.17 (1.11, 1.25)
Sorbitol	1.15 (1.13, 1.17)
Lactate	1.14 (1.12, 1.16)
Hydrocortisone	1.14 (1.09, 1.19)
Homoserine	1.12 (1.07, 1.16)
Caffeine	1.11 (1.04, 1.17)

Lipids - Positive Coefficients

Lipid	OR $\pm$ 2SD (> 1)
SM(18:1)	1.51 (1.42, 1.6)
CE(16:1)	1.22 (1.2, 1.25)
CE(20:1)	1.19 (0.91, 1.54)
PC(18:0/20:3)	1.12 (0.99, 1.26)

Targeted Metabolites - Negative Coefficients

Metabolite	OR $\pm$ 2SD (< 1)
N-Acetylneuraminic acid	0.76 (0.71, 0.8)
Glycocyamine	0.8 (0.79, 0.82)
4-Aminobutyric acid	0.84 (0.81, 0.88)
Creatine	0.85 (0.8, 0.9)
Urocanic acid	0.86 (0.73, 1.01)
Homocysteine	0.88 (0.84, 0.91)
Uridine	0.89 (0.85, 0.92)

Lipids - Negative Coefficients

Lipid	OR $\pm$ 2SD (< 1)
PE(P-18:0/22:6),0.77	1.12 (0.99, 1.26)
PE(18:0/20:4),0.84	1.12 (0.99, 1.26)
PE(18:0/22:6),0.90	1.12 (0.99, 1.26)

Mutations in GBA are the strongest genetic risk factor for PD, increasing the risk of PD in heterozygous carriers by approximately 5-fold compared to non-carriers [145], and are present in 5% of all PD patients. GBA E326K is also associated with PD risk, albeit with a smaller effect size [146], and both GBA mutations and E326K are associated with more rapid progression of cognitive and motor symptoms in PD [147, 148, 149]. Our cohort of PD

patients included six GBA mutation carriers and two GBA E326K carriers. Due to the small size of both genotype groups, we pooled the six mutation carriers and two E326K carriers to improve power for our analysis. Models classifying the eight GBA variant carriers against non-carriers within the PD cohort do not perform better than chance. We also examined whether PD-related medications could be significantly influencing metabolic profiles, as 53 of the 56 PD subjects were taking dopaminergic medication at the time of LP. However, models predicting LEDD from metabolites had very little predictive power, not performing much better than a naive intercept-only model. Additionally, we performed a reanalysis of the untargeted profile after removing all metabolites within 1 m/z of levodopa, entacapone, and their drug metabolites (dopamine, dihydroxyphenylacetic acid, homovanillic acid, and 3-O-methyldopa, S-adenosyl-L-methionine, norepinephrine, epinephrine). In this reanalysis, we found no reduction in the discriminative performance of our models.

Classifying AD from PD

In addition to classifying disease subjects versus controls, models were also fit to classify AD versus PD using the metabolome. Similar to PD versus control classifications, the untargeted profile was able to distinguish between the two classes with near perfect accuracy, while the lipid models do not perform significantly better than chance. For reference, the leave one out predictions of logistic regression models using only age and sex yields an AUC of 0.716. The most significant metabolites distinguishing between AD and PD, as determined by analyzing untargeted metabolite profiles, are summarized in Table 4.4. The corresponding tables for lipids are reported in Appendix Table C.3

4.3.3 Missing Data

We noted that there was a significant percentage of missing features for targeted and untargeted aqueous metabolite and lipid profiles per group. To determine whether there could be any correlation between the pattern of missing features and AD, PD or controls, we

fit a series of models following the procedure described in Section 4.2.3 with the modifications that all metabolite abundances were replaced with 1’s if they were present, and 0 if they were missing. Age and sex were still included in this model to control for their possible associations with missingness [123]. These matrices of missingness indicators (along with sex at birth and age) were still moderately to extremely successful at classifying AD and PD against controls, reporting leave one out AUCs as high as 0.97 (classifying PD against controls in the untargeted profile). The high predictive accuracy at classifying PD against controls in the untargeted missingness indicator profile is moderately robust, as removing all significant metabolites and rerunning the analysis yields an AUC of 0.72. However, the near perfect classification result reported in Section 4.3.2 is not solely a result of differential missingness, as similar performance is achieved after removing all metabolites that were missing in more than 1% of samples. Targeted and lipidomic tables for the full models (following the same format as Tables 4.2 and 4.3), are shown in Table 4.5.

We also found that when the proportion of missing data between phenotype subjects is large, the controls tended to have the most missingness, and PD subjects have the least missingness (after controlling for age). A series of univariate binary logistic regressions were fit on the untargeted profile to classify each metabolite’s pattern of abundance missingness based on phenotype (AD, PD, or age-matched controls). All metabolites with at least one coefficient $FDR < 0.05$ are displayed in Appendix C.4.

4.3.4 Pathway and Set Enrichment Analysis

We used pathway analysis to map significant individual metabolites identified by our analysis to known metabolite networks. To annotate the metabolites obtained in an untargeted method, we used `mummichog`, which combines the tasks of pathway analysis and metabolite identification using mass:charge, retention time, MS ion mode, along with univariate classification results. Mummichog identified vitamin B3 (nicotinate and nicotinamide) metabolism and alanine and aspartate metabolite pathways with the negative ion mode metabolites

classifying PD against controls. No associations were found classifying AD against controls with negative ion mode metabolites at a false discovery rate < 0.05 . For the positive ion mode detected metabolites, urea cycle/amino group metabolism, aspartate and asparagine metabolism, C21-steroid hormone biosynthesis and metabolism, biopterin metabolism, glutathione metabolism, vitamin B12 (cyanocobalamin) metabolism, vitamin E metabolism, bile acid biosynthesis, drug metabolism -cytochrome P450, and arginine and proline metabolisms were found to be associated with PD at the false discovery rate < 0.05 . Classifying AD against controls, mummichog identified the C21-steroid hormone biosynthesis and metabolism, sialic acid metabolism, androgen and estrogen biosynthesis metabolism, starch and sucrose metabolism, and hexose phosphorylation pathways at a false discovery rate < 0.05 .

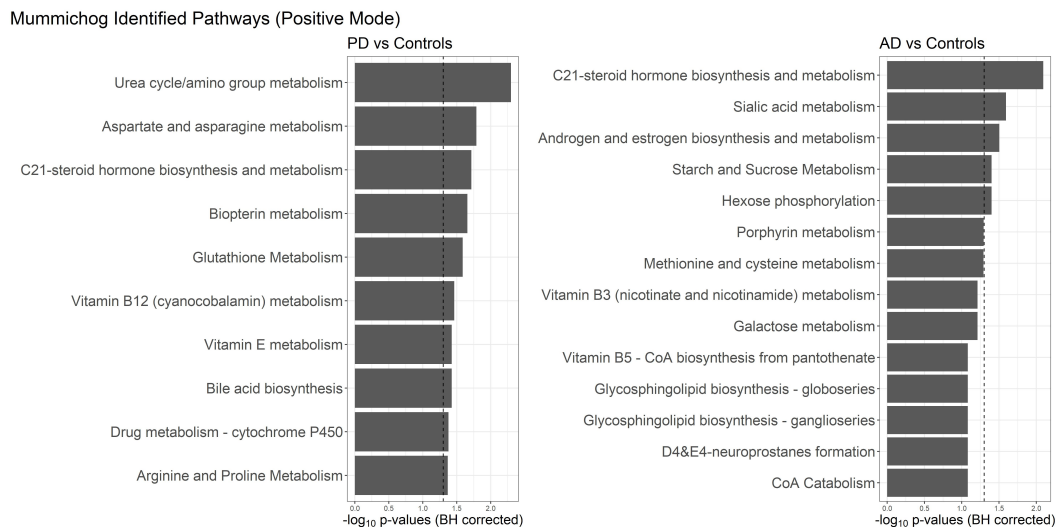


Figure 4.3: Pathway Analysis of Mummichog on the positive mode untargeted metabolites from univariate logistic models classifying PD and AD against Controls, sorted by $-\log_{10}$ Benjamini Hochberg corrected p -values, with the vertical dashed line marking $p = 0.05$

Using metabolite abundance missingness indicators to classify PD against controls in univariate analysis, we found vitamin E metabolism to be the only pathway associated with the significant metabolites at a false discovery rate < 0.05 in the positive ion mode detected metabolites. No associations were found using the negative mode metabolite missingness to classify PD against controls. No associations were found using positive or negative mode metabolite missingness to classify AD against controls. Using significant metabolites from

untargeted univariate analysis classifying AD against PD, mummichog identified the ascorbate (vitamin C) and aldarate metabolism pathways in the negative ion mode, and drug metabolism -cytochrome P450 and vitamin E metabolism pathways in the positive ion mode.

For the targeted profile, where metabolite identities have already been established, we used Metabolite Set Enrichment Analysis (MSEA), which performed overrepresentation analysis to compare significant metabolites from univariate classification against known metabolites sets compiled from literature. MSEA found no significant relationships using p -values computed from the hypergeometric distribution. Using metabolite sets from the Small Molecule Pathways Database (SMPDB) [150], we find that the targeted metabolites identified in our univariate classification of PD against controls match two out of three metabolites in the carnitine synthesis set, two out of four of the metabolites in the betaine metabolism set, and three out of eight of the metabolites in the methionine metabolism set. A table of the sets with largest overlap can be found in Appendix C.8.

4.4 Discussion

In this study we analyzed targeted and untargeted metabolomic and lipidomic CSF profiles to develop possible predictive models for AD and PD. Using elastic net regression models, we were able to classify PD against both controls and AD with a high degree of accuracy using untargeted metabolomic profiles, as well as moderate accuracy for all other models classifying phenotype using the targeted and untargeted metabolome. We found that some metabolites driving these predictive multivariate models, such as alanine and ornithine, have already been reported in the existing literature, providing further support for our results. We also performed univariate analysis coupled with pathway and set enrichment analysis and identified putative metabolomic pathways that have been previously associated with AD or PD. Associations between the presence of missing values and phenotype were also reported, suggesting the need for careful treatment of missing data in future metabolomic studies.

While several similar studies analyzing metabolomic and/or lipidomic profiles have been published, our study has several unique strengths. First, we analyzed both targeted and untargeted metabolomic profiles and were able to distinguish a specific metabolomic signature associated with neurodegeneration in both an unbiased interrogation of the metabolome as well as in metabolites of interest. In addition, most prior studies have analyzed metabolomic or lipidomic alterations in a single disease versus controls. We analyzed metabolomic and lipidomic profiles from both AD and PD in relation to each other as well as compared to controls to identify metabolomic and lipidomic changes that might be specific to each disease versus common to neurodegeneration. We were able to distinguish between the metabolome of AD and PD CSF, supporting growing evidence that neurodegenerative diseases have unique metabolomic "signatures". Finally, we analyzed CSF from living patients rather than from post mortem samples, which may be advantageous in revealing significant alterations in the metabolome. This could be relevant to developing biomarkers of disease progression, rather than alterations that are only detectable in end-stage disease.

The leave-one-out predictive accuracy of our models ranges greatly depending on the classification task and profile used. Models fit using the untargeted profile tended to be the most accurate, classifying PD against controls and PD against AD nearly perfectly, and classifying AD against controls better than chance. We also found that the models achieving near perfect accuracy are not limited to a few metabolites, and that the pattern of missing values alone is enough to distinguish these phenotypes. This suggests that features with missingness may encode important information about phenotype and should be treated with careful consideration in future metabolomics studies. There were a small number of targeted metabolites and lipids that appeared in all leave one out models, which are reported in Appendix C.11.

Interestingly, the lipidomic profiles of AD and PD subjects were nearly indistinguishable, despite our models being able to classify the lipidomic profiles of PD against controls better than chance. This suggests that there may be common alterations in lipid metabolism in both

neurodegenerative diseases. Increasing evidence indicates that alterations in lipid metabolism, particularly ceramide and sphingolipid metabolism, are present in multiple neurodegenerative diseases including AD and PD. This study is one of only a few studies able to compare AD and PD CSF lipidomic profiles obtained at the same time by the same laboratory, reducing alterations due to batch effect. Alterations in lipid metabolism could signal perturbations in multiple biological pathways, including inflammation, mitochondrial function, cell signaling, cell death, endolysosomal trafficking, and exosome biogenesis [151, 152, 153]. The two lipid species whose abundances have the strongest positive association with AD in the multivariate model are ceramide (24:1) and cholesterol ester (20:1), which are consistent with studies finding an increased abundance of long chain ceramides and cholesterol in AD [154, 155, 156, 157]. A decrease in concentration of four species of phosphatidylethanolamine (PE) were found to be associated with AD in our multivariate model (Table 1). These bioactive glycerophospholipids have been previously associated with AD and PD [158, 159], though there is conflicting evidence of decreased PE levels in AD [160], or unchanged between AD and age-matched controls [161].

Several of the targeted metabolites driving our predictive multivariate models have been previously reported in association with AD or PD. Alanine [162, 163, 164, 121], kynurenine [119], glycine [165], tryptophan [164, 166], xanthine [167], and serine [167, 164] have all been found to be associated with PD, and appear in either our multivariate analysis of the targeted profile, or the pathways and set enrichment analysis derived from univariate analysis. Ornithine, an amino acid involved in the urea cycle, is the metabolite most positively associated with PD. It has been reported to be increased in urine and serum from PD patients [168, 169, 121] but decreased in CSF of PD patients compared to controls [170]. In our sample, ornithine is strongly associated with PD in the untargeted profile as well, as `mummichog` identified the urea cycle/amino acid group metabolism as the most enriched pathway for classification between PD patients and controls. Alanine is one of the metabolites most positively associated with AD in our multivariate model (Table 1). Increased levels of this

amino acid, as well as several others including glycine, methionine, threonine, phenylalanine and citrulline have been found to be increased in CSF of AD patients [171, 172, 173]. Targeted aqueous metabolomic analysis of premortem blood and postmortem brain from individuals with AD and healthy controls found acylcarnitine to be associated with AD and cognitive changes [174]. We found decreased levels of creatine to be associated with AD in our multivariate model, and proteomic studies have found creatine kinase to be associated with AD patients [175, 176]. The positive association between fructose and AD in our multivariate model might provide support for the hypothesis in [177] suggesting that AD can be driven by an overactivation of fructose metabolism. Our pathway and set enrichment analysis identified several putative pathways previously associated with AD and/or PD, including biopterin metabolism, which is important for dopamine synthesis, and glutathione metabolism, which is critical for antioxidant activity and mitochondrial integrity. Urea cycle and amino group metabolism were also identified, which are important for fatty acid metabolism, and have been implicated in neuronal function and neurodegeneration. Finally, alanine and asparagine metabolism may be important for metabolic control of neurons [178]. The emergence of these pathways from our analysis suggests that our model correctly predicted PD compared to controls based on changes in metabolites and lipids that arise from significant alterations in cell function, and further supports this method in identifying possible biomarkers as well as biological mechanisms underlying disease pathogenesis.

4.4.1 Modeling limitations

There are several limitations to the current study, which we address here. We focus in particular on the near perfect ability to discriminate PD from both controls and AD, which might be met with some skepticism. First, it is possible that our results could be distorted by the effects of dopaminergic medication on the metabolome. While the LEDD was recorded for most PD subjects, it was not used in this study because non-zero LEDD itself is a nearly perfect predictor of PD status. Levodopa treatment has been shown to impact kynurenine,

tryptophan, and tyrosine metabolisms [109], which could explain several of the associations found in MSEA (Table 5) and multivariate analysis (Table 2). That said, there is evidence to suggest that the classification is not driven solely by the presence of drugs; we find that a Gaussian elastic net model is unable to predict LEDD better than chance in the PD cohort, and we also find that removing all untargeted metabolites within 1 m/z of levodopa, entacapone, and their drug metabolites does not change the near-perfect discrimination of PD against controls.

Another potential confounder driving the near perfect PD classification results could come from differences in sample processing. The PD samples were collected by a different study and researchers (Pacific Udall Center) than the AD and control samples (VA MIRECC; see Materials and Methods), and our models were able to classify PD against both AD and controls with high degrees of accuracy. It is possible that despite strict adherence to the same protocol, there were still differences in sample collection and processing between the two studies that our models were able to detect. One important limitation of our PD analysis is that CSF was extracted at an earlier date for the AD and control subjects than for the PD subjects (Appendix C.3), and there is some evidence to suggest that prolonged storage time can lead to significant alterations in metabolomic profiles [179]. Determining the relationship between metabolomic profiles and storage time is an active area of research, and we refer to Section 2.5 of [180] for a review of the current literature. However, the exploratory analysis presented in Appendix C.8 provides evidence to suggest that batch effects are not the sole source of predictive power in our models. In short, the metabolites distinguishing AD and PD are distinct from those distinguishing AD and controls, suggesting that the sample extraction date is not the sole confounder driving these predictions. In addition, studies such as [166] and [120] achieve similar performance in distinguishing PD from controls, which indicates that it is possible that the nearly perfect classification in our models is due to true signal in the metabolome. Further, we note that this analysis is meant to explore the predictive power of the metabolome to distinguish phenotype without relying on known associations with age

and sex. As such, the choice to detrend and exclude sex and age in the elastic net models was a conservative choice that shrinks the possibility that the predictive performance and metabolites retained were dependent on age and sex. An approach that explicitly includes age and sex would likely generate more accurate predictions.

Relatedly, the pattern of missing values in metabolomic and lipidomic profiles differed between PD and controls but not between AD and controls. A re-analysis of PD and control metabolite profiles suggests that differential missingness is not solely responsible for the predictive performance of our models; after removing nearly all metabolites that were missing in either or both groups, we still found significant differentiation between PD and controls.

More generally, a larger sample size and further investigation using experimental model systems are required to test whether the discriminatory features that we identified between AD, PD, and controls may have causal biological relationships. In particular, our exploratory analysis of the models distinguishing PD from controls shows that there is a multiplicity of good models, suggesting that the results shown in this study are neglecting potentially important predictors. While the elastic net classification models used in this study are powerful tools for interpretable high dimensional analysis, performing statistical inference (e.g. obtaining p -values for individual coefficients) in this setting is an active area of research and not yet well understood (see [181] for a recent and thorough treatment of this topic). As such, the results of this analysis should be seen as exploratory and as potential directions for future work, rather than making strong inferential claims on its own. Additionally, the null result of being unable to classify GBA variant carriers against non-carriers within the PD cohort is likely due to the small sample size. Specific lipidome alterations in GBA variant carriers with PD have been identified in CSF [182], but these observations were not replicated in our study.

A third limitation of the study is the cross-sectional nature of our data. The metabolome is known to be dynamic and volatile, so patterns identified in a cross-sectional sample might not generalize well. Longitudinal studies, similar to those done by [174] and [167], will be

needed to determine whether the discriminatory features that we identified between PD, AD, and control CSF are not only diagnostic, but change over time to reflect disease progression.

In conclusion, we identified a number of metabolites and developed an elastic net regression model that can distinguish between AD, PD, and healthy control CSF metabolite profiles, as well as PD versus control lipidomic profiles. Our study is one of the few studies to examine targeted and untargeted aqueous metabolite profiles as well as lipidomic profiles in the two most common neurodegenerative diseases, AD and PD, as well as healthy controls. While our study has some limitations, including small cohort size and lack of longitudinal data, our analyses suggest several directions for future investigation of lipidomic and metabolomic profiles as markers of disease state and progression.

Table 4.3: Names and OR for targeted metabolites and lipids retained in all five elastic net models fit on the full data to classify PD patients against controls. Only ORs with magnitude greater than 1.1 or less than 0.9 are shown here. A two standard deviation interval is shown for the OR to quantify variability across the five missing data imputations.

Targeted Metabolites - Positive Coefficients

Metabolite	OR $\pm$ 2SD (> 1)
Ornithine	2.1 (1.82, 2.41)
Glycylproline	1.75 (1.52, 2.01)
Levulinic acid	1.62 (1.43, 1.82)
Acetylglycine	1.57 (1.42, 1.73)
Glycine	1.57 (1.45, 1.7)
Creatinine	1.52 (1.46, 1.58)
Cytosine	1.48 (1.28, 1.7)
Adenosine	1.45 (1.26, 1.67)
Pentadecanoic acid	1.4 (1.32, 1.49)
Sorbitol	1.4 (1.3, 1.52)
N-Acetyethanolamine	1.39 (1.31, 1.48)
alpha-Hydroxyisovaleric acid	1.39 (1.23, 1.57)
2-aminoadipic acid	1.36 (1.16, 1.6)
Methylguanidine	1.32 (1.27, 1.38)
Xanthosine	1.25 (1.2, 1.3)
Dimethylarginine	1.22 (1.15, 1.3)
Homoserine	1.21 (1.14, 1.28)
Threonine	1.2 (1.15, 1.25)
Cystine	1.16 (1.09, 1.23)
3 α -Hydroxy-12 Ketolithocholic Acid	1.16 (1.09, 1.23)
Adenosyl-L-homocysteine	1.15 (1.08, 1.22)
6-Methyl-DL-tryptophan	1.13 (1.06, 1.2)
Anthranilic acid	1.12 (1.03, 1.21)
Fructose	1.11 (1.02, 1.2)

Lipids - Positive Coefficients

Lipid	OR $\pm$ 2SD (> 1)
PE(P-16:0/18:1)	1.54 (1.45, 1.63)
HCER(18:0)	1.49 (1.43, 1.55)
FFA(16:1)	1.46 (1.3, 1.65)
SM(18:1)	1.42 (1.26, 1.6)
FFA(24:0)	1.22 (1.11, 1.35)
PC(16:0/20:2)	1.21 (0.93, 1.57)
FFA(20:2)	1.2 (0.94, 1.52)
CE(20:1)	1.2 (0.98, 1.46)
DAG(20:0/20:0)	1.17 (0.96, 1.43)
PE(16:0/22:6)	1.16 (0.97, 1.39)
LPC(18:1)	1.11 (1.06, 1.15)

Targeted Metabolites - Negative Coefficients

Metabolite	OR $\pm$ 2SD (< 1)
Indole-3-acetic acid	0.57 (0.54, 0.61)
Serine	0.58 (0.52, 0.64)
N-Acetylneuraminic acid	0.61 (0.55, 0.68)
Urocanic acid	0.64 (0.53, 0.76)
Agmatine	0.65 (0.63, 0.68)
HIAA	0.66 (0.6, 0.73)
Glycocyamine	0.71 (0.58, 0.87)
Aspartic acid	0.76 (0.66, 0.88)
4-Methylvaleric acid	0.79 (0.73, 0.85)
Serotonin	0.82 (0.77, 0.87)
Mannose	0.82 (0.74, 0.9)
Creatine	0.83 (0.78, 0.88)
Xanthine	0.83 (0.76, 0.9)
4-Aminobutyric acid	0.86 (0.81, 0.91)
4-Methoxyphenylacetic acid	0.86 (0.81, 0.91)
Citraconic acid	0.87 (0.74, 1.02)
Decanoylcarnitine	0.89 (0.84, 0.94)

Lipids - Negative Coefficients

Lipid	OR $\pm$ 2SD (< 1)
PC(18:1/18:2)	0.49 (0.45, 0.53)
FFA(18:0)	0.64 (0.53, 0.76)
PE(18:1/18:1)	0.65 (0.48, 0.88)
FFA(24:1)	0.68 (0.61, 0.75)
PC(18:1/20:4)	0.72 (0.64, 0.81)
PC(18:0/22:6)	0.76 (0.7, 0.82)
PC(18:1/16:1)	0.88 (0.79, 0.97)

Table 4.4: Names and OR (associated with a standard deviation increase in concentration) for targeted metabolites retained in all five elastic net models fit on the full data to classify PD patients against AD patients. A metabolite with $OR > 1$ indicates that higher concentration is associated with AD in our models. Only ORs with magnitude greater than 1.1 or less than 0.9 are shown here. A two standard deviation interval is shown for the OR to quantify variability across the five missing data imputations.

Targeted Metabolites - Positive Coefficients

Metabolite	OR $\pm$ 2SD (> 1)
Serine	1.63 (1.36, 1.95)
Alanine	1.62 (1.46, 1.79)
Indole-3-acetic acid	1.52 (1.46, 1.58)
Xanthine	1.42 (1.26, 1.6)
Aspartic acid	1.4 (1.32, 1.49)
Caffeine	1.4 (1.3, 1.52)
Sarcosine	1.22 (1.15, 1.3)
HIAA	1.2 (1.11, 1.3)
n-glycyl-l-proline	1.16 (1.01, 1.34)
Glycodeoxycholic acid	1.15 (1.06, 1.25)
4-Methoxyphenylacetic acid	1.14 (1.12, 1.16)
Serotonin	1.13 (1.08, 1.17)

Targeted - Negative Coefficients

Lipid	OR $\pm$ 2SD (> 1)
Ornithine	0.52 (0.51, 0.53)
alpha-Hydroxyisovaleric acid	0.63 (0.58, 0.68)
Homocysteine	0.64 (0.59, 0.7)
Histidine	0.7 (0.62, 0.79)
Creatinine	0.72 (0.66, 0.78)
Glycylproline	0.73 (0.64, 0.82)
Levulinic acid	0.76 (0.7, 0.83)
Adenosine	0.77 (0.67, 0.89)
N-Acetyethanolamine	0.81 (0.75, 0.88)
Acetyl-L-glutamine	0.86 (0.79, 0.93)

Table 4.5: Names and OR for targeted metabolite abundance missingness indicators in AD classification against controls, lipid abundance missingness indicators in PD classification against controls, and lipid abundance missingness indicators in AD classification against PD, retained in elastic net models fit on the full data. The OR corresponds to increased odds of classifying a patient as having AD/PD/AD associated with a standard deviation increase in metabolite concentration (for left, middle, and right, respectively). Age and sex are not penalized, and are therefore guaranteed to be included in these models. Other combinations of phenotype and profile (i.e targeted-PD or lipids-AD tables) are not shown because the only retained covariates were age and sex

Metabolite	OR (AD v C)	Lipid	OR (PD v C)	Lipid	OR (AD v PD)
Citraconic acid	0.53	Male	1.85	Male	0.25
Phenylalanine	1.85	TAG46:0- FA16:0	1.77	PE(18:1/18:1)	0.68
Creatinine	1.59	DAG(18:1/22:6)	1.08	PC(16:0/14:0)	0.74
Glucosamine	1.46	Age	1.04	TAG52:4- FA16:1	1.17
Amiloride	1.42			CE(18:4)	1.10
N- Acetylneuraminic acid	0.71			Age	1.09
Mannose	0.73				
Male	0.85				
Age	1.07				
Creatine	1.02				

Chapter 5

Discussion

The methods developed over these chapters represent efforts towards more principled uncertainty quantification across the data analysis workflow in high dimensional regression tasks. The approaches taken are necessarily domain and task-specific, and the structure of the data defines the distinction between the methods in these chapters. For instance, the dimensionality in these tasks comes from different sources. In Chapter 2, past monthly precipitation is modeled as unknown parameters, so that the high dimensionality comes from the many months for which a reconstruction is desired, and structure is imposed through the use of priors on the monthly precipitation relationship, both through autocorrelation and through the use of weights to constraint which months are informed by the proxies. In Chapter 3, reconstructions of past temperature are considered at annual temporal resolution but over a spatial field, so that the high dimensionality comes not only from the number of years to be inferred, but also in the size of the spatial grid for which a reconstruction is desired. In this setting, strong assumptions on the spatial covariance of temperature were crucial for obtaining valid inference. In Chapter 4, the dimensionality comes from the large number of highly correlated metabolites which are used to classify neurodegenerative disease status. Here, the elastic net regularization structure was uniquely suited for the inferential task of biomarker identification with a small sample size relative to dimension. A ridge penalty would

not perform any variable selection and would therefore not be as useful a tool for biomarker identification, and a LASSO penalty would select no more than n metabolites, which induces too much sparsity in the small n , large p context of the untargeted and lipidomic profiles.

All chapters also feature approaches to imputing missing regressors. In the context of paleoclimate reconstruction, the climate of the past is considered a missing predictor of tree ring widths, whereas in the metabolomic analysis, missing metabolite concentrations are predictors of neurodegenerative disease status. The treatment of missingness in both of these applications are probabilistic in nature, relying on assumptions of multivariate Gaussianity. The key conceptual difference in the treatment of missing data in these applications is the role of the response variable in the imputation. The paleoclimate reconstruction task is formulated as an inverse model, which requires the the response variable, observed proxy values, to be used to reconstruct past climate. In contrast, the application with metabolomic data only relies on other metabolite concentrations in order to perform the imputation, and crucially cannot depend on neurodegenerative disease classifications in order to inform the imputation. Another distinction between the two applications is the importance of missing data imputation on the inferential task. Imputed climate values are the primary estimand of interest in the paleoclimate reconstruction task, so that many samples from the posterior distribution are drawn to characterize the distribution. In contrast, the missingness in the metabolite concentrations are viewed as a nuisance parameter in the metabolomic study, so that few samples from an approximate posterior distribution are sufficient to quantify uncertainty due to this missingness.

The methods shown here provide frameworks which should be expanded on for further use. For instance, the paleoclimate reconstruction methods shown here utilize a single proxy type to reconstruct a single climate variable. In an idealized world, paleoclimate reconstructions would be able to use a large proxy network to simultaneously learn proxy-climate relationships, invert these relationships to reconstruct multiple climate variables, and interpolate information across space to estimate full spatiotemporal fields of climate over long timescales. In addition,

the methods of the first two chapters would be combined by creating a proxy network which contains multiple proxy measurements at each site rather than containing a single site-level record. While these extensions all fit naturally into the Bayesian framework, incorporating these elements into a single model will prove to be computationally challenging, and carefully chosen assumptions will be required for each proxy type and each climate variable of interest to ensure that the model is well posed. Computational advances for Bayesian inference, including approximate and efficient high dimensional Bayesian samplers, will go a long way towards making fully probabilistic paleoclimate reconstructions feasible.

The predictive models in Chapter 4 suggest that the metabolome contains biological information that can be used to identify neurodegenerative disease. The inference performed across the three metabolomic profiles provide a starting point for identifying potential biomarkers, but represent just the beginning of the research process. Targeted studies should be conducted on the potential biomarkers identified here to understand the biological processes underlying the results shown in this predictive analysis. Longitudinal studies will need to be conducted to better control for all of the unobserved confounding effects that are implicitly present in these cross-sectional metabolomic panels. Studies similar to the one shown here should be repeated as new advances in metabolomics assays develop, as the improvements in metabolite identification will reduce sampling variability and measurement error. Libraries for targeted metabolomics will also expand over time and become more robust as scientific understanding improves, which will allow for better metabolite identification and shrink the observed gap in predictive skill between the targeted and untargeted assays. Statistical theory for inference in elastic net regression and in post-selection inference is an active area of research and should be monitored, especially for the non-asymptotic setting needed to be useful with such a small sample size as shown here.

A common theme of this work is that the methods and assumptions rely on properties of the data which are only known through subject matter expertise. Methodological progress in both paleoclimatology and high throughput biology will require close collaboration between

domain scientists and statisticians. Furthermore, the data used in these two fields of study share many statistical characteristics, suggesting that research in both fields would benefit from ongoing conversation between statisticians working in different applied fields.

Appendices

Appendix A

Chapter 2 Appendix

A.1 Details of the statistical model

The fully written out model is;

$$\log(\text{tree-ring width})_{i,t} = \beta_{0i} + \beta_{a,i}\text{age}_{i,t} + \beta_c\text{climate}_t + \eta_t + \epsilon_{i,t}$$

$$\text{climate}_t = \sum_{j=1}^{12} w_t \text{climate}_{t_j}$$

$$\text{climate}_{1_1} \sim N(\mu, \sigma_{ar}^2)$$

$$\text{climate}_{t_m} - \mu = \alpha_{ar}(\text{climate}_{t_{m-1}} - \mu) + \epsilon_{ar}$$

$$\eta_t = N(0, \sigma_{\eta_t}^2)$$

$$\sigma_{\eta_t} = \sqrt{\gamma_0 + \gamma_1 \eta_t^2 + \gamma_2 \sigma_{\eta_{t-1}}^2}$$

$$w_m = \frac{\exp(\theta_m)}{\sum_{j=1}^{12} \exp(\theta_j)}$$

$$\theta_m \sim N_{12}(w_{\text{input}}, \Sigma)$$

$$\Sigma_{j,k} = \begin{cases} \alpha^2 \exp \left\{ -\frac{1}{2\rho^2} \min\{k-j, j+12-k\} \right\} & j < k \\ \alpha^2 + \sigma_{w,j}^2 + 0.001 & j = k \end{cases}$$

$$\begin{aligned}
\alpha_{(0,1)} &\sim N(0, 1) \\
\rho &\sim \text{Inverse-Gamma}(5, 5) \\
\sigma_{w,j}^2 &\sim N(0, 1) \\
\epsilon_{ar} &\sim N(0, \sigma_{ar}^2) \\
\epsilon_{i,t} &\sim N(0, \sigma_i^2) \\
\beta_{a,i} &\sim N(0, 1) \\
\beta_{0,i} &\sim N(\mu_{b0}, \sigma_{b0}^2) \\
\mu_{b0} &\sim N(0, 5^2) \\
\beta_{c(0,\infty)} &\sim N(0, 5^2); \\
\sigma_{b0(0,\infty)}^2 &\sim N(0, 5^2) \\
\sigma_{ar(0,\infty)}^2 &\sim N(0, 5^2) \\
\sigma_{i(0,\infty)}^2 &\sim N(\mu_s, \sigma_s^2) \\
\mu_s &\sim N(0, 5^2) \\
\log(\sigma_s) &\sim N(0, 1) \\
\gamma_{0(0,\infty)} &\sim N(0, 1) \\
\gamma_{1(0,1)} &\sim N(0, 1) \\
\gamma_{2(0,1-\gamma_1)} &\sim N(0, 1)
\end{aligned}$$

for i indexing over trees, t_u indexing over months u in years t , and $j, k = 1, 2, \dots, 12$ indexing over months. The first seven equations of this model are described in the main text, with the logarithm of TRW being modeled as a function of age and a weighted average of monthly climate, with an autoregressive(1) model over the months with parameters (α, σ_{ar}^2) and a zero mean model discrepancy term η_t with a generalized autoregressive conditional heteroscedasticity(1,1) model accounting for changes in variance over time. The monthly climate variable chosen is $\log(\text{precipitation} + 0.001)$, where the offset 0.001 is used to avoid

the problem of log being undefined at 0. We found that the age coefficient was very close to zero in our case study, and so we encourage that shrinkage to 0 by using a standard normal prior for all $\beta_{a,i}$. This is in contrast to the varying intercepts $\beta_{0,i}$ and variance σ_{ar}^2 , whose pooled mean and variance is learned from the data. The subscript under each parameter indicates if its range was truncated (for example, $\alpha_{(0,1)}$ indicates that the autoregressive parameter α is restricted to lie between 0 and 1). Given the variable standardization, these prior choices (e.g. on the variances) are weakly informative. Models are run in `stan` with 4 chains of 5000 iterations for all experiments except for the sensitivity analyses In Fig S3, which are performed with 1000 iterations each. In all cases, the first half of the iterations are used as warmup.

We suspect that these monthly weights are correlated in time (e.g. the weight corresponding to July is correlated with those in June and August), so we place a hyper-prior on the logarithm of the weight parameters, namely, multivariate normal with covariance matrix (specified above by just the upper diagonal, by symmetry): This covariance specification encodes prior belief that weights in nearby months should be similar to one another, and is defined so that distances between months respects the annual cycle (e.g. the correlation between December and January is the same as the correlation between January and February). Prior information about the growth-dominant season is encoded in the hyperparameter w_{input} (a vector of length 12), which is supplied by the user. Many different choices of covariance are possible; the one presented here is an example of a squared exponential kernel. If information about the relationship between months is known, then the functional form of the covariance matrix can be modified to better encode this information into the analysis.

A.2 Eliciting informative prior weights

In our proposed model, the hyperparameter w_{input} can be specified to encode prior knowledge of the tree-ring width growth-dominant season. In the main text, an uninformative uniform prior across the months is chosen to evaluate the model’s skill at learning a target through data alone. However, when prior knowledge is available, an informative choice of w_{input} can improve both model skill and runtime. Selecting values of these parameters that accurately reflect expert knowledge can be done through prior simulation, whereby pseudoproxies are generated using the specified model parameters to be evaluated. Our model can naturally be used to generate pseudoproxies by simulating random variables from the prior distributions of the regression parameters, weights, and precipitation (done iteratively through the AR(1) process) and applying the regression equation. Repeating this procedure for different values of w_{input} allows us to evaluate the impact of w_{input} on the final range of admissible weights w that we encode in the prior. This allows for a flexible approach to finding hyperparameter values that accurately reflect prior knowledge. Here is one such example of a procedure for determining w_{input} : first, suppose I know roughly what proportion of weight I expect in each month (e.g, I expect a tree-ring width archive grows equally in DJF). This suggests that the w_{input} values for DJF should be the only nonzero entries. Suppose also that I know the maximum allowable weight I would want for non-DJF months (e.g., I really do not expect July weights to be larger than 0.5). This provides guidance for choosing the magnitude of the nonzero elements in w_{input} , as we can generate pseudoproxies with larger and larger nonzero values of w_{input} until the prior weight for non DJF months has, say, 90% of its mass under 0.5. An example of this workflow with pseudoproxy data is available in the github repository ([/notebooks/prior_sim.qmd](#)), and a review of prior elicitation procedures is available in [183].

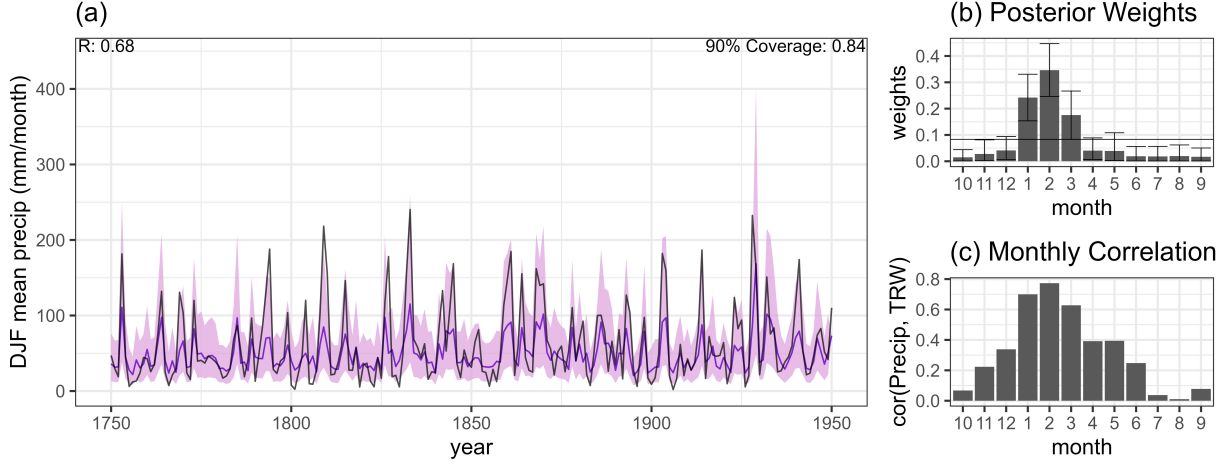


Figure A.1: Demonstration of the proposed model with VSLite pseudoproxies. (a) The reconstructed (red) and true (black, from the climate model) DJF precipitation. The reconstructed DJF precipitation is transformed from the original reconstruction target, which is in log-precipitation space. (b) The posterior mean (bars) of the learned weights for each month in the water year (October - September), with error bars representing 90% credible intervals. (c) Monthly correlations of the noiseless VSLite pseudoproxy and each month's precipitation

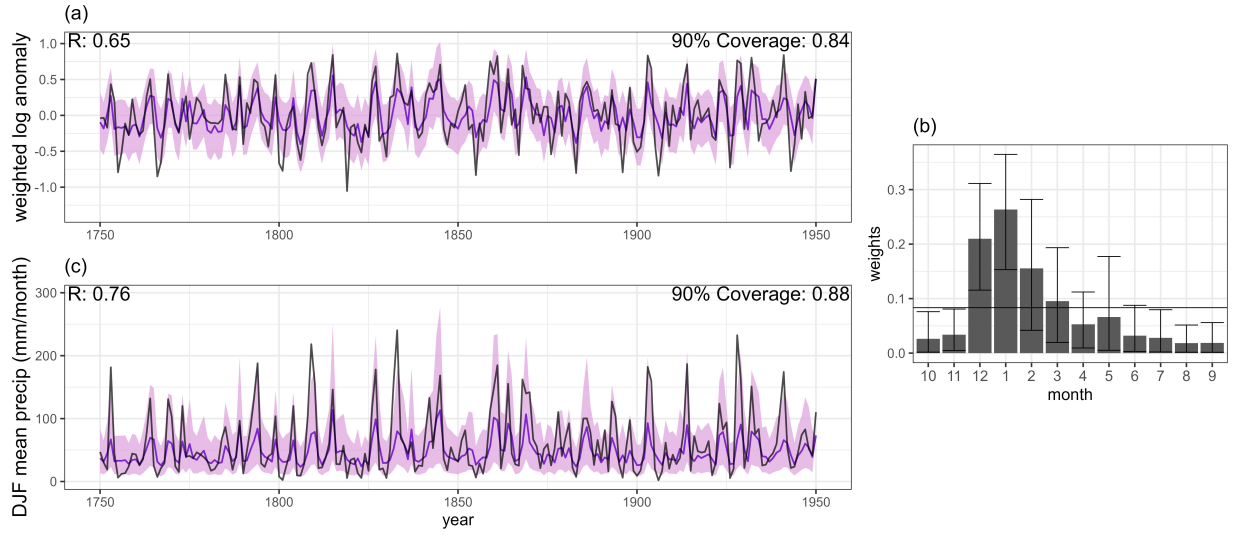


Figure A.2: Demonstration of the proposed model with simplified VS-Lite pseudoproxies. As in Fig. 2, but using VS-Lite to produce the pseudoproxies with SNR of 0.75 and setting growth outside of DJF to 0.

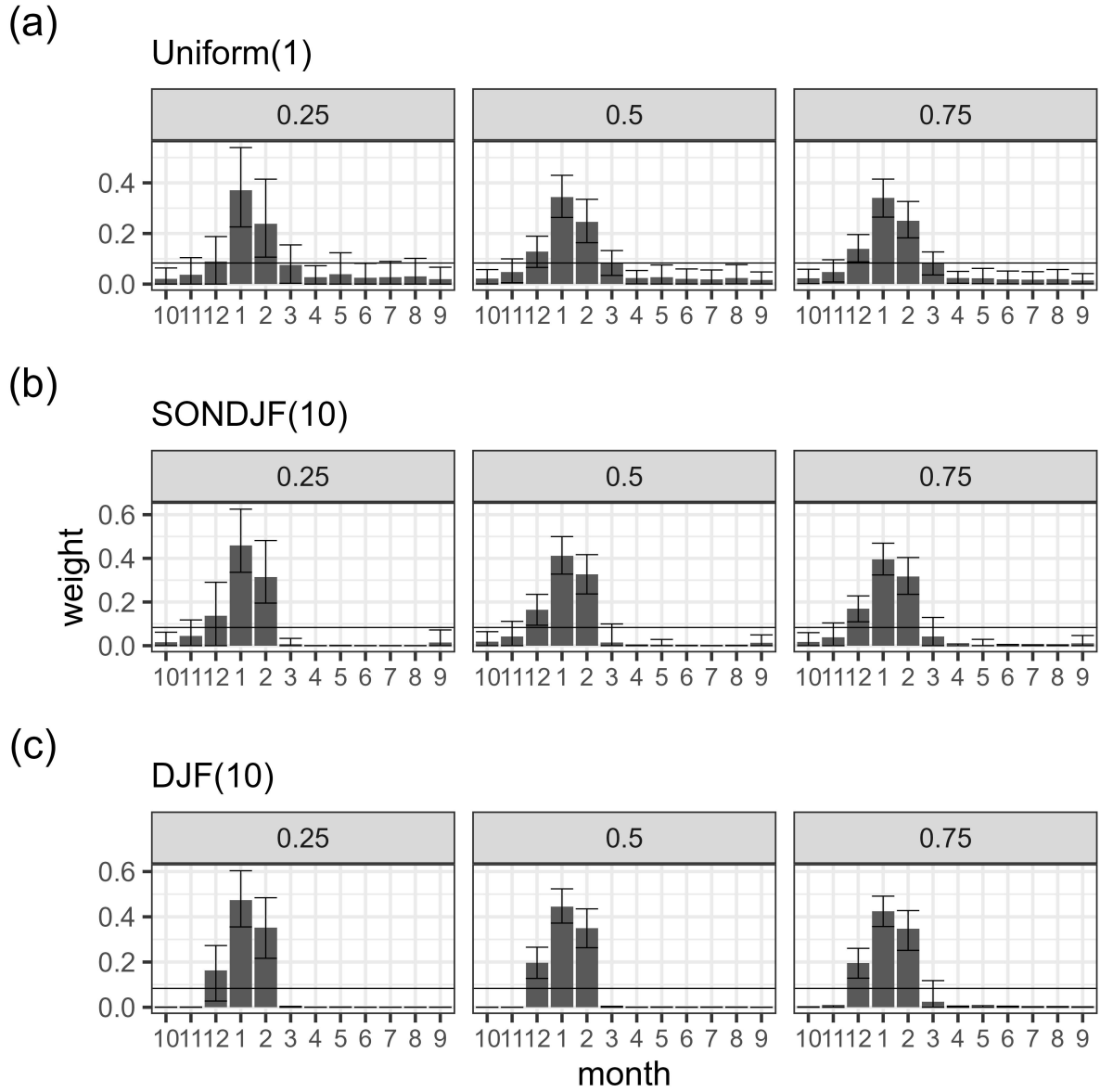


Figure A.3: Sensitivity of the weight inference to prior choice and SNR in VS-Lite pseudoproxies with growing season truncated to lie solely in DJF. The posterior mean and 90% credible intervals of monthly weights for pseudoproxies with varying SNR (0.25, 0.5, and 0.75), using (a) an uninformative prior with $w_{\text{input}} = 1$ in all months, (b) a prior with $w_{\text{input}} = 10$ in SONDJF and 1 in all other months, and (c) a prior with $w_{\text{input}} = 10$ in DJF and 1 in all other months.

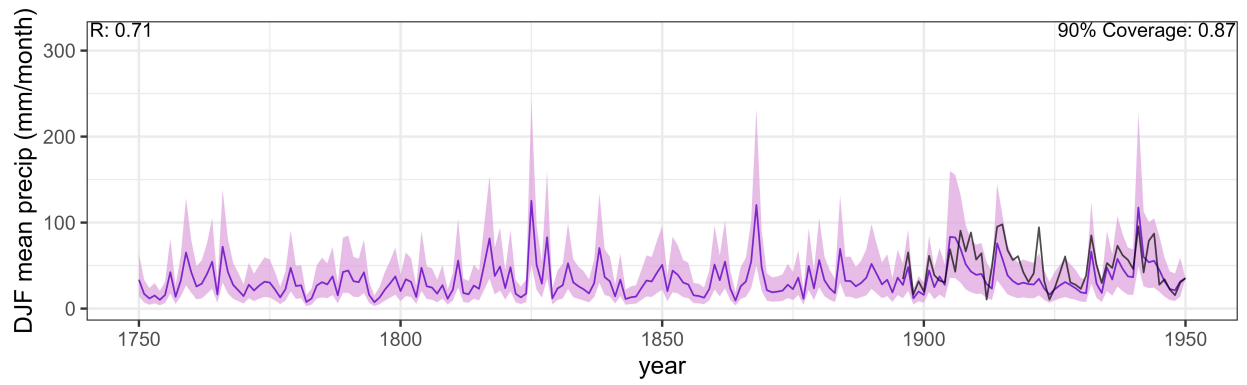


Figure A.4: Reference model for the blue oaks. The baseline model applied to the same blue oaks, choosing DJF as the reference season to facilitate comparisons with Figure 3c. Calibration and reconstruction time periods match that of Section 3.3.

Appendix B

Chapter 3 Appendix

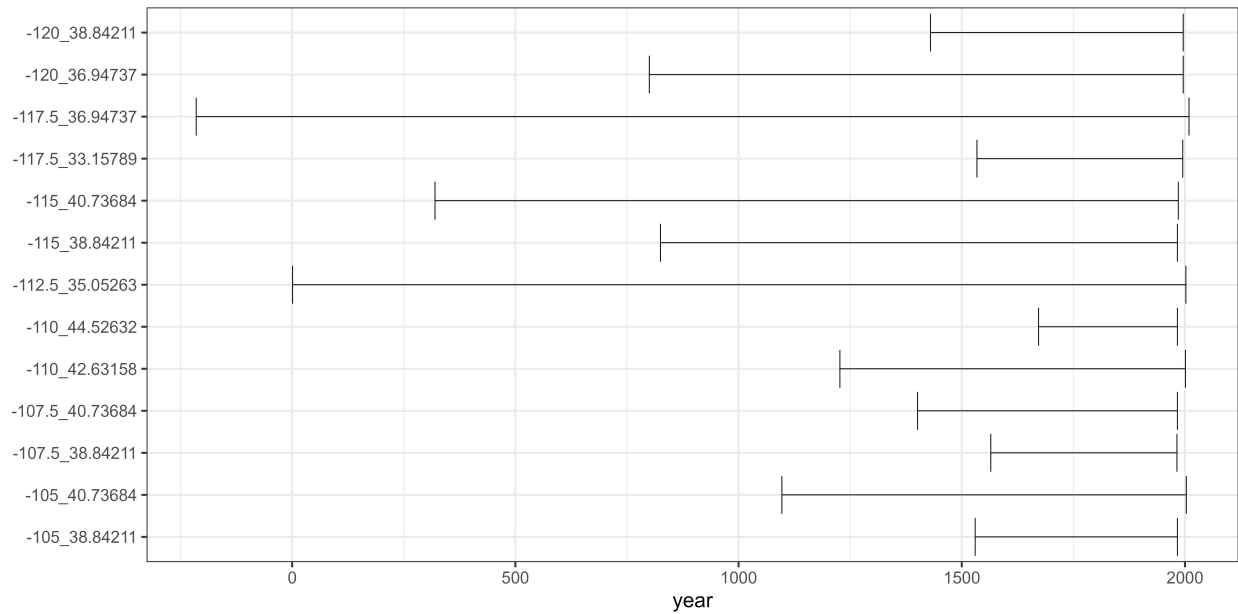


Figure B.1: Temporal availability of proxies at each gridcell in the subset of PAGES2K being used. Each row represents a gridcell containing proxy data, with the length of the bar representing the years with data in at least one chronology.

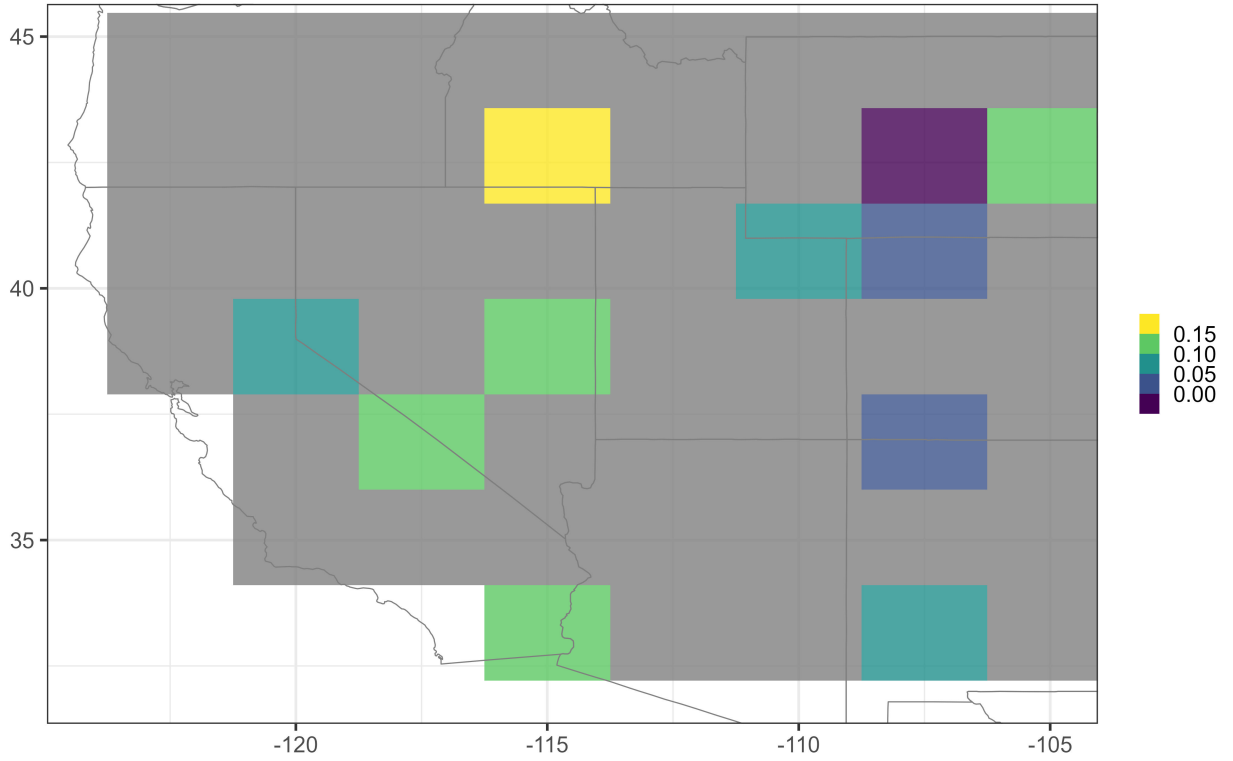


Figure B.2: Ridge regression coefficients (i.e. the posterior mean coefficients) for estimating the spatial mean over all shown gridcells, given pseudoproxies in the colored gridcells. Each location's temperature time series is normalized to have mean 0 and variance one prior to model fitting, so that the coefficient is unitless.

B.1 Prior specification

The following prior distributions are shared across all model iterations, unless specified otherwise.

$$X_1 \sim N_{n_s}(\mu_x, \sigma_x^2)$$

$$\eta_t \sim N(0, \sigma_\eta^2)$$

$$\sigma_{\eta(0,\infty)} \sim N(\mu_\eta, 5^2)$$

$$\alpha_{(0,1)} \sim N(0, 1)$$

$$\begin{aligned}
\rho &\sim \text{Inverse-Gamma}(5, 5) \\
\beta_{0,i} &\sim N(\mu_{b0}, 5^2) \\
\beta_{1,i} &\sim N(\mu_{b1}, 5^2) \\
\mu_{b0} &\sim N(0, 10^2) \\
\mu_{b1} &\sim N(0, 10^2) \\
\sigma_{i(0,\infty)}^2 &\sim N(0, 5^2) \\
\mu_x &\sim N(0, 5^2) \\
\sigma_{x(0,\infty)}^2 &\sim N(0, 5^2) \\
\mu_\eta &\sim N(0, 10^2)
\end{aligned}$$

X refers to a yearly climate variable, standardized at each spatial location to have mean zero and unit standard deviation. Subscripts on the normal distribution are used to indicate the dimension of multivariate normal distributions. For indexed variables with univariate priors, the prior distribution is iid across the indices. Interval subscripts on the variable name refer to parameters whose distributions are truncated to fit the described interval.

Models are fit using the probabilistic programming language **Stan** [184], accessed through the programming language **R**. Each model is run with 4 chains of 5000 iterations for all experiments, with the first half of the iterations in each chain used for warmup. Convergence is assessed using $\hat{R}$, B-FMI, and bulk effective size metrics, in addition to visual inspection of the draws.

Appendix C

Chapter 4 Appendix

C.1 Abbreviations

The following abbreviations are used in the manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of Open Access Journals
AD	Alzheimer's Disease
PD	Parkinson's Disease
CSF	Cerebrospinal Fluid
LEDD	Levodopa Equivalent Daily Dosage
GBA	Glucosylceramidase
OR	Odds Ratio
ROC	Receiver Operating Characteristic
AUC	Area Under the ROC Curve
MSEA	Metabolite Set Enrichment Analysis
PCA	Principal Component Analysis
CE	Cholesterol Ester
CER	Ceramides
DAG	Diacylglycerol
DCER	Dihydroceramides
FFA	Free Fatty Acids
HCER	Hexosylceramides
LCER	Lactosylceramide
LPC	Lysophosphatidylcholine
LPE	Lysophosphatidylethanolamine
PC	Phosphatidylcholine
PE	Phosphatidylethanolamine
SM	Sphingomyelin
TAG	Triacylglycerol

C.2 Summary of data by phenotype

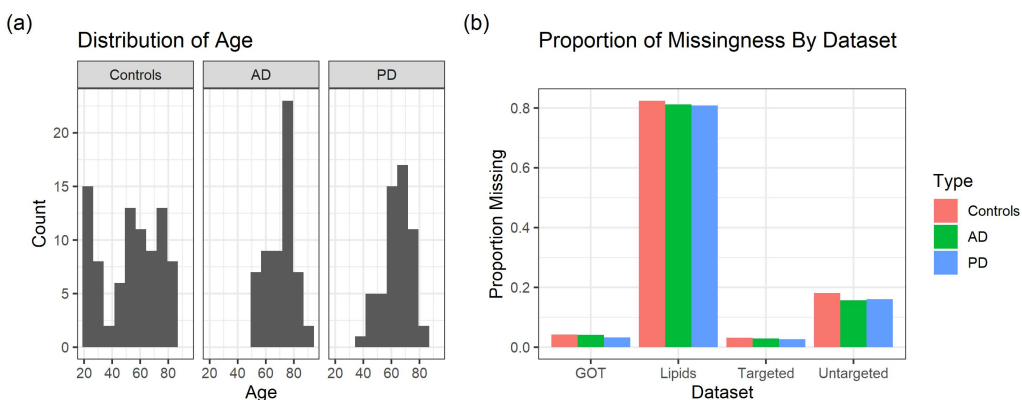


Figure C.1: (a) Distribution of subject age, split by phenotype. (b) a comparison of missingness between profiles, split by phenotype

C.3 Flowchart of analysis

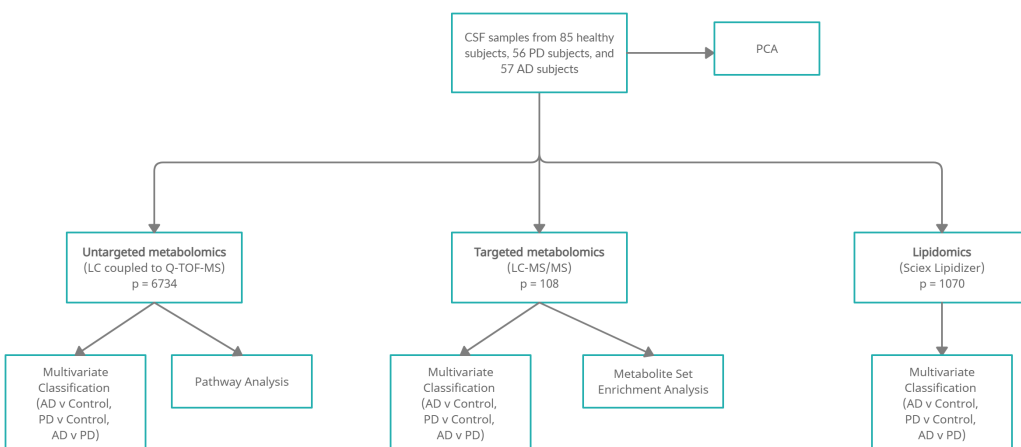


Figure C.2: Flowchart outlining the analysis performed on each of the three profiles.

C.4 Distribution in date of LP draws

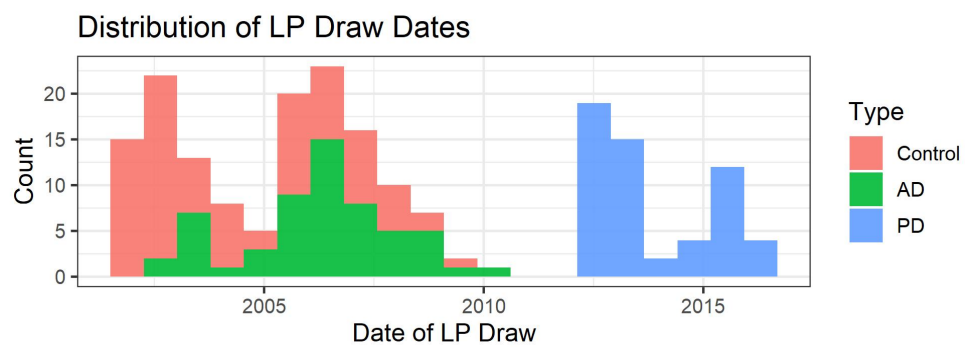


Figure C.3: Distribution of sample collection date, split by phenotype.

C.5 Univariate untargeted patterns of missingness

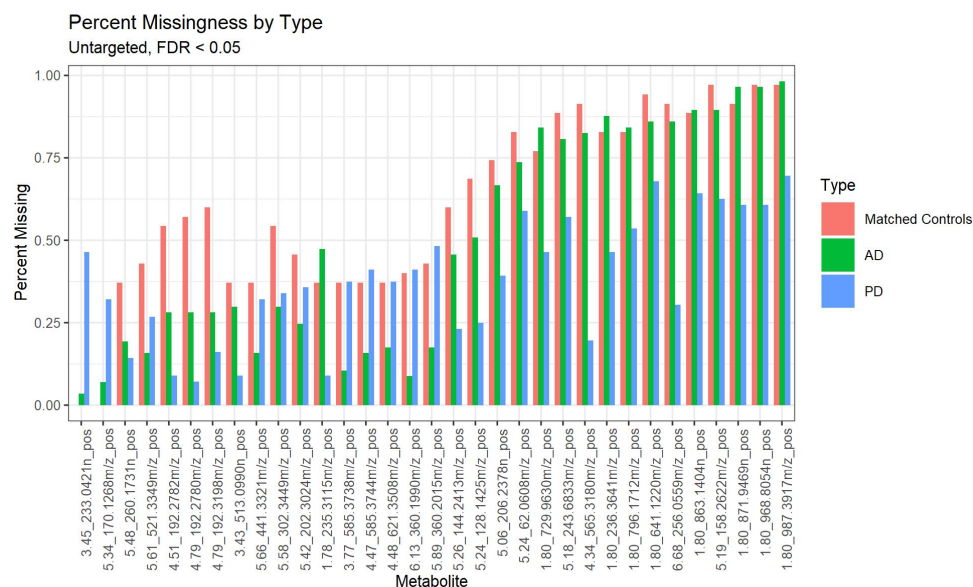


Figure C.4: For metabolites with widely varying amounts of abundance missingness by phenotype (AD, PD, age-matched controls), PD tend to have the least missing data. Plotted is the percent missingness of untargeted metabolites for which univariate logistic regressions classifying its missingness across subjects using phenotype contained a $FDR < 0.05$. The controls used for this analysis were found by matching each member of the AD/PD cohort to the control of closest age, removing duplicates. Metabolites are listed by their retention time, neutral mass (DA), and mode, with each value separated by an underscore

C.6 Precision-Recall curves and threshold-dependent metrics

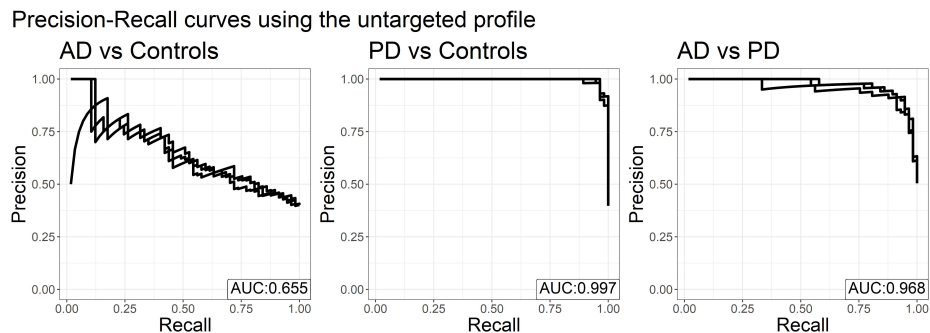


Figure C.5: Precision-Recall Curves for binomial elastic net regressions classifying (a) controls against subjects with AD, (b) controls against subjects with PD, and (c) subjects with AD against subjects with PD, using the same predictions as Figure 4.2. Each line represents models formed using one of the five missing data imputed datasets. The average Area Under the Curve (AUC) is displayed in the bottom right.

Metric	AD v C	PD v C	AD v PD
F1 score	0.63	0.97	0.93
Sensitivity	0.79	0.96	0.95
Specificity	0.52	0.99	0.91
Positive predictive value	0.52	0.98	0.92
Negative predictive value	0.79	0.98	0.94

Table C.1: A display of various cutoff-dependent metrics, using the same leave one out predictions displayed in Figures 4.2 and C.5. In each case, the cutoff is chosen to maximize F1 score.

C.7 ROC curves of predictive models including age and sex

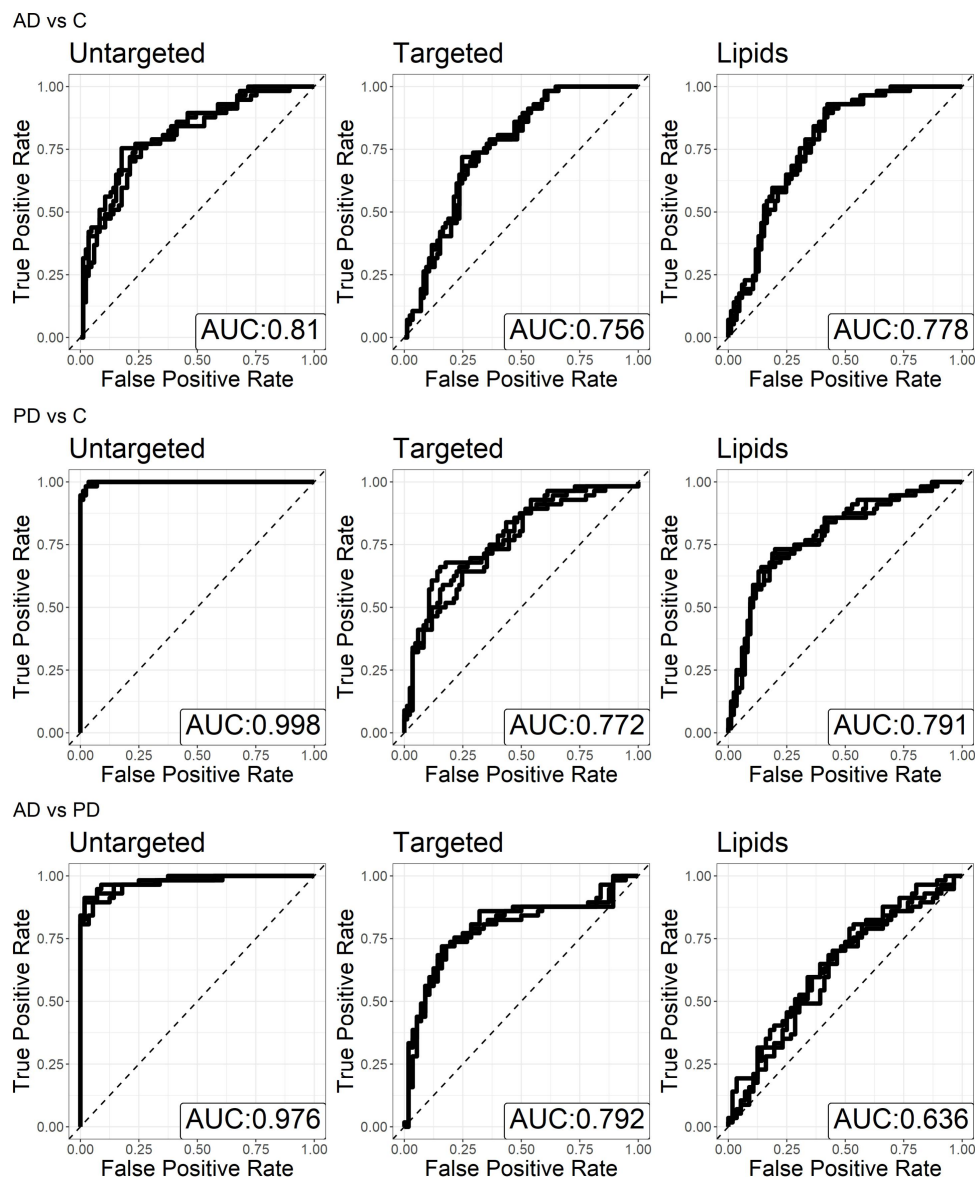


Figure C.6: ROCs displaying leave one out predictive accuracy of models including age and sex rather than detrending each metabolite. Age and sex are not penalized in the elastic net procedure to reflect the known associations between age and sex with AD and PD. As expected, we generally find that the predictive performance of these models is slightly better than those presented in the main analysis.

C.8 MSEA output tables

Table C.2: Metabolite Set Enrichment Analysis: Output of MSEA on the targeted profile based on metabolites significant for PD classification in univariate logistic regression. Set are the pre-defined metabolite sets in the SMPDB (left) and CSF disease library (right). Total is the number of metabolites in these sets, hits is the number of metabolites shared between the pre-defined sets and our input list, and expected is the expected number of metabolites shared by random chance, using the cumulative hypergeometric distribution.

SMPDB

Set	Total	Expected	Hits
Carnitine Synthesis	3	0.67	2
Betaine Metabolism	4	0.90	2
Methionine Metabolism	8	1.79	3
Tyrosine Metabolism	3	0.67	1
Glycine and Serine Metabolism	13	2.91	4
Arginine and Proline Metabolism	8	1.79	2
Tryptophan Metabolism	8	1.79	2
Histidine Metabolism	4	0.90	1
Urea Cycle	5	1.12	1

CSF disease library

Set	Total	Expected	Hits
Aging-Related Metabolites	3	1.1	2
Leukemia	13	4.8	3
Alzheimer's Disease	14	5.2	3
Different Seizure Disorders	12	4.5	1
Schizophrenia	13	4.8	1

C.9 Removing batch effects for PD vs Control analysis

Batch effects and other sources of unwanted variation are often a problem in metabolomic analyses [185]. As mentioned in the discussion, the result of near perfect classification of the PD against controls in this study is largely limited by potential artifacts introduced by the processing differences between the PD samples, compared to the AD and control samples. For example, CSF was collected from the AD and control subjects much earlier than the PD samples, which makes it difficult to disentangle the impacts of disease phenotype on metabolite profiles from the impacts of long term storage time. To explore the effect of storage time in our dataset, we isolated just the control subjects, whose samples were collected and stored over a period of 7 years (see Figure C.3). Then we used univariate linear models to regress each metabolite against storage time, and extracted p -values for the test that the slope is zero. We find that of the 41 targeted metabolites reported in Table 3, 9 of them had a relationship with storage time (i.e had a BH-corrected p value less than 0.05). However, 5 out of 9 of these metabolites exhibited the opposite relationship with time compared to what we would expect if storage time was the primary driver of the predictive power in the main analysis. For example, Table 3 reports that PD subjects in our dataset had decreased concentrations of HIAA compared to controls. If storage time was the true source of this association, then we would expect HIAA to have a negative relationship with storage time (because PD samples were collected after control samples). However, HIAA is positively associated with storage time, suggesting that storage time alone is not the sole driver of our predictive models. The four metabolites whose relationship with PD is consistent with their relationship with storage time are cystine, xanthosine, glycoamine, and 4-aminobutyric acid. We find that removing these four metabolites from our predictive models does not change the classification performance reported in Section 4.3.2, still retaining an AUC around 0.8. Future work is needed to further study the impact of long term storage time on metabolomic profiles, but this exploratory analysis of the targeted profiles of control subjects over a 9-year span provides weak evidence that our findings are not driven solely by storage time. However,

there are other potential statistical artifacts that might have been introduced by processing differences between batches, and most of them cannot be measured or corrected for in our dataset. In the remainder of this section, we explore methods of mitigating these concerns by leveraging our unique situation of having three different disease phenotypes but only two batches.

The first approach involves retroactively comparing predictions from the models presented in Section 4.2.3, taking advantage of the following line of reasoning: if the primary signal driving our model predictions is batch effects, then a model fit to classify AD against PD should tend to classify control subjects as AD, and a model fit to classify PD against controls should tend to classify AD subjects as controls (because AD and control subjects were processed together). We find that this is not the case in our data – the class probabilities in both cases are centered near 0.5, which suggests that our models are picking up more signal than just the batch effect. The second approach instead attempts to remove batch variation before fitting a model. Here, the goal is to fit a model to classify PD from controls using data which has batch effects removed. To remove batch effects from the metabolite concentrations, we add a preprocessing step to our methods to remove any signal in the dataset that classifies AD from PD. In particular, we do the following in each cross-validated loop:

for each observation (control or PD):

1. Fit an elastic net regression model to classify on AD/PD to get coefficient vector β . If the left-out observation is a PD subject, exclude it from this model.
2. Orthogonalize all control and PD observations ($\{x_i\}_{i=1}^n$) by the coefficient vector, computing

$$x_i^* := x_i - \frac{x_i \cdot \beta}{\|\beta\|^2} \beta \tag{C.1}$$

3. Use the orthogonalized data to fit a model to classify PD from controls (using the n-1 observations)
4. Use the model to predict the class probability on the left out observation

b Intuitively, the orthogonalized dataset resulting from this procedure is unable to distinguish differences between AD and PD. This includes true metabolomic signal separating the two neurodegenerative diseases, but more importantly, it includes processing differences between the Control/AD batch and the PD batch. Fitting models on this new orthogonalized dataset, we find that we mostly lose the ability to classify AD (AUC between 0.5-0.7 across the different profiles and imputations), but still retain some predictive power to classify PD from controls (AUC between 0.7-0.9). Figure C.7 displays the predictive accuracy of these models fit on the orthogonalized targeted profiles. To evaluate the effectiveness of the orthogonalization, we adapt ideas from [185], comparing the results of univariate logistic regressions fit to classify PD from controls, before and after orthogonalization. We find that of eight differentially abundant metabolites in the targeted profile before orthogonalization, four of them remain differentially abundant after orthogonalization (here, differentially abundant refers to BH-corrected p -values < 0.05). These metabolites are 2-aminoadipic acid, 6-methyl-DL-tryptophan, adenosyl-L-homocysteine, and lysine. For the other profiles, the orthogonalization process makes minimal changes to individual metabolites, hence univariate tests produce identical results before and after orthogonalization

ROC curves after batch correction (targeted profile)

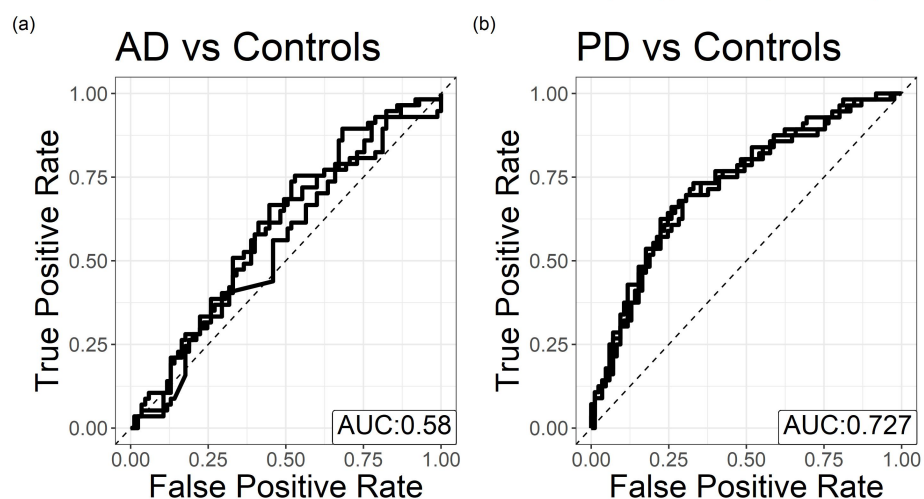


Figure C.7: ROC curves displaying leave one out predictive accuracy of models fit on the targeted profile after the batch orthogonalization procedure. We find that these models are largely unable to distinguish AD from controls after this procedure, and are still able to distinguish PD from controls (but with less predictive accuracy compared to the results in main analysis).

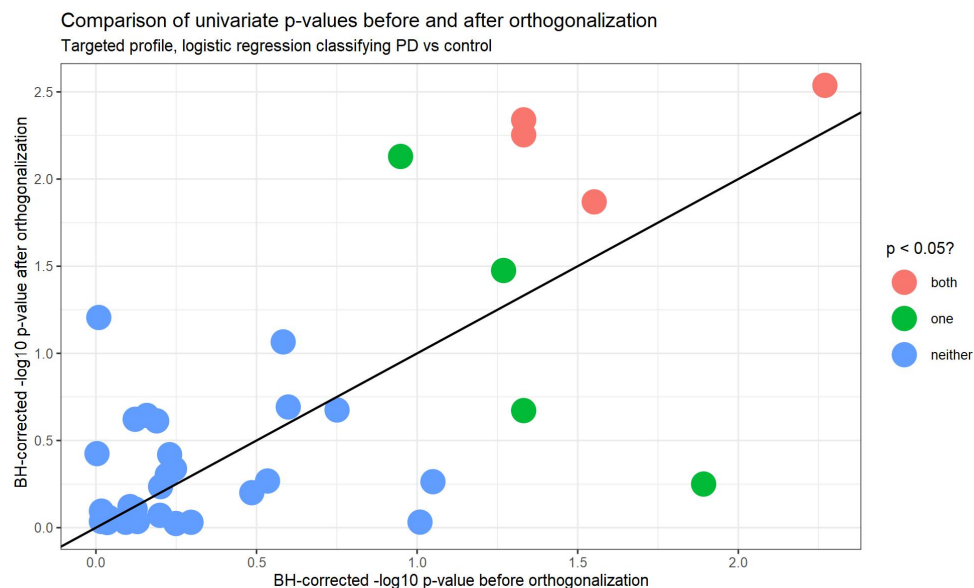


Figure C.8: Comparison of univariate p -values before and after orthogonalization, using univariate logistic regressions fit on the targeted profile to classify PD against controls. p -values are Benjamini-Hochberg corrected within each dataset, with a $-\log_{10}$ transformation applied. The $y = x$ line is also drawn, with values above the line representing metabolites which are more differentially abundant after orthogonalization. Points are colored according to false discovery rate cutoffs of 0.05, indicating whether a metabolite is considered differentially abundant in both analyses, only one analysis, or in neither analysis.

C.10 Lipids tables for discriminating AD from PD

Table C.3: Names and OR (associated with a standard deviation increase in concentration) for lipids retained in all five elastic net models fit on the full data to classify PD patients against AD patients. A metabolite with $OR > 1$ indicates that higher concentration is associated with AD in our models. Only ORs with magnitude greater than 1.1 or less than 0.9 are shown here.

Lipid	OR $\pm$ 2SD (> 1)	Lipid	OR $\pm$ 2SD (< 1)
PC(18:1/18:2)	1.79 (1.32, 2.41)	PC(18:0/18:2)	0.62 (0.38, 1)
PC(18:1/20:4)	1.79 (1.55, 2.05)	PE(P-16:0/18:1)	0.64 (0.44, 0.91)
FFA(20:3)	1.75 (1.62, 1.9)	PE(P-18:0/20:4)	0.7 (0.63, 0.79)
DAG(16:0/18:1)	1.55 (1.25, 1.93)	TAG48:1-FA16:0	0.72 (0.55, 0.93)
FFA(18:3)	1.43 (1.13, 1.82)	FFA(20:4)	0.72 (0.63, 0.83)
CE(16:1)	1.34 (1.21, 1.48)	TAG46:0-FA16:0	0.75 (0.57, 0.99)
PC(18:0/22:6)	1.27 (1.06, 1.52)	CER(14:0)	0.78 (0.61, 0.99)
TAG53:2-FA18:1	1.23 (0.99, 1.54)	PE(P-18:0/22:6)	0.78 (0.73, 0.83)
PC(16:0/22:5)	1.19 (0.95, 1.48)	HCER(18:0)	0.79 (0.72, 0.88)
CE(20:3)	1.16 (0.93, 1.45)	PE(18:0/20:4)	0.8 (0.63, 1.02)
PC(18:1/16:1)	1.14 (1.01, 1.28)	TAG48:0-FA14:0	0.82 (0.68, 0.98)
		TAG55:4-FA18:1	0.84 (0.72, 0.99)
		DAG(20:0/20:0)	0.86 (0.65, 1.14)
		PC(18:0/18:0)	0.87 (0.76, 1)
		CE(15:0)	0.88 (0.75, 1.03)

C.11 Metabolites appearing in all leave one out models

To make the most of the limited size of our dataset, discriminative performance was estimated by fitting a separate model to each sample for each of the five missing data imputations,

as described in Section 4.2-4.2.3. This results in $5n$ models for each sample in each profile, where n is the total number of samples used for that model. For example, there are 56 PD patients and 85 controls, so there are $5 * (56 + 85) = 705$ models fit to discriminate PD against controls using the targeted profile. Here, we list all targeted metabolite and lipids that appear in all $5n$ models. We note that there were no untargeted metabolites that appeared in all models, which explains their exclusion from these models. The average odds ratios across these models are reported to describe the sign and relative weight of these features in their respective profiles. A two standard deviation interval is also included to quantify uncertainty across leave one out iterations. The first phenotype listed in each table represents the “1” class in the odds ratios, e.g. threonine has a mean odds ratio greater than 1 in the PD vs control table, meaning that higher concentrations of threonine were consistently found in PD patients in our data.

Table C.4: Targeted metabolites and lipids appearing in all leave one out models, for each modeling task.

PD vs control			AD vs control			AD vs PD		
Profile	Name	Mean OR ± 2SD	Profile	Name	Mean OR ± 2SD	Profile	Name	Mean OR ± 2SD
Lipids	HCER(24:0)	1.22 (1.15, 1.3)	Targeted	1-Methyladenosine	1.27 (1.02, 1.58)	Lipids	FFA(20:3)	4.03 (2.66, 6.11)
Targeted	Threonine	1.19 (1.08, 1.32)	Targeted	Sarcosine	1.12 (1.05, 1.19)	Targeted	Glycylproline	0.53 (0.33, 0.85)
Targeted	Methylguanidine	1.19 (1.11, 1.26)	Targeted	Alanine	1.08 (1.03, 1.14)	Targeted	Lactate	1.57 (1.1, 2.23)
Lipids	PC(16:0/22:0)	1.16 (1.03, 1.3)	Lipids	CE(20:3)	1.07 (0.99, 1.15)	Targeted	Creatinine	0.75 (0.62, 0.9)
Targeted	Dimethylarginine	1.16 (1.03, 1.29)						

Bibliography

- [1] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant”. In: *Psychological Science* 22.11 (Nov. 2011). <https://doi.org/10.1177/0956797611417632>, pp. 1359–1366. ISSN: 0956-7976. DOI: [10.1177/0956797611417632](https://doi.org/10.1177/0956797611417632). (Visited on 03/03/2022).
- [2] Nate Breznau et al. “Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty”. In: *Proceedings of the National Academy of Sciences* 119.44 (Nov. 2022). <https://www.pnas.org/doi/abs/10.1073/pnas.2203150119>, e2203150119. DOI: [10.1073/pnas.2203150119](https://doi.org/10.1073/pnas.2203150119). (Visited on 05/26/2025).
- [3] Richard A. Kerr. “A North Atlantic Climate Pacemaker for the Centuries”. In: *Science* 288.5473 (June 2000). <https://www.science.org/doi/full/10.1126/science.288.5473.1984>, pp. 1984–1985. DOI: [10.1126/science.288.5473.1984](https://doi.org/10.1126/science.288.5473.1984). (Visited on 05/29/2025).
- [4] Ulf Büntgen et al. “The Influence of Decision-Making in Tree Ring-Based Climate Reconstructions”. In: *Nature Communications* 12.1 (June 2021). <https://www.nature.com/articles/s41467-021-23627-6>, p. 3411. ISSN: 2041-1723. DOI: [10.1038/s41467-021-23627-6](https://doi.org/10.1038/s41467-021-23627-6). (Visited on 02/03/2022).
- [5] Martin P. Tingley et al. “Piecing Together the Past: Statistical Insights into Paleoclimatic Reconstructions”. In: *Quaternary Science Reviews* 35 (Mar. 2012). <https://doi.org/10.1016/j.quascirev.2012.01.011>. (Visited on 03/03/2022).

- [//linkinghub.elsevier.com/retrieve/pii/S0277379112000248](https://linkinghub.elsevier.com/retrieve/pii/S0277379112000248), pp. 1–22. ISSN: 02773791. DOI: [10.1016/j.quascirev.2012.01.012](https://doi.org/10.1016/j.quascirev.2012.01.012). (Visited on 10/21/2021).
- [6] Sebastian Sippel et al. “Early-Twentieth-Century Cold Bias in Ocean Surface Temperature Observations”. In: *Nature* 635.8039 (Nov. 2024). <https://www.nature.com/articles/s41586-024-08230-1>, pp. 618–624. ISSN: 1476-4687. DOI: [10.1038/s41586-024-08230-1](https://doi.org/10.1038/s41586-024-08230-1). (Visited on 05/02/2025).
- [7] Nathan Hwangbo et al. “Predictive Modeling of Alzheimer’s and Parkinson’s Disease Using Metabolomic and Lipidomic Profiles from Cerebrospinal Fluid”. In: *Metabolites* 12.4 (Apr. 2022). <https://www.mdpi.com/2218-1989/12/4/277>, p. 277. ISSN: 2218-1989. DOI: [10.3390/metabo12040277](https://doi.org/10.3390/metabo12040277). (Visited on 03/23/2022).
- [8] Jessica E. Tierney et al. “Past Climates Inform Our Future”. In: *Science* 370.6517 (Nov. 2020), eaay3701. DOI: [10.1126/science.aay3701](https://doi.org/10.1126/science.aay3701).
- [9] Sandy P. Harrison, Patrick J. Bartlein, and I. Colin Prentice. “What Have We Learnt from Palaeoclimate Simulations?” In: *Journal of Quaternary Science* 31.4 (2016), pp. 363–385. ISSN: 1099-1417. DOI: [10.1002/jqs.2842](https://doi.org/10.1002/jqs.2842).
- [10] Tamsin L. Edwards, Michel Crucifix, and Sandy P. Harrison. “Using the Past to Constrain the Future: How the Palaeorecord Can Improve Estimates of Global Warming”. In: *Progress in Physical Geography* 31.5 (Oct. 2007). <https://doi.org/10.1177/0309133307083295>, pp. 481–500. ISSN: 0309-1333. DOI: [10.1177/0309133307083295](https://doi.org/10.1177/0309133307083295). (Visited on 04/02/2025).
- [11] Julien Emile-Geay et al. “A Global Multiproxy Database for Temperature Reconstructions of the Common Era”. In: *Scientific Data* 4.1 (July 2017). <https://www.nature.com/articles/sdata201788>, p. 170088. ISSN: 2052-4463. DOI: [10.1038/sdata.2017.88](https://doi.org/10.1038/sdata.2017.88). (Visited on 05/17/2022).
- [12] Scott St. George. “An Overview of Tree-Ring Width Records across the Northern Hemisphere”. In: *Quaternary Science Reviews* 95 (July 2014). <https://www.elsevier.com/locate/quascirev>.

- [sciencedirect.com/science/article/pii/S0277379114001607](https://www.sciencedirect.com/science/article/pii/S0277379114001607), pp. 132–150. ISSN: 0277-3791. DOI: [10.1016/j.quascirev.2014.04.029](https://doi.org/10.1016/j.quascirev.2014.04.029). (Visited on 12/06/2024).
- [13] Raphael Neukom et al. “Consistent Multidecadal Variability in Global Temperature Reconstructions and Simulations over the Common Era”. In: *Nature Geoscience* 12.8 (Aug. 2019). <https://www.nature.com/articles/s41561-019-0400-0>, pp. 643–649. ISSN: 1752-0908. DOI: [10.1038/s41561-019-0400-0](https://doi.org/10.1038/s41561-019-0400-0). (Visited on 11/20/2024).
- [14] Jianglin Wang et al. “Tree-Ring Inferred Annual Mean Temperature Variations on the Southeastern Tibetan Plateau during the Last Millennium and Their Relationships with the Atlantic Multidecadal Oscillation”. In: *Climate Dynamics* 43.3 (Aug. 2014). <https://doi.org/10.1007/s00382-013-1802-0>, pp. 627–640. ISSN: 1432-0894. DOI: [10.1007/s00382-013-1802-0](https://doi.org/10.1007/s00382-013-1802-0). (Visited on 10/02/2023).
- [15] T. J. Blasing, D. W. Stahle, and D. N. Duvick. “Tree Ring-Based Reconstruction of Annual Precipitation in the South-Central United States from 1750 to 1980”. In: *Water Resources Research* 24.1 (1988). <https://onlinelibrary.wiley.com/doi/abs/10.1029/WR024i001p00163>, pp. 163–171. ISSN: 1944-7973. DOI: [10.1029/WR024i001p00163](https://doi.org/10.1029/WR024i001p00163). (Visited on 10/02/2023).
- [16] Lucie J. Lücke et al. “Orbital Forcing Strongly Influences Seasonal Temperature Trends During the Last Millennium”. In: *Geophysical Research Letters* 48.4 (2021). <https://onlinelibrary.wiley.com/doi/abs/10.1029/2020GL088776>, e2020GL088776. ISSN: 1944-8007. DOI: [10.1029/2020GL088776](https://doi.org/10.1029/2020GL088776). (Visited on 05/17/2022).
- [17] A. P. Williams et al. “Tree Rings and Observations Suggest No Stable Cycles in Sierra Nevada Cool-Season Precipitation”. In: *Water Resources Research* 57.3 (2021), e2020WR028599. ISSN: 1944-7973. DOI: [10.1029/2020WR028599](https://doi.org/10.1029/2020WR028599).
- [18] Feng Zhu et al. “Resolving the Differences in the Simulated and Reconstructed Temperature Response to Volcanism”. In: *Geophysical Research Letters* 47.8 (2020). <https://doi.org/10.1029/2020GL088776>.

- [//onlinelibrary.wiley.com/doi/abs/10.1029/2019GL086908](https://onlinelibrary.wiley.com/doi/abs/10.1029/2019GL086908), e2019GL086908. ISSN: 1944-8007. DOI: [10.1029/2019GL086908](https://doi.org/10.1029/2019GL086908). (Visited on 01/16/2025).
- [19] Jessie K. Pearl et al. “Reconstructing Northeastern United States Temperatures Using Atlantic White Cedar Tree Rings”. In: *Environmental Research Letters* 12.11 (Nov. 2017). <https://dx.doi.org/10.1088/1748-9326/aa8f1b>, p. 114012. ISSN: 1748-9326. DOI: [10.1088/1748-9326/aa8f1b](https://doi.org/10.1088/1748-9326/aa8f1b). (Visited on 12/06/2024).
- [20] Kevin J. Anchukaitis and Jason E. Smerdon. “Progress and Uncertainties in Global and Hemispheric Temperature Reconstructions of the Common Era”. In: *Quaternary Science Reviews* 286 (June 2022), p. 107537. ISSN: 0277-3791. DOI: [10.1016/j.quascirev.2022.107537](https://doi.org/10.1016/j.quascirev.2022.107537).
- [21] Jessica E. Tierney et al. “Bayesian Calibration of the Mg/Ca Paleothermometer in Planktic Foraminifera”. In: *Paleoceanography and Paleoclimatology* 34.12 (2019), pp. 2005–2030. ISSN: 2572-4525. DOI: [10.1029/2019PA003744](https://doi.org/10.1029/2019PA003744).
- [22] Jinbao Li et al. “Tree-Ring Based Drought Reconstruction for the Central Tien Shan Area in Northwest China”. In: *Geophysical Research Letters* 33.7 (2006). <https://onlinelibrary.wiley.com/doi/abs/10.1029/2006GL025803>. ISSN: 1944-8007. DOI: [10.1029/2006GL025803](https://doi.org/10.1029/2006GL025803). (Visited on 10/02/2023).
- [23] Robert Tardif et al. “Last Millennium Reanalysis with an Expanded Proxy Database and Seasonal Proxy Modeling”. In: *Climate of the Past* 15.4 (July 2019). <https://cp.copernicus.org/articles/15/1251/2019/>, pp. 1251–1273. ISSN: 1814-9324. DOI: [10.5194/cp-15-1251-2019](https://doi.org/10.5194/cp-15-1251-2019). (Visited on 05/05/2022).
- [24] Jernej Jevšenak and Tom Levanič. “*dendroTools*: R Package for Studying Linear and Nonlinear Responses between Tree-Rings and Daily Environmental Data”. In: *Dendrochronologia* 48 (Apr. 2018). <https://www.sciencedirect.com/science/article/pii/S1125786517301613>, pp. 32–39. ISSN: 1125-7865. DOI: [10.1016/j.dendro.2018.01.005](https://doi.org/10.1016/j.dendro.2018.01.005). (Visited on 11/15/2024).

- [25] Jun Hu, Julien Emile-Geay, and Judson Partin. “Correlation-Based Interpretations of Paleoclimate Data – Where Statistics Meet Past Climates”. In: *Earth and Planetary Science Letters* 459 (Feb. 2017). <https://www.sciencedirect.com/science/article/pii/S0012821X16306823>, pp. 362–371. ISSN: 0012-821X. DOI: [10.1016/j.epsl.2016.11.048](https://doi.org/10.1016/j.epsl.2016.11.048). (Visited on 01/29/2025).
- [26] T J Blasing, A M Solomon, and D N Duval. “Response Functions Revisited”. In: *Tree-Ring Bulletin* 44 (1984), pp. 1–15.
- [27] Peter F. Craigmile et al. “Paleoclimate Reconstruction: Looking Backwards to Look Forward”. In: *Handbook of Environmental and Ecological Statistics*. 2019.
- [28] Jason E. Smerdon. “Climate Models as a Test Bed for Climate Reconstruction Methods: Pseudoproxy Experiments”. In: *WIREs Climate Change* 3.1 (2012), pp. 63–77. ISSN: 1757-7799. DOI: [10.1002/wcc.149](https://doi.org/10.1002/wcc.149).
- [29] H. L. O. McClelland et al. “Statistical Uncertainty in Paleoclimate Proxy Reconstructions”. In: *Geophysical Research Letters* 48.15 (2021). <https://onlinelibrary.wiley.com/doi/abs/10.1029/2021GL092773>, e2021GL092773. ISSN: 1944-8007. DOI: [10.1029/2021GL092773](https://doi.org/10.1029/2021GL092773). (Visited on 05/11/2022).
- [30] Bette L. Otto-Bliesner et al. “Climate Variability and Change since 850 CE: An Ensemble Approach with the Community Earth System Model”. In: *Bulletin of the American Meteorological Society* 97.5 (May 2016). <https://journals.ametsoc.org/view/journals/bams/97/5/bams-d-14-00233.1.xml>, pp. 735–754. ISSN: 0003-0007, 1520-0477. DOI: [10.1175/BAMS-D-14-00233.1](https://doi.org/10.1175/BAMS-D-14-00233.1). (Visited on 08/24/2023).
- [31] D. W. Stahle et al. “NOAA/WDS Paleoclimatology - Los Lobos Creek - QUDG - ITRDB CA660”. In: (2012). [Dataset] <https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/noaa-tree-8558/html>. DOI: [10.25921/szhc-ge03](https://doi.org/10.25921/szhc-ge03). (Visited on 03/14/2022).

- [32] D. W. Stahle et al. “NOAA/WDS Paleoclimatology - Figueroa Mountain - QUDG - ITRDB CA656”. In: (2012). [Dataset]<https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/noaa-tree-8558/html>. DOI: [10.25921/a2e3-d017](https://doi.org/10.25921/a2e3-d017). (Visited on 03/14/2022).
- [33] David Meko et al. “Inferring Precipitation-Anomaly Gradients from Tree Rings”. In: *Quaternary International*. PACLIM: Proceedings of the 24th Pacific Climate Workshop, 2009 235.1 (Apr. 2011). <https://www.sciencedirect.com/science/article/pii/S1040618210003654>, pp. 89–100. ISSN: 1040-6182. DOI: [10.1016/j.quaint.2010.09.006](https://doi.org/10.1016/j.quaint.2010.09.006). (Visited on 06/27/2024).
- [34] D. W. Stahle et al. “The Ancient Blue Oak Woodlands of California: Longevity and Hydroclimatic History”. In: *Earth Interactions* 17.12 (Aug. 2013), pp. 1–23. DOI: [10.1175/2013EI000518.1](https://doi.org/10.1175/2013EI000518.1).
- [35] Daniel Griffin and Kevin J. Anchukaitis. “How Unusual Is the 2012–2014 California Drought?” In: *Geophysical Research Letters* 41.24 (2014), pp. 9017–9023. ISSN: 1944-8007. DOI: [10.1002/2014GL062433](https://doi.org/10.1002/2014GL062433).
- [36] Christopher Daly et al. “A Knowledge-Based Approach to the Statistical Mapping of Climate”. In: *Climate Research* 22.2 (Sept. 2002). <https://www.int-res.com/abstracts/cr/v22/n2/p99-113/>, pp. 99–113. ISSN: 0936-577X, 1616-1572. DOI: [10.3354/cr022099](https://doi.org/10.3354/cr022099). (Visited on 11/04/2024).
- [37] J. Haslett et al. “Bayesian Palaeoclimate Reconstruction”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169.3 (2006). <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-985X.2006.00429.x>, pp. 395–438. ISSN: 1467-985X. DOI: [10.1111/j.1467-985X.2006.00429.x](https://doi.org/10.1111/j.1467-985X.2006.00429.x). (Visited on 04/19/2022).
- [38] Bo Li, Douglas W. Nychka, and Caspar M. Ammann. “The Value of Multiproxy Reconstruction of Past Climate”. In: *Journal of the American Statistical Association*

- 105.491 (Sept. 2010). <https://doi.org/10.1198/jasa.2010.ap09379>, pp. 883–895. ISSN: 0162-1459. DOI: [10.1198/jasa.2010.ap09379](https://doi.org/10.1198/jasa.2010.ap09379). (Visited on 10/28/2021).
- [39] Martin P. Tingley and Peter Huybers. “A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part II: Comparison with the Regularized Expectation–Maximization Algorithm”. In: *Journal of Climate* 23.10 (May 2010). <https://journals.ametsoc.org/view/journals/clim/23/10/2009jcli3016.1.xml>, pp. 2782–2800. ISSN: 0894-8755, 1520-0442. DOI: [10.1175/2009JCLI3016.1](https://doi.org/10.1175/2009JCLI3016.1). (Visited on 10/18/2021).
- [40] Martin P. Tingley and Peter Huybers. “A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part II: Comparison with the Regularized Expectation–Maximization Algorithm”. In: *Journal of Climate* 23.10 (May 2010). <https://journals.ametsoc.org/view/journals/clim/23/10/2009jcli3016.1.xml>, pp. 2782–2800. ISSN: 0894-8755, 1520-0442. DOI: [10.1175/2009JCLI3016.1](https://doi.org/10.1175/2009JCLI3016.1). (Visited on 10/18/2021).
- [41] Johannes P. Werner, Juerg Luterbacher, and Jason E. Smerdon. “A Pseudoproxy Evaluation of Bayesian Hierarchical Modeling and Canonical Correlation Analysis for Climate Field Reconstructions over Europe”. In: *Journal of Climate* 26.3 (Feb. 2013). <https://journals.ametsoc.org/view/journals/clim/26/3/jcli-d-12-00016.1.xml>, pp. 851–867. ISSN: 0894-8755, 1520-0442. DOI: [10.1175/JCLI-D-12-00016.1](https://doi.org/10.1175/JCLI-D-12-00016.1). (Visited on 07/08/2022).
- [42] Luis Barboza et al. “Reconstructing Past Temperatures from Natural Proxies and Estimated Climate Forcings Using Short- and Long-Memory Models”. In: *The Annals of Applied Statistics* 8.4 (Dec. 2014). <https://projecteuclid.org/journals/annals-of-applied-statistics/volume-8/issue-4/Reconstructing-past-temperatures-from-natural-proxies-and-estimated-climate-forcings/10.1214/14-AOS1214>

- 1214/14-AOAS785.full, pp. 1966–2001. ISSN: 1932-6157, 1941-7330. DOI: [10.1214/14-AOAS785](https://doi.org/10.1214/14-AOAS785). (Visited on 10/28/2021).
- [43] Luis A. Barboza et al. “Efficient Reconstructions of Common Era Climate via Integrated Nested Laplace Approximations”. In: *Journal of Agricultural, Biological and Environmental Statistics* 24.3 (Sept. 2019). <https://doi.org/10.1007/s13253-019-00372-4>, pp. 535–554. ISSN: 1537-2693. DOI: [10.1007/s13253-019-00372-4](https://doi.org/10.1007/s13253-019-00372-4). (Visited on 03/14/2023).
- [44] Matthew R. Schofield et al. “A Model-Based Approach to Climate Reconstruction Using Tree-Ring Data”. In: *Journal of the American Statistical Association* 111.513 (Jan. 2016). <https://doi.org/10.1080/01621459.2015.1110524>, pp. 93–106. ISSN: 0162-1459. DOI: [10.1080/01621459.2015.1110524](https://doi.org/10.1080/01621459.2015.1110524). (Visited on 03/22/2023).
- [45] Scott Steinschneider et al. “Hierarchical Regression Models for Dendroclimatic Standardization and Climate Reconstruction”. In: *Dendrochronologia* 44 (June 2017), pp. 174–186. ISSN: 1125-7865. DOI: [10.1016/j.dendro.2017.05.003](https://doi.org/10.1016/j.dendro.2017.05.003).
- [46] Ramzi Touchan, David Meko, and Malcolm K Hughes. “A 396-Year Reconstruction of Precipitation in Southern Jordan”. In: *JAWRA Journal of the American Water Resources Association* 35.1 (1999). <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1752-1688.1999.tb05451.x>, pp. 49–59. ISSN: 1752-1688. DOI: [10.1111/j.1752-1688.1999.tb05451.x](https://doi.org/10.1111/j.1752-1688.1999.tb05451.x). (Visited on 12/04/2024).
- [47] Matthew W. Salzer and Kurt F. Kipfmüller. “Reconstructed Temperature And Precipitation On A Millennial Timescale From Tree-Rings In The Southern Colorado Plateau, U.S.A.” In: *Climatic Change* 70.3 (June 2005). <https://doi.org/10.1007/s10584-005-5922-3>, pp. 465–487. ISSN: 1573-1480. DOI: [10.1007/s10584-005-5922-3](https://doi.org/10.1007/s10584-005-5922-3). (Visited on 12/04/2024).
- [48] Harold C. Fritts. “Growth-Rings of Trees: Their Correlation with Climate”. In: *Science* 154.3752 (Nov. 1966), pp. 973–979. DOI: [10.1126/science.154.3752.973](https://doi.org/10.1126/science.154.3752.973).

- [49] Marco Carrer. “Individualistic and Time-Varying Tree-Ring Growth to Climate Sensitivity”. In: *PLOS ONE* 6.7 (July 2011). <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0022813>, e22813. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0022813](https://doi.org/10.1371/journal.pone.0022813). (Visited on 01/29/2025).
- [50] Juan Diego Galván, Jesús Julio Camarero, and Emilia Gutiérrez. “Seeing the Trees for the Forest: Drivers of Individual Growth Responses to Climate in Inus Uncinata Mountain Forests”. In: *Journal of Ecology* 102.5 (2014). <https://onlinelibrary.wiley.com/doi/abs/10.1111/1365-2745.12268>, pp. 1244–1257. ISSN: 1365-2745. DOI: [10.1111/1365-2745.12268](https://doi.org/10.1111/1365-2745.12268). (Visited on 01/29/2025).
- [51] David Meko. “Applications of Box-Jenkins Methods of Time Series Analysis to the Reconstruction of Drought from Tree Rings”. PhD thesis. The University of Arizona., 1981.
- [52] David Meko et al. “Spatial Patterns of Tree-Growth Anomalies in the United States and Southeastern Canada”. In: *Journal of Climate* (Sept. 1993). https://journals.ametsoc.org/view/journals/clim/6/9/1520-0442_1993_006_1773_spotga_2_0_co_2.xml. ISSN: 1520-0442. (Visited on 01/29/2025).
- [53] Robert A. Monserud. “Time-Series Analyses of Tree-Ring Chronologies”. In: *Forest Science* 32.2 (June 1986). <https://doi.org/10.1093/forestscience/32.2.349>, pp. 349–372. ISSN: 0015-749X. DOI: [10.1093/forestscience/32.2.349](https://doi.org/10.1093/forestscience/32.2.349). (Visited on 01/31/2025).
- [54] Jenný Brynjarsdóttir and Anthony O’Hagan. “Learning about Physical Parameters: The Importance of Model Discrepancy”. In: *Inverse Problems* 30.11 (Nov. 2014). <https://iopscience.iop.org/article/10.1088/0266-5611/30/11/114007>, p. 114007. ISSN: 0266-5611, 1361-6420. DOI: [10.1088/0266-5611/30/11/114007](https://doi.org/10.1088/0266-5611/30/11/114007). (Visited on 02/14/2024).

- [55] Susan E. Tolwinski-Ward et al. “An Efficient Forward Model of the Climate Controls on Interannual Variation in Tree-Ring Width”. In: *Climate Dynamics* 36.11 (June 2011), pp. 2419–2439. ISSN: 1432-0894. DOI: [10.1007/s00382-010-0945-5](https://doi.org/10.1007/s00382-010-0945-5).
- [56] Jin Huang, Huug M. van den Dool, and Konstantine P. Georgarakos. “Analysis of Model-Calculated Soil Moisture over the United States (1931–1993) and Applications to Long-Range Temperature Forecasts”. In: *Journal of Climate* (June 1996). https://journals.ametsoc.org/view/journals/clim/9/6/1520-0442_1996_009_1350_aomcsm_2_0_co_2.xml. ISSN: 1520-0442. (Visited on 12/04/2024).
- [57] S. E. Tolwinski-Ward et al. “Probabilistic Reconstructions of Local Temperature and Soil Moisture from Tree-Ring Data with Potentially Time-Varying Climatic Response”. In: *Climate Dynamics* 44.3 (Feb. 2015). <https://doi.org/10.1007/s00382-014-2139-z>, pp. 791–806. ISSN: 1432-0894. DOI: [10.1007/s00382-014-2139-z](https://doi.org/10.1007/s00382-014-2139-z). (Visited on 10/21/2021).
- [58] John Tipton et al. “Reconstruction of Late Holocene Climate Based on Tree Growth and Mechanistic Hierarchical Models”. In: *Environmetrics* 27.1 (2016), pp. 42–54. ISSN: 1099-095X. DOI: [10.1002/env.2368](https://doi.org/10.1002/env.2368).
- [59] K.R. Briffa and P.D. Jones. “Global Surface Air Temperature Variations during the Twentieth Century: Part 2 , Implications for Large-Scale High-Frequency Palaeoclimatic Studies”. In: *The Holocene* 3.1 (Mar. 1993). <https://doi.org/10.1177/095968369300300109>, pp. 77–88. ISSN: 0959-6836. DOI: [10.1177/095968369300300109](https://doi.org/10.1177/095968369300300109). (Visited on 05/17/2022).
- [60] Rob Wilson et al. “Last Millennium Northern Hemisphere Summer Temperatures from Tree Rings: Part I: The Long Term Context”. In: *Quaternary Science Reviews* 134 (Feb. 2016). <https://www.sciencedirect.com/science/article/pii/S0277379115301888>, pp. 1–18. ISSN: 0277-3791. DOI: [10.1016/j.quascirev.2015.12.005](https://doi.org/10.1016/j.quascirev.2015.12.005). (Visited on 01/29/2025).

- [61] Kevin J. Anchukaitis et al. “Last Millennium Northern Hemisphere Summer Temperatures from Tree Rings: Part II, Spatially Resolved Reconstructions”. In: *Quaternary Science Reviews* 163 (May 2017). <https://www.sciencedirect.com/science/article/pii/S0277379117301592>, pp. 1–22. ISSN: 0277-3791. DOI: [10.1016/j.quascirev.2017.02.020](https://doi.org/10.1016/j.quascirev.2017.02.020). (Visited on 01/29/2025).
- [62] Jonathan M. King et al. “A Data Assimilation Approach to Last Millennium Temperature Field Reconstruction Using a Limited High-Sensitivity Proxy Network”. In: *Journal of Climate* (Sept. 2021). <https://journals.ametsoc.org/view/journals/clim/34/17/JCLI-D-20-0661.1.xml>. DOI: [10.1175/JCLI-D-20-0661.1](https://doi.org/10.1175/JCLI-D-20-0661.1). (Visited on 01/29/2025).
- [63] Markus Stoffel et al. “Estimates of Volcanic-Induced Cooling in the Northern Hemisphere over the Past 1,500 Years”. In: *Nature Geoscience* 8.10 (Oct. 2015). <https://www.nature.com/articles/ngeo2526>, pp. 784–788. ISSN: 1752-0908. DOI: [10.1038/ngeo2526](https://doi.org/10.1038/ngeo2526). (Visited on 01/29/2025).
- [64] S. E. Tolwinski-Ward, K. J. Anchukaitis, and M. N. Evans. “Bayesian Parameter Estimation and Interpretation for an Intermediate Model of Tree-Ring Width”. In: *Climate of the Past* 9.4 (July 2013), pp. 1481–1493. ISSN: 1814-9324. DOI: [10.5194/cp-9-1481-2013](https://doi.org/10.5194/cp-9-1481-2013).
- [65] J. Emile-Geay and M. Tingley. “Inferring Climate Variability from Nonlinear Proxies: Application to Palaeo-ENSO Studies”. In: *Climate of the Past* 12.1 (Jan. 2016). <https://cp.copernicus.org/articles/12/31/2016/cp-12-31-2016.html>, pp. 31–50. ISSN: 1814-9324. DOI: [10.5194/cp-12-31-2016](https://doi.org/10.5194/cp-12-31-2016). (Visited on 03/12/2025).
- [66] Yan Sun et al. “Statistical Treatment for the Wet Bias in Tree-Ring Chronologies: A Case Study from the Interior West, USA”. In: *Environmental and Ecological Statistics* 24.1 (Mar. 2017). <https://doi.org/10.1007/s10651-016-0363-x>, pp. 131–150. ISSN: 1573-3009. DOI: [10.1007/s10651-016-0363-x](https://doi.org/10.1007/s10651-016-0363-x). (Visited on 11/15/2024).

- [67] Gregory J. Hakim et al. “The Last Millennium Climate Reanalysis Project: Framework and First Results”. In: *Journal of Geophysical Research: Atmospheres* 121.12 (2016). <https://onlinelibrary.wiley.com/doi/abs/10.1002/2016JD024751>, pp. 6745–6764. ISSN: 2169-8996. DOI: [10.1002/2016JD024751](https://doi.org/10.1002/2016JD024751). (Visited on 11/18/2021).
- [68] Edmund M. Hart and Kendon Bell. *prism: Download data from the Oregon prism project*. R package version 0.0.6. 2015. DOI: [10.5281/zenodo.33663](https://doi.org/10.5281/zenodo.33663). URL: <https://github.com/ropensci/prism>.
- [69] Jessica E. Tierney et al. “Advances in Paleoclimate Data Assimilation”. In: *Annual Review of Earth and Planetary Sciences* (2025). <https://www.annualreviews.org/content/journals/10.1146/annurev-earth-032320-064209>. ISSN: 0084-6597. DOI: [10.1146/annurev-earth-032320-064209](https://doi.org/10.1146/annurev-earth-032320-064209).
- [70] Emily J. Judd et al. “A 485-Million-Year History of Earth’s Surface Temperature”. In: *Science* 385.6715 (Sept. 2024). <https://www.science.org/doi/10.1126/science.adk3705>, eadk3705. DOI: [10.1126/science.adk3705](https://doi.org/10.1126/science.adk3705). (Visited on 09/22/2024).
- [71] Martin P. Tingley and Peter Huybers. “A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part I: Development and Applications to Paleoclimate Reconstruction Problems”. In: *Journal of Climate* 23.10 (May 2010). <https://journals.ametsoc.org/view/journals/clim/23/10/2009jcli3015.1.xml>, pp. 2759–2781. ISSN: 0894-8755, 1520-0442. DOI: [10.1175/2009JCLI3015.1](https://doi.org/10.1175/2009JCLI3015.1). (Visited on 10/18/2021).
- [72] Martin P. Tingley and Peter Huybers. “Recent Temperature Extremes at High Northern Latitudes Unprecedented in the Past 600 Years”. In: *Nature* 496.7444 (Apr. 2013). <https://www.nature.com/articles/nature11969>, pp. 201–205. ISSN: 1476-4687. DOI: [10.1038/nature11969](https://doi.org/10.1038/nature11969). (Visited on 10/15/2024).
- [73] Matthew B. Osman et al. “Globally Resolved Surface Temperatures since the Last Glacial Maximum”. In: *Nature* 599.7884 (Nov. 2021). <https://www.nature.com/>

- [articles/s41586-021-03984-4](#), pp. 239–244. ISSN: 1476-4687. DOI: [10.1038/s41586-021-03984-4](#). (Visited on 04/26/2024).
- [74] J. Hansen et al. “Global Surface Temperature Change”. In: *Reviews of Geophysics* 48.4 (2010). <https://onlinelibrary.wiley.com/doi/abs/10.1029/2010RG000345>. ISSN: 1944-9208. DOI: [10.1029/2010RG000345](#). (Visited on 05/12/2025).
- [75] Nathan Lenssen et al. “A NASA GISTEMPv4 Observational Uncertainty Ensemble”. In: *Journal of Geophysical Research: Atmospheres* 129.17 (2024). <https://onlinelibrary.wiley.com/doi/abs/10.1029/2023JD040179>, e2023JD040179. ISSN: 2169-8996. DOI: [10.1029/2023JD040179](#). (Visited on 05/12/2025).
- [76] Nathan J. Steiger et al. “Assimilation of Time-Averaged Pseudoproxies for Climate Reconstruction”. In: *Journal of Climate* 27.1 (Jan. 2014). <https://journals.ametsoc.org/view/journals/clim/27/1/jcli-d-12-00693.1.xml>, pp. 426–441. ISSN: 0894-8755, 1520-0442. DOI: [10.1175/JCLI-D-12-00693.1](#). (Visited on 03/04/2022).
- [77] Geir Evensen. “Sequential Data Assimilation with a Nonlinear Quasi-Geostrophic Model Using Monte Carlo Methods to Forecast Error Statistics”. In: *Journal of Geophysical Research: Oceans* 99.C5 (1994). <https://onlinelibrary.wiley.com/doi/abs/10.1029/94JC00572>, pp. 10143–10162. ISSN: 2156-2202. DOI: [10.1029/94JC00572](#). (Visited on 04/11/2025).
- [78] P. L. Houtekamer and Herschel L. Mitchell. “Data Assimilation Using an Ensemble Kalman Filter Technique”. In: *Monthly Weather Review* (Mar. 1998). https://journals.ametsoc.org/view/journals/mwre/126/3/1520-0493_1998_126_0796_dauaek_2.0.co_2.xml. ISSN: 1520-0493. (Visited on 04/11/2025).
- [79] Jeffrey S. Whitaker and Thomas M. Hamill. “Ensemble Data Assimilation without Perturbed Observations”. In: *Monthly Weather Review* 130.7 (July 2002). https://journals.ametsoc.org/view/journals/mwre/130/7/1520-0493_2002_130_1913_edawpo_2.0.co_2.xml. ISSN: 1520-0493. (Visited on 04/11/2025).

- [80] Evan Kwiatkowski and Jan Mandel. “Convergence of the Square Root Ensemble Kalman Filter in the Large Ensemble Limit”. In: *SIAM/ASA Journal on Uncertainty Quantification* 3.1 (Jan. 2015). <https://epubs.siam.org/doi/abs/10.1137/140965363>, pp. 1–17. DOI: [10.1137/140965363](https://doi.org/10.1137/140965363). (Visited on 04/11/2025).
- [81] Christopher K. Wikle and L. Mark Berliner. “A Bayesian Tutorial for Data Assimilation”. In: *Physica D: Nonlinear Phenomena*. Data Assimilation 230.1 (June 2007). <https://www.sciencedirect.com/science/article/pii/S016727890600354X>, pp. 1–16. ISSN: 0167-2789. DOI: [10.1016/j.physd.2006.09.017](https://doi.org/10.1016/j.physd.2006.09.017). (Visited on 04/11/2025).
- [82] A. Apte, C. K. R. T. Jones, and A. M. Stuart. “A Bayesian Approach to Lagrangian Data Assimilation”. In: *Tellus A* 60.2 (2008). <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0870.2007.00295.x>, pp. 336–347. ISSN: 1600-0870. DOI: [10.1111/j.1600-0870.2007.00295.x](https://doi.org/10.1111/j.1600-0870.2007.00295.x). (Visited on 04/11/2025).
- [83] Oliver G. Ernst, Björn Sprungk, and Hans-Jörg Starkloff. “Bayesian Inverse Problems and Kalman Filters”. In: *Extraction of Quantifiable Information from Complex Systems*. Ed. by Stephan Dahlke et al. https://doi.org/10.1007/978-3-319-08159-5_7. Cham: Springer International Publishing, 2014, pp. 133–159. ISBN: 978-3-319-08159-5. DOI: [10.1007/978-3-319-08159-5_7](https://doi.org/10.1007/978-3-319-08159-5_7). (Visited on 04/11/2025).
- [84] M. N. Evans et al. “Applications of Proxy System Modeling in High Resolution Paleoclimatology”. In: *Quaternary Science Reviews* 76 (Sept. 2013). <https://www.sciencedirect.com/science/article/pii/S0277379113002011>, pp. 16–28. ISSN: 0277-3791. DOI: [10.1016/j.quascirev.2013.05.024](https://doi.org/10.1016/j.quascirev.2013.05.024). (Visited on 11/18/2021).
- [85] Jessica E. Tierney et al. “Glacial Cooling and Climate Sensitivity Revisited”. In: *Nature* 584.7822 (Aug. 2020). <https://www.nature.com/articles/s41586-020-2617-x>, pp. 569–573. ISSN: 1476-4687. DOI: [10.1038/s41586-020-2617-x](https://doi.org/10.1038/s41586-020-2617-x). (Visited on 10/12/2022).

- [86] Jonathan M. King et al. “A Data Assimilation Approach to Last Millennium Temperature Field Reconstruction Using a Limited High-Sensitivity Proxy Network”. In: *Journal of Climate* (Sept. 2021). <https://journals.ametsoc.org/view/journals/clim/34/17/JCLI-D-20-0661.1.xml>. DOI: [10.1175/JCLI-D-20-0661.1](https://doi.org/10.1175/JCLI-D-20-0661.1). (Visited on 01/29/2025).
- [87] Sylvia G. Dee et al. “On the Utility of Proxy System Models for Estimating Climate States over the Common Era”. In: *Journal of Advances in Modeling Earth Systems* 8.3 (2016). <https://onlinelibrary.wiley.com/doi/abs/10.1002/2016MS000677>, pp. 1164–1179. ISSN: 1942-2466. DOI: [10.1002/2016MS000677](https://doi.org/10.1002/2016MS000677). (Visited on 03/24/2022).
- [88] Jessica E. Tierney and Martin P. Tingley. “A Bayesian, Spatially-Varying Calibration Model for the TEX86 Proxy”. In: *Geochimica et Cosmochimica Acta* 127 (Feb. 2014), pp. 83–106. ISSN: 0016-7037. DOI: [10.1016/j.gca.2013.11.026](https://doi.org/10.1016/j.gca.2013.11.026).
- [89] S. Dee et al. “PRYSM: An Open-Source Framework for PROXY System Modeling, with Applications to Oxygen-Isotope Systems”. In: *Journal of Advances in Modeling Earth Systems* 7.3 (2015). <https://onlinelibrary.wiley.com/doi/abs/10.1002/2015MS000447>, pp. 1220–1247. ISSN: 1942-2466. DOI: [10.1002/2015MS000447](https://doi.org/10.1002/2015MS000447). (Visited on 03/11/2022).
- [90] Veronika Valler, Jörg Franke, and Stefan Brönnimann. “Impact of Different Estimations of the Background-Error Covariance Matrix on Climate Reconstructions Based on Data Assimilation”. In: *Climate of the Past* 15.4 (Aug. 2019). <https://cp.copernicus.org/articles/15/1427/2019/>, pp. 1427–1441. ISSN: 1814-9324. DOI: [10.5194/cp-15-1427-2019](https://doi.org/10.5194/cp-15-1427-2019). (Visited on 04/30/2025).
- [91] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. “Generating Random Correlation Matrices Based on Vines and Extended Onion Method”. In: *Journal of Multivariate Analysis* 100.9 (Oct. 2009). <https://www.sciencedirect.com/science/article/>

- [pii/S0047259X09000876](#), pp. 1989–2001. ISSN: 0047-259X. DOI: [10.1016/j.jmva.2009.04.008](#). (Visited on 05/15/2025).
- [92] John Barnard, Robert McCulloch, and Xiao-Li Meng. “Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, with Application to Shrinkage”. In: *Statistica Sinica* 10.4 (2000). <https://www.jstor.org/stable/24306780>, pp. 1281–1311. ISSN: 1017-0405. JSTOR: [24306780](#). (Visited on 05/15/2025).
- [93] Tine Nilsen et al. “Assessing the Performance of the BARCAST Climate Field Reconstruction Technique for a Climate with Long-Range Memory”. In: *Climate of the Past* 14.6 (June 2018). <https://cp.copernicus.org/articles/14/947/2018/>, pp. 947–967. ISSN: 1814-9324. DOI: [10.5194/cp-14-947-2018](#). (Visited on 04/21/2025).
- [94] Tine Nilsen, Stefanie Talento, and Johannes P. Werner. “Constraining Two Climate Field Reconstruction Methodologies over the North Atlantic Realm Using Pseudo-Proxy Experiments”. In: *Quaternary Science Reviews* 265 (Aug. 2021). <https://www.sciencedirect.com/science/article/pii/S027737912100216X>, p. 107009. ISSN: 0277-3791. DOI: [10.1016/j.quascirev.2021.107009](#). (Visited on 11/19/2024).
- [95] J Luterbacher et al. “European Summer Temperatures since Roman Times”. In: *Environmental Research Letters* 11.2 (Jan. 2016). <https://dx.doi.org/10.1088/1748-9326/11/2/024001>, p. 024001. ISSN: 1748-9326. DOI: [10.1088/1748-9326/11/2/024001](#). (Visited on 04/22/2025).
- [96] Johannes P. Werner et al. “Spatio-Temporal Variability of Arctic Summer Temperatures over the Past 2 Millennia”. In: *Climate of the Past* 14.4 (Apr. 2018). <https://cp.copernicus.org/articles/14/527/2018/>, pp. 527–557. ISSN: 1814-9324. DOI: [10.5194/cp-14-527-2018](#). (Visited on 04/01/2025).
- [97] Julien Emile-Geay et al. “Temporal Comparisons Involving Paleoclimate Data Assimilation: Challenges and Remedies”. In: *Journal of Climate* (Feb. 2025). [https:](#)

- [//journals.ametsoc.org/view/journals/clim/38/5/JCLI-D-24-0101.1.xml](https://journals.ametsoc.org/view/journals/clim/38/5/JCLI-D-24-0101.1.xml). DOI: [10.1175/JCLI-D-24-0101.1](https://doi.org/10.1175/JCLI-D-24-0101.1). (Visited on 05/12/2025).
- [98] Kevin Cowtan et al. “Statistical Analysis of Coverage Error in Simple Global Temperature Estimators”. In: *Dynamics and Statistics of the Climate System* 3.1 (Jan. 2018). <https://doi.org/10.1093/climsys/dzy003>, dzy003. ISSN: 2059-6987. DOI: [10.1093/climsys/dzy003](https://doi.org/10.1093/climsys/dzy003). (Visited on 05/02/2025).
- [99] Matthias Morzfeld and Daniel Hodyss. “A Theory for Why Even Simple Covariance Localization Is So Useful in Ensemble Data Assimilation”. In: *Monthly Weather Review* (Feb. 2023). <https://journals.ametsoc.org/view/journals/mwre/151/3/MWR-D-22-0255.1.xml>. DOI: [10.1175/MWR-D-22-0255.1](https://doi.org/10.1175/MWR-D-22-0255.1). (Visited on 05/19/2025).
- [100] Huaran Liu, Zhengyu Liu, and Feiyu Lu. “A Systematic Comparison of Particle Filter and EnKF in Assimilating Time-Averaged Observations”. In: *Journal of Geophysical Research: Atmospheres* 122.24 (2017). <https://onlinelibrary.wiley.com/doi/abs/10.1002/2017JD026798>, pp. 13, 155–13, 173. ISSN: 2169-8996. DOI: [10.1002/2017JD026798](https://doi.org/10.1002/2017JD026798). (Visited on 04/23/2025).
- [101] Robert J. Webber and Matthias Morzfeld. *Localized Covariance Estimation: A Bayesian Perspective*. <http://arxiv.org/abs/2301.04828>. Jan. 2023. DOI: [10.48550/arXiv.2301.04828](https://doi.org/10.48550/arXiv.2301.04828). arXiv: [2301.04828](https://arxiv.org/abs/2301.04828) [stat]. (Visited on 05/19/2025).
- [102] Jörg Franke et al. “The Importance of Input Data Quality and Quantity in Climate Field Reconstructions – Results from the Assimilation of Various Tree-Ring Collections”. In: *Climate of the Past* 16.3 (June 2020). <https://cp.copernicus.org/articles/16/1061/2020/>, pp. 1061–1074. ISSN: 1814-9324. DOI: [10.5194/cp-16-1061-2020](https://doi.org/10.5194/cp-16-1061-2020). (Visited on 04/30/2025).
- [103] Jonathan R. Bradley, Noel Cressie, and Tao Shi. “A Comparison of Spatial Predictors When Datasets Could Be Very Large”. In: *Statistics Surveys* 10.none (Jan. 2016). <https://projecteuclid.org/journals/statistics-surveys/volume-10/issue->

- none/A-comparison-of-spatial-predictors-when-datasets-could-be-very/
10.1214/16-SS115.full, pp. 100–131. ISSN: 1935-7516. DOI: 10.1214/16-SS115.
(Visited on 05/15/2025).
- [104] Matthew J. Heaton et al. “A Case Study Competition Among Methods for Analyzing Large Spatial Data”. In: *Journal of Agricultural, Biological and Environmental Statistics* 24.3 (Sept. 2019). <https://doi.org/10.1007/s13253-018-00348-w>, pp. 398–425. ISSN: 1537-2693. DOI: 10.1007/s13253-018-00348-w. (Visited on 05/15/2025).
- [105] Feng Zhu et al. “A Pseudoproxy Emulation of the PAGES 2k Database Using a Hierarchy of Proxy System Models”. In: *Scientific Data* 10.1 (Sept. 2023). <https://www.nature.com/articles/s41597-023-02489-1>, p. 624. ISSN: 2052-4463. DOI: 10.1038/s41597-023-02489-1. (Visited on 11/26/2024).
- [106] Melonie Heron. “Deaths: Leading Causes for 2017”. In: *Natl Vital Stat Rep* 68.6 (June 2019), pp. 1–77. ISSN: 1551-8930. pmid: 32501203.
- [107] C. Marras et al. “Prevalence of Parkinson’s Disease across North America”. In: *npj Parkinson’s Disease* 4.1 (July 2018), pp. 1–7. ISSN: 2373-8057. DOI: 10.1038/s41531-018-0058-0.
- [108] L. E. Hebert et al. “Alzheimer Disease in the United States (2010-2050) Estimated Using the 2010 Census”. In: *Neurology* 80.19 (May 2013), pp. 1778–1783. ISSN: 0028-3878, 1526-632X. DOI: 10.1212/WNL.0b013e31828726f5.
- [109] Jesper F. Havelund et al. “Changes in Kynurenine Pathway Metabolism in Parkinson Patients with L-DOPA-Induced Dyskinesia”. In: *Journal of Neurochemistry* 142.5 (2017), pp. 756–766. ISSN: 1471-4159. DOI: 10.1111/jnc.14104.
- [110] Eugenia Trushina and Michelle M. Mielke. “Recent Advances in the Application of Metabolomics to Alzheimer’s Disease”. In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. Misfolded Proteins, Mitochondrial Dysfunction, and

- Neurodegenerative Diseases 1842.8 (Aug. 2014), pp. 1232–1239. ISSN: 0925-4439. DOI: [10.1016/j.bbadis.2013.06.014](https://doi.org/10.1016/j.bbadis.2013.06.014).
- [111] David S. Wishart et al. “The Human Cerebrospinal Fluid Metabolome”. In: *Journal of Chromatography B. Hyphenated Techniques for Global Metabolite Profiling* 871.2 (Aug. 2008), pp. 164–173. ISSN: 1570-0232. DOI: [10.1016/j.jchromb.2008.05.001](https://doi.org/10.1016/j.jchromb.2008.05.001).
- [112] Martin J Romeo et al. “CSF Proteome: A Protein Repository for Potential Biomarker Identification”. In: *Expert Review of Proteomics* 2.1 (Jan. 2005), pp. 57–70. ISSN: 1478-9450, 1744-8387. DOI: [10.1586/14789450.2.1.57](https://doi.org/10.1586/14789450.2.1.57).
- [113] Susan Cheng et al. “Distinct Metabolomic Signatures Are Associated with Longevity in Humans”. In: *Nature Communications* 6.1 (Apr. 2015), p. 6791. ISSN: 2041-1723. DOI: [10.1038/ncomms7791](https://doi.org/10.1038/ncomms7791).
- [114] Richard D. Beger et al. “Metabonomic Models of Human Pancreatic Cancer Using 1D Proton NMR Spectra of Lipids in Plasma”. In: *Metabolomics* 2.3 (Sept. 2006), pp. 125–134. ISSN: 1573-3890. DOI: [10.1007/s11306-006-0026-2](https://doi.org/10.1007/s11306-006-0026-2).
- [115] Takashi Kobayashi et al. “A Novel Serum Metabolomics-Based Diagnostic Approach to Pancreatic Cancer”. In: *Cancer Epidemiol Biomarkers Prev* 22.4 (Apr. 2013), pp. 571–579. ISSN: 1055-9965, 1538-7755. DOI: [10.1158/1055-9965.EPI-12-1033](https://doi.org/10.1158/1055-9965.EPI-12-1033). pmid: [23542803](https://pubmed.ncbi.nlm.nih.gov/23542803/).
- [116] Kacey J. Prentice et al. “The Furan Fatty Acid Metabolite CMPF Is Elevated in Diabetes and Induces β Cell Dysfunction”. In: *Cell Metabolism* 19.4 (Apr. 2014), pp. 653–666. ISSN: 1550-4131. DOI: [10.1016/j.cmet.2014.03.008](https://doi.org/10.1016/j.cmet.2014.03.008). pmid: [24703697](https://pubmed.ncbi.nlm.nih.gov/24703697/).
- [117] Rima Kaddurah-Daouk et al. “Metabolomic Changes in Autopsy-Confirmed Alzheimer’s Disease”. In: *Alzheimer’s & Dementia* 7.3 (May 2011), pp. 309–317. ISSN: 1552-5260. DOI: [10.1016/j.jalz.2010.06.001](https://doi.org/10.1016/j.jalz.2010.06.001).

- [118] Mohammad Nasir Abdullah et al. “Metabolites Selection and Classification of Metabolomics Data on Alzheimer’s Disease Using Random Forest”. In: *Soft Computing in Data Science*. Ed. by Michael W. Berry, Azlinah Hj. Mohamed, and Bee Wah Yap. Vol. 652. Communications in Computer and Information Science. Singapore: Springer Singapore, 2016, pp. 100–112. ISBN: 978-981-10-2776-5 978-981-10-2777-2. DOI: [10.1007/978-981-10-2777-2_9](https://doi.org/10.1007/978-981-10-2777-2_9).
- [119] Peter A. LeWitt et al. “3-Hydroxykynurenine and Other Parkinson’s Disease Biomarkers Discovered by Metabolomic Analysis”. In: *Movement Disorders* 28.12 (2013), pp. 1653–1660. ISSN: 1531-8257. DOI: [10.1002/mds.25555](https://doi.org/10.1002/mds.25555).
- [120] Desiree Willkommen et al. “Metabolomic Investigations in Cerebrospinal Fluid of Parkinson’s Disease”. In: *PLOS ONE* 13.12 (Dec. 2018), e0208752. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0208752](https://doi.org/10.1371/journal.pone.0208752).
- [121] Stephan Klatt et al. “A Six-Metabolite Panel as Potential Blood-Based Biomarkers for Parkinson’s Disease”. In: *npj Parkinsons Dis.* 7.1 (Dec. 2021), p. 94. ISSN: 2373-8057. DOI: [10.1038/s41531-021-00239-x](https://doi.org/10.1038/s41531-021-00239-x).
- [122] Hui Zou and Trevor Hastie. “Regularization and Variable Selection via the Elastic Net”. In: *J Royal Statistical Soc B* 67.2 (Apr. 2005), pp. 301–320. ISSN: 1369-7412, 1467-9868. DOI: [10.1111/j.1467-9868.2005.00503.x](https://doi.org/10.1111/j.1467-9868.2005.00503.x).
- [123] Nathan Hwangbo et al. “A Metabolomic Aging Clock Using Human Cerebrospinal Fluid”. In: *The Journals of Gerontology: Series A* (Aug. 2021), glab212. ISSN: 1079-5006. DOI: [10.1093/gerona/glab212](https://doi.org/10.1093/gerona/glab212).
- [124] Claire L. Tomlinson et al. “Systematic Review of Levodopa Dose Equivalency Reporting in Parkinson’s Disease”. In: *Movement Disorders* 25.15 (2010), pp. 2649–2653. ISSN: 1531-8257. DOI: [10.1002/mds.23429](https://doi.org/10.1002/mds.23429).
- [125] Brenna A. Cholerton et al. “Pacific Northwest Udall Center of Excellence Clinical Consortium: Study Design and Baseline Cohort Characteristics”. In: *J Parkinsons*

- Dis* 3.2 (2013), pp. 205–214. ISSN: 1877-718X. DOI: [10.3233/JPD-130189](https://doi.org/10.3233/JPD-130189). pmid: [23938350](https://pubmed.ncbi.nlm.nih.gov/23938350/).
- [126] Hojoong M. Kim et al. “Prediction of Cognitive Progression in Parkinson’s Disease Using Three Cognitive Screening Measures”. In: *Clinical Parkinsonism & Related Disorders* 1 (Jan. 2019), pp. 91–97. ISSN: 2590-1125. DOI: [10.1016/j.prdoa.2019.08.006](https://doi.org/10.1016/j.prdoa.2019.08.006).
- [127] Min Shi et al. “Cerebrospinal Fluid Biomarkers for Parkinson Disease Diagnosis and Progression”. In: *Ann Neurol* 69.3 (Mar. 2011), pp. 570–580. ISSN: 0364-5134. DOI: [10.1002/ana.22311](https://doi.org/10.1002/ana.22311). pmid: [21400565](https://pubmed.ncbi.nlm.nih.gov/21400565/).
- [128] Ignacio F. Mata et al. “GBA Variants Are Associated with a Distinct Pattern of Cognitive Deficits in Parkinson’s Disease”. In: *Movement Disorders* 31.1 (2016), pp. 95–102. ISSN: 1531-8257. DOI: [10.1002/mds.26359](https://doi.org/10.1002/mds.26359).
- [129] Xinyu Zhang, Jiyang Dong, and Daniel Raftery. “Five Easy Metrics of Data Quality for LC–MS-Based Global Metabolomics”. In: *Anal. Chem.* 92.19 (Oct. 2020), pp. 12925–12933. ISSN: 0003-2700. DOI: [10.1021/acs.analchem.0c01493](https://doi.org/10.1021/acs.analchem.0c01493).
- [130] Cheng Zheng et al. “Evaluation of Potential Metabolomic-Based Biomarkers of Protein, Carbohydrate and Fat Intakes Using a Controlled Feeding Study”. In: *European Journal of Nutrition* 60.8 (Dec. 2021), pp. 4207–4218. ISSN: 1436-6215. DOI: [10.1007/s00394-021-02577-1](https://doi.org/10.1007/s00394-021-02577-1).
- [131] Angela J. Hanson et al. “Cerebrospinal Fluid Lipidomics: Effects of an Intravenous Triglyceride Infusion and apoE Status”. In: *Metabolomics* 16.1 (Dec. 2019), p. 6. ISSN: 1573-3890. DOI: [10.1007/s11306-019-1627-x](https://doi.org/10.1007/s11306-019-1627-x).
- [132] Mohan Ghorasaini et al. “Cross-Laboratory Standardization of Preclinical Lipidomics Using Differential Mobility Spectrometry and Multiple Reaction Monitoring”. In: *Analytical Chemistry* 93.49 (Dec. 2021), pp. 16369–16378. ISSN: 0003-2700. DOI: [10.1021/acs.analchem.1c02826](https://doi.org/10.1021/acs.analchem.1c02826).

- [133] Gerhard Liebisch et al. “Update on LIPID MAPS Classification, Nomenclature, and Shorthand Notation for MS-derived Lipid Structures”. In: *Journal of Lipid Research* 61.12 (Dec. 2020), pp. 1539–1555. ISSN: 0022-2275, 1539-7262. DOI: [10.1194/jlr.S120001025](https://doi.org/10.1194/jlr.S120001025).
- [134] Carl Morris. “A Finite Selection Model for Experimental Design of the Health Insurance Study”. In: *Journal of Econometrics* 11.1 (Sept. 1979), pp. 43–61. ISSN: 03044076. DOI: [10.1016/0304-4076\(79\)90053-8](https://doi.org/10.1016/0304-4076(79)90053-8).
- [135] Tianqi Chen and Carlos Guestrin. “Xgboost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794.
- [136] Stephen K. Van Den Eeden et al. “Incidence of Parkinson’s Disease: Variation by Age, Gender, and Race/Ethnicity”. In: *Am J Epidemiol* 157.11 (June 2003), pp. 1015–1022. ISSN: 0002-9262. DOI: [10.1093/aje/kwg068](https://doi.org/10.1093/aje/kwg068).
- [137] Glenda E. Gillies et al. “Sex Differences in Parkinson’s Disease”. In: *Frontiers in Neuroendocrinology*. Sex Differences in Neurological and Psychiatric Disorders 35.3 (Aug. 2014), pp. 370–384. ISSN: 0091-3022. DOI: [10.1016/j.yfrne.2014.02.002](https://doi.org/10.1016/j.yfrne.2014.02.002).
- [138] Sujuan Gao et al. “The Relationships Between Age, Sex, and the Incidence of Dementia and Alzheimer Disease: A Meta-Analysis”. In: *Arch Gen Psychiatry* 55.9 (Sept. 1998), p. 809. ISSN: 0003-990X. DOI: [10.1001/archpsyc.55.9.809](https://doi.org/10.1001/archpsyc.55.9.809).
- [139] James Honaker, Gary King, and Matthew Blackwell. “**Amelia** II: A Program for Missing Data”. In: *J. Stat. Soft.* 45.7 (2011). ISSN: 1548-7660. DOI: [10.18637/jss.v045.i07](https://doi.org/10.18637/jss.v045.i07).
- [140] Maytal Saar-Tsechansky and Foster Provost. “Handling Missing Values When Applying Classification Models”. In: *Journal of Machine Learning Research* 8.Jul (2007), pp. 1623–1657. ISSN: ISSN 1533-7928.

- [141] Shuzhao Li et al. “Predicting Network Activity from High Throughput Metabolomics”. In: *PLoS computational biology* 9.7 (2013).
- [142] J. Xia and D. S. Wishart. “MSEA: A Web-Based Tool to Identify Biologically Meaningful Patterns in Quantitative Metabolomic Data”. In: *Nucleic Acids Research* 38.Web Server (July 2010), W71–W77. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gkq329](https://doi.org/10.1093/nar/gkq329).
- [143] Zhiqiang Pang et al. “MetaboAnalystR 3.0: Toward an Optimized Workflow for Global Metabolomics”. In: *Metabolites* 10.5 (May 2020), p. 186. DOI: [10.3390/metabo10050186](https://doi.org/10.3390/metabo10050186).
- [144] Ian Jolliffe. “Principal Component Analysis”. In: *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd, 2005. ISBN: 978-0-470-01319-9. DOI: [10.1002/0470013192.bsa501](https://doi.org/10.1002/0470013192.bsa501).
- [145] Ellen Sidransky and Grisel Lopez. “The Link between the GBA Gene and Parkinsonism”. In: *The Lancet Neurology* 11.11 (Nov. 2012), pp. 986–998. ISSN: 1474-4422. DOI: [10.1016/S1474-4422\(12\)70190-4](https://doi.org/10.1016/S1474-4422(12)70190-4).
- [146] Nathan Pankratz et al. “Meta-Analysis of Parkinson’s Disease: Identification of a Novel Locus, RIT2”. In: *Annals of Neurology* 71.3 (2012), pp. 370–384. ISSN: 1531-8249. DOI: [10.1002/ana.22687](https://doi.org/10.1002/ana.22687).
- [147] Kathrin Brockmann et al. “GBA-Associated Parkinson’s Disease: Reduced Survival and More Rapid Progression in a Prospective Longitudinal Study”. In: *Movement Disorders* 30.3 (2015), pp. 407–411. ISSN: 1531-8257. DOI: [10.1002/mds.26071](https://doi.org/10.1002/mds.26071).
- [148] Marie Y. Davis et al. “Association of GBA Mutations and the E326K Polymorphism With Motor and Cognitive Progression in Parkinson Disease”. In: *JAMA Neurology* 73.10 (Oct. 2016), pp. 1217–1224. ISSN: 2168-6149. DOI: [10.1001/jamaneurol.2016.2245](https://doi.org/10.1001/jamaneurol.2016.2245).

- [149] Albert A. Davis et al. “Variants in GBA, SNCA, and MAPT Influence Parkinson Disease Risk, Age at Onset, and Progression”. In: *Neurobiology of Aging* 37 (Jan. 2016), 209.e1–209.e7. ISSN: 0197-4580. DOI: [10.1016/j.neurobiolaging.2015.09.014](https://doi.org/10.1016/j.neurobiolaging.2015.09.014).
- [150] Alex Frolkis et al. “SMPDB: The Small Molecule Pathway Database”. In: *Nucleic acids research* 38.suppl_1 (2010), pp. D480–D487.
- [151] Ju Youn Lee, Hee Kyung Jin, and Jae-Sung Bae. “Sphingolipids in Neuroinflammation: A Potential Target for Diagnosis and Therapy”. In: *BMB Rep* 53.1 (Jan. 2020), pp. 28–34. ISSN: 1976-670X. pmid: [31818364](https://pubmed.ncbi.nlm.nih.gov/31818364/).
- [152] Sam Vanherle et al. “Extracellular Vesicle-Associated Lipids in Central Nervous System Disorders”. In: *Adv Drug Deliv Rev* 159 (2020), pp. 322–331. ISSN: 1872-8294. DOI: [10.1016/j.addr.2020.04.011](https://doi.org/10.1016/j.addr.2020.04.011). pmid: [32360577](https://pubmed.ncbi.nlm.nih.gov/32360577/).
- [153] Yusuf A. Hannun and Lina M. Obeid. “Sphingolipids and Their Metabolism in Physiology and Disease”. In: *Nat Rev Mol Cell Biol* 19.3 (Mar. 2018), pp. 175–191. ISSN: 1471-0080. DOI: [10.1038/nrm.2017.107](https://doi.org/10.1038/nrm.2017.107). pmid: [29165427](https://pubmed.ncbi.nlm.nih.gov/29165427/).
- [154] Roy G. Cutler et al. “Involvement of Oxidative Stress-Induced Abnormalities in Ceramide and Cholesterol Metabolism in Brain Aging and Alzheimer’s Disease”. In: *PNAS* 101.7 (Feb. 2004), pp. 2070–2075. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.0305799101](https://doi.org/10.1073/pnas.0305799101).
- [155] Xianlin Han et al. “Substantial Sulfatide Deficiency and Ceramide Elevation in Very Early Alzheimer’s Disease: Potential Role in Disease Pathogenesis”. In: *Journal of Neurochemistry* 82.4 (2002), pp. 809–818. ISSN: 1471-4159. DOI: [10.1046/j.1471-4159.2002.00997.x](https://doi.org/10.1046/j.1471-4159.2002.00997.x).
- [156] Guanghu Wang et al. “Long-Chain Ceramide Is Elevated in Presenilin 1 (PS1M146V) Mouse Brain and Induces Apoptosis in PS1 Astrocytes”. In: *Glia* 56.4 (2008), pp. 449–456. ISSN: 1098-1136. DOI: [10.1002/glia.20626](https://doi.org/10.1002/glia.20626).

- [157] Michelle M. Mielke et al. “Plasma Ceramides Are Altered in Mild Cognitive Impairment and Predict Cognitive Decline and Hippocampal Volume Loss”. In: *Alzheimer’s & Dementia* 6.5 (2010), pp. 378–385. ISSN: 1552-5279. DOI: [10.1016/j.jalz.2010.03.014](https://doi.org/10.1016/j.jalz.2010.03.014).
- [158] Elizabeth Calzada, Ouma Onguka, and Steven M. Claypool. “Chapter Two - Phosphatidylethanolamine Metabolism in Health and Disease”. In: *International Review of Cell and Molecular Biology*. Ed. by Kwang W. Jeon. Vol. 321. Academic Press, Jan. 2016, pp. 29–88. DOI: [10.1016/bs.ircmb.2015.10.001](https://doi.org/10.1016/bs.ircmb.2015.10.001).
- [159] Kinga Czubowicz et al. “The Role of Ceramide and Sphingosine-1-Phosphate in Alzheimer’s Disease and Other Neurodegenerative Disorders”. In: *Mol Neurobiol* 56.8 (Aug. 2019), pp. 5436–5455. ISSN: 1559-1182. DOI: [10.1007/s12035-018-1448-3](https://doi.org/10.1007/s12035-018-1448-3). pmid: [30612333](https://pubmed.ncbi.nlm.nih.gov/30612333/).
- [160] Zhizhong Guan et al. “Decrease and Structural Modifications of Phosphatidylethanolamine Plasmalogen in the Brain with Alzheimer Disease.” in: *Journal of Neuropathology and Experimental Neurology* 58.7 (July 1999), pp. 740–747. ISSN: 0022-3069. DOI: [10.1097/00005072-199907000-00008](https://doi.org/10.1097/00005072-199907000-00008).
- [161] Xianlin Han, David M. Holtzman, and Daniel W. McKeel. “Plasmalogen Deficiency in Early Alzheimer’s Disease Subjects and in Animal Models: Molecular Characterization Using Electrospray Ionization Mass Spectrometry”. In: *Journal of Neurochemistry* 77.4 (2001), pp. 1168–1180. ISSN: 1471-4159. DOI: [10.1046/j.1471-4159.2001.00332.x](https://doi.org/10.1046/j.1471-4159.2001.00332.x).
- [162] Anna Wuolikainen et al. “Multi-Platform Mass Spectrometry Analysis of the CSF and Plasma Metabolomes of Rigorously Matched Amyotrophic Lateral Sclerosis, Parkinson’s Disease and Control Subjects”. In: *Mol. BioSyst.* 12.4 (Mar. 2016), pp. 1287–1298. ISSN: 1742-2051. DOI: [10.1039/C5MB00711A](https://doi.org/10.1039/C5MB00711A).

- [163] Anders Öhman and Lars Forsgren. “NMR Metabonomics of Cerebrospinal Fluid Distinguishes between Parkinson’s Disease and Controls”. In: *Neuroscience Letters* 594 (May 2015), pp. 36–39. ISSN: 0304-3940. DOI: [10.1016/j.neulet.2015.03.051](https://doi.org/10.1016/j.neulet.2015.03.051).
- [164] Miles Trupp et al. “Metabolite and Peptide Levels in Plasma and CSF Differentiating Healthy Controls from Patients with Newly Diagnosed Parkinson’s Disease”. In: *Journal of Parkinson’s Disease* 4.3 (2014), pp. 549–560. ISSN: 1877718X, 18777171. DOI: [10.3233/JPD-140389](https://doi.org/10.3233/JPD-140389).
- [165] Hemi Luan et al. “Comprehensive Urinary Metabolomic Profiling and Identification of Potential Noninvasive Marker for Idiopathic Parkinson’s Disease”. In: *Scientific Reports* 5.1 (Sept. 2015), p. 13888. ISSN: 2045-2322. DOI: [10.1038/srep13888](https://doi.org/10.1038/srep13888).
- [166] Hemi Luan et al. “LC–MS-Based Urinary Metabolite Signatures in Idiopathic Parkinson’s Disease”. In: *J. Proteome Res.* 14.1 (Jan. 2015), pp. 467–478. ISSN: 1535-3893. DOI: [10.1021/pr500807t](https://doi.org/10.1021/pr500807t).
- [167] Peter A. LeWitt et al. “Metabolomic Biomarkers as Strong Correlates of Parkinson Disease Progression”. In: *Neurology* 88.9 (Feb. 2017), pp. 862–869. ISSN: 0028-3878, 1526-632X. DOI: [10.1212/WNL.0000000000003663](https://doi.org/10.1212/WNL.0000000000003663). pmid: [28179471](https://pubmed.ncbi.nlm.nih.gov/28179471/).
- [168] Sadhana Kumari et al. “Identification of Potential Urine Biomarkers in Idiopathic Parkinson’s Disease Using NMR”. In: *Clinica Chimica Acta* 510 (Nov. 2020), pp. 442–449. ISSN: 0009-8981. DOI: [10.1016/j.cca.2020.08.005](https://doi.org/10.1016/j.cca.2020.08.005).
- [169] Anna Picca et al. “Circulating Amino Acid Signature in Older People with Parkinson’s Disease: A Metabolic Complement to the EXosomes in PARKinson Disease (EXPAND) Study”. In: *Exp Gerontol* 128 (Dec. 2019), p. 110766. ISSN: 1873-6815. DOI: [10.1016/j.exger.2019.110766](https://doi.org/10.1016/j.exger.2019.110766). pmid: [31666195](https://pubmed.ncbi.nlm.nih.gov/31666195/).
- [170] J. A. Molina et al. “Decreased Cerebrospinal Fluid Levels of Neutral and Basic Amino Acids in Patients with Parkinson’s Disease”. In: *J Neurol Sci* 150.2 (Sept.

- 1997), pp. 123–127. ISSN: 0022-510X. DOI: [10.1016/S0022-510X\(97\)00069-5](https://doi.org/10.1016/S0022-510X(97)00069-5). pmid: [9268238](https://pubmed.ncbi.nlm.nih.gov/9268238/).
- [171] Y. Mochizuki et al. “Amino Acid Concentration in Dementia of the Alzheimer Type and Multi-Infarct Dementia”. In: *Ann Clin Lab Sci* 26.3 (May 1996), pp. 275–278. ISSN: 0091-7370, 1550-8080. pmid: [8726221](https://pubmed.ncbi.nlm.nih.gov/8726221/).
- [172] George H. Fisher et al. “Free D-Aspartate and D-Alanine in Normal and Alzheimer Brain”. In: *Brain Research Bulletin* 26.6 (June 1991), pp. 983–985. ISSN: 0361-9230. DOI: [10.1016/0361-9230\(91\)90266-M](https://doi.org/10.1016/0361-9230(91)90266-M).
- [173] Domenico Ciavardelli et al. “Medium-Chain Plasma Acylcarnitines, Ketone Levels, Cognition, and Gray Matter Volumes in Healthy Elderly, Mildly Cognitively Impaired, or Alzheimer’s Disease Subjects”. In: *Neurobiol Aging* 43 (July 2016), pp. 1–12. ISSN: 1558-1497. DOI: [10.1016/j.neurobiolaging.2016.03.005](https://doi.org/10.1016/j.neurobiolaging.2016.03.005). pmid: [27255810](https://pubmed.ncbi.nlm.nih.gov/27255810/).
- [174] Zhiguang Huo et al. “Brain and Blood Metabolome for Alzheimer’s Dementia: Findings from a Targeted Metabolomics Analysis”. In: *Neurobiol Aging* 86 (Feb. 2020), pp. 123–133. ISSN: 1558-1497. DOI: [10.1016/j.neurobiolaging.2019.10.014](https://doi.org/10.1016/j.neurobiolaging.2019.10.014). pmid: [31785839](https://pubmed.ncbi.nlm.nih.gov/31785839/).
- [175] Alessandra Castegna et al. “Proteomic Identification of Oxidatively Modified Proteins in Alzheimer’s Disease Brain. Part I: Creatine Kinase BB, Glutamine Synthase, and Ubiquitin Carboxy-Terminal Hydrolase L-1”. In: *Free Radical Biology and Medicine* 33.4 (Aug. 2002), pp. 562–571. ISSN: 0891-5849. DOI: [10.1016/S0891-5849\(02\)00914-0](https://doi.org/10.1016/S0891-5849(02)00914-0).
- [176] Michael Aksenov et al. “Oxidative Modification of Creatine Kinase BB in Alzheimer’s Disease Brain”. In: *Journal of Neurochemistry* 74.6 (2000), pp. 2520–2527. ISSN: 1471-4159. DOI: [10.1046/j.1471-4159.2000.0742520.x](https://doi.org/10.1046/j.1471-4159.2000.0742520.x).
- [177] Richard J. Johnson et al. “Cerebral Fructose Metabolism as a Potential Mechanism Driving Alzheimer’s Disease”. In: *Front. Aging Neurosci.* 12 (Sept. 2020), p. 560865. ISSN: 1663-4365. DOI: [10.3389/fnagi.2020.560865](https://doi.org/10.3389/fnagi.2020.560865).

- [178] A. Schousboe, Ursula Sonnewald, and Helle S. Waagepetersen. “Differential Roles of Alanine in GABAergic and Glutamatergic Neurons”. In: *Neurochem Int* 43.4-5 (2003 Sep-Oct), pp. 311–315. ISSN: 0197-0186. DOI: [10.1016/s0197-0186\(03\)00017-2](https://doi.org/10.1016/s0197-0186(03)00017-2). pmid: [12742074](https://pubmed.ncbi.nlm.nih.gov/12742074/).
- [179] Mark Haid et al. “Long-Term Stability of Human Plasma Metabolites during Storage at -80 °C”. In: *Journal of Proteome Research* 17.1 (Jan. 2018), pp. 203–211. ISSN: 1535-3893. DOI: [10.1021/acs.jproteome.7b00518](https://doi.org/10.1021/acs.jproteome.7b00518).
- [180] Victoria L. Stevens et al. “Pre-Analytical Factors That Affect Metabolite Stability in Human Urine, Plasma, and Serum: A Review”. In: *Metabolites* 9.8 (Aug. 2019), p. 156. ISSN: 2218-1989. DOI: [10.3390/metabo9080156](https://doi.org/10.3390/metabo9080156).
- [181] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Feb. 2019. ISBN: 978-1-108-49802-9.
- [182] Stefanie Lerche et al. “The Mutation Matters: CSF Profiles of GCase, Sphingolipids, α -Synuclein in PDGBA”. In: *Movement Disorders* 36.5 (2021), pp. 1216–1228. ISSN: 1531-8257. DOI: [10.1002/mds.28472](https://doi.org/10.1002/mds.28472).
- [183] Anthony O’Hagan. “Expert Knowledge Elicitation: Subjective but Scientific”. In: *The American Statistician* 73.sup1 (Mar. 2019). <https://doi.org/10.1080/00031305.2018.1518265>, pp. 69–81. ISSN: 0003-1305. DOI: [10.1080/00031305.2018.1518265](https://doi.org/10.1080/00031305.2018.1518265). (Visited on 12/06/2024).
- [184] Bob Carpenter et al. “Stan: A Probabilistic Programming Language”. In: *Journal of Statistical Software* 76 (Jan. 2017). <https://doi.org/10.18637/jss.v076.i01>, pp. 1–32. ISSN: 1548-7660. DOI: [10.18637/jss.v076.i01](https://doi.org/10.18637/jss.v076.i01). (Visited on 05/16/2025).
- [185] Alysha M. De Livera et al. “Statistical Methods for Handling Unwanted Variation in Metabolomics Data”. In: *Anal. Chem.* 87.7 (Apr. 2015), pp. 3606–3615. ISSN: 0003-2700, 1520-6882. DOI: [10.1021/ac502439y](https://doi.org/10.1021/ac502439y).