WIENER-CHAOS ANALYSIS ON BAYESIAN MODELS WITH APPLICATIONS IN AGRICULTURE AND CLIMATOLOGY

By

Han Wang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics — Doctor of Philosophy

2020

ABSTRACT

WIENER-CHAOS ANALYSIS ON BAYESIAN MODELS WITH APPLICATIONS IN AGRICULTURE AND CLIMATOLOGY

By

Han Wang

Understanding the challenges to increasing maize productivity in sub-Saharan Africa has important implications for policies to reduce national and global food insecurity. There is insufficient research on the key agronomic and environmental factors that influence maize yield in a smallholder-farm environment. We implement a Bayesian analysis with longitudinal household survey data covering 1,197 plots among 320 farms in central Malawi. The results reveal a high positive association between a leaf chlorophyll indicator and yield, with significance levels exceeding 95% Bayesian credibility at all sites, and the posterior mean of the regression coefficient ranging from 28% to 42% on a relative scale. A parasitic weed, *Striga asiatica*, is the variable that negatively associated with yield of high intensity. The impact of rainfall varies by site and season, either directly or indirectly. We conclude that the determinants preventing striga infestation and enhancing nitrogen fertility will lead to higher maize yield in Malawi. To improve plant nitrogen status, fertilizer is effective at higher-productivity sites, whereas soil carbon and organic inputs are important at marginal sites. Uniquely, the Bayesian approach allows differentiation of response by site for a modest-sample-size study. Considering the biophysical constraints, our findings highlight area-specific recommendations as well as management strategies for crop yield.

Quantifying the sensitivity of climate forcing factors such as greenhouse gas concentration and solar irradiation, is critical in comprehending the evolution of the Earth's climate. There exists a variety of statistical methods to reconstruct temperature in the past, but the same is not true for projecting future temperatures. We produce a multi-level stochastic model to systematically reconstruct and project the northern-hemisphere average temperature anomalies, for the past millennium (1000-1999) and the next century (2019-2100), by coordinating with climatic forcings and natural proxies from diverse data sources. Additive noises are applied to the model to capture the unaccounted variability. Model parameters are estimated using Bayesian-inference techniques, resulting in complete distributional information. Reconstructions with memory features (no, short, long) are evaluated through selected validation metrics, and the results constitute evidence in favor of using a moderate-memory length. For the purpose of temperature projections, we incorporate realistic climate forcing uncertainties to Year 2100. Similarly, we include an uncertainty component on top of using representative carbon pathway scenarios for global greenhouse gases. Our projections' posterior means show a great level of agreement with the 95% confidence interval provided by the Coupled Model Intercomparison Project, while featuring differences in most cases.

The models described above are both implemented via Gibbs sampler with 10,000 iterations. In order to avoid its potential computational heft, we combine the use of maximum likelihood estimators for regression elements with properties of Wiener chaos, to approximate the predictive samples with specific chaos distributions that do not require sampling via numerics. Some of the approximations' statistics, such as error variances are also explicitly provided. The precision are relatively high (nearly 0.1% and 0.5%) depending on dimension circumstances. This allows practitioners to estimate approximation accuracy and convergence rates in practice, with no resort to heavy computational demands.

ACKNOWLEDGMENTS

First and foremost, I would love to express my deep gratitude to Dr. Frederi Viens. Not only is he a knowledgeable advisor providing constant guidance throughout the program, but also he is a great mentor who teaches me how to become a better person.

Also, I am grateful for Dr. Sieglinde Snapp who helps me open the window of research world. It is my great pleasure to collaborate with her and get to learn how to do research meticulously. I'm thankful that Dr. R.V. Ramamoorthi and Dr. Shlomo Levental agree to be on my committee, offering crucial insights and expertise.

Moreover, a special thank you to my parents' consistent love and support. I would never be where I am today without them. My girlfriend Mengzi's companionship through the years is very much appreciated as well.

Last but not least, I'd like to thank all the staff members and my fellow PhD colleagues, with whom I have countless discussions on a variety of topics.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Bayesian Regression Modeling

Bayesian statistics is based on the concept of probability to predict the future, where probability is interpreted as a degree of belief in an event. The degree of belief can come from previous knowledge or personal beliefs about the event. Bayesian statistics was named after Thomas Bayes, who first described the outcome of a coin-toss experiment using Bayes's theorem on his paper published in 1763.

From the late 1700s to the early 1800s, Pierre-Simon Laplace established the Bayesian interpretation of probability, and many other Bayesian methods were developed by other scholars during that time. Over much of the 20th century, the widely-used statistical methods were based on the frequentist interpretation. Bayesian methods were not in favor of, even controversial to many statisticians, due to computational reliance and practical restrictions.

Since the 1950s, researchers began to apply Bayes's theorem to account for model uncertainties, by incorporating educated guesses about the likelihood of something happening, and then making predictions. However, it is difficult to implement these types of guesses. In the recent two decades, particularly with the emergence of high-performance computers and new repetitive algorithms like Markov chain Monte Carlo (MCMC), Bayesian methods have been more recognized as powerful tools in the scientific community after overcoming the implementation difficulties.

Bayesian linear regression is a modeling approach in which the statistical analysis is undertaken within the context of Bayesian inference. When the regression model has an error following a normal distribution, and a particular form of prior distribution is assumed, the full posterior probability distributions for every model parameter are explicitly available. Not only does it provide more information than point estimates like means and variances in classical frequentist statistics, but credible intervals are straightforward to interpret even by non-statisticians. P-values can be computed in a Bayesian way, with more power and flexibility in assessing the significance of explanatory variables [1], avoiding misinterpretations of p-values [2, 3]. Also, the Bayesian approach allows background knowledge from domain specialists to be incorporated into the analysis, as a type of participatory model building, improving the accuracy and credibility of the estimations [4].

Last but not least, Bayesian statistics has an advantage in statistical power for data-limited studies [5, 6]. There is an often quoted but rarely if ever formally cited rule of thumb, by which the number of parameters (or degrees of freedom) that one can estimate reliably (e.g., with credibility level higher than 90%) in a linear model, is a third of the total number of data points, compared to a tenth in ordinary frequentist linear regression [7].

1.2 Chaos Structure of Wiener Space

Polynomial chaos (also called Wiener chaos expansion) is a non-sampling-based method to determine evolution of uncertainty in a dynamic system in which the parameters have probabilistic uncertainties. It was first introduced by Norbert Wiener in 1938, where Hermite polynomials were used to model stochastic processes with Gaussian random variables [8].

When the second moment is finite, such an expansion converges in the L^2 sense for any arbitrary stochastic process, which is applicable in most physics systems.

In any real and separable Hilbert space $\mathcal{H}$, there exists an isonormal Gaussian process of a centered Gaussian family $(G(\varphi), \varphi \in \mathcal{H})$ of random variables on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, such that

$$\mathbb{E}[G(\varphi)G(\psi)] = \langle \varphi, \psi \rangle_{\mathcal{H}} \tag{1.1}$$

The Wiener chaos of order n is defined as the closure in $L^2(\Omega)$, a linear span of the random variables $H_n(G(\varphi))$, where H_n is the Hermite polynomial of degree n, and $||\varphi||_{\mathcal{H}} = 1$. For any $F \in L^2(\Omega)$, there is a unique sequence of functions $f_n \in H^{\odot n}$ (symmetric tensor product) such that

$$F = \sum_{n=0}^{\infty} I_n(f_n) \tag{1.2}$$

where I_n is the Wiener stochastic integral with respect to G, $I_0(f_0) := \mathbb{E}[F]$, and all of them are mutually orthogonal in $L^2(\Omega)$. This is the fundamental decomposition of $L^2(\Omega)$ as a direct sum of all Wiener-chaos terms [9].

Moreover, $L^2(\Omega)$ is closed under multiplication, for any p, q and $f \in \mathcal{H}^{\odot p}$, $g \in \mathcal{H}^{\odot q}$ (symmetric), the product of Wiener integrals is calculated by:

$$I_p(f)I_q(g) = \sum_{n=1}^{p \wedge q} r! C_p^r C_q^r I_{p+q-2r}(f \otimes_r g) \tag{1.3}$$

where the contraction $f \otimes_r g$ is an element of $\mathcal{H}^{\otimes(p+q-2r)}$. In particular, the special case when $p = q = 1$ is mostly used as follows:

$$I_1(f)I_1(g) = \frac{1}{2} I_2(f \otimes g + g \otimes f) + \langle f, g \rangle_{\mathcal{H}} \tag{1.4}$$

1.3 Dissertation Overview

This dissertation focuses on the applications of Bayesian regression models, specifically in agriculture and climatology, as well as Wiener-chaos analysis on the predictive distribution of the response variable under linear-regression-model setup. Here is the breakdown:

- Chapter 2 introduces a multi-linear regression model to mainly explore the determinants of maize yield stability in Malawi.

- Chapter 3 develops a hierarchical Bayesian model to investigate the reconstruction and the projection of temperature anomalies in northern hemisphere.

- Chapter 4 proposes an estimation framework of predictive distribution using Wiener-chaos-expansion technique, which can give rise to a better understanding on the risk of a linear-regression model and the approximation error.

- Chapter 5 discusses the possible work directions in the future.

Chapter 2

Maize Yield Determinants and Management Strategies

2.1 Introduction

Yield gaps in African smallholder agriculture are pervasive and large. The yields achieved on the vast majority of African farms are 10-30% of their genetic potential [10]. Yield-limiting factors have been identified, such as environment, sub-optimal planting in terms of timing and spacing, deficiencies in soil nutrients, moisture, as well as damage from weeds and pests [11,12]. Agricultural economists commonly emphasize market prices, farmer education, and related socio-economic factors to influence on-farm production [13]. There are many challenges to carrying out effective diagnostic analysis of yield gaps, and often the focus has been on the size of the gap. Yet if research priorities and agronomic recommendations are to address farm-level constraints, there is urgent need for evidence-based examination of the main determinants of yield in specific contexts.

This is the first study to understand maize yield determinants by applying a Bayesian approach to a unique survey dataset from central Malawi. Crop simulation models are often used for gap analysis, and are suited to providing insights into yield potential as well as technology response to weather variability; however, they do not reveal the drivers of yield gaps [14]. In field experimentation, trials are often run under conditions that are not representative of on-farm conditions. Smallholders in sub-Saharan Africa often have marginal soils with less intensive management, facing weed, disease, and other pest problems [15]. The disconnect between soil conditions at research stations and those on smallholder farms is illustrated by a nationwide assessment in Malawi, where soil organic matter levels at research stations are 1.5 to 2 times as high as those observed on smallholder farms [16]. Researchers generally choose a field site and invest resources, so as to ensure a homogeneous environment, within which to evaluate a practice or to address a specific research question. Thus field research sites tend to be flat, uniform, and high-potential, given that conventional research experimentation usually tests one or two component practices while controlling other sources of variability [17].

The overall objective of the project is to conduct a Bayesian analysis of household-survey data that comprises multiple visits to maize-focal plots in central Malawi, to determine which variables influence the maize yield [18]. Specifically, we examine the predictive ability of time-series environmental factors and management practices regarding field observations of maize yield. Furthermore, we assess leaf chlorophyll status and parasitic weed incidence to provide site-specific models and recommendations.

2.2 Materials

2.2.1 Study Sites

Central Malawi agriculture is dominated by mixed maize production systems with limited livestock presence, which is broadly typical for poor-resource smallholder farms in southern Africa [19]. Administrative units in the Malawi government are comprised of region, agricultural development division, and extension planning area (EPA). The study sites are chosen using a stratified random sampling scheme, where all EPAs within central Malawi are classified using the strata of marginal, moderate or mesic for plant growth based on rainfall and evapotranspiration. There is one marginal site that contains two adjacent EPAs–Golomoti and Mtakataka (referred herein as Golomoti), two moderate-potential sites which are Kandeu and Nsipe, and one high-potential site called Linthipe, a total of 22 village clusters included within these five EPAs [19]. Golomoti is located near the lakeshore at a low elevation and a high evapotranspiration, with a mix of soil types dominated by Eutric Cambisols and Eutric Fluvisols. Linthipe has well-distributed rainfall and a long history of maize-dominated agriculture. Soils in Linthipe are primarily Ferric Luvisols, whereas in Kandeu and Nsipe, soils are mixed with Chromic Luvisols and Orthic Ferrasols [20]. Market access also varies across locations, with Kandeu and Nsipe being moderately remote, Golomoti and Linthipe being proximate.

2.2.2 Data Collection

The data are from a panel of 320 farm households, two maize plots per household surveyed in 2014/2015 and in 2015/2016, with a survey instrument approved through the Michigan State University Human Research Protection Program in the Office of Regulatory Affairs, following a human subjects' protocol with informed consent obtained from all farmers, translated into local languages. The farmers are asked to choose two maize plots at random, and the same plots and farmers are then revisited and surveyed at preseason (October 2014 and 2015), mid-season (March 2015 and 2016), and harvest (May 2015 and 2016). Enumerators are trained over a one-week period, and supervised by graduate students on the field. The data collection process involves close attention to data entry and quality control. The information on the survey is voluntary and every effort is carried out to maintain confidentiality.

The survey topics address crop production, socioeconomic information (household size and composition, household head's educational level), farm management and practices (labor, seed, planting dates, plot history, residue, crop grown, time of sowing and weeding, fertilizer application), and soil characteristics (pH, total carbon, permanganate oxidizable carbon). Mid-season survey is to assess maize planting arrangements (including row spacing), and maize leaf chlorophyll, which is based on soil plant analysis development (SPAD) absorbance. Enumerators record three reading replicates per plant for four plants at each of the eight locations. To avoid edge effects, the enumerators observe at least two ridges from the plot border, and randomly choose three locations (two-ridge apart) along a diagonal transect. The spacing is from the center of one ridge to the center of the adjacent one.

Additionally, two types of measurements are made to investigate the incidence of *Striga asiatica* (L.) Kuntze, commonly known as witchweed, a genus of parasitic plants. One is directly asking farmers if they have a problem with striga on a given maize plot. The other one is obtained by enumerators who make eight observations per plot at random sites along rows following a prescribed procedure; thus, striga information is recorded from 0 to 8 for each plot. At harvest, a survey is provided to measure maize yield by weighing biomass of stover and grain from three-square-meter plots per field, where grain is removed from cobs to allow for a dry weight basis.

In summary, there are 1,197 plots in total, which are geographically located with GPS coordinates in October of 2014, involving a small amount of missing data. The plot-level data is longitudinal, and combine socioeconomic characteristics, maize production, plant and soil information, as well as farm management.

2.3 Modeling Framework

2.3.1 Explanatory Variables

Environmental factors (rainfall and soil properties) are continuous variables, which are standardized for comparison with the estimated coefficients on the same scale. Rainfall data come from the Climate Hazards InfraRed Precipitation with Station (CHIRPS) resource, which is a public quasi-global rainfall data set starting from 1981. This is the only comprehensive precipitation data source that is available for Malawi, and has been previously validated by comparisons to local rainfall records for three of the five EPAs [19]. We include three variables in the model that indicate the amount of rainfall (in millimeter) for

(a) December, January, and February (the sowing period)

(b) March (the end of the rainy season)

(c) April and May (the harvest period)

to explore the association between seasonal-rainfall variability and maize yield by measuring at three critical stages of the maize growing season. Using more than one subset of semi-annual rainfall is common in certain agricultural studies and practices, such as the definition and calculation of rain-index-based crop insurance [21]. Regarding soil properties, permanganate oxidizable carbon (POXC), which is a sensitive indicator of active soil organic carbon, and pH are selected as the main soil factors in the model.

There are five farm-management variables, namely ridge (row) spacing, total ridge-weed biomass, fertilizer application, intercropping, and manure/compost use. Although fertilizer and SPAD are highly correlated, they both have explanatory power for maize yield. Ridge spacing (in centimeter) is the distance between two ridges at each of three locations within a field. Total ridge-weed biomass (in quadrat of 0.5 m^2) is weighted in the field at harvest, and treated as an index of weeding effectiveness. Intercropping and manure/compost use are both binary variables, and we do not consider the density or the genre of the crop that farmers intercropped with maize. Fertilizer is calculated by the amount of nitrogen applied with any type of fertilizer amendment. Endogeneity bias is possible, particularly for these choice variables. However, we are unable to address endogeneity because our data set does not include suitable instrumental variables, that strongly predict the endogenous explanatory variables but do not directly affect the response variables. As a result, coefficient estimates should be interpreted as indicating association rather than causality.

2.3.2 Regression Model

A Bayesian framework is applied to estimate the statistical relationship in a linear regression setting, and an agronomy perspective based on expert knowledge is used to form the basis of this model as the following:

$$Y_{ijk} = \alpha_i + X_{ijk}\beta_i + \sigma\varepsilon_{ijk} \tag{2.1}$$

- Y_{ijk} (response variable): maize yield (in kilogram per hectare) of plot k managed by farm household j at EPA site i
- α_i (y-intercept): can either be a constant or vary from location to location
- X_{ijk} (design matrix): incorporates all the data from 12 explanatory variables
- β_i (regression coefficients): measure how much of the variation of maize yield (Y_{ijk}) is explained by the covariates (X_{ijk})

- ε_{ijk} (noise terms): independently and identically distributed $\mathcal{N}(0,1)$ representing the error of the model

- σ: scale of the error

To evaluate the determinants of maize yield, we compare the effects of each explanatory variable on the response in any one of our three models (i.e., yield, SPAD, striga). Because all variables in the models have been standardized, the estimated regression coefficients (β_i) of each variable give a magnitude of influence, which can be compared with the scale of the noise terms (σ), to find out which predictor(s) has the strongest effect. Usually, σ is quite large relative to a single regression coefficient, but one should add the absolute magnitudes of several independent variables for a more meaningful comparison, since the statistical error needs to be compared to the strength of all the explanatory factors combined. For each model, the set of explanatory variables is chosen to be consistent with agronomists' beliefs about what factors may influence yield, SPAD, or striga. Each of the three models is specified in the simple linear framework, with appropriate logistic modifications in the case of the the striga model, to distinguish between incidence of striga and levels of striga. Such a linear framework can be seen as a first-order approximation for each response.

We can gauge the random effect of location since α_i and β_i both depend on the EPA index i. Pooling the two-year data avoids model misspecification, and has the added benefit of increasing statistical power. Further models test response variables SPAD and the parasitic weed striga, to uncover the underlying key drivers of maize yield, where we expect striga to be a negative factor and SPAD to be positive. The noise terms (ε_{ijk}) are assumed to be independent across all three models in order to minimize the number of parameters needed to be estimated. It also avoids the use of a large number of correlation parameters (hyperparameters) at the prior level in the Bayesian context, which need to be consistent when investigating a system of 3 equations with 11 common explanatory variables between any pair of models. Consequently unobserved factors that might simultaneously affect more than one model are not taken into account.

The eight striga records have been reverted to a scale of quantitative values from 0 to 8; with 0 indicating the absence of striga, and 1 to 8 revealing the level of striga infestation. The striga model can be thought of as a two-step procedure:

- First, the zero and non-zero values provide two alternatives which allow us to estimate the influence of the presence or absence of striga via logistic regression, speaking for the possibility of prevention.

- Next, when conditioning on the presence of striga (1–8), ordinary linear regression is employed, which links to the insights on effectiveness of striga control

Although having a large number of observations equal to zero (known as zero inflation) may induce biased results, the proposed strategy mitigates this problem, since the standard linear regression model is relative to the non-zero striga values.

In addition to the models described above, we conduct two sets of analysis to consider if socioeconomic indicators—educational level of the household head and total dependency ratio of each household—are of importance in predicting maize yield, by adding these two variables to the covariate matrix X. The available data is also used to investigate non-specific household effects, which utilizes simplified versions of the three linear models, allowing the y-intercept (α) to depend on the household identifier (dummy variable) to help determine whether households are predictive of maize yield, SPAD, or striga. This may be interpreted as pointing indirectly to socioeconomic effects, or directly to effects of farmer skill.

We apply the package ‘PyMC3’ built into Python to implement the methodology. This package provides a way to estimate the posterior distribution of our model parameters (i.e., regression coefficients and error terms) by implementing a sampling mechanisum for these distributions. It uses the ordinary Gibbs sampler to produce samples, with a burn-in period of 500 initial samples, and an additional 10,000 iterations with two independent MCMC chains after each burn-in. Without having prior knowledge on the distribution of the parameters, we employ the classical weakly-informative prior distributions: the standard normal distributions (for β_i) and inverse-gamma distribution (for σ). These prior choices present numerical advantages in terms of conjugacy [1,22]. We monitor the convergence of the procedure by keeping track of the discrepancy between the two aforementioned chains using R-hat statistics (a widely-accepted convergence diagnosis). All the R-hat values are below the acceptable threshold of 1.1, which implies that the chains successfully converge, producing excellent parameters’ estimates as well as their credible intervals.

In a word, we focus on biophysical determinants, yet there is more to be explored in the future regarding socioeconomic drivers. Given the limited amount of data, this project does not delve into the higher-complexity models, since they lie beyond the scope of our data set and thus of our analysis. Hence, the three structural models are uniquely characterized by their respective response variables and explanatory variables.

2.4 Results

2.4.1 Descriptive Statistics

Table 2.1 provides descriptive statistics for the variables in the model categorized by EPA location. Figure 2.1 shows the precipitation for all study regions during the maize-growing season over two years. Mean values for rainfall are consistent with earlier characterization of Golomoti as a low-rainfall (marginal) site. The other locations differ in terms of mean rainfall for March, with Linthipe (mesic site) having the highest rainfall level.

	Golomoti	**Linthipe**	**Kandeu**	**Nsipe**
Environment				
Dec.–Feb. precipitation (mm)	563.0 (78.0)	624.5 (27.7)	629.5 (94.6)	636.8 (132.8)
March precipitation (mm)	78.6 (26.8)	128.9 (42.0)	102.8 (26.5)	116.8 (25.4)
Apr.–May precipitation (mm)	25.2 (10.4)	43.8 (11.6)	39.1 (9.1)	44.4 (8.8)
POXC (mg carbon/kg soil)	278.9 (152.43)	466.9 (220.5)	390.41 (191.15)	340.70 (160.22)
Soil pH	6.56 (0.61)	6.09 (0.46)	6.10 (0.53)	6.32 (0.61)
Crop performance				
Maize yield (kg/ha)	1567.44 (1039.3)	2636.3 (1526.0)	2069.4 (1471.5)	2320.9 (1452.9)
SPAD	41.20 (8.85)	46.98 (7.15)	46.00 (8.62)	41.84 (8.11)
Management practice				
Maize spacing (m)	0.897 (0.11)	0.927 (0.11)	0.970 (0.13)	0.914 (0.14)
Weed biomass (kg/m^2)	0.183 (0.16)	0.079 (0.07)	0.201(0.15)	0.246 (0.16)
Fertilizer (kg)	8.908 (13.60)	13.763 (25.48)	15.901 (18.04)	12.411 (12.86)
Striga (binary)	0.22 (0.42)	0.31 (0.46)	0.16 (0.37)	0.28 (0.45)
Intercrop (binary)	0.66 (0.47)	0.77 (0.42)	0.74 (0.44)	0.60 (0.49)
Compost (binary)	0.41 (0.49)	0.45 (0.50)	0.31 (0.46)	0.27 (0.45)
Total number of observations	312	282	298	305

Table 2.1: Mean and (standard error) of all model variables by location

Soils are generally marginal at Golomoti sites, as evidenced by low mean value for soil active carbon (POXC), in accordance with previous reports [19]. The highest POXC value appears to be in Linthipe. Soil pH varies little among locations, and mean values are consistent with moderate acidity, and thus non-limiting pH conditions for the crops grown.

Crop response consists of maize yield and leaf nitrogen content, as indicated by SPAD values. Average maize yield is the lowest in Golomoti, followed by Nsipe and Kandeu, and the highest is in Linthipe. SPAD data has a similar but not identical pattern: low in Golomoti and Nsipe, and high in Kandeu and Linthipe.

Striga incidence and weed biomass are distributed across EPAs with no clear spatial pattern. Farmer-reported striga problems on about 16-30% of sampled fields, and from 0.18 to 0.25 kg/m^2 dry-weight weed biomass remaining in the field at harvest. The latter is an

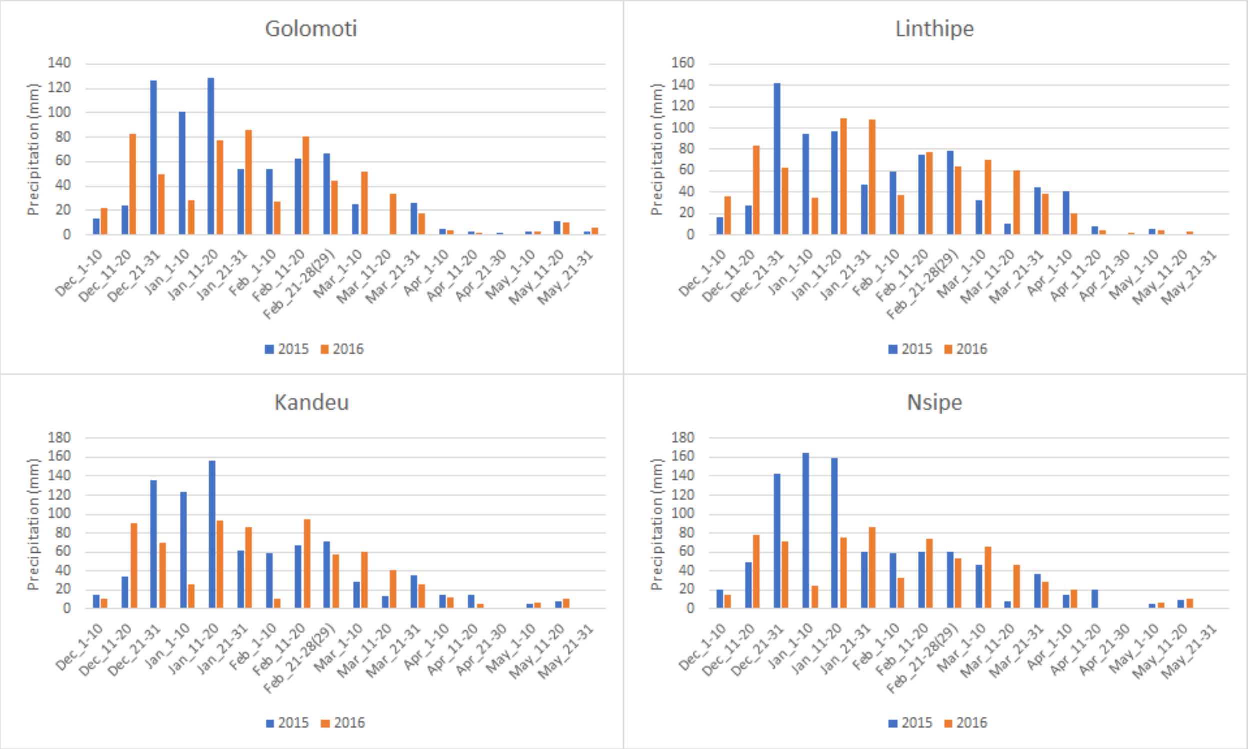

Figure 2.1: 10-day precipitation (CHIRPS) from December to May at all EPA sites

indicator of how effective farmers' weed management is, although the endogenous infestation levels of weeds at a site could also contribute to observed presence.

Overall, fertilizer use is lower at Golomoti site than the other ones. This is accordant with the intuition that farmers in marginal environments have less motivation to invest in their lands and crops. Fertilizer application is similar in Linthipe, Kandeu, and Nsipe. Compost use is higher in Golomoti and Linthipe than it is in Kandeu and Nsipe. As expected, intercropping is more frequently practiced in Linthipe and in Kandeu, where many farmers grow bean and cowpea (legume) in mixed stands with maize. Average plant spacing is lower in Golomoti and Nsipe, and the highest in Kandeu.

2.4.2 Bayesian Inference

Statistical significance is determined under a linear Bayesian circumstance. For example, an explanatory variable (such as SPAD) for a response variable (such as yield) is statistically significant at 95% Bayesian credibility if its regression coefficient has a posterior probability of being on one side of zero which exceeds 95%. It is true as soon as the 95% credible interval of that variable's regression coefficient lies on either side of zero. If this condition fails to happen, then strictly speaking, one may accuse it of not being statistically significant. This is a steep threshold to apply in most cases, since a variable with 90% credibility, still holds some predictive information. In this project, however, when an explanatory variable fails to be significantly associated with the model's response variable, the failure occurs at a much lower level than 90%. The size of an association needs to be distinguished from its significance. For variables which are significant in terms of size (or intensity), the significance are measured by their regression coefficients' posterior mean.

2.4.2.1 Maize yield

Figure 2.2 presents the 95% credibility intervals for the determinants of maize yield at different sites. In general, maize yield is positively associated with SPAD and negatively associated with plant spacing; the magnitude of the coefficients for SPAD are particularly large. There is also a small but positive direct relationship between fertilizer and maize yield for Kandeu and Nsipe. Yield is not consistently affected by soil pH, weed biomass, intercropping, or March rainfall. The α values differ markedly from each other by EPA, which suggests that location has an important influence on maize yield, independently of other factors in the model. Since those factors are capable of exacerbating differences in maize yield by location, we turn to consider specific locations from now on.

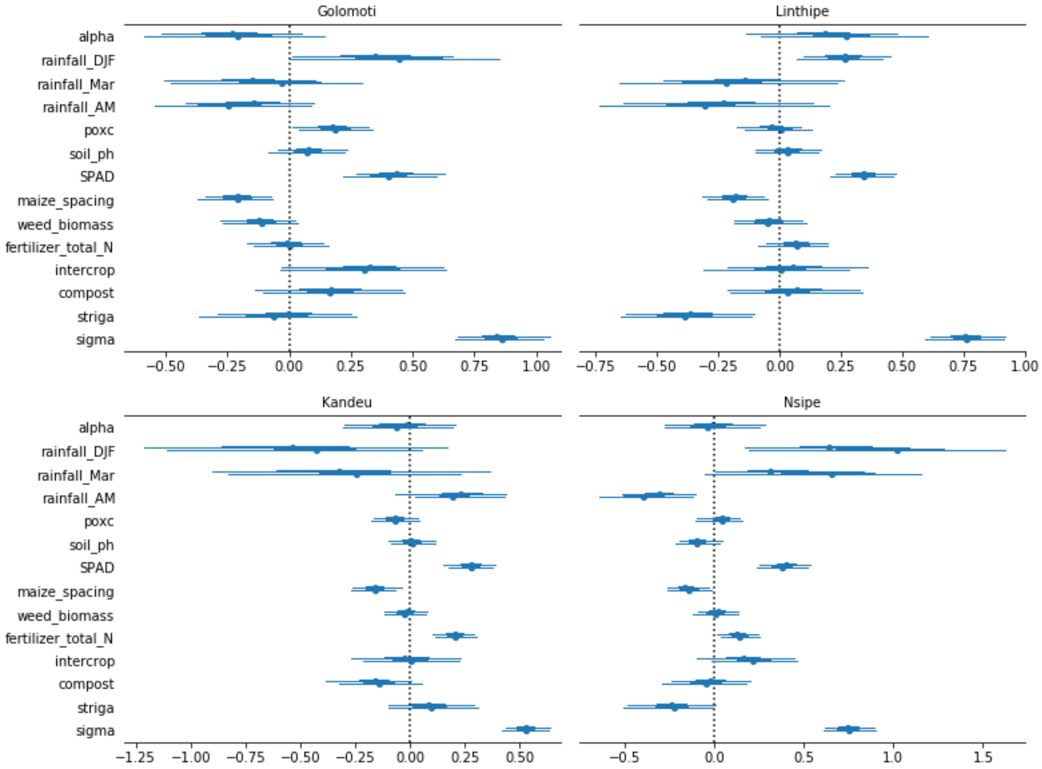

Figure 2.2: 95% credible intervals for the determinants of **maize yield**

First, rainfall levels are only significant factors for yield in Linthipe and Nsipe. In both areas, early-season rainfall (i.e., December to February) has a positive relation with maize yield. In Nsipe, high rainfall in April/May is associated with low maize yield, perhaps a reflection of late-season disease harming the crop such as Fusarium ear rot [23]. Secondly, soil active carbon and compost application are positively associated with maize yield at Golomoti. Third, in Linthipe and Nsipe, striga has a large negative effect on maize yield, whereas it is not significantly related with maize yield in Golomoti or Kandeu. Slightly-streamlined yield models are also investigated, where either SPAD or striga is removed. It turns out that these reduced models resulted in decreasing explanatory power for all variables, which could be easily assessed via the posterior distributions of regression coefficients.

Overall, the results indicate that maize leaf nitrogen (SPAD) and striga are the strongest determinants of maize yield at all sites, while early and late rainfall are significant at some sites. As farmer behavior can directly influence striga and SPAD, we evaluate separate models to uncover the drivers of these two critical inputs.

2.4.2.2 SPAD

As we can see from Figure 2.3, there are three main predictors for SPAD: rainfall, POXC, and application of fertilizer or compost. Rainfall is generally found to be influential on SPAD, except in Kandeu, where rainfall variables are not statistically significant. In Golomoti, SPAD increases with March rainfall, whereas in Linthipe and Nsipe it is the early rainfall that has a positive impact on SPAD. Also in Nsipe, a negative association with SPAD is observed for late-season rainfall.

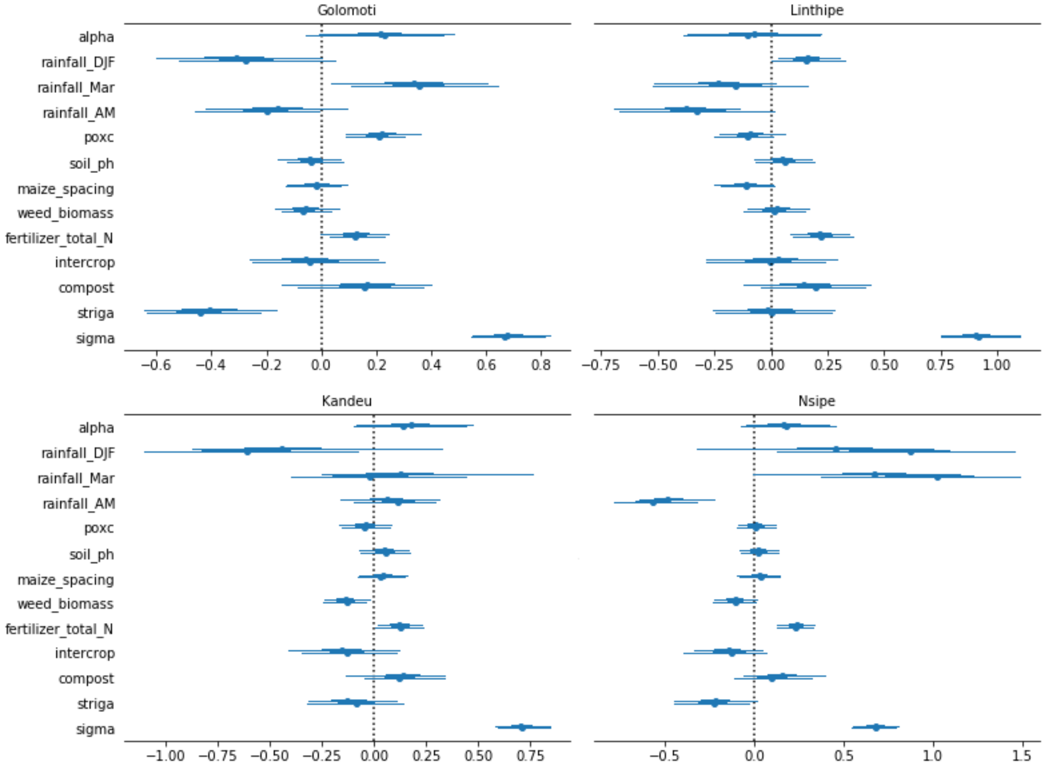

Figure 2.3: 95% credible intervals for the determinants associated with **SPAD**

Fertilizer quantity is an important driver of SPAD at all sites except Golomoti, which in turn is highly predictive of maize yield. The magnitude of its effect on SPAD is higher in Kandeu and Nsipe than it is in Linthipe. In dry and marginal sites, plant growth and response to fertilizer are often limited by insufficient soil moisture, thus fertilizer application does not necessarily lead to plant uptake of nitrogen (or yield). The results in the yield model also show a lack of response to fertilizer in Golomoti (see Figure 2.2), where compost/manure and POXC are positively related to SPAD. Fertilizer impacts on both SPAD and yield are substantial, however, suggesting the need to improve the effectiveness of fertilizer applied.

2.4.2.3 Striga

Factors that influence striga incidence (Figure 2.4) and level of striga infestation (Figure 2.5) are shown below. Among farm management and soil properties, most are either not significant or of small magnitude in relationship to striga. There is some evidence of soil fertility amendments being useful in prevention, as fertilizer application is associated with striga absence (as reported by farmers) in Kandeu and in Linthipe. Fertilizer is also an apparent control factor in Linthipe, where it is a predictor of low striga-infestation level. Compost is related with both striga absence and low incidence in Nsipe, with contrasting results observed in Kandeu. At Golomoti site, fertility amendments have no striga control benefits, and the only farm-management effect is wide ridge spacing, which is negatively correlated to striga presence and intensity. In addition, intercropping is neither helpful nor harmful to striga prevention and control. The only exception is Linthipe where intercrops are associated with striga problems.

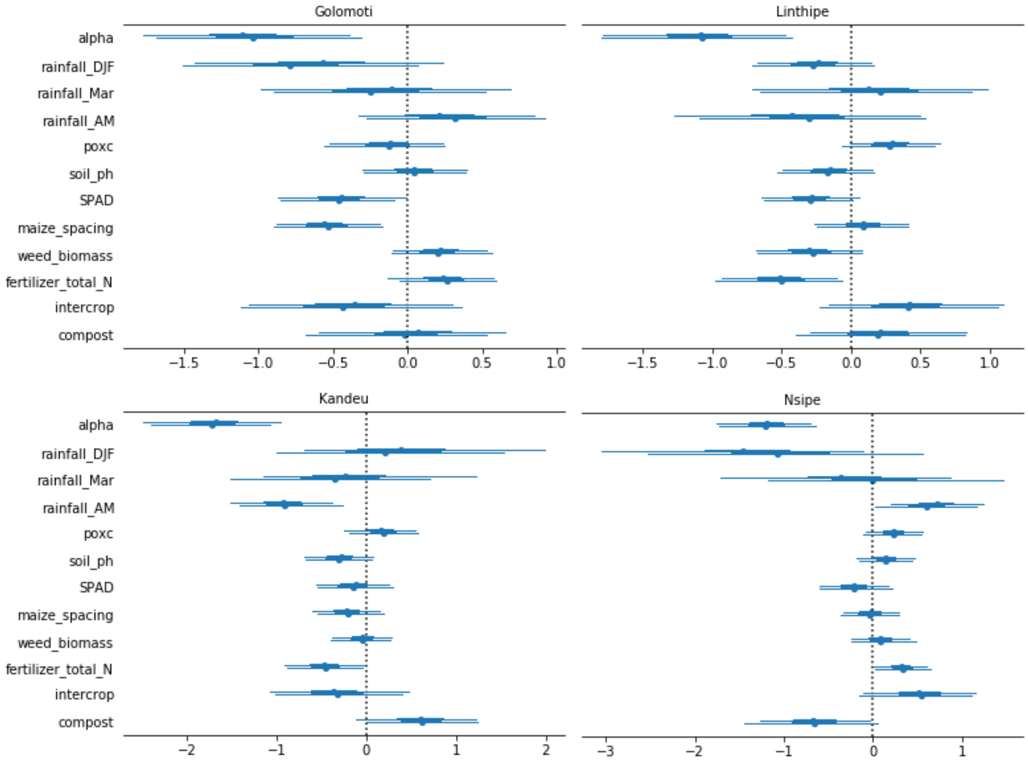

Figure 2.4: 95% credible intervals for the drivers of **striga prevention** (logistic)

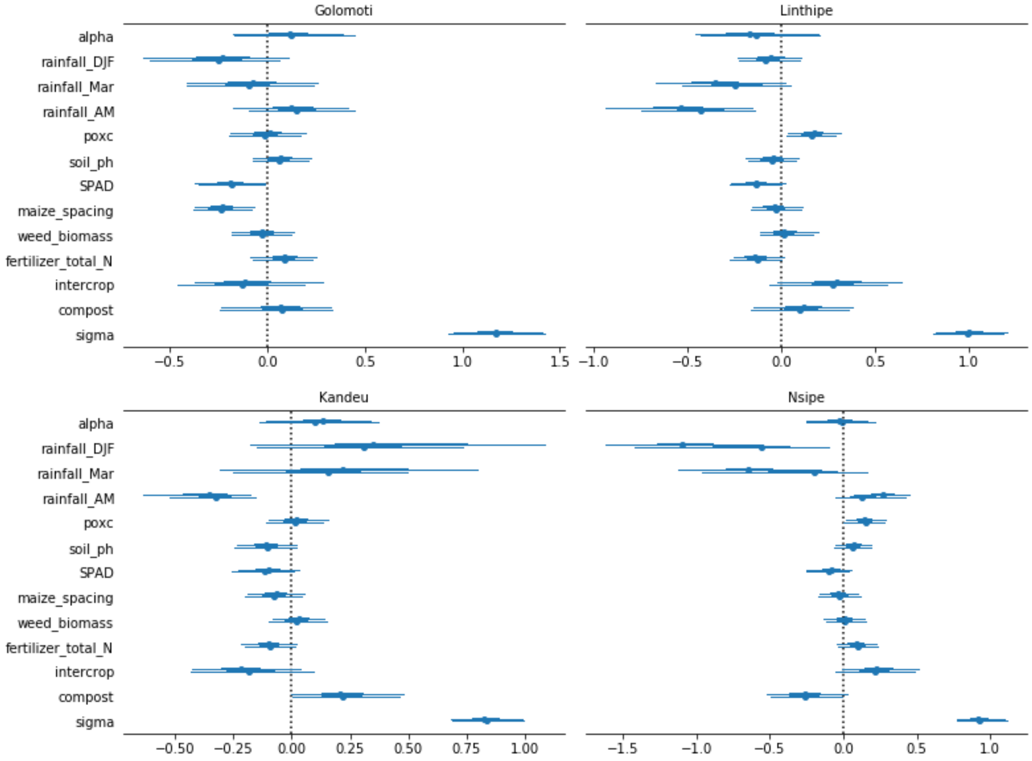

Figure 2.5: 95% credible intervals for the determinants of **striga control**

2.4.2.4 Household (farmer) effect

With most of the 320 households, and up to four plots per household (two plots per household in two years), we find that it is not possible to determine if there is a connection between any particular farmer and the corresponding data. However, some statistical significance is extracted from roughly 10% of cases, meaning that the variability among the four data points of such households is almost certainly not due to chance alone. Moreover, the effect is most likely to identify a favorable household environment. Specifically, setting the significance level at 5%, we compute the number of households for which the regression intercept α is away from zero with posterior probability at least 95% (see Table 2.2).

Model	**Positive** effect	**Negative** effect
Yield	23 (7.5%)	5 (1.6%)
SPAD	11 (3.6%)	7 (2.3%)
Striga	26 (8.5%)	0 (0.0%)

Table 2.2: Number and (proportion) of households where the y-intercept is significantly non-zero at 95% credibility level in all models

Despite the limited data per household, we are able to detect such effects with nearly 10% of the households. In the yield model, the total number of significant households exceeds 9%. Interestingly, among these households, there are far more cases where the yield is higher than it is lower (7.5% against 1.6%). In other words, when the data towards a farmer having a significant effect on yield, the odds are about 5:1 that this is a skilled farmer with high maize yield. For SPAD, about 6% of households have SPAD levels which cannot be explained by chance. The odds of having high SPAD against low SPAD is about 3 to 2.

As for the striga model, there is no case that a farmer could avoid striga entirely, while 8.5% of farmers are likely to be associated with a striga problem. It may seem surprising to say that we cannot determine any farmer with the skills or the knowledge of practices to be superior to others in preventing striga. This result is not to be taken as a discouraging fact. Instead, it reflects that over two thirds of all plots in the study regions are striga-free. Thus, while one third of the plots being infected with striga reaches epidemic levels, having two thirds of plots without striga makes its absence so prevalent that the 320 households cannot identify anyone with unusual striga-prevention skills. Readers are referred to the full striga model for more precise recommendations on striga.

2.4.2.5 Socioeconomic factors

The credible intervals for the two socioeconomic drivers (2nd and 3rd line in Figure 2.6) overlap rather heavily with the zero vertical line in most cases, and are small in magnitude compared to other determinants, indicating inconclusive significance and small impact. It also shows that the regression coefficients of other variables are insensitive to whether or not one includes the additional two socioeconomic indicators. This could be a sign of insufficient data to draw conclusions on either rejecting or asserting socioeconomic importance at any reasonable level of significance (e.g 80% credibility or higher). We also carry out additional models by removing insignificant variables from X, and the remaining analysis (not reported) are largely unaffected, which are similar to Figure 2.6. In sum, these evidence imply that our regression model is robust.

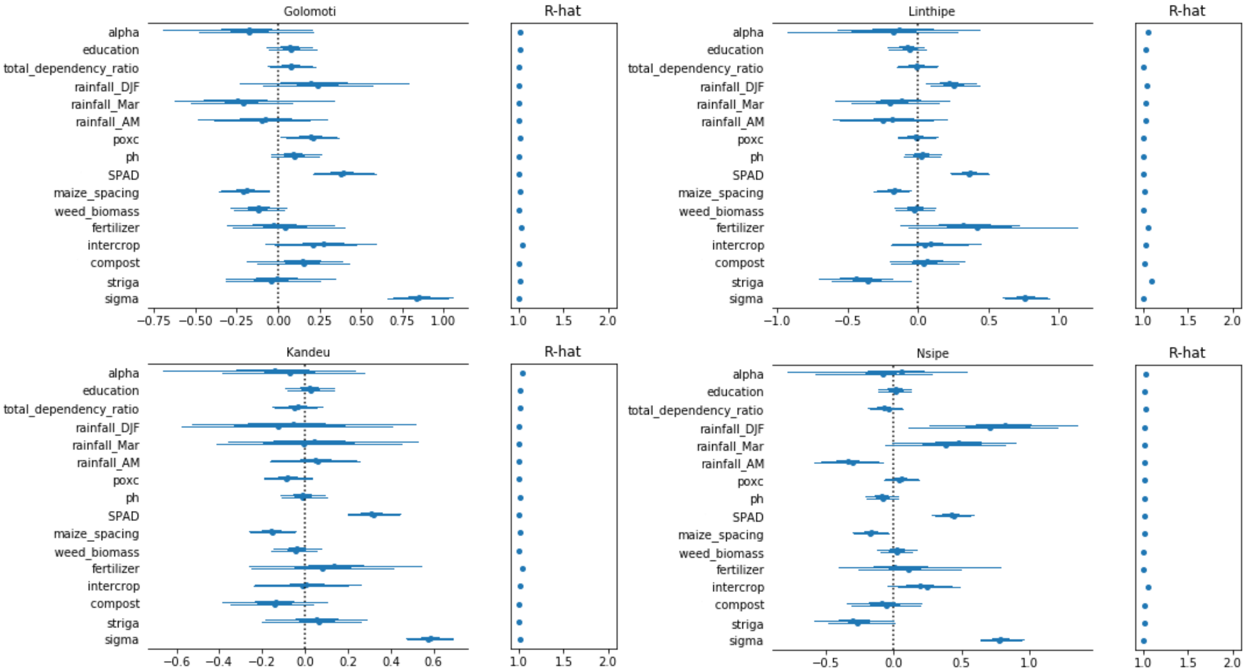

Figure 2.6: 95% credible intervals for determinants including two **socioeconomic** factors associated with maize yield

2.5 Discussions

The plot-level longitudinal data set and Bayesian regression models place biophysical constraints in sharp focus in the analysis of what influences maize yield in central Malawi. Some have strong effects with very high credibility, notably SPAD, striga, and sub-seasonal rainfall pattern, which is consistent with much of the literature on small-scale, mixed-maize production systems [12,15]. Yet, fertilizer application does not necessarily lead to improving leaf nitrogen nutrition (SPAD) or in subsequent maize yields. Hence, the results highlight the challenges to ensure effective nitrogen uptake and translation into grain, particularly in marginal environments.

Fertilizer is generally associated with high plant nitrogen tissue at all areas but Golomoti. Soil organic carbon fractions tend to be sensitive for crop response at lower values, as Golomoti is 20-40% lower compared to the other locations, which both recommend farmers to apply manure or legume rotations at marginal sites, to build nitrogen supply capacity [14]. Integrated management of soil fertility combining manure and fertilizer has been shown previously to be highly effective and profitable for raising maize yields [24]. In concurrence, a study finds out that maize yield response to nitrogen fertilizer application is low when soil organic matter is low [25].

Rainfall is another crucial factor, particularly early-to-mid season, which supports a positive association between seasonal rainfall and maize yield based on an econometric analysis of farm-level data [24]. At Kandeu, we find a negative relationship of late-season rainfall to yield and SPAD. This could reflect a leaching problem, with rainfall inducing soil inorganic nitrogen losses and thus limiting nitrogen availability during the critical maize grain filling period, which requires high nitrogen availability [26]. It could also be related to Fusarium or other infections of the corn ear, induced by a late-season moist environment causing grain spoilage and yield loss [5]. To our knowledge, it is the first study to produce this type of differentiated conclusions based on panel survey data.

Although weed biomass alone has no discernible effect on maize yield, parasitic weed striga is a strong negative determinant, especially in Linthipe and Nsipe. It is a bit surprising that the phenomena are only observed in the two relatively high-potential locations, as the presence of striga is typical throughout the study sites and is indeed ubiquitous nationwide [27]. This implies that a barrier to crop production that has been largely overlooked by agricultural research and policy makers, where the focus has often been on subsidized access to hybrid maize seeds and fertilizers. Effective and affordable means of striga prevention and control are in need.

In Malawi, farmers primarily rely on hand weeding for striga control, which appears to be apparently ineffective due to the parasitic nature of striga. Therefore, maize growth suppression has already occurred by the time hand weeding is done [28]. The utility of fertilizer and manure/compost links low-nitrogen soil to higher striga incidence [29, 30]. Similarly, farmers in a recent survey rank manure application as the best option for striga control [31]. Another study reveals complex relationships that early application of fertilizer helps maize plants overcome early effects of striga attachment, whereas late application is associated with worse striga [32]. Given the observed differences in soil fertility and fertilizer across the EPAs, we expect higher striga infestation in Golomoti over the other areas, instead there is a widespread presence of striga and lack of uniformity in what works for prevention and control. For instance, fertilizer is found to be a critical factor for striga, but only in Kandeu and Linthipe. Manure/compost appears to have merits for reducing striga problems in Nsipe only. In short, our findings suggest a complex mode of action with difficult-to-predict reactions of maize to striga presence.

Overall, the results from this project call for area-specific recommendations. The Malawi government's suggestions for hybrid maize production have mainly focused on nitrogen rate [33]. We find evidence that shows targeting complementary investments and timing of application could potentially add value. For example, maize yield at the mesic sites would gain benefits on early and judicious use of fertilizer for not only striga control, but also for nitrogen nutrition. Whereas for the marginal sites, response is markedly different, with no fertilizer or striga drivers observed on yield, and instead soil active carbon and compost are positive determinants of yield, recommending practices should focus on managing soil organic matter. Soil carbon accumulation provides important environmental services at all sites; however, marginal locations benefit from stable production over time and space to substantial gains in nitrogen efficiency, and the incremental gains in soil carbon are high.

Compost preparation and utilization at modest amounts have beneficial influence on some sites, with gains in plant health as indicated by nitrogen status and striga suppression. This management practice is especially helpful during a poor-rainfall season in a marginal area. With little to no downside, government policies, extensions, and educational efforts by civil society, can all play crucial roles in building appreciation for compost advantages.

Chapter 3

Global Temperature's Reconstruction and Projection

3.1 Introduction

Understanding the evolution of the climate on Earth is one of the major scientific challenges for this century. Quantifying the link between the sensitivity of the climate system constituents, and its response to the climate forcing factors can be critical to discover the underlying relationship in climate change [34]. Although the most updated instrument-base observational temperature has a record since 1850, it is desirable to extend the temperature-reconstruction methodology to make use of climate proxies, hence better capturing the climate variability on different time scales [35].

The Intergovernmental Panel on Climate Change (IPCC) is dedicated to providing an objective and comprehensive view of climate change to the world, and pushes forward the development of accurate climatic models like General Circulation Models (GCMs), which are used for forecasting global and regional climate change [36]. Coupled Model Intercomparison Project Phase 6 (CMIP6) is another ongoing collaborative framework organized by the World Climate Research Program (WCRP), aiming to gather the efforts of the international climate research community, to improve the design of global climate model [37,38]. In addition, there are plenty of research articles relating the past temperature to proxies [35,39,40], such as tree ring, ice cores, lacustrine deposits, etc., implying that the natural proxies are good indicators for paleoclimate because they accurately reflect a wide range of climate sensitivity [41].

Previous approaches used in temperature reconstruction including principal component regression (PCR) [42,43], canonical correlation analysis (CCA) [43–45], regularized expectation maximization (RegEM) [46–48], and linear regression with various kinds of regularizations [49,50]. These literatures significantly improve the comprehension of the climate in the past. However, the lack of spatiotemporal covariance specification and uncertainty quantification continue to be the statistical challenges [51–53]. Several recommendations like model simplification, model averaging, have been proposed in order to address these issues [54–56].

The idea of using Bayesian model to reconstruct past temperature first appears in [57], and further studied by [40, 58, 59]. An advantage of applying hierarchical modeling framework to paleo-climate reconstruction problem is that it can incorporate different sources of information (proxies, forcings), to capture the variability of temperature. Furthermore, parameters estimates are systematically updated within Bayesian inference, which simultaneously provides the full posterior distributions of each model parameter to better understand the relation between each factor and the mean temperature, as well as reduces and quantifies the reconstruction uncertainty in a more statistically meaningful way [35, 40, 53].

Stochastic models have been applied to hydrology and coral archives involving proxies and paleo-data [60,61]. In this study, we develop a multi-level stochastic model to reconstruct northern hemisphere (NH) temperature over the past millennium (1000-1999), and extend the three forcing time series (solar, volcanic, greenhouse gases) to project the temperature in this century (2019-2100). The first level (data) embeds the climate proxies into the underlying temperature anomalies, and the second level (process) linearly relates the temperature to solar irradiance, volcanic activity, and greenhouse gas concentration. Finally, we implement the Bayesian technique to estimate all the parameters in the model, due to a limited amount of data in the calibration interval.

3.2 Data Source

There are three data sources employed in the project. Four main types of proxies including tree ring widths, lacustrine sediment cores, ice cores and speleothems (cave formations) between 1000 and 1999. Observational surface temperature over the period 1900-1999, and estimates of three natural climate forcings (solar, volcanic, greenhouse gas) from 1000 to 1999 (for reconstruction), as well as from 2019 to 2100 (for projection).

The original proxy data comprises of 1209 annually and decadal series [62]. Owing to the focus of NH average temperatures reconstruction instead of spatial analysis, it is reasonable to reduce the dimension of proxies and maintain as much climatic information as possible at the same time. Moreover, proxy reduction can avoid model overfitting and decrease computational time of parameters estimations, yield a more parsimonious model due to a limiting calibration period. All the natural proxies are aggregated into a single variable through a rigorous selection and averaging process [41, 58, 63].

HadCRUT4 is the longest and most up-to-date global temperature data set, which combines observational near-surface air temperature [64] and sea-surface temperature [65] anomalies evolution since 1850 [66]. Anomalies are calculated relative to the average of 1961-1990 [65, 67]. Because the climate proxies may intrinsically carry notable noise, and the data before 1900 may not be reliable enough, we choose 1900-1999 to be the calibration time to circumvent the potential significant amplitude attenuation issue [68].

Total solar irradiance (TSI) is the measurement of solar power per unit area (W/m2) on the Earth's upper atmosphere. The solar activity is a complex phenomenon, whose fluctuation is considered to be around an overall constant 1361.0 W/m2 [38, 69]. We take a proxy-base reconstruction of TSI [70] that is in accordance with the climate reconstruction in [52]. Its relationship with NH temperature is studied in [71]. In terms of projection, we use the dataset recommended for CMIP6, where it calculates the weighted average of three statistical models, namely, analogue forecast (non-parametric), autoregressive model (parametric), and harmonic model (non-linear) [38]. The volcanic forcing is proposed based on ice core aerosol proxies, which is generated by the ejected particles during the explosive volcanic eruptions [72]. The impact of volcanism on global and regional temperature are also investigated [73, 74] .

Greenhouse gas concentration (measured by parts per million (ppm), denoted by CO2), is the most dominant forcing account for climate variations since 1950 [72]. According to the level of greenhouse gas (equivalent to radiative forcing in W/m2) at Year 2100 (Figure 3.1), the IPCC chooses and names four possible representative concentration pathways (RCP) for the evolution of the global greenhouse gases, depending on how fast the human civilization adapts itself to reduce the emission of greenhouse gas [75, 76]. The first scenario (RCP2.6) describes a trajectory where the emissions stay very low, as a result of the environment-policy changes drastically made by governments and firms [77]. The medium scenarios are RCP4.5 and RCP6.0, a situation where the greenhouse gases increase but still stabilize before or after 2100 [78]. RCP8.5 is the worst scenario where the greenhouse gas concentration keeps increasing without stabilization [79].

All three forcings are available over 1000 till 1999 for reconstruction purpose. The full description regarding their derivation is in [72].

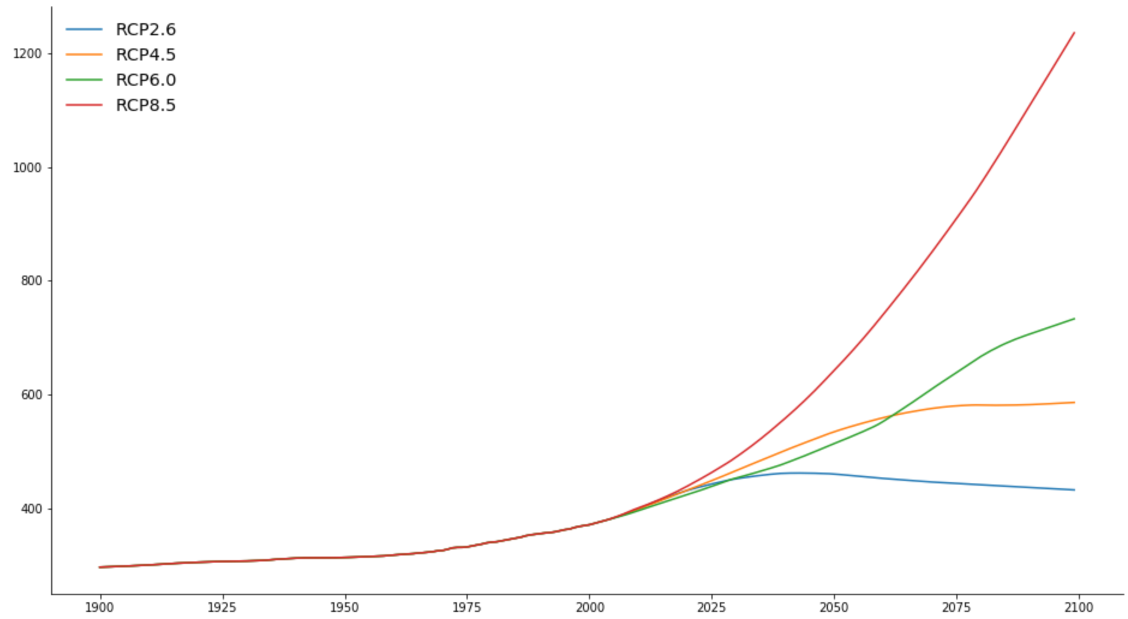

Figure 3.1: Greenhouse gas projections of four RCP scenarios

3.3 Model Specification

3.3.1 Forcing Transformation and Extension

The volcanic influence can be extremely strong among the years following eruptions, but fast decays over non-volcanic years, leading to a cooling effect in the long run [74,80]. Hence, we apply a decreasing logarithm transformation

$$\tilde{V}_t = \log(-V_t + 1) \tag{3.1}$$

to the original data to mitigate the impact of large volcanic eruptions during calibration period [35]. Also, we consider the simplest transformation

$$\tilde{C}_t = \log(C_t) \tag{3.2}$$

given that the radiative forcing depends on greenhouse gases logarithmically [81, 82].

In order to project the NH mean temperature for the century, we need to expand the three forcing series to 2100. Following [38], we remove the 11-year cycle in the solar forcing via a moving-average process, since this periodic cycle does not appear in the reconstructed data that we use. In addition, the solar constant of the series has to be adjusted by about 5 W/m2, because of the technological update on its evaluation [38, 69].

As for greenhouse gas forcing, we add a random drifting noise term to each RCP scenario, to capture the uncertainty in the global evolution with respect to the climate policies' change. Specifically, we have:

$$C'_{2019+t} = C_{2019+t} + \frac{t}{82}\sigma_C \varepsilon_C, t \in [0, 81] \tag{3.3}$$

with $\varepsilon_C \sim \mathcal{N}(0, 1)$ and σ_C is decided by

$$\sigma_C = \frac{1}{10}\big(\max C_t - \min C_t\big), 2019 \leq t \leq 2100 \tag{3.4}$$

which yields an uncertainty that grows with the RCPs' magnitudes. Various uncertainties have been applied to test robustness of the model (refer Appendix A).

To our knowledge, there has not been any literature about the long-term volcanic prediction on a global scale. Most studies concentrate on either a specific volcano or forecasting eruptions short-periodically ahead [83–85]. Therefore, we treat volcanic activity as a stochastic process, allowing to take volcanic uncertainty into account for the temperature's projection. Note that the volcanic aerosol concentration decays similarly (up to a rescaling factor, see Figure 3.2) after eruptions, we thus model the volcanic time series as a succession of spikes with random amplitudes and time intervals, which can be written as:

$$V_t = V_0 + \tau_t + \sum_{i=0}^{\infty} \mathbf{1}_{\{t \geq T_i\}} A_i P_{t-T_i} \tag{3.5}$$

where

- V_0 is an overall constant
- τ_t is a small random fluctuation
- T_i are the dates of eruptions with amplitude A_i
- P_{t-T_i} is the decreasing pattern of the aerosol concentration after each spike

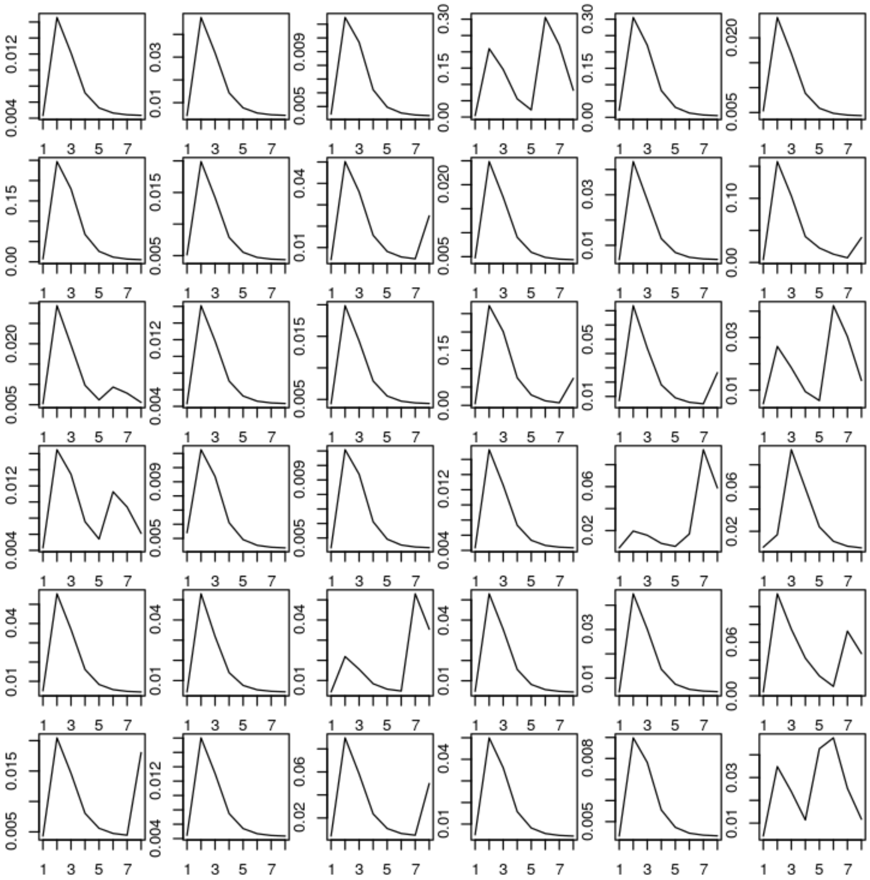

Figure 3.2: The first 36 re-normalized volcanic eruptions' decay

Moreover, the time increment ($T_{i+1} - T_i$) and the spike amplitude (A_i) are assumed to be independent and identically distributed. This assumption is verified by high p-values of Ljung-Box test (autocorrelations of a time series) applied to both series [86–88]. Eventually, the three natural forcing factors are normalized to make their influence on the temperature comparable among each other.

3.3.2 Multi-level Autoregressive Model

The hierarchical stochastic model is established as follows:

$$\begin{cases} P_t = aP_{t-1} + \alpha_0 + \alpha_1 T_t + \sigma_P \varepsilon_t \\ T_t = bT_{t-1} + \beta_0 + \beta_S S_{t-1} + \beta_V V_{t-1} + \beta_C C_{t-1} + \sigma_T \eta_t \end{cases} \tag{3.6}$$

where

- ε_t, η_t: white noises (mean 0, standard deviation 1) without any correlation structure
- a, b: autoregressive parameters
- α, β: coefficients associated with temperature and forcings

At any given year, the non-stationary model has the advantage of taking into account the impact of forcings in the past years (with an exponential decay) to the current temperature (see Appendix B for further details). From a Bayesian perspective, we assign the following prior distributions to the parameters in the model:

$$\begin{aligned}
\alpha = (\alpha_0, \alpha_1) &\sim \mathcal{N}([0,1], I_2) \\
\beta = (\beta_0, \beta_S, \beta_V, \beta_C) &\sim \mathcal{N}([0,1,1,1], I_4) \\
\sigma_P^2, \sigma_T^2 &\sim \mathcal{IG}(2, 0.1) \\
a, b &\sim \mathcal{U}(0,1)
\end{aligned} \tag{3.7}$$

I_n is an n-dimensional identity matrix. $\mathcal{IG}$, $\mathcal{U}$ represent inverse gamma distribution and uniform distribution respectively.

The model is run by Gibbs sampling algorithm with 10,000 iterations, but we only take the last 5,000 samples since the beginning of the chain (burn-in period) is often discarded, owing to its lacking accuracy to represent the desired distribution. The choice of prior distributions (normal and inverse gamma) for the linear coefficients and variances, is for the sake of conjugacy (posterior distributions yield same distributions as priors with different parameters), thereby avoiding computational burden [89, 90]. The complete computation of posterior conditional densities for the model may be found in Appendix C. It appears that small variations in the memory coefficients (a and b) lead to convergence instability for other model coefficients (α and β). Thus, we explore all the combinations of $(a, b) \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}^2$ (no memory to strong memory), and run the Gibbs sampler for all other parameters. The model is implemented in Python using a few fundamental libraries such as NumPy and pandas.

3.4 Results

3.4.1 Validation Metrics

Several metrics are produced to assess the quality of the temperature's reconstructions as well as to evaluate the posterior distributions of the parameters, applying them as criteria to compare with other works.

We first calculate the root mean square error (RMSE) by using the mean of the posterior distributions. It is a frequently employed measure of the differences between observed values and predicted values [91–93]. The empirical coverage probability (ECP) indicates the proportion of true value of interest (temperature) fall into the confidence interval [89]. Lower RMSE implies better model fit.

Additionally, we provide two scoring rules which are interval score (IS) and continuous ranked probability score (CRPS) [94–97]. IS rates the posterior confidence interval in function of the quantiles of posterior distribution for each observation, rewarding sharp prediction intervals and penalizing uncovered observations [96]. ECP and IS are both computed and compared at the levels of 80% and 95%. CRPS is defined as a squared distance between the cumulative distribution function $F(y)$ and the indicator function $\mathbf{1}_{y \geq x}$ of the predictive distribution as below:

$$CRPS(F, x) = \int_{-\infty}^{\infty} \left[F(y) - \mathbf{1}_{y \geq x}\right]^2 dy \tag{3.8}$$

Both scoring rules and ECP are positive-oriented with the designated model (higher values indicate a more appropriate model).

A Markov chain tends to run through as much state space as possible, and continuously reduces its variability to focus on the areas of the space with high density for an invariant distribution, which it converges toward in the end. Therefore, the total variance of the chain shrinks until it asymptotically converges to the variance of the invariant distribution. In order to determine the performance of the MCMC's convergence, we compute the potential scale reduction factor (PSRF) for the posterior samples [98, 99]. The PSRF (optimal value is 1) estimates how much variance needs to be reduced before achieving convergence.

3.4.2 Reconstruction (1000-1999)

Smaller memories are related to lower RMSE (better model performance) is not surprising, because the no-memory model is equivalent to a linear regression, which comes down to minimizing the squared error (Table 3.1).

a b	0	0.1	0.3	0.5	0.7	0.9
0	0.157	0.160	0.169	0.182	0.204	0.219
0.1	0.148	0.151	0.161	0.175	0.199	0.215
0.3	0.139	0.142	0.152	0.166	0.189	0.208
0.5	0.136	0.138	0.146	0.159	0.187	0.212
0.7	0.138	0.143	0.153	0.168	0.203	0.238
0.9	0.175	0.190	0.221	0.288	0.341	0.375

Table 3.1: Root mean square error (RMSE) for all combinations of memory coefficients

For the ECPs, the credible intervals are widened at both levels 80% and 95% when increasing memory parameters, and a good compromise can be found with a between 0.5 and 0.7, b between 0.3 and 0.7 (Table 3.2). The IS (80% level) does not exhibit any specific trend, except for extreme memory value ($b = 0.9$) returning unsatisfactory results. Whereas at level of 95%, enlarge a and diminish b would lower the negative interval score overall indicating a better model fit (Table 3.3). The continuous ranked probability scores are quite similar for a and b below 0.7, but become larger for higher memory values (Table 3.4).

Since the best performances are achieved with a and b at the order of magnitude described

a b	0	0.1	0.3	0.5	0.7	0.9
0	71	69	67	73	81	88
0.1	68	67	66	69	81	87
0.3	60	62	67	68	80	84
0.5	58	57	64	70	80	88
0.7	50	52	62	71	79	96
0.9	42	42	41	47	65	96

a b	0	0.1	0.3	0.5	0.7	0.9
0	89	90	91	96	99	99
0.1	84	86	89	94	98	99
0.3	78	80	82	94	97	99
0.5	74	74	76	91	97	99
0.7	72	72	75	84	98	99
0.9	58	60	63	65	83	100

Table 3.2: Empirical coverage probability (ECP) at levels 80% (up) and 95% (down)

regarding the ECPs (i.e. the uncertainty quantification), we run another set of simulations refining this particular area ($a \in [0.5, 0.7]$ and $b \in [0.3, 0.7]$), results in Table 3.5). Among the new simulations, the IS(80) and IS(95) are best at low memory parameters until $a = 0.6$ and $b = 0.55$. The CRPS does not vary much but seems to decrease as a and b increase. The highest ECP (80% level) is always attained when $a = 0.7$. However, the best ECP at level 95 is obtained for lower values of a. Eventually, considering all the other validation metrics, we choose the pair $a = 0.6, b = 0.5$ to be the memory coefficients. It confirms that even though the no-memory model achieves a better fit, it is necessary to include memories to properly address model uncertainties [58].

In [58], the authors build a hierarchical Bayesian model with eight scenarios based on memory or memoryless feature (controlled by two Hurst parameters H and K), with or without external forcings, and error terms' structures: fractional Gaussian (fGn) or autoregressive (AR). We compute all of the validation measurements for one case (scenario D in [58]

a b	0	0.1	0.3	0.5	0.7	0.9
0	0.184	0.184	0.186	0.188	0.193	0.198
0.1	0.188	0.189	0.189	0.186	0.188	0.193
0.3	0.201	0.199	0.193	0.186	0.178	0.189
0.5	0.221	0.215	0.204	0.185	0.178	0.192
0.7	0.254	0.251	0.234	0.218	0.204	0.215
0.9	0.363	0.382	0.415	0.489	0.432	0.344

a b	0	0.1	0.3	0.5	0.7	0.9
0	0.062	0.061	0.059	0.058	0.062	0.067
0.1	0.066	0.065	0.061	0.058	0.061	0.066
0.3	0.085	0.084	0.067	0.057	0.060	0.065
0.5	0.108	0.101	0.081	0.057	0.061	0.068
0.7	0.137	0.132	0.107	0.076	0.067	0.078
0.9	0.214	0.221	0.219	0.225	0.153	0.129

Table 3.3: Negative interval score (IS) at both 80% (up) and 95% (down) levels

a b	0	0.1	0.3	0.5	0.7	0.9
0	0.075	0.076	0.076	0.081	0.083	0.086
0.1	0.074	0.075	0.076	0.082	0.084	0.086
0.3	0.075	0.074	0.076	0.077	0.079	0.084
0.5	0.078	0.078	0.076	0.074	0.078	0.081
0.7	0.084	0.086	0.083	0.079	0.079	0.081
0.9	0.117	0.126	0.138	0.173	0.168	0.114

Table 3.4: Negative continuous ranked probability score (CRPS) for different memories

vs. $a = 0.6, b = 0.5$ in this work) in each model to compare the reconstructions (see Table 3.6 and Table 3.7).

In general, scenario D performs the best among other scenarios, and is also the closest one to the model in this project, which put short memories in both equations. It turns out that our model obtains more precise ECPs at both 80% and 95% levels, and the CRPS is less than half as much as the model in [58]. The interval scores stay quite similar between both models, and the RMSE is a bit better for [58], but RMSE is not necessarily a performance indicator since it favors the memoryless models.

a b	0.50	0.55	0.60	0.65	0.70
0.30	68	69	76	78	80
0.35	68	70	71	79	80
0.40	70	71	73	79	82
0.45	70	71	74	78	81
0.50	70	71	76	78	80
0.55	72	74	76	79	81
0.60	74	74	76	79	79
0.65	70	73	75	77	79
0.70	71	72	76	78	79

a b	0.50	0.55	0.60	0.65	0.70
0.30	94	95	96	97	97
0.35	93	94	97	97	98
0.40	94	94	97	97	97
0.45	94	94	97	98	98
0.50	91	94	95	98	97
0.55	90	93	97	98	98
0.60	88	92	97	98	98
0.65	86	88	92	97	98
0.70	84	88	90	95	98

Table 3.5: ECP at two levels 80% (up) and 95% (down) for specific calibrations

Additionally, both models appear to converge towards limit posterior distributions numerically well (PSRFs in Table 3.7), but our model seems to converge slightly better. This could be explained by the smaller variability in the Gibbs sampling process for our model, given that a and b are fixed to be 0.6 and 0.5 respectively, whereas the memory parameters (H and K) are not constants in scenario D [58].

	RMSE	ECP (80)	ECP (95)	IS (80)	IS (95)	CRPS
Scenario D	0.162	74.7	90.9	0.176	0.063	0.209
Model ($a = 0.6, b = 0.5$)	0.174	76.0	95.0	0.181	0.056	0.074

Table 3.6: Validation metrics (negative IS and CRPS) for both models

We also compare our reconstruction results with another regression method in climatology domain, namely errors in variables (EIV), which is a scaling regularization allowing for errors

	α_0	α_1	β_0	β_S	β_V	β_C	σ_P^2	σ_T^2
Scenario D	1.01	1.00	1.05	1.07	1.00	1.03	1.09	1.00
Model ($a = 0.6, b = 0.5$)	1.00	1.01	1.01	1.01	1.00	1.01	1.00	1.01

Table 3.7: The diagnosis of MCMC's convergence (PSRF) for both models

in both explanatory and response variables [41,100]. The comparison may not be perfect, since our model also incorporates the reconstructions of the three climate forcings, which is proved to outperform those reconstitutions that only involving proxies [58,101].

From the curves (displayed in Figure 3.3), we can see the trend is lower on our side as opposed to [41], implying a more radical temperature in the past millennium, which confirms that forcings' incorporation helps produce a cooler reconstruction [58]. The local variations are rather similar, thanks to the common proxy data used in both methods. The gap between the two curves is more distinct during the period of notably low solar activity (the Maunder Minimum, see [102] for more details), which can be easily detected in the solar forcing series (S_t), because sunspot number is strongly correlated to the TSI [103,104].

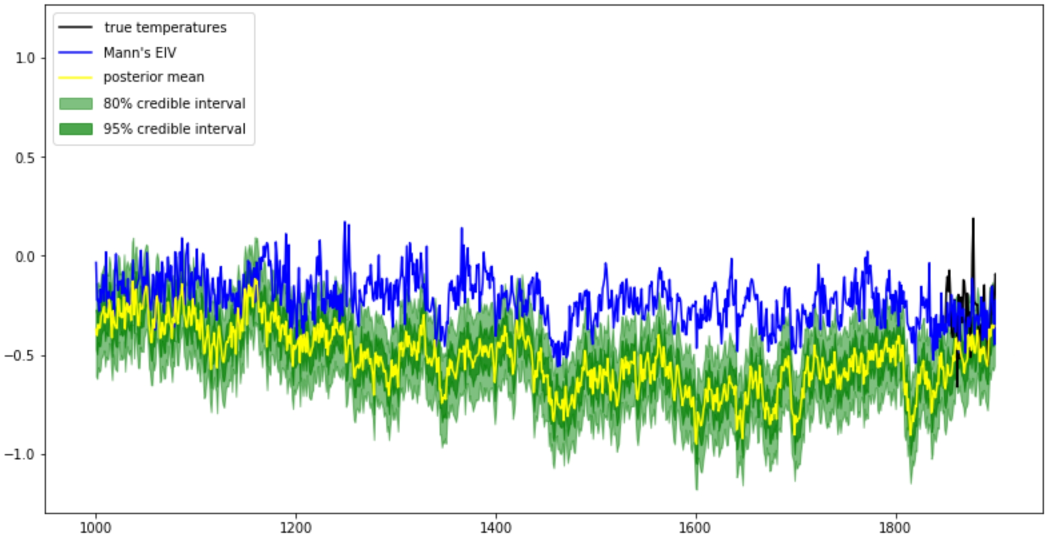

Figure 3.3: Temperature anomalies' reconstructions between our model and EIV

Furthermore, the memory included in our model takes into account the influence of forcings of any given year on the following year. It is especially true for volcanism where eruptions may yield cooling periods, but whose spikes only last around two years. In [105], the authors show that tree rings underestimate the cooling effect of large volcanic events, which means that tree-ring based reconstructions (particularly proxy-only) may not properly address volcanic eruptions. In this study, the proxy is an aggregation of multiple proxy time series, less than half of which are tree rings. Hence, our reconstruction compensates this phenomenon by adding moderate memories in the model.

3.4.3 Projection (2019-2100)

As part of the CMIP5, many institutes publish the result of climate simulations for this century using different models and data [106]. For the four RCPs from 2006 to 2100, we compute the means of the surface temperature on the northern hemisphere from different models (listed in Table 3.8), followed by converting them to temperature anomalies. Although the models and data exhibit difference, the sample paths of the various simulations have almost identical trends (see Figure 3.4 for an example). Thus, for each RCP, we consider the temperature projection is equal to the mean of all model simulations, which allowing us to compare our results with this multi-model average method (Figure 3.5 and Figure 3.6).

Except RCP8.5, the temperature's posterior means in our model exactly fall into the 95% confidence intervals of CMIP5 projections. For RCP4.5 and RCP6.0, we can also see that the 95% credible intervals cover CMIP5 means very well, especially for RCP6.0, the temperature's posterior means are almost the same as the CMIP5 projections. We cannot explain why both our model and CMIP5 have a jump at the beginning, however, we do not intend to model the high-frequency fluctuations from year to year. In the initial 13-year (2006-2018) projections, we note that the discrepancy between the NH average temperature's

Modeling center or group (**Location**)	**Model** (acronym)
AORI, NIES, JAMSTEC (Japan)	MIROC-ESM-CHEM MIROC-ESM MIROC5
Canadian Centre for Climate Modelling and Analysis (Canada)	CanESM2
Centre national de recherche météorologique (France)	CNRM-CM5
Commonwealth Scientific and Industrial Research Organisation (Australia)	ACCESS1.0 ACCESS1.3 CSIRO-Mk3.6.0 CSIRO-Mk3L
EC-Earth (Europe)	EC-EARTH
Institute of Atmospheric Physics, Chinese Academy of Sciences (China)	FGOALS-s2
Institut Pierre Simon Laplace (France)	IPSL-CM5A-LR IPSL-CM5A-MR IPSL-CM5B-LR
Max Planck Institute for Meteorology (Germany)	MPI-ESM-LR MPI-ESM-MR
Meteorological Research Institute (Japan)	MRI-CGCM3 MRI-ESM1
Met Office Hadley Centre for Climate Science and Services (UK)	HadGEM2-AO
NASA Goddard Institute for Space Studies (USA)	GISS-E2-H-CC GISS-E2-H
Norwegian Climate Centre (Norway)	NorESM1-ME NorESM1-M

Table 3.8: List of CMIP5 models used for multi-averaging temperature calculations

annual fluctuations and CMIP5, our projection is smoother with greater magnitude than this initial jump, which could be viewed as part of the fact that the feature of yearly fluctuations are missing from our model and from CMIP5.

In terms of the model robustness, not only do we add various choices of σ_C to expand greenhouse gas concentration forcings (Appendix A), but we compute the differences over two separate projection periods (2006-2100 and 2019-2100) using the same model (Table 3.9). The trends in the graphs are relatively similar to each other (the progression results can be found in the supplemental materials), and the values in Table 3.9 are rather minimal compared with the projected temperature anomalies in the corresponding RCP scenarios. Therefore, both methods justify that our model is quite robust and the uncertainty

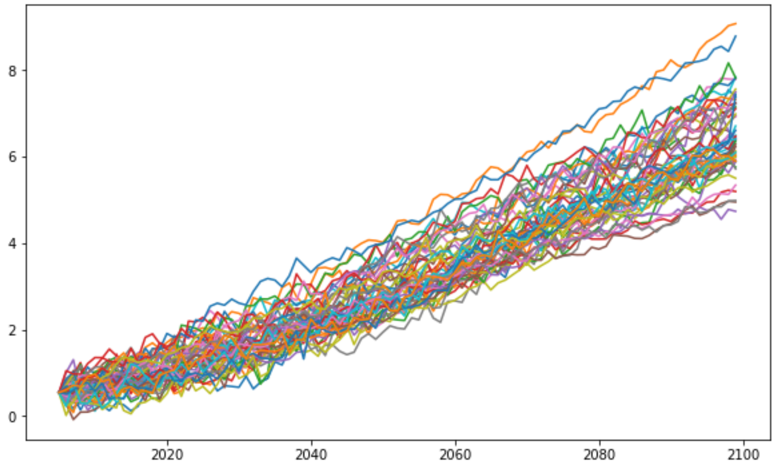

Figure 3.4: Temperature prediction for RCP8.5 between 2006 and 2100 (CMIP5)

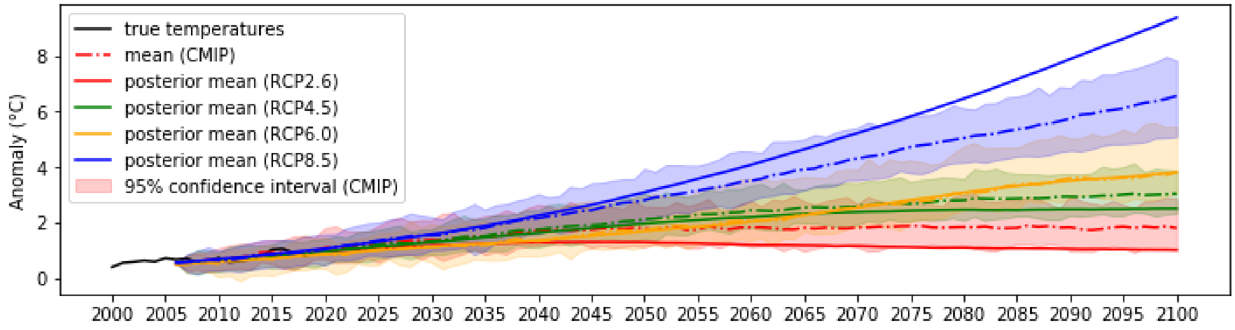

Figure 3.5: Temperature projections until 2100 for all four RCPs: comparisons with CMIP5 multi-model average

quantification is reasonable, which is not allowed to be evaluated by the model averaging methodology of CMIP5.

Besides, following [35], we extend and test all memory combinations-white noise (WN, no memory), AR(1) and AR(2) (short-term memory), and fGn (long-term memory) for the purpose of projection in [58]. The outcomes are shown in Table 3.10 below.

In similar to the reconstruction results, the best one is the no-memory model (WN-WN) except the ECPs, where the AR(1)-fGn achieves better at both levels. Models with stronger memory yield worse RMSE, since memoryless model corresponds to the classical framework of linear regression, which directly minimizes the mean squared error. The scoring measures

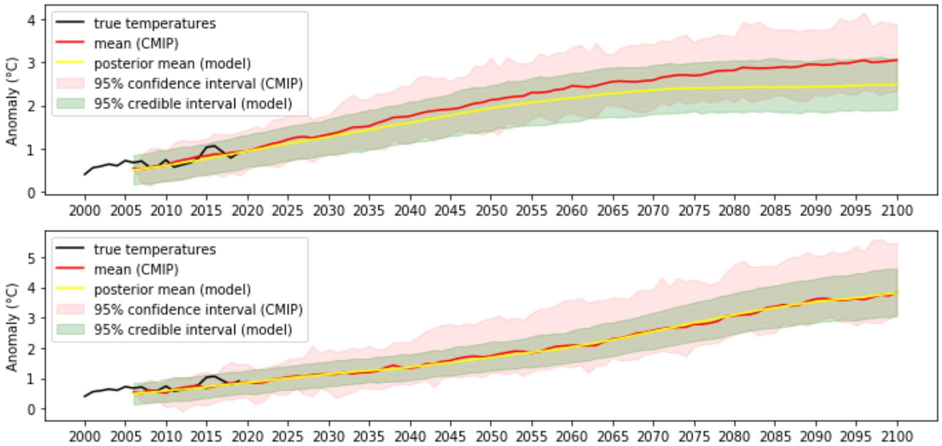

Figure 3.6: Temperature projections' comparisons with CMIP5 for medium scenarios (up: RCP4.5; down: RCP6.0)

Year	RCP2.6	RCP4.5	RCP6.0	RCP8.5
2020	0.039	0.026	0.061	0.196
2030	0.044	0.032	0.076	0.284
2040	0.041	0.029	0.104	0.400
2050	0.052	0.031	0.120	0.540
2060	0.034	0.025	0.127	0.717
2070	0.066	0.031	0.168	0.914
2080	0.050	0.048	0.214	1.134
2090	0.045	0.046	0.239	1.354
2100	0.042	0.039	0.244	1.627
Average	0.045	0.033	0.148	0.775
Maximum	0.066	0.052	0.266	1.627

Table 3.9: Temperature differences (in absolute values) between two projection periods

(IS and CRPS) do not have explicit pattern. Scenario AR(1)-AR(1) model attains decent validation metrics, and beats WN-WN model with respect to the ECPs, which assures to incorporate memories to more suitably quantify uncertainties.

For the coefficients part, the values of α_1 are much larger for the models with a white noise in the proxy equation, because the noise is anticipated to take a part of the signal. Among the forcings, the greenhouse gas concentration (β_C) is always the highest with low standard deviation, which identifies its statistical significance. This is also reasonable because it explains the temperature's increase over the last century [72, 100]. Solar coefficient (β_S) is

H	K	α_0	α_1	β_0	β_S	β_V	β_C
WN	WN	-0.015 (0.008)	0.684 (0.008)	-0.474 (0.018)	0.026 (0.018)	0.010 (0.018)	0.127 (0.018)
	AR(1)	-0.029 (0.005)	0.573 (0.005)	-0.538 (0.038)	0.046 (0.038)	-0.008 (0.038)	0.153 (0.038)
	AR(2)	-0.028 (0.005)	0.587 (0.005)	-0.526 (0.045)	0.041 (0.045)	-0.008 (0.045)	0.158 (0.045)
	fGn	-0.024 (0.005)	0.629 (0.005)	-0.444 (0.210)	0.007 (0.210)	-0.005 (0.210)	0.151 (0.210)
AR(1)	WN	-0.234 (0.034)	0.230 (0.034)	-0.053 (0.986)	-0.012 (0.986)	-0.413 (0.986)	0.895 (0.986)
	AR(1)	-0.188 (0.028)	0.266 (0.028)	-0.550 (0.047)	0.084 (0.047)	-0.017 (0.047)	0.113 (0.047)
	AR(2)	-0.076 (0.022)	0.253 (0.022)	-1.006 (0.180)	0.071 (0.180)	-0.019 (0.180)	0.278 (0.180)
	fGn	-0.171 (0.034)	0.285 (0.034)	-0.574 (0.086)	0.071 (0.086)	-0.018 (0.086)	0.130 (0.086)
AR(2)	WN	-0.204 (0.040)	0.224 (0.040)	-0.543 (0.043)	0.089 (0.043)	-0.018 (0.043)	0.107 (0.043)
	AR(1)	-0.179 (0.055)	0.268 (0.055)	-0.535 (0.055)	0.077 (0.055)	-0.017 (0.055)	0.113 (0.055)
	AR(2)	-0.185 (0.044)	0.257 (0.044)	-0.550 (0.047)	0.083 (0.047)	-0.018 (0.047)	0.114 (0.047)
	fGn	-0.170 (0.062)	0.282 (0.062)	-0.531 (0.122)	0.048 (0.122)	-0.017 (0.122)	0.132 (0.122)
fGn	WN	-0.167 (0.313)	0.190 (0.313)	-0.542 (0.044)	0.078 (0.044)	-0.015 (0.044)	0.113 (0.044)
	AR(1)	-0.153 (0.159)	0.202 (0.159)	-0.648 (0.155)	0.065 (0.155)	-0.020 (0.155)	0.159 (0.155)
	AR(2)	-0.081 (0.103)	0.214 (0.103)	-1.086 (0.249)	0.075 (0.249)	-0.022 (0.249)	0.291 (0.249)
	fGn	-0.133 (0.158)	0.247 (0.158)	-0.618 (0.142)	0.048 (0.142)	-0.018 (0.142)	0.154 (0.142)

H	K	RMSE	ECP (80)	ECP (95)	IS (80)	IS (95)	CRPS
WN	WN	0.126	0.72	0.65	0.126	0.036	0.056
	AR(1)	0.118	0.50	0.63	0.224	0.122	0.074
	AR(2)	0.115	0.52	0.65	0.213	0.114	0.071
	fGn	0.106	0.46	0.68	0.185	0.090	0.064
AR(1)	WN	0.169	0.63	0.90	0.204	0.072	0.082
	AR(1)	0.158	0.69	0.92	0.172	0.053	0.071
	AR(2)	0.297	0.64	0.90	0.347	0.110	0.138
	fGn	0.167	0.76	0.95	0.169	0.050	0.073
AR(2)	WN	0.157	0.70	0.92	0.170	0.055	0.070
	AR(1)	0.156	0.82	0.97	0.147	0.049	0.064
	AR(2)	0.153	0.76	0.97	0.147	0.048	0.064
	fGn	0.172	0.88	0.99	0.152	0.052	0.067
fGn	WN	0.159	0.73	0.92	0.167	0.049	0.070
	AR(1)	0.203	0.90	0.99	0.176	0.065	0.095
	AR(2)	0.310	0.82	0.92	0.319	0.106	0.012
	fGn	0.178	0.83	0.99	0.161	0.054	0.070

Table 3.10: Gibbs sampler convergence for the posterior means and (standard deviations) of model coefficients in Scenario D as well as validation metrics for each memory combination

the same order of magnitude even though it is smaller than β_C, which is in accordance with the solar activity has a non-negligible influence on the climate before greenhouse gas explodes [71]. The volcanic coefficient (β_V) is always negative with probability greater than 99% in most models. Considering we apply a decreasing transformation on volcanism $\log(-V_t + 1)$, which means that the volcanic activity is expected to slightly warm the planet. However, the volcanic eruptions are known to produce an overall global cooling effect [105]. It cannot have any impact beyond five or six years since eruptions are represented as short-memory spikes (in V_t). Therefore, long-period volcanic activity is more appropriately addressed by autoregression proposed in our study.

In sum, increasing the memory parameters would decrease all the coefficients except for volcanism. It is coherent because increasing memory grants more weights to the past climate forcing factors and temperature in the model. The stability of the volcanic coefficient indicates that stronger memories help taking the impact of past eruptions over the following years into consideration.

3.5 Conclusions and Discussions

In the study, we build a multi-level stochastic model to reconstruct NH temperature anomalies over the past millennium and to make projections for this century, by incorporating natural climate forcings (solar irradiance, volcanism, greenhouse gases) with memories in both levels. Bayesian inference is adopted to systematically estimate the magnitude of all unknown quantities as well as model uncertainties.

Moderate memory is suggested by the validation metrics in comparisons from both reconstructions and projections. The memory inclusion in our model (autoregression) not only allows for the influence of external forcings at any given year on the years after, but also strengthens the overall decreasing trend appeared in reconstructions (Figure 3.3). It generates lower temperatures (especially before Year 1900) compared to [41], noticing the model and data are different though. The projection deviation between this paper and CMIP5 is proportional to the greenhouse gas growth amplitude, which may be a consequence of the solar activity prediction took from the CMIP6, where the solar decay over the next century is a novelty that CMIP5 did not take into account [38]. We believe that the solar activity's decay is more impactful on the temperature when the greenhouse gas concentration is smaller, which explains the evolution of the gap.

There are a few possible extensions of this paper. One is to simultaneously evaluate memory parameters (a and b) with other prior distributions involving model coefficients (α and β), which most likely would cause higher computational demand. Another possibility is to include spatiotemporal component by designing a spatial pattern for proxies over time [59], providing more smooth reconstructions [107]. This implementation requires more comprehensive understandings of climatic system to propose a more proper and feasible temporal covariance structure for the model, which probably might simplify computations and create further scientific insights.

Chapter 4

Predictive-distribution Approximation via Wiener Chaos

4.1 Preliminaries

Assume we have the simplest linear regression model:

$$Y_i = \sum_{j=1}^{k} X_{ij}\beta_j + \sigma\varepsilon_i \tag{4.1}$$

where $i = 1, 2, ..., n$, $j = 1, 2, ..., k$, and $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Let $\mathbf{X}$, $\mathbf{Y}$, $\boldsymbol{\beta}$, $\boldsymbol{\varepsilon}$ denote X_{ij}, Y_i, β_j, ε_i respectively in matrix forms. The model (4.1) is rewritten as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon} \tag{4.2}$$

The log-likelihood function of the parameters $(\boldsymbol{\beta}, \sigma^2)$ up to a constant term is:

$$l(\boldsymbol{\beta}, \sigma^2) = -nln(\sigma) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^2 \tag{4.3}$$

Then the maximum likelihood estimator (MLE) of ($\boldsymbol{\beta}$, σ^2) can be derived from (4.3) by letting $\frac{\partial l(\boldsymbol{\beta},\sigma^2)}{\partial \boldsymbol{\beta}} = 0$ and $\frac{\partial l(\boldsymbol{\beta},\sigma^2)}{\partial \sigma^2} = 0$. Thus, the MLEs of the parameters are

$$\begin{cases} \hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y} \\ \widehat{\sigma^2} = \dfrac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^2 \end{cases} \tag{4.4}$$

$\hat{\boldsymbol{\beta}}$ is known as the ordinary linear squared (OLS) estimator, which is unbiased for $\boldsymbol{\beta}$ (i.e., $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$). However, $\widehat{\sigma^2}$ is biased for σ^2. An unbiased estimator of σ^2 (i.e., $\mathbb{E}(s^2) = \sigma^2$) is

$$s^2 = \frac{\mathbf{Y}^\mathsf{T}\mathbf{Y} - \hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{Y}}{n - (k+1)} \tag{4.5}$$

where $\hat{\boldsymbol{\beta}}$ is independent from s^2 and

$$\frac{(n-(k+1))s^2}{\sigma^2} \sim \chi^2_{n-(k+1)} \tag{4.6}$$

The variance-covariance matrices of ($\hat{\boldsymbol{\beta}}$, s^2) are

$$\begin{cases} \mathbb{C}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbb{V}(\mathbf{Y})((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})^\mathsf{T} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} := \mathbf{C} \\ \mathbb{V}(s^2) = \mathbb{V}\big(\dfrac{\sigma^2}{n-(k+1)}\chi^2_{n-(k+1)}\big) = \dfrac{\sigma^4}{(n-(k+1))^2}\mathbb{V}\big(\chi^2_{n-(k+1)}\big) = \dfrac{2\sigma^4}{n-(k+1)} \end{cases} \tag{4.7}$$

4.2 Approximation Error

4.2.1 Wiener-chaos Expansion

Intuitively, s^2 is an unbiased estimator of σ^2, so we decide to use s^2 to replace σ^2 in $\mathbb{C}(\hat{\boldsymbol{\beta}})$ and in $\mathbb{V}(s^2)$. Then the error of the approximation e_{s^2} for $\mathbb{V}(s^2)$ is

$$e_{s^2} = \frac{2\sigma^4}{n-(k+1)} - \frac{2s^4}{n-(k+1)} \tag{4.8}$$

Now let $d = n-(k+1)$. (4.6) becomes

$$Z := \frac{ds^2}{\sigma^2} \sim \chi^2_d \tag{4.9}$$

where $\mathbb{E}(Z) = d$, $\mathbb{V}(Z) = 2d$. Then

$$\tilde{Z} := Z - d \sim \chi^2_d \tag{4.10}$$

is a centered chi-squared distribution (i.e., $\mathbb{E}(\tilde{Z}) = 0$) with degrees of freedom d.

In the classic Wiener space $L^2[0,1]$, a centered Gaussian family $G(\psi)$, $\psi \in L^2[0,1]$ of random variables can be identified as the stochastic differential of a Wiener process W:

$$G(\psi) := \int_0^1 \psi(s)dW(s) \tag{4.11}$$

In order to find how big the approximation error in (4.8) is, it is convenient to represent

$\tilde{Z}$ in terms of Wiener integrals for computational sake, that is

$$\tilde{Z} = \frac{ds^2}{\sigma^2} - d \overset{\mathcal{D}}{=} \sum_{i=1}^{d} (G_i^2 - 1) \tag{4.12}$$

where

$$G_i = \int_0^1 \varepsilon_i(s) dW(s) := W(\varepsilon_i) = I_1^W(\varepsilon_i) \tag{4.13}$$

and $\varepsilon_i, i \geq 1$ are orthonormal family existed on $L^2[0,1]$. By contraction rule,

$$W^2(\varepsilon_i) - 1 = I_2^W(\varepsilon_i^{\otimes 2}) \tag{4.14}$$

Hence, applying product formula and (4.14) to $\tilde{Z}^2$, we have

$$\begin{aligned}
\tilde{Z}^2 - \mathbb{E}(\tilde{Z}^2) &\overset{\mathcal{D}}{=} \sum_{i=1}^{d}\sum_{j=1}^{d}(G_i^2 - 1)(G_j^2 - 1) = \sum_{i=1}^{d}\sum_{j=1}^{d}\big(W^2(\varepsilon_i) - 1\big)\big(W^2(\varepsilon_j) - 1\big) \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} I_2^W(\varepsilon_i^{\otimes 2}) I_2^W(\varepsilon_j^{\otimes 2}) = \frac{1}{4}\sum_{i=1}^{d}\sum_{j=1}^{d}\left[I_2^W(\varepsilon_i^{\otimes 2} + \varepsilon_j^{\otimes 2})^2 - I_2^W(\varepsilon_i^{\otimes 2} - \varepsilon_j^{\otimes 2})^2\right] \\
&= \frac{1}{4}\sum_{i=1}^{d}\sum_{j=1}^{d}\Big[I_4^W\big((\varepsilon_i^{\otimes 2} + \varepsilon_j^{\otimes 2}) \otimes (\varepsilon_i^{\otimes 2} + \varepsilon_j^{\otimes 2})\big) + 4I_2^W\big((\varepsilon_i^{\otimes 2} + \varepsilon_j^{\otimes 2}) \\
&\quad \otimes_1 (\varepsilon_i^{\otimes 2} + \varepsilon_j^{\otimes 2})\big)\Big] - \frac{1}{4}\sum_{i=1}^{d}\sum_{j=1}^{d}\Big[I_4^W\big((\varepsilon_i^{\otimes 2} - \varepsilon_j^{\otimes 2}) \otimes (\varepsilon_i^{\otimes 2} - \varepsilon_j^{\otimes 2})\big) \\
&\quad + 4I_2^W\big((\varepsilon_i^{\otimes 2} - \varepsilon_j^{\otimes 2}) \otimes_1 (\varepsilon_i^{\otimes 2} - \varepsilon_j^{\otimes 2})\big)\Big] \\
&= \frac{1}{4}\sum_{i=1}^{d}\sum_{j=1}^{d}\left[4I_4^W(\varepsilon_i^{\otimes 2} \otimes \varepsilon_j^{\otimes 2}) + 16I_2^W(\varepsilon_i^{\otimes 2} \otimes_1 \varepsilon_j^{\otimes 2})\right] \\
&= I_4^W\left(\sum_{i=1}^{d}\sum_{j=1}^{d} \varepsilon_i \otimes \varepsilon_i \otimes \varepsilon_j \otimes \varepsilon_j\right) + 4I_2^W\left(\sum_{i=1}^{d}\sum_{j=1}^{d} \varepsilon_i^{\otimes 2} \otimes_1 \varepsilon_j^{\otimes 2}\right)
\end{aligned} \tag{4.15}$$

where

$$\varepsilon_i^{\otimes 2} \otimes_1 \varepsilon_j^{\otimes 2}(s_1, s_2) = \int_0^1 \varepsilon_i(s_1)\varepsilon_i(s)\varepsilon_j(s_2)\varepsilon_j(s)ds = \begin{cases} 0 & i \neq j \\ \varepsilon_i(s_1)\varepsilon_j(s_2) = \varepsilon_i \otimes \varepsilon_i & i = j \end{cases} \tag{4.16}$$

Note that $\mathbb{E}(\tilde{Z}^2) = \mathbb{V}(\tilde{Z}) + \mathbb{E}^2(\tilde{Z}) = 2d$. Therefore, $\tilde{Z}^2$ becomes

$$\tilde{Z}^2 \overset{\mathcal{D}}{=} I_4^W\Big(\sum_{i=1}^{d}\sum_{j=1}^{d} \varepsilon_i \otimes \varepsilon_i \otimes \varepsilon_j \otimes \varepsilon_j\Big) + 4I_2^W\Big(\sum_{i=1}^{d} \varepsilon_i \otimes \varepsilon_i\Big) + 2d \tag{4.17}$$

4.2.2 Error Magnitude

The error (4.8) in the replacement of σ^2 by s^2 for $\mathbb{V}(s^2)$ is equal to

$$\begin{aligned} e_{s^2} &= \frac{2\sigma^4}{d} - \frac{2}{d}\Big(\frac{\sigma^2(\tilde{Z}+d)}{d}\Big)^2 = -\frac{2\sigma^4}{d^3}(\tilde{Z}^2 + 2\tilde{Z}d) \\ &\overset{\mathcal{D}}{=} -\frac{2\sigma^4}{d^3}\Big(I_4^W\Big(\sum_{i=1}^{d}\sum_{j=1}^{d} \varepsilon_i \otimes \varepsilon_i \otimes \varepsilon_j \otimes \varepsilon_j\Big) + (2d+4)I_2^W\Big(\sum_{i=1}^{d} \varepsilon_i \otimes \varepsilon_i\Big)\Big) \end{aligned} \tag{4.18}$$

with $\mathbb{E}(e_{s^2}) = 0$ because it only contains the sums of chaos terms.

For any function g and $q \geq 1$,

$$\mathbb{V}(I_q(g)) = q!\|g\|^2_{L^2([0,1]^q)} \tag{4.19}$$

Then the variance of the approximation given (4.19) is

$$\begin{aligned}\mathbb{V}(\tilde{Z}^2) &= 4!\Big\|\sum_{i=1}^{d}\sum_{j=1}^{d}\varepsilon_i\otimes\varepsilon_i\otimes\varepsilon_j\otimes\varepsilon_j\Big\|^2_{L^2([0,1]^4)} + 16\times 2!\Big\|\sum_{i=1}^{d}\varepsilon_i\otimes\varepsilon_i\Big\|^2_{L^2([0,1]^2)}\\ &= 24\sum_{i=1}^{d}\sum_{i'=1}^{d}\sum_{j=1}^{d}\sum_{j'=1}^{d}1 + 32\sum_{i=1}^{d}\sum_{i'=1}^{d}1\\ &= 24d^2+32d\end{aligned}\tag{4.20}$$

Apply the results in (4.20), the variance of e_{s^2} is

$$\begin{aligned}\mathbb{V}(e_{s^2}) &= \Big(-\frac{2\sigma^4}{d^3}\Big)^2\big(24d^2+(2d+4)^2\times 2d\big)\\ &= \frac{4\sigma^8}{d^6}\big(8d^3+56d^2+32d\big)\\ &= \frac{32\sigma^8}{d^3}+O\big(\frac{1}{d^4}\big)\end{aligned}\tag{4.21}$$

Similarly, we employ the same methodology to $\mathbf{C}$ to find its approximation error $e_{\hat{\boldsymbol{\beta}}}$.

$$\begin{aligned}e_{\hat{\boldsymbol{\beta}}} &= (\sigma^2-s^2)(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\\ &= \Big(\sigma^2-\frac{\sigma^2(\tilde{Z}+d)}{d}\Big)(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\\ &= -\frac{\sigma^2}{d}\tilde{Z}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\\ &\overset{\mathcal{D}}{=} -\frac{\sigma^2}{d}I_2^W\Big(\sum_{i=1}^{d}\varepsilon_i\otimes\varepsilon_i\Big)(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\end{aligned}\tag{4.22}$$

where $\mathbb{E}(e_{\hat{\boldsymbol{\beta}}})=0$ since it only includes the second-chaos terms. The variance of $e_{\hat{\boldsymbol{\beta}}}$ is

$$\begin{aligned}\mathbb{V}(e_{\hat{\boldsymbol{\beta}}}) &= \Big(-\frac{\sigma^2}{d}\Big)^2\times 2\times\mathbb{V}\big((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\big)\\ &= \frac{2\sigma^4}{d^2}\mathbb{V}\big((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\big)\end{aligned}\tag{4.23}$$

4.3 Prediction Evaluation

4.3.1 Explicit Boundaries

Considering prediction for a new observation X^*. The density of $\sqrt{\sigma^2}\varepsilon^*$ is

$$\begin{aligned}
law\ \sqrt{\sigma^2}\varepsilon^* &= law\left(\sqrt{\mathcal{N}\big(\sigma^2, \frac{2\sigma^4}{d}\big)} \times \varepsilon^*\right) \\
&= law\left(\sqrt{\sigma^2 + \sqrt{\frac{2\sigma^4}{d}}\eta} \times \varepsilon^*\right) \\
&\approx law\left(\sigma\Big(1 + \frac{1}{2}\sqrt{\frac{2}{d}}\eta\Big) \times \varepsilon^*\right) \\
&= law\left(\sigma\varepsilon^*\Big(1 + \frac{1}{\sqrt{2d}}\eta\Big)\right)
\end{aligned} \tag{4.24}$$

where $\eta \sim \mathcal{N}(0,1)$. ε^* and $\boldsymbol{\varepsilon}$ are independent, which is consistent with $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Based on the approximation error calculated above, replacing σ^2 with s^2 (data-base value), the approximate distribution of Y^* from (4.2) is

$$\widehat{law}\, Y^* = law\ (X^*\hat{\boldsymbol{\beta}} + \sqrt{s^2}\varepsilon^*) \tag{4.25}$$

with

$$\begin{cases} \hat{\boldsymbol{\beta}} \sim \mathcal{N}\big((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}, \widehat{\mathbf{C}}\big) \\ s^2 \sim \mathcal{N}\big(s^2, \frac{2s^4}{d}\big) \end{cases} \tag{4.26}$$

where $\widehat{\mathbf{C}} = s^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$.

Thus, the predictive distribution of Y^* in (4.25) is approximated by convolution as

$$\widehat{law}\, Y^* \approx \mathcal{N}\big(\mu_{Y^*}, \sigma^2_{Y^*}\big) \circledast law\ \Big(s\varepsilon^*\big(1 + \frac{1}{\sqrt{2d}}\eta\big)\Big) \tag{4.27}$$

with

$$\begin{cases} \mu_{Y^*} = X^*\big(\mathbf{X}^\mathsf{T}\mathbf{X}\big)^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y} \\ \sigma^2_{Y^*} = X^*\widehat{\mathbf{C}}X^{*\mathsf{T}} \end{cases} \tag{4.28}$$

Define

$$U_1 = s\varepsilon^* \sim \mathcal{N}(0, s^2),\ U_2 = 1 + \frac{1}{\sqrt{2d}}\eta \sim \mathcal{N}(1, \frac{1}{2d}) \tag{4.29}$$

U_1 and U_2 are independent by definition. Then

$$U := U_1 \times U_2 = s\varepsilon^* + \frac{s}{\sqrt{2d}}\varepsilon^*\eta \tag{4.30}$$

is a product of two independent normal random variables.

Lemma 4.3.1. *$\eta\varepsilon^*$ is a product of two independent standard normal variables, which can be represented in the second Wiener chaos as follows:*

$$\eta\varepsilon^* = \frac{1}{2}\big((\eta')^2 - (\varepsilon^{*\prime})^2\big)$$

where $\eta = \frac{1}{\sqrt{2}}(\eta' + \varepsilon^{*\prime})$ and $\varepsilon^* = \frac{1}{\sqrt{2}}(\eta' - \varepsilon^{*\prime})$. η', $\varepsilon^{*\prime} \sim \mathcal{N}(0,1)$ are independent.

The distribution of U can be seen as a non-convolution sum of two coupled (due to the common ε^*) random variables. The first piece is $\mathcal{N}(0, s^2)$, the second piece is a product normal, which can be characterized by two independent chi-square distributions (from Lemma 4.3.1), each with degree of freedom 1 (i.e., $\chi^2(1)$), and a scaling factor $\frac{s}{\sqrt{8d}}$.

Corollary 4.3.2. *When $\sigma^2 = 1$ (i.e., X is standard normal), for $\epsilon > 0$*

$$\frac{1}{2}e^{-\frac{(\epsilon+1)^2}{2}} \leq \mathbb{P}(X \geq \epsilon) \leq \frac{1}{2}e^{-\frac{\epsilon^2}{2}}$$

Proof.

$$\begin{aligned}
\mathbb{P}(X \geq \epsilon) &= \int_{\epsilon}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+\epsilon)^2}{2}} dx \\
&\geq \int_{0}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+\epsilon)^2}{2}} dx \\
&\geq 0.34 e^{-\frac{\epsilon^2+2\epsilon}{2}} \\
&\geq \frac{1}{2} e^{-\frac{(\epsilon+1)^2}{2}}
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(X \geq \epsilon) &= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+\epsilon)^2}{2}} dx \\
&\leq \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2+\epsilon^2}{2}} dx \\
&= \frac{1}{2} e^{-\frac{\epsilon^2}{2}}
\end{aligned}$$

□

Equivalently,

$$\mathbb{P}(X \geq \epsilon) \leq \frac{1}{2} e^{-\frac{\epsilon^2}{2}} \leq \mathbb{P}(X \geq \epsilon - 1) \tag{4.31}$$

showing the upper bound is between the tail probabilities within one standard deviation.

Hence, the cumulative distribution function (CDF) of U, namely $\Phi_U(u) = \mathbb{P}(U \leq u), u > 0$ is computed as

$$\begin{aligned}
\Phi_U(u) = 1 - \mathbb{P}(U \geq u) &= 1 - \mathbb{E}_\eta \Big[\mathbb{P}_{\varepsilon^*}\big(s\varepsilon^* + \frac{s}{\sqrt{2d}} \eta \varepsilon^* \geq u \big| \eta \big) \Big] \\
&= 1 - \mathbb{E}_\eta \Big[\mathbb{P}_{\varepsilon^*}\big(\varepsilon^* \geq \frac{u}{s(1+\frac{\eta}{\sqrt{2d}})} \big| \eta \big) \mathbf{1}_{\{\eta > -\sqrt{2d}\}} \\
&\quad + \mathbb{P}_{\varepsilon^*}\big(\varepsilon^* < \frac{u}{s(1+\frac{\eta}{\sqrt{2d}})} \big| \eta \big) \mathbf{1}_{\{\eta < -\sqrt{2d}\}} \Big]
\end{aligned} \tag{4.32}$$

Remark. $\eta \sim \mathcal{N}(0,1)$. *Suppose* $d > 4$,

$$\mathbb{P}_{\varepsilon^*}\big(\varepsilon^* < \frac{u}{s(1+\frac{\eta}{\sqrt{2d}})}\big|\eta\big) < \mathbb{P}_{\varepsilon^*}(\varepsilon^* < 0) < 0.5$$

$$\mathbb{P}(\eta < -\sqrt{2d}) < \mathbb{P}(\eta < -3) < 0.0015$$

which makes it negligible compared to the first term inside $\mathbb{E}_\eta$.

Based on Corollary 4.3.2, the bounds of $\Phi_U(u)$ are

$$\Phi_1(u,d) \le \Phi_U(u) \le \Phi_2(u,d) \tag{4.33}$$

with

$$\begin{cases} \Phi_1(u,d) = 1 - \mathbb{E}_\eta\Big[\frac{1}{2}e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{\eta}{\sqrt{2d}})^2}}\Big] = 1 - \frac{1}{2}\int_{-\sqrt{2d}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{x}{\sqrt{2d}})^2}}dx \\ \Phi_2(u,d) = 1 - \mathbb{E}_\eta\Big[\frac{1}{2}e^{-\frac{[u+s(1+\frac{\eta}{\sqrt{2d}})]^2}{2s^2(1+\frac{\eta}{\sqrt{2d}})^2}}\Big] = 1 - \frac{1}{2}\int_{-\sqrt{2d}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}e^{-\frac{[u+s(1+\frac{x}{\sqrt{2d}})]^2}{2s^2(1+\frac{x}{\sqrt{2d}})^2}}dx \end{cases} \tag{4.34}$$

Theorem 4.3.3. $s = \sqrt{\frac{\mathbf{Y}^\top\mathbf{Y} - \hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{Y}}{d}}$ *defined in* (4.5).

$$1 - \frac{1}{2}e^{-\frac{u^2}{2s^2}} \le \lim_{d\to\infty}\Phi_U(u) \le 1 - \frac{1}{2}e^{-\frac{(u+s)^2}{2s^2}}$$

Proof. $\Phi_1(u,d)$: For each $x \in (-\sqrt{2d}, \infty)$,

$$\lim_{d\to\infty} e^{-\frac{x^2}{2}}e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{x}{\sqrt{2d}})^2}} = e^{-\frac{x^2}{2}}e^{-\frac{u^2}{2s^2}}$$

Also,

$$\left|e^{-\frac{x^2}{2}} e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{x}{\sqrt{2d}})^2}}\right| \leq e^{-\frac{x^2}{2}}, \quad \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} = \sqrt{2\pi} < \infty$$

By Dominated Convergence Theorem,

$$\lim_{d\to\infty} \Phi_1(u,d) = 1 - \frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \lim_{d\to\infty} e^{-\frac{x^2}{2}} e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{x}{\sqrt{2d}})^2}} dx = 1 - \frac{1}{2}e^{-\frac{u^2}{2s^2}}$$

In similar, for $\Phi_2(u,d)$, using Dominated Convergence Theorem,

$$\lim_{d\to\infty} e^{-\frac{x^2}{2}} e^{-\frac{[u+s(1+\frac{x}{\sqrt{2d}})]^2}{2s^2(1+\frac{x}{\sqrt{2d}})^2}} = e^{-\frac{x^2}{2}} e^{-\frac{(u+s)^2}{2s^2}}, \quad \left|e^{-\frac{x^2}{2}} e^{-\frac{[u+s(1+\frac{x}{\sqrt{2d}})]^2}{2s^2(1+\frac{x}{\sqrt{2d}})^2}}\right| \leq e^{-\frac{x^2}{2}}$$

$$\Rightarrow \lim_{d\to\infty} \Phi_2(u,d) = 1 - \frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \lim_{d\to\infty} e^{-\frac{x^2}{2}} e^{-\frac{[u+s(1+\frac{x}{\sqrt{2d}})]^2}{2s^2(1+\frac{x}{\sqrt{2d}})^2}} dx = 1 - \frac{1}{2}e^{-\frac{(u+s)^2}{2s^2}}$$

Therefore, when $d \longrightarrow \infty$,

$$1 - \frac{1}{2}e^{-\frac{u^2}{2s^2}} = \lim_{d\to\infty} \Phi_1(u,d) \leq \lim_{d\to\infty} \Phi_U(u) \leq \lim_{d\to\infty} \Phi_2(u,d) = 1 - \frac{1}{2}e^{-\frac{(u+s)^2}{2s^2}}$$

□

Lemma 4.3.4. *Let $X \sim \mathcal{N}(\mu,\sigma^2)$ be a Gaussian random variable. For any non-negative integer m, the m-th moment of X is*

$$\mathbb{E}(X^m) = \begin{cases} \mu\sigma^{m-1}2^{\frac{m+1}{2}}\frac{\Gamma(\frac{m}{2}+1)}{\sqrt{\pi}}M(\frac{1-m}{2},\frac{3}{2},-\frac{\mu}{2\sigma^2}), & m \text{ is odd} \\ \sigma^m 2^{\frac{m}{2}}\frac{\Gamma(\frac{m+1}{2})}{\sqrt{\pi}}M(-\frac{m}{2},\frac{1}{2},-\frac{\mu}{2\sigma^2}), & m \text{ is even} \end{cases} \tag{4.35}$$

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$$

is the gamma function and

$$M(a, b, z) = \sum_{n=0}^{\infty} \frac{a(a+1)\cdots(a+n-1)z^n}{b(b+1)\cdots(b+n-1)n!}$$

is the Kummer's confluent hypergeometric function. In particular, if $\mu = 0$,

$$\begin{aligned}\mathbb{E}(X^m) &= \int_{-\infty}^{\infty} x^m \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \sigma^m (m-1)!!\end{aligned} \tag{4.36}$$

when m *is even.* !! *denotes double factorial, all odd moments are* 0.

Theorem 4.3.5. $s = \sqrt{\frac{\mathbf{Y}^\mathsf{T}\mathbf{Y} - \hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{Y}}{d}}$ *(same as in Theorem 4.3.3). For any* $u > 0$,

$$\left| \Phi_U(u) - \left(1 - \frac{1}{2} e^{-\frac{(u+s)^2}{2s^2}} - \frac{u(u+s)(2-e^{-d})}{4s^2\sqrt{\pi d}} e^{-\frac{(u+s)^2}{2s^2}}\right) \right| \leq \delta_2(u, d)$$

where

$$\delta_2(u, d) = \frac{u^2\left[3 + (\frac{u}{s} + 1)^2\right]}{8s^2 d} e^{-\frac{(u+s)^2}{2s^2}}$$

The proof uses second-order Taylor series expansion on $e^{-\frac{(u+s)^2}{2s^2}}$ and Lemma 4.3.4 (see Appendix D for further reference). Finally, Y^* can be divided as the following:

$$\begin{aligned} Y^* &\approx X^* \hat{\boldsymbol{\beta}} + \sqrt{s^2}\varepsilon^* \\ &= \mu_{Y^*} + \sigma_{Y^*}\varepsilon' + s\varepsilon^*\big(1 + \frac{1}{\sqrt{2d}}\eta\big) \\ &= \mu_{Y^*} + \big(\sigma_{Y^*}\varepsilon' + s\varepsilon^*\big) + \frac{s}{\sqrt{2d}}\eta\varepsilon^* \end{aligned} \tag{4.37}$$

where $\varepsilon' \sim \mathcal{N}(0,1)$ is independent from ε^* and η. The CDF of

$$V := \sigma_{Y^*}\varepsilon' + U \tag{4.38}$$

is $\Phi_V(v) = \mathbb{P}(V \leq v), v > 0$ derived as

$$\Phi_V(v) = \mathbb{P}(\sigma_{Y^*}\varepsilon' + U \leq v) = \mathbb{E}_{\varepsilon'}\Big[\mathbb{P}(U \leq v - \sigma_{Y^*}\varepsilon')\Big] = \mathbb{E}_{\varepsilon'}\Big[\Phi_U\big(v - \sigma_{Y^*}\varepsilon'\big)\Big]$$

Thus, according to Theorem 4.3.5, the bound of $\Phi_V(v)$ is

$$\begin{aligned}
\Phi_V(v) &\leq \mathbb{E}_x\Big[1 - \frac{1}{2}e^{-\frac{(v - x\sigma_{Y^*} + s)^2}{2s^2}} + \delta_2'(v - x\sigma_{Y^*}, d)\Big] \\
&= 1 - \frac{1}{2}\int_{-\infty}^{\infty} e^{-\frac{(v+s-x\sigma_{Y^*})^2}{2s^2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx + \mathbb{E}_x\big[\delta_2'(v - x\sigma_{Y^*}, d)\big] \\
&= 1 - \frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\big[\big(1+\frac{\sigma_{Y^*}^2}{s^2}\big)x^2 - 2\frac{(v+s)\sigma_{Y^*}}{s^2}x + \frac{(v+s)^2}{s^2}\big]}\,dx + \mathbb{E}_x\big[\delta_2'(v - x\sigma_{Y^*}, d)\big] \\
&= 1 - \frac{e^{-\frac{(v+s)^2}{2(s^2+\sigma_{Y^*}^2)}}}{2\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\big[\frac{s^2+\sigma_{Y^*}^2}{s^2}\big(x - \frac{(v+s)\sigma_{Y^*}}{s^2+\sigma_{Y^*}^2}\big)^2\big]}\,dx + \mathbb{E}_x\big[\delta_2'(v - x\sigma_{Y^*}, d)\big] \\
&= 1 - \frac{se^{-\frac{(v+s)^2}{2(s^2+\sigma_{Y^*}^2)}}}{2\sqrt{s^2 + \sigma_{Y^*}^2}} + \mathbb{E}_x\big[\delta_2'(v - x\sigma_{Y^*}, d)\big]
\end{aligned} \tag{4.39}$$

given

$$\delta_2'(u, d) = \delta_2(u, d) + \frac{u(u+s)(2 - e^{-d})}{4s^2\sqrt{\pi d}} e^{-\frac{(u+s)^2}{2s^2}} \tag{4.40}$$

Denote the density function of $X' \sim \mathcal{N}(\mu_{X'}, \sigma^2_{X'})$ as

$$f_{X'}(x) = \frac{1}{\sigma_{X'}\sqrt{2\pi}} e^{-\frac{(x-\mu_{X'})^2}{2\sigma^2_{X'}}}$$

$$\begin{cases} \mu_{X'} = \dfrac{(v+s)\sigma_{Y*}}{s^2+\sigma^2_{Y*}} \\ \sigma^2_{X'} = \dfrac{s^2}{s^2+\sigma^2_{Y*}} \end{cases} \tag{4.41}$$

Using Lemma 4.3.4, we have

$$\begin{aligned} \mathbb{E}_x\big[\delta_2'(v - x\sigma_{Y*}, d)\big] &= \int_{-\infty}^{\infty} \frac{(v-x\sigma_{Y*})e^{-\frac{(v-x\sigma_{Y*}+s)^2}{2s^2}}}{4s^2\sqrt{d}} \Big(\frac{(v-x\sigma_{Y*}+s)(2-e^{-d})}{\sqrt{\pi}} \\ &\quad + \frac{v-x\sigma_{Y*}}{2\sqrt{d}}\Big[3+\big(\frac{v-x\sigma_{Y*}}{s}+1\big)^2\Big]\Big)\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}dx \\ &= \frac{e^{-\frac{(v+s)^2}{2(s^2+\sigma^2_{Y*})}}}{4s\sqrt{d(s^2+\sigma^2_{Y*})}} \int_{-\infty}^{\infty} (v-x\sigma_{Y*})\Big(\frac{(v-x\sigma_{Y*}+s)(2-e^{-d})}{\sqrt{\pi}} \\ &\quad + \frac{v-x\sigma_{Y*}}{2\sqrt{d}}\Big[3+\big(\frac{v-x\sigma_{Y*}}{s}+1\big)^2\Big]\Big) f_{X'}(x)dx \\ &= \frac{e^{-\frac{(v+s)^2}{2(s^2+\sigma^2_{Y*})}}}{4s\sqrt{d(s^2+\sigma^2_{Y*})}} \Big\{ \frac{2-e^{-d}}{\sqrt{\pi}}\Big[v(v+s) - (2v+s)\sigma_{Y*}\mu_{X'} + \sigma^2_{Y*}\big(\mu^2_{X'} \\ &\quad + \sigma^2_{X'}\big)\Big] + \frac{1}{2\sqrt{d}}\Big[4\Big(v^2 - 2v\sigma_{Y*}\mu_{X'} + \sigma^2_{Y*}\big(\mu^2_{X'}+\sigma^2_{X'}\big)\Big) + \frac{2}{s}\Big(v^3 \\ &\quad - 3v^2\sigma_{Y*}\mu_{X'} + 3v\sigma^2_{Y*}\big(\mu^2_{X'}+\sigma^2_{X'}\big) - \sigma^3_{Y*}\big(\mu^3_{X'} + 3\mu_{X'}\sigma^2_{X'}\big)\Big) \\ &\quad + \frac{1}{s^2}\Big(v^4 - 4v^3\sigma_{Y*}\mu_{X'} + 6v^2\sigma^2_{Y*}\big(\mu^2_{X'}+\sigma^2_{X'}\big) - 4v\sigma^3_{Y*}\big(\mu^3_{X'} \\ &\quad + 3\mu_{X'}\sigma^2_{X'}\big) + \sigma^4_{Y*}\big(\mu^4_{X'} + 6\mu^2_{X'}\sigma^2_{X'} + 3\sigma^4_{X'}\big)\Big)\Big]\Big\} \end{aligned} \tag{4.42}$$

Hence,

$$\left|\Phi_V(v) - \left(1 - \frac{se^{-\frac{(v+s)^2}{2(s^2+\sigma_{Y*}^2)}}}{2\sqrt{s^2+\sigma_{Y*}^2}} - \frac{(2-e^{-d})e^{-\frac{(v+s)^2}{2(s^2+\sigma_{Y*}^2)}}}{4s\sqrt{\pi d(s^2+\sigma_{Y*}^2)}}\Big[v(v+s) - (2v+s)\sigma_{Y*}\mu_{X'} + \sigma_{Y*}^2\big(\mu_{X'}^2 + \sigma_{X'}^2\big)\Big]\right)\right| \leq \delta(v,d) \tag{4.43}$$

with a precision as

$$\delta(v,d) = \frac{e^{-\frac{(v+s)^2}{2(s^2+\sigma_{Y*}^2)}}}{8sd\sqrt{s^2+\sigma_{Y*}^2}}\Big[4\Big(v^2 - 2v\sigma_{Y*}\mu_{X'} + \sigma_{Y*}^2\big(\mu_{X'}^2+\sigma_{X'}^2\big)\Big) + \frac{2}{s}\Big(v^3 - 3v^2\sigma_{Y*}\mu_{X'} + 3v\sigma_{Y*}^2\big(\mu_{X'}^2+\sigma_{X'}^2\big) - \sigma_{Y*}^3\big(\mu_{X'}^3 + 3\mu_{X'}\sigma_{X'}^2\big)\Big) + \frac{1}{s^2}\Big(v^4 - 4v^3\sigma_{Y*}\mu_{X'} + 6v^2\sigma_{Y*}^2\big(\mu_{X'}^2+\sigma_{X'}^2\big) - 4v\sigma_{Y*}^3\big(\mu_{X'}^3 + 3\mu_{X'}\sigma_{X'}^2\big) + \sigma_{Y*}^4\big(\mu_{X'}^4 + 6\mu_{X'}^2\sigma_{X'}^2 + 3\sigma_{X'}^4\big)\Big)\Big] \tag{4.44}$$

and $\mu_{X'}, \sigma_{X'}^2$ defined in (4.41). The sharpness of the bound (4.43) is proved in Appendix E.

Let

$$g(v,d) = v(v+s) - (2v+s)\sigma_{Y*}\mu_{X'} + \sigma_{Y*}^2\big(\mu_{X'}^2 + \sigma_{X'}^2\big) \tag{4.45}$$

be the non-normal correction for $\Phi_V(v)$. For the sake of simplicity and also without loss of generality, removing e^{-d}, we have

$$e_V = 1 - \frac{e^{-\frac{(v+s)^2}{2(s^2+\sigma_{Y*}^2)}}}{2\sqrt{s^2+\sigma_{Y*}^2}}\Big(s + \frac{1}{s\sqrt{\pi d}}g(v,d)\Big) \tag{4.46}$$

Plug in (4.41) to expand $g(v, d)$.

$$\begin{aligned} g(v,d) &= \frac{s^2}{(s^2+\sigma_{Y*}^2)^2}\left[s^2v^2 + (s^3 - s\sigma_{Y*}^2)v + \sigma_{Y*}^4\right] \\ &= \frac{s^2}{(s^2+\sigma_{Y*}^2)^2}\Big[s^2\big(v + \frac{s^2-\sigma_{Y*}^2}{2s}\big)^2 - \frac{(s^2-\sigma_{Y*}^2)^2}{4} + \sigma_{Y*}^4\Big] \\ &= \frac{s^4}{(s^2+\sigma_{Y*}^2)^2}\big(v + \frac{s^2-\sigma_{Y*}^2}{2s}\big)^2 + \frac{s^2(3\sigma_{Y*}^2 - s^2)}{4(s^2+\sigma_{Y*}^2)} \end{aligned} \tag{4.47}$$

is increasing in v when $v > -\frac{s^2-\sigma_{Y*}^2}{2s}$, so is e_V.

Denote the precision term $\delta(v, d)$ in (4.44) as

$$\delta(v,d) = \frac{e^{-\frac{(v+s)^2}{2(s^2+\sigma_{Y*}^2)}}}{8sd\sqrt{s^2+\sigma_{Y*}^2}}\Big(h_1(v,d) + h_2(v,d) + h_3(v,d)\Big) \tag{4.48}$$

where

$$\begin{aligned} h_1(v,d) &= 4\Big[v^2 - 2v\sigma_{Y*}\mu_{X'} + \sigma_{Y*}^2\big(\mu_{X'}^2 + \sigma_{X'}^2\big)\Big] \\ h_2(v,d) &= \frac{2}{s}\Big[v^3 - 3v^2\sigma_{Y*}\mu_{X'} + 3v\sigma_{Y*}^2\big(\mu_{X'}^2 + \sigma_{X'}^2\big) - \sigma_{Y*}^3\big(\mu_{X'}^3 + 3\mu_{X'}\sigma_{X'}^2\big)\Big] \\ h_3(v,d) &= \frac{1}{s^2}\Big[v^4 - 4v^3\sigma_{Y*}\mu_{X'} + 6v^2\sigma_{Y*}^2\big(\mu_{X'}^2 + \sigma_{X'}^2\big) - 4v\sigma_{Y*}^3\big(\mu_{X'}^3 + 3\mu_{X'}\sigma_{X'}^2\big) \\ &\quad + \sigma_{Y*}^4\big(\mu_{X'}^4 + 6\mu_{X'}^2\sigma_{X'}^2 + 3\sigma_{X'}^4\big)\Big] \end{aligned} \tag{4.49}$$

Using the similar expansion in (4.47),

$$\begin{aligned}
h_1(v,d) &= \frac{4s^2}{(s^2+\sigma^2_{Y*})^2}(s^2v^2 - 2\sigma^2_{Y*}sv + s^2\sigma^2_{Y*} + 2\sigma^4_{Y*}) \\
&= \frac{4s^4}{(s^2+\sigma^2_{Y*})^2}\big(v - \frac{\sigma^2_{Y*}}{s}\big)^2 + \frac{4s^2\sigma^2_{Y*}}{s^2+\sigma^2_{Y*}} \\
h_2(v,d) &= \frac{2}{s(s^2+\sigma^2_{Y*})^3}\Big[\big(s^6 + s^4\sigma^2_{Y*} + 2s^2\sigma^4_{Y*} + \sigma^6_{Y*}\big)v^3 - 3s^5\sigma^2_{Y*}v^2 \\
&\quad + 3s^4\sigma^2_{Y*}\big(s^2 + 2\sigma^2_{Y*}\big)v - s^3\sigma^4_{Y*}\big(3s^2 + 4\sigma^2_{Y*}\big)\Big] \\
h_3(v,d) &= \frac{1}{s^2(s^2+\sigma^2_{Y*})^4}\Big[s^8v^4 - 4\big(s^7\sigma^2_{Y*} + 3s\sigma^8_{Y*}\big)v^3 + 6\big(s^8\sigma^2_{Y*} + 2s^6\sigma^4_{Y*}\big)v^2 \\
&\quad - 4\big(3s^7\sigma^4_{Y*} + 4s^5\sigma^6_{Y*}\big)v + s^4\sigma^4_{Y*}\big(3s^4 + 12s^2\sigma^2_{Y*} + 10\sigma^4_{Y*}\big)\Big]
\end{aligned} \tag{4.50}$$

Plug (4.50) into (4.48), $\delta(v,d)$ is

$$\begin{aligned}
\delta(v,d) &= \frac{e^{-\frac{(v+s)^2}{2(s^2+\sigma^2_{Y*})}}}{8s^3d(s^2+\sigma^2_{Y*})^{\frac{9}{2}}}\Big[s^8v^4 + 2\big(s^9 + 3s^5\sigma^4_{Y*} + 3s^3\sigma^6_{Y*} - 5s\sigma^8_{Y*}\big)v^3 \\
&\quad + 2\big(2s^{10} + 10s^8\sigma^2_{Y*} + 11s^6\sigma^4_{Y*}\big)v^2 - 2\big(s^9\sigma^2_{Y*} + 5s^7\sigma^4_{Y*} + 6s^5\sigma^6_{Y*}\big)v \\
&\quad + s^4\sigma^2_{Y*}\big(4s^6 + 13s^4\sigma^2_{Y*} + 18s^2\sigma^4_{Y*} + 10\sigma^6_{Y*}\big)\Big]
\end{aligned} \tag{4.51}$$

which shows that except for the well-expected $\frac{1}{d}$, $\delta(v,d)$ is free of d.

4.3.2 Simulation Results

Now we choose different values for s and σ^2_{Y*} under certain circumstances. Note that $\delta(v,d)$ is homogeneous degree 0 (i.e., dimension-free). Given a standardized data set, the values of v are one-standard-deviation percentile and two-standard-deviation percentile at $\alpha = 2.5\%, 16\%$. As we can see from Table 4.1, when n is large, the precision of the approximation is quite high because $\delta(v, d)$ is extremely small (roughly 0.1% and 0.5% of e_V in each respective scenario). We may lose some degree of accuracy when the significance level α is increasing. Even if n is not large enough compared to the number of explanatory variables k, our approximation is still reliably precise (later blocks in Table 4.1). The non-normal correction e_V is monotonously increasing when the magnitude of new noise σ^2_{Y*} becomes bigger. The largest percentile v appears at s^2 is comparable with σ^2_{Y*}, but overall not affected much by the the noise scale, which indicates its consistency.

4.4 Discussions

In this chapter, not only do we construct an explicit sequence of closed-form functions to approximate the true predictive distribution of the response variable, but also we provide a comprehensive analysis on the approximation based on Wiener's polynomial chaos. The approximation is with a great deal of accuracy, and the unbiased estimation convergences in a relatively-fast fashion.

The biggest advantage of this approximation methodology is to avoid numeric sampling algorithms such as Markov chain Monte Carlo, which reduces the computational burden to a certain level [108]. Also, the boundary formulations as well as the asymptotic properties yield the benefits to closely monitor the performance of predictions.

Second Wiener chaos is a linear space, and because the model is linear, its solution lies in the same chaos. In many practical situations, however, incomplete or inaccurate statistical knowledge about parameters' uncertainties limits the utility of high-order polynomial chaos expansions [109]. Fortunately, in order to create a finite-order expansion, we just need some reliable information on the probability measure that can be represented by a finite number of moments. One of the possible extensions is to explore the behavior in a nonlinear-model setting such as involving stochasticity. This may require the derivation of maximum likelihood estimators under approximated log-likelihood functions [110].

α	n	k	s^2	σ^2_{Y*}	v	e_V	$\delta(v,d)$
0.5%	1000	3	0.15	0.01	-0.3944	0.5153	0.0005
				0.05	-0.3944	0.5650	0.0006
				0.13	-0.3944	0.6310	0.0006
				0.48	-0.3934	0.7527	0.0007
				1.50	-0.3934	0.8468	0.0006
2.5%	1000	3	0.15	0.01	-0.3943	0.5153	0.0005
				0.05	-0.3945	0.5650	0.0006
				0.13	-0.3939	0.6310	0.0006
				0.48	-0.3949	0.7527	0.0007
				1.50	-0.3949	0.8468	0.0006
10%	1000	3	0.15	0.01	-0.3945	0.5153	0.0005
				0.05	-0.3949	0.5650	0.0006
				0.13	-0.3941	0.6310	0.0006
				0.48	-0.3946	0.7527	0.0007
				1.50	-0.3932	0.8468	0.0006
16%	1000	3	0.15	0.01	-0.3944	0.5153	0.0005
				0.05	-0.3945	0.5650	0.0006
				0.13	-0.3949	0.6310	0.0006
				0.48	-0.3974	0.7527	0.0007
				1.50	-0.4025	0.8468	0.0006
0.5%	200	12	1.50	0.15	-1.2763	0.5211	0.0028
				0.47	-1.2763	0.5591	0.0033
				1.52	-1.2763	0.6401	0.0034
				4.33	-1.2763	0.7386	0.0037
				12.5	-1.2450	0.8303	0.0033
2.5%	200	12	1.50	0.15	-1.2787	0.5211	0.0028
				0.47	-1.2787	0.5591	0.0033
				1.52	-1.2738	0.6402	0.0034
				4.33	-1.3031	0.7385	0.0038
				12.5	-1.3031	0.8303	0.0035
10%	200	12	1.50	0.15	-1.2783	0.5211	0.0028
				0.47	-1.2783	0.5591	0.0033
				1.52	-1.2763	0.6401	0.0034
				4.33	-1.2867	0.7386	0.0037
				12.5	-1.3063	0.8303	0.0036
16%	200	12	1.50	0.15	-1.2780	0.5211	0.0028
				0.47	-1.2709	0.5591	0.0033
				1.52	-1.2765	0.6401	0.0034
				4.33	-1.2769	0.7386	0.0037
				12.5	-1.2775	0.8303	0.0035

Table 4.1: Approximation results for various combinations of (s^2, σ^2_{Y*}) under different significance levels (α), dimension (n), and number of coefficients (k)

Chapter 5

Future Work Directions

In the agriculture study, one way to improve the analysis is to include more data points from later years. We could also remove the irrelevant explanatory variables from the model, or to develop a link among maize yield, SPAD, and striga to systematically analyze the determinants of each individual response. The latter proposal is most definitely going to generate more coefficients needed to be estimated [18].

For the climatology project, assigning more proper prior distributions to the model coefficients may be able to achieve their estimates all at once. However, it is very likely to be accompanied with more computational burden. Another possibility is to incorporate a spatial pattern for the natural proxies over time [59], which needs an extensive understandings of climatic system to propose a more feasible temporal covariance structure, thus simplifying calculations and bringing more meaningful insights.

There are several directions to extend the work in Chapter 4. For example, we can utilize Taylor-series expansion with higher order, so to obtain a more statistical power and faster convergence [111]. Moreover, we can either compare our estimates with the actual MCMC-method results, or calculate its total-variation distance between chaos terms and the normal law [112], to see how far the approximations are away from the true values.

APPENDICES

APPENDIX A

Model Robustness

In (3.4), σ_C is defined to grow with the RCPs' magnitudes for each scenario. Here, we test the model robustness by applying different fraction values (i.e., 0, 0.5, 2, 3) in σ_C. Apparently, based on the prediction graphs (Figures A.1, A.2, A.3, A.4), the patterns are quite similar to Figure 3.6 except for reasonable fluctuations when σ_C increases. The convergence diagnosis (Figure A.5) shows that the Gibbs samplers indeed converge (and quite fast). Hence, the projection model is not heavily affected by the added greenhouse gas uncertainty with various magnitudes (robust in other words).

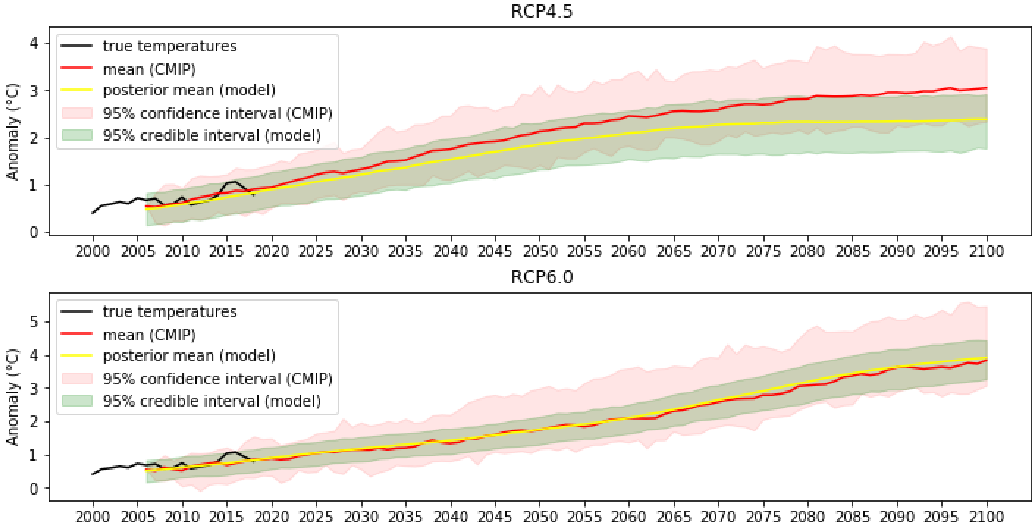

Figure A.1: Temperature projections' comparisons (CMIP5) with **no** σ_C

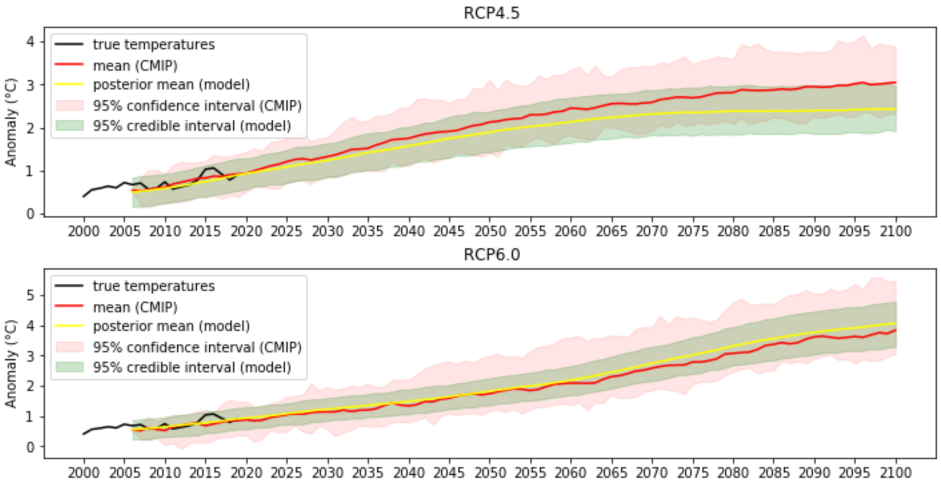

Figure A.2: Temperature projections' comparisons (CMIP5) with $\boldsymbol{\frac{1}{2}}\sigma_C$

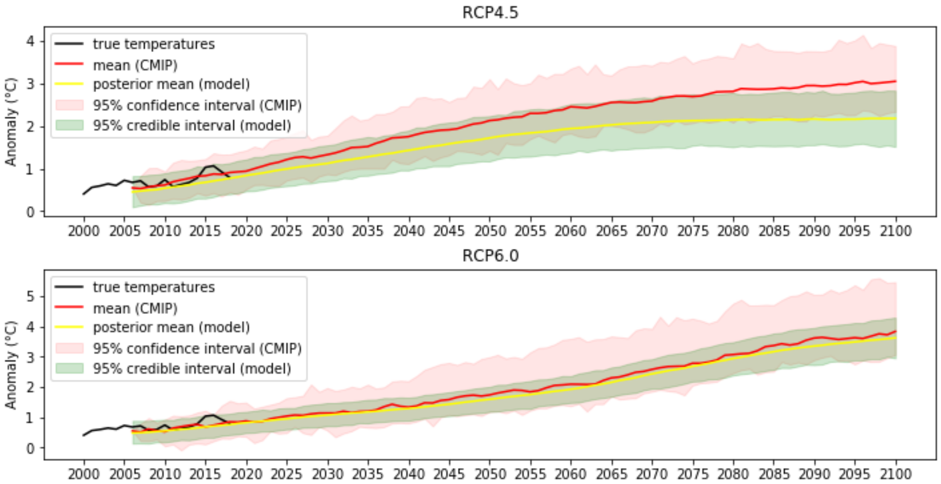

Figure A.3: Temperature projections' comparisons (CMIP5) with $\mathbf{2}\sigma_C$

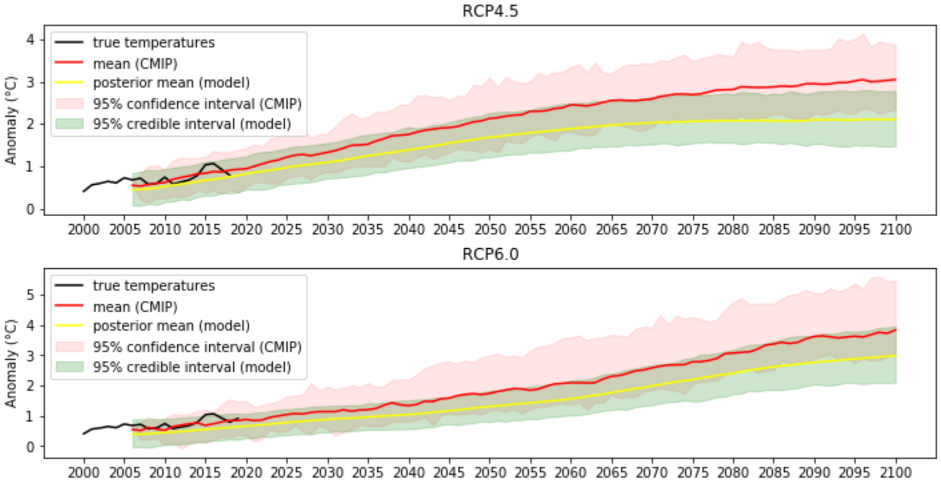

Figure A.4: Temperature projections' comparisons (CMIP5) with $\mathbf{3}\sigma_C$

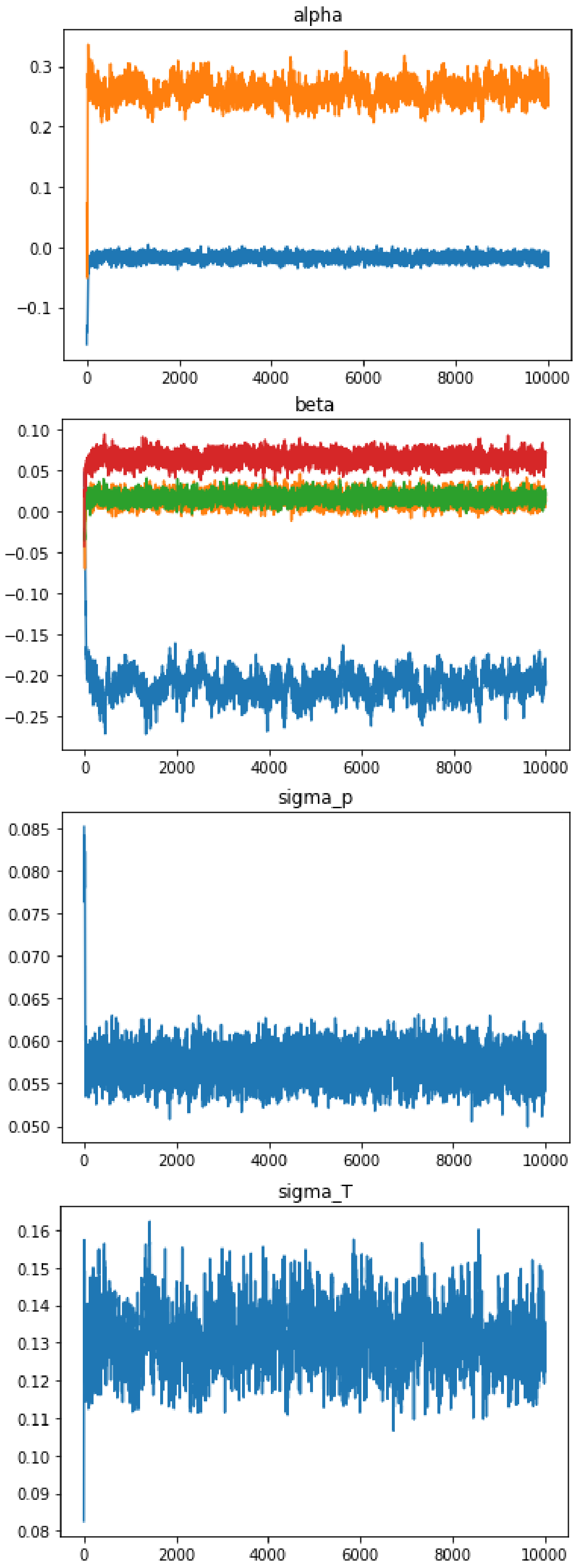

Figure A.5: 10,000 MCMC samples of all model parameters (from top to bottom): α_1, α_0, β_C, β_V, β_S, β_0, σ_P^2, σ_T^2

APPENDIX B

Exponential Decay

The hierarchical stochastic model described in (3.6) is equivalent as follows after iteration:

$$\begin{cases} P_t = a^t P_0 + \sum_{i=0}^{t-1} a^i \big(\alpha_0 + \alpha_1 T_{t-i}\big) + \sum_{i=0}^{t-1} a^i \sigma_P \varepsilon_{t-i} \\ \quad = a^t P_0 + (u\alpha)_t + \sum_{i=0}^{t-1} a^i \sigma_P \varepsilon_{t-i} \\ T_t = b^t T_0 + \sum_{i=0}^{t-1} b^i \big(\beta_0 + \beta_S S_{t-i} + \beta_V V_{t-i} + \beta_C C_{t-i}\big) + \sum_{i=0}^{t-1} b^i \sigma_T \eta_{t-i} \\ \quad = b^t T_0 + (v\beta)_t + \sum_{i=0}^{t-1} b^i \sigma_T \eta_{t-i} \end{cases} \tag{B.1}$$

Hereinafter, $u \in \mathcal{M}_{T,2}(\mathbb{R})$ such that $\forall t \leq T$:

$$u_{t,0} = \sum_{i=0}^{t-1} a^i, \;\; u_{t,1} = \sum_{i=0}^{t-1} a^i T_{t-i} \tag{B.2}$$

and $v \in \mathcal{M}_{T,4}(\mathbb{R})$ is definite and same for all three forcings ($t \leq T$) such that

$$v_{t,0} = \sum_{i=0}^{t-1} a^i, \;\; v_{t,1} = \sum_{i=0}^{t-1} a^i S_{t-i}, \;\; v_{t,2} = \sum_{i=0}^{t-1} a^i V_{t-i}, \;\; v_{t,3} = \sum_{i=0}^{t-1} a^i C_{t-i} \tag{B.3}$$

These formalizations show that proxy (**P**) and temperature (**T**) at time t, take all temperature and forcings before t into consideration respectively with exponential decay.

Moreover, the time series P_t and T_t are non-stationary, whose expectations depend on t.

When t is large enough, assume $P_0 = T_0 = 0$ and let $\sigma_P = 1 - a^2$, $\sigma_T = 1 - b^2$. From now on, we use $[\mathbf{Y}|\mathbf{X}]$ to represent the conditional probability distribution of the random variable $\mathbf{Y}$ given $\mathbf{X}$. Then the variances of $[\mathbf{P}|\mathbf{T}, a, \alpha, \sigma_P]$ and $[\mathbf{T}|b, \beta, \sigma_T]$ are approximately equal to constants σ_P and σ_T separately. Namely, the conditional densities of $\mathbf{P}$ and $\mathbf{T}$ are two normal distributions with known covariance structures:

$$\begin{cases} \big[\mathbf{P}\big|\mathbf{T}, a, \alpha, \sigma_P\big] \sim \mathcal{N}\big(u\alpha, \sigma_P^2 \Sigma_P\big) \\ \big[\mathbf{T}\big|b, \beta, \sigma_T\big] \sim \mathcal{N}\big(v\beta, \sigma_T^2 \Sigma_T\big) \end{cases} \tag{B.4}$$

where Σ_P and Σ_T are the covaraince matrices of AR(1) processes with parameters a and b.

APPENDIX C

Posterior-distribution Computation

Recall the prior distributions: normal ($\mathcal{N}$), inverse gamma ($\mathcal{IG}$), and uniform ($\mathcal{U}$) that are assigned to the parameters in the model (3.6):

$$\begin{aligned} \alpha &\sim \mathcal{N}(\mu_\alpha, I_2) \\ \beta &\sim \mathcal{N}(\mu_\beta, I_4) \\ \sigma^2 &\sim \mathcal{IG}(q, r) \\ a, b &\sim \mathcal{U}(0, 1) \end{aligned} \tag{C.1}$$

where

$$\alpha = (\alpha_0, \alpha_1),\ \beta = (\beta_0, \beta_S, \beta_C, \beta_V),\ \sigma^2 = (\sigma_P^2, \sigma_T^2)$$

$$q = (q_P, q_T),\ r = (r_P, r_T)$$

I_n is an identity matrix of n dimensions. In particular, we also assign the initial values as:

$$\begin{aligned} \mu_\alpha &= (0, 1) \\ \mu_\beta &= (0, 1, 1, 1) \\ q &= (2, 2) \\ r &= (0.1, 0.1) \end{aligned} \tag{C.2}$$

After some derivations, we obtain the full posterior distributions for all the parameters.

Regression coefficients: $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$

$$\begin{aligned}\big[\boldsymbol{\alpha}\big|\mathbf{P},\mathbf{T},a,\sigma_P^2\big] &\propto \big[\mathbf{P}\big|\boldsymbol{\alpha},\mathbf{T},a,\sigma_P^2\big]\big[\boldsymbol{\alpha}\big|\mathbf{T},a,\sigma_P^2\big]\\ &\propto \exp\left\{-\frac{1}{2\sigma_P^2}\big(\mathbf{P}-u\alpha\big)^\mathsf{T}\Sigma_P^{-1}\big(\mathbf{P}-u\alpha\big)\right\}\exp\left\{-\frac{1}{2}\|\alpha\|^2\right\}\\ &\propto \exp\left\{-\frac{1}{2\sigma_P^2}\alpha^\mathsf{T}\big(u^\mathsf{T}\Sigma_P^{-1}u+\sigma_P^2I_2\big)\alpha+\frac{1}{\sigma_P^2}\alpha^\mathsf{T}u^\mathsf{T}\Sigma_P^{-1}\mathbf{P}\right\}\\ &\propto \exp\left\{-\frac{1}{2}(\alpha-\mu_\alpha)^\mathsf{T}\Omega_\alpha^{-1}(\alpha-\mu_\alpha)\right\}\\ &\sim \mathcal{N}\big(\mu_\alpha,\Omega_\alpha\big)\end{aligned} \tag{C.3}$$

where exp means exponential distribution and

$$\begin{cases}\mu_\alpha = \dfrac{1}{\sigma_P^2}\Omega_\alpha u^\mathsf{T}\Sigma_P^{-1}\mathbf{P}\\ \Omega_\alpha^{-1} = \dfrac{1}{\sigma_P^2}u^\mathsf{T}\Sigma_P^{-1}u+I_2\end{cases} \tag{C.4}$$

$$\begin{aligned}\big[\boldsymbol{\beta}\big|\mathbf{T},b,\sigma_T^2\big] &\propto \big[\mathbf{T}|\boldsymbol{\beta},b,\sigma_T^2\big]\big[\boldsymbol{\beta}\big|b,\sigma_T^2\big]\\ &\propto \exp\left\{-\frac{1}{2\sigma_T^2}\big(\mathbf{T}-v\beta\big)^\mathsf{T}\Sigma_T^{-1}\big(\mathbf{T}-v\beta\big)\right\}\exp\left\{-\frac{1}{2}\|\beta\|^2\right\}\\ &\propto \exp\left\{-\frac{1}{2\sigma_T^2}\beta^\mathsf{T}\big(v^\mathsf{T}\Sigma_T^{-1}v+\sigma_T^2I_4\big)\beta+\frac{1}{\sigma_T^2}\beta^\mathsf{T}v^\mathsf{T}\Sigma_T^{-1}\mathbf{T}\right\}\\ &\propto \exp\left\{-\frac{1}{2}(\beta-\mu_\beta)^\mathsf{T}\Omega_\beta^{-1}(\beta-\mu_\beta)\right\}\\ &\sim \mathcal{N}\big(\mu_\beta,\Omega_\beta\big)\end{aligned} \tag{C.5}$$

with

$$\begin{cases}\mu_\beta = \dfrac{1}{\sigma_T^2}\Omega_\beta v^\mathsf{T}\Sigma_T^{-1}\mathbf{T}\\ \Omega_\beta^{-1} = \dfrac{1}{\sigma_T^2}v^\mathsf{T}\Sigma_T^{-1}v+I_4\end{cases} \tag{C.6}$$

Scale of noise terms: $\boldsymbol{\sigma^2_P}$, $\boldsymbol{\sigma^2_T}$

$$\begin{aligned}
\big[\boldsymbol{\sigma^2_P}\big|\mathbf{P},\mathbf{T},a,\alpha\big] &\propto \big[\mathbf{P}\big|\boldsymbol{\sigma^2_P},\mathbf{T},a,\alpha,\big]\big[\boldsymbol{\sigma^2_P}|\mathbf{T},a,\alpha\big] \\
&\propto \left(\frac{1}{\boldsymbol{\sigma_P}}\right)^{\dim\mathbf{T}} \exp\left\{-\frac{1}{2\boldsymbol{\sigma^2_P}}\big(\mathbf{P}-u\alpha\big)^{\mathsf{T}}\Sigma_P^{-1}\big(\mathbf{P}-u\alpha\big)\right\}\left(\frac{1}{\boldsymbol{\sigma^2_P}}\right)^{q_P+1} \\
&\quad\times\exp\left\{-\frac{r_P}{\boldsymbol{\sigma^2_P}}\right\} \\
&\propto \left(\frac{1}{\boldsymbol{\sigma^2_P}}\right)^{q_P+\frac{\dim\mathbf{T}}{2}+1} \exp\left\{-\frac{1}{\boldsymbol{\sigma^2_P}}\left[r_P+\frac{1}{2}\big(\mathbf{P}-u\alpha\big)^{\mathsf{T}}\Sigma_P^{-1}\big(\mathbf{P}-u\alpha\big)\right]\right\} \\
&\sim \mathcal{IG}\big(q'_P,r'_P\big)
\end{aligned} \tag{C.7}$$

where dim is short for dimension and

$$\begin{cases} q'_P = q_P + \dfrac{\dim\mathbf{T}}{2} \\ r'_P = r_P + \dfrac{1}{2}\big(\mathbf{P}-u\alpha\big)^{\mathsf{T}}\Sigma_P^{-1}\big(\mathbf{P}-u\alpha\big) \end{cases} \tag{C.8}$$

$$\begin{aligned}
\big[\boldsymbol{\sigma^2_T}\big|\mathbf{T},b,\beta\big] &\propto \big[\mathbf{T}\big|\boldsymbol{\sigma^2_T},b,\beta\big]\big[\boldsymbol{\sigma^2_T}|b,\beta\big] \\
&\propto \left(\frac{1}{\boldsymbol{\sigma_T}}\right)^{\dim\mathbf{T}} \exp\left\{-\frac{1}{2\boldsymbol{\sigma^2_T}}\big(\mathbf{T}-v\beta\big)^{\mathsf{T}}\Sigma_T^{-1}\big(\mathbf{T}-v\beta\big)\right\}\left(\frac{1}{\boldsymbol{\sigma^2_T}}\right)^{q_T+1} \\
&\quad\times\exp\left\{-\frac{r_T}{\boldsymbol{\sigma^2_T}}\right\} \\
&\propto \left(\frac{1}{\sigma^2_T}\right)^{q_T+\frac{\dim\mathbf{T}}{2}+1} \exp\left\{-\frac{1}{\boldsymbol{\sigma^2_T}}\left[r_T+\frac{1}{2}\big(\mathbf{T}-v\beta\big)^{\mathsf{T}}\Sigma_T^{-1}\big(\mathbf{T}-v\beta\big)\right]\right\} \\
&\sim \mathcal{IG}\big(q'_T,r'_T\big)
\end{aligned} \tag{C.9}$$

with

$$\begin{cases} q'_T = q_T + \dfrac{\dim\mathbf{T}}{2} \\ r'_T = r_T + \dfrac{1}{2}\big(\mathbf{T}-v\beta\big)^{\mathsf{T}}\Sigma_T^{-1}\big(\mathbf{T}-v\beta\big) \end{cases} \tag{C.10}$$

Autoregressive coefficients: $\boldsymbol{a}$, $\boldsymbol{b}$

$$\begin{aligned}
\left[\boldsymbol{a}\middle|\mathbf{P},\mathbf{T},\alpha,\sigma_P^2\right] &\propto \left[\mathbf{P}|\boldsymbol{a},\mathbf{T},\alpha,\sigma_P^2\right]\left[\boldsymbol{a}|\mathbf{T},\alpha,\sigma_P^2\right] \\
&\propto \frac{1}{\sqrt{\det(\Sigma_P)}}\exp\left\{-\frac{1}{2\sigma_P^2}\big(\mathbf{P}-u\alpha\big)^{\mathsf{T}}\Sigma_P^{-1}\big(\mathbf{P}-u\alpha\big)\right\} \\
\left[\boldsymbol{b}\middle|\mathbf{T},\beta,\sigma_T^2\right] &\propto \left[\mathbf{T}|\boldsymbol{b},\beta,\sigma_T^2\right]\left[\boldsymbol{b}|\beta,\sigma_T^2\right] \\
&\propto \frac{1}{\sqrt{\det(\Sigma_T)}}\exp\left\{-\frac{1}{2\sigma_T^2}\big(\mathbf{T}-v\beta\big)^{\mathsf{T}}\Sigma_T^{-1}\big(\mathbf{T}-v\beta\big)\right\}
\end{aligned} \tag{C.11}$$

where det stands for determinant (of a matrix). Finally, we derive the posterior distribution for temperature ($\mathbf{T}$), from which to draw samples using Gibbs sampler.

$$\begin{aligned}
\left[\mathbf{T}\middle|\mathbf{P},a,b,\alpha,\beta,\sigma_P^2,\sigma_T^2\right] &\propto \left[\mathbf{P}\middle|\mathbf{T},a,b,\alpha,\beta,\sigma_P^2,\sigma_T^2\right]\left[\mathbf{T}|b,\beta,\sigma_T^2\right] \\
&\propto \exp\left\{-\frac{1}{2\sigma_P^2}\big(\mathbf{P}-u\alpha\big)^{\mathsf{T}}\Sigma_P^{-1}\big(\mathbf{P}-u\alpha\big)\right\} \\
&\times \exp\left\{-\frac{1}{2\sigma_T^2}\big(\mathbf{T}-v\beta\big)^{\mathsf{T}}\Sigma_T^{-1}\big(\mathbf{T}-v\beta\big)\right\}
\end{aligned} \tag{C.12}$$

where u relies on $\mathbf{T}$. Let us re-write $(P-u\alpha)$ at time t by expanding u as defined above:

$$\begin{aligned}
\big(P-u\alpha\big)_t &= P_t-\alpha_0\sum_{i=0}^{t-1}a^iT_{t-i}-\alpha_1\sum_{i=0}^{t-1}a^iT_{t-i} \\
&= P_t-\alpha_0\sum_{i=0}^{T}a^{t-i}\mathbf{1}_{\{i\le t\}}-\alpha_1\sum_{i=0}^{T}a^{t-i}T_i\mathbf{1}_{\{i\le t\}} \\
&= \mathbf{P}-\alpha_0\mathbf{M}e-\alpha_1\mathbf{M}\mathbf{T}
\end{aligned} \tag{C.13}$$

where $\mathbf{M}\in\mathcal{M}_{T,T}(\mathbb{R})$ with $M_{i,j}=\mathbf{1}_{\{j\le i\}}a^{i-j}$, and e is a vector whose entries are all ones.

Therefore,

$$
\begin{aligned}
\big[\mathbf{T}\big|\mathbf{P}, a, b, \alpha, \beta, \sigma_P^2, \sigma_T^2\big] &\propto \exp\left\{-\frac{1}{2\sigma_P^2}\big(\mathbf{P}-\alpha_0\mathbf{M}e-\alpha_1\mathbf{M}\mathbf{T}\big)^{\mathsf{T}}\Sigma_P^{-1}\big(\mathbf{P}-\alpha_0\mathbf{M}e-\alpha_1\mathbf{M}\mathbf{T}\big)\right\} \\
&\quad\times \exp\left\{-\frac{1}{2\sigma_T^2}\big(\mathbf{T}-v\beta\big)^{\mathsf{T}}\Sigma_T^{-1}\big(\mathbf{T}-v\beta\big)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\mathbf{T}^{\mathsf{T}}\Big(\frac{\alpha_1^2}{\sigma_P^2}\mathbf{M}^{\mathsf{T}}\Sigma_P^{-1}\mathbf{M}+\frac{1}{\sigma_T^2}\Sigma_T^{-1}\Big)\mathbf{T}\right\} \\
&\quad\times \exp\left\{\frac{\alpha_1}{\sigma_P^2}\mathbf{T}^{\mathsf{T}}\mathbf{M}^{\mathsf{T}}\Sigma_P^{-1}\big(\mathbf{P}-\alpha_0\mathbf{M}e\big)+\frac{1}{\sigma_T^2}\mathbf{T}^{\mathsf{T}}\Sigma_T^{-1}v\beta\right\} \\
&\propto \exp\left\{-\frac{1}{2}\big(\mathbf{T}-\mu_T\big)^{\mathsf{T}}\Omega_T^{-1}\big(\mathbf{T}-\mu_T\big)\right\} \\
&\sim \mathcal{N}\big(\mu_T, \Omega_T\big)
\end{aligned}
\tag{C.14}
$$

with

$$
\begin{cases}
\mu_T = \dfrac{\alpha_1}{\sigma_P^2}\Omega_T\mathbf{M}^{\mathsf{T}}\Sigma_P^{-1}\big(\mathbf{P}-\alpha_0\mathbf{M}e\big)+\dfrac{1}{\sigma_T^2}\Omega_T\Sigma_T^{-1}v\beta \\
\Omega_T^{-1} = \dfrac{\alpha_1^2}{\sigma_P^2}\mathbf{M}^{\mathsf{T}}\Sigma_P^{-1}\mathbf{M}+\dfrac{1}{\sigma_T^2}\Sigma_T^{-1}
\end{cases}
\tag{C.15}
$$

Theorem C.0.1. *Assume*

$$
X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{bmatrix}\right)
$$

where $\Sigma_{21} = \Sigma_{12}^{\mathsf{T}}$. *Then*

$$
\big[X_1\big|X_2\big] \sim \mathcal{N}\Big(\mu_1 + \Sigma_{12}\Sigma_2^{-1}\big(X_2-\mu_2\big), \Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{21}\Big)
$$

Consider $X_1 = \mathbf{T}_1$ (past) and $X_2 = \mathbf{T}_2$ (calibration). According to Theorem C.0.1, the posterior distribution of past temperature given the data is:

$$\begin{cases} \big[\mathbf{T}_1 \big| \mathbf{T}_2, \mathbf{P}, a, b, \alpha, \beta, \sigma_P^2, \sigma_T^2\big] \sim \mathcal{N}\big(\mu, \Omega\big) \\ \mu = \mu_T^{(1)} + \Omega_T^{(1)(2)}\big(\Omega_T^{(2)(2)}\big)^{-1}\big(\mathbf{T}_2 - \mu_T^{(2)}\big) \\ \Omega = \Omega_T^{(1)(1)} - \Omega_T^{(1)(2)}\big(\Omega_T^{(2)(2)}\big)^{-1}\big(\Omega_T^{(2)(1)}\big) \end{cases} \tag{C.16}$$

where $\Omega_T^{(i)(j)}(i = 1, 2$ and $j = 1, 2)$ is the partitioned matrix. The projection (denoted by $\mathbf{T}_3$) can be done via the second level in (3.6) after all climate forcings extended to Year 2100.

APPENDIX D

Proof of Theorem 4.3.5

Proof. Let $H(x) = e^{-\frac{1}{2}[\frac{u}{s(1+\frac{x}{\sqrt{2d}})}+1]^2}$. The first and second derivative of $H(x)$ are

$$\begin{aligned}
H'(x) &= e^{-\frac{1}{2}[\frac{u}{s(1+\frac{x}{\sqrt{2d}})}+1]^2} \times \Big(-\big[\frac{u}{s(1+\frac{x}{\sqrt{2d}})}+1\big]\Big) \times \frac{u}{s}\big[-\frac{1}{(1+\frac{x}{\sqrt{2d}})^2}\frac{1}{\sqrt{2d}}\big] \\
&= e^{-\frac{1}{2}[\frac{u}{s(1+\frac{x}{\sqrt{2d}})}+1]^2} \times \frac{u}{s\sqrt{2d}}\frac{1}{(1+\frac{x}{\sqrt{2d}})^2}\big[\frac{u}{s(1+\frac{x}{\sqrt{2d}})}+1\big] \\
&:= H(x)H_1(x)
\end{aligned}$$

$$\begin{aligned}
H''(x) &= H'(x)H_1(x) + H(x)H_1'(x) = H(x)\big[H_1^2(x) + H_1'(x)\big] \\
&= e^{-\frac{1}{2}[\frac{u^2}{s(1+\frac{x}{\sqrt{2d}})}+1]^2} \frac{u^2}{2s^2d(1+\frac{x}{\sqrt{2d}})^4}\big[\frac{u}{s(1+\frac{x}{\sqrt{2d}})}+1\big]^2 \\
&\quad - e^{-\frac{1}{2}[\frac{u^2}{s(1+\frac{x}{\sqrt{2d}})}+1]^2} \frac{u}{2sd(1+\frac{x}{\sqrt{2d}})^3}\Big(\frac{u}{s(1+\frac{x}{\sqrt{2d}})} + 2\big[\frac{u}{s(1+\frac{x}{\sqrt{2d}})}+1\big]\Big) \qquad \text{(D.1)} \\
&= e^{-\frac{1}{2}[\frac{u^2}{s(1+\frac{x}{\sqrt{2d}})}+1]^2} \frac{u}{2sd(1+\frac{x}{\sqrt{2d}})^3}\Bigg\{\frac{u[\frac{u}{s(1+\frac{x}{\sqrt{2d}})}+1]^2}{s(1+\frac{x}{\sqrt{2d}})} \\
&\quad - \Big(\frac{u}{s(1+\frac{x}{\sqrt{2d}})} + 2\big[\frac{u}{s(1+\frac{x}{\sqrt{2d}})}+1\big]\Big)\Bigg\}
\end{aligned}$$

Use Taylor expansion on $H(x)$ at point a ($a \in \mathbb{R}$),

$$
\begin{aligned}
H(x) &= \sum_{0}^{\infty} \frac{H^{(n)}(a)}{n!}(x-a)^n \\
&= e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{a}{\sqrt{2d}})^2}} + e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{a}{\sqrt{2d}})^2}}(x-a)\frac{u^2}{s^2\sqrt{2d}}\frac{1}{(1+\frac{a}{\sqrt{2d}})^3} + R_1(x)
\end{aligned}
$$

where $R_1(x)$ is the mean-value (or Lagrange) form of the remainder:

$$
\begin{aligned}
R_1(x) &= \frac{H''(\xi)}{2!}(x-a)^2 \\
&= \frac{(x-a)^2}{2} e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{\xi}{\sqrt{2d}})^2}} \frac{u^2}{2s^2 d}\frac{1}{(1+\frac{\xi}{\sqrt{2d}})^4}\Big[\frac{u^2}{s^2(1+\frac{\xi}{\sqrt{2d}})^2} - 3\Big]
\end{aligned}
$$

for $\xi \in [a, x]$. When $a = 0$ (i.e., Maclaurin series),

$$
H(x) = e^{-\frac{(u+s)^2}{2s^2}}\Big[1 + \frac{ux}{s\sqrt{2d}}\big(\frac{u}{s}+1\big)\Big] + R_H(x) \tag{D.2}
$$

and

$$
\begin{aligned}
R_H(x) &= \frac{x^2}{2!}H''(\xi) \\
&= \frac{x^2}{2} e^{-\frac{1}{2}[\frac{u^2}{s(1+\frac{\xi}{\sqrt{2d}})}+1]^2} \frac{u}{2sd(1+\frac{\xi}{\sqrt{2d}})^3}\Bigg\{\frac{u[\frac{u}{s(1+\frac{\xi}{\sqrt{2d}})}+1]^2}{s(1+\frac{\xi}{\sqrt{2d}})} \\
&\quad - \Big(\frac{u}{s(1+\frac{\xi}{\sqrt{2d}})} + 2\big[\frac{u}{s(1+\frac{\xi}{\sqrt{2d}})}+1\big]\Big)\Bigg\}
\end{aligned} \tag{D.3}
$$

Note that

$$
\Phi_2(u,d) = 1 - \frac{1}{2}\int_{-\sqrt{2d}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}e^{-\frac{[u+s(1+\frac{x}{\sqrt{2d}})]^2}{2s^2(1+\frac{x}{\sqrt{2d}})^2}}dx, \quad \lim_{d\to\infty}\Phi_2(u,d) = 1 - \frac{1}{2}e^{-\frac{(u+s)^2}{2s^2}}
$$

Hence, plug in (D.2) and by triangle inequality,

$$\begin{aligned}
\Big|\Phi_2(u,d)-\big(1-\frac{1}{2}e^{-\frac{(u+s)^2}{2s^2}}\big)\Big| &\le \frac{1}{2}\int_{-\sqrt{2d}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\Big|e^{-\frac{1}{2}[\frac{u}{s(1+\frac{x}{\sqrt{2d}})}+1]^2}-e^{-\frac{(u+s)^2}{2s^2}}\Big|dx \\
&\le \frac{1}{2}\int_{-\sqrt{2d}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\big|H(x)-H(0)\big|dx \\
&\le \frac{1}{2}\int_{-\sqrt{2d}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}e^{-\frac{(u+s)^2}{2s^2}}\frac{u|x|}{s\sqrt{2d}}\big(\frac{u}{s}+1\big)dx \\
&\quad+\frac{1}{2}\int_{-\sqrt{2d}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\big|R_H(x)\big|dx \\
&:=\Phi_{21}(u,d)+\Phi_{22}(u,d)
\end{aligned}$$

$$\begin{aligned}
\Phi_{21}(u,d) &= \frac{1}{2}\Big(\int_{0}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}e^{-\frac{(u+s)^2}{2s^2}}\frac{ux}{s\sqrt{2d}}\big(\frac{u}{s}+1\big)dx \\
&\quad+\int_{-\sqrt{2d}}^{0}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}e^{-\frac{(u+s)^2}{2s^2}}\frac{-ux}{s\sqrt{2d}}\big(\frac{u}{s}+1\big)dx\Big) \\
&= \frac{1}{2\sqrt{2\pi}}\frac{u(u+s)e^{-\frac{(u+s)^2}{2s^2}}}{s^2\sqrt{2d}}\Big(\int_{0}^{\infty}xe^{-\frac{x^2}{2}}dx+\int_{0}^{\sqrt{2d}}xe^{-\frac{x^2}{2}}dx\Big) \qquad \text{(D.4)} \\
&= \frac{u(u+s)e^{-\frac{(u+s)^2}{2s^2}}}{4s^2\sqrt{\pi d}}\Big((-e^{-\frac{x^2}{2}})\Big|_0^{\infty}+(-e^{-\frac{x^2}{2}})\Big|_0^{\sqrt{2d}}\Big) \\
&= \frac{u(u+s)e^{-\frac{(u+s)^2}{2s^2}}}{4s^2\sqrt{\pi d}}\big(2-e^{-d}\big)
\end{aligned}$$

$H''(\xi)$ is continuous for all $\xi\in[0,x], x\in\mathbb{R}$ and

$$\begin{aligned}
\big|H''(\xi)\big| &\le e^{-\frac{(u+s)^2}{2s^2}}\frac{u}{2sd}\Big|\frac{u(\frac{u}{s}+1)^2}{s}-\Big[\frac{u}{s}+2\big(\frac{u}{s}+1\big)\Big]\Big| \\
&= e^{-\frac{(u+s)^2}{2s^2}}\frac{u}{2sd}\Big|\frac{u}{s}\big(\frac{u}{s}+1\big)^2-\frac{3u}{s}-2\Big|
\end{aligned} \qquad \text{(D.5)}$$

Plug (4.36), (D.5) into (D.3),

$$\begin{aligned}
\Phi_{22}(u,d) = \frac{1}{2}\int_{-\sqrt{2d}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{x^2}{2} e^{-\frac{1}{2}[\frac{u^2}{s(1+\frac{\xi}{\sqrt{2d}})}+1]^2} & \frac{u}{2sd(1+\frac{\xi}{\sqrt{2d}})^3}\Bigg\{\frac{u[\frac{u}{s(1+\frac{\xi}{\sqrt{2d}})}+1]^2}{s(1+\frac{\xi}{\sqrt{2d}})} \\
& - \Big(\frac{u}{s(1+\frac{\xi}{\sqrt{2d}})} + 2\big[\frac{u}{s(1+\frac{\xi}{\sqrt{2d}})}+1\big]\Big)\Bigg\}dx \\
& \le \frac{ue^{-\frac{(u+s)^2}{2s^2}}}{4sd}\Big|\frac{u}{s}\big(\frac{u}{s}+1\big)^2 - \frac{3u}{s} - 2\Big|\Big(\int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{x^2}{2}dx \\
& + \int_{-\sqrt{2d}}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{x^2}{2}dx\Big) \\
& \le \frac{u^2 e^{-\frac{(u+s)^2}{2s^2}}}{8s^2 d}\Big[3 + \big(\frac{u}{s}+1\big)^2\Big]
\end{aligned} \tag{D.6}$$

From (D.4) and (D.6),

$$\left|\Phi_2(u,d) - \Big(1 - \frac{1}{2}e^{-\frac{(u+s)^2}{2s^2}} - \frac{u(u+s)(2-e^{-d})}{4s^2\sqrt{\pi d}}e^{-\frac{(u+s)^2}{2s^2}}\Big)\right| \le \frac{u^2\big[3+(\frac{u}{s}+1)^2\big]}{8s^2 d}e^{-\frac{(u+s)^2}{2s^2}} \tag{D.7}$$

Therefore,

$$\begin{aligned}
& \left|\Phi_U(u) - \Big(1 - \frac{1}{2}e^{-\frac{(u+s)^2}{2s^2}} - \frac{u(u+s)(2-e^{-d})}{4s^2\sqrt{\pi d}}e^{-\frac{(u+s)^2}{2s^2}}\Big)\right| \\
\le & \left|\Phi_2(u,d) - \Big(1 - \frac{1}{2}e^{-\frac{(u+s)^2}{2s^2}} - \frac{u(u+s)(2-e^{-d})}{4s^2\sqrt{\pi d}}e^{-\frac{(u+s)^2}{2s^2}}\Big)\right| \\
\le & \frac{u^2\big[3+(\frac{u}{s}+1)^2\big]}{8s^2 d}e^{-\frac{(u+s)^2}{2s^2}} := \delta_2(u,d)
\end{aligned}$$

□

APPENDIX E

Boundaries Justification

Moreover, let $G(x) = e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{x}{\sqrt{2d}})^2}}$.

$$
\begin{aligned}
G'(x) &= e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{x}{\sqrt{2d}})^2}} \times \big(-\frac{u^2}{2s^2}\big)\Big[-2\frac{1}{(1+\frac{x}{\sqrt{2d}})^3}\frac{1}{\sqrt{2d}}\Big] \\
&= e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{x}{\sqrt{2d}})^2}} \times \frac{u^2}{s^2\sqrt{2d}}\frac{1}{(1+\frac{x}{\sqrt{2d}})^3} \\
&:= G(x)G_1(x)
\end{aligned}
$$

$$
\begin{aligned}
G''(x) &= G'(x)G_1(x) + G(x)G_1'(x) \\
&= e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{x}{\sqrt{2d}})^2}}\Big[\Big(\frac{u^2}{s^2\sqrt{2d}}\frac{1}{(1+\frac{x}{\sqrt{2d}})^3}\Big)^2 - \frac{3u^2}{s^2\sqrt{2d}}\frac{1}{(1+\frac{x}{\sqrt{2d}})^4}\frac{1}{\sqrt{2d}}\Big] \qquad \text{(E.1)} \\
&= e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{x}{\sqrt{2d}})^2}} \times \frac{u^2}{2s^2 d}\frac{1}{(1+\frac{x}{\sqrt{2d}})^4}\Big[\frac{u^2}{s^2(1+\frac{x}{\sqrt{2d}})^2} - 3\Big]
\end{aligned}
$$

Then the Maclaurin series of $G(x)$ is

$$
G(x) = e^{-\frac{u^2}{2s^2}}\Big(1 + \frac{u^2 x}{s^2\sqrt{2d}}\Big) + R_G(x) \qquad \text{(E.2)}
$$

and the Lagrange form of the remainder for $\xi \in [0,x]$ is

$$\begin{aligned} R_G(x) &= \frac{x^2}{2!}G''(\xi) \\ &= \frac{x^2}{2}e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{\xi}{\sqrt{2d}})^2}}\frac{u^2}{2s^2d}\frac{1}{(1+\frac{\xi}{\sqrt{2d}})^4}\Big[\frac{u^2}{s^2(1+\frac{\xi}{\sqrt{2d}})^2}-3\Big],\ \xi \in [0,x] \end{aligned} \tag{E.3}$$

Recall

$$\Phi_1(u,d) = 1 - \frac{1}{2}\int_{-\sqrt{2d}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{x}{\sqrt{2d}})^2}}dx, \quad \lim_{d\to\infty}\Phi_1(u,d) = 1-\frac{1}{2}e^{-\frac{u^2}{2s^2}}$$

$$\Rightarrow \Phi_1(u,d) - \big(1-\frac{1}{2}e^{-\frac{u^2}{2s^2}}\big) = \frac{1}{2}\int_{-\sqrt{2d}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\Big[e^{-\frac{u^2}{2s^2}} - e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{x}{\sqrt{2d}})^2}}\Big]dx$$

By triangle inequality,

$$\begin{aligned} \Big|\Phi_1(u,d) - \big(1-\frac{1}{2}e^{-\frac{u^2}{2s^2}}\big)\Big| &\le \frac{1}{2}\int_{-\sqrt{2d}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\Big|e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{x}{\sqrt{2d}})^2}} - e^{-\frac{u^2}{2s^2}}\Big|dx \\ &\le \frac{1}{2}\int_{-\sqrt{2d}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\big|G(x)-G(0)\big|dx \\ &\le \frac{1}{2}\int_{-\sqrt{2d}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}e^{-\frac{u^2}{2s^2}}\frac{u^2|x|}{s^2\sqrt{2d}}dx \\ &\quad + \frac{1}{2}\int_{-\sqrt{2d}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\big|R_G(x)\big|dx \\ &:= \Phi_{11}(u,d) + \Phi_{12}(u,d) \end{aligned}$$

$$\begin{aligned}\Phi_{11}(u,d) &= \frac{1}{2}\Big(\int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} e^{-\frac{u^2}{2s^2}} \frac{u^2 x}{s^2\sqrt{2d}} dx + \int_{-\sqrt{2d}}^0 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} e^{-\frac{u^2}{2s^2}} \frac{u^2(-x)}{s^2\sqrt{2d}} dx\Big) \\ &= \frac{1}{2\sqrt{2\pi}} \frac{u^2 e^{-\frac{u^2}{2s^2}}}{s^2\sqrt{2d}} \Big(\int_0^\infty x e^{-\frac{x^2}{2}} dx + \int_0^{\sqrt{2d}} x e^{-\frac{x^2}{2}} dx\Big) \\ &= \frac{u^2 e^{-\frac{u^2}{2s^2}}}{4s^2\sqrt{\pi d}} \Big((-e^{-\frac{x^2}{2}})\Big|_0^\infty + (-e^{-\frac{x^2}{2}})\Big|_0^{\sqrt{2d}}\Big) \\ &= \frac{u^2 e^{-\frac{u^2}{2s^2}}}{4s^2\sqrt{\pi d}} \big(2 - e^{-d}\big)\end{aligned} \tag{E.4}$$

For all $\xi \in [0, x]$, $x \in \mathbb{R}$, $G''(\xi)$ is continuous and

$$\big|G''(\xi)\big| \le e^{-\frac{u^2}{2s^2}} \frac{u^2}{2s^2 d} \Big|\frac{u^2}{s^2} - 3\Big| \tag{E.5}$$

Applying the same calculation in (D.6) to $\Phi_{12}(u,d)$,

$$\begin{aligned}\Phi_{12}(u,d) &= \frac{1}{2}\int_{-\sqrt{2d}}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{|x|^2}{2} e^{-\frac{u^2}{2s^2}\frac{1}{(1+\frac{\xi}{\sqrt{2d}})^2}} \frac{u^2}{2s^2 d} \frac{1}{(1+\frac{\xi}{\sqrt{2d}})^4} \Big|\frac{u^2}{s^2(1+\frac{\xi}{\sqrt{2d}})^2} - 3\Big| dx \\ &\le \frac{1}{2} e^{-\frac{u^2}{2s^2}} \frac{u^2}{2s^2 d} \Big|\frac{u^2}{s^2} - 3\Big| \Big(\int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{x^2}{2} dx + \int_{-\sqrt{2d}}^0 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{x^2}{2} dx\Big) \\ &\le \frac{u^2 e^{-\frac{u^2}{2s^2}}}{8s^2 d} \Big(3 + \frac{u^2}{s^2}\Big)\end{aligned} \tag{E.6}$$

So

$$\Big|\Phi_1(u,d) - \Big(1 - \frac{1}{2} e^{-\frac{u^2}{2s^2}} - \frac{u^2(2 - e^{-d})}{4s^2\sqrt{\pi d}} e^{-\frac{u^2}{2s^2}}\Big)\Big| \le \frac{u^2(3 + \frac{u^2}{s^2})}{8s^2 d} e^{-\frac{u^2}{2s^2}} \tag{E.7}$$

From (E.7),

$$\begin{aligned}
&\left|\Phi_U(u) - \Big(1 - \frac{1}{2}e^{-\frac{u^2}{2s^2}} - \frac{u^2(2-e^{-d})}{4s^2\sqrt{\pi d}}e^{-\frac{u^2}{2s^2}}\Big)\right| \\
\geq &\left|\Phi_1(u,d) - \Big(1 - \frac{1}{2}e^{-\frac{u^2}{2s^2}} - \frac{u^2(2-e^{-d})}{4s^2\sqrt{\pi d}}e^{-\frac{u^2}{2s^2}}\Big)\right| \\
\geq &\frac{u^2(3+\frac{u^2}{s^2})}{8s^2 d}e^{-\frac{u^2}{2s^2}} := \delta_1(u,d)
\end{aligned}$$

which shows that $\delta_2(u,d)$ in Theorem 4.3.5 is sharp enough.

By the similar computation in (4.39) and (4.42),

$$\begin{aligned}
\Phi_V(v) &\geq \mathbb{E}_x\Big[1 - \frac{1}{2}e^{-\frac{(v-x\sigma_{Y^*})^2}{2s^2}} - \delta_1(v - x\sigma_{Y^*}, d)\Big] \\
&= 1 - \frac{1}{2}\int_{-\infty}^{\infty} e^{-\frac{(v-x\sigma_{Y^*})^2}{2s^2}} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx - \mathbb{E}_x\big[\delta_1(v - x\sigma_{Y^*}, d)\big] \\
&= 1 - \frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\big[\frac{s^2+\sigma_{Y^*}^2}{s^2}\big(x - \frac{v\sigma_{Y^*}}{s^2+\sigma_{Y^*}^2}\big)^2\big]} e^{-\frac{v^2}{2(s^2+\sigma_{Y^*}^2)}}\,dx \\
&\quad - \mathbb{E}_x\big[\delta_1(v - x\sigma_{Y^*}, d)\big] \\
&= 1 - \frac{s}{2\sqrt{s^2+\sigma_{Y^*}^2}}e^{-\frac{v^2}{2(s^2+\sigma_{Y^*}^2)}} - \mathbb{E}_x\big[\delta_1(v - x\sigma_{Y^*}, d)\big]
\end{aligned}$$

The density function of $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ is

$$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}}e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}$$

$$\begin{cases} \mu_X = \dfrac{v\sigma_{Y^*}}{s^2+\sigma_{Y^*}^2} \\ \sigma_X^2 = \dfrac{s^2}{s^2+\sigma_{Y^*}^2} \end{cases} \tag{E.8}$$

$$\begin{aligned}
\Rightarrow \mathbb{E}_x\big[\delta_1(v - x\sigma_{Y*}, d)\big] &= \int_{-\infty}^{\infty} \frac{(v - x\sigma_{Y*})^2 e^{-\frac{(v - x\sigma_{Y*})^2}{2s^2}}}{4s^2\sqrt{d}} \Big(\frac{2 - e^{-d}}{\sqrt{\pi}} \\
&\quad + \frac{3 + \frac{(v - x\sigma_{Y*})^2}{s^2}}{2\sqrt{d}}\Big) \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx \\
&= \frac{e^{-\frac{v^2}{2(s^2+\sigma_{Y*}^2)}}}{4s\sqrt{d(s^2 + \sigma_{Y*}^2)}} \Big[\int_{-\infty}^{\infty} \Big(\frac{2 - e^{-d}}{\sqrt{\pi}} + \frac{3}{2\sqrt{d}}\Big)(v - x\sigma_{Y*})^2 f_X(x) dx \\
&\quad + \int_{-\infty}^{\infty} \frac{1}{2s^2\sqrt{d}} (v - x\sigma_{Y*})^4 f_X(x) dx\Big] \\
&= \frac{e^{-\frac{v^2}{2(s^2+\sigma_{Y*}^2)}}}{4s\sqrt{d(s^2 + \sigma_{Y*}^2)}} \Big\{\Big(\frac{2 - e^{-d}}{\sqrt{\pi}} + \frac{3}{2\sqrt{d}}\Big)\Big[v^2 - 2v\sigma_{Y*}\mu_X + \sigma_{Y*}^2\big(\mu_X^2 \\
&\quad + \sigma_X^2\big)\Big] + \frac{1}{2s^2\sqrt{d}}\Big[v^4 - 4v^3\sigma_{Y*}\mu_X + 6v^2\sigma_{Y*}^2\big(\mu_X^2 + \sigma_X^2\big) \\
&\qquad - 4v\sigma_{Y*}^3\big(\mu_X^3 + 3\mu_X\sigma_X^2\big) + \sigma_{Y*}^4\big(\mu_X^4 + 6\mu_X^2\sigma_X^2 + 3\sigma_X^4\big)\Big]\Big\}
\end{aligned} \tag{E.9}$$

Therefore,

$$\begin{aligned}
&\Bigg|\Phi_V(v) - \Bigg(1 - \frac{se^{-\frac{v^2}{2(s^2+\sigma_{Y*}^2)}}}{2\sqrt{s^2 + \sigma_{Y*}^2}} - \frac{(2 - e^{-d})e^{-\frac{v^2}{2(s^2+\sigma_{Y*}^2)}}}{4s\sqrt{\pi d(s^2 + \sigma_{Y*}^2)}}\Big[v^2 \\
&\qquad - 2v\sigma_{Y*}\mu_X + \sigma_{Y*}^2\big(\mu_X^2 + \sigma_X^2\big)\Big]\Bigg)\Bigg| \geq \delta'(v, d)
\end{aligned} \tag{E.10}$$

$$\begin{aligned}
\delta'(v, d) &= \frac{e^{-\frac{v^2}{2(s^2+\sigma_{Y*}^2)}}}{8sd\sqrt{s^2 + \sigma_{Y*}^2}}\Big[3\Big(v^2 - 2v\sigma_{Y*}\mu_X + \sigma_{Y*}^2\big(\mu_X^2 + \sigma_X^2\big)\Big) + \frac{1}{s^2}\Big(v^4 - 4v^3\sigma_{Y*}\mu_X \\
&\quad + 6v^2\sigma_{Y*}^2\big(\mu_X^2 + \sigma_X^2\big) - 4v\sigma_{Y*}^3\big(\mu_X^3 + 3\mu_X\sigma_X^2\big) + \sigma_{Y*}^4\big(\mu_X^4 + 6\mu_X^2\sigma_X^2 + 3\sigma_X^4\big)\Big)\Big]
\end{aligned} \tag{E.11}$$

and μ_X, σ_X^2 defined in (E.8), which implies the bounds of $\Phi_V(v)$ cannot do any better.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. Bayesian Analysis, 1(3), 515-533.

[2] Haller, H. and Krauss, S. (2002). Misinterpretations of significance: a problem students share with their teachers. Methods of Psychological Research, 7(1), 1-20.

[3] McShane, B. B. and Gal, D. (2017). Statistical significance and the dichotomization of evidence. Journal of the American Statistical Association, 112(519), 885-895.

[4] Dunson, D. B. (2001). Commentary: practical advantages of Bayesian analysis of epidemiologic data. American journal of Epidemiology, 153(12), 1222-1226.

[5] Baldos, U. L. C., Viens, F. G., Hertel, T. W., and Fuglie, K. O. (2019). R&D spending, knowledge capital, and agricultural productivity growth: a Bayesian approach. American Journal of Agricultural Economics, 101(1), 291-310.

[6] Zhang, Z., Hamagami, F., Wang, L., Nesselroade, J. R., and Grimm, K. J. (2005). Bayesian analysis of longitudinal data using growth curve models. International Journal of Behavioral Development, 31(4), 374-383.

[7] Neufcourt, L., Cao, Y., Nazarewicz, W., and Viens, F. (2018). Bayesian approach to model-based extrapolation of nuclear observables. Physical Review C, 98(3), 034318.

[8] Wiener, N. (1938). The homogeneous chaos. American Journal of Mathematics, 60(4), 897-936.

[9] Nourdin, I. and Peccati, G. (2009). Noncentral convergence of multiple integrals. The Annals of Probability, 37(4), 1412-1426.

[10] Lobell, D. B., Cassman, K. G., and Field, C. B. (2009). Crop yield gaps: their importance, magnitudes, and causes. Annual Review of Environment and Resources, 34, 179-204.

[11] Fermont, A. M., van Asten, P. J. A., Tittonell, P., van Wijk, M. T., and Giller, K. E. (2009). Closing the cassava yield gap: an analysis from smallholder farms in East Africa. Field Crops Research, 112(1), 24-36.

[12] Tamene, L., Mponela, P., Ndengu, G., and Kihara, J. (2016). Assessment of maize yield gap and major determinant factors between smallholder farmers in the Dedza district of Malawi. Nutrient Cycling in Agroecosystems, 105, 291-308.

[13] van Dijk, M., Morley, T., Jongeneel, R., van Ittersum, M., Reidsma, P., and Ruben, R. (2017). Disentangling agronomic and economic yield gaps: an integrated framework and application. Agricultural Systems, 154, 90-99.

[14] Liu, Z., Yang, X., Hubbard, K. G., and Lin, X. (2012). Maize potential yields and yield gaps in the changing climate of northeast China. Global Change Biology, 18(11), 3441-3454.

[15] Tittonell, P. and Giller, K. E. (2013). When yield gaps are poverty traps: The paradigm of ecological intensification in African smallholder agriculture. Field Crops Research, 143, 76-90.

[16] Snapp, S. S. (1998). Soil nutrient status of smallholder farms in Malawi. Communications in Soil Science and Plant Analysis, 29(17-18), 2571-2588.

[17] Batie, S. S. (2008). Wicked problems and applied economics. American Journal of Agricultural Economics, 90(5), 1176-1191.

[18] Wang, H., Snapp, S. S., Fisher, M., and Viens, F. (2019). A Bayesian analysis of longitudinal farm surveys in central Malawi reveals yield determinants and site-specific management strategies. PLoS ONE, 14(8), e0219296.

[19] Mungai, L. M., Snapp, S., Messina, J. P., Chikowo, R., Smith, A., Anders, E., Richardson, R. B., and Li, G. (2016). Smallholder farms and the potential for sustainable intensification. Frontiers in Plant Science, 7, 1720.

[20] Lowole, M. W. (1983). Soil Map of Malawi. Department of Agricultural Research, Lilongwe, Malawi.

[21] Giné, X., Townsend, R., and Vickery, J. (2007). Statistical analysis of rainfall insurance payouts in southern India. American Journal of Agricultural Economics, 89(5), 1248-1254.

[22] Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. The Annals of Statistics, 7(2), 269-281.

[23] Beukes, I., Rose, L. J., Shephard, G. S., Flett, B. C., and Viljoen, A. (2017). Mycotoxigenic Fusarium species associated with grain crops in South Africa – a review. South African Journal of Science, 113(3-4), 1-12.

[24] Sauer, J. and Tchale, H. (2009). The economics of soil fertility management in Malawi. Review of Agricultural Economics, 31(3), 535-560.

[25] Marenya, P. P. and Barrett, C. B. (2009). State-conditional fertilizer yield response on western Kenyan farms. American Journal of Agricultural Economics, 91(4), 991-1006.

[26] Scharf, P. C., Wiebold, W. J., and Lory, J. A. (2002). Corn yield response to nitrogen fertilizer timing and deficiency level. Agronomy Journal, 94(3), 435-441.

[27] Kabambe, V. H., Nambuzi, S. C., and Kauwa A. E. (2008). Integrated management of witchweed (*Striga asiatica* [L.] Kuntze) by means of maize-legume rotations and intercropping systems in Malawi. Bunda Journal of Agriculture, Environmental Science and Technology, 3(2), 35-42.

[28] Berner, D. K., Kling, J. G. and Singh, B. B. (1995). Striga research and control: a perspective from Africa. Plant Disease, 79(7), 652-660.

[29] Jamil, M., Kanampiu, F. K., Karaya, H., Charnikhova, T., and Bouwmeester, H. J. (2012). Striga hermonthica parasitism in maize in response to N and P fertilisers. Field Crops Research, 134, 1-10.

[30] Sauerborn, J., Kranz, B., and Mercer-Quarshie, H. (2003). Organic amendments mitigate heterotrophic weed infestation in savannah agriculture. Applied Soil Ecology, 23(2), 181-186.

[31] Atera, E. A., Itoh, K., Azuma, T. and Ishii, T. (2012). Farmers perception and constraints to the adoption of weed control option: the case of *Striga asiatica* in Malawi. The Journal of Agricultural Science, 4(5), 41.

[32] Shaxson, L. and Riches, C. (1998). Where once there was grain to burn: a farming system in crisis in eastern Malawi. Outlook on Agriculture, 27(2), 101-105.

[33] Snapp, S. S. , Blackie, M. J., and Donovan, C. (2003). Realigning research and extension to focus on farmers' constraints and opportunities. Food Policy, 28(4), 349-363.

[34] Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K. , Boschung, J., Nauels, A., Xia, Y., Bex V., and Midgley, P. M. (2013). Climate change 2013: the physical science basis. Working group I contribution to the fifth assessment report of the Intergovernmental Panel on Climate Change. Cambridge, New York: Cambridge University Press.

[35] Li, B., Nychka, D. W., and Ammann, C. M. (2010). The value of multiproxy reconstruction of past climate. Journal of the American Statistical Association, 105(491), 883-895.

[36] Grotch, S. L. and MacCracken, M. C. (1991). The use of general circulation models to predict regional climatic change. Journal of Climate, 4, 286-303.

[37] Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase

6 (CMIP6) experimental design and organization. Geoscientific Model Development, 9(5), 1937-1958.

[38] Matthes, K., Andersson, F. B., Barnard, M. E., Beer, L., Charbonneau, J., Clilverd, P., Dudok de Wit, M. A., Haberreiter, T., Hendry, M., Jackman, A., Kretzschmar, C. H., Kruschke, M., Kunze, T., Langematz, M., Marsh, U.,Maycock, D. R., Misios, A. C. (2017). Solar Forcing for CMIP6 (v3.2). Geoscientific Model Development, 10(6), 2247-2302.

[39] Jones, P. D., Briffa, K. R., Osborn, T. J., Lough, J. M., van Ommen, T. D., Vinther, B. M., Luterbacher, J., Wahl, E. R., Zwiers, F. W., Mann, M. E., Schmidt, G. A., Ammann, C. M., Buckley, B. M., Cobb, K. M., Esper, J., Goosse, H., Graham, N., Jansen, E., Kiefer, T., Kull, C., and Kütte, M. (2009). High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects. The Holocene, 19, 3-49.

[40] Werner, J. P. and Tingley, M. P. (2015). Technical note: probabilistically constraining proxy agedepth models within a Bayesian hierarchical reconstruction model. Climate of the Past, 11, 533-545.

[41] Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S., and Ni, F. (2008). Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. Proceedings of the National Academy of Sciences, 105(36), 13252-13257.

[42] Mann, M. E., Bradley, R. S., and Hughes, M. K. (1998). Global-scale temperature patterns and climate forcing over the past six centuries. Nature, 392, 779787.

[43] Yu, Z., Chu, P., and Schroeder, T. (1997). Predictive skills of seasonal to annual rainfall variations in the U.S. affiliated Pacific islands: canonical correlation analysis and multivariate principal component regression approaches. Journal of Climate, 10, 2586-2509.

[44] Smerdon, J. E., Kaplan, A., Chang, D., and Evans, M. N. (2010). A pseudoproxy evaluation of the CCA and RegEM methods for reconstructing climate fields of the last millennium. Journal of Climate, 23, 4856-4880.

[45] Werner, J. P., Luterbacher, J., and Smerdon, J. E. (2013). A pseudoproxy evaluation of Bayesian hierarchical modeling and canonical correlation analysis for climate field reconstructions over Europe. Journal of Climate, 26, 851-867.

[46] Schneider, T. (2001). Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. Journal of Climate, 14, 853-871.

[47] Rutherford, S., Mann, M. E., Delworth, T. L., and Stouffer, R. J. (2003). Climate field reconstruction under stationary and nonstationary forcing. Journal of Climate, 16, 462-479.

[48] Zhang, Z., Mann, M. E., and Cook, E. R. (2004). Alternative methods of proxy-based climate field reconstruction: application to summer drought over the conterminous United States back to AD 1700 from tree-ring data. The Holocene, 14(4), 502-516.

[49] Lee, T. C. K., Zwiers, F. W., and Tsao, M. (2008). Evaluation of proxy-based millennial reconstruction methods. Climate Dynamics, 31, 263-281.

[50] Christiansen, B., Schmith, T., and Thejll, P. (2009). A surrogate ensemble study of climate reconstruction methods: stochasticity and robustness. Journal of Climate, 22, 951-976.

[51] Jones, P. D., Briffa, K. R., Barnett, T. P., and Tett, S. F. B. (1998). High-resolution palaeoclimatic records for the last millennium: interpretation, integration and comparison with general circulation model control-run temperatures. The Holocene, 8(4), 455-471.

[52] Mann, M. E., Bradley, R. S., and Hughes, M. K. (1999). Northern hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations. Geophysical Research Letters, 26(6), 759-762.

[53] Tingley, M. P., Craigmile, P. F., Haran, M., Li, B., Mannshardt, E., and Rajaratnam, B. (2012). Piecing together the past: statistical insights into paleoclimatic reconstructions. Quaternary Science Reviews, 35, 1-22.

[54] Hoeting, J. A., Madigan, D., Raft, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. Statistical Science, 14(4), 382-417.

[55] Katz, R. W. (2002). Techniques for estimating uncertainty in climate change scenarios and impact studies. Climate Research, 20, 167-185.

[56] Cheaib, A. , Badeau, V., Boe, J., Chuine, I., Delire, C., Dufrêne, E., François, C., Gritti, E. S., Legay, M., Pagé, C., Thuiller, W., Viovy, N., and Leadley, P. (2012). Climate change impacts on tree ranges: model intercomparison facilitates understanding and quantification of uncertainty. Ecology Letters, 15, 533-544.

[57] Haslett, J. , Whiley, M., Bhattacharya, S., SalterTownshend, M., Wilson, S. P., Allen, J. R., Huntley, B., and Mitchell, F. J. (2006). Bayesian palaeoclimate reconstruction. Journal of the Royal Statistical Society: Series A (Statistics in Society), 169(3), 395-438.

[58] Barboza, L., Li, B., Tingley, M. P., and Viens, F. G. (2014). Reconstructing past temperatures from natural proxies and estimated climate forcings using short- and long-memory models. The Annals of Applied Statistics, 8, 1966-2001.

[59] Tingley, M. P. and Huybers, P. (2010). A Bayesian algorithm for reconstructing climate anomalies in space and time. Part I: development and applications to paleoclimate reconstruction problems. Journal of Climate, 23, 2759-2781.

[60] Henley, B. J., Thyer, M. A., Kuczera, G., and Franks, S. W. (2011). Climate-informed stochastic hydrological modeling: Incorporating decadal-scale variability using paleo data. Water Resources Research, 47, W11509.

[61] Comboul, M., Emile-Geay, J., Evans, M. N., Mirnateghi, N., Cobb, K. M., and Thompson, D. M. (2014). A probabilistic model of chronological errors in layer-counted climate proxies: applications to annually banded coral archives. Climate of the Past, 10, 825-841.

[62] Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., Shindell, D., Ammann, C., Faluvegi, G., and Ni, F. (2009). Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly. Science, 326, 1256-1260.

[63] Loso, M. G. (2009). Summer temperatures during the Medieval Warm Period and Little Ice Age inferred from varved proglacial lake sediments in southern Alaska. Journal of Paleolimnology, 41, 117.

[64] Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F. B., and Jones, P. D. (2006). Uncertainty estimates in regional and global observed temperature changes: a new data set from 1850. Journal of Geophysical Research, 111, D12106.

[65] Rayner, N. A., Brohan, P., Parker, D. E., Folland, C. K., Kennedy, J. J., Vanicek, M., Ansell, T. J., and Tett, S. F. B. (2006). Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: the HadSST2 dataset." Journal of Climate, 19, 446-469.

[66] Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D. (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 data set. Journal of Geophysical Research, 117, D08101.

[67] Jones, P. D., Lister, D. H., Osborn, T. J., Harpham, C., Salmon, M., and Morice, C. P. (2012). Hemispheric and large-scale land surface air temperature variations: an extensive revision and an update to 2010. Journal of Geophysical Research, 117, D05127.

[68] Ammann, C. M., Genton, M. G., and Li, B. (2010). Technical note: correcting for signal attenuation from noisy proxy data in climate reconstructions. Climate of the Past, 6, 273-279.

[69] Gueymard, C. A. (2018). A reevaluation of the solar constant based on a 42-year total solar irradiance time series and a reconciliation of spaceborne observation. Solar Energy, 168, 2-9.

[70] Bard, E., Raisbeck, G., Yiou, F., and Jouzel, J. (2000). Solar irradiance during the last 1200 years based on cosmogenic nuclides. Tellus B, 52, 985-992.

[71] Lean, J., Beer, J., and Bradley, R. (1995). Reconstruction of solar irradiance since 1610: implications for climate change. Geophysical Research Letters, 22(23), 3195-3198.

[72] Ammann, C. M., Joos, F., Schimel, D. S., Otto-Bliesner, B. L., and Tomas, R. A. (2007). Solar influence on climate during the past millennium: results from transient simulations with the NCAR climate system model. Proceedings of the National Academy of Sciences, 104(10), 3713-3718.

[73] Shindell, D. T., Schmidt, G. A., Miller, R. L., and Rind, D. (2001). Northern hemisphere winter climate response to greenhouse gas, ozone, solar, and volcanic forcing. Journal of Geophysical Research, 106(D7), 7193-7210.

[74] Shindell, D. T. and Schmidt, G. A. (2003). Volcanic and solar forcing of climate change during the preindustrial era." Journal of Climate, 16, 4094-4107.

[75] Moss, R. H., Edmonds, J. A., Hibbard, K. A., Manning, M. R., Rose, S. K., van Vuuren, D. P., Carter, T. R., Emori, S., Kainuma, M., Kram, T., Meehl, G. A., Mitchell, J. F. B., Nakicenovic, N., Riahi, K., Smith, S. J., Stouffer, R. J., Thomson, A. M., Weyant, J. P., and Wilbanks, T. J. (2010). The next generation of scenarios for climate change research and assessment. Nature, 463, 747-756.

[76] van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G. C., Kram, T., Krey, V., Lamarque, J. F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S. J., Rose, S. K. (2011). The representative concentration pathways: an overview. Climatic Change, 109, 5-31.

[77] van Vuuren, D. P., den Elzen, M. G. J., Lucas, P. L., Eickhout, B., Strengers, B. J., van Ruijven, B., Wonink, S., and van Houdt, R. (2007). Stabilizing greenhouse gas concentrations at low levels: an assessment of reduction strategies and costs. Climatic Change, 81, 119-159.

[78] Hijioka, Y., Matsuoka, Y., Nishimoto, H., Masui, T., and Kainuma, M. (2008). Global GHG emission scenarios under GHG concentration stabilization targets. Journal of Global Environment Engineering, 13, 97-108.

[79] Riahi, Keywan, Grübler, A., and Nakicenovicac, N. (2007). Scenarios of long-term socio-economic and environmental development under climate stabilization. Technological Forecasting and Social Change, 74(7), 887-935.

[80] Free, M. and Robock, A. (1999). Global warming in the context of the Little Ice Age. Journal of Geophysical Research, 104, 19057-19070.

[81] Myhre, G., Highwood, E. J., Shine, K. P., and Stordal, F. (1998). New estimates of radiative forcing due to well mixed greenhouse gases. Geophysical Research Letters, 25, 2715-2718.

[82] Collins, W. D., Ramaswamy, V., Schwarzkopf, M. D., Sun, Y., Portmann, R. W., Fu, Q., Casanova, S. E. B., Dufresne, J. L., Fillmore, D. W., Forster, P. M. D., Galin, V. Y., Gohar, L. K., Ingram, W. J., Kratz, D. P., Lefebvre, M. P., Li, J., Marquet, P., Oinas, V., and Tsush, Y. (2006). Radiative forcing by well-mixed greenhouse gases: estimates from climate models in the intergovernmental panel on climate change (IPCC) fourth assessment report (AR4). Journal of Geophysical Research, 111, D14317.

[83] Chouet, B. A. (1996). Long-period volcano seismicity: its source and use in eruption forecasting. Nature, 380, 309-316.

[84] Oppenheimer, C. (2003). Climatic, environmental and human consequences of the largest known historic eruption: Tambora volcano (Indonesia) 1815. Progress in Physical Geography, 27 (2), 230-259.

[85] Man, W., Zhou, T., and Jungclaus, J. H. (2014). Effects of large volcanic eruptions on global summer climate and East Asian monsoon changes during the last millennium: analysis of MPI-ESM simulations. Journal of Climate, 27, 7394-7409.

[86] Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. Biometrika, 65(2), 297-303.

[87] Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. Journal of the American Statistical Association, 89(425), 208-218.

[88] Eitrheim, Ø. and Teräsvirta, T. (1996). Testing the adequacy of smooth transition autoregressive models. Journal of Econometrics, 74 (1), 59-75.

[89] Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. Journal of the American Statistical Association, 78(381), 47-55.

[90] Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. Statistical Science, 7(4), 457-472.

[91] McKeen, S., Wilczak, J., Grell, G., Djalalova, I., Peckham, S., Hsie, E. Y., Gong, W., Bouchet, V., Menard, S., Moffet, R., McHenry, J., McQueen, J., Tang, Y., Carmichael, G. R., Pagowski, M., Chan, A., Dye, T., Frost, G., Lee, P., and Mathur, R. (2005). Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. Journal of Geophysical Research, 110, D21307.

[92] Savage, N. H., Agnew, P., Davis, L. S., Ordóñnez, C., Thorpe, R., Johnson, C. E., O'Connor, F. M., and Dalvi, M. (2013). Air quality modelling using the met office unified model (AQUM OS24-26): model description and initial evaluation. Geoscientific Model Development, 6, 353-372.

[93] Chai, T. and Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7, 1247-1250.

[94] Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting, 15, 559-570.

[95] Grimit, E. P., Gneiting, T., Berrocal, V. J., and Johnson, N. A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. Quarterly Journal of the Royal Meteorological Society, 132, 2925-2942.

[96] Gschlößl, S. and Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. Scandinavian Actuarial Journal, 2007(3), 202-225.

[97] Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359-378.

[98] Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association, 91(434), 883-904.

[99] Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics, 7(4), 434-455.

[100] Hegerl, G. C., Brönnimann, S., Schurer, A., and Cowan, T. (2018). The early 20th century warming: anomalies, causes, and consequences. Wiley Interdisciplinary Reviews: Climate Change, 9(4), e522.

[101] Phipps, S. J., McGregor, H. V., Gergis, J., Gallant, A. J. E., Neukom, R., Stevenson, S., Ackerley, D., Brown, J. R., Fischer, M. J., and van Ommen, T. D. (2013). Paleoclimate datamodel comparison and the role of climate forcings over the past 1500 years. Journal of Climate, 26, 6915-6936.

[102] Eddy, J. A. (1976). The Maunder Minimum. Science, 192(4245), 1189-1202.

[103] Hempelmann, A. and Weber, W. (2012). Correlation between the sunspot number, the total solar irradiance, and the terrestrial insolation. Solar Physics, 277, 417-430.

[104] Kopp, G., Krivova, N., Wu, C. J., and Lean, J. (2016). The impact of the revised sunspot record on solar irradiance reconstructions. Solar Physics, 291, 2951-2965.

[105] Mann, M. E., Fuentes, J. D., and Rutherford, S. (2012). Underestimation of volcanic cooling in tree-ring-based reconstructions of hemispheric temperatures. Nature Geoscience, 5, 202-205.

[106] Taylor, K. E., Stouffer, R. J., and Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. Bulletin of the American Meteorological Society, 93, 485-498.

[107] Holmström, L., Ilvonen, L., Seppä, H., and Veski, S. (2015). A Bayesian spatiotemporal model for reconstructing climate from multiple pollen records. The Annals of Applied Statistics, 9(3), 1194-1225.

[108] Kim, K. K., Shen, D. E., Nagy, Z. K., and Braatz, R. D. (2013). Wiener's polynomial chaos for the analysis and control of nonlinear dynamical systems with probabilistic uncertainties [historical perspectives]. IEEE Control Systems, 33, 58-67.

[109] Oladyshkin, S. and Nowak, W. (2018). Incomplete statistical information limits the utility of high-order polynomial chaos expansions. Reliability Engineering & System Safety, 169, 137-148.

[110] Pedersen, A. R. (1995). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. Scandinavian Journal of Statistics, 22(1), 55-71.

[111] Aït-Sahalia, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: a closedform approximation approach. Econometrica, 70(1), 223-262.

[112] Nourdin, I. and Peccati, G. (2015). The optimal fourth moment theorem. Proceedings of the American Mathematical Society, 143, 3123-3133.