A Bayesian approach to the interpretation of climate model ensembles

by

Marianna Demetriou

Department of Statistical Science
University College London (UCL)

A thesis submitted for the degree of
Doctor of Philosophy

December 2015

Declaration

I, Marianna Demetriou confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature .....

Abstract

This thesis is concerned with uncertainty quantification when interpreting ensembles of climate models (MMEs), to learn about the true climate. The work improves a recent Bayesian framework for MME interpretation and applies it for first time on real data. The original framework accounts for shared simulator discrepancy from reality. Inference for the true climate is provided through a posterior distribution. The posterior is obtained based on a computationally efficient implementation which assigns estimates to the covariance matrices expressing variability due to different sources of uncertainty in the MME structure. Three framework improvements are considered. Firstly, an improved estimator of the covariance matrix expressing variability due to shared simulator discrepancy from reality is proposed. It relates future to historical discrepancy using bootstrapping from earlier data, instead of subjectively setting a parameter to relate them. The second improvement incorporates prior information about the framework's covariance matrices to account for uncertainty in their values, leading to a fully Bayesian implementation. Two such implementations are introduced, to explore sensitivity of the true-climate posterior to different priors. The third improvement proposes an extended framework which accounts for simulator grouping. A random effects model is also proposed to estimate within-group variability when singleton groups exist. The improvements are assessed through an application on global surface air temperature, using observations and simulator outputs. Results suggest that ignoring simulator grouping and uncertainty in the values of the framework's covariances does not seriously underestimate the uncertainty in inference for real climate. A simulation study is also presented, to compare the performance of the extended framework relative to the simpler. Results suggest that given small shared discrepancy and a well-defined grouping structure, accounting for grouping improves uncertainty quantification in inference for true climate, provided the bias in estimation of shared discrepancy is small.

Acknowledgements

There are no words to describe how thankful I am to my principal PhD supervisor, Prof. Richard Chandler, for his continuous support and guidance during my PhD. He has been always providing valuable feedback, motivation and encouragement. He contributed greatly to my development, both as a researcher but as a person too. His professionalism, kindness and generous personality will form a paradigm to follow for the rest of my life.

I would also like to thank my secondary PhD supervisor Prof. Mark Girolami as well as Dr Ricardo Silva, for their constructive feedback during my Upgrade examination. Additional thanks to Prof. Tom Fearn for his useful advice on my work and the rest of the academic and administrative staff for providing a friendly and supportive environment during my studies in the Department of Statistical Science. I am also grateful to UCL and EPSRC (Engineering and Physical Sciences Research Council) for the financial support during my PhD.

My life the last four years would not be the same without my fellow PhD students. The exchange of knowledge, experiences, support and advice, as well as the various social gatherings, were vital in making this journey exciting and stimulating. Special thanks to my very good friends and neighbours in the office Nayia and Eleftheria, who were always listening to my concerns, advised and tolerated me during my bad days.

I am also very grateful to all my friends who supported me during my PhD life, by ensuring that I have a good work-life balance and many fun moments. Special thanks to my precious friends Myria and Yiannis, with whom I shared common experiences of PhD life, endless discussions (serious and not), travelling, moving and sharing of flats and most importantly, who have been always there for me in good and bad times. Last but not least, I am very thankful to my partner Vassilis for two years of endless support, understanding, encouragement and love.

I would have never made it this far if it wasn't for my parents. They have always supported my decisions in many ways, believed in me, cheered me up in difficulties and celebrated with me in successes. There are no words to express my gratitude

to them for their unconditional caring and love. I am also thankful to my cousin Michalis, who made my weekends more cheerful as my neighbour in London. Finally, I am mostly grateful to my beloved sister Yiota, for constantly being by my side in this journey as my closest family member in the UK, my best friend and an amazing flatmate.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Modern computing power provides the opportunity to simulate systems of ever-increasing complexity, one of which is the climate system. Climate simulators are complex computer models which mimic the behaviour of the climate system. Inference from simulators is used to inform high-impact strategies such as climate change adaptation and risk assessment of natural hazards. An emerging area of research aims to quantify the uncertainties associated with simulation of complex systems. A large source of uncertainty arises from the choice of simulator, referred to as "model uncertainty". A widely used approach for quantifying model uncertainty is to combine information from different simulators and use it to form a so called "multi-model ensemble" (MME) to learn about real climate.

Interpretation of MMEs is a non-trivial task. Various approaches have been developed, although few of them account satisfactorily for all relevant features of MMEs. Inter-simulator dependence and the presence of shared simulator discrepancies from reality are two such features that are rarely handled explicitly. This motivates the development of frameworks which make explicit realistic assumptions about the MME structure and also quantify representatively the underlying uncertainties.

This thesis improves a recently introduced probabilistic, Bayesian framework for uncertainty quantification in MMEs which is introduced in Chandler (2013). The framework provides inference about the real climate using observations and simulator outputs. It explicitly accounts for the presence of shared simulator discrepancies from reality and exploits the strengths of each ensemble member while simultaneously discounting their weaknesses in simulating true climate. The improvements of the framework considered in this thesis are motivated by existing challenges in MME interpretation, focused in two particular directions. The first relates to quantification

of uncertainties in the MME structure and how this is incorporated in the implementation of the framework. The second is relevant to assumptions about the relations between the ensemble members in the MME structure.

The contributions of this thesis are outlined in the next section.

1.2 Thesis statement

This thesis improves the framework of Chandler (2013) for quantifying uncertainty in MMEs. The improvements are summarised in three main points which outline the contributions of the thesis. They are introduced below:

- Improved estimation of shared simulator discrepancy from reality.

  The framework of Chandler (2013) explicitly accounts for the presence of shared simulator discrepancy of simulators from reality in the MME structure. Variability due to shared simulator discrepancy is expressed in the framework through a covariance matrix $\mathbf{\Lambda}$. The computationally efficient "poor man's" (PM) implementation introduced in Chandler (2013), assigns estimates to the covariance matrices expressing variability due to the different sources of uncertainty. However, it is not trivial how to estimate $\mathbf{\Lambda}$ for future climate, due to the absence of observations. Therefore, a parameter K is introduced and is subjectively chosen to relate historical to future shared simulator discrepancy from reality.

  In order to avoid the limitation of the subjective choice of K, a new more robust estimator is proposed in this thesis by exploiting information from earlier data (simulator outputs and observations), through the method of bootstrapping. The methodology is illustrated through a proposed "revised poor man's" (RPM) framework implementation (Section 3.5.1) using simulator outputs and observations, for inference in global surface air temperature.

- Incorporate prior information about the sources of uncertainty in the levels of the MME structure.

  The PM implementation in Chandler (2013) assigns estimates to the unknown covariance matrices expressing variability due to different sources of uncertainty in the framework. Although this procedure is computationally convenient, it has the limitation of ignoring the uncertainty in the values of these unknown covariance matrices.

  To overcome the limitation, this thesis exploits prior information about the sources of variability in the framework and therefore assigns prior distributions

instead of estimates to the relevant covariance matrices, leading to a fully-Bayesian implementation. Two such implementations are proposed based on different prior choices, to explore sensitivity of inference about true climate to different prior choices.

The proposed fully-Bayesian implementations are applied to an analysis of global surface air temperature (Sections 3.5.2-3.5.3). Results are also compared to the computationally simpler RPM implementation of Section 3.5.1, to explore the impact of failing to account for uncertainty in the estimation of covariance matrices values in the latter.

- Develop an extended framework which explicitly accounts for potential simulator grouping in the MME.

  It is widely recognized that there are similarities between simulators (see Sections 2.4,4.1). Sources of these similarities are for example the sharing of code for their design, or the common modelling centre to which they belong. The framework of Chandler (2013) however, does not explicitly account for the presence of potential simulator grouping arising from similarities between the ensemble members.

  An extended framework is therefore proposed in Chapter 4, which adds levels to the hierarchical framework of Chandler (2013), to account for potential nested grouping of the ensemble members. The extended framework in principle enables an arbitrary number of levels and accommodates unbalanced and sparse ensemble structures. Additionally, a random effects model is proposed (Section 4.6) to allow sharing of information among groups based on exchangeability assumptions about the MME structure, in order to estimate variability arising from simulator grouping. Implementation of the extended framework is also illustrated in an application to global surface air temperature, where simulator grouping is determined from expert judgement. Results are compared to those from implementations of the simpler framework, in order to explore the effect of accounting for simulator grouping in inference about the true climate (Section 5.3).

1.3 Outline

The thesis is organised as follows: Chapter 2 firstly provides an overview of the different sources of uncertainty in climate modelling and the approaches for uncertainty quantification. Particular emphasis is given on the issues relevant to multi-model

ensemble (MME) interpretation (Section 2.4) and the frameworks developed to interpret MMEs (Section 2.5). Next, the framework of Chandler (2013) is introduced in Section 2.6.

Chapter 3 illustrates the first two of the contributions in this thesis. This is achieved by performing three proposed implementations of the framework of Chandler (2013) for uncertainty quantification, in an application to global surface air temperature. The datasets used in the application are introduced in Section 3.2. The three proposed implementations are described in Sections 3.5.1-3.5.3, after introducing the statistical model describing simulator outputs and observations (Section 3.3) and some specifications required for the implementations (Section 3.4). The results from the three proposed implementations are presented in Section 3.6.

Chapter 4 introduces the extended framework which accounts for simulator grouping. The conceptual framework is described in Section 4.2 and the mathematical analysis based on Gaussian distributional assumptions is shown in Section 4.3. The challenges in the implementation are discussed in Section 4.4, followed by a review of the existing relevant techniques for dealing with them in Section 4.5. The proposed random effects model for estimating within-group variability is presented in Section 4.6. Model inference is considered in Section 4.7. Finally, an algorithm for implementing the framework is introduced in Section 4.8.

Chapter 5 firstly presents a simulation study (Section 5.2) which explores the performance of the extended framework relative to the simpler framework (which does not account for simulator grouping), for different settings. Then, an application to temperature data is presented in Section 5.3, where the extended framework is implemented and the results are compared to those from implementations of the simpler framework.

Finally, Chapter 6 is a discussion about the work presented in this thesis. The main conclusions from the contribution of the thesis are summarized in Section 6.1, followed by a discussion about relevant potential future work, in Section 6.2.

Chapter 2

Climate modelling and uncertainty

This chapter reviews the literature relevant to quantification of uncertainty in climate modelling. Particular emphasis is given to the framework of Chandler (2013), which this thesis improves.

Firstly, the different sources of uncertainty which occur in climate modelling are discussed in Section 2.1. Then, two schemes for combining information among climate models, namely the "Perturbed physics ensembles" and "Multi-model ensembles" are presented in Sections 2.2-2.3 respectively. The issues regarding interpretation of MMEs are considered in Section 2.4. Next, some of the existing frameworks for MME interpretation are reviewed in Section 2.5, ranging from explicit weighting schemes (Section 2.5.1) which can be either deterministic (Section 2.5.1.1) or probabilistic (2.5.1.2), to model-based approaches (Section 2.5.2). From the model-based approaches, frameworks for partitioning uncertainty are reviewed in Section 2.5.2.1, and those based on the truth-plus error interpretation and exchangeability assumptions are discussed in Sections 2.5.2.2 and 2.5.2.3 respectively. Finally, Section 2.5.2.4 considers how shared simulator discrepancy from reality is treated in existing frameworks.

The framework for quantifying uncertainty (Chandler, 2013) which this thesis extends is presented in Section 2.6. The conceptual framework is introduced in Section 2.6.1 and the mathematical analysis based on Gaussian distributional assumptions is provided in Section 2.6.2. Finally, the PM implementation introduced in Chandler (2013) is described in Section 2.6.3. Section 2.7 summarizes the chapter.

2.1 Sources of uncertainty in climate modelling

Effects of climate change have become a global issue nowadays, concerning governments worldwide. Lack of water availability, flooding and disease are only some of the

potential impacts of climate change. Political authorities are developing national and international strategies to deal with these issues. Examples include the UK government funded climate projections tool UKCP09 (UK climate projections) (UKCP09, 2014), as well as the United Nations Framework Convention on Climate Change (UNFCCC), an international environmental treaty aimed in reducing the anthropogenic greenhouse gas emissions (UNFCCC, 2014). The severity of the situation has been addressed since 1992, during the United Nations Conference on Environment and Development. In Principle 15 it is stated that "where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation" (UN, 1992). This also illustrates the involvement of the scientific community in handling the issues raised by climate change and recognizes the uncertainty encountered in providing scientific evidence about the environment. The question now arises as to the way this scientific evidence is provided through climate modelling. According to Cox and Stephenson (2007), advice from climate scientists has become necessary in order for society to prepare for the upcoming climate change and avoid its potential dangerous impacts.

Providing this evidence in a way that is representative of the actual climate is not a trivial task. Lucarini (2002) highlights the complexity of the climate system, which makes it difficult to be interpreted by single elements and processes. According to Knutti (2008), in order to guide decision-making of policy-makers and provide advice on high-impact strategies, there is a need to "rely on complex numerical computer models ". Climate models, alternatively named as "simulators", aim to capture the main characteristics of the physical and chemical processes relevant to climate, in order (as far as possible) to reproduce the current climate and also provide future projections. According to Collins et al. (2012), a climate model is a "mathematical description of the climate system where the output is controlled by internal model parameters and external forcing (e.g. greenhouse gases)". Climate models can be regarded as aiming to provide some understanding of the ongoing processes in an attempt to reduce complexity of the climate system. Depending on the scientific question of interest, the observational restrictions as well as the constraints in computational resources, several types of simulators are available, ranging from the "one-dimensional energy balance models through models of intermediate complexity to fully coupled atmosphere-ocean general circulation models (AOGCMs)" (Sansó et al., 2008; Knutti, 2008).

"General Circulation Models" (GCMs) are considered as "the most advanced tools currently available for simulating the response of the global climate system to increasing greenhouse gas concentrations" (IPCC-DDC, 2013). Such models are defined by

coupled sets of differential equations discretized onto a grid, with parametrisations which are often used to represent small-scale physical processes (e.g. convection within clouds) that cannot be modelled due to their fine resolution. The equations are propagated forwards numerically from some initial conditions and with some prespecified "forcing", e.g. solar radiation, greenhouse gases and aerosols (Stephenson et al., 2012). Some consider GCMs as being the only tools (possibly together with nested regional models) that provide geographically and physically adequate estimates of regional climate change (IPCC-DDC, 2013). Although being computationally demanding during both their design and their implementation, the increase in computer capabilities nowadays provides the opportunity for the GCMs to include additional processes and increase the resolution, while retaining the same computational time (Chandler et al., 2010).

Despite the ability of GCMs to adequately capture and reproduce complicated climate processes, they are still climate *simulators*; in other words they cannot perfectly represent reality (Chandler et al., 2010; Kennedy and O'Hagan, 2001; Smith, 2002). Knutti (2008) distinguishes between the model components which are based on first principles, such as equations of motion and conservation of energy, and those which cannot be fully understood and therefore need to be simplified using parametrizations. For the former, there is capability for improvement by increasing computational capacity. However, for the parametrized processes, the improvement is limited due to our lack of understanding rather than the computational restrictions. Climate simulators can therefore be considered as "credible approximations to the description of the climate system given our limited understanding, the lack of complete observations and the simplifications that need to be made due to computational constraints" (Knutti, 2008). The inevitable uncertainty in simulating future climate arises from various sources. Chandler et al. (2010) categorize the sources of uncertainty into "structural", "input" and "parametric" uncertainty.

"Structural uncertainty" is defined as the uncertainty caused by approximating and simplifying actual climate in order to be able to simulate it. According to Lucarini (2002), there might be particular climate characteristics that are not yet discovered, such as non-linear behaviour, which can sometimes lead to rapid climate changes. Furthermore, there is still a high degree of uncertainty in determining natural emissions of greenhouse gases and describing the carbon cycle. Additionally, the fact that various physical processes occur at small time-scales, thus requiring averaging of known properties over larger scales, can be considered as another source of structural uncertainty (IPCC-DDC, 2013).

On the other hand, "input uncertainty" (Chandler et al., 2010) refers to the uncertainty caused by imperfect knowledge of the quantities that describe the state of the

system at particular time points. Those quantities are used as inputs in the climate simulators. They can be values of earth-system components that are regarded as fixed in the simulator, values of historical forcing variables or values of future forcing variables, expressed in terms of emissions scenarios. Additionally, the choice of initial values for the dynamical equations that are implemented in simulations is another source of input uncertainty (Chandler et al., 2010). The chaotic structure of the climatic dynamics results in a high degree of climate variability. As a result, GCM outputs can be highly sensitive to the choice of initial conditions (Stainforth et al., 2005; Murphy et al., 2004; Stephenson et al., 2012). An example of input uncertainty is considered in Lucarini (2002), who refers to the important uncertainties surrounding the values assigned to climate sensitivity and the efficiency of the oceanic heat uptake from the atmosphere.

Finally, "parametric uncertainty" is the uncertainty invoked by specifying values for parameters in the simulators, either because they are not perfectly known, or because they do not have an analogue in the actual climate. This relates to parameters for sub-grid scale modelling for example, for which computation is infeasible (Chandler et al., 2010).

In view of the issues raised above, a natural question that arises is: To what extend should model outputs be trusted for future projections? Various studies exist in the literature which aim to answer this critical question (Räisänen, 2007; Knutti and Sedláček, 2013; Tebaldi et al., 2011). Findings from these studies suggest that although simulator projections can be informative about future climate, in order to interpret the available information in a useful way, it is necessary to perform uncertainty quantification. Räisänen (2007) uses three criteria for assessing model reliability: model agreement with observed historical climate, ability of models to simulate current climate and inter-model agreement in future projections. The author concludes that although the evidence supports the use of climate models to inform about future climate, model reliability is not yet quantified, which urges the need for uncertainty estimation in simulator projections. Tebaldi et al. (2011) argue that testing for inter-model agreement makes sense only if the projections are not within the bounds of internal variability, which again suggests the need to quantify the effect of internal variability, relative to model uncertainty. In other words, it is non-trivial to quantify model agreement in the presence of large internal variability. Furthermore, Knutti and Sedláček (2013), in a comparison of simulator performance in simulating current temperature and precipitation change in two phases of the Coupled Model Intercomparison Project (CMIP3 and CMIP5), observe agreement in local model spread between the developed CMIP5 generation and the earlier CMIP3. This implies that the increase in development of simulators is not guaranteed to improve inter-

model agreement and therefore model uncertainty should not be ignored.

The non-effective assessment of uncertainty can lead to wrong interpretations of model predictions (Rougier et al., 2013). According to Rougier et al. (2013) and Lucarini (2002), the pressure brought by policy makers, who expect to get quick and clear scientific evidence for implementing policies, re-directs the initial scope of climate models from being largely explanatory about representing actual climate to being predictive. According to Oreskes (2000), the quality of climate simulators should not be solely assessed based on their predictive capability and attention should be paid on simulator outputs, the theory and the empirical information behind simulator design. Reilly et al. (2001) contend that it is essential to provide a thorough representation of the uncertainties in simulating the climate, in order to guide decision-makers towards reducing risks related to climate change.

Initially, in order to characterize parametric uncertainty, this was performed by running and analysing Perturbed Physics Ensembles (PPEs), which are discussed next.

2.2 Perturbed physics ensembles (PPEs)

According to Tebaldi and Knutti (2007), PPEs are sets of simulations including a single model, where each simulation is run under different choices for various parameters. There are many climate-related examples in the literature: see, for example, Collins et al. (2011); Stainforth et al. (2005); Murphy et al. (2004); Sansó et al. (2008). PPEs are useful for model evaluation and calibration. The consistency of each ensemble member with observations is investigated (evaluation), which then informs the choice of model parametrisation (calibration). This is often performed by defining a statistical model defined as an "emulator" (Sansó et al., 2008; Williamson et al., 2014) which is computationally cheaper than the climate simulators, to represent the model output. The emulator, defined to be a Gaussian process (Sansó et al., 2008), is then used to constrain the parameter space based on matching simulator output to historical observations, a technique known as "history matching" (Williamson et al., 2013, 2014). This is done in a Bayesian framework, where historical observations are expressed as a function of the emulator, which itself is a function of the climate model parameters of interest (Sansó et al., 2008; Murphy et al., 2007). Another attempt to quantify parametric uncertainty using a single simulator was performed in the UKCP09. Information is provided from a collection of simulations from a particular GCM, where the basic parameters have been varied "within a range of uncertainty determined by expert judgement" (Chandler et al., 2010).

Tebaldi and Knutti (2007), in reviewing the PPE approach, recognize its substantial contribution to quantifying parametric uncertainty. They consider it as an efficient way of exploring the parameter space by varying a single parameter at a time, or as a successful way of producing probability density functions (PDFs) of climate quantities, by varying multiple parameters simultaneously. On the other hand, they consider the PPE approach as being "limited in its ability to capture the full range of uncertainties". One reason for this is the complexity of the climate system, as a result of which any attempt to describe the climate using a single simulator is over-simplified (Chandler et al., 2010). It is true that the existence of structural uncertainty in addition to computational constraints often creates limitations to the climate simulator, regardless of the choice of parametrisation.

The above limitations highlight the additional need to characterize the collective uncertainty in climate modelling, as it evolves from its various sources. A well-established approach to achieve that is through the use of a collection of climate simulators, defined as a "Multi-Model Ensemble" (MME). It is described in the next section.

2.3 Multi-model ensembles (MMEs)

MMEs are constructed based on combining information collectively from a set of climate simulators, instead of considering the models one-at-a-time (Rougier et al., 2013). Each member is considered as having the "best combination of resolved and parametrized processes, as limited by practical considerations of run time" (Collins et al., 2012).

The use of MMEs in climate change projections has gained increasing recognition in the scientific community (Räisänen, 2007; Knutti, 2008; Knutti and Sedláček, 2013). Knutti (2008) justifies the use of MMEs by observing that although models might differ considerably in structure, their consistency with observations within a plausible range of uncertainty makes them all potentially "useful representations" of reality. Furthermore, the use of MMEs allows for simultaneously accounting for initial value and forcing scenario uncertainty in addition to model uncertainty. This can be achieved by merging variants of the same simulators but with different initial values (Hagedorn et al., 2005) or forcing scenario (Knutti and Sedláček, 2013) in the MME.Räisänen (2007) emphasizes the importance of using MMEs by arguing that inter-model variability potentially gives the "most meaningful" estimates of the existing uncertainty.

Climate scientists often use the mean of the ensemble outputs as an estimate of the

true climate state, under the assumption that the ensemble members are i.i.d (Flato et al., 2013, Chapter 9.2.2). The associated estimation uncertainty is then determined as the standard error of the ensemble mean. In mathematical terms, denoting the true unknown climate state by x_0 and the output of the ensemble member i by x_i for a MME of m members, then x_0 is estimated as

$$\hat{x}_0 = \frac{\sum_{i=1}^{m} x_i}{m}. \tag{2.1}$$

Based on the formula for the standard error of the sample mean, the uncertainty u in estimation of true climate is determined as

$$u = \frac{s}{\sqrt{n}}, \tag{2.2}$$

where s is the sample standard deviation of the ensemble outputs, defined as $s = \sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2/(m-1)}$, with $\bar{x} = \sum_{i=1}^{m} x_i/m$ being the ensemble mean.

However, this approach is often characterised as "naive", especially regarding the assumption of *i.i.d.* ensemble members, for reasons discussed in the next sections. In general, although the use of MMEs is "promising" in terms of more robust quantification of uncertainty in climate modelling, their interpretation is not straightforward and requires closer collaboration between climate scientists and statisticians (Williams et al., 2013). Interpretation of MMEs has been an active research area for statisticians in climate science (Chandler et al., 2010; Tebaldi et al., 2005; Jun et al., 2008; Giorgi and Mearns, 2002). The importance of developing formal statistical frameworks for MME interpretation has been widely recognized (Knutti, 2008; Heideman et al., 1993; Williams et al., 2013). Williams et al. (2013) highlight that the "development and evaluation of appropriate statistical frameworks is required in order to create credible and reliable probability distributions of real-world observables from ensembles of climate predictions". Furthermore, Heideman et al. (1993) stress the need to allocate resources on interpretation of the available information from simulators, rather than solely focusing on increasing the amount of information. Knutti (2008) adds to this the requirement for additional resources on how to make optimal use of the available information and characterization of the uncertainties in large amounts of data. The next section reviews some of the issues related to MME interpretation.

2.4 Issues in multi-model ensemble (MME) interpretation

When it comes to MME interpretation, several different techniques exist which are routinely used to analyse MMEs. These vary from schemes which explicitly assign weights to the ensemble members based on various heuristic criteria (Section 2.5.1), to model-based approaches (Section 2.5.2) which are constructed based on assumptions about the MME structure. In explicit weighting schemes, the weights can be either deterministic, such as simple or weighted averages (Section 2.5.1.1) or probabilistic, based on frequency counting of event occurrences using ensemble outputs (Section 2.5.1.2). On the other hand, in model-based approaches, weights are implicitly assigned to ensemble members, usually through a Bayesian framework where a statistical model describing simulator outputs is defined and its parameters are assumed to be random. Inference about the true climate is then obtained from the posterior distribution of model parameters.

The question of how to weight ensemble members in a MME is a natural one (Williams et al., 2013). Knutti (2010) highlights the difficulty in determining weights in MMEs, by characterizing it as a process of deciding about various potential strategies based on incomplete understanding of a very complex system. One issue relates to the way MMEs are assembled. According to Tebaldi and Knutti (2007), the MMEs are often characterised as "ensembles of opportunity". The size and composition of the MMEs is determined by various non-scientific factors, like the interest of modelling centres, funding and computing availability. As a result, the sampling of ensemble members when forming a MME is not random and inference from different MMEs can vary considerably, even if prior information about climate is unchanged.

Another key issue is the fact that the ensemble members are not independent (Tebaldi and Knutti, 2007; Chandler, 2013; Rougier et al., 2013). Inter-simulator similarities arise from various sources, such as similar resolution between simulators, which determines whether they are capable of reproducing similar processes or not, the use of identical theoretical arguments for the parametrizations and shared errors caused by similarities in parametrizations (Tebaldi and Knutti, 2007). Adding to this the relatively limited number of simulators, there is no guarantee that a collection of them will capture the full range of modelling uncertainty (Räisänen, 2007; Chandler, 2013). Additionally, numerical methods, grids and even entire model components are often shared between modelling groups. Finally, the existence of different variants of particular simulators is another source of dependence within a MME (Tebaldi and Knutti, 2007). This inter-model dependence suggests that model spread is not

necessarily a reasonable measure for uncertainty quantification (Knutti and Sedláček, 2013). A further consequence is that all climate simulators share discrepancies with the actual climate system (Chandler, 2013; Räisänen, 2007; Sanderson and Knutti, 2012), leading to the conclusion that simulators are likely to be much more similar to each other than each of them independently is to reality (Rougier et al., 2013). This point is also made by Knutti (2008), who claims that interpretation of MMEs is often based on the misleading assumption that simulators are all centred on a "perfect model of the climate system". The presence of inter-simulator similarities is often not accounted for explicitly in existing frameworks for MME interpretation. The proposed framework of Chapter 4 deals with this, since it enables potential simulator grouping in the MME structure.

An additional consideration regarding MME interpretation is that simulators are often weighted based solely on calibration from observations. This is problematic in long-term future projections in which case it is not possible to calibrate models, due to the absence of observations. Additionally, there is no guarantee that the skill of a simulator in reproducing current climate will be the same in the future, mainly due to the presence of internal variability and uncertainty in external forcing and observations (Räisänen, 2007). Furthermore, Tebaldi et al. (2011) raise concerns about whether it is reasonable to reject or downweight an ensemble member only because it disagrees with the projections from the rest of the ensemble members. The authors note that the simple criteria used for mapping model agreement in MMEs can often be misleading, since they interpret the lack of model consensus as lack of information from the ensemble. Also, many studies highlight that different simulators have different credibility in representing different aspects of the climate system, which adds another challenge to MME interpretation and in particular to the weighting of each ensemble member when different climate variables are studied simultaneously (Chandler, 2013; Knutti, 2008; Räisänen, 2007). Finally, Sanderson and Knutti (2012) consider the different plausible assumptions about the MME structure as another important issue in probabilistic interpretation of MMEs, where model-based approaches are developed. Of the assumptions reviewed, the authors conclude that none is entirely satisfactory and hence it is not clear that a user-relevant probabilistic interpretation can be justified.

It is therefore clear that careful thought is needed regarding how to best exploit the collective amount of information from MMEs in order to quantify reliably the full range of the underlying uncertainties. Section 2.5 reviews the different frameworks.

2.5 Existing frameworks for interpreting MMEs

Different frameworks for MME interpretation exist in the literature, which attempt to address the issues discussed in Section 2.4. Many of them propose explicit weighting schemes (Section 2.5.1) for the ensemble members, based on heuristic criteria. These can be either deterministic (Section 2.5.1.1), or probabilistic (Section 2.5.1.2), in which case PDFs of the climate quantities of interest are produced based on frequencies of events in the ensemble members. On the other hand, various limitations of explicit weighting schemes gave rise to an emerging class of probabilistic frameworks which provide a statistical representation of the MME structure using model-based approaches (see Section 2.5.2). Note that this thesis is concerned with improvements of a model-based framework, and is therefore more relevant to the content of Section 2.5.2. The reader can refer to Section 2.5.1 however for a review of the existing explicit weighting schemes and the underlying limitations of these approaches.

2.5.1 Explicit weighting schemes

2.5.1.1 Deterministic weighting schemes

Many frameworks for MME interpretation consist of weighting schemes assigning deterministic weights to the ensemble members, based on different heuristic criteria to determine their quality in the ensemble (Palmer et al., 2005; Giorgi and Mearns, 2002; Christensen et al., 2010; Kharin and Zwiers, 2002; Doblas-Reyes et al., 2005). The simplest of those weighting schemes assigns equal weights to the ensemble members, leading to simple averaging of the simulator outputs (see Expressions (2.1)-(2.2)) to learn about reality (Weisheimer et al., 2009; Hagedorn et al., 2005; Barnston et al., 2003; Palmer et al., 2005). The motivation behind simple averaging is that it provides a consistent way of combining information from the ensemble, avoiding the subjective choice of metric for assigning unequal weights. Since different weighting schemes can yield different results (Lenderink, 2010; Kjellström et al., 2010), the arbitrary choice of the most "appropriate" weighting scheme becomes another source of uncertainty, according to Kjellström and Giorgi (2010). On the other hand, this approach is often treated with scepticism, since it is known that simulators are different in their skill, depending on the climate variable of interest while also having similarities in structural errors and other assumptions (Knutti, 2008; Doblas-Reyes et al., 2005). A study presented in Jun et al. (2008) for simulated mean and trend of temperature among various simulators reveals that the simulators have different credibility in simulating mean temperature, compared to the corresponding temperature trend.

Leith and Chandler (2010) suggest that, in order to quantify uncertainty in a coherent way, it is crucial that for each quantity of interest, the strengths of each simulator are exploited, while at the same time its weaknesses are discounted.

As an alternative, many frameworks consider weighted averages, where weights are mainly determined by comparing simulator outputs with historical climate observations and the ensemble consensus (Giorgi and Mearns, 2002; Christensen et al., 2010; Palmer et al., 2005). Krishnamurti et al. (2000) assign weights to ensemble members on a test period, by applying multiple regression of the ensemble forecasts towards observations in a training period. The weights vary depending on the climate variable of interest, the geographical location and the vertical resolution of the climate models. The forecast skill of the weighted ensemble for the test period is shown to be superior than the simple averaging ensemble weights. The "reliability ensemble averaging" (REA) method introduced by Giorgi and Mearns (2002), assigns weights to simulators based on their performance in reproducing current climate and in the degree of deviation of each simulator output from the corresponding average over all simulators. Another example is the weighting scheme introduced by Christensen et al. (2010), where weights based on different criteria are combined in multiple ways, in order to obtain single weights for each ensemble member. The rationale behind this technique is that a "trustful" simulator should have a reasonable performance in a range of different metrics, in order to reduce potential effects of systematic model biases. However, the authors recognize that some schemes are not proved to be much better than simpler ones, that not all types of model quality assessment are captured and that there is a significant amount of subjectivity in the choice of combination of metrics.

Similarly, Tebaldi and Knutti (2007) suggest that ensemble members should be judged based on their ability to represent the full structure (i.e. mean and variability) of a time series for the climate quantity under study. Based on this approach, Elvidge et al. (2013) construct an ensemble average by weighting each ensemble member based on a skill metric which assesses simulator performance in reproducing the mean and variance of an observed time series of atmospheric density. The comparison of the performance of the weighted average to the unweighted, suggests that the former performs better in reproducing mean and variability for most of the time period studied. However, the method indicates that in particular periods when a model has serious bias, simple averaging has a better performance. According to Weigel et al. (2010), in cases where the model weights are not representing the underlying uncertainties, simple averaging might perform favourably compared to weighted average.

It is therefore obvious that there is no single deterministic weighting scheme that is proved to be superior both in realistically representing the ensemble structure

and improving performance in reproducing current climate. The situation now is characterised by lack of consensus in the method of determining simulator weights. The limitations of the existing deterministic weighting schemes led to the development of probabilistic weighting schemes, by treating weights "as defining a probability distribution" Chandler (2013). This is achieved by constructing probability density functions (PDFs) for the climate quantities of interest, based on frequency counting of probabilistic event occurrence in the ensemble members. The ensemble members are then weighted by assessing their skill on reproducing the observed climate. Some examples of the existing probabilistic weighting schemes are reviewed next.

2.5.1.2 Probabilistic weighting schemes

Déqué and Somot (2010) use an ensemble of 17 members to construct PDFs for temperature and precipitation for a particular reference period. The ensemble PDF is constructed by a weighted sum of the model-specific PDFs, each produced by counting the proportion of daily occurrences of temperature and precipitation events of interest for a single ensemble member. The weights are then determined based on consensus of model outputs with observations. In comparison with the reference climate PDF, the method is not proved to be worse or superior to equal weighting. The authors note also that their method explores only some of the aspects of a weighting system and therefore should not be used to guide decision-making and acknowledge the challenge of interpreting probabilistic weights based on multiple diagnostics.

Watterson (2008) produce probabilistic projections of the net climate change based on GCM outputs from a MME, by constructing a joint PDF for the product of two factors, expressing respectively "global mean warming" and "the standardized regional change per degree of warming", which are assumed to be independent. The joint PDF can then be derived from the product of the PDFs of the two factors. The latter are constructed in different ways, by fitting continuous distributions to the weighted sum of the ensemble outputs. The weights are themselves determined by calculating the "skill" of each GCM in reproducing patterns in observed regional climate changes. The authors however recognize the subjectivity in the choice of the weights as a "caveat" in their method.

Although probabilistic weighting schemes attempt to provide an estimate of the underlying uncertainty in climate projections, their interpretation is yet not clear. According to Chandler (2013), it is unclear how arbitrarily chosen weights can be treated as probabilities. Additionally, due to the existence of only a few climate simulators, treating the corresponding weights as defining a probability distribution will underestimate uncertainty, due to underestimating the range of possible modelling

decisions. Besides, most existing deterministic and probabilistic weighting schemes derive their weights by assessing forecasting skill for historical and current climate. However, as already discussed in Section 2.4, it is not enough to consider earlier forecasting skill, especially when interest is on making future projections based on combining information from a MME.

It is widely recognized that formal statistical frameworks are essential, mainly to enable explicit quantification of the uncertainties associated with climate projections (Sansom et al., 2013; Zappa et al., 2013; Min and Hense, 2007; Buser et al., 2009; Siegert et al., 2015). Siegert et al. (2015) argue that the ignorance of uncertainty in frameworks based on heuristic weighting schemes becomes more serious when it comes to short-term predictions or small ensemble sizes, in which case the underlying uncertainties are "substantial". Additionally, according to Chandler (2013), the employment of a formal statistical framework for ensemble projections allows for a transparent and coherent representation of the underlying assumptions and limitations. Section 2.5.2 reviews some of the existing model-based approaches used for MME interpretation.

2.5.2 Model-based approaches

2.5.2.1 Frameworks for partitioning uncertainty

Ideally, an ensemble of different runs, forcing scenarios and models could be considered simultaneously to quantify the relevant uncertainties. However, the fact that GCMs are computationally expensive to run, generates the need to identify the dominant sources of uncertainty in order to decide on the allocation of resources for characterizing the uncertainty in future projections of climate. Hawkins and Sutton (2009) attempted to partition total uncertainty by smoothing temperature predictions using a polynomial trend model and to quantify the various sources of variability based on heuristic criteria. A more formal statistical approach for partitioning variability is the analysis of variance (ANOVA). This method is based on decomposing variability from an ensemble of climate simulators, forcing scenarios and initial conditions into different contributions, and identifies the dominant sources of uncertainty. According to Yip et al. (2011), this approach allows for flexibility in the model assumptions, since it does not require assuming constant internal variability in time, or a particular form for the noise and trend distribution. Various studies have performed uncertainty quantification using ANOVA-based approaches (Northrop and Chandler, 2014; Sansom et al., 2013; Yip et al., 2011; Geinitz et al., 2015; Zappa et al., 2013).

Results in most cases suggest that for short term predictions, model uncertainty

dominates relative to scenario and internal variability, in contrast to long-term predictions where scenario uncertainty is the dominant source. Sansom et al. (2013), by introducing a nested family of ANOVA models consider the most appropriate model based on assumptions about model consensus relative to reality and its magnitude relative to internal variability. They conclude that when model uncertainty is sufficiently small relative to internal variability, estimation of model differences can be ignored. The two-way ANOVA model with interactions is then reduced to the simpler 2-way additive ANOVA model, which assumes that all models are indistinguishable in simulating climate response, with common internal variability. This implies that any uncertainty in simulating the climate response is attributable to the common internal variability. In this case however, the authors assume that any shared simulator discrepancy from reality is constant between historical and future scenarios, which is not always realistic. This issue is handled in Section 3.5.1, where a robust estimator of shared simulator discrepancy from reality based on bootstrapping is proposed and which forms the first of the three contributions of this thesis. In a similar spirit, Zappa et al. (2013), by calculating the signal-to-noise ratio, i.e. the ratio of mean climate response over internal variability, observe that, provided model uncertainty is small relative to internal variability, models can be informative even when the signal-to-noise ratio is small.

According to Northrop and Chandler (2014), the ANOVA designs can be regarded as a preliminary analysis for guiding the design of experiments involving MMEs by identifying the dominant sources of uncertainties that need to be accounted for. However, for multiple-impact experiments such as the CMIP5, it is useful to develop more detailed statistical frameworks which make explicit assumptions about the MME structure. These frameworks, which are mostly Bayesian, implicitly assign weights to the ensemble members, based on a posterior distribution which enables inference about the true climate (Tebaldi et al., 2005). The Bayesian approach allows incorporating prior information about the various sources of uncertainty in the MME structure. This can be achieved either by assigning estimates or distributions to variability due to the different sources of uncertainty. Both approaches for quantifying the various sources of uncertainty are illustrated in the temperature application of Chapter 3, under three proposed implementations (Sections 3.5.1-3.5.3).

2.5.2.2 Truth-plus-error interpretation

One very popular interpretation of the MME structure is the so called "truth-plus-error" interpretation (Sanderson and Knutti, 2012). The ensemble members are assumed to be sampled from a population of models which are centred "about the true

climate" (Sanderson and Knutti, 2012), or in other words, each model is assumed to be an approximation of reality, with some noise. This interpretation considers each ensemble member as an independent approximation of reality. This section reviews some frameworks based on the truth-plus-error interpretation.

Tebaldi et al. (2005) assume that historical and future temperature simulator outputs are symmetrically distributed around the true temperature for each period respectively. Each simulator output deviates individually from reality. Future deviation is assumed to be proportional to that in the historical period, by a factor which is common to all simulators, due to the limited number of data. Historical observations are also assumed to be centred on reality, with individual variability. The model is formulated by assigning Gaussian distributions to the simulator outputs and observations. A Bayesian framework is deployed, using uninformative conjugate priors for the model parameters, apart from variability of observations, which is fixed using estimates of natural variability. The joint posterior distribution of temperature change is then derived, for different seasons and regions. A similar approach is followed in Tebaldi and Sansó (2009), for deriving joint PDFs of temperature and precipitation change. A hierarchical Bayesian framework is deployed in this study, by assigning distributions to the prior parameters to account for uncertainty in their values. Buser et al. (2009) additionally account for time-varying additive and multiplicative model biases in their Bayesian framework, considering that model biases vary with climate state. The authors also stress the importance of transparent prior judgements, in order to allow for discrimination of model biases from climate change. The issue of incorporating transparent prior information for to the various sources of variability in the MME structure is addressed in the proposed fully Bayesian implementations of Sections 3.5.2-3.5.3, which illustrate the second of the three contributions of this thesis.

Greene et al. (2006) generate projections of regional temperature change for the 21^{st} century, by adopting a Bayesian linear model fitted to observations and simulator outputs of the 20^{th} century. Different model variants are considered, all of which assume that observations are centred on the consensus of simulator outputs, the latter being expressed as a weighted sum of the outputs. This assumption however, ignores the existence of shared simulator discrepancies from reality, which is known to exist (see Section 2.4) . The "weights" are assigned prior distributions, which vary depending on whether they are assumed to be independent *a priori* or correlated. Results reveal an improvement in model fit for the 20^{th} century when structured correlation is assumed between the model "weights". This is unsurprising, considering the known similarities between ensemble members (see Section 2.4). Uncertainty in projections is however comparable to that of the unweighted mean of ensemble

members. Min and Hense (2007) perform Bayesian Model Averaging (BMA) by weighting each ensemble member by its Bayes factor. The latter is calculated relative to a reference model, which has mean equal to the observations and variability equal to that of a pre-industrial control run of the ensemble member. Similarly to Greene et al. (2006), this approach relies on calibration of simulator outputs based solely on observations and therefore ignores the presence of shared simulator discrepancies from reality.

Apart from the widely applied truth-plus-error interpretation of the MME structure, an alternative interpretation has been adopted recently (Annan and Hargreaves, 2010). This interpretation is based on exchangeability assumptions regarding the relation between ensemble members, as well as their relation with real climate. The motivation for an alternative interpretation was generated by observing that if an ensemble is centred on truth, the ensemble mean should rapidly converge to the observations as more simulators are added to the ensemble (Annan and Hargreaves, 2010). However, this is not always the case in practice (Knutti et al., 2010; Jun et al., 2008). In addition, Sanderson and Knutti (2012) argue that the presence of shared simulator discrepancies from reality and the presence of uncertainty in the future ensemble spread makes the truth-plus-error interpretation challenging. The next section reviews some frameworks based on exchangeability assumptions about the MME structure.

2.5.2.3 Interpretation based on exchangeability assumptions

Annan and Hargreaves (2010), by observing the limitations of the truth-plus-error representation, consider the ensemble members to be exchangeable, i.e. "statistically indistinguishable" with reality. In other words, reality is treated as being another simulator (Chandler, 2013), implying that simulators and observations are thought to be drawn from the same distribution. The authors validate the exchangeability assumption by assessing forecast reliability of the exchangeable ensemble (that now also includes the observations), using simulators from the CMIP3 experiment, for a historical period. Results reveal good forecast quality. The authors note however that their interpretation, which implicitly assigns equal weights to ensemble members, has potential for improvement towards assigning non-uniform weighting, to account for the heterogeneous nature of the ensemble members in MMEs. Furthermore, Sanderson and Knutti (2012) observe that the presence of evident "strong similarities" between some ensemble members somehow violates the exchangeability assumption.

On the other hand, Rougier et al. (2013) consider a second-order exchangeability framework (by considering means and variances) to model ensemble members and

reality. They consider the ensemble members to be statistically indistinguishable, centred on their own consensus, to reflect "ignorance" as to whether one simulator is superior to another. Reality is then assumed to be exchangeable with the ensemble consensus, to express that simulators are considered to be of equal performance in representing reality. In model terms, reality is expressed as a sum of the ensemble consensus and a random error term expressing shared simulator discrepancy from reality.

Siegert et al. (2015) introduce a probabilistic latent variable model for representing the MME structure and perform a Bayesian implementation to produce posterior predictive distributions of the North-Atlantic Oscillation. Their model assumes that observations and ensemble members are linked through a common unobserved predictable component that they share in their distributions. The framework also explicitly accounts for uncertainties due to model and measurement error as well as internal variability, by assigning random noise terms in the model representation for ensemble members and observations. The unknown model parameters are assigned prior distributions and MCMC is used to obtain their posterior distribution. Exchangeability between observations and simulator outputs, as well as the truth-plus-error model are considered as special cases of the proposed framework. However, results suggest that the assumptions of both interpretations are not compatible with the application considered in Siegert et al. (2015).

Although model-based approaches for representing ensemble structures seem to be promising in terms of providing more coherent quantification of the underlying uncertainties, it is widely recognized that there are remaining challenges in the probabilistic interpretation of MMEs (Knutti, 2008; Tebaldi et al., 2005; Sanderson and Knutti, 2012). Sanderson and Knutti (2012) conclude that neither the "truth-plus-error" nor the" statistically indistinguishable" interpretations are fully representative of the CMIP5 ensembles. A main reason for this is the existence of shared structural errors between the ensemble members, which are often indistinguishable from model variability. According to Chandler (2013), ignorance of shared simulator discrepancies from reality prevents representative sampling of the space of possible modelling decisions regarding construction of a MME. The next section discusses how the presence of shared simulator discrepancy from reality is handled in existing frameworks.

2.5.2.4 Accounting for shared simulator discrepancy from reality

The existence of shared simulator discrepancies from reality is implicitly recognized in many studies (Sansom et al., 2013; Räisänen and Palmer, 2001; Rougier et al., 2013; Chandler, 2013). According to Sansom et al. (2013), the presence of shared simulator

discrepancy from reality is expected, since simulators share a lot of characteristics, some of which were discussed in Section 2.4. Rougier and Goldstein (2014) stress the importance of making reasonable judgements of the discrepancies of each simulator from reality and the interconnections between them.

Although widely recognized, the presence of shared simulator discrepancy from reality is not handled satisfactorily within the existing statistical frameworks for MME interpretation (Chandler, 2013), with a few exceptions. Buser et al. (2009) make explicit assumptions about biases of simulator outputs and incorporate them in their Bayesian framework for deriving the posterior distribution of future temperature. The approach has the benefit of making a clear distinction between additive model biases in the location and multiplicative biases in the spread of the distribution of true climate. However, it assumes that subject to small individual simulator biases, all simulators "correctly" predict the mean climate, suggesting a truth-plus error interpretation of the ensemble structure. Although they allow *a priori* judgements about correlation of biases among simulators, they do not explicitly incorporate shared simulator discrepancy in their framework for MME interpretation. On the other hand, Rougier et al. (2013) explicitly incorporate "ensemble discrepancy" as a random component representing shared errors among ensemble members. In contrast with Chandler (2013) however, their framework restricts all simulators to equally deviate from their consensus. This is not always realistic, since expert judgement might suggest grouping simulators into subsets, such that simulators in each subset are "equally credible" with respect to the overall consensus (Chandler, 2013). Additionally, Rougier et al. (2013) estimate the variance of shared simulator discrepancy by subjectively relating it to the ensemble variance. Chandler (2013) shows that there is more work to be done in this direction, by estimating historical discrepancy using historical observations and simulator outputs and accounting for difference between historical and future discrepancy.

To summarize, it is clear that a perfect reproduction of real climate is impossible, even using the highest quality climate simulators and sophisticated frameworks to combine information from a collection of them. It is therefore necessary to account for the resulting uncertainties when interpreting information from a MME. The existence of shared simulator discrepancies from reality must be recognized and at the same time special attention has to be paid regarding weighting of ensemble members, by exploiting their strengths while simultaneously discounting their weaknesses in simulating the quantities of interest. Chandler (2013) aims to deal with these issues by introducing a framework for quantifying uncertainty in climate simulations, under a Bayesian point of view.

2.6 Framework for quantifying uncertainty

This thesis aims to improve the framework of Chandler (2013) which is now discussed in detail. Firstly, the conceptual framework is introduced in Section 2.6.1, followed by the mathematical analysis under Gaussian distributional assumptions in Section 2.6.2. Finally, Section 2.6.3 describes the PM implementation introduced in Chandler (2013).

2.6.1 Conceptual framework

Chandler (2013) derives a probabilistic Bayesian framework for quantifying uncertainty in MMEs. The framework aims to take into account the issues discussed in Section 2.4 regarding MME interpretation. It pays particular attention to the fact that "all available simulators are imperfect" and that they do not cover the whole range of possible modelling decisions. Consequently, the probabilistic assessment of uncertainty in climate modelling should be treated with special care, since otherwise it will be unrepresentative of the full range of uncertainty. In addition, Chandler (2013) highlights the issue of simulators having different "quality" in reproducing different quantities. Therefore, his framework aims to exploit the strengths of each simulator, related to the climate parameters that it describes.

The intuition behind the framework arose from Leith and Chandler (2010), who consider the statistical properties of simulator outputs as parameter estimates of a model describing the outputs. Since simulators aim to represent the same climate-related dynamical processes, it is sensible to adopt a statistical model of the same form for all simulator outputs. Following the terminology of Chandler (2013), this model is referred to as a "mimic". The statistical parameters in the mimic are collected in a vector $\boldsymbol{\theta}$, which is called a "descriptor". The differences between the simulators and between simulators and reality are summarised via different descriptors for each simulator and for reality.

Returning to the issue of currently available simulators not covering the whole range of uncertainty, Leith and Chandler (2010) accounted for it, by suggesting that the descriptors $\boldsymbol{\theta}_i$ corresponding to different climate simulators i all belong to an underlying joint probability distribution. Information for each descriptor is provided by the output of the corresponding climate simulator, and in Chandler (2013) it is summarised by the maximum likelihood estimates (MLEs) of $\boldsymbol{\theta}_i$, denoted by $\hat{\boldsymbol{\theta}}_i$ (descriptor estimators). According to Cox and Hinkley (1974, p. 307), in large samples, the MLEs of the descriptors are asymptotically sufficient statistics for the raw data. Therefore, minimal information about the descriptors is lost by replacing the raw

data with the MLEs of the descriptors. This information is hierarchically linked to the corresponding descriptor by considering the MLE to be drawn from a distribution that is centred on the true descriptor $\boldsymbol{\theta}_i$, which is itself considered as random. Denoting by $\boldsymbol{\theta}_0$ the descriptor which summarises the true climate, observations of climate are incorporated in the framework through the descriptor estimate $\hat{\boldsymbol{\theta}}_0$, which is assumed to be centred on $\boldsymbol{\theta}_0$. Similarly to simulator outputs, $\hat{\boldsymbol{\theta}}_0$ is defined as the MLE of $\boldsymbol{\theta}_0$, obtained by fitting the mimic to observations.

For reasons discussed extensively in Section 2.4, Chandler (2013) considers all the simulator descriptors as being centred on $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$. Here $\boldsymbol{\omega}$ represents the shared discrepancy from reality of all the simulators that could be included in the MME. Since in general $\boldsymbol{\omega}$ is unknown, it is of crucial importance to make reasonable subjective judgements about it, in order to be able to quantify uncertainty in a meaningful probabilistic way.

Figure 2.1 shows a schematic representation of the structure of the proposed framework, as it appears in Chandler (2013).

The aim is to use all the available data to learn about actual climate. It is clear from Figure 2.1 that the data consist of the simulator outputs from the MME, expressed through the descriptor estimators $\{\hat{\boldsymbol{\theta}}_i, i \geq 0\}$ and climate observations, expressed through $\hat{\boldsymbol{\theta}}_0$. Under the Bayesian approach adopted in this framework, the required information about true climate is summarized in a posterior distribution $\pi(\boldsymbol{\theta}_0 \mid \hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m)$, where $\pi(\cdot)$ denotes a generic probability density and the vertical bar "|" denotes a conditional distribution, in this case conditioned on the values of $\{\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m\}$, m being the number of available simulators.

Figure 2.1 can be interpreted as a directed acyclic graph, which itself encodes conditional independence statements that can be used in the derivation of the generic posterior $\pi(\boldsymbol{\theta}_0 \mid \hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m)$. The result that is used for this derivation states that two nodes which are connected only through a common ancestor are conditionally independent of each other, given the value of the ancestor (Koski and Noble, 2009, Sections 2.3 and 2.8). This implies that the descriptor estimators $\{\hat{\boldsymbol{\theta}}_i, i > 0\}$ are conditionally independent of $\hat{\boldsymbol{\theta}}_0$, given $\boldsymbol{\theta}_0$.

The generic posterior of $\boldsymbol{\theta}_0$ can now be derived, using Bayes' theorem and the conditional independence result stated above. It is shown in Chandler (2013, suppl. material S1) to be:

$$\pi(\boldsymbol{\theta}_0 \mid \hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m) \propto \pi(\boldsymbol{\theta}_0)\pi(\hat{\boldsymbol{\theta}}_0 \mid \boldsymbol{\theta}_0)\pi(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m \mid \boldsymbol{\theta}_0). \tag{2.3}$$

The prior $\pi(\boldsymbol{\theta}_0)$ provides the opportunity also to make judgements *a priori* about the actual climate, instead of relying solely on observed values and simulator outputs.

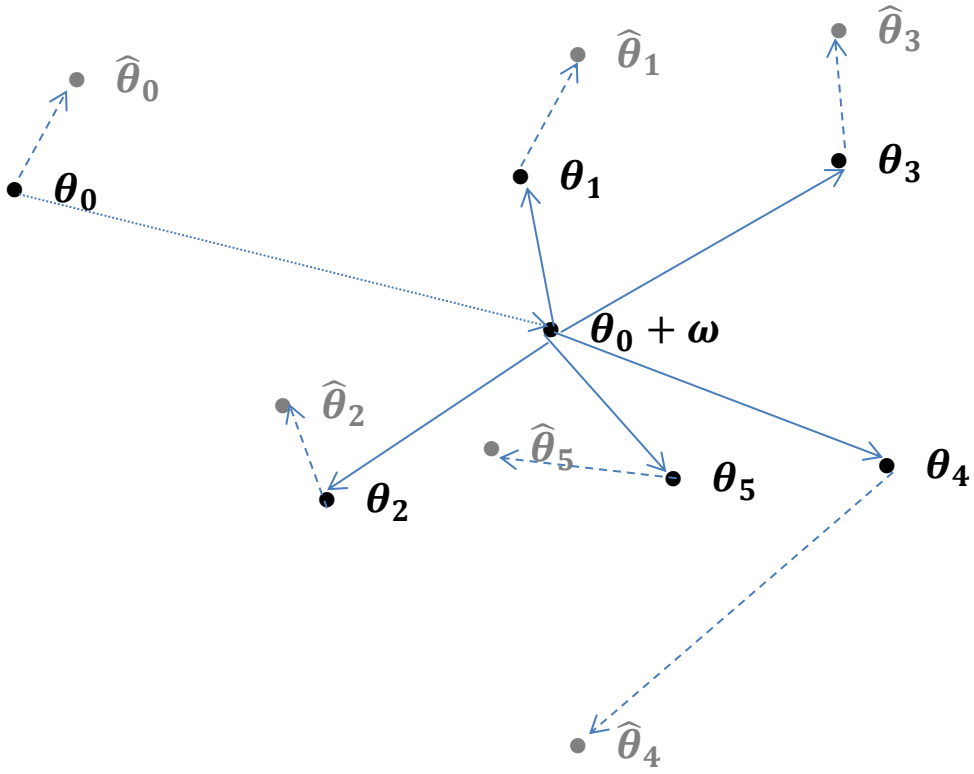


Figure 2.1: Geometrical representation of the proposed MME framework. $\boldsymbol{\theta}_0$ denotes the descriptor for the true climate; $\{\boldsymbol{\theta}_i, i > 0\}$ are descriptors for simulators; and $\{\hat{\boldsymbol{\theta}}_i, i \geq 0\}$ are descriptor estimates obtained from data ($i = 0$) and simulator outputs ($i > 0$). Dashed lines represent estimation errors; dotted line represents shared simulator discrepancy from reality, with simulator descriptors centred on $\boldsymbol{\theta}_0 + \omega$. Arrows indicate direction of causal relationships in which an intervention at the "parent" node is expected to produce a change at the "child" node.

The likelihood is partitioned into two components, $\pi(\hat{\boldsymbol{\theta}}_0 \mid \boldsymbol{\theta}_0)$ and $\pi(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m \mid \boldsymbol{\theta}_0)$. The first component of the likelihood represents the contribution of actual climate observations to the posterior and the second is the contribution of simulator outputs.

In Chandler (2013), an explicit analytical expression for the posterior distribution $\pi(\boldsymbol{\theta}_0 \mid \hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m)$ ((2.3)) is derived under multivariate Gaussian distributional assumptions for all the random quantities involved. This analytical expression is then used to formulate the PM implementation of the Bayesian analysis. Sections 2.6.2-2.6.3 summarise the ideas.

2.6.2 Mathematical analysis of the Gaussian specification

In this section, all the random quantities involved in the framework (illustrated in Figure 2.1) are assigned multivariate normal distributions, for purposes of simplicity and computational efficiency. Alternative distributions are considered in the proposed fully Bayesian implementation of Section 3.5.3. The quantities that are regarded as random in the framework are: The descriptor $\boldsymbol{\theta}_0$ of true climate, the descriptors $\{\boldsymbol{\theta}_i, i = 1, \ldots, m\}$ of simulator outputs, the descriptor estimators $\{\hat{\boldsymbol{\theta}}_i, i = 0, \ldots, m\}$ and the shared discrepancy $\boldsymbol{\omega}$. The distributional assumptions are shown below:

- Distribution of descriptor estimators $\{\hat{\boldsymbol{\theta}}_i, i = 0, \ldots, m\}$

$$\hat{\boldsymbol{\theta}}_i | \boldsymbol{\theta}_i \sim MVN(\boldsymbol{\theta}_i, \boldsymbol{J}_i) \qquad (i = 0, \ldots, m). \tag{2.4}$$

  The $\{\hat{\boldsymbol{\theta}}_i,\ i = 1, \ldots, m\}$ are the MLEs of $\{\boldsymbol{\theta}_i,\ i = 1, \ldots, m\}$ and therefore have asymptotically a multivariate normal distribution. Under the framework of Chandler (2013), the estimators $\{\hat{\boldsymbol{\theta}}_i,\ i = 1, \ldots, m\}$ are considered to be independent, conditional on the true descriptors $\{\boldsymbol{\theta}_i,\ i = 1, \ldots, m\}$. $\boldsymbol{J}_i$ is the covariance matrix of the MLE for the i^{th} descriptor. $\boldsymbol{J}_i$ represents uncertainty due to internal (natural) variability in data source i.

  Often, interest is on summarizing projected climate and comparing it with the corresponding historical performance. In this case, the descriptor vector $\boldsymbol{\theta}$ can be split into historical and future components, as follows:

$$\boldsymbol{\theta} = (\boldsymbol{x}^{\prime\,(\text{hist})}, \boldsymbol{x}^{\prime\,(\text{fut})})',$$

  where $\boldsymbol{x}$ is a vector of the climatological parameters of interest.

  Chandler (2013) notes that since we don't have future observations of the true climate, the future components of $\hat{\boldsymbol{\theta}}_0$ are undefined and are thus estimated with

zero precision. In this context, $\boldsymbol{J}_0$ is most conveniently defined via its inverse $\boldsymbol{J}_0^{-1}$, which can be written in the form:

$$\boldsymbol{J}_0^{-1} = \begin{pmatrix} \boldsymbol{J}_0^{-1\,(\text{hist})} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{J}_0^{-1\,(\text{fut})} \end{pmatrix} = \begin{pmatrix} \boldsymbol{J}_0^{-1\,(\text{hist})} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{2.5}$$

Setting $\boldsymbol{J}_0^{-1\,(\text{fut})}$ in (2.5) equal to zero, forces the off-diagonal blocks of the precision matrix $\boldsymbol{J}_0^{-1}$ also to be zero.

Equation (2.5) shows that having set the future precision $\boldsymbol{J}_0^{-1\,(\text{fut})}$ to zero, it does not matter what value is chosen for $\hat{\boldsymbol{\theta}}_0^{(fut)}$. For implementation therefore, it can be set to any value. It is convenient to set $\hat{\boldsymbol{\theta}}_0^{(\text{fut})} = \mathbf{0}$.

- Distribution of simulator outputs' descriptors $\{\boldsymbol{\theta}_i, i = 1, \ldots, m\}$

  The distribution of $\boldsymbol{\theta}_i$ is expressed through discrepancies from the actual climate, defined as $\{\boldsymbol{\delta_i} = \boldsymbol{\theta_i} - \boldsymbol{\theta_0} : i = 1, \ldots, m\}$. The $\{\boldsymbol{\delta_i},\ i = 1, \ldots, m\}$ are assumed to be independent, conditional on the shared simulator discrepancy $\boldsymbol{\omega}$. However, it is questionable whether this assumption is always realistic, e.g. if an ensemble contains several simulators belonging to the same modelling group, or if different variants of the same simulator exist. This issue is handled in Section 4, where an extended framework which accounts for simulator grouping is proposed and forms the third contribution of this thesis. According to the framework described in Section 2.6.1 and as also illustrated in Figure 2.1:

$$\boldsymbol{\delta}_i | \boldsymbol{\omega} \sim MVN(\boldsymbol{\omega}, \boldsymbol{C}_i) \qquad (i = 1, \ldots, m). \tag{2.6}$$

  Equivalently, the descriptors $\{\boldsymbol{\theta_i} = \boldsymbol{\theta_0} + \boldsymbol{\delta_i},\ i = 1, \ldots, m\}$ are centred on a "simulator consensus" $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is defined as the "shared simulator discrepancy from reality" and represents the tendency of the simulators collectively to depart from the true climate $\boldsymbol{\theta_0}$.

  $\boldsymbol{C}_i$ represents the propensity of simulator i to deviate from the simulator consensus $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$.

- Distribution of shared simulator discrepancy $\boldsymbol{\omega}$ (from reality $\boldsymbol{\theta_0}$)

$$\boldsymbol{\omega} \sim MVN(\mathbf{0}, \boldsymbol{\Lambda}). \tag{2.7}$$

  Due to lack of knowledge, *a priori*, about the direction of the discrepancy $\boldsymbol{\omega}$, its expected value is set to zero. Therefore, $\boldsymbol{\Lambda}$ represents the propensity of simulators collectively to deviate from the actual climate.

- Distribution of true-climate descriptor $\boldsymbol{\theta}_0$

$$\boldsymbol{\theta}_0 \sim MVN(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0). \tag{2.8}$$

Here, $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are determined based on *a priori* judgements about real climate.

Let $\boldsymbol{D}_i$ be defined as $\boldsymbol{C}_i + \boldsymbol{J}_i$. Additionally denote by $\boldsymbol{I}$ the $p \times p$ identity matrix, where p is the number of elements in $\boldsymbol{\theta}_0$. Then, it can be shown (Chandler, 2013, suppl. material S2) that the posterior distribution of $\boldsymbol{\theta}_0$ under the Gaussian distributional assumptions above is:

$$\boldsymbol{\theta}_0 | \hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m \sim MVN(\boldsymbol{\tau}, \boldsymbol{S}), \tag{2.9}$$

where

$$\boldsymbol{S^{-1}} = \boldsymbol{\Sigma}_0^{-1} + \boldsymbol{J}_0^{-1} + \left[\boldsymbol{\Lambda} + \left(\sum_{k=1}^{m} \boldsymbol{D}_k^{-1} \right)^{-1} \right]^{-1}, \tag{2.10}$$

and

$$\boldsymbol{\tau} = \boldsymbol{S} \left[\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \boldsymbol{J}_0^{-1} \hat{\boldsymbol{\theta}}_0 + \left(\boldsymbol{I} + \sum_{k=1}^{m} \boldsymbol{D}_k^{-1} \boldsymbol{\Lambda} \right)^{-1} \sum_{i=1}^{m} \boldsymbol{D}_i^{-1} \hat{\boldsymbol{\theta}}_i \right]. \tag{2.11}$$

From the expression of $\boldsymbol{\tau}$ in (2.11), the posterior mean takes the form of a matrix-weighted average of the prior mean $\boldsymbol{\mu}_0$, the estimated descriptor $\hat{\boldsymbol{\theta}}_0$ obtained from real climate observations and the estimated descriptors $\{\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m\}$ obtained from the MME simulator outputs. The matrix-valued weights ensure that each simulator contributes to the posterior mean only for the components of $\boldsymbol{\theta}_0$ for which it is informative. This approach thereby succeeds in exploiting the strengths of simulators while simultaneously discounting their weaknesses in terms of describing each of the components of $\boldsymbol{\theta}_0$. The structure of $\boldsymbol{S^{-1}}$ implies that perfect knowledge of reality would make the use of simulators unnecessary. If the available relevant observations could be increased indefinitely, $\boldsymbol{J}_0^{-1}$ would become arbitrarily large and therefore dominate in the expression for $\boldsymbol{S^{-1}}$, suggesting infinite reduction of uncertainty in real climate. On the other hand, an attempt to increase the number of simulators (i.e. let $m \to \infty$) or the length and number of simulator runs (i.e. let $\boldsymbol{J}_i \to 0$), or the consensus between the simulators (i.e. let $\boldsymbol{C}_i \to 0$), would not achieve infinite reduction of uncertainty, due to the presence of shared simulator discrepancy (expressed as $\boldsymbol{\Lambda}$ in (2.10)).

It is clear that in order to calculate the values of $\boldsymbol{\tau}$ and $\boldsymbol{S}^{-1}$, the values of the covariance matrices $\boldsymbol{J}_0, \boldsymbol{\Lambda}$ and $\boldsymbol{D}_i = \boldsymbol{C}_i + \boldsymbol{J}_i$ are needed. In practice, those covariance matrices are unknown and must be estimated.

A fully Bayesian implementation of the proposed framework accounts for uncertainty in the values of the covariance matrices $\boldsymbol{J}_i$, $\boldsymbol{\Lambda}$ and $\boldsymbol{C}_i$ by building a hierarchical model, where hyperprior distributions are assigned to the unknown covariance matrices. However, under this approach an exact analytical derivation of the posterior is generally not possible and, consequently, it requires the use of alternative methods, such as the Markov Chain Monte Carlo technique. These methods increase the computational burden of the problem and require careful thought in terms of the judgements relevant to the choice of hyperprior distributions and their corresponding parameters (see Sections 3.5.2 - 3.5.3).

In view of this, Chandler (2013) presents a less computationally-demanding method, the PM implementation, to provide estimates of the unknown covariance matrices. Then the expressions for the posterior mean ((2.11)) and precision matrix ((2.10)) can be evaluated with a small amount of computational effort. This approach can be considered as a form of empirical Bayes analysis, where the hyperparameters of the prior distributions are estimated from the data. However, this approach, although being computationally efficient, tends to ignore part of the overall uncertainty, by not accounting for the uncertainty in the values of the unknown covariance matrices. This issue is addressed in the two proposed fully Bayesian implementations of Sections 3.5.2 - 3.5.3, which form the second contribution of this thesis. The next section outlines the PM implementation.

2.6.3 Poor man's (PM) implementation

This section presents the proposed estimators of the covariance matrices $\boldsymbol{J}_i$, $\boldsymbol{C}_i$ and $\boldsymbol{\Lambda}$ that were used in the PM implementation of Chandler (2013), under the Gaussian specification described in Section 2.6.2.

It is important to note that in order for the derived posterior distribution of $\boldsymbol{\theta}_0$ to be valid, the posterior variance $\boldsymbol{S}$ in (2.9) must be non-negative definite.

The following estimators are suggested for $\boldsymbol{J}_i$, $\boldsymbol{C}_i$ and $\boldsymbol{\Lambda}$ corresponding to descriptor i:

- Estimator of $\boldsymbol{J}_i$

  Since $\boldsymbol{J}_i$ is the covariance matrix for the MLE of the i^{th} descriptor $\boldsymbol{\theta}_i$, standard statistical practice and specifically the Fisher Information matrix (Davison, 2003, Section 4.4) can be used to evaluate its theoretical value. Practically, $\boldsymbol{J}_i$ can be easily estimated from statistical software, after fitting the mimic to the data.

- Estimator of $\boldsymbol{C}_i$

  The simulators can be partitioned into groups such that all the simulators of each group can be considered as being equally credible with respect to the simulator consensus. Then, simulators of each group can be considered as sharing a common covariance matrix $\boldsymbol{C}_S$, for group S, say. A natural estimator $\hat{\boldsymbol{C}}_S$ of $\boldsymbol{C}_S$ is the covariance matrix of the descriptor estimates $\hat{\boldsymbol{\theta}}_i$ that belong to group S. Equivalently:

$$\hat{\boldsymbol{C}}_S = \frac{1}{m_S - 1} \sum_{i \in S} \left(\hat{\boldsymbol{\theta}}_i - \bar{\hat{\boldsymbol{\theta}}} \right) \left(\hat{\boldsymbol{\theta}}_i - \bar{\hat{\boldsymbol{\theta}}} \right)', \tag{2.12}$$

  where m_S is the number of simulators in S. In this expression, $\bar{\hat{\boldsymbol{\theta}}}$ is the overall mean of the estimated simulator descriptors, defined to be:

$$\bar{\hat{\boldsymbol{\theta}}} = m^{-1} \sum_{i=1}^{m} \hat{\boldsymbol{\theta}}_i.$$

  It is proved in Chandler (2013) that the proposed estimator overestimates $\boldsymbol{C}_S$ on average, in which case a fully Bayesian implementation (Sections 3.5.2-3.5.3), which would assign a distribution to $\boldsymbol{C}_S$, could be considered as an alternative. However, these bias effects can be regarded as insignificant compared to the overall under-estimation of uncertainty caused by plugging-in estimates of the unknown covariance matrices. Another limitation of the proposed estimator is that it is not easy to handle in cases where some components of $\hat{\boldsymbol{\theta}}_i$ are undefined. This can happen for example when the descriptors $\{\boldsymbol{\theta}_i,\ i = 1, \ldots, m\}$ have components that correspond to multiple time-periods and there are missing simulator runs for some of those periods (this occurs in the application described in Chapter 3). Under these circumstances, an alternative form of $\hat{\boldsymbol{C}}_S$ must be proposed, which is non-trivial to derive analytically from first principles, such as maximum-likelihood estimation (see also Rukhin (2013) and Appendix A).

- Estimator of $\boldsymbol{\Lambda}$

  Chandler (2013) proposes two estimators of $\boldsymbol{\Lambda}$, depending on whether observations provide information about all, or only some of the components of $\boldsymbol{\theta}_0$. In the fist case, $\boldsymbol{\Lambda}$ is estimated empirically, as $\hat{\boldsymbol{\Lambda}} = \hat{\boldsymbol{\omega}}\hat{\boldsymbol{\omega}}'$, where $\hat{\boldsymbol{\omega}} := \bar{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{\theta}}_0$. However, if the observations cannot provide at all any information for some components of the descriptor vector, for example when the descriptor has future components in addition to historical, the above estimator can be used to estimate only the

historical block of $\boldsymbol{\Lambda}$, i.e. $\hat{\boldsymbol{\Lambda}}^{(\text{hist})} = \hat{\omega}^{(\text{hist})}\hat{\omega}^{(\text{hist})\prime}$, where $\hat{\omega}^{(\text{hist})} = \bar{\hat{\boldsymbol{\theta}}}^{(\text{hist})} - \hat{\boldsymbol{\theta}}_0^{(\text{hist})}$. Since no observations are available to estimate the rest of the matrix $\boldsymbol{\Lambda}$, it has to be estimated using subjective judgements. In order to account for potential changes in future discrepancy relative to historical, it can be judged that $\boldsymbol{\omega}^{(\text{fut})} = \boldsymbol{\omega}^{(\text{hist})} + \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is a zero-mean vector, independent of $\boldsymbol{\omega}^{(\text{hist})}$, with covariance matrix $K\boldsymbol{\Lambda}^{(\text{hist})}$, for some constant $K > 0$. Under these assumptions, $\boldsymbol{\Lambda}$ is estimated by:

$$\hat{\boldsymbol{\Lambda}} = \begin{pmatrix} \hat{\boldsymbol{\Lambda}}^{(\text{hist})} & \hat{\boldsymbol{\Lambda}}^{(\text{hist})} \\ \hat{\boldsymbol{\Lambda}}^{(\text{hist})} & (1+K)\hat{\boldsymbol{\Lambda}}^{(\text{hist})} \end{pmatrix}. \tag{2.13}$$

The factor $1 + K$ in the future diagonal block of $\hat{\boldsymbol{\Lambda}}$ supports the assumption that shared simulator discrepancy will increase in the future. The extent of this increase relative to the historical discrepancy depends on the magnitude of K.

This empirical way of estimating $\boldsymbol{\Lambda}$ neglects the uncertainty induced by estimation and this is one argument in favour of the fully Bayesian implementation. Additionally, the subjectivity invoked by relating future discrepancy to historical, through the choice of value for K is an important limitation of this method of estimating $\boldsymbol{\Lambda}$. Of course, subjectivity is inevitable in making judgements about unobserved quantities like the future shared simulator discrepancy. However, it would be of interest to perform a sensitivity test on the choice of K, in order to investigate to what extent it affects the posterior of $\boldsymbol{\theta}_0$. On the other hand, alternative methods, relying on more robust criteria for making judgements about future discrepancy can be explored. An example is the proposed bootstrap estimator of $\boldsymbol{\Lambda}$ based on earlier data (Section 3.5.1), which forms the first contribution of this thesis.

To summarize, the above suggested estimators for $\boldsymbol{J}_i$, $\boldsymbol{C}_i$ and $\boldsymbol{\Lambda}$ are easy to calculate from the available data and therefore act as devices for a computationally-efficient analytical solution of the posterior of $\boldsymbol{\theta}_0$, under the PM implementation. A revised version of the "poor man's" implementation is applied to global surface air-temperature data in Section 3.5.1, where the limitations discussed here are taken into consideration in the proposed estimators of the unknown covariance matrices.

2.7 Summary

This chapter provided a review of the main sources of uncertainty in climate modelling and the ways this is handled in the literature. Particular emphasis was given to the use of MMEs and the different techniques which are often used for MME interpre-

tation. It was illustrated that explicit weighting schemes of the ensemble members often do not capture the underlying uncertainties representatively. This motivates the use of model-based probabilistic frameworks, which are constructed based on assumptions about the MME structure. It was observed that the reviewed model-based approaches, although sometimes recognizing the presence of shared simulator discrepancy from reality, often fail to account satisfactorily for its presence in their frameworks.

The probabilistic, Bayesian framework of Chandler (2013) which was presented in Section 2.6, explicitly accounts for the presence of shared simulator discrepancy in the MME structure. The conceptual framework (Section 2.6.1) regards the statistical properties of simulator outputs as parameter estimates for a common statistical model ("mimic") which describes the outputs. The parameters are collected in the descriptor vector, which differs for each ensemble member and for real climate. The descriptors are assumed to be random, to account for the fact that the currently available simulators do not cover the whole range of uncertainty. It is assumed that simulator descriptors are exchangeable, conditional on their consensus, which deviates from reality due to the presence of shared simulator discrepancy. Descriptor estimates are obtained by fitting the mimic to simulator outputs and observations.

The aim is to derive the posterior distribution of real-climate descriptor, given simulator outputs and observations. In Section 2.6.2, the posterior of real-climate descriptor was presented, based on Gaussian distributional assumptions for the different levels of the hierarchical framework. Calculation of the posterior parameters requires knowledge of the unknown covariance matrices expressing the different sources of variability in the levels of the hierarchical framework. The PM implementation which was presented in Section 2.6.3 provides estimates of these covariance matrices, thus ignoring the uncertainty in their values, in an attempt to provide a computationally efficient framework implementation.

The first of the contributions of this thesis is an improved estimator of the covariance matrix $\mathbf{\Lambda}$ expressing variability due to shared simulator discrepancy from reality. The proposed estimator of $\mathbf{\Lambda}$ avoids the subjective choice of the parameter K (which relates future and historical discrepancy) and provides a more robust estimator instead, by exploiting information from historical data through bootstrapping. The methodology is illustrated in the RPM implementation described in Section 3.5.1.

The second contribution relates to the uncertainty in the values of the covariance matrices expressing variability in the framework of Chandler (2013). The PM implementation in Section 2.6.3, in an attempt to provide a computationally efficient method to obtain the posterior of true-climate descriptor $\boldsymbol{\theta}_0$, assigns estimates to the involved covariance matrices. However, this approach has the limitation of ignoring

the underlying uncertainty in the values of the covariance matrices. To account for this uncertainty, two fully Bayesian implementations are proposed in Sections 3.5.2-3.5.3, which assign prior distributions instead of estimates to the covariance matrices. The two implementations differ in the prior choices, to explore sensitivity of the posterior of $\boldsymbol{\theta}_0$ to the prior judgements.

Finally, the third contribution is the development of an extended framework which accounts for potential simulator grouping that is known to often exist in MMEs. The assumed independence of simulator descriptors conditional on their consensus in the framework of Chandler (2013) (see (2.6)), ignores any potential grouping of the ensemble members. An extended framework is therefore proposed in Chapter 4, which accounts for simulator grouping, as determined by expert judgement. A random effects model is also proposed (Section 4.6), to allow estimation of within-group variability in groups of ensemble members.

The next chapter illustrates the first two improvements, through the three proposed framework implementations in Sections 3.5.1-3.5.3, in an application for inference on global surface air temperature data, using simulator outputs and observations.

Chapter 3

Application to global surface air temperature

3.1 Outline of the study

The focus of this chapter is to illustrate the first two improvements of the framework of Chandler (2013) which are considered in this thesis.The first improvement is illustrated through the RPM implementation of Section 3.5.1, which assigns estimates to the framework's covariance matrices. The implementation is computationally efficient, since it provides an analytical expression for the posterior of true-climate descriptor $\boldsymbol{\theta}_0$. However, it has the limitation of ignoring uncertainty in the values of the framework's covariance matrices. In order to account for this uncertainty, distributions instead of estimates are assigned to the covariance matrices, leading to a fully Bayesian implementation, which forms the second contribution of this thesis. However, the posterior of $\boldsymbol{\theta_0}$ cannot be evaluated analytically in this case, requiring the use of numerical algorithms and precisely the Gibbs sampler. This makes the fully Bayesian implementations more computationally demanding, thus adding to the complexity induced by specifying distributions instead of estimates to the unknown quantities.

Another issue to consider is the subjective choice of distributional assumptions, caused by adding another level of hierarchy under the fully Bayesian implementation. Careful thought is required regarding the choice of distributions that are specified under the fully Bayesian framework. The choice of appropriate distributions, as well as parameters for those distributions, is an extensive and broad area of research in statistics (Wikle et al., 1998; Wikle, 2003; Gelman, 2006; Tebaldi and Sansó, 2009). The decision could be based, for example, on commonly used hyperpriors for the

parameters of interest in the relevant literature (Gelman, 2006). Alternatively, it could be based on studies of datasets from an earlier period than the one considered in the application. Inevitably, in any approach of specifying hyperpriors and their corresponding hyperprior parameters for unknown quantities, there is always a degree of subjectivity. Therefore, it is also of interest to perform a sensitivity analysis of the posterior, based on different distributional assumptions for the unknown quantities. In order to examine sensitivity of the posterior to prior choices, two fully Bayesian implementations are proposed, which are presented in Sections 3.5.2-3.5.3.

The posterior of $\boldsymbol{\theta}_0$ under the three proposed implementations is derived, which is used to make inference for global surface air temperature, using simulator outputs and observations. Interest is on learning about yearly mean global surface air temperature, for the historical and future 20-year periods $1986-2005$ and $2016-2035$ respectively. The datasets used in the application are real climate observations from the HadCRUT3 dataset (for the historical period), as well as simulator outputs from the CMIP5 experiment (both for the historical and future periods).

The comparison of the two versions of the fully Bayesian implementation provides a sensitivity analysis on the choice of distributional assumptions for the unknown quantities, for this particular study. Ultimately, results from both versions are compared with those from the RPM implementation, in an attempt to investigate whether the more simplified and computationally easier RPM implementation yields adequate approximations to the posteriors of interest. The results are presented in Section 3.6.

Section 3.2 describes the datasets (observations and simulator outputs) used in the application. In Section 3.3, the "mimic', i.e the statistical model representing the structure of observations and simulator outputs is introduced. Section 3.4 introduces some general conventions and notation that are helpful in describing the three implementations that follow; these implementations are described in Sections 3.5.1-3.5.3. Finally, results and conclusions are presented in Sections 3.6 and 3.7 respectively.

3.2 Description of the datasets

It is of interest to make inference not only about historical yearly mean global surface air temperature, but also about its change in the future relative to the present. For this reason, the study covers two 20-year periods, one corresponding to the present denoted as "historical" period and one corresponding to the future, denoted as "future" period.

The historical and future periods are chosen to be the periods: **1986−2005** and **2016−2035** respectively. The two periods are chosen to have a duration of 20 years

since any trends over 20-year periods can be regarded as being approximately linear. This simplifies the statistical modelling, so that focus remains on the aims of the study. Furthermore, the particular time periods (historical period from the 20^{th} century; future period from the 21^{st} century) are chosen based on time horizons which were used by many organizations (Vincent and Gullett, 1999; Solomon et al., 2007). Of course, the study could be extended so as to cover longer or overlapping periods (see Section 6.2).

Yearly mean global surface air temperature observations for the historical period 1986-2005 are obtained from the HadCRUT3 (Brohan et al., 2006) and "absolute surface air temperature climatology" (Jones et al., 1999) datasets both accessed through the Climatic Research Unit website (Jones et al., 2014). Those datasets were also used in the fourth assessment report (AR4) of the Intergovernmental Panel on Climate Change (IPCC) (Solomon et al., 2007). The HadCRUT3 dataset is an updated version of HadCRUT dataset, the latter presenting combined land and marine monthly mean surface air temperature departures (anomalies) from the period $1961 - 1990$ on a $5°C \times 5°C$ grid-box basis from year 1850 until October 2011. The HadCRUT3 dataset extends the improvements made to the sea-surface data, to the global dataset. According to Brohan et al. (2006), since the mid 20^{th} century (i.e. including the historical period of interest), the uncertainties in global and hemispheric mean temperatures in the HadCRUT3 anomalies are small compared to the increase in temperature during that period. Figure 10 in Brohan et al. (2006) suggests a standard error of about $\pm 0.05°C$ in the HadCRUT3 yearly global temperature anomaly time series for the historical period $1986 - 2005$. The "absolute surface air temperature climatology" dataset contains the absolute monthly surface air temperatures for the base period $1961 - 1990$, on a $1°C \times 1°C$ grid-box basis. Rigorous assessment of the uncertainty in the global values from the "absolute surface air temperature climatology" dataset is difficult to obtain from the existing literature.

In both datasets, the global monthly temperatures are obtained from the gridded values. Because of the spherical shape of the earth, even if the earth grids are regularly spaced in terms of latitude and longitude lines, grid boxes have larger area as we approach the equator. Consequently, to compute a global average, a weighted average of the grid boxes is required, the weight being the cosine of the latitude value at the centres of the grid box (see Figure 3.1). According to Jones et al. (1999), this point is chosen to be the central point of the $5°C \times 5°C$ grid box. After the grid boxes have been weighted, the global monthly means are obtained by averaging the hemispheric monthly means. This, according to Brohan et al. (2006) is done in order to prevent the northern hemisphere from dominating the average due to having more observations than the southern hemisphere. Then, the monthly global values from the "absolute

surface air temperature climatology" dataset for the period $1961 - 1990$ are added to the HadCRUT3 monthly global anomalies, in order to get the absolute mean monthly global values for the historical period $1986 - 2005$. Finally, the yearly global means for the historical period are calculated from the corresponding monthly global means.

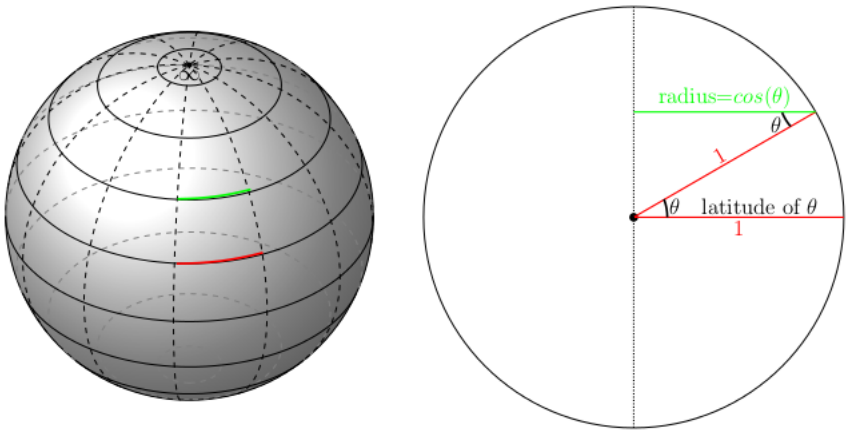


Figure 3.1: Illustration of the Earth's grid weighting. Left: Schematic representation of the Earth's grid. Vertical dashed lines: Longitude lines. Solid lines: Latitude lines. Right: Illustration of the cosine weighting in two successive grid boxes of different latitude.

Global surface air temperature simulator outputs are obtained from an MME consisting of 32 GCMs from the CMIP5 project. The CMIP5 project is the 5^{th} phase of CMIP (Coupled Model Intercomparison Project,Meehl et al. 2000) and according to Taylor et al. (2007), it aims to "deal with scientific questions related to the IPCC AR4 (Intergovernmental Panel on Climate change 4^{th} Assessment Report), improve the understanding of climate and give estimates of future climate change". Under the CMIP5 experimental framework, various climate-related experiments are performed, for both short-term and long-term predictions and different ensemble sizes. Experiments are initialised under different scenarios about future radiative forcing and CO_2 concentration.

In the GCMs used, change in global surface air temperature is considered to be driven by radiative forcing, the latter being used as the input of the models. According to the IPCC expert meeting report (Moss et al., 2008), a set of "benchmark emissions scenarios", referred to as "Representative Concentration Pathways" (RCPs) is identified, in order to initiate future climate model simulations. The primary purpose of the RCPs is to "provide time-dependent projections of atmospheric greenhouse gas concentrations". The main types of RCP scenarios based on approximate radiative forcing levels and CO_2 concentrations are indicated in Table 3.1.

Name	Radiative Forcing	Concentration	Pathway shape
RCP8.5	$> 8.5W/m^2$ in 2100	$>\sim 1370C0_2$ -eq in 2100	Rising
RCP6	$\sim 6W/m^2$ at stabilization after 2100	$\sim 850CO_2$ -eq (at stabilization after 2100)	Stabilization without overshoot
RCP4.5	$\sim 4.5W/m^2$ at stabilization after 2100	$\sim 650C0_2$ -eq (at stabilization after 2100)	Stabilization without overshoot
RCP3-PD	peak at $\sim 3W/m^2$ before 2100 and then decline	peak at $\sim 490CO_2$ -eq before 2100 and then decline	Peak and decline

Table 3.1: Types of representative concentration pathways (Moss et al., 2008, Table 1).

For the purposes of this study, runs driven by the RCP8.5 forcing scenario only are considered. According to Taylor et al. (2007), RCP8.5 is officially defined as: "The representative concentration pathway that approximately results in a radiative forcing of $8.5Wm^{-2}$ at year 2100, relative to pre-industrial conditions". In practice, compared to the other scenarios illustrated in Table 3.1, the RCP8.5 scenario predicts the maximum amount of radiation and CO_2 concentration by year 2100.

The initial aim was to use all the 45 GCMs participating in the CMIP5 experiment, for the historical period and for the future period under the RCP8.5 scenario. However, there are some particular GCMs that produced output for the historical period, but not for the future period under the RCP8.5 scenario. In this case, the estimated framework covariance matrix $\hat{\boldsymbol{J}}_i$ could be easily modified by considering its inverse $\hat{\boldsymbol{J}}_i^{-1}$ ((see (2.5)) and setting the future components of the latter to zero, for the simulators without future components. However, regarding the estimation of $\boldsymbol{C}_i$ (see (2.12)), in the case of missing future components of some descriptors $\{\boldsymbol{\theta}_i, i = 1, \ldots, m\}$, the attempt to find an appropriate estimator from first principles (i.e. maximum likelihood estimation) analytically, seems to be very complicated, as also discussed in Section 2.6.3. For the purposes of this study, the GCMs that produced historical outputs only, leading to descriptors without future components, are discarded. Note however, that the simulators with missing outputs are also informative in estimating the historical simulator consensus and therefore they should ideally be included in the analysis. It is also worth mentioning that for some runs of two GCMs (HadGEM2-CC and HadGEM2-ES), there are particularly low values for year 2005, the last year of the historical period. This is possibly because of missing output values for December of 2005, for those particular runs. Those GCMs are also discarded from the analysis.

The 32 remaining GCMs form the MME which provides the surface air temperature outputs for the two periods of interest. A table with the names of the GCMs, together with the corresponding modelling centres, is presented in Appendix B. It is

important to note that different GCMs provide different numbers of runs. However, this does not require any modification of the current framework, since outputs from multiple runs of the same simulator are pooled together and fitted in the mimic to give a single estimate of $\boldsymbol{\theta}_i$, corresponding to simulator i. More details about fitting the mimic are given in Section 3.3.

The outputs of the involved GCMs can be accessed from the PCMDI (Program for Climate Model Diagnosis and Inter-comparison) gateway (`https://pcmdi9.llnl.gov/projects/esgf-llnl/`). It is important to note that the time variable is measured according to different calendar-types (e.g. 365-day, 360-day, Gregorian) among the data files. Additionally, not all the data files have the same earth-grid representations. However, they all give an indication of the approximate grid locations. Similarly to the historical observations, a weighted average over all the gridded outputs is calculated for each monthly time entry of each GCM, the weight being the cosine of the latitude value at the approximate locations of each grid box. In contrast to the historical observations' dataset, in the CMIP5 experiment, the gridded GCM outputs are uniformly spread along the two hemispheres. Therefore, for the CMIP5 datasets, there is no need to average the hemispheric monthly means in order to get the global ones. The yearly mean global outputs are therefore extracted from the corresponding global monthly mean values, for the historical and future periods of interest. From now on, for consistency with the terminology in the framework, the GCMs will be named as "simulators", since they are designed to simulate true climate. Conditioning on the information obtained from climate simulators for the historical and future periods under study, together with the observed global surface air temperature obtained from the HadCRUT3 data set, for the historical period 1986-2005, the posterior $\boldsymbol{\theta}_0$ of the true-climate descriptor is derived, by implementing the framework described in Section 2.6.

The first step in implementing the framework is to develop a "mimic", i.e. a statistical model that adequately describes the structure of both the observations and the simulator outputs described above. This is considered in the next section.

3.3 The mimic

By definition (see Section 2.6.1), the mimic is a statistical model which aims to represent the structure both of the simulator outputs and of the climate observations. Then, the descriptors of observations and simulator outputs can be regarded as being the parameters of the mimic, having different values for the actual climate observations and for each simulator. The mimic can then be fitted separately to the

observations and to the output from each simulator, to obtain estimates of the corresponding descriptors. Therefore, in order to make a reasonable choice of the form of the mimic, it is important to explore the data and observe their behaviour. Figure 3.2 shows observations and simulator outputs for the historical and future periods of interest.

The choice of the most "appropriate" mimic is quite subjective. The model should reproduce the main characteristics of simulator outputs and at the same time it should allow a computational efficient estimation of model parameters. According to Figure 3.2, the use of a linear time trend model as mimic does not seem unreasonable, since it is obvious that global surface air temperature increases almost linearly in the historical observations and in the majority of simulators. This also shows that it is not unrealistic to model the descriptors of actual climate and simulator outputs using the same mimic, following the conceptual framework in Section 2.6.1. It is evident that there is variability, both between the simulator outputs and between the outputs and the observations (in the historical period). This variability is also accounted for in the framework, by assigning different values to the descriptors, each time the mimic is fitted to the output from a distinct simulator or to the observations. The only obvious feature not captured by the linear trend mimic is the temperature drop around the years $1991 - 1993$. This is caused by the eruption of Mount Pinatubo in Philippines, in 1991, which is the second largest volcanic eruption of the 20^{th} century (Newhall et al., 2005).

In mathematical terms, the chosen mimic has the following form:

$$Y_t = \alpha + \beta(t - \bar{t}) + \epsilon_t. \tag{3.1}$$

The mimic is fitted separately to observations and to each simulator output. In addition, it is fitted separately to simulator outputs for the historical and future periods. Note also that the runs of each simulator are firstly pooled together and then the mimic is fitted to them. For the historical period, values of t represent the years 1986-2005, repeated for all the runs of that particular simulator and Y_t is the yearly mean global surface air temperature value at time t. Similarly for the future period 2016-235. $\bar{t}$ is the mean of the values of t within each period. The centering of the time index in (3.1) ensures that conditional on the regression parameter values α and β for a particular data source, the corresponding estimates $\hat{\alpha}$, $\hat{\beta}$ are independent (Chandler and Scott, 2011). The errors ϵ_t are assumed to form a sequence of independent and identically distributed random variables with mean zero and variance σ^2. The parameter α represents the mean global surface air temperature at time $t = \bar{t}$, i.e at the average time of the series, whereas β expresses

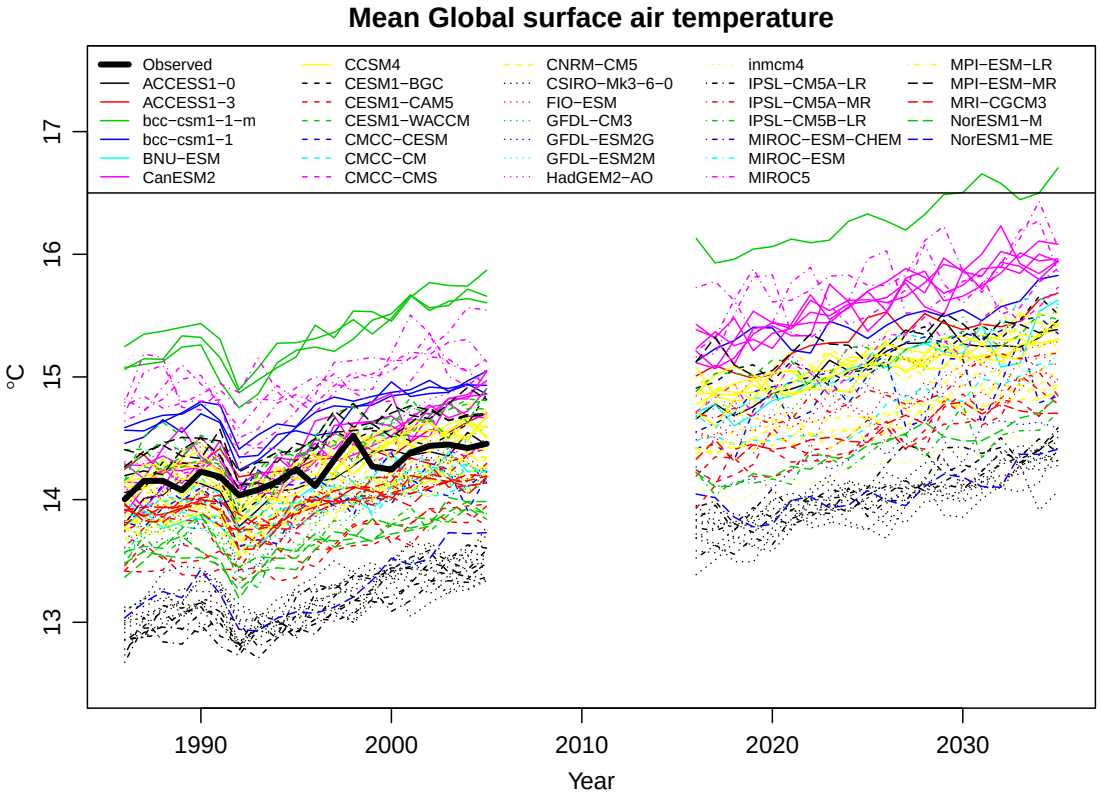


Figure 3.2: Mean global surface air temperature ($°C$) for the historical period 1986-2005 and the future period 2016-2035. The black line is the observed global surface air temperature, obtained from the HadCRUT3 observations. The rest of the lines represent the outputs from the 32 GCMs of the CMIP5 experiment used in this study. Lines with the same shape and colour correspond to output from different runs of the same GCM.

the mean annual change in global surface air temperature. As mentioned before, the mimic is fitted separately to each simulator and to the observations, leading to the parameters of interest having different values in each case. It is worth noting that based on the residual correlograms shown in Appendix C, there is evidence of residual autocorrelation in some cases. Ignoring this correlation has the limitation of potentially underestimating the variances of α, β and σ^2 and consequently the components of $\{\boldsymbol{J}_i, i = 0, \ldots, m\}$.

To summarize, according to this particular choice of mimic, the descriptors used to summarize yearly mean global surface air temperature for each time period are: α, β and σ^2. As mentioned at the beginning of Section 3.2, interest is on making inference about mean yearly global surface air temperature for the historical period $1986-2005$ as well as for the change in temperature between the future period $2016-2035$ and the historical. This is achieved by deriving a posterior distribution for the descriptor/s of actual climate, corresponding to the historical period and the change between future and historical periods. For this distribution to be derived, information from the descriptor estimators corresponding to observations and simulators is used, which is obtained by fitting the mimic to the data sources. Note that, as discussed in Section 2.6.1, the information from the data sources (observations and simulator outputs) is incorporated in the framework through the MLEs of α, β and σ^2, instead of directly using the raw data (Y_t in (3.1)) of observations and simulator outputs.

The three proposed implementations presented in Section 3.5 combine the information gained from fitting the mimic to observations and simulator outputs, to derive the required posterior distribution, under the framework of Section 2.6. Before the proposed implementations are described, some general conventions and notation are introduced in the next section.

3.4 General Conventions

For a fully Bayesian implementation the descriptors α, β, σ^2 in each period can be used directly, although it will be convenient to reparametrise in some places (see below). For all implementations, it is helpful to be able to assign distributions independently (as far as possible) to different quantities; and this requires different parametrisations at different levels in the hierarchical framework. For the first two proposed implementations, where the parameters of interest are assigned a joint multivariate normal distribution, the parameters are collected in a vector $\boldsymbol{\theta}_i$, for each simulator i ($i = 0, \ldots, m$). Note also that the error variance σ^2 is more appropriately expressed in logarithmic terms, in order to avoid negative values for the variance,

under normality assumptions. Consequently, $\boldsymbol{\theta}_i$ contains the historical values of α, β and $\log(\sigma^2)$ and the differences between future and historical values of those parameters, corresponding to simulator i. This is because the differences are reasonably considered as being independent of the corresponding historical values, conditional on the simulator consensus $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$.

The first proposed implementation is a modified version of the PM implementation of Section 2.6.3. It specifies multivariate Normal distributions for the descriptors, the descriptor estimators and the shared simulator discrepancy from reality. The second implementation is a fully Bayesian implementation, based on the same distributional assumptions as above. In the extra level of hierarchy, all the hyperpriors of the covariance matrices are assigned Wishart distributions. The third implementation is a fully Bayesian implementation, where a multivariate Normal distribution is specified jointly for (α, β) and a Gamma distribution for $1/\sigma^2$. The same holds for the distributions of descriptor estimators and the shared simulator discrepancy from reality corresponding to (α, β) and $1/\sigma^2$. In the extra level of hierarchy, all the hyperpriors relevant to (α, β) are assigned Wishart distributions and the ones relevant to $1/\sigma^2$ are assigned Gamma distributions. So, for the two jointly Gaussian implementations, the descriptors are defined as follows:

$$\boldsymbol{\theta_i} := \left(\alpha_i^{(\text{hist})}, \beta_i^{(\text{hist})}, \log\left(\sigma_i^{2\ (\text{hist})}\right), \alpha_i^{(\text{fut})} - \alpha_i^{(\text{hist})}, \beta_i^{(\text{fut})} - \beta_i^{(\text{hist})}, \log\left(\sigma_i^{2\ (\text{fut})}/\sigma_i^{2\ (\text{hist})}\right)\right) (i = 0, \ldots, m). \quad (3.2)$$

Although $\boldsymbol{\theta_i}$ is chosen as the descriptor vector of substantive interest, because the final three components relate to future climate change, for some of the work presented below it will be convenient to work instead with a different parametrisation:

$$\boldsymbol{\theta}_i^* = \left(\alpha_i^{(\text{hist})}, \beta_i^{(\text{hist})}, \log\left(\sigma_i^{2\ (\text{hist})}\right), \alpha_i^{(\text{fut})}, \beta_i^{(\text{fut})}, \log\left(\sigma_i^{2\ (\text{fut})}\right)\right) (i = 0, \ldots, m). \quad (3.3)$$

Then, it holds that:

$$\boldsymbol{\theta}_i = \boldsymbol{A}\boldsymbol{\theta_i^*}, \quad (3.4)$$

where:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix}. \quad (3.5)$$

This enables specifications of means and covariance matrices for $\boldsymbol{\theta}^*$ to be translated easily back to those for $\boldsymbol{\theta}$ (and vice versa).

In the third implementation, the descriptors are defined to be the parameters α, β and $1/\sigma^2$ that summarize yearly mean global surface air temperature for the historical period and the change between future and historical periods. Therefore, for the third implementation it is convenient to define the following notation:

$$\boldsymbol{\mu}_i := (\alpha_i, \beta_i), \tag{3.6}$$

and

$$\psi_i := 1/\sigma_i^2. \tag{3.7}$$

In the absence of a shared simulator discrepancy $\boldsymbol{\omega}$, the framework illustrated in Figure 2.1 would be a standard hierarchical model in which it is conventional (Tebaldi and Sansó, 2009; Gelman, 2006) to refer to distributions at the "data" level (variation of $\hat{\boldsymbol{\theta}}_i$ about $\boldsymbol{\theta}_i$), the "prior" level (variation of $\boldsymbol{\theta}_i$ about the simulator consensus) and "hyperprior" level (judgements about the consensus itself). However, the shared simulator discrepancy from reality - and the potential for additional levels in the hierarchy - means that this terminology is not appropriate for this particular problem. Therefore, in this study, the term "distribution" refers to the distributional assumptions about the descriptors, the descriptor estimators and the shared simulator discrepancy from reality, which are the random variables considered in Section 2.6.2. The distributional assumptions about any other quantities considered as random in the fully Bayesian implementations are referred to as "priors".

Sections 3.5.1-3.5.3 describe in detail the three proposed implementations.

3.5 The proposed implementations

Table 3.2 at the end of Section 3.5.3 summarizes the distributional assumptions for the random quantities considered in each of the three proposed implementations.

3.5.1 "Revised poor man's" (RPM) implementation

The first proposed implementation is a slightly modified version of the one described in Section 2.6.3, mainly because it proposes an alternative more robust estimator of the covariance matrix $\boldsymbol{\Lambda}$ expressing variability due to shared simulator discrepancy from reality.

The RPM implementation obtains estimates for the simulator descriptors and the framework's covariance matrices as described in Section 2.6.3 (with the exception of $\boldsymbol{\Lambda}$, in which case a more robust estimator is proposed here), by fitting the mimic to HadCRUT3 observations and simulator outputs from the CMIP5 experiment. Similarly, the prior parameters of $\boldsymbol{\theta}_0$ are determined using earlier observations and simulator outputs. The resulting descriptor estimates, together with the estimates of the covariance matrices and the prior parameters are plugged in (2.10)-(2.11), in order to obtain approximations of the posterior precision matrix $\boldsymbol{S^{-1}}$ and mean $\boldsymbol{\tau}$ of $\boldsymbol{\theta}_0$.

Expressions (3.8)-(3.14) below, illustrate the RPM implementation applied to global surface air temperature observations and simulator outputs. The Gaussian distributional assumptions introduced in Section 2.6.2 are retained in this implementation.

- Distribution of descriptor estimators $\{\hat{\boldsymbol{\theta}}_i, i = 0, \ldots, m\}$

  By definition, the MLEs $\{\hat{\boldsymbol{\theta}}_i,\ i = 1, \ldots, m\}$ have a multivariate Gaussian distribution, as shown in (2.4).

  The descriptor estimates $\{\hat{\boldsymbol{\theta}}_i,\ i = 1, \ldots, m\}$ are obtained by fitting the mimic to the observations and simulator outputs, separately for the historical and future periods. The estimates $\hat{\alpha}_i$ and $\hat{\beta}_i$ for each data source i are the MLEs of the regression coefficients, which are the least squares estimates of α and β respectively. The MLE of the residual variance in a regression model fitted to a sample of size n is $n^{-1}\sum_{t=1}^{n} e_t^2$, where e_t is the t^{th} residual. However, since this estimator is slightly biased, it is common practice to use the unbiased estimator $(n-q)^{-1}\sum_{t=1}^{n} e_t^2$, where q is the number of regression coefficients estimated. There is no loss of information by using the unbiased estimator instead of the MLE, since the latter is a one-to-one transformation of the MLE.

  Having obtained the estimates of all the components of each descriptor $\boldsymbol{\theta}_i$, the corresponding estimates of the covariance matrices $\{\boldsymbol{J}_i,\ i = 1, \ldots, m\}$ are easily deduced, using standard least-squares theory (Davison, 2003, Section 8.3). According to the definition of $\boldsymbol{\theta}_i$, ((3.2)), the off-diagonal blocks of the covariance matrices $\{\boldsymbol{J}_i,\ i = 0, \ldots, m\}$ will be non-zero, since the components of $\boldsymbol{\theta}_i$ corresponding to historical period and the change between future and historical period are not independent. This requires deriving expressions for the covariance of the "historical" and "historical-to-future change" components of $\boldsymbol{\theta}_i$ for each i. In order to avoid this task, it is more convenient to firstly determine the values of $\{\boldsymbol{J}_i^*,\ i = 1, \ldots, m\}$, the covariance matrices of $\{\hat{\boldsymbol{\theta}}_i^*,\ i = 1, \ldots, m\}$ (see (3.3)).

  Consider fixing i. By centering the linear time-trend regression model on $\bar{t}$, the

estimates $\hat{\alpha}_i^{(\text{hist})}$ and $\hat{\beta}_i^{(\text{hist})}$ conditional on the true regression coefficients $\alpha_i^{(\text{hist})}$ and $\beta_i^{(\text{hist})}$ respectively, become independent; moreover, the residual variance is independent of the regression coefficients. The same holds for the future period. Consequently, both the historical and future blocks of $\boldsymbol{J}_i^*$ have zero off-diagonal entries.

Standard results for least squares estimation (Chandler and Scott, 2011, Section 3.1.3) applied to the mimic structure in (3.1), show that the variances of the parameter estimates are:

$$Var(\hat{\alpha}_i) = \frac{\sigma_i^2}{T_i}, \tag{3.8}$$

$$Var(\hat{\beta}_i) = \frac{12\sigma_i^2}{T_i(T_i^2 - 1)}, \tag{3.9}$$

and

$$Var(\hat{\sigma}_i^2) = \frac{2\sigma_i^4}{T_i - 2}. \tag{3.10}$$

Here T_i is the number of data values to which the mimic is fitted for the i^{th} data source. For example, if a simulator provides 2 runs, then for the historical 20-year period, say, the output from the 2 runs is pooled to give $2 \times 20 = 40$ data values, giving $T_i = 40$.

In order to deduce the expression for $Var(log(\hat{\sigma}^2))$ from $Var(\hat{\sigma}^2)$, the method of "propagation of error" (Rice, 2007, p. 162) is used. It is based on a 2^{nd} order Taylor expansion of a function $g(Y)$ of a random variable Y around $E[Y]$, giving the following expression for $Var(g(Y))$:

$$Var(g(Y)) \simeq Var(Y)[g'(E[Y])]^2. \tag{3.11}$$

So, with $Y = \hat{\sigma}_i^2$, $g(Y) = \log(Y)$ and $E(Y) = \sigma_i^2$, $Var(\log(\hat{\sigma}_i^2))$ is approximated as:

$$Var\left(log\left(\hat{\sigma}_i^2\right)\right) \simeq \frac{2}{T_i - 2}. \tag{3.12}$$

Since historical and future periods are far apart, it is assumed that estimates of historical and future descriptor components are independent of each other, conditional on the actual descriptors. The validity of this assumption is checked by producing scatter plots of historical versus future values of the estimated descriptor

components $\hat{\boldsymbol{\alpha}}_i$ and $\hat{\boldsymbol{\beta}}_i$, corresponding to the simulators $i = 1, \ldots, 32$. These are shown in Figure D.1 of Appendix D. The scatter plots do not reveal any violation of the assumption of independence in the two periods, with the exception of $\hat{\alpha}_i^{(\text{hist})}$ Vs $\hat{\alpha}_i^{(\text{fut})}$, in which case a linear relationship is suggested. In physical terms, this suggests a linear relationship in the mean temperature at the time average, between the historical and future periods. This is possibly since the simulators predict a linear increase in mean temperature for the time periods considered in this study. Ignoring this correlation is considered as a limitation of the proposed implementations. However, since the covariance matrices $\{\boldsymbol{J}_i,\ i = 1, \ldots, m\}$ represent uncertainty due to natural variability, which is not considered to be a dominant source of uncertainty in the framework (according to Section 2.5.2.1), violation of the assumption is not expected to seriously affect the posterior of $\boldsymbol{\theta}_0$. For convenience therefore, all the off-diagonal entries of $\boldsymbol{J}_i^*$ can be set to zero. The diagonal entries are determined from (3.8)- (3.12). $\boldsymbol{J}_i^*$ is most conveniently defined via its inverse, i.e. the precision matrix $\boldsymbol{J}_i^{*\,-1}$, which takes the following form:

$$\boldsymbol{J}_i^{*\,-1} = \begin{pmatrix} \frac{T_i^{(\text{hist})}}{\sigma_i^{2\ (\text{hist})}} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{T_i^{(\text{hist})}\left(T_i^{2\ (\text{hist})}-1\right)}{12\sigma_i^{2\ (\text{hist})}} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{T_i^{(\text{hist})}-2}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{T_i^{(\text{fut})}}{\sigma_i^{2(\text{fut})}} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{T_i^{(\text{fut})}\left(T_i^{2\ (\text{fut})}-1\right)}{12\sigma_i^{2\ (\text{fut})}} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{T_i^{(\text{fut})}-2}{2} \end{pmatrix}. \tag{3.13}$$

Since $\sigma_i^{2\ (\text{hist})}$ and $\sigma_i^{2\ (\text{fut})}$ are unknown, $\boldsymbol{J}_i^{*\,-1}$ is estimated by $\hat{\boldsymbol{J}}_i^{*\,-1}$, where the historical and future components of σ_i^2 are substituted by their estimates.

Note also that the future components of the descriptor estimate $\hat{\boldsymbol{\theta}}_0^*$ are undefined due to the lack of future observations. This is equivalent to setting the future block of $\boldsymbol{J}_0^{*\,-1}$ to zero, as also illustrated in (2.5). By doing this, $\hat{\boldsymbol{\theta}}_0^{(fut)}$ can be set equal to a zero vector without influencing the framework, since otherwise $\hat{\boldsymbol{\theta}}_0^*$ would consist of fewer components than the rest of the descriptors, leading to computational problems in the software implementation for the calculation of the posterior.

Having obtained $\hat{\boldsymbol{J}}_i^{*\,-1}$, the corresponding matrix $\hat{\boldsymbol{J}}_i^{-1}$ can be calculated using (3.4) and (3.5) as $\hat{\boldsymbol{J}}_i^{-1} = \left(\boldsymbol{A}^T\right)^{-1} \hat{\boldsymbol{J}}_i^{*\,-1} \boldsymbol{A}^{-1}$.

- Distribution of simulator outputs' descriptors $\{\boldsymbol{\theta}_i,\ i = 1, \ldots, m\}$

The distribution of the descriptors $\{\boldsymbol{\theta}_i,\ i = 1, \ldots, m\}$ is expressed through discrepancies $\{\boldsymbol{\delta}_i,\ i = 1, \ldots, m\}$, as in Expression (2.6) of Section 2.6.2.

In the present application, all the simulators are assumed to deviate equally from simulator consensus. To judge whether this assumption is reasonable requires careful understanding of how each simulator works. Additionally, as suggested in Chandler (2013), simulators belonging to the same modelling group could be grouped together by adding additional levels on the hierarchical framework. This is handled in the proposed extended framework described in Chapter 4. For the present chapter however, since all the simulators used are state of the art GCMs and for the sake of simplicity, all the simulators are assigned a common $\boldsymbol{C}$. In this case, $\boldsymbol{C}$ is estimated as in (2.12), with m_S being the total number of simulators.

- Distribution of shared simulator discrepancy $\boldsymbol{\omega}$

  In (2.6), the quantity $\boldsymbol{\omega}$ represents the shared discrepancy of simulators from reality. Following the Gaussian assumptions of Section 2.6.2, the shared simulator discrepancy is assumed to be distributed as in (2.7).

  The covariance matrix $\boldsymbol{\Lambda}$ in the above expression represents the propensity of simulators collectively to deviate from reality. Chandler (2013) suggested the estimator given in (2.13). However, the choice of K in this estimator is quite subjective, since in the absence of observations of future reality, it is not easy to determine the degree of dependence of future discrepancy upon historical. Chandler (2013) pointed out that subjective judgements are unavoidable when attempting to make inference about the future, and recommended carrying out a sensitivity analysis to explore to what extent the choice of K affects the posterior of interest.

  In order to avoid the subjective choice of K, an alternative could be to estimate $\boldsymbol{\Lambda}$ as $\hat{\boldsymbol{\Lambda}} = \hat{\boldsymbol{\omega}}\hat{\boldsymbol{\omega}}'$, where $\hat{\boldsymbol{\omega}} = \bar{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{\theta}}_0$ (see Section 2.6.3), using a pair of past 20-year periods, for which there exist both observations of real climate, as well as simulator outputs. Then all the components of $\hat{\boldsymbol{\omega}}$ could be estimated and therefore $\hat{\boldsymbol{\Lambda}}$ could be obtained without requiring the use of K. However, this would give a non-invertible matrix $\hat{\boldsymbol{\Lambda}}$, since it would be attempted to estimate a 6×6 matrix using only a 6-element vector ($\hat{\boldsymbol{\omega}}$).

  To overcome these limitations, a new estimator of $\boldsymbol{\Lambda}$ is proposed in this thesis. The proposed estimator generates realisations from a population of pairs of past periods and uses them to obtain individual estimates $\hat{\boldsymbol{\omega}}_i$, each corresponding to the i^{th} pair of periods. For consistency with the current application, the two periods in each pair are treated analogously to the historical and future periods of interest. Similarly to the periods under study, they are both 20-year periods and have equal

year-difference to that between the historical and future periods of interest. For convenience, the earliest period of the pair will be referred to as the "historical" period and the other as the "future" period. In order to avoid any effects of particular pairs of periods in estimating $\mathbf{\Lambda}$, the bootstrapping method is applied.

Bootstrapping is a useful method for various statistical inference problems and has the important advantage of being independent of the assumptions underlying the distribution of the data (Lahiri, 2003). It was first introduced by Efron (1979) for the purpose of estimating the sampling distribution of a random variable, given observed data. The idea behind bootstrapping is simple. According to Lahiri (2003) it attempts to make inference about the population, using the observed data, which are considered as realizations of the underlying population. This is achieved by resampling with replacement from the observed data in a way such that the relation between the "resamples" or, bootstrap replicates and the observed data is representative of the relation between the population and the observed data.

To obtain the bootstrap pairs of periods, the "historical " periods are obtained by sampling with replacement 100 samples of 20-year periods, each period starting from one of the years $1860 - 1920$. The corresponding "future" periods are then obtained. This gives a bootstrap sample of 100 pairs of periods. For each bootstrap pair i, the corresponding $\hat{\boldsymbol{\omega}}_i$ is obtained, using observations from the HadCRUT3 dataset and simulator outputs from the 32 GCMs involved in the study (see Section 3.2). For each i, in order to avoid any effects of particular simulators in the calculation of $\hat{\boldsymbol{\omega}}_i$, bootstrapping among the 32 available simulators is performed. A bootstrap sample of 32 simulators is obtained by sampling with replacement from the original 32 distinct simulators.

The steps for obtaining $\hat{\boldsymbol{\omega}}_i$ for a bootstrap pair i of periods are outlined below:

1. For the "historical" and "future" periods of the bootstrap pair i, evaluate $\hat{\boldsymbol{\theta}}_0$, by separately fitting the mimic ((3.1), p.40) to the HadCRUT3 observations corresponding to those periods.
2. Sample with replacement from the 32 distinct GCMs, to obtain a bootstrap sample of 32 GCMs.
3. Fit the mimic to each bootstrap simulators' output for the same pair of periods, to obtain $\{\hat{\boldsymbol{\theta}}_j, j = 1, \ldots, 32\}$.
4. Calculate $\hat{\boldsymbol{\omega}}_i = \bar{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{\theta}}_0$, where:

$$\bar{\hat{\boldsymbol{\theta}}} = m^{-1} \sum_{j=1}^{m} \hat{\boldsymbol{\theta}}_j,$$

and m is the number of simulators involved for calculating $\bar{\hat{\boldsymbol{\theta}}}$, i.e. $m = 32$.

The procedure described above is repeated for every bootstrap pair i ($i = 1, \ldots, 100$) of periods. Then $\boldsymbol{\Lambda}$ is estimated by $\hat{\boldsymbol{\Lambda}}_{boot}$, the sample covariance matrix of the $\{\hat{\boldsymbol{\omega}}_i,\ i = 1, \ldots, 100\}$.

Now that the distribution of shared simulator discrepancy $\boldsymbol{\omega}$ is specified (together with a bootstrap estimate of its covariance matrix $\boldsymbol{\Lambda}$), it remains to specify a prior distribution for the true climate descriptor $\boldsymbol{\theta}_0$.

- Prior distribution of true-climate descriptor $\boldsymbol{\theta}_0$

  The prior distribution of $\boldsymbol{\theta}_0$ is defined as in (2.8). Knowledge about past climate can be exploited, in order to make realistic assumptions about the distribution of $\boldsymbol{\theta}_0$, by specifying values for $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0^{-1}$. These values are obtained from the HadCRUT3 dataset. The idea is to obtain descriptor estimates by fitting the mimic to observations from multiple periods earlier than the periods of interest and use their mean and precision matrix to determine $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0^{-1}$ respectively, as follows:

  1. Extract the HadCRUT3 global surface air temperature observations from the following pairs of periods:
     $(1850 - 1869, 1880 - 1899)$,
     $(1870 - 1889, 1900 - 1919)$,
     $(1890 - 1909, 1920 - 1939)$,
     $(1910 - 1929, 1940 - 1959)$,
     $(1930 - 1949, 1960 - 1979)$.

     For consistency with the current application, each pair of periods consists of two 20-year periods, separated by the same number of years as the "historical-future" pair under study. The above periods are all such non-overlapping pairs prior to the periods of interest, for which HadCRUT3 observations of global surface air temperature exist.

  2. For each pair, fit the mimic to the corresponding HadCRUT3 observations, to obtain estimates of the descriptors, denoted as $\{\hat{\boldsymbol{\theta}}_{past}[i],\ i = 1, \ldots, 5\}$.

     The first three components of $\hat{\boldsymbol{\theta}}_{past}[i]$, correspond to the first period in each pair ("historical") and the rest to the second period ("future").

  3. Evaluate the mean of $\hat{\boldsymbol{\theta}}_{past}[i]$, denoted as $\bar{\hat{\boldsymbol{\theta}}}_{past}$ and defined as:

     $$\bar{\hat{\boldsymbol{\theta}}}_{past} = m^{-1} \sum_{i=1}^{m} \hat{\boldsymbol{\theta}}_{past}[i], \tag{3.14}$$

     where m is the number of pairs, i.e. $m = 5$ here.

4. Obtain $\boldsymbol{\Sigma}_0^{-1}$ using $\{\hat{\boldsymbol{\theta}}_{past}[i],\ i = 1, \ldots, 5\}$

   It is assumed *a priori* that the historical components of $\boldsymbol{\theta}_0$ (i.e the first three components) are independent of the components of change between historical and future periods. Note that *a posteriori* they are expected to be dependent. The assumption of independence *a priori* aims to make the prior less informative, so as it does not dominate in the posterior and therefore facilitates illustration of the differences in the performance of the three proposed implementations.

   The off-diagonal blocks of $\boldsymbol{\Sigma}_0^{-1}$ are therefore set equal to zero. The diagonal blocks are set equal to the corresponding diagonal blocks of $\hat{\boldsymbol{\Sigma}}_{past}^{-1}$, where $\boldsymbol{\Sigma}_{past}^{-1}$ is estimated as the sample precision matrix of $\{\hat{\boldsymbol{\theta}}_{past}[i],\ i = 1, \ldots, 5\}$. However, higher uncertainty is expected in specifying the distribution of the real-climate descriptor for the historical and future periods of interest, compared to the corresponding distribution for the earlier periods used to derive $\boldsymbol{\Sigma}_0^{-1}$. This expected increase in uncertainty is expressed by dividing $\hat{\boldsymbol{\Sigma}}_{past}^{-1}$ by 25 to allow for increased uncertainty and to ensure that the *a priori* beliefs about $\boldsymbol{\theta}_0$ do not dominate the data $\{\hat{\boldsymbol{\theta}}_i,\ i = 0, \ldots, m\}$ in the derivation of the posterior for $\boldsymbol{\theta}_0$.

Therefore, the prior parameters of the true-climate descriptor are set to be: $\boldsymbol{\mu}_0 = \bar{\hat{\boldsymbol{\theta}}}_{past}$ and

$$\boldsymbol{\Sigma}_0^{-1} = \begin{pmatrix} \frac{\hat{\boldsymbol{\Sigma}}_{past}^{-1}[1:3,1:3]}{25} & \underline{\mathbf{0}} \\ \underline{\mathbf{0}} & \frac{\hat{\boldsymbol{\Sigma}}_{past}^{-1}[4:6,4:6]}{25} \end{pmatrix},$$

where $\hat{\boldsymbol{\Sigma}}_{past}^{-1}[1:3, 1:3]$ and $\hat{\boldsymbol{\Sigma}}_{past}^{-1}[4:6, 4:6]$ are the diagonal blocks of $\hat{\boldsymbol{\Sigma}}_{past}^{-1}$, corresponding to the historical and the historical-to-future change components of $\boldsymbol{\theta}_0$ respectively. The zero off-diagonal blocks are denoted by $\underline{\mathbf{0}}$.

The estimates described above are plugged in Expressions (2.10)-(2.11) of Section 2.6.2, to derive analytical expressions for the posterior mean and precision matrix of the true-climate descriptor $\boldsymbol{\theta}_0$.

To summarize, the RPM implementation is computationally efficient, since it enables deriving analytical expressions for the posterior parameters of $\boldsymbol{\theta}_0$, by assigning estimates to the framework's covariance matrices. This however ignores uncertainty in the values of the covariance matrices, which is a limitation of the implementation. The issue is addressed in the second contribution of this thesis, which involves two proposed fully Bayesian implementations. The first, namely the "Gaussian fully Bayesian" (GFB) implementation, assigns priors instead of estimates to the unknown

precision matrices $\boldsymbol{C}^{-1}$ and $\boldsymbol{\Lambda}^{-1}$, while retaining the multivariate Gaussian distributions for the random variables considered in the RPM implementation. The proposed implementation is presented in the next section.

3.5.2 Gaussian Fully Bayesian (GFB) Implementation

In the GFB implementation, the multivariate Gaussian assumptions of the RPM implementation for the involved random quantities are retained. However, an additional level of hierarchy is inserted in the framework, by specifying priors for the unknown covariance matrices (or, equivalently, for the precision matrices). In the second version of the fully Bayesian implementation (FB implementation), the jointly multivariate Gaussian assumption is relaxed for the descriptors $\{\boldsymbol{\theta}_i, i = 0, \ldots, m\}$. Here, the GFB implementation is described.

For the GFB implementation, the descriptors $\{\boldsymbol{\theta}_i,\ i = 0, \ldots, m\}$ are the same as in the RPM implementation, defined in (3.2). The distributional assumptions for their estimators are defined below:

- Distribution of descriptor estimators $\{\hat{\boldsymbol{\theta}}_i,\ i = 0, \ldots, m\}$

  The MLEs $\{\hat{\boldsymbol{\theta}}_i,\ i = 0, \ldots, m\}$ have approximately a multivariate Gaussian distribution, as shown in (2.4).

  Since this is a fully Bayesian implementation, it is expected to assign distributions to the precision matrices $\boldsymbol{J}_i^{-1}$, in contrast with the RPM implementation in Section 3.5.1, where each $\boldsymbol{J}_i^{-1}$ is estimated by $\hat{\boldsymbol{J}}_i^{-1}$. Note however that each $\boldsymbol{J}_i^{-1}$ is a function of $\boldsymbol{\theta}_i$ (i.e. $\boldsymbol{J}_i^{-1} = \boldsymbol{J}^{-1}(\boldsymbol{\theta}_i)$). Thus, uncertainty in the value of each $\boldsymbol{J}_i^{-1}$ is expressed through the distribution of the corresponding descriptor $\boldsymbol{\theta}_i$. Therefore, it is not necessary to assign priors for $\{\boldsymbol{J}_i^{-1},\ i = 0, \ldots, m\}$.

  According to the Gaussian specification of Section 2.6.2, the distributional assumptions about each descriptor $\boldsymbol{\theta}_i$ is expressed as discrepancy ($\boldsymbol{\delta}_i$) from reality (see (2.6), p.27).

- Distribution of simulator outputs' descriptors $\{\boldsymbol{\theta_i}, i = 1, \ldots, m\}$

  The distribution of $\{\boldsymbol{\theta_i}, i = 1, \ldots, m\}$ is defined as in (2.6).

  In the GFB implementation, the unknown precision matrix $\boldsymbol{C}_i^{-1}$ instead of the covariance matrix $\boldsymbol{C}$ is assigned a prior distribution. This is because working with precision instead of covariance matrices often simplifies the mathematical analysis. For this study, it is conventional for performing the fully Bayesian implementations using the OpenBUGS software (Lunn et al., 2009). A commonly-used distribution

assigned as a prior for precision matrices of multivariate normally distributed random vectors is the Wishart distribution (or, equivalently, the inverse-Wishart distribution as a hyperpior for covariance matrices) (Gelman, 2006; Leith and Chandler, 2010; Liechty et al., 2004).

The Wishart distribution is defined as follows (Gelman et al., 2014):

Definition 1. *Suppose $\boldsymbol{W}$ is a symmetric, positive-definite $k \times k$ matrix. Then $\boldsymbol{W}$ has the Wishart distribution with $k \times k$ positive definite scale matrix $\boldsymbol{R}$ and degrees of freedom $v > 0$, if it has a probability density function given by:*

$$p(\boldsymbol{W}|\boldsymbol{R}, v) \propto |\boldsymbol{R}|^{-v/2}|\boldsymbol{W}|^{(v-k-1)/2} exp\left(-\frac{1}{2}tr(\boldsymbol{R}^{-1}\boldsymbol{W})\right).$$

The distribution has expectation $E[\boldsymbol{W}] = v\boldsymbol{R}$, and it is proper only if $v > k - 1$.

For consistency with the RPM implementation, for this particular application, the simulators are assumed to deviate equally from the overall simulator consensus. Therefore, a common $\boldsymbol{C}^{-1}$, which is assumed to follow a Wishart distribution, is assigned to all $\{\boldsymbol{\delta}_i,\ i = 1, \ldots, m\}$, as follows:

- Prior for $\boldsymbol{C}^{-1}$

$$\boldsymbol{C}^{-1} \sim Wishart(\boldsymbol{R}_1, v_1). \tag{3.15}$$

$\boldsymbol{R}_1$ is the scale matrix of the Wishart distribution and v_1 the corresponding degrees of freedom. To ensure that the resulting distribution is proper as in Definition 1, v_1 must be at least equal to the number of rows /columns of the square scale matrix $\boldsymbol{R}_1$. According to Definition 1, $\boldsymbol{R}_1$ has the same size as the 6×6 precision matrix $\boldsymbol{C}^{-1}$. Therefore, v_1 must be at least 6. To ensure that the prior for $\boldsymbol{C}^{-1}$ is not too informative and hence that it does not dominate in the posterior of interest, the degrees of freedom are chosen to be as few as possible, i.e. $v_1 = 6$.

It remains to specify values for the entries of the scale-matrix $\boldsymbol{R}_1 = {}^{E[\boldsymbol{C}^{-1}]}/_{v_1}$ (according to Definition 1), or equivalently, $\boldsymbol{R}_1^{-1} = v_1 E[\boldsymbol{C}]$, following the default parametrisation in OpenBUGS (Lunn et al., 2009). Since interest is on specifying the prior distribution of $\boldsymbol{C}^{-1}$, $E[\boldsymbol{C}]$ in this case represents the *a priori* expectation of $\boldsymbol{C}$. It is determined by considering the characterisation of $\boldsymbol{C}$ as representing the propensity of a simulator to deviate from the simulator consensus, based on the particular choice of the mimic for this study ((3.1), p.40). $E[\boldsymbol{C}]$ is calculated as the corresponding $\hat{\boldsymbol{C}}$ used in the RPM implementation, but for two 20-year periods

earlier than the periods considered in the study; the periods chosen are $1860-1879$ (treated as "historical" period) and $1890-1909$ (treated as "future" period). The period $1860-1879$ is randomly chosen. For consistency with the current application, the period $1890-1909$ is chosen so as it has the same time-lag from $1890-1909$ as the historical and future periods of interest. The resulting matrix is defined as $\boldsymbol{C}_{prior}$. The matrix $\boldsymbol{R}_1^{-1}$ therefore becomes: $\boldsymbol{R}_1^{-1} = v_1 E[\boldsymbol{C}] = 6\boldsymbol{C}_{prior}$.

The next step is to assign a distribution to the shared discrepancy $\boldsymbol{\omega}$.

- Distribution of shared simulator discrepancy $\boldsymbol{\omega}$

  This is defined to be as in (2.7).

  In the GFB implementation, a prior distribution is assigned to the unknown precision matrix $\boldsymbol{\Lambda}^{-1}$. Similarly to the prior for $\boldsymbol{C}^{-1}$, $\boldsymbol{\Lambda}^{-1}$ is assumed to follow the Wishart distribution:

- Prior for $\boldsymbol{\Lambda}^{-1}$

$$\boldsymbol{\Lambda}^{-1} \sim Wishart(\boldsymbol{R}_2, v_2). \tag{3.16}$$

  The number of degrees of freedom is chosen to be the least possible, again to ensure that the prior for $\boldsymbol{\Lambda}^{-1}$ is not too informative and hence that it does not dominate the posterior of interest. So $v_2 = 6$. In this case, $\boldsymbol{R}_2^{-1} = v_2 E[\boldsymbol{\Lambda}]$, according to Definition 1. For consistency with the judgements about discrepancy in the RPM implementation, $E[\boldsymbol{\Lambda}]$ is estimated by $\hat{\boldsymbol{\Lambda}}_{boot}$, the bootstrap estimate of $\boldsymbol{\Lambda}$, obtained in the RPM implementation of Section 3.5.1. Consequently, $\boldsymbol{R}_2^{-1} = 6\hat{\boldsymbol{\Lambda}}_{boot}$.

  It remains to specify a distribution for the true-climate descriptor $\boldsymbol{\theta}_0$.

- Distribution of true-climate descriptor $\boldsymbol{\theta}_0$

  The distribution is set to be as in (2.8). To retain consistency between the implementations in the *a priori* judgements about reality, the values of $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are set to be equal to those in the "revised poor-man's" implementation (see end of Section 3.5.1).

The model for the GFB implementation is now fully defined, since all the unknown random quantities in the framework have been assigned distributions and all the unknown precision matrices have been assigned priors. It remains to define the FB implementation, where the multivariate Gaussian distributional assumption does not apply to all the components of $\{\boldsymbol{\theta}_i\}$. A detailed description of the proposed implementation is provided in the next section.

3.5.3 Fully Bayesian (FB) Implementation

In both implementations described so far, the descriptors are assumed to be drawn from the multivariate Gaussian distribution. An alternative implementation could be to treat some components of the descriptor $\boldsymbol{\theta}_i$ independently and assign to them a different distribution. For example, it is common in the literature to assign a Gamma distribution to the precision of Normally distributed data (or, equivalently, an inverse Gamma distribution to the variance) (Gelman, 2006; Ghosh and Dunson, 2009). The Gamma prior is particularly convenient in a fully Bayesian analysis, since it is a conjugate prior for Gaussian data. Note also that the log-normal distribution assigned to σ^2 in the GFB implementation can often exhibit long skewness and therefore potentially wider tails than the inverse-Gamma distribution. Therefore, it would be of interest to perform a sensitivity analysis based on these two prior choices for σ^2, in order to investigate their effect on probabilities of exceedances when moving away from the distribution mean. Under the above considerations, the proposed implementation aims to provide a fully Bayesian analysis, by treating the residual variance parameter of the mimic (σ^2) separately from the other two parameters (α and β) and assign Gamma priors to the residual precision $1/\sigma^2$ (or, equivalently, inverse-Gamma priors to σ^2).

As introduced in (3.6), for the subsequent discussion, it is convenient to collect together the regression parameters α and β in a vector $\boldsymbol{\mu}_i := (\alpha_i, \beta_i)$.

The distributional assumptions for the descriptor estimators are shown below:

- Distribution of descriptor estimators

  As also discussed in the RPM implementation (Section 3.5.1), the historical and future estimators are assumed to be independent, conditional on the actual descriptors. This allows expressing the distributional assumptions for the descriptor estimators by considering the estimators corresponding to the historical and future periods individually.

  For the estimators $\hat{\boldsymbol{\mu}}_i^{(\text{hist})} = \left(\alpha_i^{(\text{hist})}, \beta_i^{(\text{hist})}\right)$ and $\hat{\boldsymbol{\mu}}_i^{(\text{fut})} = \left(\alpha_i^{(\text{fut})}, \beta_i^{(\text{fut})}\right)$, the multivariate Normal distribution is assigned. The estimators are centred on the true descriptors $\boldsymbol{\mu}_i^{(\text{hist})}$ and $\boldsymbol{\mu}_i^{(\text{fut})}$ respectively, according to the framework. So,

$$\hat{\boldsymbol{\mu}}_i^{(\text{hist})} | \boldsymbol{\mu}_i^{(\text{hist})} \sim MVN\left(\boldsymbol{\mu}_i^{(\text{hist})}, \boldsymbol{L}_i^{(\text{hist})}\right) (i = 0, \ldots, m),$$

  and

$$\hat{\boldsymbol{\mu}}_i^{(\text{fut})} | \boldsymbol{\mu}_i^{(\text{fut})} \sim MVN\left(\boldsymbol{\mu}_i^{(\text{fut})}, \boldsymbol{L}_i^{(\text{fut})}\right) (i = 0, \ldots, m).$$

The covariance matrices $\boldsymbol{L}_i^{(\text{hist})}$ and $\boldsymbol{L}_i^{(\text{fut})}$ for each data source i are determined according to standard least squares theory for the variance of regression parameters (Davison, 2003, Section 8.3). These considerations were also made for determining the corresponding sub-blocks of $\boldsymbol{J}_i^{*\,-1}$ in the RPM implementation ((3.13), p.47). In particular,

$$\boldsymbol{L}_i^{(\text{hist})} = \begin{pmatrix} \frac{T_i^{(\text{hist})}}{\sigma_i^{2\ (\text{hist})}} & 0 \\ 0 & \frac{T_i^{(\text{hist})}\left(T_i^{2\ (\text{hist})}-1\right)}{12\sigma_i^{2\ (\text{hist})}} \end{pmatrix},$$

and

$$\boldsymbol{L}_i^{(\text{fut})} = \begin{pmatrix} \frac{T_i^{(\text{fut})}}{\sigma_i^{2\ (\text{fut})}} & 0 \\ 0 & \frac{T_i^{(\text{fut})}\left(T_i^{2\ (\text{fut})}-1\right)}{12\sigma_i^{2\ (\text{fut})}} \end{pmatrix}.$$

$T_i^{(\text{hist})}$ and $T_i^{(\text{fut})}$ represent the number of data values to which the mimic is fitted for the i^{th} data source, for the historical and future periods respectively.

For the residual variances $\sigma_i^{2\ (\text{hist})}$ and $\sigma_i^{2\ (\text{fut})}$ in each data source i, the standard result about the distribution of the corresponding estimators $\hat{\sigma}_i^{2\ (\text{hist})}$ and $\hat{\sigma}_i^{2\ (\text{fut})}$ is used (Rice, 2007, p. 197) . According to this result, the estimators have a sampling distribution proportional to chi-squared, or equivalently an exact gamma distribution. Specifically,

$$\hat{\sigma}_i^{2\ (\text{hist})} | \sigma_i^{2\ (\text{hist})} \sim Gamma\left(\frac{T_i^{(\text{hist})}-2}{2}, \frac{T_i^{(\text{hist})}-2}{2\sigma_i^{2\ (\text{hist})}}\right) (i = 0, \ldots, m),$$

and

$$\hat{\sigma}_i^{2\ (\text{fut})} | \sigma_i^{2\ (\text{fut})} \sim Gamma\left(\frac{T_i^{(\text{fut})}-2}{2}, \frac{T_i^{(\text{fut})}-2}{2\sigma_i^{2\ (\text{fut})}}\right) (i = 0, \ldots, m),$$

where $T_i^{(\text{hist})}$ and $T_i^{(\text{fut})}$ represent the number of Y_t output values from simulator i to which the mimic is fitted, for the historical and future periods respectively.

The next step is to assign distributions to the actual descriptors of the simulator outputs.

- Distribution of simulator outputs' descriptors

As in the previous two implementations, the distributional assumptions of the simulator outputs' descriptors are expressed via discrepancies from reality, under the assumption that for each descriptor, conditional on the shared simulator discrepancy $\boldsymbol{\omega}$ corresponding to each descriptor, the discrepancies are independent between the simulators.

The Gaussian assumptions are retained for $\{\boldsymbol{\mu}_i^{(\text{hist})},\ i = 1, \ldots, m\}$. Similarly to the RPM implementation, the change between historical and future values of $\boldsymbol{\mu}_i$ is considered additively, as: $\boldsymbol{\mu}_i^{(\text{fut})} - \boldsymbol{\mu}_i^{(\text{hist})}$. Consequently, the change is also assigned a multivariate Normal distribution. Identically to the RPM implementation, the shared simulator discrepancies of $\boldsymbol{\mu}_i^{(\text{hist})}$ and $\boldsymbol{\mu}_i^{(\text{fut})} - \boldsymbol{\mu}_i^{(\text{hist})}$ from reality, defined as $\boldsymbol{\omega}_{\boldsymbol{\mu}^{(\text{hist})}}$ and $\boldsymbol{\omega}_{\boldsymbol{\Delta\mu}}$ respectively, are considered additively. Equivalently, $E\left[\boldsymbol{\mu}^{(\text{hist})}\right] = \boldsymbol{\mu}_0^{(\text{hist})} + \boldsymbol{\omega}_{\boldsymbol{\mu}^{(\text{hist})}}$ and $E\left[\boldsymbol{\mu}^{(\text{fut})} - \boldsymbol{\mu}^{(\text{hist})}\right] = \left(\boldsymbol{\mu}_0^{(\text{fut})} - \boldsymbol{\mu}_0^{(\text{hist})}\right) + \boldsymbol{\omega}_{\boldsymbol{\Delta\mu}}$. Note also that, conditional on the simulator consensus corresponding to each descriptor, the historical discrepancies can be considered as independent of the corresponding discrepancies for the change between historical and future periods. This allows assigning distributions to the quantities $\boldsymbol{\mu}_i^{(\text{hist})} - \boldsymbol{\mu}_0^{(\text{hist})}$ and $\left(\boldsymbol{\mu}_i^{(\text{fut})} - \boldsymbol{\mu}_i^{(\text{hist})}\right) - \left(\boldsymbol{\mu}_0^{(\text{fut})} - \boldsymbol{\mu}_0^{(\text{hist})}\right)$ individually.

All the above lead to the following expressions:

$$\boldsymbol{\mu}_i^{(\text{hist})} - \boldsymbol{\mu}_0^{(\text{hist})} | \boldsymbol{\omega}_{\boldsymbol{\mu}^{(\text{hist})}} \sim MVN\left(\boldsymbol{\omega}_{\boldsymbol{\mu}^{(\text{hist})}}, \boldsymbol{\phi}_{\boldsymbol{\mu}^{(\text{hist})}}\right) (i = 1, \ldots, m), \tag{3.17}$$

and

$$\left(\boldsymbol{\mu}_i^{(\text{fut})} - \boldsymbol{\mu}_i^{(\text{hist})}\right) - \left(\boldsymbol{\mu}_0^{(\text{fut})} - \boldsymbol{\mu}_0^{(\text{hist})}\right) | \boldsymbol{\omega}_{\boldsymbol{\Delta\mu}} \sim MVN\left(\boldsymbol{\omega}_{\boldsymbol{\Delta\mu}}, \boldsymbol{\phi}_{\boldsymbol{\Delta\mu}}\right) (i = 1, \ldots, m). \tag{3.18}$$

The distributional assumptions about historical and future precision $\psi_i^{(\text{hist})}$ and $\psi_i^{(\text{fut})}$ (see (3.7), p.44) are also expressed as discrepancies of descriptors from reality. In addition, any change in precision is considered multiplicatively in this study. Therefore, interest is on the distributions of $\psi_i^{(\text{hist})}/\psi_0^{(\text{hist})}$ and $\left(\psi_i^{(\text{fut})}/\psi_i^{(\text{hist})}\right)/\left(\psi_0^{(\text{fut})}/\psi_0^{(\text{hist})}\right)$. Since, conditional on the simulator consensus, any change between historical and future precision can be considered as independent of the historical precision, the quantities $\psi_i^{(\text{hist})}/\psi_0^{(\text{hist})}$ and $\left(\psi_i^{(\text{fut})}/\psi_i^{(\text{hist})}\right)/\left(\psi_0^{(\text{fut})}/\psi_0^{(\text{hist})}\right)$ can be modelled individually.

A Gamma distribution is assigned to the historical precision, such that: $E\left[\psi_i^{(\text{hist})}\right] = \omega_{\psi^{(\text{hist})}} \times \psi_0^{(\text{hist})}$, where $\omega_{\psi^{(\text{hist})}}$ is the shared simulator discrepancy corresponding to $\psi_i^{(\text{hist})}$. Similarly for the distribution of future-to-historical precision ratio. In an obvious notation, the shared simulator discrepancy corresponding to $\psi_i^{(\text{fut})}/\psi_i^{(\text{hist})}$

is defined as $\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}}$.

The expressions below illustrate the distributional assumptions related to the descriptors of the precision:

$$\psi_i^{(\text{hist})}/\psi_0^{(\text{hist})}|\omega_{\psi^{(\text{hist})}} \sim Gamma\left(v_3, \frac{v_3}{\omega_{\psi^{(\text{hist})}}}\right)(i = 1, \ldots, m), \tag{3.19}$$

and

$$\frac{\psi_i^{(\text{fut})}/\psi_i^{(\text{hist})}}{\psi_0^{(\text{fut})}/\psi_0^{(\text{hist})}}|\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}} \sim Gamma\left(v_4, \frac{v_4}{\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}}}\right)(i = 1, \ldots, m). \tag{3.20}$$

According to the framework, the shared simulator discrepancies $\boldsymbol{\omega}_{\boldsymbol{\mu}^{(\text{hist})}}$, $\boldsymbol{\omega}_{\boldsymbol{\Delta\mu}}$, $\omega_{\psi^{(\text{hist})}}$ and $\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}}$ of (3.17)-(3.20) are regarded as random quantities and are thus assigned distributions.

The unknowns $\boldsymbol{\phi}_{\boldsymbol{\mu}^{(\text{hist})}}$, $\boldsymbol{\phi}_{\boldsymbol{\Delta\mu}}$, v_3 and v_4 characterize the variation of the simulators from their consensus, which is assumed to be identical for all the simulators $\{i,\ i = 1, \ldots, m\}$. In a fully Bayesian implementation, those quantities are treated as random and, consequently, priors are assigned to them.

Firstly, the priors related to simulator consensus are defined:

- Priors for simulator consensus

  Since both $\boldsymbol{\phi}^{-1}_{\boldsymbol{\mu}^{(hist)}}$ and $\boldsymbol{\phi}^{-1}_{\boldsymbol{\Delta\mu}}$ are the precision matrices of Gaussian quantities, they are assumed to follow Wishart distributions, similarly to the precision matrix $\boldsymbol{C}^{-1}$ considered in the GFB implementation, i.e.

$$\boldsymbol{\phi}^{-1}_{\mu^{(\text{hist})}} \sim Wishart\left(\boldsymbol{R}_3, s_1\right),$$

  and

$$\boldsymbol{\phi}^{-1}_{\boldsymbol{\Delta\mu}} \sim Wishart\left(\boldsymbol{R}_4, s_2\right).$$

  For consistency with uncertainty judgements related to the distribution of $\boldsymbol{C}^{-1}$ in the GFB implementation ((3.15), p.53), the degrees of freedom s_1 and s_2 are chosen to be identical to v_1 in the GFB implementation. Therefore, $s_1 = s_2 = 6$. Also, the values for $\boldsymbol{R}_3^{-1}$ and $\boldsymbol{R}_4^{-1}$ are set to be equal to $\boldsymbol{R}_1^{-1}[1:2, 1:2]$ and $\boldsymbol{R}_1^{-1}[4:5, 4:5]$ respectively, where $\boldsymbol{R}_1$ was defined in p.53. This ensures that the prior expectations of $\boldsymbol{\phi}^{-1}_{\boldsymbol{\mu}^{(\text{hist})}}$ and $\boldsymbol{\phi}^{-1}_{\boldsymbol{\Delta\mu}}$ are identical to the corresponding quantities in the GFB implementation.

In order to specify priors for v_3 and v_4, it is important to interpret their role as parameters of the Gamma distribution. From the definition of the moments of a Gamma distribution, it turns out that $v_3 = {}^1\!/\!_{CV^2}\left(\psi_i^{(\text{hist})}\right)$, where $CV\left(\psi_i^{(\text{hist})}\right) = {}_{SD}\left(\psi_i^{(\text{hist})}\right)/_E\left(\psi_i^{(\text{hist})}\right)$ is the coefficient of variation of the historical precision.

By definition, v_3 and v_4 can only take positive values, thus Gamma distributions can be assigned to them:

$$v_3 \sim Gamma\left(\rho_1, \frac{\rho_1}{\eta_1}\right), \tag{3.21}$$

and

$$v_4 \sim Gamma\left(\rho_2, \frac{\rho_2}{\eta_2}\right). \tag{3.22}$$

By definition of the expectation of a Gamma distribution, the parameters η_1 and η_2 are equal to the expected values of v_3 and v_4 respectively, i.e. $\eta_1 = E\left[{}^1\!/\!_{CV^2}\left(\psi_i^{(\text{hist})}\right)\right]$ and $\eta_2 = E\left[{}^1\!/\!_{CV^2}\left({}_{\psi_i^{(\text{fut})}}/_{\psi_i^{(\text{hist})}}\right)\right]$. η_1 and η_2 are the *a priori* expectations of ${}^1\!/\!_{CV^2}\left(\psi_i^{(\text{hist})}\right)$ and ${}^1\!/\!_{CV^2}\left({}_{\psi_i^{(\text{fut})}}/_{\psi_i^{(\text{fut})}}\right)$ respectively. In order to retain consistency with *a priori* judgements about consensus in the GFB implementation, the same data sources used for specifying the parameters of the distribution of $\boldsymbol{C}^{-1}$ ((3.15)) are also used here. The parameters η_1 and η_2 are determined by calculating ${}^1\!/\!_{CV^2_{prior}}\left(\psi_i^{(\text{hist})}\right)$ and ${}^1\!/\!_{CV^2_{prior}}\left({}_{\psi_i^{(\text{fut})}}/_{\psi_i^{(\text{hist})}}\right)$, where $CV^2_{prior}\left(\psi_i^{(\text{hist})}\right)$ and $CV^2_{prior}\left({}_{\psi_i^{(\text{fut})}}/_{\psi_i^{(\text{hist})}}\right)$ are calculated using the 32 CMIP5 GCMs involved in the study, for the 20-year periods $1860-1879$ and $1890-1909$ treated as the "historical" and "future" periods respectively.

The shape parameter ρ_1(of the distribution of v_3) is set equal to 1, again to ensure consistency in *a priori* judgements about consensus in the two fully Bayesian implementations. Here, consistency is checked by considering the distributions of $\boldsymbol{C}[3,3]$ ((3.15), p.53) and v_3 ((3.21), p.59) in the two fully Bayesian implementations. More details are given in Appendix E. Similarly, ρ_2 is also set equal to 1.

Next, the distributional assumptions about shared discrepancy of the descriptors from reality are specified:

- Distribution of shared discrepancy

  Since $\boldsymbol{\mu}^{(\text{hist})}$ and $\boldsymbol{\mu}^{(\text{fut})} - \boldsymbol{\mu}^{(\text{hist})}$ are assumed to follow multivariate Gaussian distributions, it is sensible to consider their discrepancies from reality to also be multivariate Gaussian. For consistency with the previous implementations, the discrepancies are assumed to be additive for those two descriptors, and according to the framework, have expected values equal to zero. Therefore,

$$\boldsymbol{\omega}_{\boldsymbol{\mu}^{(\text{hist})}} \sim MVN\left(\mathbf{0}, \boldsymbol{\gamma}_1\right), \tag{3.23}$$

and

$$\boldsymbol{\omega}_{\boldsymbol{\Delta\mu}} \sim MVN\left(\mathbf{0}, \boldsymbol{\gamma}_2\right). \tag{3.24}$$

The discrepancies $\omega_{\psi^{(\text{hist})}}$ and $\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}}$ of $\psi^{(\text{hist})}$ and $\psi^{(\text{fut})}/\psi^{(\text{hist})}$ respectively are considered as the collective multiplicative deviations of the simulators from reality. Since the discrepancies here are related to precisions, they take positive values only, therefore allowing the use of Gamma distributions to model them. Because the direction of discrepancy is unknown *a priori*, $E[\omega_{\psi^{(\text{hist})}}]$ and $E[\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}}]$ become equal to one, i.e. the shape and scale parameters of the Gamma priors are identical. So,

$$\omega_{\psi^{(\text{hist})}} \sim Gamma\left(\gamma_3, \gamma_3\right), \tag{3.25}$$

and

$$\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}} \sim Gamma\left(\gamma_4, \gamma_4\right). \tag{3.26}$$

The unknown parameters of (3.23)-(3.26) are assigned priors in a fully Bayesian analysis, which are defined as follows:

- Priors for shared discrepancy

  $\boldsymbol{\gamma}_1^{-1}$ and $\boldsymbol{\gamma}_2^{-1}$ are precision matrices of Gaussian variables. These are assigned Wishart distributions:

$$\boldsymbol{\gamma}_1^{-1} \sim Wishart\left(\boldsymbol{R}_5, s_5\right),$$

and

$$\boldsymbol{\gamma}_2^{-1} \sim Wishart\left(\boldsymbol{R}_6, s_6\right).$$

For consistency with *a priori* judgements about the shared simulator discrepancy in the two fully Bayesian implementations, the degrees of freedom s_5 and s_6 are chosen to be identical to v_2 in the prior for $\boldsymbol{\Lambda}^{-1}$ in the GFB implementation ((3.16), p.54). Therefore, $s_5 = s_6 = 6$. The values for $\boldsymbol{R}_5^{-1}$ and $\boldsymbol{R}_6^{-1}$ are then obtained from the sub-blocks of $\boldsymbol{R}_2^{-1}$ ((3.16)) in the GFB implementation that correspond to $\boldsymbol{\mu}_i^{(\text{hist})}$ and $\boldsymbol{\mu}_i^{(\text{fut})} - \boldsymbol{\mu}_i^{(\text{hist})}$ respectively. Precisely, $\boldsymbol{R}_5^{-1} = \boldsymbol{R}_2^{-1}[1:$

$2,1:2]$ and $\boldsymbol{R}_6^{-1} = \boldsymbol{R}_2^{-1}[4:5,4:5]$, where $\boldsymbol{R}_2$ was defined in p. 54. This ensures that the prior expectations of $\boldsymbol{\gamma}_1^{-1}$ and $\boldsymbol{\gamma}_2^{-1}$ are the same with those of the corresponding quantities in the GFB implementation.

It remains to specify the priors for the parameters γ_3 and γ_4 of the distributions of $\omega_{\psi^{(\text{hist})}}$ and $\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}}$ respectively. According to (3.25)-(3.26), $\gamma_3 = Var^{-1}(\omega_{\psi^{(\text{hist})}})$ and $\gamma_4 = Var^{-1}(\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}})$. Consequently, since they express precisions, they are assigned Gamma distributions:

$$\gamma_3 \sim Gamma\left(\rho_3, \frac{\rho_3}{\eta_3}\right), \tag{3.27}$$

and

$$\gamma_4 \sim Gamma\left(\rho_4, \frac{\rho_4}{\eta_4}\right). \tag{3.28}$$

By definition of the expectation of a Gamma distribution, $\eta_3 = E[\gamma_3] = E[Var^{-1}(\omega_{\psi^{(\text{hist})}})]$ and $\eta_4 = E[\gamma_4] = E[Var^{-1}(\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}})]$. η_3 and η_4 represent the *a priori* expectations of $Var^{-1}(\omega_{\psi^{(\text{hist})}})$ and $Var^{-1}(\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}})$ respectively. In order to retain consistency in the *a priori* judgements about discrepancy between the two fully Bayesian implementations, the same data sources used for specifying the parameters of the distribution of $\boldsymbol{\Lambda}^{-1}$ ((3.16), p.54) are also used here. Precisely, η_3 is determined by calculating the precision of the discrepancies of $\psi_i^{(\text{hist})}$, calculated from the same data used to determine $\boldsymbol{R_2}^{-1}$ in the GFB implementation ((3.16)). The dataset is the sample used for evaluating $\hat{\boldsymbol{\Lambda}}_{boot}$ in the RPM implementation (Section 3.5.1). The value of η_4 is calculated in a similar way.

For consistency in the *a priori* judgements about discrepancy between the two fully Bayesian implementations, the shape parameter ρ_3 is chosen to be equal to 10. The check for consistency is based on considering the distributions of $\boldsymbol{\Lambda}^{-1}[3,3]$ ((3.16), p.54) and γ_3 ((3.27), p.61) in the two implementations. More details are provided in Appendix E. Similarly, $\rho_4 = 10$.

According to the framework, it remains to specify distributions for the true-climate descriptors, namely $\boldsymbol{\mu}_0^{(\text{hist})}, \boldsymbol{\mu}_0^{(\text{fut})} - \boldsymbol{\mu}_0^{(\text{hist})}, \psi_0^{(\text{hist})}$ and $\psi_0^{(\text{fut})}/\psi_0^{(\text{hist})}$.

- Distribution of true-climate descriptors

  Similarly to the descriptors of simulator outputs, $\boldsymbol{\mu}_0^{(\text{hist})}$ and $\boldsymbol{\mu}_0^{(\text{fut})} - \boldsymbol{\mu}_0^{(\text{hist})}$ are assigned multivariate Gaussian distributions:

$$\boldsymbol{\mu}_0^{(\text{hist})} \sim MVN\left(\boldsymbol{\lambda}_1, \boldsymbol{\kappa}_1\right),$$

and

$$\boldsymbol{\mu}_0^{(\text{fut})} - \boldsymbol{\mu}_0^{(\text{hist})} \sim MVN\left(\boldsymbol{\lambda}_2, \boldsymbol{\kappa}_2\right).$$

For consistency in the prior judgements about the true-climate descriptors between the three implementations, the values of $\boldsymbol{\lambda}_1$, $\boldsymbol{\kappa}_1^{-1}$, $\boldsymbol{\lambda}_2$ and $\boldsymbol{\kappa}_2^{-1}$ are determined from the corresponding components of $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0^{-1}$, the mean vector and precision matrix of the distribution of true climate in the previous two implementations. Precisely,

$\boldsymbol{\lambda}_1 = \boldsymbol{\mu}_0[1:2, 1:2]$,
$\boldsymbol{\kappa}_1^{-1} = \boldsymbol{\Sigma}_0^{-1}[1:2, 1:2]$,
$\boldsymbol{\lambda}_2 = \boldsymbol{\mu}_0[4:5, 4:5]$,
$\boldsymbol{\kappa}_2^{-1} = \boldsymbol{\Sigma}_0^{-1}[4:5, 4:5]$.

It remains to specify distributions for the descriptors $\psi_0^{(\text{hist})}$ and $\psi_0^{(\text{fut})}/\psi_0^{(\text{hist})}$. Similarly to the corresponding descriptors from simulator outputs, they are both assigned Gamma distributions:

$$\psi_0^{(\text{hist})} \sim Gamma\left(v_5, \frac{v_5}{w_1}\right),$$

and

$$\frac{\psi_0^{(\text{fut})}}{\psi_0^{(\text{hist})}} \sim Gamma\left(v_6, \frac{v_6}{w_2}\right).$$

By definition of the moments of a Gamma distribution,

$w_1 = E\left[\psi_0^{(\text{hist})}\right]$,
$w_2 = E\left[\psi_0^{(\text{fut})}/\psi_0^{(\text{hist})}\right]$,
$v_5 = w_1^2/Var\left(\psi_0^{(\text{hist})}\right)$,
$v_6 = w_2^2/Var\left(\psi_0^{(\text{fut})}/\psi_0^{(\text{hist})}\right)$.

For consistency with the previous implementations, the quantities above are determined using the same datasets used for determining $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0^{-1}$ in the previous implementations. The procedure for obtaining w_1 and v_5 is described below:

1. Obtain estimates of $\{\hat{\psi}_{past}^{(\text{hist})}[j],\ j = 1, \ldots, 5\}$ by fitting the mimic to the HadCRUT3 observations, for the 5 pairs of periods considered in the previous implementations.
2. Determine $w_1 = E\left[\psi_0^{(\text{hist})}\right]$ to be the mean of $\{\hat{\psi}_{past}^{(\text{hist})}[j],\ j = 1, \ldots, 5\}$.

3. Determine $Var\left(\psi_0^{(\text{hist})}\right)$ using $\{\hat{\psi}_{past}^{(\text{hist})}[j],\ j = 1, \ldots, 5\}$.
   Define $s^2_{past} := Var\left(\hat{\psi}_{past}^{(\text{hist})}[j]\right),\ j = 1, \ldots, 5$. Then, for consistency with the previous implementations, $Var\left(\psi_0^{(\text{hist})}\right)$ is set equal to $25s^2_{past}$. This is because higher uncertainty is expected for the historical and future periods under study, compared to those used to determine the priors for true-climate descriptors. Therefore, the calculated variance s^2_{past} is multiplied by 25, in order to ensure that the *a priori* beliefs about the true-climate descriptors do not dominate over the data (descriptor estimators) in the derivation of the posterior of the true-climate descriptors.
4. Evaluate $v_5 = {}^{w_1^2}\!/_{25s^2_{past}}$.

Values for w_2 and v_6 are obtained in an similar way.

Table 3.2 summarizes the distributional assumptions about the random quantities considered in each of the three implementations.

Now that all the three proposed implementations are introduced, the results of the derived posteriors from each case can be analysed and compared in order to investigate whether the computationally efficient RPM implementation yields adequate approximations to the posterior of interest. The results are presented in the next section.

	IMPLEMENTATION		
	RPM	**Gaussian full-Bayes**	**Full-Bayes**
Data	$\hat{\boldsymbol{\theta}}_i \vert \boldsymbol{\theta}_i \sim MVN(\boldsymbol{\theta}_i, \boldsymbol{J}_i)$	$\hat{\boldsymbol{\theta}}_i \vert \boldsymbol{\theta}_i \sim MVN(\boldsymbol{\theta}_i, \boldsymbol{J}_i)$	$\hat{\boldsymbol{\mu}}_i^{(\text{hist})} \mid \boldsymbol{\mu}_i^{(\text{hist})} \sim MVN\left(\boldsymbol{\mu}_i^{(\text{hist})}, \boldsymbol{L}_i^{(\text{hist})}\right)$ $\hat{\boldsymbol{\mu}}_i^{(\text{fut})} \mid \boldsymbol{\mu}_i^{(\text{fut})} \sim MVN\left(\boldsymbol{\mu}_i^{(\text{fut})}, \boldsymbol{L}_i^{(\text{fut})}\right)$ $1/\hat{\psi}_i^{(\text{hist})} \mid 1/\psi_i^{(\text{hist})} \sim Gamma\left(\frac{T_i^{(\text{hist})}-2}{2}, \frac{\left(T_i^{(\text{hist})}-2\right)\psi_i^{(\text{hist})}}{2}\right)$ $1/\hat{\psi}_i^{(\text{fut})} \mid 1/\psi_i^{(\text{fut})} \sim Gamma\left(\frac{T_i^{(\text{fut})}-2}{2}, \frac{\left(T_i^{(\text{fut})}-2\right)\psi_i^{(\text{fut})}}{2}\right)$
Simulator descriptors	$\boldsymbol{\delta}_i \mid \boldsymbol{\omega} \sim MVN(\boldsymbol{\omega}, \boldsymbol{C})$	$\boldsymbol{\delta}_i \mid \boldsymbol{\omega} \sim MVN(\boldsymbol{\omega}, \boldsymbol{C})$	$\boldsymbol{\mu}_i^{(\text{hist})} \mid \boldsymbol{\omega}_{\boldsymbol{\mu}(\text{hist})} \sim MVN\left(\boldsymbol{\mu}_0^{(\text{hist})} + \boldsymbol{\omega}_{\boldsymbol{\mu}(\text{hist})}, \boldsymbol{\phi}_{\mu(\text{hist})}\right)$ $\boldsymbol{\mu}_i^{(\text{fut})} - \boldsymbol{\mu}_i^{(\text{hist})} \mid \boldsymbol{\omega}_{\boldsymbol{\Delta\mu}} \sim MVN\left(\boldsymbol{\mu}_0^{(\text{fut})} - \boldsymbol{\mu}_0^{(\text{hist})} + \boldsymbol{\omega}_{\boldsymbol{\Delta\mu}}, \boldsymbol{\phi}_{\boldsymbol{\Delta\mu}}\right)$ $\psi_i^{(\text{hist})} \mid \omega_{\psi(\text{hist})} \sim Gamma\left(v_3, \frac{v_3}{\psi_0^{(\text{hist})}\omega_{\psi(\text{hist})}}\right)$ $\psi_i^{(\text{fut})}/\psi_i^{(\text{hist})} \vert \omega_{\psi(\text{fut})/\psi(\text{hist})} \sim Gamma\left(v_3, \frac{v_3\psi_0^{(\text{hist})}}{\psi_0^{(\text{fut})}\omega_{\psi(\text{fut})/\psi(\text{hist})}}\right)$
Priors for consensus	–	$\boldsymbol{C}^{-1} \sim Wishart(\boldsymbol{R}_1, v_1)$	$\boldsymbol{\phi}_{\mu(\text{hist})}^{-1} \sim Wishart\left(\boldsymbol{R}_3, s_1\right) \quad \boldsymbol{\phi}_{\boldsymbol{\Delta\mu}}^{-1} \sim Wishart\left(\boldsymbol{R}_4, s_2\right)$ $v_3 \sim Gamma\left(\rho_1, \frac{\rho_1}{\eta_1}\right) \quad v_4 \sim Gamma\left(\rho_2, \frac{\rho_2}{\eta_2}\right)$
Discrepancy	$\boldsymbol{\omega} \sim MVN(\boldsymbol{0}, \boldsymbol{\Lambda})$	$\boldsymbol{\omega} \sim MVN(\boldsymbol{0}, \boldsymbol{\Lambda})$	$\boldsymbol{\omega}_{\boldsymbol{\mu}(\text{hist})} \sim MVN\left(\boldsymbol{0}, \boldsymbol{\gamma}_1\right) \quad \boldsymbol{\omega}_{\boldsymbol{\Delta\mu}} \sim MVN\left(\boldsymbol{0}, \boldsymbol{\gamma}_2\right)$ $\omega_{\psi(\text{hist})} \sim Gamma\left(\gamma_3, \gamma_3\right) \quad \omega_{\psi(\text{fut})/\psi(\text{hist})} \sim Gamma\left(\gamma_4, \gamma_4\right)$
Priors for discrepancy	–	$\boldsymbol{\Lambda}^{-1} \sim Wishart(\boldsymbol{R}_2, v_2)$	$\boldsymbol{\gamma}_1^{-1} \sim Wishart\left(\boldsymbol{R}_5, s_5\right) \quad \boldsymbol{\gamma}_2^{-1} \sim Wishart\left(\boldsymbol{R}_6, s_6\right)$ $\gamma_3 \sim Gamma\left(\rho_3, \frac{\rho_3}{\eta_3}\right) \quad \gamma_4 \sim Gamma\left(\rho_4, \frac{\rho_4}{\eta_4}\right)$
True-climate descriptor	$\boldsymbol{\theta}_0 \sim MVN(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$	$\boldsymbol{\theta}_0 \sim MVN(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$	$\boldsymbol{\mu}_0^{(\text{hist})} \sim MVN\left(\boldsymbol{\lambda}_1, \boldsymbol{\kappa}_1\right) \quad \boldsymbol{\mu}_0^{(\text{fut})} - \boldsymbol{\mu}_0^{(\text{hist})} \sim MVN\left(\boldsymbol{\lambda}_2, \boldsymbol{\kappa}_2\right)$ $\psi_0^{(\text{hist})} \sim Gamma\left(v_5, \frac{v_5}{w_1}\right) \quad \psi_0^{(\text{fut})}/\psi_0^{(\text{hist})} \sim Gamma\left(v_6, \frac{v_6}{w_2}\right)$

Table 3.2: Summary of the different distributional assumptions and hierarchical levels in the three implementations.

3.6 Results

3.6.1 Software Implementation

The RPM implementation is performed using the statistical software R (R Core Team, 2012). The calculation time for $\boldsymbol{\tau}$ ((2.11, p.28)) and $\boldsymbol{S^{-1}}$ ((2.10), p.28) is less than a second. This shows that the compromise of making assumptions which make the RPM approach simplified and easy to implement, does indeed lead to a computationally efficient, analytical calculation of the posterior parameters for the true-climate descriptor/s.

MCMC for both the fully Bayesian implementations is implemented using the R2OpenBUGS software (Lunn et al., 2009). BUGS provides inference based on the Gibbs sampler algorithm, which in this study gives samples approximately from the posterior of $\boldsymbol{\theta}_0$, by iteratively sampling directly from the full-conditional distributions of the random variables in the framework. Four chains with different initial values are run for each implementation. The initial values are chosen from a diverse range, in order to ensure that convergence to the posterior distribution is attained. This is achieved either by varying the initial values manually or using random generators. Note however, that they are chosen to be within a plausible range from the expected values of the random variables involved (i.e values close to the boundary of the support of the random variable are not used). This is done in order to avoid starting points with very low posterior probabilities, since this will not allow the MCMC algorithm to move away from those points. 500 000 iterations are performed in order to reach a reasonable degree of convergence, from which the first 200 000 iterations are discarded (burn-in period). Whether the chains have converged for each iteration or not can be checked using various criteria. For both implementations, the deviance convergence among the chains, as well as the "potential scale reduction factor " (Gelman and Rubin, 1992) are used to assess convergence. Trace plots of the 4 chains for every descriptor of the true climate are produced, in order to also assess convergence visually. Some MCMC diagnostics are presented in Tables F.1-F.2 of Appendix F, for the GFB and FB implementations respectively. The calculation time for the derived posteriors in the fully Bayesian implementations is significantly higher compared to the RPM implementation. It takes around 50 minutes for the GFB implementation and 30 minutes for the FB implementation. All calculations were done on a PC with CPU Intel $i7$ at $3.4GHz$ and $8GB$ of RAM, running with Windows 7, with R version 3.2.0 and OpenBUGS version 3.2.1.

The next section presents the derived posterior means and standard deviations of the true-climate descriptors, together with plots of the posterior distributions as well

as the predictive distributions of mean global surface air temperature (Y_t), for the historical and future periods under study.

3.6.2 Analysis of results

The results presented in this section constitute the output from implementing the framework of Chandler (2013) using real data for the first time. The posterior parameters of the true-climate descriptor/s under the three implementations are summarised in Table 3.3. The top block of Table 3.3 shows the true-climate descriptor estimates $\left(\hat{\boldsymbol{\theta}}_0\right)$ obtained by fitting the mimic to historical observations and estimates of shared simulator discrepancy ($\hat{\boldsymbol{\omega}}$) from reality $\left(\hat{\boldsymbol{\theta}}_0\right)$, for the historical descriptor components. They are calculated by subtracting the estimates of true-climate descriptors from the estimates of simulator consensus $\left(\bar{\hat{\boldsymbol{\theta}}}^{(hist)} - \hat{\boldsymbol{\theta}}_0^{(hist)}\right)$. The bottom block shows the posterior means and standard deviations (in parentheses) of the true climate descriptor $\boldsymbol{\theta}_0$, as derived from the three proposed implementations (RPM, GFB and FB). For comparison, the posterior parameters obtained from the PM implementation with $K = 0, 0.2$ and 1 are also provided. The choices of K are set to be the same as those in the artificial data example in Chandler (2013). According to the author, $K = 0$ represents the judgement that historical and future discrepancies are identical, $K = 0.2$ suggests that there is a slight increase in future relative to historical discrepancy, whereas $K = 1$ represents "scepticism" as to whether historical GCM outputs are "informative about future climate". Additionally, results from the naive approach which estimates the true climate descriptor as the ensemble mean ($\bar{\hat{\boldsymbol{\theta}}}$) are also provided for comparison with the proposed implementations. The uncertainty about the ensemble-mean estimate for each component of $\boldsymbol{\theta}$ is calculated using the standard result for the ensemble mean standard error (i.e. $s/\sqrt{n}$, where s is the standard deviation of the GCM descriptor estimates and $n = 32$ is the ensemble size). The "naive ensemble mean" approach is abbreviated as NEM in Table 3.3.

For the historical period, the derived posterior means of $\alpha_0^{(\text{hist})}$, $\beta_0^{(\text{hist})}$ and $\log\left(\sigma_0^{2\ (\text{hist})}\right)$ are very similar for the proposed RPM, GFB and FB implementations. For the descriptors $\alpha_0^{(\text{hist})}$ and $\beta_0^{(\text{hist})}$ they are identical up to 2 decimal places.

It is also of interest to observe that, for the historical period, all three proposed implementations yield posterior means that are closer to the true-climate descriptor estimates ($\hat{\boldsymbol{\theta}}_0$) than to the simulator consensus ($\bar{\hat{\boldsymbol{\theta}}}$). This is unsurprising, since it is expected that climate observations (if available), are informative about the true climate. The fact that the derived posterior means from all the implementations do not deviate a lot from the observations, reveals that the framework in general yields satisfactory approximations to the climate parameters of interest.

	Historical			**Change**		
	$\alpha_0^{(\text{hist})}$	$\beta_0^{(\text{hist})}$	$\log\left(\sigma_0^{2\,(\text{hist})}\right)$	$\alpha_0^{(\text{fut})} - \alpha_0^{(\text{hist})}$	$\beta_0^{(\text{fut})} - \beta_0^{(\text{hist})}$	$\log\left(\sigma_0^{2\,(\text{fut})}/\sigma_0^{2\,(\text{hist})}\right)$
Estimates						
$\hat{\boldsymbol{\theta}}_0$	14.25	0.02	-4.77	-	-	-
$\hat{\boldsymbol{\omega}}^{(hist)}$	-0.14	0.004	0.79	-	-	-
Inference for $\boldsymbol{\theta}_0$						
NEM	14.10 (0.082)	0.03 (0.001)	-3.99 (0.094)	0.82 (0.032)	0.01(0.002)	-0.54 (0.092)
PM(K=0)	14.24 (0.020)	0.02 (0.002)	-4.54 (0.205)	0.92 (0.050)	0.00 (0.002)	-1.05 (0.231)
PM (K=0.2)	14.24 (0.020)	0.02 (0.002)	-4.57 (0.209)	0.88 (0.068)	0.00 (0.003)	-0.88 (0.334)
PM (K=1)	14.24 (0.020)	0.02 (0.002)	-4.60 (0.213)	0.84 (0.087)	0.01 (0.003)	-0.65 (0.440)
RPM	14.24 (0.020)	0.02 (0.003)	-4.30 (0.186)	0.69 (0.108)	0.01 (0.009)	-0.47 (0.352)
GFB	14.24 (0.025)	0.02 (0.004)	-4.44 (0.241)	0.64 (0.175)	0.01 (0.016)	-0.41(0.570)
FB	14.24 (0.027)	0.02 (0.004)	-4.21 (0.220)	0.78 (0.165)	-0.001(0.013)	-0.73 (0.570)

Table 3.3: Analysis of yearly mean global surface air temperature data from HadCRUT3 observations and CMIP5 simulator outputs. Top block: parameter estimates of the mimic fitted to observations $\left(\hat{\boldsymbol{\theta}}_0\right)$ and of historical shared simulator discrepancy $\left(\hat{\boldsymbol{\omega}}^{(hist)}\right)$ from reality. Bottom block: Estimate of $\boldsymbol{\theta}_0$ and the associated uncertainty (in parentheses) based on the NEM approach and posterior means and standard deviations (in parentheses) of $\boldsymbol{\theta}_0$ derived from the PM with K=0, 0.2 and 1, RPM, GFB and FB implementations.

As far as the historical posterior standard deviations are concerned, it can be clearly seen that for the descriptors $\alpha_0^{(\text{hist})}$ and $\beta_0^{(\text{hist})}$, they are higher for the fully Bayesian implementations than for the RPM implementation. This is expected, since the latter ignores part of the uncertainty (by assigning estimates instead of distributions to the unknown distributional parameters), thus underestimating the overall uncertainty in the posterior. A comparison of the posterior standard deviations of both descriptors between the two fully Bayesian implementations yields the lowest posterior standard deviations in the Gaussian implementation. In general, it could be argued that the results are similar for all the three implementations, for both descriptors. This is not unreasonable, considering the same distributional assumptions (multivariate Gaussian distribution) and the same judgements for values of distributional parameters are assigned to them in all the implementations. For $\log\left(\sigma_0^{2\ (\text{hist})}\right)$ however, more deviations in the posterior standard deviations are observed between the three implementations, showing that the residual variance might be more sensitive to the choice of estimate/distributional assumptions in the three implementations, compared to the other descriptors.

For the true-climate descriptors corresponding to the change between historical and future period, the posterior means deviate more in the three implementations. In general, they tend to have more similarities in the RPM and GFB implementations, compared to the third implementation. This is expected, since the same distributional assumptions are assigned to the descriptors, in both implementations. For $\alpha_0^{(\text{fut})} - \alpha_0^{(\text{hist})}$, the third implementation yields higher posterior mean compared to the other two implementations. In contrast, the posterior means corresponding to $\beta_0^{(\text{fut})} - \beta_0^{(\text{hist})}$ and $\log\left(\sigma_0^{2\ (\text{fut})}/\sigma_0^{2\ (\text{hist})}\right)$ are smaller in the FB implementation than the posterior means obtained from the other two implementations.

The posterior standard deviations of the true-climate descriptors corresponding to the historical-to-future change are higher than the historical. This is explained by the absence of observations, leading to increased uncertainty in the future period. The lower posterior standard deviations in the RPM implementation compared to the fully Bayesian implementations, are indicative of the underestimation of uncertainty in the former. In contrast to the historical period, the posterior standard deviations of $\alpha_0^{(\text{fut})} - \alpha_0^{(\text{hist})}$ and $\beta_0^{(\text{fut})} - \beta_0^{(\text{hist})}$ under the RPM implementation are closer to the FB implementation. Additionally, a comparison between the two fully Bayesian implementations reveals that for both descriptors, the lowest posterior standard deviations are achieved under the FB implementation. All the above suggest potential sensitivity of the relevant posterior standard deviations to the choice of distributional assumptions assigned to the descriptors.

It is also of interest to observe what the results suggest in terms of the physical

interpretation of the true-climate descriptors. Firstly, since all the derived posterior means of $\alpha_0^{(\text{fut})} - \alpha_0^{(\text{hist})}$ are positive, the three implementations suggest an increase in the mean global surface air temperature in the future period 2016-2035, compared to the historical period 1986-2005. Moreover, the positive values of the posterior means of $\beta_0^{(\text{fut})} - \beta_0^{(\text{hist})}$ suggest that the rate of increase of mean global surface air temperature (per time unit) will increase in the future period, relative to the historical period. As far as the residual standard deviation (σ) is concerned, according to the implementations it is expected to decrease. This, in practice, means that less inter-annual temperature variability is expected in the future period considered, compared to the historical period. However, in order to make clear statements about the true-climate descriptors based on the implementations, it is important to consider the derived posterior means in conjunction with the corresponding posterior standard deviations.

A comparison of the results from the three proposed implementations compared to the simpler NEM and PM implementations suggests the following: The estimates of $\boldsymbol{\theta}_0$ in the NEM approach differ from the posterior means in the remaining implementations. The uncertainty is in general estimated to be smaller for most of the descriptor components. This is attributed to the fact that the NEM approach ignores the uncertainties in the MME structure, which are accounted for in the Bayesian framework followed by the remaining implementations. The posterior means from the PM implementations are similar with these from the RPM implementation for the historical components of $\boldsymbol{\theta}_0$, for all choices of K. Similarly for the posterior standard deviations. For the components $\alpha_0^{(\text{fut})} - \alpha_0^{(\text{hist})}$ and $\log\left(\sigma_0^{2\,(\text{fut})}/\sigma_0^{2\,(\text{hist})}\right)$ however, the posterior means from the PM implementations deviate from those in the RPM implementation. Results from the RPM approach are agree more with those from the PM with $K = 1$, relative to the other choices of K. This implies that earlier data used to estimate $\boldsymbol{\Lambda}$ in the RPM implementation suggest that historical discrepancy is not very informative about future discrepancy. Similarly for the posterior standard deviations, which are in general larger for PM with $K = 1$ and RPM, compared to PM with $K = 0$ and $K = 0.2$. This is expected, since the more future discrepancy is assumed to differ from historical, the more uncertainty is expected in the posteriors of true climate, since $\boldsymbol{\Lambda}$ is a substantial source of uncertainty in the framework.

In Figure 3.3, the posterior distributions under the three implementations are plotted, for each true-climate descriptor, in order to graphically illustrate the differences in the posteriors of interest between the three implementations.

The posterior for each descriptor under the RPM implementation is obtained as the corresponding marginal distribution of the joint posterior defined in (2.9)-(2.11). Plots of the posterior distributions derived from the two fully Bayesian implemen-

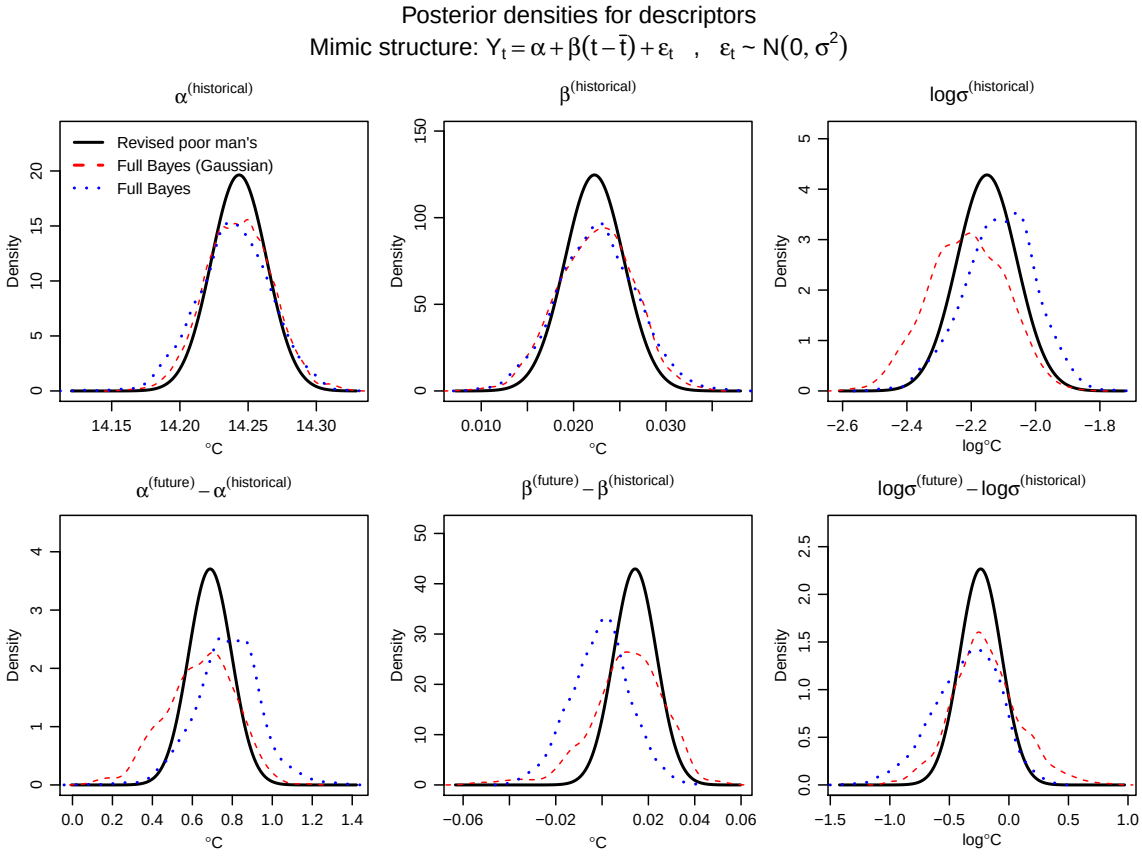


Figure 3.3: Posterior densities for true-climate descriptors, for the three implementations. Top row: historical descriptors. Bottom row: change between historical and future descriptors.

tations are obtained by plotting the density of 1000 descriptor values obtained from the MCMC simulations. The 1000 sampled descriptors in each case are obtained by sampling 1 descriptor in every 300 from the posterior. This is done to ensure that a sample of approximately independent descriptors from the posterior distribution is used for producing each posterior distribution.

It is evident from Figure 3.3 that the posterior distributions are more similar for the descriptors corresponding to the historical period than for those corresponding to the change between historical and future periods, in agreement with the results of Table 3.2. The posteriors seem to achieve the highest degree of similarity for the descriptor $\beta^{(\text{hist})}$. As far as the descriptors corresponding to the change are concerned, larger deviations in the posterior distributions under the three implementations are observed, compared to the historical period. The fact that the posterior distributions under the RPM implementation are narrower is explained by the fact that the RPM implementation ignores part of the underlying uncertainty. This can have interesting implications for policy-making, especially when calculating probabilities for extreme climate events. Consider for example the goal of retaining global warming below $2^{\circ}C$ during the 2^{nd} half of the 21^{st} century relative to pre-industrial conditions, which was set during the 2015 United Nations Climate Change Conference (UN, 2015). Based on the PDFs for the posterior distributions of $\alpha_0^{(\text{fut})} - \alpha_0^{(\text{hist})}$ in Figure 3.3, a fully Bayesian implementation, which has wider tails, would assign much higher probability of global warming exceeding $2^{\circ}C$ relative to the RPM implementation.

However, it can be argued from Figure 3.3 that any deviations of the posterior distribution under the RPM implementation are not serious, compared to those from the fully Bayesian implementations. Furthermore, the posterior distributions under the two fully Bayesian implementations seem to agree more with each other than each of them with the posterior distribution under the RPM implementation. This suggests that any potential sensitivities of the posterior to distributional assumptions might not seriously affect the uncertainty in the derived posterior under the RPM implementation.

Apart from making inference about the true-climate descriptors which originally are not of direct interest, it is of particular interest to observe what the three implementations reveal about the actual climate. For this reason, plots of the predictive distributions of mean global surface air temperature (Y_t) are presented in Figures 3.4a and 3.4b for the two periods respectively.

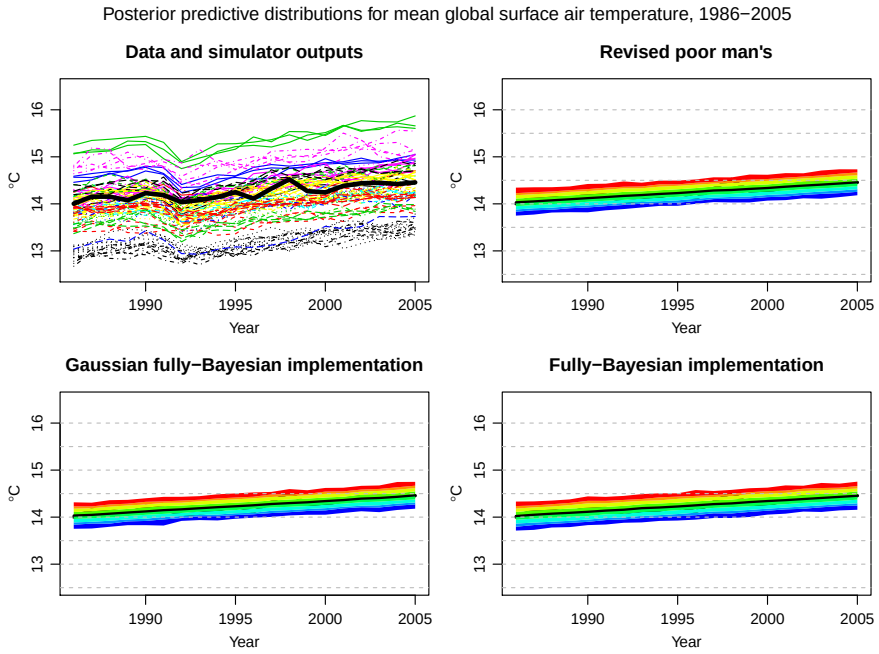
The posterior predictive distribution of Y_t for the RPM implementation, for the historical period, is evaluated by randomly sampling 1000 values of the true-climate descriptors, from the multivariate-Gaussian distribution defined in (2.9)-(2.11). For the fully Bayesian implementations, the descriptors are obtained directly from the

derived posteriors from the MCMC simulations. Then, for each descriptor, the mimic ((3.1), p.40) is used to simulate a 20-year sequence of yearly mean global surface air temperature, corresponding to the historical period $1986 - 2005$. Therefore, 1000 predicted values of Y_t are obtained for every value of t. Figures 3.4a-3.4b show, for each year of the historical and future periods respectively, the percentiles of the resulting simulated (posterior predictive) distribution of global mean temperature. Colours change depending on the choices of percentiles considered for constructing the plots.

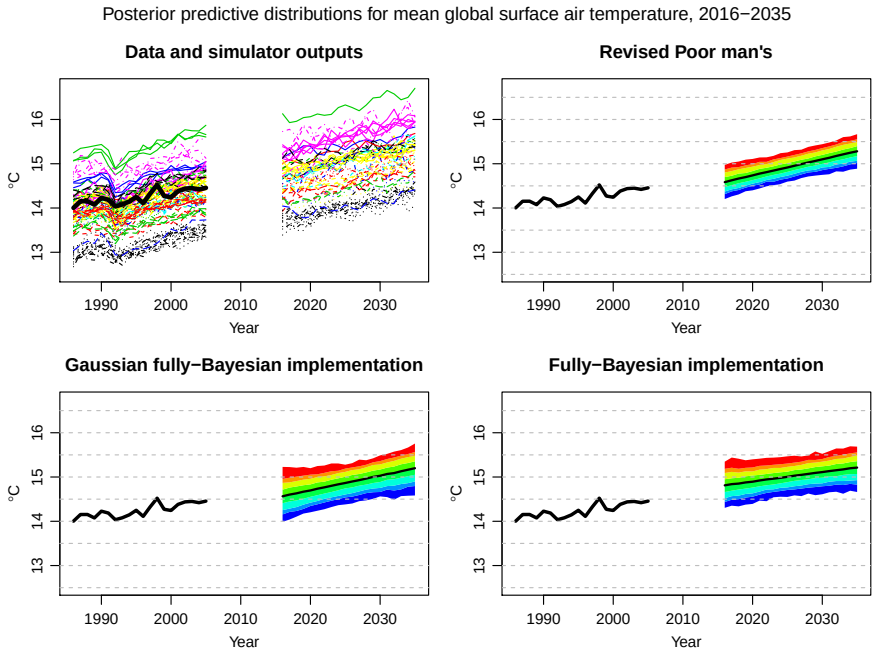
For the historical period, it is evident that the mean global surface air temperature increases approximately linearly, according to the derived posteriors from all the implementations. Any differences in the width of the inter-percentile ranges between the poor man's and the fully Bayesian implementations are small, suggesting that the uncertainty ignored in the RPM implementation is not very serious in the historical period. Additionally, the predictive distributions under the two fully Bayesian implementations are very similar, implying that the predictive distributions are not seriously affected by relaxing the Gaussian assumption for the residual variance in the FB implementation.

The posterior predictive distributions for the future period are evaluated similarly to the ones for the historical period. They also suggest an approximately linear increase in mean global surface air temperature. The slight underestimation of uncertainty under the RPM implementation is evident from the narrower inter-percentile ranges compared to the fully Bayesian implementations. Also, there seem to be larger differences between the three predictive distributions, compared to the historical period, as also suggested from Table 3.3 and Figure 3.3. Although the posterior mean under the FB implementation is higher relative to the corresponding posterior means under the other two implementations, the inter-percentile ranges are very similar under the two fully Bayesian implementations, suggesting that the Gaussian assumptions might not cause a serious effect on the uncertainty estimation. It is also worth noting the increased uncertainty in the edges of the distribution. Although this is a generic property of all linear trend models, for the beginning of the future period, it can be partly since inference for the 20-year periods under study is considered independently from anything prior to it. To overcome this limitation, a more realistic, sophisticated mimic could be deployed, enabling representation of temperature in a continuous time-frame (instead of considering two disjoint periods). This is left as future work; it is further discussed in Section 6.2.

To summarize, it is clear from the results illustrated in Table 3.3 and Figures 3.3-3.4b that the RPM implementation slightly underestimates the underlying uncertainty in the derived posterior of the true-climate descriptors and the predictive



(a) Posterior predictive distributions of yearly mean global surface air temperature ($°C$), for the historical period $1986 - 2005$.



(b) Posterior predictive distributions of yearly mean global surface air temperature ($°C$), for the future period $2016 - 2035$.

Figure 3.4: For each period: Top-left: Plots of observations (black solid line) and simulator outputs (coloured lines) of yearly mean global surface air temperature. Top-right: Predictive distribution of yearly mean global surface air temperature, obtained from the derived posteriors in the RPM implementation. Bottom-left: Predictive distribution under the GFB implementation. Bottom-right: Predictive distribution under the FB implementation. Black line: Posterior mean of yearly mean global surface air temperature. Coloured segments: Partition the predictive distribution based on the $1^{st}, 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}$ and 95^{th} and 99^{th} percentiles, from the bottom to the top.

distributions of global mean temperature. However, this underestimation is barely noticeable in the historical period and not very serious in the historical-to-future change, for this particular application. Results from the historical period reveal more consensus of posteriors between the three implementations, whereas in the absence of observations for the future period, the derived posteriors become more sensitive to the choice of distributional assumptions. The sensitivity is more evident in the posterior means for the change and less obvious in the corresponding posterior standard deviations.

In order to examine whether the prior distributional assumptions seriously affect the resulting uncertainty in the posteriors, a closer look at the results from the two fully Bayesian implementations is required. Regarding sensitivity of the posterior to the Gaussian assumption about residual variance, it can be argued that for the historical period, results do not seem to be very sensitive to the assumption. The differences in the posterior means for the historical-to-future change suggest that in the absence of future observations, results might be more sensitive to the choice of distributional assumptions. It is not very clear however, whether the Gaussian or Gamma distribution for the residual variance is more appropriate in terms of improving approximations of real climate, since the posterior standard deviations corresponding to the change between historical and future periods are very similar. It is worth exploring this further however, in order to investigate to what extent the underestimation of uncertainty in the RPM implementation is caused by the underlying Gaussian assumptions.

3.7 Conclusions

The work presented in this chapter demonstrated three different implementations of the conceptual framework of Chandler (2013), to make inference about yearly mean global surface air temperature, for the periods $1986-2005$ and $2016-2035$. Information from simulator outputs forming an MME of 32 GCMs from the CMIP5 experiment was combined with real-climate observations, to derive the posterior distributions of real yearly global surface air temperature, under the probabilistic, Bayesian framework. Additionally, predictive distributions of yearly mean global surface air temperature were produced for the periods under study, in order to get an indication of the underlying uncertainty in the temperature estimates obtained under the three implementations.

The RPM implementation was computationally simplified, since it assigned estimates to the unknown covariance matrices in the framework (including a proposed

estimator of variability due to shared simulator discrepancy from reality), thus allowing the derivation of an analytical solution to the posterior distribution of real climate. However, it ignored part of the uncertainty, by plugging-in estimates for the unknown covariance matrices. As an alternative, two versions of the fully Bayesian implementations were proposed, under different distributional assumptions for the residual variance σ^2 in the mimic. By assigning distributions instead of estimates to the unknowns, the underlying uncertainties were fully captured in the implemented framework. The fully Bayesian implementations were more computationally-demanding and required the use of numerical techniques (Gibbs sampler), to derive the posterior of interest. Ultimately, results from the fully Bayesian implementations were compared with those from the RPM implementation, in an attempt to investigate whether the more simplified and computationally easier RPM implementation yields adequate approximations to the posteriors of real-climate. It was also of interest to deduce the effect of the Gaussian assumptions for residual variance in the RPM implementation, in the quantification of the underlying uncertainties. This was checked by comparing the results in the two fully Bayesian implementations, since the latter relaxed the Gaussian assumption for the residual variance.

Results revealed that in general, the RPM implementation slightly underestimates the uncertainty in the derived posteriors of true climate, as expected. However, the underestimation did not seem to be very serious, thus providing an important argument in favour of the simplified, computationally-efficient RPM implementation. Regarding sensitivity of the posterior to the Gaussian assumption about residual variance, this was more evident in the change between historical and future periods. However, it was not very clear whether this seriously affected estimation of uncertainty in the RPM implementation. Therefore, it could be argued that the RPM implementation showed a good performance in terms of estimating the underlying uncertainties in the global surface air temperature application considered in this study. However, the Gaussian assumptions about the random quantities involved must be handled with care, depending on the application of interest.

The above conclusions suggest that: given sufficient computational resources, a fully Bayesian implementation is recommended, since it is expected to yield more accurate estimates of the uncertainty in real climate compared to the RPM implementation. Regarding user convenience in setting priors, the GFB implementation is recommended over the FB implementation, since it jointly considers the components of the descriptor vector, thus reducing the number of distributions to be specified. Given sufficient time however, it would be advised to perform both the fully Bayesian implementations, since there might be considerable sensitivity of the real climate posterior to the prior choices. If there is need for a computationally efficient imple-

mentation, the simpler RPM implementation is to be preferred, which is not expected to seriously underestimate uncertainty in the posterior of real climate, according to previous findings.

Chapter 4

Framework for simulator grouping

4.1 Motivation

The work presented in Chapter 3 was based on the conceptual framework of Chandler (2013) for combining information from multiple climate simulators in a MME, as described in Section 2.6.1. One of the framework assumptions is that the descriptors $\{\boldsymbol{\theta}_i,\ i = 1, \ldots, m\}$ for each ensemble member in a MME, are centred on the simulator consensus $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$, where $\boldsymbol{\theta}_0$ is the reality descriptor and $\boldsymbol{\omega}$ is the shared simulator discrepancy from reality. Conceptually, the framework considers all simulators as being exchangeable, conditional on their consensus. In probabilistic terms, exchangeability implies that the joint distribution of a collection of random variables is invariant under any permutation of the variable indices (Gelman et al., 2014, Section 1.2). Exchangeability of simulator descriptors then suggests that their joint distribution, conditional on their consensus, is invariant under any permutation of the simulator labels. This is imposed by (2.6) of Section 2.6.2, which assumes that the discrepancies $\{\boldsymbol{\delta}_i = \boldsymbol{\theta}_i - \boldsymbol{\theta}_0,\ i = 1, \ldots, m\}$ are i.i.d., conditional on $\boldsymbol{\omega}$. According to Gelman et al. (2014, Section 5.2), exchangeability is a reasonable assertion when there is no reason to believe that any ordering or grouping between simulators in a MME takes place, in which case it is reasonable to assume "symmetry" among them. However, it is well-known in the literature (Chandler, 2013; Masson and Knutti, 2011; Knutti et al., 2013; Abramowitz, 2010; Pirtle et al., 2010) that inter-simulator similarities exist, which generates doubts as to whether assuming simulator exchangeability in the framework is always reasonable (see also Section 2.4).

Chandler (2013) recognizes that it would be unrealistic to make the above assumption in MMEs involving more than one simulator from the same modelling group. The same also holds in cases of perturbed physics ensembles (PPEs) (see Section 2.2), where simulator runs belong to the same GCM, run under different parametrisations.

This is illustrated in Masson and Knutti (2011), who also express scepticism as to whether ensemble members sharing components, code or numerical schemes can be considered as independent in their departures from other models. They cluster the GCMs from the CMIP5 project in a genealogical tree, to express similarities between model outputs for surface temperature and precipitation. "Similarity" in this case is determined using a metric of the distances between GCM monthly outputs, which considers various characteristics of their behaviour, such as seasonality, trend and spatial correlation. Their results reveal that the modelling centre and the atmospheric model component are dominant sources of similarities. Knutti et al. (2013), in a similar attempt to determine similarities between GCMs of the CMIP5 project, note that although GCMs share biases from reality, the reasons for this might be different for different GCMs. Their results confirm that the modelling group and sharing of code are important sources of similarities between GCMs. Pirtle et al. (2010) also consider use of identical datasets for validating the GCMs as another reason of inter-model similarities. According to Abramowitz (2010), the fact that the design of a newer version of a particular GCM is affected by earlier versions of it, is another important source of dependence between GCMs.

It is clear from the above that ignoring similarities between ensemble members is unrealistic in the representation of MME structures. It is also recognized that unrealistic representations of inter-dependence of simulators in MMEs may lead to non-coherent quantification of uncertainty. Masson and Knutti (2011) argue that ignoring similarities of GCMs in frameworks for quantifying uncertainty in MMEs, makes interpretation of results "challenging". Pirtle et al. (2010) also mention that accounting for simulator dependence in MME interpretation will help clarify the "meaning of agreement" between models. In view of these, in this chapter the conceptual framework of Section 2.6.1 is modified, in order to explicitly account for similarities among the ensemble members. An extended framework is proposed (illustrated in Figure 4.1, p.80), to accommodate potential clustering of ensemble members, in an attempt to more realistically represent MME structures. The rest of this chapter is devoted to the conceptual and mathematical formulation of the proposed "extended" framework (Sections 4.2-4.3). Sections 4.4-4.8 deal with the implementation of the extended framework, including a proposed random effects model for estimation of within-group variability (Section 4.6).

4.2 Conceptual framework

According to the discussion in Section 4.1, the proposed framework revisits the exchangeability assumption around the whole collection of ensemble members, conditional on their consensus, aiming for a more realistic representation of the MME structure which explicitly accounts for inter-simulator similarities. Suppose that expert judgement suggests that it is reasonable to collect MME members into groups with similar characteristics. This grouping is explicitly incorporated in the framework, by assigning random descriptors representing each group member, while group descriptors are themselves centred on a random descriptor representing the *group consensus*. Deviation of descriptors from their group consensus is expressed through a covariance matrix, assumed for simplicity to be identical among descriptors of the same group. The covariance matrix is assumed to be random, for reasons explained in Sections 4.3 and 4.6. Further grouping can be potentially applied in the same way within each group, by repeating this grouping process recursively. This results in a nested MME structure, which is incorporated in the framework by adding levels to the hierarchy, corresponding to the different grouping stages.

Figure 4.1 illustrates the idea for a MME with 9 members, grouped according to simulator variants, nested within simulators, nested within modelling centres (referred to as "families"). The indices $\{i,\ i = 1, \ldots, m\}$, $\{j,\ j = 1, \ldots, n_i\}$ and $\{k,\ k = 1, \ldots, n_{ij}\}$ where $m = 4$, $n_i = \{3, 2, 1, 1\}$ and $n_{ij} = \{2, 1, 1, 1, 1, 2, 1\}$ correspond respectively to simulator families, simulators nested within families and simulator variants nested within simulators. The structure is represented in the framework in the form of descriptors at each level, clustered around their consensus descriptor (*group consensus*) in the upper level. Deviation from their consensus is determined by random covariance matrices. Those are $\{\boldsymbol{C}\}, \{\boldsymbol{C}_i\}$ and $\{\boldsymbol{C}_{ij}\}$, for the sets of indices defined above. The available data are expressed through the 9 descriptor estimators $\left\{\hat{\boldsymbol{\theta}}_{ijk}\right\}$ from the simulators, as well as the estimator $\left\{\hat{\boldsymbol{\theta}}_0\right\}$ from real-climate observations. Each descriptor estimator is centred on the corresponding descriptor $\boldsymbol{\theta}_{ijk}$ (or $\boldsymbol{\theta}_0$ for observations), with deviations determined by the corresponding covariance matrix $\boldsymbol{J}_{ijk}$ (or $\boldsymbol{J}_0$ for observations). The descriptors at the top level of grouping, i.e. at the *family level* in Figure 4.1, are centred on the *overall family consensus* $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is defined to be the shared discrepancy of family descriptors from reality. This contrasts with the simpler framework of Section 2.6, in which $\boldsymbol{\omega}$ was defined to be the shared simulator discrepancy from reality.

Figure 4.1 can be interpreted as a nested multilevel tree with "nodes" and "branches", the latter connecting "parent" nodes to their corresponding "child" nodes. Each parent node collects two random quantities: First, the random descriptor acting as the

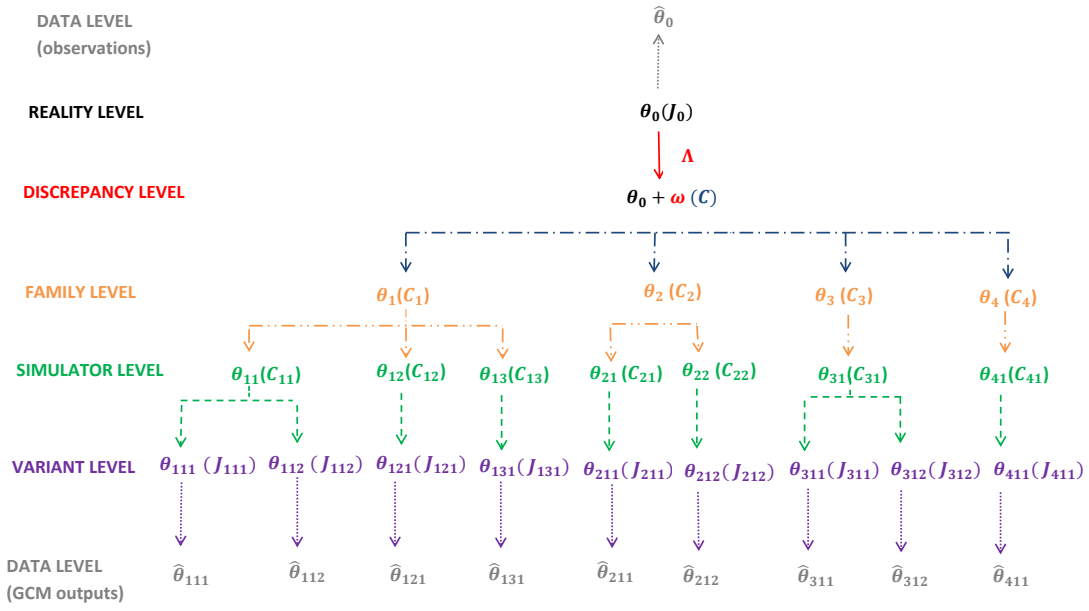


Figure 4.1: Geometrical representation of the proposed framework, for a MME with 9 members, grouped according to simulator variants, nested within simulators, nested within families. $\boldsymbol{\theta}_0$: True-climate descriptor; For $\{i, j, k,\ i = 1, \ldots, 4, j = 1, \ldots, n_i, k = 1, \ldots, n_{ij}\}$, where $n_i = \{3, 2, 1, 1\}$ and $n_{ij} = \{2, 1, 1, 1, 1, 2, 1\}$, $\boldsymbol{\theta}_i$: Family descriptor of family i; $\boldsymbol{\theta}_{ij}$: Simulator descriptor j of family i; $\boldsymbol{\theta}_{ijk}$: Variant descriptor k of simulator j of family i; $\hat{\theta}_{ijk}$: Estimator of $\boldsymbol{\theta}_{ijk}$ obtained from simulator output; $\hat{\boldsymbol{\theta}}_0$: Estimator of $\boldsymbol{\theta}_0$, obtained from observations; $\boldsymbol{\omega}$: Shared discrepancy of family descriptors from reality, determined by $\boldsymbol{\Lambda}$; $(\cdots)$: error in estimating $\boldsymbol{\theta}_0$ by $\hat{\theta}_0$, determined by $\boldsymbol{J}_0$; $(\cdots)$: error in estimating $\boldsymbol{\theta}_{ijk}$ by $\hat{\theta}_{ijk}$, determined by $\boldsymbol{J}_{ijk}$; $(-)$: shared simulator bias; $(- \cdot -\cdot)$: Deviation of family descriptor $\boldsymbol{\theta}_i$ from the overall simulator consensus $\boldsymbol{\theta}_0 + \omega$, determined by $\boldsymbol{C}$; $(- \cdot \cdot -)$: Deviation of simulator descriptor $\boldsymbol{\theta}_{ij}$ from family descriptor $\boldsymbol{\theta}_i$, determined by $\boldsymbol{C}_i$; $(- - -)$: Deviation of variant descriptor $\boldsymbol{\theta}_{ijk}$ from simulator descriptor $\boldsymbol{\theta}_{ij}$, determined by $\boldsymbol{C}_{ij}$; Arrows indicate direction of causal relationships in which an intervention at the "parent" node is expected to produce a change at the "child" node.

group consensus of its child descriptors and secondly, the random covariance matrix expressing deviation of child descriptors from their parent descriptor. Additionally, nodes sharing the same parent node, are defined as "sibling" nodes. The framework assumes that the child nodes are exchangeable, conditional on their parent. This assumption is well-justified under the proposed framework, since any identified differences (according to expert judgement) between nodes are accounted for by applying nested grouping to them. Consequently, there is no obvious reason to distinguish between sibling nodes. To summarize, the proposed framework extends the framework of Section 2.6, by explicitly accounting for similarities of ensemble members, through justifiable exchangeability assumptions around groups of descriptors at each level of the hierarchy. It therefore provides a more realistic representation of MME structures, particularly in the presence of complex inter-simulator dependencies.

In order to make inference about the true climate, the posterior for the descriptor $\boldsymbol{\theta}_0$ under the proposed framework must be derived, conditional on the available data. The representation of the MME structure under the framework allows treating Figure 4.1 as a directed acyclic graph (DAG), similarly to the simpler framework. The useful property of DAGs for the derivation of the posterior is that "sibling" nodes which are otherwise unconnected, are assumed to be independent, conditional on their parent nodes. For an arbitrary-level MME structure, denote the descriptor estimators from GCM outputs by $\left\{\hat{\boldsymbol{\theta}}^{(1)} \dots \hat{\boldsymbol{\theta}}^{(N)}\right\}$, where N denotes the number of ensemble members, or equivalently the number of GCM outputs in the MME considered as distinct data sources. For example, in the MME structure illustrated in Figure 4.1, $N = \sum_{i=1}^{m} \sum_{j=1}^{n_i} n_{ij} = 9$. The derivation of the generic posterior summarized for the simpler framework (Chandler, 2013, Supplementary material S1), shows that the form of the posterior remains unchanged under the proposed framework. With the suggested notation, this becomes,

$$\pi\left(\boldsymbol{\theta}_0|\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}^{(1)}, \dots, \hat{\boldsymbol{\theta}}^{(N)}\right) \propto \pi(\boldsymbol{\theta}_0)\pi(\hat{\boldsymbol{\theta}}_0|\boldsymbol{\theta}_0)\pi\left(\hat{\boldsymbol{\theta}}^{(1)}, \dots, \hat{\boldsymbol{\theta}}^{(N)}|\boldsymbol{\theta}_0\right). \tag{4.1}$$

Implementation of the framework requires assigning distributional assumptions for the random quantities involved in all the levels of the hierarchical framework. They are presented in the next section.

4.3 Mathematical analysis of the Gaussian specification

Each node in the framework consists of a descriptor which acts as a parent descriptor, and a covariance matrix expressing the deviation of the sibling descriptors from their parent. Motivated by de Finetti's theorem (Hoff, 2009, Section 2.8), the exchangeable sibling descriptors (based on the exchangeability assumptions introduced in Section 4.2) are modelled as i.i.d., conditional on their parent descriptor. The conditional distributions are assumed to be multivariate normal, similarly to the Gaussian specification in Section 2.6.2. For simplicity, this section presents the distributional assumptions underlying the 6 levels of the MME structure illustrated in Figure 4.1 of Section 4.2. However, equivalent assumptions can be made for structures with additional levels.

Equations (4.2)-(4.7) below summarize the underlying distributional assumptions of the extended framework.

- Data level

  Similarly to the simpler framework, the MLEs $\{\hat{\boldsymbol{\theta}}_{ijk},\ i = 1, \ldots, m, j = 1, \ldots, n_i, k = 1, \ldots, n_{ij}\}$ are assumed to be centred on the corresponding variant descriptors $\{\boldsymbol{\theta}_{ijk},\ i = 1, \ldots, m, j = 1, \ldots, n_i, k = 1, \ldots, n_{ij}\}$, as follows:

  $$\hat{\boldsymbol{\theta}}_{ijk} | \boldsymbol{\theta}_{ijk}, \boldsymbol{J}_{ijk} \sim MVN\left(\boldsymbol{\theta}_{ijk}, \boldsymbol{J}_{ijk}\right)\ (i = 1, \ldots, m, j = 1, \ldots, n_i, k = 1, \ldots, n_{ij}). \tag{4.2}$$

  The covariance matrices $\{\boldsymbol{J}_{ijk},\ i = 1, \ldots, m, j = 1, \ldots, n_i, k = 1, \ldots, n_{ij}\}$ represent internal variability in data source i. Additionally, the descriptor estimators are assumed to be independent, conditional on the corresponding variant descriptors.

  Similarly, the true-climate descriptor estimator $\hat{\boldsymbol{\theta}}_0$ is assumed to be centred on the true-climate descriptor $\boldsymbol{\theta}_0$, as in (2.4) with $i = 0$.

- Variant level

  The variant descriptors $\{\boldsymbol{\theta}_{ijk},\ i = 1, \ldots, m, j = 1, \ldots, n_i, k = 1, \ldots, n_{ij}\}$ are distributed as:

$$\boldsymbol{\theta}_{ijk}|\boldsymbol{\theta}_{ij}, \boldsymbol{C}_{ij} \sim MVN\left(\boldsymbol{\theta}_{ij}, \boldsymbol{C}_{ij}\right) \ (i = 1, \ldots, m, j = 1, \ldots, n_i, k = 1, \ldots, n_{ij}). \tag{4.3}$$

- Simulator level

  The distribution of simulator descriptors $\{\boldsymbol{\theta}_{ij},\ i = 1, \ldots, m, j = 1, \ldots, n_i\}$ is:

$$\boldsymbol{\theta}_{ij}|\boldsymbol{\theta}_i, \boldsymbol{C}_i \sim MVN\left(\boldsymbol{\theta}_i, \boldsymbol{C}_i\right) \ (i = 1, \ldots, m, j = 1, \ldots, n_i). \tag{4.4}$$

- Family level

  The family descriptors $\{\boldsymbol{\theta}_i,\ i = 1, \ldots, m\}$ are distributed as follows:

$$\boldsymbol{\theta}_i|\boldsymbol{\omega}, \boldsymbol{C} \sim MVN\left(\boldsymbol{\theta}_0 + \boldsymbol{\omega}, \boldsymbol{C}\right) \ (i = 1, \ldots, m). \tag{4.5}$$

- Discrepancy level

  Similarly to the simpler framework, the shared family discrepancy $\boldsymbol{\omega}$ from reality is regarded as random, distributed as in (2.7).

$$\boldsymbol{\omega} \sim MVN(\mathbf{0}, \boldsymbol{\Lambda}), \tag{4.6}$$

  where $\boldsymbol{\Lambda}$ now represents the propensity of family descriptors collectively to deviate from reality.

  It remains to specify the prior distribution of the true-climate descriptor $\boldsymbol{\theta}_0$.

- Reality level

  As in the simpler framework, the real-climate descriptor is assumed to be multivariate normally distributed:

$$\boldsymbol{\theta}_0 \sim MVN(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0). \tag{4.7}$$

  The prior parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ represent the *a priori* beliefs about real climate.

It is important to note that the attempt to realistically represent the MME structure can often lead to data sparseness (see for example Masson and Knutti (2011, Figure 1) and Knutti et al. (2013, Figure 1), where simulator grouping based on expert judgement yields singleton groups). Note that throughout this chapter, data sparseness refers to the situation of having many groups of small sample size and

in some cases, even singleton groups. This generates the need to share information between siblings, to improve estimation. To achieve this, the covariance matrices $\{\boldsymbol{C}_{ij}, \boldsymbol{C}_i,\ i = 1, \ldots, m, j = 1, \ldots, n_i\}$ and $\boldsymbol{C}$ in (4.3)-(4.5) respectively, expressing deviation of the sibling descriptors from their parent descriptors, are assumed to be random in the proposed framework, as also discussed in Section 4.2. Details about the distribution of the covariance matrices are given in Section 4.6.

Under the multivariate normal assumptions of (4.2)-(4.7), the expressions $\pi(\boldsymbol{\theta}_0)$ and $\pi(\hat{\boldsymbol{\theta}}_0|\boldsymbol{\theta}_0)$ in the generic posterior defined in (4.1) are,

$$\exp\left[-\frac{1}{2}\left(\boldsymbol{\theta}_0 - \boldsymbol{\mu}_0\right)' \boldsymbol{\Sigma}_0^{-1} \left(\boldsymbol{\theta}_0 - \boldsymbol{\mu}_0\right)\right], \tag{4.8}$$

and

$$\exp\left[-\frac{1}{2}\left(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0\right)' \boldsymbol{J}_0^{-1} \left(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0\right)\right], \tag{4.9}$$

respectively, as in the simpler framework (Chandler, 2013, suppl. material S1).

The remaining term $\pi\left(\hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(N)}|\boldsymbol{\theta}_0\right)$ in (4.1) is multivariate normal (MVN), since the estimators are themselves assumed to be MVN (see (4.2)). In order to determine the mean and covariance matrix of $\pi\left(\hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(N)}|\boldsymbol{\theta}_0\right)$, each of the descriptor estimators $\left\{\hat{\boldsymbol{\theta}}^{(l)},\ l = 1, \ldots, N\right\}$ is expressed as a sum of deviations of child descriptors from their parents in moving from $\hat{\boldsymbol{\theta}}^{(i)}$ in the data level to $\boldsymbol{\theta}_0$ in the reality level. For the example of Figure 4.1, the estimators $\left\{\hat{\boldsymbol{\theta}}^{(l)},\ l = 1, \ldots, N\right\}$ correspond to $\left\{\hat{\boldsymbol{\theta}}_{ijk},\ i = 1, \ldots, m, j = 1, \ldots, n_i, k = 1, \ldots, n_{ij}\right\}$ with m, n_i and n_{ij} defined in Section 4.2. The latter are then expressed as,

$$\hat{\boldsymbol{\theta}}_{ijk} = \boldsymbol{\theta}_0 + (\boldsymbol{\theta}_i - \boldsymbol{\theta}_0) + (\boldsymbol{\theta}_{ij} - \boldsymbol{\theta}_i) + (\boldsymbol{\theta}_{ijk} - \boldsymbol{\theta}_{ij}) + (\hat{\boldsymbol{\theta}}_{ijk} - \boldsymbol{\theta}_{ijk}), \tag{4.10}$$

for $i = 1, \ldots, m,\ j = 1, \ldots, n_i$ and $k = 1, \ldots, n_{ij}$.

It follows immediately from the distributional assumptions of the framework's random variables discussed earlier that all the discrepancies in (4.10) have zero expectation conditional on $\boldsymbol{\theta_0}$, which leads to $E\left(\hat{\boldsymbol{\theta}}_{ijk}|\boldsymbol{\theta}_0\right) = \boldsymbol{\theta}_0$. It can be easily deduced that this result can be generalised for a MME structure of arbitrary number of levels. For the vector $\hat{\boldsymbol{\Theta}} = \left(\hat{\boldsymbol{\theta}}^{(1)'} \ldots \hat{\boldsymbol{\theta}}^{(N)'}\right)'$ concatenating the simulator descriptor estimators from a MME, it holds that,

$$E\left(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta}_0\right) = \boldsymbol{\Theta}_0, \tag{4.11}$$

where $\boldsymbol{\Theta}_0 = (\boldsymbol{\theta}_0' \cdots \boldsymbol{\theta}_0')'$.

On the other hand, derivation of the covariance matrix $\pi(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta}_0)$ requires evaluating $Cov\left(\hat{\boldsymbol{\theta}}^{(i)}, \hat{\boldsymbol{\theta}}^{(j)}|\boldsymbol{\theta}_0\right)$ for $\{i, j = 1, \ldots, N\}$. Under the framework of Section 2.6, these calculations result in a block matrix $\boldsymbol{K}$ with identical off-diagonal blocks, as shown in Chandler (2013, Supplementary material S2). Under the extended framework however, the presence of nested grouping of descriptors suggests that the form of $\boldsymbol{K}$ is non-trivial and becomes structurally more complicated as more levels are added to the hierarchy.

Denote by $\tilde{\boldsymbol{\Lambda}}_{x,y}$ the $(xp) \times (yp)$ block matrix, where each block consists of the $p \times p$ matrix $\boldsymbol{\Lambda}$:

$$\tilde{\boldsymbol{\Lambda}}_{x,y} = \underbrace{\left(\begin{array}{ccc} \boldsymbol{\Lambda} & \ldots & \boldsymbol{\Lambda} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Lambda} & \ldots & \boldsymbol{\Lambda} \end{array} \right)}_{y \text{ times}} \Bigg\} x \text{ times}$$

Similarly, $\tilde{\boldsymbol{C}}_{x,y}$ is the $(xp) \times (yp)$ block matrix, where each block consists of the $p \times p$ matrix $\boldsymbol{C}$, i.e

$$\tilde{\boldsymbol{C}}_{x,y} = \underbrace{\left(\begin{array}{ccc} \boldsymbol{C} & \ldots & \boldsymbol{C} \\ \vdots & \ddots & \vdots \\ \boldsymbol{C} & \ldots & \boldsymbol{C} \end{array} \right)}_{y \text{ times}} \Bigg\} x \text{ times}$$

Denote also by $n_{i.}$ the sum of $\{n_{ij}\}$ over j.

Following the above notation, for the example in Figure 4.1, the resulting $\boldsymbol{K}$ is an $Np \times Np$ block matrix having the following form:

$$\boldsymbol{K} = \begin{pmatrix} \tilde{\boldsymbol{\Lambda}}_{n_{1.},n_{1.}} + \tilde{\boldsymbol{K}}_1 & \tilde{\boldsymbol{\Lambda}}_{n_{1.},n_{2.}} & \tilde{\boldsymbol{\Lambda}}_{n_{1.},n_{3.}} & \tilde{\boldsymbol{\Lambda}}_{n_{1.},n_{4.}} \\ \tilde{\boldsymbol{\Lambda}}_{n_{2.},n_{1.}} & \tilde{\boldsymbol{\Lambda}}_{n_{2.},n_{2.}} + \tilde{\boldsymbol{K}}_2 & \tilde{\boldsymbol{\Lambda}}_{n_{2.},n_{3.}} & \tilde{\boldsymbol{\Lambda}}_{n_{2.},n_{4.}} \\ \tilde{\boldsymbol{\Lambda}}_{n_{3.},n_{1.}} & \tilde{\boldsymbol{\Lambda}}_{n_{3.},n_{2.}} & \tilde{\boldsymbol{\Lambda}}_{n_{3.},n_{3.}} + \tilde{\boldsymbol{K}}_3 & \tilde{\boldsymbol{\Lambda}}_{n_{3.},n_{4.}} \\ \tilde{\boldsymbol{\Lambda}}_{n_{4.},n_{1.}} & \tilde{\boldsymbol{\Lambda}}_{n_{4.},n_{2.}} & \tilde{\boldsymbol{\Lambda}}_{n_{4.},n_{3.}} & \boldsymbol{\Lambda} + \boldsymbol{K}_4 \end{pmatrix}, \tag{4.12}$$

where,

$$\tilde{\boldsymbol{K}}_1 = \begin{pmatrix} \tilde{\boldsymbol{C}}_{n_{11},n_{11}} + \tilde{\boldsymbol{K}}_{11} & \tilde{\boldsymbol{C}}_{n_{11},n_{12}} & \tilde{\boldsymbol{C}}_{n_{11},n_{13}} \\ \tilde{\boldsymbol{C}}_{n_{12},n_{11}} & \boldsymbol{C} + \boldsymbol{K}_{12} & \boldsymbol{C} \\ \tilde{\boldsymbol{C}}_{n_{13},n_{11}} & \boldsymbol{C} & \boldsymbol{C} + \boldsymbol{K}_{13} \end{pmatrix},$$

$$\tilde{\boldsymbol{K}}_2 = \begin{pmatrix} \boldsymbol{C} + \boldsymbol{C}_2 + \boldsymbol{C}_{21} + \boldsymbol{J}_{211} & \boldsymbol{C} \\ \boldsymbol{C} & \boldsymbol{C} + \boldsymbol{C}_2 + \boldsymbol{C}_{22} + \boldsymbol{J}_{212} \end{pmatrix},$$

$$\tilde{\boldsymbol{K}}_3 = \begin{pmatrix} \boldsymbol{C} + \boldsymbol{C}_3 + \boldsymbol{C}_{31} + \boldsymbol{J}_{311} & \boldsymbol{C} + \boldsymbol{C}_3 \\ \boldsymbol{C} + \boldsymbol{C}_3 & \boldsymbol{C} + \boldsymbol{C}_3 + \boldsymbol{C}_{31} + \boldsymbol{J}_{312} \end{pmatrix},$$

and

$$\boldsymbol{K}_4 = \boldsymbol{C} + \boldsymbol{C}_4 + \boldsymbol{C}_{41} + \boldsymbol{J}_{411},$$

with,

$$\tilde{\boldsymbol{K}}_{11} = \begin{pmatrix} \boldsymbol{C}_1 + \boldsymbol{C}_{11} + \boldsymbol{J}_{111} & \boldsymbol{C}_1 \\ \boldsymbol{C}_1 & \boldsymbol{C}_1 + \boldsymbol{C}_{12} + \boldsymbol{J}_{112} \end{pmatrix},$$

$$\boldsymbol{K}_{12} = \boldsymbol{C}_1 + \boldsymbol{C}_{12} + \boldsymbol{J}_{121}$$

and

$$\boldsymbol{K}_{13} = \boldsymbol{C}_1 + \boldsymbol{C}_{13} + \boldsymbol{J}_{131}.$$

All the involved covariance matrices have dimension $p \times p$, where p denotes the length of the descriptor vector $\boldsymbol{\theta}$. The form of (4.12) illustrates that $\boldsymbol{K}$ is constructed from smaller block matrices, whose structure is determined from the grouping applied to the descriptors at the different levels of the framework in Figure 4.1. The diagonal blocks of $\boldsymbol{K}$ are a sum of the different sources of variability assigned to each data source, conditional on reality descriptor $\boldsymbol{\theta}_0$. In other words, they show how uncertainty at each level of the hierarchical framework is propagated through from the data level to the reality level. On the other hand, the off-diagonal blocks illustrate the grouping structure of the simulators in the framework, through the variability components which are common between the simulators. The shared family discrepancy from reality, expressed through the covariance matrix $\boldsymbol{\Lambda}$ appears in all the blocks of $\boldsymbol{K}$, as in the simpler framework. This implies that simulators are always imperfect in reproducing reality and share common discrepancies from it, regardless of the connections between them. As also illustrated in Figure 4.1, each diagonal block i of $\boldsymbol{K}$ is a sum of the covariance matrices determining the length of "branches" that intervene between $\hat{\boldsymbol{\theta}}^{(i)}$ and $\boldsymbol{\theta}_0$ in the tree. On the other hand, each $(i,j)^{th}$ off-diagonal block is a sum of covariance matrices corresponding to the "branches" that are shared between $\hat{\boldsymbol{\theta}}^{(i)}$ and $\hat{\boldsymbol{\theta}}^{(j)}$ in moving from the data level to the reality level of the tree.

After obtaining the mean $\boldsymbol{\Theta}_0$ ((4.11)) and covariance matrix $\boldsymbol{K}$ ((4.12)) of the

multivariate normal distribution $\pi(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta}_0)$, its density becomes proportional to

$$\exp\left[-\frac{1}{2}\left(\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0\right)' \boldsymbol{K}^{-1}\left(\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0\right)\right]. \tag{4.13}$$

Equations (4.8), (4.9) and (4.13) provide the expressions for the terms required to obtain the generic posterior shown in (4.1). Substituting these expressions in (4.1) yields,

$$\begin{aligned}\pi\left(\boldsymbol{\theta}_0|\hat{\boldsymbol{\theta}}_0,\hat{\boldsymbol{\theta}}^{(1)},\ldots,\hat{\boldsymbol{\theta}}^{(N)}\right) \propto \exp\Bigg\{-\frac{1}{2}\Bigg[&\left(\boldsymbol{\theta}_0-\boldsymbol{\mu}_0\right)' \boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{\theta}_0-\boldsymbol{\mu}_0\right)+\\ &\left(\hat{\boldsymbol{\theta}}_0-\boldsymbol{\theta}_0\right)' \boldsymbol{J}_0^{-1}\left(\hat{\boldsymbol{\theta}}_0-\boldsymbol{\theta}_0\right)+\left(\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0\right)'\boldsymbol{K}^{-1}\left(\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0\right)\Bigg]\Bigg\}.\end{aligned} \tag{4.14}$$

Expression (4.14) suggests that an explicit expression for the posterior requires inverting the matrix $\boldsymbol{K}$. In the framework of Section 2.6, the convenient structure of $\boldsymbol{K}$ allows obtaining an analytical expression for $\boldsymbol{K}^{-1}$ using the Woodbury formula (Press et al., 1992, Section 2.7) (see Chandler (2013, suppl. material S1)). In the extended framework, $\boldsymbol{K}$ could probably be derived via repeated application of the Woodbury formula (by inverting recursively the blocks of $\boldsymbol{K}$). However, it is not clear whether this will reveal any insight or save much computational time compared to numerical inversion of $\boldsymbol{K}$ (especially for a large number of levels in the MME structure) and therefore in this study, $\boldsymbol{K}$ is inverted numerically.

Consider partitioning $\boldsymbol{K}^{-1}$ into N^2 $p \times p$ sub-blocks $\widetilde{\boldsymbol{K}}_{ij}$, as follows,

$$\boldsymbol{K}^{-1}=\begin{pmatrix}\widetilde{\boldsymbol{K}}_{11} & \ldots & \widetilde{\boldsymbol{K}}_{1N}\\ \vdots & \ddots & \vdots\\ \widetilde{\boldsymbol{K}}_{N1} & \ldots & \widetilde{\boldsymbol{K}}_{NN}\end{pmatrix}. \tag{4.15}$$

Using the standard representation of a quadratic form (Schott, 2005, p.15), the expression $\left(\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0\right)' \boldsymbol{K}^{-1}\left(\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0\right)$ in (4.14) can be expressed in the form of sums of squares involving $\boldsymbol{\theta}_0$:

$$\left(\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0\right)' \boldsymbol{K}^{-1}\left(\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0\right)=\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\hat{\boldsymbol{\theta}}^{(i)}-\boldsymbol{\theta}_0\right)' \widetilde{\boldsymbol{K}}_{ij}\left(\hat{\boldsymbol{\theta}}^{(j)}-\boldsymbol{\theta}_0\right),$$

with $\widetilde{\boldsymbol{K}}_{ij}$ as defined in (4.15).

Substituting this expression back into (4.14) and rearranging yields,

$$\pi\left(\boldsymbol{\theta}_0|\hat{\boldsymbol{\theta}}_0,\hat{\boldsymbol{\theta}}^{(1)},\ldots,\hat{\boldsymbol{\theta}}^{(N)}\right) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\theta}_0-\boldsymbol{\tau}\right)'\boldsymbol{S}^{-1}\left(\boldsymbol{\theta}_0-\boldsymbol{\tau}\right)+\text{terms not involving }\boldsymbol{\theta}_0\right], \tag{4.16}$$

where,

$$\boldsymbol{S^{-1}} = \boldsymbol{\Sigma}_0^{-1} + \boldsymbol{J}_0^{-1} + \sum_{i=1}^{N}\sum_{j=1}^{N} \widetilde{\boldsymbol{K}}_{ij}, \tag{4.17}$$

and

$$\boldsymbol{\tau} = \boldsymbol{S}\left[\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{J}_0^{-1}\hat{\boldsymbol{\theta}}_0 + \sum_{i=1}^{N}\sum_{j=1}^{N} \widetilde{\boldsymbol{K}}_{ij}\hat{\boldsymbol{\theta}}^{(j)}\right]. \tag{4.18}$$

The only distribution whose form is proportional to the expression in (4.16) is a multivariate normal distribution, with mean $\boldsymbol{\tau}$ and covariance matrix $\boldsymbol{S}$.

The posterior precision $\boldsymbol{S}^{-1}$ ((4.17)) is a sum of precision matrices representing the different sources of variability in the framework. The first two terms represent contribution of the prior precision and precision in estimating real climate through observations, as in the simpler framework. The last term results from the numerical inversion of the covariance matrix $\boldsymbol{K}$ and therefore expresses the contribution of the precision in $\pi\left(\hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(N)} | \boldsymbol{\theta}_0\right)$. In other words, it shows contribution of the sources of variability propagated through from each data source in the data level, up to the reality level. This also includes the shared components between different variant descriptor estimators (through the off-diagonal blocks of $\boldsymbol{K}$). It is also important to note that the contribution of shared family discrepancy from reality, expressed through the covariance matrix $\boldsymbol{\Lambda}$, is implicitly present in the posterior precision, since $\boldsymbol{\Lambda}$ is involved in all the blocks of $\left\{\widetilde{\boldsymbol{K}}_{ij},\ i, j = 1, \ldots, N\right\}$, as shown in (4.12).

The form of the posterior mean (4.18) is a matrix-weighted average of the prior mean $\boldsymbol{\mu}_0$, the real-climate descriptor $\hat{\boldsymbol{\theta}}_0$ and the variant descriptor estimators $\left\{\hat{\boldsymbol{\theta}}^{(l)},\ l = 1, \ldots, N\right\}$, similarly to the simpler framework. The weight of each $\hat{\boldsymbol{\theta}}^{(l)}$ is now expressed through the contribution of $\hat{\boldsymbol{\theta}}^{(l)}$ to the blocks of $\boldsymbol{K}^{-1}$. This includes the various sources of variability propagated through from $\hat{\boldsymbol{\theta}}^{(l)}$ in the data level to $\boldsymbol{\theta}_0$ in reality level, as well as the variability components shared between $\hat{\boldsymbol{\theta}}^{(l)}$ and the other variant descriptor estimators.

The expressions (4.17) and (4.18) suggest that if the covariance matrices $\boldsymbol{\Sigma}_0$, $\boldsymbol{J}_0$ and $\boldsymbol{K}$, as well as the prior mean $\boldsymbol{\mu}_0$ were known, then the parameters of the posterior $\pi\left(\boldsymbol{\theta}_0 | \hat{\boldsymbol{\theta}}^{(0)}, \ldots, \hat{\boldsymbol{\theta}}^{(N)}\right)$ could be calculated analytically, allowing inference about true-climate descriptor $\boldsymbol{\theta}_0$, conditional on climate observations $\left(\hat{\boldsymbol{\theta}}_0\right)$ and simulator outputs $\left(\hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(N)}\right)$ from the underlying MME. In practice however, those quantities are unknown.

A RPM implementation would assign estimates to the unknown quantities, whereas a fully Bayesian implementation would account for uncertainty in the unknown quan-

tities by specifying distributions to them. However, since a fully Bayesian implementation requires the use of numerical algorithms (e.g. Gibbs sampler) for calculating the posterior, it is expected to be computationally costly if applied to the proposed framework with a large number of levels in the MME structure. Besides, each additional level would require making prior judgements for a collection of covariance matrices expressing within-group variability, which even for the simpler framework is shown to be a lengthy process (see Sections 3.5.2-3.5.3). Considering also the good performance of the RPM implementation in quantifying uncertainty under the simpler framework (see Section 3.7), an implementation similar to the one in Section 3.5.1 is adopted for the extended framework, named as "revised poor man with groups" (RPMG) implementation. Details about the implementation are discussed in Section 4.4.

4.4 Challenges in implementation

According to (4.17)-(4.18), the unknown quantities in the posterior parameters consist of: The parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ in the prior of real-climate descriptor $\boldsymbol{\theta}_0$ ((4.7), p.83), the covariance matrix $\boldsymbol{J}_0$ in the distribution of $\hat{\boldsymbol{\theta}}_0$ (2.4) and the block matrices $\left\{\widetilde{\boldsymbol{K}}_{ij},\ i=1,\ldots,N, j=1,\ldots,N\right\}$ ((4.15), p.87). Similarly to the simpler framework, the prior parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are estimated from earlier data. Additionally, the covariance matrix $\boldsymbol{J}_0$ is obtained based on fitting climate observations to the chosen mimic, as shown in Section 3.5.1. What remains is estimation of $\left\{\widetilde{\boldsymbol{K}}_{ij},\ i=1,\ldots,N, j=1,\ldots,N\right\}$, the block matrices of $\boldsymbol{K}^{-1}$ ((4.15), p.87). As also discussed in Section 4.3, the blocks of $\boldsymbol{K}$ consist of the covariance matrices expressing sources of variation in all the levels of the proposed framework. Estimation of $\left\{\widetilde{\boldsymbol{K}}_{ij},\ i,j=1,\ldots,N\right\}$ therefore requires estimates for the underlying covariance matrices. The rest of the section deals with estimation under the framework structure illustrated in Figure 4.1. However, the arguments can be generalised for a framework structure with an arbitrary number of levels. Section 4.8 provides an algorithm for the estimation procedure in the general case of a framework structure with arbitrary number of levels.

For the example in Figure 4.1, the covariance matrices involved in the blocks of $\boldsymbol{K}$ are: $\{\boldsymbol{J}_{ijk}\}$, $\{\boldsymbol{C}_{ij}\}$, $\{\boldsymbol{C}_i\}$, $\boldsymbol{C}$ and $\boldsymbol{\Lambda}$, for $\{i=1,\ldots,m, j=1,\ldots,n_i, k=1,\ldots,n_{ij}\}$, where $m=4$, $n_i=\{3,2,1,1\}$ and $n_{ij}=\{2,1,1,2,2,1\}$, starting from the data level and moving up to the reality level. The covariance matrices $\{\boldsymbol{J}_{111},\ldots,\boldsymbol{J}_{411}\}$ in the variant level are estimated by fitting the chosen mimic to the corresponding GCM outputs, as shown in Section 3.5.1.

For simplicity, the covariance matrix $\boldsymbol{\Lambda}$ is estimated using $\hat{\boldsymbol{\Lambda}}$ obtained in Section 3.5.1 for the simpler framework. Note however the slightly different interpretation of $\boldsymbol{\omega}$ in the extended framework compared to the simpler. The term $\boldsymbol{\omega}$ here represents the propensity of family descriptors $\{\boldsymbol{\theta}_i,\ i = 1, \ldots, m\}$ collectively to deviate from reality $\boldsymbol{\theta}_0$ (see Section 4.2). On the other hand, in the simpler framework, it represents the collective propensity of the simulator descriptors to deviate from reality. Therefore, the use of $\hat{\boldsymbol{\Lambda}}$ as obtained from the simpler framework is considered as a naive choice of estimator. Alternatively, $\hat{\boldsymbol{\Lambda}}$ could be estimated following the process described in Section 3.5.1, but based on bootstrapping of the grouped instead of the ungrouped simulator outputs and using the family instead of the simulator descriptors to estimate $\boldsymbol{\omega}_i$ for each bootstrap sample i. This is left as future work.

The remaining covariance matrices to be estimated correspond to the intermediate levels between variant level and reality level. They all express deviation of sibling descriptors from their parent descriptor. A natural estimator in this case could be the sample covariance matrix, similarly to estimation of $\boldsymbol{C}$ in the simpler framework (see ((2.12), p.30). Consider, for example, estimation of the covariance matrix $\boldsymbol{C}_{11}$ in the simulator level of Figure 4.1, expressing deviation of sibling variant descriptors $\boldsymbol{\theta}_{111}$ and $\boldsymbol{\theta}_{112}$ from their parent descriptor $\boldsymbol{\theta}_{11}$. Provided the unknown variant descriptors are estimated from their corresponding estimates $\hat{\boldsymbol{\theta}}_{111}$ and $\hat{\boldsymbol{\theta}}_{112}$ respectively in the variant level, $\boldsymbol{C}_{11}$ can be estimated as:

$$\hat{\boldsymbol{C}}_{11} = \sum_{k=1}^{2} \left(\hat{\boldsymbol{\theta}}_{11k} - \bar{\hat{\boldsymbol{\theta}}}_{11}\right)\left(\hat{\boldsymbol{\theta}}_{11k} - \bar{\hat{\boldsymbol{\theta}}}_{11}\right)', \tag{4.19}$$

where $\bar{\hat{\boldsymbol{\theta}}}_{11} = \frac{1}{2}\sum_{k=1}^{2} \hat{\boldsymbol{\theta}}_{11k}$ is the sample mean of the variant descriptors which are centred on $\boldsymbol{\theta}_{11}$.

It can be argued that the sample covariance matrix can be used in all the levels of the hierarchy in the same way as shown in (4.19), provided that at least two sibling descriptors exist, in order for their sample covariance matrix to be estimated. However, the potential for grouping of simulators under the proposed framework does not exclude the possibility of having singleton groups, i.e. descriptors without siblings in the hierarchical structure. As illustrated in Masson and Knutti (2011, Figure 1), grouping of CMIP3 simulators for surface temperature and precipitation according to institution or component of the same atmospheric model yields singleton groups. Similarly, Knutti et al. (2013, Figure 1) suggest that this is also the case for CMIP5 simulators grouped according to similarities of code. In the absence of siblings, the covariance matrix expressing deviation of the descriptor from its parent descriptor cannot be obtained using the formula for the sample covariance matrix illustrated

in (4.19). In the MME structure of Figure 4.1 for example, estimation of $\boldsymbol{C}_{12}$ in the simulator level expressing deviation of the variant descriptor $\boldsymbol{\theta}_{121}$ from its parent descriptor $\boldsymbol{\theta}_{12}$ cannot be obtained as illustrated in (4.19), since $\boldsymbol{\theta}_{121}$ has no siblings.

An alternative could be to remove all child descriptors without siblings from the framework and therefore avoid the issue of estimating variability in singleton groups. For example, removing the variant descriptor $\boldsymbol{\theta}_{121}$ (and consequently $\boldsymbol{C}_{12}$ in the parent node) in Figure 4.1, implies that $\hat{\boldsymbol{\theta}}_{121}$ is centred on $\boldsymbol{\theta}_{12}$ and its deviation from it is now determined by $\boldsymbol{J}_{121}$. In mathematical terms, this is expressed as $\hat{\boldsymbol{\theta}}_{121}|\boldsymbol{\theta}_{12}, \boldsymbol{J}_{121} \sim MVN\left(\boldsymbol{\theta}_{12}, \boldsymbol{J}_{121}\right)$. Therefore, the covariance matrix $\boldsymbol{C}_{12}$ in the simulator level is now substituted by $\boldsymbol{J}_{121}$. However, these modifications would contradict the assumption of exchangeability of sibling nodes, conditional on their parent node, since $\boldsymbol{C}_{11}$ and $\boldsymbol{J}_{121}$ have different interpretations. $\boldsymbol{C}_{11}$ represents deviation of variant descriptors $\{\boldsymbol{\theta}_{11k}\}$ from their parent simulator node $\boldsymbol{\theta}_{11}$, whereas $\boldsymbol{J}_{121}$ now represents deviation of the descriptor estimator $\hat{\boldsymbol{\theta}}_{121}$ from the parent $\boldsymbol{\theta}_{12}$. It is therefore not reasonable to assume the two covariance matrices as belonging to exchangeable sibling nodes. In general, retaining the exchangeability assumptions in the proposed framework is important, since it justifies the use of a common distribution for the sibling descriptors, conditional on their parent. Additionally, it ensures that the interpretation of the nodes at each level of the hierarchy is consistent.

To summarize: nested simulator grouping potentially results in sparse descriptor structures, i.e. groups of sibling descriptors with small sample size, including singleton descriptor groups. This creates a challenge for the estimation of covariance matrices expressing deviation of child descriptors from their parents. At a fundamental level, the problem is one of estimating within-group variability in sparse data structures. Additional challenges include the fact that the proposed framework structure is multilevel (i.e. nested) and can potentially have an arbitrary number of levels. Furthermore, the data (and consequently the descriptors at each level) are assumed to be multivariate and the groups of sibling descriptors at each level do not necessarily have the same sample size, leading to an unbalanced structure. Finally, the covariance matrices expressing within-group variability among sibling descriptors at each level in the framework are not necessarily equal.

Consequently, in order to implement the framework, it is required to propose a method for estimating within-group variability in data structures of high complexity, such as those suggested by the proposed framework. The resulting estimates at each level of the framework will then be used to estimate the remaining covariance matrices required to estimate $\boldsymbol{K}^{-1}$ in (4.15). This will complete estimation of the terms required to obtain an analytic expression for the true-climate posterior

$\pi\left(\boldsymbol{\theta}_0|\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(N)}\right)$ ((4.16)), under the RPMG implementation.

Section 4.5 reviews the existing literature on estimating variability in nested, sparse data structures such as the one illustrated in Figure 4.1. Section 4.6 defines the proposed model for estimating within-group variability. Model inference is considered in Section 4.7. Section 4.8 describes an algorithm which summarizes the main steps for framework implementation and provides an overview of the proposed method for estimating within-group variability.

4.5 Variance component estimation in nested, sparse data structures

The proposed framework is an example of a nested data structure. The conventional approach for the statistical analysis of this kind of data structures is by representing the structure as a multilevel model. Multilevel modelling has gained increasing popularity in a wide variety of disciplines (Fielding and Goldstein, 2006, Section 7), ranging from sociology, epidemiology and demography (Clarke and Wheaton, 2007), to neuroimaging (Friston et al., 2002; Baldi and Long, 2001; Worsley et al., 2002) and environmental sciences (Clark and Gelfand, 2006). The main statistical model for representing multilevel structures is according to Snijders and Bosker (2002, Section 4) the "hierarchical linear model", its simplest form being the "one-way random effects ANOVA model". The latter is employed for representing two-level structures of individuals nested within groups, one level being the "group level" and the other being the "individual level". The model partitions total variance into within and between-group components. The wide applicability of multilevel modelling in modern applications however, suggests that extensions beyond the standard one-way random effects model are required. Fielding and Goldstein (2006, Section 2.2) characterise the two-level representation as often being "unduly simplistic" and give examples of educational and social data structures represented by three or four-level nested models. Ecochard and Clayton (1998) illustrate that the three-level representation can be extended in many ways, in order to fit data structures of increased complexity such as the MME structure under the framework of Section 4.2. Apart from the extension to more levels (Snijders and Bosker, 2002, Section 5.5), explanatory variables with fixed or random coefficients and interaction terms can be added to the model to explain part of the variability (Snijders and Bosker, 2002, Section 4.4). Additional extensions involve representing complex covariance structures (Goldstein, 1995, Section 3), multivariate responses (Snijders and Bosker, 2002, Section 13) and binary or discrete data (Searle et al., 2006, Section 10).

On the other hand, interpretation and inference of complex multilevel models is a challenging task, which makes model specification a non-trivial decision (Gelman and Hill, 2007, Section 18.5). According to Snijders and Bosker (2002, Section 6.4), the challenge is to achieve a "satisfactory" representation of the data structure, while simultaneously considering the statistical concerns regarding inference of the resulting model. For example, the number of parameters to be estimated increases considerably for multiple-level models, while at the same time the group sample sizes at finer levels decrease, therefore leading to increased computational complexity and reduced estimation precision. For this reason, Hox (2010, Section 2.4.1) suggests keeping the model as simple as possible, especially at the finer levels and impose random effects only when there is "strong theoretical or empirical justification". Gelman and Hill (2007, Section 12.9) also argue that it might not worth the effort fitting complex models with small datasets, because of the resulting high uncertainties in inference. In a similar spirit, Snijders and Bosker (2002, Section 5.5) note that formulation of models at least as complex as three-level models requires either substantial prior knowledge about the underlying parameters, or a "good theory".

Because of the complexity of multilevel models, traditional estimation methods such as maximum-likelihood (ML) and method of moments, sometimes fail to provide parameter estimates. Rao et al. (1981) illustrate that this is the case for ANOVA and ML estimation of within-group variance in the one-way random effects model with unequal within-group variances. In ANOVA, the method of moments estimator of within-group variances turns out to be the sample variance, which is undefined for singleton groups. Similarly, the maximum-likelihood equations show clearly that an ML estimate of within-group variability cannot be obtained. Therefore, both methods would fail to provide estimates of within-group variability for potential singleton groups in the proposed MME structure.

An additional class of estimators of variance components is considered to be "MINQUE" (minimum norm quadratic unbiased estimators) family of estimators. MINQUE estimators are derived by minimising a weighted Euclidean norm resulting from the difference of two quadratic forms of the data acting as estimators of variance components, where one of the two is constructed based on prior values for the variance components (Searle et al., 2006, Section 11.3). The minimisation is performed using constraints determined based on desired properties for the resulting estimators, such as unbiasedness and minimum variance. Rao et al. (1981, Section 2.5) present the MINQE non-negative estimator of within-group variabilities for the one-way random effects model with unequal variances when the unbiasedness constraint is removed, and set prior values of within-group variability being equal to between-group variability, for all groups. Although the resulting estimator of within-group

variability is defined even for singleton groups, the prior assumption of equal within and between-group variabilities in all groups is not realistic for the framework of Section 4.2, assuming that simulator grouping exists. Besides, the sensitivity of the MINQUE family of estimators on prior values for the unknown variance components often creates scepticism (Searle et al., 2006, Section 11.3c), since it is non-trivial to attain reasonable prior judgements for variance components, especially in multiple-level frameworks with unbalanced structures and singleton groups like the framework of Section 4.2.

An alternative approach is considered in Goldstein (1995, Section 2.5) to be the iterative generalized least squares (IGLS) method. The method is based on iteratively producing successive estimates of the fixed and random parameters in the one-way random effects model based on least squares principle, until convergence is achieved. However, under the framework's normality assumptions introduced in Section 4.2, the estimates reduce to the ML estimates, which cannot be obtained for the within-group variance component in singleton groups.

Additional difficulties arise when attempting to fit a model with an arbitrary number of levels, to represent structures such as that of the framework in Section 4.2 (Rabe-Hesketh et al., 2002; Fielding and Goldstein, 2006; Beckmann et al., 2003). Many approaches employ generalized linear models for representing multilevel structures of arbitrary number of levels. Rabe-Hesketh et al. (2002) represent such a model as a generalised linear model and review various methods for parameter estimation. Marginal and penalised quasi-likelihood (MQL and PQL respectively) methods apply Taylor series expansion to approximate the expectation of the response, which is analytically intractable. The PQL method provides an improved expansion of the Taylor series compared to MQL, by estimating the random effects instead of setting them to zero in the expansion. The model parameters are then estimated using IGLS. On the other hand, Gaussian and adaptive Gaussian quadrature avoid the IGLS algorithm by collecting all the unknown parameters in a parameter vector and derive the likelihood as a product of conditional likelihoods at each level. In Gaussian quadrature, the resulting integrals are approximated using Cartesian product quadrature, whereas in adaptive Gaussian quadrature the approximation is improved based on a posterior distribution for the random effects. In a comparison of methods, the authors note that MQL and PQL are superior in terms of computational efficiency, whereas adaptive quadrature is less sensitive to normality assumptions and computationally more efficient than ordinary quadrature. However, Fielding and Goldstein (2006, Section 6.4) warn that quasi-likelihood might fail to achieve convergence for some types of data structures, while also still being computationally demanding.

A framework attempting to reduce the computational burden of estimation in

multiple-level models is illustrated in Beckmann et al. (2003), assuming known variances. The idea relies on the so called "summary statistics" approach (Holmes and Friston, 1998). Initially assuming a two-level design, the method fits a regression model to the raw data of each group in the individual level. The estimated regression parameters corresponding to each group are then interpreted as being the raw data to fit a regression model on the group level. The method can be generalised to models with arbitrary number of levels, achieving therefore parameter estimation in each level by using only the parameter estimates from the coarser levels and thus avoiding the need to revisit the original dataset. The idea of using parameter estimates from coarser levels as raw data for inference in finer levels is quite appealing for implementing frameworks like that of Section 4.2, which can potentially consist of multiple levels. The method of Beckmann et al. (2003) is subsequently extended by Woolrich et al. (2004), who follow a fully Bayesian approach to estimate group effects and between-group variability in the group level, using non-informative reference priors for the model parameters. The posterior of the parameters is then derived by incorporating the resulting posterior from the individual level model. Finally, model parameters are estimated based on approximating the marginal posterior distribution of the regression parameter with a t-distribution. The method can be generalised to structures of arbitrary number of levels, by retaining the approximation to the posterior in every level and therefore incorporating the posterior parameters from the coarser level to inform the posterior in the finer level. The authors however emphasize the importance of prior choice in finer levels, when fewer data are available. Additionally, they do not explicitly consider estimation of within-group variability, especially in the presence of singleton groups, which is of primary interest in the implementation of the framework of Section 4.2.

It is widely recognised that variance component estimation in sparse data structures is an important issue in statistical analysis (Clarke and Wheaton, 2007; Bell et al., 2008, 2010). Beckmann et al. (2003) characterise the group sample size as often being "restrictive" in the model choice for nested structures. The authors refer to the situation of having only singleton groups in a particular level of a multilevel model assuming unequal within-group variances, in which case the model is not estimable. The issue of data sparseness is also common in multilevel modelling of data from population-based surveys (Clarke and Wheaton, 2007; Bell et al., 2008). According to Clarke and Wheaton (2007), data sparseness induces bias in estimation and potential lack of model performance, as well as convergence issues in numerical algorithms for parameter estimation in complex models. The authors refer to the use of clustering techniques to overcome the problem, by aggregating existing groups together into larger groups, according to some pre-specified "similarity" metrics. A serious

implication however, is that clustering methods can reduce the effective sample size or the group effects of original interest in the analysis. Besides, if interest is on implementing frameworks where clustering is determined based on expert judgement, such as the framework of Section 4.2, any modification of the clustering would deviate from the initial scope of the study.

The issue of estimating within-group variability in the presence of small groups is frequently encountered in neuroimaging (Lönnstedt and Speed, 2002; Wright and Simon, 2003; Baldi and Long, 2001). Gene replication data produced in microarray experiments are used to infer about potential gene differences. The analysis is usually performed using multilevel models. However, the small number of gene replications, leading to low degrees of freedom in estimating within-gene variability, limits the power of the statistical tests for gene differentiation (Wright and Simon, 2003; Lönnstedt and Speed, 2002; Baldi and Long, 2001). Furthermore, the assumption of equal within-gene variance is not considered to be realistic (Wright and Simon, 2003). The situation of small group sizes with the additional assumption of unequal within-group variances is very similar to the underlying assumptions regarding grouping of descriptors in the framework of Section 4.2. The problem is usually approached by "borrowing strength" from neighbouring groups to improve the required estimate, an effect often referred to as "shrinkage" (Stein, 1981). Shrinkage is often applied implicitly through parametric empirical Bayes estimation, as well as through non-Bayesian shrinkage estimators.

A Bayesian interpretation allows for sharing information among sibling groups at a particular level of the hierarchy, by assuming that individual group parameters are exchangeable among groups and therefore a common prior is assigned to them. In the empirical Bayes approach, the hyperparameters are estimated from the data (Casella, 1985). For example, Lönnstedt and Speed (2002) model gene replication data using a Normal distribution with different mean and variance for each gene. They assign the Gamma distribution with fixed scale parameter as a common prior for the within-group precisions. A Normal distribution is assigned to each group mean, conditional on the corresponding group precision. The unknown hyperpriors for the means are specified based on subjective judgements. The shape parameter in the prior for the precision is estimated from the empirical within-gene variances using a method of moments. However, if singleton groups are present, an estimate for the shape parameter cannot be obtained empirically, therefore requiring alternative approaches. Baldi and Long (2001) follow a similar approach which provides estimates of within-group variability even in singleton groups, by assigning a scaled inverse gamma prior to the variance of each group. The within-group variance is then estimated to be the mean from the corresponding marginal posterior. The resulting form of the estimate

allows the prior for the variance to completely dominate the likelihood in case of singleton groups. The hyperparameters to be estimated consist of a prior value for the within-group variance as well as a parameter determining the degree of confidence for the prior variance, relative to the empirical one. The prior value for the variance is obtained by pooling empirical variances among "neighbouring groups" belonging to a window of a pre-specified length, or equivalently sibling groups in the framework of Section 4.2.

The specification of a common prior for variance components within siblings provides a convenient way to share information between them. Furthermore, the form of the within-group estimate resulting from the posterior provides a justifiable automatic weighting of the relevant contributions from the empirical and prior variances. However, determining the hyperparameters requires care. An attempt to make realistic *a priori* judgements about group variances when grouping is based on expert judgement such as in the framework of Section 4.2 seems non-trivial, especially in the presence of singleton groups. Sections 3.5.2-3.5.3 illustrating the fully Bayesian implementations of the simpler framework show clearly that choosing "reasonable" priors for complex models involving many parameters is not straightforward. The process is expected to be even more complicated for multivariate data under the proposed framework, which can potentially consist of many levels, or if a fully Bayesian implementation is attempted. A natural solution is to seek priors that are non-informative in the sense that they do not influence the results unduly. However, Gelman (2006) points out that standard "non-informative" prior choices for variance components can have a large influence on the results, in situations involving sparse data.

Alternative approaches exist, which apply shrinkage to the within-group variances, while avoiding any *a priori* assumptions about model parameters. The proposed estimator in Cui et al. (2005) borrows strength from all the individual within-group variances by shrinking them towards their "common corrected geometric mean". The degree of shrinkage is determined by the degree of homogeneity of the individual variances, in the sense that the less homogeneous the variances are, more weight is given to the individual variances instead. The resulting shrinkage estimator involves natural logarithms of the residual sums of squared errors within each group, which is a potential limitation of the method in the presence of singleton groups, where the residual sums of squared errors is zero. Wright and Simon (2003) assign a common inverse-Gamma distribution to the within-group variances of the standard linear model in order to share information among groups but perform parameter estimation from a frequentist instead of a Bayesian perspective. They perform ML estimation based on the suggested random variance model. The parameters of the inverse-Gamma distribution are estimated by fitting an F-distribution to the empirical estimates of

within-group variance and using ML estimation over the parameters of interest. However, in the presence of singleton groups, when the empirical within-group variance is undefined, such an approach would fail to provide estimates of the parameters of the inverse-Gamma distribution and consequently an estimate for within-group variability.

To summarize, the methods reviewed above approach some of the issues relevant to the implementation of the framework of Section 4.2, especially group inhomogeneity and sparse data structures. Particular emphasis is given to estimation of within-group variability in sparse data, which is considered to be the main challenge in implementation of the proposed framework. Most of the reviewed methods handle the presence of sparse data in estimation of multilevel models by allowing for sharing of information among groups at each individual level. For estimation of within-group variance, this is usually achieved by assigning a common distribution to the within-group variances at each level and therefore creating an additional level to the assumed model for the data. Estimation of the model parameters can be handled either in a Bayesian or in a frequentist framework. A Bayesian approach usually leads to empirical Bayes methods, since it allows estimating hyperparameters from the data and therefore avoids the more complicated fully Bayesian implementation, which requires setting hyperpriors to them. However, even the empirical Bayes methods require subjective judgements about the model parameters, which is expected to be non-trivial for the proposed framework structure. On the other hand, the reviewed frequentist approaches do not handle the presence of singleton groups, since they require values of the empirical within-group variance in their proposed estimators for each group.

It can be deduced from the above discussion that none of the reviewed methods addresses exactly all the underlying challenges in the implementation of the framework in Section 4.2. Also, the reviewed approaches do not explicitly handle the presence of singleton groups in ways which can be conveniently generalised to multiple-level and multivariate data structures. A new methodology is proposed here which addresses the issues discussed in Section 4.4 regarding the implementation of the framework of Section 4.2. The proposed method builds on some of the reviewed idea of information sharing among sibling groups. The proposed method is shown to work in the presence of data sparseness, as well as singleton groups. It can be also generalized to structures with an arbitrary number of levels, as well as for multivariate data. Motivated by the summary statistics approach, inference is performed recursively to allow for a structure of arbitrary number of levels. Firstly, sibling descriptors in a particular group are collected together to form individual groups. A random effects model (defined in Section 4.6) is then fitted to descriptor groups sharing the same "grand-

parents", or equivalently to groups whose parent nodes are assumed as exchangeable by the framework. Inference on the proposed model (Section 4.7) yields estimates of within-group variability required for the framework implementation. Additionally, parameter estimates are used to estimate the descriptors at the next level up. The model is fitted recursively from the level above the data level, up to the reality level, in order to obtain the required estimates. Algorithm 1 of Section 4.8 describes the steps in detail. Data sparseness is handled by exploiting the exchangeability assumptions among sibling nodes, as set out in Section 4.2. An approach similar to the empirical Bayes idea is adopted for inference, in order to avoid the complication of making prior judgements about model parameters. The method can be easily generalized to multilevel data as shown in the temperature application presented in Section 5.3.

Section 4.6 describes the proposed model for sharing information among groups of descriptors in a particular level, to estimate within-group variability in the extended framework.

4.6 Proposed model for sharing information

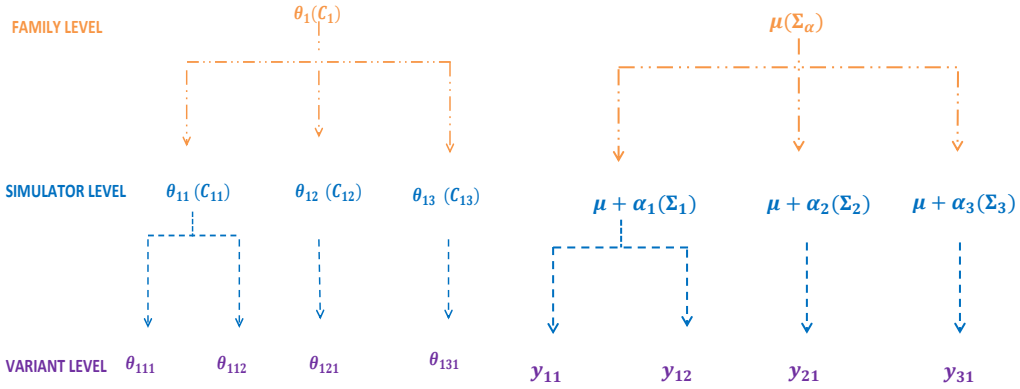
This section defines a model which allows estimation of within-group variability, for descriptor groups sharing common grand-parents, or equivalently groups whose parent nodes are assumed by the framework to be exchangeable, conditional on their own parents (see Section 4.2). The model deals with sparseness, by sharing information among groups, provided at least one group is not singleton. The model is fitted recursively to the framework levels, starting from the variant level and moving up to the discrepancy level (see Figure 4.1). It can accommodate an arbitrary number of levels expressing nested grouping between the data level and the discrepancy level.

Since the descriptors in the framework levels above the data level are unknown, they need to be estimated, in order to fit the proposed model to them. Descriptors at the variant level are estimated from their corresponding estimates at the data level. For the remaining levels, descriptor estimates are automatically produced by fitting the proposed model in the lower level, as described in Algorithm 1 of Section 4.8. In the remainder of this section however, the descriptors in each level of the framework are assumed as known, for simplicity.

The hierarchical data structure is such that the analysis can be performed recursively in the nested MME structure. Accordingly, this section concentrates on the variant level and considers fitting the proposed model, to a set of descriptor groups sharing common grand-parents. Algorithm 1 in Section 4.8 generalises the idea to all the descriptor groups, in all the levels between variant level and discrepancy level.

Concentrating on a single level, the structure becomes identical to the one-way layout illustrated in Figure 4.2.

This structure could be represented by fitting a standard one-way random effects model (Searle et al., 2006, Section 3.1) to the observations $\{\boldsymbol{y}_{ij},\ i = 1,\ldots,k,\ j = 1,\ldots,n_i\}$, where each observation $\boldsymbol{y}_{ij}$ is expressed as $\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij}$. The random exchangeable group effects $\{\boldsymbol{\alpha}_i,\ i = 1,\ldots,k\}$ are assumed to be i.i.d., distributed as $\boldsymbol{\alpha}_i \sim MVN\,(0, \boldsymbol{\Sigma}_\alpha)$. The residual errors $\{\boldsymbol{\epsilon}_{ij},\ i = 1,\ldots,k, j = 1,\ldots,n_i\}$ are assumed to be normally distributed, with mean zero and a common covariance matrix $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{\epsilon}_{ij} \sim MVN\,(0, \boldsymbol{\Sigma})$. The covariance matrix $\boldsymbol{\Sigma}$ in this case represents within-group variability, assumed to be equal among groups. However, under the proposed framework, it is unrealistic to assume that all groups have the same within-group variability. It is therefore reasonable to assign a group-specific within-group variability $\boldsymbol{\Sigma}_i$ for each group i. Figures 4.2a-4.2b illustrate the equivalence in framework and model terms, when the one-way random effects model with unequal within-group variances is fitted to the three groups of descriptors $\{\boldsymbol{\theta}_{111}, \boldsymbol{\theta}_{112}\}$, $\{\boldsymbol{\theta}_{121}\}$ and $\{\boldsymbol{\theta}_{131}\}$ in the variant level of Figure 4.1 (p.80), which share the common grand-parent descriptor $\boldsymbol{\theta}_1$ in the family level.



(a) Framework representation (b) Model representation

Figure 4.2: Framework and model representation when fitting the proposed model to the groups of descriptors $\{\boldsymbol{\theta}_{111}, \boldsymbol{\theta}_{112}\}$, $\{\boldsymbol{\theta}_{121}\}$ and $\{\boldsymbol{\theta}_{131}\}$ at the variant level.

It is evident from Figure 4.2 that the parameters $\{\boldsymbol{\Sigma}_i,\ i = 1,\ldots,3\}$ in the one-way model can be used as estimates for the unknown framework covariance matrices $\boldsymbol{C}_{11}, \boldsymbol{C}_{12}$ and $\boldsymbol{C}_{13}$ respectively. However, the presence of singleton groups creates challenges in estimating the corresponding within-group covariance matrices $\{\boldsymbol{\Sigma}_i,\ i = 2,3\}$, as also discussed in Section 4.4. To deal with this, it is common practice (see Section 4.5) to consider the within-group covariance matrices $\{\boldsymbol{\Sigma}_i,\ i = 1,\ldots,k\}$ as random and exchangeable. This assumption is justified for the proposed framework, since, as also illustrated in Figure 4.2, the corresponding covariance matrices

belong to nodes that are assumed to be exchangeable, conditional on their consensus (see Section 4.2). Exchangeability allows the $\{\boldsymbol{\Sigma}_i,\ i = 1, \ldots, k\}$ to be modelled as i.i.d., which allows sharing of information among groups and thus enables parameter estimation, even in the presence of singleton groups (provided there is at least one non-singleton group), as illustrated in Section 4.7. A conventional choice of distribution for covariance matrices of Gaussian data is the Inverse-Wishart distribution, which is also employed in the proposed model. The proposed one-way random effects model for multivariate data $\left\{\boldsymbol{y}_{ij},\ i = 1, \ldots, k, j = 1, \ldots, n_i\right\}$ is thus summarized in Equations (4.20)-(4.23):

$$\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij}\ (i = 1, \ldots, k,\ j = 1, \ldots, n_i), \tag{4.20}$$

$$\boldsymbol{\alpha}_i \overset{i.i.d.}{\sim} MVN\left(0, \boldsymbol{\Sigma}_\alpha\right)\ (i = 1, \ldots, k), \tag{4.21}$$

$$\boldsymbol{\epsilon}_{ij} \sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}_i\right)\ (i = 1, \ldots, k,\ j = 1, \ldots, n_i), \tag{4.22}$$

and

$$\boldsymbol{\Sigma}_i \overset{i.i.d.}{\sim} IW\left(\boldsymbol{R}, v\right). \tag{4.23}$$

The model terms $\left\{\boldsymbol{y}_{ij}, \boldsymbol{\mu}, \boldsymbol{\alpha}_i, \boldsymbol{\epsilon}_{ij},\ i = 1, \ldots, k, j = 1, \ldots, n_i\right\}$ are $p \times 1$ vectors, where p is the number of components of each data vector. The random effects $\{\boldsymbol{\alpha}_i,\ i = 1, \ldots, k\}$ are assumed to be mutually independent, and independent of the residual errors $\{\boldsymbol{\epsilon}_{ij},\ i = 1, \ldots, k, j = 1, \ldots, n_i\}$, as in the standard one-way random effects model (Snijders and Bosker, 2002, Section 4.3). The terms $\boldsymbol{\Sigma}_\alpha$ and $\boldsymbol{\Sigma}_i$ are $p \times p$ covariance matrices. The parameters $\boldsymbol{R}$ and v are the $p \times p$ scale matrix (restricted to be positive definite) and degrees of freedom respectively of the common distribution for the within-group covariance matrices $\{\boldsymbol{\Sigma}_i,\ i = 1, \ldots, k\}$. To facilitate parameter estimation, it is convenient to define a parametrisation for $\boldsymbol{\xi} := E\left(\boldsymbol{\Sigma}_i\right)$ (note the difference in notation between $\boldsymbol{\xi}$ here and $\boldsymbol{\tau}$ in (4.18)). Therefore,

$$\boldsymbol{\xi} := E\left(\boldsymbol{\Sigma}_i\right) = \frac{\boldsymbol{R}}{v - p - 1}, \tag{4.24}$$

based on the theoretical mean of the inverse-Wishart distribution (Gelman et al., 2014, p.576-577). Additionally, in order to perform parameter estimation, it is required to have $v > p + 3$ (see Appendix H). Estimation of model parameters is discussed next.

4.7 Model inference

4.7.1 Estimation (univariate data)

4.7.1.1 Random parameters

For simplicity, this section considers parameter estimation of the proposed model for the case when the data are scalar-valued, i.e. $p = 1$. In this case, the model equations (4.20)-(4.23) become:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \ (i = 1, \ldots, k, j = 1, \ldots, n_i), \tag{4.25}$$

$$\alpha_i \overset{i.i.d.}{\sim} N\left(0, \sigma_\alpha^2\right) \ (i = 1, \ldots, k), \tag{4.26}$$

$$\epsilon_{ij} \sim N\left(0, \sigma_i^2\right) \ (i = 1, \ldots, k, j = 1, \ldots n_i), \tag{4.27}$$

and

$$\sigma_i^2 \sim IG(\tilde{v}, t) \ (i = 1, \ldots, k), \tag{4.28}$$

where $\tilde{v} = {}^{v}/_{2}$ and $t = {}^{R}/_{2}$, based on the univariate analogue of the inverse-Wishart distribution defined in (4.23). The parameters $\tilde{v}$ and t in (4.28) correspond respectively to the shape and rate parameters of the inverse-Gamma distribution. To facilitate parameter estimation, it is convenient to define parametrisations for $\xi := E\left(\sigma_i^2\right)$ and $\psi := Var\left(\sigma_i^2\right)$, as follows:

$$\xi = \frac{t}{\tilde{v} - 1}, \tag{4.29}$$

and

$$\psi = \frac{t^2}{(\tilde{v} - 1)^2(\tilde{v} - 2)}, \tag{4.30}$$

based on the theoretical first and second moments of an inverse-Gamma distribution (Gelman et al., 2014, p.576-577).

A Bayesian approach to inference of the above model would regard (4.26) and (4.28) as priors for the unknown random parameters α_i and σ_i^2 respectively, for a particular group i. For reasons discussed in Section 4.5, any prior assumptions about the parameters σ_α^2, $\tilde{v}$ and t are avoided in inference of the proposed model. Alternatively, an approach similar to the idea of empirical Bayes is followed. Firstly,

Bayes' theorem is used in order to derive the distributions for the random parameters $\{\sigma_i^2,\ i = 1, \ldots, k\}$ and $\{\alpha_i,\ i = 1, \ldots, k\}$, given the data $\{y_{ij},\ i = 1, \ldots, k, j = 1, \ldots, n_i\}$. Then, the fixed parameters $\sigma_\alpha^2, \tilde{v}$ and t are estimated from the data, as shown in Section 4.7.1.2. For convenience, the fixed parameters are considered as known in the remaining of this section.

Regarding estimation of $\{\sigma_i^2,\ i = 1, \ldots, k\}$, it is natural to consider the distribution of σ_i^2 for a particular group i, conditional on the group observations $\{y_{ij},\ j = 1, \ldots, n_i\}$. Denoting the observations in group i as $\boldsymbol{y}_i$, Bayes' theorem yields,

$$\pi\left(\sigma_i^2|\boldsymbol{y}_i\right) \propto \pi\left(\sigma_i^2\right)\pi\left(\boldsymbol{y}_i|\sigma_i^2\right),$$

where $\pi\left(\sigma_i^2\right)$ is the density of the assumed Inverse-Gamma distribution in (4.28) and from (4.25)-(4.27), $\pi\left(\boldsymbol{y}_i|\sigma_i^2\right)$ is the density of a multivariate normal distribution with mean vector $(\mu \ldots \mu)'$ and covariance matrix

$$\begin{pmatrix} \sigma_\alpha^2 + \sigma_i^2 & \sigma_\alpha^2 & \ldots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \ddots & \ldots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \ldots & \ldots & \sigma_\alpha^2 + \sigma_i^2 \end{pmatrix}.$$

Ideally, the resulting distribution for $\sigma_i^2|\boldsymbol{y}_i$ would be of a known form, allowing σ_i^2 to be estimated as, for example $E\left(\sigma_i^2|\boldsymbol{y}_i\right)$. However, it turns out that the presence of σ_α^2 in the expression does not allow an analytical expression for $E\left(\sigma_i^2|\boldsymbol{y}_i\right)$.

On the other hand, conditioning on both $\boldsymbol{y}_i$ and α_i yields an analytical expression for $\pi\left(\sigma_i^2|\boldsymbol{y}_i, \alpha_i\right)$. Using Bayes' theorem,

$$\pi\left(\sigma_i^2|\boldsymbol{y}_i, \alpha_i\right) \propto \pi\left(\sigma_i^2\right)\pi\left(\boldsymbol{y}_i|\sigma_i^2, \alpha_i\right). \tag{4.31}$$

The term $\pi\left(\boldsymbol{y}_i|\sigma_i^2, \alpha_i\right)$ in the RHS of (4.31) is the product of joint univariate normal densities, since according to the proposed model,

$$y_{ij}|\sigma_i^2, \alpha_i \sim N\left(\mu + \alpha_i, \sigma_i^2\right)\ (i = 1, \ldots, k, j = 1, \ldots, n_i). \tag{4.32}$$

Thus,

$$\pi\left(\boldsymbol{y}_i|\sigma_i^2, \alpha_i\right) = \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{1}{2\sigma_i^2}\left(y_{ij} - \mu - \alpha_i\right)^2\right\}.$$

Multiplying this with the Inverse-Gamma density of $\pi\left(\sigma_i^2\right)$ gives,

$$\pi\left(\sigma_i^2|\boldsymbol{y}_i,\alpha_i\right) \propto \frac{t^{\tilde{v}}}{\Gamma(\tilde{v})}\left(\sigma_i^2\right)^{-(\tilde{v}+1)}\exp\left(-\frac{t}{\sigma_i^2}\right)\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left\{-\frac{1}{2\sigma_i^2}\left(y_{ij}-\mu-\alpha_i\right)^2\right\}$$
$$\propto \left(\sigma_i^2\right)^{-\frac{n_i}{2}-\tilde{v}-1}\exp\left(-\frac{1}{\sigma_i^2}\left(t+\frac{\sum_{j=1}^{n_i}\left(y_{ij}-\mu-\alpha_i\right)^2}{2}\right)\right). \tag{4.33}$$

The form of (4.33) is proportional to the PDF of an Inverse-Gamma distribution for $\sigma_i^2|\boldsymbol{y}_i,\alpha_i$:

$$\sigma_i^2|\boldsymbol{y}_i,\alpha_i \sim IG(a,b), \tag{4.34}$$

where:

$$a = \frac{n_i}{2}+\tilde{v}, \tag{4.35}$$

and

$$b = t+\frac{\sum_{j=1}^{n_i}\left(y_{ij}-\mu-\alpha_i\right)^2}{2}. \tag{4.36}$$

Equation (4.34) illustrates that conditioning on α_i yields a distribution of known form for $\sigma_i^2|\boldsymbol{y}_i,\alpha_i$. Consequently, if the shape and rate parameters a and b of the distribution were known, an estimate for the within-group variability σ_i^2 could be defined to be the expectation of $\sigma_i^2|\boldsymbol{y}_i,\alpha_i$.

The parameters a and b are unknown, since they involve the unknown fixed quantities $\mu, t, \tilde{v}$ and the random component α_i, induced from conditioning. Estimation of the fixed parameters μ, t and $\tilde{v}$ is considered in Section 4.7.1.2. One way to estimate the random parameter α_i, could be to derive an analytical form for its distribution, conditional solely on the group data $\boldsymbol{y}_i$, as was initially attempted for σ_i^2. This would require integrating out σ_i^2 from the posterior $\pi\left(\sigma_i^2,\alpha_i|\boldsymbol{y}_i\right)$. However, it turns out that the resulting integral cannot be solved analytically, which suggests that obtaining an analytic expression for $\pi\left(\alpha_i|\boldsymbol{y}_i\right)$ is not feasible.

On the other hand, in a similar vein to considerations for estimating σ_i^2, conditioning on both $\boldsymbol{y}_i$ and σ_i^2 yields a distribution of known form for $\alpha_i|\boldsymbol{y}_i,\sigma_i^2$. Bayes' theorem suggests that,

$$\pi\left(\alpha_i|\boldsymbol{y}_i,\sigma_i^2\right) \propto \pi\left(\alpha_i\right)\pi\left(\boldsymbol{y}_i|\sigma_i^2,\alpha_i\right).$$

With $y_{ij}|\sigma_i^2,\alpha_i$ and α_i distributed according to (4.32) and (4.26) respectively,

$$\pi\left(\alpha_i|\boldsymbol{y}_i,\sigma_i^2\right) \propto \frac{1}{\sqrt{2\pi\sigma_\alpha^2}}\exp\left(-\frac{1}{2\sigma_\alpha^2}\alpha_i^2\right)\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left\{-\frac{1}{2\sigma_i^2}\left(y_{ij}-\mu-\alpha_i\right)^2\right\}.$$

After some straightforward manipulations and removing the terms not proportional to α_i, it turns out that,

$$\pi\left(\alpha_i|\boldsymbol{y}_i,\sigma_i^2\right) \propto \exp\left(-\frac{1}{2}\left(\sigma_\alpha^{-2}+n_i\sigma_i^{-2}\right)\left(\alpha_i-\frac{\left(\sum_{j=1}^{n_i}y_{ij}-n_i\mu\right)\sigma_i^{-2}}{\sigma_\alpha^{-2}+n_i\sigma_i^{-2}}\right)^2\right). \quad (4.37)$$

The form of (4.37) is proportional to the PDF of a Normal distribution for $\alpha_i|\boldsymbol{y}_i,\sigma_i^2$, as follows:

$$\alpha_i|\boldsymbol{y}_i,\sigma_i^2 \sim N\left(r,s\right), \quad (4.38)$$

where:

$$r=\frac{n_i\sigma_\alpha^2\left(\bar{y}_{i\cdot}-\mu\right)}{\sigma_i^2+n_i\sigma_\alpha^2}, \quad (4.39)$$

and

$$s=\frac{\sigma_i^2\sigma_\alpha^2}{\sigma_i^2+n_i\sigma_\alpha^2}. \quad (4.40)$$

The term $\bar{y}_{i\cdot}$ in (4.39) denotes the sample mean of group observations collected in $\boldsymbol{y}_i$, i.e.,

$$\bar{y}_{i\cdot}=\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij}.$$

To summarize, so far it is shown that there exist analytical expressions for the full conditional distributions of $\sigma_i^2|\boldsymbol{y}_i,\alpha_i$ and $\alpha_i|\boldsymbol{y}_i,\sigma_i^2$, as shown in (4.34) and (4.38) respectively. Given both the full conditional distributions in (4.34) and (4.38), samples of σ_i^2 and α_i from the joint distribution $\pi\left(\sigma_i^2,\alpha_i|\boldsymbol{y}_i\right)$ can be obtained straightforwardly using the Gibbs sampler; given a large enough number of samples for σ_i^2 and α_i, the corresponding sample means (after burn-in) can be used to derive $E\left(\sigma_i^2|\boldsymbol{y}_i\right)$ and $E\left(\alpha_i|\boldsymbol{y}_i\right)$, and therefore estimate σ_i^2 and α_i respectively.

4.7.1.2 Fixed parameters

The development so far has assumed that the parameters μ, $\tilde{v}$, t and σ_α^2 are known. In practice however, they are unknown and must therefore be estimated. This is considered next.

It is common practice (Searle, 1987, Section 2.3) to estimate μ in random effects models using least-squares estimation. The least squares estimate of μ, denoted by $\hat{\mu}$, is:

$$\hat{\mu} = \bar{y}_{..},$$

where $\bar{y}_{..}$ denotes the overall sample mean of the group observations $\{y_{ij},\ i = 1, \ldots, k, j = 1, \ldots, n_i\}$, i.e:

$$\bar{y}_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij},$$

and $N = \sum_{i=1}^{k} n_i$, i.e. the total number of observations.

Estimation is performed using the method of moments, similarly to standard ANOVA estimation (Searle et al., 2006, Section 3.6). The method of moments is an appealing means of estimation, because of its simple implementation and computational convenience. It requires choosing a set of data sums of squares such that their theoretical expected values involve the parameters of interest. The expressions for the resulting expected values are then set equal to their empirical values, i.e. the chosen sums of squares, which are functions of the data only. The resulting linear system of equations is then solved in terms of the unknown parameters, in order to obtain their estimates. Although the method of moments is computationally convenient since it provides analytical estimates of the unknown parameters, it also has some limitations (Searle et al., 2006, Section 2.3b). Firstly, it can produce negative variance estimates. Secondly, in the case of unbalanced data, i.e. groups of data with unequal sample size, like the situation in the proposed framework, the sum of squares decomposition is not unique. This implies that the property of uniformly best unbiasedeness of ANOVA estimates, which holds for balanced data, is lost in the unbalanced case. The issues are further discussed in Section 4.9.

For estimating $\tilde{v}$ and t, the parameters ξ and ψ (defined in (4.29) and (4.30) respectively) are estimated first. Then, $\tilde{v}$ and t are estimated based on (4.41)-(4.42) below:

$$\tilde{v} = \frac{\xi^3 + 2\xi\psi}{\xi\psi}, \tag{4.41}$$

and

$$t = \frac{\xi^3 + \xi\psi}{\psi}, \tag{4.42}$$

which are obtained by rearranging (4.29) and (4.30).

By analogy with standard ANOVA practice (Searle et al., 2006, Section 2.1), in the first instance it is natural to consider the within- and between-groups sums of squares as candidate functions for the estimation process. Writing each observation y_{ij} as a sum of deviations:

$$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.}), \ (i = 1, \ldots, k, j = 1, \ldots, n_i),$$

allows expressing total variation of observations about the overall mean as a sum of squares:

$$\begin{aligned} \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 &= \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2, \\ &= S_G + S_E, \text{ say.} \end{aligned} \tag{4.43}$$

The RHS of (4.43) partitions variability of observations around the overall mean into two sources of variation. The first term expresses between-group variability and the second term the within-group variability. The next step is to derive $E(S_G)$ and $E(S_E)$ under the model defined in (4.25)-(4.28).

Based on (4.25), the expressions $\bar{y}_{i.}$ and $\bar{y}_{..}$ in (4.43) can be written as

$$\bar{y}_{i.} = \mu + \alpha_i + \bar{\epsilon}_{i.}, \ (i = 1, \ldots, k), \tag{4.44}$$

where $\bar{\epsilon}_{i.} = n_i^{-1} \sum_{j=1}^{n_i} \epsilon_{ij}$ is the mean of the residuals in group i.

Additionally,

$$\bar{y}_{..} = \mu + \bar{\alpha} + \bar{\epsilon}_{..}, \tag{4.45}$$

where $\bar{\alpha} = k^{-1} \sum_{i=1}^{k} \alpha_i$ is the sample mean of the random effects $\{\alpha_i, \ i = 1, \ldots, k\}$ and $\bar{\epsilon}_{..} = N^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \epsilon_{ij}$ is the sample mean of the residuals $\{\epsilon_{ij}, \ i = 1, \ldots, k, j = 1, \ldots, n_i\}$.

The remaining steps of the derivations are shown in Appendix G. The resulting expressions are:

$$E(S_G) = \frac{(k-1)}{k} \left(N\sigma_\alpha^2 + k\xi\right), \tag{4.46}$$

and

$$\begin{aligned} E\left(S_E\right) &= \sum_{i=1}^{k}(n_i-1)\xi \\ &= (N-k)\xi. \end{aligned} \tag{4.47}$$

It is required to estimated three unknown parameters in total, namely σ_α^2, ξ and ψ. However, up to this point there exist only two equations (4.46) and (4.47), neither of which involves the unknown ψ. A third function of the data is thus required, similarly to S_G and S_E (defined in (4.43)), whose expected value involves ψ. The resulting expectation can be used in conjunction with (4.46) and (4.47), to form a linear system of three equations with three unknowns σ_α^2, ξ and ψ. The method of moments then suggests equating the three expectations with their observed values to derive the corresponding estimators $\hat{\sigma}_\alpha^2$, $\hat{\xi}$ and $\hat{\psi}$ respectively.

The additional function of the data is chosen based on the interpretation of ψ as the variance of within-group variability. This suggests that it is reasonable to consider a function similar to $S_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2$, as the latter expresses within-group variability. Since $E\left(S_E\right)$ involves ξ, the first moment of σ_i^2, for an expression whose expectation involves ψ, the second moment of σ_i^2, it is reasonable to consider raising the sum of squared deviations $\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2$ in S_E to a higher power. An intuitive choice is therefore $\sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2\right]^2$.

Next, it is shown that the expectation of the chosen function involves ψ and therefore allows for its estimation.

Precisely,

$$\begin{aligned} E\left(\sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2\right]^2\right) &= E\left(\sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}\left(\epsilon_{ij}-\bar{\epsilon}_{i.}\right)^2\right]^2\right), \text{ using (4.25) and (4.44)} \\ &= E\left(\sum_{i=1}^{k}\left[\left(n_i-1\right)s_i^2\right]^2\right) \\ &= \sum_{i=1}^{k}E\left[\left(n_i-1\right)^2 s_i^4\right], \\ &= \sum_{i=1}^{k}\left(n_i-1\right)^2 E\left(s_i^4\right), \end{aligned} \tag{4.48}$$

where,

$$\begin{aligned} E\left(s_i^4\right) &= E\left[E\left(s_i^4|\sigma_i^2\right)\right], \text{ using the iterated expectations formula} \\ &= E\left[Var\left(s_i^2|\sigma_i^2\right) + E^2\left(s_i^2|\sigma_i^2\right)\right]. \end{aligned} \tag{4.49}$$

From the standard results on the distribution of the sample variance for Gaussian observations (Rice, 2007, p.197), it holds that

$$E\left(s_i^2|\sigma_i^2\right) = \sigma_i^2,$$

and

$$Var\left(s_i^2|\sigma_i^2\right) = \frac{2\sigma_i^4}{n_i - 1}.$$

Thus

$$\begin{aligned} E\left(s_i^4\right) &= E\left(\frac{2\sigma_i^4}{n_i - 1} + \sigma_i^4\right) \\ &= \frac{n_i + 1}{n_i - 1} E\left(\sigma_i^4\right) \\ &= \frac{n_i + 1}{n_i - 1}\left(Var\left(\sigma_i^2\right) + E^2\left(\sigma_i^2\right)\right) \\ &= \frac{n_i + 1}{n_i - 1}\left(\psi + \xi^2\right), \text{ as defined in (4.29) and (4.30).} \end{aligned} \tag{4.50}$$

Equation (4.48) now yields

$$\begin{aligned} E\left(\sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2\right]^2\right) &= \left(\psi + \xi^2\right)\sum_{i=1}^{k}\left(n_i^2 - 1\right) \\ &= \left(\psi + \xi^2\right)\left(\sum_{i=1}^{k} n_i^2 - k\right), \end{aligned} \tag{4.51}$$

which indeed involves the parameter ψ of interest.

Consequently, equation (4.51), together with (4.46) and (4.47) form the system of three equations needed for deriving the estimators $\hat{\sigma}_\alpha^2$, $\hat{\xi}$ and $\hat{\psi}$, using the method of moments. Equating the theoretical expectations to their empirical values, the system of equations summarised in (4.52)-(4.54) is obtained:

$$\sum_{i=1}^{k} n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 = \frac{(k-1)}{k}\left(N\hat{\sigma}_\alpha^2 + k\hat{\xi}\right), \tag{4.52}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.}\right)^2 = (N-k)\hat{\xi}, \tag{4.53}$$

and

$$\sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2\right]^2 = \left(\hat{\psi} + \hat{\xi}^2\right)\left(\sum_{i=1}^{k} n_i^2 - k\right). \tag{4.54}$$

Solving the linear system in (4.52)-(4.54) in terms of σ_α^2, ξ and ψ yields the estimators,

$$\hat{\sigma}_\alpha^2 = \frac{k(N-k)\sum_{i=1}^{k} n_i \left(\bar{y}_{i.} - y_{..}\right)^2 - k(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2}{N(N-k)(k-1)}, \tag{4.55}$$

$$\hat{\xi} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2}{N-k},$$

and

$$\hat{\psi} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^4}{\sum_{i=1}^{k} n_i^2 - k} - \left(\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2}{N-k}\right)^2.$$

To summarise, this section illustrated a method to obtain estimates for the parameters σ_α^2, ξ and ψ involved in the full conditional distributions of $\sigma_i^2|y_{ij}, \alpha_i$ and $\alpha_i|y_{ij}, \sigma_i^2$ defined in (4.34)-(4.36) and (4.38)-(4.40) respectively. The resulting estimates of ξ and ψ can be plugged in to (4.41)-(4.42), in order to obtain estimates of $\tilde{v}$ and t. The resulting estimates, together with the estimate of σ_α^2 shown in (4.55) are plugged in the corresponding terms in the full conditional distributions of $\sigma_i^2|y_{ij}, \alpha_i$ and $\alpha_i|y_{ij}, \sigma_i^2$. Then, Gibbs sampling is performed in order to numerically evaluate estimates of $\{\sigma_i^2,\ i = 1, \ldots, k\}$ and $\{\alpha_i,\ i = 1, \ldots, k\}$, as discussed in Section 4.7.1.1.

4.7.2 Multivariate data

The estimates in the previous sections are derived from scalar data values $\{y_{ij},\ i = 1, \ldots, k, j = 1, \ldots, n_i\}$. However, the data are multivariate in practice, since they correspond to multivariate descriptor estimates, as discussed in Section 4.6. This section summarizes the steps for the estimation process in case of multivariate data.

Details about the estimation process are provided in Appendix H.

According to Section 4.6, it is of interest to estimate the unknown parameters $\left\{\boldsymbol{\Sigma}_i^2,\ i=1,\ldots,k, j=1,\ldots,n_i\right\}$ and $\{\boldsymbol{\mu}+\boldsymbol{\alpha}_i,\ i=1,\ldots,k, j=1,\ldots,n_i\}$ in order to implement the framework of Section 4.2. The estimation process follows that described in Section 4.7.1 for estimating $\{\sigma_i^2,\ i=1,\ldots,k\}$ and $\{\mu+\alpha_i,\ i=1,\ldots,k\}$ in the scalar case.

The first step is to derive the full conditional distributions $\boldsymbol{\Sigma}_i|\boldsymbol{y}_i,\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_i|\boldsymbol{y}_i,\boldsymbol{\Sigma}_i$, for a particular group i $(i=1,\ldots,k)$. They are derived based on the assumptions of the model in (4.20)-(4.23) of Section 4.6 and Bayes' theorem, analogously to the procedure described in Section 4.7.1.1. The resulting full conditional distributions are:

$$\boldsymbol{\Sigma}_i|\boldsymbol{y}_i,\boldsymbol{\alpha}_i \sim IW\left(\boldsymbol{R}+\boldsymbol{S}, v+n_i\right), \tag{4.56}$$

where

$$\boldsymbol{S}=\sum_{j=1}^{n_i}\left(\boldsymbol{y}_{ij}-(\boldsymbol{\mu}+\boldsymbol{\alpha}_i)\right)\left(\boldsymbol{y}_{ij}-(\boldsymbol{\mu}+\boldsymbol{\alpha}_i)\right)',$$

and

$$\boldsymbol{\alpha}_i|\boldsymbol{y}_i,\boldsymbol{\Sigma}_i \sim MVN\left(\boldsymbol{\mu}_\alpha,(\boldsymbol{T}_\alpha+n_i\boldsymbol{T}_i)^{-1}\right), \tag{4.57}$$

where $\boldsymbol{T}_\alpha=\boldsymbol{\Sigma}_\alpha^{-1}$, $\boldsymbol{T}_i=\boldsymbol{\Sigma}_i^{-1}$ and $\boldsymbol{\mu}_\alpha=(\boldsymbol{T}_\alpha+n_i\boldsymbol{T}_i)^{-1}\boldsymbol{T}_i\sum_{j=1}^{n_i}\left(\boldsymbol{y}_{ij}-\boldsymbol{\mu}\right)$.

Therefore, given estimates of the unknown parameters in the distributions in (4.56) and (4.57), $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\alpha}_i$ can be estimated by sampling approximately from the joint distribution $\pi\left(\boldsymbol{\Sigma}_i,\boldsymbol{\alpha}_i|\boldsymbol{y}_i\right)$ using Gibbs sampler on the pair of the full-conditional distributions (see discussion at the end of Section 4.7.1.1). The next step is therefore to estimate the fixed unknown parameters $\boldsymbol{\mu}$, $\boldsymbol{R}$, v and $\boldsymbol{T}_\alpha$ involved in the expressions for the full-conditional distributions.

As usual, $\hat{\boldsymbol{\mu}}$ is defined to be the least squares estimate of $\boldsymbol{\mu}$ in the random effects model defined in (4.20)-(4.23) of Section 4.6. This is equal to the overall mean of observations $\left\{\boldsymbol{y}_{ij},\ i=1,\ldots,k, j=1,\ldots,n_i\right\}$, denoted as $\bar{\boldsymbol{y}}_{..}$, i.e.

$$\hat{\boldsymbol{\mu}}=\bar{\boldsymbol{y}}_{..}=\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\boldsymbol{y}_{ij}.$$

By generalising from the univariate case ((4.43), p.107), the between- and within-groups sums of squares for multivariate data are defined as,

$$\boldsymbol{S}_G = \sum_{i=1}^{k} n_i \left(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..}\right)\left(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..}\right)',$$

and

$$\boldsymbol{S}_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.}\right)\left(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.}\right)',$$

respectively, where $\bar{\boldsymbol{y}}_{i.} = {}^{1}/_{n_i} \sum_{j=1}^{n_i} \boldsymbol{y}_{ij}$.

Then, by generalising the expressions in (4.46) and (4.47), their expectations become,

$$E\left(\boldsymbol{S}_G\right) = \frac{(k-1)}{k}\left(N\boldsymbol{\Sigma}_\alpha + k\boldsymbol{\xi}\right),$$

and

$$E\left(\boldsymbol{S}_E\right) = (N-k)\boldsymbol{\xi},$$

with $\boldsymbol{\xi}$ as defined in (4.24).

Method of moments suggests equating the above expectations to their empirical values $\boldsymbol{S}_G$ and $\boldsymbol{S}_E$ and solving the resulting linear system in terms of the unknown parameters $\boldsymbol{\Sigma}_\alpha$ and $\boldsymbol{\xi}$, in order to obtain their estimates $\hat{\boldsymbol{\Sigma}}_\alpha$ and $\hat{\boldsymbol{\xi}}$ respectively. This gives,

$$\hat{\boldsymbol{\Sigma}}_\alpha = \frac{k(N-k)\sum_{i=1}^{k} n_i \left(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..}\right)\left(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..}\right)' - k(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.}\right)\left(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.}\right)'}{N(N-k)(k-1)}, \tag{4.58}$$

and

$$\hat{\boldsymbol{\xi}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.}\right)\left(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.}\right)'}{N-k}. \tag{4.59}$$

The expression for $\hat{\boldsymbol{\Sigma}}_\alpha$ in (4.58) suggests that depending on the available data $\{\boldsymbol{y}_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots n_i\}$, $\hat{\boldsymbol{\Sigma}}_\alpha$ might not be positive-definite, as required. This can be caused for example when $\hat{\boldsymbol{\Sigma}}_\alpha$ is rank deficient, i.e. when $k < p$. In this cases, $\hat{\boldsymbol{\Sigma}}_\alpha$ is substituted by its nearest positive definite matrix, which is obtained by implementing the algorithm of Higham (2002) (using the command *nearPD* from the *Matrix* package in R (R Core Team, 2012)).

The remaining parameters to estimate are $\boldsymbol{R}$ and v. Equation (4.24) implies that given an estimate $\hat{v}$ of v, an estimate $\hat{\boldsymbol{R}}$ can also be obtained ($\hat{\boldsymbol{R}} = (\hat{v} -$

$p-1)\hat{\boldsymbol{\xi}}$). In order to estimate v using method of moments, it is required to derive a function of the data whose expectation involves v. The analogous expression to $\sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2\right]^2$ considered in the scalar case (see (4.48), p. 108), becomes $\sum_{i=1}^{k}(n_i-1)^2\boldsymbol{S}_i\boldsymbol{S}_i'$ for multivariate data, where $\boldsymbol{S}_i=\frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(\boldsymbol{y}_{ij}-\bar{\boldsymbol{y}}_{i.}\right)\left(\boldsymbol{y}_{ij}-\bar{\boldsymbol{y}}_{i.}\right)'$. Since v is a scalar quantity, the trace of $\sum_{i=1}^{k}(n_i-1)^2\boldsymbol{S}_i\boldsymbol{S}_i'$ is considered. Analogously to (4.51) (Section 4.7.1.2) in the scalar case, $E\left[tr\left(\sum_{i=1}^{k}(n_i-1)^2\boldsymbol{S}_i\boldsymbol{S}_i'\right)\right]$ is expected to involve v, in which case the latter can be estimated using method of moments. It is proved in (H.13) of Appendix H that this is the case, which then allows setting the empirical value $tr\left(\sum_{i=1}^{k}(n_i-1)^2\boldsymbol{S}_i\boldsymbol{S}_i'\right)$ equal to the theoretical expression for $E\left[tr\left(\sum_{i=1}^{k}(n_i-1)^2\boldsymbol{S}_i\boldsymbol{S}_i'\right)\right]$. This leads to a quadratic equation for v, for which at most one root satisfies the required constraint $v>p+3$ (see also Appendix I). Once $\hat{v}$ is obtained, $\hat{\boldsymbol{R}}$ is calculated based on (4.24).

Having obtained estimates of $\boldsymbol{\mu}$, $\boldsymbol{R}$, v and $\boldsymbol{T}_\alpha$, Gibbs sampler is applied to the full-conditional distributions $\boldsymbol{\Sigma}_i|\boldsymbol{y}_i,\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_i|\boldsymbol{y}_i,\boldsymbol{\Sigma}_i$ (shown in (4.56) and (4.57) respectively), in order to estimate $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\alpha}_i$ for each group i ($i=1,\ldots,k$). The resulting estimates are then used for the implementation of the extended framework, as described in Section 4.8. The next section considers some cases of data structures which require special treatment when the model of Section 4.6 is fitted to them.

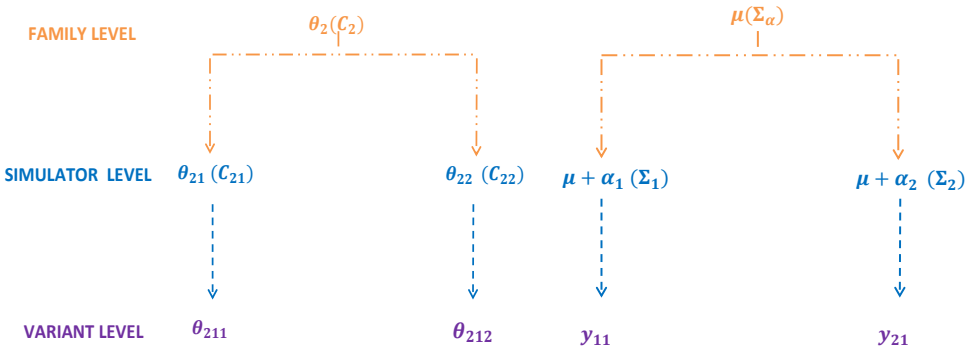
4.7.3 Special cases

This section considers some particular types of nesting structures for which model inference cannot be performed as described in Section 4.7.2. These special cases arise when fitting the model to descriptor groups which are all singleton, or to a single group with multiple observations, or a single observation. In the presence of only singleton groups, sharing of information cannot be performed, since singleton groups provide no information about the corresponding group effect and within-group variability. On the other hand, the presence of a single group only, implies that there is no other group to share information with, as suggested by the model of Section 4.6. Handling of each special case is considered individually. For illustration, examples from the variant level of Figure 4.1 (p.80) are considered.

- **Special Case 1:** Multiple groups, all singleton

  Consider fitting the model of Section 4.6 to the descriptors $\{\boldsymbol{\theta}_{211}\}$ in the variant level of Figure 4.1, since they share the common grand-parent descriptor $\boldsymbol{\theta}_2$ in the family level. The descriptors are considered to belong to two different singleton groups in the model, since they are not sibling descriptors, i.e. they

don't have a common parent descriptor. The structure (in model and framework terms) is illustrated in Figures 4.3a-4.3b:



(a) Framework representation (b) Model representation

Figure 4.3: Framework and model representation when fitting the proposed model to the singleton groups of descriptors $\{\boldsymbol{\theta}_{211}\}$ and $\{\boldsymbol{\theta}_{212}\}$ at the variant level.

In model terms, there exist two singleton groups, with data vectors $\boldsymbol{y}_{11}$ and $\boldsymbol{y}_{21}$ respectively. This implies that $N = \sum_{i=1}^{2} n_i = k$, where k is the number of groups, and $n_1 = n_2 = 1$, following the notation introduced in Section 4.6. As shown in (4.58)-(4.59) of Section 4.7.2, $\hat{\boldsymbol{\Sigma}}_\alpha$ and $\hat{\boldsymbol{\xi}}$ are undefined for $N = k$ and therefore $\{\boldsymbol{\Sigma}_i,\ i = 1, \ldots, k\}$ and $\{\boldsymbol{\alpha}_i,\ i = 1, \ldots, k\}$ cannot be estimated as described in Section 4.7.2. Besides, as also discussed, sharing of information among groups is meaningless, unless at least one group has more than 1 observations. In this case, the parameters $\left\{\hat{\boldsymbol{\Sigma}}_i,\ i = 1, \ldots, k\right\}$ and $\{\hat{\boldsymbol{\alpha}}_i,\ i = 1, \ldots, k\}$ are set to zero.

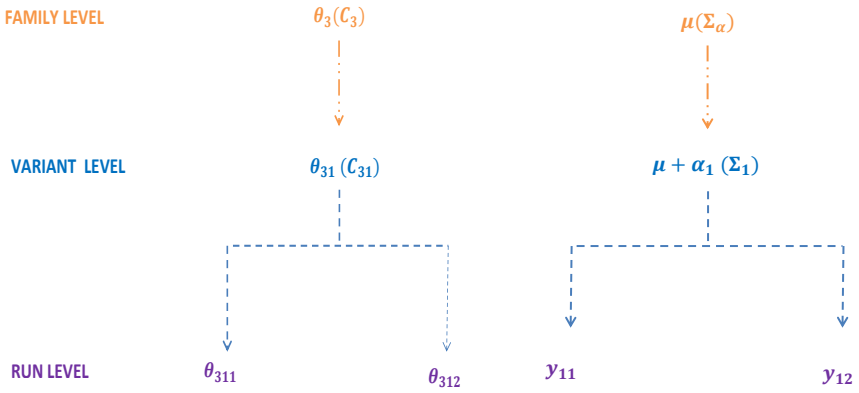
- **Special Case 2:** Single group

  It is clear that in the presence of a single group only, there are not enough groups to share information with. Therefore, it is expected that when fitting the model of Section 4.6 with k, the number of groups, being equal to 1, model inference considered in Section 4.7.2 would fail to provide the required parameter estimates. In fact, $k = 1$ implies that $\hat{\boldsymbol{\Sigma}}_\alpha$ in (4.58) becomes $0/0$, i.e. undefined. Therefore, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\alpha}_1$ cannot be estimated as described in Section 4.7.2. The issue of estimation is treated differently, depending on whether the underlying group is singleton ($n_1 = 1$) or not ($n_1 > 1$), as shown below:

  - **Special Case 2a:** Multiple observations

    Consider fitting the model to the group of sibling descriptors $\{\boldsymbol{\theta}_{311}, \boldsymbol{\theta}_{312}\}$ in the variant level of Figure 4.1 (Section 4.2), which share the grandparent descriptor $\boldsymbol{\theta}_3$ in the family level. Since there is no other group in

the variant level sharing the same grand-parent descriptor, the model is fitted to the single group $\{\boldsymbol{\theta}_{311}, \boldsymbol{\theta}_{312}\}$ only, as shown in Figures 4.4a-4.4b:



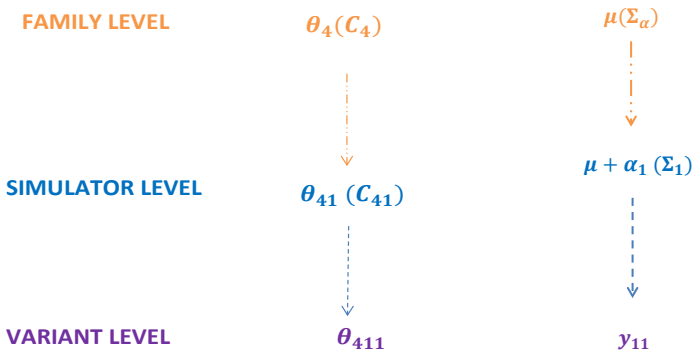
(a) Framework representation (b) Model representation

Figure 4.4: Framework and model representation when fitting the proposed model to the group of descriptors $\{\boldsymbol{\theta}_{311}, \boldsymbol{\theta}_{312}\}$ at the variant level.

The presence of multiple observations allows using the sample covariance matrix to estimate $\boldsymbol{\Sigma}_1$ in the proposed model of Section 4.6. Also, in the presence of a single group, the group effect $\boldsymbol{\alpha}_1$ is set to zero, since there are no other groups in order for group effects to be estimated.

– **Special Case 2b:** Single observation

Consider finally fitting the model to the descriptor $\boldsymbol{\theta}_{411}$ in the variant level of Figure 4.1. The descriptor has no other sibling descriptors and additionally, there are no other groups sharing the same grand-parent descriptor $(\boldsymbol{\theta}_4)$ with $\boldsymbol{\theta}_{411}$. In this case, the model is fitted to a single group, with a single observation, as shown in Figures 4.5a-4.5b:



(a) Framework representation (b) Model representation

Figure 4.5: Framework and model representation when fitting the proposed model to the descriptor $\boldsymbol{\theta}_{411}$ at the variant level.

In the presence of a single group with sample size $n_1 = 1$, the sample covariance matrix is undefined and therefore cannot be used to estimate $\boldsymbol{\Sigma}_1$ in the model of Section 4.6. The estimate is therefore set to be zero. Similarly, $\boldsymbol{\alpha}$ is estimated to be zero.

To summarize, it is worth noting that the proposed estimators of within-group variability for the special cases considered in this section can be considered as being slightly naive, in the sense that they have very little statistical foundation. This is because of the absence of enough data which could be informative in estimation of within-group variability based on the model of Section 4.6 or other standard statistical approaches to variance component estimation. As such, they are inevitably based on heuristic assumptions and therefore it is important to explore their performance. This is left as future work.

4.8 Implementation algorithm

Once the parameters of the proposed random effects model (Section 4.6) are estimated as described in Section 4.7, they are incorporated in the implementation of the extended framework, to estimate the relevant framework covariance matrices. Algorithm 1 illustrates how this is achieved by applying the random effects model recursively in the framework levels.

In each level, the proposed random effects model is fitted repeatedly to groups of descriptor estimates sharing a common grand-parent descriptor. This process is repeated recursively in moving from level 1 to level $D-2$ (following to the notation in Algorithm 1), to enable successive estimation of the framework's covariance matrices at each level. Each time the model is fitted, the resulting estimates $\{\hat{\boldsymbol{\Sigma}}_i, i = 1, \ldots, k\}$ are used to estimate the corresponding framework covariance matrices $\{\hat{\boldsymbol{C}}^{(i)}_{parent}, i = 1, \ldots, k\}$ in the next level up. Additionally, the parameter estimates $\{\boldsymbol{\mu} + \boldsymbol{\alpha}_i, i = 1, \ldots, k\}$ are used to estimate the framework descriptors in the next level up, which is necessary in order to fit the model recursively. When the model is fitted in level 1, the estimates $\{\hat{\boldsymbol{\theta}}^{(l)}, l = 1, \ldots, N\}$ in level 0, obtained by fitting the mimic to the GCM outputs, are used to estimate the corresponding descriptors at level 1, to which the model is fitted. When the model is fitted to descriptor estimates at level $D-2$, the resulting estimate $\boldsymbol{\Sigma}_\alpha$ is used to estimate the framework's covariance matrix $\boldsymbol{C}$.

Apart from the unknown covariance matrices which are estimated by fitting the proposed random effects model recursively in the framework, in order to obtain the posterior of $\boldsymbol{\theta}_0$, it is also required to: Estimate the unknown covariance matrices $\{\hat{\boldsymbol{J}}^{(l)}, l = 1, \ldots, N\}$ and $\boldsymbol{\Lambda}$, calculate the descriptor estimates $\hat{\boldsymbol{\theta}}_0$ and $\{\hat{\boldsymbol{\theta}}^{(l)}, l =$

Algorithm 1 Framework implementation

Additional notation

Level 0: Data level
Level D: Discrepancy level
$\boldsymbol{\theta}^{(i)}_{parent}$: Parent descriptor of descriptor group i $(i = 1, \ldots, k)$
$\hat{\boldsymbol{\theta}}^{(i)}_{parent}$: Estimate of $\boldsymbol{\theta}^{(i)}_{parent}$ $(i = 1, \ldots, k)$
$\underset{\sim}{\boldsymbol{\Theta}}_l$: Set of descriptors at level l $(l = 1, \ldots, D)$
$\hat{\underset{\sim}{\boldsymbol{\Theta}}}_l$: Set of descriptor estimates at level l $(l = 0, \ldots, D)$
$\boldsymbol{C}^{(i)}_{parent}$: Covariance matrix in the parent node of descriptor group i $(i = 1, \ldots, k)$
$\hat{\boldsymbol{C}}^{(i)}_{parent}$: Estimate of $\boldsymbol{C}^{(i)}_{parent}$ $(i = 1, \ldots, k)$
$\underset{\sim}{\boldsymbol{C}}_l$: Set of covariance matrices in level l $(l = 2, \ldots, D)$
$\hat{\underset{\sim}{\boldsymbol{C}}}_l$: Set of estimated covariance matrices in level l $(l = 2, \ldots, D)$

Inputs

$\{\hat{\boldsymbol{\theta}}^{(l)},\ l = 1, \ldots, N\}$: Descriptor estimates from simulator outputs
$\hat{\boldsymbol{\theta}}_0$: Descriptor estimate from real-climate observations
$\{\hat{\boldsymbol{J}}^{(l)},\ l = 1, \ldots, N\}$: Estimated covariance matrices of $\{\hat{\boldsymbol{\theta}}^{(l)},\ l = 1, \ldots, N\}$
$\hat{\boldsymbol{J}}_0$: Estimated covariance matrix of $\hat{\boldsymbol{\theta}}_0$
$\hat{\boldsymbol{\Lambda}}$: Estimated covariance matrix of $\boldsymbol{\omega}$
$\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$: Prior mean and covariance matrix for $\boldsymbol{\theta}_0$

Implementation Steps

In Level 1:
Collect siblings in distinct groups
Obtain $\hat{\underset{\sim}{\boldsymbol{\Theta}}}_1$ from $\{\hat{\boldsymbol{\theta}}^{(l)},\ l = 1, \ldots, N\}$ in level 0
Fit the model of Section 4.6 to groups $\{i,\ i = 1, \ldots, k\}$ of $\hat{\underset{\sim}{\boldsymbol{\Theta}}}_l$ sharing the same grand-parent
Each time the model is fitted:
for each group i, $(i = 1, \ldots, k)$ **do**
    Obtain $\hat{\boldsymbol{\Sigma}}_i$ and $\hat{\boldsymbol{\alpha}}_i$, as in Section 4.7
    Set $\hat{\boldsymbol{C}}^{(i)}_{parent} \equiv \hat{\boldsymbol{\Sigma}}_i$ and $\hat{\boldsymbol{\theta}}^{(i)}_{parent} \equiv \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\alpha}}_i$
end for

for each level l, $(l = 2, \ldots, (D-2))$ **do**
    Collect siblings in distinct groups
    Obtain $\hat{\underset{\sim}{\boldsymbol{\Theta}}}_l$ from $\{\hat{\boldsymbol{\theta}}^{(i)}_{parent},\ i = 1, \ldots, k\}$ obtained when model was fitted to level $l-1$
    Fit the model of Section 4.6 to groups $\{i,\ i = 1, \ldots, k\}$ of $\hat{\underset{\sim}{\boldsymbol{\Theta}}}_l$ sharing the same grand-parent
    Each time the model is fitted:
    for each group i, $(i = 1, \ldots, k)$ **do**
        Obtain $\hat{\boldsymbol{\Sigma}}_i$ and $\hat{\boldsymbol{\alpha}}_i$, as in Section 4.7
        Set $\hat{\boldsymbol{C}}^{(i)}_{parent} \equiv \hat{\boldsymbol{\Sigma}}_i$ and $\hat{\boldsymbol{\theta}}^{(i)}_{parent} \equiv \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\alpha}}_i$
    end for

Algorithm 1 Framework implementation (continued)

17: **if** $l = D - 2$ **then**
18: Obtain $\hat{\boldsymbol{\Sigma}}_\alpha$ as in Section 4.7
19: Set $\hat{\boldsymbol{C}} \equiv \hat{\boldsymbol{\Sigma}}_\alpha$
20: **end if**
21: **end for**

Evaluate $\{\tilde{\boldsymbol{K}}_{ij},\ i,j = 1,\ldots,N\}$**:**

22: Calculate $\boldsymbol{K}$, using $\{\hat{\boldsymbol{J}}^{(l)},\ l = 1,\ldots,N\}$, $\{\hat{\underline{\boldsymbol{C}}}_l,\ l = 2,\ldots,D\}$ (obtained in 1-21) and $\hat{\boldsymbol{\Lambda}}$, as described in Section 4.3
23: Numerically invert $\boldsymbol{K}$
24: Evaluate the blocks $\{\tilde{\boldsymbol{K}}_{ij},\ i,j = 1,\ldots,N\}$ of $\boldsymbol{K}^{-1}$ ((4.15), p.87)

Derive the parameters of the true-climate posterior $\pi\left(\boldsymbol{\theta}_0|\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}^{(1)},\ldots,\hat{\boldsymbol{\theta}}^{(N)}\right)$

25: Calculate $\boldsymbol{S}^{-1}$ and $\boldsymbol{\tau}$ according to (4.17) -(4.18) in Section 4.3, using: $\boldsymbol{\Sigma}_0$, $\boldsymbol{\mu}_0$, $\hat{\boldsymbol{\theta}}_0$, $\hat{\boldsymbol{J}}_0$, $\{\hat{\boldsymbol{\theta}}^{(l)},\ l = 1,\ldots,N\}$ and $\{\tilde{\boldsymbol{K}}_{ij},\ i,j = 1,\ldots,N\}$
26: The values of $\boldsymbol{S}^{-1}$ and $\boldsymbol{\tau}$ become the precision and mean respectively of the MVN posterior distribution of $\boldsymbol{\theta}_0$ ((4.16), p.87)

$1,\ldots,N\}$ and set values for the prior parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ of $\boldsymbol{\theta}_0$. For the RPMG implementation, these are all obtained as described in Section 3.5.1. Then, the posterior mean $\boldsymbol{\tau}$ and precision matrix $\boldsymbol{S}^{-1}$ of $\boldsymbol{\theta}_0$ can be calculated according to (4.17)-(4.18).

4.9 Summary

This chapter introduced a framework for combining information among simulators in a MME, which extends the conceptual framework of Section 2.6 in order to additionally account for the presence of simulator grouping. The extended framework is motivated by recognizing that simulators in a MME share similarities with each other and therefore it is not realistic to consider them collectively as being independent, conditional on their consensus. Instead, the extended framework incorporates grouping of simulators as determined by expert judgement, leading to a nested group MME structure enabling nested grouping of simulators for an arbitrary number of levels. Simulator grouping in each level is expressed as groups of sibling descriptors centred around their parent descriptors in the next level up.

The aim was to derive the posterior of true-climate descriptor under the proposed framework, given a set of outputs from the simulators in the MME, as well as climate observations. In order to achieve that, it was required to make distributional assumptions about the descriptors at each level of the framework, as shown in Section

4.3. However, the complexity of the framework implies that its implementation is challenging. The main challenge was to estimate the covariance matrices expressing within-group variability of sibling descriptors with regards to their parent descriptor. The difficulty in estimation particularly arises due to the presence of sparse structures and especially singleton groups in the framework levels, which the majority of methods reviewed in Section 4.5 fail to capture. A model which allows sharing of information among groups was therefore proposed (Section 4.6), which enables estimation of within-group variability, regardless of the group size.

The proposed random effects model is based on the assumption that the covariance matrices expressing within-group variability in exchangeable groups are randomly sampled from a common distribution. This allows sharing of information between the exchangeable groups and therefore enables estimation of within-group variability in sparse data structures, including singleton groups. The assumption of random within-group variability requires an additional level in the conventional one-way random effects model. However, Section 4.7.1.1 suggests that estimation of the random model parameters can be achieved by deploying the Gibbs sampler algorithm on two full conditional distributions whose analytical form is straightforward to derive. This however also requires method of moments estimation of the fixed model parameters which are also involved in the two full conditional distributions.

The use of method of moments for estimation of the fixed parameters was motivated from the empirical Bayes approach, where prior parameters are estimated from the data: prior judgements are thereby avoided, in contrast to hierarchical Bayesian implementation. Besides, it is far from clear how to make reasonable prior judgements about within-group variability for any level in the framework, since descriptors at each level represent different quantities and also depend on the grouping criteria. The empirical Bayes approach is commonly used in model inference for estimation of within-group variability in sparse data structures, as also discussed in Section 4.5. The method of moments is a computationally efficient method of estimation, since it provides analytical expressions for the required parameter estimates, by solving a simple linear system of equations. However, it has the limitation of potentially producing estimates outside the parameter support, in some cases. In the model of Section 4.6, this particularly refers to the possibility of getting a negative-definite estimate of between-group variability $\mathbf{\Sigma}_\alpha$. According to Searle et al. (2006, Section 4.4), this might indicate that between-group variability is less than within-group variability, suggesting that more data are required in order to make group-effects non-negligible. If no more data are available, the estimate can be slightly modified in order to become positive-definite. In this case however, the unbiasedness property of the estimate is lost. Another limitation of the ANOVA method is that, since there

is no uniformly "best" choice of sums of squares in unbalanced data, the uniformly best unbiasedeness property of ANOVA estimators under normality assumptions in balanced data does not hold for unbalanced designs. Besides, Searle et al. (2006, Section 2.3) argue that ANOVA estimators lack any analytical properties which help identify "optimality" of one over another. For the above reasons, it is usually more justifiable in terms of assessing theoretical properties of estimators, to perform ML estimation instead of ANOVA for unbalanced data (Searle et al., 2006, Section 6.8). Despite the limitations however, it can be argued that the method of moments is a computationally convenient alternative to the maximum-likelihood approach in the presence of singleton groups, in which case the latter fails to provide estimates of within-group variability. Another limitation of the implementation is the bias of the framework's covariance matrices, induced from using descriptor estimates instead of the actual descriptors in the implementation, as shown in Algorithm 1. Chandler (2013) derives an expression for the magnitude of this bias effect, using the analytic formula for estimating $\boldsymbol{C}$ in the simpler framework of Section 2.6. However, it is not straightforward how to achieve that under the extended framework, where estimation of the covariance matrices is performed numerically and after sharing information among groups.

Once the fixed model parameters are estimated, a Gibbs sampler can be used to estimate the random model parameters. The proposed random effects model is fitted recursively in the framework levels, as described in Section 4.8. The recursive structure allows framework implementation for an arbitrary number of levels, providing flexibility to accommodate a wide range of nested group MME structures. It provides a consistent way of estimating the framework's covariance matrices expressing within-group variability at each level. The recursive algorithm is efficient in the sense that each time the proposed random effects model is fitted in a particular level, it achieves estimation of both the covariance matrices and the corresponding parent descriptors in the parent node. The latter will act as data to which the proposed random effects model will be fitted in the next level up. Finally, the recursive algorithm avoids the dimensionality issues relevant to handling of large matrices which would potentially arise if the whole structure was represented by a single multilevel model, when the number of levels is high.

The fact that implementation of the extended framework is limited in its ability to provide analytical estimates of within-group variability, thus requiring the use of Gibbs sampler, could be considered as a disadvantage, mainly in terms of computational efficiency and potential convergence issues. For the real data application of Section 5.3 however, diagnostics suggest that convergence is fast. An additional limitation in the implementation of the framework is considered to be the suggested

estimators of within-group variability in the special cases considered in Section 4.7.3, since they are based on heuristic criteria and therefore their properties such as bias and root mean squared error need to be assessed. It is also worth noting that estimates $\{\hat{\boldsymbol{\Sigma}}_i,\ i = 1, \ldots, k\}$ of within-group variabilities in a particular level of the framework could alternatively be obtained by equating them to $\hat{\boldsymbol{\Sigma}}_\alpha$ resulting from fitting the model of Section 4.6 to the lower level. However, this would not achieve sharing of information between exchangeable groups in each level, which provides more justifiable estimates of within-group variability in sparse data structures.

The next chapter presents a simulation study and an application to temperature data, to assess the performance of the extended framework in inference for true climate, relative to the simpler framework of Section 2.6.

Chapter 5

Simulation study and application to temperature

5.1 Introduction

The aim of this study is to explore the performance of the extended framework described in Chapter 4, on both simulated and real data. An extensive simulation study is presented in Section 5.2, followed by an application to temperature data in Section 5.3.

The main objective of the simulation study is to evaluate the performance of the extended framework in inference about real climate, when grouping of simulators is assumed to exist. This is achieved by comparing its performance in estimating the true-climate descriptor $\boldsymbol{\theta}_0$, which is assumed to be known in the study, relative to the simpler framework (Section 2.6), given synthetic data (acting as "observations' and "simulator outputs"). The performance of the methodology in Sections 4.6-4.7 for estimating variance components under the extended framework is also assessed. Section 5.2.1 describes the process for generation of synthetic data. Two simulation sets, namely set A and B are produced. Simulation set A aims in drawing some generic conclusions about the performance of the extended framework relative to the simpler, assuming mild conditions for the data structure (e.g. large, balanced groups). On the other hand, simulation set B explores sensitivity of the extended framework's performance when these conditions are relaxed. Section 5.2.2 defines the parameter settings which are common in both simulation sets. Next, estimation of $\boldsymbol{\theta}_0$ under the two frameworks to be compared is described in Section 5.2.3. The metrics deployed for assessing the relative performance of the extended framework in estimating $\boldsymbol{\theta}_0$ are described in Section 5.2.4. The parameter settings and results of simulation sets A

and B are presented in Sections 5.2.5 and 5.2.6 respectively. Finally, the conclusions are summarised in Section 5.2.7.

The application on temperature data is presented in Section 5.3. A RPMG implementation is performed, based on grouping of the CMIP5 GCMs according to expert judgement. The posterior parameters of $\boldsymbol{\theta}_0$ are calculated and compared to those under the RPM and GFB implementations, along with posterior distributions for the components of $\boldsymbol{\theta}_0$ and predictive distributions of temperature. An overview of the application is provided in Section 5.3.1 and the data are described in Section 5.3.2. Details about the implementation are given in Section 5.3.3 and the results are presented in Section 5.3.4.

5.2 Simulation study

In the simulation study, estimates of $\boldsymbol{\theta}_0$ are obtained from calculating the posterior of $\boldsymbol{\theta}_0$ given the data, i.e. simulator outputs and observations, under the two frameworks, for a variety of synthetic data settings where simulator grouping takes place. Different scenarios are considered, to investigate sensitivity of results to various data attributes (e.g. presence of unbalanced and singleton groups, group size) and to the contribution of different sources of variability (e.g. shared simulator discrepancy from reality, within-group and between-group variability) in the posterior of $\boldsymbol{\theta}_0$. Each scenario corresponds to different parameter settings, as shown in Tables 5.1 and 5.7. For each scenario, 1000 synthetic datasets are produced. The next section describes the process for generating the synthetic datasets.

5.2.1 Synthetic data generation

The synthetic data are generated by sampling from the hierarchical model defined under the Gaussian specification of the extended framework, shown in Section 4.3. It is assumed that N simulators are grouped into m families, each with n_i simulators $(i = 1, \ldots, m)$, such that $N = \sum_{i=1}^{m} n_i$. For simplicity, there is no variant level in the simulation study. The data consist of: the real climate observations, which are represented by $\hat{\boldsymbol{\theta}}_0$ and the estimates $\{\hat{\boldsymbol{\theta}}_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i\}$ of the simulator descriptors $\{\boldsymbol{\theta}_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i\}$.

The process for generating the synthetic data is summarised below:

1. Fix N, the total sample size, m, the number of simulator families and $\{n_i,\ i = 1, \ldots, m\}$ the family sizes corresponding to the m families.

2. Fix $\boldsymbol{\Lambda}$ and sample $\boldsymbol{\omega}$ from $\boldsymbol{\omega} \sim MVN(\mathbf{0}, \boldsymbol{\Lambda})$.

3. Fix $\boldsymbol{C}$ and $\boldsymbol{\theta}_0$ and sample $\{\boldsymbol{\theta}_i\ i = 1, \ldots, m\}$ from $\boldsymbol{\theta}_i|\boldsymbol{\omega}, \boldsymbol{C} \sim MVN\left(\boldsymbol{\theta}_0 + \boldsymbol{\omega}, \boldsymbol{C}\right)$ $(i = 1, \ldots, m)$.

   Note that $\boldsymbol{C}$ is interpreted as the between-families variability.

4. Fix $\boldsymbol{\xi} = E\left(\boldsymbol{C}_i\right)$ and v and sample $\{\boldsymbol{C}_i,\ i = 1, \ldots, m\}$ from $\boldsymbol{C}_i \overset{i.i.d.}{\sim} IW\left(\boldsymbol{R}, v\right)$, where $\boldsymbol{R} = (v - p - 1)$
   $mdbold\xi$.

   Note that $\boldsymbol{C}_i$, which expresses within-family variability, is equivalent to $\boldsymbol{\Sigma}_i$ in the random effects model of Section 4.6, for each family i. Based on the assumed distribution for $\boldsymbol{\Sigma}_i$ in (4.23), it follows that $\boldsymbol{C}_i \overset{i.i.d.}{\sim} IW\left(\boldsymbol{R}, v\right)$, where $\boldsymbol{R}$ is the scale matrix and v are the degrees of freedom.

5. Sample $\{\boldsymbol{\theta}_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i\}$ from $\boldsymbol{\theta}_{ij}|\boldsymbol{\theta}_i, \boldsymbol{C}_i \sim MVN\left(\boldsymbol{\theta}_i, \boldsymbol{C}_i\right)$.

6. Fix $\{\boldsymbol{J}_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i\}$ and sample $\{\hat{\boldsymbol{\theta}}_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i\}$ from $\hat{\boldsymbol{\theta}}_{ij}|\boldsymbol{\theta}_{ij}, \boldsymbol{J}_{ij} \sim MVN\left(\boldsymbol{\theta}_{ij}, \boldsymbol{J}_{ij}\right)$.

7. Fix $\boldsymbol{J}_0$ and sample $\hat{\boldsymbol{\theta}}_0$ from $\hat{\boldsymbol{\theta}}_0|\boldsymbol{\theta}_0, \boldsymbol{J}_0 \sim MVN\left(\boldsymbol{\theta}_0, \boldsymbol{J}_0\right)$.

Synthetic data generation thus requires specifying values for the parameters: N, m, $\{n_i,\ i = 1, \ldots, m\}$, $\boldsymbol{\Lambda}$, $\boldsymbol{C}$, $\boldsymbol{\theta}_0$, $\boldsymbol{\xi}$, v, $\{\boldsymbol{J}_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i\}$ and $\boldsymbol{J}_0$.

Simulation experiments are performed in two sets, namely sets A and B (Sections 5.2.5 and 5.2.6 respectively), which examine the performance of the extended framework relative to the simpler, for a range of different scenarios corresponding to different settings for the above parameters. Simulation set A examines the performance of the extended framework in an "idealised" data structure scenario, where the total sample size is large, data are balanced and within-family variabilities are similar between groups (i.e. families are homogeneous). This allows some generic conclusions to be driven about the effect of accounting for simulator grouping in learning about $\boldsymbol{\theta}_0$. Then, simulation set B explores sensitivity of the conclusions to scenarios corresponding to alternative data structures that often occur in real climate applications, such as group inhomogeneity (small v), small sample size N or number of groups m, unbalancedness ($n_i \neq n_j$) and the presence of singleton groups.

The next section describes the parameter settings for synthetic data generation which are common in all scenarios of simulation sets A and B. The individual parameter settings for each simulation set are discussed in Sections 5.2.5.1 and 5.2.6.1, for sets A and B respectively.

5.2.2 Common parameter settings

The parameter settings for $\boldsymbol{\theta}_0$, $\boldsymbol{C}$, $\{\boldsymbol{J}_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,n_i\}$ and $\boldsymbol{J}_0$ are invariant in all scenarios of both simulation sets. Without loss of generality, $\boldsymbol{\theta}_0$ is set for convenience to be a $p \times 1$ zero vector with $p = 6$, similarly to the application to temperature data in Chapter 3. Again for consistency with the application, $\boldsymbol{\theta}_0$ is assumed to be split into "historical" components, corresponding to a period for which observations are available, and components representing the change between future components, for which no observations are available, and historical components.

The covariance matrix $\boldsymbol{C}$ representing between-family variability is chosen to be a 6×6 identity matrix. The magnitude of $\boldsymbol{C}$ provides a reference against which other sources of variability (shared discrepancy and within-family variation) can be assessed, by varying the magnitudes of the associated covariance matrices relative to $\boldsymbol{C}$. The $p \times p$ covariance matrices $\{\boldsymbol{J}_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,n_i\}$ expressing internal variability within each data source are all set to be $0.1 \times \boldsymbol{I}$, where $\boldsymbol{I}$ is the 6×6 identity matrix. Note that the matrices are chosen to be small relative to $\boldsymbol{C}$, since in general natural variability is not expected to be the dominant source of uncertainty in MME structures. Since $\boldsymbol{\theta}_0$ consists of historical and "change" components, so does $\boldsymbol{J}_0$, which represents deviation of $\hat{\boldsymbol{\theta}}_0^{(true)}$ from $\boldsymbol{\theta}_0$ attributable to internal variability in real climate. As in (2.5) of Section 2.6.2, $\boldsymbol{J}_0$ is more conveniently defined via its inverse $\boldsymbol{J}_0^{-1}$, the historical components of which are set to be $10 \times \boldsymbol{I}$. The precision for the change is set to zero, to express the absence of information from observations in the future. The historical components of the precision are set equal to the equivalent in $\{\boldsymbol{J}_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,n_i\}$ defined above. In other words, it is assumed that uncertainty due to natural variability in observations is equal to that of historical simulator outputs.

The remaining parameters N, m, $\{n_i,\ i = 1,\ldots,m\}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\xi}$ and v which vary between simulation sets are described in Sections 5.2.5.1-5.2.6.1, for simulation set A and B respectively.

5.2.3 Estimation of $\boldsymbol{\theta}_0$

For each scenario, the posterior mean $\boldsymbol{\tau}$ and precision matrix $\boldsymbol{S}^{-1}$ of $\boldsymbol{\theta}_0$ are calculated under each framework, for each synthetic dataset. This requires estimates of the framework covariance matrices $\boldsymbol{J}_0$, $\{\boldsymbol{J}_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,n_i\}$, $\boldsymbol{\Lambda}$, $\boldsymbol{C}$ and $\{\boldsymbol{C}_i,\ i = 1,\ldots,m\}$ (in the extended framework), as well as parameter values for $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ in the prior distribution of $\boldsymbol{\theta}_0$.

Note that any bias in the estimation of $\boldsymbol{J}_0$, $\{\boldsymbol{J}_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,n_i\}$ and

$\boldsymbol{\Lambda}$ would potentially interfere with the relative performance of the two frameworks in estimating $\boldsymbol{\theta}_0$. Therefore, since the aim of this simulation study is primarily in examining the effect of accounting for simulator grouping in estimating $\boldsymbol{\theta}_0$, the true values of $\boldsymbol{J}_0$, $\{\boldsymbol{J}_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,n_i\}$ and $\boldsymbol{\Lambda}$ (as determined from the simulation parameter settings) are used instead. Ignorance of the uncertainty in $\boldsymbol{J}_0$ and $\{\boldsymbol{J}_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,n_i\}$, which represent internal variability in each data source, is not expected to seriously violate conclusions, since the parameter settings for $\boldsymbol{J}_0$ and $\{\boldsymbol{J}_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,n_i\}$ are set such that internal variability is small relative to other sources of variability (see Section 5.2.2). Besides, estimation of $\boldsymbol{J}_0$ and $\{\boldsymbol{J}_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,n_i\}$ requires fitting a mimic to observations and simulator outputs respectively, as shown in Section 2.6.3. However, the simulation study does not involve a mimic and therefore it is not clear how to obtain reasonable estimates.

On the other hand, ignoring uncertainty in $\boldsymbol{\Lambda}$, which represents variability due to shared simulator discrepancy from reality, could be considered as a limitation in the simulation study, since shared discrepancy is considered to be an important source of uncertainty in the MME structure. However, the proposed estimator of $\boldsymbol{\Lambda}$ in Section 3.5.1 cannot be used here, since it is based on bootstrapping applied to earlier data, which cannot be obtained in the simulation study. Alternatively, the estimator introduced in Chandler (2013) ((2.13), p.31) could be used, but this would also require subjectively choosing the value of K. Some simulations are run with $K = 0$, i.e. assuming that historical and future shared simulator discrepancy from reality are equal, to investigate whether bias in estimating $\boldsymbol{\Lambda}$ affects the relative performance of the two frameworks in estimating $\boldsymbol{\theta}_0$. Simulation results obtained when $\boldsymbol{\Lambda}$ is estimated are not reported in this chapter. However, they are briefly discussed in Section 5.2.7.

The prior mean $\boldsymbol{\mu}_0$ is set to be a $p \times 1$ zero vector and the prior covariance matrix $\boldsymbol{\Sigma}_0$ is specified as $10^4 \times \boldsymbol{I}$, leading to a non-informative prior for $\boldsymbol{\theta}_0$. This choice is made to ensure that the prior does not affect the performance measures (which will be introduced later) in assessing the aspects of the framework which are relevant to the scope of the study.

For the simpler framework, given a synthetic dataset and the chosen parameter settings for $\boldsymbol{\Lambda}$, $\boldsymbol{J}_0$, $\{\boldsymbol{J}_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,n_i\}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_0$, the posterior parameters are calculated according to (2.10)-(2.11) of Section 2.6.2, under the assumption that the deviation of simulator descriptors from their consensus is represented by a common covariance matrix $\boldsymbol{C}$, estimated as shown in (2.12) of Section 2.6.3.

For the extended framework, Algorithm 1 (p.118) is implemented given the synthetic dataset and the parameter specifications for $\boldsymbol{\Lambda}$, $\boldsymbol{J}_0$, $\{\boldsymbol{J}_{ij},\ i = 1,\ldots,m,\ j =$

$1, \ldots, n_i\}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_0$, to calculate the posterior parameters. According to the algorithm, the covariance matrices $\boldsymbol{C}$ and $\{\boldsymbol{C}_i,\ i = 1, \ldots, m\}$ are estimated using the random effects model of Section 4.6. Estimation of $\{\boldsymbol{C}_i,\ i = 1, \ldots, m\}$ requires the use of Gibbs sampler, which is run for 4 chains with diverse initial values; each chain consists of 1000 iterations, with a burn-in sample of 500.

For each scenario, the above process is repeated for the 1000 synthetic datasets, yielding 1000 sets of posterior parameters of $\boldsymbol{\theta}_0$, under each framework. Denote by G the extended ("Group") framework and by NG the simpler ("No Group") framework. The posterior parameters for each scenario are denoted as $\{(\boldsymbol{\tau}^i_G, \boldsymbol{S}^i_G),\ i = 1, \ldots, 1000\}$ and $\{(\boldsymbol{\tau}^i_{NG}, \boldsymbol{S}^i_{NG}),\ i = 1, \ldots, 1000\}$ for the extended and simpler frameworks respectively, where $\boldsymbol{\tau}$ denotes the posterior mean and $\boldsymbol{S}$ the posterior covariance matrix in each case. For each simulation run i, $\boldsymbol{\theta}_0$ is estimated for each framework by the corresponding posterior mean, following standard Bayesian point estimation theory.

5.2.4 Performance metrics

In order to compare the relative performance of the two frameworks in estimating $\boldsymbol{\theta}_0$ for each scenario, a series of performance metrics is deployed. Denote the posterior mean for each run i by $\boldsymbol{\tau}^i_{\cdot}$, where $\cdot$ stands for either G or NG, based on whether the posterior mean was calculated under the extended or simpler framework respectively. Similarly, $\boldsymbol{S}^i_{\cdot}$ denotes the posterior covariance matrix for run i.

In order to examine, on average, how close the estimates $\{\boldsymbol{\tau}^i_{\cdot},\ i = 1, \ldots, 1000\}$ of $\boldsymbol{\theta}_0$ are to $\boldsymbol{\theta}_0$, the mean bias of $\boldsymbol{\tau}_{\cdot}$ (over the 1000 simulation runs) is calculated as: $\overline{Bias}(\boldsymbol{\tau}_{\cdot}) = \sum_{i=1}^{1000}(\boldsymbol{\tau}^i_{\cdot} - \boldsymbol{\theta}_0)/1000$. Additionally, an indication of the variability in estimating $\boldsymbol{\theta}_0$ is obtained by calculating the root mean squared error (RMSE) of the estimates $\{\boldsymbol{\tau}^i_{\cdot},\ i = 1, \ldots, 1000\}$, for each component j of $\boldsymbol{\theta}_0$, as: $RMSE(\boldsymbol{\tau}_{\cdot}[j]) = \sqrt{\sum_{i=1}^{1000}(\boldsymbol{\tau}^i_{\cdot}[j] - \boldsymbol{\theta}_0[j])^2/1000}$. If the extended framework performs at least as well as the simpler framework in estimating $\boldsymbol{\theta}_0$, it is expected to yield at least as small values of mean biases and RMSEs, relative to those calculated under the simpler framework.

The uncertainty in estimation of $\boldsymbol{\theta}_0$ is assessed by calculating the coverage probabilities of 95% and 99% credible intervals for the individual components of $\boldsymbol{\theta}_0$, and coverages of 95% and 99% credible regions for the whole vector of $\boldsymbol{\theta}_0$. Firstly, the individual component coverage probabilities of 95% and 99% standard Normal credible intervals for $\boldsymbol{\theta}_0$ are evaluated. For a particular component j of $\boldsymbol{\theta}_0$, if $\boldsymbol{\theta}_0[j] \sim MVN(\boldsymbol{\tau}^i_{\cdot}[j], \boldsymbol{S}^i_{\cdot}[jj])$, then each 95% coverage probability is defined as the proportion (among the 1000 simulation runs) of $(\boldsymbol{\theta}_0[j] - \boldsymbol{\tau}^i_{\cdot}[j])/\sqrt{\boldsymbol{S}^i_{\cdot}[jj]}$ which lie between $[-z_{0.95}, z_{0.95}]$; here $z_{0.95}$ denotes the 95% quantile of the $N(0,1)$ distribution, $\boldsymbol{\tau}^i_{\cdot}[j]$ denotes the j^{th} component of $\boldsymbol{\tau}^i_{\cdot}$ and $\boldsymbol{S}^i_{\cdot}[jj]$ denotes the j^{th} diagonal element of

$\boldsymbol{S}^i_{\cdot}$. Likewise, the coverage probabilities are calculated for the 99% credible intervals. Furthermore, to allow involvement of the whole vector of $\boldsymbol{\theta}_0$ in the coverage probability calculation (instead of solely the individual components), the standard result for the distribution of a quadratic form is used (Mathai and Provost, 1992, Theorem 5.1.1). From this, it can be deduced that if $\boldsymbol{\theta}_0 \sim MVN(\boldsymbol{\tau}^i_{\cdot}, \boldsymbol{S}^i_{\cdot})$, then $[\boldsymbol{\theta}_0 - \boldsymbol{\tau}^i_{\cdot}]\, \boldsymbol{S}^{-1\ (i)}_{\cdot}\, [\boldsymbol{\theta}_0 - \boldsymbol{\tau}^i_{\cdot}]' \sim \chi^2(6)$, where $\chi^2(6)$ denotes the Chi-squared distribution with 6 degrees of freedom. Each coverage probability is therefore obtained by calculating the proportion (among the 1000 simulation runs) of values of the above expression which lie in $[0,\chi^2_{0.95}(6)]$, where $\chi^2_{0.95}(6)$ is the 95% quantile of the $\chi^2(6)$ distribution. Likewise, coverage probabilities of 99% credible regions are obtained. Ideally, if a framework yields precise estimates for $\boldsymbol{\theta}_0$, the corresponding coverage probabilities are expected to be close to 0.95 and 0.99, for the 95% and 99% confidence levels respectively.

Finally, the relative sizes of the credible intervals/regions between the two frameworks are determined for the Normal credible intervals and Chi-squared credible regions, for each of the considered scenarios. Smaller intervals or regions can be considered more useful (providing they have the correct coverages), since they offer increased precision when making statements about true climate. For the Normal credible intervals, the interval sizes are evaluated at the 95% level. The mean interval length $\bar{l}_{\cdot}$ is determined for each component j of $\boldsymbol{\theta}_0$, calculated as $\bar{l}_{\cdot}[j] := \sum_{i=1}^{1000} l^i_{\cdot}[j] / 1000$, where $l^i_{\cdot}[j]$ is the individual length for run i, calculated as $l^i_{\cdot}[j] := 2 \times z_{0.95} \times \sqrt{\boldsymbol{S}^i_{\cdot}[jj]}$. For the Chi-squared credible regions, since the whole vector $\boldsymbol{\theta}_0$ is used for calculating the coverage probability, the size of each credible region is defined as the volume of the ellipsoidal credible region for $\boldsymbol{\theta}_0$, determined by the inequality: $[\boldsymbol{\theta}_0 - \boldsymbol{\tau}^i_{\cdot}]\, \boldsymbol{S}^{-1\ (i)}_{\cdot}\, [\boldsymbol{\theta}_0 - \boldsymbol{\tau}^i_{\cdot}]' \leq \chi^2_{\alpha}(6)$; here $\chi^2_{\alpha}(6)$ denotes the percentage point of the $\chi^2(6)$ distribution, at confidence level α. Since the volume is proportional to the determinant of $\boldsymbol{S}^i_{\cdot}$ for each run i, evaluation of the size of each credible region in this study is determined by the mean determinant of $\boldsymbol{S}^i_{\cdot}$ (over the 1000 runs of each scenario). This is calculated as $\bar{d}_{\cdot} := \sum_{i=1}^{1000} d^i_{\cdot} / 1000$, where $d^i_{\cdot}$ is the determinant of the posterior covariance matrix for $\boldsymbol{\theta}_0$, for simulation run i, i.e $d^i_{\cdot} := det\left(\boldsymbol{S}^i_{\cdot}\right)$. Ideally, the extended framework should yield at least as small sizes of credible intervals/regions as the simpler framework.

It is also of interest to assess the performance of the random effects model of Section 4.6 in variance component estimation, especially in estimating the within-family variabilities $\{\boldsymbol{C}_i,\ i = 1, \ldots, m\}$ in the extended framework. This is performed by calculating the mean bias (over the 1000 simulations) in estimation of $\{\boldsymbol{C}_i,\ i = 1, \ldots, m\}$. In order to examine the performance of the random effects model of Section 4.6 in more detail, the mean biases of the parameters v, $\boldsymbol{\xi}$, $\boldsymbol{\Sigma}_{\alpha}$, $\boldsymbol{S}_E$, $\boldsymbol{S}_G$ and $\{\boldsymbol{\mu} + \boldsymbol{\alpha}_i,\ i = 1, \ldots, m\}$ (following the notation in Sections 4.6-4.7) are also calculated.

Some of these results are presented in Appendix J.

Section 5.2.5 describes the parameter settings and presents the results for simulation set A.

5.2.5 Simulation set A

5.2.5.1 Parameter settings

For simulation set A, $N = 100$ simulators are considered in total, falling into $m = 10$ groups. The total sample size and family size are both fairly large, therefore. The individual group sizes $\{n_i,\ i = 1, \ldots, m\}$ are all set to 10, to give a balanced data structure. To preserve homogeneity in families, v, the degrees of freedom in the common inverse-Wishart distribution of within-group variabilities $\{\boldsymbol{C}_i,\ i = 1, \ldots, m\}$, needs to be large and therefore is set to be 100.

The performance of the extended framework is assessed for different sizes of $\boldsymbol{\xi}$, the expected value of within-group variability. The settings for $\boldsymbol{\xi}$ are: $0.1 \times \boldsymbol{I}$ (*small*), $\boldsymbol{I}$ (*moderate*) and $10 \times \boldsymbol{I}$ (*large*). Recall from Section 5.2.2 that $\boldsymbol{C}$ is fixed throughout the simulations to be a 6×6 identity matrix. Thus, small $\boldsymbol{xi}$ implies that within-family variability is considerably smaller than between-family variability $\boldsymbol{C}$, suggesting a well-defined group structure. As $\boldsymbol{\xi}$ increases, simulator grouping becomes less obvious.

For each choice of $\boldsymbol{\xi}$, three different settings are considered for $\boldsymbol{\Lambda}$: $0.1 \times \boldsymbol{I}$ (*small*), $\boldsymbol{I}$ (*moderate*) and $10 \times \boldsymbol{I}$ (*large*). This allows examining the performance of the extended framework when variability due to shared simulator discrepancy ($\boldsymbol{\omega}$) from reality varies, for different degrees of grouping. Small $\boldsymbol{\Lambda}$ implies that variability due to $\boldsymbol{\omega}$ is small relative to between-family variability $\boldsymbol{C}$, whereas large $\boldsymbol{\Lambda}$ implies that variability due to $\boldsymbol{\omega}$ is the same as between-family variability.

Table 5.1 presents the parameter settings for simulation set A.

Scenario	$\boldsymbol{\Lambda}$	$\boldsymbol{\xi}$
A1	Large ($10\boldsymbol{I}$)	Small ($0.1\boldsymbol{I}$)
A2	"	Moderate ($\boldsymbol{I}$)
A3	"	Large ($10\boldsymbol{I}$)
A4	Moderate ($\boldsymbol{I}$)	Small ($0.1\boldsymbol{I}$)
A5	"	Moderate ($\boldsymbol{I}$)
A6	"	Large ($10\boldsymbol{I}$)
A7	Small ($0.1\boldsymbol{I}$)	Small ($0.1\boldsymbol{I}$)
A8	"	Moderate ($\boldsymbol{I}$)
A9	"	Large ($10\boldsymbol{I}$)

Table 5.1: Parameter Settings for simulation set A, where $N = 100$, $m = 10$ and $\{n_i = 10,\ i = 1, \ldots, m\}$.

The next section presents the results from simulation set A.

5.2.5.2 Results

Firstly, the mean biases in estimation of $\boldsymbol{\theta}_0^{(true)}$ under the two frameworks in each scenario are shown in Table 5.2.

		Historical			**Change**		
Scenario	*Framework*	$\boldsymbol{\tau}.[1]$	$\boldsymbol{\tau}.[2]$	$\boldsymbol{\tau}.[3]$	$\boldsymbol{\tau}.[4]$	$\boldsymbol{\tau}.[5]$	$\boldsymbol{\tau}.[6]$
A1	G	0.009	-0.016	0.012	-0.118	0.214	0.012
	NG	0.009	-0.016	0.012	-0.118	0.213	0.012
A2	G	-0.003	0.008	0.001	0.016	0.010	0.043
	NG	-0.003	0.008	0.001	0.015	0.011	0.044
A3	G	-0.001	0.024	0.008	-0.069	-0.052	0.107
	NG	-0.001	0.025	0.008	-0.070	-0.051	0.109
A4	G	0.009	-0.018	0.013	-0.056	0.068	0.016
	NG	0.009	-0.018	0.013	-0.057	0.068	0.016
A5	G	-0.001	0.007	-0.001	0.008	0.004	0.012
	NG	-0.001	0.007	-0.002	0.006	0.005	0.013
A6	G	-0.001	0.023	0.004	-0.012	-0.001	0.036
	NG	-0.001	0.023	0.004	-0.015	0.000	0.036
A7	G	0.006	-0.016	0.009	-0.035	0.021	0.016
	NG	0.005	-0.017	0.008	-0.037	0.022	0.017
A8	G	-0.001	0.003	-0.007	0.007	0.001	0.003
	NG	0.001	0.002	-0.009	0.003	0.003	0.003
A9	G	-0.002	0.019	-0.002	0.006	0.012	0.017
	NG	-0.001	0.017	-0.003	0.002	0.015	0.014

Table 5.2: Mean biases of $\{\boldsymbol{\tau}_{\cdot}^{i},\ i = 1, \ldots, 1000\}$, for simulation set A.

The biases are very similar between the two frameworks in all scenarios and for all components of $\boldsymbol{\theta}_0$, suggesting that the extended framework is not in general outperformed by the simpler framework in this respect. Results are insensitive to the magnitude of $\boldsymbol{\Lambda}$. Any differences between the two frameworks are attributable to sampling uncertainty, which is calculated to be in the range of $0.001 - 0.008$ and $0.014 - 0.102$ for the historical and future components respectively.

Results for the RMSE are shown in Table 5.3.

		Historical			**Change**		
Scenario	*Framework*	$\boldsymbol{\tau}.[1]$	$\boldsymbol{\tau}.[2]$	$\boldsymbol{\tau}.[3]$	$\boldsymbol{\tau}.[4]$	$\boldsymbol{\tau}.[5]$	$\boldsymbol{\tau}.[6]$
A1	G	0.33	0.33	0.31	3.22	3.25	3.19
	NG	0.33	0.33	0.31	3.22	3.25	3.19
A2	G	0.32	0.31	0.32	3.22	3.14	3.16
	NG	0.32	0.31	0.32	3.22	3.14	3.16
A3	G	0.32	0.30	0.31	3.17	3.16	3.13
	NG	0.32	0.30	0.31	3.17	3.16	3.13
A4	G	0.32	0.32	0.30	1.07	1.06	1.06
	NG	0.32	0.32	0.30	1.07	1.06	1.06
A5	G	0.30	0.30	0.31	1.07	1.02	1.05
	NG	0.30	0.30	0.31	1.07	1.02	1.05
A6	G	0.31	0.29	0.30	1.09	1.07	1.09
	NG	0.31	0.29	0.30	1.08	1.07	1.08
A7	G	0.27	0.27	0.26	0.46	0.45	0.46
	NG	0.28	0.28	0.26	0.46	0.44	0.46
A8	G	0.26	0.27	0.27	0.48	0.45	0.48
	NG	0.27	0.28	0.27	0.46	0.44	0.47
A9	G	0.28	0.27	0.28	0.56	0.56	0.57
	NG	0.28	0.27	0.28	0.55	0.54	0.55

Table 5.3: Root mean squared errors (RMSEs) of $\{\boldsymbol{\tau}^i_{\cdot},\ i = 1, \ldots, 1000\}$, for simulation set A.

It can be argued that both frameworks in general perform similarly. Any differences in the RMSE values lie approximately within the bounds of sampling variability and are therefore considered not significant. Typical sampling uncertainty varies between 0.006 and 0.007 for historical components and between 0.01 and 0.073 for future components.

Next, the coverages of the 95% $N(0, 1)$ credible intervals for each component of $\boldsymbol{\theta}_0$ are presented in Table 5.4.

		95%						**99%**					
		Historical			**Change**			**Historical**			**Change**		
Scenario	*Framework*	$\boldsymbol{\theta}_0[1]$	$\boldsymbol{\theta}_0[2]$	$\boldsymbol{\theta}_0[3]$	$\boldsymbol{\theta}_0[4]$	$\boldsymbol{\theta}_0[5]$	$\boldsymbol{\theta}_0[6]$	$\boldsymbol{\theta}_0[1]$	$\boldsymbol{\theta}_0[2]$	$\boldsymbol{\theta}_0[3]$	$\boldsymbol{\theta}_0[4]$	$\boldsymbol{\theta}_0[5]$	$\boldsymbol{\theta}_0[6]$
A1	G	0.94	0.95	0.95	0.95	0.94	0.94	0.99	0.99	0.99	0.99	0.98	0.99
	NG	0.94	0.95	0.95	0.95	0.94	0.94	0.99	0.99	0.99	0.98	0.98	0.98
A2	G	0.94	0.95	0.94	0.95	0.95	0.95	0.99	0.99	0.99	0.98	0.99	0.99
	NG	0.94	0.95	0.94	0.95	0.95	0.95	0.99	0.99	0.99	0.98	0.99	0.99
A3	G	0.95	0.95	0.95	0.94	0.96	0.96	0.99	0.99	0.99	0.99	0.99	1.00
	NG	0.95	0.95	0.95	0.94	0.96	0.95	0.99	0.99	0.99	0.99	0.99	1.00
A4	G	0.93	0.93	0.96	0.95	0.94	0.93	0.98	0.98	0.99	0.99	0.99	0.99
	NG	0.93	0.93	0.96	0.95	0.93	0.93	0.98	0.98	0.99	0.98	0.98	0.98
A5	G	0.95	0.96	0.94	0.95	0.96	0.95	0.99	0.99	0.98	0.98	0.99	0.99
	NG	0.95	0.96	0.94	0.94	0.95	0.94	0.99	0.99	0.99	0.98	0.99	0.99
A6	G	0.94	0.95	0.95	0.95	0.95	0.96	0.99	0.99	0.99	0.98	0.99	0.99
	NG	0.94	0.95	0.95	0.95	0.95	0.95	0.99	0.99	0.99	0.98	0.99	0.99
A7	G	0.92	0.90	0.94	0.88	0.89	0.89	0.98	0.98	0.99	0.98	0.99	0.98
	NG	0.90	0.89	0.92	0.85	0.84	0.84	0.97	0.96	0.98	0.94	0.95	0.94
A8	G	0.94	0.92	0.92	0.89	0.92	0.89	0.98	0.98	0.98	0.98	0.99	0.98
	NG	0.92	0.91	0.90	0.86	0.88	0.86	0.97	0.97	0.97	0.95	0.97	0.95
A9	G	0.94	0.95	0.94	0.92	0.92	0.91	0.99	0.99	0.99	0.98	0.98	0.98
	NG	0.94	0.94	0.93	0.90	0.90	0.90	0.98	0.99	0.99	0.97	0.97	0.96

Table 5.4: Coverages of the 95% and 99% credible intervals for each component of $\boldsymbol{\theta}_0$, for simulation set A.

It is observed that the extended framework attains satisfactory coverages for all scenarios, and is not in general outperformed by the simpler framework. Unless $\boldsymbol{\Lambda}$ is small, i.e in scenarios A1-A6, coverage performance is similar between the two frameworks. Any differences could be easily attributable to sampling variability, which is calculated to be in the ranges of $0.006 - 0.012$ and $0.002 - 0.008$ for 95% and 99% credible intervals respectively. When $\boldsymbol{\Lambda}$ is small however (scenarios A7-A9), the extended framework shows evidence of superiority over the simpler framework, especially for small $\boldsymbol{\xi}$ (scenario A7). The superiority becomes less obvious as the group structure becomes less clear, i.e. as $\boldsymbol{\xi}$ increases (scenarios A8-A9). It is also worth mentioning that the coverages for the "change" components of $\boldsymbol{\theta}_0$ are generally improved in both frameworks as $\boldsymbol{\xi}$ increases, since it dominates over $\boldsymbol{J}_0$ and $\{\boldsymbol{J}_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i\}$ and therefore reduces the contribution of natural variability in estimation of $\boldsymbol{\theta}_0$.

The mean lengths of the 95% intervals are presented in Table 5.5.

		Historical			**Change**		
Scenario	*Framework*	$\boldsymbol{\theta}_0[1]$	$\boldsymbol{\theta}_0[2]$	$\boldsymbol{\theta}_0[3]$	$\boldsymbol{\theta}_0[4]$	$\boldsymbol{\theta}_0[5]$	$\boldsymbol{\theta}_0[6]$
A1	G	1.23	1.23	1.23	12.45	12.45	12.45
	NG	1.23	1.23	1.23	12.40	12.40	12.40
A2	G	1.23	1.23	1.23	12.46	12.46	12.46
	NG	1.23	1.23	1.23	12.40	12.40	12.40
A3	G	1.23	1.23	1.23	12.53	12.52	12.52
	NG	1.23	1.23	1.23	12.46	12.46	12.46
A4	G	1.19	1.19	1.19	4.11	4.11	4.11
	NG	1.18	1.18	1.18	3.94	3.94	3.94
A5	G	1.19	1.19	1.19	4.13	4.12	4.13
	NG	1.18	1.18	1.18	3.96	3.96	3.96
A6	G	1.19	1.19	1.19	4.31	4.31	4.31
	NG	1.19	1.19	1.19	4.13	4.13	4.13
A7	G	1.00	1.00	1.00	1.71	1.71	1.71
	NG	0.90	0.90	0.90	1.31	1.31	1.31
A8	G	1.01	1.01	1.01	1.74	1.74	1.74
	NG	0.92	0.92	0.92	1.36	1.36	1.36
A9	G	1.07	1.07	1.07	2.12	2.11	2.12
	NG	1.02	1.02	1.02	1.80	1.79	1.79

Table 5.5: Mean lengths of the 95% credible intervals for each component of $\boldsymbol{\theta}_0$, for simulation set A.

Table 5.5 shows that the interval lengths are comparable between the two frameworks when $\boldsymbol{\Lambda}$ is small and medium (scenarios A1-A6), for the historical components of $\boldsymbol{\theta}_0$. For the future components, the lengths increase, due to the absence of observations. Additionally, the simpler framework achieves slightly smaller interval sizes than the extended. As $\boldsymbol{\Lambda}$ becomes smaller, the interval sizes reduce substantially, especially for the future components of $\boldsymbol{\theta}_0$. When $\boldsymbol{\Lambda}$ is small (scenarios A7-A9), the interval lengths are still comparable between the two frameworks, but slightly increase in both frameworks as $\boldsymbol{\xi}$ becomes larger. The mutual increase in interval sizes can be explained by the fact that the group structure becomes less well-defined as $\boldsymbol{\xi}$ increases. It can be generally argued that the extended framework yields consistently however not substantially higher interval lengths, which is unsurprising, considering that the extended framework additionally accounts for variability attributable to simulator grouping. For scenario A7, in which case the extended framework achieves substantially higher coverages than the simpler framework, interval sizes are compa-

rable between the two frameworks.

Next, the coverages of the 95% and 99% $\chi^2(6)$ credible regions are presented in Table 5.6, together with the mean determinants $\bar{d}_G = \sum_{i=1}^{1000} det(\boldsymbol{S}_G^i)/1000$, $\bar{d}_{NG} = \sum_{i=1}^{1000} det(\boldsymbol{S}_{NG}^i)/1000$ (and their ratio $\bar{d}_G/\bar{d}_{NG}$), used to evaluate the sizes of confidence regions in the two frameworks.

	95%		**99%**				
	G	**NG**	**G**	**NG**	$\bar{\boldsymbol{d}}_{\boldsymbol{G}}$	$\bar{\boldsymbol{d}}_{\boldsymbol{NG}}$	$\bar{d}_G/\bar{d}_{NG}$
Scenario							
A1	0.94	0.94	0.99	0.99	0.998	0.971	1.03
A2	0.95	0.95	0.99	0.99	1.00	0.974	1.03
A3	0.95	0.95	0.99	0.99	1.03	1.00	1.03
A4	0.94	0.93	0.99	0.98	0.001	0.001	1.31
A5	0.95	0.93	0.99	0.98	0.001	0.001	1.30
A6	0.95	0.94	0.99	0.99	0.001	0.001	1.30
A7	0.90	0.75	0.97	0.89	1.7^{-6}	2.06^{-7}	8.26
A8	0.89	0.76	0.97	0.91	1.96^{-6}	2.87^{-7}	6.81
A9	0.91	0.88	0.97	0.95	9.29^{-6}	2.81^{-6}	6.81

Table 5.6: Coverages of the 95% and 99% $\chi^2(6)$ credible regions for $\boldsymbol{\theta}_0$ and values of $\bar{d}_G$, $\bar{d}_{NG}$ and the ratio $\bar{d}_G/\bar{d}_{NG}$, for simulation set A.

Table 5.6 suggests clear evidence of the superiority of the extended framework when $\boldsymbol{\Lambda}$ and $\boldsymbol{\xi}$ are both small (scenario A7), since the coverages of both the 95% and 99% credible regions for $\boldsymbol{\theta}_0$ under the extended framework are much closer to the correct ones, relative to those under the simpler framework. This however becomes less obvious as $\boldsymbol{\xi}$ increases (scenarios A8-A9). When $\boldsymbol{\Lambda}$ is medium or large (scenarios A1-A6), any difference between the two frameworks falls within the bounds of sampling error and therefore is not considered as being significant.

Regarding the relative sizes of confidence regions (see last three columns in Table 5.6), it is clear that in both frameworks, they are substantially smaller as $\boldsymbol{\Lambda}$ decreases. Additionally, both frameworks show similar sizes of confidence regions for scenarios A1-A6, i.e. when $\boldsymbol{\Lambda}$ is large and medium relative to $\boldsymbol{C}$. When $\boldsymbol{\Lambda}$ is small however, the sizes of confidence regions in the extended framework are substantially higher than in the simpler framework, since the latter does not account for variability attributable to simulator grouping. The small credible regions are the reason for getting much lower coverages than the expected in the simpler framework. It is also worth mentioning that the large differences between the frameworks are also attributable to the fact that they represent volume in 6 dimensions and therefore their relative magnitudes

are not directly comparable to the lengths of the Normal credible intervals.

The results of simulation set A can be summarized as follows: The performance of the extended framework is very similar to the simpler framework when $\boldsymbol{\Lambda}$ is not small (scenarios A1-A6), in all the performance metrics considered. On the other hand, when both $\boldsymbol{\Lambda}$ and $\boldsymbol{\xi}$ are small (scenario A7), the extended framework achieves coverages closer to the correct ones relative to the simpler, without much increase in the size of the credible intervals/regions. Superiority of the extended framework is less obvious as $\boldsymbol{\xi}$ increases (scenarios A8-A9). The two frameworks perform similarly in terms of mean biases and RMSEs for scenarios A7-A9.

Section 5.2.6 presents the parameter settings and the results for simulation set B.

5.2.6 Simulation set B

5.2.6.1 Parameter settings

Since results of simulation set A suggest that unless $\boldsymbol{\xi}$ and $\boldsymbol{\Lambda}$ are small, the two frameworks perform equally well, the simulations in set B are only run for small $\boldsymbol{\xi}$ and $\boldsymbol{\Lambda}$. Also, alternative settings are considered for the parameters N, m, $\{n_i,\ i = 1,\ldots,m\}$ and v, as shown in Table 5.7. Firstly, a scenario where families are not necessarily homogeneous, implying increased variability in $\{\boldsymbol{C}_i,\ i = 1\ldots,m\}$ and therefore smaller v is examined in Scenario B1. The choice of $v = 10$ is made such that it is close to its lower bound (since $v > 9$ when $p = 6$ according to Section 4.6) and thus substantially different from the choice $v = 100$ in simulation set A. To explore the effect of total sample size, simulation B3 is run with $N = 30$ instead of 100, which is closer to the number of simulator outputs in the real data application of Section 5.3.2. The effect of m, the number of groups, is explored by comparing the results of scenarios B2 and B3, since in the former, $m = 6$ instead of 10. Data unbalancedness is considered in scenarios B4 and B5, where B5 also accounts for the presence of singleton groups. The group sizes $\{n_i,\ i = 1,\ldots,m\}$ in scenarios B4 and B5 are chosen such that they are similar to those in the application of Section 5.3.2. The parameter settings are presented in Table 5.7.

Scenario	$\boldsymbol{N}$	$\boldsymbol{m}$	$\{\boldsymbol{n_i},\ \boldsymbol{i=1,\ldots,m}\}$	$\boldsymbol{v}$
B1	100	10	$\{10,\ldots,10\}$	10
B2	30	6	$\{5,\ldots,5\}$	100
B3	"	10	$\{3,\ldots,3\}$	"
B4	"	"	$\{2,2,2,2,3,3,3,3,5,5\}$	"
B5	"	"	$\{1,1,1,1,1,3,5,5,6,6\}$	"

Table 5.7: Parameter Settings for simulation set B, where $\boldsymbol{\Lambda}$ and $\boldsymbol{\xi}$ are both small.

Section 5.2.6.2 presents the results from simulation set B.

5.2.6.2 Results

Firstly, the mean biases of the estimates of $\boldsymbol{\theta}_0$ are presented in Table 5.8.

		Historical			**Change**		
Scenario	*Framework*	$\boldsymbol{\tau}.[1]$	$\boldsymbol{\tau}.[2]$	$\boldsymbol{\tau}.[3]$	$\boldsymbol{\tau}.[4]$	$\boldsymbol{\tau}.[5]$	$\boldsymbol{\tau}.[6]$
B1	G	-0.001	0.003	0.002	-0.014	0.005	0.004
	NG	-0.003	0.004	0.004	-0.009	0.007	0.002
B2	G	-0.006	0.003	0.002	-0.037	-0.004	0.009
	NG	-0.008	0.008	-0.004	-0.035	-0.010	0.013
B3	G	-0.003	0.007	-0.013	0.022	0.017	-0.009
	NG	-0.002	0.010	-0.012	0.023	0.018	-0.008
B4	G	0.007	-0.002	-0.003	-0.006	0.014	0.013
	NG	0.005	-0.006	-0.003	-0.006	0.014	0.007
B5	G	-0.010	0.007	0.003	-0.035	0.022	0.021
	NG	-0.007	0.009	0.009	-0.030	0.021	0.012

Table 5.8: Mean biases of $\{\boldsymbol{\tau}^i_{\cdot}, \, i = 1, \ldots, 1000\}$, for simulation set B.

Similarity in biases between the two frameworks is observed for all scenarios; none of the differences are significant, with no consistent evidence of superiority of one framework over another. The biases are relatively small, suggesting that the satisfactory performance of the extended framework is robust to changes in group homogeneity, total sample size, group size, unbalancedness and presence of singleton groups.

Table 5.9 presents the RMSEs for the estimates of $\boldsymbol{\theta}_0$ under the two frameworks.

		Historical			**Change**		
Scenario	*Framework*	$\boldsymbol{\tau}_\cdot[1]$	$\boldsymbol{\tau}_\cdot[2]$	$\boldsymbol{\tau}_\cdot[3]$	$\boldsymbol{\tau}_\cdot[4]$	$\boldsymbol{\tau}_\cdot[5]$	$\boldsymbol{\tau}_\cdot[6]$
B1	G	0.25	0.26	0.25	0.47	0.46	0.45
	NG	0.26	0.27	0.26	0.45	0.45	0.44
B2	G	0.28	0.28	0.28	0.57	0.57	0.54
	NG	0.29	0.30	0.29	0.53	0.54	0.51
B3	G	0.26	0.26	0.26	0.46	0.47	0.48
	NG	0.27	0.27	0.27	0.45	0.46	0.47
B4	G	0.25	0.27	0.26	0.46	0.47	0.46
	NG	0.26	0.27	0.27	0.47	0.48	0.47
B5	G	0.27	0.26	0.27	0.48	0.48	0.48
	NG	0.28	0.27	0.29	0.50	0.50	0.50

Table 5.9: Root mean squared errors (RMSEs) of $\{\boldsymbol{\tau}_\cdot^i,\ i = 1, \ldots, 1000\}$, for simulation set B.

In all scenarios, the extended framework always performs at least as well as the simpler one, in terms of RMSE for the historical components of $\boldsymbol{\theta}_0$. For the "change" components however, the extended framework is slightly outperformed in scenarios B1-B3, suggesting a small sensitivity of its performance when groups are inhomogeneous (scenario B1) and when the total sample size decreases (scenarios B2-B3). There is no evidence of sensitivity when data are unbalanced (scenario B4) and singleton groups exist (scenario B5). A comparison of scenarios B2 and B3 implies that the performance of both frameworks is negatively affected when fewer groups exist.

Next, the coverage probabilities of the 95% and 99% $N(0, 1)$ credible intervals for the individual components of $\boldsymbol{\theta}_0$ are presented in Table 5.10.

		95%						**99%**					
		Historical			**Change**			**Historical**			**Change**		
Scenario	*Framework*	$\boldsymbol{\theta}_0[1]$	$\boldsymbol{\theta}_0[2]$	$\boldsymbol{\theta}_0[3]$	$\boldsymbol{\theta}_0[4]$	$\boldsymbol{\theta}_0[5]$	$\boldsymbol{\theta}_0[6]$	$\boldsymbol{\theta}_0[1]$	$\boldsymbol{\theta}_0[2]$	$\boldsymbol{\theta}_0[3]$	$\boldsymbol{\theta}_0[4]$	$\boldsymbol{\theta}_0[5]$	$\boldsymbol{\theta}_0[6]$
B1	G	0.94	0.92	0.94	0.89	0.89	0.90	0.99	0.99	0.99	0.98	0.98	0.98
	NG	0.91	0.90	0.92	0.84	0.86	0.86	0.97	0.97	0.98	0.94	0.94	0.95
B2	G	0.91	0.91	0.91	0.86	0.85	0.88	0.98	0.98	0.98	0.97	0.96	0.96
	NG	0.89	0.89	0.89	0.83	0.81	0.84	0.96	0.96	0.97	0.94	0.92	0.93
B3	G	0.94	0.93	0.94	0.92	0.92	0.89	0.99	0.99	0.99	0.99	0.98	0.98
	NG	0.93	0.92	0.93	0.89	0.89	0.87	0.99	0.98	0.98	0.97	0.96	0.95
B4	G	0.94	0.93	0.95	0.91	0.90	0.90	0.99	0.99	0.99	0.99	0.98	0.99
	NG	0.94	0.91	0.93	0.89	0.87	0.88	0.99	0.98	0.98	0.96	0.96	0.96
B5	G	0.93	0.93	0.93	0.89	0.88	0.88	0.98	0.99	0.99	0.98	0.97	0.98
	NG	0.92	0.91	0.90	0.86	0.84	0.86	0.97	0.98	0.98	0.94	0.93	0.94

Table 5.10: Coverages of the 95% and 99% credible intervals for each component of $\boldsymbol{\theta}_0$, for simulation set B.

Results show that for all scenarios, the extended framework always performs at least as well as the simpler one, in terms of credible interval coverage, for all components of $\boldsymbol{\theta}_0$. However, there is a clear reduction in the performance of both frameworks when the number of groups decreases from 10 (scenario B3) to 6 (scenario B2). Furthermore, slightly reduced performance is also observed in scenario B5 when singleton groups are present and data are highly unbalanced.

Table 5.11 shows the mean lengths of the 95% credible intervals.

		Historical			**Change**		
Scenario	*Framework*	$\boldsymbol{\theta}_0[1]$	$\boldsymbol{\theta}_0[2]$	$\boldsymbol{\theta}_0[3]$	$\boldsymbol{\theta}_0[4]$	$\boldsymbol{\theta}_0[5]$	$\boldsymbol{\theta}_0[6]$
B1	G	1.01	1.00	1.00	1.71	1.72	1.72
	NG	0.90	0.90	0.90	1.31	1.31	1.31
B2	G	1.03	1.03	1.04	1.89	1.88	1.90
	NG	0.94	0.94	0.94	1.45	1.44	1.45
B3	G	1.00	1.01	1.01	1.74	1.73	1.74
	NG	0.95	0.95	0.95	1.46	1.46	1.46
B4	G	1.01	1.01	1.00	1.74	1.73	1.74
	NG	0.94	0.95	0.95	1.46	1.46	1.46
B5	G	1.00	1.00	1.01	1.72	1.72	1.72
	NG	0.94	0.94	0.94	1.45	1.45	1.45

Table 5.11: Mean lengths of the 95% credible intervals for each component of $\boldsymbol{\theta}_0$, for simulation set B.

The interval lengths are similar for both frameworks for all scenarios and do not differ substantially from those in scenario A7 (shown in Table 5.5), which has the same settings for $\boldsymbol{\Lambda}$ and $\boldsymbol{\xi}$ as the scenarios in simulation set B. This suggests that changes in the group structure examined in simulation set B do not seem to seriously affect the interval length, which is rather controlled by variation of $\boldsymbol{\Lambda}$ and $\boldsymbol{\xi}$ in the different scenarios. The extended framework achieves in general slightly higher interval lengths than the simpler, which is more evident in the future components of $\boldsymbol{\theta}_0$.

Table 5.12 presents the coverages of the 95% and 99% $\chi^2(6)$ credible regions for $\boldsymbol{\theta}_0$, together with values of the mean determinants $\bar{d}_G$ and $\bar{d}_{NG}$ (and their ratio $\bar{d}_G/\bar{d}_{NG}$), which evaluate the sizes of the confidence regions under the two frameworks.

	95%		**99%**				
	G	**NG**	**G**	**NG**	$\bar{d}_G$	$\bar{d}_{NG}$	$\bar{d}_G/\bar{d}_{NG}$
Scenario							
B1	0.91	0.77	0.98	0.91	1.73^{-6}	2.06^{-7}	8.38
B2	0.81	0.68	0.91	0.83	3.09^{-6}	4.6^{-7}	6.70
B3	0.91	0.84	0.96	0.93	1.91^{-6}	5.12^{-7}	3.74
B4	0.91	0.82	0.98	0.93	1.87^{-6}	5.05^{-7}	3.70
B5	0.86	0.74	0.96	0.89	1.71^{-6}	4.73^{-7}	3.61

Table 5.12: Coverages of the 95% and 99% $\chi^2(6)$ credible regions for $\boldsymbol{\theta}_0$ and values of $\bar{d}_G$, $\bar{d}_{NG}$ and the ratio $\bar{d}_G/\bar{d}_{NG}$, for simulation set B.

Similarly to the Normal credible intervals, the extended framework achieves higher coverages in all scenarios, suggesting that its performance relative to the simpler framework is robust to changes in data structure (e.g. group inhomogeneity, unbalancedness, reduction of groups and presence of singleton groups) when $\boldsymbol{\Lambda}$ and $\boldsymbol{\xi}$ are small. However, there is a clear reduction in coverages for both frameworks when the number of group decreases from 10 (scenario B3) to 6 (scenario B2) and when singleton groups are present (scenario B5).

The last three columns of Table 5.12 suggest that the extended framework yields higher credible region sizes compared to the simpler framework, for all scenarios of simulation set B. Furthermore, the confidence region sizes in both frameworks are substantially increased when the number of groups decreases from 10 (scenario B3) to 6 (scenario B2). Additionally, the sizes of confidence regions in the extended framework seem to be more robust to decrease of total sample size from $N = 100$ (scenario A7, Table 5.6) to $N = 30$ (scenario B3), unbalancedness (scenario B4) and presence of singleton groups (scenario B5), compared to the simpler framework.

5.2.7 Conclusions

To conclude, the results from the simulation study can be summarised as follows: When variability due to shared simulator discrepancy from reality is small relative to between-group variability, and there is a well-defined group structure, the extended framework is superior in quantifying uncertainty about $\boldsymbol{\theta}_0$ more accurately, and without the size of confidence intervals increasing too much. Regarding bias and RMSE, its performance is similar to that of the simpler framework. When $\boldsymbol{\Lambda}$ is larger or when the group structure becomes less clear, the two frameworks have similar performance and in most cases the extended framework performs at least as well as the simpler. There is consistency in the conclusions when groups are inhomogeneous, unbalanced,

or when fewer or singleton groups exist, although the performance of both frameworks reduces in some cases, especially when the group number is smaller and in the presence of highly unbalanced structures with singleton groups.

The results reported in Sections 5.2.5.2 and 5.2.6.2 use the true instead of an estimated value of $\boldsymbol{\Lambda}$ in obtaining the posterior of $\boldsymbol{\theta}_0$. In order to get an indication of the potential effect of bias in estimation of $\boldsymbol{\Lambda}$ to the above conclusions, simulation for scenario A7 is repeated using the estimate of $\boldsymbol{\Lambda}$ shown in (2.13) of Section 2.6.3, with $K = 0$. Results revealed that the coverages of credible intervals/regions under both frameworks deviate a lot from the desired values and there is substantial increase in the sizes of the credible intervals/regions. Moreover, the mean bias of $\hat{\boldsymbol{\Lambda}}$ (over 1000 simulation runs) was almost twice the true value of $\boldsymbol{\Lambda}$. The key message is therefore that in the presence of substantial bias in estimation of $\boldsymbol{\Lambda}$, accounting for simulator grouping does not improve estimation of $\boldsymbol{\theta}_0$, even when $\boldsymbol{\Lambda}$ is small and there is a well-defined grouping structure (scenario A7). This suggests that variability due to shared simulator discrepancy from reality seems to be the limiting factor in uncertainty quantification of MMEs; determining accurately this variability is probably more important than representing in detail the ensemble structure. Obviously, this also depends on the size of the bias of $\hat{\boldsymbol{\Lambda}}$. Section 6.2 provides recommendations for future research on how uncertainty in $\boldsymbol{\Lambda}$ could be incorporated in the simulation study.

Although estimation of the variance components was not the main focus of this study, because they are merely a means to an end in the context of the application, it is potentially of interest to assess the performance of the estimation methodology proposed in Section 4.7. This has been done; details are in Appendix J. The conclusions can be summarised as follows: Large variability is observed in estimating v, owed to the constraints set for its estimation (see Figure J.1). According to Figure J.2, bias in estimation of $\boldsymbol{\xi}$ increases considerably (relative to the true value of $\boldsymbol{\xi}$) as $\boldsymbol{\xi}$ becomes smaller. This is attributable to the increasing contribution of natural variability in simulator outputs (expressed through $\{\boldsymbol{J}_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots n_i\}$), as $\boldsymbol{\xi}$ decreases. The biases in $\{\boldsymbol{C}_i,\ i = 1, \ldots, m\}$ (expressing within-family variability), do not seem to be affected by the magnitude of $\boldsymbol{\xi}$; their magnitude is close to that of $\boldsymbol{\xi}$ when the latter is small. On the other hand, although the bias of between-family variability $\boldsymbol{C}$ increases as the group structure becomes less well-defined (i.e. as $\boldsymbol{\xi}$ increases), it can generally be considered to be negligible relative to the true value for $\boldsymbol{C}$ (see Figures J.2-J.3). To conclude, parameter estimation of the model in Section 4.6 occasionally leads to substantial biases in the simulation study. This suggests that if primary interest is on estimation of variance components using the model of Section 4.6, results are expected to show sensitivity to the data structure and the degree of grouping. However, the simulation results reported above suggest that any

biases in variance component estimation do not adversely affect the estimation of $\boldsymbol{\theta}_0$.

Section 5.3 illustrates the performance of the extended framework in an application for inference about global surface air temperature.

5.3 Application to global surface air temperature

5.3.1 Overview of the study

This application focuses on implementing the extended framework for inference about global surface air temperature using observations and simulator outputs, in the presence of simulator grouping. The results are compared with these from the implementation of the simpler framework (Section 2.6,) which does not account for simulator grouping, to explore the effect of accounting for simulator grouping in inference about global surface air temperature, in this particular application. The datasets (observations and simulator outputs) used are identical to those described in Section 3.2 and also used for implementation of the simpler framework. In the current application however, in order to investigate the performance of the extended framework in learning about $\boldsymbol{\theta}_0$, simulators are grouped into families, based on expert judgement, as described in Section 5.3.2.

The RPMG implementation is performed by applying Algorithm 1 of Section 4.8, to calculate the posterior parameters of $\boldsymbol{\theta}_0$. The posterior parameters are then compared with the equivalent from the simpler framework, to investigate whether accounting for simulator grouping in this particular application has any effect on the posterior. Results are also compared with those from the GFB implementation, to assess the effect of accounting for simulator grouping, relative to that of accounting for uncertainty in the simpler framework's covariance matrices. The posterior densities for each component of $\boldsymbol{\theta}_0$ under the three implementations are also compared. Finally, predictive distributions of yearly mean global surface air temperature are produced and compared with those produced under the simpler framework. Results are reported in Section 5.3.4.

5.3.2 Description of the datasets

The datasets used for the implementation of the extended framework (observations and GCM outputs) and the periods under study are the same as those used in the implementation of the simpler framework (see Section 3.2). In the current study, the 32 simulators are grouped into families as determined by expert judgement, based on the dendrogram shown in Figure 1*a* of Knutti et al. (2013).

In Knutti et al. (2013), simulators are grouped into families (coloured-coded in their Figure 1a), based on whether they share code or belong to the same modelling centre. A hierarchical clustering analysis of their outputs produces very similar grouping, but it is slightly less interpretable; the colour-coded families are therefore used here. This results in grouping the 32 simulators in 11 families, as shown in Table K.1. Note that a few of the 32 simulators in Table K.1 which do not appear in Figure 1a of Knutti et al. (2013), are assigned to a family involving other members of the modelling centre to which they belong. The grouping structure is unbalanced, involving also some singleton groups.

The next section outlines the RPMG implementation using the available datasets and based on simulator grouping illustrated in Table K.1 of Appendix K.

5.3.3 "Revised poor man's with groups" (RPMG) implementation

From Table K.1, there are 11 simulator families. Let n_i denote the number of simulators in family i ($i = 1, \ldots, 11$). Under the extended framework, grouping in the MME structure is represented by the simulator descriptors $\{\boldsymbol{\theta}_{ij},\ i = 1, \ldots, 11,\ j = 1, \ldots, n_i\}$ in the simulator level, which are centred on the corresponding family descriptors $\{\boldsymbol{\theta}_i,\ i = 1, \ldots, 11\}$ in the family level (see Figure 4.1, p.80).

Data are incorporated in the framework through the descriptor estimates $\hat{\boldsymbol{\theta}}_0$ and $\{\hat{\boldsymbol{\theta}}_{ij},\ i = 1, \ldots, 11,\ j = 1, \ldots n_i\}$, corresponding to observations and simulator outputs respectively. The estimates are the same as in the simpler framework, obtained by fitting the mimic of Section 3.3 (see (3.1), p.40) to observations and simulator outputs. In order to implement Algorithm 1 (p.118) for obtaining the posterior of $\boldsymbol{\theta}_0$, the estimated covariance matrices $\hat{\boldsymbol{J}}_0$, $\{\hat{\boldsymbol{J}}_{ij},\ i = 1, \ldots, 11,\ j = 1, \ldots, n_i\}$ and $\hat{\boldsymbol{\Lambda}}$ must be known. Additionally, the prior mean $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$ for $\boldsymbol{\theta}_0$ need to be set. The matrices $\hat{\boldsymbol{J}}_0$, $\{\hat{\boldsymbol{J}}_{ij},\ i = 1, \ldots, 11,\ j = 1, \ldots, n_i\}$ are already calculated for the simpler framework, as described in Section 3.5.1. $\hat{\boldsymbol{\Lambda}}$ is set to be equal to the proposed bootstrap estimate used in the RPM implementation in Section 3.5.1. Finally, $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are set to be equal to those in the simpler framework (see end of Section 3.5.1).

Given $\hat{\boldsymbol{\theta}}_0$, $\{\hat{\boldsymbol{\theta}}_{ij},\ i = 1, \ldots, 11,\ j = 1, \ldots n_i\}$, $\hat{\boldsymbol{J}}_0$, $\{\hat{\boldsymbol{J}}_{ij},\ i = 1, \ldots, 11,\ j = 1, \ldots, n_i\}$ and $\hat{\boldsymbol{\Lambda}}$, Algorithm 1 (p.118) is implemented to yield the posterior parameters of $\pi\left(\boldsymbol{\theta}_0 | \hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(32)}\right)$, where $\{\hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(32)}\}$ denote the descriptor estimates for the 32 simulator outputs. Similarly to the simulation study, the Gibbs sampler (used to estimate $\{\boldsymbol{C}_i,\ i = 1, \ldots, 11\}$ based on the model of Section 4.6) is run for 4 chains with diverse initial values, where each chain consists of 1000 iterations, with

a burn-in sample of 500. The software implementation time is about 1.5 minutes, using the R statistical software (R Core Team, 2012). Convergence of the MCMC is investigated by calculating the "potential scale reduction factor " (Gelman-Rubin diagnostic) (Gelman and Rubin, 1992), for all the components of $\boldsymbol{\alpha}_i$ and $\boldsymbol{\Sigma}_i$ (following the notation of Section 4.6), for each group i. The maximum value of the diagnostic in the study is 1.009 (rounded to 4 significant digits).

The next section presents the results of the implementation of the extended framework, and compares them with those from the simpler framework shown in Section 3.6.

5.3.4 Results

Firstly, the posterior means and standard deviations are presented for the RPM and RPMG implementations. Results from the GFB implementation are also shown, in order to investigate the relative effects of accounting for simulator grouping and for the uncertainty in the covariance matrices of the simpler framework.

According to Table 5.13, the posterior parameters under the RPMG implementation are in general similar to those under the RPM implementation, suggesting that accounting for simulator grouping in this particular application does not significantly affect inference for the true-climate descriptor $\boldsymbol{\theta}_0$. Similarities are more obvious in the historical components and particularly in $\alpha_0^{(\text{hist})}$ and $\beta_0^{(\text{hist})}$. On the other hand, the extended framework yields a slightly lower posterior mean for ${\sigma_0^2}^{(\text{hist})}$ compared to the simpler framework, but with higher standard deviation, possibly since the extended framework accounts also for the variability due to simulator grouping.

Differences in RPM and RPMG implementations are more evident for the "change" components relative to the historical, suggesting that in the absence of observations, accounting for simulator grouping has higher impact in inference about true climate. The extended framework yields lower/equal posterior means for the "change" components, implying a smaller/equal change in yearly mean global surface air temperature between the historical and future period, compared to that implied in the RPM implementation. The posterior standard deviations are slightly larger in the extended framework, since the latter additionally accounts for variability due to simulator grouping.

The posterior parameters under the RPM and RPMG implementations are in general more similar to each other than they are to the GFB implementation (with the exception of the posterior mean for $\alpha_0^{(\text{fut})} - \alpha_0^{(\text{hist})}$). From this, it can be argued that for the current application, uncertainty in the covariance matrices involved in the simpler framework is more significant than variability due to simulator grouping.

	Historical			**Change**		
	$\alpha_0^{(\text{hist})}$	$\beta_0^{(\text{hist})}$	$\log\left(\sigma_0^{2\ (\text{hist})}\right)$	$\alpha_0^{(\text{fut})} - \alpha_0^{(\text{hist})}$	$\beta_0^{(\text{fut})} - \beta_0^{(\text{hist})}$	$\log\left(\sigma_0^{2\ (\text{fut})}/\sigma_0^{2\ (\text{hist})}\right)$
Observations						
$\hat{\boldsymbol{\theta}}_0$	14.25	0.02	-4.77	-	-	-
Consensus						
$\bar{\hat{\boldsymbol{\theta}}}_{\boldsymbol{ij}}$	14.11	0.03	-3.99	0.82	0.01	-0.54
$\bar{\hat{\boldsymbol{\theta}}}_{\boldsymbol{i}}$	14.19	0.03	-4.21	0.83	0.01	-0.41
Posterior						
RPM	14.24 (0.020)	0.02 (0.003)	-4.30 (0.186)	0.69 (0.110)	0.01 (0.009)	-0.47 (0.352)
GBF	14.24 (0.025)	0.02 (0.004)	-4.44 (0.240)	0.64 (0.175)	0.01 (0.016)	-0.41 (0.570)
RPMG	14.24 (0.020)	0.02 (0.003)	-4.34 (0.193)	0.65 (0.116)	0.01 (0.010)	-0.48 (0.353)

Table 5.13: Analysis of yearly mean global surface air temperature data from HadCRUT3 observations and CMIP5 simulator outputs, under the simpler and the extended frameworks. Top block: parameter estimates of the mimic fitted to observations ($\hat{\boldsymbol{\theta}}_0$). Middle block: estimates of simulator consensus $\bar{\hat{\boldsymbol{\theta}}}_{ij} = \sum_{i=1}^{11}\sum_{j=1}^{n_i}\hat{\boldsymbol{\theta}}_{ij}/32$ and family consensus $\bar{\hat{\boldsymbol{\theta}}}_i = \sum_{i=1}^{11}\hat{\boldsymbol{\theta}}_i/11$; $\hat{\boldsymbol{\theta}}_i$ is the estimated descriptor of family i, defined to be $\hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\alpha}}_i$, after applying the model of Section 4.6. Bottom block: Posterior means and standard deviations (in parentheses) derived from the RPM, GFB and RPMG implementations.

Equivalently, ignoring the uncertainty in the covariance matrices has greater effect in learning about true climate, than not accounting for simulator grouping.

Figure 5.1 presents the posterior densities for each component of $\boldsymbol{\theta}_0$, produced under the RPM,GFB and RPMG implementations. The densities are produced as described in Section 3.6.2.

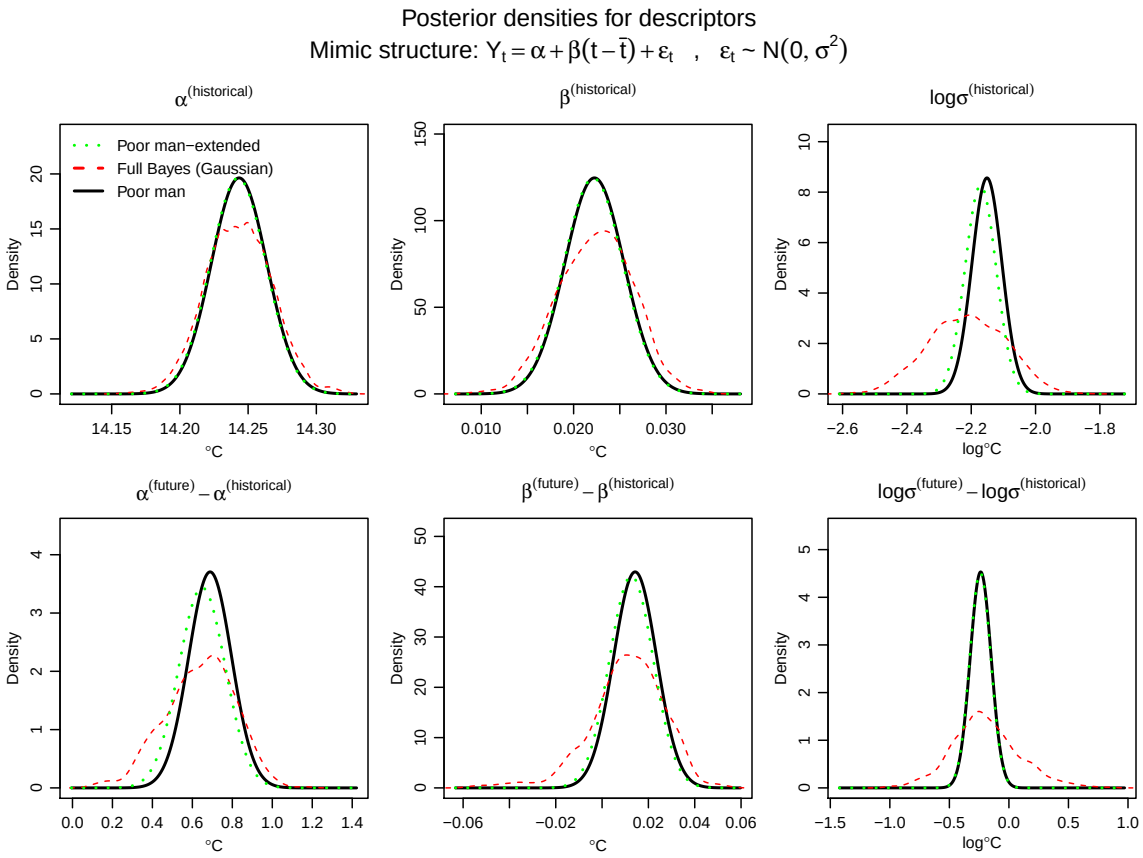


Figure 5.1: Posterior densities for true-climate descriptors, for the RPM, GBF and RPMG implementations. Top row: historical descriptors. Bottom row: change between historical and future descriptors.

Figure 5.1 suggests that the RPM and RPMG implementations have generally very similar performance compared to the GFB implementation. This confirms previous findings, i.e that accounting for uncertainty in the covariance matrices involved in the simpler framework has a larger effect than accounting for simulator grouping. It is also worth mentioning the slightly lower posterior means of $\log\left(\sigma_0^{2\ (\text{hist})}\right)$ and $\alpha_0^{(\text{fut})} - \alpha_0^{(\text{hist})}$ which are evident in the plots of their posterior densities.

Figure 5.2 shows the predictive distributions under the three implementations of interest, which are produced as described in Section 3.6.2.

The predictive distribution of temperature under the extended framework shows strong similarities with the RPM implementation. Additionally, the variability in

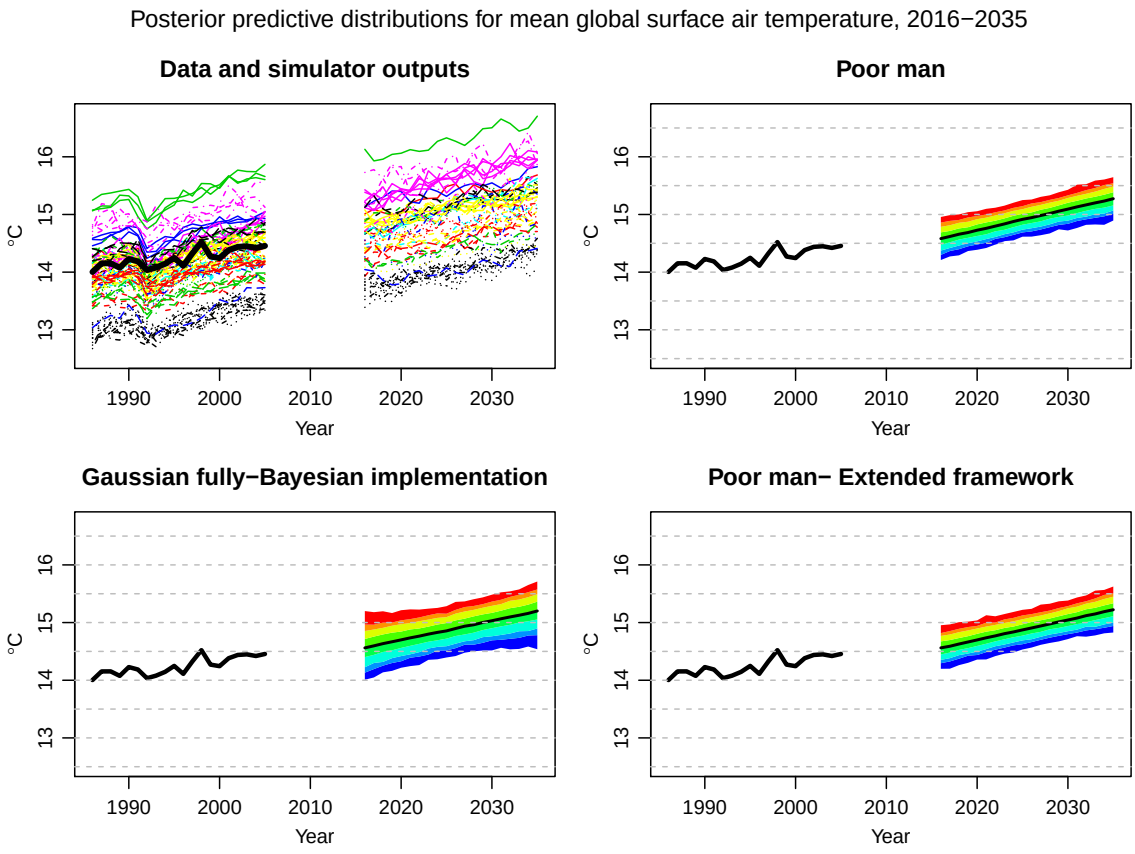


Figure 5.2: Posterior predictive distributions of yearly mean global surface air temperature ($°C$), for the future period $2016 - 2035$. Top-left: Plots of observations (black solid line) and simulator outputs (coloured lines) of yearly mean global surface air temperature for the historical and future periods. Top-right: Predictive distribution of yearly mean global surface air temperature for the future period, obtained from the derived posteriors in RPM implementation. Bottom-left: Predictive distribution under the GFB implementation. Bottom-right: Predictive distribution under the RPMG implementation. Black line: Posterior mean of yearly mean global surface air temperature. Coloured segments: Partition the predictive distribution based on the 1^{st} , 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, 95^{th} and 99^{th} percentiles, from the bottom to the top.

the distribution is still substantially lower than that in the GFB implementation, which is consistent with previous conclusions about the small effect of accounting for simulator grouping in this particular application. Despite the obvious similarities in the plots corresponding to the RPM and RPMG implementations, a closer look reveals slightly larger variability in the predictive distribution under the extended framework at the edges of the distribution, i.e. between the 1^{st} and 5^{th} percentiles and the 95^{th} and 99^{th} percentiles. This suggests that when accounting for simulator grouping, it is expected to have increased uncertainty in the edges of the predictive distribution. Furthermore, there is increased uncertainty at the beginning and end of the future period compared to the simpler framework, suggesting that the extended framework is more sensitive to the absence of any information prior to 2016 and after 2035 relevant to the future period under study.

The results from the application to temperature data can be summarised as follows: The RPMG implementation has very similar performance to the RPM implementation, implying that in this particular application, accounting for simulator grouping does not seriously affect inference for temperature, for the periods under study. A possible explanation for this is that the simulator grouping from Knutti et al. (2013) is not reflected in the descriptor estimates. Recall that the simulation study showed the performance of the extended framework to be noticeably better than the simpler version only in cases where $\boldsymbol{\xi}$ is substantially less than $\boldsymbol{C}$ so that groups are well defined. Expressions (K.1)-(K.2) of Appendix K show however, that this does not appear to be the case for the data considered here. Furthermore, the maximum eigenvalues of $\hat{\boldsymbol{\xi}}$, $\hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{\Lambda}}$ are calculated, as an indication of the relative contributions of the covariance matrices in overall variability. Rounded to 2 decimal places, they are respectively 0.32, 0.16 and 0.19. This reassures that there is no clear grouping structure (since the maximum eigenvalue of $\hat{\boldsymbol{\xi}}$ is not smaller than that of $\hat{\boldsymbol{C}}$) and that variability due to shared simulator discrepancy is not substantially smaller than the other sources of variability (since the maximum eigenvalue of $\hat{\boldsymbol{\Lambda}}$ is not substantially smaller than that of the others).

To conclude, it is natural in general to worry about the structure of the CMIP ensemble and what are the consequences of failing to take it into account. There are often examples of such considerations in the literature (Pirtle et al., 2010; Abramowitz, 2010; Knutti et al., 2010, 2013). The current analysis however, shows that for this particular application at least, ignoring simulator grouping does not make much difference in inference about reality; besides, findings from the earlier simulation study (Section 5.2) clearly indicate why this is the case. Obviously, accounting for simulator grouping might have a bigger effect in other applications, where the grouping structure is more obvious and variability due to simulator grouping is more significant.

Chapter 6

Discussion

6.1 Conclusions

The work in this thesis was motivated by challenges related to uncertainty quantification when combining information from an ensemble of climate simulators, namely a multi-model ensemble (MME). These challenges relate to making realistic assumptions about the MME structure, as well as representatively characterizing the associated uncertainties. These are both non-trivial tasks, especially in the presence of dependencies between simulators and the existence of shared simulator discrepancy from reality. The problems are most effectively tackled using probabilistic frameworks, which make explicit assumptions about the MME structure. These frameworks are in their majority Bayesian, to allow incorporation of prior information about the different sources of uncertainty. Inference for true climate is expressed in the form of a posterior distribution, to characterize the underlying uncertainty in climate projections.

The contribution of this thesis was concerned with improving the probabilistic, Bayesian framework of Chandler (2013) for MME interpretation and implemented the framework for the first time on real data. The improvements regard both assumptions about the MME structure in the framework, as well as characterization of the underlying sources of uncertainty. They are summarized in three points.

Firstly, an improved estimator of $\mathbf{\Lambda}$, the variability due to shared simulator discrepancy from reality, was proposed. In the absence of observations for the future, Chandler (2013) proposes estimating $\mathbf{\Lambda}$ by estimating the historical discrepancy and introducing a parameter K which is subjectively chosen to relate historical to future discrepancy. An improved estimator of $\mathbf{\Lambda}$ was proposed in this thesis, which uses a more robust technique to estimate $\mathbf{\Lambda}$, avoiding thus the subjective choice of K. The estimator was obtained using bootstrapping from earlier data than the pe-

riods of study. Bootstrapping is a widely used technique which has the advantage of constructing realizations of the underlying population, by simply sampling with replacement from the available data and without being restricted by the distributional assumptions about the data. In the context of the current study, pairs of earlier periods for which both observations and simulator outputs are available were sampled with replacement and acted as realizations of a population of "historical" and "future" periods from which shared simulator discrepancy from reality and therefore $\mathbf{\Lambda}$ was estimated.

The second contribution of this thesis incorporated uncertainty in the values of the covariance matrices expressing variability arising from the different sources of uncertainty in the framework of Chandler (2013). The PM implementation of Chandler (2013) assigns estimates to the unknown covariance matrices, in an attempt to provide a computationally cheap implementation. This has the limitation however, of ignoring the underlying uncertainty in the values of these covariance matrices. In this thesis, two fully Bayesian implementations were performed, which assigned prior distributions instead of estimates to the unknown covariance matrices. Different prior choices were considered in the two fully Bayesian implementations, to explore sensitivity of true-climate posterior to the prior choices.

The implementations were illustrated in an application to global surface air temperature using observations and simulator outputs from the latest suite of climate models. The aim was to make inference for yearly mean global surface air temperature, for the 20-year periods $1986-2005$ and $2016-2035$. To investigate the effect of accounting for uncertainty in the covariance matrices, the resulting true-climate posteriors from the two fully Bayesian implementations were compared to the RPM implementation (using the improved estimator of $\mathbf{\Lambda}$ discussed earlier), which is computationally more efficient but ignores uncertainty about the framework's covariance matrices (by assigning estimates instead of distributions to them). The RPM implementation yielded an analytical expression for the posterior of true-climate, whereas in the fully Bayesian implementations, the posterior was calculated numerically, by deploying the Gibbs sampler algorithm. Results revealed that in this particular application, the effect of ignoring uncertainty in the covariance matrices was not so serious, which argues in favour of the computationally simpler RPM implementation. Some slight sensitivity to the prior choices was evident in inference for temperature, for the change between historical and future periods.

The third contribution of this thesis extended the framework of Chandler (2013), to account for potential simulator grouping in the MME. Similarities between simulators are often recognized but rarely accounted for in existing frameworks. The framework of Chandler (2013) treats the ensemble members as independent, conditional

on their consensus. However, considering the known similarities between simulators, this assumption is not always realistic. The proposed extended framework developed in this thesis enables nested simulator grouping in the MME structure, where the latter is determined by expert judgement. Theoretically, an arbitrary number of levels is allowed in the structure. This was achieved by adding levels to the hierarchical framework of Chandler (2013), where each level represents groups of simulators, centred on their own consensus. This was repeated recursively in the MME structure, to represent nested grouping. For the purposes of framework implementation, a random effects model was proposed, to enable estimation of within-group variability in sparse structures, including singleton groups. This was achieved by exploiting the exchangeability assumptions between simulator groups, to allow for sharing of information between them. However, estimation required the use of Gibbs sampler, which can be considered as a disadvantage of the method in terms of computational efficiency. A recursive algorithm was developed, which enables estimation of within-group variability in all the levels of the MME structure involving simulator grouping.

A simulation study was performed to investigate the performance of the extended framework relative to the simpler, in inference for reality. Results revealed that when variability due to shared simulator discrepancy from reality is small and there is a well-defined simulator grouping structure, the extended framework quantifies uncertainty more accurately than the simpler, without considerable increase of the size of uncertainty intervals. However, in the presence of substantial biases in estimation of variability due to shared discrepancy, the performance of the extended framework is seriously affected. This highlights the importance of focusing on robust estimation of uncertainty due to shared simulator discrepancy from reality, which is probably more crucial than representing in detail the ensemble structure. Results were similar between the two frameworks in terms of bias and root mean squared error (RMSE) of the estimates of reality. Regarding parameter estimation of the proposed random effects model, substantial biases were observed in some cases, but without affecting inference for reality. The presence of substantial biases could be considered as a drawback of the proposed random effects model, if focus is primarily on estimation of the model parameters.

The RPMG implementation was also illustrated in an application to the temperature data considered earlier, assuming simulator grouping as determined by expert judgement. The posterior of temperature for the periods under study was compared with that determined under the RPM implementation, when no simulator grouping was assumed. Results suggested that for this particular application, accounting for simulator grouping did not seriously affect inference for temperature. This is explained by the simulation study, since there was not very clear grouping structure.

Nevertheless, the extended framework is expected to provide insights into other applications, where simulator grouping structure is more evident in the framework.

Some potential areas of future work relevant to the contribution of this thesis are presented in the next section.

6.2 Future work

The proposed future work can be summarized in three remarks.

Remark 1. The application to temperature data presented in this thesis was focused in inference about temperature, for two disjoint 20-year periods, in order to facilitate illustration of the proposed methodology. In high-impact studies however, which aim to inform strategies of climate-change adaptation for example, it is of interest to provide projections that can be used at any time over some specified horizon. Simple statistical representations, such as linear trends over non-overlapping time periods (like the mimic used in the temperature application in this thesis), are typically not sufficiently realistic for this purpose. A potential area of future research is therefore to allow more flexible, realistic representations of system behaviour when interpreting MMEs.

This would require the choice of sophisticated mimics that will capture the long-term behaviour of the quantities of interest, such as the existence of non-linear trends and seasonality. There exist various methodologies in the literature which could be used to achieve this (see Chandler and Scott (2011, Chapters 4-5) for an extensive analysis of methodologies for modelling trends in the context of environmental applications). In order to allow for greater flexibility in the structure of the underlying trend, non-parametric trend modelling is preferred compared to parametric. According to Chandler and Scott (2011, Sections 4-5), it can be assumed that the trends are either deterministic, and modelled using local linear smoothing of spline smoothing for example, or stochastic, as in the local linear trend model. For modelling of stochastic trends, state space representation could be deployed, which explicitly allows predicting the future state of the system under study, via the Kalman filter.

Remark 2. As already mentioned in Section 6.1, the use of Gibbs sampler for estimation of within-group variability in the extended framework could be considered as a limitation in the implementation of the framework. This is mainly because it can potentially become computationally demanding if the nested MME grouping structure consists of many levels. On the other hand, the existing methodologies for variance components estimation in unbalanced, sparse structures (also including singleton groups) reviewed in Section 4.5 do not provide much insight into alternative

techniques which avoid the use of numerical algorithms. An exception is the MINQE estimator in Rao et al. (1981, Section 2.5) which provides an analytical expression for the estimator of within-group variability that is defined even for singleton groups. However, as also discussed in Section 4.5, this estimator is defined based on a priori weights which assume equal within-group and between-group variability (which is unrealistic when simulator grouping is known to exist) and also has the limitations of being very sensitive to the choice of prior weights (Searle et al., 2006, Section 11.3c). An alternative approach is proposed in Rao et al. (1981, Section 2.4), where the initial weights can be refined iteratively, using the resulting MINQE estimates in each iteration as "prior" weights for the next iteration. It would therefore be interesting to apply this technique for estimation of within-group variability in the extended framework and compare its performance in terms of estimation and computational efficiency to that of the Gibbs sampler.

Remark 3. In the simulation study of Section 5.2, ignorance of uncertainty in $\boldsymbol{\Lambda}$ in estimation of $\boldsymbol{\theta}_0$ was considered as a limitation. On the other hand, the use of $\hat{\boldsymbol{\Lambda}}$ as defined in (2.13) with $K = 0$ (i.e. assuming equal historical and future shared simulator discrepancy from reality), yielded large biases, which were shown to seriously affect inference for $\boldsymbol{\theta}_0$. One alternative solution which provides a more robust way of accounting for uncertainty in $\boldsymbol{\Lambda}$ is to assign a prior distribution to it, similarly to the GFB implementation of Section 3.5.2.

Another limitation relevant to estimation of $\boldsymbol{\Lambda}$ arises in the temperature application of Section 5.3. $\hat{\boldsymbol{\Lambda}}$ is set to be equal to that in Section 3.5.1, which was concerned with implementation of the simpler framework. Note however that $\boldsymbol{\Lambda}$ in the extended framework has a different interpretation than in the simpler framework, since it represents variability due to shared discrepancy of *family* descriptors (and not *simulator* descriptors) from reality. Thus, ideally, $\hat{\boldsymbol{\Lambda}}$ could be obtained by applying bootstrapping to grouped simulators in the extended framework. Of course this will become computationally challenging as more levels are added to the nested grouped MME structure.

Although the simulation study in Section 5.2 is comprehensive, there are additional things that can be done. One proposal for future research relates to the parameter settings for the covariance matrix $\boldsymbol{\Lambda}$ expressing variability due to shared simulator discrepancy from reality. In the simulation study of Section 5.2, the historical and future blocks of $\boldsymbol{\Lambda}$ where identical, assuming equal historical and future shared simulator discrepancy from reality. In practice however, it is expected that historical and future discrepancy differ. It would therefore be of interest to examine sensitivity of the simulation results to alternative parameter settings for $\boldsymbol{\Lambda}$, with dif-

ferent values for the historical, future and off-diagonal blocks, to express differences in shared simulator discrepancy from reality in the two periods under study.

Appendix A

Sketch of derivation of the MLE of C for $\{\hat{\theta}_i \sim N(\theta_i, C + J_i),\ i = 1, \ldots, m\}$, when $p = 1$

Assume for simplicity, that the descriptor estimates $\{\hat{\theta}_i,\ i = 1, \ldots, m\}$ in the framework of Section 2.6 are univariate and that they correspond to the raw data $\{y_i,\ i = 1, \ldots, m\}$. Assume also that C is the variance representing common deviation of simulator descriptors from their consensus and J_i is the variance representing natural variability in data source i (following the notation of Section 2.6.2). Then, based on the Gaussian distributional assumptions of Section 2.6.2, it can be deduced that $\{\hat{\theta}_i \sim N(\theta_i, C + J_i),\ i = 1, \ldots, m\}$.

This appendix illustrates the attempt to derive an estimate of C from first principles, and in particular using maximum-likelihood estimation.

For $\{\hat{\theta}_i \sim N(\theta_i, C + J_i),\ i = 1, \ldots, m\}$, the joint likelihood $L_{\hat{\theta}_1,..,\hat{\theta}_m}(\theta_i, C, J_i)$ of $\hat{\theta}_1, \ldots, \hat{\theta}_m$ is derived to be:

$$L_{\hat{\theta}_1,\ldots,\hat{\theta}_m}(\theta_i, C, J_i) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi(C + a_i)}} \exp\left(- \sum_{i=1}^{m} \frac{\left(\hat{\theta}_i - \theta_i\right)^2}{2(C + J_i)} \right). \qquad \text{(A.1)}$$

The log-likelihood $l_{\hat{\theta}_1,\ldots,\hat{\theta}_m}(\theta_i, C, J_i)$ is then:

$$l_{\hat{\theta}_1,\ldots,\hat{\theta}_m}(\theta_i, C, J_i) = \ln\left(2\pi^{-\frac{m}{2}}\right) + \sum_{i=1}^{m} \ln\left(C + J_i\right)^{-\frac{1}{2}} - \frac{1}{2}\sum_{i=1}^{m} \frac{\left(\hat{\theta}_i - \theta_i\right)^2}{C + J_i}, \tag{A.2}$$

and the derivative of (A.2) with respect to C is:

$$-\frac{1}{2}\sum_{i=1}^{m} \frac{1}{(C + J_i)} + \frac{1}{2}\sum_{i=1}^{m} \frac{(\hat{\theta}_i - \theta_i)^2}{(C + J_i)^2}. \tag{A.3}$$

Setting (A.3) equal to zero yields:

$$\sum_{i=1}^{m} \frac{(\hat{\theta}_i - \theta_i)^2 - (C + J_i)}{(C + J_i)^2} = 0. \tag{A.4}$$

In order to obtain the MLE of C, it is required to solve (A.4) in terms of C, treating $\hat{\theta}_i$, θ_i and J_i as constants. However, it is not straightforward how this can be achieved analytically for this particular expression. This suggests that an analogous attempt to find the MLE of $\boldsymbol{C}$ in the multivariate case will not be a trivial task, as already discussed in Section 2.6.3.

Appendix B

Official model and group names of the GCMs participating in the CMIP5 experiment

Modeling Centre(or Group)	**Institute ID**	**Model Name**
Commonwealth Scientific and Industrial Research Organization(CSIRO) and Bureau of Meteorology(BOM),Australia	CSIRO-BOM	ACCESS1.0 ACCESS1.3
Beijing Climate Centre,China Metereological Administration	BCC	BCC-CSM1.1
College of Global Change and Earth System Science, Beijing Normal University	GCESS	BNU-ESM
Canadian Centre for Climate Modelling and Analysis	CCMA	CanESM2 CanCM4 CanAM4
University of Miami-RSMAS	RSMAS	CCSM4(RSMAS)

Modeling Centre(or Group)	**Institute ID**	**Model Name**
National Centre for Atmospheric Research	NCAR	CCSM4
Community Earth System Model Contributors	NSF-DOE-NCAR	CESM1(BGC) CESM1(CAM5) CESM1(CHEM,CAM5) CESM1(CHEM) CESM1(WACCM)
National Centres for Environmental Prediction	NCEP	CFSv2-2011
Centro Euro-Mediterraneo per I Cambiamenti Climatici	CMCC	CMCC-CESM CMCC-CM CMCC-CMS
Centre National de Recherches Meteorologiques/Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique	CNRM-CERFACS	CNRM-CM5
Commonwealth Scientific and Industrial Research Organization in collaboration with Queensland Climate Change Centre of Excellence	CSIRO-QCCCE	CSIRO-Mk3.6.0
EC-EARTH consortium	EC-EARTH	EC-EARTH
LASG, Institute of Atmospheric Physic, Chinese Academy of Sciences and CESS,Tsinghua University	LASG-CESS	FGOALS-g2
LASG,Institute of Atmospheric Physics,Chinese Academy of Sciences	LASG-IAP	FGOALS-gl FGOALS-s2

Modeling Centre(or Group)	**Institute ID**	**Model Name**
The first Institute of Oceanography,SOA,China	FIO	FIO-ESM
NASA Global Modeling and Assimilation Office	NASA GMAO	GEOS-5
NOAA Geophysical Fluid Dynamics Laboratory	NOAA GFDL	GFDL-CM2.1 GFDL-CM3 GFDL-ESM2G GFDL-ESM2M GFDL-HIRAM-C180 GFDL-HIRAM-C360
NASA Goddard Institute for Space Studies	NASA GISS	GISS-E2-H GISS-E2-H-CC GISS-E2-R GISS-E2-R-CC GIS-E2CS-H GISS-E2CS-R
National Institute of Meteorological Research/Korea Meteorological Administration	NMR/KMA	HadGEM2-AO
Met Office Hadley Centre	MOCH	HadCM3 HadCM3Q HadGEM2-CC HadGEM2-ES HadGEM2-A
Natural and Environmental Research Council/Met Office Hadley Centre	undeclared	HiGEM1.2
Institute of Numerical Mathematics	INM	INM-CM4

Modeling Centre(or Group)	**Institute ID**	**Model Name**
Institut Pierre-Simon Laplace	IPSL	IPSL-CM5A-LR IPSL-CM5A-MR IPSL-CM5B-LR
Japan Agency for Marine-Earth Science and Technology,Atmosphere and Ocean Research Institute(The University of Tokyo),and National Institute for Environmental Studies	MIROC	MIROC-ESM MIROC-ESM-CHEM
Atmosphere and Ocean Research Institute(The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	MIROC	MIROC4h MIROC4m MIROC5
Max Planck Institute for Meteorology	MPI-M	MPI-ESM-HR MPI-ESM-MR MPI-ESM-LR MIP-ESM-P
Meteorological Research Institute	MRI	MRI-AGCM3.2H MRI-AGCM3.2S MRI-CGCM3 MRI-ESM1
Norwegian Climate Centre	NCC	Nor-ESM1-M NorESM1-ME
-	-	tas _ Amon _ ens _ rcp85 _ 29.nc [1]

[2]

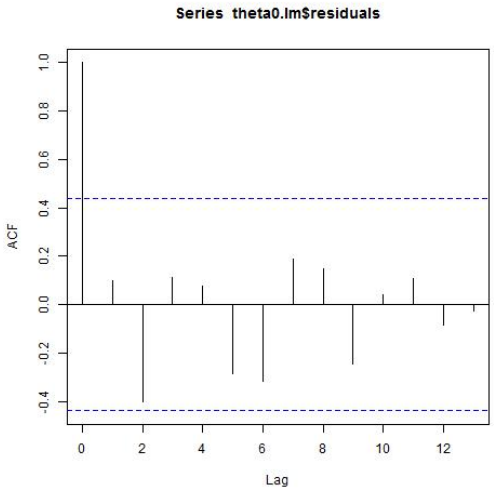
[1]From the available information in CMIP5 model description, this model is assumed to be the 29^{th} ensemble member from a multi-model ensemble, having the RCP8.5 forcing scenario as input and average monthly surface air temperature values as output.

[2]Source:`http://cmip-pcmdi.llnl.gov/cmip5/docs/CMIP5_modeling_groups.pdf`

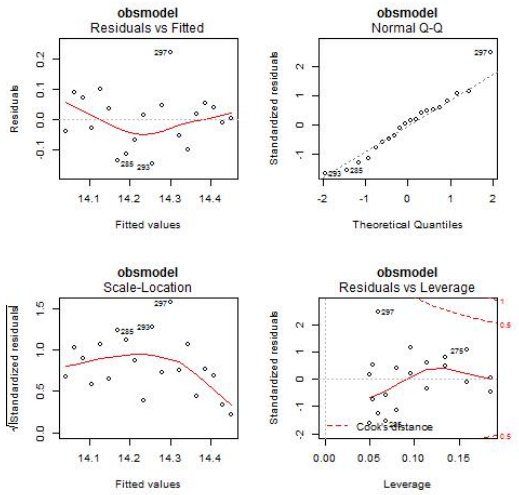
Appendix C

Residual correlograms for mimic fitted to HadCRUT3 observations and GCM outputs

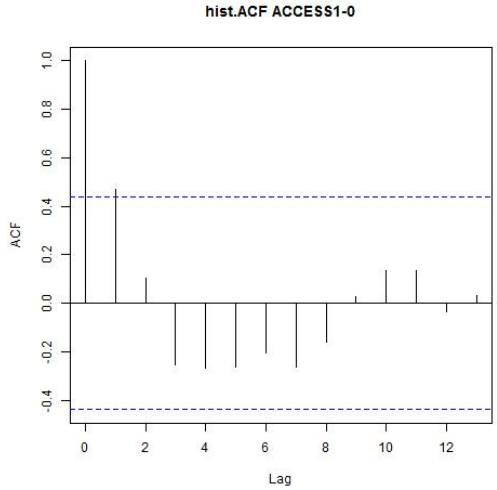
Below is a sample of diagnostic plots of the residuals, which were performed in order to check the fit of the mimic ((3.1)) to the HadCRUT3 observations (historical period) and to the CMIP5 GCM outputs (historical and future periods):



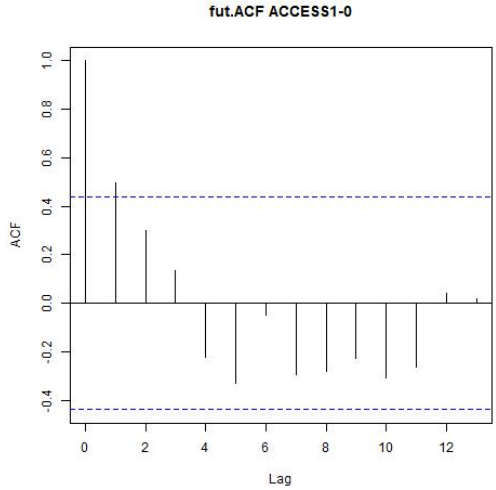
(a) Residual correlogram from fitting the mimic to HadCRUT3 observations, for the historical period 1986-2005.



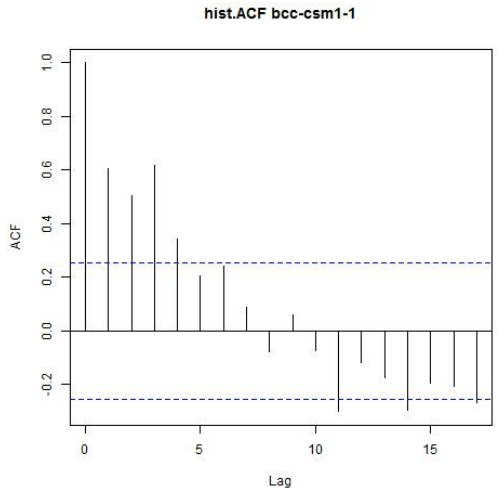
(b) Residual plots from fitting the mimic to HadCRUT3 observations, for the historical period 1986-2005.



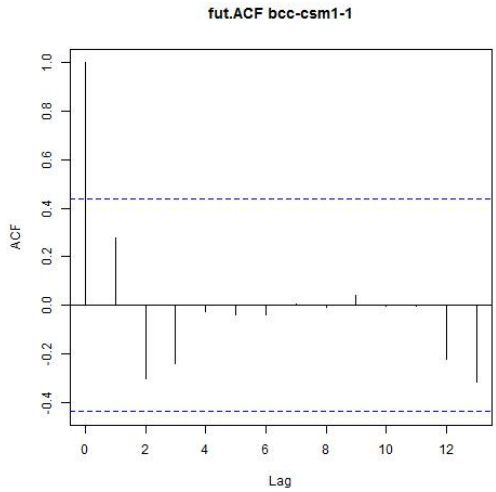
(a) Residual correlogram from fitting the mimic to $ACCESS1-0$ output, for the historical period 1986-2005.



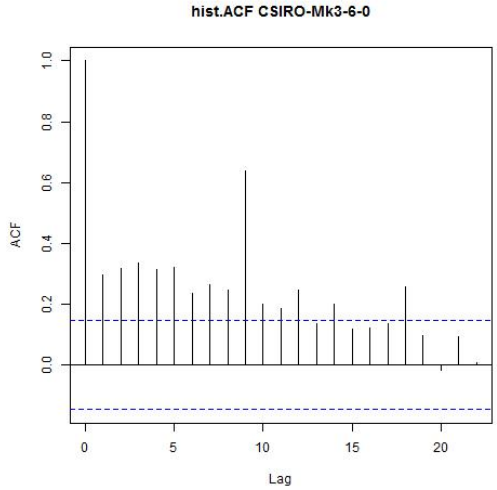
(b) Residual correlogram from fitting the mimic to $ACCESS1-0$ output, for the future period 2016-2035, under the RCP8.5 forcing scenario.



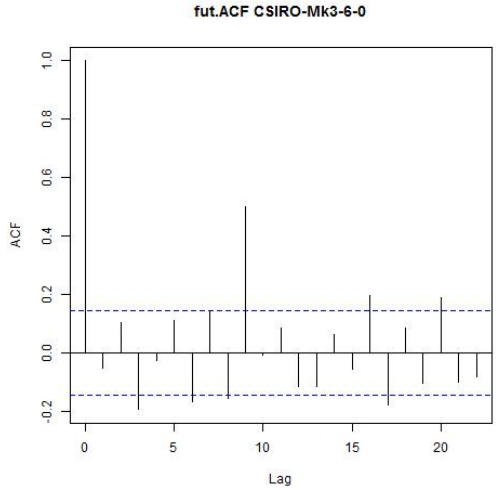
(a) Residual correlogram from fitting the mimic to $bcc-csm1-1$ output, for the historical period 1986-2005.



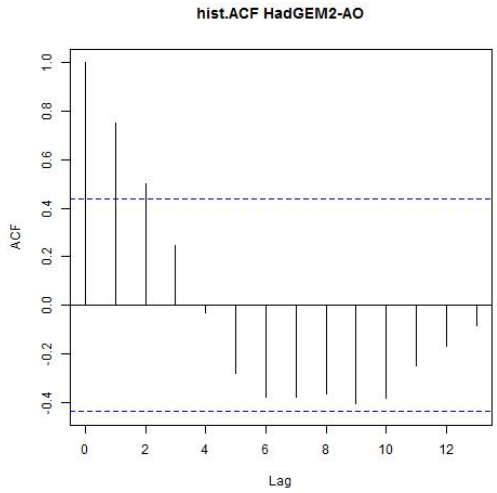
(b) Residual correlogram from fitting the mimic to $bcc-csm1-1$ output, for the future period 2016-2035, under the RCP8.5 forcing scenario.



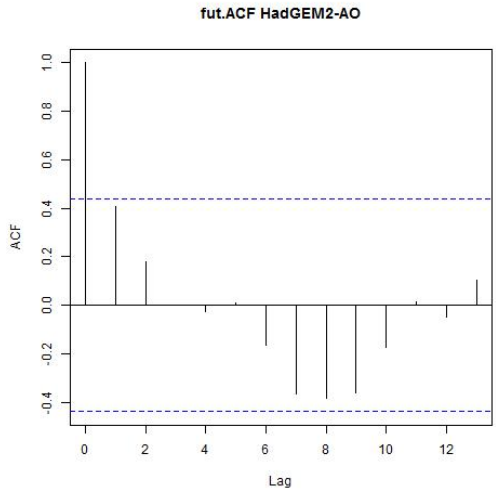
(a) Residual correlogram from fitting the mimic to $CSIRO - Mk3 - 6 - 0$ output, for the historical period 1986-2005.



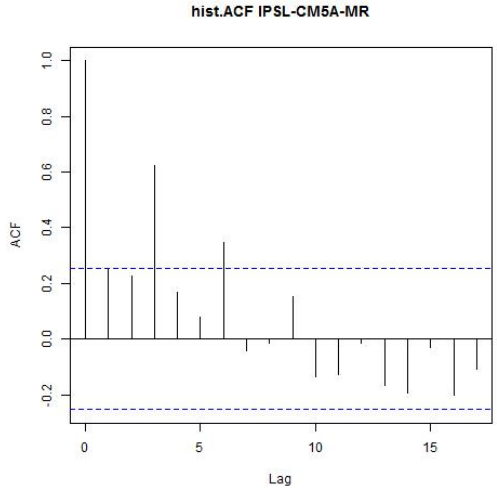
(b) Residual correlogram from fitting the mimic to $CSIRO-Mk3-6-0$ output, for the future period 2016-2035, under the RCP8.5 forcing scenario.



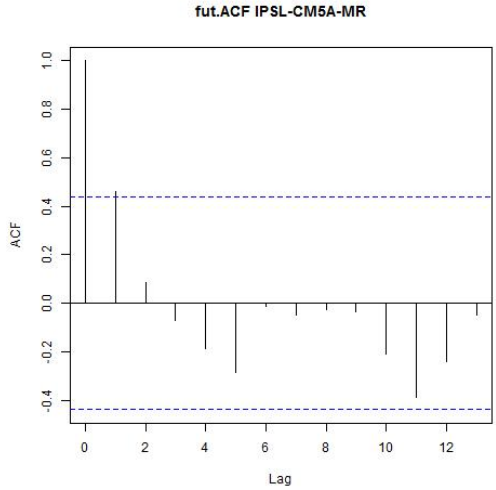
(a) Residual correlogram from fitting the mimic to $HadGEM2 - AO$ output, for the historical period 1986-2005.



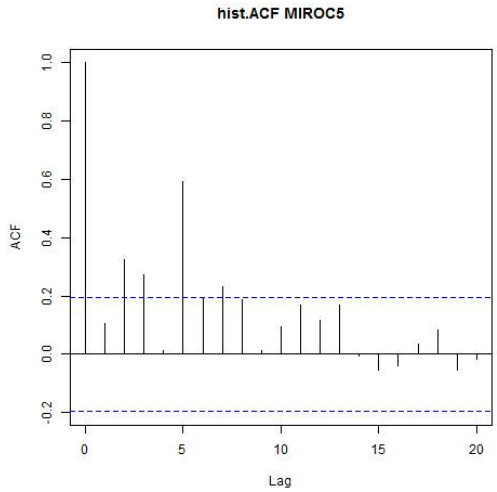
(b) Residual correlogram from fitting the mimic to $HadGEM2 - AO$ output, for the future period 2016-2035, under the RCP8.5 forcing scenario.



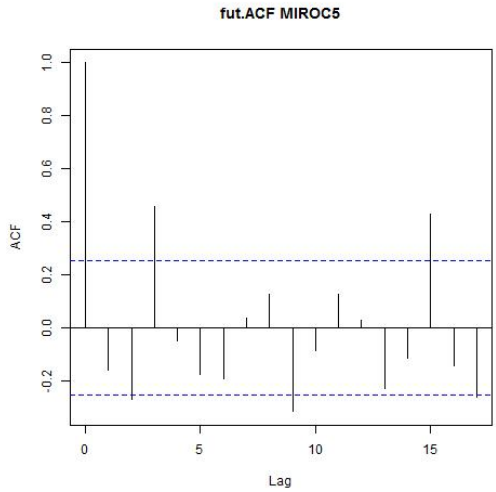
(a) Residual correlogram from fitting the mimic to $IPSL - CM5A - MR$ output, for the historical period 1986-2005.



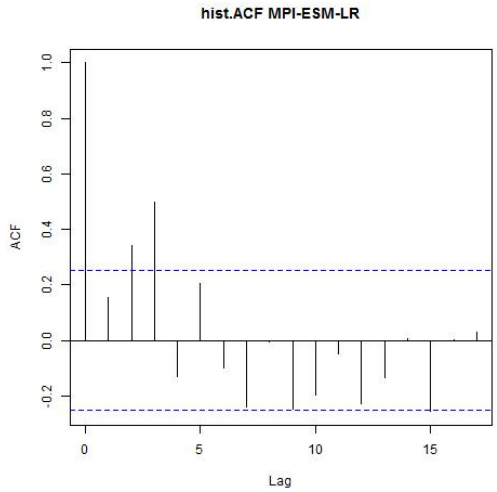
(b) Residual correlogram from fitting the mimic to $IPSL - CM5A - MR$ output, for the future period 2016-2035, under the RCP8.5 forcing scenario.



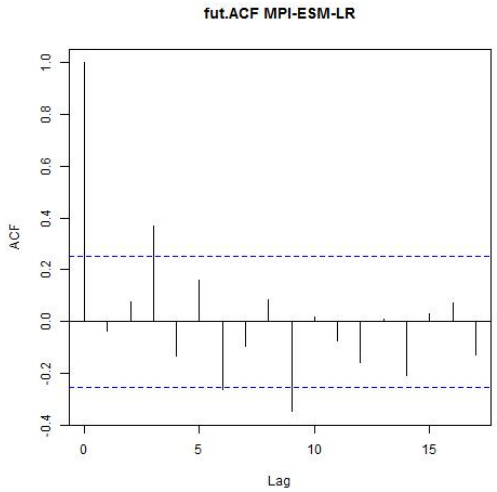
(a) Residual correlogram from fitting the mimic to $MIROC5$ output, for the historical period 1986-2005.



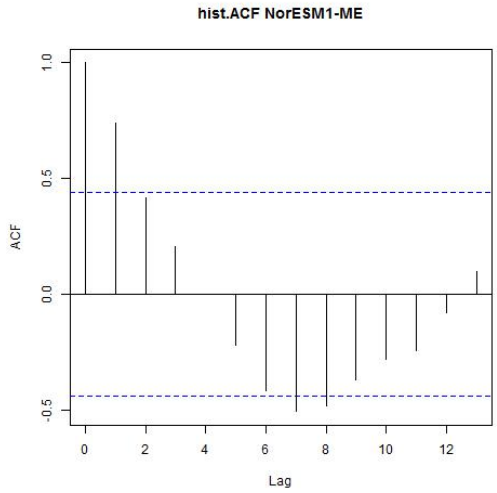
(b) Residual correlogram from fitting the mimic to $MIROC5$ output, for the future period 2016-2035, under the RCP8.5 forcing scenario.



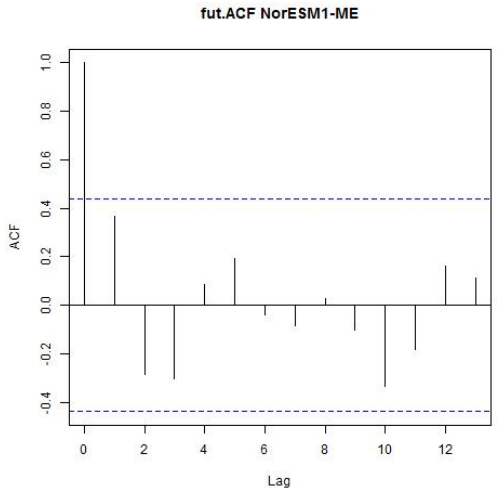
(a) Residual correlogram from fitting the mimic to $MPI-ESM-LR$ output, for the historical period 1986-2005.



(b) Residual correlogram from fitting the mimic to $MPI-ESM-LR$ output, for the future period 2016-2035, under the RCP8.5 forcing scenario.



(a) Residual correlogram from fitting the mimic to $NorESM1-ME$ output, for the historical period 1986-2005.



(b) Residual correlogram from fitting the mimic to $NorESM1-ME$ output, for the future period 2016-2035, under the RCP8.5 forcing scenario.

Appendix D

Scatter plots of historical Vs future descriptor estimates $\hat{\alpha}$ and $\hat{\beta}$

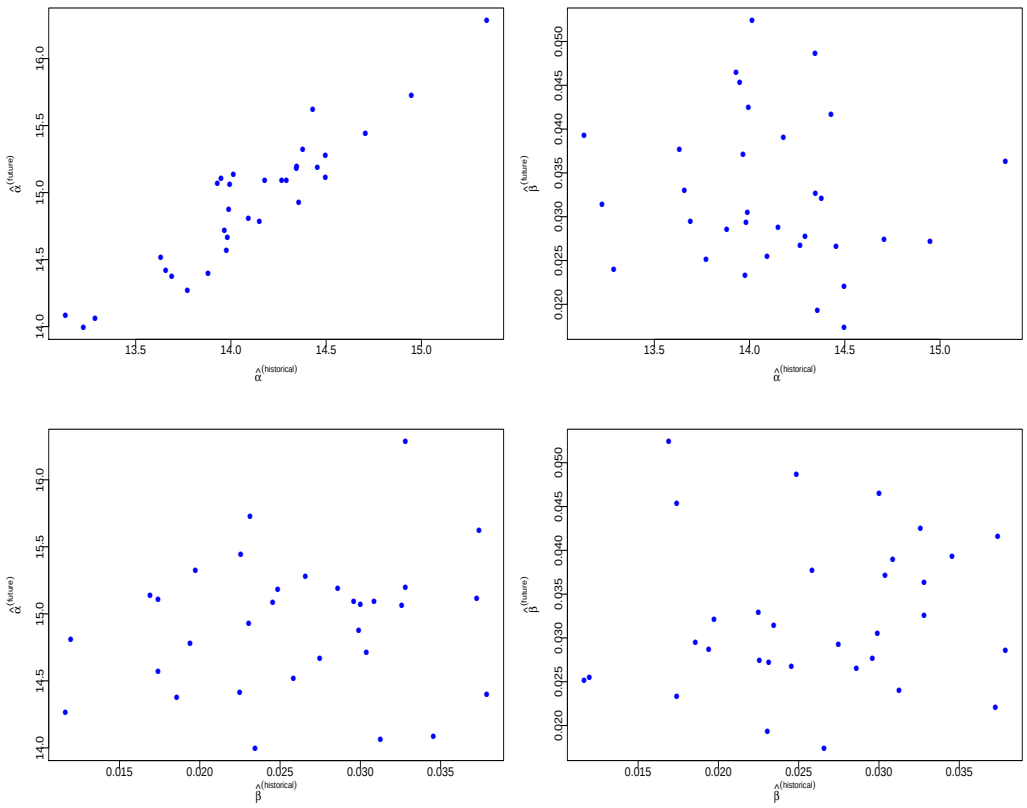


Figure D.1: Scatter plots of the historical versus future components of the descriptor estimates $\hat{\boldsymbol{\alpha}}_i$ and $\hat{\boldsymbol{\beta}}_i$, for $i = 1, \ldots, 32$.

Appendix E

Comparison of distributions between the two fully Bayesian implementations

As discussed in Section 3.5.3, the values for the shape parameters of the distributions of v_3, v_4, γ_3 and γ_4 are chosen so as to retain consistency in the *a priori* judgements about simulator consensus and shared simulator discrepancy of σ^2, between the two fully Bayesian implementations. In practice, this is achieved by comparing the relevant distributions in the two implementations, as shown below:

- Specify the shape parameter ρ_1 of the distribution of v_3.

  Interest is on specifying a value for the shape parameter ρ_1 in the distribution of v_3 ((3.21)), where v_3 is the shape parameter of the distribution of $\psi_i^{(\text{hist})} = {}^{1}/_{\sigma_i^{2\ (\text{hist})}}$ ((3.19)).

  According to (3.19),

  $$ {}^{\psi_i^{(\text{hist})}}/_{\psi_0^{(\text{hist})}} | \omega_{\psi^{(\text{hist})}} \sim Gamma\left(v_3, \frac{v_3}{\omega_{\psi^{(\text{hist})}}}\right) (i = 1, .., m). $$

  Therefore, $\psi_i^{(\text{hist})}$ can be expressed as:

  $$ \psi_i^{(\text{hist})} = \omega_{\psi^{(\text{hist})}} \times \psi_0^{(\text{hist})} \times Y, $$

  where $Y \sim Gamma(v_3, v_3)$.

  To retain consistency in the *a priori* judgements about simulator consensus for historical residual variance/precision between the fully Bayesian implemen-

tations, the distribution of $\psi_i^{(\text{hist})}$ in the second version of the full Bayesian implementation is compared to the marginal distribution of $\log\left(\sigma_i^{2\ (\text{hist})}\right)$ in the GFB implementation. The latter is determined as:

$$\log\left(\sigma_i^{2\ (\text{hist})}\right)|\omega_{\log\left(\sigma_i^{2\ (\text{hist})}\right)}, \log\left(\sigma_0^{2\ (\text{hist})}\right) \sim N\left(\log\left(\sigma_0^{2\ (\text{hist})}\right) + \omega_{\log\left(\sigma_i^{2\ (\text{hist})}\right)}, \boldsymbol{C}[3,3]\right),$$

where $\omega_{\log\left(\sigma_i^{2\ (\text{hist})}\right)}$ is defined as the shared simulator discrepancy of $\log\left(\sigma_i^{2\ (\text{hist})}\right)$ from $\log\left(\sigma_0^{2\ (\text{hist})}\right)$.

This allows expressing $\log\left(\sigma_i^{2\ (\text{hist})}\right)$ as:

$$\log\left(\sigma_i^{2\ (\text{hist})}\right) = \omega_{\log\left(\sigma_i^{2\ (\text{hist})}\right)} + \log\left(\sigma_0^{2\ (\text{hist})}\right) + Z,$$

where $Z \sim N(0, \boldsymbol{C}[3,3])$.

The next step is to determine the relation between Y and Z, in order to compare the relevant distributions from the two implementations. Some straightforward manipulations show that $Y = \exp(-Z)$. It remains to sample from both Y and $\exp(-Z)$ and specify the value of ρ_1 that gives similar distributions of Y and $\exp(-Z)$. The procedure for a particular choice of ρ_1 is briefly described below:

1. Sample 1000 values from $v_3 \sim Gamma\left(\rho_1, \frac{\rho_1}{\eta_1}\right)$ ((3.21)), where η_1 is already specified (See Section 3.5.3).
2. Using the sampled values of v_3, sample 1000 values from $Y \sim Gamma(v_3, v_3)$.
3. Sample 1000 values from the distribution of $\boldsymbol{C}^{-1}[3,3]$ (see (3.15)) in the GFB implementation.

   The distribution of $\boldsymbol{C}^{-1}[3,3]$ can be obtained from the marginal distribution of $\boldsymbol{C}^{-1}$. According to Rao (1965, p. 452), this is $\boldsymbol{R}_1[3,3] \times \chi^2_{v_1}$, $\boldsymbol{R}_1$ and v_1 being the scale matrix and degrees of freedom of the distribution of $\boldsymbol{C}$ respectively (see (3.15)).
4. Using the sampled values of $\boldsymbol{C}^{-1}[3,3]$, sample 1000 values from $Z \sim N(0, {}^1/_{\boldsymbol{C}^{-1}[3,3]})$.
5. Produce boxplots of the sampled distributions from Y and $\exp(-Z)$ and compare them.

Repeat the above procedure for a range of values of ρ_1 and choose a value that provides similar boxplots and therefore, sampled distributions of Y and $\exp(-Z)$. It

turns out that a choice of $\rho_1 = 1$ gives similar sampled distributions for the comparable variables. Those are shown in Figure E.1.

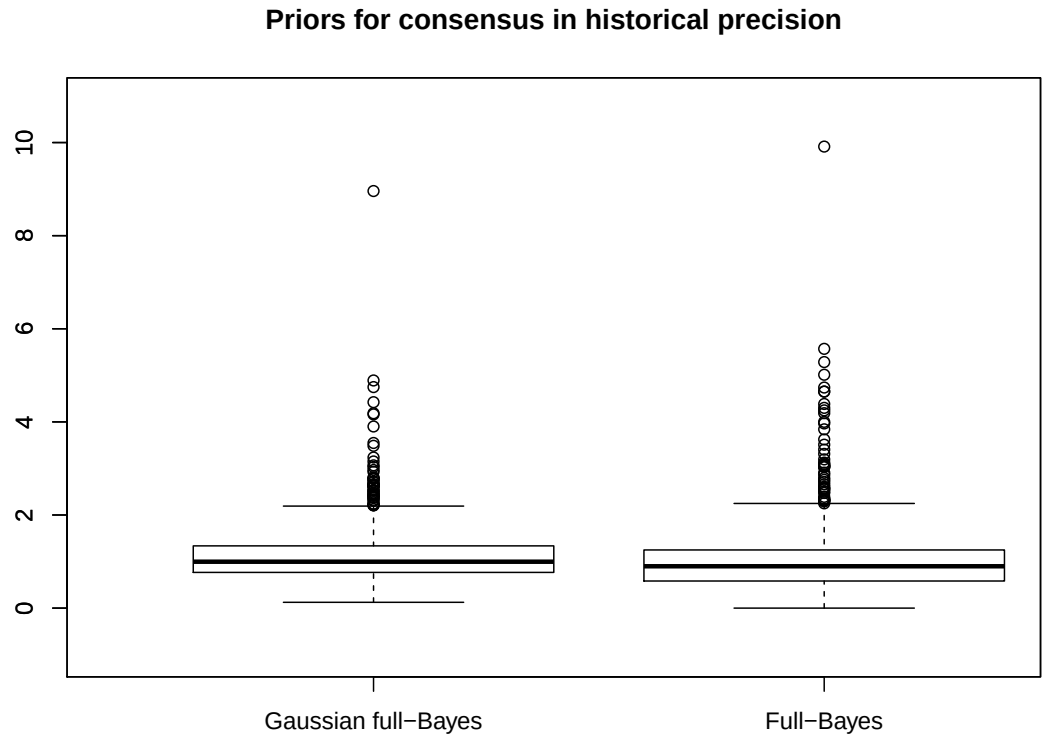


Figure E.1: Boxplots for comparing sampling distributions of $\exp(-Z)$ and Y in the GFB (left) and FB (right) implementations respectively, with $\rho_1 = 1$.

By following a similar procedure, the value of the shape parameter ρ_2 of the distribution of v_4 ((3.22)), which is relevant to the simulator consensus about $\psi_i^{(\text{fut})}/\psi_i^{(\text{hist})}$ ((3.20)), is also chosen to be 1. The relevant boxplots are shown in Figure E.2.

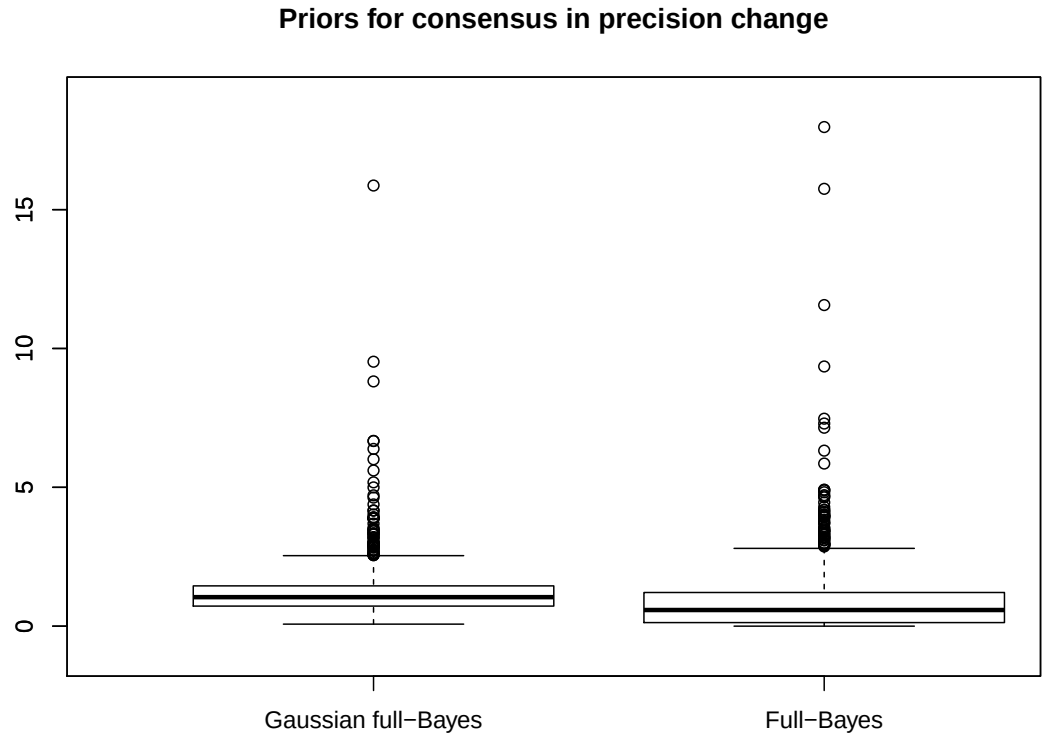


Figure E.2: Boxplots for comparing sampling distributions relevant to $\log\left(\sigma_i^{2\ (\text{fut})}/\sigma_i^{2\ (\text{hist})}\right)$ and $\psi_i^{(\text{fut})}/\psi_i^{(\text{hist})}$ in the GFB (left) and FB (right) implementations respectively, with $\rho_2 = 1$.

It remains to show how values for the shape parameters ρ_3 and ρ_4 of γ_3 and γ_4 ((3.27)-(3.28)) are chosen, where $\gamma_3 = Var^{-1}(\omega_{\psi^{(\text{hist})}})$ and $\gamma_4 = Var^{-1}(\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}})$ (see (3.25)-(3.26)). As also discussed in Section 3.5.3, the shape parameters are chosen so as to retain consistency in the *a priori* judgements about shared simulator discrepancies $\omega_{\psi^{(\text{hist})}}$ and $\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}}$, between the two fully Bayesian implementations. The procedure for specifying a value for ρ_3 is demonstrated below:

- Specify the shape parameter ρ_3 of the distribution of γ_3.

  Interest is on specifying a value for the shape parameter ρ_3 in the distribution of γ_3 ((3.27)), where $\gamma_3 = Var^{-1}(\omega_{\psi^{(\text{hist})}})$ ((3.25)).

  Precisely, according to (3.25),

  $$\omega_{\psi^{(\text{hist})}} \sim Gamma\left(\gamma_3, \gamma_3\right).$$

  To retain consistency in the *a priori* judgements about shared simulator discrepancy for historical residual variance/precision between the fully Bayesian

implementations, the distribution of $\omega_{\psi^{(\text{hist})}}$ in FB implementation is compared to the marginal distribution of $\omega_{\log\left(\sigma^{2\ (\text{hist})}\right)}$ in the GFB implementation. The latter can be obtained from the distribution of $\boldsymbol{\omega}$ in the GFB implementation, which suggests that $\omega_{\log\left(\sigma^{2\ (\text{hist})}\right)}$ is distributed as:

$$\omega_{\log\left(\sigma^{2\ (\text{hist})}\right)} \sim N(\mathbf{0}, \boldsymbol{\Lambda}[3,3]).$$

For convenience, the following notation is introduced:

$$Y := \omega_{\psi^{(\text{hist})}},$$

and

$$Z := \omega_{\log\left(\sigma^{2\ (\text{hist})}\right)}.$$

The next step is to determine the relation between Y and Z, in order to compare the relevant distributions from the two implementations. Some straightforward manipulations show that $1/Y = \exp(Z)$. It remains to sample from both $1/Y$ and $\exp(Z)$ and specify the value of ρ_3 that gives similar distributions of $1/Y$ and $\exp(Z)$. The procedure for a particular choice of ρ_3 is briefly described below:

1. Sample 1000 values from $\gamma_3 \sim Gamma\left(\rho_3, \frac{\rho_3}{\eta_3}\right)$ ((3.27)), where η_3 is already specified (See Section 3.5.3).
2. Using the sampled values of γ_3, sample 1000 values from $Y \sim (\gamma_3, \gamma_3)$.
3. Sample 1000 values from the distribution of $\boldsymbol{\Lambda}^{-1}[3,3]$ ((3.16)) in the GFB implementation.

   The distribution of $\boldsymbol{\Lambda}^{-1}[3,3]$ can be obtained from the marginal distribution of $\boldsymbol{\Lambda}^{-1}$. According to Rao (1965, p. 452), this is $\boldsymbol{R}_2[3,3] \times \chi^2_{v_2}$, $\boldsymbol{R}_2$ and v_2 being the scale matrix and degrees of freedom of the distribution of $\boldsymbol{\Lambda}$ respectively (see (3.16)).
4. Using the sampled values of $\boldsymbol{\Lambda}^{-1}[3,3]$, sample 1000 values from $Z \sim N(0, 1/\boldsymbol{\Lambda}[3,3])$.
5. Produce boxplots of the sampled distributions from $1/Y$ and $\exp(Z)$ and compare them.

Repeat the above procedure for a range of values of ρ_3 and choose a value that provides similar boxplots and therefore, sampled distributions of $1/Y$ and $\exp(Z)$.

It turns out that a choice of $\rho_3 = 10$ gives similar sampled distributions for the comparable variables. Those are shown in Figure E.3.

By following a similar procedure, the value of the shape parameter ρ_4 of the distribution of γ_4 ((3.28)), which is relevant to the shared simulator discrepancy $\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}}$ ((3.26)) from $\omega_{\psi_0^{(\text{fut})}/\psi_0^{(\text{hist})}}$, is also chosen to be 10. The relevant boxplots are shown in Figure E.4.

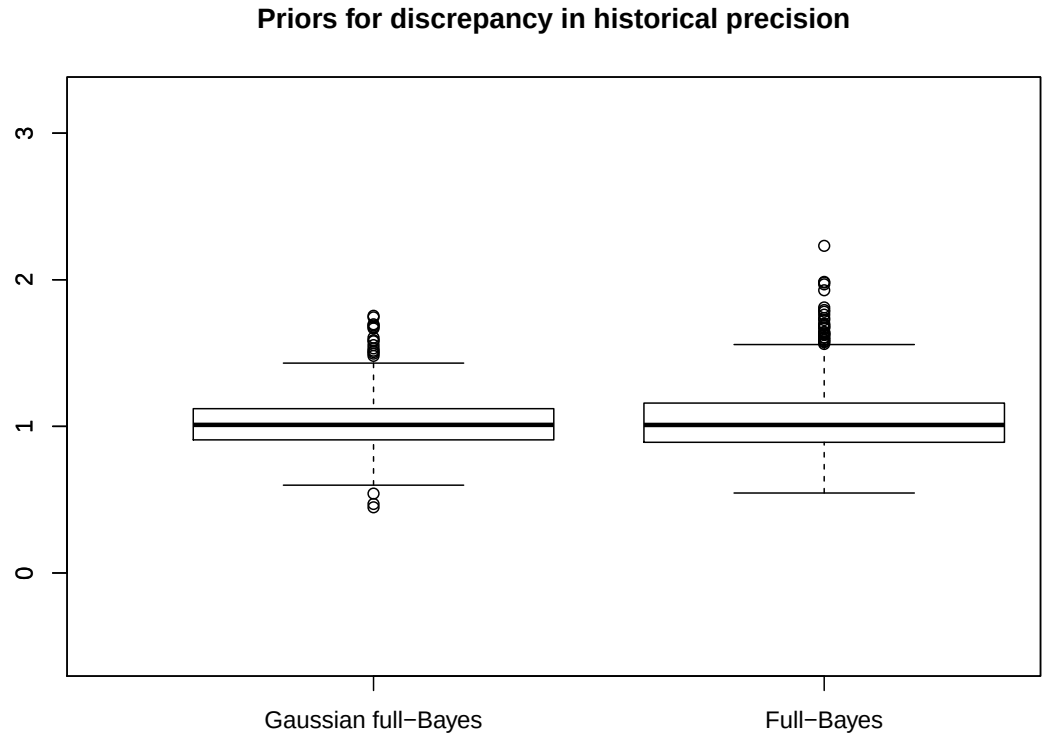


Figure E.3: Boxplots for comparing sampling distributions of $\exp\left(\omega_{\log\left(\sigma^{2\ (\text{hist})}\right)}\right)$ and $1/\omega_{\psi^{(\text{hist})}}$ in the GFB (left) and FB (right) implementations respectively, with $\rho_3 = 10$.

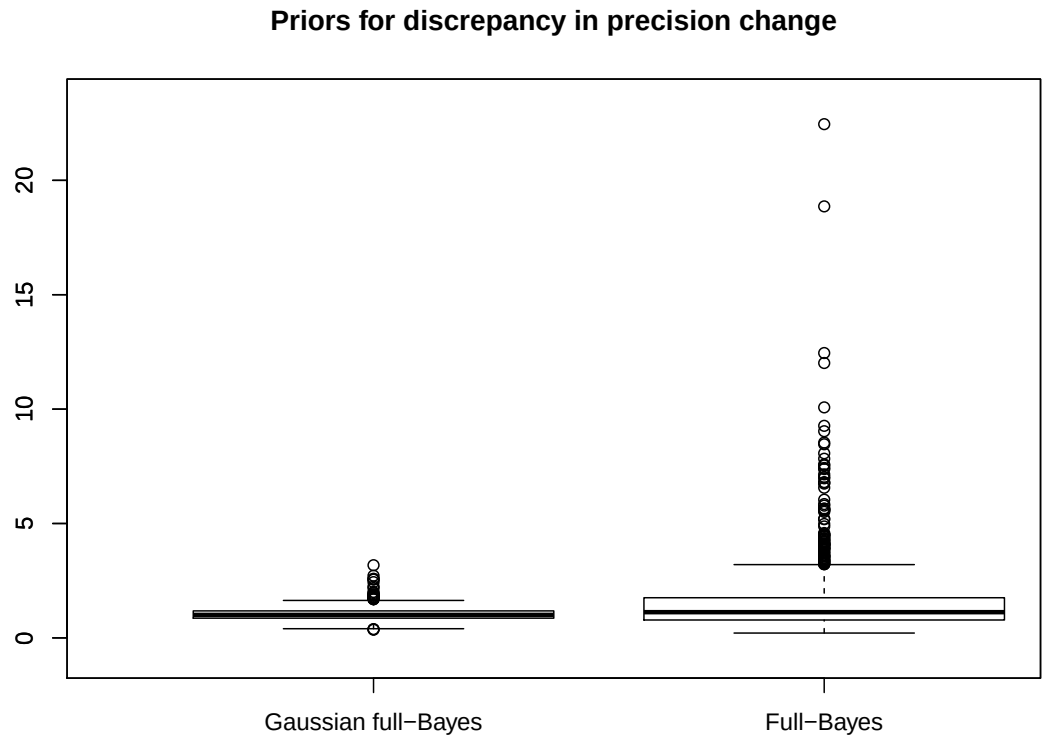


Figure E.4: Boxplots for comparing sampling distributions of $\exp\left(\omega_{\log(\sigma^{2\,(\mathrm{fut})}/\sigma^{2\,(\mathrm{hist})})}\right)$ and $1/\omega_{\psi^{(\mathrm{fut})}/\psi^{(\mathrm{hist})}}$ in the GFB (left) and FB (right) implementations respectively, with $\rho_4 = 10$.

Appendix F

Supplement to the MCMC implementation under the simpler framework

Tables F.1-F.2 present summary statistics (mean, standard deviation, quantiles), the Gelman-Rubin diagnostic ($\hat{R}$) and the effective sample size (n.eff) of the MCMC samples for the posterior parameters of $\boldsymbol{\theta}_0$, under the GFB and FB implementations respectively. The results are based on the samples generated from 4 MCMC chains, giving a total of 1200000 iterations after burn-in. The deviance information criterion (DIC) is -590.8 and -1324 with corresponding estimated number of parameters equal to 129.3 and 164.2, for the GFB and FB implementations respectively.

	mean	**sd**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**	$\hat{R}$	**n.eff**
$\alpha_0^{(\text{hist})}$	14.2	0.0	14.2	14.2	14.2	14.3	14.3	1	110000
$\beta_0^{(\text{hist})}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	290
$\log\left(\sigma_0^{2\ (\text{hist})}\right)$	-4.4	0.2	-4.9	-4.6	-4.4	-4.3	-4.0	1	10000
$\alpha_0^{(\text{fut})} - \alpha_0^{(\text{hist})}$	0.6	0.2	0.3	0.5	0.7	0.8	1.0	1	1900
$\beta_0^{(\text{fut})} - \beta_0^{(\text{hist})}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	75
$\log\left(\sigma_0^{2\ (\text{fut})}/\sigma_0^{2\ (\text{hist})}\right)$	-0.5	0.6	-1.5	-0.8	-0.5	-0.1	0.7	1	130
Deviance	-720.1	17.7	-753.3	-732.3	-720.7	-708.4	-683.8	1	180000

Table F.1: Summary statistics (mean, standard deviation, quantiles) and MCMC convergence diagnostics ($\hat{R}$ and effective sample size) for the posterior parameters and deviance, under the GFB implementation.

	mean	**sd**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**	$\hat{R}$	**n.eff**
$\alpha_0^{(\text{hist})}$	14.2	0.0	14.2	14.2	14.2	14.3	14.3	1	1200000
$\beta_0^{(\text{hist})}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	410000
$\log\left(\sigma_0^{2\ (\text{hist})}\right)$	-4.2	0.2	-4.7	-4.4	-4.2	-4.1	-3.8	1	8300
$\alpha_0^{(\text{fut})} - \alpha_0^{(\text{hist})}$	0.8	0.2	0.4	0.7	0.8	0.9	1.1	1	5300
$\beta_0^{(\text{fut})} - \beta_0^{(\text{hist})}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	4000
$\log\left(\sigma_0^{2\ (\text{fut})}/\sigma_0^{2\ (\text{hist})}\right)$	-0.7	0.6	-2.0	-1.1	-0.7	-0.3	0.3	1	5800
Deviance	-1488.0	20.6	-1527.0	-1502.0	-1489.0	-1474.0	-1446.0	1	81000

Table F.2: Summary statistics (mean, standard deviation, quantiles) and MCMC convergence diagnostics ($\hat{R}$ and effective sample size) for the posterior parameters and deviance, under the FB implementation.

Appendix G

Derivation of $E(S_G)$ and $E(S_E)$

This appendix shows the derivation of $E(S_G)$ and $E(S_E)$ required for estimating the fixed parameters of the random effects model defined in (4.25)-(4.28) of Section 4.7.1.1.

It holds that,

$$\begin{aligned}
E(S_G) &= E\left(\sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}_{..})^2\right) \\
&= E\left(\sum_{i=1}^{k} n_i (\mu + \alpha_i + \bar{\epsilon}_{i.} - \mu - \bar{\alpha} - \bar{\epsilon}_{..})^2\right) \text{ using (4.44 and 4.45} \\
&= E\left(\sum_{i=1}^{k} n_i \{(\alpha_i - \bar{\alpha}) + (\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})\}^2\right) \\
&= E\left(\sum_{i=1}^{k} n_i (\alpha_i - \bar{\alpha})^2\right) + 2E\left(\sum_{i=1}^{k} n_i (\alpha_i - \bar{\alpha})(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})\right) + E\left(\sum_{i=1}^{k} n_i (\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})^2\right).
\end{aligned} \tag{G.1}$$

The first term in the RHS of (G.1) can be expressed as:

$$\begin{aligned}
\sum_{i=1}^{k} n_i E\left[(\alpha_i - \bar{\alpha})^2\right] &= \sum_{i=1}^{k} n_i E\left[\alpha_i^2 - 2\alpha_i\bar{\alpha} + \bar{\alpha}^2\right] \\
&= \sum_{i=1}^{k} n_i \left[E\left(\alpha_i^2\right) - 2E\left(\alpha_i\bar{\alpha}\right) + E\left(\bar{\alpha}^2\right)\right] \\
&= \sum_{i=1}^{k} n_i \left[Var\left(\alpha_i\right) - 2E\left(\alpha_i \frac{\sum_{i=1}^{k}\alpha_i}{k}\right) + Var\left(\frac{\sum_{i=1}^{k}\alpha_i}{k}\right)\right], \\
&\qquad\qquad \text{since } E\left(\alpha_i\right) = 0 \\
&= \sum_{i=1}^{k} n_i \left[\sigma_\alpha^2 - \frac{2}{k}E\left(\alpha_i^2\right) + \frac{1}{k^2}\sum_{i=1}^{k} Var\left(\alpha_i\right)\right], \text{since } \alpha_i \perp\!\!\!\perp \alpha_j \\
&= \sum_{i=1}^{k} n_i \left[\sigma_\alpha^2 - \frac{2}{k}\sigma_\alpha^2 + \frac{1}{k}\sigma_\alpha^2\right] \\
&= \frac{N(k-1)}{k}\sigma_\alpha^2. \qquad \text{(G.2)}
\end{aligned}$$

The second term in the RHS of (G.1) can be written as

$$2\sum_{i=1}^{k} n_i E\left(\alpha_i - \bar{\alpha}\right) E\left(\epsilon_{i.} - \bar{\epsilon}_{..}\right),$$

since $(\alpha_i - \bar{\alpha}) \perp\!\!\!\perp (\epsilon_{i.} - \bar{\epsilon}_{..})$ according to Section 4.6.

The expectation $E\left(\epsilon_{i.} - \bar{\epsilon}_{..}\right)$ is zero, since $E\left(\epsilon_{ij}\right) = 0$ according to (4.27). Therefore the corresponding expression in (G.1) vanishes.

Finally, the third term in the RHS of (G.1) can be expressed as a sum of three terms:

$$\begin{aligned}
E\left(\sum_{i=1}^{k} n_i\left(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..}\right)^2\right) &= E\left(\sum_{i=1}^{k} n_i \bar{\epsilon}_{i.}^2\right) - 2E\left(\sum_{i=1}^{k} n_i \bar{\epsilon}_{i.}\bar{\epsilon}_{..}\right) + E\left(\sum_{i=1}^{k} n_i \bar{\epsilon}_{..}^2\right) \\
&= \sum_{i=1}^{k} n_i E\left(\bar{\epsilon}_{i.}^2\right) - 2E\left(\bar{\epsilon}_{..}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\epsilon_{ij}\right) + \sum_{i=1}^{k} n_i E\left(\bar{\epsilon}_{..}^2\right) \\
&= \sum_{i=1}^{k} n_i E\left(\bar{\epsilon}_{i.}^2\right) - 2NE\left(\bar{\epsilon}_{..}^2\right) + NE\left(\bar{\epsilon}_{..}^2\right) \\
&= \sum_{i=1}^{k} \left\{n_i E\left(\bar{\epsilon}_{i.}^2\right)\right\} - NE\left(\bar{\epsilon}_{..}^2\right). \qquad \text{(G.3)}
\end{aligned}$$

In the first term, it holds that

$$\begin{aligned} E\left(\bar{\epsilon}_{i.}^2\right) &= E\left[E\left(\bar{\epsilon}_{i.}^2|\sigma_i^2\right)\right], \text{ using the law of iterated expectation} \\ &= E\left[Var\left(\bar{\epsilon}_{i.}|\sigma_i^2\right) + E^2\left(\bar{\epsilon}_{i.}|\sigma_i^2\right)\right] \\ &= E\left[Var\left(\bar{\epsilon}_{i.}|\sigma_i^2\right)\right], \end{aligned} \tag{G.4}$$

since $E\left(\epsilon_{ij}|\sigma_i^2\right) = 0$ implies $E\left(\bar{\epsilon}_{i.}|\sigma_i^2\right) = 0$.
Because $\epsilon_{ij}|\sigma_i^2$ are i.i.d. $\forall j$, distributed according to (4.27) in Section 4.7.1.1,

$$\begin{aligned} E\left[Var\left(\bar{\epsilon}_{i.}|\sigma_i^2\right)\right] &= E\left[Var\left(\sum_{j=1}^{n_i}\frac{1}{n_i}\epsilon_{ij}|\sigma_i^2\right)\right] \\ &= E\left[\frac{1}{n_i^2}\sum_{j=1}^{n_i}Var\left(\epsilon_{ij}|\sigma_i^2\right)\right] \text{ since } \epsilon_{ij}|\sigma_i^2 \text{ are independent } \forall j \\ &= E\left(\frac{1}{n_i}\sigma_i^2\right), \text{ since } Var\left(\epsilon_{ij}|\sigma_i^2\right) = \sigma_i^2, \text{ according to (4.27)} \\ &= \frac{\xi}{n_i}, \text{ from (4.24).} \end{aligned} \tag{G.5}$$

The first term in (G.3) is thus equal to $\sum_{i=1}^{k} {}^{n_i\xi}/_{n_i} = k\xi$. Turning now to the second term, it holds that

$$\begin{aligned} E\left(\bar{\epsilon}_{..}^2\right) &= E\left[E\left(\bar{\epsilon}_{..}^2|\sigma_1^2,\ldots,\sigma_k^2\right)\right] \text{ using the law of iterated expectation} \\ &= E\left[Var\left(\bar{\epsilon}_{..}|\sigma_1^2,\ldots,\sigma_k^2\right) + E^2\left(\bar{\epsilon}_{..}|\sigma_1^2,\ldots,\sigma_k^2\right)\right] \\ &= E\left[Var\left(\bar{\epsilon}_{..}|\sigma_1^2,\ldots,\sigma_k^2\right)\right] \text{ since } E\left(\bar{\epsilon}_{..}|\sigma_1^2,\ldots,\sigma_k^2\right) = 0 \\ &= E\left[\frac{1}{N^2}\sum_{i=1}^{k}n_i^2 Var\left(\bar{\epsilon}_{i.}|\sigma_1^2,\ldots,\sigma_k^2\right)\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{k}n_i^2\frac{\xi}{n_i}, \text{ from (G.5)} \\ &= \frac{\xi}{N}. \end{aligned} \tag{G.6}$$

Returning to (G.3) gives,

$$E\left(\sum_{i=1}^{k} n_i \left(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..}\right)^2\right) = \sum_{i=1}^{k}\left\{n_i \frac{\xi}{n_i}\right\} - N\frac{\xi}{N}$$
$$= (k-1)\xi, \tag{G.7}$$

and combining this with (G.2), (G.1) becomes

$$E\left(S_G\right) = \frac{(k-1)}{k}\left(N\sigma_\alpha^2 + k\xi\right). \tag{G.8}$$

The next step is to derive $E\left(S_E\right)$, the expected value of within-group variability defined at (4.43). It holds that,

$$\begin{aligned} E\left(S_E\right) &= E\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2\right) \\ &= E\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\epsilon_{ij} - \bar{\epsilon}_{i.}\right)^2\right), \text{ using (4.25) and (4.44)} \\ &= \sum_{i=1}^{k} E\left(\sum_{j=1}^{n_i}\left(\epsilon_{ij} - \bar{\epsilon}_{i.}\right)^2\right) \\ &= \sum_{i=1}^{k}(n_i - 1)E\left(s_i^2\right), \end{aligned} \tag{G.9}$$

where s_i^2 is the unbiased sample variance of the $\{\epsilon_{ij},\ j = 1, \ldots, n_i\}$.
Using the law of iterated expectation,

$$\begin{aligned} E\left(s_i^2\right) &= E\left(E\left(s_i^2|\sigma_i^2\right)\right) \\ &= E\left(\sigma_i^2\right) \\ &= \xi, \text{ as defined earlier.} \end{aligned} \tag{G.10}$$

Therefore, $E\left(S_E\right)$ in (G.9) becomes,

$$\begin{aligned} E\left(S_E\right) &= \sum_{i=1}^{k}(n_i - 1)\xi \\ &= (N-k)\xi. \end{aligned} \tag{G.11}$$

Appendix H

Estimation of v in the extended framework(multivariate data)

According to Section 4.7.2, it is expected that $E\left[tr\left(\sum_{i=1}^{k}(n_i-1)^2\boldsymbol{S}_i\boldsymbol{S}_i'\right)\right]$ involves v. Here, it is proved that this statement is true, which then allows the derivation of a moment estimator of v, defined in (H.15)-(H.17).

In mathematical terms, it is firstly required to derive,

$$E\left(tr\left(\sum_{i=1}^{k}(n_i-1)^2\boldsymbol{S}_i\boldsymbol{S}_i'\right)\right)=\sum_{i=1}^{k}(n_i-1)^2E\left(tr\left(\boldsymbol{S}_i\boldsymbol{S}_i'\right)\right). \tag{H.1}$$

In fact, it is easy to prove that: $E\left(tr\left(\boldsymbol{S}_i\boldsymbol{S}_i'\right)\right)=tr\left(E\left(vec\left(\boldsymbol{S}_i\right)vec\left(\boldsymbol{S}_i\right)'\right)\right)$.

Proof. $\boldsymbol{S}_i\boldsymbol{S}_i'$ has $(m,n)^{th}$ element, $\sum_{k=1}^{p}s_{mk}s_{kn}=\sum_{k=1}^{p}s_{mk}s_{nk}$, since $\boldsymbol{S}_i$ is symmetric. Therefore,

$$tr\left(\boldsymbol{S}_i\boldsymbol{S}_i'\right)=\sum_{l=1}^{p}\left(\boldsymbol{S}_i\boldsymbol{S}_i'\right)_{ll}=\sum_{l=1}^{p}\sum_{k=1}^{p}s_{lk}s_{lk}=\sum_{l=1}^{p}\sum_{k=1}^{p}s_{lk}^2,$$

which gives,

$$E\left(tr\left(\boldsymbol{S}_i\boldsymbol{S}_i'\right)\right)=E\left(\sum_{l=1}^{p}\sum_{k=1}^{p}s_{lk}^2\right). \tag{H.2}$$

On the other hand, $vec\left(\boldsymbol{S}_i\right)$ is defined as

$$vec\left(\boldsymbol{S}_i\right)=\begin{bmatrix}\boldsymbol{s}_1\\ \boldsymbol{s}_1\\ \vdots\\ \boldsymbol{s}_p\end{bmatrix},$$

where $\{\boldsymbol{s}_l,\ l = 1, \ldots, p\}$ correspond to the columns of $\boldsymbol{S}_i$.

Therefore, the $(l,l)^{th}$ block of $vec\left(\boldsymbol{S}_i\right) vec\left(\boldsymbol{S}_i\right)'$ is $\boldsymbol{s}_l \boldsymbol{s}_l'$, with $(k,k)^{th}$ element $s_{lk} s_{lk}'$. Thus,

$$tr\left(vec\left(\boldsymbol{S}_i\right) vec\left(\boldsymbol{S}_i\right)'\right) = \sum_{l=1}^{p} tr\left(\boldsymbol{s}_l \boldsymbol{s}_l'\right) = \sum_{l=1}^{p}\sum_{k=1}^{p} s_{lk}^2. \tag{H.3}$$

Consequently,

$$\begin{aligned} tr\left(E\left(vec\left(\boldsymbol{S}_i\right) vec\left(\boldsymbol{S}_i\right)'\right)\right) &= E\left(tr\left(vec\left(\boldsymbol{S}_i\right) vec\left(\boldsymbol{S}_i\right)'\right)\right) \\ &= E\left(\sum_{l=1}^{p}\sum_{k=1}^{p} s_{lk}^2\right), \end{aligned} \tag{H.4}$$

which according to (H.2) is equal to $E\left(tr\left(\boldsymbol{S}_i \boldsymbol{S}_i'\right)\right)$.

□

This result, applied to the RHS of (H.1) gives the equality

$$\sum_{i=1}^{k} (n_i - 1)^2 E\left(tr\left(\boldsymbol{S}_i \boldsymbol{S}_i'\right)\right) = \sum_{i=1}^{k} (n_i - 1)^2 tr\left(E\left(vec\left(\boldsymbol{S}_i\right) vec\left(\boldsymbol{S}_i\right)'\right)\right). \tag{H.5}$$

This is a useful result since (H.6)-(H.13) below prove that the theoretical value of $tr\left(E\left(vec\left(\boldsymbol{S}_i\right) vec\left(\boldsymbol{S}_i\right)'\right)\right)$ in the RHS of (H.5) involves v, which then allows deriving a moment estimator for v.

In fact, using the formula of iterated expectations,

$$\begin{aligned} tr\left(E\left(vec\left(\boldsymbol{S}_i\right) vec\left(\boldsymbol{S}_i\right)'\right)\right) &= tr\left(E\left(E\left(vec\left(\boldsymbol{S}_i\right) vec\left(\boldsymbol{S}_i\right)' | \boldsymbol{\Sigma}_i\right)\right)\right) \\ &= tr\left(E\left(Var\left(vec\left(\boldsymbol{S}_i\right) | \boldsymbol{\Sigma}_i\right) + E\left(vec\left(\boldsymbol{S}_i\right) | \boldsymbol{\Sigma}_i\right) E\left(vec\left(\boldsymbol{S}_i\right) | \boldsymbol{\Sigma}_i\right)'\right)\right) \\ &= tr\left(E\left(Var\left(vec\left(\boldsymbol{S}_i\right) | \boldsymbol{\Sigma}_i\right) + vec\left(\boldsymbol{\Sigma}_i\right) vec\left(\boldsymbol{\Sigma}_i\right)'\right)\right) \\ &= E\left(tr\left(Var\left(vec\left(\boldsymbol{S}_i\right) | \boldsymbol{\Sigma}_i\right)\right)\right) + E\left(tr\left(vec\left(\boldsymbol{\Sigma}_i\right) vec\left(\boldsymbol{\Sigma}_i\right)'\right)\right). \end{aligned} \tag{H.6}$$

In order to derive an expression for the first term in the RHS of (H.6), the following identity is used:

According to Schott (2005, p.426),

$$Var\left(vec\left(\boldsymbol{S}_i\right)\right) = \frac{2}{n_i - 1} \boldsymbol{N}_p\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{\Sigma}_i\right), \tag{H.7}$$

where $\boldsymbol{N}_p = \frac{1}{2}\left(\boldsymbol{I}_{p^2} + \boldsymbol{K}_{pp}\right)$ (Schott, 2005, p.311), $\boldsymbol{I}_{p^2}$ the $p^2 \times p^2$ identity matrix and $\boldsymbol{K}_{pp}$ the $p^2 \times p^2$ commutation matrix (Schott, 2005, p.306).

Therefore, the term $E\left(tr\left(Var\left(vec\left(\boldsymbol{S}_i\right)|\boldsymbol{\Sigma}_i\right)\right)\right)$ in the RHS of (H.6) becomes

$$\begin{aligned} E\left(tr\left(Var\left(vec\left(\boldsymbol{S}_i\right)|\boldsymbol{\Sigma}_i\right)\right)\right) &= E\left(tr\left(\frac{2}{n_i-1}\boldsymbol{N}_p\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{\Sigma}_i\right)\right)\right) \\ &= \frac{2}{n_i-1}E\left(tr\left(\boldsymbol{N}_p\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{\Sigma}_i\right)\right)\right), \end{aligned} \tag{H.8}$$

where, using properties of traces (Schott, 2005, Theorem 8.26),

$$\begin{aligned} tr\left(\boldsymbol{N}_p\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{\Sigma}_i\right)\right) &= tr\left(\frac{1}{2}\left(\boldsymbol{I}_{p^2} + \boldsymbol{K}_{pp}\right)\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{\Sigma}_i\right)\right) \\ &= \frac{1}{2}\left(tr\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{\Sigma}_i\right) + tr\left(\boldsymbol{K}_{pp}\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{\Sigma}_i\right)\right)\right) \\ &= \frac{1}{2}\left(tr\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{\Sigma}_i\right) + tr\left(\boldsymbol{\Sigma}_i^2\right)\right). \end{aligned} \tag{H.9}$$

Therefore,

$$\begin{aligned} E\left(tr\left(\boldsymbol{N}_p\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{\Sigma}_i\right)\right)\right) &= \frac{1}{2}\left(E\left(tr\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{\Sigma}_i\right)\right) + E\left(tr\left(\boldsymbol{\Sigma}_i^2\right)\right)\right) \\ &= \frac{1}{2}\left(tr\left(E\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{\Sigma}_i\right)\right) + tr\left(E\left(\boldsymbol{\Sigma}_i^2\right)\right)\right). \end{aligned} \tag{H.10}$$

When $\boldsymbol{\Sigma}_i \sim IW(\boldsymbol{R}, v)$, expressions for $E\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{\Sigma}_i\right)$ and $E\left(\boldsymbol{\Sigma}_i^2\right)$ are explicitly defined in von Rosen (1988, Theorem 3.1,Corollary 3.1), provided $v - p - 3 > 0$. Therefore, an expression for $E\left(tr\left(Var\left(vec\left(\boldsymbol{S}_i\right)|\boldsymbol{\Sigma}_i\right)\right)\right)$ in (H.8) is obtained, as follows:

$$E\left(tr\left(Var\left(vec\left(\boldsymbol{S}_i\right)|\boldsymbol{\Sigma}_i\right)\right)\right) = \frac{1}{n_i-1}\left[(c_1+c_2)tr^2\left(\boldsymbol{R}\right) + (c_1+2c_2)\,tr\left(\boldsymbol{R}^2\right) + c_2 tr\left(\boldsymbol{R}\boldsymbol{R}'\right)\right], \tag{H.11}$$

where $c_2^{-1} = (v-p)(v-p-1)(v-p-3)$, $c_1 = (v-p-2)c_2$, and $\boldsymbol{R}$ is the scale matrix in the distribution of $\boldsymbol{\Sigma}_i$ ((4.23)). Therefore, the first term in the RHS of (H.6) is determined.

The next step is to derive $E\left(tr\left(vec\left(\boldsymbol{\Sigma}_i\right)vec\left(\boldsymbol{\Sigma}_i\right)'\right)\right)$, the second term in the RHS of (H.6). It is easy to see from (H.3) that $tr\left(vec\left(\boldsymbol{\Sigma}_i\right)vec\left(\boldsymbol{\Sigma}_i\right)'\right) = tr\left(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_i'\right)$, which simplifies calculations since

$$\begin{aligned} E\left(tr\left(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_i'\right)\right) &= tr\left(E\left(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_i'\right)\right) \\ &= tr\left(E\left(\boldsymbol{\Sigma}_i^2\right)\right) \\ &= (c_1+c_2)\,tr\left(\boldsymbol{R}^2\right) + c_2 tr^2\left(\boldsymbol{R}\right), \text{ according to von Rosen (1988, Corollary 3.1).} \end{aligned} \tag{H.12}$$

Plugging (H.11) and (H.12) in the RHS of (H.6) and simplifying the expression yields,

$$tr\left(E\left(vec\left(\boldsymbol{S}_i\right)vec\left(\boldsymbol{S}_i\right)'\right)\right) = \frac{1}{n_i-1}\left[(c_1+n_ic_2)tr^2\left(\boldsymbol{R}\right) + (n_ic_1+(n_i+1)c_2)\,tr\left(\boldsymbol{R}^2\right) + c_2 tr\left(\boldsymbol{R}\boldsymbol{R}'\right)\right].$$

Since, as already proved, $E\left(tr\left(\boldsymbol{S}_i\boldsymbol{S}_i'\right)\right) = tr\left(E\left(vec\left(\boldsymbol{S}_i\right)vec\left(\boldsymbol{S}_i\right)'\right)\right)$, the term $\sum_{i=1}^{k}(n_i-1)^2E\left(tr\left(\boldsymbol{S}_i\boldsymbol{S}_i'\right)\right)$ can be written as

$$\sum_{i=1}^{k}(n_i-1)\left[(c_1+n_ic_2)tr^2\left(\boldsymbol{R}\right) + (n_ic_1+(n_i+1)c_2)\,tr\left(\boldsymbol{R}^2\right) + c_2 tr\left(\boldsymbol{R}\boldsymbol{R}'\right)\right]. \tag{H.13}$$

The RHS of (H.13) involves the unknown terms v (through c_1 and c_2) and $\boldsymbol{R}$. As also discussed in Section 4.7.2, $\boldsymbol{R}$ can be substituted by $\hat{\boldsymbol{R}} = (\hat{v}-p-1)\hat{\boldsymbol{\xi}}$, therefore leaving $\hat{v}$ to be the only unknown term in the RHS of (H.13). As the method of moments suggests, $E\left(tr\left(\boldsymbol{S}_i\boldsymbol{S}_i'\right)\right)$ in the LHS of (H.13) can be substituted by its empirical value $tr\left(\boldsymbol{S}_i\boldsymbol{S}_i'\right)$ which can be determined directly from the observations $\{\boldsymbol{y}_{ij},\ i=1,\ldots k, j=1,\ldots,n_i\}$. The resulting simplified expression then becomes,

$$(\hat{v}-p)(\hat{v}-p-3)Q = (\hat{v}-p-1)\sum_{i=1}^{k}(n_i-1)\{[n_i(\hat{v}-p-1)+2]tr\left(\hat{\boldsymbol{\xi}}^2\right)+(n_i+\hat{v}-p-2)tr^2(\hat{\boldsymbol{\xi}}) \tag{H.14}$$

Expression (H.14) reduces to a quadratic equation for v. Solving this quadratic equation yields the estimates:

$$\hat{v}_{1,2} = \frac{-b \pm \sqrt{b^2-4ac}}{2a}, \tag{H.15}$$

where,

$$
\begin{aligned}
a &= A_1 tr\left(\hat{\boldsymbol{\xi}}^2\right) + B_1 tr^2\hat{\boldsymbol{\xi}} - Q \\
b &= A_2 tr\left(\hat{\boldsymbol{\xi}}^2\right) + B_2 tr^2\hat{\boldsymbol{\xi}} + (2p+3)Q \\
c &= A_3 tr\left(\boldsymbol{\xi}^2\right) + B_3 tr^2\boldsymbol{\xi} - p(p+3)Q,
\end{aligned} \tag{H.16}
$$

with,

$$
\begin{aligned}
A_1 &= \sum_{i=1}^{k} n_i(n_i-1) \\
B_1 &= N-k \\
A_2 &= 2(N-k) - 2(p+1)\sum_{i=1}^{k} n_i(n_i-1) \\
B_2 &= \sum_{i=1}^{k}(n_i-1)^2 - 2(p+1)(N-k) \\
A_3 &= (p+1)^2\sum_{i=1}^{k} n_i(n_i-1) - 2(p+1)(N-k) \\
B_3 &= (p+1)^2(N-k) - (p+1)\sum_{i=1}^{k}(n_i-1)^2 \\
Q &= \sum_{i=1}^{k}(n_i-1)^2 tr\left(\boldsymbol{S}_i\boldsymbol{S}_i'\right).
\end{aligned} \tag{H.17}
$$

The estimate $\hat{v}$ must also satisfy $\hat{v} > p+3$ in order for both $E\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{\Sigma}_i\right)$ and $E\left(\boldsymbol{\Sigma}_i^2\right)$ in (H.10) to be defined, as previously discussed. The choice between $\hat{v}_1$ and $\hat{v}_2$ in order to get an estimate for v in the required range $(p+3, \infty)$ is discussed in Appendix I.

Appendix I

Choice of $\hat{v}$

After some straightforward manipulations, it can be shown that (H.14) can be expressed as:

$$A_l(\hat{v}-p-3)^2 + B_l(\hat{v}-p-3) = A_r(\hat{v}-p-3)^2 + B_r(\hat{v}-p-3) + C_r, \qquad \text{(I.1)}$$

where:

$$A_l = \sum_{i=1}^{k}(n_i-1)^2 tr\left(\boldsymbol{S}_i^2\right),$$

$$B_l = 3\sum_{i=1}^{k}(n_i-1)^2 tr\left(\boldsymbol{S}_i^2\right), \qquad \text{(I.2)}$$

$$A_r = tr^2(\hat{\boldsymbol{\xi}})(N-k) + tr\left(\hat{\boldsymbol{\xi}}^2\right)\sum_{i=1}^{k} n_i(n_i-1),$$

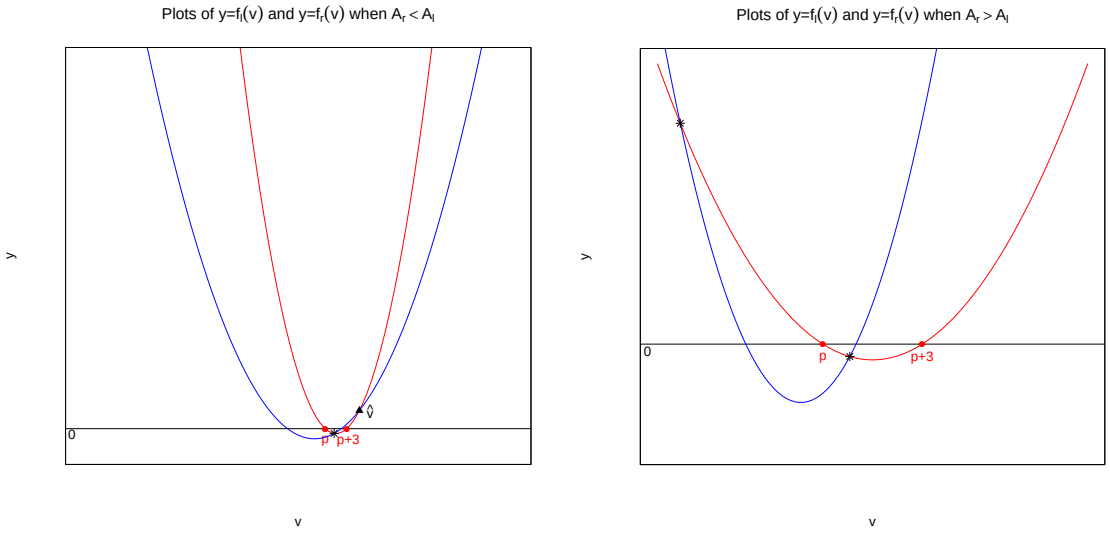
$$B_r = tr\left(\hat{\boldsymbol{\xi}}^2\right)\left[2(N-k) + 4\sum_{i=1}^{k} n_i(n_i-1)\right] + tr^2(\hat{\boldsymbol{\xi}})\left[4(N-k) + \sum_{i=1}^{k}(n_i-1)^2\right], \qquad \text{(I.3)}$$

and

$$C_r = 4\left[tr^2(\hat{\boldsymbol{\xi}})(N-k) + tr\left(\hat{\boldsymbol{\xi}}^2\right)\sum_{i=1}^{k} n_i(n_i-1)\right]. \qquad \text{(I.4)}$$

Both the L.H.S. and the R.H.S of (I.1) represent quadratic functions of v. Define the two quadratic functions as $f_l(v)$ and $f_r(v)$ respectively. The functions can be illustrated graphically as parabolas. Both of the parabolas open upwards, since their

leading coefficients (A_l and A_r respectively) are positive. Additionally, $f_l(v)$ is zero for $\hat{v} = p + 3$ and monotonically increasing in $(p + 3, +\infty)$ (since both A_l and B_l are positive). On the other hand, $f_r(v)$ is non-negative for $\hat{v} = p + 3$ and monotonically increasing for $v > p + 3$ (since A_r and B_r are positive). Therefore, a sufficient condition for obtaining a root of (I.1) in the required range $(p+3, \infty)$ is that $A_l > A_r$. Equivalently, it is required that the parabola $y = f_l(v)$ is "steeper" than $y = f_r(v)$, such that the two graphs intersect within the required range. Figure I.1a illustrates the idea.



(a) (b)

Figure I.1: Schematic illustration of $y = f_l(v)$ and $y = f_r(v)$ for determining $\hat{v}$ under different scenarios. Left: Plot of $y = f_l(v)$, $y = f_r(v)$ and the resulting $\hat{v}$ when $A_l > A_r$ Right: Plots of $y = f_l(v)$, $y = f_r(v)$ and the resulting $\hat{v}$ when $A_r > A_l$. (−): Plot of $y = f_l(v)$; (−): Plot of $y = f_r(v)$; $(*)$: Points of intersection of $y = f_l(v)$ and $y = f_r(v)$; (▲): $\hat{v}$.

In order to show that there is exactly one root in the required range when $A_l > A_r$, it suffices to prove that when $v = p$, $f_l(v)$ is zero and $f_r(v)$ is non-positive. In this case, there must be always one root in the range $[p, p + 3]$ (see Figure I.1a), which proves that there is exactly one root in the range $(p + 3, \infty)$ when $A_l > A_r$.

To prove that when $v = p$, $f_l(v)$ is zero and $f_r(v)$ is non-positive, it is more convenient to consider the form of the two quadratic functions $f_l(v)$ and $f_r(v)$ as shown in the L.H.S. and R.H.S. of (H.14) respectively. From the L.H.S. of (H.14), it is trivial to see that $f_l(v)$ is zero when $v = p$. Now, in order to prove that $f_r(v)$ is non-positive when $v = p$, it is sufficient to consider the form of the R.H.S. of (H.14) when $v = p$. It turns out to be:

$$-\sum_{i=1}^{k}(n_i-1)\left\{(n_i-2)\left[tr^2\left(\hat{\boldsymbol{\xi}}\right)-tr\left(\hat{\boldsymbol{\xi}}\right)\right]\right\}. \tag{I.5}$$

The Cauchy-Schwarz inequality (Stapleton, 2009, Chapter 1) in fact implies that $tr^2\left(\hat{\boldsymbol{\xi}}\right) \geq tr\left(\hat{\boldsymbol{\xi}}\right)$. Since also $n_i \geq 1\ \forall i$, the expression in (I.5) is always non-positive, suggesting that there is always one root of (I.1) in the range $[p, p+3]$. This completes the proof that when $A_l > A_r$, there is exactly one root in the required range $(p+3, \infty)$. In this case, $\hat{v}$ is chosen to be the largest of the two roots.

In the alternative scenario when $A_r > A_l$, the slope of the parabola $y = f_r(v)$ is "steeper" than that of $y = f_l(v)$, implying that there exists one root of (I.1) in the range $(-\infty, p]$ (see Figure I.1b). With the other root being in the range $[p, p+3]$ as previously shown, there is no root in the required range $(p+3, \infty)$. In this case, $\hat{v}$ is subjectively chosen to be the smallest integer value within the required range, i.e. $\hat{v} = p+4$.

Finally, when $A_l = A_r$, Equation (I.1) reduces to a linear equation and can be solved in terms of $\hat{v}$ to give,

$$\hat{v} = \frac{C_r}{B_l - B_r} + p + 3,$$

where B_l, B_r and C_r are defined in (I.2),(I.3) and (I.4) respectively.

It is also worth noting that the scenario of existence of complex roots in (I.1) is not examined, since complex roots appear in pairs, and as previously proved, there is always a real solution in the range $[p, p+3]$. The following table summarizes the choices for $\hat{v}$ in the different scenarios considered.

$\boldsymbol{A_l > A_r}$	$\boldsymbol{A_l < A_r}$	$\boldsymbol{A_l = A_r}$
$\hat{v} = max(\hat{v}_1, \hat{v}_2)$	$\hat{v} = p+4$	$\hat{v} = \frac{C_r}{B_l - B_r} + p + 3$

Table I.1: Choice of $\hat{v}$ based on different scenarios

Appendix J

Supplementary simulation results

This appendix presents a selection of simulation results about the bias of parameters from the random effects model of Section 4.6, which provides estimates of $\{\boldsymbol{C}_i,\ i = 1,\ldots,m\}$ in the extended framework. Specifically, results for the parameters which are more relevant to variance components (i.e. between -group variability and within-group variability) are presented here. These are: v, the degrees of freedom in the common distribution of $\{\boldsymbol{C}_i,\ i = 1,\ldots,m\}$, $\boldsymbol{\xi} = E(\boldsymbol{C}_i)$, $\boldsymbol{C}$ and $\{\boldsymbol{C}_i,\ i = 1,\ldots,m\}$, the within-group variabilities. For simplicity, since the mean biases (over the 1000 simulation runs) for $\{\hat{\boldsymbol{C}}_i,\ i = 1,\ldots,m\}$ are similar between groups, the average bias over all groups $\{i,\ i = 1,\ldots,m\}$ is shown instead.

Results are presented for a selection of scenarios considered in the simulation study of Section 5.2. Three scenarios from simulation set A (scenarios A7-A9) are selected covering all different settings for the true $\boldsymbol{\xi}$, to illustrate its effect on the bias of the variance components. Results are also shown for all scenarios of simulation set B.

Variability in the estimated v is expressed through boxplots of the 1000 estimates for each scenario, in Figure J.1. Additionally, the mean biases (among the 1000 simulation runs) of $\hat{\boldsymbol{\xi}}$ and $\hat{\boldsymbol{C}}$ (denoted as $\overline{Bias}(\hat{\boldsymbol{\xi}})$ and $\overline{Bias}(\hat{\boldsymbol{C}})$ respectively), as well as the average of mean biases of $\{\hat{\boldsymbol{C}}_i,\ i = 1,\ldots,m\}$ (denoted as $\overline{Bias}(\hat{\boldsymbol{C}}_1,\ldots,\hat{\boldsymbol{C}}_m)$) are calculated. Matrix biases are illustrated by coloured maps, generated by representing the biases of the matrix entries using colours. The colour range varies from dark red (negative bias) to dark blue (positive bias). The coloured maps are shown in Figures J.2-J.3, for simulation sets A and B respectively.

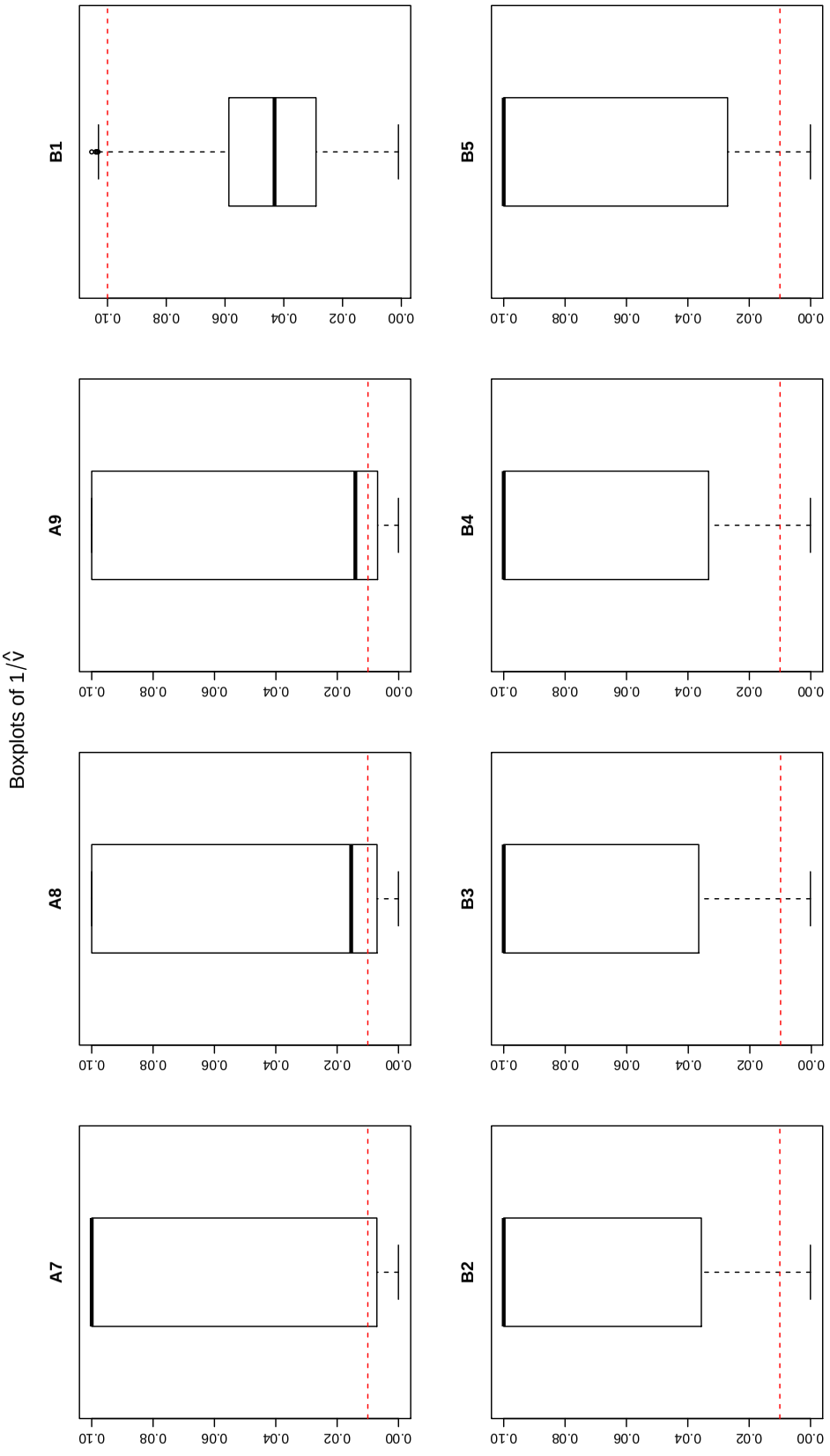


Figure J.1: Boxplots of $1/\hat{v}$ (among 1000 simulation runs), for scenarios A7- A9 and B1-B5; (– – –): True value of $1/v$.

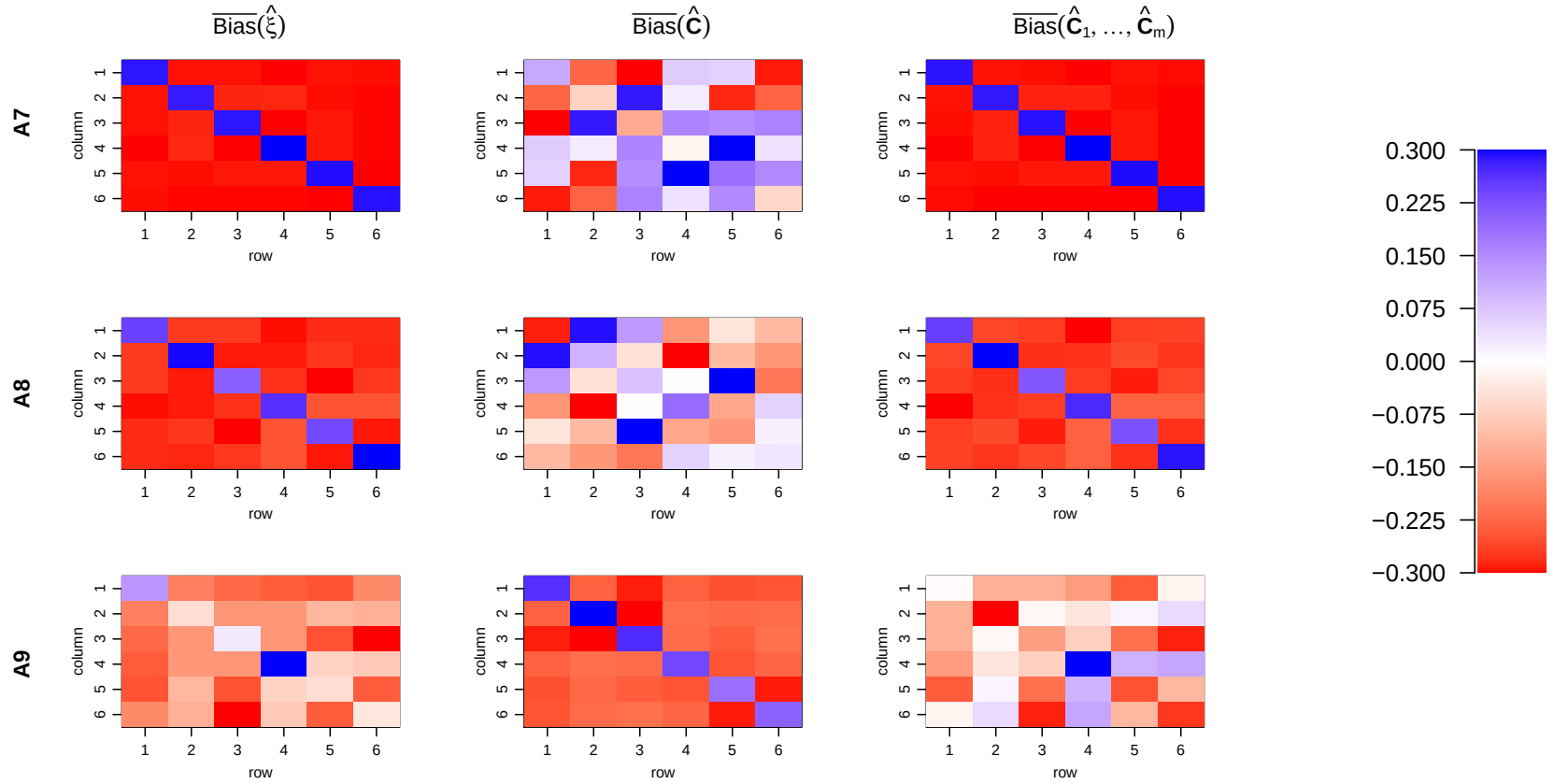


Figure J.2: Maps of $\overline{Bias}(\hat{\boldsymbol{\xi}})$(leftmost column), $\overline{Bias}(\hat{\boldsymbol{C}})$ (middle column) and $\overline{Bias}(\hat{\boldsymbol{C}}_1, \ldots, \hat{\boldsymbol{C}}_m)$ (rightmost column), for scenarios A7 (top row), A8 (middle row) and A9 (bottom row). Real values: $\boldsymbol{\xi}$= $0.1 \times \boldsymbol{I}$ (A7), $\boldsymbol{I}$ (A8), $10 \times \boldsymbol{I}$ (A9); $\boldsymbol{C} = \boldsymbol{I}$. Note that $E\left(\overline{Bias}(\hat{\boldsymbol{C}}_1, \ldots, \hat{\boldsymbol{C}}_m)\right) = \boldsymbol{\xi}$ (from (4.24), p. 101).

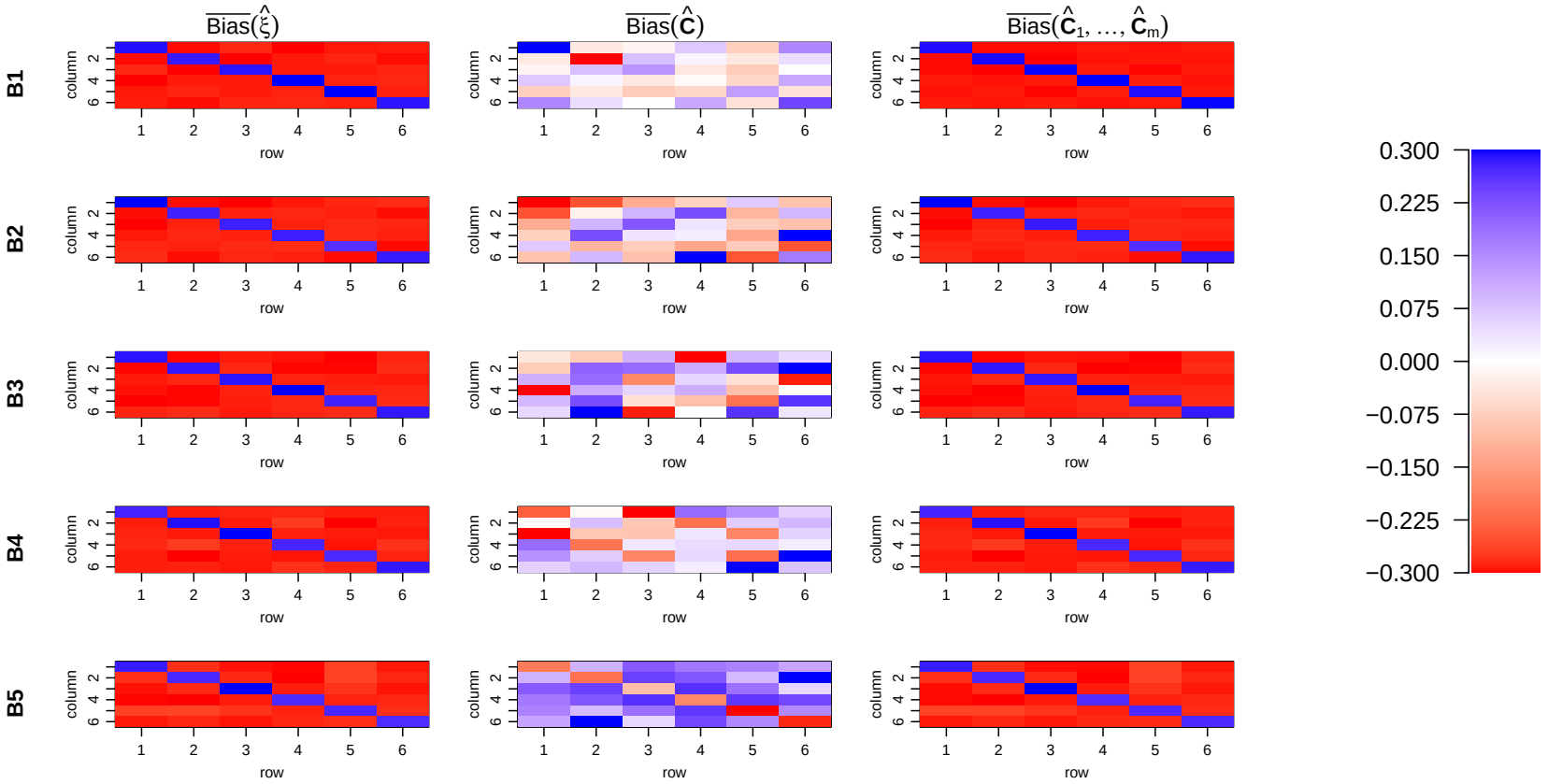


Figure J.3: Maps of $\overline{Bias}(\hat{\boldsymbol{\xi}})$(leftmost column), $\overline{Bias}(\hat{\boldsymbol{C}})$ (middle column) and $\overline{Bias}(\hat{\boldsymbol{C}}_1, \ldots, \hat{\boldsymbol{C}}_m)$ (rightmost column), for scenarios B1 (top row) to B5 (bottom row). Real values: $\boldsymbol{\xi} = \boldsymbol{I}$; $\boldsymbol{C} = \boldsymbol{I}$. Note that $E\left(\overline{Bias}(\hat{\boldsymbol{C}}_1, \ldots, \hat{\boldsymbol{C}}_m)\right) = \boldsymbol{\xi}$ (from (4.24), p. 101).

Appendix K

Supplement to application on temperature data under the extended framework

Table K.1 shows the simulators from the CMIP5 experiment whose outputs are used in the application described in Section 5.3, grouped according to Figure 1*a* of Knutti et al. (2013).

Family	***Simulators***
1	ACCESS1-0 ACCESS1-3 HadGEM2-AO
2	BCC-CSM1-1-M BCC-CSM1-1 BNU-ESM CCSM4 CESM1-BGC CESM1-CAM5 CESM1-WACCM FIO-ESM NorESM1-M NorESM1-ME
3	CMCC-CESM CMCC-CM CMCC-CMS MPI-ESM-LR MPI-ESM-MR
4	CNRM-CM5
5	CSIRO-Mk3-6-0
6	GFDL-CM3 GFDL-ESM2G GFDL-ESM2M
7	INMCM4
8	IPSL-CM5A-LR IPSL-CM5A-MR IPSL-CM5B-LR
9	MIROC-ESM-CHEM MIROC-ESM MIROC5
10	MRI-CGCM3
11	CanESM2

Table K.1: Grouping of the 32 ensemble members into families.

In Figure K.1, the corresponding simulator outputs (grouped according to families) are shown, for the historical and "change" components of $\boldsymbol{\theta}_0$.

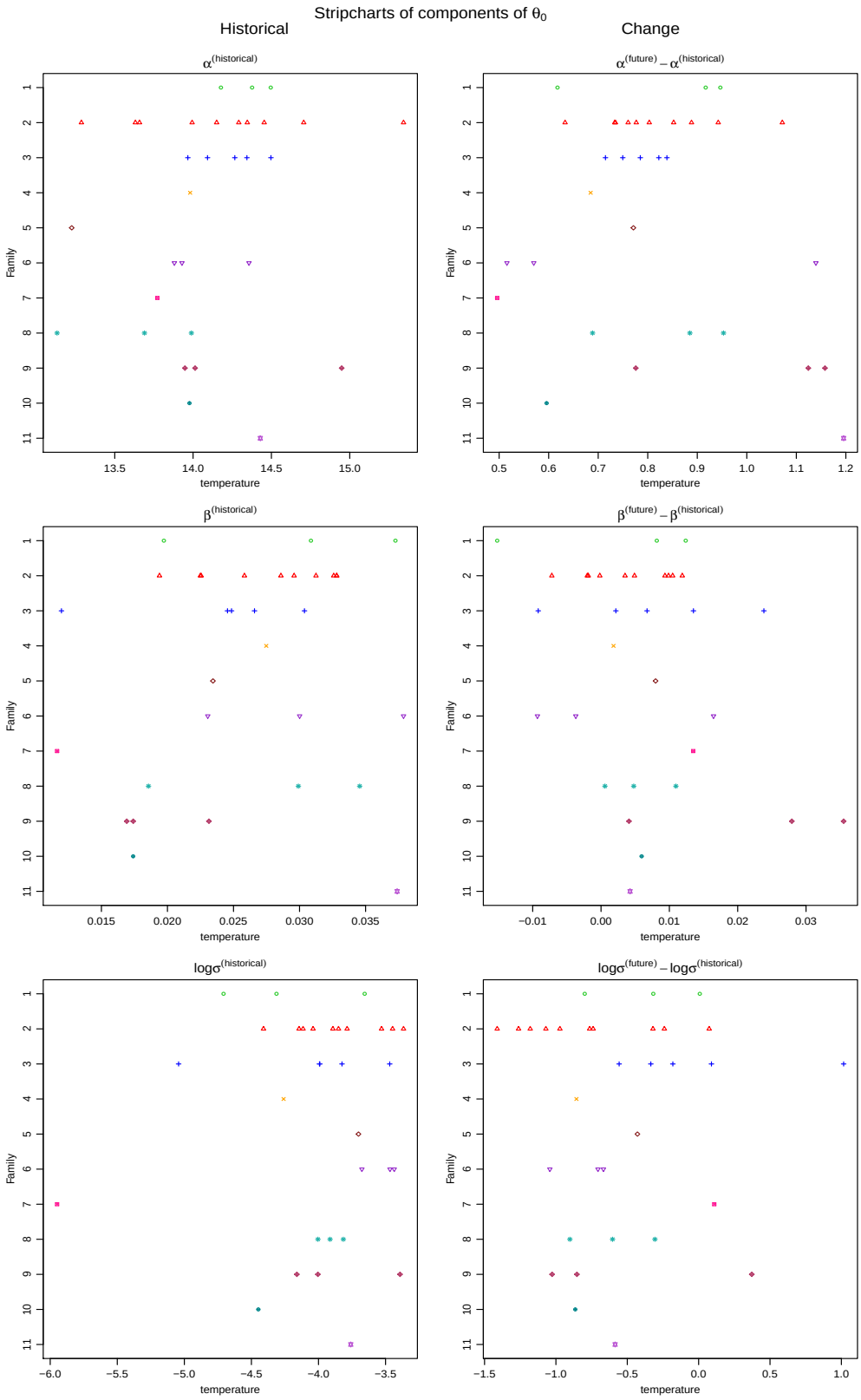


Figure K.1: Stripcharts of the historical (left column) and "change" components (right column) of descriptor estimates grouped in families, obtained after fitting the mimic to the simulator outputs from the 32 GCMs.

The estimated $\boldsymbol{\xi}$, $\boldsymbol{C}$ and $\boldsymbol{\Lambda}$ (rounded to 2 decimal places) are:

$$\hat{\boldsymbol{\xi}} = \begin{pmatrix} 0.22 & 0.00 & 0.06 & -0.01 & 0.00 & -0.03 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.06 & 0.00 & 0.16 & -0.01 & 0.00 & -0.06 \\ -0.01 & 0.00 & -0.01 & 0.03 & 0.00 & -0.01 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.03 & 0.00 & -0.06 & -0.01 & 0.00 & 0.26 \end{pmatrix}, \tag{K.1}$$

$$\hat{\boldsymbol{C}} = \begin{pmatrix} 0.01 & 0.00 & -0.01 & 0.01 & 0.00 & 0.01 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.01 & 0.00 & 0.13 & 0.03 & 0.00 & -0.05 \\ 0.01 & 0.00 & 0.03 & 0.02 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.01 & 0.00 & -0.05 & 0.00 & 0.00 & 0.02 \end{pmatrix}, \tag{K.2}$$

and

$$\hat{\boldsymbol{\Lambda}} = \begin{pmatrix} 0.01 & 0.00 & -0.01 & -0.01 & 0.00 & -0.01 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.01 & 0.00 & 0.06 & 0.01 & 0.00 & -0.05 \\ -0.01 & 0.00 & 0.01 & 0.02 & 0.00 & 0.01 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.01 & 0.00 & -0.05 & 0.01 & 0.00 & 0.17 \end{pmatrix}.$$

Appendix L

List of abbreviations

Abbreviation	***Meaning***
ANOVA	Analysis of variance
AOGCM	Atmosphere-Ocean General Circulation Model
AR4	4^{th} Assessment Report
CMIP3/5	Coupled Model Intercomparison Project phase 3/5
DAG	Directed Acyclic Graph
FB	Fully Bayesian
G	Group
GCM	General Circulation Model
GFB	Gaussian fully Bayesian
IPCC	Inter-governmental Panel on Climate Change
LHS	Left-hand side
MCMC	Markov chain Monte Carlo
MINQE	Minimum norm quadratic estimator
MINQUE	Minimum norm quadratic unbiased estimator
ML	Maximum-likelihood
MLE	Maximum-likelihood estimate
MME	Multi-model ensemble
MVN	Multivariate Normal
NG	No-group
PCMDI	Program for Climate Model Diagnosis and Intercomparison
PDF	Probability density function
PM	Poor man's
PPE	Perturbed Physics Ensemble
REA	Reliability ensemble averaging
RHS	Right-hand side

RMSE	Root-mean squared error
RCP	Representative Concentration Pathway
RPMG	Revised poor man's with groups
RPM	Revised poor man's
i.i.d.	Independent, identically distributed
UKCP09	UK Climate Projections
UNFCC	United Nations Framework Convention on Climate Change

Table L.1

Appendix M

Mathematical glossary

Symbol	***Meaning***
$\boldsymbol{\theta}_0$	Reality descriptor
$\boldsymbol{\omega}$	Shared simulator discrepancy from reality
$\boldsymbol{\theta}_i,\ i > 0$	Descriptors which are centred on $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$
$\hat{\boldsymbol{\theta}}_i$	Estimator of $\boldsymbol{\theta}_i$
$\boldsymbol{J}_i$	Covariance matrix representing uncertainty due to natural variability in data source i
$\hat{\boldsymbol{J}}_i$	Estimator of $\boldsymbol{J}_i$
$\boldsymbol{C}$	Covariance matrix representing deviation of descriptors from $\boldsymbol{\theta}_0 + \boldsymbol{\omega}$
$\hat{\boldsymbol{C}}$	Estimator of $\boldsymbol{C}$
$\boldsymbol{\Lambda}$	Covariance matrix representing uncertainty due to shared simulator discrepancy from reality
$\hat{\boldsymbol{\Lambda}}$	Estimator of $\boldsymbol{\Lambda}$
$\boldsymbol{\mu}_0$	Prior mean for $\boldsymbol{\theta}_0$
$\boldsymbol{\Sigma}_0$	Prior covariance matrix of $\boldsymbol{\theta}_0$
$\boldsymbol{\tau}$	Posterior mean of $\boldsymbol{\theta}_0$
$\boldsymbol{S}$	Posterior covariance matrix of $\boldsymbol{\theta}_0$
Y_t	Yearly mean global surface air temperature at year t
ϵ_t	Residual error in the mimic at time t
$\alpha_i^{(hist)}, \alpha_i^{(fut)}$	Historical and future mean global surface air temperature at $t = \bar{t}$, for data source i
$\beta_i^{(hist)}, \beta_i^{(fut)}$	Mean annual change in global surface air temperature, for data source i
$\sigma_i^{2\ (hist)}, \sigma_i^{2\ (fut)}$	Historical and future residual variance

$\boldsymbol{\mu}_i$	Vector collecting α_i and β_i
ψ_i	$1/\sigma_i^2$
$\boldsymbol{R}_1$	Scale matrix in prior for $\boldsymbol{C}$
v_1	Degrees of freedom in prior for $\boldsymbol{C}$
$\boldsymbol{R}_2$	Scale matrix in prior for $\boldsymbol{\Lambda}$
v_2	Degrees of freedom in prior for $\boldsymbol{\Lambda}$
$\boldsymbol{L}_i$	Covariance matrix in distribution of $\boldsymbol{\mu}_i$
$\boldsymbol{\omega}_{\boldsymbol{\mu}^{(\text{hist})}}$	Shared discrepancy of $\boldsymbol{\mu}_i^{(\text{hist})}$ from reality
$\boldsymbol{\phi}_{\boldsymbol{\mu}^{(\text{hist})}}$	Covariance matrix in distribution of $\boldsymbol{\mu}_i^{(\text{hist})} - \boldsymbol{\mu}_0^{(\text{hist})}$
$\boldsymbol{\omega}_{\boldsymbol{\Delta\mu}}$	Shared discrepancy of $\boldsymbol{\mu}_i^{(\text{fut})} - \boldsymbol{\mu}_i^{(\text{hist})}$ from reality
$\boldsymbol{\phi}_{\boldsymbol{\Delta\mu}}$	Covariance matrix in the distribution of $\left(\boldsymbol{\mu}_i^{(\text{fut})} - \boldsymbol{\mu}_i^{(\text{hist})}\right) - \left(\boldsymbol{\mu}_0^{(\text{fut})} - \boldsymbol{\mu}_0^{(\text{hist})}\right)$
v_3	Shape parameter in the distribution of $\psi_i^{(\text{hist})}/\psi_0^{(\text{hist})}$
v_4	Shape parameter in the distribution of $\frac{\psi_i^{(\text{fut})}/\psi_i^{(\text{hist})}}{\psi_0^{(\text{fut})}/\psi_0^{(\text{hist})}}$
$\omega_{\psi^{(\text{hist})}}$	Shared discrepancy of $\psi_i^{(\text{hist})}$ from reality
$\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}}$	Shared discrepancy of $\psi_i^{(\text{fut})}/\psi_i^{(\text{hist})}$ from reality
$\boldsymbol{R}_3$	Scale matrix in the distribution of $\boldsymbol{\phi}^{-1}_{\boldsymbol{\mu}^{(\text{hist})}}$
s_1	Degrees of freedom in the distribution of $\boldsymbol{\phi}^{-1}_{\boldsymbol{\mu}^{(\text{hist})}}$
$\boldsymbol{R}_4$	Scale matrix in the distribution of $\boldsymbol{\phi}^{-1}_{\boldsymbol{\Delta\mu}}$
s_2	Degrees of freedom in the distribution of $\boldsymbol{\phi}^{-1}_{\boldsymbol{\Delta\mu}}$
ρ_1	Shape parameter in the distribution of v_3
η_1	Expected value of v_3
ρ_2	Shape parameter in the distribution of v_4
η_2	Expected value of v_4
$\boldsymbol{\gamma}_1$	Covariance matrix in the distribution of $\boldsymbol{\omega}_{\boldsymbol{\mu}^{(\text{hist})}}$
$\boldsymbol{\gamma}_2$	Covariance matrix in the distribution of $\boldsymbol{\omega}_{\boldsymbol{\Delta\mu}}$
γ_3	Shape and scale parameter in distribution of $\omega_{\psi^{(\text{hist})}}$
γ_4	Shape and scale parameter in distribution of $\omega_{\psi^{(\text{fut})}/\psi^{(\text{hist})}}$
$\boldsymbol{R}_5$	Scale matrix in the distribution of $\boldsymbol{\gamma}_1^{-1}$
s_5	Degrees of freedom in the distribution of $\boldsymbol{\gamma}_1^{-1}$
$\boldsymbol{R}_6$	Scale matrix in the distribution of $\boldsymbol{\gamma}_2^{-1}$
s_6	Degrees of freedom in the distribution of $\boldsymbol{\gamma}_2^{-1}$
ρ_3	Shape parameter in the distribution of γ_3
η_3	Expected value in the distribution of γ_3
ρ_4	Shape parameter in the distribution of γ_4
η_4	Expected value in the distribution of γ_4
$\boldsymbol{\lambda}_1$	Mean in the distribution of $\boldsymbol{\mu}_0^{(\text{hist})}$

$\boldsymbol{\kappa}_1$	Covariance matrix in the distribution of $\boldsymbol{\mu}_0^{(\text{hist})}$
$\boldsymbol{\lambda}_2$	Mean in the distribution of $\boldsymbol{\mu}_0^{(\text{fut})} - \boldsymbol{\mu}_0^{(\text{hist})}$
$\boldsymbol{\kappa}_2$	Covariance matrix in the distribution of $\boldsymbol{\mu}_0^{(\text{fut})} - \boldsymbol{\mu}_0^{(\text{hist})}$
v_5	Shape parameter in the distribution of $\psi_0^{(\text{hist})}$
w_1	Expected value of $\psi_0^{(\text{hist})}$
v_6	Shape parameter in the distribution of $\frac{\psi_0^{(\text{fut})}}{\psi_0^{(\text{hist})}}$
w_2	Expected value of $\frac{\psi_0^{(\text{fut})}}{\psi_0^{(\text{hist})}}$
$\boldsymbol{\theta}_{ij},\ i,j>0$	Descriptor of simulator j belonging to family i
$\boldsymbol{\theta}_{ijk},\ i,j,k>0$	Descriptor of variant k of simulator j of family i
$\hat{\boldsymbol{\theta}}_{ijk},\ i,j,k>0$	Estimator of $\boldsymbol{\theta}_{ijk},\ i,j,k>0$
$\boldsymbol{C}_i,\ i>0$	Covariance matrix expressing deviation of simulator descriptors in family i from their consensus
$\boldsymbol{C}_{ij},\ i,j>0$	Covariance matrix expressing deviation of variant descriptors of simulator j from their consensus
$\boldsymbol{J}_{ijk},\ i,j,k>0$	Covariance matrix expressing deviation of $\hat{\boldsymbol{\theta}}_{ijk}$ from $\boldsymbol{\theta}_{ijk}$
$\boldsymbol{y}_{ij}$	Observation vector in the proposed random effects model
$\boldsymbol{\mu}$	Mean of $\boldsymbol{y}_{ij}$ in the proposed random effects model
$\boldsymbol{\alpha}_i$	Effect of group i in the proposed random effects model
$\boldsymbol{\epsilon}_{ij}$	Residual errors in the proposed random effects model
$\boldsymbol{\Sigma}_\alpha$	Covariance matrix in the distribution of $\boldsymbol{\alpha}_i$
$\hat{\boldsymbol{\Sigma}}_\alpha$	Estimator of $\boldsymbol{\Sigma}_\alpha$
$\boldsymbol{\Sigma}_i$	Group-specific covariance matrix in the distribution of $\boldsymbol{\epsilon}_{ij}$
$\boldsymbol{R}$	Scale matrix in the distribution of $\boldsymbol{\Sigma}_i$
$\hat{\boldsymbol{R}}$	Estimator of $\boldsymbol{R}$
v	Degrees of freedom in the distribution of $\boldsymbol{\Sigma}_i$
$\hat{v}$	Estimator of v
$\boldsymbol{\xi}$	Expected value of $\boldsymbol{\Sigma}_i$
$\hat{\boldsymbol{\xi}}$	Estimator of $\boldsymbol{\xi}$
y_{ij}	Observation in the proposed random effects model (univariate data)
μ	Mean of y_{ij} in the proposed random effects model (univariate data)
$\hat{\mu}$	Estimator of μ
α_i	Effect of group i in the proposed random effects model (univariate data)
ϵ_{ij}	Residual errors in the proposed random effects model (univariate data)

σ_i^2	Variance of ϵ_{ij} for group i
σ_α^2	Variance of α_i
$\hat{\sigma}_\alpha^2$	Estimator of σ_α^2
$\tilde{v}$	Shape parameter the distribution of σ_i^2
t	Rate parameter in the distribution of σ_i^2
ξ	Mean of σ_i^2
$\hat{\xi}$	Estimator of ξ
ψ	Variance of σ_i^2
$\hat{\psi}$	Estimator of ψ
a	Shape parameter in the full-conditional distribution of σ_i^2
b	Rate parameter in the full-conditional distribution of σ_i^2
r	Mean in the full-conditional distribution of α_i
s	Variance in the full-conditional distribution of β_i
$\boldsymbol{\tau}_G^i$	Posterior mean of $\boldsymbol{\theta}_0$ for synthetic dataset i, under the group framework
$\boldsymbol{S}_G^i$	Posterior covariance matrix of $\boldsymbol{\theta}_0$ for synthetic dataset i, under the group framework
$\boldsymbol{\tau}_{NG}^i$	Posterior mean of $\boldsymbol{\theta}_0$ for synthetic dataset i, under the no group framework
$\boldsymbol{S}_{NG}^i$	Posterior covariance matrix of $\boldsymbol{\theta}_0$ for synthetic dataset i, under the no group framework

Table M.1

Bibliography

Abramowitz, G. (2010). Model independence in multi-model ensemble prediction. *Australian Meteorological and Oceanographic Journal*, 59:3–6.

Annan, J. D. and Hargreaves, J. C. (2010). Reliability of the CMIP3 ensemble. *Geophys. Res. Lett.*, 37(2). Wiley Online Library.

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519.

Barnston, A. G., Mason, S. J., Goddard, L., Dewitt, D. G., and Zebiak, S. E. (2003). Multimodel ensembling in seasonal climate forecasting at IRI. *Bull. Amer. Meteor. Soc.*, 84(12):1783–1796.

Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *NeuroImage*, 20:1052–1063.

Bell, B. A., Ferron, J. M., and Kromrey, J. D. (2008). Cluster size in multilevel models: the impact of sparse data structures on point and interval estimates in two-level models. In *Proceedings of the Joint Statistical Meetings*, pages 1122–1129.

Bell, B. A., Morgan, G. B., Kromrey, J. D., and Ferron, J. M. (2010). The impact of small cluster size on multilevel models: A Monte Carlo examination of two-level models with binary and continuous predictors. *JSM Proceedings, Survey Research Methods Section*, pages 4057–4067.

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 dataset from 1850. *J. Geophys. Res.*, 111(D12). doi:10.1029/2005JD006548.

Buser, C. M., Künsch, H., Lüthi, D., Wild, M., and Schär, C. (2009). Bayesian multimodel projection of climate: Bias assumptions and interannual variability. *Clim. Dynam.*, 33(6):849–868.

Casella, G. (1985). An introduction to Empirical Bayes data analysis. *The American Statistician*, 39(2):83–87.

Chandler, R., Rougier, J., and Collins, M. (2010). Climate change: Making certain what the uncertainties are. *Significance*, 7(1):9–12. doi: 10.1111/j.1740-9713.2010.00403.x.

Chandler, R. E. (2013). Exploiting strength, discounting weakness: Combining information from multiple climate simulators. *Phil. Trans. R. Soc. A*, 371(1991):20120388. doi:10.1098/rsta.2012.0388.

Chandler, R. E. and Scott, E. M. (2011). *Statistical Methods for Trend Detection and Analysis In the Environmental Sciences*. Statistics in Practice. Wiley, United Kingdom.

Christensen, J. H., Kjellström, E., Giorgi, F., Lenderink, G., and Rummukainen, M. (2010). Weight assignment in regional climate models. *Clim. Res.*, 44:179–194. doi:10.3354/cr00916.

Clark, J. S. and Gelfand, A. E., editors (2006). *Hierarchical Modelling for the environmental sciences*. Oxford University Press.

Clarke, P. and Wheaton, B. (2007). Addressing data sparseness in contextual population research. *Sociological Methods & Research*, 35(3):311–351.

Collins, M., Booth, B. B. B., Bhaskaran, B., Harris, G. R., Murphy, J. M., Sexton, D. M. H., and Webb, M. J. (2011). Climate model errors, feedbacks and forcings: A comparison of perturbed physics and multi-model ensembles. *Clim. Dynam.*, 36:1737–1766.

Collins, M., Chandler, R. E., Cox, P. M., Huthnance, J. M., Rougier, J., and Stephenson, D. B. (2012). Quantifying future climate change. *Nature Clim. Change*, 2(6):403–409.

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Cox, P. and Stephenson, D. (2007). A changing climate for prediction. *Science*, 317:207–208. doi:10.1126/ science.1145956.

Cui, X., Hwang, J. T. G., Qiu, J., Blades, N. J., and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6(1):59–75.

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press, Cambridge.

Déqué, M. and Somot, S. (2010). Weighted frequency distributions express modelling uncertainties in the ENSEMBLES regional climate experiments. *Clim. Res.*, 44:195–209.

Doblas-Reyes, F., Hagedorn, R., and Palmer, T. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting–II. Calibration and combination. *Tellus A*, 57(3):234–252.

Ecochard, R. and Clayton, D. G. (1998). Multi-level modelling of conception in artificial insemination by donor. *Statistics in medicine*, 17:1137–1156.

Efron, B. (1979). Bootstrap methods: Another look at the Jacknife. *Ann. Statist.*, 7(1):1–26.

Elvidge, S., Godinez, H., Angling, M. J., and Koller, J. (2013). Improved modelling of upper atmospheric densities using multi-model ensembles. Technical report, Los Alamos Space Weather Summer School. Available at: `http://ima.org.uk/_db/_documents/Elvidge%20paper.pdf`.

Fielding, A. and Goldstein, H. (2006). *Cross-classified and Multiple Membership Structures in Multilevel Models: An Introduction and Review*. London. Research Report, RR791. Available at: `http://www.education.gov.uk/publications/eorderingdownload/rr791.pdf`.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., W.Collins, Cox, P., Driouech, F., S.Emori, Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., C.Reason, and Rummukainen, M. (2013). *Evaluation of Climate Models*. Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: Theory. *NeuroImage*, 16:465–483.

Geinitz, S., Furrer, R., and Sain, S. R. (2015). Bayesian multilevel analysis of variance for relative comparison across sources of global climate model variability. *Int. J. Climatol.*, 35(3):433–443.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.*, 1(3):515–533.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Chapman and Hall, Third edition.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Ghosh, J. and Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *J. Comput. Graph. Stat.*, 18(2):306–320.

Giorgi, F. and Mearns, L. (2002). Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging" (REA) method. *J. Clim.*, 15:1141–1158.

Goldstein, H. (1995). *Multilevel Statistical Models*, volume 3 of *Kendall's Library of Statistics*. Arnold, London, Second edition.

Greene, A., Goddard, L., and Lall, U. (2006). Probabilistic multimodel regional temperature change projections. *J. Climate*, 19(17):4326–4343.

Hagedorn, R., Doblas-Reyes, F., and Palmer, T. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting-I. Basic concept. *Tellus A*, 57(3):219–233.

Hawkins, E. and Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bull. Amer. Meteor. Soc.*, 90(8):1095–1107.

Heideman, K. F., Stewart, T. R., Moninger, W. R., and Reagan-Cirincione, P. (1993). The weather information and skill experiment (WISE): The effect of varying levels of information on forecast skill. *Weather Forecast.*, 8(1):25–36.

Higham, N. (2002). Computing the nearest correlation matrix- a problem from finance. *IMA J Numer Anal*, 22:329–343.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer Science & Business Media.

Holmes, A. P. and Friston, K. J. (1998). Generalisability, random effects & population inference. *Neuroimage*, 7:S754.

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications*. Quantitative methodology series. Routledge, New York, Second edition.

IPCC-DDC (2013). What is a GCM? Available at: `http://www.ipcc-data.org/guidelines/pages/gcm_guide.html`.

Jones, P., Harpham, C., Osborn, T., and Salmon, M. (2014). Temperature. Available at: `http://www.cru.uea.ac.uk/cru/data/temperature/`.

Jones, P. D., New, M., Parker, D., Martin, S., and Rigor, I. (1999). Surface air temperature and its changes over the past 150 years. *Rev. Geophys.*, 37(2):173–199.

Jun, M. R., Knutti, R., and Nychka, D. (2008). Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *J. Amer. Statist. Assoc.*, 103(483):934–947. doi:10.1198/016214507000001265.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *J. Roy. Stat. Soc. B*, 63:425–464.

Kharin, V. V. and Zwiers, F. W. (2002). Climate predictions with multimodel ensembles. *J. Climate*, 15(7):793–799.

Kjellström, E., Boberg, F., Castro, M., Christensen, J. H., Nikulin, G., and Sánchez, E. (2010). Daily and monthly temperature and precipitation statistics as performance indicators for regional climate models. *Clim. Res.*, 44:135–150. doi:10.3354/cr00932.

Kjellström, E. and Giorgi, F. (2010). Introduction to special issue "Regional climate model evaluation and weighting". *Clim. Res.*, 44(2-3):117–119.

Knutti, R. (2008). Should we believe model predictions of future climate change? *Phil. Trans. R. Soc. A: Mathematical, Physical and Engineering Sciences*, 366(1885):4647–4664.

Knutti, R. (2010). The end of model democracy? *Clim. Chang.*, 102:395–404. doi: 10.1007/s10584-010-9800-2.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. (2010). Challenges in combining projections from multiple climate models. *J. Climate*, 23(10):2739–2758. doi:10.1175/2009JCLI3361.1.

Knutti, R., Masson, D., and Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophys. Res. Lett.*, 40:1194–1199.

Knutti, R. and Sedláček, J. (2013). Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Clim. Change*, 3(4):369–373.

Koski, T. and Noble, J. M. (2009). *Bayesian networks*. Wiley, Chichester.

Krishnamurti, T. N., Kishtawal, C., Zhang, C. M., Larow, T., Bachiochi, D., Williford, E., Gadgil, S., and Surendrn, S. (2000). Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, 13(2000):4196–4216. doi:10.1175/1520-0442(2000)013!4196:MEFFWAO2.0.CO;2.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York.

Leith, N. A. and Chandler, R. E. (2010). A framework for interpreting climate model outputs. *J. R. Stat. Soc. C*, 59(2):279–296.

Lenderink, G. (2010). Exploring metrics of extreme daily precipitation in a large ensemble of regional climate model simulations. *Clim. Res.*, 44(2):151–166.

Liechty, J. C., Liechty, M. W., and Müller, P. (2004). Bayesian correlation estimation. *Biometrica*, 91(1):1–14.

Lönnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistica Sinica*, 12:31–46.

Lucarini, V. (2002). Towards a definition of climate science. *Int. J. Environ. Pollut.*, 18(5):413–422.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Stat. Med.*, 28(25):3049–3067.

Masson, D. and Knutti, R. (2011). Climate model genealogy. *Geophys. Res. Lett.*, 38(L08703).

Mathai, A. M. and Provost, S. B. (1992). *Quadratic forms in random variables: theory and applications*, volume 126 of *STATISTICS: Textbooks and Monographs*. Marcel Dekker, New York.

Meehl, G., Boer, G., Covey, C., Latif, M., and Stouffer, R. (2000). The coupled model intercomparison project CMIP. *Bull.Amer.Meteor.Soc.*, 81(2):313–318.

Min, S.-K. and Hense, A. (2007). Hierarchical evaluation of IPCC AR4 coupled climate models with systematic consideration of model uncertainties. *Clim. Dynam.*, 29(7-8):853–868.

Moss, R., Babiker, M., Brinkman, S., Calvo, E., Carter, T., Edmonds, J., Elgizouli, I., Emori, S., Erda, L., Hibbard, K., Jones, R., Kainuma, M., Kelleher, J., Lamarque, J. F., Manning, M., Matthews, B., Meehl, J., Meyer, L., Mitchell, J., Nakicenovic, N., O'Neill, B., Pichs, R., Riahi, K., Rose, S., Runci, P., Stouffer, R., van Vuuren, D., Weyant, J., Wilbanks, T., van Ypersele, J. P., and Zurek, M. (2008). Towards new scenarios for analysis of emissions, climate change, impacts, and response strategies. Technical summary, Intergovernmental Panel on Climate Change, Geneva. 25pp.

Murphy, J. M., Booth, B. B., Collins, M., Harris, G. R., Sexton, D. M., and Webb, M. J. (2007). A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Phil. Trans. R. Soc. A: Mathematical, Physical and Engineering Sciences*, 365(1857):1993–2028.

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., and Stainforth, D. A. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, 430(7001):768–772.

Newhall, C., II, J. W. H., and Stauffer, P. H. (2005). The Cataclysmic 1991 Eruption of Mount Pinatubo, Philippines. Available at: `http://pubs.usgs.gov/fs/1997/fs113-97/`.

Northrop, P. J. and Chandler, R. E. (2014). Quantifying sources of uncertainty in projections of future climate. *J. Climate*, 27(23):8793–8808.

Oreskes, N. (2000). *Why predict? Historical perspectives on prediction in Earth Science*. Prediction: Science, Decision Making, and the Future of Nature. Island Press.

Palmer, T., Doblas-Reyes, F. J., Hagedorn, R., and Weisheimer, A. (2005). Probabilistic prediction of climate using multi-model ensembles: From basics to applications. *Phil. Trans. R. Soc. B*, 360(1463). doi:10.1098/rstb.2005.1750.

Pirtle, Z., Meyer, R., and Hamilton, A. (2010). What does it mean when climate models agree? *Environ. Sci. Policy*, 13:351–361.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in FORTRAN*. Cambridge University Press, second edition.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at: `http://www.R-project.org`.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2:1–21.

Räisänen, J. (2007). How reliable are climate models? *Tellus A*, 59(1):2–29.

Räisänen, J. and Palmer, T. (2001). A probability and decision-model analysis of a multimodel ensemble of climate change simulations. *J. Climate*, 14(15):3212–3226.

Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*. Probability and Mathematical Statistics. John Wiley and Sons, Inc., USA.

Rao, P. S. R. S., Kaplan, J., and Cohran, W. G. (1981). Estimators for the one-way random effects model with unequal error variance. *Journal of the American Statistical Association*, 76(373):89–97.

Reilly, J., Stone, P. H., Forest, C. E., Webster, M. D., Jacoby, H. D., and Prinn, R. G. (2001). Uncertainty and climate change assessments. *Science*, 293(5529):430–433.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*. Brooks/Cole, USA, third edition.

Rougier, J. and Goldstein, M. (2014). Climate Simulators and Climate Projections. *Annu. Rev. Stat. Appl.*, 1:103–123.

Rougier, J., Goldstein, M., and House, L. (2013). Second-order exchangeability analysis for multi-model ensembles. *J. Am. Stat. Assoc.*, 108(503):852–863.

Rukhin, A. L. (2013). Estimating heterogeneity variance in meta-analysis. *J. R. Statist. Soc. B*, 75(3):451–469.

Sanderson, B. M. and Knutti, R. (2012). On the interpretation of constrained climate model ensembles. *Geophys. Res. Lett.*, 39(16).

Sansó, B., Forest, C. E., Zantedeschi, D., et al. (2008). Inferring climate system properties using a computer model. *Bayesian Anal.*, 3(1):1–37.

Sansom, P. G., Stephenson, D. B., Ferro, C. A., Zappa, G., and Shaffrey, L. (2013). Simple uncertainty frameworks for selecting weighting schemes and interpreting multimodel ensemble climate change experiments. *J. Climate*, 26(12):4017–4037.

Schott, J. R. (2005). *Matrix Analysis for Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey, Second edition.

Searle, S. R. (1987). *Linear Models for Unbalanced Data*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Canada.

Searle, S. R., Casella, G., and McCulloch, C. E. (2006). *Variance Components*. Wiley Series in Probability and Statistics. John Wiley and Sons, Inc., Hoboken, New Yersey.

Siegert, S., Stephenson, D. B., Sansom, P. G., Scaife, A. A., Eade, R., and Arribas, A. (2015). A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability? *arXiv preprint arXiv:1504.01933*.

Smith, L. A. (2002). What might we learn from climate forecasts? *Proc. Natl. Acad. Sci.*, 99(suppl 1):2487–2492.

Snijders, T. A. B. and Bosker, R. J. (2002). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*. Sage.

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

Stainforth, D. A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D. J., Kettleborough, J. A., Knight, S., Martin, A., Murphy, J. M., Piani, C., Sexton, D., Smith, L. A., Spicer, R. A., Thorpe, A. J., and Allen, M. R. (2005). Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, 433(7024):403–406.

Stapleton, J. H. (2009). *Linear statistical models*, volume 719. John Wiley & Sons.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Stephenson, D. B., Collins, M., Rougier, J. C., and Chandler, R. E. (2012). Statistical problems in the probabilistic prediction of climate change. *Environmetrics*, 23(5):364–372. doi:10.1002/env.2153.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. (2007). A summary of the CMIP5 experimental design. *World*, 4:1–33.

Tebaldi, C., Arblaster, J. M., and Knutti, R. (2011). Mapping model agreement on future climate projections. *Geophys. Res. Lett.*, 38(23).

Tebaldi, C. and Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Phil. Trans. R. Soc. A*, 365(1857):2053–2075.

Tebaldi, C. and Sansó, B. (2009). Joint projections of temperature and precipitation change from multiple climate models: A hierarchical Bayesian approach. *J. Roy. Stat. Soc. A*, 172(1):83–106.

Tebaldi, C., Smith, R. L., Nychka, D., and Mearns, L. O. (2005). Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *J. Climate*, 18(10):1524–1540.

UKCP09 (2014). Available at: `http://ukclimateprojections.metoffice.gov.uk`.

UN (1992). Principle 15, United Nations Conference on Environment and Development. Available at: `http://www.un.org/documents/ga/conf151/aconf15126-1annex1.htm`.

UN (2015). United nations conference on climate change. Available at: `http://www.cop21.gouv.fr/en/`.

UNFCCC (2014). United Nations Framework Convention on Climate Change. Available at: `http://unfccc.int/home/items/2784.php`.

Vincent, L. and Gullett, D. (1999). Canadian historical and homogeneous temperature datasets for climate change analyses. *Int. J. Climatol.*, 19(12):1375–1388.

von Rosen, D. (1988). Moments for the inverted Wishart distribution. *Scandinavian Journal of Statistics*, 2(2):97–109.

Watterson, I. G. (2008). Calculation of probability density functions for temperature andprecipitation change under global warming. *J. Geophys. Res.*, 113(D12106).

Weigel, A., Knutti, R., Liniger, M., and Appenzeller, C. (2010). Risks of model weighting in multimodel climate projections. *J. Climate*, 23(15):4175–4191.

Weisheimer, A., Doblas-Reyes, F., Palmer, T., Alessandri, A., Arribas, A., Déqué, M., Keenlyside, N., MacVean, M., Navarra, A., and Rogel, P. (2009). ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions - skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophys. Res. Lett.*, 36(21).

Wikle, C. K. (2003). Hierarchical models in environmental science. *Int. Stat. Rev.*, 71(2):181–199.

Wikle, C. K., Berliner, L. M., and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environ. Ecol. Stat.*, 5(2):117–154.

Williams, P. D., Cullen, M. J., Davey, M. K., and Huthnance, J. M. (2013). Mathematics applied to the climate system: Outstanding challenges and recent progress. *Phil. Trans. R. Soc. A: Mathematical, Physical and Engineering Sciences*, 371(1991):20120518.

Williamson, D., Blaker, A. T., Hampton, C., and Salter, J. (2014). Identifying and removing structural biases in climate models with history matching. *Clim. Dynam.*, pages 1–26.

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K. (2013). History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Clim. Dynam.*, 41(7-8):1703–1729.

Woolrich, M. W., Behrens, T. E., Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. *NeuroImage*, 21:1732–1747.

Worsley, K. J., Liao, C., Aston, J., Petre, V., Duncan, G. H., Morales, F., and Evans, A. C. (2002). A general statistical analysis for fMRI data. *NeuroImage*, 15:1–15.

Wright, G. W. and Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, 19(18):2448–2455.

Yip, S., Ferro, C. A., Stephenson, D. B., and Hawkins, E. (2011). A simple, coherent framework for partitioning uncertainty in climate predictions. *J. Climate*, 24(17):4634–4643.

Zappa, G., Shaffrey, L. C., Hodges, K. I., Sansom, P. G., and Stephenson, D. B. (2013). A multimodel assessment of future projections of North Atlantic and European extratropical cyclones in the CMIP5 climate models. *J. Climate*, 26(16):5846–5862.