# Uncertainty Quantification in Palaeoclimate Reconstruction

Jake Carson, MSci.

Thesis submitted to The University of Nottingham
for the degree of Doctor of Philosophy

July 2015

To my parents.

# Abstract

Studying the dynamics of the palaeoclimate is a challenging problem. Part of the challenge lies in the fact that our understanding must be based on only a single realisation of the climate system. With only one climate history, it is essential that palaeoclimate data are used to their full extent, and that uncertainties arising from both data and modelling are well characterised. This is the motivation behind this thesis, which explores approaches for uncertainty quantification in problems related to palaeoclimate reconstruction.

We focus on uncertainty quantification problems for the glacial-interglacial cycle, namely parameter estimation, model comparison, and age estimation of palaeoclimate observations. We develop principled data assimilation schemes that allow us to assimilate palaeoclimate data into phenomenological models of the glacial-interglacial cycle. The statistical and modelling approaches we take in this thesis means that this amounts to the task of performing Bayesian inference for multivariate stochastic differential equations that are only partially observed.

One contribution of this thesis is the synthesis of recent methodological advances in approximate Bayesian computation and particle filter methods. We provide an up-to-date overview that relates the different approaches and provides new insights into their performance. Through simulation studies we compare these approaches using a common benchmark, and in doing so we highlight the relative strengths and weaknesses of each method.

There are two main scientific contributions in this thesis. The first is that by using inference methods to jointly perform parameter estimation and model comparison, we demonstrate that the current two-stage practice of first estimating observation times, and then treating them as fixed for subsequent analysis, leads to conclusions that are not robust to the methods used for estimating the observation times. The second main contribution is the development of a novel age model based on a linear sediment accumulation model. By extending the target of the particle filter we are able to jointly perform parameter estimation, model comparison, and observation age estimation. In doing so, we are able to perform palaeoclimate reconstruction using sediment core data that takes age uncertainty in the data into account, thus solving the problem of dating uncertainty highlighted above.

# Acknowledgements

First and foremost, a huge thanks to my supervisors, Richard Wilkinson and Simon Preston. Both have dedicated a huge amount of time, effort, and encouragement over the past four years. I truly couldn't have hoped for better supervisors.

I am particularly grateful to Michel Crucifix, who has been our primary collaborator in this project. He has given invaluable guidance throughout my PhD, and ensured that the project has been valuable scientifically, as well as statistically.

Special thanks to my external examiner Caitlin Buck, my internal examiner Theo Kypraios, and my internal assessors over the past four years, Chris Brignell, Sergey Utev, and Christopher Fallaize, for providing valuable and constructive suggestions.

Thanks to my fellow statistics and probability PhD students: Rosanna Cassidy, Jonathan Davies, Benjamin Davis, Richard Haydock, Anthony Hennessey, Lisa Mott, Phillip Paine, Iker Pérez Lopéz, Heather Pettitt, Laurence Shaw, Kamonrat Suphawan, Michael Thomson, and Vytaute Zabarskaite, many of whom I count as good friends. They have always been there to help, and to offer insight when it was needed.

A big thanks to the rest of office C11: Yefei Cao, Victory Ezeofor, Hugo Ferreira, Anthony Hennessey, Sunny 'Boss' Modhara, Thomas Oliver, Sara Tavares, Michael Thomson, and Vladimir Toussaint. They have been amazing company.

Thanks to Jack Wade, Jamie Watkin, and David Webb, who have gone above and beyond in supporting me during my PhD, and cheering me up when I needed it most.

A heartfelt thanks to my parents, for encouraging me to never stop learning, and for always offering support when it was required.

This project was funded by the Engineering and Physical Sciences Research Council.

# Contents

# List of Figures

# List of Tables

# List of Algorithms

CHAPTER 1

# Introduction

The current ice age began approximately 2.5-3.5 Myr (million years) BP (before present) during the late Pliocene [2, 3]. Since then, the climate has fluctuated between cold periods, which see a large increase in ice volume in the northern hemisphere, and warm periods in which these ice sheets retreat [1, 4]. This process is called the glacial-interglacial cycle. The glacial-interglacial cycle is observable in many palaeoclimate records. One such example is the LR04 stack [1], shown in Figure 1.1.

A glacial period begins with a glacial inception, in which glaciers begin to advance, and ends with a glacial termination, where glaciers start to retreat. Between a glacial termination and a glacial inception is a warm period, called an interglacial period. In the early Pleiostocene (approximately 2.6 Myr BP onwards) each glacial cycle would occur over approximately 40 kyr (thousand years), but during the mid-Pleistocene (approximately 800 kyr BP) this pattern changed to a 100 kyr cycle. This shift is known as the mid-Pleistocene transition. The most recent glacial cycles exhibit a saw-tooth shaped structure, with ice slowly advancing over 80 kyr, and melting rapidly over 20 kyr. The last glacial cycle terminated approximately 11 kyr BP, at which point we entered an interglacial period called the Holocene. Detailed records of the last glacial cycle indicate complex sub-millennial dynamics throughout glacial periods [5]. Dansgaard-Oeschger events are a well known example of these sub-millennial dynamics, where by, during glaciation, the temperature can raise several degrees over a few decades, followed by a longer cooling trend [6]. Possibly linked to Dansgaard-Oeschger events are Heinrich events, in which glaciers around the North Atlantic rapidly collapse [7].

The most commonly accepted mechanism to explain this long term climate variation is that changes in the Earth's orbit alter the seasonal and spatial distribution of incoming solar radiation (insolation). In particular Milkankovitch theory (named for geophysicist and astronomer Multin Milankovitch) mathematically describes the variation of insolation in terms of three orbital elements: obliquity, precession and eccentricity. Under the Milankovitch theory, cold northern-

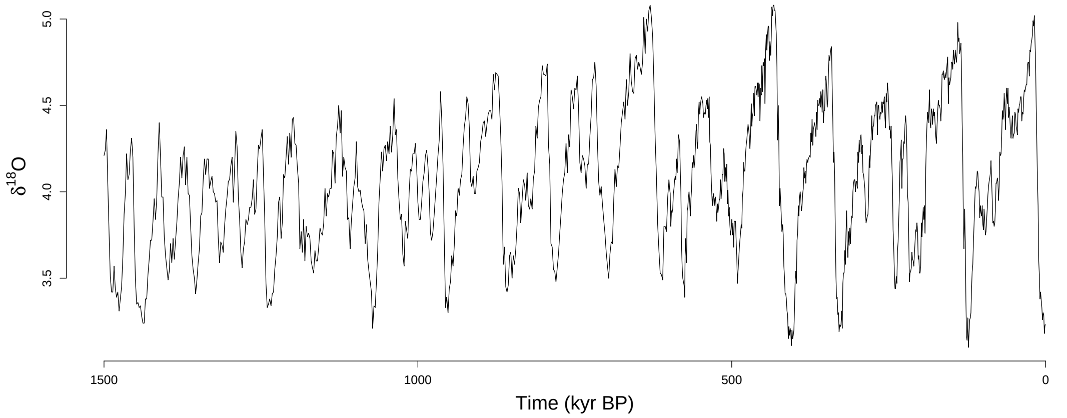

**Figure 1.1:** Observed $\delta^{18}$O, a proxy of global temperature and ice volume, from the LR04 stack corresponding to the past 1.5 Myr [1]. The LR04 stack is constructed from 57 individual records. Large values of $\delta^{18}$O indicate cold periods with a large ice volume, whereas small values indicate warm periods with little ice.

hemisphere summers are necessary for a glacial inception, which requires a low insolation over the summer season. Milankovitch measured insolation by integrating the daily mean insolation over the caloric summer (the 180 days of highest insolation assuming a 360 day year [8]) at latitude 65°N. This theory was supported by Hays et al. [9] who demonstrated that variations in the Earth's orbital parameters corresponded to variations in the climate over the past 500 kyr. Variation in insolation alone fails to explain a number of features of the glacial-interglacial cycle. The dominant periods in the insolation signal are $\sim$21 kyr and $\sim$40 kyr, matching the dominant periods of precession and obliquity respectively, but the more recent glacial cycles have a period of $\sim$100 kyr. This is a major period of eccentricity, but there are no 400 kyr cycles, which is the dominant period of eccentricity [10]. The variation in insolation also does not explain the saw-tooth shape of the more recent cycles.

The internal dynamics of the Earth's climate are also expected to have a large influence on the glacial-interglacial cycle. For example, during glaciation atmospheric $CO_2$ decreases to approximately 200 ppm (parts per million), and during deglaciation atmospheric $CO_2$ rises to approximately 280 ppm. Since $CO_2$ is a powerful greenhouse gas, these variations almost certainly influence the glacial-interglacial cycle.

Scientists have long constructed mathematical models of the climate to understand the effects of both internal climate dynamics and the astronomical forcing on the glacial-interglacial cycle (as well as many other climate processes). The complexity of climate models covers a wide spectrum, from using only a handful of climate variables to attempting to model as many physical and biochemical processes as possible [10]. Given the large timescales involved in the dynamics of the glacial-interglacial cycle, the system is often studied using *phenomenological* models. Phenomenological models are low-dimensional models that are consistent with the underlying dynamics, but not derived from the underlying physical processes of a system [11, 12]. By focussing on a limited number of climate variables it is possible to design models for the purpose of finding out which climate processes are important over long timescales. These models are more formally introduced in Chapter 2.

Even the most advanced climate models have no hope of capturing the full complexity of the climate. If we are to trust in the predictions of any climate model we need to make use of palaeoclimate data. With only one climate history, it is important to extract as much information from the data as possible. Historically, the complexity of even the simplest climate models has made a careful statistical analysis very difficult, and so approaches have been somewhat unprincipled. For example, model parameters are manually adjusted until a good fit to the data is obtained [13]. This approach leaves many questions unanswered, such as whether another set of parameters might give an equally good, or better, fit to the data, and can lead to poor characterisation of the uncertainty in model predictions. However, with recent advances in Monte Carlo methodology, phenomenological models are simple enough to allow palaeoclimate data to be assimilated into models in a principled way. The primary aim of this thesis is to develop principled data assimilation schemes for phenomenological models of the glacial-interglacial cycle, and apply them to uncertainty quantification problems in palaeoclimate science, specifically, parameter estimation, state estimation, and model selection.

## 1.1 Palaeoclimate Data

Since there are no direct observations of the palaeoclimate, the history of the climate over long timescales needs to be inferred through *proxy* measurements. Proxy measurements are measurements of a physical quantity that contains information about another quantity of interest. In the context of studying the glacial-interglacial cycle, proxy measurements are often taken from sediment and ice cores. Sediment cores are obtained by drilling into sediment using a hollow drill, called a core drill. When the drill is raised, it extracts a long cylinder of sediment, which is called a core section. The drill can then be lowered into the hole that was left in order to recover additional core sections at increasing depths. The core sections are then combined to form a single core. Finally, the core is sliced at regular intervals, within which the quantity of interest is measured. This provides a series of measurements according to depth. For the data to be useful in understanding the palaeoclimate, the depth scale needs to be converted into a time scale, so that we can see how the climate evolves through time. We know that the top of the core (depth 0) was recently deposited, and, for the most part, the ages of the measurements are an increasing function of depth, although post-depositional effects, such as sediment shifts, can alter this. Apart from this general trend there is typically very little time information that can be extracted from a core. Consequently, dating observations is extremely difficult. Following Milakovitch theory this has historically been done by aligning features in the dataset, such as glacial inceptions and terminations, with the insolation signal, the timescale of which is well understood over the past several million years from gravitational theory [14]. This approach was taken when dating the LR04 stack [1]. However, if one of the inferences that we wish to make is the influence of variations in the Earth's orbit on the glacial-interglacial cycle, then using a dataset that has been dated in this manner is undesirable; any influence that is detected could be due to the dating assumptions. Alternatively, the timescale can be assigned by conditioning on parts of the record that have been dated with high accuracy. For instance, geomagnetic reversals are evident in many sediment cores, and can often be dated with relatively small uncertainty by using, for example, radiometric dating techniques [15, 16]. The most recent geomagnetic reversal was the Brunhes-Matuyama (BM) reversal, which has been dated at $780\pm2$ kyr BP [17]. The rarity of these events means that any inferred timescale will have considerable uncertainty. This approach was taken when dating the H07 stack [4], in which the timescale between magnetic reversals was inferred by fitting a deterministic piecewise linear model between reversals that related depth to age. The dating uncertainties in both the LR04 and H07 stacks are estimated to be approximately 10 kyr, but have yet to be accurately quantified [1, 4]. The large dating uncertainties can make inferences about the palaeoclimate challenging. For example, the precession varies quasi-periodically over approximately 21 kyr, making the precession signal almost impossible to distinguish in the record. Consequently, the influence of precession is difficult to determine [4, 18].

One of the proxies most used to study the climate history is $\delta^{18}$O [1, 4]. This is a measure of the ratio $^{18}$O:$^{16}$O, two stable isotopes of oxygen. This type of data is frequently taken from sediment cores (divided into benthic (ocean floor) or planktic cores) containing fossilised foraminifer shells. Foraminifera are small marine organisms that create calcium carbonate shells through a process that requires the incorporation of oxygen from the surrounding water. Foraminiferal $\delta^{18}$O (the ratio $^{18}$O:$^{16}$O contained in the fossil shells) depends on the local sea-water $\delta^{18}$O, salinity, and temperature at the time the foraminifera was last metabolising in the water. In turn these quantities depend on global evaporation, precipitation, ice volume, and temperature, indicating the conditions of the past climate [1]. However, the numerous factors affecting foraminiferal $\delta^{18}$O means that it is rarely possible to assess the relative contributions of any single quantity [19]. Alternatively, $\delta^{18}$O measurements can be taken from ice cores. In contrast to foraminiferal $\delta^{18}$O measurements, which are indirect observations from organisms that lived in the relevant environment, ice core measurements are made directly on the core itself. Other common proxies include deuterium and $CO_2$ from ice core samples. In Figure 1.2 foraminiferal $\delta^{18}$O data from the LR04 stack [1] are superimposed with $CO_2$ data from the

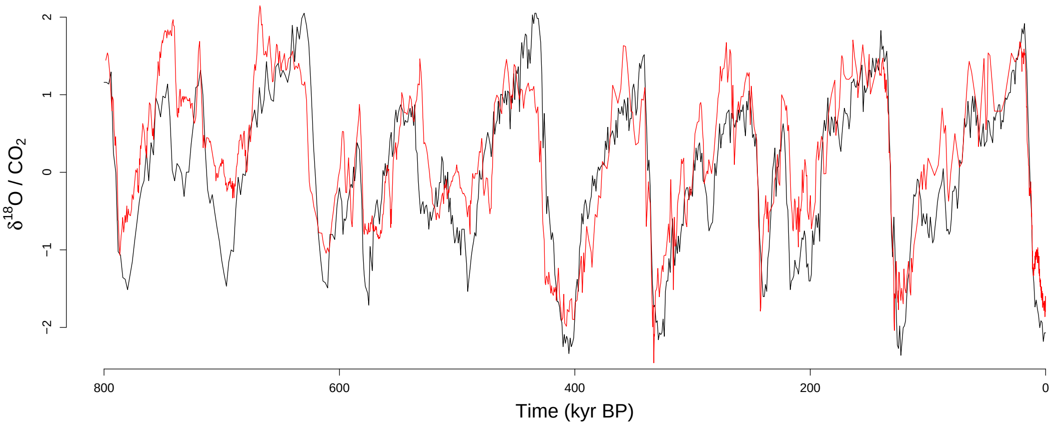

**Figure 1.2:** Observed $\delta^{18}$O from the LR04 stack [1] (black) superimposed with $CO_2$ from the Dome C ice core [20] (red) over the past 800 kyr. Both records have been scaled to highlight similarities.

Dome C ice core [20], where it can be seen that the two proxy datasets share many similar features. Each of these proxy datasets is collected as a function of depth. As such, there is uncertainty in both how the proxy measurements relate to the state of the climate and the date they represent.

Paleoclimate data are often combined into stacks. Stacks are averages over multiple datasets from different drill sites. By averaging over multiple datasets the signal-noise ratio of the climate signal is improved, and if the sites represent a wide geographical area then local trends should be suppressed. Constructing a stack typically involves correlating features in the datasets. For instance, glacial terminations are usually well defined in individual datasets, and so the terminations in each record can be aligned. However, there can be some large variations between datasets. For example, missing data and duplicated sections are common due to sediment shifts. There are numerous approaches to averaging over the datasets, but the most common approach, which was used in the construction of the H07 stack [4], is to interpolate each dataset onto a common set of time points (every 1 kyr for example), and then take the sample average at each time point.

## 1.2 Inference

Numerous phenomenological models have been proposed in the literature [21]. These models take a set of conditions (i.e. parameter values and initial conditions) and generate the evolution of specified climate variables according to the model. The aim of this thesis is to look at the inverse problem. We aim to assimilate palaeoclimate data, representing the historical evolution of climate variables through time, into the models. In doing so, we aim to learn about the best set of conditions for a model (parameter estimation), or the best models (model selection), in order to study the dynamics of the climate.

In this thesis we follow the Bayesian approach to statistics. In the Bayesian setting uncertainty is described probabilistically. Beginning with a model parameterised by some parameter vector, $\boldsymbol{\theta} \in \mathcal{R}^v$, we express our initial belief about $\boldsymbol{\theta}$ as a probability distribution, called the prior distribution, $\pi(\boldsymbol{\theta})$. The prior distribution reflects our initial uncertainty about the value of $\boldsymbol{\theta}$. When we are confident in a value for $\boldsymbol{\theta}$, then the prior distribution has a narrow support, favouring a small range of values. In contrast, a prior distribution with a wide support will be used when we are uncertain about what value $\boldsymbol{\theta}$ takes. We then wish to update our beliefs about $\boldsymbol{\theta}$ through assimilation of some set of data, here assumed to be $M$ discrete observations, $\boldsymbol{Y}_{1:M}$, where $\boldsymbol{Y}_m \in \mathcal{R}^w$. This is done through the likelihood function, $\pi(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta})$, which describes how well the data are explained for a given $\boldsymbol{\theta}$. Mathematically, we update our beliefs using Bayes formula:

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}) = \frac{\pi(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}{\pi(\boldsymbol{Y}_{1:M})}, \tag{1.2.1}$$

where the denominator $\pi(\boldsymbol{Y}_{1:M})$ is termed the model evidence, and is obtained by integrating out $\boldsymbol{\theta}$,

$$\pi(\boldsymbol{Y}_{1:M}) = \int \pi(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{1.2.2}$$

Bayes formula is very intuitive. The values of $\boldsymbol{\theta}$ that best match our initial beliefs, and best explain the data, are given the highest posterior probability density.

Bayes theorem allows the assimilation of palaeoclimate data in a principled framework. However, for the models in this thesis, the likelihood, $\pi(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta})$, is intractable, meaning that it is not available in analytical form. This restricts our approach to methods that either remove the likelihood term from the necessary calculations, or methods that use some approximation of the likelihood. Some applicable methods are introduced in Chapter 3.

## 1.3 Thesis Structure and Major Contributions

This thesis studies a number of uncertainty quantification problems related to palaeoclimate reconstruction. In particular, looking at parameter estimation, model selection, and dating uncertainties in sediment cores. The contribution to the scientific community can be summarised as follows:

- Numerous phenomonelogical models have been proposed. I have demonstrated how phenomenological models can be embedded in a state space model framework, allowing us to apply standard data assimilation schemes.
- The parameters in phenomonelogical models are often chosen by hand, so that the output of the model resembles the gross structure of palaeoclimate data. I provide an overview of approximate Bayesian computation (ABC) and particle filter methods for parameter estimation and model selection. These methods are compared in a simulation study, which can be considered as a benchmark.
- The design of efficient algorithms is explored. In particular I show how to design efficient proposals in the particle filter, and how to adaptively choose between accuracy and computational expense in an ABC-SMC algorithm.
- I apply these methods to palaeoclimate data using three models from the literature. In doing so, I show how a Bayesian model selection approach can be used to investigate ongoing questions in palaeoclimate science. In particular, I show whether palaeoclimate data support one phenomenological model over others, and whether the astronomical forcing adds explanatory power in these models. I demonstrate that the results of these experiments are sensitive to the methods in which sediment cores are dated.
- Motivated by the sensitivity of the model comparison experiments to the dating methods, I design a novel age model, and extend the particle filter to also perform filtering on the ages of observations. This algorithm provides a statistically rigorous characterisation of age model uncertainty in sediment cores. Using a model comparison experiment, I demonstrate that at least one dataset supports astronomically forced models when accounting for age model uncertainty.

A complete breakdown of this Thesis is as follows:

**Chapter 1**

We introduce the glacial-interglacial cycle and Milankovitch theory, which states that the glacial-interglacial cycle is paced by variations in the Earth's orbit over large timescales. We introduce $\delta^{18}$O, measurements of which provide the data on which we will perform inference. Finally, Bayes theorem is introduced, giving the foundation of the inference methods used throughout this thesis.

**Chapter 2**

We discuss the phenomenological modelling approach, and show how any of the phenomenological models in the literature can be embedded in a state space model (SSM) framework. We provide an introduction to SSMs, oscillators, and stochastic differential equations (SDEs), which are necessary to understand the phenomenological modelling approach. An observation process is developed in order to compare phenomenological models with observations of $\delta^{18}$O. We introduce three phenomenological models, which will act as benchmarks to our inference methods in Chapter 3 and Chapter 4.

**Chapter 3**

We introduce Approximate Bayesian Computation (ABC) and particle filter methods for parameter estimation for the phenomenological models of the climate discussed in Chapter 2. This problem is difficult, as the likelihood term required in Bayes theorem is intractable. ABC utilises repeated simulations from the model in order to remove the likelihood term from any analytical calculations, whereas particle filter methods provide an unbiased estimate of the likelihood that can then be used in calculations. We design a simulation study to compare the strengths and weaknesses, in particular comparing accuracy and computational cost, of each of these methods.

**Chapter 4**

We extend the inference methods introduced in Chapter 3 to perform model selection. We use this methodology to compare phenomenological models with different underlying dynamics. A simulation study is again used to assess the performance of each algorithm. We apply the methods to real-world data, showing that phenomenological models that undergo self-sustaining oscillations are more strongly supported by the data than steady-state models.

**Chapter 5**

We discuss the problems associated with using palaeoclimate data. In particular, we advocate using individual sediment cores rather than stacks. We apply the model selection methodology from Chapter 4 to investigate palaeoclimate model selection problems. Specifically, we assess the influence of the astronomical forcing in phenomenological models, and select between phenomenological models that have been proposed in the literature. It is shown that different dating assumptions have a strong impact on the conclusions.

**Chapter 6**

We introduce a novel age model based on a stochastic linear sediment accumulation model with down-core compaction. The particle filter is extended to include observation times in the filtering distribution. The concept of age-control points is also introduced, extending the observation process to include time information. This allows us to jointly perform parameter estimation, model selection, and quantification of the dating uncertainty. The new method performs well in a simulation study and is applied to two real-world datasets. The dating results are compared with the age estimates of the LR04 and H07 stacks. Our method gives results that are largely consistent with these estimates. A model selection experiment shows that one real-world dataset more strongly supports astronomically forced phenomenological models over unforced models.

**Chapter 7**

Concludes the work in this thesis and suggests future directions of research.

CHAPTER 2

# Phenomenological Modelling of the Palaeoclimate

The aim of this thesis is to develop methods for uncertainty quantification in problems related to palaeoclimate reconstruction. In this chapter we discuss the wide-range of models available to study climate dynamics, both short and long term. We motivate the use of phenomenological models, which are relatively simple in the small number of state variables and system parameters. These models are at the limit of complexity for which data can be assimilated in a formal Bayesian framework over the necessary timescales. To do so requires extending the models in the literature in two ways: Stochastic effects are added to account for model discrepancy, and the phenomenological models are embedded in a state space model (SSM) framework in order to relate the models with observations.

The chapter is set out as follows. In Section 2.1 we discuss modern approaches to climate modelling, and explain why we are limited in our choice of models if we want to perform inference about the evolution of the climate over millions of years. In Section 2.2 we give an overview of SSMs and discuss how phenomenological models of the climate can be embedded in this framework. In Section 2.3 we explain the main features of phenomenological models in the literature. We extend the phenomenological models to account for model discrepancy, and introduce an observation model in order to relate phenomenological models to palaeoclimate observations. We introduce three new models as a benchmark for the inference approaches that we will introduce in Chapter 3 and Chapter 4. In Section 2.4 we discuss the inference challenges, and in Section 2.5 we conclude the chapter.

## 2.1 Modelling the Climate

In recent decades scientists have aimed to understand the climate by constructing large numerical models that include as many of the physical processes of the climate system as possible. These models are referred to as General Circulation Models, or GCMs. GCMs work by dividing the Earth into a number of grid-cells. Physical processes are resolved within each cell, and interactions are modelled between neighbouring cells. GCMs aim to have a high resolution (a large number of cells) to enable physical processes that occur over relatively small spatial areas to be incorporated into the simulator. Many of the processes operating over small spatio-temporal intervals (either occurring over an area smaller than a cell, or within one integration timestep of the model), such as turbulence and cloud dynamics, are more crudely approximated. The complexity of GCMs means that they are computationally expensive, and so are typically designed to simulate only a few hundred years of the climate [10]. This approach has largely been driven by the need to understand the effect of the recent increase in the levels of greenhouse gases, with the aim of assessing the impact of climate change on a regional level over the next few decades.

The goal of palaeoclimate science is different. The aim is to understand processes that occur over thousands or millions of years. GCMs are ill-suited to study processes that evolve over a timescale of 100 kyr, such as the glacial-interglacial cycle. For this reason, much simpler models known as Earth Models of Intermediate Complexity (EMICs) have been proposed to model over longer timescales. Compared to GCMs, these models typically reduce the temporal and spatial resolutions dramatically [10]. Huge savings in computation time are obtained with such simplifications, allowing the study of large timescale elements of the climate. However, even at this stage the models may be considered oversimplified. Many of the dynamical processes that influence the glacial-interglacial cycle are poorly represented, if at all. It is impossible to account for all of the physical processes accurately enough to reproduce the glacial-interglacial cycles.

Since we lack the ability to exactly simulate the glacial-interglacial cycles, observations of previous glacial cycles need to be incorporated into the models. A standard approach is to vary the parameters of the model until the output is in good agreement with some training data [13]. This process is referred to as tuning, which is distinct from the more principled approach of parameter estimation (sometimes referred to as model calibration). Confidence in the model is gained if it is then able to reproduce observations that were not included in the training data. If this is not the case then the training dataset can be extended to include more observations before choosing a new set of parameters. There are a number of issues with the standard tuning approach. There may be multiple sets of parameters that give a convincing fit to the data, particularly when the data are sparse and noisy, and it can be difficult to determine if one set of parameters better explain the data than another. How the data are being incorporated, and deciding when a model is well tuned is often not made explicit. This is a cause for concern in the output of these models, as it can be difficult to trace conclusions back to the initial assumptions. Finally, by only incorporating part of the available data, important information may be omitted.

Recently interest has returned to using low-dimensional dynamical models, which can be considered as phenomenological models of the climate [11, 12]. In Figure 1.2 we can see that the glacial-interglacial cycle has a regular structure. Rather than model all of the physical and biological processes that contribute to glacial-interglacial cycles, phenomenological models allow us to model this regular behaviour directly. In such cases we only need to learn about processes on the large timescales of interest, ignoring the interactions between the short timescale elements of the climate system. Low-dimensional dynamical models have been more thoroughly explored than the large numerical climate models, and there is a huge amount of theory relating to the assimilation of data into such models owing to the low computational expense of performing simulations. Whilst this does not solve our inability to model the climate, it does give us well developed techniques that allow the available data to be incorporated in a principled framework. In particular, phenomenological models enable us to perform data assimilation in such a way as to accurately quantify our uncertainty, which is currently impossible for EMICs. A popular approach is to use a class of models known as oscillators. This modelling approach was introduced by Saltzman and Maasch, who proposed using an oscillator with three states representing ice volume, carbon dioxide concentration and deep-ocean temperature [22, 23]. The following decades saw the introduction of dozens of models following this approach [21, 24, 25]. Most of these have not been calibrated in a principled framework, but rather the parameters are altered until one state of the system broadly resembles observations of $\delta^{18}$O, much like the tuning approach used in EMICs and GCMs.

Whilst not explicitly discussed in the palaeoclimate literature, the proposed models can be embedded within an SSM framework. An SSM is comprised of a hidden Markov process such as the phenomenological models of the climate, and an observation process relating the hidden Markov process to observations. In regards to palaeoclimate inference, this observation process relates the phenomenological models of the climate to sediment and ice-core records. This is an ideal modelling approach, as we can relate the phenomenological models to observations through the observation process, even when components of the phenomenological model offer no physical interpretation, and we observe long-term climate variation only through proxy records. In the next section we more formally explain SSMs, and relate this modelling approach to phenomenological models in the literature.

## 2.2 State Space Models

We represent the phenomenological models of the climate as SSMs (also called Hidden Markov Models, HMMs, usually in the case of discrete state variables). SSMs describe the evolution of a system, and how the system relates to observations. The evolution of the system is represented by a hidden (indirectly observed) Markov process $\boldsymbol{X}_{1:M} = \{\boldsymbol{X}\left(t_1\right), ..., \boldsymbol{X}\left(t_M\right)\}$, where $t_1, ..., t_M$ represent a series of $M$ observation times and $\boldsymbol{X}_m \in \mathbb{R}^u$ is the state of the system at time $t_m$. The Markov process is described by an initial density $\pi(\boldsymbol{X}_1 \mid \boldsymbol{\theta})$ and transition probability density

$$\pi(\boldsymbol{X}_{m+1} \mid \boldsymbol{X}_{1:m}, \boldsymbol{\theta}) = \pi(\boldsymbol{X}_{m+1} \mid \boldsymbol{X}_m, \boldsymbol{\theta}), \qquad m \geq 1, \tag{2.2.1}$$

for some given static parameter $\boldsymbol{\theta} \in \mathbb{R}^v$. The Markov process $\boldsymbol{X}_{1:M}$ is observed indirectly through a set of observations $\boldsymbol{Y}_{1:M} = \{\boldsymbol{Y}\left(t_1\right), ..., \boldsymbol{Y}\left(t_M\right)\}$, where each $\boldsymbol{Y}_m \in \mathbb{R}^w$ are assumed to be conditionally independent given $\boldsymbol{X}_{1:M}$, and are related to the state of the system through an observation process with probability density

$$\pi(\boldsymbol{Y}_m \mid \boldsymbol{X}_{1:M}, \boldsymbol{\theta}) = \pi(\boldsymbol{Y}_m \mid \boldsymbol{X}_m, \boldsymbol{\theta}), \qquad 1 \leq m \leq M. \tag{2.2.2}$$

In the context of this thesis the glacial-interglacial cycle is represented by a dynamical model with state $\boldsymbol{X}_m = \left(X_1\left(t_m\right), ..., X_u\left(t_m\right)\right)^T$ representing climate variables of interest. Phenomenological models of the climate are typically low-dimensional, with the state of the system containing only $u = 2$ or $u = 3$ dimensions. The observations $\boldsymbol{Y}_m = Y\left(t_m\right)$ will be measurements of $\delta^{18}$O obtained from sediment cores. The state is related to the observations through $X_1(t)$, representing ice volume. The unobservable components of the state $X_2(t), ..., X_u(t)$ may have a physical interpretation, or may be physically undefined, but used to replicate features of the glacial-interglacial cycle in the evolution of $X_1(t)$.

## 2.3 Phenomenological Models of the Glacial-Interglacial Cycle

Many phenomenological models of the glacial-interglacial cycle follow the modelling approach of Saltzman and Maach [22, 23]. It is thus not surprising that they share several features:

- Only a few climate variables are explicitly included in the model.
- The climate variables evolve according to a dynamical system that exhibits self-sustaining oscillations. These oscillations stem either from nonlinearities in the equations governing the model [22, 23, 25], or from threshold criteria altering the system dynamics [26–28].

- A forcing term term is included in the model to account for the variation in energy received from the Sun due to changes in the Earth's orbit.

The concepts required to understand the construction of these models are introduced below.

### 2.3.1 Preliminaries

#### Oscillators

Here we include only such information as is necessary to understand the models discussed in this thesis. For the interested reader, an excellent introduction to oscillators and dynamical systems is available in [29], while a more rigorous approach to oscillator theory is available in [30].

In the context of dynamical systems theory an oscillator is a system that exhibits *self-sustaining* oscillations. Self-sustaining means that the system requires no external forcing to oscillate. Oscillators are characterised by a stable limit cycle, which is a closed trajectory in phase space to which all neighbouring trajectories are attracted. To demonstrate this, consider the Van der Pol oscillator [31], which is described by the system of ordinary differential equations (ODEs)

$$\begin{aligned} \frac{dX_1}{dt} &= -\frac{1}{\tau}X_2, \\ \frac{dX_2}{dt} &= \frac{\alpha}{\tau}\left(X_1 + X_2 - \frac{X_2^3}{3}\right), \end{aligned} \tag{2.3.1}$$

where $\tau$ is the timescale of the system and $\alpha$ separates the timescale of the two state variables. Dependence of $X_1$ and $X_2$ on $t$ has been suppressed for notational convenience. Figure 2.1 demonstrates how trajectories progress towards the limit cycle, the shape of which depends on the values of $\tau$ and $\alpha$.

The Van der Pol oscillator is an example of a *relaxation oscillator*. Relaxation oscillators are characterised by alternating relaxation and destabilisation dynamics [29]. We can see from Equation 2.3.1 that when the state is in the vicinity of $X_1 = \frac{X_2^3}{3} - X_2$ then $\frac{dX_2}{dt}$ will be very small, and the system will evolve slowly. This is the Van der Pol oscillator's destabilisation process, highlighted in Figure 2.1 and Figure 2.2. The two regions of destabilisation are often referred to as the *branches* of the limit cycle. In the destabilisation regime, the system gradually becomes unstable until it is ejected into a relaxation regime. During the relaxation process, the system is attracted to a region of phase space known as a relaxation state. Unlike the destabilisation process, the relaxation process is very quick. Such models are said to exhibit *slow-fast dynamics*. Relaxation oscillators do not necessarily need to exhibit slow-fast dynamics, but it is a common feature. Once the system has relaxed, it once again begins to destabilise. In the context of climate modelling, a number of relaxation oscillators are discussed in [21].

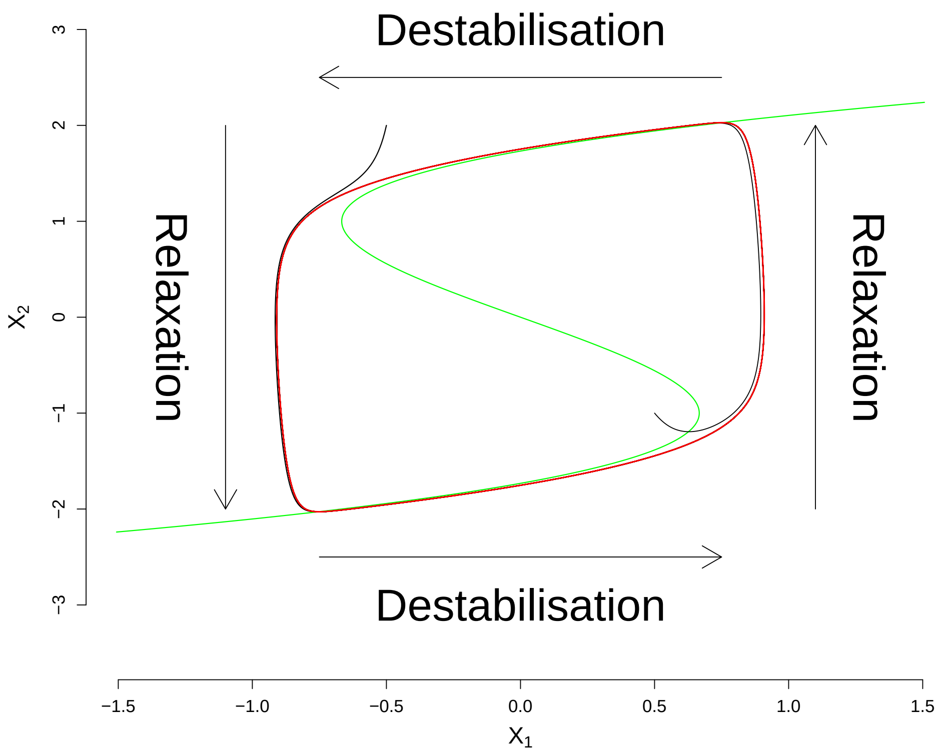

**Figure 2.1:** Phase space representation of two trajectories (black lines) of the Van der Pol oscillator described by Equation 2.3.1 with different initial conditions. The limit cycle is shown in red. The green line shows $X_1 = \frac{X_2^3}{3} - X_2$, near which $\frac{dX_2}{dt}$ is small, and the system is in a slow destabilisation regime.

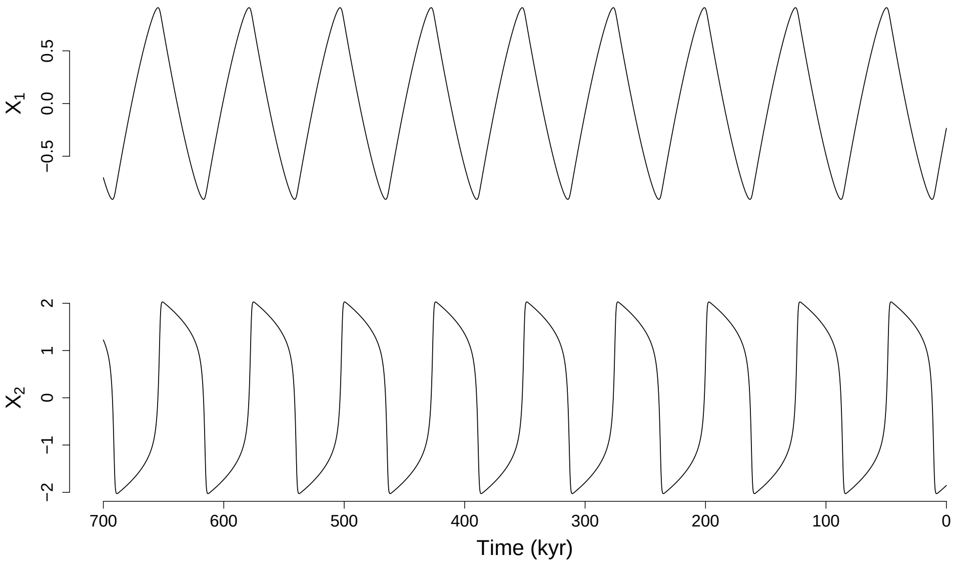

**Figure 2.2:** Trajectory of the Van der Pol oscillator described by Equation 2.3.1 over 700 kyr with $\tau = 31000$ and $\alpha = 20$.

When an oscillator is forced by a periodic or quasi-periodic forcing function the oscillator may become synchronised on the forcing. The forcing function will attract the trajectories of the oscillator into specific regions of state space. This means that if many trajectories are generated from the forced oscillator using the same parameters, but with different initial conditions, the trajectories will merge over time until only a few distinct trajectories remain. An example is provided in the next section. The remaining trajectories show the local pullback attractors of the system. The number and structure of the local pullback attractors are parameter dependent. No knowledge of pullback attractors is necessary for this thesis. It suffices to know that the astronomical forcing guides trajectories of oscillating systems into specific regions of state space, but for the interested reader a thorough discussion of pullback attractors in dynamical models of the palaeoclimate is given in [25].

### Astronomical Forcing

The astronomical theory of palaeoclimates is among the most popular ways of explaining the glacial-interglacial cycle. We only give an introduction here, but a historical perspective on the topic can be found in [2]. The theory holds that long-term variations in the energy received by the Earth from the Sun has a direct effect on pacing the glacial-interglacial cycles. These long-term spatial and seasonal variations are predominantly explained by variation in the Earth's orbital parameters, namely eccentricity, obliquity, and precession. These parameters are discussed in turn below [2, 32]:

- Eccentricity ($e$) is a measure of how much the Earth's orbit around the Sun differs from a perfect circle. A perfect circle is given by $e = 0$, and a parabolic trajectory is given by $e = 1$. For values $0 < e < 1$ the orbit forms an ellipse. Over time the eccentricity of the Earth's orbit varies from near circular ($e \gtrsim 0$) to slightly elliptical ($e \approx 0.07$). This leads to an increase in insolation during the Earth's closest approach to the Sun (perihelion), as well as altering the time spent in each season due to the Earth's orbital velocity increasing as the Earth-Sun distance decreases. The dominant period of eccentricity is $\sim$400 kyr, with additional $\sim$100 kyr cycles.

- Obliquity ($E$) is the angle between the plane of the Earth's orbit and the equatorial plane. The Earth's obliquity varies between approximately 22° (low obliquity) and 24.5° (high obliquity). A high obliquity leads to a higher amount of insolation being received during the summer season (for both hemispheres), and a lower amount of insolation over the winter season. The period of obliquity is $\sim$41 kyr.

- Precession ($\Pi = e \sin \tilde{\omega}$, where $e$ is eccentricity, and $\tilde{\omega}$ is the longitude of perihelion, which is a measure of the angular distance between perihelion and the vernal (spring) equinox) describes the variation in the direction of the Earth's axis of rotation over time, combining two phenomena. The first is that the Earth's axis of rotation over time behaves like a spinning top, such that the motion of the North pole describes a circle in space. The second is that the elliptical orbit of the Earth rotates in the orbital plane over time. If the Earth's axis of rotation points towards the Sun during perihelion there is a greater discrepancy between the amount of insolation received over the Northern Hemisphere summer than over the Northern Hemisphere winter. Conversely, if the axis of rotation points away from the Sun then there is less variation between the two seasons. The dominant period of precession is $\sim$23 kyr.

The astronomical forcing terms are well approximated by trigonometric series expansions of the form [2, 32]:

$$e = e^* + \sum_{i=1}^{n_e} e_i \cos\left(\omega_{e,i} t + \varphi_{e,i}\right), \tag{2.3.2}$$

$$\Pi = \sum_{i=1}^{n_P} P_i \sin\left(\omega_{P,i} t + \varphi_{P,i}\right), \tag{2.3.3}$$

$$E = E^* + \sum_{i=1}^{n_E} E_i \cos\left(\omega_{E,i} t + \varphi_{E,i}\right), \tag{2.3.4}$$

where the values of the amplitudes ($e_i$, $P_i$, $E_i$,), frequencies ($\omega_{e,i}$, $\omega_{P,i}$, $\omega_{E,i}$) and phases ($\varphi_{e,i}$, $\varphi_{P,i}$, $\varphi_{E,i}$) are given in [32] and [33]. It is standard to take at least 30 components in each sum for an accurate approximation.

When the astronomical theory of the glacial-interglacial cycle was first proposed it was thought that long cold winters and short hot summers were necessary for a glacial inception [2]. However, the prevailing viewpoint today is that long mild summers and short mild winters give preferential conditions for inception. The reasoning is that winter ice build-up is less likely to melt over the summer. This creates a positive feedback, as the additional sea-ice increases the surface albedo. This is most commonly referred to as the Milankovitch viewpoint [2].

Any model of the glacial-interglacial cycle needs to take in to account the astronomical forcing, so that the chance of an inception increases during favourable seasonal conditions. In dynamical systems, the Milankovitch viewpoint is included by forcing the system with some measure of the variation of insolation, known as the astronomical (or orbital) forcing. The most common approach is to use summer solstice insolation at 65°N, but alternative measures of insolation have been proposed that are still in-line with the Milankovitch viewpoint. For example, the insolation integrated over the caloric summer (defined as the half-year with the largest values of insolation) [2, 8], and insolation integrated over all days for which insolation exceeds some threshold [34] have both been used. These measures of insolation are well approximated by a

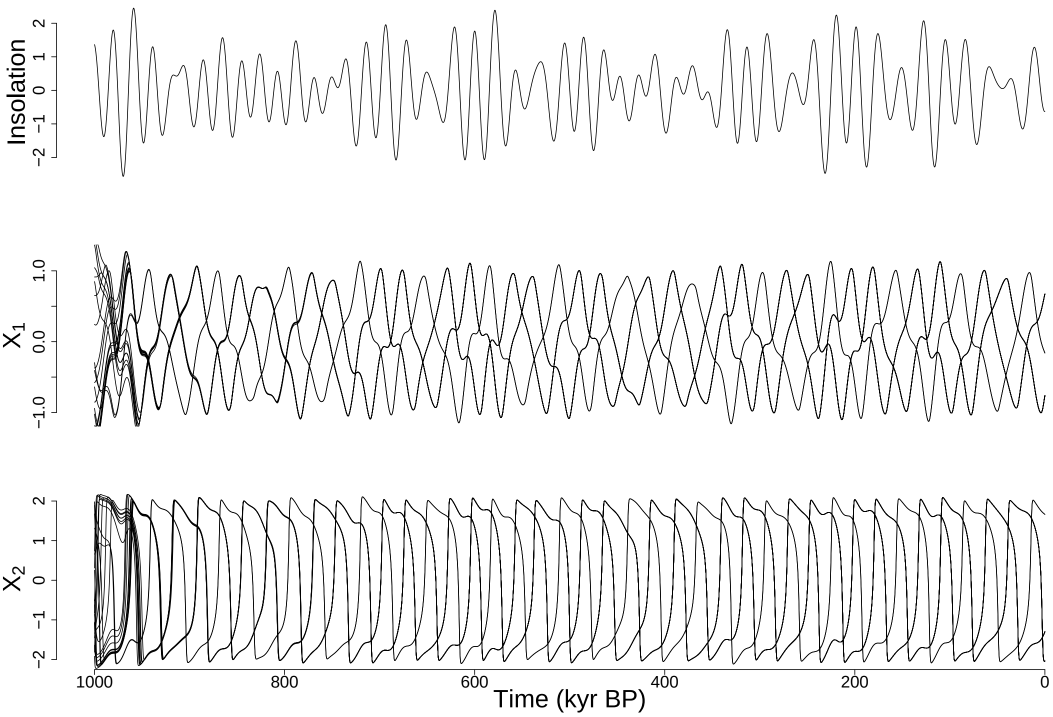

**Figure 2.3:** Top: Normalised summer solstice insolation at 65°N obtained using Equation 2.3.5 with $\gamma_P = 0.78$, $\gamma_C = 0$, and $\gamma_E = 0.38$, and the orbital solutions calculated in [32]. Middle and bottom: Trajectories of the forced Van der Pol oscillator with 20 random sets of initial conditions. The astronomical forcing attracts trajectories into specific regions of phase space, so that after 1 Myr only three distinct trajectories remain.

linear combination of precession, coprecession ($\mathit{\Pi} = e\cos\tilde{\omega}$, effectively controlling the phase of precession) and obliquity [25]:

$$I = \gamma_P \bar{\Pi} + \gamma_C \bar{\mathit{\Pi}} + \gamma_E \bar{E}, \tag{2.3.5}$$

where $\bar{\Pi}$, $\bar{\mathit{\Pi}}$ and $\bar{E}$ are normalised precession, coprecession and obliquity respectively, and $\gamma_P$, $\gamma_C$ and $\gamma_E$ are adimensional scaling parameters. Eccentricity is included only through the precession and coprecession terms. The coprecession term is often excluded, fixing the phase of precession, in which case $\gamma_C$ is set to 0 [25]. Summer solistice isolation at 65°N is well approximated by $\gamma_P = 0.8949$, $\gamma_C = 0$, and $\gamma_E = 0.4346$ [25]. The astronomical forcing is a complex, aperiodic signal, as can be seen in Figure 2.3. Note that the parameters have been rescaled to $\gamma_P = 0.78$, $\gamma_C = 0$, and $\gamma_E = 0.38$, in order to normalise the insolation signal.

Consider the Van der Pol oscillator as a model for the glacial-interglacial cycle. We assume that $X_1$ represents some measure of ice volume, and $X_2$ is left physically undefined. The oscillator needs to be forced in such a way that high values of insolation promote ice reduction, and low values of insolation promote ice growth. This is achieved by including the astronomical forcing in the ice volume equation with a negative coefficient. The ODEs now take the form

$$\begin{aligned}\frac{dX_1}{dt} &= -\frac{1}{\tau}\left(X_2 + I\left(\gamma_P, \gamma_C, \gamma_E\right)\right), \\ \frac{dX_2}{dt} &= \frac{\alpha}{\tau}\left(X_1 + X_2 - \frac{X_2^3}{3}\right).\end{aligned} \tag{2.3.6}$$

In Figure 2.3 the trajectories of the forced Van der Pol oscillator with parameters $\tau = 31000$, $\alpha = 20$, $\gamma_P = 0.78$, $\gamma_C = 0$, and $\gamma_E = 0.38$, are plotted with 20 random sets of initial conditions. The oscillator synchronises on the astronomical forcing, and after 1 Myr only three distinct trajectories remain, showing the local pullback attractors of the system.

The Van der Pol oscillator spends approximately the same amount of time on each branch of its limit cycle. If $X_1$ is taken to be ice volume then this model fails to capture the asymmetry between slow ice build-up and rapid melting shown in Figure 1.2. A modified version of the Van der Pol oscillator, dubbed CR12, included an additional asymmetry parameter [21, 25]:

$$\begin{aligned}\frac{dX_1}{dt} &= -\frac{1}{\tau}\left(\beta + X_2 + I\left(\gamma_P, \gamma_C, \gamma_E\right)\right), \\ \frac{dX_2}{dt} &= \frac{\alpha}{\tau}\left(X_1 + X_2 - \frac{X_2^3}{3}\right),\end{aligned} \tag{2.3.7}$$

where $\beta$ controls the relative time spent on each branch of the limit cycle. Selecting the parameters $\beta = 0.8, \tau = 31000$, $\alpha = 20$, $\gamma_P = 0.78$, and $\gamma_E = 0.38$ captures the saw-tooth shaped structure of the last 7 glacial-interglacial cycles, as shown in Figure 2.4. For the given parameters there is only a single pullback attractor.

### Stochastic Differential Equations

Phenomenological models have long been used to study the glacial-interglacial cycle [25]. Such simple models do not correspond particularly well to reality, due to the huge number of physical processes that are not included within them. It has previously been proposed that long timescale "climate" variations could be modelled explicitly, while short timescale "weather" variations could be approximated as stochastic perturbations [35]. This thesis follows the same approach, extending the dynamical systems of the climate by including stochastic fluctuations to account for the discrepancy between the models and real-world process. Our models are now represented by $u$ dimensional stochastic differential equations (SDEs) of the general form:

$$d\boldsymbol{X}\left(t\right) = \boldsymbol{\mu}\left(\boldsymbol{X}\left(t\right), \boldsymbol{\theta}\right)dt + \boldsymbol{\Sigma}_X^{\frac{1}{2}}\left(\boldsymbol{X}\left(t\right), \boldsymbol{\theta}\right)d\boldsymbol{W}\left(t\right), \tag{2.3.8}$$

where $\boldsymbol{\mu}\left(\boldsymbol{X}\left(t\right), \boldsymbol{\theta}\right) \in \mathbb{R}^u$ is called the drift function, $\boldsymbol{\Sigma}_X^{\frac{1}{2}}\left(\boldsymbol{X}\left(t\right), \boldsymbol{\theta}\right) \in \mathbb{R}^{u\times u}$ is called the diffusion function, and $\boldsymbol{W}\left(t\right) \in \mathbb{R}^u$ is a vector of independent Brownian motions. An understanding

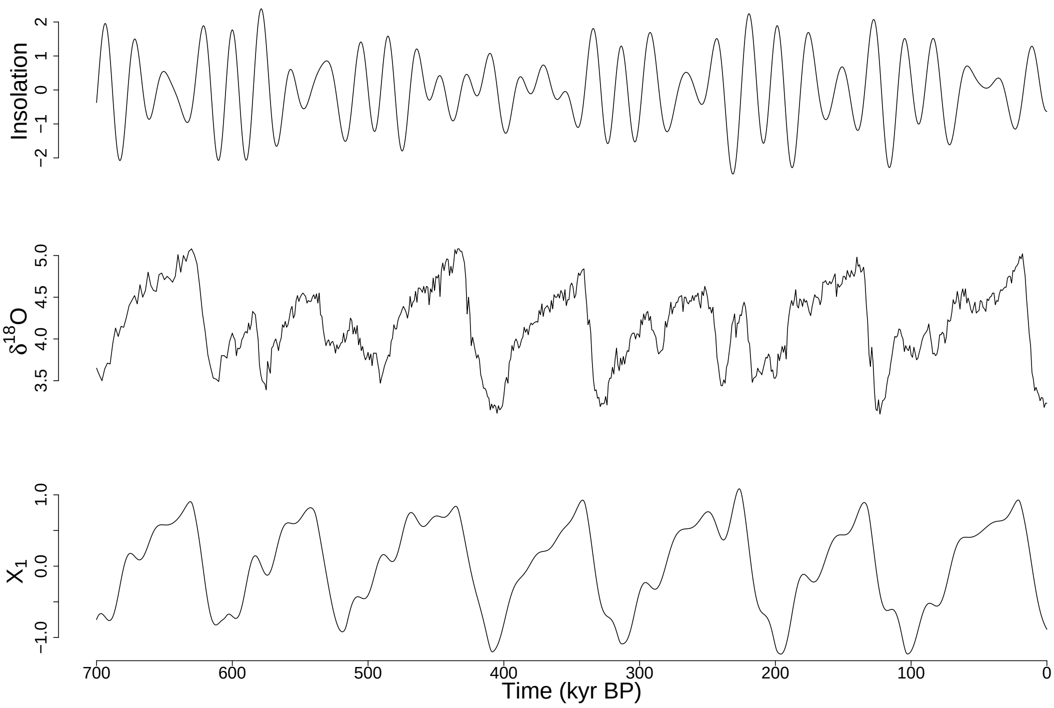

**Figure 2.4:** Top: Normalised summer solstice insolation at 65°N obtained using Equation 2.3.5 with $\gamma_P = 0.78$, $\gamma_C = 0$, and $\gamma_E = 0.38$, and the orbital solutions calculated in [32]. Middle: Observed $\delta^{18}$O over 700 kyr from the LR04 stack [1]. Bottom: Observable state of CR12, as described by Equation 2.3.7, with parameters are $\beta = 0.8$, $\tau = 31000$, and $\alpha = 20$. This set of parameters captures the structure of the recent glacial-interglacial cycles.

of the theoretical properties of SDEs is not required to understand this thesis, but an introduction to stochastic calculus is available in [36] for interested readers. For all cases in this thesis the diffusion function is independent of the state of the system and uncorrelated, so that

$$\boldsymbol{\Sigma}_X\left(\boldsymbol{X}\left(t\right),\boldsymbol{\theta}\right) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_u^2 \end{pmatrix}$$

is a $u \times u$ dimensional diagonal matrix.

Within this framework, consider adding stochastic perturbations to CR12, so that the equations become

$$\begin{aligned} dX_1 &= -\frac{1}{\tau}\left(\beta + X_2 + I\left(\gamma_P, \gamma_C, \gamma_E\right)\right) dt + \sigma_1 dW_1, \\ dX_2 &= \frac{\alpha}{\tau}\left(X_1 + X_2 - \frac{X_2^3}{3}\right) dt + \sigma_2 dW_2, \end{aligned} \tag{2.3.9}$$

which now contains two additional parameters, $\sigma_1$ and $\sigma_2$, that dictate the strength of the stochastic perturbations for each state variable. The model is no longer deterministic. Even

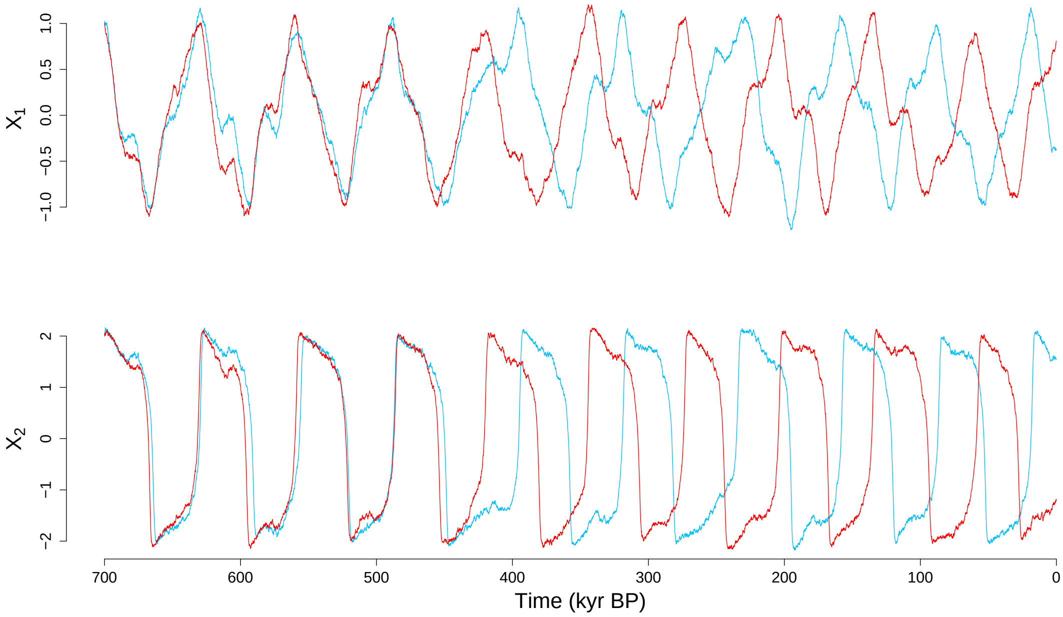

**Figure 2.5:** Two trajectories from CR12 with stochastic perturbations, as described by Equation 2.3.9, with parameters $\beta = 0.8$, $\tau = 31000$, $\alpha = 20$, $\sigma_1 = 0.2$, $\sigma_2 = 0.3$, $\gamma_P = 0.78$, $\gamma_C = 0$, and $\gamma_E = 0.38$. The two trajectories go out of phase near 400 kyr BP, but maintain the gross structure of the model.

with fixed parameters and initial conditions, two trajectories from this model will never be identical. However, when the stochastic forcing is relatively weak (small values of $\sigma_1$ and $\sigma_2$) the gross structure of the model is unchanged, as shown for CR12 in Figure 2.5. This is a consequence of the system's limit cycle; if the state of the system is perturbed away from the limit cycle then the following trajectory will be attracted back towards the limit cycle.

Since the astronomical forcing attracts the trajectory of the system into specific regions of state-space, it should also help to preserve the gross structure of the model when stochastic perturbations are added. However, in regimes in which there are multiple local pullback attractors, the stochastic perturbations may shift the trajectory from one region of attraction to another. This is demonstrated in Figure 2.6, where the system follows the trajectory of a local pullback attractor, undergoes a period of desynchronisation, and eventually synchronises on a different local pullback attractor.

When $\sigma_1$ and $\sigma_2$ are increased, the stochastic perturbations can cause a relaxation oscillator to enter a relaxation regime early, or cause a delay in entering one. Picturing the system's limit cycle, this can be seen as being ejected from a branch of the limit cycle earlier (noise-induced escape), or later (noise-induced delay) than the deterministic model. These phenomena are demonstrated for the Van der Pol oscillator in Figure 2.7. Typically, noise-induced escapes will dominate, reducing the period of oscillation [12]. In the context of the glacial-interglacial cycle, noise-induced phenomena can alter the glacial inception and termination times dramatically.

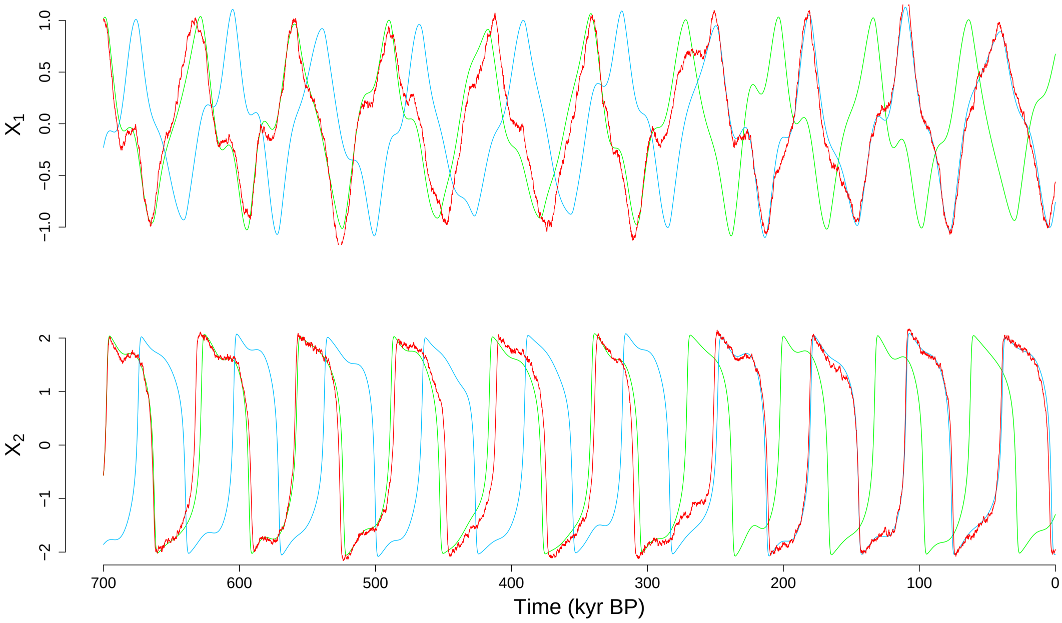

**Figure 2.6:** Green and blue lines: Two local pullback attractors of the deterministic CR12 model with parameters $\beta = 0$, $\tau = 31000$, $\alpha = 20$, $\gamma_P = 0.78$, $\gamma_C = 0$, and $\gamma_E = 0.38$. Red line: Realisation of the stochastic CR12 model with parameters $\sigma_1 = 0.2$, and $\sigma_2 = 0.3$. Stochastic perturbations cause the trajectory to switch synchronisation between two different local pullback attractors.

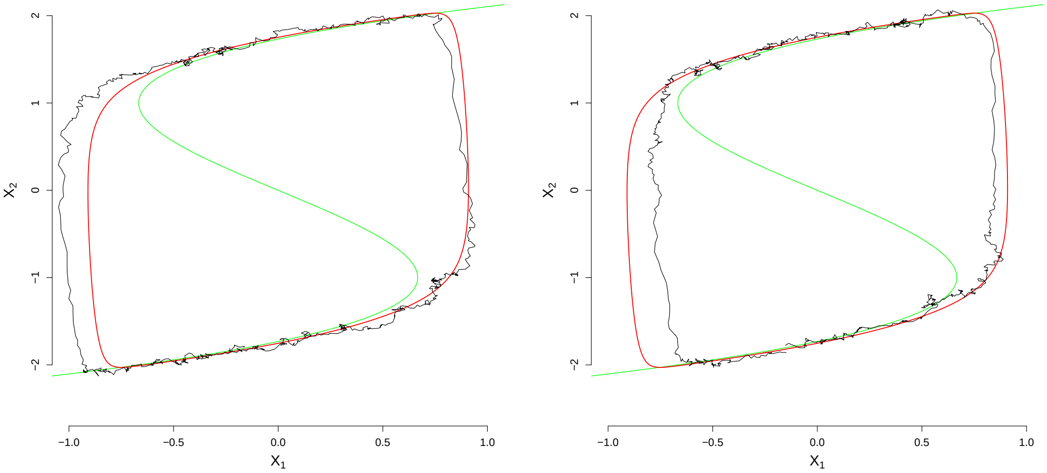

**Figure 2.7:** Left: An example of a noise-induced delay on the upper-left part of the limit cycle. Right: An example of a noise-induced escape on the upper-left part of the limit cycle. Red line shows the limit cycle of the Van der Pol oscillator, and the green line shows the solution to $X_1 = \frac{X_2^3}{3} - X_2$.

**Algorithm 2.1** The Euler-Murayama method.

```
Partition the time interval $[0,T]$ into $J$ equal subintervals of width $\Delta t = \frac{T}{J}$:
    $$(t_0 = 0, t_1 = \Delta t, ..., t_J = J\Delta t).$$
Select the initial conditions, $\boldsymbol{X}(t_0)$, and parameters, $\boldsymbol{\theta}$.
for $j = 1, ..., J$ do
    Set
    $$\boldsymbol{X}(t_j) = \boldsymbol{X}(t_{j-1}) + \boldsymbol{\mu}(\boldsymbol{X}(t_{j-1}), \boldsymbol{\theta})\Delta t + \boldsymbol{\Sigma}_X^{\frac{1}{2}}(\boldsymbol{X}(t_{j-1}), \boldsymbol{\theta})\sqrt{\Delta t}\boldsymbol{\epsilon}_{j-1},$$
    where $\boldsymbol{\epsilon}_{j-1} \in \mathbb{R}^u$ is a vector of independent standard Gaussian random variables.
end for
```

### Simulating SDE trajectories

The inference methods in this thesis require that we are at least able to simulate from the model. For SDEs an analytical solution is available in only a small number of cases. In recent years exact algorithms have been proposed that offer approaches to exactly simulate from some SDEs without analytical solutions [37, 38], but these can not be applied for the majority of the models in this thesis. However, there are numerous methods available to approximately simulate from an SDE. Since the inference methods introduced later in the thesis require a large number of simulations from the model it is important to use a computationally efficient approach. We will be using a commonly used, first order, discrete-time approximation, called the Euler-Maruyama method [36, 39]. For a generic SDE of the form given in Equation 2.3.8, an approximate numerical solution over $[0,T]$ is calculated according to Algorithm 2.1. As an example, a trajectory from CR12 is generated using the discretised equations

$$\begin{aligned} X_1(t+\Delta t) &= -\frac{1}{\tau}(\beta + X_2(t) + I(t;\gamma_P,\gamma_C,\gamma_E))\Delta t + \sigma_1\sqrt{\Delta t}\epsilon_1(t), \\ X_2(t+\Delta t) &= \frac{\alpha}{\tau}\left(X_1(t) + X_2(t) - \frac{X_2(t)^3}{3}\right)\Delta t + \sigma_2\sqrt{\Delta t}\epsilon_2(t), \end{aligned} \tag{2.3.10}$$

where $\epsilon_1(t)$ and $\epsilon_2(t)$ are independent realisations of a standard Gaussian random variable. The Euler-Maruyama method is exact in the limit $\Delta t \to 0$, with the accuracy of the approximation decreasing as $\Delta t$ increases. A standard method to choose the step size involves pre-generating a Brownian motion and comparing a simulated trajectory for different step sizes, repeated over a range of parameter values. The step size is chosen so that reducing the step size has little impact on the resulting trajectory. On CR12 it was shown that the Euler-Maruyama method provides suitable accuracy as long as the system is stable, in the sense that the numerical approximation does not diverge to $\pm\infty$ [40]. This is achieved with a time-step of $\Delta t = 0.1$ kyr.

### 2.3.2 Dynamical Models of the Palaeoclimate

CR12 is a typical phenomenological model of the glacial-interglacial cycle. There are an abundance of models in the literature that treat the glacial-interglacial cycle as a limit cycle synchronised on the astronomical forcing. These models almost always follow the Milankovitch viewpoint, but the dynamics of the models can differ greatly. However, it has been suggested that any dynamical model with self-sustaining oscillations with a period of roughly 100 kyr that is forced by the astronomical forcing can reasonably resemble the features of the last few glacial cycles [10]. With this in mind, we introduce three models that capture much of the variety of the models in the literature. These models were developed by Michel Crucifix, and so are referred to as CR14-a, CR14-b, and CR14-c [41]. They will be used as a benchmark for the inference methods introduced in Chapter 3, and allows us to test whether any one modelling approach is better supported by the data (Chapter 4).

#### CR14-a

$$\begin{aligned} dX_1 &= -\left(\beta_0 + \beta_1 X_1 + \beta_2\left(X_1^3 - X_1\right) + \delta X_2 + I(\gamma_P, \gamma_C, \gamma_E)\right) dt + \sigma_1 dW_1 \\ dX_2 &= \alpha\delta\left(X_1 + X_2 - \frac{X_2^3}{3}\right) dt + \sigma_2 dW_2 \end{aligned} \tag{2.3.11}$$

CR14-a builds on the CR12 model shown in Equation 2.3.7 by adding additional dynamics to the ice volume equation, and mirrors numerous phenomenological models exhibiting slow-fast dynamics. The introduction of $\beta_1$ adds a linear response to the ice volume, while $\beta_2$ is introduced to keep the system stable when $\beta_1$ is negative. This model has ten tunable parameters, $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \alpha, \delta, \gamma_P, \gamma_C, \gamma_E, \sigma_1, \sigma_2)^T$.

#### CR14-b

$$\begin{aligned} dX_1 = &-\left(\beta_0 + \beta_1 X_1 + \beta_2\left(X_1^3 - X_1\right) + I(\gamma_P, \gamma_C, \gamma_E) + \right. \\ &\qquad\qquad\qquad\qquad \left. \delta\mathcal{H}\left(X_2 - \kappa_0 - \kappa_1 X_1\right)\right) dt + \sigma_1 dW_1 \\ dX_2 = &\ \alpha\left(X_1 - X_2\right) dt + \sigma_2 dW_2 \end{aligned} \tag{2.3.12}$$

CR14-b represents another popular class of models in which oscillations are induced through a threshold function. In CR14-b the threshold function is introduced through the Heaviside step function, $\mathcal{H}$, whose argument is a measure of the difference between the two state variables. Beyond this threshold the system loses ice volume at rate $\delta$. This can be thought of as a “flushing” mechanism: ice volume accumulates until a critical value is reached and the system flushes the ice from the system, returning to low ice volume conditions. This model has twelve tunable parameters, $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \alpha, \delta, \kappa_0, \kappa_1, \gamma_P, \gamma_C, \gamma_E, \sigma_1, \sigma_2)^T$, two more than the slow-fast model.

It is important to note that although the parameters share the same symbols as the slow-fast system, many of the parameters are non-comparable as they play a different role depending on the dynamics of the system.

**CR14-c**

$$\begin{aligned} dX_1 &= -\left(\beta_0 + \beta_1 X_1 + \beta_2\left(X_1^3 - X_1\right) + \right.\\ &\qquad\left. \delta\mathcal{H}\left(X_2 - \kappa_0 - \kappa_1 X_1 + I(\gamma_P, \gamma_C, \gamma_E)\right)\right) dt + \sigma_1 dW_1 \\ dX_2 &= \alpha\left(X_1 - X_2\right) dt + \sigma_2 dW_2 \end{aligned} \tag{2.3.13}$$

CR14-c is a minor modification to CR14-b. The forcing function is now included as part of the threshold function, so that ice volume responds nonlinearly to insolation. As with CR14-b there are twelve tunable parameters, $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \alpha, \delta, \kappa_0, \kappa_1, \gamma_P, \gamma_C, \gamma_E, \sigma_1, \sigma_2)^T$. Again, note that the astronomical forcing parameters will likely take different values than CR12-b, due to the different responses between ice volume and insolation.

### 2.3.3 Relating to Observations

Phenomenological models have a limited ability to represent the climate. Even the most complex numerical simulators that include as many physical processes as possible need to be tuned using some kind of data. Using a dynamical model within an SSM framework allows us to relate the state of the system to observations through an observation model. The observation model is designed to represent measurement error and/or uncertainty in how observations relate to the state of the system. By far the most common observation model is to assume that observations are a noisy version of the observable states, that is,

$$\boldsymbol{Y}_m = \boldsymbol{h}^T \boldsymbol{X}_m + \boldsymbol{\Sigma}_Y^{\frac{1}{2}} \boldsymbol{\eta}_m, \tag{2.3.14}$$

where $\boldsymbol{h} \in \mathbb{R}^u$ is a vector comprised of ones for observable states and zeros for unobservable states, $\boldsymbol{\eta}_m \in \mathbb{R}^w$ is a vector of standard Gaussian random variables, and $\boldsymbol{\Sigma}_Y \in \mathbb{R}^{w\times w}$ is a diagonal matrix scaling the measurement error.

The observations used in this thesis are measurements of $\delta^{18}$O, whereas the ice volume response is modelled on a variety of different scales. For example, the models introduced in this chapter, among many others, are adimensional (in that the components of the state vector have no physical dimensions), aiming to capture general features observed in past glacial-interglacial cycles. Even in models where the observable state aims to represent ice volume physically there are multiple possible representations, such as ice volume ($km^3$) [27], or equivalent sea-level change ($m$) [28]. The observation model above needs to be extended in order to account for the different scales of observations and models. In vector form we use

$$\boldsymbol{Y}_m = \boldsymbol{D} + \boldsymbol{C}^T \boldsymbol{X}_m + \boldsymbol{\Sigma}_Y^{\frac{1}{2}} \boldsymbol{\eta}_m, \tag{2.3.15}$$

which is reduced to scalar form for most of our applications, so the observation process becomes

$$Y_m = D + \boldsymbol{C}^T \boldsymbol{X}_m + \Sigma_Y^{\frac{1}{2}} \eta_m, \tag{2.3.16}$$

where we have 3 tunable parameters: $D$ displaces the observable component, $\boldsymbol{C}^T = (C, 0, ..., 0)$ rescales it, and $\Sigma_Y = \sigma_Y^2$ scales the measurement error.

Using an observation model that depends on a subset of the state variables allows the unobservable variables to be physically undefined. This is common in regards to phenomenological models of the glacial-interglacial cycle, as it allows the study of the dynamics of glacial-interglacial cycles without the concern of linking the model to specific physical processes. For models which explicitly include more than one climate variable, it is possible to extend the observation model to include additional data sources. Examples would be SM90 and SM91, which explicitly model variation in $CO_2$, and so $CO_2$ records could be included [22, 23].

## 2.4 Inference Challenges

This chapter has dealt with the development and motivation of phenomenological models of the palaeoclimate. A model allows us to specify a number of parameters and obtain a set of possible observations. The aim of this thesis is to study the inverse problem: given a set of data we want to determine which sets of parameters are best supported by the data. In doing so, we aim to infer information about the dynamics of the climate over large timescales. In a Bayesian context this means estimating the posterior distribution given in Equation 1.2.1. In an SSM, the joint posterior distribution of the parameters $\boldsymbol{\theta}$ and the state of the system $\boldsymbol{X}_{1:M}$ at observation times $t_1, ..., t_M$ is

$$\pi\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:M} \mid \boldsymbol{Y}_{1:M}\right) \propto \pi\left(\boldsymbol{\theta}\right) \pi\left(\boldsymbol{X}_1 \mid \boldsymbol{\theta}\right) \prod_{m=2}^{M} \pi\left(\boldsymbol{X}_m \mid \boldsymbol{X}_{m-1}, \boldsymbol{\theta}\right) \prod_{m=1}^{M} \pi\left(\boldsymbol{Y}_m \mid \boldsymbol{X}_m, \boldsymbol{\theta}\right), \tag{2.4.1}$$

from which we obtain the marginal posterior distribution of the parameters by integrating out the state:

$$\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right) = \int \pi\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:M} \mid \boldsymbol{Y}_{1:M}\right) d\boldsymbol{X}_{1:M}. \tag{2.4.2}$$

Calculating the posterior distribution is non-trivial. For multivariate nonlinear SDEs the transition density $\pi\left(\boldsymbol{X}_m \mid \boldsymbol{X}_{m-1}, \boldsymbol{\theta}\right)$ is not available in closed-form, and when using palaeoclimate data the observations are too sparsely distributed (typically one observation every 2-3 kyr) to approximate the transition density using the Euler-Maruyama method with interval size $\Delta t = t_m - t_{m-1}$. Estimating the transition density requires partitioning the interval $\Delta t$ into further sub-intervals of length $\Delta\tau = \frac{\Delta t}{J}$, over which the first-order approximation provides a reasonable estimate. Doing so introduces $(J-1) \times u$ random variables, $\boldsymbol{X}(t_{m-1} + \Delta\tau), ..., \boldsymbol{X}(t_{m-1} + (J-1)\Delta\tau)$, that need to be integrated out:

$$\pi(\boldsymbol{X}_m \mid \boldsymbol{X}_{m-1}, \boldsymbol{\theta}) \approx \int \prod_{j=1}^{J} \pi(\boldsymbol{X}(t_{m-1} + j\Delta\tau) \mid \boldsymbol{X}(t_{m-1} + (j-1)\Delta\tau), \boldsymbol{\theta}) \\ \times d(\boldsymbol{X}(t_{m-1} + \Delta\tau), ..., \boldsymbol{X}(t_{m-1} + (J-1)\Delta\tau)).$$

To give a sense of scale, recall that $u$ is small (usually 2-3), and note that 20-30 sub-intervals will typically be required between every pair of observations so that $\Delta\tau = 0.1$ kyr. Monte Carlo methods are usually necessary to evaluate this integral. These are a collection of algorithms that use random sampling to obtain numerical solutions, some examples of which are provided in the next chapter.

Even with a reasonable approximation to the transition density, the fact that the system is only partially observed, and observed with measurement error, means that neither the likelihood, $\pi(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta})$, or the model evidence, $\pi(\boldsymbol{Y}_{1:M})$, are available in closed form. Such distributions are called intractable. This restricts the availability of methods to evaluate, or sample from, the posterior distribution. Fortunately, with the recent growth in available computer power, many computational approaches to sample from the posterior distribution have been proposed in recent years [42–44]. These methods form the basis of our inference approach in this thesis.

## 2.5 Chapter Summary

In this Chapter we have described modern approaches to climate modelling. In particular, we have motivated the use of a phenomenological modelling approach. Phenomenological models are relatively simple, modelling only a small number of climate variables explicitly. We have described the key components in the construction of phenomenological models, and demonstrated how these can be embedded within an SSM framework, which allows us to formally relate the phenomenological models to palaeoclimate data. The SSM framework provides the foundation from which we perform statistical inference using these models throughout this thesis.

Even though phenomenological models are relatively simple, performing inference on them is extremely challenging, owing to the intractable likelihood. In the next two chapters we introduce state of the art inference methods, for parameter estimation and model comparison respectively, that can be used without the need for a tractable likelihood. We will then use three phenomenological models: CR14-a, CR14-b, and CR14-c, introduced in this chapter, to compare the relative performance of the proposed inference methods using a common benchmark.

CHAPTER 3

# Inference Methods

The focus in this chapter is parameter estimation for the class of models described in Chapter 2. This is a complex problem, as an analytical solution to the posterior distribution is unavailable. Statisticians have long employed Monte Carlo methods in an attempt to solve such problems. These are a collection of computer algorithms that use random sampling to obtain numerical solutions. However, the intractable likelihood in our models means that many commonly used Monte Carlo methods can not be applied. Fortunately, over the last decade, numerous approaches have been proposed for performing inference for models with intractable likelihoods, owing to the recent surge in the availability of computing power. We present two of these approaches in this chapter. The first is approximate Bayesian computation (ABC), which is a Monte Carlo approach that can be used in intractable problems where it is possible to simulate observations from the model for a given set of parameters. The second is to use the particle filter to estimate the likelihood, and is more tailored towards SSMs.

The chapter is divided as follows. In Section 3.1 we describe the statistical tools necessary to understand the inference methods used in this chapter. In Section 3.2 we design a simulation study in order to test and compare the different inference methods. In Section 3.3, we introduce ABC, and consider a number of variants based on different sampling techniques. In Section 3.4 we describe the particle filter, as well as inference methods that make use of likelihood estimation in order to sample from the posterior distribution. In Section 3.5 we summarise the chapter, and discusses the relative strengths and weaknesses of each approach.

## 3.1 Preliminaries

The inference methods introduced in this chapter expand on a number of well-known sampling methods. These are introduced below for reference later. We present each sampling scheme as a method to obtain a sample of size $N$ from the posterior distribution, $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right)$, where the data are assumed to be ordered in time. More general forms of each algorithm can be found in the references provided.

### 3.1.1 Rejection Sampling

The rejection algorithm [45] is a Monte Carlo approach for sampling from a probability distribution, which we will take to be the posterior distribution, $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right)$. Parameter values are drawn from a proposal distribution $q\left(\boldsymbol{\theta}\right)$, and accepted with probability

$$\mathcal{P} = \frac{\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right)}{Cq\left(\boldsymbol{\theta}\right)},$$

where $C$ is a constant such that $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right) < Cq\left(\boldsymbol{\theta}\right)$, for all $\boldsymbol{\theta}$. When the model evidence, $\pi\left(\boldsymbol{Y}_{1:M}\right)$, is unknown, the sampled parameters can be accepted with probability

$$\mathcal{P} = \frac{\pi\left(\boldsymbol{\theta}\right)\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}\right)}{C'q\left(\boldsymbol{\theta}\right)},$$

where $C'$ is a constant such that $\pi\left(\boldsymbol{\theta}\right)\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}\right) < C'q\left(\boldsymbol{\theta}\right)$, for all $\boldsymbol{\theta}$. The pseudocode is presented in Algorithm 3.1. The collection of accepted parameter values are independent samples from the posterior distribution $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right)$.

The advantages of the rejection algorithm are that it is easy to implement and it provides independent samples, so that it can be run in parallel on multiple computer cores. The major drawback is the difficulty in designing efficient proposal distributions. For example, a common choice when targeting the posterior distribution is to sample from the prior distribution, $\pi\left(\boldsymbol{\theta}\right)$, so that the acceptance probability is

$$\mathcal{P} = \frac{\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}\right)}{C'}.$$

This choice will often lead to many proposed values in regions of low posterior probability density, giving a very low acceptance probability. The greater the disparity between the prior and posterior distributions, the more samples are required in order to obtain a sample of the desired size.

**Algorithm 3.1** Rejection sampling algorithm targeting $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right)$.

```
  Set $n = 1$.
  while $n \leq N$ do
    Sample $\boldsymbol{\theta}^*$ from the proposal distribution, $q\left(\boldsymbol{\theta}\right)$.
    With probability
        $\mathcal{P} = \frac{\pi\left(\boldsymbol{\theta}^*\right)\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}^*\right)}{C' q\left(\boldsymbol{\theta}^*\right)}$
    set $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}^*$ and $n = n + 1$.
  end while
```

### 3.1.2 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) [46] is among the most popular Monte Carlo approaches for performing inference. The two most famous MCMC algorithms are the Metropolis Hastings sampler, and the Gibbs sampler (which is a special case of the Metropolis Hastings sampler). A Metropolis Hastings algorithm for sampling from the posterior distribution, $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M})$, is presented in Algorithm 3.2.

The presented Metropolis Hastings algorithm produces a Markov chain, $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(N)}$, which converges to the stationary distribution, $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M})$, under mild conditions. A common way to check that the posterior distribution is a stationary distribution of the Markov chain is to ensure that the detailed balance equation is satisfied [47]. The detailed balance equations are

$$\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right)\mathcal{P}\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}\right) = \pi\left(\boldsymbol{\theta}^* \mid \boldsymbol{Y}_{1:M}\right)\mathcal{P}\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*\right), \qquad \forall \boldsymbol{\theta}, \boldsymbol{\theta}^*, \tag{3.1.1}$$

where

$$\mathcal{P}\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}\right) = \lambda\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}\right) q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}\right) + \left(1 - \lambda\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}\right)\right)\delta_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^*\right) \tag{3.1.2}$$

is the transition kernel of the resulting Markov chain. Satisfying the detailed balance equation is a stronger condition than is necessary, but is a simple check, and the one most commonly used. The stationary distribution is unique as long as the resulting Markov chain is ergodic. This is ensured, for example, by choosing a proposal distribution so that $q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}\right) > 0$, for all $\boldsymbol{\theta}^*$ with posterior support, meaning $\pi(\boldsymbol{\theta}^* \mid \boldsymbol{Y}_{1:M}) > 0$. Note that this is a sufficient condition, but not necessary.

These guidelines produce a Markov chain with the posterior distribution, $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M})$, as the unique stationary distribution, from an arbitrary initial value, $\boldsymbol{\theta}^{(1)}$. However, if the initial parameter value is in a region of low posterior probability density, then it will take time for the chain to converge, and the initial samples might poorly represent the stationary distribution. In this case a burn-in period will be necessary, in which a number of samples at the beginning of the chain are discarded. It is up to the user to determine when a chain has converged to the sta-

**Algorithm 3.2** Metropolis Hastings algorithm targeting $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M})$.

```
Initialise $\boldsymbol{\theta}^{(1)}$.
for $n = 2, ..., N$ do
    Propose move to $\boldsymbol{\theta}^*$ according to proposal density $q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(n-1)}\right)$.
    With probability
        $$\lambda\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(n-1)}\right) = \min\left(1, \frac{\pi\left(\boldsymbol{\theta}^*\right)\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}^*\right) q\left(\boldsymbol{\theta}^{(n-1)} \mid \boldsymbol{\theta}^*\right)}{\pi\left(\boldsymbol{\theta}^{(n-1)}\right)\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}^{(n-1)}\right) q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(n-1)}\right)}\right)$$
    set $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}^*$. Otherwise set $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}^{(n-1)}$.
end for
```

tionary distribution, and hence, how long of a burn-in period to use. A number of convergence diagnostics are discussed in [47]. One standard approach is to run multiple MCMC chains with different initial values. After some burn-in period the output from each chain should appear similar. Alternatively, it might be possible to obtain an initial sample from the posterior distribution using another sampling scheme, avoiding the need for a burn-in period.

The Metropolis Hastings algorithm requires the user to choose a number of settings. This is referred to as 'tuning' the algorithm, and is usually dependent on the specific model and dataset. Most notably, the user is required to choose the proposal distribution, $q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}\right)$. Common choices for the proposal distribution are:

- A random-walk sampler, usually with symmetric proposals, so that $q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}\right) = q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*\right)$. Using a symmetric proposal has the advantage of removing the transition density from the acceptance probability ratio. Choosing the variance of the proposal distribution can be difficult, as a small variance can lead to the chain becoming trapped in local modes, and a large variance can lead to many proposals in regions of low posterior probability density, lowering the acceptance rate, and slowing convergence. Guidelines for choosing the variance of a Gaussian proposal distribution can be found in [48].

- An independence sampler, where proposals are independent of the current state of the Markov chain, so that $q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}\right) = q\left(\boldsymbol{\theta}^*\right)$. Selecting an efficient distribution can be difficult. If many proposals lie in regions of low posterior probability density, or if few proposals lie in regions of high posterior probability density, then it will take a long time for the chain to converge to the stationary distribution. The optimal proposal distribution is $q\left(\boldsymbol{\theta}^*\right) = \pi(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M})$, for which the acceptance probability is one (direct sampling of the posterior distribution). For situations in which MCMC is applied this choice is not available, but choosing proposal distributions resembling the posterior distribution should give good performance.

The proposal distribution does not need to remain the same over the entire chain. For example, random-walk moves can alter between low-variance and large-variance proposal distributions. When the parameter is multidimensional, the proposal distribution can be designed to update subsets of the parameter vector in every iteration, rather than the entire parameter vector. Updating the components of the parameter vector one at a time, using the full conditional distribution of the parameter given all of the other parameters and the data, gives a specialised case of the Metropolis Hastings algorithm called the Gibbs Sampler. In a Gibbs sampler the acceptance probability is always one, so that every proposal is accepted. However, the full conditional distributions are often not available, and even when they are the mixing of the chain can be poor [47].

Compared to the rejection algorithm, MCMC methods increase the number of proposals in regions of high posterior probability density. However, samples from an MCMC chain are correlated. This means that an MCMC sample of the posterior distribution will contain less information than an independent sample of the same size. Ideally, the chain should have a small auto-correlation, and if so, it is said that the chain is fast mixing. Conversely, a large auto-correlation shows that the chain is slow mixing. Mixing is a consequence of the choice of proposal distribution, and so designing an efficient MCMC algorithm often requires multiple training runs in order to fine-tune the proposal distribution. If the chain is mixing slowly, it is common to thin the output by taking every $k$th value, and so to obtain a sample of size $N$ the chain will need to be of length $Nk$ following the burn-in.

### 3.1.3 Importance Sampling

Importance sampling (IS) [46] is used for estimating expectations with respect to some target distribution which is known pointwise up to a normalising constant, and from which obtaining samples is difficult. When targeting the posterior distribution, $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right)$, IS can be used when $\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)$ is known pointwise, but the model evidence, $\pi\left(\boldsymbol{Y}_{1:M}\right)$, is unknown. Suppose we wish to evaluate the expectation of some test function, $\varphi\left(\boldsymbol{\theta}\right)$, with respect to the posterior distribution. We define an an importance distribution, $\eta\left(\boldsymbol{\theta}\right)$, with the property that $\eta\left(\boldsymbol{\theta}\right) > 0$ if $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right) > 0$. Then the following identities hold:

$$\begin{aligned}
\mathbb{E}_{\pi}\left(\varphi\left(\boldsymbol{\theta}\right)\right) &= \pi\left(\boldsymbol{Y}_{1:M}\right)^{-1} \int \varphi\left(\boldsymbol{\theta}\right) \pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}, \\
&= \pi\left(\boldsymbol{Y}_{1:M}\right)^{-1} \int \varphi\left(\boldsymbol{\theta}\right) w\left(\boldsymbol{\theta}\right) \eta\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}, \\
&= \mathbb{E}_{\eta}\left(\varphi\left(\boldsymbol{\theta}\right) w\left(\boldsymbol{\theta}\right)\right)
\end{aligned} \tag{3.1.3}$$

**Algorithm 3.3** Importance sampling algorithm targeting $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right)$.

```
for $n = 1, ..., N$ do
    Sample $\boldsymbol{\theta}^{(n)}$ from the importance distribution $\eta\left(\boldsymbol{\theta}\right)$.
    Set the importance weight
        $$w^{(n)} = w\left(\boldsymbol{\theta}^{(n)}\right) = \frac{\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}^{(n)}\right)\pi\left(\boldsymbol{\theta}^{(n)}\right)}{\eta\left(\boldsymbol{\theta}^{(n)}\right)}.$$
end for
```

and

$$\begin{aligned} \pi\left(\boldsymbol{Y}_{1:M}\right) &= \int \pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta} \\ &= \int w\left(\boldsymbol{\theta}\right)\eta\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta}, \end{aligned} \tag{3.1.4}$$

where

$$w\left(\boldsymbol{\theta}\right) = \frac{\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)}{\eta\left(\boldsymbol{\theta}\right)} \tag{3.1.5}$$

are called the (unnormalised) importance weights. The IS algorithm samples $N$ parameter values (termed particles) from the importance distribution, $\eta\left(\boldsymbol{\theta}\right)$, giving the Monte Carlo approximation

$$\hat{\eta}\left(\boldsymbol{\theta}\right) \approx \frac{1}{N}\sum_{n=1}^{N}\delta_{\boldsymbol{\theta}^{(n)}}\left(\boldsymbol{\theta}\right), \tag{3.1.6}$$

where $\delta\left(\cdot\right)$ is the Dirac delta distribution. This Monte Carlo approximation is then substituted into Equations 3.1.3 and 3.1.4 to obtain estimates of $\mathbb{E}_{\pi}\left(\varphi\left(\boldsymbol{\theta}\right)\right)$ and $\pi\left(\boldsymbol{Y}_{1:M}\right)$ respectively. The pseudocode is presented in Algorithm 3.3. The weighted particles give the following Monte Carlo approximation of the target distribution

$$\hat{\pi}\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right) \approx \sum_{n=1}^{N} W^{(n)}\delta_{\boldsymbol{\theta}^{(n)}}\left(\boldsymbol{\theta}\right), \tag{3.1.7}$$

where $W^{(n)}$ are the normalised importance weights, defined as

$$W^{(n)} = \frac{w^{(n)}}{\sum_{n=1}^{N} w^{(n)}}. \tag{3.1.8}$$

The choice of importance distribution is determined by the user, and should be given careful consideration. If the importance distribution is heavy tailed with respect to the posterior distribution, then a large number of proposals will lie in regions of low posterior probability density and have low weights. On the other hand, if the importance distribution has light tails in regions of large posterior probability density then any proposals from the tail of the importance distribution will be given very large weights. In either case, the Monte Carlo approximation will be dominated by a small number of high-weight particles, giving a poor approximation. The variance of the weights is often monitored by using the effective sample size (ESS) [49], defined as

$$\text{ESS} = \left( \sum_{n=1}^{N} \left( W^{(n)} \right)^2 \right)^{-1}. \tag{3.1.9}$$

The ESS takes a value between one, where a single particle has all of the weight, to $N$, for an equally weighted sample. Hence, a small ESS indicates that many of the particles have very little weight, and contain little information. This means that an importance distribution should strongly resemble the target distribution, in order to minimise the variance of the weights [50]. In most cases selecting an efficient importance distribution is very difficult, leading to importance sampling approaches being less widely used than MCMC methods.

### 3.1.4 Sequential Importance Sampling

The difficulty in finding efficient importance distributions motivated the development of a sequential importance sampling (SIS) approach [50]. A sequence of intermediary distributions $\pi_s, s = 1, ..., S$, are introduced with

$$\pi_s = \frac{\gamma_s}{Z_s}, \tag{3.1.10}$$

where each $\gamma_s$ is known pointwise, and the normalising constant $Z_s$ is unknown. As $s$ increases, the intermediary distributions, $\pi_s$, evolve from an initial distribution, $\pi_1$, from which it is easy to sample, to the target distribution, $\pi_S$. The $\pi_s$ are chosen such that successive distributions are similar. In our context, $\pi_1$ will typically be the prior distribution, $\pi(\boldsymbol{\theta})$, and $\pi_S$ will be the posterior distribution, $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M})$. A sequence of importance distributions, $\eta_s, s = 1, ..., S$, are then used to obtain weighted samples from each target distribution using IS. By targeting the intermediary distributions sequentially, it is possible to design efficient importance distributions by using the previous sample. In other words, since successive distributions are similar, it is possible to move a sample targeting $\pi_{s-1}$ into regions of high probability density in $\pi_s$.

In the first iteration of the SIS algorithm, IS is used to sample $N$ particles (called the first population), $\boldsymbol{\theta}_1^{(1:N)}$, with importance weights $w_1^{(1:N)}$, from the initial distribution $\pi_1$. In the next iteration these particles are then perturbed using a Markov transition kernel $K_2$, with associated transition density $K_2(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)$. The new collection of particles (population two) are then marginally distributed according to

$$\eta_2(\boldsymbol{\theta}_2) = \int \eta_1(\boldsymbol{\theta}_1) K_2(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1. \tag{3.1.11}$$

When $\eta_2(\boldsymbol{\theta})$ can be evaluated pointwise, it can be used as an importance distribution for targeting $\pi_2(\boldsymbol{\theta})$. This process repeats until a population is sampled from the target distribution, $\pi_S$,

**Algorithm 3.4** Sequential importance sampling algorithm targeting $\pi_S\left(\boldsymbol{\theta}\right)$.

```
for $n = 1, ..., N$ do
  Sample $\boldsymbol{\theta}_1^{(n)}$ from the importance distribution $\eta_1\left(\boldsymbol{\theta}\right)$.
  Set the importance weight
      $$w_1^{(n)} = \frac{\gamma_1\left(\boldsymbol{\theta}_1^{(n)}\right)}{\eta_1\left(\boldsymbol{\theta}_1^{(n)}\right)}.$$
end for
for $s = 2, ..., S$ do
  for $n = 1, ..., N$ do
    Sample $\boldsymbol{\theta}_s^{(n)}$ from the importance distribution
      $$\eta_s\left(\boldsymbol{\theta}_s\right) = \int \eta_{s-1}\left(\boldsymbol{\theta}_{s-1}\right) K_s\left(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_{s-1}\right) d\boldsymbol{\theta}_{s-1}.$$
    Set the importance weight
      $$w_s^{(n)} = \frac{\gamma_s\left(\boldsymbol{\theta}_s^{(n)}\right)}{\eta_s\left(\boldsymbol{\theta}_s^{(n)}\right)}.$$
  end for
end for
```

as shown in Algorithm 3.4.

In most cases, being able to evaluate $\eta_s\left(\boldsymbol{\theta}\right)$ pointwise requires independent Markov transition densities, such that $K_s\left(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_{s-1}\right) = K_s\left(\boldsymbol{\theta}_s\right)$. This is highly restrictive, as in many situations it may be advantageous to use local moves, such as a random-walk proposal. Sometimes an approximation is used in place of $\eta_s\left(\boldsymbol{\theta}\right)$, for example

$$\hat{\eta_s}\left(\boldsymbol{\theta}\right) = \frac{1}{N}\sum_{n=1}^{N} K_s\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{s-1}^{(n)}\right), \tag{3.1.12}$$

which can be evaluated for random-walk proposals. However, it is not possible to evaluate $K_s\left(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_{s-1}\right)$ pointwise in many situations. Moreover, the approximation is computationally expensive as the sum needs to be evaluated for every particle.

### 3.1.5 Sequential Monte Carlo Samplers

A Sequential Monte Carlo (SMC) approach was introduced as a way to overcome the inability to evaluate $\eta_s\left(\boldsymbol{\theta}\right)$ for $s \geq 2$ in SIS [50]. Since the initial importance distribution is tractable, SMC samplers use IS to target the extended distributions

$$\tilde{\pi}_s\left(\boldsymbol{\theta}_{1:s}\right) = \frac{\tilde{\gamma}_s\left(\boldsymbol{\theta}_{1:s}\right)}{Z_s}, \tag{3.1.13}$$

allowing the intractable importance distributions to be related to the initial (tractable) importance distribution through the auxiliary variables $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{s-1}$. This is achieved by introducing artificial backward Markov kernels $L_{s-1}$ with density $L_{s-1}(\boldsymbol{\theta}_{s-1} \mid \boldsymbol{\theta}_s)$, chosen by the user, so that

$$\tilde{\gamma}_s\left(\boldsymbol{\theta}_{1:s}\right) = \gamma_s\left(\boldsymbol{\theta}_s\right) \prod_{i=2}^{s} L_{s-1}\left(\boldsymbol{\theta}_{s-1} \mid \boldsymbol{\theta}_s\right). \tag{3.1.14}$$

Each $\tilde{\pi}_s\left(\boldsymbol{\theta}_{1:s}\right)$ admits $\pi_s\left(\boldsymbol{\theta}_s\right)$ as a marginal distribution, obtained by integrating out the auxiliary variables,

$$\pi_s\left(\boldsymbol{\theta}_s\right) = \int \tilde{\pi}_s\left(\boldsymbol{\theta}_{1:s}\right) d\boldsymbol{\theta}_{1:s-1}.$$

In the SMC algorithm, the initial population is obtained using IS, as in the SIS algorithm. In later iterations, the path of each particle is extended using a Markov kernel $K_s$ with transition density $K_s\left(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_{s-1}\right)$. That is, sample $\boldsymbol{\theta}_s^* \sim K_s\left(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_{s-1}\right)$ and set $\boldsymbol{\theta}_{1:s}^{(n)} = \left\{\boldsymbol{\theta}_{1:s-1}^{(n)}, \boldsymbol{\theta}_s^*\right\}$. This gives an importance distribution, $\eta_s\left(\boldsymbol{\theta}_{1:s}^{(n)}\right)$, used to perform importance sampling on $\tilde{\pi}_s\left(\boldsymbol{\theta}_{1:s}^{(n)}\right)$, where the importance weights are given by

$$\begin{aligned} w_s^{(n)} &= \frac{\tilde{\gamma}_s\left(\boldsymbol{\theta}_{1:s}^{(n)}\right)}{\eta_s\left(\boldsymbol{\theta}_{1:s}^{(n)}\right)} && (3.1.15) \\ &= w_{s-1}^{(n)} \frac{\gamma_s\left(\boldsymbol{\theta}_s^{(n)}\right) L_{s-1}\left(\boldsymbol{\theta}_{s-1}^{(n)} \mid \boldsymbol{\theta}_s^{(n)}\right)}{\gamma_{s-1}\left(\boldsymbol{\theta}_{s-1}^{(n)}\right) K_s\left(\boldsymbol{\theta}_s^{(n)} \mid \boldsymbol{\theta}_{s-1}^{(n)}\right)}. && (3.1.16) \end{aligned}$$

The pseudocode is presented in Algorithm 3.5.

As $s$ increases, it becomes harder to design a good importance distribution. As described in Section 3.1.3, this can lead to a small number of particles with relatively large weights. In the SMC framework this is known as degeneracy. The level of degeneracy is often monitored by calculating the ESS (refer to Equation 3.1.9) in every iteration, before the particles are perturbed. If the ESS falls below some threshold (usually set as half of the sample size, $\frac{N}{2}$) then a resampling step is applied. In the resampling step, $N$ particles are sampled with replacement from the current population, $\boldsymbol{\theta}_{1:s}^{(1:N)}$, according to their weights $W_s^{(1:N)}$. Possible resampling methods are discussed in [49]. The resampled set of particles are given weights $\frac{1}{N}$, so that the ESS returns to $N$. The resampling step adds computational expense, but by replacing particles that poorly represent the target distribution, the approximation in later populations should be improved (but not the current population). It is usually not optimal to resample in every population, as if there is little variability between the weights then a resampling step will reduce the number of distinct particles, discarding some information. Choosing to only resample when the ESS is below some threshold also has the effect of reducing the additional computational expense. It is important to monitor the ESS for every population. A particularly small ESS will

**Algorithm 3.5** Sequential Monte Carlo sampling algorithm targeting $\pi_S(\boldsymbol{\theta})$.

```
for $n = 1, ..., N$ do
    Sample $\boldsymbol{\theta}_1^{(n)}$ from the importance distribution $\eta_1(\boldsymbol{\theta})$.
    Set the importance weight
        $$w_1^{(n)} = \frac{\gamma_1\left(\boldsymbol{\theta}_1^{(n)}\right)}{\eta_1\left(\boldsymbol{\theta}_1^{(n)}\right)}.$$
end for
Normalise the weights. For $n = 1, ..., N$
        $$W_1^{(n)} = \frac{w_1^{(n)}}{\sum_{i=1}^{N} w_1^{(i)}}.$$
for $s = 2, ..., S$ do
    if ESS<$\frac{N}{2}$ then
        Resample $\boldsymbol{\theta}^{(1:N)}$ according to weights $W_{s-1}^{(1:N)}$.
        Set the importance weights. For $n = 1, ..., N$
            $$W_{s-1}^{(n)} = \frac{1}{N}.$$
    end if
    for $n = 1, ..., N$ do
        Draw $\boldsymbol{\theta}_s^{(n)}$ from transition density $K_s\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{s-1}^{(n)}\right)$.
        Set the importance weight
            $$w_s^{(n)} = W_{s-1}^{(n)} \frac{\gamma_s\left(\boldsymbol{\theta}_s^{(n)}\right) L_{s-1}\left(\boldsymbol{\theta}_{s-1}^{(n)} \mid \boldsymbol{\theta}_s^{(n)}\right)}{\gamma_{s-1}\left(\boldsymbol{\theta}_{s-1}^{(n)}\right) K_s\left(\boldsymbol{\theta}_s^{(n)} \mid \boldsymbol{\theta}_{s-1}^{(n)}\right)}.$$
    end for
    Normalise the weights. For $n = 1, ..., N$
        $$W_s^{(n)} = \frac{w_s^{(n)}}{\sum_{i=1}^{N} w_s^{(i)}}.$$
end for
```

likely mean that the approximations given in later iterations are poor. In extreme cases, if the ESS becomes too low, then the SMC sampler should be restarted or redesigned.

The backward Markov kernels $L_{s-1}$ are arbitrary, but should be chosen to give the optimal performance of the algorithm with respect to $K_s$. For example, choosing $L_{s-1} = K_s$ simplifies the weight calculations, but leads to poor performance in most cases [50, 51]. The optimal sequence of backward Markov kernels $L_{s-1}$ minimises the variance of the weights. This requires backward Markov kernels with transition densities

$$L_{s-1}\left(\boldsymbol{\theta}_{s-1} \mid \boldsymbol{\theta}_s\right) = \frac{\eta_{s-1}\left(\boldsymbol{\theta}_{s-1}\right) K_s\left(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_{s-1}\right)}{\eta_s\left(\boldsymbol{\theta}_s\right)}, \tag{3.1.17}$$

which give the importance weights

$$w_s^{(n)} = \frac{\gamma_s\left(\boldsymbol{\theta}_s\right)}{\eta_s\left(\boldsymbol{\theta}_s\right)}. \tag{3.1.18}$$

Recalling that this algorithm was proposed as $\eta_s(\boldsymbol{\theta}_s)$ is typically intractable for $s \geq 2$, the optimal backwards Markov kernels are impossible to calculate in most cases. However, choosing backward Markov kernels that are approximations of the optimal backward Markov kernels should still lead to good performance [50].

One possibility is to substitute $\pi_{s-1}\left(\boldsymbol{\theta}\right)$ in place of $\eta_{s-1}\left(\boldsymbol{\theta}\right)$. This gives a backwards Markov kernel of the form

$$L_{s-1}\left(\boldsymbol{\theta}_{s-1} \mid \boldsymbol{\theta}_s\right) = \frac{\pi_{s-1}\left(\boldsymbol{\theta}_{s-1}\right) K_s\left(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_{s-1}\right)}{\int \pi_{s-1}\left(\boldsymbol{\theta}_{s-1}\right) K_s\left(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_{s-1}\right) d\boldsymbol{\theta}_{s-1}}, \tag{3.1.19}$$

which eliminates $\eta_s(\boldsymbol{\theta}_s)$ from the importance weight calculation, which becomes

$$w_s^{(n)} = W_{s-1}^{(n)} \frac{\gamma_s\left(\boldsymbol{\theta}_s\right)}{\int \gamma_{s-1}\left(\boldsymbol{\theta}_{s-1}\right) K_s\left(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_{s-1}\right) d\boldsymbol{\theta}_{s-1}}. \tag{3.1.20}$$

Alternatively, if the forward Markov kernels, $K_s$, are MCMC kernels with stationary distribution $\pi_s\left(\boldsymbol{\theta}\right)$, then a good choice for the backward Markov kernels are the reversed MCMC kernels associated with $K_s$. The backward Markov transition density then takes the form

$$L_{s-1}\left(\boldsymbol{\theta}_{s-1} \mid \boldsymbol{\theta}_s\right) = \frac{\pi_s\left(\boldsymbol{\theta}_{s-1}\right) K_s\left(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_{s-1}\right)}{\pi_s\left(\boldsymbol{\theta}_s\right)}, \tag{3.1.21}$$

and the importance weights become

$$w_s^{(n)} = W_{s-1}^{(n)} \frac{\gamma_s\left(\boldsymbol{\theta}_{s-1}\right)}{\gamma_{s-1}\left(\boldsymbol{\theta}_{s-1}\right)}. \tag{3.1.22}$$

Since the importance weights do not depend on $\boldsymbol{\theta}_s$, the sampling and resampling stages can be reversed. This is expected to be a good approximation to the optimal backwards kernel if $\pi_{s-1}(\boldsymbol{\theta}) \approx \pi_s(\boldsymbol{\theta})$ [50].

By propagating particles through a sequence of intermediary distributions that close in on the target distribution, SMC gives improved efficiency over rejection sampling and importance sampling. SMC also has a number of advantages over MCMC methods:

- There is no need to assess convergence in SMC applications, which can be problematic when using MCMC algorithms.
- SMC is less likely to become trapped in regions of low posterior probability density, or in local modes. These are common problems in MCMC when, for instance, the variance of proposal distribution is too small.
- Samples from $\tilde{\pi}_{s-1}(\boldsymbol{\theta}_{1:s-1})$ can be informative in designing a proposal distribution for $\tilde{\pi}_{1:s}(\boldsymbol{\theta}_s)$. This allows the choice of proposal distributions to be more easily automated than in MCMC.

The main drawback is that degeneracy is often a problem, and so it is important to monitor the ESS for every population.

## 3.2 Simulation Study

In this chapter we have introduced some standard Monte Carlo samplers. In the following sections we extend these samplers so that they can be used on problems with intractable likelihoods. It is important to assess the accuracy of the proposed inference algorithms to gain confidence in the results when they are applied to real-world data. In this chapter, the various proposed algorithms will be tested in a simulation study. A dataset is generated using a model with known parameters, and the inference methods are applied to this dataset using the correct model, but with uncertain model parameters. We simulate the dataset using CR14-a, reproduced below,

$$\begin{aligned} dX_1 &= -\left(\beta_0 + \beta_1 X_1 + \beta_2\left(X_1^3 - X_1\right) + \delta X_2 + I(\gamma_P, \gamma_C, \gamma_E)\right) dt + \sigma_1 dW_1 \\ dX_2 &= \alpha\delta\left(X_1 + X_2 - \frac{X_2^3}{3}\right) dt + \sigma_2 dW_2, \end{aligned} \tag{3.2.1}$$

and draw observations from the observation process

$$Y(t) = D + CX_1(t) + \sigma_Y \eta_t. \tag{3.2.2}$$

In order to represent a typical sediment core, we take observations every 2 kyr over a period of 780 kyr, giving 391 observations. Since we are interested in the limit cycle and synchroni-

| Parameter | True Value | Prior Distribution |
|---|---|---|
| $\beta_0$ | 0.65 | $\mathcal{N}\left(0.4, 0.3^2\right)$ |
| $\beta_1$ | 0.2 | $\mathcal{N}\left(0, 0.4^2\right)$ |
| $\beta_2$ | 0.5 | $exp\left(1/0.5\right)$ |
| $\delta$ | 0.5 | $exp\left(1/0.5\right)$ |
| $\alpha$ | 11 | $\Gamma\left(10, 2\right)$ |
| $\gamma_P$ | 0.2 | $exp\left(1/0.3\right)$ |
| $\gamma_C$ | 0.1 | $exp\left(1/0.3\right)$ |
| $\gamma_E$ | 0.3 | $exp\left(1/0.3\right)$ |
| $\sigma_1$ | 0.2 | $exp\left(1/0.3\right)$ |
| $\sigma_2$ | 0.5 | $exp\left(1/0.5\right)$ |
| $\sigma_y$ | 0.1 | $exp\left(1/0.1\right)$ |
| $D$ | 4.1 | $\mathcal{U}\left(3, 5\right)$ |
| $C$ | 0.8 | $\mathcal{U}\left(0.5, 2\right)$ |
| $X_1\left(t_1\right)$ | -1.02 | $\mathcal{U}\left(-1.5, 1.5\right)$ |
| $X_2\left(t_1\right)$ | 0.33 | $\mathcal{U}\left(-2.5, 2.5\right)$ |

**Table 3.1:** List of parameters used to generate data for the simulation study, and the associated prior distributions used in the statistical analysis. The prior distributions have been chosen in order to give a high probability of the model being in an oscillating regime, and to ensure that the astronomical forcing parameters are positive, following Milankovitch theory.

sation properties of the model, we begin the trajectory at 1 Myr BP and discard the first 220 kyr (called a spin-up period, similar to an MCMC burn-in period), so that the trajectory has had time to synchronise on the astronomical forcing prior to the first observation. If the inference methods are performing as intended then we expect that the 'true' parameter values (the values used to generate the dataset) lie in regions of high posterior probability density. They may not be the modal values, as the data may more strongly support different parameter values, the prior density could be larger for other parameter values, and we expect Monte Carlo variation (the sampling variability from using a finite sample size) in the results. On the other hand, if poor estimates are obtained, then it is easier to understand where and why problems occur than if we were using real data.

We also compare the results for the different inference schemes, where we check that each scheme provides consistent posterior density estimates, and compare the proposed schemes for efficiency in generating a sample from the posterior distribution. Since these algorithms have been implemented on a variety of hardware platforms, we compare the computational efficiency of each algorithm using the number of required simulations of the model, since this is the most computationally expensive component in each of the algorithms.

The selected parameter values are shown in Table 3.1 along with the associated prior distributions. Since palaeoclimate data are typically sparse and noisy, we ideally want to incorporate expert knowledge into the prior distributions to augment the limited information contained in palaeoclimate records. This process is known as elicitation, and an overview can be found in [52, 53]. We elicit prior distributions from an expert in dynamical systems theory [40]. This is a difficult and imprecise process for phenomenological models. Many of the model parameters do not represent physical quantities, and so prior distributions must be based on the experts' knowledge of how the model response depends upon the parameters, and how the model then relates to climate. Consequently the prior distributions are all somewhat vague. However, using prior distributions that are less informative than we might initially have hoped for is preferable to using overly-confident distributions. In general, the prior distributions are chosen in order to give a high probability of the model being in an oscillating (rather than excitable) regime. The astronomical forcing parameters are positive, following Milankovitch theory, with the possibility that the system is unforced. These choices are based on the theory that the glacial-interglacial cycle originates from internal climate dynamics, and paced by the astronomical forcing.

## 3.3 Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) is a class of methods for sampling from a posterior distribution when the likelihood is intractable, but where it is still possible to simulate from the model. ABC was initially developed in the field of genetics [54, 55]. Since the likelihood is not required to be known analytically, ABC approaches are often termed "likelihood-free" methods.

ABC utilises repeated simulations from the model, which are compared with observations. Consider the rejection sampling algorithm given in Section 3.1.1. This algorithm can not be used in cases where the likelihood is intractable, as it is required to calculate the acceptance probability. However, the rejection sampling algorithm can be modified to remove the need for a tractable likelihood by using a two-step process that simulates from the model, and compares the simulated values, $\tilde{\boldsymbol{Y}}_{1:M}$, with the observed dataset, $\boldsymbol{Y}_{1:M}$. Assuming parameter values are sampled from the prior distribution, if there is equality between the simulated values and the observed dataset then the proposed parameters are accepted, otherwise they are rejected. The pseudocode is presented in Algorithm 3.6. As with the rejection sampling algorithm, the accepted parameter values form a sample from the posterior distribution.

This modified rejection algorithm can only be applied in cases where equality between simulations from the model and the observed dataset occurs with non-zero probability. Even in such situations the probability that simulated values will exactly match the observed data is typically very small. This makes the acceptance rate of the modified rejection sampling algorithm prohibitively poor, with a huge number of simulations being required to obtain a reasonably sized sample. The first ABC algorithm was proposed to overcome these difficulties. Based on the

**Algorithm 3.6** Modified rejection sampling algorithm targeting $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right)$.

```
Set $n = 1$.
while $n \leq N$ do
    Sample $\boldsymbol{\theta}^*$ from the prior distribution, $\pi(\boldsymbol{\theta})$.
    Simulate value $\tilde{\boldsymbol{Y}}^*_{1:M}$ from the model using parameter $\boldsymbol{\theta}^*$.
    if $\tilde{\boldsymbol{Y}}^*_{1:M} = \boldsymbol{Y}_{1:M}$ then
        Set $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}^*$ and $n = n + 1$.
    end if
end while
```

rejection sampling algorithm, it replaces the acceptance condition in Algorithm 3.6 to accept a proposal if the simulated values are 'close' to the observed dataset [55]. Closeness is determined according to some distance measure, $\rho\left(\tilde{\boldsymbol{Y}}_{1:M}, \boldsymbol{Y}_{1:M}\right)$, between the simulated values and the observed data, which needs to be less than some tolerance, $\epsilon \geq 0$, for an acceptance. By setting $\epsilon > 0$ the algorithm can be used even when equality occurs with probability zero. The modification can be seen in Algorithm 3.7.

The accepted parameter values and corresponding simulations form a sample from the joint posterior distribution $\pi_\epsilon\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)$, defined by

$$\begin{aligned} \pi_\epsilon\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right) &= \pi\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \rho\left(\tilde{\boldsymbol{Y}}_{1:M}, \boldsymbol{Y}_{1:M}\right) \leq \epsilon\right) && (3.3.1) \\ &\propto \pi(\boldsymbol{\theta})\, \pi\left(\tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{\theta}\right) \mathbb{I}_{\rho\left(\tilde{\boldsymbol{Y}}_{1:M}, \boldsymbol{Y}_{1:M}\right) \leq \epsilon}, && (3.3.2) \end{aligned}$$

where $\mathbb{I}$ denotes the indicator function

$$\mathbb{I}_{\rho\left(\tilde{\boldsymbol{Y}}_{1:M}, \boldsymbol{Y}_{1:M}\right) \leq \epsilon} = \begin{cases} 1 & \text{if } \rho\left(\tilde{\boldsymbol{Y}}_{1:M}, \boldsymbol{Y}_{1:M}\right) \leq \epsilon, \\ 0 & \text{if } \rho\left(\tilde{\boldsymbol{Y}}_{1:M}, \boldsymbol{Y}_{1:M}\right) > \epsilon. \end{cases}$$

The simulated values can be marginalised out to give an approximate marginal posterior distribution of the parameters, $\pi_\epsilon\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right)$, given by

$$\pi_\epsilon\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right) = \int \pi_\epsilon\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right) d\tilde{\boldsymbol{Y}}_{1:M}. \qquad (3.3.3)$$

In the ABC rejection algorithm this means that the simulated vales, $\tilde{\boldsymbol{Y}}_{1:M}$, do not need to be stored in order to obtain an approximate sample from the marginal posterior distribution of the parameters, greatly reducing the necessary memory requirements.

The ABC posterior is an approximation to the true posterior distribution, $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right)$, the accuracy of which depends on the tolerance and distance measure used [56]. Lowering the tolerance gives improved approximations, as the simulated values are closer to the observed dataset. With many commonly used distance measures, for example, distance metrics, such as

**Algorithm 3.7** ABC rejection sampling algorithm targeting $\pi_\epsilon\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)$.

```
Set $n = 1$.
while $n \leq N$ do
    Sample $\boldsymbol{\theta}^*$ from the prior distribution, $\pi(\boldsymbol{\theta})$.
    Simulate values $\tilde{\boldsymbol{Y}}^*_{1:M}$ from the model using parameter $\boldsymbol{\theta}^*$.
    if $\rho\left(\tilde{\boldsymbol{Y}}^*_{1:M}, \boldsymbol{Y}_{1:M}\right) \leq \epsilon$ then
        Set $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}^*$, $\tilde{\boldsymbol{Y}}^{(n)}_{1:M} = \boldsymbol{Y}^*_{1:M}$, and $n = n + 1$.
    end if
end while
```

the Euclidean distance, setting $\epsilon = 0$ indicates equality between the simulated values and the observed dataset. In this setting, the modified rejection algorithm is recovered, and samples are once again from the true posterior distribution. However, for low tolerances it will be too computationally costly to obtain a sample large enough to represent the posterior distribution. At the other extreme, setting $\epsilon = \infty$ means that every proposal is accepted, giving a sample from the prior distribution, $\pi(\boldsymbol{\theta})$. While it is easy to obtain a sample for large values of $\epsilon$, the resulting sample will be a poor approximation of the true posterior distribution. Thus, the tolerance can be thought of as a trade-off between obtaining a good approximation and the time taken to obtain the sample. The choice of tolerance will depend on both the user's computer resources and the desired accuracy. In the rejection sampling setting the choice of tolerance can be determined after running the algorithm. First run a suitably large number of iterations without the accept-reject step, instead recording all of the proposed parameters $\boldsymbol{\theta}^*$ and associated distances $\rho\left(\tilde{\boldsymbol{Y}}^*_{1:M}, \boldsymbol{Y}_{1:M}\right)$, then the sample can be sorted by distance, and the tolerance can be set to accept a specified proportion (say $\frac{k_N}{N}$) of the closest simulated realisations. This can be seen as a $k_N$-nearest neighbour algorithm [57].

### Results

The ABC rejection sampling method is applied to the simulation study dataset, described in Section 3.2, with 20 million simulations. Proposals are drawn from the prior distribution, and we chose the Euclidean distance between the simulated values and the observed data, defined by

$$\rho\left(\boldsymbol{Y}^*_{1:M}, \boldsymbol{Y}_{1:M}\right) = \sqrt{\sum_{i=1}^{M} \left(\boldsymbol{Y}^*_i - \boldsymbol{Y}_i\right)^2}, \tag{3.3.4}$$

as the distance metric. The closest 1000 simulations are accepted, giving a tolerance level of $\epsilon = 128.35$. The code was written in c and R, and has a runtime of approximately 10 hours on a 3 GHz processor. The marginal posterior distributions of the parameters are shown in Figure 3.1.

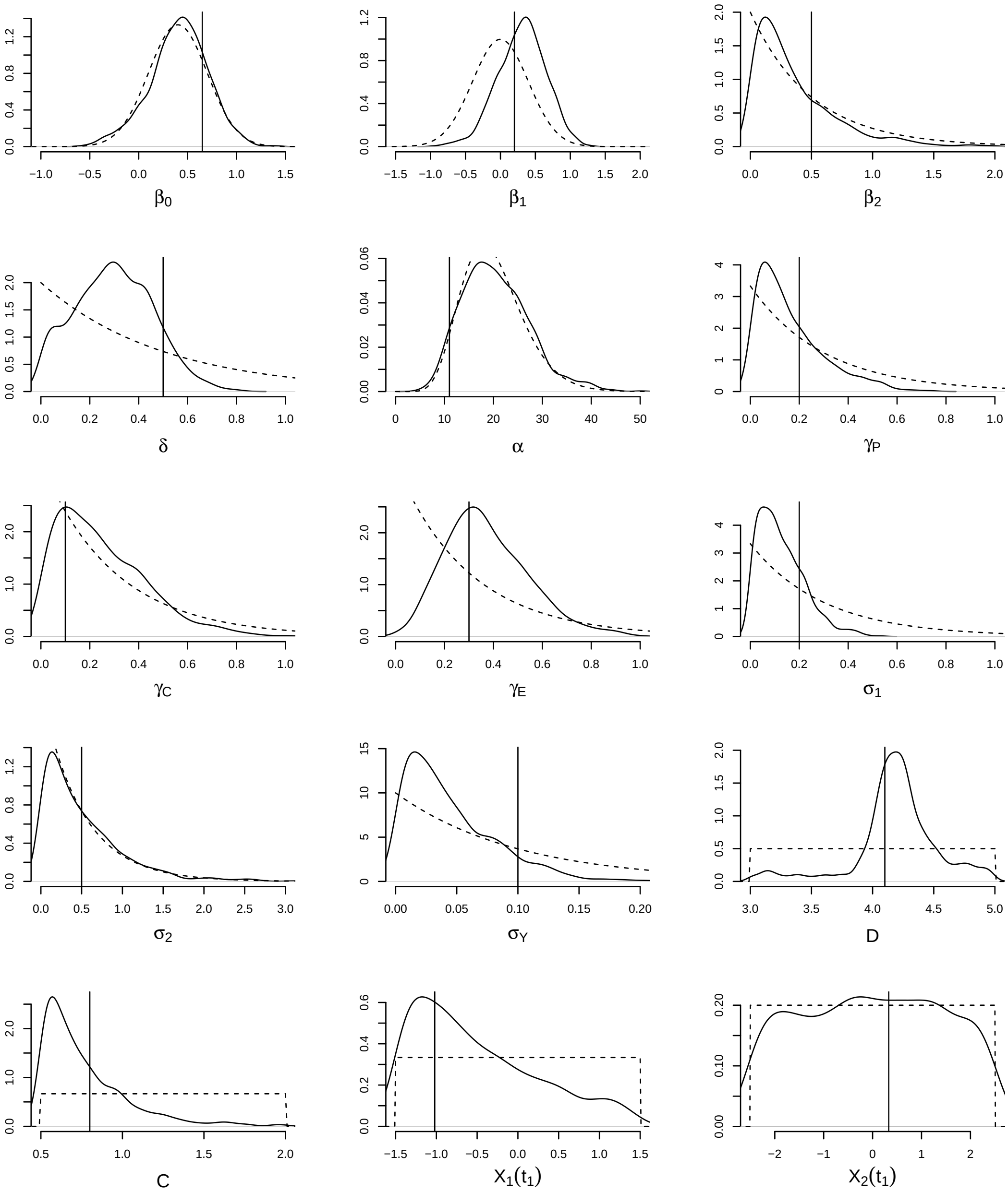

**Figure 3.1:** Marginal posterior distributions of the parameters of CR14-a, obtained using ABC rejection sampling in the simulation study. Vertical lines indicate the values used to generate the data. Dashed lines show the prior distributions. In many cases the approximate posterior distribution strongly resembles the prior distribution, indicating that we are learning little about those parameters. We seem to detect the influence of the astronomical forcing, as there is little posterior mass around zero for the obliquity scaling term, $\gamma_E$. The estimates of the stochastic scaling terms, $\sigma_1$, $\sigma_2$, and $\sigma_Y$, are underestimated due to the ABC approximation.

For many of the model parameters, the marginal posterior distributions do not deviate much from the prior distribution, indicating that we are not learning much about the parameters from the data. With a large tolerance it is not surprising that a wide range of parameter values are accepted. A notable exception to this is the obliquity scaling term, $\gamma_E$, where the true value lies near the mode of the posterior distribution. Being synchronised on the same forcing function as the observed data should lead to a good agreement between the simulated values and the observed dataset, so this is not surprising. The true values of the displacement and scaling terms in the observation model, $D$ and $C$, also lie in regions of relatively high posterior probability density. These terms have a large impact on the distance between the simulated values and the observed dataset, and are mostly insensitive to stochastic perturbations. The estimates of the stochastic scaling terms, $\sigma_1$, $\sigma_2$, and $\sigma_Y$, are underestimated due to the ABC approximation. At sufficiently large tolerance values, near-deterministic trajectories will always be considered close to the observed data, but a highly volatile trajectory might differ greatly. In this case, low values of $\sigma_1$, $\sigma_2$, and $\sigma_Y$, will lead to a greater acceptance probability than large values, giving a greater posterior probability density. The opposite is true as the tolerance is lowered. Stochastic perturbations will be necessary to simulate values sufficiently close to observations, and so near-deterministic trajectories will usually be rejected. The true value of the initial condition of the observable state is near the mode of the posterior distribution, but the initial condition of the unobservable state is similar to the prior distribution. Given that $X_2$ is unobserved, and is quickly attracted towards the limit cycle, this is not surprising. The accepted parameter values are mostly uncorrelated, with the exception of $\beta_0$ and $\delta$, which are strongly correlated with a correlation coefficient of 0.76, and $\beta_0$ and $D$, which have a correlation coefficient of 0.5. Incorporating these correlations into the proposal distribution would give a more efficient algorithm.

### 3.3.1 ABC-MCMC

Since the ABC rejection sampling algorithm proposes parameters from the prior distribution, $\pi(\boldsymbol{\theta})$, it can suffer from poor acceptance rates if many proposals lie in regions of low posterior probability density. An MCMC implementation of ABC targeting $\pi_\epsilon\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right)$, called ABC-MCMC, was proposed as a more efficient sampling method [58]. ABC-MCMC replaces the likelihood term in the Metropolis Hastings acceptance ratio with a comparison between simulated values and the observed dataset, as in ABC rejection sampling. The pseudocode is presented in Algorithm 3.8.

The chain is initialised with a sample obtained using the ABC rejection sampling approach. This ensures that the chain is initialised with a value drawn from the stationary distribution, avoiding the need for a burn-in period. The tolerance needs be chosen at initialisation, unlike

**Algorithm 3.8** ABC-MCMC algorithm targeting $\pi_\epsilon\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)$.

```
Sample $\boldsymbol{\theta}^{(1)}$ and $\tilde{\boldsymbol{Y}}_{1:M}^{(1)}$ from $\pi_\epsilon\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)$ using the ABC rejection sampling algorithm.
for $n = 2, ..., N$ do
    Propose move to $\boldsymbol{\theta}^*$ according to proposal density $q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(n-1)}\right)$.
    Simulate values $\tilde{\boldsymbol{Y}}_{1:M}^{*}$ from the model using parameter $\boldsymbol{\theta}^*$.
    With probability
        $\lambda\left(\boldsymbol{\theta}^*, \tilde{\boldsymbol{Y}}_{1:M}^{*} \mid \boldsymbol{\theta}^{(n-1)}, \tilde{\boldsymbol{Y}}_{1:M}^{(n-1)}\right) = \min\left(1, \frac{\pi\left(\boldsymbol{\theta}^*\right) q\left(\boldsymbol{\theta}^{(n-1)} \mid \boldsymbol{\theta}^*\right) \mathbb{I}_{\rho\left(\tilde{\boldsymbol{Y}}_{1:M}^{*}, \boldsymbol{Y}_{1:M}\right) \leq \epsilon}}{\pi\left(\boldsymbol{\theta}^{(n-1)}\right) q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(n-1)}\right)}\right)$
    set $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}^*$ and $\tilde{\boldsymbol{Y}}_{1:M}^{(n)} = \tilde{\boldsymbol{Y}}_{1:M}^{*}$. Otherwise set $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}^{(n-1)}$ and $\tilde{\boldsymbol{Y}}_{1:M}^{(n)} = \tilde{\boldsymbol{Y}}_{1:M}^{(n-1)}$.
end for
```

with the ABC rejection sampling approach. This can be problematic, as if the tolerance is too small the chain will suffer from slow mixing, taking a long time to explore the parameter space.

We can check that the algorithm targets the posterior distribution, $\pi_\epsilon\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)$, by showing that the detailed balance equation is satisfied. Assuming that $\rho\left(\tilde{\boldsymbol{Y}}_{1:M}, \boldsymbol{Y}_{1:M}\right) \leq \epsilon$, and $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$,

$$
\begin{aligned}
&\frac{\pi_\epsilon\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right) \mathcal{P}\left(\boldsymbol{\theta}^*, \tilde{\boldsymbol{Y}}_{1:M}^{*} \mid \boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M}\right)}{\pi_\epsilon\left(\boldsymbol{\theta}^*, \tilde{\boldsymbol{Y}}_{1:M}^{*} \mid \boldsymbol{Y}_{1:M}\right) \mathcal{P}\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{\theta}^*, \tilde{\boldsymbol{Y}}_{1:M}^{*}\right)} \\
= & \left(\frac{\pi_\epsilon\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)}{\pi_\epsilon\left(\boldsymbol{\theta}^*, \tilde{\boldsymbol{Y}}_{1:M}^{*} \mid \boldsymbol{Y}_{1:M}\right)}\right)\left(\frac{q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}\right) \pi\left(\tilde{\boldsymbol{Y}}_{1:M}^{*} \mid \boldsymbol{\theta}^*\right) \lambda\left(\boldsymbol{\theta}^*, \tilde{\boldsymbol{Y}}_{1:M}^{*} \mid \boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M}\right)}{q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*\right) \pi\left(\tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{\theta}\right) \lambda\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{\theta}^*, \tilde{\boldsymbol{Y}}_{1:M}^{*}\right)}\right) \\
= & \frac{\pi\left(\boldsymbol{\theta}\right) q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}\right) \lambda\left(\boldsymbol{\theta}^*, \tilde{\boldsymbol{Y}}_{1:M}^{*} \mid \boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M}\right)}{\pi\left(\boldsymbol{\theta}^*\right) q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*\right) \mathbb{I}_{\rho\left(\tilde{\boldsymbol{Y}}_{1:M}^{*}, \boldsymbol{Y}_{1:M}\right) \leq \epsilon} \lambda\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{\theta}^*, \tilde{\boldsymbol{Y}}_{1:M}^{*}\right)} \\
= & \frac{\pi\left(\boldsymbol{\theta}\right) q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}^*\right) q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*\right) \mathbb{I}_{\rho\left(\tilde{\boldsymbol{Y}}_{1:M}^{*}, \boldsymbol{Y}_{1:M}\right) \leq \epsilon}}{\pi\left(\boldsymbol{\theta}^*\right) q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*\right) \pi\left(\boldsymbol{\theta}\right) q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}\right) \mathbb{I}_{\rho\left(\tilde{\boldsymbol{Y}}_{1:M}^{*}, \boldsymbol{Y}_{1:M}\right) \leq \epsilon}} \\
= & \ 1
\end{aligned}
$$

| Parameter | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\delta$ | $\alpha$ | $\gamma_P$ | $\gamma_C$ | $\gamma_E$ |
|---|---|---|---|---|---|---|---|---|
| $\tau$ | 0.03 | 0.03 | 0.03 | 0.02 | 0.7 | 0.02 | 0.02 | 0.02 |

| Parameter | $\sigma_1$ | $\sigma_2$ | $\sigma_Y$ | $D$ | $C$ | $X_1(t_1)$ | $X_2(t_1)$ |
|---|---|---|---|---|---|---|---|
| $\tau$ | 0.01 | 0.05 | 0.004 | 0.03 | 0.03 | 0.08 | 0.15 |

**Table 3.2:** Standard deviation ($\tau$) associated with each of the parameters in the Gaussian random walk proposal for ABC-MCMC. These values were selected based on several trial runs.

#### Results

For comparison with the ABC rejection sampling algorithm we use the distance metric given by Equation 3.3.4, and set the tolerance to $\epsilon = 128.35$. The ABC-MCMC algorithm is performed with a chain length of 3 million. Since the ABC rejection algorithm has a 1 in 50 000 acceptance rate at this tolerance level, obtaining the initial sample is expected to add another 50 000 simulations from the model. For the proposal distribution, $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{n-1})$, we use a multivariate Gaussian distribution with mean $\boldsymbol{\theta}_{n-1}$ and a diagonal covariance matrix. The variance terms, denoted here as $\tau_i^2$ for component $i$ of the parameter vector, are provided in Table 3.2. These were chosen based on several trial runs. The code was written in c and R, and has a runtime of approximately 2 hours on a 3 GHz processor

The resulting chain has an acceptance rate of approximately 0.05. We thin the chain so that we have a sample size of 1000 by taking every 3000th value. A smaller chain length may have been acceptable, but even with 3 million iterations we have reduced the computation time compared to the ABC rejection sampling scheme dramatically. The marginal posterior distributions of the parameters are shown in Figure 3.2. There are strong similarities with the posterior distributions obtained via ABC rejection sampling, as should be the case given that both algorithms target the same distribution. The correlation between $\beta_0$ and $\delta$, and $\beta_0$ and $D$ are slightly lower than was the case in the ABC rejection sampling case, with correlation coefficients of 0.68 and 0.4 respectively.

### 3.3.2 ABC-SMC

ABC can be naturally generalised to an SMC approach, referred to as ABC-SMC. When $\epsilon = \infty$ a sample from the prior distribution is obtained, and when $\epsilon = 0$ the true posterior distribution is sampled from. Everything in between is a trade-off between computability and the accuracy of the approximation to the posterior distribution. In SMC, we select a series of tolerances $\epsilon_1 > \epsilon_2 > ... > \epsilon_S$, so that the intermediary distributions are the approximate posterior distributions $\pi_{\epsilon_1}\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right), ..., \pi_{\epsilon_S}\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)$. Typically, $\epsilon_1 = \infty$ and $\epsilon_S$ is as small as is computationally feasible. The transition density of the Markov kernel takes the form

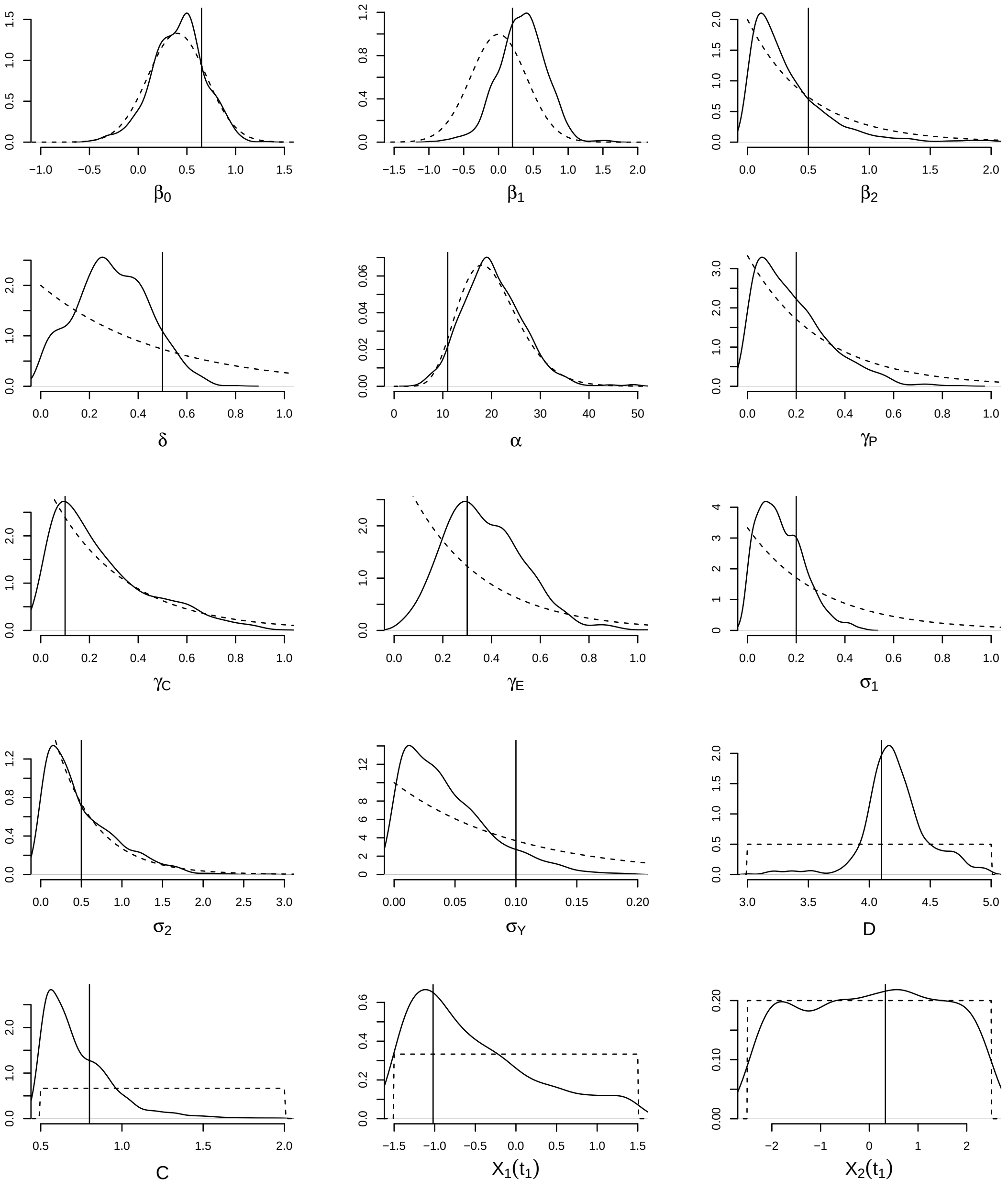

**Figure 3.2:** Marginal posterior distributions of the parameters of CR14-a, obtained using ABC-MCMC in the simulation study. Vertical lines indicate the values used to generate the data. Dashed lines show the prior distributions. These distributions are very similar to those shown in Figure 3.1, as both algorithms target the same approximate posterior distribution.

$K_s\left(\boldsymbol{\theta}_s, \tilde{\boldsymbol{Y}}_{1:M,s} \mid \boldsymbol{\theta}_{s-1}, \tilde{\boldsymbol{Y}}_{1:M,s-1}\right) = K_s\left(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_{s-1}\right)\pi\left(\tilde{\boldsymbol{Y}}_{1:M,s} \mid \boldsymbol{\theta}_s\right)$. In other words, the parameter values are being perturbed by a Markov kernel, and a simulation is performed with the perturbed parameters. This choice is necessary to remove the likelihood term from the importance weight calculation.

### ABC-Partial Rejection Control

The first ABC-SMC algorithm used backward Markov kernels $L_{s-1}\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*\right) = K_s\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}\right)$ [59], which is an inefficient choice in most circumstances. In the ABC framework this also led to biased estimates [51]. The original paper was updated to use instead an approximation to the optimal backward kernel of the form given in Equation 3.1.19, which became a popular choice in the literature [51, 60, 61]. The importance weight is estimated using a Monte Carlo approximation to Equation 3.1.20, as follows,

$$\begin{aligned} w_s^{(n)} &= W_{s-1}^{(n)} \frac{\pi\left(\boldsymbol{\theta}_s^{(n)}\right)\pi\left(\tilde{\boldsymbol{Y}}_{1:M,s}^{(n)} \mid \boldsymbol{\theta}^{(n)}\right)\mathbb{I}_{\rho\left(\tilde{\boldsymbol{Y}}_{1:M,s}^{(n)}, \boldsymbol{Y}_{1:M}\right) \le \epsilon_s}}{\sum_{i=1}^N W_{s-1}^{(i)} K_s\left(\boldsymbol{\theta}_s^{(n)} \mid \boldsymbol{\theta}_{s-1}^{(i)}\right)\pi\left(\tilde{\boldsymbol{Y}}_{1:M,s}^{(n)} \mid \boldsymbol{\theta}^{(n)}\right)} \\ &= W_{s-1}^{(n)} \frac{\pi\left(\boldsymbol{\theta}_s^{(n)}\right)\mathbb{I}_{\rho\left(\tilde{\boldsymbol{Y}}_{1:M,s}^{(n)}, \boldsymbol{Y}_{1:M}\right) \le \epsilon_s}}{\sum_{i=1}^N W_{s-1}^{(i)} K_s\left(\boldsymbol{\theta}_s^{(n)} \mid \boldsymbol{\theta}_{s-1}^{(i)}\right)}. \end{aligned} \tag{3.3.5}$$

Note that since we simulate from the model, the likelihood does not appear in the weight calculation. This Monte Carlo approximation is computationally costly for large $N$, as the sum needs to be evaluated for every particle.

As the tolerance is lowered, the number of particles with weight zero will increase. This will potentially lead to poor approximations at low tolerances. Partial rejection control (PRC) is a way of resampling these particles [62]. In PRC, when a particle is proposed with an importance weight below some threshold it is probabilistically either rejected or given a larger weight. Any proposed particles with a weight above the threshold are always accepted. In ABC, the natural choice is to accept any particle with non-zero weight and reject particles with weight zero [51, 60]. To ensure that the sample size remains at $N$, particles are no longer propagated systematically. Instead, particles are sampled with replacement according to their weights, propagated, and then either accepted or rejected according to the PRC step until we have $N$ acceptances. With PRC implementations there is no longer a resampling step if the ESS falls below its threshold, as the PRC approach effectively resamples the particles in every population. The ABC-PRC algorithm is reproduced in Algorithm 3.9. Note that the importance weight in the algorithm takes a different form than Equation 3.3.5. Firstly, due to the resampling in PRC, the importance weights, $W_{s-1}^{(1:N)}$, are all set to $\frac{1}{N}$. Secondly, since only simulations within

**Algorithm 3.9** ABC-PRC sampling algorithm targeting $\pi_\epsilon\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)$.

```
for $n = 1, ..., N$ do
    Sample $\boldsymbol{\theta}_1^{(n)}$ from the prior distribution, $\pi(\boldsymbol{\theta})$.
    Simulate values $\tilde{\boldsymbol{Y}}_{1:M,1}^{(n)}$ from the model using parameter $\boldsymbol{\theta}_1^{(n)}$.
    Set the importance weight
        $W_1^{(n)} = \frac{1}{N}.$
end for
for $s = 2, ..., S$ do
    while $n \leq N$ do
        Sample $\boldsymbol{\theta}_s^{**}$ from the previous population, $\boldsymbol{\theta}_{s-1}^{(1:N)}$, according to weights $W_{s-1}^{(1:N)}$.
        Sample $\boldsymbol{\theta}_s^{*}$ from the transition density $K_s\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{s-1}^{**}\right)$.
        if $\pi(\boldsymbol{\theta}_s^{*}) > 0$ then
            Simulate values $\tilde{\boldsymbol{Y}}_{1:M,s}^{*}$ from the model using parameter $\boldsymbol{\theta}^{*}$.
            if $\rho\left(\tilde{\boldsymbol{Y}}_{1:M,s}^{*}, \tilde{\boldsymbol{Y}}_{1:M}\right) \leq \epsilon_s$ then
                Set $\boldsymbol{\theta}_s^{(n)} = \boldsymbol{\theta}^{*}$ and $\tilde{\boldsymbol{Y}}_{1:M,s}^{(n)} = \tilde{\boldsymbol{Y}}_{1:M,s}^{*}$.
                Set the importance weight
                    $w_s^{(n)} = \frac{\pi\left(\boldsymbol{\theta}_s^{(n)}\right)}{\sum_{i=1}^{N} W_{s-1}^{(i)} K_s\left(\boldsymbol{\theta}_s^{(n)} \mid \boldsymbol{\theta}_{s-1}^{(i)}\right)}.$
                Set $n = n + 1$.
            end if
        end if
    end while
    Normalise the weights. For $n = 1, ...N$
        $W_s^{(n)} = \frac{w_s^{(n)}}{\sum_{i=1}^{N} w_s^{(i)}}.$
end for
```

the tolerance are accepted, the indicator function is removed from the numerator.

The Markov kernels, $K_s\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{s-1}\right)$, can be chosen using the population of particles at iteration $s-1$. For example, in [51] $K_s\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{s-1}\right)$ is a multivariate Gaussian distribution with mean $\boldsymbol{\theta}_{s-1}$ and a diagonal covariance matrix, where the diagonal terms are twice the empirical variance of the corresponding components of the parameter vector. This provides a proposal distribution that is automatically adapted for each iteration of the algorithm.

A disadvantage of this approach is that the sequence of tolerance values can not be chosen during post-processing. Selecting a sequence of tolerances at initialisation can be challenging, as reducing the tolerance too slowly will result in a computationally expensive algorithm, but reducing the tolerance too quickly will lead to a poor approximation of the posterior distribution. Fortunately, the tolerance only needs to be selected one iteration ahead, mitigating the problem. The tolerance scheme should be chosen to limit the amount of degeneracy, measured by the ESS, in each sample. At some point the tolerance will be lowered to a point that sampling becomes too computationally costly, providing a convenient method for selecting the final tolerance value, unlike ABC-MCMC.

### Results

The ABC-PRC algorithm is run with 1000 particles, with the tolerance scheme shown in Table 3.3. This tolerance scheme was selected based on several trial runs. The perturbation kernels were taken to be multivariate Gaussian random walks, with twice the empirical variance of the current sample.

The parameter marginal posterior distributions are shown in Figure 3.3, and the number of simulations required in each iteration are shown in Table 3.3, showing that in later iterations it takes a large computational cost to reduce the tolerance even a small amount. The code was written in c and R, and has a runtime of approximately 3 hours on a 3 GHz processor. This tolerance scheme takes approximately 2 million simulations to sample from the approximate posterior distribution for $\epsilon = 125$. This is much more efficient than the ABC rejection scheme, which required 20 million simulations to sample from the approximate posterior distribution with $\epsilon = 128.35$. The required number of simulations is comparable to the ABC-MCMC approach. The weight calculations in ABC-PRC leads to additional computational time in comparison to ABC-MCMC, but simulations from the model remain the dominant computational expense. The ESS falls to approximately 40 early on, and recovers to near 200 in the final population. This means that the early sample had an effective size of 40, and the final iteration had an effective size of 200. It is possible that this is too low to be confident that we have a good approximation, but the posterior distributions are consistent with the previous methods. It might

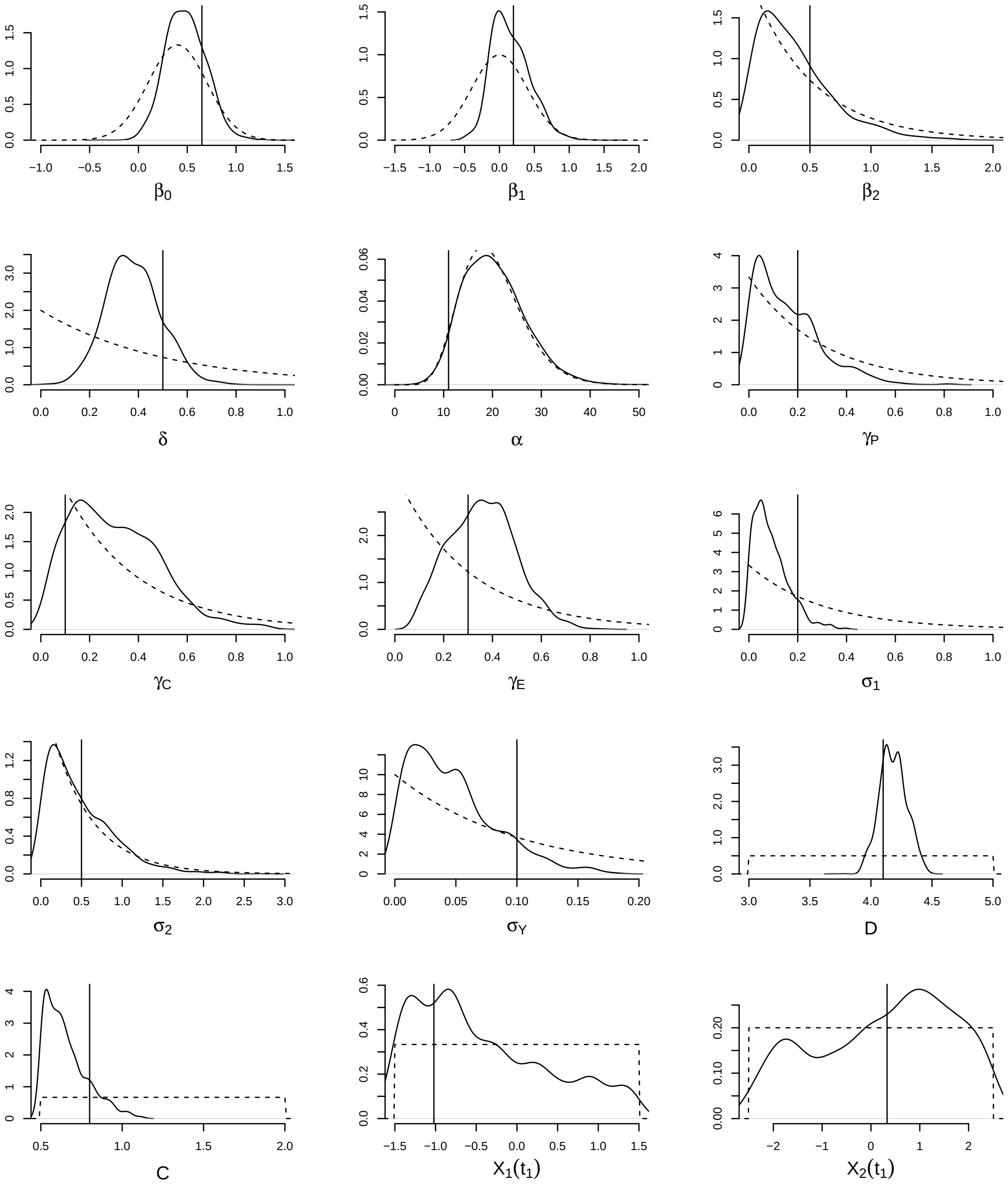

**Figure 3.3:** Marginal posterior distributions of the parameters of CR14-a, obtained using ABC-PRC in the simulation study. Vertical lines indicate the values used to generate the data. Dashed lines show the prior distributions. In comparison to the marginal posterior distributions obtained using ABC rejection (shown in Figure 3.1), and ABC-MCMC (shown in Figure 3.2), many of the marginal posterior distributions have closed in on the true values. This is due to reaching a lower tolerance, meaning that simulations must be closer to the observed data in order for the proposed parameter values to be accepted.

| Iteration | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ | $\infty$ | 320 | 220 | 180 | 160 | 145 | 130 | 125 | 120 | 115 |
| Simulations $\left(\times 10^3\right)$ | 1.0 | 3.5 | 7.0 | 11.9 | 12.3 | 43.6 | 641 | 1224 | 1682 | 1451 |

**Table 3.3:** Tolerance scheme used in ABC-PRC in the simulation study, with the number of simulations required at each tolerance level. This tolerance scheme was selected based on several trial runs. As the tolerance is lowered, the number of required simulations rises sharply.

be sensible to increase the number of iterations, bringing down the tolerance more gradually to ensure that the ESS remains higher. This would provide a better approximation, but would increase the computational cost of the algorithm. The posterior variance is smaller than the ABC rejection and ABC-MCMC posterior distributions. This can be caused by both the lower tolerance in the final population, and the low ESS at intermediate stages putting small mass on some regions of the parameter space. When the degeneracy is severe, most of the posterior mass is in a small region of parameter space, but the specific region changes between runs. On the other hand, if the posterior variance is lower due to the smaller tolerance, repeating the experiment should give consistent posterior density estimates. Repeating the experiment gives posterior distributions consistent with Figure 3.3, unless the ESS falls to very low ($< 10$) values, suggesting that the lower posterior variance is due to the lower tolerance, rather than degeneracy. The correlation between $\beta_0$ and $\delta$ is larger than our previous methods with a coefficient of 0.96, but comparable on the 7th and 8th iterations, which are closer to the tolerance used in the ABC rejection, and ABC-MCMC schemes.

### Adaptive ABC-SMC

Alternatively, an ABC-SMC approach has been proposed that uses MCMC kernels, as in Equation 3.1.21 [63]. The MCMC proposals follow the ABC-MCMC approach, discussed in Section 3.3.1, to target the approximate posterior distribution in each iteration. The importance weights become

$$\begin{aligned} w_s^{(n)} &= W_{s-1}^{(n)} \frac{\pi\left(\boldsymbol{\theta}_{s-1}^{(n)}\right)\pi\left(\tilde{\boldsymbol{Y}}_{1:M,s-1}^{(n)} \mid \boldsymbol{\theta}_{s-1}^{(n)}\right)\mathbb{I}_{\rho\left(\tilde{\boldsymbol{Y}}_{1:M,s-1}^{(n)},\tilde{\boldsymbol{Y}}_{1:M}\right)\leq\epsilon_s}}{\pi\left(\boldsymbol{\theta}_{s-1}^{(n)}\right)\pi\left(\tilde{\boldsymbol{Y}}_{1:M,s-1}^{(n)} \mid \boldsymbol{\theta}_{s-1}^{(n)}\right)\mathbb{I}_{\rho\left(\tilde{\boldsymbol{Y}}_{1:M,s-1}^{(n)},\tilde{\boldsymbol{Y}}_{1:M}\right)\leq\epsilon_{s-1}}} \\ &= W_{s-1}^{(n)} \frac{\mathbb{I}_{\rho\left(\tilde{\boldsymbol{Y}}_{1:M,s-1}^{(n)},\tilde{\boldsymbol{Y}}_{1:M}\right)\leq\epsilon_s}}{\mathbb{I}_{\rho\left(\tilde{\boldsymbol{Y}}_{1:M,s-1}^{(n)},\tilde{\boldsymbol{Y}}_{1:M}\right)\leq\epsilon_{s-1}}}, \end{aligned} \tag{3.3.6}$$

which are obtained by substituting the intermediary approximate posterior distributions into Equation 3.1.22. The prior distributions and likelihoods cancel, leaving the ratio of the indicator functions multiplied by the previous weights. Since the importance weights do not depend on the perturbed parameters, the ordering of the propagation and weighting steps can be swapped. The tolerance scheme can then be adaptively chosen based on a controlled decline of the ESS [63]. Hence, this method can be considered to be an adaptive SMC approach (ASMC). The ESS of the sample at iteration $s$ is dependant on the weights of the sample, and the tolerance value used. A scaling parameter, $a$, dictates the rate at which the ESS, and hence the tolerance, declines. Specifically, for a fixed choice of $a$ we choose $\epsilon_s$ so that

$$\text{ESS}\left(W_s, \epsilon_s\right) = a\text{ESS}\left(W_{s-1}, \epsilon_{s-1}\right) \tag{3.3.7}$$

in every iteration. Large values of $a$ mean that the tolerance evolves slowly, giving a good approximation, but with a high computational cost, as more iterations are required to reach the target. The value of $a$ should be chosen as large as computational resources allow for the best approximation. In the version of the algorithm given in this section the ESS is simply the number of particles with non-zero weight (called ‘alive’ particles). The choice of $\epsilon_s$ is then selected to give the proportion $a$ closest alive particles weight 1, and the rest 0.

The acceptance rate of the MCMC kernels is monitored in each iteration, denoted as $R_s$. It is suggested in [63] to terminate the algorithm when the acceptance rate falls below some set threshold $\hat{R}$. This provides an automated method to chose the final tolerance value in addition to the intermediary tolerance values. The value of $\hat{R}$ is chosen based on the available computer resources. Smaller values of $\hat{R}$ will lead to improved approximations as a lower tolerance will be reached.

It is important to choose an appropriate Markov chain length. A small chain length will mean few simulations are required, but the diversity of the particles (the number of unique particles) will remain small if the chain is slow mixing. In other words, particles may be resampled numerous times in the resampling step and are then unlikely to move in short MCMC chains, giving a poor approximation to the target distribution. A larger chain will increase the chance of the particles having good diversity, but will increase computational expense. Since the MCMC acceptance rate depends on the tolerance, it makes sense to begin with a small chain length and increase it in later populations. This can be selected adaptively by monitoring the MCMC acceptance rate. If an MCMC chain of length $L$ has acceptance rate $R_s$, then the probability of not moving along the entire chain $\mathcal{Q}$ is

$$\mathcal{Q} = (1 - R_s)^L$$

An appropriate chain length can then be estimated adaptively by using the acceptance rate of the previous population, $R_{s-1}$, and by selecting a value for $\mathcal{Q}$, such that

$$L = \frac{\log(\mathcal{Q})}{\log(1 - R_{s-1})}$$

Since this is proportional to $\log(\mathcal{Q})$, it makes sense to take $\mathcal{Q}$ small. For example, by setting $\mathcal{Q} = 0.001$ instead of $\mathcal{Q} = 0.01$, the computational expense is only increased by a factor of 1.5, yet the particle diversity is increased dramatically. This approach has been independently developed in a similar algorithm [64], where it was suggested to set $\mathcal{Q} = \frac{1}{N}$. This would mean that, on average, only a single particles will not move along the length of the chain. In practice, this will not be the case, as we know the acceptance rate will be lower than the previous iteration. We advocate using a smaller value of $\mathcal{Q}$, as there is a good trade-off between particle diversity and added computational expense. The ABC-ASMC algorithm, with the addition of an adaptive Markov chain length, is shown in Algorithm 3.10.

As well as checking that the ESS and particle diversity remain high, it is also important to choose the proposal distribution of the MCMC kernel appropriately. A standard approach is to use a Gaussian random walk proposal with variance proportional to the empirical variance of the particles. If this choice is made, then it is possible to end up with a very small variance for the proposal distribution, and the proposed particles are considered distinct even though they are arbitrarily close. This is undesirable, and can be avoided by ensuring that the variance of the proposal distribution remains relatively large.

**Algorithm 3.10** ABC-ASMC sampling algorithm targeting $\pi_\epsilon\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)$.

```
for $n = 1, ...N$ do
  Sample $\boldsymbol{\theta}_1^{(n)}$ from the prior distribution, $\pi\left(\boldsymbol{\theta}\right)$.
  Simulate values $\tilde{\boldsymbol{Y}}_{1:M,1}^{(n)}$ from the model using parameter $\boldsymbol{\theta}_1^{(n)}$.
  Set the importance weight
      $$W_1^{(n)} = \frac{1}{N}.$$
end for
while $R_s > \hat{R}$ do
  Set $s = s + 1$.
  Determine $\epsilon_s$ by solving ESS $\left(W_s^{(1:N)}, \epsilon_s\right) = a$ESS $\left(W_{s-1}^{(1:N)}, \epsilon_{s-1}\right)$, where
      $$w_s^{(n)} = W_{s-1}^{(n)} \frac{\mathbb{I}_{\rho\left(\tilde{\boldsymbol{Y}}_{1:M,s-1}^{(n)}, \tilde{\boldsymbol{Y}}_{1:M}\right) \leq \epsilon_s}}{\mathbb{I}_{\rho\left(\tilde{\boldsymbol{Y}}_{1:M,s-1}^{(n)}, \tilde{\boldsymbol{Y}}_{1:M}\right) \leq \epsilon_{s-1}}}.$$
  Normalise the weights. For $n = 1, ..., N$
      $$W_s^{(n)} = \frac{w_s^{(n)}}{\sum_{i=1}^{N} w_s^{(i)}}.$$
  if ESS<$\frac{N}{2}$ then
    Resample $\boldsymbol{\theta}^{(1:N)}$ and $\tilde{\boldsymbol{Y}}_{1:M,s}^{(1:N)}$ according to weights $W_s^{(1:N)}$.
    Set the importance weights. For $n = 1, ..., N$
        $$W_s^{(n)} = \frac{1}{N}.$$
  end if
  Update Markov chain length $L = \frac{Q}{\log(1-R_{s-1})}$.
  for $n = 1, ..., N$ do
    Sample $\boldsymbol{\theta}_s^{(n)}$ and $\tilde{\boldsymbol{Y}}_{1:M,s}^{(n)}$ from an ABC-MCMC chain targeting $\pi_{\epsilon_s}\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)$.
  end for
  Evaluate the acceptance rate, $R_s$.
end while
```

| Iteration | 1 | 11 | 21 | 31 | 41 | 51 | 61 |
|---|---|---|---|---|---|---|---|
| $\epsilon$ | $\infty$ | 221.5 | 155.4 | 138.5 | 129.8 | 119.6 | 105.7 |
| Chain Length | 1 | 355 | 690 | 1154 | 2852 | 4744 | 8191 |
| Simulations $\left(\times 10^3\right)$ | 1.0 | 54.2 | 71.0 | 251 | 434 | 1242 | 1460 |

**Table 3.4:** Sample of the tolerance scheme (shown for every 10th iteration) obtained in ABC-ASMC in the simulation study, with the number of simulations required at each tolerance level, and the associated Markov chain length. As the tolerance is lowered, the number of simulations required rises sharply.

### Results

The ABC-ASMC algorithm is run with 1000 particles and terminates when when the Markov chain acceptance rate falls below 0.001. We select $a = 0.8$, which seems to be a good trade off between computational expense and accuracy. This means that a resampling step is required every four iterations. The MCMC proposal distribution was taken to be a multivariate Gaussian independence sampler, where the mean and covariance are calculated empirically from the current sample. We take $\mathcal{Q} = 10^{-4}$ to ensure a good particle diversity. This gives 61 iterations, where the tolerance and Markov chain length for every 10th iteration are given in Table 3.4.

The parameter marginal posterior distributions are shown in Figure 3.4. The marginal posterior distributions are in good agreement with the previous approaches, but have a lower posterior variance due to the lower tolerance. A total of 34 million simulations was required to reach the final tolerance, and 9.3 million simulations were required to reach a tolerance of 128.1. The code was written in c and R, and has a runtime of approximately 18 hours on a 3 GHz processor. This remains much more efficient than the ABC rejection sampling scheme, but is more computationally expensive than the ABC-MCMC and ABC-PRC approaches. This is also true when considering the cheaper weight calculations in comparison to ABC-PRC. The algorithm can be sped up by reducing $a$, which was chosen to have a rather large value here. Unlike the ABC-PRC scheme, a good ESS was maintained throughout by design. Particle diversity remained high throughout, with only a handful of MCMC chains remaining stationary over the entire length $L$. As with ABC-PRC, the correlation between $\beta_0$ and $\delta$ is much stronger at lower tolerances, with a coefficient of 0.95, but comparable to ABC rejection sampling and ABC-MCMC at similar tolerance levels.

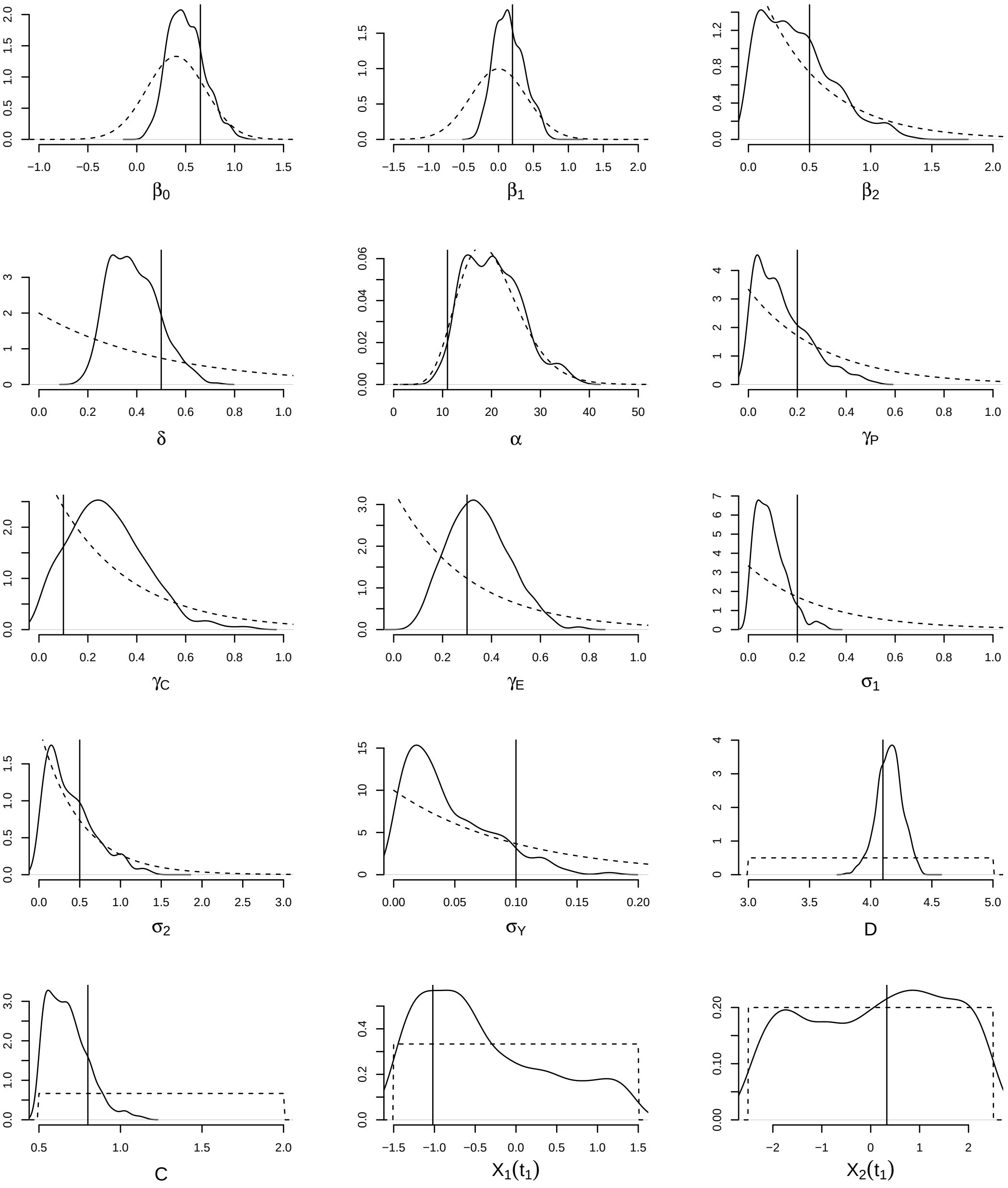

**Figure 3.4:** Marginal posterior distributions of the parameters of CR14-a, obtained using ABC-ASMC in the simulation study. Vertical lines indicate the values used to generate the data. Dashed lines show the prior distributions. In comparison to the marginal posterior distributions obtained using ABC-PRC (shown in Figure 3.3), the distributions have further contracted around the the true values due to reaching a lower tolerance.

### 3.3.3 ABC Extensions

**Summary Statistics**

As the dimension of the data increases, the probability of simulating values close to the observed dataset decreases dramatically. In ABC this is problematic as the computational expense required to generate a sample at a desired tolerance level becomes prohibitive. One solution is to reduce the dimension of the problem by summarising datasets using a collection of summary statistics, $\boldsymbol{s}\left(\boldsymbol{Y}_{1:M}\right) = \left\{s_1\left(\boldsymbol{Y}_{1:M}\right), ..., s_L\left(\boldsymbol{Y}_{1:M}\right)\right\}$. ABC can then be performed by considering the distance between the collection of summary statistics for the simulated values and the observed dataset. This is done by replacing the distance measure, $\rho\left(\tilde{\boldsymbol{Y}}_{1:M}, \boldsymbol{Y}_{1:M}\right)$, in the above algorithms with $\rho_s\left(\boldsymbol{s}\left(\tilde{\boldsymbol{Y}}_{1:M}\right), \boldsymbol{s}\left(\boldsymbol{Y}_{1:M}\right)\right)$. The joint posterior density approximation becomes

$$\begin{aligned} \pi_{s,\epsilon}\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{Y}_{1:M}\right) &= \pi\left(\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{1:M} \mid \rho_s\left(\boldsymbol{s}\left(\tilde{\boldsymbol{Y}}_{1:M}\right), \boldsymbol{s}\left(\boldsymbol{Y}_{1:M}\right)\right) \leq \epsilon\right) & (3.3.8) \\ &\propto \pi\left(\boldsymbol{\theta}\right) \pi\left(\tilde{\boldsymbol{Y}}_{1:M} \mid \boldsymbol{\theta}\right) \mathbb{I}_{\rho_s\left(\boldsymbol{s}\left(\tilde{\boldsymbol{Y}}_{1:M}\right), \boldsymbol{s}\left(\boldsymbol{Y}_{1:M}\right)\right) \leq \epsilon}. & (3.3.9) \end{aligned}$$

The new distance measure must be chosen carefully. For example, if the Euclidean distance between summary statistics is used then it is possible that a single summary statistic will dominate the measure. Instead a weighted measure, such as the Mahalanobis distance [65, 66], should be used.

Ideally the set of summary statistics should be chosen so that $\boldsymbol{s}\left(\boldsymbol{Y}_{1:M}\right)$ is a sufficient statistic, satisfying $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{s}\left(\boldsymbol{Y}_{1:M}\right)\right) = \pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right)$. In other words, the sufficient statistics contains all of the information in $\boldsymbol{Y}_{1:M}$. In this case we recover the classical ABC target, and an exact algorithm can be recovered by setting $\epsilon = 0$. However, when the likelihood is intractable there is no method to determine whether a statistic is sufficient. Instead, the summary statistics should be chosen to include the important aspects of the data, hopefully so that $\pi\left(\boldsymbol{\theta} \mid \boldsymbol{s}\left(\boldsymbol{Y}_{1:M}\right)\right) \approx \pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:M}\right)$.

How to choose such summary statistics is a challenging problem. One possible approach is to test a large number of summary statistics, which are considered for inclusion sequentially [67], but such schemes are often sensitive to the order in which the summary statistics are considered, and depend on having good summary statistics in the initial set [44]. Even then the best choice of summary statistics may be dataset dependent, and so this process would need to be repeated every time a new dataset was introduced. Different methods of choosing between a subset of summary statistics are discussed in [68].

### ABC with model error

In Chapter 2 we assumed that the data we observe are subject to measurement error. In the ABC methods described above, we simulate from a phenomenological model according to $\pi\left(\boldsymbol{X}_{1:M} \mid \boldsymbol{\theta}\right)$, and then draw observations from the observation process $\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{X}_{1:M}, \boldsymbol{\theta}\right)$ in order to make a comparison to an observed dataset. Alternatively, we can compare simulations from the phenomenological model, $\tilde{\boldsymbol{X}}_{1:M}$, with the observed dataset directly, with the knowledge that there will be a strictly positive lower bound on the tolerance. In other words, even if we simulated the exact underlying trajectory that we are observing with noise the distance between the simulated values, $\tilde{\boldsymbol{X}}_{1:M}$, and the observed dataset would be greater than zero. This can be incorporated into any of the above algorithms by replacing simulations from $\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}\right)$ with $\pi\left(\boldsymbol{X}_{1:M} \mid \boldsymbol{\theta}\right)$, and $\rho\left(\tilde{\boldsymbol{Y}}_{1:M}, \boldsymbol{Y}_{1:M}\right)$ with $\rho\left(\tilde{\boldsymbol{X}}_{1:M}, \boldsymbol{Y}_{1:M}\right)$. Classically, this is perceived as performing approximate inference on the model of interest. However, it can be thought of as exact inference where the observation error is implicitly assumed to be uniform [69]. This will be a poor assumption in most situations. Alternatively, as our measurement error is assumed to be Gaussian, then by weighting the simulated draws from $\pi\left(\boldsymbol{X}_{1:M} \mid \boldsymbol{\theta}\right)$ according to a Gaussian density then we can (in principle) perform exact inference on our model without knowledge of the likelihood. However, the variance of the weights will be very high, and so an MCMC scheme will mix poorly. An alternative approach would be to use an SMC scheme with intermediary distributions

$$\pi_{\kappa_s}\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:M}\right) \propto \pi\left(\boldsymbol{\theta}\right) \pi\left(\boldsymbol{X}_{1:M} \mid \boldsymbol{\theta}\right) \pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{X}_{1:M}, \boldsymbol{\theta}\right)^{\frac{1}{\kappa_s}},$$

for $\kappa_1 = \infty, ..., \kappa_S = 1$, such that the initial sample is from the prior distribution and the final sample is from the exact posterior distribution. The observations gradually become more and more influential as in classical ABC. This mirrors an annealed importance sampling approach, where the likelihood is said to be ‘heated’ for $\kappa > 1$ [70]. As with classic ABC the computational cost increases dramatically as $\kappa$ is lowered. From our experience, this approach performs as well as the standard ABC approach, in that we obtain similar posterior distributions with a similar computational cost.

## 3.4 Particle Filter Approaches

ABC methods bypass the problem of intractability by simulating from the model, removing the likelihood term from any analytical calculations. Alternatively, we can consider using approximations to the likelihood. In other words, for a given parameter vector $\boldsymbol{\theta}$, we want to approximate $\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}\right)$, and use this approximation in place of the likelihood term. In SSMs a particle filter can provide unbiased estimates of the likelihood. The particle filter is described below.

There are a number of variants of the particle filter. When first proposed the particle filter was based on sequential importance resampling (a variant of SIS with a resampling step in each iteration) [71]. The particle filter introduced here is based on an SMC framework targeting the the intermediary distributions

$$\tilde{\pi}_m\left(\boldsymbol{X}_{1:m}\right) = \pi\left(\boldsymbol{X}_{1:m} \mid \boldsymbol{Y}_{1:m}, \boldsymbol{\theta}\right), \tag{3.4.1}$$

so that each successive intermediary distribution includes an additional observation. The algorithm ends when all of the observations have been assimilated, so that we are left with a sample from $\pi\left(\boldsymbol{X}_{1:M} \mid \boldsymbol{Y}_{1:M}, \boldsymbol{\theta}\right)$. The intermediary distributions can be written

$$\frac{\tilde{\gamma}_m\left(\boldsymbol{X}_{1:m}\right)}{Z_m} = \frac{\pi\left(\boldsymbol{X}_{1:m}, \boldsymbol{Y}_{1:m} \mid \boldsymbol{\theta}\right)}{\pi\left(\boldsymbol{Y}_{1:m} \mid \boldsymbol{\theta}\right)}, \tag{3.4.2}$$

giving the distributions used within an SMC framework. Note that for $m = 1$

$$\pi\left(\boldsymbol{X}_1, \boldsymbol{Y}_1 \mid \boldsymbol{\theta}\right) = \pi\left(\boldsymbol{X}_1 \mid \boldsymbol{\theta}\right) \pi\left(\boldsymbol{Y}_1 \mid \boldsymbol{X}_1, \boldsymbol{\theta}\right), \tag{3.4.3}$$

and for $m \geq 2$

$$\pi\left(\boldsymbol{X}_{1:m}, \boldsymbol{Y}_{1:m} \mid \boldsymbol{\theta}\right) = \pi\left(\boldsymbol{X}_{1:m-1}, \boldsymbol{Y}_{1:m-1} \mid \boldsymbol{\theta}\right) \pi\left(\boldsymbol{X}_m \mid \boldsymbol{X}_{m-1}, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{Y}_m \mid \boldsymbol{X}_m, \boldsymbol{\theta}\right). \tag{3.4.4}$$

To sample from Equation 3.4.1 using IS, we must select an appropriate importance distribution. The optimal importance distribution is $\pi\left(\boldsymbol{X}_{1:m} \mid \boldsymbol{Y}_{1:m}, \boldsymbol{\theta}\right)$, but this is not available in most situations. However, where possible, the importance sampling distributions should be conditional on the next observation, otherwise there will be many proposals in regions of low posterior probability density. When $m = 1$, this suggests using an importance distribution of the form

$$\eta_1\left(\boldsymbol{X}_1\right) = r_1\left(\boldsymbol{X}_1 \mid \boldsymbol{Y}_1, \boldsymbol{\theta}\right), \tag{3.4.5}$$

so that the importance weights for a set of particles numbered $k = 1, ..., N_X$ become

$$\omega_1^{(k)} = \frac{\pi\left(\boldsymbol{X}_1^{(k)} \mid \boldsymbol{\theta}\right)\pi\left(\boldsymbol{Y}_1 \mid \boldsymbol{X}_1^{(k)}, \boldsymbol{\theta}\right)}{r_1\left(\boldsymbol{X}_1^{(k)} \mid \boldsymbol{Y}_1, \boldsymbol{\theta}\right)}. \tag{3.4.6}$$

When $m \geq 2$ the existing particles are resampled to avoid degeneracy. In the particle filter, this is done by sampling the ancestor particle index, $a_{m-1}^{(k)}$, of particle $k$, where $a_{m-1}^{(k)}$ takes the value $1, ..., N_X$ according to weights $\omega_{m-1}^{(1:N_X)}$, i.e. $\mathbb{P}\left(a_{m-1}^{(k)} = j\right) = \Omega_{m-1}^{(j)}$, where

$$\Omega_{m-1}^{(k)} = \frac{\omega_{m-1}^{(k)}}{\sum_{i=1}^{N_X} \omega_{m-1}^{(i)}}, \tag{3.4.7}$$

are the normalised importance weights. The resampled particles are extended using a proposal distribution of the form an $r_m\left(\boldsymbol{X}_m \mid \boldsymbol{X}_{m-1}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)$, so that the importance distribution takes the form

$$\eta_m\left(\boldsymbol{X}_{1:m}\right) = \pi\left(\boldsymbol{X}_{1:m-1}, \boldsymbol{Y}_{1:m-1} \mid \boldsymbol{\theta}\right) r_m\left(\boldsymbol{X}_m \mid \boldsymbol{X}_{m-1}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right), \tag{3.4.8}$$

so that it is conditioned on the next observation. The importance weights are

$$\begin{aligned} \omega_m^{(k)} &= \frac{\pi\left(\boldsymbol{X}_{1:m-1}^{(a_{m-1}^{(k)})}, \boldsymbol{Y}_{1:m-1} \mid \boldsymbol{\theta}\right)\pi\left(\boldsymbol{X}_m^{(k)} \mid \boldsymbol{X}_{m-1}^{(a_{m-1}^{(k)})}, \boldsymbol{\theta}\right)\pi\left(\boldsymbol{Y}_m \mid \boldsymbol{X}_m^{(k)}, \boldsymbol{\theta}\right)}{\pi\left(\boldsymbol{X}_{1:m-1}^{(a_{m-1}^{(k)})}, \boldsymbol{Y}_{1:m-1} \mid \boldsymbol{\theta}\right) r_m\left(\boldsymbol{X}_m^{(k)} \mid \boldsymbol{X}_{m-1}^{(a_{m-1}^{(k)})}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)} \\ &= \frac{\pi\left(\boldsymbol{X}_m^{(k)} \mid \boldsymbol{X}_{m-1}^{(a_{m-1}^{(k)})}, \boldsymbol{\theta}\right)\pi\left(\boldsymbol{Y}_m \mid \boldsymbol{X}_m^{(k)}, \boldsymbol{\theta}\right)}{r_m\left(\boldsymbol{X}_m^{(k)} \mid \boldsymbol{X}_{m-1}^{(a_{m-1}^{(k)})}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)}. \end{aligned} \tag{3.4.9}$$

The full pseudocode is presented in Algorithm 3.11.

The particle filter can provide an unbiased estimate of the likelihood, $\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}\right)$, via the identity

$$\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}\right) = \pi\left(\boldsymbol{Y}_1 \mid \boldsymbol{\theta}\right) \prod_{m=2}^{M} \pi\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta}\right). \tag{3.4.10}$$

It was shown in [72] that each component of the product, $\pi\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta}\right)$, can be replaced with an unbiased approximation, $\hat{\pi}\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta}\right)$, to obtain an unbiased estimate of the

**Algorithm 3.11** Particle filter targeting $\pi\left(\boldsymbol{X}_{1:M} \mid \boldsymbol{Y}_{1:M}, \boldsymbol{\theta}\right)$.

```
for $k = 1, ..., N_X$ do
    Sample $\boldsymbol{X}_1^{(k)} \sim r_1\left(\boldsymbol{X}_1 \mid \boldsymbol{Y}_1, \boldsymbol{\theta}\right)$.
    Set the importance weight
        $\omega_1^{(k)} = \frac{\pi\left(\boldsymbol{X}_1^{(k)} \mid \boldsymbol{\theta}\right) \pi\left(\boldsymbol{Y}_1 \mid \boldsymbol{X}_1^{(k)}, \boldsymbol{\theta}\right)}{r_1\left(\boldsymbol{X}_1^{(k)} \mid \boldsymbol{Y}_1, \boldsymbol{\theta}\right)}.$
end for
Normalise the weights. For $k = 1, ..., N_X$
    $\Omega_1^{(k)} = \frac{\omega_1^{(k)}}{\sum_{i=1}^{N_X} \omega_1^{(i)}}.$
for $m = 2, ..., M$ do
    for $k = 1, ..., N_X$ do
        Sample ancestor particle index $a_{m-1}^{(k)}$ according to weights $\Omega_{m-1}^{(1:N_X)}$.
        Sample $\boldsymbol{X}_m^{(k)} \sim r_m\left(\boldsymbol{X}_m \mid \boldsymbol{X}_{m-1}^{\left(a_{m-1}^{(k)}\right)}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)$.
        Extend particle trajectory $\boldsymbol{X}_{1:m}^{(k)} = \left(\boldsymbol{X}_{1:m-1}^{\left(a_{m-1}^{(k)}\right)}, \boldsymbol{X}_m^{(k)}\right)$.
        Set the importance weight
            $\omega_m^{(k)} = \frac{\pi\left(\boldsymbol{X}_m^{(k)} \mid \boldsymbol{X}_{m-1}^{\left(a_{m-1}^{(k)}\right)}, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{Y}_m \mid \boldsymbol{X}_m^{(k)}, \boldsymbol{\theta}\right)}{r_m\left(\boldsymbol{X}_m^{(k)} \mid \boldsymbol{X}_{m-1}^{\left(a_{m-1}^{(k)}\right)}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)}.$
    end for
    Normalise the weights. For $k = 1, ..., N_X$
        $\Omega_m^{(k)} = \frac{\omega_m^{(k)}}{\sum_{i=1}^{N_X} \omega_m^{(i)}}.$
end for
```

likelihood, denoted by $\hat{\pi}\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}\right)$. The components of the product are given by

$$\begin{aligned} &\pi(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta}) \\ &\qquad = \int \pi\left(\boldsymbol{Y}_m \mid \boldsymbol{X}_m, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{X}_m \mid \boldsymbol{X}_{m-1}, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{X}_{1:m-1} \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta}\right) d\boldsymbol{X}_{1:m} \\ &\qquad = \int \omega_m\left(\boldsymbol{X}_{1:m}\right) r_m\left(\boldsymbol{X}_m \mid \boldsymbol{X}_{m-1}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{X}_{1:m-1} \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta}\right) d\boldsymbol{X}_{1:m}, \end{aligned}$$

which can be estimated in each iteration of the particle filter via IS, specifically,

$$\hat{\pi}\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta}\right) = \frac{1}{N_X} \sum_{j=1}^{N_X} \omega_m^{(j)}, \tag{3.4.11}$$

provides an unbiased estimate of $\pi(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta})$. It is shown in [73] that the likelihood in Monte Carlo computations can be replaced with an unbiased approximation, and the resulting algorithm will still be a valid Monte Carlo sampler. This is known as the pseudo-marginal approach, and has been exploited in PMCMC [43] and SMC$^2$ [74] by replacing the likelihood with the unbiased estimate obtained from a particle filter in an MCMC scheme and SMC scheme respectively.

A well known draw back of the above particle filter (and SMC methods in general) is that as $M$ increases the joint posterior density approximations of $\pi\left(\boldsymbol{X}_{1:M} \mid \boldsymbol{Y}_{1:M}, \boldsymbol{\theta}\right)$ become poorer. Resampling in every iteration means that the sample at iteration $M$ can be traced back to a handful of ancestor particles from early iterations. In other words, the marginal density $\pi\left(\boldsymbol{X}_m \mid \boldsymbol{Y}_{1:M}, \boldsymbol{\theta}\right)$ will be approximated through only a handful of particles when $m$ is small. This problem will be made worse if $N_X$ is small. As a rule of thumb, $N_X$ should be greater than $M$ for the inference methods discussed in this chapter [43, 74] .

### Designing an efficient importance distribution

When the transition density is unavailable in closed form, it is standard to choose proposal distributions $r_m\left(\boldsymbol{X}_m \mid \boldsymbol{X}_{m-1}^{\left(a_{m-1}^{(k)}\right)}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right) = \pi\left(\boldsymbol{X}_m^{(k)} \mid \boldsymbol{X}_{m-1}^{\left(a_{m-1}^{(k)}\right)}, \boldsymbol{\theta}\right)$, so that we are simulating from the model, and the intractable term cancels out in the importance weight calculations. This choice will typically lead to many proposals in regions of low posterior probability density.

When using the Euler-Maruyama approximation to simulate from the underlying SDE, it is possible to condition the proposal distribution on the next observation, increasing the number of proposals in regions of high posterior probability density. As a reminder, the Euler-Maruyama approximation simulates over some time interval, $\Delta t$, from an SDE of the general form

$$d\boldsymbol{X}\left(t\right)=\boldsymbol{\mu}\left(\boldsymbol{X}\left(t\right),\boldsymbol{\theta}\right)dt+\boldsymbol{\Sigma}_{X}^{\frac{1}{2}}\left(\boldsymbol{X}\left(t\right),\boldsymbol{\theta}\right)d\boldsymbol{W}\left(t\right), \tag{3.4.12}$$

by partitioning the time interval into $J$ intervals of time $\Delta\tau=\frac{\Delta t}{J}$, giving the discrete time equation

$$\boldsymbol{X}\left(t+\Delta\tau\right)=\boldsymbol{\mu}\left(\boldsymbol{X}\left(t\right),\boldsymbol{\theta}\right)\Delta\tau+\boldsymbol{\Sigma}_{X}^{\frac{1}{2}}\left(\boldsymbol{X}\left(t\right),\boldsymbol{\theta}\right)\sqrt{\Delta\tau}\boldsymbol{\epsilon}_{t}, \tag{3.4.13}$$

where $\boldsymbol{\epsilon}_t$ is vector of independent standard Gaussian random variables. Denoting $\boldsymbol{X}_{m,j}=\boldsymbol{X}\left(t_m+j\Delta\tau\right)$, simulating from the discrete time equation between two observation times, $t_{m-1}$ and $t_m$, introduces $(J-1)\times u$ latent variables, $\boldsymbol{X}_{m-1,1},...,\boldsymbol{X}_{m-1,J-1}$. We can extend the importance distribution to also sample from the latent variables, by using a proposal distribution of the form $\tilde{r}_m\left(\boldsymbol{X}_{m-1,1},...,\boldsymbol{X}_{m-1,J}\mid\boldsymbol{Y}_m,\boldsymbol{\theta}\right)$. The importance weight calculation in the particle filter becomes

$$\omega_m^{(k)}=\frac{\prod_{j=1}^{J}\pi\left(\boldsymbol{X}_{m-1,j}\mid\boldsymbol{X}_{m-1,j-1},\boldsymbol{\theta}\right)\pi\left(\boldsymbol{Y}_m\mid\boldsymbol{X}_m,\boldsymbol{\theta}\right)}{\tilde{r}_m\left(\boldsymbol{X}_{m-1,1},...,\boldsymbol{X}_{m-1,J}\mid\boldsymbol{Y}_m,\boldsymbol{\theta}\right)}, \tag{3.4.14}$$

where the $\pi\left(\boldsymbol{X}_{m-1,j}\mid\boldsymbol{X}_{m-1,j-1},\boldsymbol{\theta}\right)$ are now assumed to be Gaussian. The latent variables can be iteratively sampled from $\pi\left(\boldsymbol{X}_{m-1,j}\mid\boldsymbol{X}_{m-1,j-1},\boldsymbol{\theta}\right)$, in which case the importance weights simplify as $\omega_m^{(k)}=\pi\left(\boldsymbol{Y}_m\mid\boldsymbol{X}_m,\boldsymbol{\theta}\right)$.

Recall that the general observation process is

$$\boldsymbol{Y}_m=\boldsymbol{C}\boldsymbol{X}_m+\boldsymbol{D}+\Sigma_Y\boldsymbol{\eta}_m. \tag{3.4.15}$$

In order to guide the particles in to regions of high likelihood, we can consider conditioning the value of $\boldsymbol{X}_{m-1,j}$ on a future observation, $\boldsymbol{Y}_m$, by approximating the distribution of $\boldsymbol{Y}_m$ conditional on $\boldsymbol{X}_{m-1,j-1}$ using a single Euler-Murayama step of size $\widetilde{\Delta t}=dt-(j-1)\Delta\tau$. First note that under an Euler-Murayama step of interval size $\widetilde{\Delta t}$

$$\boldsymbol{X}_m\mid\boldsymbol{X}_{m-1,j-1},\boldsymbol{\theta}\sim\mathcal{N}_u\left(\boldsymbol{X}_{m-1,j-1}+\boldsymbol{\mu}_{m-1,j-1}\widetilde{\Delta t},\Sigma_{m-1,j-1}\widetilde{\Delta t}\right), \tag{3.4.16}$$

and so

$$\boldsymbol{Y}_m\mid\boldsymbol{X}_{m-1,j-1},\boldsymbol{\theta}\sim\mathcal{N}_w\left(\boldsymbol{C}\left(\boldsymbol{X}_{m-1,j-1}+\boldsymbol{\mu}_{m-1,j-1}\widetilde{\Delta t}\right)+\boldsymbol{D},\right.$$
$$\left.\boldsymbol{C}\Sigma_{m-1,j-1}\boldsymbol{C}^T\widetilde{\Delta t}+\Sigma_Y\right), \tag{3.4.17}$$

where $\boldsymbol{\mu}_{m,j}=\boldsymbol{\mu}\left(\boldsymbol{X}_{m,j},\boldsymbol{\theta}\right)$ and $\Sigma_{m,j}=\Sigma_X\left(\boldsymbol{X}_{m,j},\boldsymbol{\theta}\right)$. In other words, we use a single Euler-Murayama step to predict the value of $\boldsymbol{X}_m$, which is assumed to be Gaussian, and then add the

Gaussian observation error after scaling. This approach was taken in [42], but with no scaling or displacement terms in the observation process. The joint distribution of $\boldsymbol{X}_{m-1,j}$ and $\boldsymbol{Y}_m$, given $\boldsymbol{X}_{m-1,j-1}$, is then

$$\begin{pmatrix} \boldsymbol{X}_{m-1,j} \\ \boldsymbol{Y}_m \end{pmatrix} \mid \boldsymbol{X}_{m-1,j-1}, \boldsymbol{\theta} \sim \mathcal{N}_{u+w} \left( \begin{pmatrix} \boldsymbol{X}_{m-1,j-1} + \boldsymbol{\mu}_{m-1,j-1}\Delta\tau \\ \boldsymbol{C}\left(\boldsymbol{X}_{m-1,j-1} + \boldsymbol{\mu}_{m-1,j-1}\widetilde{\Delta t}\right) + \boldsymbol{D} \end{pmatrix}, \right.$$

$$\left. \begin{pmatrix} \boldsymbol{\Sigma}_{m-1,j-1}\Delta\tau & \boldsymbol{\Sigma}_{m-1,j-1}\boldsymbol{C}^T\Delta\tau \\ \boldsymbol{C}\boldsymbol{\Sigma}_{m-1,j-1}\Delta\tau & \boldsymbol{C}\boldsymbol{\Sigma}_{m-1,j-1}\boldsymbol{C}^T\widetilde{\Delta t} + \boldsymbol{\Sigma}_Y \end{pmatrix} \right). \quad (3.4.18)$$

Using standard multivariate Gaussian conditioning rules [75] to condition on $\boldsymbol{Y}_m$ gives

$$\boldsymbol{X}_{m-1,j} \mid \boldsymbol{X}_{m-1,j-1}, \boldsymbol{Y}_m, \boldsymbol{\theta} \sim \mathcal{N}_u\left(\boldsymbol{\mathcal{M}}_{m-1,j-1}, \boldsymbol{\mathcal{S}}_{m-1,j-1}\right), \quad (3.4.19)$$

where

$$\boldsymbol{\mathcal{M}}_{m-1,j-1} = \boldsymbol{X}_{m-1,j-1} + \boldsymbol{\mu}_{m-1,j-1}\Delta\tau +$$
$$\boldsymbol{B}^T\boldsymbol{A}^{-1}\left(\boldsymbol{Y}_m - \left(\boldsymbol{C}\left(\boldsymbol{X}_{m-1,j-1} + \boldsymbol{\mu}_{m-1,j-1}\widetilde{\Delta t}\right) + \boldsymbol{D}\right)\right),$$

and

$$\boldsymbol{\mathcal{S}}_{m-1,j-1} = \boldsymbol{\Sigma}_{m-1,j-1}\Delta\tau - \boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B},$$

with

$$\boldsymbol{A} = \left(\boldsymbol{C}\boldsymbol{\Sigma}_{m-1,j-1}\boldsymbol{C}^T\widetilde{\Delta t} + \boldsymbol{\Sigma}_Y\right),$$

and

$$\boldsymbol{B} = \boldsymbol{C}\boldsymbol{\Sigma}_{m-1,j-1}\Delta\tau.$$

To understand why this is a more efficient proposal distribution than simulating from the model, we consider an alternative derivation of the above. Consider the case in this chapter, where our observations are scaled and noisy versions of the observable state, $X_1$. For $X_1$, if we simulate from the model under the Euler-Murayama approximation,

$$X_{1;m-1,j} \mid X_{1;m-1,j-1} \sim \mathcal{N}\left(X_{1;m-1,j-1} + \mu_{1;m-1,j-1}\Delta\tau, \sigma_1^2\Delta\tau\right). \quad (3.4.20)$$

Also consider predicting the value of $X_1$ at observation time $t_m$ using a single Euler-Murayama step over the interval $\widetilde{\Delta t}$. Denoting the predicted value as $\hat{X}_m$,

$$\hat{X}_m \mid X_{1;m-1,j-1} \sim \mathcal{N}\left(X_{1;m-1,j-1} + \mu_{1;m-1,j-1}\widetilde{\Delta t}, \sigma_1^2\widetilde{\Delta t}\right). \quad (3.4.21)$$

Scaling $\hat{X}_m$ allows us to compare our prediction with the observation, and the difference is given by

$$Y_m - \left(C\hat{X}_m + D\right). \tag{3.4.22}$$

We can then consider adjusting the mean of the proposal distribution through a small perturbation of the form

$$\mathcal{K}C^{-1}\left(Y_m - \left(C\hat{X}_m + D\right)\right), \tag{3.4.23}$$

so that we propose $X_{1;m-1,j}$ from

$$X_{1;m-1,j} \mid X_{1;m-1,j-1}, Y_m \sim \mathcal{N}\left(X_{1;m-1,j-1} + \mu_{1;m-1,j-1}\Delta\tau + \right.$$
$$\left. \mathcal{K}C^{-1}\left(Y_m - \left(C\hat{X}_m + D\right)\right), \sigma_1^2\Delta\tau\right), \tag{3.4.24}$$

where $\mathcal{K}$ determines the size of the perturbation. The optimal choice of $\mathcal{K}$ will depend on the relative difference of the model discrepancy variance over one time step and the observation error. Considering a perturbation over one step of size $\widetilde{\Delta t}$, it is desired that $\mathcal{K} \to 1$ as $\frac{\sigma_Y^2}{C^2\sigma_1^2\widetilde{\Delta t}} \to 0$, and $\mathcal{K} \to 0$ as $\frac{C^2\sigma_1^2\widetilde{\Delta t}}{\sigma_Y^2} \to 0$. In other words, when the observation error is relatively small, the perturbation should give proposals closer to the observation, and when the observation error is relatively large the proposals do not deviate strongly from the model. A suitable choice is

$$\mathcal{K} = \frac{C^2\sigma_1^2\widetilde{\Delta t}}{C^2\sigma_1^2\widetilde{\Delta t} + \sigma_Y^2}, \tag{3.4.25}$$

which is the ratio of the variance from the model, and the combined variance of the model and observation process. In each iteration of step size $\Delta\tau$, we scale the perturbation by $\frac{\Delta\tau}{\widetilde{\Delta t}}$ to give the correct proportion, giving

$$\mathcal{K} = \frac{C^2\sigma_1^2\Delta\tau}{C^2\sigma_1^2\widetilde{\Delta t} + \sigma_Y^2}. \tag{3.4.26}$$

This suggests using a proposal distribution of the form

$$X_{1;m-1,j} \mid X_{1;m-1,j-1}, Y_m \sim \mathcal{N}\left(X_{1;m-1,j-1} + \mu_{1;m-1,j-1}\Delta\tau + \right.$$
$$\left. \frac{C^2\sigma_1^2\Delta\tau}{C^2\sigma_1^2\widetilde{\Delta t} + \sigma_Y^2}C^{-1}\left(Y_m - \left(C\hat{X}_m + D\right)\right), \sigma_1^2\Delta\tau\right), \tag{3.4.27}$$

For the unobservable states, proposals are made by simulating from the model under the Euler-Murayama approximation. This can be considered as a scalar version of Equation 3.4.19, giving insight in to why conditioning on a future observation in a proposal distribution leads to improved efficiency.

Note that through a formal conditioning on the observation, $\boldsymbol{Y}_m$, the variance of the proposal distribution is reduced as the observation time is neared. For the observable state the proposal variance becomes

$$\frac{C^2\sigma_1^2\left(\widetilde{\Delta t}-\Delta\tau\right)+\sigma_Y^2}{C^2\sigma_1^2\widetilde{\Delta t}+\sigma_Y^2}\sigma_1^2\Delta\tau, \tag{3.4.28}$$

which can be considered as scaling the proposal variance by the variance remaining after the integration step has been performed relative to the total variance. This ratio is close to 1 for $\Delta\tau \ll \widetilde{\Delta t}$, and for $C^2\sigma_x^2\widetilde{\Delta t} \ll \sigma_y^2$. Whereas in the case $\Delta\tau = \widetilde{\Delta t}$, and $\Sigma_y \ll C^2\Sigma_x\Delta t$, the proposal variance is approximately the observation variance. This is expected to be beneficial for informative observations, as ensuring the state of the system is near an observation before reaching it prevents rapid state changes, which have low likelihoods and thus lead to poor acceptance rates in MCMC, and a poor ESS in IS.

### 3.4.1 Particle MCMC

Particle MCMC (PMCMC) samplers are a collection of methods embedding the particle filter into an MCMC algorithm. We focus on a PMCMC algorithm designed to sample from $\pi\left(\boldsymbol{\theta},\boldsymbol{X}_{1:M}\mid\boldsymbol{Y}_{1:M}\right)$, called the particle marginal Metropolis Hastings (PMMH) sampler [43]. Consider the case where it is possible to sample from $\pi\left(\boldsymbol{X}_{1:M}\mid\boldsymbol{Y}_{1:M},\boldsymbol{\theta}\right)$, and note that

$$\pi\left(\boldsymbol{\theta},\boldsymbol{X}_{1:M}\mid\boldsymbol{Y}_{1:M}\right)=\pi\left(\boldsymbol{\theta}\mid\boldsymbol{Y}_{1:M}\right)\pi\left(\boldsymbol{X}_{1:M}\mid\boldsymbol{Y}_{1:M},\boldsymbol{\theta}\right). \tag{3.4.29}$$

In this situation it makes sense to use proposal distributions of the form

$$q\left(\boldsymbol{\theta}^*,\boldsymbol{X}^*_{1:M}\mid\boldsymbol{\theta},\boldsymbol{X}_{1:M}\right)=q\left(\boldsymbol{\theta}^*\mid\boldsymbol{\theta}\right)\pi\left(\boldsymbol{X}^*_{1:M}\mid\boldsymbol{Y}_{1:M},\boldsymbol{\theta}^*\right). \tag{3.4.30}$$

The acceptance rate of the Metropolis Hastings acceptance ratio is then

$$\frac{\pi\left(\boldsymbol{\theta}^*,\boldsymbol{X}^*_{1:M}\mid\boldsymbol{Y}_{1:M}\right)q\left(\boldsymbol{\theta},\boldsymbol{X}_{1:M}\mid\boldsymbol{\theta}^*,\boldsymbol{X}^*_{1:M}\right)}{\pi\left(\boldsymbol{\theta},\boldsymbol{X}_{1:M}\mid\boldsymbol{Y}_{1:M}\right)q\left(\boldsymbol{\theta}^*,\boldsymbol{X}^*_{1:M}\mid\boldsymbol{\theta},\boldsymbol{X}_{1:M}\right)}=\frac{\pi\left(\boldsymbol{\theta}^*\right)\pi\left(\boldsymbol{Y}_{1:M}\mid\boldsymbol{\theta}^*\right)q\left(\boldsymbol{\theta}\mid\boldsymbol{\theta}^*\right)}{\pi\left(\boldsymbol{\theta}\right)\pi\left(\boldsymbol{Y}_{1:M}\mid\boldsymbol{\theta}\right)q\left(\boldsymbol{\theta}^*\mid\boldsymbol{\theta}\right)}, \tag{3.4.31}$$

showing that the algorithm is sampling $\boldsymbol{\theta}$ from the correct marginal distribution $\pi\left(\boldsymbol{\theta}\mid\boldsymbol{Y}_{1:M}\right)$. It can also be shown that this algorithm satisfies detailed balance [43]. It is rarely possible to sample from $\pi\left(\boldsymbol{X}_{1:M}\mid\boldsymbol{Y}_{1:M},\boldsymbol{\theta}\right)$, and in our case the likelihood, $\pi\left(\boldsymbol{Y}_{1:M}\mid\boldsymbol{\theta}\right)$, is intractable. PMMH uses the particle filter approximations of these quantities, denoted $\hat{\pi}\left(\boldsymbol{X}_{1:M}\mid\boldsymbol{Y}_{1:M},\boldsymbol{\theta}\right)$ and $\hat{\pi}\left(\boldsymbol{Y}_{1:M}\mid\boldsymbol{\theta}\right)$. Despite using the particle filter approximations, the resulting Markov chain leaves the posterior distribution, $\pi\left(\boldsymbol{\theta},\boldsymbol{X}_{1:M}\mid\boldsymbol{Y}_{1:M}\right)$, invariant. This has given rise to the term "exact-approximations" to describe such algorithms.

In each iteration of the algorithm, $\boldsymbol{X}^*_{1:M}$ is sampled by sampling particle index $b_M$ according to weights $\Omega_M^{(1:N_X)}$, from the particle filter. The ancestry of this particle can then be defined deterministically through the recursive relation $b_{m-1} = a_{m-1}^{(b_m)}$, where the $a_m^{(k)}$ are the ancestor particle indices. The proposed trajectory is then given by the combined ancestry $b_{1:M} = \{b_1, ..., b_M\}$ as $\boldsymbol{X}_{1:M}^{(b_M)} = \left\{\boldsymbol{X}_1^{(b_1)}, ..., \boldsymbol{X}_M^{(b_M)}\right\}$.

Formally, the PMMH algorithm is an MCMC algorithm leaving the extended distribution $\pi\left(\boldsymbol{\theta}, b_M, \boldsymbol{X}_{1:M}^{1:N_X}, a_{1:M-1}^{1:N_X}\right)$ invariant, where the extended distribution is defined as

$$\pi\left(\boldsymbol{\theta}, b_M, \boldsymbol{X}_{1:M}^{1:N_X}, a_{1:M-1}^{1:N_X}\right) = \frac{\pi\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:M}^{(b_M)} \mid \boldsymbol{Y}_{1:M}\right)}{N_X^M}$$

$$\times \left\{ \frac{\left\{\prod_{i=1}^{N_X} r_1\left(\boldsymbol{X}_1^{(i)} \mid \boldsymbol{Y}_1, \boldsymbol{\theta}\right)\right\}\left\{\prod_{m=2}^{M}\prod_{i=1}^{N_X} \Omega_{m-1}^{\left(a_{m-1}^{(i)}\right)} r_m\left(\boldsymbol{X}_m^{(i)} \mid \boldsymbol{X}_{m-1}^{\left(a_{m-1}^{(i)}\right)}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)\right\}}{r_1\left(\boldsymbol{X}_1^{(b_1)} \mid \boldsymbol{Y}_1, \boldsymbol{\theta}\right)\prod_{m=2}^{M}\Omega_{m-1}^{(b_{m-1})} r_m\left(\boldsymbol{X}_m^{(b_m)} \mid \boldsymbol{X}_{m-1}^{(b_{m-1})}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)} \right\}, \tag{3.4.32}$$

which admits $\pi\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)$ as a marginal distribution. The extended proposal distribution is given by

$$q\left(\boldsymbol{\theta}, b_M, \boldsymbol{X}_{1:M}^{1:N_X}, a_{1:M-1}^{1:N_X} \mid \boldsymbol{\theta}^*, b_M^*, \boldsymbol{X}_{1:M}^{*1:N_X}, a_{1:M-1}^{*1:N_X}\right) = q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*\right)\Omega_M^{(b_M)}$$

$$\times \left\{\prod_{i=1}^{N_X} r_1\left(\boldsymbol{X}_1^{(i)} \mid \boldsymbol{Y}_1, \boldsymbol{\theta}\right)\right\}\left\{\prod_{m=2}^{M}\prod_{i=1}^{N_X}\Omega_{m-1}^{\left(a_{m-1}^{(i)}\right)} r_m\left(\boldsymbol{X}_m^{(i)} \mid \boldsymbol{X}_{m-1}^{\left(a_{m-1}^{(i)}\right)}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)\right\}, \tag{3.4.33}$$

so that

$$\frac{\pi\left(\boldsymbol{\theta}, b_M, \boldsymbol{X}_{1:M}^{1:N_X}, a_{1:M-1}^{1:N_X}\right)}{q\left(\boldsymbol{\theta}, b_M, \boldsymbol{X}_{1:M}^{1:N_X}, a_{1:M-1}^{1:N_X} \mid \boldsymbol{\theta}^*, b_M^*, \boldsymbol{X}_{1:M}^{*1:N_X}, a_{1:M-1}^{*1:N_X}\right)} \propto \frac{\pi\left(\boldsymbol{\theta}\right)\prod_{m=1}^{M}\frac{1}{N_X}\sum_{i=1}^{N_X} w_m^{(i)}}{q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*\right)}.$$

In other words, the target distribution of the Markov chain is extended to include all of the random variables used in the particle filter. Under mild assumptions $\prod_{m=1}^{M}\frac{1}{N_X}\sum_{i=1}^{N_X} w_m^{(i)}$ is a consistent estimator for $\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}\right)$, and the PMMH acceptance probability tends to Equation 3.4.31 as $N_X \to \infty$. For any number of particles, $N_X \geq 1$, the PMMH algorithm leaves $\pi\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)$ invariant. The pseudocode is presented in Algorithm 3.12.

There are a number of decisions that the user needs to make. The number of iterations to be performed, the length of the burn-in period, and the proposal distributions for the parameters are typical MCMC tuning parameters. Additionally, it is up to the user to decide the number of particles. A rule of thumb is to choose $N_X = \mathcal{O}(M)$ for a well mixing algorithm [43].

**Algorithm 3.12** PMCMC sampler targeting $\pi\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)$.

```
Initialise $\boldsymbol{\theta}^{(1)}$.
Run the particle filter targeting $\pi\left(\boldsymbol{X}_{1:M} \mid \boldsymbol{\theta}^{(1)}, \boldsymbol{Y}_{1:M}\right)$. Sample trajectory $\boldsymbol{X}_{1:M}^{(1)}$ from $\hat{\pi}\left(\boldsymbol{X}_{1:M} \mid \boldsymbol{\theta}^{(1)}, \boldsymbol{Y}_{1:M}\right)$ and record the marginal likelihood estimate $\hat{\pi}\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}^{(1)}\right)$.
for $n = 2, ..., N_\theta$ do
    Propose move to $\boldsymbol{\theta}^*$ according to proposal density $q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(n-1)}\right)$.
    Run the particle filter targeting $\pi\left(\boldsymbol{X}_{1:M} \mid \boldsymbol{\theta}^*, \boldsymbol{Y}_{1:M}\right)$. Sample trajectory $\boldsymbol{X}_{1:M}^*$ from $\hat{\pi}\left(\boldsymbol{X}_{1:M} \mid \boldsymbol{\theta}^*, \boldsymbol{Y}_{1:M}\right)$ and record the likelihood estimate $\hat{\pi}\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}^*\right)$.
    With probability
        $$\lambda\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(n-1)}\right) = \min\left(1, \frac{\pi\left(\boldsymbol{\theta}^*\right) \hat{\pi}\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}^*\right) q\left(\boldsymbol{\theta}^{(n-1)} \mid \boldsymbol{\theta}^*\right)}{\pi\left(\boldsymbol{\theta}^{(n-1)}\right) \hat{\pi}\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}^{(n-1)}\right) q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(n-1)}\right)}\right)$$
    set $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}^*$ and $\boldsymbol{X}_{1:M}^{(n)} = \boldsymbol{X}_{1:M}^*$. Otherwise set $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}^{(n-1)}$ and $\boldsymbol{X}_{1:M}^{(n)} = \boldsymbol{X}_{1:M}^{(n-1)}$.
end for
```

### Results

The PMCMC algorithm is performed over 45000 iterations, and the first 15000 samples are discarded as a burn-in. We choose $N_X = 1000$, so that in total the algorithm requires the equivalent of 45 million simulations from the model. The code was written in c and R, and has a runtime of approximately 24 hours on a 3 GHz processor. While this is more computationally expensive than ABC-MCMC, PMCMC is targeting the posterior distribution, rather than an approximation to it.

The proposal distribution, $q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(n-1)}\right)$, is a multivariate Gaussian distribution with mean $\boldsymbol{\theta}^{(n-1)}$ and a diagonal covariance matrix where the variance terms are provided in Table 3.5. These values were selected based on several trial runs. Within the particle filter, values of $X_2\left(t_1\right)$ are proposed from the prior distribution, and values of $X_1\left(t_1\right)$ are proposed from a Gaussian distribution with mean $\frac{1}{C}\left(Y_1 - D\right)$ and variance $\frac{\sigma_Y^2}{C^2}$. These proposal distributions ensure a good agreement with the first observation, whereas drawing $X_1$ from the prior distribution would lead to many low-weight particles. For $m \geq 2$ the proposal distributions are conditioned on the next observation, as described in the previous section.

The resulting MCMC chain has an acceptance rate of 0.38 after burn-in. The chain is thinned to give a sample size of 1000 by taking every 20th value. The parameter marginal distributions are shown in Figure 3.5. The posterior variance is much lower than the ABC posteriors, which is to be expected given that ABC is targeting an approximation to the posterior distribution. Most of the true values of the parameters are in regions of high posterior probability density. In this case the values of the coprecession scaling term $\gamma_C$ and the stochastic scaling term $\sigma_1$ are in the tails of the posterior distribution, but this seems to be a feature of the particular simulated dataset. The posterior variance of $X_2\left(t_1\right)$ is still quite large. The start of the simulated dataset

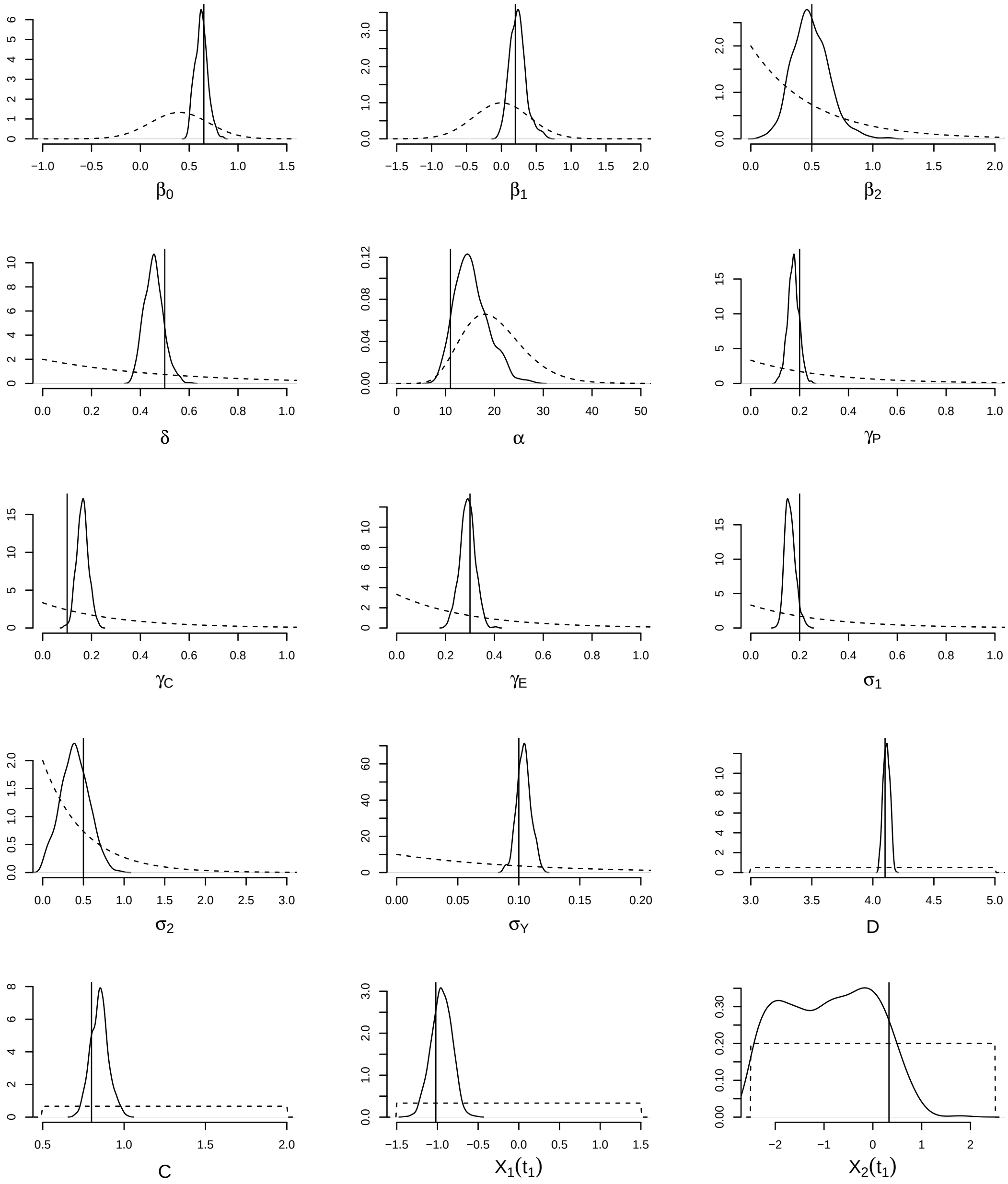

**Figure 3.5:** Marginal posterior distributions of the parameters of CR14-a, obtained using PMCMC in the simulation study. Vertical lines indicate the values used to generate the data. Dashed lines show the prior distributions. Comparing these distributions with the approximate posterior distributions obtained using ABC methods, shown, for example, in Figure 3.4, suggests that we are learning more about the parameters using PMCMC. It appears that using the likelihood approximation in PMCMC extracts more information from the observations than using the ABC approximation. The values of the coprecession scaling term $\gamma_C$ and the stochastic scaling term $\sigma_1$ are in the tails of the posterior distribution, but this seems to be a feature of the particular simulated dataset.

| Parameter | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\delta$ | $\alpha$ | $\gamma_P$ | $\gamma_C$ |
|---|---|---|---|---|---|---|---|
| $\tau$ | 0.014 | 0.023 | 0.03 | 0.009 | 0.71 | 0.005 | 0.005 |

| Parameter | $\gamma_E$ | $\sigma_1$ | $\sigma_2$ | $\sigma_Y$ | $D$ | $C$ |
|---|---|---|---|---|---|---|
| $\tau$ | 0.006 | 0.004 | 0.02 | 0.001 | 0.005 | 0.01 |

**Table 3.5:** Standard deviation ($\tau$) associated with each of the parameters in the Gaussian random walk proposal for PMCMC. These values were selected based on several trial runs.

is undergoing a transition from the positive branch of $X_2$ to the negative branch. As such any trajectories from negative values of $X_2\left(t_1\right)$ are quickly drawn to the limit cycle, so that it is easy to detect if $X_2\left(t_1\right)$ is positive or negative, but difficult to find the true value used.

As the PMCMC algorithm targets a different distribution than ABC methods, the correlations in the accepted sample are different. In the PMCMC sample $\beta_0$ is still strongly correlated with $\delta$ and $D$, with coefficients 0.93 and 0.54 respectively, but also with $\alpha$ and $C$, with correlation coefficients -0.65 and -0.61. Furthermore $\beta_1$ is correlated with $\beta_2$ (0.84), $\delta$ (0.52), and $C$ (0.72); $\delta$ with $\alpha$ (-0.72), and $C$ (-0.68); $\alpha$ with $C$ (0.73); and $\gamma_E$ with $C$ (-0.61).

### 3.4.2 SMC$^2$

The SMC$^2$ algorithm embeds the particle filter within an SMC algorithm targeting the sequence of intermediary distributions $\pi_0, ..., \pi_M$ defined by

$$\pi_0 = \pi(\boldsymbol{\theta}), \qquad \pi_m = \pi(\boldsymbol{\theta}, \boldsymbol{X}_{1:m} \mid \boldsymbol{Y}_{1:m}), \quad m \geq 1.$$

This is achieved by sampling $N_\theta$ parameter particles from the prior distribution when initialising the algorithm. A particle filter is then attached to each parameter particle, giving a total of $N_\theta \times N_X$ particles. In every iteration the particle filter is performed up until the next observation [74]. Formally the target density in each iteration is given by

$$\pi_m\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:m}^{1:N_X}, a_{1:m-1}^{1:N_X}\right) =$$

$$\frac{\pi\left(\boldsymbol{\theta} \mid \boldsymbol{Y}_{1:m}\right)}{N_X^k} \sum_{j=1}^{N_X} \frac{\pi\left(\boldsymbol{X}_{1:m}^{(b_m^j)} \mid \boldsymbol{Y}_{1:m}, \boldsymbol{\theta}\right)}{r_1\left(\boldsymbol{X}_1^{(b_m^j)} \mid \boldsymbol{Y}_1, \boldsymbol{\theta}\right) \prod_{k=2}^m \Omega_{k-1}^{(b_{k-1}^j)} r_k\left(\boldsymbol{X}_k^{(b_k^j)} \mid \boldsymbol{X}_{k-1}^{(b_{k-1}^j)}, \boldsymbol{Y}_k, \boldsymbol{\theta}\right)} \times$$

$$\left\{\prod_{i=1}^{N_X} r_1\left(\boldsymbol{X}_1^{(i)} \mid \boldsymbol{Y}_1, \boldsymbol{\theta}\right)\right\}\left\{\prod_{k=2}^{m}\prod_{i=1}^{N_X} \Omega_{k-1}^{\left(a_{k-1}^{(i)}\right)} r_k\left(\boldsymbol{X}_k^{(i)} \mid \boldsymbol{X}_{k-1}^{\left(a_{k-1}^{(i)}\right)}, \boldsymbol{Y}_k, \boldsymbol{\theta}\right)\right\}, \quad (3.4.34)$$

where $b_1^j, ..., b_m^j$ denotes the ancestry of particle $\boldsymbol{X}_m^j$. Note that this distribution differs from the target distribution of the PMCMC algorithm as $b_m$ is not sampled. The proposal density is given by

$$\eta_m\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:m}^{1:N_X}, a_{1:m-1}^{1:N_X}\right) = \pi\left(\boldsymbol{\theta}\right) \times$$

$$\left\{\prod_{i=1}^{N_X} r_1\left(\boldsymbol{X}_1^{(i)} \mid \boldsymbol{Y}_1, \boldsymbol{\theta}\right)\right\}\left\{\prod_{k=2}^{m}\prod_{i=1}^{N_X}\Omega_{k-1}^{\left(a_{k-1}^{(i)}\right)} r_k\left(\boldsymbol{X}_k^{(i)} \mid \boldsymbol{X}_{k-1}^{\left(a_{k-1}^{(i)}\right)}, \boldsymbol{Y}_k, \boldsymbol{\theta}\right)\right\}, \quad (3.4.35)$$

so that

$$\frac{\pi_m\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:m}^{1:N_X}, a_{1:m-1}^{1:N_X}\right)}{\eta_m\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:m}^{1:N_X}, a_{1:m-1}^{1:N_X}\right)} \propto \prod_{k=1}^{m} \frac{1}{N_X}\sum_{i=1}^{N_X} w_k^{(i)}.$$

A resampling step is added when the ESS falls below $\frac{N_\theta}{2}$. As the parameter particles are not perturbed between iterations, resampling causes a diminished particle diversity. The particle diversity is improved by running a PMCMC algorithm targeting $\pi(\boldsymbol{\theta}, \boldsymbol{X}_{1:m} \mid \boldsymbol{Y}_{1:m})$ after each resampling step. Note that resampling effectively alters the proposal distribution, as rather than being distributed according to the prior distribution, the parameter particles are distributed according to the mixture distribution

$$\sum_{n=1}^{N_\theta} W_{m-1}^{(n)}\pi\left(\boldsymbol{\theta}^{(n)}\right), \quad (3.4.36)$$

where $W_m$ are the normalised weights of the parameter particles in iteration $m$. The pseudocode is presented in Algorithm 3.13.

The user choices are the number of particles, $N_\theta$ and $N_x$, the length of the PMCMC chain in the resampling steps, and the proposal distributions in the PMCMC steps. Typically, $N_\theta$ will be decided by the available computational resources and desired sample size. It is not necessary to keep $N_X$ fixed. It has been suggested that $N_X$ should be $\mathcal{O}(m)$, implying that few state particles are required in early iterations. This means that $N_X$ can be automatically calibrated, for example by doubling $N_X$ whenever the acceptance rate of the PMCMC algorithm falls below some set tolerance [74]. Having a collection of parameter particles in iteration $m$ also allows the PMCMC proposal distributions to be automatically calibrated by using the sample mean and covariance to design a random-walk proposal, or independent Gaussian proposals, for example.

A notable property of SMC$^2$ is the memory cost involved when storing the complete trajectory of every particle, which is $\mathcal{O}\left(MN_\theta \times N_X\right)$. When interest lies only in the parameter marginal distribution, $\boldsymbol{X}_{1:m-2}$ can be discarded, as for iteration $m$ only $\boldsymbol{X}_{m-1}$ are required due to the Markov property of the model. This reduces the memory requirements to $\mathcal{O}\left(N_\theta N_X\right)$.

**Algorithm 3.13** SMC$^2$ algorithm targeting $\pi\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)$.

```
for $n = 1, ..., N_\theta$ do
    Sample $\boldsymbol{\theta}^{(n)}$ from the prior distribution, $\pi\left(\boldsymbol{\theta}\right)$.
    Set the importance weight
        $$W_0^{(n)} = \frac{1}{N_\theta}.$$
end for
for $m = 1, ..., M$ do
    if ESS$< \frac{N_\theta}{2}$ then
        for $n = 1, ..., N_\theta$ do
            Sample $\boldsymbol{\theta}^{*(n)}$ and $\boldsymbol{X}_{1:m-1}^{*(1:N_X,n)}$ from $\boldsymbol{\theta}^{(1:N_\theta)}$ and $\boldsymbol{X}_{1:m-1}^{(1:N_X,1:N_\theta)}$, according to weights $W_{m-1}^{(1:N_\theta)}$.
            Sample $\boldsymbol{\theta}^{**(n)}$ and $\boldsymbol{X}_{1:m-1}^{**(1:N_X,n)}$ from a PMCMC algorithm initialised with $\boldsymbol{\theta}^{*(n)}$ and $\boldsymbol{X}_{1:m-1}^{*(1:N_X,n)}$, and targeting $\pi\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:m-1} \mid \boldsymbol{Y}_{1:m-1}\right)$.
        end for
        Set $\boldsymbol{\theta}^{(1:N_\theta)} = \boldsymbol{\theta}^{**(1:N_\theta)}$ and $\boldsymbol{X}_{1:m-1}^{(1:N_X,1:N_\theta)} = \boldsymbol{X}_{1:m-1}^{**(1:N_X,1:N_\theta)}$.
        Set the importance weights. For $n = 1, ..., N_\theta$
            $$W_{m-1}^{(n)} = \frac{1}{N_\theta}.$$
    end if
    for $n = 1, ..., N_\theta$ do
        Sample $\boldsymbol{X}_{1:m}^{(1:N_X,n)}$ by performing iteration $m$ of the particle filter, and record estimates of $\hat{\pi}\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta}^{(n)}\right)$ and $\hat{\pi}\left(\boldsymbol{Y}_{1:m} \mid \boldsymbol{\theta}^{(n)}\right)$.
        Set the importance weights
            $$w_m^{(n)} = w_{m-1}^{(n)} \hat{\pi}\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta}^{(n)}\right).$$
    end for
    Normalise the weights. For $n = 1, ..., N_\theta$
        $$W_m^{(n)} = \frac{w_m^{(n)}}{\sum_{i=1}^{N_\theta} w_m^{(i)}}.$$
end for
```

### Results

The SMC$^2$ algorithm is performed with $N_\theta = 1000$ parameter particles and $N_X = 1000$ state particles, giving 1 million particles in total. The parameter proposal distribution in the PMCMC resampling steps is taken to be independent Gaussian, with mean and variance equal to the mean and variance of the resampled values. We choose a PMCMC chain length of 10, which seems sufficient to maintain a good particle diversity. When the particle filter is first initialised we sample initial conditions in the same manner as the PMCMC algorithm. In the resampling stages the initial conditions are sampled from an independent Gaussian distribution, with the same mean and variance as the current sample, so that our proposal distributions are adaptive.

The full algorithm required the equivalent of 120 million simulations, making it the most expensive algorithm in this chapter. The code was written in c and R, and has a runtime of approximately 60 hours on a 3 GHz processor. The final sample had 984 distinct particles, suggesting that the chain length could be reduced while keeping a large number of distinct particles, which would reduce the number of required simulations.

The parameter marginal distributions are shown in Figure 3.6. There is very strong agreement between the SMC$^2$ posterior and PMCMC posterior, as would be expected as both algorithms target the same distribution. The interpretation of the posterior sample is therefore not repeated. There is also a strong agreement with the correlation in the posterior sample. $\beta_0$ is correlated with $\delta$ (0.95), $\alpha$ (-0.73), $D$ (0.52), and $C$ (-0.66); $\beta_1$ is correlated with $\beta_2$ (0.85), $\delta$ (0.6), and $C$ (0.77); $\delta$ with $\alpha$ (-0.78), and $C$ (-0.72); $\alpha$ with $C$ (0.78); and $\gamma_E$ with $C$ (-0.6).

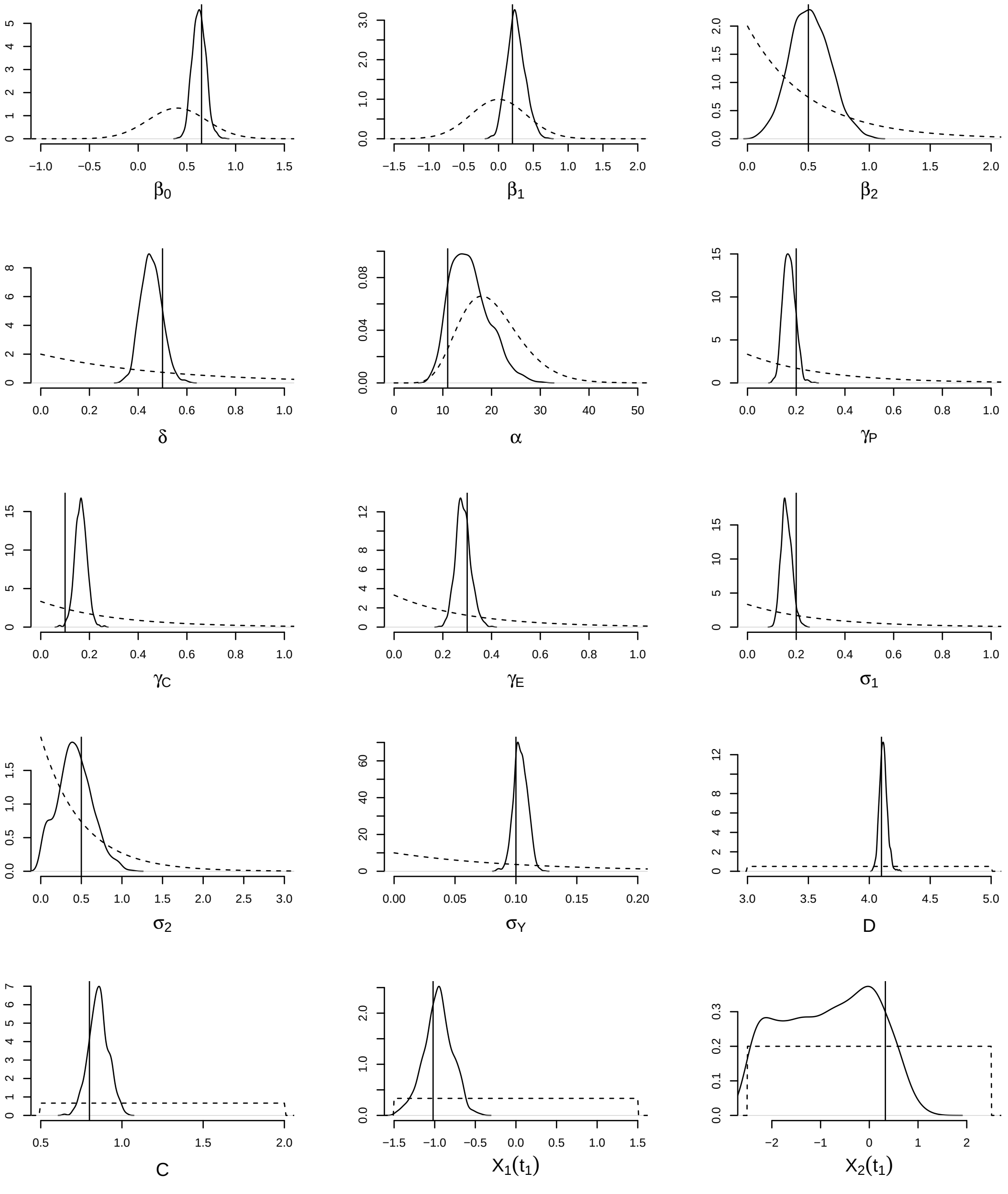

**Figure 3.6:** Marginal posterior distributions of the parameters of CR14-a, obtained using SMC$^2$ in the simulation study. Vertical lines indicate the values used to generate the data. Dashed lines show the prior distributions. These distributions strongly resemble those obtained using PMCMC, shown in Figure 3.5.

## 3.5 Chapter Summary

In this chapter we introduced ABC and particle filter methods for sampling from a posterior distribution when the likelihood term is intractable. Despite being presented as two separate entities, the methodologies are related. The proposal distributions in the particle filter can be simulations from the model, as in ABC methods. With a single particle these algorithms become ABC approaches with a Gaussian error distribution, as discussed in Section 3.3.3.

ABC methods were introduced from their inception to recent developments. ABC is one of the most useful tools for performing inference in models where the likelihood function is intractable and has applications in a wide array of fields. Developments such as ABC-MCMC, ABC-SMC, and post-processing techniques have improved the efficiency of ABC dramatically. However, some important issues still need to be addressed. Both the choice of tolerance, $\epsilon$, and distance metric, $\rho$, affect the posterior approximation in complex ways. This problem is amplified with the use of summary statistics, as it is often the case that sufficiency can not be determined, and it is difficult to assess if the summary statistics capture all of the important aspects of the data, or if valuable information is being discarded. How the various choices impact the ABC approximation remains an open question. At this point, there is also no satisfying way to automate the construction of useful summary statistics, but this is an active topic in the literature [76, 77].

PMCMC and SMC$^2$ were introduced as two recent methods that make use of the unbiased likelihood estimate of the particle filter. Unlike ABC approaches, these methods target the correct posterior distribution, but this comes at a potentially greater computational expense. From our simulation study it seems that SMC$^2$ has a greater computational cost than PMCMC. However, one benefit of SMC$^2$ is that it is easy to design adaptive proposal distributions for both the parameter particles and state particles. Given the large number of phenomenological models of the glacial-interglacial cycle, and the number of available datasets from sediment cores, SMC$^2$ has the advantage that it can be applied with little user input. PMCMC on the other hand would require tuning for each model and each dataset.

It should be stressed that the aim in this chapter was to demonstrate how ABC and particle filter based methods can be applied to parameter estimation problems, with a focus on the models discussed in Chapter 2. As such, this is by no means a complete account of ABC and particle methods. Thorough discussions can be found in the references given throughout this chapter. Additionally, recent reviews are available for both ABC [44] and particle methods [78].

A comparison of the marginal posterior distributions of the parameters, obtained using ABC rejection, ABC-ASMC, and SMC$^2$, is shown in Figure 3.7. There is a visible benefit to using ABC-ASMC over ABC rejection, as reaching a lower tolerance causes the posterior distributions to narrow in on the true parameters. However, ABC performs poorly in comparison to SMC$^2$. The loss of information due to using the ABC approximation gives poor approximations to the posterior distributions. The choice of algorithm therefore depends on the problem. The comparative benefits of ABC methods are that they are easy to implement, and are not restricted to performing inference on SSMs. It should also be stressed that the particle filter methods perform less well when only inefficient proposal distributions, such as simulations from the model, are available. In such cases, either the number of particles, or the Markov chain length, will need to be increased, adding computational expense. To conclude, in situations in which only approximate samples of the posterior distribution are required it will be much quicker to design and implement an ABC algorithm than it will a particle filter algorithm. On the other hand, if samples from the correct posterior distribution are required then particle filter methods should be used.

This chapter has shown how state of the art inference methods can be used to estimate parameters in phenomenological models using palaeoclimate data. The algorithms as presented, however, do not indicate which models are most supported by the data, nor do they formally test whether the model parameters have explanatory power. These can be considered as model comparison problems. In the next chapter we build on the inference methods introduced here, in order to perform model comparison experiments.

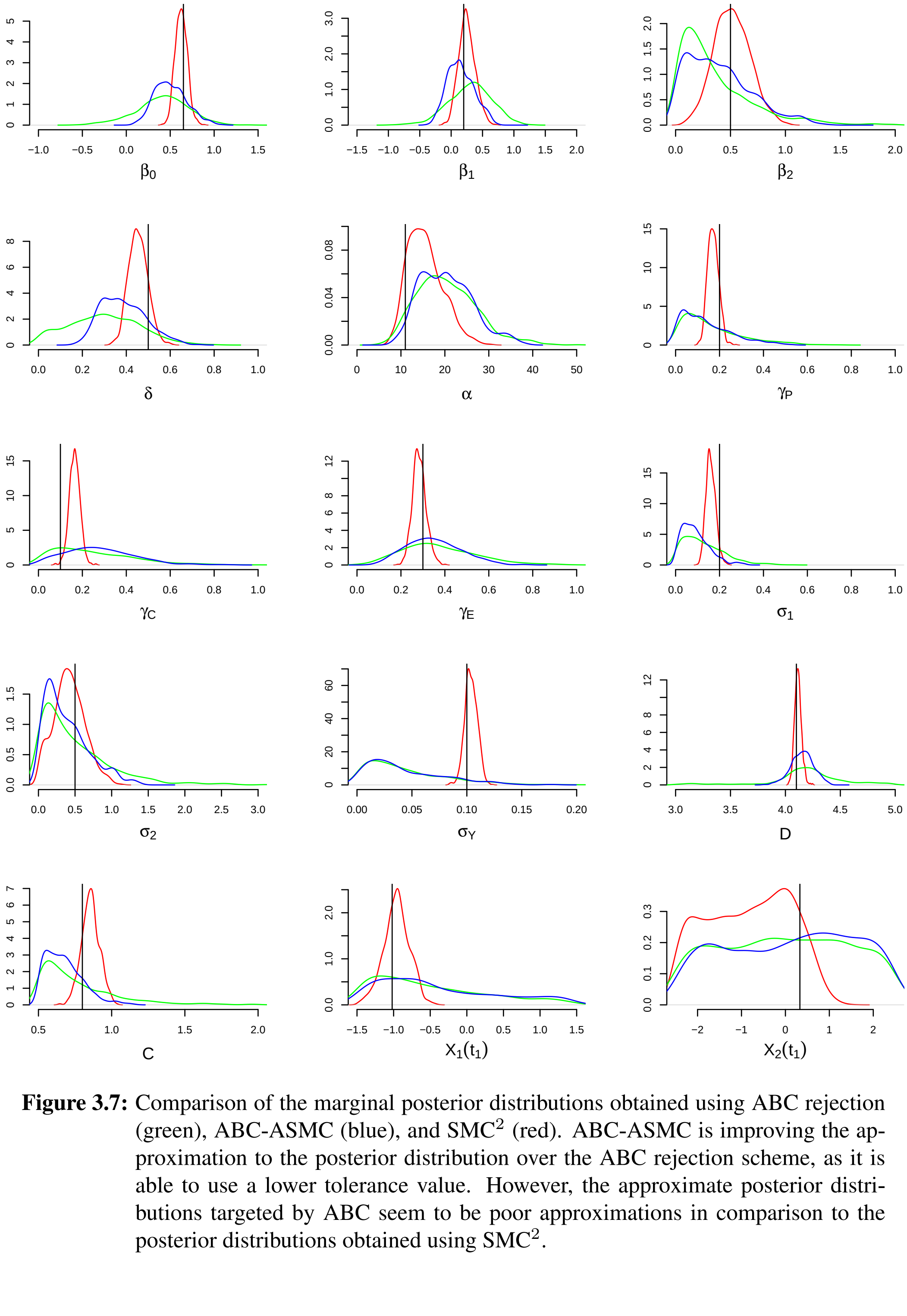

**Figure 3.7:** Comparison of the marginal posterior distributions obtained using ABC rejection (green), ABC-ASMC (blue), and SMC$^2$ (red). ABC-ASMC is improving the approximation to the posterior distribution over the ABC rejection scheme, as it is able to use a lower tolerance value. However, the approximate posterior distributions targeted by ABC seem to be poor approximations in comparison to the posterior distributions obtained using SMC$^2$.

CHAPTER 4

# Model Comparison

In this chapter we consider the problem of model comparison for phenomenological models of the glacial-interglacial cycle. In a Bayesian setting, model comparison is often done by evaluating the normalising constant in Equation 1.2.1, $\pi\left(\boldsymbol{Y}_{1:M}\right)$, termed the model evidence, for each model under consideration. The ratio of the model evidence terms between two models, termed the Bayes factor, indicates which model is more strongly supported by the data.

Evaluating the model evidence is a challenging problem, as it requires integration over the entire parameter space. As with the likelihood, the model evidence is intractable in our models of interest. Fortunately, the inference methods introduced in Chapter 3 can be extended to perform model comparison via estimation of the Bayes factors, without the need for a tractable likelihood.

We focus on SMC implementations, which provide estimates of normalising constants, such as the model evidence, with relative ease [79]. As discussed in Chapter 3, SMC approaches also allow automated calibration of many tuning parameters, such as proposal distributions. This is a useful feature when numerous models are under consideration, as it would take time to design efficient implementations of an algorithm for each model by hand.

The chapter is divided as follows. In Section 4.1 we discuss how model comparison is performed in a Bayesian setting. We introduce Bayes factors, and the different approaches that can be taken to evaluate them. In Section 4.2 we design a new simulation study to compare different model comparison approaches on synthetic data, where we know the true model. We compare two different approaches that are extensions to the inference methods introduced in Chapter 3. The first is an extension to the ABC-PRC algorithm that jointly targets all models under consideration, and the second utilises an estimate of the model evidence from SMC$^2$ for each model. In Section 4.4 we repeat the experiment on real-world data in order to test if the data more strongly support oscillators than steady-state models. Finally, in Section 4.5 we conclude the chapter with a discussion of the results.

## 4.1 Bayes Factors

In model comparison problems we have a collection of models $\mathcal{M}_l$, where $l = 1, ..., L$, which aim to explain a set of data, $\boldsymbol{Y}_{1:M}$. Each model, $\mathcal{M}_l$, has parameters $\boldsymbol{\theta}_l \in \mathbb{R}^{v_l}$. Prior probabilities, $\pi(\mathcal{M}_l)$, are assigned to the models, and prior distributions, $\pi(\boldsymbol{\theta}_l \mid \mathcal{M}_l)$, are assigned to the parameters for each model. The goal is to evaluate the posterior model probabilities, $\pi(\mathcal{M}_l \mid \boldsymbol{Y}_{1:M})$, showing which models are more strongly supported by the data. Using Bayes theorem, we aim to evaluate

$$\pi(\mathcal{M}_l \mid \boldsymbol{Y}_{1:M}) = \frac{\pi(\boldsymbol{Y}_{1:M} \mid \mathcal{M}_l)\,\pi(\mathcal{M}_l)}{\pi(\boldsymbol{Y}_{1:M})}, \tag{4.1.1}$$

where

$$\pi(\boldsymbol{Y}_{1:M} \mid \mathcal{M}_l) = \int \pi(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}_l, \mathcal{M}_l)\,\pi(\boldsymbol{\theta}_l \mid \mathcal{M}_l)\,d\boldsymbol{\theta}_l \tag{4.1.2}$$

is the model evidence for model $\mathcal{M}_l$. The normalising constant,

$$\pi(\boldsymbol{Y}_{1:M}) = \sum_{i=1}^{L} \pi(\boldsymbol{Y}_{1:M} \mid \mathcal{M}_l)\,\pi(\mathcal{M}_l), \tag{4.1.3}$$

is easily calculated if the model evidence for each model can be evaluated.

Models are often compared by evaluating the ratio of the posterior model probabilities between each pair of models. For example, if we have two competing models, $\mathcal{M}_1$ and $\mathcal{M}_2$, then we aim to calculate the ratio

$$\frac{\pi(\mathcal{M}_1 \mid \boldsymbol{Y}_{1:M})}{\pi(\mathcal{M}_2 \mid \boldsymbol{Y}_{1:M})} = \frac{\pi(\boldsymbol{Y}_{1:M} \mid \mathcal{M}_1)}{\pi(\boldsymbol{Y}_{1:M} \mid \mathcal{M}_2)}\frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)}, \tag{4.1.4}$$

which is the ratio of the prior model probabilities multiplied by

$$B_{12} = \frac{\pi(\boldsymbol{Y}_{1:M} \mid \mathcal{M}_1)}{\pi(\boldsymbol{Y}_{1:M} \mid \mathcal{M}_2)}. \tag{4.1.5}$$

The ratio of the model evidence terms, $B_{12}$, is termed the Bayes factor for model $\mathcal{M}_1$ against $\mathcal{M}_2$ [80]. When the models have equal prior probability, then the Bayes factor is equal to the ratio of the posterior model probabilities of the two models. Note that since the model evidence evaluation requires integration over the parameter space, Bayes factors tend to penalise the model with the larger number of parameters, particularly if the additional parameters add little explanatory power.

The Bayes factor summarises the support from the data in favour of one model over another, and in doing so allows competing scientific hypotheses, represented by different models, to be compared against one another. A common interpretation of Bayes factors is given in Table 4.1.

| $B_{12}$ | Evidence against $\mathcal{M}_2$ |
|---|---|
| 1 to 3 | Barely worth mentioning |
| 3 to 20 | Worth mentioning |
| 20 to 150 | Strong |
| $> 150$ | Decisive |

**Table 4.1:** Interpretation of Bayes factors [80]

This interpretation is intended as a rough guide, rather than to strictly categorise results as significant, or non-significant, as in frequentist hypothesis testing. An advantage of a Bayesian model comparison approach is that it indicates which models are more strongly supported by the data, whereas in hypothesis testing one model is considered as the null hypothesis, and evidence is only ever weighed against it. In hypothesis testing, a large $p$ value does not indicate that the null model is more strongly supported by the data, or that two models are equally well supported, but only that there is insufficient evidence to choose between them. Likewise, smaller $p$ values do not indicate that models are more strongly supported by the data, only that the null model lacks explanatory power in comparison. A Bayesian approach can also be used in cases where the models are not nested, unlike the commonly used frequentist likelihood ratio test. This is an essential property to be able to select between the numerous proposed phenomenological models in the literature.

There are three distinct approaches for estimating Bayes factors [79]. These are:

- All-in-one approach: Calculate the posterior model probabilities, $\pi\left(\mathcal{M}_l \mid \boldsymbol{Y}_{1:M}\right)$.
- Evidence calculation approach: Calculate the model evidence, $\pi\left(\boldsymbol{Y}_{1:M} \mid \mathcal{M}_l\right)$, for each model.
- Evidence ratio calculation approach: Directly calculate the Bayes factor for every pair of models.

Each approach is described in more detail below, along with some of the advantages and disadvantages of each method.

### All-in-One Approach

The all-in-one approach aims to sample from the joint posterior distribution

$$\pi\left(\mathcal{M}_l, \boldsymbol{\theta}_l \mid \boldsymbol{Y}_{1:M}\right) \propto \pi\left(\mathcal{M}_l\right) \pi\left(\boldsymbol{\theta}_l \mid \mathcal{M}_l\right) \pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}_l, \mathcal{M}_l\right). \tag{4.1.6}$$

Targeting the joint posterior distribution allows inter-model relationships to be exploited. For example, reversible-jump MCMC (RJMCMC) [81] uses the Metropolis Hastings algorithm on an extended space that admits the posterior distribution of both the models and model parameters. When inter-model relationships exist, say two models share one or more parameters, then more efficient proposal distributions can be designed to transition between the two models than if the relationship was ignored. The construction of efficient proposal distributions for RJMCMC is discussed in [82]. Likewise, exploiting inter-model relationships allows the design of efficient proposal distributions in many other Monte Carlo methods, such as SMC. Additionally, computation time is not spent on improbable models. This is beneficial in the sense that the algorithms sample from the posterior distribution more quickly, but with the downside that the posterior distributions of the parameters are poorly characterised for improbable models (the sample size will be too small to accurately resemble the posterior distribution). This can be problematic in SMC methods. If a model has low posterior probability in early iterations, then the accuracy of the posterior distribution of the parameters might remain poor in later iterations, even if the posterior probability of the model increases. For example, a small sample in an early iteration can lead to the design of poor importance sampling distributions when proposal distributions are automated, impacting the accuracy of later iterations.

### Evidence Calculation Approach

The evidence calculation approach evaluates the model evidence of each model, from which we can obtain the Bayes factors. This requires estimating the posterior distribution,

$$\pi\left(\boldsymbol{\theta}_l \mid \boldsymbol{Y}_{1:M}, \mathcal{M}_l\right) \propto \pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}_l, \mathcal{M}_l\right) \pi\left(\boldsymbol{\theta}_l \mid \mathcal{M}_l\right), \tag{4.1.7}$$

for each model, and then integrating over the parameter space to obtain the model evidence,

$$\pi\left(\boldsymbol{Y}_{1:M} \mid \mathcal{M}_l\right) = \int \pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}_l, \mathcal{M}_l\right) \pi\left(\boldsymbol{\theta}_l \mid \mathcal{M}_l\right) d\boldsymbol{\theta}_l. \tag{4.1.8}$$

The evidence calculation approach targets the posterior distribution of each model separately, and so inter-model relationships can not be exploited. However, a benefit of this is that proposal distributions are not required for transitions between models. Designing such proposal distributions can be difficult, and so this approach offers relatively simple implementations. A further benefit of this approach is that whenever a new model needs to be considered, then only the model evidence for the new model needs to be evaluated (in contrast to the all-in-one approach, where the entire extended distribution would need to be targeted). This can save computational expense in situations in which new models are frequently introduced. Additionally, the posterior distributions of the parameters are well characterised for every model (in that every model has a large number of samples, giving a good approximation to the posterior distribution), at the cost of additional computational expense compared to the all-in-one approach. Whether this is considered as an advantage or a disadvantage depends on the specific problem.

**Evidence Ratio Calculation Approach**

The final approach is to evaluate the Bayes factors directly for each pair of models. That is, for $i, j = 1, ..., L$, with $i \neq j$, evaluate

$$B_{ij} = \frac{\pi\left(\boldsymbol{Y}_{1:M} \mid \mathcal{M}_i\right)}{\pi\left(\boldsymbol{Y}_{1:M} \mid \mathcal{M}_j\right)}, \tag{4.1.9}$$

which is the Bayes factor in favour of model $\mathcal{M}_i$ against model $\mathcal{M}_j$. This approach allows inter-model relationships to be exploited, as in the all-in-one approach, but the proposal distributions are only required to support transitions between the two models under evaluation, making the proposal distributions easier to design. When a new model is considered, the Bayes factor needs to be evaluated in favour of the new model against each of the existing models. Hence, adding a new model will have a lower computational cost than the all-in-one approach, but a greater cost than evaluating each of the model evidence terms individually. As with the evidence calculation approach, the posterior distributions of the parameters are well characterised for every model.

### 4.1.1 Model Comparison with Intractable Likelihoods

The methods used to evaluate the Bayes factors that are described in Section 4.1 each require the likelihood , $\pi\left(\boldsymbol{Y}_{1:M} \mid \boldsymbol{\theta}_l, \mathcal{M}_l\right)$, and the model evidence, $\pi\left(\boldsymbol{Y}_{1:M} \mid \mathcal{M}_l\right)$. For phenomenological models of the glacial-interglacial cycle, both of these terms are intractable. Performing model comparison with these models requires inference methods that either simulate from the model, or approximate the likelihood, such as the methods introduced in Chapter 3. Both ABC and particle filter methods admit extensions to perform model comparison. We study two extensions in this chapter: An extension to ABC-PRC, discussed in Section 4.3.1, and an extension to SMC$^2$, discussed in Section 4.3.2. Before reviewing these methods, we introduce a simulation study with which the performances of the model comparison approaches will be compared.

## 4.2 Simulation Study

In this chapter we extend the inference approaches introduced in Chapter 3 to perform model comparison. As in Chapter 3, we design a simulation study to assess the accuracy and performance of the proposed inference methods. We use the same simulated dataset from CR14-a, as discussed in Section 3.2. The models under consideration are the three oscillators described in Section 2.3.2 (which includes the true model), as well as two steady-state models [41]. Here, steady-state refers to the deterministic ($\boldsymbol{\Sigma}_X = \mathbf{0}$), unforced ($\gamma_E = \gamma_P = \gamma_C = 0$), component of the model. A steady-state is a value of $\boldsymbol{X}$ for which $\frac{d\boldsymbol{X}}{dt} = \mathbf{0}$, so that the system remains at the steady-state unless perturbed, by stochastic perturbations or a forcing function, for example. A steady-state is called a stable steady-state if the system is attracted to the steady-state, and an unstable steady-state if it is non-attracting. Below, we consider a model with a single stable steady-state, and a model with two stable steady-states [41].

**Energy Balance Model (EBM)**

$$dX_1 = -\left(\beta_0 + \beta_1 X_1 + I(\gamma_P, \gamma_C, \gamma_E)\right) dt + \sigma_1 dW_1 \tag{4.2.1}$$

EBM has a single state variable, $X_1$, representing ice volume. Considering the deterministic ($\sigma_1 = 0$), unforced ($\gamma_E = \gamma_P = \gamma_C = 0$) case, the system has a single steady-state at $X_1 = -\frac{\beta_0}{\beta_1}$, which is stable as long as $\beta_1 > 0$. The system is constantly perturbed from the steady-state by the astronomical forcing and the Brownian motion. The astronomical forcing promotes ice growth at low values of insolation, and ice reduction at large values of insolation. EBM has 6 tunable parameters, $\boldsymbol{\theta} = (\beta_0, \beta_1, \gamma_P, \gamma_C, \gamma_E, \sigma_1)^T$, which is fewer than the oscillators described in Section 2.3.2.

**Two Stable Steady-State Model (TSS)**

$$dX_1 = -\left(\beta_1 X_1 + \beta_2 \left(X_1^3 - X_1\right) + I(\gamma_P, \gamma_C, \gamma_E)\right) dt + \sigma_1 dW_1 \tag{4.2.2}$$

As with EBM, there is only a single state variable, $X_1$, representing ice volume. However, in the absence of the astronomical forcing ($\gamma_E = \gamma_P = \gamma_C = 0$) and stochastic perturbations ($\sigma_1 = 0$), there are three steady-states, which are $X_1 = 0$ and $X_1 = \pm\sqrt{\frac{\beta_1+\beta_2}{\beta_2}}$. The steady-state at $X_1 = 0$ is an unstable steady-state, and the two non-zero solutions are stable steady-states as long as $\beta_2 > 0$ and $\beta_1 > -\beta_2$. If the system is at $X_1 = 0$, a positive perturbation will cause the state of the system to be attracted to the positive steady-state, and vice versa. The astronomical forcing promotes crossing towards negative values of $X_1$ (ice reduction) when the insolation is high, and towards positive values (ice growth) when the insolation is low. TSS has 6 tunable parameters, $\boldsymbol{\theta} = (\beta_0, \beta_2, \gamma_P, \gamma_C, \gamma_E, \sigma_1)^T$.

The prior distributions for each of the models are given in Table 4.2. We assess the accuracy of the model comparison methods by their ability to favour the correct model, and to recover the true parameters in the correct model (as in Chapter 3).

| Parameter | 1. CR14-a | 2. CR14-b | 3. CR14-c | 4. TSS | 5. EBM |
|---|---|---|---|---|---|
| $\beta_0$ | $\mathcal{N}\left(0.4, 0.3^2\right)$ | $\mathcal{N}\left(0, 0.4^2\right)$ | $\mathcal{N}\left(0, 0.4^2\right)$ | | $\mathcal{N}\left(0, 0.4^2\right)$ |
| $\beta_1$ | $\mathcal{N}\left(0, 0.4^2\right)$ | $\mathcal{N}\left(0, 0.4^2\right)$ | $\mathcal{N}\left(0, 0.4^2\right)$ | $\mathcal{N}\left(0, 0.3^2\right)$ | $exp\left(1/0.4\right)$ |
| $\beta_2$ | $exp\left(1/0.5\right)$ | $exp\left(1/0.5\right)$ | $exp\left(1/0.5\right)$ | $exp\left(1/0.5\right)$ | |
| $\delta$ | $exp\left(1/0.5\right)$ | $\Gamma\left(10, 0.1\right)$ | $\Gamma\left(10, 0.1\right)$ | | |
| $\alpha$ | $\Gamma\left(10, 2\right)$ | $exp\left(1/0.5\right)$ | $exp\left(1/0.5\right)$ | | |
| $\kappa_0$ | | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ | | |
| $\kappa_1$ | | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ | | |
| $\gamma_P$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ |
| $\gamma_C$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ |
| $\gamma_E$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ |
| $\sigma_1$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ | $exp\left(1/0.3\right)$ |
| $\sigma_2$ | $exp\left(1/0.5\right)$ | $exp\left(1/0.5\right)$ | $exp\left(1/0.5\right)$ | | |
| $\sigma_y$ | $exp\left(1/0.1\right)$ | $exp\left(1/0.1\right)$ | $exp\left(1/0.1\right)$ | $exp\left(1/0.1\right)$ | $exp\left(1/0.1\right)$ |
| $D$ | $\mathcal{U}\left(3, 5\right)$ | $\mathcal{U}\left(3, 5\right)$ | $\mathcal{U}\left(3, 5\right)$ | $\mathcal{U}\left(3, 5\right)$ | $\mathcal{U}\left(2.5, 4.5\right)$ |
| $C$ | $\mathcal{U}\left(0.5, 2\right)$ | $\mathcal{U}\left(0.5, 2\right)$ | $\mathcal{U}\left(0.5, 2\right)$ | $\mathcal{U}\left(0.5, 2\right)$ | $\mathcal{U}\left(0.5, 2\right)$ |
| $X_1\left(t_1\right)$ | $\mathcal{U}\left(-1.5, 1.5\right)$ | $\mathcal{U}\left(-1.5, 1.5\right)$ | $\mathcal{U}\left(-1.5, 1.5\right)$ | $\mathcal{U}\left(-1.5, 1.5\right)$ | $\mathcal{U}\left(-1.5, 1.5\right)$ |
| $X_2\left(t_1\right)$ | $\mathcal{U}\left(-2.5, 2.5\right)$ | $\mathcal{U}\left(-2.5, 2.5\right)$ | $\mathcal{U}\left(-2.5, 2.5\right)$ | | |

**Table 4.2:** List of prior distributions for the model comparison experiment.

To simplify notation when presenting the Bayes factors, we number the models as follows:

| | | |
|---|---|---|
| 1. CR14-a | 2. CR14-b | 3. CR14-c |
| 4. TSS | 5. EBM | |

## 4.3 Model Comparison Methods

We now present extensions to two of the algorithms introduced in Chapter 3, in order to perform model comparison. The first is an extension to the ABC-PRC algorithm [60], which follows the all-in-one approach to estimating Bayes factors, and the second extends SMC$^2$ [74], which evaluates unbiased estimates of the model evidence for each model. These methods are compared in the simulation study designed in Section 4.2. Note that many more extensions are available, but with these two approaches we can compare the advantages and disadvantages of using an all-in-one approach over a model evidence calculation approach, and of using the ABC approximation over a likelihood estimation scheme.

### 4.3.1 ABC-PRC

The ABC-PRC algorithm [60] can be extended to perform model comparison by including a discrete model-parameter, $\mathcal{M}_l$, where $l = 1, ..., L$ . Following the all-in-one approach, the algorithm targets the extended posterior distribution $\pi_\epsilon\left(\mathcal{M}_l, \boldsymbol{\theta}_l, \tilde{\boldsymbol{Y}}_{1:M,l} \mid \boldsymbol{Y}_{1:M}\right)$, where $\pi_\epsilon$ represents the ABC posterior distribution discussed in Section 3.3.

The proposal distribution is extended to include the model-parameter, $\mathcal{M}_l$. Specifically, $\mathcal{M}_l$ is sampled from the prior model probabilities, $\pi(\mathcal{M}_l)$, in every proposal. If a model has zero weight then it is excluded, and only the remaining models are considered. In the first iteration of the algorithm, parameters are then sampled from the prior distribution, $\pi(\boldsymbol{\theta}_l \mid \mathcal{M}_l)$. In later iterations the parameters are sampled from the particles associated with the proposed model, and perturbed according to a Markov kernel with transition density $K_{l,s}(\boldsymbol{\theta}_{l,s} \mid \boldsymbol{\theta}_{l,s-1})$, which depends on the proposed model, $\mathcal{M}_l$, and the iteration number, $s$. Finally, simulated values are generated and compared with observations. The importance weights are the same as in the implementation presented in Section 3.3. An ABC approximation to the posterior model probabilities, $\pi_\epsilon(\mathcal{M}_l \mid \boldsymbol{Y}_{1:M})$, is obtained in each iteration by summing the normalised importance weights of the particles associated with each model.

The pseudocode is presented in Algorithm 4.1. The notation in the algorithm follows the notation introduced in Section 3.3. Additionally, we introduce 'counters', $N_{l,s}$, which track how many particles are associated with model $\mathcal{M}_l$ in population $s$.

As with the implementation of the ABC-PRC algorithm presented in Section 3.3, certain choices can be automated. For example, the perturbation kernels $K_{l,s}(\boldsymbol{\theta}_{l,s} \mid \boldsymbol{\theta}_{l,s-1})$ can be chosen based on the current particles for each model, and the tolerance scheme can be chosen one iteration ahead until the computational cost becomes too high.

The more strongly favoured models will typically contain the highest number of particles, giving a good approximation of the posterior distributions of the parameters. For other models, the ABC-PRC algorithm needs to be run for every model independently to obtain the posterior distributions. It is also important to monitor the number of particles associated with each model in every iteration. If there is strong evidence against models in early iterations, then the number of particles will be small, and the resulting approximate posterior distribution might be a poor approximation to the true posterior distribution in later iterations.

An alternative implementation was presented in [83], in which the model-parameter is sampled from the ABC posterior model probabilities, $\pi_{\epsilon_s}(\mathcal{M}_l \mid \boldsymbol{Y}_{1:M})$, in each iteration, rather than the prior probabilities. This can give improved performance when models have low posterior probabilities in every iteration, as more particles are proposed for the more strongly supported models. However, in cases where models have low posterior probability in early iterations, but large posterior probability in later iterations, the alternative proposal distribution might be less efficient.

**Algorithm 4.1** ABC-PRC sampling algorithm targeting $\pi_\epsilon\left(\mathcal{M}_l, \boldsymbol{\theta}_l, \tilde{\boldsymbol{Y}}_{1:M,l} \mid \boldsymbol{Y}_{1:M}\right)$.

```
Set $N_{1,1}, ..., N_{L,1} = 0$.
for $n = 1, ..., N$ do
  Sample $\mathcal{M}_l$ from $\pi(\mathcal{M}_l)$, and set $N_{l,1} = N_{l,1} + 1$.
  Sample $\boldsymbol{\theta}_{l,1}^{(N_{l,1})}$ from the prior distribution, $\pi(\boldsymbol{\theta}_l \mid \mathcal{M}_l)$.
  Simulate values $\tilde{\boldsymbol{Y}}_{1:M,l,1}^{(N_{l,1})}$ from model $\mathcal{M}_l$ using parameter $\boldsymbol{\theta}_{l,1}^{(N_{l,1})}$.
end for
Set the importance weights. For $l = 1, ..., L$ and $j = 1, ..., N_{l,1}$ set
    $$W_{l,1}^{(j)} = \frac{1}{N_{l,1}}.$$
for $s = 2, ..., S$ do
  Set $N_{1,s}, ..., N_{L,s} = 0$.
  while $n \leq N$ do
    Sample $\mathcal{M}_l$ from $\pi(\mathcal{M}_l)$.
    if $N_{l,s-1} > 0$ then
      Sample $\boldsymbol{\theta}^{**}$ from the previous population, $\boldsymbol{\theta}_{l,s-1}^{1:N_{l,s-1}}$, according to weights $W_{l,s-1}^{1:N_{l,s-1}}$.
      Sample $\boldsymbol{\theta}^{*}$ from the transition density $K_{l,s}(\boldsymbol{\theta}_{l,s} \mid \boldsymbol{\theta}^{**})$.
      if $\pi(\boldsymbol{\theta}^{*} \mid \mathcal{M}_l) > 0$ then
        Simulate values $\tilde{\boldsymbol{Y}}_{1:M}^{*}$ from model $\mathcal{M}_l$ using parameter $\boldsymbol{\theta}^{*}$.
        if $\rho\left(\tilde{\boldsymbol{Y}}_{1:M}^{*}, \tilde{\boldsymbol{Y}}_{1:M}\right) \leq \epsilon_s$ then
          Set $N_{l,s} = N_{l,s} + 1$, $\boldsymbol{\theta}_{l,s}^{(N_{l,s})} = \boldsymbol{\theta}^{*}$, and $\tilde{\boldsymbol{Y}}_{1:M,l,s}^{(N_{l,s})} = \tilde{\boldsymbol{Y}}_{1:M}^{*}$.
          Set the importance weight
              $$w_{l,s}^{(N_{l,s})} = \frac{\pi\left(\boldsymbol{\theta}_{l,s}^{(N_{l,s})}\right)}{\sum_{i=1}^{N_{l,s-1}} W_{l,s-1}^{(i)} K_{l,s}\left(\boldsymbol{\theta}_{l,s}^{(N_{l,s})} \mid \boldsymbol{\theta}_{l,s-1}^{(i)}\right)}.$$
          Set $n = n + 1$.
        end if
      end if
    end if
  end while
  for $l = 1, ..., L$ do
    Evaluate
        $$\pi_{\epsilon_s}(\mathcal{M}_l \mid \boldsymbol{Y}_{1:M}) = \frac{\sum_{j=1}^{N_{l,s}} w_{l,s}^{(j)}}{\sum_{i=1}^{L} \sum_{j=1}^{N_{i,s}} w_{i,s}^{(j)}}.$$
  end for
  Normalise the weights. For $l = 1, ..., L$ and $j = 1, ..., N_{l,s}$ set
      $$W_{l,s}^{(j)} = \frac{w_{l,s}^{(j)}}{\sum_{i=1}^{N_{l,s}} w_{l,s}^{(i)}}.$$
end for
```

| Iteration | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ | 320 | 220 | 180 | 160 | 145 | 130 | 125 | 120 | 115 |
| Simulations $\left(\times 10^3\right)$ | 23.8 | 25.0 | 33.8 | 45.0 | 136 | 2408 | 2315 | 3456 | 18422 |

**Table 4.3:** Tolerance scheme used in ABC-PRC in the model comparison simulation study, with the number of simulations required at each tolerance level.

### Results on the Simulation Study Data

We run the ABC-PRC algorithm with 5000 particles, where the tolerance scheme is shown in Table 4.3. The Markov kernels are multivariate Gaussian random walks, with zero mean and variance equal to twice the sample variance of the current sample for each model. The prior model probabilities are uniformly distributed.

The model posterior probabilities are shown in Figure 4.1. Recall that the true model is CR14-a, and so we would hope that the Bayes factors favour this model. There is a lot of variation in the Bayes factors in early iterations, but by the final population the correct model is preferred with the Bayes factors $\beta_{12} = 2.35$, $\beta_{13} = 25.3$, $\beta_{14} = 6.27$, $\beta_{15} = 17.6$. Hence, CR14-a is only slightly more preferred than CR14-b. A possible reason for this is that both models are oscillators in which the ice volume responds linearly to the astronomical forcing, making them the most similar models. The forcing in CR14-c is in the threshold function, and so ice volume responds nonlinearly to the forcing. TSS is slightly preferred over EBM.

The parameter marginal posterior distributions of CR14-a are shown in Figure 4.2. The resulting posterior distributions strongly resemble the posterior distributions shown in Figure 3.3. There are 2054 particles associated with CR14-a in the final population, and so the marginal posterior distributions of the parameters should be well characterised. This number fell to 304 in the third iteration, but this seems to have had little impact on the final distribution.

The total number of simulations in each iteration is given in Table 4.3. Naively, since we are using five times the number of particles as in the simulation study in Chapter 3, we might expect to need five times the number of simulations. However, the number of simulations required is influenced by a number of factors. For example, including models that are more likely to generate simulated values close to the observations means that the acceptance rate should be improved. On the other hand, including models that are less likely to generate simulated values close to the observations lowers the acceptance rate. In the final iteration, all of the additional models are less likely to generate trajectories that are close to the data than the true model, increasing the number of simulations required. Additionally, designing the proposal distributions on an existing sample of particles can lead to poorly designed proposal distributions if the sample is small. For example, if there are too few particles in regions of high posterior probability density, then exploration of the parameter space may be slowed down.

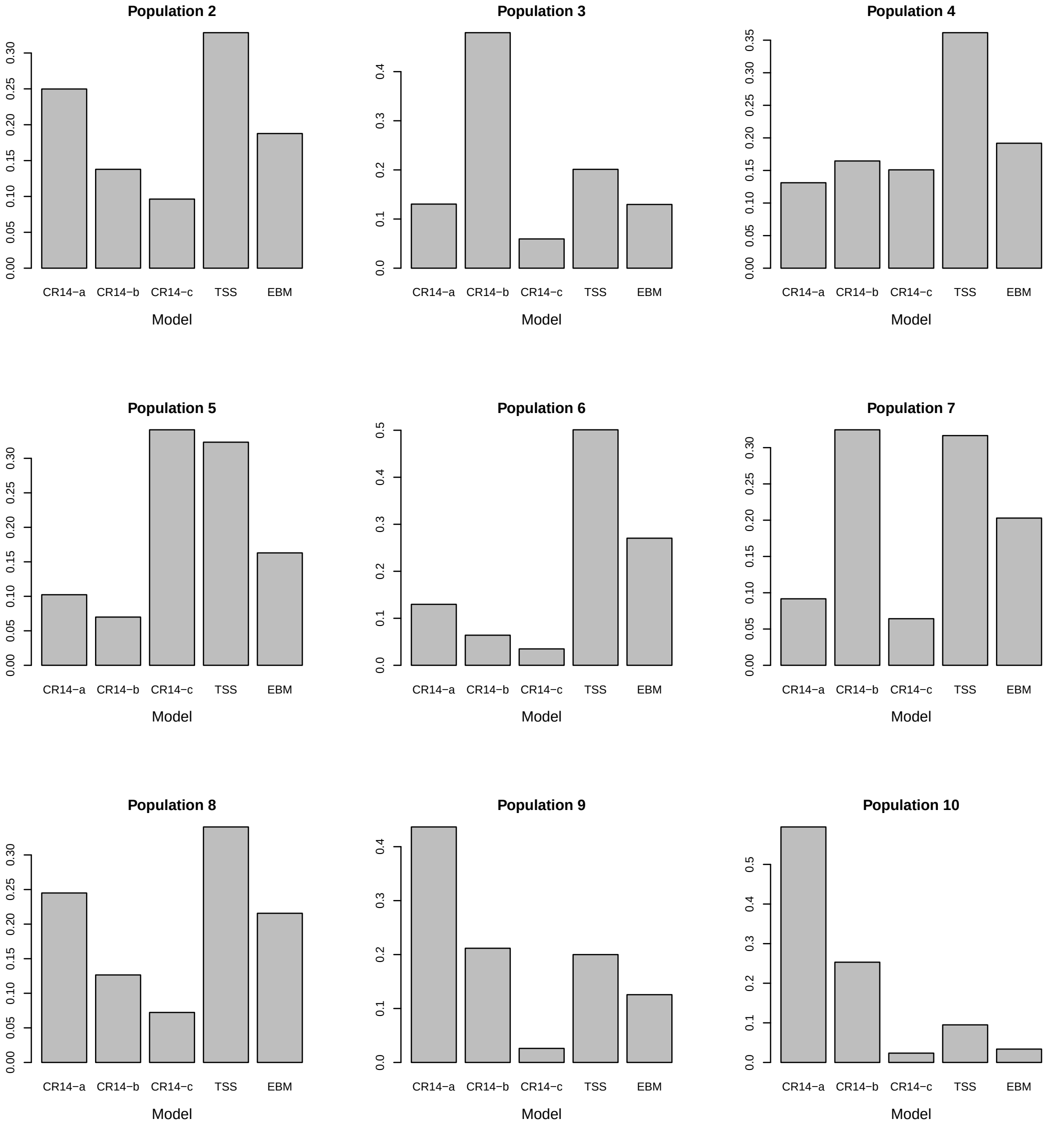

**Figure 4.1:** Posterior model probabilities for the model comparison simulation study obtained using ABC-PRC. CR14-a is the correct model. These results show that the correct model is favoured at lower tolerances. However, the posterior probability for CR14-a drops to low values in early iterations of the algorithm, which will impact the accuracy of the approximate posterior distributions of the parameters. CR14-b is second most probable model, and most resembles CR14-a. The steady-state models have a high posterior probability in early iterations, but a very low posterior probability in the final iteration, showing that they rarely simulate values very close to the observed data. Repeated experiments all favour the correct model in the final iteration, but favour different models in earlier iterations.

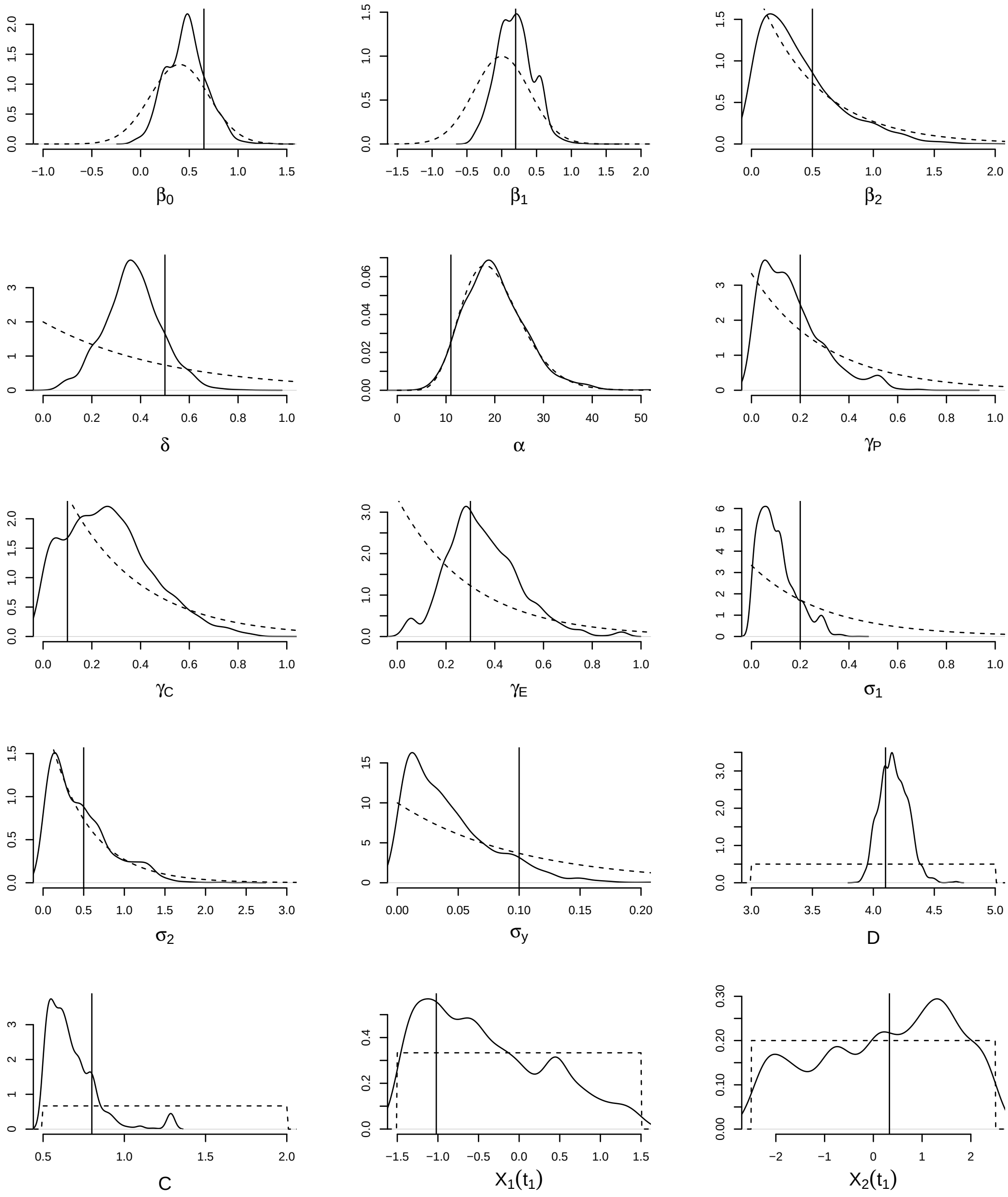

**Figure 4.2:** Marginal posterior distributions of the parameters of CR14-a, obtained using ABC-PRC in the model comparison simulation study. Vertical lines indicate the values used to generate the data. Dashed lines show the prior distributions. These results are consistent with running the ABC-PRC algorithm using only CR14-a, as shown in Figure 3.3.

### 4.3.2 SMC$^2$

As discussed in Section 3.4, as each observation is assimilated the particle filter provides an unbiased estimate of $\pi\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta}\right)$. Using this property, an unbiased estimate of the model evidence can be obtained in SMC$^2$ [74]. The model evidence can be decomposed as

$$\pi\left(\boldsymbol{Y}_{1:M}\right) = \pi\left(\boldsymbol{Y}_1\right) \prod_{m=2}^{M} \pi\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}\right). \tag{4.3.1}$$

SMC$^2$ gives an unbiased estimate, $\hat{\pi}\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}\right)$, of each component of the product by averaging over the parameter particles in each iteration, i.e.

$$\hat{\pi}\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}\right) = \sum_{i=1}^{N_\theta} W_m^{(i)} \hat{\pi}\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta}^{(i)}\right), \tag{4.3.2}$$

where $W_m^{(i)}$ are the normalised importance weights. The product of these terms provides an unbiased estimate of the model evidence [74]. The pseudocode is presented in Algorithm 4.2, where the notations are described in Section 3.4.

#### Results on the Simulation Study Data

We run the SMC$^2$ algorithm with $N_\theta = 1000$ parameter particles and $N_X = 1000$ state particles for every model. A PMCMC chain length of 10 is used in the resampling steps, and the parameter proposal distribution is independent Gaussian, with mean and variance equal to the mean and variance of the resampled parameters.

The model evidence for each model is given in Table 4.4. The true model is preferred, with the Bayes factors $\beta_{12} = \mathcal{O}(10^4)$, $\beta_{13} = \mathcal{O}(10^{27})$, $\beta_{14} = \mathcal{O}(10^{26})$, $\beta_{15} = \mathcal{O}(10^{28})$. Rerunning the algorithm 20 times on CR14-a shows that the model evidence varies by a factor of 10 due to Monte Carlo error (the sampling variability from using a finite sample size). Hence, we need to update the interpretation given in Table 4.1. In this case we say that, as a rule of thumb, a result is worth mentioning if the Bayes factor is over $10^3$. With this interpretation the correct model is the most strongly favoured by the data. As in the ABC-PRC results, CR14-b is the next favoured model. Taking the Monte Carlo variability into consideration means that it is difficult to order the remaining models, which are potentially equally well supported by the data. These Bayes factors are much larger than those obtained by the ABC-PRC algorithm. The ABC-PRC algorithm ranks models by how likely they are to generate values close to the observations. At large tolerances, the incorrect models still have a reasonable chance of generating values close to the observations, which is reflected in the Bayes factors. On the other hand, SMC$^2$ estimates how likely the observations are to have come from the model under evaluation, and is able to

**Algorithm 4.2** SMC$^2$ algorithm targeting $\pi\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:M} \mid \boldsymbol{Y}_{1:M}\right)$.

```
for $n = 1, ..., N_\theta$ do
    Sample $\boldsymbol{\theta}^{(n)}$ from the prior distribution, $\pi\left(\boldsymbol{\theta}\right)$.
    Set the importance weight
        $$W_0^{(n)} = \frac{1}{N_\theta}.$$
end for
for $m = 1, ..., M$ do
    if ESS$< \frac{N_\theta}{2}$ then
        for $n = 1, ..., N_\theta$ do
            Sample $\boldsymbol{\theta}^{*(n)}$ and $\boldsymbol{X}_{1:m-1}^{*(1:N_X,n)}$ from $\boldsymbol{\theta}^{(1:N_\theta)}$ and $\boldsymbol{X}_{1:m-1}^{(1:N_X,1:N_\theta)}$, according to weights $W_{m-1}^{(1:N_\theta)}$.
            Sample $\boldsymbol{\theta}^{**(n)}$ and $\boldsymbol{X}_{1:m-1}^{**(1:N_X,n)}$ from a PMCMC algorithm initialised with $\boldsymbol{\theta}^{*(n)}$ and $\boldsymbol{X}_{1:m-1}^{*(1:N_X,n)}$, and targeting $\pi\left(\boldsymbol{\theta}, \boldsymbol{X}_{1:m-1} \mid \boldsymbol{Y}_{1:m-1}\right)$.
        end for
        Set $\boldsymbol{\theta}^{(1:N_\theta)} = \boldsymbol{\theta}^{**(1:N_\theta)}$ and $\boldsymbol{X}_{1:m-1}^{(1:N_X,1:N_\theta)} = \boldsymbol{X}_{1:m-1}^{**(1:N_X,1:N_\theta)}$.
        Set the importance weights. For $n = 1, ..., n_\theta$
            $$W_{m-1}^{(n)} = \frac{1}{N_\theta}.$$
    end if
    for $n = 1, ..., N_\theta$ do
        Sample $\boldsymbol{X}_{1:m}^{(1:N_X,n)}$ by performing iteration $m$ of the particle filter, and record estimates of $\hat{\pi}\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta}^{(n)}\right)$ and $\hat{\pi}\left(\boldsymbol{Y}_{1:m} \mid \boldsymbol{\theta}^{(n)}\right)$.
        Set the importance weights
            $$w_m^{(n)} = w_{m-1}^{(n)} \hat{\pi}\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta}^{(n)}\right).$$
    end for
    Normalise the weights. For $n = 1, ..., N_\theta$
        $$W_m^{(n)} = \frac{w_m^{(n)}}{\sum_{i=1}^{N_\theta} w_m^{(i)}}.$$
    Evaluate
        $$\hat{\pi}\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}\right) = \sum_{i=1}^{N_\theta} W_m^{(i)} \hat{\pi}\left(\boldsymbol{Y}_m \mid \boldsymbol{Y}_{1:m-1}, \boldsymbol{\theta}^{(i)}\right).$$
end for
```

| Model | Evidence |
|---|---|
| CR14-a | $9.1 \times 10^{84}$ |
| CR14-b | $2.1 \times 10^{80}$ |
| CR14-c | $3.5 \times 10^{57}$ |
| TSS | $7.8 \times 10^{58}$ |
| EBM | $2.3 \times 10^{56}$ |

**Table 4.4:** The model evidence for each model obtained using SMC$^2$ in the simulation study. The favoured model is shown in red. In this case, we favour the correct model.

give much stronger results.

The number of simulations for each model was 120 million for CR14-a, 134 million for CR14-b, 131 million for CR14-c, 101 million for TSS, and 93 million for EBM, which seems to scale with the number of parameters in each model. The total, 579 million, far exceeds the ~27 million required in our ABC-PRC implementation. However, all of the marignal posterior distributions of the parameters are well characterised (with over 900 distinct particles in the final iteration for each model), and the correct posterior distributions are being targeted, rather than ABC approximations. Since the marginal posterior distributions of the parameters for CR14-a are shown in Figure 3.6, this figure is not reproduced here.

## 4.4 An Experiment on Real-World Data

In this chapter we have compared different model comparison approaches in a simulation study, and shown that the true model is favoured. When using real-world data, no model will be the true model, but the data may more strongly support some models over others. Here, we repeat our experiments on a real-world dataset to highlight the additional challenges of performing inference on real-world data. Since many of the phenomenological models of the glacial-interglacial cycle in the literature are oscillators, we are also interested in testing whether real-world data more strongly support oscillators over steady-state models.

There are hundreds of possible datasets derived from sediment cores from which we can choose. We can also choose between stacks (averages over multiple datasets) and the individual datasets. The observation times of a dataset depend on the age model used to convert observation depths to time. In some cases datasets have been assigned multiple age models, prompting an additional choice. A discussion of different datasets, stacks, and age models is included in the next chapter. For now we use ODP677 [84], which has been dated without astronomical tuning assumptions [4]. ODP677 includes a marker for the Brunhes-Matuyama (BM) reversal, which occurred around 780 kyr BP (±2 kyr). There are 363 observations between the BM reversal and the present, which is similar in number to our simulation study data.

| Iteration | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ | 320 | 220 | 180 | 160 | 145 | 130 | 125 | 120 | 115 |
| Simulations $\left(\times 10^3\right)$ | 24.8 | 27.5 | 30.5 | 36.6 | 60.8 | 269 | 686 | 4311 | 17530 |

**Table 4.5:** Tolerance scheme used in ABC-PRC for ODP677, with the number of simulations required at each tolerance level.

### 4.4.1 ABC-PRC Results on ODP677

The ABC-PRC algorithm is implemented on ODP677 with 5000 particles. The sequence of tolerances used in the simulation study seems to work well on ODP677, and is given in Table 4.5. This is likely a consequence of choosing parameters for the simulation study based on previous studies. The Markov kernels are again taken to be multivariate Gaussian random walks, with zero mean and twice the sample variance of the current sample for each model. The prior model probabilities are uniformly distributed.

The model posterior probabilities are shown in Figure 4.3. TSS has the largest posterior probability in early iterations, but CR14-c becomes the most supported model in the later iterations. The Bayes factors for CR14-c are $\beta_{31} = 7.25$, $\beta_{32} = 11.2$, $\beta_{34} = 12.9$, $\beta_{35} = 65.6$. The Bayes factors indicate that the data support a nonlinear response to the astronomical forcing. CR14-a, CR14-b, and TSS seem to have equal support from the data. However, the Bayes factors vary between repeated experiments due to Monte Carlo error. The general trend between independent trials is large Bayes factors against EBM, but the favoured model varies between the alternative models.

For comparison, the parameter marginal posterior distributions of CR14-a are shown in Figure 4.4. Since the parameters used in the simulation study were based on previous trials, the posterior distributions are similar. The posterior distributions appear to have a lower variance than in the simulation study (particularly noticeable for $\delta$ and $\gamma_C$). This could be because a narrower range of parameters are likely to generate trajectories that are close to the data, or because of poor exploration of the parameter space. The number of particles associated with CR14-a fell to 297 in iteration four, and recovered to 1586 in the final iteration. This is similar to the simulation study, in which the posterior distributions were consistent with the parameter estimation results from the simulation study in chapter 3, suggesting that 300 particles can be sufficient to explore the parameter space. Additionally, repeating the experiment yields similar posterior distributions, and the discrepancy can be accounted for by Monte Carlo error. Hence, we conclude that the lower posterior variance is due to the data, rather than degeneracy.

The required number of simulations in each iteration are given in Table 4.5. The total number of simulations is comparable to the simulation study.

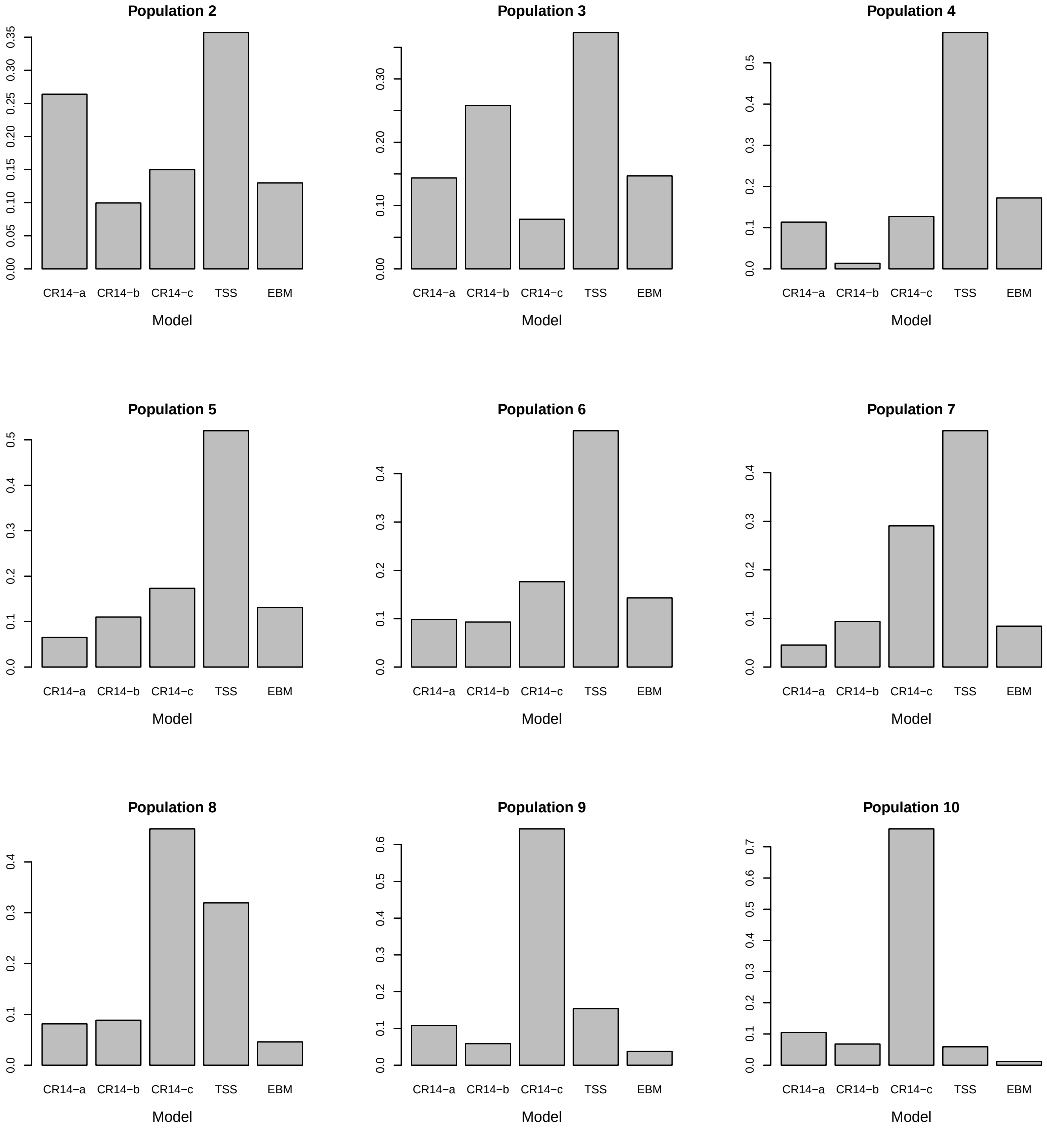

**Figure 4.3:** Posterior model probabilities for ODP677 obtained using ABC-PRC. These results show that CR14-c is favoured in the final iteration. The other oscillators have a low posterior probability. The fact that they have a low posterior probability in early iterations might mean that the parameter space is poorly explored, which could prevent them being favoured in later iterations. EBM has a low posterior probability throughout the algorithm, but TSS is favoured in early iterations. The results are not consistent when repeating the experiment. Any of the models, with the exception of EBM, can be favoured between trials, but usually by a small margin.

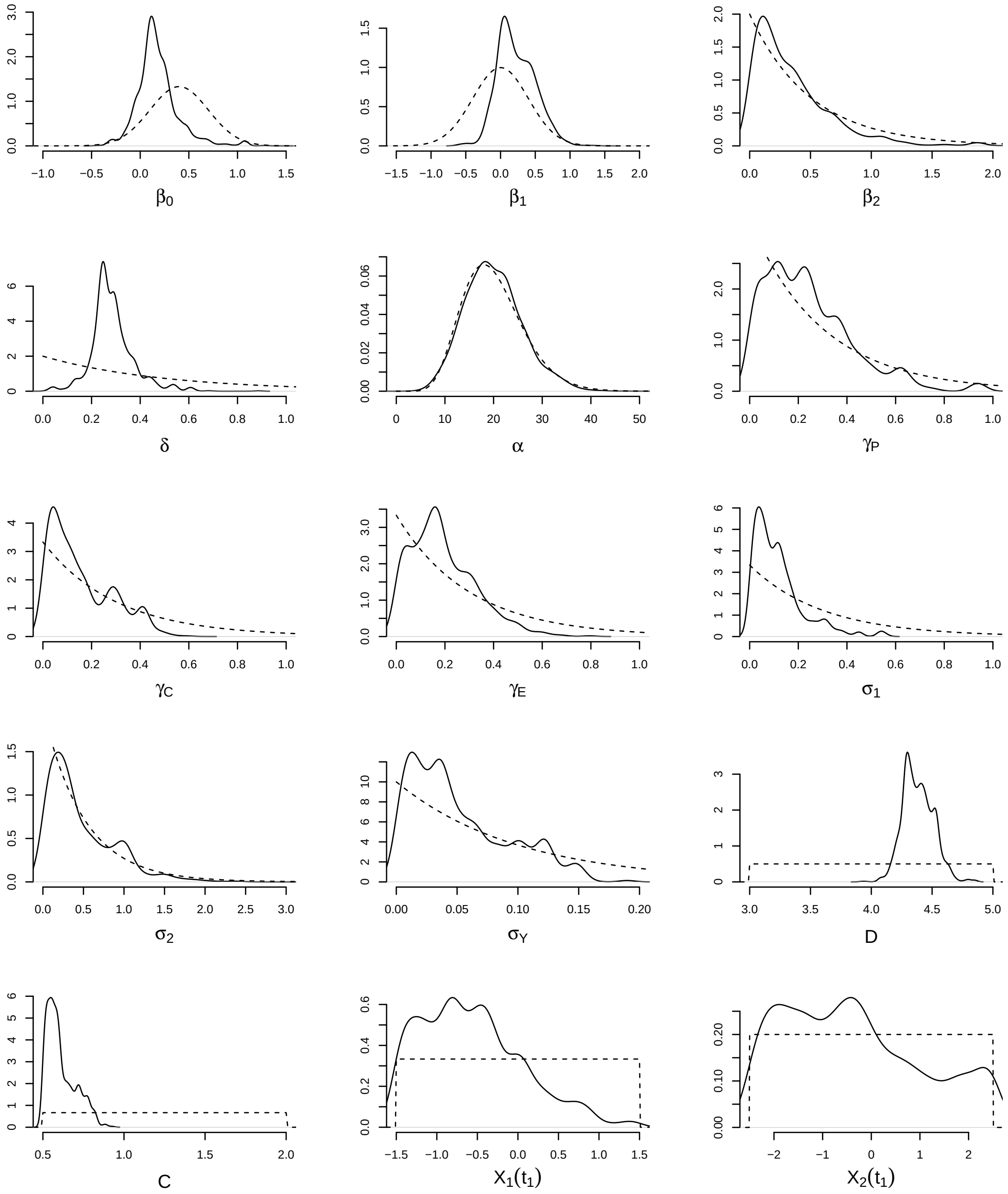

**Figure 4.4:** Marginal posterior distributions of the parameters of CR14-a, obtained using ABC-PRC for ODP677. Dashed lines show the prior distributions. The numerous local maxima indicate highly weighted particles, showing degeneracy. In comparison to the posterior distributions obtained in the simulation study, shown in Figure 4.2, many of the distributions appear to have a lower variance. Repeated trials yield consistent posterior distributions.

| Model | Evidence |
|---|---|
| CR14-a | $1.8 \times 10^{28}$ |
| CR14-b | $3.9 \times 10^{27}$ |
| CR14-c | $1.8 \times 10^{26}$ |
| TSS | $8.4 \times 10^{20}$ |
| EBM | $1.1 \times 10^{21}$ |

**Table 4.6:** The model evidence for each model obtained using SMC$^2$ for ODP677. The favoured models, with consideration of the Monte Carlo error, are shown in red. Each of the oscillators seem to be equally supported by the data.

### 4.4.2 SMC$^2$ Results on ODP677

The SMC$^2$ algorithm is implemented on ODP677 with $N_\theta = 1000$ parameter particles and $N_X = 1000$ state particles for every model. The parameter proposal distributions are independent Gaussian, with mean and variance equal to the mean and variance of the resampled parameters. The PMCMC chain length in the resampling step is 10.

The model evidence estimates for each model are given in Table 4.6. The Bayes factors for CR14-a are $\beta_{12} = \mathcal{O}(10)$, $\beta_{13} = \mathcal{O}(10^2)$, $\beta_{14} = \mathcal{O}(10^7)$, $\beta_{15} = \mathcal{O}(10^7)$. Accounting for Monte Carlo error, each of the oscillators seem to have roughly equal support from the data. However, the oscillators have a much larger posterior model probability than the steady-state models. This conclusion differs from that of the ABC-PRC results, which suggest equal support for TSS. It is possible that the tolerance is large enough so that simulations from TSS are frequently considered close to the data, and that reducing the tolerance further would lead to the oscillators being more strongly supported than TSS. Since SMC$^2$ targets the correct posterior distributions, and doesn't suffer from the problems associated with the all-in-one approach to estimating Bayes factors, we consider the SMC$^2$ conclusions to be more reliable. The results support the use of oscillators in phenomenological modelling of the glacial-interglacial cycle.

The marginal posterior probabilities for CR14-a are shown in Figure 4.5. The posterior distributions are mostly consistent with the posterior distributions generated by the ABC-PRC algorithm. The scaling terms for the Brownian motions and the observation error are centred around larger values than in ABC-PRC. This is a result of the ABC approximation, which favours lower values for these parameters at large tolerances, as discussed in Chapter 3. The values of $\beta_0$ and $\delta$ are also larger in comparison, which may also be a result of the ABC approximation. Finally, the astronomical forcing terms are all close to zero. In the next chapter we demonstrate that this is influenced by the choice of age model.

The number of required simulations for each model was 121 million for CR14-a, 133 million for CR14-b, 131 million for CR14-c, 95 million for TSS, and 90 million for EBM, which are very close to the number required in the simulation study. Each model had at least 835 distinct

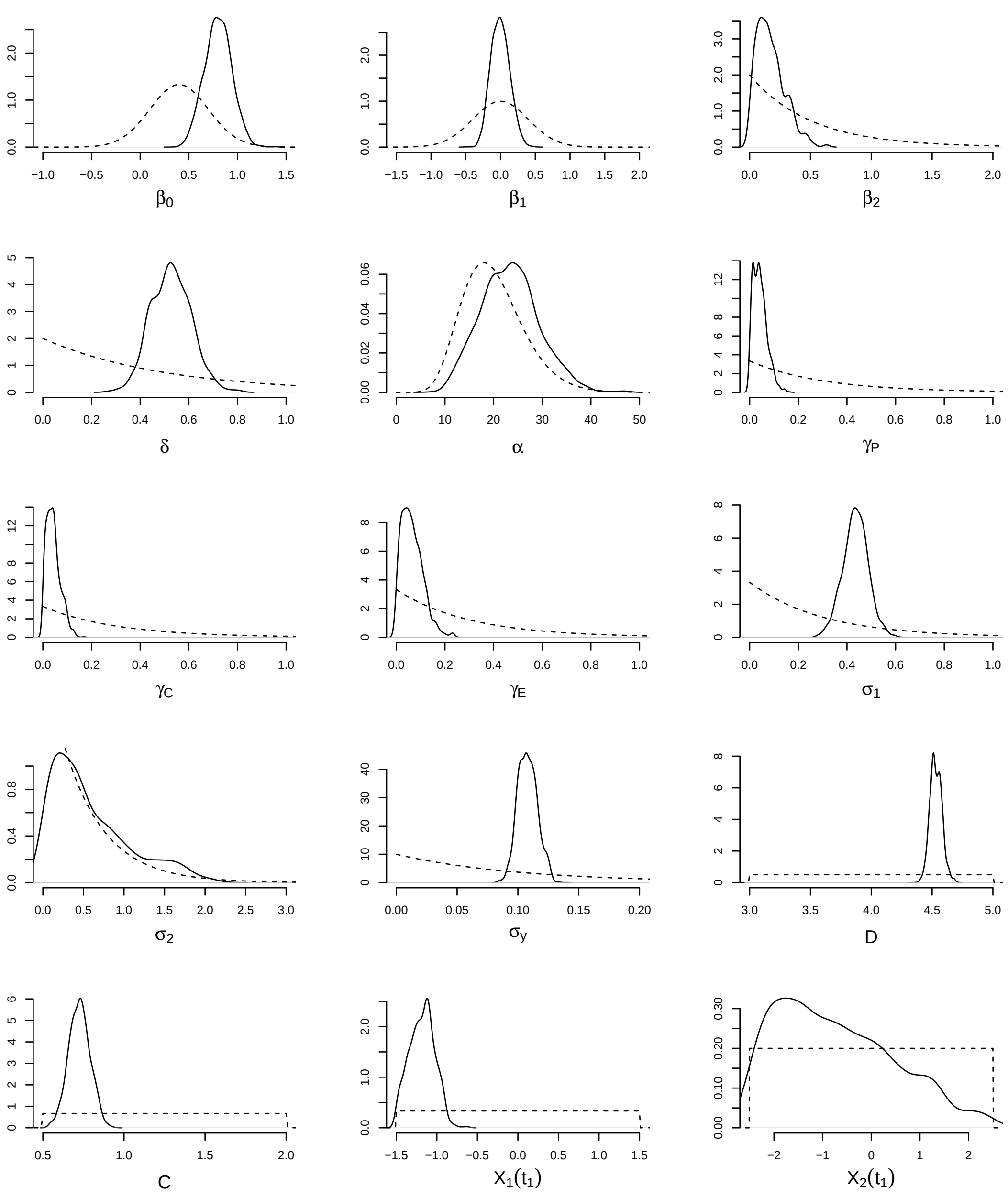

**Figure 4.5:** Marginal posterior distributions of the parameters of CR14-a, obtained using SMC$^2$ for ODP677. Dashed lines show the prior distributions. In comparison to the posterior distributions obtained in the simulation study, shown in Figure 3.6, we seem to be learning less about the parameters.

particles in the final iteration, so that the posterior distributions are well characterised for every model.

## 4.5 Chapter Summary

In this chapter we have introduced Bayes factors as a method to perform model comparison. Two of the inference methods in Chapter 3 have been extended to evaluate Bayes factors for phenomenological models of the glacial-interglacial cycle. We have focussed on SMC approaches, which naturally provide estimates of normalising constants, such as the model evidence, from which we can compute Bayes factors.

We designed a simulation study to assess the performance of the proposed inference methods, specifically ABC-PRC and SMC$^2$, which give results consistent with each other. The SMC$^2$ algorithm targets the correct posterior distribution for each model, but has a much greater computational expense than ABC-PRC. Applied to real-world data, the two methods give slightly different results. The output of SMC$^2$ suggests that the data more strongly support oscillators over steady-state models, but does not distinguish between oscillators. The ABC-PRC results suggest that the data gives equal support to the oscillators and a two stable steady-state model, but weak support to a single steady-state model. The discrepancy likely lies in the nature of the ABC approximation to the posterior distribution. With a sufficiently large tolerance, every model is likely to generate simulated values that are considered close to the observations. In our implementation, it seems that the tolerance has been lowered to the point that EBM is less likely to generate trajectories that are close to the data than the other models. Reducing the tolerance further would likely result in TSS being less likely to generate accepted simulations than the oscillators.

Alternative extensions to the methods described in Chapter 3 are possible. The MCMC approaches can be extended to perform model comparison by using RJMCMC [81], which targets an extended distribution that includes all models and model parameters. RJMCMC can also be used to design an ABC-ASMC algorithm for model comparison, or to design an SMC$^2$ algorithm that uses the all-in-one approach to evaluate Bayes factors. However, designing efficient proposal distributions in an RJMCMC algorithm can be difficult, whereas the SMC approaches we have discussed allow this process to be easily automated. Finally, ABC-SMC extensions to estimate the model evidence are described in [64].

We conclude that SMC algorithms, modified to perform inference with intractable likelihoods, are promising inference tool to study model comparison problems in palaeoclimate science. The fact that the selection of the tuning parameters in these algorithms can be automated is extremely beneficial, due to the large number of phenomenological models and palaeoclimate datasets that can be used. However, the approximation used in ABC approaches seems to discard a lot of information in comparison to the likelihood estimation approach used in SMC$^2$. In the following chapter we study two important model comparison problems in palaeoclimate science using SMC$^2$.

CHAPTER 5

# Model Comparison Problems in Palaeoclimate Science

In this chapter we study two model comparison problems in palaeoclimate science. Specifically, we aim to test whether palaeoclimate data support some phenomenological models more strongly than others, and whether the data favour astronomically forced models over unforced ones. As a results of the experiments in Chapter 4, we decide to use SMC$^2$ to evaluate the model evidence of each model, as the correct posterior distribution is targeted, the posterior distributions are well characterised for every model, and little tuning is required when applying the algorithm to different models and datasets.

The chapter is divided as follows. In Section 5.1 we discuss the model comparison problems of interest, and give an overview of previous studies. In Section 5.2 we give an overview of the models that will be used in this chapter. In Section 5.3 we design a simulation study in order to test our ability to study the specified model comparison problems. In Section 5.4 we discuss palaeoclimate data, and motivate using individual datasets rather than stacks. We then repeat the model comparison experiment on real-world data. Finally, in Section 5.5 we conclude the chapter, and give a discussion of the results.

## 5.1 Topical Model Comparison Problems

Phenomenological models of the glacial-interglacial cycle are mathematical representations of hypotheses about long-term climate dynamics. Through the model comparison methods described in Chapter 4, we can test competing hypotheses. We focus on two topical problems: selecting between competing phenomenological models of the glacial-interglacial cycle, and assessing the impact of the astronomical forcing. The problems are discussed in more detail below.

### 5.1.1 Selecting Between Phenomenological Models

Numerous phenomenological models of the glacial-interglacial cycle have been proposed over the last few decades. The performance of these models is often assessed by superimposing the output of the model onto observations from sediment cores over the timescales of interest. While this approach demonstrates that each model can reproduce much of the climate variation over long timescales, it offers no insight into the relative explanatory power between models.

A formal approach to select between competing models was taken in [24]. Six models were considered as representatives of different modelling approaches. All of the models were deterministic, and the residuals were modelled as an autoregressive process. The model parameters were chosen to give the best fit to the data, which was interpolated on to a 5 kyr grid. The mean square prediction error was then evaluated for each model, in order to determine which of the models provide a better fit to the data. An F-test was then constructed to investigate whether there was significant evidence to reject each model in favour of the best fitting model. The conclusion was that the data does not support any one model over another.

The Bayesian model comparison framework described in Chapter 3 has several advantages when compared with the frequentist approach used in [24]. Firstly, we characterise parameter uncertainty, rather than choosing the best parameters for each model. Models with large likelihoods over a wide range of parameters are more strongly supported than those with large likelihoods over a narrow range of parameters, due to averaging over the parameter space. We can also include expert knowledge in the prior probabilities of the models, and prior distributions of the parameters. It is concluded in [24] that since the data does not support any particular model, that they need to be considered on physical grounds. For example, the authors rejected one model (a linear combination of the orbital parameters) as the optimal parameter choice is unrealistic. A Bayesian approach allows this knowledge to be formally incorporated into the procedure.

### 5.1.2 Assessing the Impact of the Astronomical Forcing

There has been considerable interest in determining whether Milankovitch theory, which suggests that the glacial-interglacial cycle is paced by variation in the Earth's orbital parameters over time, is supported by palaeoclimate data, and if it is, which of the orbital parameters are most influential in pacing the glacial-interglacial cycle [4, 18, 85, 86]. This problem has recently been studied using frequentist hypothesis testing. Some examples are reviewed below.

An early experiment reduced a palaeoclimate dataset to seven termination times, with the aim of testing for correlation in the phase of obliquity at those times [4, 18]. The stability of the phase was measured by Rayleigh's $R$, defined as

$$R = \frac{1}{N}\left|\sum_{n=1}^{N} \cos\phi_n + i\sin\phi_n\right|, \tag{5.1.1}$$

where $\phi_n$ is the phase of obliquity at termination $n$. Rayleigh's $R$ takes a value between 0 and 1, where a value of 1 indicates that the phase is equal at each termination time. The intuition behind using $R$ was that if obliquity drives the glacial-interglacial cycle, then the terminations would likely occur at a similar point in the obliquity phase, which would be indicated by a large value of $R$. In order to determine if the observed $R$ value was large, a null model was proposed in which the terminations are independent of obliquity. The observed $R$ value was then be compared with $R$ values under the null hypothesis to determine if it was significantly larger. In [18], the distribution of $R$ under the null hypothesis was estimated using the random walk model

$$V_{t+1} = V_t + \eta_t, \tag{5.1.2}$$

where $V_t$ is the ice-volume at time $t$, and $\eta_t \sim \mathcal{N}\left(1, 2^2\right)$. One time step is 1 kyr. At some predefined threshold the ice volume returns to 0 linearly over 10 kyr. The random walk model was used to generate $10^4$ values of $R$, which were used to form a histogram approximating the probability distribution of $R$ under the null hypothesis. This distribution was then used to decide whether the observed $R$ value provides significant evidence to reject the null hypothesis.

The null hypothesis was rejected in the case of obliquity, but further hypothesis tests for precession and eccentricity gave a negative result in each case. These results also held when the distribution under the null hypothesis was generated by drawing phases from a uniform distribution at each termination time [4, 18]. Additional insight was gained through the construction of the distribution of $R$ under the alternative hypothesis. This was done by considering the age model uncertainty at the termination times, which was assumed to be 10 kyr (although the distribution was not specified). Realisations of $R$ under the alternative hypothesis were generated by perturbing the glacial termination times according to the age model uncertainty. The distribution of $R$ under the alternative hypothesis was then approximated by generating a histogram using $10^4$ realisations. The alternative distribution allowed the power of the test to be evaluated. The power was high for obliquity and eccentricity, but very small for precession. This was because the age model uncertainty was approximately half of the period of precession. In other words, even if the glacial-interglacial cycle is strongly influenced by precession, this test would likely not reject the null hypothesis.

Similar experiments have since been performed. For example, in [85], cross-wavelet analysis was used in place of phase measurements at instantaneous termination times. With this alteration, the null hypothesis was rejected for eccentricity. In [86], obliquity and precession were linearly combined into an insolation function, similar to Equation 2.3.5. Focus was shifted to the magnitude of insolation peaks associated with glacial terminations, rather than the phase of the orbital parameters. The conclusion was that both parameters are influential, and that the combination of precession and obliquity has significantly more explanatory power than either of the parameters individually.

We assess the influence of the astronomical forcing by selecting between forced and unforced phenomenological models within a Bayesian model comparison framework. While this does not single out each parameter, the same approach can be used with phenomenological models forced purely by obliquity or precession. The main benefit of our approach is that we use all of the data, rather than summarising the data as a series of termination times as in the described frequentist methods. In this chapter we assume that the observation times are known. Incorporating age model uncertainty is discussed in Chapter 6.

## 5.2 Models

Here we consider three phenomenological models from the literature. The reader is referred to the original papers for interpretations of the parameters.

### SM91

SM91 was among the first models that represented the glacial-interglacial cycle as an oscillating system synchronised on the astronomical forcing [23]. SM91 explicitly models three climate variables: Ice-volume ($X_1$), carbon dioxide concentration ($X_2$), and deep-ocean temperature

($X_3$). The full system of equations are

$$\begin{aligned} dX_1 &= -\left(X_1 + X_2 + vX_3 + I(\gamma_P, \gamma_C, \gamma_E)\right) dt + \sigma_1 dW_1, \\ dX_2 &= \left(rX_2 - pX_3 - sX_2^2 - X_2^3\right) dt + \sigma_2 dW_2, \\ dX_3 &= -q\left(X_1 + X_3\right) dt + \sigma_3 dW_3. \end{aligned}$$

Nonlinearity is introduced through the carbon dioxide equation, which is the cause of the oscillations. Similar variants of this model exist with different carbon dioxide equations, such as SM90 [22]. The astronomical forcing, $I(\gamma_P, \gamma_C, \gamma_E)$, is included in the ice-volume equation, pacing the oscillations. The unit of time is 10 kyr. SM91 has a total of eleven tunable parameters, $\boldsymbol{\theta} = (p, q, r, s, v, \gamma_P, \gamma_C, \gamma_E, \sigma_1, \sigma_2, \sigma_3)^T$.

**T06**

T06 is a hybrid model, in that the system is a combination of continuous and discrete state variables [27]. The continuous state variable ($X_1$) represents ice-volume (in units of $10^{15}$ m$^3$), and the discrete state variable ($X_2$) represents the absence ($X_2 = 0$) or presence ($X_2 = 1$) of Arctic sea-ice. The system of equations is

$$\begin{aligned} dX_1 &= \left((p_0 - KX_1)(1 - \alpha X_2) - (s + I(\gamma_P, \gamma_C, \gamma_E))\right) dt + \sigma_1 dW_1, \\ X_2 &: \text{ switches from 0 to 1 when } X_1 \text{ exceeds some threshold } X_u, \\ X_2 &: \text{ switches from 1 to 0 when } X_1 \text{ decreases below } X_l. \end{aligned}$$

This model was used to demonstrate nonlinear phase locking to the astronomical forcing, and to highlight that a good fit to the ice volume record can be obtained as long as the glacial mechanism is strongly nonlinear. As with SM91, the unit of time is 10 kyr. T06 contains the fewest parameters of the models considered in this chapter, with a total of ten tunable parameters, $\boldsymbol{\theta} = (p_0, K, s, \alpha, x_l, x_u, \gamma_P, \gamma_C, \gamma_E, \sigma_1)^T$.

**PP12**

PP12 models distinct glaciation and deglaciation phases [28]. During the glaciation phase the ice-volume (expressed as equivalent sea-level) slowly increases. The build-up is influenced by the astronomical forcing, with lower values of insolation increasing the rate of ice accumulation. During the deglaciation phase the ice-volume decreases rapidly. The phase changes occur mainly due to the astronomical forcing. The system is described below.

Define a truncation function:

$$f(x) = \begin{cases} x + \sqrt{4a^2 + x^2} - 2a & \text{if } x > 0, \\ x & \text{otherwise,} \end{cases}$$

with parameter $a$. Define rescaled precession and coprecession parameters

$$\begin{aligned} \Pi^\dagger &= (f(\bar{\Pi}) - 0.148)/0.808, \\ \mathit{\Pi}^\dagger &= (f(\bar{\mathit{\Pi}}) - 0.148)/0.808, \end{aligned}$$

and a rule defining the transition between glacial states ($g$) and interglacial states ($d$)

$$\begin{cases} d \to g & \text{if } \kappa_P \bar{\Pi} + \kappa_C \bar{\mathit{\Pi}} + \kappa_E \bar{O} < v_1, \\ g \to d & \text{if } \kappa_P \bar{\Pi} + \kappa_C \bar{\mathit{\Pi}} + \kappa_E \bar{O} + X_1 > v_0. \end{cases}$$

The ice-volume then evolves according to

$$dX_1 = -(\gamma_P \Pi^\dagger + \gamma_C \mathit{\Pi}^\dagger + \gamma_E \bar{O} - A)dt + \sigma_1 dW_1,$$

where

$$A = \begin{cases} -a_d - \frac{X_1}{\tau} & \text{if in state } d, \\ a_g & \text{if in state } g. \end{cases}$$

Due to the truncation of the forcing in the ice volume equation, this model responds nonlinearly to variation in insolation. The unit of time is 1 kyr. PP12 has the greatest number of parameters out of the models considered here, with a total of thirteen tunable parameters, $\boldsymbol{\theta} = (a, a_d, a_g, \kappa_P, \kappa_C, \kappa_E, \tau, v_0, v_1, \gamma_P, \gamma_C, \gamma_E, \sigma_1)^T$.

## 5.3 Simulation Study

We design a simulation study to assess our ability to study the model comparison problems described above. We generate two datasets from SM91, one which has been obtained using the forced model, and one which has been obtained from the unforced model. These are denoted SM91-f and SM91-u respectively. The observation model takes the same form as Equation 2.3.16, which scales and displaces the state representing ice-volume. The selected parameter values are: $p = 0.8$, $q = 1.6$, $r = 0.6$, $s = 1.4$, $v = 0.3$, $\sigma_1 = 0.2$, $\sigma_2 = 0.3$, $\sigma_3 = 0.3$, $\gamma_P = 0.3$, $\gamma_C = 0.1$, $\gamma_E = 0.4$, $D = 3.8$, $C = 0.8$, $\sigma_Y = 0.1$. To generate a dataset from the unforced model we set $\gamma_P = \gamma_C = \gamma_E = 0$, thus giving no astronomical forcing. Observations are taken every 3 kyr over the past 780 kyr, giving 261 observations in each dataset. This is

| SM91 | | T06 | | PP12 | |
|---|---|---|---|---|---|
| $\gamma_P$ | $exp(1/0.3)$ | $\gamma_P$ | $exp(1/0.6)$ | $\gamma_P$ | $exp(1/1.5)$ |
| $\gamma_C$ | $exp(1/0.3)$ | $\gamma_C$ | $exp(1/0.6)$ | $\gamma_C$ | $exp(1/1.5)$ |
| $\gamma_E$ | $exp(1/0.3)$ | $\gamma_E$ | $exp(1/0.6)$ | $\gamma_E$ | $exp(1/1.5)$ |
| | | | | | |
| $p$ | $\Gamma(2, 1.2)$ | $p_0$ | $exp(1/0.3)$ | $a$ | $\Gamma(8, 0.1)$ |
| $q$ | $\Gamma(7, 3)$ | $K$ | $exp(1/0.1)$ | $a_d$ | $exp(1)$ |
| $r$ | $\Gamma(2, 1.2)$ | $s$ | $exp(1/0.3)$ | $a_g$ | $exp(1)$ |
| $s$ | $\Gamma(2, 1.2)$ | $\alpha$ | $Beta(40, 30)$ | $\kappa_P$ | $exp(1/20)$ |
| $v$ | $exp(1/0.3)$ | $x_l$ | $exp(1/3)$ | $\kappa_C$ | $exp(1/20)$ |
| $\sigma_1$ | $exp(1/0.3)$ | $x_u$ | $\Gamma(90, 0.5)$ | $\kappa_E$ | $exp(1/20)$ |
| $\sigma_2$ | $exp(1/0.3)$ | $\sigma_1$ | $exp(1/2)$ | $\tau$ | $exp(1/10)$ |
| $\sigma_3$ | $exp(1/0.3)$ | | | $v_0$ | $\Gamma(220, 0.5)$ |
| | | | | $v_1$ | $exp(1/5)$ |
| | | | | $\sigma_1$ | $exp(1/5)$ |
| | | | | | |
| $D$ | $\mathcal{U}(3, 5)$ | $D$ | $\mathcal{U}(2.5, 4.5)$ | $D$ | $\mathcal{U}(2.5, 4.5)$ |
| $S$ | $\mathcal{U}(0.25, 1.25)$ | $S$ | $\mathcal{U}(0.02, 0.05)$ | $S$ | $\mathcal{U}(0.01, 0.03)$ |
| $\sigma_Y$ | $exp(1/0.1)$ | $\sigma_Y$ | $exp(1/0.1)$ | $\sigma_Y$ | $exp(1/0.1)$ |
| | | | | | |
| $X_1(t_1)$ | $\mathcal{U}(-1.5, 1.5)$ | $X_1(t_1)$ | $\mathcal{U}(3, 45)$ | $X_1(t_1)$ | $\mathcal{U}(0, 120)$ |
| $X_2(t_1)$ | $\mathcal{U}(-1.5, 1.5)$ | | | | |
| $X_3(t_1)$ | $\mathcal{U}(-1.5, 1.5)$ | | | | |

**Table 5.1:** List of prior distributions for each model in the model comparison experiment. Sections indicate parameters used to scale the astronomical forcing (absent in unforced models), parameters of the phenomenological model, observation model, and initial conditions respectively.

equivalent to a reasonably low resolution sediment core.

A total of five models are considered: SM91 (forced and unforced), T06 (forced and unforced), and PP12. An unforced variant of the PP12 model is not considered, as the deglaciation-glaciation switch is controlled entirely by the astronomical forcing. This is not the case for SM91 and T06, which both oscillate in the absence of any external forcing. The prior distributions used for each model are summarised in Table 5.1.

The goals of the simulation study are to determine whether we are able to select the correct model for each dataset, and whether the forced variant of the T06 model is favoured when the dataset has been generated using the forced variant of the SM91 model. In other words, we are interested in determining whether the astronomical forcing adds explanatory power to the model, even when the model is wrong.

| Model | | Evidence | |
|---|---|---|---|
| | | SM91-u | SM91-f |
| SM91 | Forced | $5.6 \times 10^{28}$ | $1.4 \times 10^{41}$ |
| | Unforced | $1.1 \times 10^{30}$ | $2.4 \times 10^{18}$ |
| T06 | Forced | $3.6 \times 10^{20}$ | $2.6 \times 10^{30}$ |
| | Unforced | $1.1 \times 10^{22}$ | $2.9 \times 10^{14}$ |
| PP12 | Forced | $2.8 \times 10^{8}$ | $2.1 \times 10^{18}$ |

**Table 5.2:** The estimated model evidence for each model for different datasets in the simulation study. In the left-hand column the dataset was generated using the SM91 model with no astronomical forcing. In the right-hand column the dataset was generated using the forced SM91 model. The largest model evidence for each dataset is shown in red, indicating the favoured model. In both cases we select the correct model. Note that in the right-hand column, the evidence for the forced T06 model (shown in blue) is greater than the unforced T06 model. In other words, we are able to detect the influence of the astronomical forcing, even when the wrong model is used.

### 5.3.1 Results on the Simulation Study Data

We run the SMC$^2$ algorithm with $N_X = 1000$ state particles and $N_\theta = 1000$ parameter particles for each model. The proposal distribution in the PMCMC chain is a Gaussian independence sampler with mean and covariance equal to that of the current sample. The PMCMC chain length in the resampling stages is set to 10, which gives a good particle diversity.

The estimated model evidence of each model, for both datasets, is given in Table 5.2. The correct model is preferred in both cases. For the dataset generated using the forced model, the forced version of SM91 is strongly preferred, with the model evidence eleven orders of magnitude greater than the next best model. The forced version of T06 is more strongly supported by the data than the unforced variant, showing that we favour forced models even when the model is wrong. For the dataset generated using the unforced model, the unforced variant of the SM91 model is only slightly more strongly supported than the forced variant. This is because the unforced model is nested within the forced model, and so the forced model is being penalised for having extra parameters that add little explanatory power. The same is true for T06. PP12 is relatively weakly supported by the data in each case. This is likely because SM91 and T06 are both oscillators paced by the astronomical forcing, whereas the glacial-interglacial switch in PP12 is controlled by the astronomical forcing.

The parameter marginal posterior distributions for the forced SM91 model are shown in Figure 5.1 for SM91-f, and Figure 5.2 for SM91-u. The true values lie in regions of large posterior probability density in each case.

In Figure 5.3 we compare the posterior distributions of the ratios of the astronomical forcing scaling parameters for SM91 and T06. Since these models use different measures for the ice volume (the ice volume state variable is adimensional in SM91, and has units of $10^{15}$ m$^3$ in

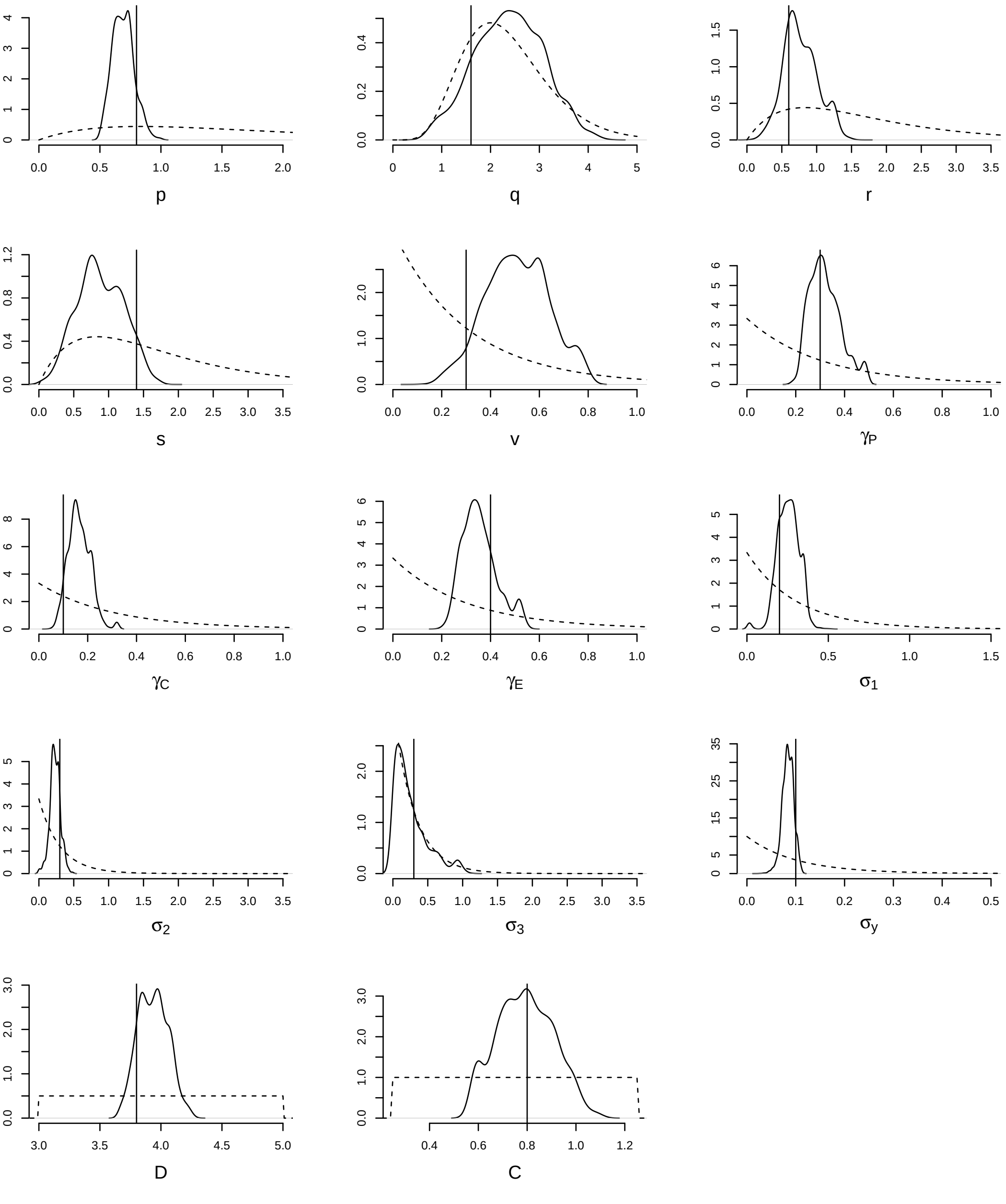

**Figure 5.1:** Marginal posterior distributions of the parameters of the SM91 model for the SM91-f dataset. Dashed lines show the prior distributions. Vertical lines show the values used to generate the data. The true values lie in regions of high posterior probability density.

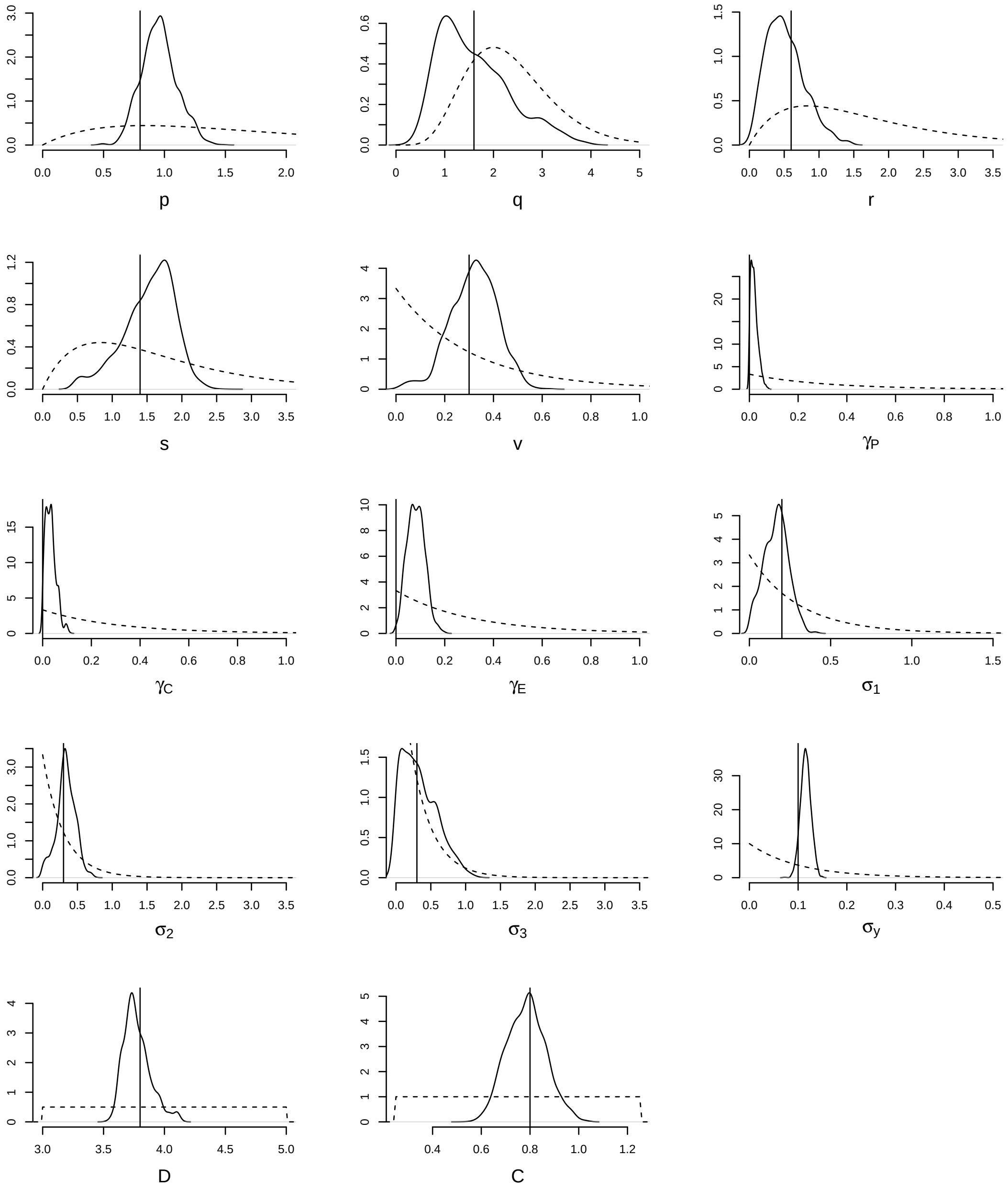

**Figure 5.2:** Marginal posterior distributions of the parameters of the SM91 model for the SM91-u dataset. Dashed lines show the prior distributions. Vertical lines show the values used to generate the data. The true values lie in regions of high posterior probability density.

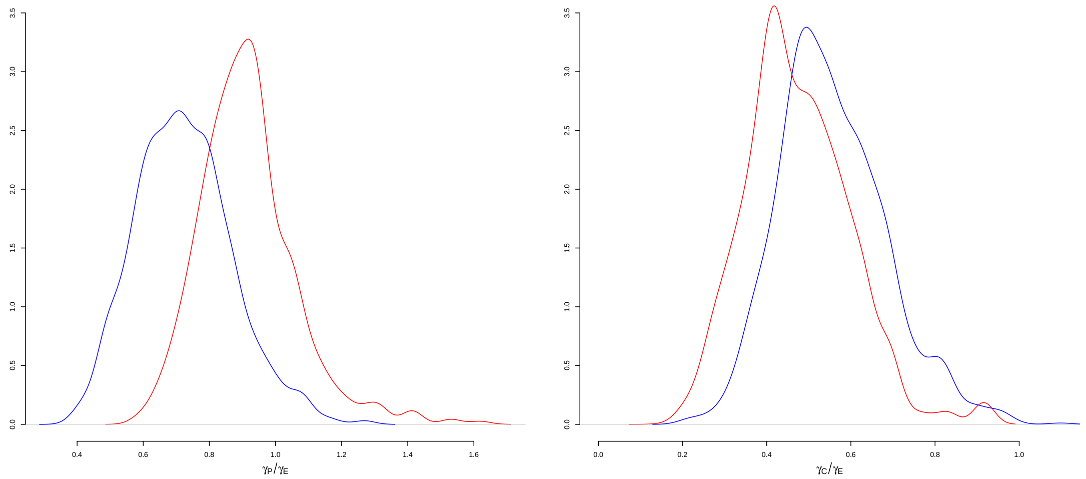

**Figure 5.3:** The posterior distributions for the ratio of the astronomical forcing terms for the SM91-f dataset. The red lines show the posterior distributions from the SM91 model, and the blue lines show the posterior distributions from the T06 model. The posterior distributions between the two models are similar enough to suggest that both models are synchronising on the same forcing function. The PP12 model is omitted as the ice volume equation responds nonlinearly to the astronomical forcing, making the parameters non-comparable.

T06), it is expected that the scaling parameters for the astronomical forcing scale accordingly. However, if the models synchronise on the same forcing signal, then the ratio between any pair of parameters should be identical, since the ratio of the astronomical forcing scaling parameters defines the shape of the forcing function. The posterior distributions are similar, suggesting that this could be the case. PP12 is omitted as the ice volume responds nonlinearly to the astronomical forcing, and so the parameters are not comparable.

## 5.4 Palaeoclimate Data

We have demonstrated that, in a simulation study, our methodology is able to select the correct phenomenological model, and determine whether or not the data are from a forced model, even when the wrong model is used. We now return to using real-world data to address the motivating scientific questions.

As described in Section 1.1, ice volume and temperature during the palaeoclimate are often studied using records of $\delta^{18}$O, which is a measure of the ratio $^{18}$O:$^{16}$O. Datasets from individual drill sites often provide sparse and noisy data, and are also influenced by local climate variation. For these reasons, datasets from sediment cores are often combined into stacks, which are averages over multiple datasets. Two recent stacks are described below.

### LR04

The LR04 stack is an average of 57 datasets extending over 5.3 Myr [1]. The first step in the stack's construction was to correlate each of the datasets, for example, by aligning peaks. This process was mostly automated, but required user input to identify missing or repeated data. The data were dated by aligning the stack to a phenomenological model of ice volume, which was forced by summer solstice insolation at 65°N. The stacking procedure involved dividing the time domain into small time intervals. Within each interval the observations from every dataset were averaged. It should be noted that, due to the sparsity of the data, each point in the stack was generated by averaging over a different subset of datasets. In other words, at any point in the stack, information has been discarded from multiple cores.

### H07

The H07 stack is an average of 14 datasets extending over 2 Myr [4]. The motivation behind H07 was to reconstruct long-term climate variation without astronomical tuning assumptions. The times of the geomagnetic reversals and the last glacial termination were treated as known, and denoted as age control points (ACPs). At first, each dataset was assigned a deterministic piecewise linear age model accounting for compaction (this is discussed in Chapter 6) between the ACPs. The piecewise linear age model provided dating estimates of common events (glacial terminations, peaks, etc.) that were identifiable in each of the datasets, and which were assumed to have occurred synchronously in each dataset. Whilst individual estimates of the event times are likely to be poor, it was assumed that averaging the estimates from multiple datasets gives a reasonable approximation. The average event ages were then fixed and treated as additional ACPs. A deterministic piecewise linear age model between the new ACPs was then assigned to each dataset. The stacking procedure was done by linearly interpolating each dataset onto a 1 kyr grid, before averaging over each of the datasets. It should be noted that interpolation can cause problems when there are large sections of missing data in a dataset, as interpolation over large timescales will poorly represent the missing data. In other words, at any point in the stack, false information is being included.

Both of the stacking procedures described above alter the information contained in the palaeoclimate datasets, potentially affecting any inference about the dynamics of the palaeoclimate. For this reason, we choose to use individual datasets to study the model comparison problems. However, it is reasonable to assume that dating palaeoclimate datasets as part of a stack provides better estimates than dating each dataset individually, as missing and repeated data are more easily identified. In this chapter we use ODP677 [84], which has been dated as part of the LR04 and H07 stacks. The Brunhes-Matuyama (BM) reversal can be identified, and

| Model | | Evidence | |
|---|---|---|---|
| | | ODP677-u | ODP677-f |
| SM91 | Forced | $4.0 \times 10^{24}$ | $1.1 \times 10^{28}$ |
| | Unforced | $3.5 \times 10^{26}$ | $1.6 \times 10^{18}$ |
| T06 | Forced | $3.3 \times 10^{25}$ | $4.5 \times 10^{29}$ |
| | Unforced | $1.7 \times 10^{28}$ | $3.3 \times 10^{21}$ |
| PP12 | Forced | $1.5 \times 10^{22}$ | $1.8 \times 10^{34}$ |

**Table 5.3:** The estimated model evidence for each model for the ODP677-u and ODP677-f datasets using SMC$^2$. The largest model evidence for each dataset, with consideration of the Monte Carlo error, is shown in red, indicating the favoured model. When using ODP677-u, we prefer unforced models, but it is not possible to confidently select a single best model due to the Monte Carlo error. When using ODP677-f, PP12 is strongly favoured. However, note that for the other models, the forced variants are more strongly supported by the data (shown in blue). Our results are highly sensitive to the age model used to estimate the observation times.

so we use the latest 780 kyr of the dataset, in which there are 363 observations. As well as the model comparison goals described earlier in this chapter, we aim to test whether the dating method influences which models are more strongly supported by the data. We denote the dataset dated as part of the LR04 stack as ODP677-f, and the dataset dated as part of the H07 stack as ODP677-u, highlighting whether or not the dataset has been astronomically tuned.

### 5.4.1 Results on ODP677

We run the SMC$^2$ algorithm using the settings given in Section 5.3.1. The model evidence of each model is given in Table 5.3.

For ODP677-u the unforced T06 model is the most strongly supported model, followed by the unforced SM91 model. The Bayes factor in favour of the unforced T06 model against the unforced SM91 model is $\sim$50, which, as described in Section 4.3.2, cannot be seen as conclusive due to our Monte Carlo error. The forced variants of these models are less strongly supported, due to containing additional parameters that have little explanatory power. Both SM91 and T06 are more strongly supported than PP12. This same dataset was used in Section 4.4, where it was shown that CR14-a, CR14-b, and CR14-c have comparable support.

For ODP677-f PP12 is the most strongly supported model. The explanation may lie in how the astronomical forcing affects each model. In SM91 and T06, the astronomical forcing acts in a similar fashion to a pullback attractor, attracting the state of the system into specific regions of state space. In PP12, the astronomical forcing controls the transition from the glaciated state to the deglaciated state. As such, we might expect the output of PP12 to be more strongly correlated to the astronomical forcing than SM91 and T06, giving better agreement to a dataset that has been astronomically tuned. Additionally, the forced variants of SM91 and T06 are more

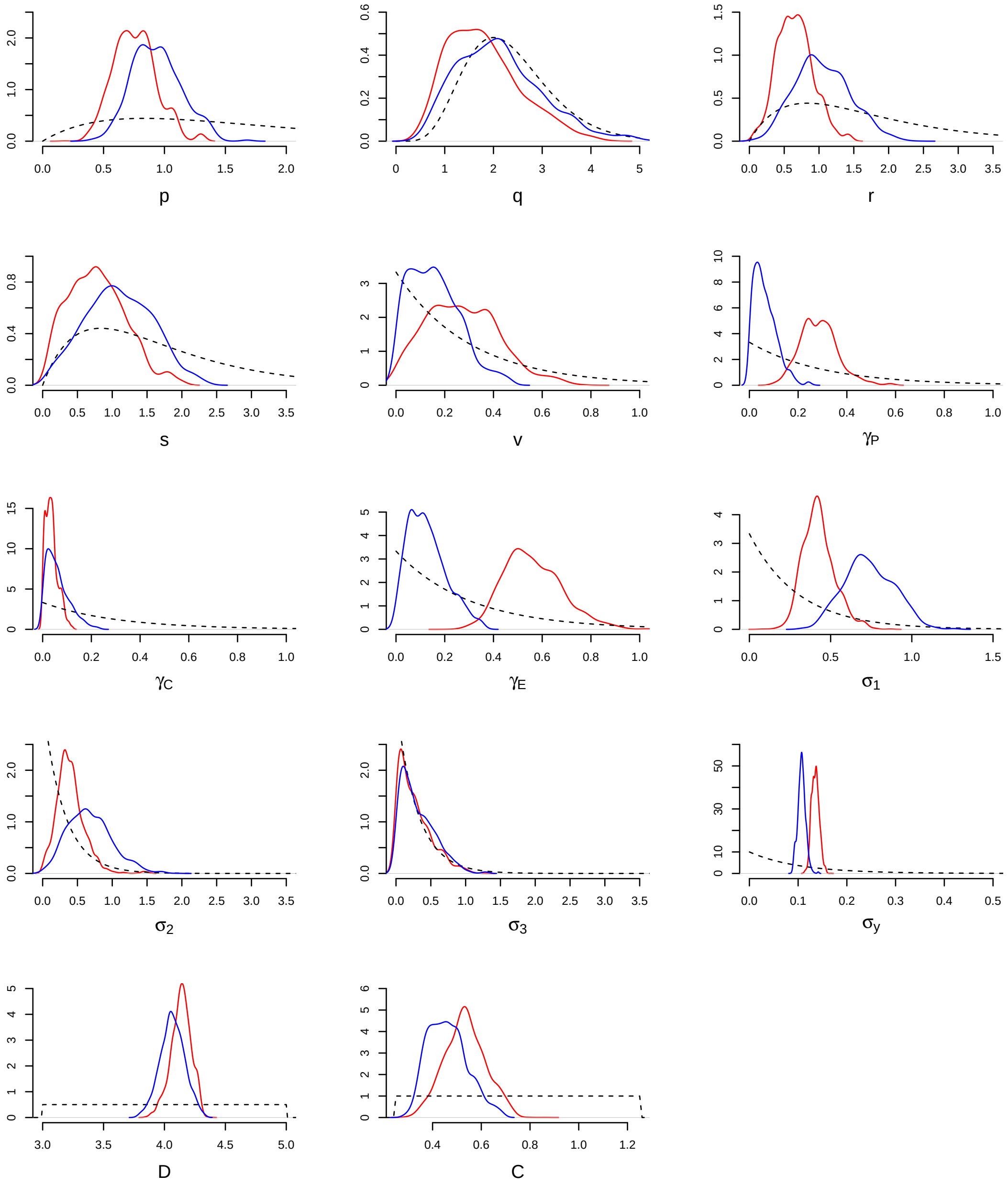

**Figure 5.4:** Marginal posterior distributions for the SM91 model for ODP677-u (blue) and ODP677-f (red). Dashed lines show the prior distributions. The two sets of distributions are similar, except for the astronomical forcing scaling terms, $\gamma_P$, $\gamma_C$, and $\gamma_E$, for which larger estimates are preferred for ODP677-f, where astronomical tuning assumptions were used in the age model.

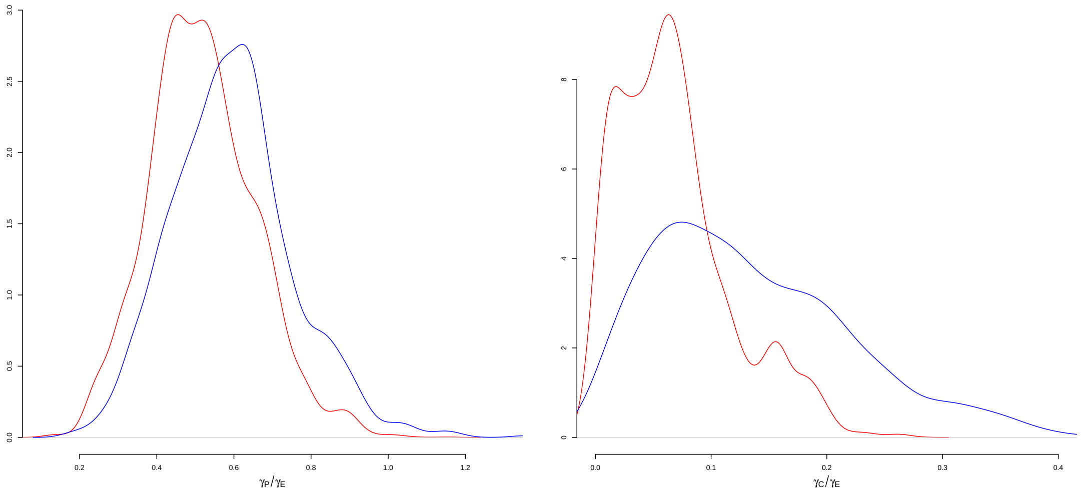

**Figure 5.5:** The posterior distributions for the ratio of the astronomical forcing terms for the ODP677-f dataset. The red lines show the posterior distributions from the SM91 model, and the blue lines show the posterior distributions from the T06 model. The posterior distributions between the two models are similar enough to suggest that both models are synchronising on the same forcing function. The PP12 model is omitted as the ice volume equation responds nonlinearly to the astronomical forcing, making the parameters non-comparable.

strongly supported than the unforced variants. As with ODP677-u, it is difficult to determine if T06 is more strongly supported than SM91 due to the Monte Carlo error.

These results show that inference is strongly affected by the age model used, and that modelling assumptions in the dating methods should be understood when performing inference on palaeoclimate data. In other words, the results are not robust across different dating strategies, and, in particular, using astronomical tuning assumptions leads to astronomically forced models (perhaps wrongly) being strongly supported by the data.

The marginal posterior distributions of the parameters of the forced SM91 model are shown in Figure 5.4. The marginal posterior distributions are similar for both ODP677-u and ODP677-f for many of the parameters. The scaling terms for obliquity ($\gamma_E$) and precession ($\gamma_P$) are larger in ODP677-f, since it has been astronomically tuned. The scaling term $\sigma_1$ is larger for ODP677-u, suggesting that larger stochastic perturbations are required to fit the model output to the dataset. This is partly a trade-off with a reduction in $\sigma_Y$, suggesting that the trajectory of the model tends to be closer to the observations.

The posterior distributions of $\frac{\gamma_P}{\gamma_E}$ and $\frac{\gamma_C}{\gamma_E}$ for SM91 and T06 on ODP677-f are shown in Figure 5.5. The posterior distributions are similar enough to suggest that both models are synchronising on the same forcing signal. The scaling term for obliquity is greater than that of precession and coprecession, suggesting that obliquity is the dominant term in the insolation signal. For both $\gamma_E$ and $\gamma_P$ there is little posterior mass near zero, suggesting that both obliquity and precession add explanatory power. The posterior distribution of $\gamma_C$ suggests that it is very small, or possibly zero. A model comparison experiment will likely favour a forced model without coprecession.

### 5.4.2 A Note on Utilising Multiple Datasets

In this chapter we choose to use an individual dataset rather than a stack, due to standard stacking approaches having the potential to influence information from paleaoclimate records. However, individual datasets contain observations that are sparsely distributed in time, and subject to local variation in climate. Where possible, it is beneficial to average over multiple datasets to improve the signal-noise ratio, and to better represent global climate conditions. In an ideal setting, such as in a simulation study, where each dataset is generated using the true model, it is trivial to extend SMC$^2$ (or any of the inference methods discussed in this thesis) to assimilate observations from multiple sediment cores. The observation model is extended to

$$Y_m^{(n)} = D_n + \boldsymbol{C}_n^T \boldsymbol{X}_m + \Sigma_{Y,n}^{\frac{1}{2}} \eta_m, \tag{5.4.1}$$

where $Y_m^{(n)}$ is observation $m$ from sediment core $n$. $D_n$, $\boldsymbol{C}_n$, and $\Sigma_{Y,n}^{\frac{1}{2}}$ are the displacement, scaling, and observation error terms for core $n$ respectively. Hence, each additional dataset requires the estimation of three additional parameters. In SMC$^2$, the particle filter is run until the next observation time, and compared to the observation using the observation model parameters associated with the same core as the observation. If there are multiple observations at the same point in time then they can be assimilated simultaneously. However, reality is very different, as datasets from different sediment cores represent local climate variation, rather than global variation. Thus, while it is possible that our approach could be used across multiple sediment cores, doing so would require extending the phenomenological models to account for local effects. This is beyond the scope of this thesis, but a brief discussion is given in Chapter 7.

## 5.5 Chapter Summary

We have demonstrated how a Bayesian model comparison approach can be used to study model comparison problems in palaeoclimate science. In a simulation study, we have shown that we are capable of finding the true model from a collection of phenomenological models, and of detecting the influence of the astronomical forcing, even when using the wrong model.

Our major scientific finding from this chapter is that, when using real-world data, the results depend strongly on the approach used to date observations. Forced models are more strongly supported by a dataset that has been astronomically tuned, whereas unforced models are preferred when the ages of the observations are depth-derived. The favoured model also depends on the chosen age model, with PP12 being favoured when the dataset has been astronomically tuned, and SM91 and T06 having roughly equal support when the depth-derived age model has been used. It is noted in the original publications of both the LR04 and H07 stacks that the observation times are subject to age model uncertainty, which is estimated to be $\sim$10 kyr [1, 4]. With this uncertainty in mind, it is possible that the LR04 and H07 observation times are realisations of the same posterior distribution. To study these model comparison problems more thoroughly requires that this age model uncertainty is well characterised and accounted for.

This has important implications for how we believe these experiments should be performed. The results show that the current practice of performing each stage of the analysis independently of the others can lead to incorrect conclusions. In this case, obtaining the observation time estimates first, and then treating them as fixed for subsequent analysis, leads to conclusions that are not robust to the methods used in the first stage of the analysis. This suggests that a better approach would be to solve the full joint Bayesian problem of simultaneously estimating the observations times, parameters, and model evidence. This extension is developed in the next chapter.

CHAPTER 6

# Quantifying Chronological Uncertainty

In this chapter we study the problem of quantifying age model uncertainty in sediment cores. Up until now we have assumed that the observation times are known. However, in the previous chapter we showed that our model selection approach is sensitive to the choice of age model, which relates the depth of an observation to the time at which the sediment was deposited. In order to overcome this problem we extend our modelling approach to include a novel age model, which allows us to estimate the ages of observations from a sediment core. Whilst it is common in palaeoclimate science to identify factors contributing age model uncertainty, our approach allows us to characterise age model uncertainty in a statistically principled framework.

The chapter is divided as follows. In Section 6.1 we demonstrate how our modelling and inference approaches can be extended to estimate the observation times in a sediment core. We introduce a novel age model based on a linear sediment accumulation model with compaction adjustment. We then extend the particle filter to perform filtering on the observation times using the age model. In Section 6.2 we design a simulation study to test our inference approach. In Section 6.3 we repeat the experiment on two real-world datasets. Finally, in Section 6.4, we conclude the chapter with a discussion of the scientific implications.

## 6.1 Dating a Single Sediment Core

We begin by extending the state of the models to include a state variable representing time. We let $T_m$, where $m = 1, ..., M$, denote the time that the sediment in observation $m$ was laid down. We then target the extended distribution

$$\begin{aligned}\pi\left(\boldsymbol{\theta}, T_{1:M}, \boldsymbol{X}_{1:M}|\boldsymbol{Y}_{1:M}\right) &= \pi\left(\boldsymbol{\theta}|\boldsymbol{Y}_{1:M}\right)\pi\left(T_{1:M}, \boldsymbol{X}_{1:M}|\boldsymbol{Y}_{1:M}, \boldsymbol{\theta}\right) \\ &= \pi\left(\boldsymbol{\theta}|\boldsymbol{Y}_{1:M}\right)\pi\left(T_{1:M}|\boldsymbol{Y}_{1:M}, \boldsymbol{\theta}\right)\pi\left(\boldsymbol{X}_{1:M}|T_{1:M}, \boldsymbol{Y}_{1:M}, \boldsymbol{\theta}\right),\end{aligned} \tag{6.1.1}$$

so that we are estimating the parameters of the model, the state of the system at observation times, and the observation times. This is a far more challenging problem than we have dealt with so far in this thesis, as the dimension of the problem has increased dramatically. To target this extended distribution we need to extend our modelling approach to describe how observation times change according to depth in a sediment core (an age model), and also extend our inference approach to include the new state variables. We describe these extensions below.

### 6.1.1 Constructing an Age Model

Bayesian age models are widely used to infer the chronologies of sediment cores in which age estimates of some of the core slices are available from radiocarbon dating [87–90]. Radiocarbon dating requires that the ages are no older than 50 kyr [90], and so in comparison to the benthic cores we are interested in, the core depth and associated timescale are much shorter. Additionally, a greater proportion of core slices will have age estimates. Despite these differences many of these age models could be used in the inference approach described in this chapter. Notable Bayesian age models include BChron [87], which constructs a monotone Markov process via a discrete renewal process, giving a piecewise linear model, P Sequence in Oxcal [88, 89], which models sediment accumulation as a Poisson process, and Bacon [90], which uses an autoregressive gamma process to control sediment accumulation rates. However, we instead develop a novel age model based on a continuous stochastic sediment accumulation model. Our age model, described below, has many of the desired features described in [87], in that it is continuous, age increases monotonously with depth, and age uncertainty increases away from observed age estimates.

It is more natural to think of sediment changes over time than it is to think about how time changes throughout a sediment core. We begin by considering a linear sediment model where the amount of sediment (in metres), $S$, varies over time according to

$$dS = \mu_s dT + \sigma_s dW_s, \tag{6.1.2}$$

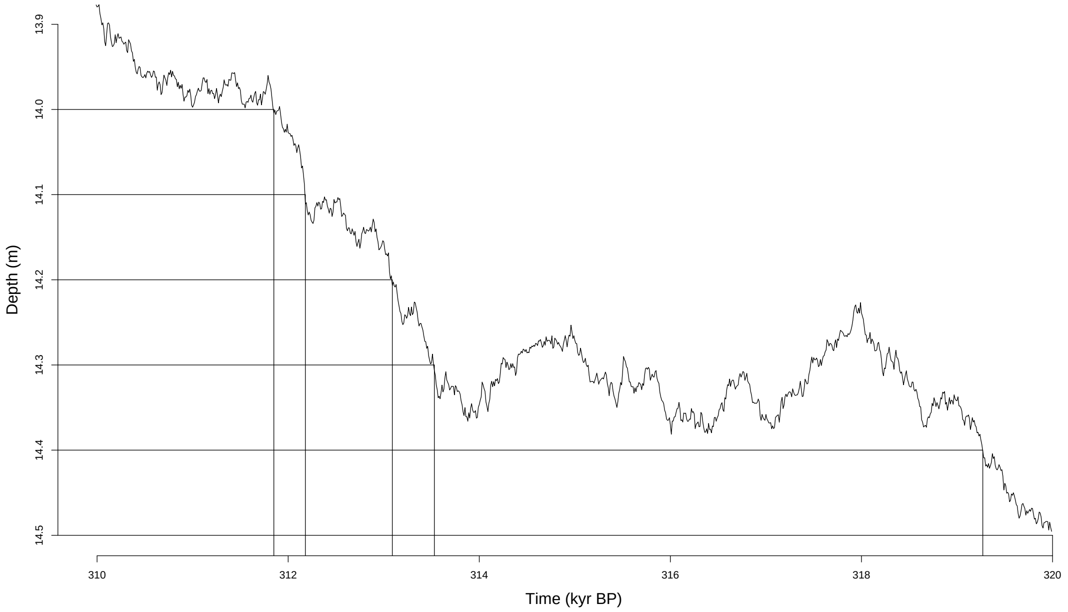

**Figure 6.1:** Demonstration of core sampling. The change in sediment over time has been plotted. The horizontal lines are the sampling depths, and the vertical lines are the sampled times.

where $\mu_s > 0$ is the mean sediment accumulation rate, $W_s$ is a standard Brownian motion, and $\sigma_s$ scales the Brownian motion. In this model, sediment accumulates over time, but can undergo periods of erosion as a result of stochastic perturbations. Since we are interested in the history of the sediment, we model sediment variation backwards in time. The reverse process is governed by the equation

$$dS = -\mu_s dT + \sigma_s dW_s, \tag{6.1.3}$$

where $dT > 0$ is now a step backwards in time. We set $T = 0$ to be the present, $T < 0$ to be times before present (BP), and $S(T = 0) = 0$ to be the present level of sediment.

Deriving an age model from a sediment accumulation model requires consideration as to how a dataset is constructed from a sediment core. A dataset is constructed by taking a series of measurements, $Y_{1:M}$, at core depths $H_{1:M}$. We assume that the final observation in time, $Y_M$, is sampled from the top of the core, so that $H_M = 0$. Consider a realisation of Equation 6.1.2, shown in Figure 6.1. It can be seen that when a core is sampled at depth $H_m$, the recorded climate information corresponds to the latest time at which $S = -H_m$ (noting that a positive change in depth corresponds to a negative change in the level of sediment). If the sediment accumulated beyond this level at a previous time, then the information it contained has been eroded away. Hence, considering the problem backwards in time, we have a first passage time problem. Given depth $H_m$ corresponds to time $T_m$, the distribution of $T_{m-1}$, the time at which $S$ first passes $H_{m-1}$, follows the inverse Gaussian distribution (also known as the Wald distribution)

[91],

$$T_{m-1}|T_m \sim T_m - IG\left(\frac{H_{m-1}-H_m}{\mu_s}, \frac{(H_{m-1}-H_m)^2}{\sigma_s^2}\right). \tag{6.1.4}$$

Combined with the initial condition $\pi(T_M) = \delta_{T_M}(0)$, the age model is fully described. However, in order to include the observation times in the target distribution of the particle filter, we require a model going forward in time. This can be achieved with an application of Bayes theorem, so that

$$\pi(T_m|T_{m-1}) = \frac{\pi(T_{m-1}|T_m)\pi(T_m)}{\pi(T_{m-1})}, \tag{6.1.5}$$

where

$$T_m \sim -IG\left(\frac{H_m}{\mu_s}, \frac{H_m^2}{\sigma_s^2}\right) \tag{6.1.6}$$

is simply an application of Equation 6.1.4 conditional on $T_M = 0$. This model is not entirely realistic. In practice, sediment is affected by a number of post-depositional effects, such as bioturbation (mixing of sediment by plants and animals), sediment shifts, and core compaction. We consider the inclusion of core compaction in the next section, but do not consider additional modelling extensions in this chapter.

The observation model also needs to be extended to include information regarding the observation times. An example that we have previously discussed (see Section 1.1 and Section 5.4) is the Brunhes-Matuyama (BM) reversal, which occurred $780 \pm 2$ kyr BP [17]. In this chapter we use the BM reversal to define the start of the dataset. Hence, the initial observation, $\boldsymbol{Y}_1$, is extended to include the 780 kyr BP estimate, and the observation model is extended as

$$\boldsymbol{Y}_1 \sim \left(\begin{array}{c} \mathcal{N}\left(D + \boldsymbol{CX}, \sigma_Y^2\right) \\ \mathcal{N}\left(-780000, 2^2\right) \end{array}\right). \tag{6.1.7}$$

Beginning the dataset at the BM reversal is not necessary, but offers two benefits. Firstly, since the BM reversal occurred later than the mid-Pleistocene transition, we can continue using phenomenological models that were designed to capture the behaviour of the 100 kyr cycle. If the dataset was extended, we would need to extend the phenomenological models to incorporate the mid-Pleistocene transition. Secondly, if the BM reversal occurred part way along our dataset, then it would be a possible source of degeneracy in the particle filter. Hence, we would need to design proposal distributions conditioned on the time of the BM reversal.

### 6.1.2 Compaction Adjustment

The above model assumes that sediment remains undisturbed once deposited, other than the possibility of being eroded away over time. In reality, sediment is subject to numerous post-depositional effects. In particular, core compaction is where sediment is compressed due to the weight of the sediment above it [92]. As we go deeper, the volume, and hence weight, of above sediment increases, and so the amount of compaction also tends to increase with depth. Hence, at large depths, a specified depth interval will reflect a larger change in time than at shallower depths. Using a linear model that does not account for compaction will, therefore, typically give biased estimates.

A consequence of core compaction is the expulsion of pore water from the sediment. This can be seen in the porosity profile of the core, in which porosity tends to decrease with depth. In the construction of H07, simple models of porosity were tuned using porosity measurements to evaluate the level of compaction in each sediment core [4, 93]. We take a similar approach in this chapter. Phenomenological models of porosity suggest a linear decline at small depths, and an exponential decline at large depths [92]. Since we use data from sediment cores over only 780 kyr (corresponding to depths $< 40$ m), we model the porosity using the linear model

$$\phi_m = \phi_0 - cH_m, \tag{6.1.8}$$

where $\phi_m$ is the porosity at depth $H_m$. This model has two parameters that need to be estimated: The intercept, $\phi_0$, and the gradient, $c$. In H07 these parameters are estimated by finding the line of best fit for the porosity measurements over 400 m. Here, we incorporate them into our parameter estimation framework. Note that we use no measurements of porosity, as measurements are sparse, noisy, and not available in many cores. In principle, were porosity measurements available, they could be included by further extending the state vector and observation model. However, it should be emphasised that we are not aiming to model porosity explicitly, but rather to use a phenomenological model of porosity to adjust the linearity assumption in our age model.

Following the approach in [93], we apply a compaction correction based on conservation of dry sediment volume, so that

$$\hat{H}_m = \frac{1-\phi_m}{1-\phi_0} H_m, \tag{6.1.9}$$

where $\hat{H}_m$ is the uncompacted depth of observation $m$. Note that $\hat{H}_M = H_M = 0$. Substituting in Equation 6.1.8 yields

$$\hat{H}_m = H_m + \frac{c}{1-\phi_0} H_m^2. \tag{6.1.10}$$

Therefore, a depth interval $H_{m-1} - H_m$ increases to

$$\begin{aligned} \hat{H}_{m-1} - \hat{H}_m &= H_{m-1} + \frac{c}{1-\phi_0} H_{m-1}^2 - \left( H_m + \frac{c}{1-\phi_0} H_m^2 \right) \\ &= (H_{m-1} - H_m) + \frac{c}{1-\phi_0} (H_{m-1}^2 - H_m^2). \end{aligned} \tag{6.1.11}$$

Applying our previous age model to the uncompacted depth scale gives the compaction adjusted age model

$$T_{m-1} | T_m \sim T_m - IG \left( \frac{\hat{H}_{m-1} - \hat{H}_m}{\mu_s}, \frac{(\hat{H}_{m-1} - \hat{H}_m)^2}{\sigma_s^2} \right), \tag{6.1.12}$$

so that, on average, there is a larger change in time for a specified depth interval at large depths than small depths.

### 6.1.3 Extending The Particle Filter

Having developed an age model, we are now interested in inferring the observation times using the statistical methodology developed in this thesis. We achieve this by modifying the particle filter to target the extended distribution $\pi\left(T_{1:M}, \boldsymbol{X}_{1:M} \mid \boldsymbol{Y}_{1:M}, \boldsymbol{\theta}\right)$. The proposal distribution in each iteration is extended to follow a two step process. First, $T_m$ is sampled from a new proposal distribution, $b_m\left(T_m \mid T_{m-1}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)$, and second, $X_m$ is sampled from $r_m\left(X_m \mid T_{m-1}, T_m, X_{m-1}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)$, which is now conditional on the random variables $T_m$ and $T_{m-1}$. The pseudocode is presented in Algorithm 6.1.

The extended particle filter can be embedded in PMCMC or SMC$^2$, in order to jointly estimate the model parameters and observation times. SMC$^2$ can also be used to estimate the model evidence, allowing us to perform model comparison. We test this approach in a simulation study in the next section, and then apply it to real-world data in Section 6.3.

**Algorithm 6.1** Particle filter targeting $\pi\left(T_{1:M}, \boldsymbol{X}_{1:M} \mid \boldsymbol{Y}_{1:M}, \boldsymbol{\theta}\right)$.

```
for $k = 1, ..., N_X$ do
    Sample $T_m^{(k)} \sim b_m\left(T_m \mid \boldsymbol{Y}_m, \boldsymbol{\theta}\right)$.
    Sample $\boldsymbol{X}_m^{(k)} \sim r_m\left(\boldsymbol{X}_m \mid T_m^{(k)}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)$.
    Set the importance weight
        $\omega_m^{(k)} = \dfrac{\pi\left(T_m^{(k)} \mid \boldsymbol{\theta}\right) \pi\left(\boldsymbol{X}_m^{(k)} \mid T_m^{(k)}, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{Y}_m \mid T_m^{(k)}, \boldsymbol{X}_m^{(k)}, \boldsymbol{\theta}\right)}{b_m\left(T_m^{(k)} \mid \boldsymbol{Y}_m, \boldsymbol{\theta}\right) r_m\left(\boldsymbol{X}_m^{(k)} \mid T_m^{(k)}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)}.$
end for
Normalise the weights. For $k = 1, ..., N_X$
        $\Omega_1^{(k)} = \dfrac{\omega_1^{(k)}}{\sum_{i=1}^{N_X} \omega_1^{(i)}}.$
for $m = 2, ..., M$ do
    for $k = 1, ..., N_X$ do
        Sample ancestor particle index $a_{m-1}^{(k)}$ according to weights $\Omega_{m-1}^{(1:N_X)}$.
        Sample $T_m^{(k)} \sim b_m\left(T_m \mid T_{m-1}^{\left(a_{m-1}^{(k)}\right)}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)$.
        Sample $\boldsymbol{X}_m^{(k)} \sim r_m\left(\boldsymbol{X}_m \mid T_{m-1}^{\left(a_{m-1}^{(k)}\right)}, T_m^{(k)}, \boldsymbol{X}_{m-1}^{\left(a_{m-1}^{(k)}\right)}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)$.
        Extend particle trajectory $\left\{T_{1:m}^{(k)}, \boldsymbol{X}_{1:m}^{(k)}\right\} = \left\{\left(T_{1:m-1}^{(a_{m-1}^{(k)})}, T_m^{(k)}\right), \left(\boldsymbol{X}_{1:m-1}^{(a_{m-1}^{(k)})}, \boldsymbol{X}_m^{(k)}\right)\right\}$.
        Set the importance weight
    $\omega_m^{(k)} = \dfrac{\pi\left(T_m^{(k)} \mid T_{m-1}^{\left(a_{m-1}^{(k)}\right)}, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{X}_m^{(k)} \mid \boldsymbol{X}_{m-1}^{\left(a_{m-1}^{(k)}\right)}, T_m^{(k)}, T_{m-1}^{\left(a_{m-1}^{(k)}\right)}, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{Y}_m \mid \boldsymbol{X}_m^{(k)}, \boldsymbol{\theta}\right)}{b_m\left(T_m^{(k)} \mid T_{m-1}^{\left(a_{m-1}^{(k)}\right)}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right) r_m\left(\boldsymbol{X}_m^{(k)} \mid T_{m-1}^{\left(a_{m-1}^{(k)}\right)}, T_m^{(k)}, \boldsymbol{X}_{m-1}^{\left(a_{m-1}^{(k)}\right)}, \boldsymbol{Y}_m, \boldsymbol{\theta}\right)}.$
    end for
    Normalise the weights. For $k = 1, ..., N_X$
        $\Omega_m^{(k)} = \dfrac{\omega_m^{(k)}}{\sum_{i=1}^{N_X} \omega_m^{(i)}}.$
end for
```

| Parameter | True Value | Prior Distribution |
|---|---|---|
| $\beta_0$ | 0.65 | $\mathcal{N}\left(0.4, 0.3^2\right)$ |
| $\beta_1$ | 0.2 | $\mathcal{N}\left(0, 0.4^2\right)$ |
| $\beta_2$ | 0.5 | $exp\left(1/0.5\right)$ |
| $\delta$ | 0.5 | $exp\left(1/0.5\right)$ |
| $\alpha$ | 11 | $\Gamma\left(10, 2\right)$ |
| $\gamma_P$ | 0.2 | $exp\left(1/0.3\right)$ |
| $\gamma_C$ | 0.1 | $exp\left(1/0.3\right)$ |
| $\gamma_E$ | 0.3 | $exp\left(1/0.3\right)$ |
| $\sigma_1$ | 0.2 | $exp\left(1/0.3\right)$ |
| $\sigma_2$ | 0.5 | $exp\left(1/0.5\right)$ |
| $\sigma_Y$ | 0.1 | $exp\left(1/0.1\right)$ |
| $D$ | 4.2 | $\mathcal{U}\left(3, 5\right)$ |
| $C$ | 0.8 | $\mathcal{U}\left(0.5, 2\right)$ |
| $\mu_s$ | $4.5 \times 10^{-5}$ | $\Gamma\left(180, 1/4 \times 10^6\right)$ |
| $\sigma_s$ | $2 \times 10^{-3}$ | $exp(500)$ |
| $\phi_0$ | 0.8 | $\beta(45, 15)$ |
| $c$ | $3.5 \times 10^{-4}$ | $exp(4000)$ |
| $X_1(t_1)$ | $-1$ | $\mathcal{U}\left(-1.5, 1.5\right)$ |
| $X_2(t_1)$ | $-1.5$ | $\mathcal{U}\left(-2.5, 2.5\right)$ |

**Table 6.1:** List of parameters used to generate data for the simulation study, and the associated prior distributions used in the statistical analysis.

## 6.2 Simulation Study

We design a simulation study to assess our ability to infer observation times from a sediment core. Observation times were drawn from a core of length 32 m, sampled at 0.1 m intervals, giving $M = 321$ observations. Observations were generated using CR14-a, described in Section 2.3.2, and the observation process described in Section 2.3.3. The first observation contains a noisy measurement of the true time, where the noise is sampled from a Gaussian distribution with mean zero and a standard deviation of 2 kyr, representing the dating of a geomagnetic reversal. The chosen parameter values are shown in Table 6.1, along with the prior distributions.

### 6.2.1 Results on the Simulation Study Data

We run the SMC$^2$ algorithm with $N_X = 1000$ state particles and $N_\theta = 1000$ parameter particles. The proposal distribution in the PMCMC chain is an independent Gaussian distribution with mean and covariance equal to that of the current sample. The PMCMC chain length in the resampling stages is set to 10. In the particle filter, we propose $T_1$ from a Gaussian distribution centred on the observation time of the simulated geomagnetic reversal with a standard deviation of 2 kyr. In later iterations we propose new times by drawing $T_m$ from

$$T_m \sim T_{m-1} + IG\left(\frac{\hat{H}_{m-1} - \hat{H}_m}{\mu_s}, \frac{(\hat{H}_{m-1} - \hat{H}_m)^2}{\sigma_s^2}\right), \tag{6.2.1}$$

and sample $\boldsymbol{X}_m$ using the proposal distribution developed in Section 3.4.

The sequence of estimated 95% highest density region (HDR) intervals for the observation times are shown in Figure 6.2. Since it is difficult to discern features along the time axis, we also show the 95% HDR intervals with the trend removed. Observe that the generated sequence of observation times differs greatly from a linear age-depth relationship. In particular, there is a large period of time in which little sediment is deposited in the middle of the dataset. Despite this, the true observation times are in regions of high posterior probability density throughout the dataset, showing that we are able to recover the observation times. The standard deviation of the filtering distribution oscillates over time. This could be a result of CR14-a being more sensitive to the astronomical forcing (which provides time information) at some points more than others. There is a large spike in the posterior variance in the middle of the record, corresponding to the period in which little sediment was deposited.

The generated observations, and a sequence of 95% HDR intervals for the state of the system are shown in Figure 6.3. The true values for both the observable and unobservable state variables lie within the 95% HDR intervals throughout the core. As would be expected, the posterior variance for the unobservable state is larger than the observable state, particularly when the system switches between glacial and interglacial periods.

The marginal posterior distributions of the parameters are shown in Figure 6.4. The true parameter values lie in regions of high posterior probability density. The posterior distribution of the gradient of the porosity model, $c$, supports a wide range of values, suggesting that it is difficult to determine porosity trends using our modelling approach.

We repeat the analysis using an unforced variant of CR14-a. The Bayes factor in favour of the forced model against the unforced model is $\mathcal{O}(10^9)$. Even with the age model uncertainty, the data strongly support the forced model.

To conclude, we have shown that in an ideal setting, it is possible to jointly perform parameter estimation, observation age estimation, and model comparison. Somewhat to our surprise, we are able to determine the influence of the astronomical forcing, even with the added uncertainty of the age model.

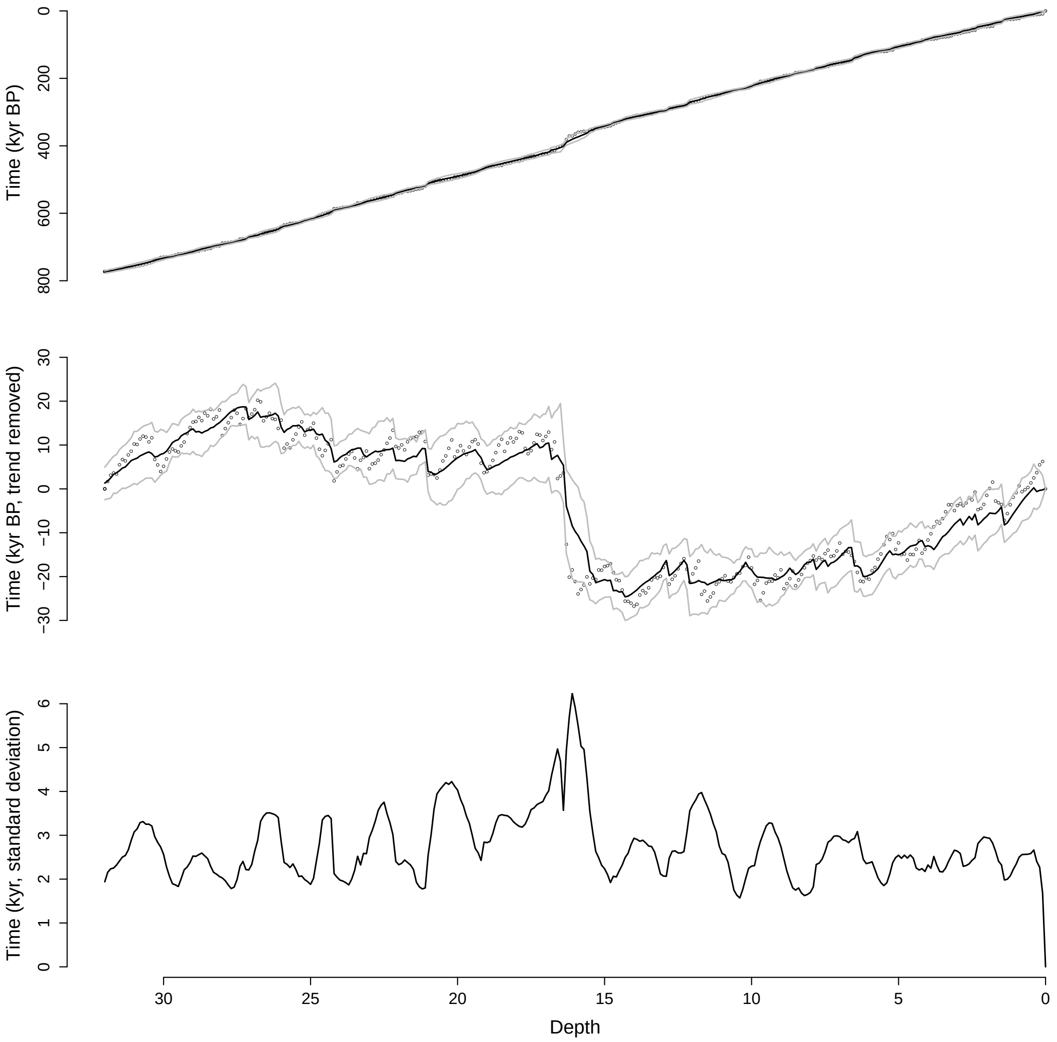

**Figure 6.2:** Top: Sequence of 95% HDR intervals for the observation times in the simulation study. Points show the true values. The black line shows the mean values of the marginal posterior distributions of the observation times, and grey lines show the 95% HDR intervals. Middle: Sequence of 95% HDR intervals for the observation times with the trend removed. The majority of the points lie within the credible regions. Note that a linear sediment-depth relationship would be a line through zero. Bottom: Sequence of standard deviations of the marginal posterior distributions of the observation times.

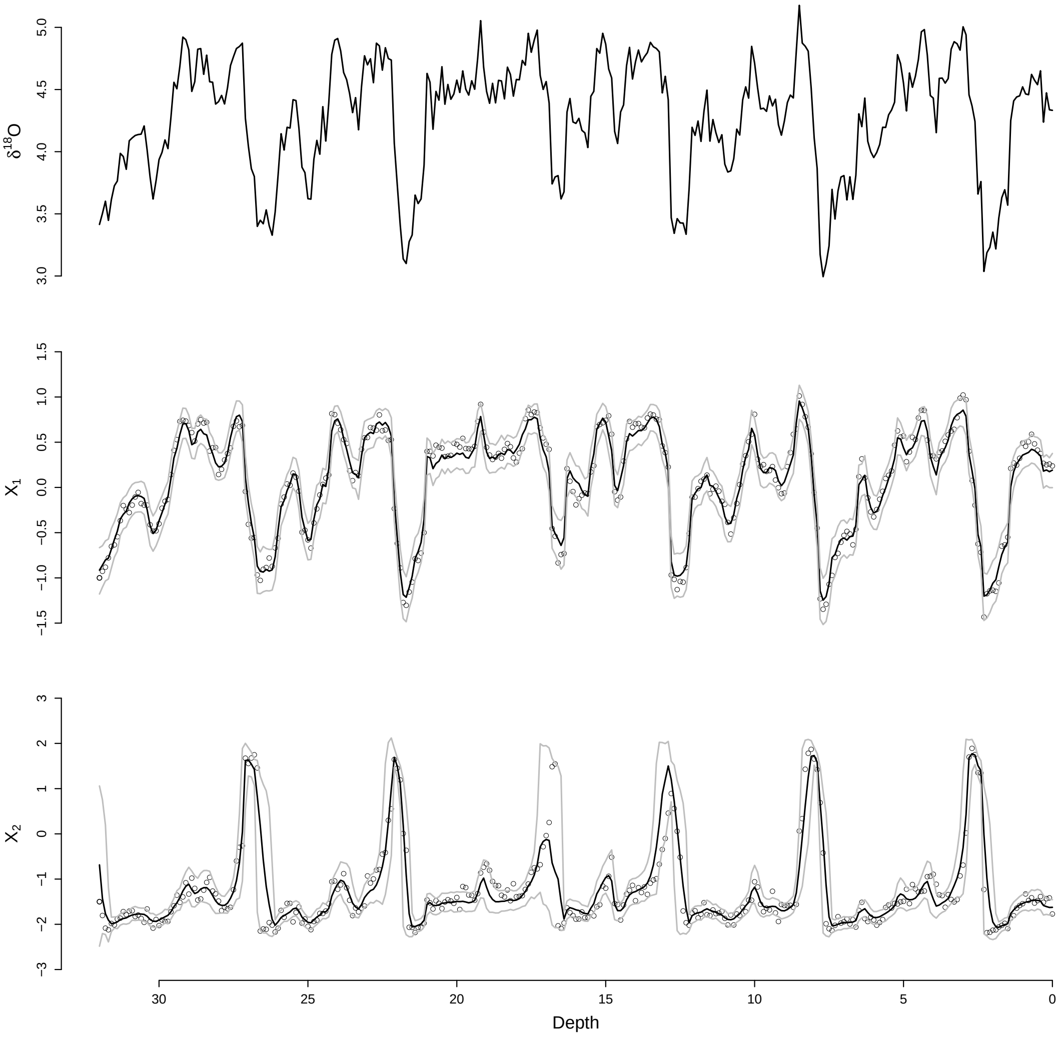

**Figure 6.3:** Top: Generated dataset in the simulation study. Middle and bottom: Sequence of 95% HDR intervals for the state of CR14-a for the observable and unobservable states respectively. Points show the true values. The black line shows the mean values of the marginal posterior distributions of the states, and grey lines show the 95% HDR intervals. The majority of the points lie within the credible regions.

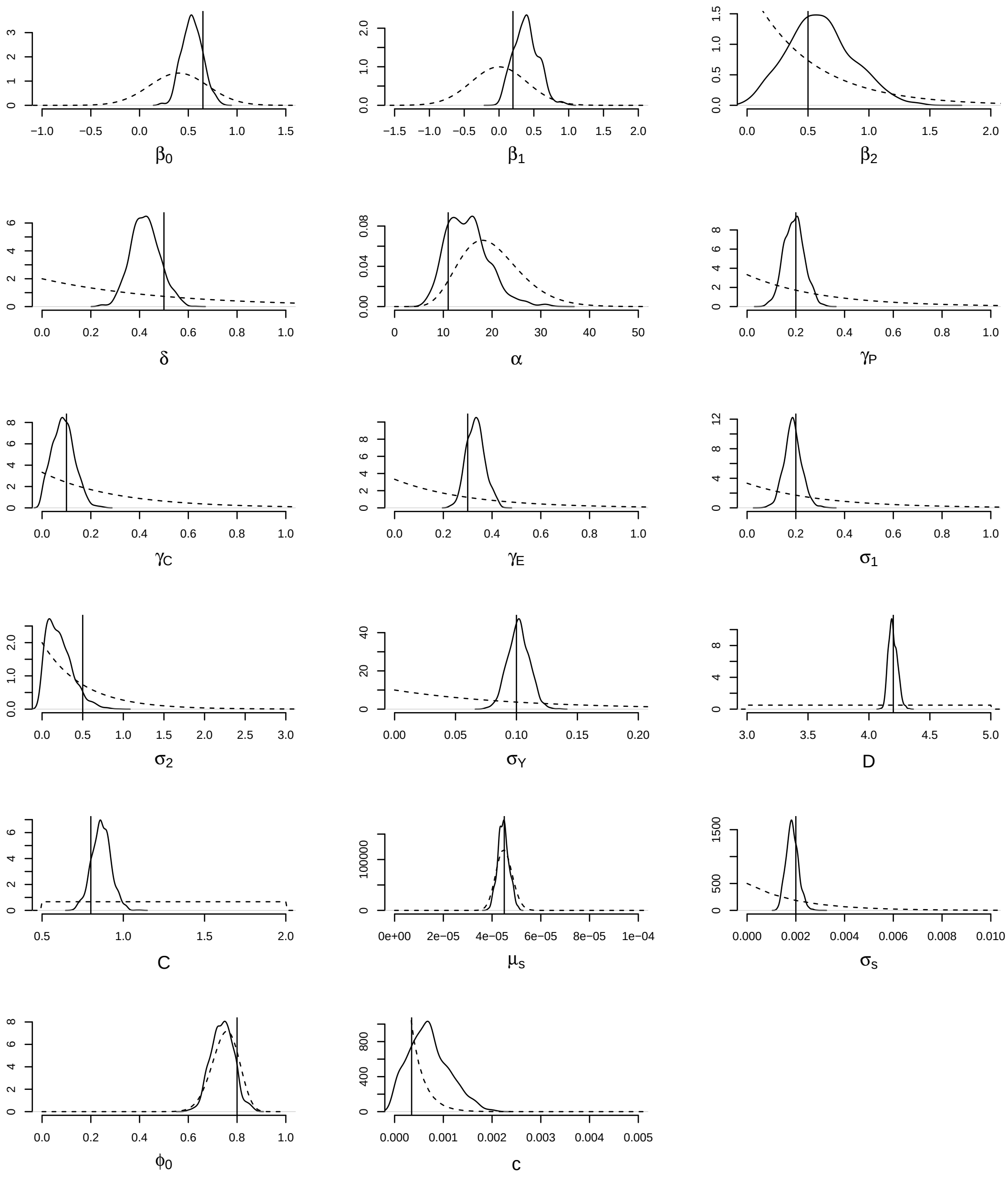

**Figure 6.4:** Marginal posterior distributions of the parameters of CR14-a in the simulation study. Dashed lines show the prior distributions, and vertical lines show the true values. The true values lie in regions of high posterior probability density.

## 6.3 Dating ODP677 and ODP846

Having demonstrated our methodology in a simulation study, we now analyse two real-world datasets. We use ODP677 [84] in order to compare the posterior distributions of the parameters with those obtained in Chapter 4. Additionally, we use ODP846 [94], in which compaction has previously been measured using a linear porosity model [4, 93], giving a point of comparison. The BM reversal is identifiable in both ODP677 (at 30.4 m) and ODP846 (at 28.7 m). The number of observations since the BM reversal are 363 and 308, for ODP677 and ODP846 respectively.

### 6.3.1 Results on ODP677 and ODP846

We run the SMC$^2$ algorithm with $N_X = 1000$ state particles and $N_\theta = 1000$ parameter particles for each dataset, using the same settings as in the simulation study.

The sequence of 95% HDR intervals for the observation times are shown in Figure 6.5 for ODP677 and Figure 6.8 for ODP846. The dating uncertainties are larger than in the simulation study, and again oscillate over time. The posterior variance is larger in ODP846 (mean standard deviation of $\sim$6.5 kyr) than ODP677 (mean standard deviation of $\sim$3.5 kyr). These dating uncertainties are smaller than previous estimates [1, 4, 93]. Additionally, the most uncertain estimates are not necessarily at the mid-point between ACPs (such as the present, or geomagnetic reversals), which has previously been assumed [4, 93]. Rather, the ACPs only seem to affect the posterior variance within a few metres. The 95% HDR intervals obtained for the observation times of ODP677 seem consistent with the LR04 estimates, which lie in credible intervals throughout the sediment core. On the other hand, the H07 estimates deviate greatly from our estimates between 11 m and 16 m. The 95% HDR intervals obtained for ODP846 seem consistent with both estimates, partly due to the increased posterior variance. However, the LR04 estimates are closer to the posterior means. A notable discrepancy with LR04 is the rapid decrease in age at 26 m, suggesting little sediment was deposited over a large time period. Comparing the ODP846 dataset to the LR04 reconstruction, shown in Figure 1.1, suggests that a section of data is missing at this depth, validating the LR04 estimate. Since we are dating a single core, rather than correlating multiple cores, recognising when data are missing is a more difficult problem. For both datasets it can be seen that using a linear age-depth relationship, even when accounting for compaction, will lead to poor estimates of observation times.

The sequence of 95% HDR intervals for the state of the system are shown in Figure 6.6 for ODP677, and Figure 6.9 for ODP846, along with the observations. The two sets of state reconstructions are broadly similar, but $X_2$ switches between positive and negative values more often in ODP846 than ODP677, showing that the state reconstruction can differ somewhat between datasets. As with the estimates of the observation times, the posterior variance is greater in ODP846 than ODP677.

The marginal posterior distributions of the parameters are shown in Figure 6.7 for ODP677, and Figure 6.10 for ODP846. The two sets of posterior distributions are similar, but there are notable differences for some of the parameters. For example, larger values for the obliquity scaling term, $\gamma_E$, are favoured for ODP677, which indicates a stronger synchronisation to the astronomical forcing in ODP677 than ODP846. However, the posterior distributions for both datasets have little mass around $\gamma_E = 0$ and $\gamma_P = 0$, suggesting that the astronomical forcing adds some explanatory power. This explains the strong agreement with the LR04 age model, which uses the astronomical forcing to constrain observatiom times, and gives a possible explanation for the smaller variance in the filtering results for ODP677. Specifically, since the astronomical forcing provides time information, and attracts the system into specific regions of state-space, we expect the uncertainty around these quantities to be smaller the more strongly the system is forced. For both sets of data, large values of $c$ are supported, indicating large changes in porosity, and hence, a large amount of compaction in the sediment over $\sim$30 m. The value used in H07 was $\sim 5 \times 10^{-4}$, which seems to be consistent with the porosity profile over 200 m [4, 93]. The large values indicated by our posterior distributions seem to be physically unrealistic in comparison. However, we again note that our porosity model is used as a phenomenological approach to modify a linear sediment model, rather than to model porosity explicitly. Comparing the marginal posterior distributions of the parameters obtained for ODP677 with Figure 4.5 shows an increase in posterior variance due to the added uncertainty of having unknown observation times. Additionally, there are some notable differences between the two sets of posterior distributions. For example, larger values of $\delta$ are supported when the observation times have been fixed using the depth-derived age model. Hence, fixing the observation times can alter inference about the model parameters.

Repeating the experiment using an unforced variant of CR14-a yields Bayes factors in favour of the forced model against the unforced model of $\mathcal{O}(10^5)$ in ODP677, and $\mathcal{O}(1)$ in ODP846. In other words, the forced model is more strongly supported by ODP677, but both models are equally well supported by ODP846.

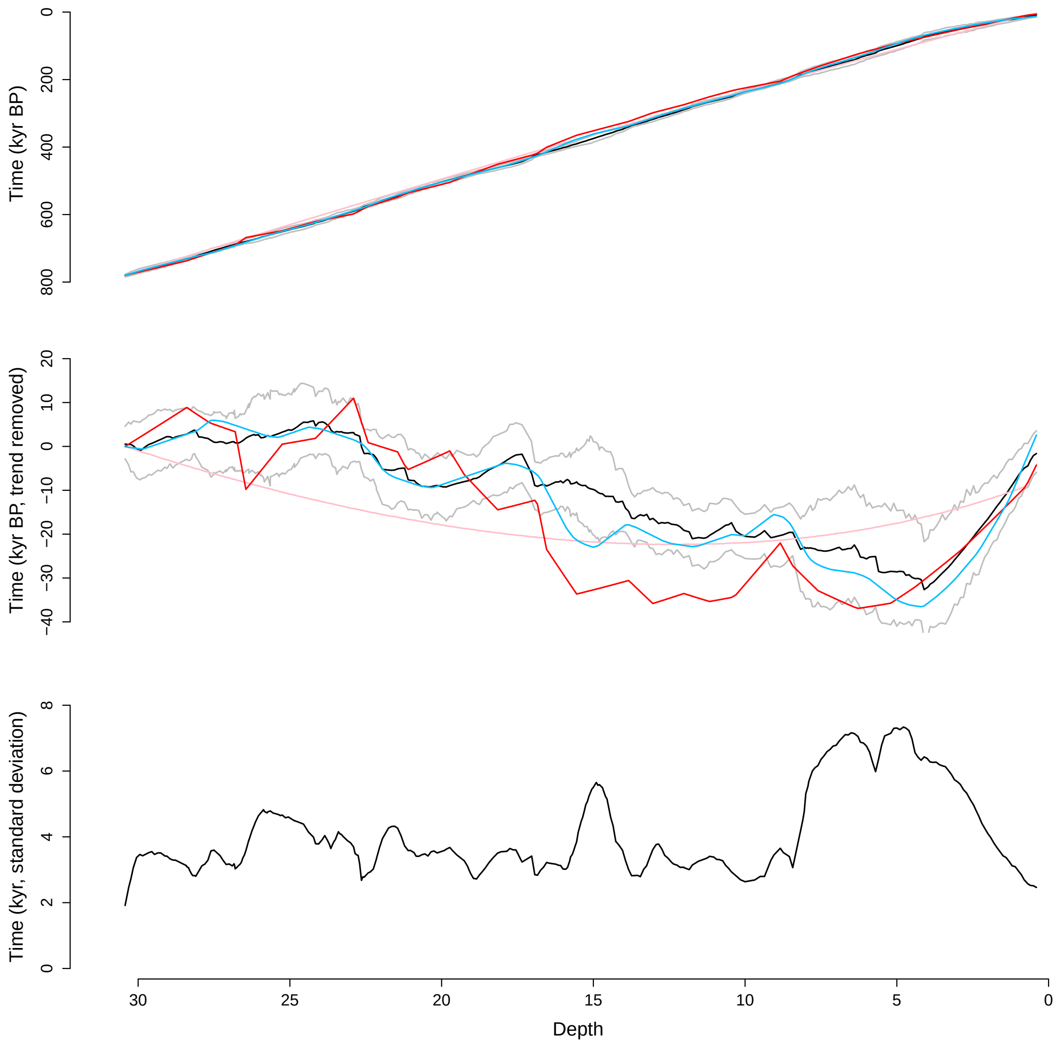

**Figure 6.5:** Top: Sequence of 95% HDR intervals for the observation times for ODP677. The black line shows the mean values of the marginal posterior distributions of the observation times, and grey lines show the 95% HDR intervals. The pink line shows the linear sediment-age relationship accounting for compaction used in [4], and the red line shows the estimates from complete depth-derived age model. The blue line shows the estimates from the LR04 stack [1]. Middle: Sequence of 95% HDR intervals for the observation times with the trend removed. Note that a linear sediment-depth relationship would be a line through zero. Our estimates agree strongly with the LR04 estimates, but there are large deviations with the H07 estimates between 9 and 16 metres. Bottom: Sequence of standard deviations of the marginal posterior distributions of the observation times.

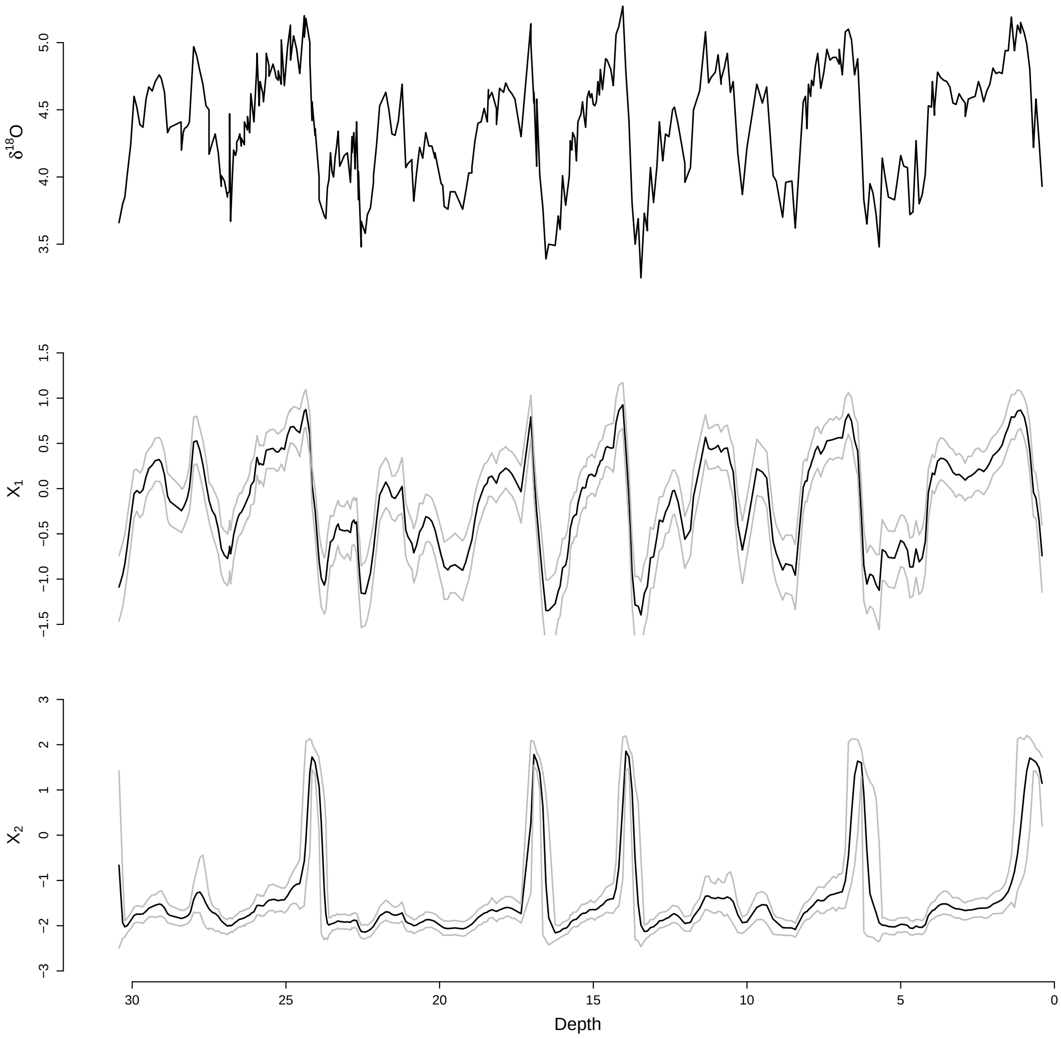

**Figure 6.6:** Top: The ODP677 dataset. Middle and bottom: Sequence of 95% HDR intervals for the state of CR14-a for the observable and unobservable states respectively. The black line shows the mean values of the marginal posterior distributions of the states, and grey lines show the 95% HDR intervals.

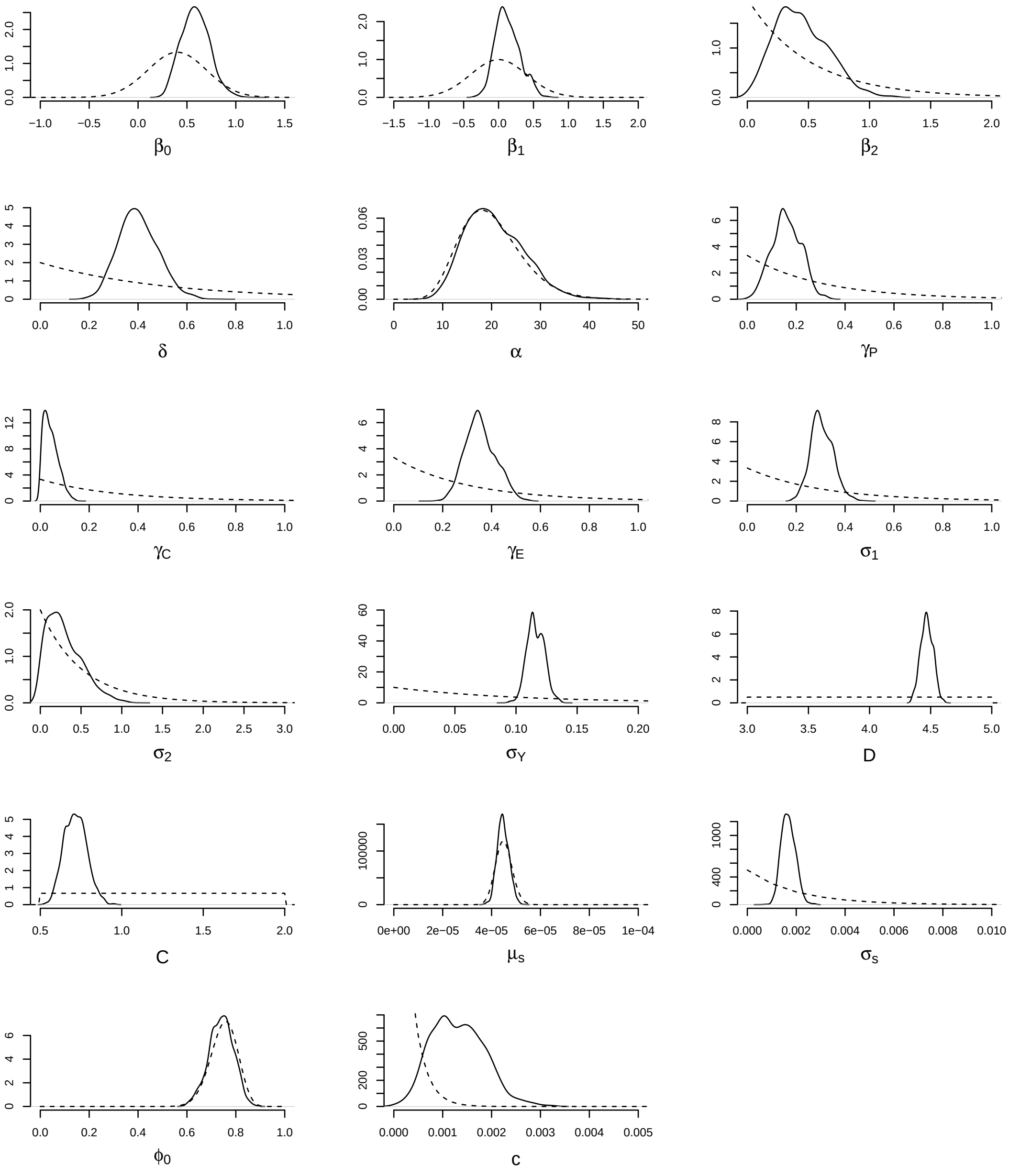

**Figure 6.7:** Marginal posterior distributions of the parameters of CR14-a for ODP677. Dashed lines show the prior distributions.

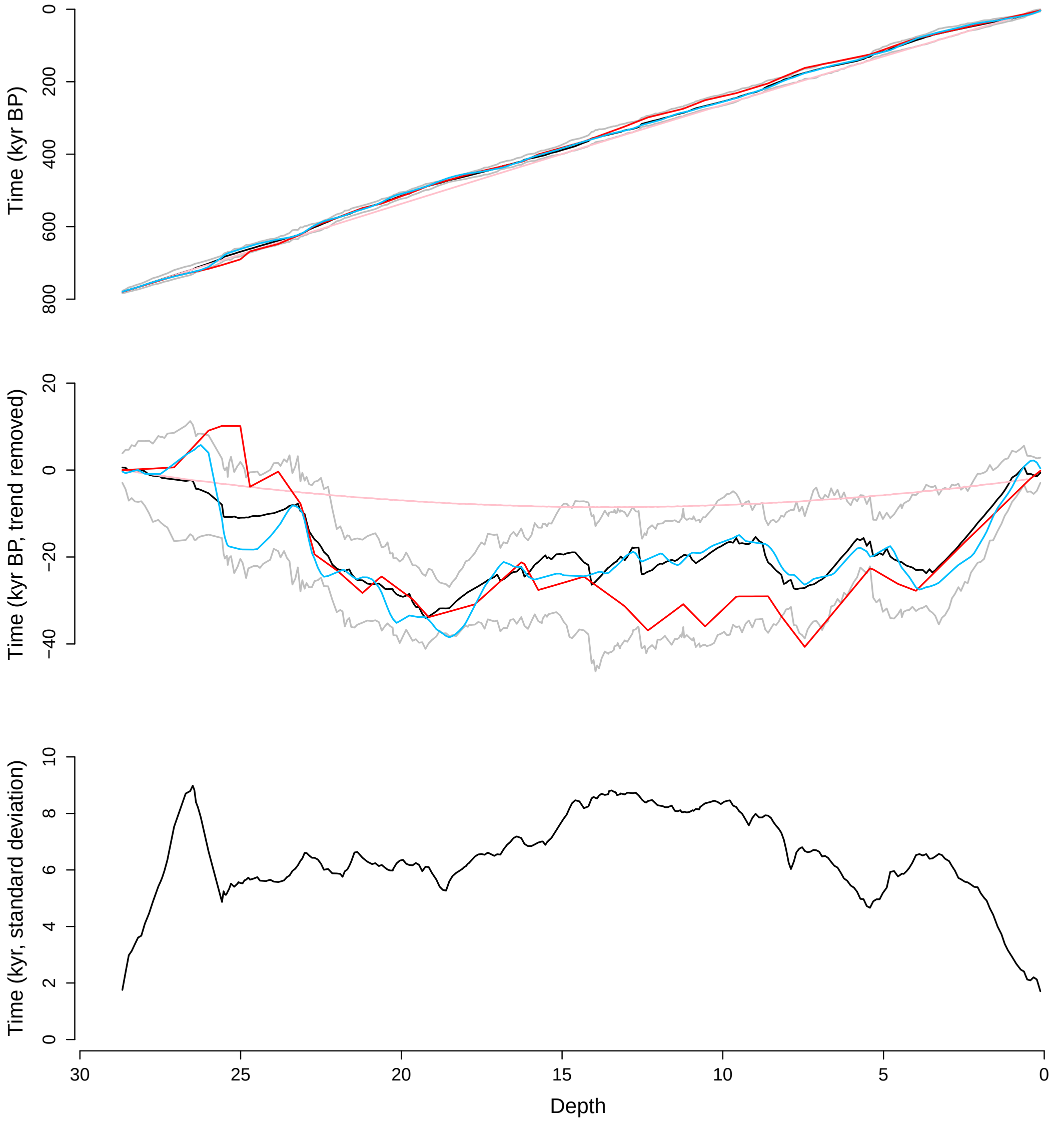

**Figure 6.8:** Top: Sequence of 95% HDR intervals for the observation times for ODP846. The black line shows the mean values of the marginal posterior distributions of the observation times, and grey lines show the 95% HDR intervals. The pink line shows the linear sediment-age relationship accounting for compaction used in [4], and the red line shows the estimates from complete depth-derived age model. The blue line shows the estimates from the LR04 stack [1]. Middle: Sequence of 95% HDR intervals for the observation times with the trend removed. Note that a linear sediment-depth relationship would be a line through zero. Our estimates agree strongly with both the LR04 and H07 estimates. Bottom: Standard deviation of the posterior distribution.

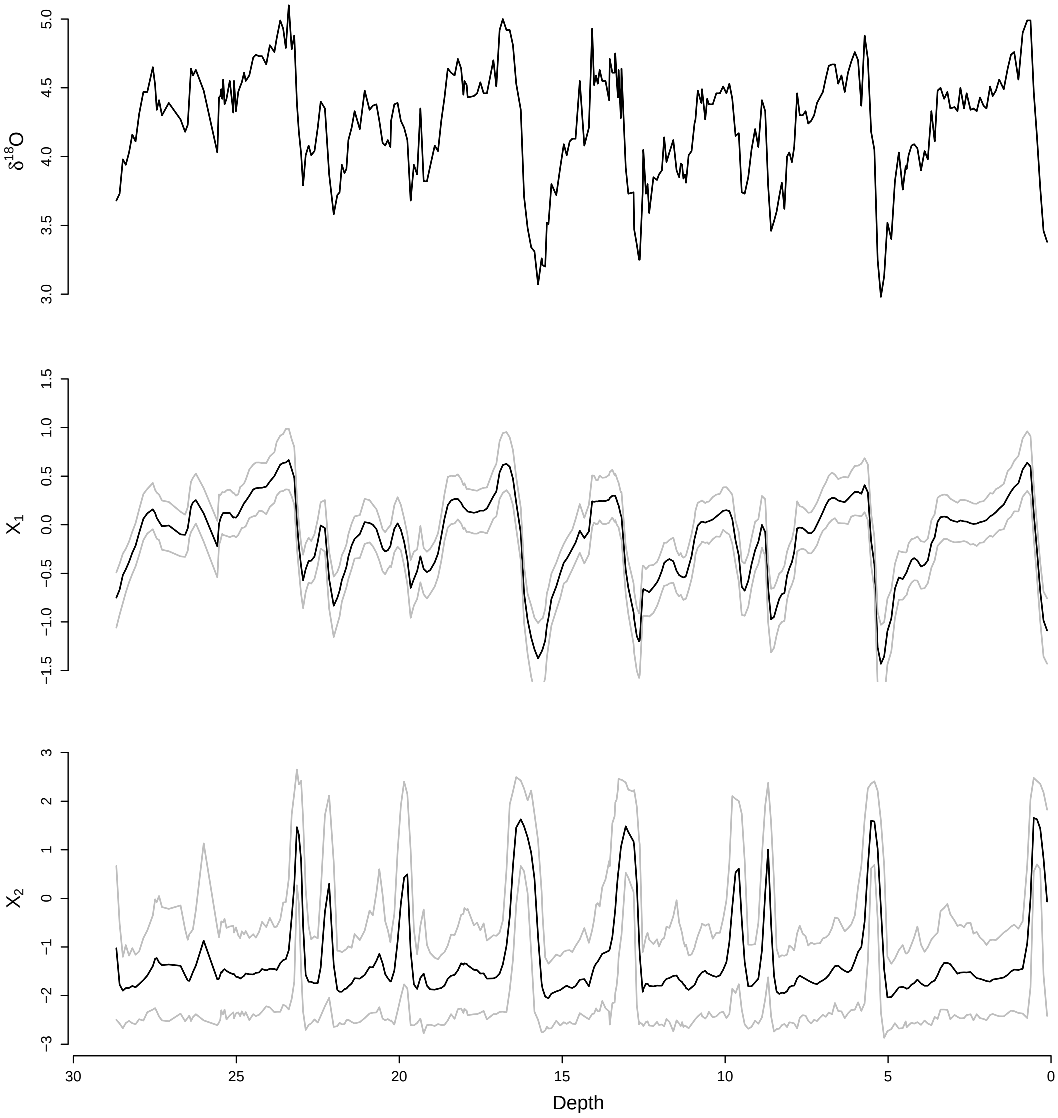

**Figure 6.9:** Top: The ODP846 dataset. Middle and bottom: Sequence of 95% HDR intervals for the state of CR14-a for the observable and unobservable states respectively. The black line shows the mean values of the marginal posterior distributions of the states, and grey lines show the 95% HDR intervals.

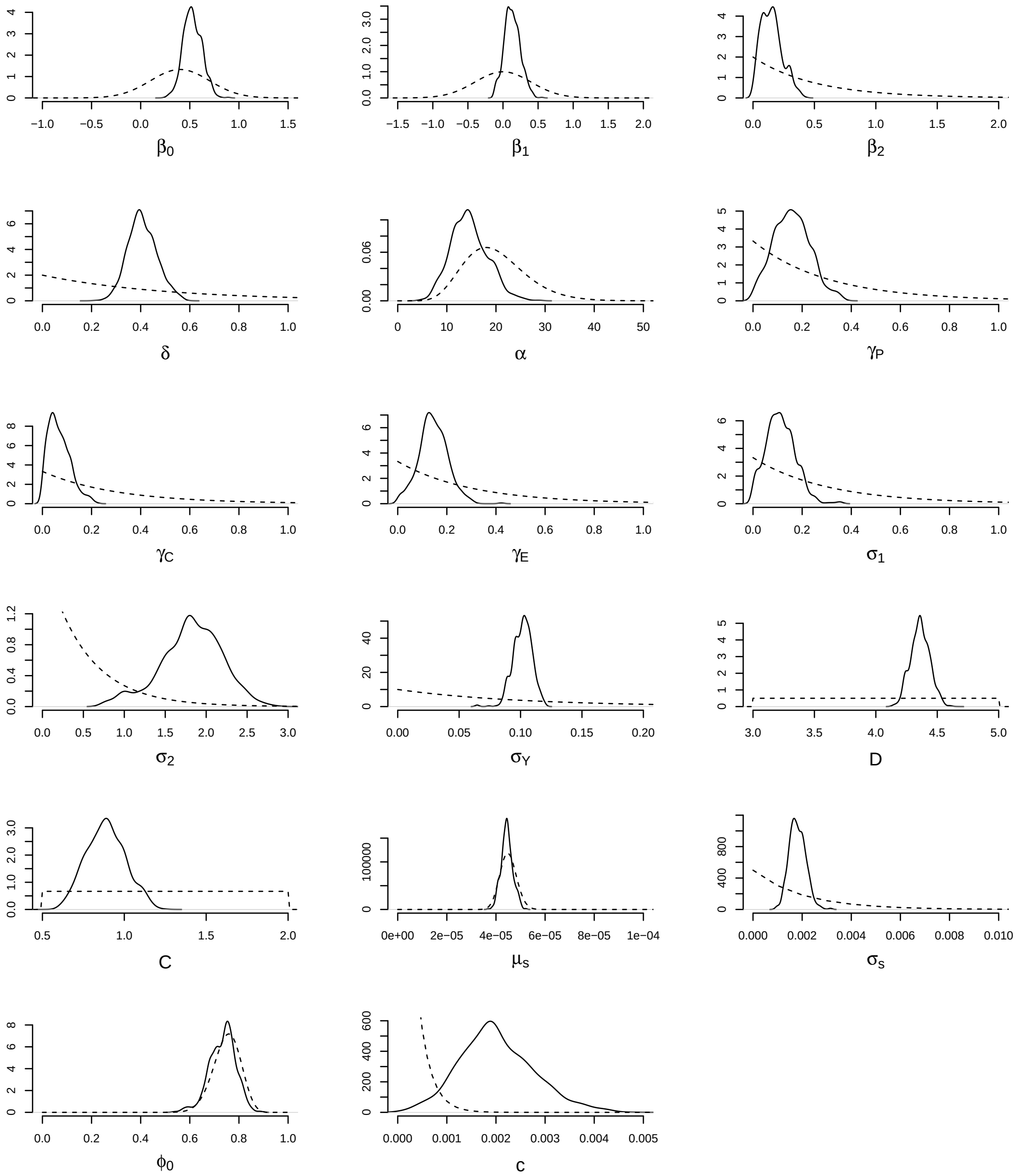

**Figure 6.10:** Marginal posterior distributions of the parameters of CR14-a for ODP846. Dashed lines show the prior distributions.

## 6.4 Chapter Summary

We have introduced a novel age model and extended our inference approach to quantify the chronological uncertainty in dating sediment cores. Our approach performed well in a simulation study, and we were able to detect the influence of the astronomical forcing with the added dating uncertainty.

We applied our methodology to two real-world datasets, and showed that our results are consistent with previous age models. Whereas previous age models determined sources of uncertainty, and estimated the uncertainty for each component, we were able to quantify the age model uncertainty in a principled statistical framework. A forced model was strongly supported by the data in favour of an unforced model when using ODP677, but not when using ODP846, showing significant variation between datasets. Therefore, when testing hypotheses about climate dynamics, multiple datasets should be considered. Ideally, observations from multiple cores should be used in a single experiment. A discussion about extending our work in this direction is given in Chapter 7.

CHAPTER 7

# Conclusions

Studying the dynamics of the palaeoclimate is a difficult problem. The climate is a highly complex system. Modelling every process on a sufficiently refined spatio-temporal scale to reconstruct the climate over millions of years is far beyond our capabilities. Even the most advanced numerical simulators in use today require tuning on palaeoclimate data, in order to reconstruct the history of the climate. Many of the methods in use today, such as tuning parameters by hand, make poor use of the information from palaeoclimate datasets, and usually lead to poor characterisation of the uncertainty in the simulator predictions. In this thesis we have focussed on using relatively simple models, termed phenomenological models, which are simple enough to enable data assimilation to be performed through statistically rigorous approaches, characterising the uncertainty arising from the data and modelling approaches.

Even with such relatively simple models, performing statistical inference in this setting is a challenging problem. The likelihood of these models is intractable, requiring state of the art statistical methods in order to perform inference. Many of these methods were published only very recently, and remain relatively untested in challenging real-world settings. A consistent theme maintained in this thesis was assessing the performance of our inference approaches in simulation studies, in which the correct model, and model parameters, are known. The fact that they are shown to perform well in this idealised setting lends confidence to the results obtained on real-world data.

There are a number of key scientific contributions from this thesis. The first is that palaeoclimate records seem to contain more information than is typically assumed. The prevailing viewpoint is that palaeoclimate data do not contain enough information to select between competing phenomenological models [24]. In our model comparison experiments, it is often the case that the data support some models more strongly than others. In some cases, a limitation in drawing strong conclusions has been the Monte Carlo variation encountered when estimating Bayes factors, which is a direct result of having limited computational resources. Given that our current implementations of SMC$^2$ require between 3 and 8 days runtime, depending on the model used and whether the age model is included, it seems unlikely that the Monte Carlo variation can be significantly reduced whilst still obtaining Bayes factor estimates within a reasonable time. However, the constant growth in the availability of computer power offers the possibility that these experiments can be repeated, with stronger conclusions, within a few years. With SMC methods, it is also possible to employ parallel processing. For example, we were able to obtain a $25\times$ speedup of SMC$^2$ using a Tesla K20 GPU.

The second contribution is that we have demonstrated that estimating observation times independently of any further analysis leads to strong sensitivity in the results of future analyses on the assumptions made when estimating the observation times. For example, model comparison experiments on data that have been dated using astronomical tuning assumptions consistently found strong support in favour of forced models. This was not the case when the age estimates were obtained without astronomical tuning assumptions.

The third is that we have developed a method to jointly estimate the model parameters, model evidence, and observation times, in individual sediment cores. Whilst our approach gave age estimates consistent with previous age models, our approach fully characterises the age model uncertainty, unlike previous methods. We have shown in a model selection experiment that one palaeoclimate dataset more strongly supports a forced phenomenological model over an unforced phenomenological model. This reinforces recent findings that palaeoclimate data support Milankovitch theory [18, 86].

A summary of each chapter and the major conclusions is given below.

## 7.1 Thesis Summary

In Chapter 2 we discussed approaches to climate modelling, and justified the use of phenomenological models to study the glacial-interglacial cycle. The key concepts necessary to understand the phenomenological modelling approach, such as relaxation oscillators, the astronomical forcing, and synchronisation, were explained. We demonstrated that phenomenological models, embedded in an SSM framework, enable us to perform inference on the dynamics of the glacial-interglacial cycle via the assimilation of proxy data from sediment cores. We highlighted the inference challenges of using this approach.

In Chapter 3 we compared ABC and particle filter approaches to performing parameter estimation in SSMs. ABC can be used in any situation in which the model can be simulated from, and is not restricted to SSMs. We gave an overview of ABC methods, leading up to recent developments, such as ABC-SMC, which reduce the computational cost in comparison to the original ABC rejection algorithm. The particle filter is an SMC algorithm targeting the posterior distribution of the state of an SSM, from which we can obtain unbiased likelihood estimates. We gave an overview of two pseudo-marginal approaches, PMCMC and SMC$^2$, that embed the particle filter in an MCMC algorithm and SMC algorithm respectively, in order to sample from the posterior distribution of the parameters. We described the statistical tools utilised by these inference methods, and designed a simulation study as a way to compare the relative performance between the different approaches. In each case, we focussed on developing computationally efficient implementations, such as adaptively selecting the Markov chain length in ABC-ASMC, and designing efficient proposal distributions in the particle filter. We showed that the ABC approaches gave a poor approximation to the posterior distribution in comparison to PMCMC and SMC$^2$.

In Chapter 4 we extended the inference methods introduced in Chapter 3, in order to perform model comparison. We gave an overview of Bayes factors as a method to perform model comparison, and described the different approaches that can be used for evaluating Bayes factors. We focussed on SMC approaches, which naturally provide estimates of normalising constants, such as the model evidence. In particular, we extended an ABC-SMC approach to target the joint posterior distribution of all models, and model parameters, under consideration, and extended SMC$^2$ to evaluate the model evidence for each model independently. We developed a simulation study to compare each approach, and once again found that ABC performed poorly in comparison to SMC$^2$. Repeating the experiment on real-world data showed that oscillators were more strongly supported by the data than one-dimensional steady-state models.

In Chapter 5 we applied the methodology developed throughout the thesis to two model comparison problems in palaeoclimate science. Specifically, we aimed to select between competing phenomenological models of the glacial-interglacial cycle, and to test whether forced models are more strongly supported by the data than unforced models. We designed a simulation study to assess our ability to solve these problems, and found that we selected the correct model, both when the astronomical forcing had influenced the data, and when it had not. Additionally, we discovered that forced models were preferred when the model used to generate the data was forced, demonstrating that we are able to detect the influence of the astronomical forcing, even when the model is wrong. We gave an overview of different dating methods for sediment cores, and chose to repeat our experiment on a dataset with two sets of age estimates. The first had been obtained with astronomical tuning assumptions, whereas the second was purely depth-derived. We discovered that the results were not robust to the age model used. The astronomically tuned data more strongly supported forced models, and the depth derived data more strongly supported unforced models. In some cases it was difficult to determine if the data more strongly supported particular models due to Monte Carlo variation. However, the partial positive result indicates that palaeoclimate data holds more information than is typically assumed, as a prevailing viewpoint is that palaeoclimate datasets are too sparse and noisy to select between competing phenomenological models.

In Chapter 6 we extended the modelling and inference approaches, in order to jointly estimate the observation times, model parameters, and model evidence. We developed a novel age model based on a linear sediment accumulation model. This model was then adjusted to account for down-core compaction. We extended the particle filter to jointly target the observation times and the state-space of the phenomenological model. We designed a simulation study to assess our ability to recover the observation times, and demonstrated that our algorithm performs well, even when the age-depth relationship is highly nonlinear. We then applied our inference approach to two sediment cores. In each case, our age estimates agreed strongly with the astronomically tuned LR04 age estimates, but had a slight discrepancy with the depth-derived H07 age estimates. Whereas previous approaches gave point estimates for the observation times, and described possible sources for age model error, our approach generates a sample from a posterior distribution, characterising the age model uncertainty. These results also suggest that the amount of information in sediment cores is often underestimated. The previous studies obtained age estimates by averaging over dozens of cores, whereas we obtained strong agreements using only a single sediment core in each case. The posterior variance was also smaller than previously predicted. We repeated our model comparison experiment between a forced and unforced phenomenological model on each sediment core. One of the cores more strongly supported the forced model, providing some evidence in favour of Milankovitch theory.

## 7.2 Directions for Future Research

When using real-world data, we have only employed our inference approaches on individual sediment cores. Ideally, this should be extended to perform inference jointly over multiple sediment cores, so that local variation is averaged out. However, there are a number of challenges in doing so. The first is that the local climate variations need to be modelled in each core, requiring a significant modelling extension. The second is that the order of the observations in time is unknown. In the particle filter, if the next observation to be assimilated occurred previously in time, the transition density of the phenomenological model will be required. That is, if $t_{n-1} < t_{n+1} < t_n$, where observation $n + 1$ is to be assimilated, then the target distribution is proportional to $\pi \left( \boldsymbol{X}_{n+1} \mid \boldsymbol{X}_{n-1} \right) \pi \left( \boldsymbol{X}_n \mid \boldsymbol{X}_{n+1} \right)$, whereas the proposal distribution is proportional to $\pi \left( \boldsymbol{X}_n \mid \boldsymbol{X}_{n-1} \right)$. With a single transition density we can extend the particle filter's proposal distribution to also target the auxiliary variables, but this is not possible with the ratio of two transition densities. ABC methods offer a possible solution, as it will always be possible to simulate a sequence of observation times for each sediment core, and then simulate from the phenomenological model for the sampled times. However, obtaining an accurate approximation will require a large computational expense.

Additionally, when we have used real-world data, we have constrained ourselves to the last 780 kyr, marked by the Brunhes-Matuyama (BM) reversal. Many sediment cores have observations stretching back over millions of years, with indications of previous geomagnetic reversals present. It is trivial to extend any of our methods over this time period. However, the phenomenological models used must then account for the mid-Pleistocene transition at approximately 800 kyr BP. It is plausible that inference methods that assimilate observations sequentially, such as PMCMC and SMC$^2$, will perform poorly in this situation, primarily due to model error, as some of the parameter space might be ruled out during the 40 kyr cycle, which is then necessary in explaining the 100 kyr cycle. The inference methods developed in this thesis also suffer from the curse of dimensionality, in that the computational cost required to obtain accurate results scales nonlinearly with increases in the number of data points.

We have also used observations of only a single proxy variable, $\delta^{18}$O. As demonstrated in Figure 1.2, other proxy variables, such as CO$_2$, show variation over the glacial-interglacial cycle. These variables can be included in our inference approach by extending the observation model. Suitable phenomenological models will need to be used, such as SM90 and SM91 [22, 23], for example, when jointly assimilating observations of both $\delta^{18}$O and CO$_2$. The model comparison methods developed in this thesis present a way to select between hypothesised relationships between climate variables, by formulating those hypotheses into phenomenological models.

Choosing suitable prior distributions is essential if we are to trust in the results of our inference schemes. In our experience, the posterior distributions of the parameters are robust to changes in the prior distributions, so long as there is non-negligible mass in regions of posterior support. The exceptions are those parameters for which we learn little from the data, as the posterior distributions will always strongly resemble the prior distributions. Bayes factors, on the other hand, are often sensitive to the prior distributions, as they require integration over the parameter space. This is a problem for relatively small Bayes factors, where altering the prior distributions can easily alter the conclusions. In this thesis we have elicited prior distributions from an expert in dynamical systems theory. Ideally, prior distributions should be elicited from a number of experts, and sensitivity analyses carried out to assess the robustness of the results to changes in the prior distributions.

Owing to computational limitations we have focussed on using phenomenological models of the glacial-interglacial cycle. There is a wide spectrum of models, ranging from simple linear combinations of the astronomical forcing terms, to GCMs. The inference approaches developed in this thesis are quite general, and can be used with any of these models. Currently, the complexity of the models that can realistically be used is limited by the availability of computational resources. Thus, as time progresses, and more computational power becomes readily available, more complex models can be considered.

A further modelling extension can be made regarding the observation errors. A standard approach, as used in this thesis, is to assume that observation errors are Gaussian. This is often unrealistic. Alternative observation error models can be considered in any of the inference approaches developed here.

There are a number of sources of uncertainty that have not been considered in this thesis. In Chapter 6 we treated the observation depths as known. In reality, there is a small amount of error when measuring the depth of an observation. Since these measurement errors do not alter the order of the observations, they can be included in any of the inference schemes described in the thesis. On a related note, the depths of geomagnetic reversals are not precisely known, and so quantifying the depth uncertainty will likely lead to time estimates that are less constrained than in Chapter 6.

The variation of eccentricity and precession through time are also treated as known, whereas small errors are present. This can be observed in the discrepancies between different orbital solutions [32, 33, 95–97]. There is very strong agreement between the different orbital solutions over the past 1 Myr, and so the assumption that the variation of the orbital parameters are known is unlikely to invalidate the conclusions of this thesis. However, the discrepancy grows over time, and so care should be taken when performing inference over longer timescales. A possible solution is to explicitly model the orbital parameters, and include small stochastic perturbations to account for model discrepancy. Alternatively, a model comparison experiment can be carried out on competing orbital solutions to test whether palaeoclimate data more strongly support one solution over others. However, given the minor variation between the different orbital solutions, obtaining a positive result seems unlikely.

## 7.3 Concluding Remarks

Ultimately, it will be straight forward for palaeoclimate experts to criticise our results on the basis of our choices of models, data, and prior distributions. However, that would miss the point of this thesis. We have shown that statistical methodology can be developed to allow careful statistical inference in these problems, and demonstrated what we believe is the appropriate way to perform analysis, hopefully adding a valuable tool to the scientist's armoury. The work can be extended and improved in many ways, and we hope that there is sufficient interest from researchers in the field that this will be done.

# References

[1] L. E. Lisiecki and M. E. Raymo. A Pliocene-Pleistocene stack of 57 globally distributed benthic $\delta^{18}$O records. *Paleoceanography*, 20:PA1003, 2005.

[2] A. Berger and M. F. Loutre. Astronomical theory of climate change. *Journal de Physique IV*, 121:1–35, 2004.

[3] S. R. Meyers and L. A. Hinnov. Northern hemisphere glaciation and the evolution of Plio-Pleistocene climate noise. *Paleoceanography*, 25:PA3207, 2010.

[4] P. Huybers. Glacial variability over the last two million years: an extended depth-derived age model, continuous obliquity pacing, and the Pleistocene progression. *Quaternary Science Reviews*, 26:37–55, 2007.

[5] North Greenland Ice Core Project members. High-resolution record of northern hemisphere climate extending into the last interglacial period. *Nature*, 431:147–151, 2004.

[6] W. Dansgaard, S. J. Johnsen, H. B. Clausen, D. Dahl-Jensen, N. S. Gundestrup, C. U. Hammer, C. S. Hvidborg, J. P. Steffensen, A. E. Sveinbjörnsdottir, J. Jouzel, and G. Bond. Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature*, 364:218–220, 1993.

[7] H. Heinrich. Origin and consequences of cyclic ice-rafting in the northeast atlantic ocean during the past 130,000 years. *Quaternary Research*, 29:142–152, 1988.

[8] A. L. Berger. Long term variations of caloric insolation resulting from the Earth's orbital elements. *Journal of Atmospheric Sciences*, 35:2362–2367, 1978.

[9] J. D. Hays, J. Ibrie, and N. J. Shackleton. Variations in the Earth's orbit: pacemaker of the ice ages. *Science*, 194:1121–1132, 1976.

[10] M. A. Cane, P. Braconnot, A. Clement, H. Gildor, S. Joussaume, Kageyama M., M. Khodri, D. Paillard, S. Tett, and E. Zorita. Progress in paleoclimate modelling. *Journal of Climate*, 19:5031–5057, 2006.

[11] M. Crucifix and J. Rougier. On the use of simple dynamical systems for climate predictions. *European Physics Journal - Special Topics*, 174:11–31, 2009.

[12] M. Crucifix. How can a glacial inception be predicted? *The Holocene*, 21:831–842, 2011.

[13] T. Mauritsen, B. Stevens, E. Roeckner, T. Crueger, M. Esch, M. Giorgetta, H. Haak, J. Jungclaus, D. Klocke, D. Matei, U. Mikolajewicz, D. Notz, R. Pincus, H. Schmidt, and L. Tomassini. Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems*, 4:1942–2466, 2012.

[14] J. Laskar. The limits of earth orbital calculations for geological time-scale use. *Philosophical Transactions of the Royal Society of London A*, 357:1735–1759, 1999.

[15] W. Berggren, F. Hilgen, C. Langereis, D. Kent, J. Obradovich, I. Raffi, M. Raymo, and N. Schackleton. Late Neogene chronology: new perspectives in high resolution stratigraphy. *Geological Society of America Bulletin*, 107:1272–1287, 1995.

[16] S. Cande and D. Kent. Revised calibration of the geomagnetic polarity timescale for the late Cretaceous and Cenozoic. *Journal of Geophysical Research*, 100:6093–6096, 1995.

[17] B. Singer and M. Pringle. Age and duration of the Matuyama-Brunhes geomagnetic polarity reversal from 40Ar/39Ar incremental heating analyses of lavas. *Earth and Planetary Science Letters*, 139:47–61, 1996.

[18] P. Huybers and C. Wunsch. Obliquity pacing of late Pleistocene terminations. *Nature*, 434:491–494, 2005.

[19] B. E. Bemis, H. J. Spero, J. Bijma, and D. W. Lea. Reevealuation of the oxygen isotopic composition of planktonic foraminifera: experimental results and revised paleotemperature equations. *Paleoceanography*, 13:150–160, 1998.

[20] D. Lüthi, M. Le Floch, B. Bereiter, T. Blunier, J-M. Barnola, U. Siegenthaler, D. Raynaud, J. Jouzel, H. Fischer, K. Kawamura, and T. F. Stocker. High-resolution carbon dioxide concentration record 650,000-800,000 years before present. *Nature*, 453:379–382, 2008.

[21] M. Crucifix. Oscillators and relaxation phenomena in Pleistocene climate theory. *Transactions of the Philosophical Transactions of the Royal Society A*, 370:1140–1165, 2012.

[22] B. Saltzman and K. A. Maasch. A first-order global model of late Cenozoic climate. *Transactions of the Royal Society of Edinburgh: Earth Sciences*, 81:315–325, 1990.

[23] B. Saltzman and K. A. Maasch. A first-order global model of late Cenozoic climate. II further analysis based on simplification of the $CO_2$ dynamics. *Climate Dynamics*, 5:201–210, 1991.

[24] G. H. Roe and M. R. Allen. A comparison of competing explanations for the 100,000-yr ice age cycle. *Geophysical Research Letters*, 26:2259–2262, 1999.

[25] M. Crucifix. Why could ice ages be unpredictable. *Climate of the Past*, 9:2253–2267, 2013.

[26] D. Paillard and F. Parrenin. The Antarctic ice sheet and the triggering of deglaciations. *Earth and Planetary Science Letters*, 227:263–271, 2004.

[27] E. Tziperman, M. E. Raymo, P. Huybers, and C. Wunsch. Consequences of pacing the Pleistocene 100 kyr ice ages by nonlinear phase locking to Milankovitch forcing. *Paleoceanography*, 21:PA4206, 2006.

[28] F. Parrenin and D. Paillard. Terminations VI and VIII ($\sim$530 and $\sim$720 kyr BP) tell us the importance of obliquity and precession in the triggering of deglaciations. *Climate of the Past*, 8:2031–2037, 2012.

[29] S. H. Strogatz. *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*. Westview Press, 2000.

[30] J. Guckenheimer and P. Holmes. *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, volume 42 of *Applied Mathematical Sciences*. Springer-New York, 1983.

[31] B. Van der Pol. On “relaxation-oscillators”. *The London, Edeinburgh and Dublin philosophical magazine and journal of science*, 2:978–992, 1926.

[32] A. L. Berger. Long term variations of daily insolation and Quaternary climate changes. *Journal of Atmospheric Sciences*, 35:2362–2367, 1978.

[33] A. Berger and M. F. Loutre. Insolation values for the climate of the last 10 million years. *Quaternary Science Reviews*, 10:297–317, 1991.

[34] P. Huybers and E. Tziperman. Integrated summer insolation forcing and 40,000-year glacial cycles: The perspective from an ice-sheet/energy-balance model. *Paleoceanography*, 23:PA1208, 2008.

[35] K. Hasselmann. Stochastic climate models. *Tellus*, 28:473–485, 1976.

[36] B. Oksendal. *Stochastic differential equations: An introduction with applications*. Springer-Verlag, 1985.

[37] A. Beskos, O. Papaspiliopoulos, and G. Roberts. Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli*, 12:1077–1098, 2006.

[38] A. Beskos, O. Papaspiliopoulos, and G. Roberts. A factorisation of diffusion measure and finite sample path constructions. *Methodology and Computing in Applied Probability*, 10: 85–104, 2008.

[39] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*. Springer-Berlin, 1992.

[40] M. Crucifix and C. Almeida. Report on videoconference of 21 Oct 2011 : Nottingham and LLN. Technical report, Univerisite Catholique de Louvain, 2011.

[41] M. Crucifix. Model selection. Technical report, Univerisite Catholique de Louvain, 2013.

[42] A. Golightly and D. J. Wilkinson. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statististics & Data Analysis*, 52:1674–1693, 2008.

[43] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B*, 72:269–342, 2010.

[44] J-M. Marin, P. Pudlo, C. P. Robert, and R. Ryder. Approximate Bayesian computation methods. *Statistics and Computing*, 22:1167–1180, 2012.

[45] J. von Neumann. Various techniques used in connection with random digits. *National Bureau of Standards Applied Mathematics Series*, 12:36–38, 1951.

[46] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC Press, 2013.

[47] D. Gamerman and H. F. Lopes. *Markov Chain Monte Carlo: stochastic Simulation for Bayesian Inference*. CRC Press, 2nd edition, 2006.

[48] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7:110–120, 1997.

[49] J. S. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93:1032–1044, 1998.

[50] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society B*, 68:411–436, 2006.

[51] M. A. Beaumont, J. Cornuet, J. Marin, and C. P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96:983–990, 2009.

[52] P. H. Garthwaite, J. B. Kadane, and A. O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100:680–701, 2005.

[53] A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, Garthwaite P. H., D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain judgements: eliciting experts' probabilities*. Wiley-Blackwell, 2006.

[54] S. Tavaré, D. J. Balding, R. C. Griffith, and P. Donnelley. Inferring coalescence times from dna sequence data. *Genetics*, 145:505–518, 1997.

[55] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. T. Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16:1791–1798, 1999.

[56] T. McKinley, A. R. Cook, and R. Deardon. Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5:24, 2009.

[57] G. Biau, F. Cérou, and A. Guyader. New insights into approximate Bayesian computation. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 2014. In press.

[58] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 100:15324–15328, 2003.

[59] S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 104: 1760–1765, 2007.

[60] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6:187–202, 2009.

[61] S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential Monte Carlo without likelihoods: errata. *Proceedings of the National Academy of Sciences of the United States of America*, 106:16889–16890, 2009. Correction to original paper.

[62] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer-New York, 2001.

[63] P. Del Moral, A. Doucet, and A. Jasra. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22:1009–1020, 2012.

[64] C. C. Drovandi and A. N. Pettitt. Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, 67:224–233, 2011.

[65] P. C. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2:49–55, 1936.

[66] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50:1–18, 2000.

[67] P. Joyce and P. Marjoram. Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7:Article 26, 2008.

[68] M. G. B. Blum, M. A. Nunes, D. Prangle, and S. A. Sisson. A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28:189–208, 2013.

[69] R. D. Wilkinson. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, 12:129–141, 2013.

[70] R. M. Neal. Annealed importance sampling. Technical report, University of Toronto, 1998.

[71] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings F*, 140:107–113, 1993.

[72] P. Del Moral. *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Springer-New York, 2004.

[73] C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37:697–725, 2009.

[74] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos. SMC$^2$: an efficient algorithm for sequential analysis of state-space models. *Journal of the Royal Statistical Society B*, 75:397–426, 2013.

[75] M. L. Eaton. *Multivariate Statistics: a vector space approach*. John Wiley and Sons, 1983.

[76] M. G. B. Blum and O. François. Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, 20:63–73, 2010.

[77] P. Fernhead and D. Prangle. Constructing summary statistics for approximate Bayesian compuation: semi-automatic ABC. *Journal of the Royal Statistical Society B*, 74:419–474, 2012.

[78] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: fifteen years later. In D. Crisan and B. Rozavsky, editors, *Oxford handbook of nonlinear filtering*. Oxford University Press, 2011.

[79] Y. Zhou, A. M. Johansen, and J. A. D. Aston. Towards automatic model comparison: An adaptive sequential Monte Carlo approach. 2013. arXiv.org:1303.3123.

[80] R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[81] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[82] S. P. Brooks, P. Giudici, and G. O. Roberts. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society B*, 65:3–39, 2003.

[83] T. Toni and M. Stumpf. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26:104–110, 2010.

[84] N. J. Shackleton, A. Berger, and W. R. Peltier. An alternative astronomical calibration of the lower Pleistocene timescale based on ODP site 677. *Transactions of the Royal Society of Edinburgh: Earth Sciences*, 81:251–261, 1990.

[85] L. E. Lisiecki. Links between eccentricity forcing and the 100,000-year glacial cycle. *Nature Geoscience*, 3:349–352, 2010.

[86] P. Huybers. Combined obliquity and precession pacing of late Pleistocene deglaciations. *Nature*, 480:229–232, 2011.

[87] J. Haslett and A. Parnell. A simple monotone process with application to radiocarbon-dated depth chronologies. *Journal of the Royal Statistical Society C*, 57:399–418, 2008.

[88] C. Bronk Ramsey. Deposition models for chronological records. *Quaternary Science Reviews*, 27:42–60, 2008.

[89] C. Bronk Ramsey. Bayesian analysis of radiocarbon dates. *Radiocarbon*, 51:337–360, 2009.

[90] M. Blaauw and A. Christen. Flexible paleoclimate age-depth models using an autoregressive gamma process. *Bayesian Analysis*, 6:457–474, 2011.

[91] A. Molini, P. Talkner, G. G. Katul, and A. Porporato. First passage time statistics of Brownian motion with purely time dependent drift and diffusion. *Physica A*, 390:1841–1852, 2011.

[92] D. B. Bahr, E. W. H. Hutton, J. P. M. Syvitski, and L. F. Pratson. Exponential approximations to compacted sediment porosity profiles. *Computers and Geosciences*, 27:691–700, 2001.

[93] P. Huybers and C. Wunsch. A depth-derived Pleistocene age model: uncertainty estimates, sedimentation variability, and nonlinear climate change. *Paleoceanography*, 19:PA1028, 2004.

[94] A. Mix, J. Le, and N. Shackleton. Benthic foraminifer stable isotope stratigraphy of site 846: 0-1.8 ma. *Proceedings of the Ocean Drilling Program, Scientific Results*, 138:839–854, 1995.

[95] J. Laskar, F. Joutel, and F. Boudin. Orbital, precessional and insolation quantities for the Earth from -20 Myr to +10 Myr. *Astronomy and Astrophysics*, 270:522–533, 1993.

[96] J. Laskar, P. Robutel, F. Joutel, M. Gastineau, A. C. M. Correia, and B. Levrard. A long term numerical solution for the insolation quantities of the Earth. *Astronomy and Astrophysics*, 428:261–285, 2004.

[97] J. Laskar, A. Fienga, M. Gastineau, and H. Manche. La2010: A new orbital solution for the long-term motion of the Earth. *Astronomy and Astrophysics*, 532:A89, 2011.