Model selection is possible with palaeoclimate data and models of the glacial-interglacial cycle What drives the glacial-interglacial cycle? A Bayesian approach to a long-standing problem

Carson, Crucifix, Preston, Wilkinson
Order can/may change

November 27, 2014

Comment by MC: We'll probably have to think about the title but that's not urgent

Abstract

VERY ROUGH - IGNORE FOR NOW.

The prevailing viewpoint in palaeoclimate science, is that a single palaeoclimate record contains insufficient information to discriminate between competing explanatory models. Our aim here is to show that this is not the case. Recent developments in Monte Carlo methodology, and advances in computer power, mean that for a wide class of phenomenological models, we are now able to perform filtering, calibration, and Bayesian model selection. Using the SMC^2 algorithm of Chopin and ? 2014, combined with Brownian bridge proposals for the state trajectories, we show that even with relatively short time series, it is possible to estimate Bayes factors to sufficient accuracy to be able to discriminate between competing models. This suggests a new possibility for those working in climate science - careful empirical comparison of competing hyptotheses is now possible.... Our results highlight the problem of using pre-dated palaeoclimate records. By analysing a dataset that has been dated in two different ways, we are led to completely contradictory conclusions. Thus indicating the current practice of first dating a core, and then analysing it to draw further scientific conclusions in an independent analysis is,.... and these tasks...

1 Introduction

Comment by Rich: This is not a modelling or a science paper. It is a proof of concept: we can do model selection. Give us your models and your data, and we can say which is best supported.

Comment by Rich: Michel - the intro is now ready for you to look at - no need to edit the language at this stage, just the message.

The history of the Earth's climate can be inferred by studying the oxygen isotope composition of microbes and water from sediment and ice core samples [1, 2]. A quantity commonly used as a "proxy measurement" of the Earth's climate is $\delta^{18}O$, a function of the ratio between oxygen isotopes ^{18}O and ^{16}O. The $\delta^{18}O$ level in microbes depends on the temperature and level of $\delta^{18}O$ in the water at the time the microbes formed, and the level of $\delta^{18}O$ in seawater depends on salinity and global ice volume [1]. Larger values of δ^{18}O in microbes from core samples broadly indicate a colder climate with greater ice volume.

Such data suggest that the Earth entered into its current ice-age (characterised by persistent ice caps at the poles) approximately 3 million years (Myr) ago [ref?]. Since then the climate has fluctuated between cold periods, in which glaciers expand, and warm periods in which the glaciers retreat. These are known as glacial–interglacial cycles. In the early Pleistocene—a period from around 2.5 Mya ago until 11 thousand years (kya) ago—the average period of the glacial–interglacial cycle was around 40 kyr. However, around 1 Mya ago the average period became around 100 kyr, a change in behaviour known as the mid-Pleistocene transition. There is much interest in the mechanisms underlying the 100 kyr glacial–interglacial cycles.

NEW The aim of this paper is to demonstrate a Bayesian approach that can be used over a broad range of model selection problems in palaeoclimatology. We focus on determining the influence of the astronomical forcing, and distinguishing between competing phenomenological (consistent with, but not derived from the physical theory of a system) models of the glacial-interglacial cycle.

An important driver of the cycles is the variability in incoming solar radiation, termed "insolation". Insolation is variable on account of the geometry of the Earth's orbit; for a historical discussion see [3]. The effect of the orbit on insolation, known as "astronomical forcing", has been studied extensively, following from work of Milankovitch [ref] who decomposed the orbit into three orbital characteristics, namely eccentricity, obliquity, and precession [4]. These characteristics, which we explain in more detail in §, are each cyclical and have different periods and amplitudes in their contribution to insolation. The relative importance of the characteristics to the glacial–interglacial cycles is a topic of ongoing interest [5, 6, 7] [any other refs? Papers that use Rayleigh's R stat? see line -12 on p229 of Huybers 2011]. Notably, the 100kyr period of glacial–interglacial cycles corresponds closely with the period of the eccentricity cycle, even though of the three orbital characteristics eccentricity has the smallest effect on the amplitude of insolation variability [adapted from Wiki article on M cylces: it this worded appropriately?]. At first this seems counterintuitive, but it is not necessarily surprising: the Earth's climate is a complex dynamical system to which astronomical forcing is just an input. The emergent periodicity is a result of interactions between the forcing and climate processes, which may themselves be oscillatory.

Hence in this paper we consider the Earth's climate as a forced dynamical system. Our goal is to use a Bayesian approach to select amongst candidate dynamical models and to assess the importance of the various orbital characteristics to the glacial–interglacial cycle.

The literature contains a vast array of models of the Earth's climate, ranging from modern and very complicated models that aim to include as many physical processes as possible, to simple dynamical models that involve few variables and aim only to describe the main modes of the dynamics; see for example [Crucifix?]. The limited nature of the data at our disposal (described later in §) in addressing the above-mentioned goal leads us to favouring simpler models. The three models we consider (detailed in §), each involves a variable representing ice volume, and either one or two other variables representing other aspects of the climate. In each model the ice volume is assumed to be forced directly by insolation, plus a Brownian motion to describe other extraneous forcing (and broadly account for error in the model specification).

Consequently, the statistical challenge that results from these modelling choices is to make inference for partially observed, forced non-linear stochastic differential equations (SDEs). A major challenge in SDE inference is that the transition density, and therefore the likelihood function, is not available is closed form, termed "intractable". This is problematic for classical Bayesian inference techniques such Markov Chain Monte Carlo (MCMC) that require evaluations of the likelihood. A powerfool tool for problems with intractable likelihoods is the particle filter, and in this paper we employ the state-of-the-art SMC^2 approach recently introduced by [ref] [perhaps need to say more here].

Before formulating further the statistical problem, we briefly review earlier research that investigated the role of the orbital characteristics on the glacial–interglacial cycle. Several papers develop frequentist hypothesis tests [5, 6, 7], based on comparing "termination times", which mark where individual glacial cycles finish, with the times of the maxima of the insolation function. In [7], for instance, the null hypothesis, H_0, is that the termination times are independent of the timings of the maxima for a given insolation function, and the alternative hypothesis, H_1, that the terminations tend to occur when the maxima are anomalously large. The test statistic used is the difference between the medians of the forcing maxima associated with terminations and with those not associated with maxima, with the null distribution generated from random simulations of termination times under a model consistent with H_0. Such an approach leads to a p-value that characterises the strength of evidence against H_0 for a given insolation function. As always with such frequentist approaches the interpretation of p-values requires care: a "non-significant" p-value reflects that the data provide insufficient evidence to reject H_0 in favour of H_1, not that H_0 should be favoured over H_1. (This is in contrast to the Bayesian approach which we adopt in this paper, in which evidence for competing hypotheses can be directly compared.) Moreover, differences in the details about how the foregoing frequentist tests are constructed substantially effect the conclusions, with different studies finding different orbital characteristics being significant (obliquity in [5], eccentricity in [6], and a combination of precession and obliquity in [7]).

Comment by SPP: JAKE: We need a para here about the Phil Trans paper comparing models using information criteria

This paper is structures as follows. In Section 2 we describe in more detail the $\delta^{18}O$ dataset on which we will base inference, and we detail the models we consider for astronomical forcing, for the Earth's climate dynamics, and for the observations. Section 4 includes a formulation of the Bayesian approach and brief review of the particle-filter methods that we extensively use. In Section 5 we present a simulation study to assess performance of the algorithms on synthetic data, and an analysis of the real $\delta^{18}O$ dataset. In Section 7 we offer some thoughts on the practical implementation of the particle-filter methods for such problems, discuss the scientific conclusions, and suggest some future directions for research.

Comment by SPP: I agree with MC's comment that such discussion would be helpful there and should fit well in RSSC - certainly more so than attempts to make scientific conclusions which might sound overstretched

---

Many models follow the viewpoint that the astronomical forcing acts as a pacemaker to the climate by synchronising a self-oscillating or excitable system with the astronomical forcing [8, 9].

Comment by MC: I wrote this, but I would soften the statement: see for example the interesting, unfortunately rejected, paper by Daruka and Ditlevsen (climate of the past discussion) were they use a Duffing oscillator. The Duffing oscillation does not have a limit cycle, even not an internal period, and yet does the job

A common viewpoint is that the information contained in a single record do not contain enough information to distinguish between the numerous proposed models [10, 11].

Comment by MC: This is a tricky point: there is a distinction between identifying the dynamics (is there a limit cycle; is the system unstable etc. and identifying physical principle, e.g.: does the formation of antarctic bottom water off the Antarctic shelve play a critical role in the dynamics? In the latter, complementary discriminating evidence is provided by the use of simulation (e.g.: global climate models) and the joint use of several climate datasets. We are here in the framework of a proof of concept and be happy to concentrate on, after all, a simple problem: consider the physcial information as fairly vague (vague priors on parameters, and a single parameter).

The phenomenological modelling approach enables us to consider the importance of the astronomical forcing terms as a model selection problem with unknown parameters. In other words a number of competing models are proposed and we are interested in determining which model best explains the available data, as well as estimating the values of parameters in each model. A Bayesian inference process provides a principled approach to such problems, and has a number of advantages over frequentist hypothesis testing. In frequentist hypothesis testing one model is considered as the null hypothesis, and evidence is only ever weighed against it. A large p value does not indicate that the null model is more strongly supported, or that two models are equally supported, but only that there is insufficient evidence to chose between them. In other words the null hypothesis can be rejected, but never accepted. Likewise smaller p values do not indicate better models, only that the null model lacks explanatory power. In contrast, A Bayesian approach indicates which models are more strongly supported by the data. In particular, it may be used in cases where the models are not nested, allowing us to select between the numerous proposed phenomenological models. Finally, the use of SSMs allows us to assimilate paleaoclimate data in a principled framework, without making it necessary to use summary statistics such as termination times.

Comment by MC: though we may agree that using summary statistics is a convenient approach for implicitly accounting for model discrepancy

2 Data and models

Attempts to understand the dynamical behaviour of the palaeoclimate typically involve four components. These are i) data consisting of palaeoclimate records; ii) models of the climate; iii) drivers/forcings of the climate (such as CO_2 emissions, or more pertinently for palaeoclimate, the solar forcing), and iv) a statistical model relating the three previous components. The final component necessary is then the statistical methodology necessary to combine these components to find answers that combine them all in an appropriate way.

The aim of this paper is to demonstrate that the kind of questions we imagine palaeoclimate scientists may wish to ask, can now be answered in a way statisticians may wish to answer them. That is to say, given some data and a selection of models, we are able to fit these models, and assess which model is best supported by the data. Scientific aspects of the approach can (and we hope will) be improved upon: different datasets, richer models, and more accurate statistical modelling may all follow. Here, we show that the methodology and computer power is now of sufficiently advanced state that the Bayesian inference problems can be answered.

Comment by Rich: Michel - I'm trying (and failing) to get towards the holy trinity diagram you have been using, with the triangle of approaches, with the aim to unify some of them using stats. I'm sure you can phrase this better.

In the next section we present methodology for

2.1 Data

The data we will use are based on measurements of $\delta^{18}O$ at different depths in sediment and ice cores sampled from various geographical locations [more details?]. In climatology, a set of such measurements is known as a "record" [?], and an average over multiple records is known as a "stack". The $\delta^{18}O$ in deeper parts of a core corresponds to climatic conditions further back in time. However, beyond monotonicity there is no simple relationship between core depth and age. This is because the accumulation of sediment [what to say about ice?] results from a combination of complicated physical processes including sedimentation (which occurs at at variable rates), erosion, and core compaction. A model for the relationship between depth and age is known as an "age model", and many such models have been proposed in the literature [refs]. A common stategy in developing an age model is to align features of records to important events, such as magnetic reversals, whose dates are accurately known. Also common is to align features of records to aspects of the astronomical forcing [refs], a process known as astronomical tuning. Investigating age models is beyond the scope of this paper, so we take as a starting point a stack for which an age model has already been applied; this gives a data set which is a time series $\{t, Y_t\}$ in which Y_t denotes the level of $\delta^{18}O$ at time t.

Figure 1 about here.

In this article we consider a stack (or is it a core) known as ODP677 which has not been astronomically tuned [2] [comment on the tuning used?]. This stack is shown in Figure 1 [comments?]. [I'm stuggling a bit with terminology here: is it okay to use "stack" to mean something that has already been dated?] We also consider a variant of this ODP677 stack, based on the same underlying records, which has been astronomically tuned [1], to highlight how conclusions of inference can be affected by incorrectly "double counting" the influence of astonomically forcing.

Comment by Rich: I'm not sure this is correct

Comment by SPP: Anything alse needed from below?

Paleoclimates are among other reconstructed usingg records of δ^{18}O, which is a measure of the ratio ^{18}O :^{16} O. This information can be extracted from formainifera and ice-cores, where the variability of δ^{18}O in response to depth indicates the evolution of ice volume and temperature over time. In particular, larger $\delta^{18}O$ in benthic foraminifera shells may result either from lower deep ocean temperature, higher continental ice volume or, more likely, a combination of both. Larger values of δ^{18}O indicate a colder climate with greater ice build-up.

Datasets from individual drill sites typically contain observations that are noisy, and sparsely distributed in time. Geological conditions can also lead to incomplete records, where there are large lengths of time with no observations, and repeated data, where the same events can be identified numerous times in a single record. These problems have motivated the development of stacks [1, 2]. A stack is a composite of multiple records, which are correlated and averaged over to give a single complete dataset with an improved signal-to-noise ratio. By using records from numerous sites, stacks are also more likely to represent global conditions, as local variation is averaged out. However, the averaging methods typically employed can lead unprincipled inference as it is not clear how the particular averaging scheme used has altered the information for each record.

On the other hand, converting the depth scale of sediment and ice-cores into a time scale is non-trivial. There are very few points in the sediment cores that can be dated confidently, using either biostratigraphic markers or identified magnetic reversals. Namely, the most recent magnetic reversal is the Bruhnes-Matuyama event 780 kyr BP. However, variations in sedimentation rates, down-core compaction, and numerous other physical processes imply that observation times do not change linearly through the core. The study of the age-depth relationship and age-depth uncertainty has led to numerous proposed age models. By far the most common approach is to align features of the astronomical forcing to the record of interest; a process known as orbital tuning. Orbitally tuned data are undesirable when studying the influence of the astronomical forcing, as the forcing itself would be double counted. Any results suggesting a strong influence from the astronomical forcing may in fact be highlighting the dating assumptions. It is therefore naturally better to use data which has not been orbitally tuned to operate model selection and, in particular, estimate the influence of astronomical factors on the dynamics of glacial interglacial cycles.

For these reasons, we use in this article the the ODP677 record [12], shown in Figure 1. ODP677 has been dated both as part of an orbitally tuned scheme [1], and a non-orbitally tuned scheme [2]. We use age-models based on stacks as correlation between different datasets is expected to highlight areas of missing data, which may not be identified if dating a single core. We focus on the last 780 kyr of this record (since the B-M), which contains 363 observations.

2.2 A model for astronomical forcing

Comment by SPP: This needs some work. Is it all standard Milankovitch Theory? Is there a standard ref? The important thing from the reader's point of view is that F is a function of t (which depends on some other parameters): this needs to be made clear notationally. It also needs: - precise defs (with intuition) of precession, coprrecession, obliquity, - definitions of all quantities/params, including defns of how ϖ, e, etc depend on t, - units of quantities, - interpretation about what parameters mean, e.g the γs, - a mathematical definition of the unit-variance scaling and explanation for why we are doing this.

Different measures of insolation can be well approximated by a linear combination of precession ($\Pi := e \sin \varpi / a_1$), coprecession ($\tilde{\Pi} := e \cos \varpi / a_2$) and obliquity ($O := (\varepsilon - \varepsilon_0)/b_1$) terms. Precession, coprecession and obliquity themselves are well approximated by a sum of sines and cosines, the values of which are provided in [13]. The astronomical forcing can therefore be represented as:

$$F(\gamma_P, \gamma_C, \gamma_E) = \gamma_P \bar{\Pi} + \gamma_C \bar{\tilde{\Pi}} + \gamma_E \bar{O} \tag{1}$$

where $\bar{\Pi}$, $\bar{\tilde{\Pi}}$ and $\bar{O}$ represent the precession, coprecession and obliquity signals scaled to have unit variance. 65°N insolation at the summer solstice is recovered by setting $\gamma_P = 0.8949$, $\gamma_C = 0$ and $\gamma_E = 0.4346$. Pure precession, coprecession or obliquity signals can be recovered by setting the other scaling parameters to 0.

2.3 Models for climate dynamics

The three models that we consider for underlying climate dynamics are as follows. Each model has a component, X_1, representing global ice volume, and one or two other components representing glaciation state. Each models the glacial–interglacial cycle using a qualitatively different dynamical mechanism [right?], as explained further below [need to add more explanation below]. For an overview of oscillators in palaeoclimate modelling see [8]).

Model SM91: Saltzmann and Maasch (1991)

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

This models glacial–interglacial cycles as a forced van der Pol oscillator [ref?] with the variables subjected to Brownian motion. Variables X_2 and X_3 respectively represent CO2 concentration [in what?] and deep-sea ocean temperature. [Need to define parameters and give interpretation of them] Example trajectories are shown in Figure [? - comment].

Model T06: [authors, year]

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

This is an example of a "hybrid" model coupling X_1, which is governed by a stochastic differential equation, to a binary indicator variable X_2 representing absence (0) or presence (1) of Arctic sea ice. Variable X_2 switches values when X_1 passes through threshold values that that are different for the $0 \to 1$ and $1 \to 0$ switches, introducing "hysteresis" which causes strong phase locking to insolation. [Description of params]. Figure [] shows examples trajectories [comment].

Model PP12: [authors, year]

Comment by SPP: We need the notation here to make clear what this has in common with T06. I've tried this below. But I could only go so far as the defns of the various quantities seem quite arbitrary. What is the physical interpretation of the scaling and truncation of the forcing? *The defn needs checking*

$$\begin{aligned}
dX_1 &= -(\gamma_P \Pi^\dagger + \gamma_C \tilde{\Pi}^\dagger + \gamma_E \bar{O} - a_g + (a_g + a_d + X_1/\tau) X_2) dt + \sigma_1 dW_1, \\
X_2 &: \quad \text{switches from 0 to 1 when } F(\kappa_P, \kappa_C, \kappa_E) \text{ is less than some threshold } v_{\mathrm{l}} \\
X_2 &: \quad \text{switches from 1 to 0 when } F(\kappa_P, \kappa_C, \kappa_E) + X_1 \text{ is greater than some } v_{\mathrm{u}}
\end{aligned}$$

where $\Pi^\dagger$ and $\tilde{\Pi}^\dagger$ are transformed precession and coprecession components defined as [needs explanation as this seems totally arbitrary!]

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

with

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

having the effect of truncation of its argument [expand/explain why?]

As with T06, this is a hybrid model with X_1 governed by an SDE and with X_2 a binary variable, in this model representing whether the climate is in a period of glaciation (0) or a period of deglaciation (1). During the glaciation phase ice volume increases according to variation in insolation [really?] Due to the truncation of the forcing in the ice volume equation this model responds nonlinearly to variation in insolation. During the deglatiation phase the system relaxes towards a deglatiated state. The phase changes occur mainly due to the astronomical forcing [meaning?], with ice volume only appearing in the glaciation-deglaciation switch [also in the drift for X_1, no?]1. Example trajectories are shown in Figure [] [comments]

2.4 An observation model

The final modelling ingredient is a model relating the unobserved variables, X_t, in the dynamical models (...) to the observed dataset $\{t, Y_t\}$. We will use the model

$$Y_t \sim \mathcal{N}(d + hX_{1t}, \sigma_y^2),$$

where d, h, σ_y^2 are scalar parameters, i.e., each Y_t is a scaled and shifted version of the value X_1t of the ice volume variable in the underlying dynamical model, observed with noise.

Comment by SPP: Is this "scalar" version general enough for the paper? We need to sort out notation (currently X_1 variously denotes first element of vector X, scalar X evaluated at t=1, and scalar X evaluated at first time index) - need to talk about this before making a decision.

3 State Space Models

Describe different dynamics: T06 (Van der Pol), PP12 (Flushing), SM91 (Hysterisis).

The low-order dynamical models that have been proposed can be represented as SSMs, which describe the evolution of the system and how to relate observations to the system. The evolution of the system is represented by a hidden (not directly observed) Markov process $X_{1:T}$, where $t = 1, ..., T$ are observation times. This Markov process can be fully described by initial density $\pi(X_1|\theta)$, and transition density

$$\pi(X_{t+1}|X_{1:t}, \theta) = \pi(X_{t+1}|X_t, \theta), \qquad t \geq 1, \tag{2}$$

for some parameter θ. For phenomenological models of the glacial-interglacial cycle the system dynamics are commonly represented by oscillators, and in particular relaxation oscillators (for an overview of oscillators in palaeocliamte modelling see [8]). Oscillators are systems that undergo self sustaining oscillations in absence of any external forcing. Relaxation oscillators are a particular kind of oscillator characterised by a relaxation process, in which the system is attracted to some region of phase-space, followed by a switch to a destabilisation process that ejects the system from its current relaxation state. Following the destabilisation the system enters another (possibly the same) relaxation state, continuing the cycle.

Saltzmann proposed that the glacial-interglacial cycle could be modelled as an oscillating system synchronised on the astronomical forcing [14, 15], and many models have since followed this approach [8, 9].

The climate is an incredibly complex system, with physical processes occuring over a wide range of timescales. One can not hope to capture the complexity of the climate using phenomenological models. It is important to account for the discrepancy between the relatively simple models and the complex system under study. This is often achieved by including stochastic contributions in the model, which approximate processes on smaller timescales than are modelled explicitly [16]. Such models can then be represented as stochastic differential equations of the form

$$dX = \mu(X, F(\gamma_P, \gamma_C, \gamma_E), \theta)dt + \Sigma_x^{\frac{1}{2}}(X, \theta)dW,$$

where μ is called the drift and $\Sigma_x^{\frac{1}{2}} dW$ is called the volatility. In the context of this article μ is a set of equations representing a relaxation oscillator forced by F, which is some representation of the astronomical forcing. W is a standard multivariate Brownian motion, approximating physical processes that occur over small timescales. Introducing stochastic components in to relaxation oscillators has a number of effects [17]. The random fluctuations mean that two runs of the model in general lead to different trajectories. This has obvious implications regarding the predictability of glacial cycles [18, 19]. Some predictability is maintained due to the system synchronising on the astronomical forcing, but experiments show that additive noise can notably influence the timing of glacial inceptions and terminations in these models [8].

Here we consider 3 phenomenological models from the literature. The reader is referred to the original papers for interpretations of the parameters.

SM91

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

SM91 [15] is one of the first models to model glacial cycles as an oscillating system synchronised on the astronomical forcing. It describes the evolution of ice volume (X_1), carbon dioxide concentration (X_2), and deep-ocean temperature (X_3). SM91 was proposed by Saltzmann and Maasch as a modification to the earlier SM90 [14], which included alternate carbon-dioxide cycle dynamics. Non-linearity is introduced through the carbon dioxide equation, which is the cause of the oscillations.

T06

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

T06 [20]. is an example of a hybrid model, as it combines a differential equation and a discrete variable. X_1 again represents ice-volume. X_2 represents the absence (0) or presence (1) of Arctic sea-ice. 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.

PP12

First 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

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

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

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

The ice volume then evolves according to

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

where

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

PP12 [21] has distinct glaciation and deglaciation phases. During the glaciation phase ice volume increases according to variation in insolation. Due to the truncation of the forcing in the ice volume equation this model responds nonlinearly to variation in insolation. During the deglatiation phase the system relaxes towards a deglatiated state. The phase changes occur mainly due to the astronomical forcing, with ice volume only appearing in the glaciation-deglaciation switch.

The second component of a SSM relates observations to the system. The hidden Markov process is indirectly observed at discrete time intervals according to an observation process with density

$$\pi(Y_t|Y_{1:t-1}, X_{1:t}, \theta) = \pi(Y_t|X_t, \theta), \qquad t \geq 1. \tag{3}$$

The state variables used in phenomenological models are often abstractions of physical quantities. As a result the output of each model needs to be scaled and shifted in order to make a comparison with observations of δ^{18}O. The observation error also needs to be taken in to account. As such we relate an observation at time t to the observable component of the model using the observation process

$$Y_t \sim \mathcal{N}(D + HX_t, \Sigma_y)$$

where $H = (S, \mathbf{0})$ scales the output the observable state X, and D displaces it.

4 Methodology

Comment by Rich: Ignore this section for now. I'll edit this next

MOVED

Formally, we have a set of models $\{\mathcal{M}\}_{m=1,\ldots,M}$, where each model $\mathcal{M}_m$ has parameter θ_m. A standard Bayesian model selection approach involves first calculating the posterior distribution of the parameters given the set of data $Y_{1:T}$ for each model:

$$\pi(\theta_m|Y_{1:T}, \mathcal{M}_m) = \frac{\pi(\theta_m|\mathcal{M}_m)\pi(Y_{1:T}|\theta_m, \mathcal{M}_m)}{\pi(Y_{1:T}|\mathcal{M}_m)}, \tag{4}$$

where $\pi(\theta_m|\mathcal{M}_m)$ is the prior distribution of the parameters given the model. For most of the proposed models the likelihood term $\pi(Y_{1:T}|\theta_m, \mathcal{M}_m)$ is intractable (meaning it is not available in closed form or it is too computationally expensive to estimate), and so we can not calculate an analytical solution to Equation 4. This also prevents the use of many traditional methods for Bayesian inference, such as Markov chain Monte Carlo (MCMC). However, there have recently been numerous methodological advancements for inference in models in which the likelihood term is intractable, particularly for state-space models [22, 23]. The normalising constant $\pi(Y_{1:T}|\mathcal{M}_m)$ is termed the model evidence, and is used in evaluating the posterior model probabilities $\pi(\mathcal{M}_m|Y_{1:T}) \propto \pi(\mathcal{M}_m)\pi(Y_{1:T}|\mathcal{M}_m)$, where $\pi(\mathcal{M}_m)$ is the prior probability of model $\mathcal{M}_m$. Thus, a Bayesian model selection approach calibrates the parameters for each model, and indicates which models have the greatest explanatory power.

END OF MOVED MATERIAL

4.1 Inference in SSMs

We now return to the problem of selecting between multiple models $\{\mathcal{M}_m\}_{m=1,\dots,M}$, where each model has parameter space θ_m. In the context of SSMs a Bayesian inference approach involves estimating the posterior distribution

$$\begin{aligned}\pi(\theta_m, X_{1:t}|Y_{1:T}, \mathcal{M}_m) &\propto \pi(\theta_m|\mathcal{M}_m)\pi(X_{1:T}|\theta_m, \mathcal{M}_m)\pi(Y_{1:T}|X_{1:T}, \theta_m, \mathcal{M}_m) \\ &\propto \pi(\theta_m|\mathcal{M}_m)\pi(X_1|\theta_m, \mathcal{M}_m)\prod_{i=1}^{T-1}\pi(X_{i+1}|X_i, \theta_m, \mathcal{M}_m)\prod_{i=1}^{T}\pi(Y_i|X_i, \theta_m, \mathcal{M}_m),\end{aligned}$$

for each model $\mathcal{M}_m$. For the models of interest the transition densities $\pi(X_{i+1}|X_i, \theta_m, \mathcal{M}_m)$ are not available in closed form, and $X_{1:T}$ are only observed through the observations $Y_{1:T}$. As such it is impossible to calculate an analytical solution to $\pi(\theta_m, X_{1:t}|Y_{1:T}, \mathcal{M}_m)$, and we must rely on approximations. Monte Carlo methods can be used to obtain random samples approximating distributions of interest, and have become popular approaches to perfom inference in a wide range of fields, including SSMs. Unfortunately, the fact that the likelihood for each model is intractable means that many of the classical Monte Carlo approaches, such as MCMC can not be implemented. However, in recent years there has been a lot of development towards inference in SSMs (and models in general) with intractable likelihoods. It is not possible to review all of the possible approaches here, instead we describe one approach tailored towards SSMs that can be implemented with little tuning.

For Bayesian model selection we are also interested in obtaining the model evidence $\pi(Y_{1:T}|\mathcal{M}_m)$ for each model $m = 1, ..., M$. The ratio of the evidence for two models, say $\mathcal{M}_1$ and $\mathcal{M}_2$ is called the Bayes factor

$$B_{12} = \frac{\pi(Y_{1:T}|\mathcal{M}_1)}{\pi(Y_{1:T}|\mathcal{M}_2)} = \frac{\pi(\mathcal{M}_1|Y_{1:T})/\pi(\mathcal{M}_2|Y_{1:T})}{\pi(\mathcal{M}_1)/\pi(\mathcal{M}_2)}, \tag{5}$$

where $\pi(\mathcal{M}_m)$ is the prior for model $\mathcal{M}_m$. The Bayes factor summarises the evidence from the data in support of one model over another. If the prior probabilities for each model are equal then the Bayes factor is equivalent to the ratio of the posterior model probabilities. Estimating the Bayes factor is a challenging problem, as first we need to estimate the posterior distribution over the parameters

$$\pi(\theta_m|Y_{1:T},\mathcal{M}_m)=\int\pi(\theta_m,X_{1:t}|Y_{1:T},\mathcal{M}_m)dX_{1:T}, \tag{6}$$

from which we need to calculate the normalising constant

$$\pi(Y_{1:T}|\mathcal{M}_m)=\int\pi(\theta_m|\mathcal{M}_m)\pi(Y_{1:T}|\theta_m,\mathcal{M}_m)d\theta. \tag{7}$$

4.2 SMC

SMC approaches are well suited to our model selection problem; they provide estimates of normalising constants with relative ease, and the selection of tuning parameters can be mostly automated. We do not include the theoretical justifications of SMC here, the reader is referred to [24] for an overview. SMC approaches are population-based sampling methods aimed at obtaining a sample from some target distribution that is difficult to sample from directly, which will be denoted π_T. A series of intermediary distributions $\{\pi_t\}_{t=1,...,T}$ are designed that 'close-in' on the target distribution from some easily sampled distribution π_1. First, a collection of particles are sampled from π_1. The particles are then propagated through the intermediary distributions using a combination of importance sampling and resampling. The gradual changes between successive intermediary distributions mean that is possible to design efficient proposals at every iteration. In other words π_t is quite informative about π_{t+1}, even if π_1 and π_T are dissimilar.

One of the earliest and well known implementations of SMC is a collection of algorithms designed to sample from the sequence of posterior distributions $\pi_t=\pi(X_t|\theta,Y_{1:t})$, where the parameters θ are assumed known [25]. This is known as the filtering problem, and the SMC algorithms that have been proposed are termed particle filters (PFs). A widely used PF works as follows. First, a sample of N_x particles are sampled from the initial density $\pi(X_1|\theta)$, and given importance weight $\pi(Y_1|X_1,\theta)$. These particles are then repeatedly resampled, propagated and weighted, such that for each successive iteration the particles are a weighted sample of the posterior $\pi(X_t|\theta,Y_{1:t})$. The algorithm is as follows:

Particle Filter

- **PF 1.** Sample state particles $X_1^n\sim q_1(\cdot|\theta,Y_1)$
- **PF 2.** Weight state particles

$$w_1^n(X_1^n)=\frac{\pi(X_1^n|\theta)\pi(Y_1|X_1^n)}{q_1(X_1^n|\theta,Y_1)},\qquad W_1^n=\frac{w_1^n(X_1^n)}{\sum w_1^n(X_1^n)}$$

- **PF 3.** For time index $t=2,...,T$
  - **PF 3.1.** Sample ancestor particle index $\mathcal{A}_{t-1}^n\sim\mathcal{F}(W_{t-1}^n)$
  - **PF 3.2.** Propogate state particles $X_t^n\sim q_t(\cdot|X_{t-1}^{\mathcal{A}_{t-1}^n},\theta,Y_t)$ and extend trajectory $X_{1:t}^n=\left(X_{1:t-1}^{\mathcal{A}_{t-1}^n},X_t^n\right)$

– **PF 3.3.** Weight state particles

$$w_t^n(X_{1:t}^n) = \frac{\pi(X_t^n|X_{t-1}^{\mathcal{A}_{t-1}^n},\theta)\pi(Y_t|X_t^n)}{q_t(X_t^n|X_{t-1}^{\mathcal{A}_{t-1}^n},\theta,Y_t)}, \qquad W_t^n = \frac{w_t^n(X_{1:t}^n)}{\sum w_t^n(X_{1:t}^n)}$$

---

where superscript n indicates that the operation is performed for all $n = 1, ..., N_x$. Here $\mathcal{F}(W_{t-1}^{1:N_x})$ samples $n = 1, ..., N_x$ according to weights W_{t-1}^n. We use q_t to denote the proposal distribution at time t. For problems in which the transition density is intractable the term $\pi(X_t|X_{t-1},\theta)$ needs to cancel in step **PF 3.3**. This can achieved by setting $q_t = \pi(X_t|X_{t-1},\theta)$ for $t > 1$, such that proposals are just simulations from the model.

Comment by MC: I believe *can* is here crucial. Isn't the essence of the Golighty-Wilkinson proposal to propose an alternative to this proposal, but yet keeping a ratio $\frac{q}{\pi}$ which is analytical ? Should this be mentioned at this stage?

This choice will typically lead to many proposals being far from the observations, such that the small number of proposals in the region of the observation gain most of the weight. This is known as degeneracy, and is a common problem in PFs (and SMC algorithms in general). Resampling the state particles in each iteration improves the approximation of later iterations as only important particles are propagated forward. Possible resampling strategies are discussed in [26].

Where possible, including information from the next observation in the proposals should lead to more equal weights. For the SSMs considered in this article it is possible to condition on the next observation under the Euler-Maruyama approximation, which should move more particles to regions of high posterior probability. Starting with the Euler-Maruyama approximation

$$X_{t+\Delta t} \sim \mathcal{N}(X_t + \mu\Delta t, \Sigma_x \Delta t),$$

we want to design a proposal that moves the observable state closer an observation at time T, where $T - t$ is usually too large for a single Euler step.

Firstly, we can consider adding a small perturbation to the mean of the proposal of the form

$$JS^{-1}(Y_T - (D + S(X_t + \mu(T - t)))),$$

where a single Euler step is used to predict the value of X_T, which is then compared to the the observation (residual nudging reference?). J determines how strongly the observation influences the proposal.

Comment by MC: is it useful here to write the analytical form of $\frac{\pi_t}{q}$ or is it considered obvious enough for the audience?

The optimal choice will depend on the relative variance of the model, and the observation. For the case $\Delta t = T - t$ it is desired that $J \to 1$ as $\Sigma_y \big/ S^2\Sigma_x(T-t) \to 0$, and $J \to 0$ as $S^2\Sigma_x(T-t) \big/ \Sigma_y \to 0$. This suggests $J = S^2\Sigma_x(T-t) \big/ \left(S^2\Sigma(T-t) + \Sigma_y\right)$, a form shared with the Kalman gain matrix. When using smaller integration time steps J is multiplied by $\Delta t/(T - t)$ to give the proportion of the perturbation that occurs over the smaller time step.

A similar proposal has been more formally derived by using Brownian bridges conditioned on observations [27], which also reduces the variance of the proposals as the observation is neared. 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 will have low likelihood). The specific variance reduction (when taking the scaling term in to account) is given by:

$$\left(\frac{S^2\Sigma_x(T-t) - S^2\Sigma_x\Delta t + \Sigma_y}{S^2\Sigma_x(T-t) + \Sigma_y}\right) S^2\Sigma_x\Delta t,$$

which scales down the variance based on the model variance, the observation variance, the time until the observation, and the integration time step.

Comment by MC: I am unsure but I believe the S^2 before Σ_x is superfluous.

The ratio can be considered as the variance remaining after the integration step has been performed relative to the total variance. This ratio is close to 1 for $\Delta t \ll (T-t)$, and for $S^2\Sigma_x\Delta t \ll \Sigma_y$. Whereas in the case $\Delta t = \Delta T$, and $\Sigma_y \ll S^2\Sigma_x\Delta t$ the proposal variance is approximately the observation variance.

Combining these changes gives a proposal of the form

$$\begin{aligned} X_{t+\Delta t} \sim \mathcal{N}(X_t + \mu\Delta t + \frac{S^2\Sigma_x\Delta t}{S^2\Sigma_x(T-t)+\Sigma_y} S^{-1}(Y_T - (D + S(X_t + \mu(T-t)))), \\ \left(\frac{S^2\Sigma_x(T-t) - S^2\Sigma_x\Delta t + \Sigma_y}{S^2\Sigma_x(T-t) + \Sigma_y}\right) S^2\Sigma_x\Delta t), \end{aligned} \tag{8}$$

Comment by MC: I am unsure but I believe the S^2 before Σ_x is superfluous.

such that for uninformative observations the proposal is approximately the standard Euler-Maruyama approximation, and for informative observations the value of X_T is drawn from $\mathcal{N}((Y_T - D)/S, \Sigma_y/S^2)$. In other words the proposal will be centered on the observation with variance equal to the observation error.

An important aspect of the PF is that it provides an unbiased estimate of the marginal likelihood $\pi(Y_{1:T}|\theta)$ as follows

$$\hat{\pi}(Y_{1:T}|\theta) = \hat{\pi}(Y_1) \prod \hat{\pi}(Y_t|Y_{1:t-1}, \theta), \tag{9}$$

where the components of the product are unbiased estimates of $\pi(Y_t|Y_{1:t-1}, \theta)$, obtained by averaging over the unnormalised weights in each iteration of the algorithm

$$\hat{\pi}(Y_t|Y_{1:t-1}, \theta) = \frac{1}{N_x} \sum w_t^n(X_{1:t}^n). \tag{10}$$

It is possible to use these unbiased estimates in place of the corresponding intractable quantities in a number of Monte Carlo algorithms, for example PMCMC [22] and SMC^2 [23] use this replacement in an MCMC and an SMC framework respectively.

4.3 SMC^2

The SMC^2 algorithm [23] embeds the particle filter within an SMC algorithm targetting the sequence of posteriors

$$\pi_0 = \pi(\theta), \qquad \pi_t = \pi(\theta, X_{1:t}|Y_{1:t}), t \geq 1$$

This is achieved by sampling N_θ parameter particles from the prior. A PF of N_x particles is attached to each parameter particle, providing unbiased estimates to intractable quantities as the parameter particles are propagated through each intermediary distribution. The amount of degeneracy in the importance weights for the parameter particles is monitored by calculating the effective sample size (ESS)

$$\text{ESS} = \left(\sum (\Omega^m)^2\right)^{-1}$$

where Ω^m are the normalised weights. When the ESS falls below some threshold (usually $N_\theta/2$) the particles are resampled to discard low-weight particles. However, resampling alone would lead to few unique particles in the parameter space. Particle diversity is improved by running a PMCMC algorithm that leaves $\pi(\theta, X_{1:t}|Y_{1:t})$ invariant, specifically the PMMH algorithm [22]. The algorithm is given below, but the reader is referred to the original paper for the theoretical justification [23]. Note that superscript j indicates that the operation is performed for all $j = 1, ..., N_\theta$, and as before superscript n indicates that the operation is performed for all $n = 1, ..., N_x$.

SMC^2

- **SMC^2 1.** Sample parameter particles $\theta^j \sim \pi(\theta)$
- **SMC^2 2.** Set importance weights

$$\omega^j = 1, \qquad \Omega^j = \frac{\omega^j}{\sum \omega^j}$$

- **SMC^2 3.** For $t = 1, ..., T$
  - **SMC^2 3.1.** For each parameter particle θ^j sample $X_{1:t}^{n,j}$ by performing iteration t of the PF.
  - **SMC^2 3.2.** Calculate $\hat{\pi}(Y_t|Y_{1:t-1}, \theta^j)$ and perform a weighted average over the parameter particles to obtain $\hat{\pi}(Y_t|Y_{t-1}) = \sum \Omega^j \hat{\pi}(Y_t|Y_{1:t-1}, \theta^j)$
  - **SMC^2 3.3.** Update the importance weights

$$\omega^j = \omega^j \hat{\pi}(Y_t|Y_{1:t-1}, \theta^j), \qquad \Omega^j = \frac{\omega^j}{\sum \omega^j}$$

  - **SMC^2 3.4.** If the ESS falls below some threshold
    - **SMC^2 3.4.1.** Resample $\left(\theta^j, X_{1:t}^{n,j}\right)^*$ according to weights Ω^j

* **SMC^2 3.4.2.** Sample $\left(\theta^p, X_{1:t}^{n,j}\right)^{**} \sim K\left(\cdot|\left(\theta^j, X_{1:t}^{n,j}\right)^*\right)$ where K is a PMCMC kernel that leaves $\pi(\theta, X_{1:t}^n|Y_{1:t})$ invariant.
* **SMC^2 3.4.3.** Set $\left(\theta^j, X_{1:t}^{1:n,j}\right) \leftarrow \left(\theta^j, X_{1:t}^{n,j}\right)^{**}$
* **SMC^2 3.4.4.** Set importance weights

$$\omega^j = 1, \qquad \Omega^j = \frac{\omega^j}{\sum \omega^j}$$

Comment by MC: I believe we need to give a bit more detail about the 3.4.1 resampling step, I mean, the details (binomial vs residual)

Comment by MC: Step 3.4.3: shouldn't it be:

1. **SMC^2 3.4.3.** Set $\left(\theta^j, X_{1:t}^{1:n,j}\right) \leftarrow \left(\theta^p, X_{1:t}^{n,j}\right)^{**}$?

The tuning parameters are the number of particles N_θ and N_x, and the proposal distributions for the PMCMC steps in **SMC^2 3.4.2.** Typically N_θ will be decided by the available computational resources. A low N_x can be used for early iterations, but must be increased for larger times. An insufficient number of state particles will have a negative impact on the PMCMC acceptance rate. Automatic calibration of N_x is discussed in [23], where it is suggested that N_x is doubled whenever the acceptance rate of the PMCMC step becomes too small. The fact that we have a collection of particles in each iteration allows automated calibration of the PMCMC proposals; for example by using the sample mean and variance to design a sensible random-walk proposal, or independent Gaussian proposals.

SMC^2 naturally provides an estimate of the model evidence. The model evidence can be decomposed as

$$\pi(Y_{1:T}) = \pi(Y_1) \prod \pi(Y_t|Y_{t-1}) \tag{11}$$

In each iteration of SMC^2 the term

$$\hat{\pi}(Y_t|Y_{t-1}) = \sum \Omega^j \hat{\pi}(Y_t|Y_{1:t-1}, \theta^j)$$

calculated in step **SMC^2 3.2** gives an unbiased estimate of $\pi(Y_t|Y_{t-1})$. An estimate of the model evidence is provided by the product of these terms.

5 Results

5.1 Simulation study

In order to gain confidence in the ability of our SMC^2 algorithm for both model selection and calibration, we begin with a simulation study. We simulate a single random trajectory from a given model and parameter setting and draw observations from the observation process. We then show that the posterior distributions recover the true value of the parameters (Figure 2), and that the Bayes factors correctly identify the true generative model (Table 2).

We present results from analysing two simulated datasets: one in which data are generated from an unforced version of SM91, and one in which a forced version of SM91 is used. We refer to these datasets as SM91-u and SM91-f respectively. The parameters used are given in Table 2 and are comparable to those estimated from real data. Crucially, realistic values of the measurement error variance are used. Observations are taken every 3kyr over the past 780kyr giving 261 observations in each dataset, which is comparable to a low resolution sediment core. The model evidence and posteriors are then calculated for each of five models. We use forced and unforced versions of SM91 and T06, as well as the forced model PP12. We do not consider an unforced PP12 model as the deglatiation-glaciation transition depends only on the astronomical forcing, whereas SM91 and T06 both oscillate in the absence of any external forcing. The models contain between 10 and 16 parameters. We then test the ability of our inference algorithms to 1) discriminate between the five models by estimating the Bayes factors; 2) recover the parameters used to generate the data; and 3) reconstruct the underlying climate trajectories $x_{1:T}$. The priors used for each model are given in Table 7.

TABLE 7 ABOUT HERE

The estimated log Bayes factors (log BF) are given in Table 2. The Bayes factor of two models is the ratio of the model evidences. However, the logarithm of the Bayes factor provides a more natural scale for interpretation, with the log BF calculated as the difference between two log evidences, and this is what is reported in Table 2. A common interpretation suggests that a log BF of 3 is strong evidence in favour of one model over another, and that a log BF of 5 is a very strong indication that one model is superior to another [28]. Conversely, a negative score indicates the same strength of evidence but in the other direction (for the other model).

For both simulated datasets, we find a strong preference for the correct model. When applied to SM91-f, the correct model (the forced SM91 model) is overwhelmingly favoured. The log BF to the next most supported model (the forced T06 model) is estimated to be 24.7, indicating decisive evidence in favour of the true model. It is interesting to note that if we remove the forced SM91 model from the analysis, we find decisive evidence in favour of the forced T06 model over any of the other unforced models (a log BF of at least 27.7), showing that the astronomical forcing has explanatory power even in the wrong model. This is not particularly surprising, because in both models the astronomical forcing acts as a synchronisation agent, controlling the timing of terminations, and has a strong effect on the likelihood. This is reassuring. It means that palaeoclimate scientists can implicitly rely upon this effect when arguing for the importance of the astronomical forcing, as it allows us to infer its importance even when using an incorrect simulator (for we surely are).

Comment by Rich: Does this make sense - it may be a little strong.

When applied to SM91-u the log BF again correctly identifies the correct generative model, although the support for the unforced and forced SM91 models is now much closer (with a log BF of 2.9 in favour of the unforced model). In cases where the forcing does not add any explanatory power this is an expected result, as the unforced version of SM91 is nested within the forced version of SM91, and is recovered by setting $\gamma_P = \gamma_C = \gamma_E = 0$. This effect is also noticeable when comparing the forced and unforced T06 models, with the unforced version being preferred with a log BF of 3.5.

These experiments clearly show that there is sufficient information in the data to easily detect the correct parametric form of the simulator in each case. Care needs to be taken when using Monte Carlo estimates of the model evidences, as the Monte Carlo error can be considerable. Experimentation suggests that the model evidence estimates have a variance of approximately an order of magnitude, and hence the log Bayes factors should be viewed as having uncertainty of approximately plus or minus 6 on the log scale. We thus suggest that we can only be confident that there is strong evidence for one model over another if the estimated log BF is at least 10. Note that our conclusions from the simulation study are mostly unaffected by this noise, aside from further confirming that the difference between the forced and unforced version of the same parametric model is small. The magnitude of the estimation error can be decreased by using more particles in the SMC^2 algorithm, but this will require very long computational runs. uUsing $N_x = N_\theta = 1000$.... JAKE CAN YOU SORT - takes ??? hours on a GPU..... and the cost increases as $O(N_x^2)$????. Nested sampling [?] is an alternative approach to estimating BFs that allows for easier estimation of sampling uncertainty. However, implementing nested sampling for intractable models with a large number of missing values (as in this case, where the trajectories are 'missing data') has not yet proven possible, but if it can be made to work, would be worth pursing.

Comment by MC: although this discussion avoids the problem of the robustness vs the choice of priors, especially if the latter are vague

TABLE 2 ABOUT HERE

The marginal posterior distributions for the parameters for the forced SM91 model applied to the SM91-f dataset are shown in Figure 2. We are able to recover the parameters used to generate the data, with the true values lying in regions of high posterior probability. The posteriors for q and σ_3 do not deviate much from the prior, suggesting that a wide range of values explain the data equally well. Further simulation studies and details are available in [29].

Figure 2 about here.

Comment by Rich: Can you add a plot showing the three SM91 trajectories, with the MC confidence interval, and the data from y_1. I think this would help illustrate, even if it is likely to end up in supplementary material.

5.2 ODP677

Comment by Rich: We need more detail on the data here.

We now analyse data from the ocean drilling programme (ODP). We focus here on ODP677, which is......JAKE..., which is of interest because it has been dated by two different groups. Analysis of other datasets can be found in [29]. When an ocean core is first extracted and analysed, the measurements are a sequence of proxy measurements ($\delta^{18}O$????) corresponding to different depths in the core. The core is then dated through the use of an age-model. When different age models are used, they can give slightly different chronologies for the same core. ODP677 is interesting in this regard, as it has been dated by two different groups. In [2], a depth-derived age model is used. 'Age-control points' are identified in the core (such as glacial terminations, magnetic reversals, etc), and then ages for all the measurements are inferred from these control points, while accounting for compression in an involved heuristic process. We will refer to the data from this study as ODP677-u, where the 'u' denotes *unforced*. In [1], the core is dated using an orbitally-tuned age model. They assume that the astronomical forcing is correlated with .... JAKE/MICHEL - can you provide a description here. We refer to this dataset as OPD677-f, where the 'f' denotes that the data have been tuned using the astronomical forcing.

The estimated log Bayes factors for each model are given in Table 3. For ODP677-u, the unforced T06 model is best supported, but the estimated BF for the unforced T06 model compared to the unforced SM91 model is within our Monte Carlo bounds, and so it is not possible to confidently assert that T06 is superior to SM91 for explaining these data. There is reasonable to strong evidence (given our Monte Carlo uncertainty) that the unforced models are preferred to the forced model, i.e., there is reasonable evidence that we do not need an astronomically forced model to explain the data. This resembles the results from SM91-u, where the forced models are being penalised for containing extra parameters with little explanatory power. Note that the two unforced models are both decisively preferred compared to PP12.

TABLE 3 ABOUT HERE

When we analyse ODP677-f, the astronomically tuned data, the results are the complete reverse. We now find that the PP12 model is strongly indicated by the data, and that the three forced models are all decisively preferred to the two unforced models, i.e., we find overwhelming evidence using these data that astronomical forcing is necessary to explain the data! The orbital tuning of ODP677-f is the most likely explanation for this. In SM91 and T06 the astronomical forcing acts similarly as a pacemaker, controlling the timing of glacial inceptions and terminations. While in PP12 the astronomical forcing dictates 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 in a similar fashion to ODP677-f. Forced SM91 and forced T06 are both more supported than the unforced versions, with strong Bayes factors. Again it is difficult to determine if T06 is more supported by the data than SM91.

This result is our second key finding. Namely, that inference about the best model is strongly affected by the age-model used to date the data. It is vital that modelling assumptions in the dating methods should be understood when performing inference on palaeoclimate data. We suggest that this demonstrates that the approach of first dating the data, and then carrying out down-stream analyses given this dating, ignoring the uncertainty, is at best harmful, and at worst, completely undermines any subsequent inference about the dynamic mechanisms at play.

Figure 3 about here.

The marginal posterior distributions of the parameters in the SM91 model when fit to the ODP677-f data are shown in Figure 3. The astronomical forcing scaling parameters γ_P and γ_E have very small posterior probabilities at 0, suggesting that both precession and obliquity are informative about ODP677-f. The scaling parameter for obliquity is typically larger than that for precession. On the other hand for the coprecession parameter γ_C, 0 is in a region of very high posterior probability, indicating that a model selection experiment might support a forced model without coprecession. This is true of the posterior values from any of the forced models. The stochastic scaling parameters are larger than in simulation study, which is expected when data has not been generated from the model.

Finally, the density of the ratios $\frac{\gamma_P}{\gamma_E}$ and $\frac{\gamma_C}{\gamma_P}$ are shown in Figure 4. While the individual parameters are not directly comparable due to each model having different spatio-temporal scales, the ratio between the astronomical scaling terms gives the shape of the forcing function, and is comparable. The forcing function in PP12 is ommitted due to the fact that the forcing is truncated, making the parameters incomparable. While there is a slight difference, the posteriors are similar enough to suggest that each model is synchronising to the same forcing, with the obliquity scaling term being dominant.

Figure 4 about here.

6 Alternative approaches

Decide what else to include, i.e.

- Kalman filter works in linear Gaussian case, and there exist a number of extensions that could be used to approximate the likelihood for nonlinear models. Less computational cost than SMC^2, but we no longer have an ‘exact approximation’.
- PMCMC could be used with reversible-jump steps. What are the pros/cons compared to SMC^2.
- ABC can be used.

  Comment by MC: Is that true for model evidence as well?

  In particular Richard has some comments on using summary statistics.
- Information criterion approaches?

What’s missing? UKF ABC comparison Details of statistical improvements - SMC^2 for model selection G and W updates.

7 Conclusions

We have two key conclusions. The first is that Monte Carlo technology and computer power are now both sufficiently advanced, that with work, it is possible to fully solve the Bayesian model selection problem for a wide class of phenomenological models of the glacial-interglacial cycle. It came as a surprise to us, that even relatively short time-series of observations contain sufficient information to discriminate between many of the models. A priori, we had expected to find that there was simply insufficient information in the data, given the level of noise, to solve simultaneously the filtering and the model calibration problems, and then still distinguish between the models. That we are able to do this, contrasts strongly with the viewpoint set out in [11]:

> Most simple models of the [...] glacial cycles have at least four degrees of freedom [parameters], and some have as many as twelve. Unsurprisingly [...this is] insufficient to distinguish between the skill of the various models

In 1999 this viewpoint may have been true. We lacked both the computer power and the algorithmic knowledge to do Bayesian inference for the parameters, never mind estimating the Bayes factors. However, recently developments in Monte Carlo methodology, and the massive increase in computing power (including the utilisation of GPUs), means that the calculations are now possible. Using only 261 observations, we are able to learn up to 16 parameters, state trajectories containing 261 × 3 values, and calculate the marginal evidence. Moreover, these evidences are sufficiently different (and able to be estimated with sufficient accuracy) that we can confidently discriminate the ability of the models to explain the data.

Comment by Rich: We don't want to be too critical of Roe and Allen, but we do want to stress why this work is cool!

Our second conclusion concerns the need to avoid theory laden data. The results from analysing the ODP677 data, show that the age model used to date the core become critical when the data are subsequently used to make scientific judgements. One age model gives overwhelming evidence that the astronomical forcing is vital for explaining the data, while another age model suggests the opposite. This suggests that analysing the data in stages, cutting feedbacks between uncertainties, is not sensible. While this observation is not new, we believe it is the first time the effect of the age model on subsequent analyses has been so clearly demonstrated. Instead of first dating the core, and then using those dates (with or without uncertainties), we instead need to jointly estimate the age model at the same time as testing further hypotheses, accounting for all the joint uncertainties.

Comment by MC: I believe one outstanding problem when we arrive to this conclusio is that the palaeclimatic conclusions all sound a bit naive. Remember that the SM90 model is a bit ancient, and that the T06 model rests an a controversial mechanism, and, again, in both cases, we are more discriminating the dynamics than the actual mechanisms that are featured by these models. With a benthic foram core, it seems indeed a bit overstretch to claim discerning the difference between a g/i mechanism resting on biogeochemical instability, and one resting on a sea-ice switch mechanism. We have to be careful not to be interpreted in this way. One necessary (and not sufficient) condition for appreciating the role of biogeochemistry or sea-ice (if we want to speak of these two models in particular) is to actually have data about these physical entities.

Comment by Rich: Michel, does the new results section and emphasis answer these concerns

Comment by MC: On the other hand, looking at the amount of code, notes written etc. that Jake and other of us have delivered in the last four years, we must have the material to say a bit more about our experience in model selection. E.g.: how many particles do we need? What is the rate of convergence? What is the improvement provided by the Golighty-Wilkinson algorithm : was it necessary to use this proposal? How did Jake chose the MH proposal when resampling particles? etc. All these elements are essential because if we admit that this was a "simple" problem using after all oversimplified models, the merit of our experience is that it allows us to envision how this work can be extended to more data (mean : CO2, ice volume, methane, . . .) and more physically explicit models. A few comments are inlined, but really we need to give this more thought.

Comment by Rich: I'll think about what we should add to Section 4

The experiments included in this paper can be extended in several ways. Firstly, we considered only a handful of models, and both the number and complexity of models can be increased. With the approach described here, extra models can be included by running the SMC^2 algorithm for each model. This has the benefit that the entire experiment does not need to be redesigned/repeated for different combinations of models. Different astronomical forcing set-ups can also be considered. For example, the astronomical forcing terms are often tested independently. This can easily be achieved by setting undesired astronomical scaling terms to 0 in our forced models. Making the forcing term state dependent, such that an increase in sea-ice increases albedo , which in turn alters the influence of variation in insolation is also a possibility.

Comment by MC: it is not so much a problem of sea-ice but, for example, the ablation area grows non-linearly with insolation; there are references for this but we can see this later.

Finally, we do not need to limit ourselves to a single dataset. The observation model can be extended to compare the state of the system to multiple cores. Likewise, multivariate observations could be used; SM91 models both ice-volume and CO_2 concentration, and records exist for both of these quantities.

Overall, we hope that this work acts as a proof of concept. Careful statistical analysis combining data and models can lead to insights in palaeoclimate science.

References

[1] L. Lisiecki and M. Raymo, “A pliocene-pleistocene stack of 57 globally distributed benthic δ^{18}O records,” *Paleoceanography*, vol. 20, p. PA1003, 2005.

[2] P. Huybers, “Glacial variability over the last two million years: an extended depth-derived agemodel, continuous obliquity pacing, and the pleistocene progression,” *Quaternary Science Reviews*, vol. 26, pp. 37–55, 2007.

[3] A. Berger and M. Loutre, “Astronomical theory of climate change,” *Journal de Physique IV*, vol. 121, pp. 1–35, 2004.

[4] J. Hays, J. Ibrie, and N. Shackleton, “Variations in the Earth’s orbit: pacemaker of the ice ages.,” *Science*, vol. 194, pp. 1121–1132, 1976.

[5] P. Huybers and C. Wunsch, “Oliquity pacing of late Pleistocene terminations,” *Nature*, vol. 434, pp. 491–494, 2005.

[6] L. Lisiecki, “Links between eccentricity forcing and the 100,000-year glacial cycle,” *Nature Geoscience*, vol. 3, pp. 349–352, 2010.

[7] P. Huybers, “Combined obliquity and precession pacing of late Pleistocene deglaciations,” *Nature*, vol. 480, pp. 229–232, 2011.

[8] M. Crucifix, “Oscillators and relaxation phenomena in Pleistocene climate theory,” *Transactions of the Philosophical Transactions of the Royal Society A*, vol. 370, pp. 1140–1165, 2012.

[9] M. Crucifix, “Why could ice ages be unpredictable,” 2013. 1302.1492, arXiv.org.

[10] M. Cane, P. Braconnot, A. Clement, H. Gildor, S. Joussaume, K. M., M. Khodri, D. Paillard, S. Tett, and E. Zorita, “Origin and consequences of cyclic ice-rafting in the northeast atlantic ocean during the past 130,000 years,” *Quaternary research*, vol. 29, pp. 142–152, 1988.

[11] G. Roe and M. Allen, “A comparison of competing explanations for the 100,000-yr ice age cycle,” *Geophysical Research Letters*, vol. 26, pp. 2259–2262, 1999.

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

[13] A. Berger, “Long term variations of daily insolation and Quaternary climate changes,” *Journal of Atmospheric Sciences*, vol. 35, pp. 2362–2367, 1978.

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

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

[16] K. Hasselmann, “Stochastic climate models,” *Tellus*, vol. 28, pp. 473–485, 1976.

[17] B. Lindner, A. García-Ojalvo, J. abd Neiman, and L. Schimansky-Geier, “Model experiments on the 100,000-yr glacial cycle,” *Physics Reports*, vol. 392, pp. 321–424, 2004.

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

[19] M. Crucifix, “How can a glacial inception be predicted?,” *The Holocene*, vol. 21, pp. 831–842, 2011.

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

[21] 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*, vol. 8, pp. 2031–2037, 2012.

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

[23] N. Chopin, P. Jacob, and O. Papaspiliopoulos, “SMC^2: an efficient algorithm for sequential analysis of state-space models,” 2012. 1101.1528, arXiv.org.

[24] P. Del Moral, A. Doucet, and A. Jasra, “Sequential Monte Carlo samplers,” *Journal of the Royal Society Series B*, vol. 68, pp. 411–436, 2006.

[25] N. Gordon, D. Salmond, and A. Smith, “Novel approach to nonlinear/non-Gaussian Bayesian state estimation,” *IEEE Proceedings F*, vol. 140, pp. 107–113, 1993.

[26] J. Liu and R. Chen, “Sequential Monte Carlo methods for dynamic systems,” *Journal of the American Statistical Association*, vol. 93, pp. 1032–1044, 1998.

[27] A. Golightly and D. Wilkinson, “Bayesian inference for nonlinear multivariate diffusion models observed with error,” *Computational Statististics & Data Analysis*, vol. 52, pp. 1674–1693, 2008.

[28] R. Kass and A. Raftery, “Bayes factors,” *Journal of the American Statistical Association*, vol. 90, pp. 773–795, 1995.

[29] J. Carson, *Uncertainty Quantification in Palaeoclimate Reconstruction.* PhD thesis, University of Nottingham, 2014.

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

Table 1: Prior distributions used for each model in both the simulation study and the analysis of ODP677. Sections indicate parameters used to scale the astronomical forcing (absent in unforced models), parameters of the phenomenological model, and observation model respectively. JAKE/MICHEL - can we say something about how these priors were chosen.

Model		Dataset	
		SM91-u	SM91-f
SM91	Forced	-2.9	0
	Unforced	0	-52.4
T06	Forced	-21.8	-24.7
	Unforced	-18.3	-61.4
PP12	Forced	-49.6	-52.5

Table 2: Log Bayes factors for comparing five different models on the two simulated datasets. In each column, the log BF is with respect to the true generative model, so that positive values indicate support for that model over the true model, and negative values indicate support for the true model. SM91-u is data generated from an unforced version of SM91, whereas SM91-f is generated from an astronomically forced version of SM91. Values of the log evidence can be reconstructed from noting that $\log Z = 69.2$ for the unforced version of SM91 on the SM91-u dataset, and that $\log Z = 94.7$ for the forced SM91 model on the SM91-f dataset. The parameter values used to generate SM91-f are: $p = 0.8$, $q = 1.6$, $r = 0.6$, $s = 1.4$, $v = 0.3$, $\sigma_1 = 0.2$, $\sigma_2 = 0.3$, $\sigma_3 = 0.3$, $\gamma_P = 0.3$, $\gamma_C = 0.1$, $\gamma_E = 0.4$, $D = 3.8$, $S = 0.8$, $\sigma_y = 0.1$. For SM91-u we set $\gamma_P = \gamma_C = \gamma_E = 0$. QUESTION: I've played around with various ways of presenting this information, including the original evidence, the log evidence, and the log BF compared to the worst model. I think this is the clearest - do you agree? WARNING: rounding errors here as I've taken Jake's Z value to two dp and logged it.

Model		Dataset	
		ODP677-u	ODP677-f
SM91	Forced	-8.4	-14.3
	Unforced	-3.9	-37.0
T06	Forced	-6.2	-10.6
	Unforced	0	-29.4
PP12	Forced	-13.9	0

Table 3: The estimated log BFs for the five different models. The Bayes factors are given in comparison to the best model for each dataset (which thus has a log BF of zero). The ODP677 dataset is analysed twice. ODP677-u refers to a dating model derived by [2] using a depth derived model, whereas ODP677-f is an astronomically tuned dating model described in [1]. Values of the log evidence can be reconstructed using the estimates $\log Z = 65.0$ for the unforced version of T06 on the ODP677-u dataset, and that $\log Z = 78.9$ for the PP12 model on the ODP677-f dataset. The prior distributions used are given in Table 7. WARNING: rounding errors here as I've taken Jake's Z value to two dp and logged it. Values of the log evidence

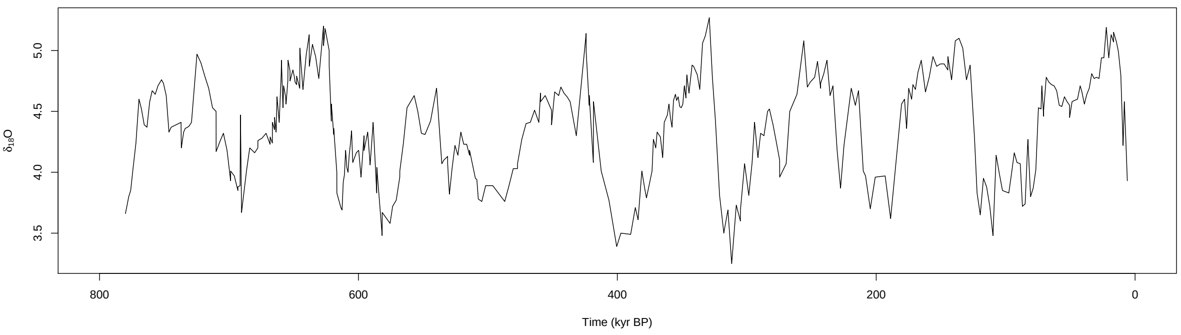

Figure 1: JAKE - Please can you make the axis labels larger? Observed δ^{18}O from ODP677 [12] corresponding to the past 780 kyr. This dataset has been dated without the use of orbital tuning [2].

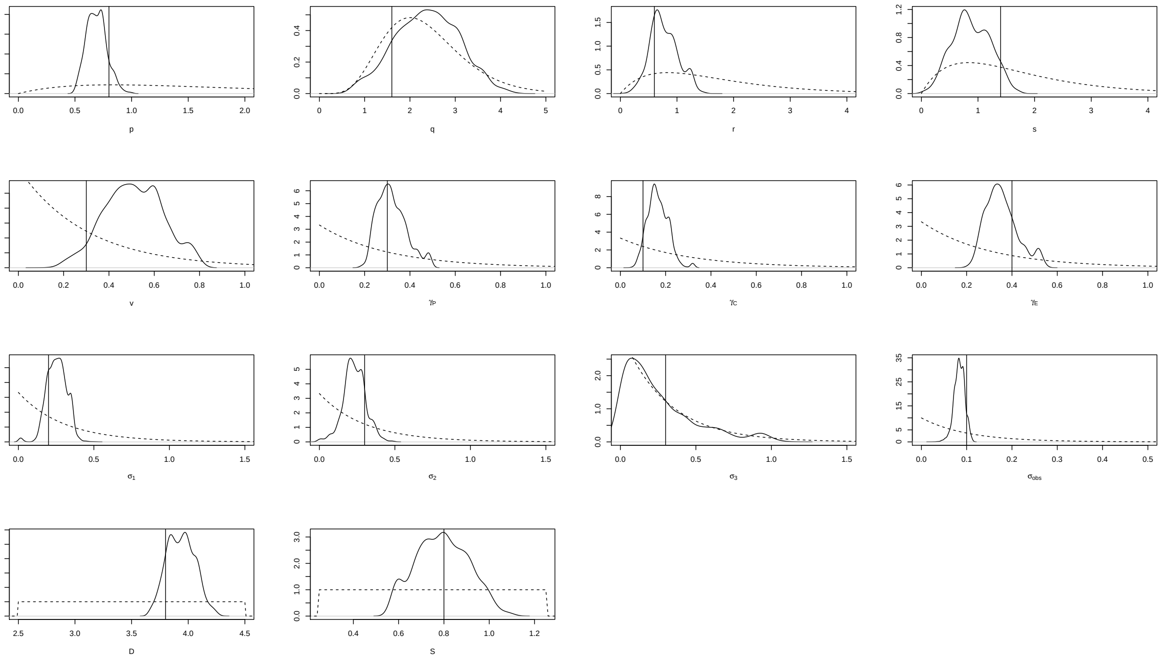

Figure 2: JAKE - Please can you make the axis labels larger? Marginal posterior distributions for the parameters of the forced SM91 model when fit to the SM91-f dataset. Vertical lines show the parameter values used to generate the data, and dashed lines represent the prior distribution.

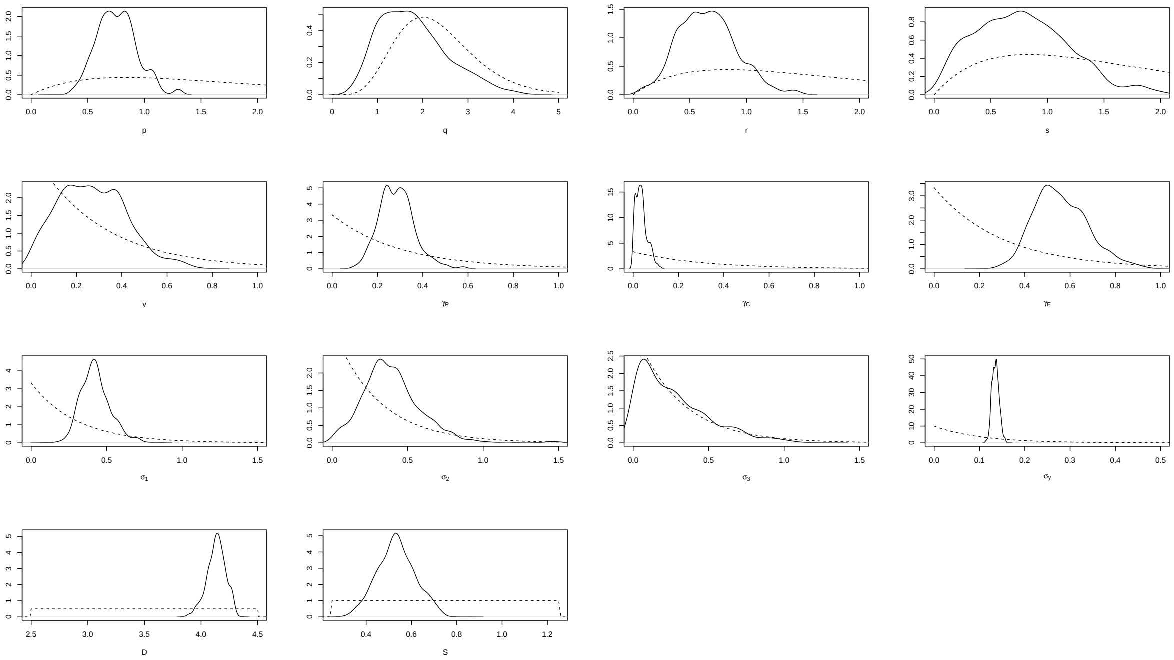

Figure 3: JAKE - Please can you make the axis labels larger? Marginal posterior distributions for the fully forced SM91 model on ODP677-f. Dashed lines represent the prior distributions, and solid lines the posteriors. The prior distributions used are given in Table refTab:Priors

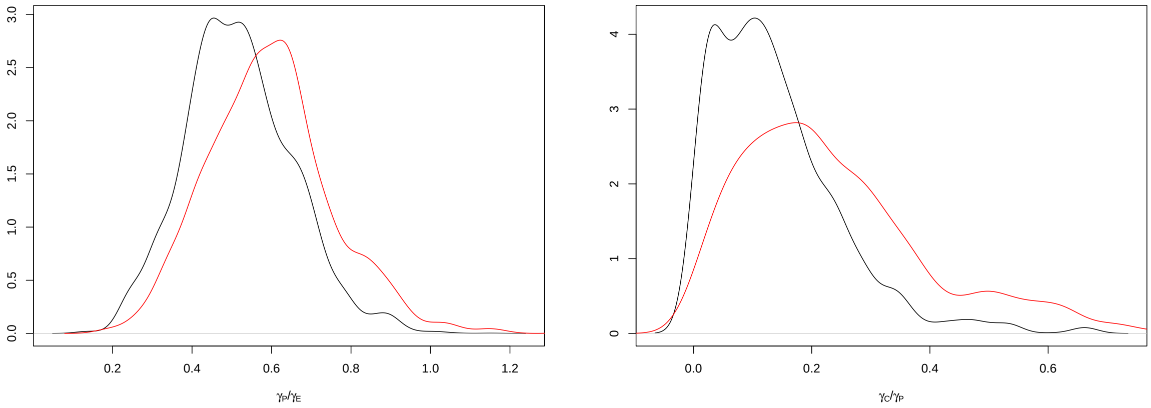

Figure 4: JAKE - Please can you make the axis labels larger? Posterior density plot of the ratio of the orbital scaling terms for the SM91 model (black line), and T06 model (red line).

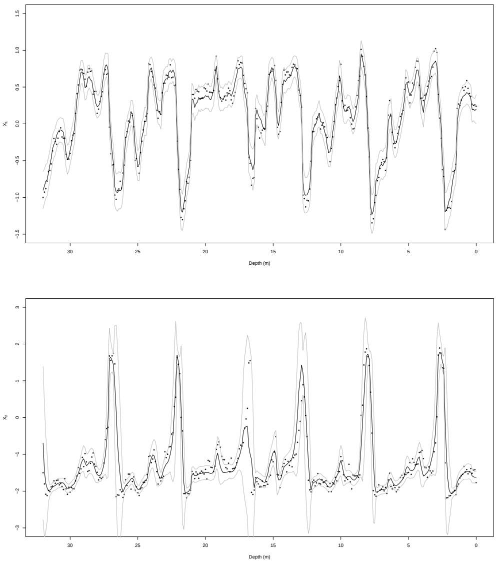

Figure 5: JAKE: These are the wrong figures - could you put in the correct figure. Add the truth on, and the data on x_1.