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

December 3, 2014

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

Abstract

The prevailing viewpoint in palaeoclimate science, is that a single palaeoclimate record contains insufficient information to discriminate between most competing explanatory models. Our aim here is to show that this is not the case. Recent developments in Monte Carlo methodology, combined with 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 *et al.* 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 select between competing models.

Our results highlight the problem of using pre-dated palaeoclimate records. By analysing a dataset that has been dated by two different authors, we reach contradictory conclusions, thus indicating that the current practice of first dating a record, and then analysing it to draw scientific conclusions about the dynamics of the system in an independent analysis is ....We conclude that dating and analysis should be done in a single process....

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 "proxy measurements" from a number of difference sources, including microbes found in sediment and water from ice cores [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 ago (Mya)[ref?]. Since then the climate has fluctuated between cold periods, in which glaciers expand, and warm periods in which the glaciers retreat. This is known as the glacial-interglacial cycle. In the early Pleistocene – a period from around 2.5 Mya ago until 11 thousand years ago (kya) – the average period of the glacial-interglacial cycle was around 40 kyr. However, around 1 Mya ago, the average period changed to approximately 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.

An important driver of the cycles is the incoming solar radiation, termed "insolation". Insolation varies through time 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, the variation in the eccentricity only accounts for about 2% of the variation in the insolation. 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.

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 phenomenological models, which are consistent with, but not necessarily derived from the physical theory of a system. The three models we consider (detailed in §), each involve 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.

A common viewpoint is that the information contained in a single proxy record is not sufficient to distinguish between the numerous proposed models [8, 9]. The aim of this paper is to demonstrate that this is not the case, and that careful empirical evaluation of candidate dynamical models is possible, including assessing the importance of the various orbital characteristics to the glacial-interglacial cycle. By treating the problem as a model selection problem with unknown parameters, we are able to estimate Bayes factors and analyse the problem in a fully Bayesian manner, determining which model best explains the data, as well as estimating the parameters for each model.

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 statistical challenge is to make inferences using partially observed, forced, non-linear stochastic differential equations (SDEs). The two major challenges in SDE inference are that the transition density, and therefore the likelihood function, is not available is closed form, and that the state trajectory is a high dimensional unobserved component that must be imputed in any Monte Carlo algorithm. A powerful tool for time-structured problems with intractable likelihoods is the particle filter, and in this paper we employ the SMC^2 approach recently introduced by [10], which is a pseudo-marginal algorithm that embeds a particle filter within a sequential Monte Carlo algorithm to do joint state and parameter estimation. A major advantage of SMC^2 over competitor methods, such as particle MCMC [11], is that it allows for easy estimation of the model evidence. We exploit this to provide estimates of the Bayes factors.

A naive implementation of SMC^2 fails due to extreme particle degeneracy. We show how guided Brownian bridge proposals can be used to maintain particle diversity with evenly distributed weights. When efficiently parallelised on a GPU, this allows inference to be performed in reasonable time (3-4 days for the results in Secion §??). A surprising result is that even though the Monte Carlo error in our Bayes factor estimates is large, the Bayes factors are sufficiently large, even for short time-seroes of data, that we can still distinguish between the competing models.

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 structured 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 3 includes a formulation of the Bayesian approach and brief review of the particle-filter methods that we extensively use. In Section 4 we present a simulation study to assess the performance of the algorithms on synthetic data, and an analysis of the real $\delta^{18}O$ dataset. In Section 6 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

2 Data and models

Our approach to understanding the dynamical behaviour of the palaeoclimate involves four components: data consisting of palaeoclimate records; models of the climate; drivers/forcings of the climate (such as CO_2 emissions, or more pertinently for palaeoclimate, the solar forcing), and a statistical model relating the three previous components.

The aim of the paper is to demonstrate the statistical methodology necessary for combining these components in order to answer the questions we imagine palaeoclimate scientists may wish to ask. That is to say, given some data and a selection of models, we show how to fit these models, and to assess which model is best supported by the data. Scientific aspects of the approach can (and we hope will) be improved upon by using different datasets, richer models, and a more realistic statistical framework. Here, our aim is merely to show that the methodology and computing power is now of sufficiently advanced state that the Bayesian machinry for inference can be applied and that it is fruitful to do so.

Comment by Rich: Michel - I was aiming to describe the holy trinity diagram you used in Uccle (the triangle of approaches) with the aim of unifying some of them using stats. Its kind of an apologia to stress that this is a proof of concept of the stats.

2.1 Data

Comment by Rich: Jake: do we just use sediment? In which case, some of this discussion can have ice removed.

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 correspond to climate 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 results from a combination of complicated physical processes including sedimentation (which occurs 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 from other sources. Also common is to align features to aspects of the astronomical forcing [refs], a process known as astronomical tuning. The result of fitting an age model is a dataset $\{\tau_t, Y_t\}$ in which Y_t denotes the measurement of $\delta^{18}O$ at time τ_t. Investigating age models is beyond the scope of this paper, so we take as a starting point a stack which has been dated by other authors. Estimating τ_t is complex and covered in depth in [**?**]. Here, in common with most other studies, we treat τ_t as a given. We comment on the wisdom of this approach in Sections §.

Comment by Rich: Should we say something about uncertainty on τ_t and that we're ignoring it? Perhaps leave this to the conclusion.

In this article, we use 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]. 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 preferable to use data which has not been orbitally tuned in model selection and, in particular, estimate the influence of astronomical factors on the dynamics of the glacial-interglacial cycle. We focus on the last 780 kyr of this record (the last magnetic reversal occured 780 kya, allowing us to date the starting point accurately), which contains 363 observations, and use it to highlight the issues surrounding double counting of the astronomical forcing.

Figure 1 about here.

2.2 The 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.

Comment by Rich: Jake, can you add a v. brief layman's explanation of each here.

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. The insolation at 65°N on the summer solstice is recovered by setting $\gamma_P = 0.8949$, $\gamma_C = 0$ and $\gamma_E = 0.4346$. This quantity is felt to be particularly influential on the glacial cycle, as it determines the degree of warming of the oceans, which controls the extent of the sea ice in the winter. Pure precession, coprecession or obliquity signals can be recovered by setting the other scaling parameters to 0.

2.3 Phenomenological models of climate dynamics

We cibsuder three models for the climate dynamics. They were each originally proposed as low order ordinary differential equations, with state vector x where the first component $x_{(1)}$ represents global ice volume. The other components represent other quantities such as glaciation state or CO_2 concentration. In order to account for model errors, we convert the models into stochastic differential equation by the addition of a Brownian motion W_t. Each model 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 [14]).

Comment by Rich: Should we add some of Jake's useful text on relaxation osciallators? 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 [14]). 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.

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. Example trajectories are shown in Figure [? - comment].

Model T06: [authors, year]

$$
\begin{aligned}
dX_{(1)} &= \big(\big(p_0 - KX_{(1)}\big)\big(1 - \alpha X_{(2)}\big) - (s + F(\gamma_P, \gamma_C, \gamma_E))\big)\,dt + \sigma_1 dW_{(1)} \\
X_{(2)} &: \quad \text{switches from 0 to 1 when } X_{(1)} \text{ exceeds some threshold } T_u \\
X_{(2)} &: \quad \text{switches from 1 to 0 when } X_{(1)} \text{ decreases below } T_l
\end{aligned}
$$

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 → 1 and 1 → 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

Comment by SPP: needs explanation as this seems totally arbitrary! Could we move the forcing description to the previous section, so that it is presented as some transformation of F from before, to keep the separation of models from forcings?

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

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]

Comment by Rich: Jake: Can you add the references back that have been list in the editing.

2.4 Statistical observation model

The final modelling ingredient is a statistical model relating the unobserved state variables in the dynamical models, X_t, to the dataset. We assume that the data are of the form $\{\tau_t, Y_t\}_{t=1}^{T}$, where τ_t is the estimated age and Y_t the measured proxy of the t^{th} data point/slice. We use the model

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

Here we use $H = (s, 0, \ldots, 0)$, so that Y_t is a scaled and shifted version of the value $X_{(1)t}$, the ice volume in the underlying dynamical model. However, vector observations can be used at no additional cost or complication to the methodology, allowing us to add in observations of other proxies if desired.

This is the measurement process, which relates true $\delta^{18}O$ to measured $\delta^{18}O$. The second part of the statistical specification is the model discrepancy. This describes the relationship between the imperfect dynamical model and true climate. To account for this error, we have added a white noise term to the models as described above. More complex discrepancy models may be sensible [15, 16], and in particular, red noise may be a more appropriate choice.

Comment by Rich: Michel: do we have a reference for this?

Comment by Rich: I've added the τ_t notation as a way to emphasise that the ages are estimates, and that t usually refers to an index, rather than a time. I'm not sure I like it yet though.

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.

Comment by Rich: Jake: Simon has changed $D- > d$, and $h- > s- > H$. I'm not sure I like this. The change will either need reversing or following through. I think Simon's reasoning for changing to d and h, is that its easier to present the method for 1d observations, in which case a lower case letter may make more sense.

3 Methodology

Our primary aim is model selection: given a collection of competing models $\{\mathcal{M}_m\}_{m=1}^M$ of dataset $y_{1:T}$, which is best supported by the data? The Bayesian approach to model selection uses Bayes factors (BF) [17, 18]. The BF for comparing two models, $\mathcal{M}_1$ and $\mathcal{M}_2$ say, is the ratio of their evidences

$$B_{12} = \frac{\pi(Y_{1:T}|\mathcal{M}_1)}{\pi(Y_{1:T}|\mathcal{M}_2)}, \tag{2}$$

where $\pi(Y_{1:T}|\mathcal{M}_m)$ is the evidence for model $\mathcal{M}_m$. The Bayes factor summarises the strength of evidence in the data in support of one model over another and is the ratio of the posterior to the prior odds in favour of $\mathcal{M}_1$ over $\mathcal{M}_2$. If the prior probabilities for each model are equal then the Bayes factor is equivalent to the ratio of the posterior model probabilities.

Secondary aims of our analysis include parameter estimation and filtering, which in this context are often called calibration and climate reconstruction (or hindcasting). Let θ_m denote the parameter for model $\mathcal{M}_m$. Calibration is the process of finding the posterior distribution of the model parameters $\pi(\theta_m|y_{1:T}, \mathcal{M}_m)$, and filtering is finding the distributions of the state variables $\pi(X_{1:T}|Y_{1:T}, \theta_m, \mathcal{M}_m)$. These three problems are of different levels of difficulty. Filtering is the simplest, but by no means simple. It involves finding the posterior distribution of a high dimensional object $X_{1:T}$ and for non-linear or non-Gaussian models, direct calculation of the filtering distributions is impossible, and so we must instead rely upon approximations. Calibration requires that we integrate out the dependence of $X_{1:T}$,

$$\pi(\theta_m|Y_{1:T}, \mathcal{M}_m) = \int \pi(\theta_m, X_{1:T}|Y_{1:T}, \mathcal{M}_m)\mathrm{d}X_{1:T}, \tag{3}$$

and so is usually considerably more difficult than filtering. Finally, model selection requires the integration of the dependence on θ,

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

and is thus more difficult than calibration.

The development of Monte Carlo methodology for solving these three problems for state space models reflects this hierarchy of difficulty. Particle filter methodology, first proposed in the 1990s [19], is able to solve the general filtering problem adequately as long as the dimension of X is not too large. Solving the calibration problem, however, has only begun to be satisfactorily answered more recently, with the development of pseudo-marginal methods such as particle-MCMC [11]. Calculating the model evidence is still very much on open problem.

Comment by Rich: We might just be inviting trouble by saying these this.

Here, we demonstrate how the recently introduced SMC^2 algorithm [10] can be used to estimate model evidences. The technique is not perfect, as it is computationally costly, and results in BF estimates with large variance, but we can think of no other approach for solving this problem, and as we will demonstrate, often the strength of evidence in the data is sufficiently strong so that the estimated BFs are large enough for the estimation variance to be of no impediment to inference.

The approach relies upon the following identities. The evidence can be decomposed as

$$\pi(Y_{1:T}) = \pi(Y_1) \prod \pi(Y_t|Y_{1:t-1}) \tag{5}$$

where we have dropped the dependence on $\mathcal{M}_m$. We can further write

$$\pi(Y_t|Y_{1:t-1}) = \int \pi(Y_t|Y_{1:t-1}, \theta)\pi(\theta|Y_{1:t-1})\mathrm{d}\theta. \tag{6}$$

SMC^2 is then used to find unbiased estimates of $\pi(\theta|Y_{1:t-1})$ and $\pi(Y_t|Y_{1:t-1}, \theta)$. Remarkably, pluggingg these estimates into Equations (5) and (6) then leads to unbiased estimates of the model evidence.

Comment by Rich: UPPER OR LOWER CASE Y?

3.1 Estimating model evidence using SMC^2

Sequential Monte Carlo algorithms [20] are population-based sampling methods aimed at obtaining a sample from some target distribution that is difficult to sample from directly. A series of intermediary distributions $\{\pi_t\}_{t=1}^T$ are chosen that 'close-in' on the target distribution, π_T, from some easily sampled distribution π_1. SMC uses a weighted collection of particles to approximate each distribution, and sequentially updates the weights and the particles. Usually the sequence of distributions is taken to be the sequence formed by adding the data a point at a time. e.g., $\pi(\theta|y_{1:T})$ or $\pi(x_t|y_{1:t})$.

One of the earliest and best known SMC algorithms [19] is designed to sample from the sequence of filtering distributions $\pi_t = \pi(X_t|\theta, Y_{1:t})$. These algorithms are termed particle filters (PFs) and work 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, where superscript n indicates that the operation is performed for $n = 1, ..., N_x$.

Particle Filter

- **PF 1.** Sample state particles $X_1^n \sim q_1(\cdot|\theta, Y_1)$
- **PF 2.** Weight the 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.** Resample the ancestor particle index $\mathcal{A}_{t-1}^n \sim \mathcal{F}(W_{t-1}^n)$
  - **PF 3.2.** Propagate the 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

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

Here, $\mathcal{F}(W_{t-1}^{1:N_x})$ denotes sampling an index from $\{1, ..., N_x\}$ according to weights $\{W_{t-1}^n\}_{n=1}^{N_x}$, and q_t denotes the proposal distribution at time t. Details of the resampling step $\mathcal{F}$ and the proposal distribution q_t are discussed in section §.

An important aspect of the PF is that an unbiased estimate of $\pi(Y_t|Y_{1:t-1}, \theta)$ can be obtained by averaging over the unnormalised weights in each iteration of the algorithm

$$\hat{\pi}(Y_t|Y_{1:t-1}, \theta) = \frac{1}{N_x} \sum_{n=1}^{N_x} w_t^n(X_{1:t}^n). \tag{7}$$

and that an unbiased estimator of the marginal likelihood $\pi(Y_{1:t}|\theta)$ can be obtained by plugging these estimates into the identity [21]

$$\hat{\pi}(Y_{1:T}|\theta) = \hat{\pi}(Y_1) \prod_{t=2}^{T} \hat{\pi}(Y_t|Y_{1:t-1}, \theta). \tag{8}$$

[22] showed that using these unbiased estimates of the likelihood in other Monte Carlo algorithms can lead to valid Monte Carlo algorithms (termed pseudo-marginal algorithms) for performing parameter estimation. For example, PMCMC [11] uses the PF within an MCMC algorithm, and SMC^2 [10] uses a PF embedded within an SMC algorithm, both with the aim of finding $\pi(\theta|Y_{1:T})$.

The SMC^2 algorithm [10] embeds the particle filter within an SMC algorithm targetting the sequence of posteriors

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

for $t = 1, \ldots, T$. This is achieved by sampling N_θ parameter particles, $\{\theta^j\}_{j=1}^{N_\theta}$ from the prior. To each θ^j, we attach a PF of N_x particles, i.e., at iteration t the PF $\{X_{1:t}^{n,j}, W_t^{n,j}\}_{n=1}^{N_x}$ is associated with θ^j, which provides an unbiased estimate of the marginal likelihood $\pi(y_{1:t}|\theta^j)$ via a version of Equation (8). To assimilate the next observation y_{t+1}, we first extend the PF for the X-states to $\{X_{1:t+1}^{n,j}, W_{t+1}^{n,j}\}_{n=1}^{N_x}$, and then estimate $\pi(Y_{1:t+1}|\theta^j)$ etc. Degeneracy occurs when the weighted particle approximation is dominated by just a few particles (i.e. a few have comparatively large weights). It can be monitored by calculating the effective sample size (ESS)[1]. 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 too 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 [11]. The algorithm is given below, but the reader is referred to the original paper for the theoretical justification [10]. Superscript j indicates that the operation is performed for all $j = 1, ..., N_\theta$.

SMC^2

[1] ESS= $\left(\sum (\Omega^m)^2\right)^{-1}$ where Ω^m are the normalised weights.

- **SMC^2 1.** Sample parameter particles $\theta^j \sim \pi(\theta)$
- **SMC^2 2.** Set importance weights $\Omega^j = \frac{1}{N_\theta}$
- **SMC^2 3.** For $t = 1, ..., T$
  - **SMC^2 3.1.** For each parameter particle θ^j, extend the associated PF $\{X^{n,j}_{1:t-1}, W^{n,j}_{t-1}\}_{n=1}^{N_x}$ to $\{X^{n,j}_{1:t}, W^{n,j}_{t}\}_{n=1}^{N_x}$ by performing iteration t of the PF.
  - **SMC^2 3.2.** Calculate $\hat{\pi}(Y_t|Y_{1:t-1}, \theta^j)$ using Equation (7) and calculate the weighted average over the parameter particles to obtain

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

  - **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, resample. For $j = 1, \ldots, N_\theta$:
    - **SMC^2 3.4.1.** Sample an index J from $\{1, \ldots, N_\theta\}$ according to weights Ω^j.
    - **SMC^2 3.4.2.** Sample $\left(\theta, \{X^n_{1:t}, W^n_t\}_{n=1}^{N_x}\right)^* \sim K\left(\cdot| \left(\theta^J, \{X^{n,J}_{1:t}, W^{n,J}_t\}_{n=1}^{N_x}\right)\right)$ where K is a PMCMC kernel that leaves $\pi(\theta, X_{1:t}|Y_{1:t})$ invariant.
    - **SMC^2 3.4.3.** Set $\left(\theta^j, X^{1:n,j}_{1:t}\right) \leftarrow (\theta, X_{1:t})^*$

    Set importance weights $\Omega^j = \frac{1}{N_\theta}$

SMC^2 naturally provides an estimate of the model evidence. The model evidence can be decomposed according to Equation (5), and 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**, is an unbiased estimate of $\pi(Y_t|Y_{t-1})$. An estimate of the model evidence is provided by the product of these terms.

3.2 Guided proposals

A further difficulty arises because for the models of interest, the transition densities $\pi(X_t|X_{t-1}, \theta)$ are not available in closed form, and so we need to choose the particle proposals q_t so that this term cancels. This can achieved by setting $q_t = \pi(X_t|X_{t-1}, \theta)$ for $t > 1$, so that proposals are just simulations from the model. However, this choice will typically lead to particle degeneracy if too many of the proposals end up being far from the observations, so that the small number of proposals in the region of the observation gain nearly all of the weight. Resampling the state particles can improve the approximation in later iterations as only important particles are propagated forward. The choice of $\mathcal{F}$ ...

Comment by Rich: Jake - say something here about the sampling mechanism

Possible resampling strategies are discussed in [23].

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?

Where possible, including information from the next observation y_t in the proposals at time t should lead to more equal weights. For the models considered in this article, it is possible to condition proposals 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.

Comment by Rich: Jake: We need to sort out notation here. The t in X_t is a counting index, rather than a time index, and so $X_{t+\Delta t}$ doesn't make sense. Perhaps we could use $X(\tau)$?

Comment by Rich: I find the rest of this section from here nearly completely impenetrable, even though we've talked through it several times. Could you have another go, and see if its readability can be improved.

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 $\Sigma_y / S^2\Sigma_x(T - t) \to 0$, and $J \to 0$ as $S^2\Sigma_x(T - t)/\Sigma_y \to 0$. This suggests $J = S^2\Sigma_x(T - t)/\,(S^2\Sigma(T - t) + \Sigma_y)$, 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 [24], 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 $\Delta t \lll (T-t)$, and for $S^2\Sigma_x\Delta t \lll \Sigma_y$. Whereas in the case $\Delta t = \Delta T$, and $\Sigma_y \lll S^2\Sigma_x\Delta t$ the proposal variance is approximately the observation variance.

Combining these changes gives a proposal of the form

$$X_{t+\Delta t} \sim \mathcal{N}(X_t + \mu\Delta t + \frac{S^2\Sigma_x\Delta t}{S^2\Sigma_x(T-t) + \Sigma_y} S^{-1}(Y_T - (D + S(X_t + \mu(T-t)))), \left(\frac{S^2\Sigma_x(T-t) - S^2\Sigma_x\Delta t + \Sigma_y}{S^2\Sigma_x(T-t) + \Sigma_y}\right) S^2\Sigma_x\Delta t), \quad (9)$$

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.

3.3 Further details

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 [10], 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.

For complex models, filtering presents significant challenges, and usually approximate techniques such as 4d-Var, and Kalman filter type algorithms are used. Can we substitute?

GPU computation and timings needed.

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)

4 Results

4.1 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 6.

TABLE 6 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 [18]. 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 [25].

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.

4.2 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 [25]. 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.

5 Alternative approaches

Comment by Rich: We can probably ditch this section, or include a little at the end of the methodology section.

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.

6 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 [9]:

> 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.

Comment by Rich: Jake: the references need sorting. Initials are being lost in names, as you need to separate them with a space, e.g., R. D. Wilkinson, not R.D. Wilkinson

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. 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.

[9] 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.

[10] 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: Series B (Statistical Methodology)*, vol. 75, no. 3, pp. 397–426, 2013.

[11] 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.

[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] 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.

[15] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 3, pp. 425–464, 2001.

[16] R. D. Wilkinson, M. Vrettas, D. Cornford, and J. E. Oakley, "Quantifying simulator discrepancy in discrete-time dynamical simulators," *Journal of agricultural, biological, and environmental statistics*, vol. 16, no. 4, pp. 554–570, 2011.

[17] H. Jeffreys, *The theory of probability.* Oxford University Press, 1939.

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

[19] 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.

[20] 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.

[21] P. Del Moral, *Feynman-Kac Formulae.* Springer, 2004.

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

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

[24] 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.

[25] 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 6. WARNING: rounding errors here as I've taken Jake's Z value to two dp and logged it. Values of the log evidence

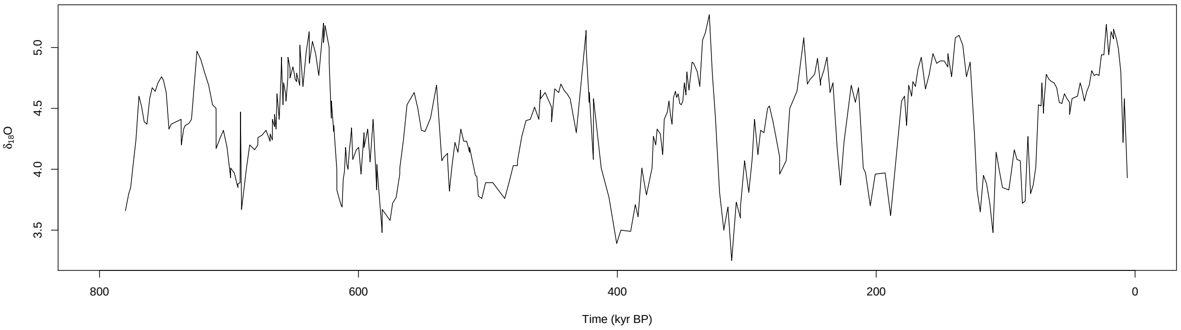

Figure 1: JAKE - Please can you make the axis labels larger? Observed δ^{18}O from ODP677 [12] corresponding to the past 780 kyr. This dataset has been dated without the use of orbital tuning [2].

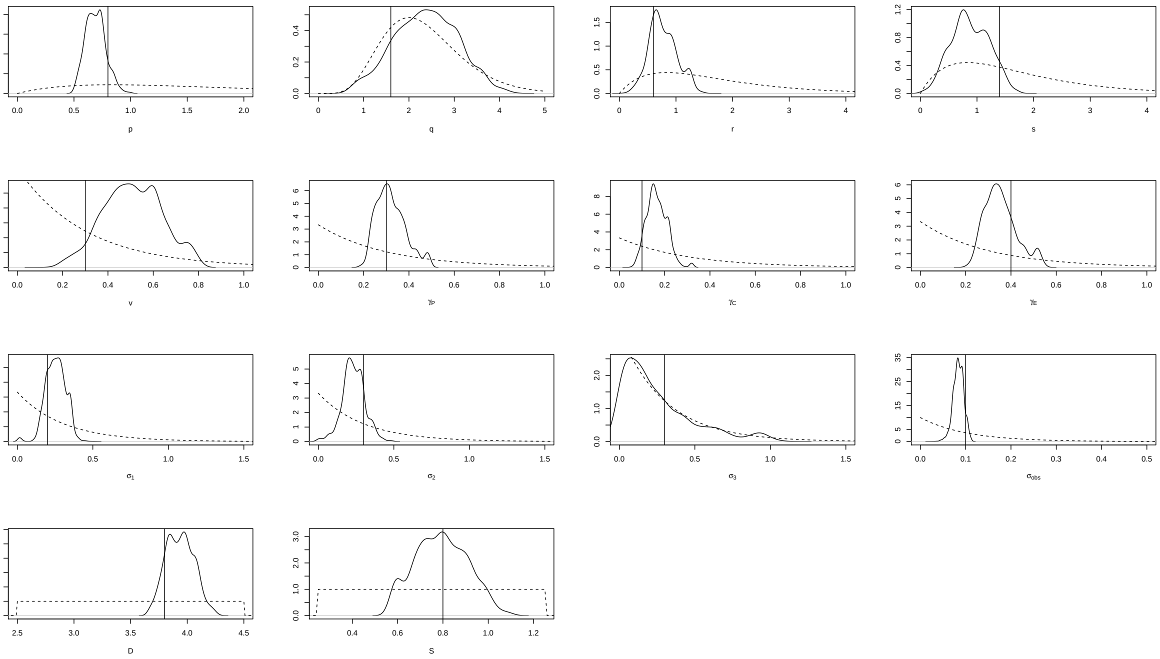

Figure 2: JAKE - Please can you make the axis labels larger? Marginal posterior distributions for the parameters of the forced SM91 model when fit to the SM91-f dataset. Vertical lines show the parameter values used to generate the data, and dashed lines represent the prior distribution.

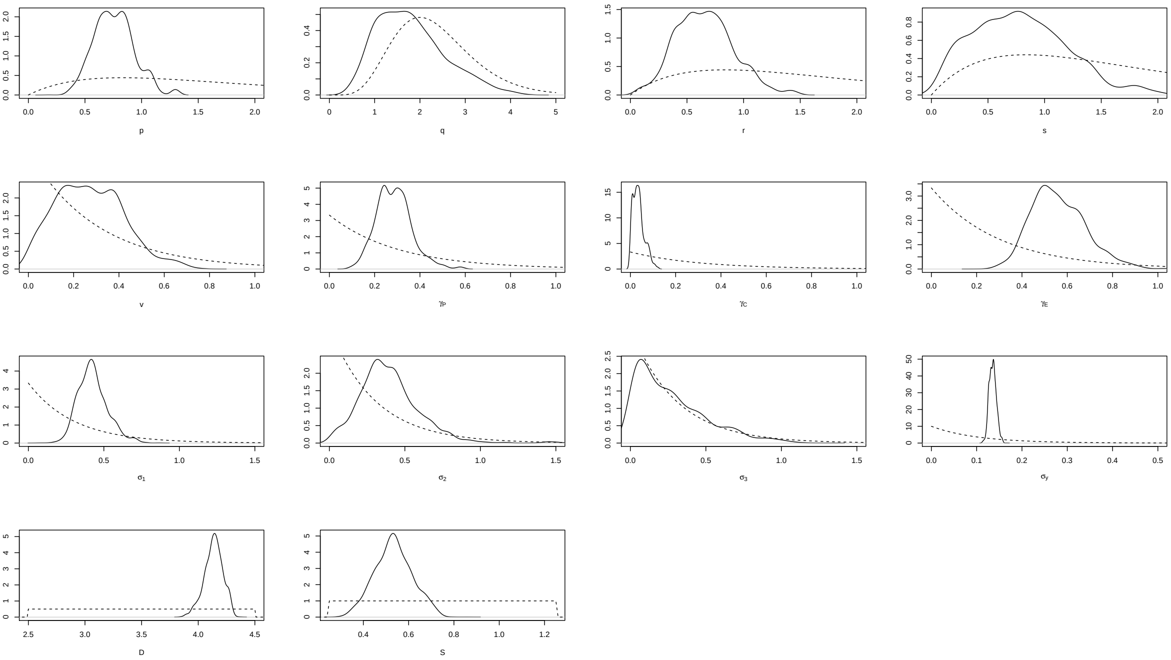

Figure 3: JAKE - Please can you make the axis labels larger? Marginal posterior distributions for the fully forced SM91 model on ODP677-f. Dashed lines represent the prior distributions, and solid lines the posteriors. The prior distributions used are given in Table refTab:Priors

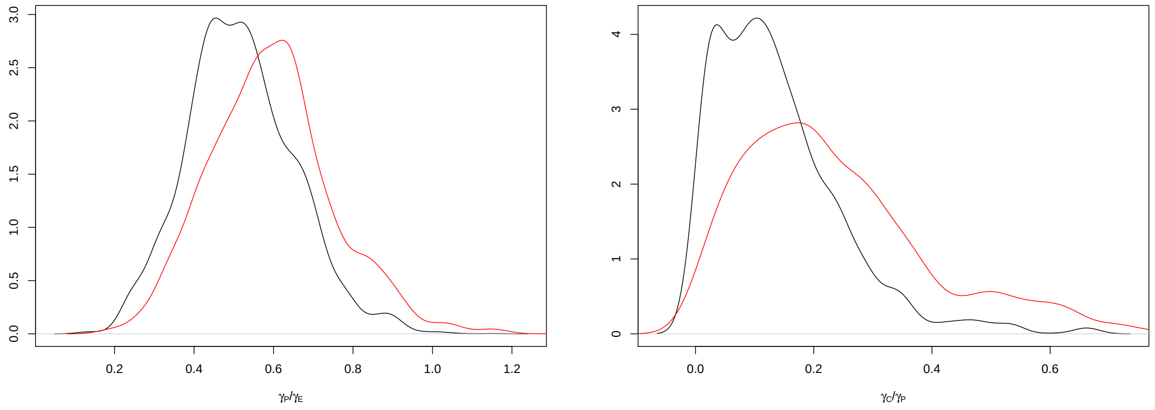

Figure 4: JAKE - Please can you make the axis labels larger? Posterior density plot of the ratio of the orbital scaling terms for the SM91 model (black line), and T06 model (red line).

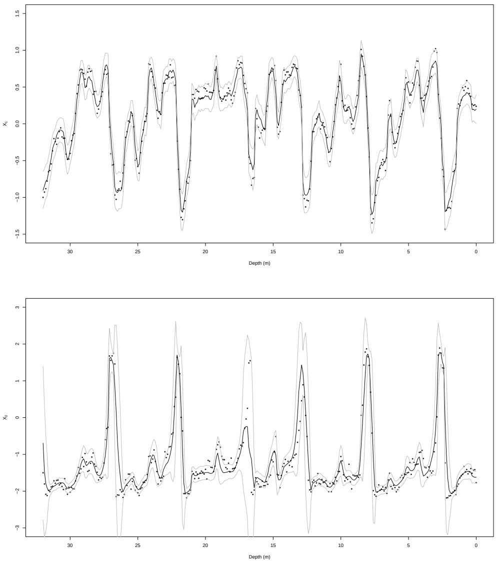

Figure 5: JAKE: These are the wrong figures - could you put in the correct figure. Add the truth on, and the data on x_1.