Bayesian Inference for Palaeoclimate with Time Uncertainty and Stochastic Volatility

Andrew C. Parnell[1,2], James Sweeney[1], Thinh K. Doan[3], Michael Salter-Townshend[4], Judy R.M. Allen[5], Brian Huntley[5], and John Haslett[3]

[1]School of Mathematical and Statistical Sciences, University College Dublin, Ireland

[2]Complex Adaptive Systems Laboratory, University College Dublin, Ireland

[3]School of Computer Science and Statistics, Trinity College Dublin, Ireland

[4]Department of Statistics, University of Oxford, UK

[5]School of Biological and Biomedical Sciences, Durham University, UK

Summary. We propose and fit a Bayesian model to infer palaeoclimate over several thousand years. The data we use arise as ancient pollen counts taken from sediment cores together with radiocarbon dates which provide (uncertain) ages. When combined with a modern pollen/climate data set, we can calibrate ancient pollen into ancient climate. We use a Normal-Inverse Gaussian process prior to model the stochastic volatility of palaeoclimate over time, and present a novel modularised MCMC algorithm to enable fast computation. We illustrate our approach with a case study from Sluggan Moss, Northern Ireland and provide an R package, `Bclim`, for use at other sites.

Keywords: Palaeoclimate Reconstruction, Normal-Inverse Gaussian Process, Modular Bayes, Hierarchical Time Series, Temporal Uncertainty

1. Introduction

In this paper we show how to perform statistical inference on palaeoclimate from pollen proxy data whilst taking account of numerous sources of uncertainty. The data we use arise from sediment cores taken from beneath lakes or bogs where pollen has accumulated over many thousands of years. The changing composition of pollen grains provides information about the climate at that location, whilst radiocarbon dates of the sediment provide information about their age. A further data set of the modern pollen/climate relationship allows for the transformation between our ancient pollen data and our inference target, ancient climate. We provide an outline of our general approach and a case study from Sluggan Moss in Northern Ireland.

After extraction, the sedimentation core will have been sliced into narrow layers, each treated as a near instantaneous snapshot of the vegetation at that

depth. From each slice a palynologist will count many different varieties of pollen and record the counts and depths. We use counts for 28 pollen varieties that have been shown to be sensitive to three carefully chosen aspects of climate (Huntley, 1993). At certain depths material will have been sent for radiocarbon dating, though the choice of depths will depend on the availability of suitable material and budgetary constraints. The number of radiocarbon dates is usually far fewer than the number of depth slices, so some interpolation is required to obtain ages at other depths. Figure 1 shows a sample of the pollen data and radiocarbon dates for our case study site. The data we use are all available online at `www.europeanpollendatabase.net` and `www.neotomadb.org`.

Palaeoclimate inference (more loosely referred to as reconstruction) is a major focus of the Intergovernmental Panel on Climate Change (Jansen et al., 2007). Public interest, however, has largely been fuelled by the 'Hockey Stick' and 'climategate' controversies, e.g. Mann et al. (1998, 1999); McShane and Wyner (2011) where reconstructions were obtained for the aggregated Northern Hemisphere. Climate changes during the past millennium are relatively small and can be inferred with reasonable precision from precisely dated proxies such as tree rings. By contrast, the much older Younger Dryas period (12.8ka to 11.5ka BP) shows a rapid switching from warm to cold to warm. During this period ice core data from Greenland show abrupt warmings of up to 16 °C within decades (Jansen et al., 2007, P435). This type of climate change is not captured well by the General Circulation Models (GCMs) which are used to predict future climate, nor by the precise proxies used to examine the past millennium. Pollen proxy data offer the best hope of resolving such sizeable past climate changes in locations other than Greenland.

Our goal is to create a posterior distribution of climate on a temporal grid given pollen and radiocarbon data at a particular site. This goal is challenging because: the relationship between pollen and climate is non-linear; the pollen data are observed irregularly and with uncertainty in time; and both climate and pollen are multivariate. To formulate such a model, we use a generalised version of the framework of Bayesian hierarchical time series models (Berliner, 1996), comprising an observation layer, a process layer, and a parameter layer. To this we add a calibration layer which allows us to learn about the link between pollen and climate. Given the complexity of the model and the size of the data sets we make a number of simplifying assumptions involving the independence of different layers, which results in a modularised Bayesian algorithm (Liu et al., 2009).

A key modelling choice in palaeoclimate inference is the selection of climate variables to reconstruct; climate often being defined as the 'average' of weather. Many authors (e.g. Mann et al., 1998; Li et al., 2010) have chosen to reconstruct mean annual temperature. This may seem like a logical choice, given that this is relatively easy to measure and relevant to human wellbeing. However, as sum-

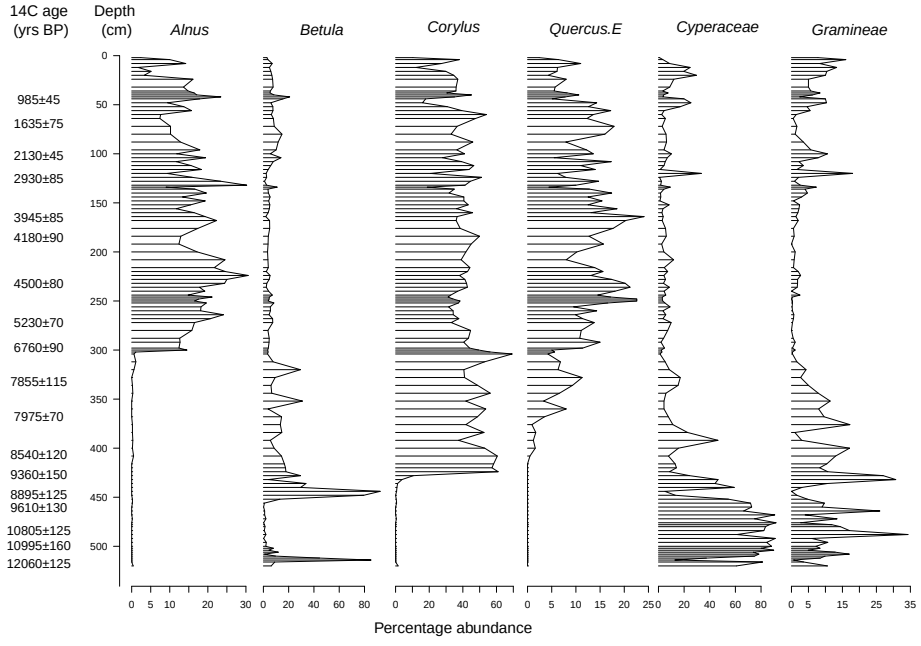

Fig. 1. 6 of the 28 pollen varieties from the Sluggan Moss core which we use as a case study. Depth is shown on the vertical axis (so that 0cm represents the surface though not necessarily the present), whilst radiocarbon ages (with 1-σ uncertainties; see Scott et al., 2010, for more details) are shown where they have been obtained further down the core. These radiocarbon ages (and their associated depths) are used to create a chronology as shown in Figure 4. For each pollen taxa, the percentage abundance is shown at each depth slice in the core (note that the core has not been sliced regularly in depth). Some pollen taxa, e.g. Alder (*Alnus*), like warmer, wetter climates, whereas others, e.g. Sedges (*Cyperaceae*) prefer cooler climates. These pollen data and their associated ages (and age uncertainties), together with the modern analogue data, form the input to our inference routine.

marised by Huntley (2012), the choice should instead depend on the aspects of climate to which the proxy is sensitive. Many biological proxies have shown that they are not sensitive to changes in mean annual temperature so reconstructions of this sort will be characterised by large uncertainties. Instead, more relevant variables were proposed by, amongst others, Huntley (1993):

- the Mean Temperature of the Coldest Month (MTCO) in Celsius, a measure of the harshness of the winter,
- the Growing Degree Days above 5°C (GDD5, calculated as the sum of daily temperatures above 5°C over a year), a measure of the warmth of the growing season,
- the ratio of Actual to Potential Evapotranspiration (AET/PET), a measure of the available moisture (Prentice et al., 1993).

We use these three dimensions as our climate variables throughout this paper. We infer these three variables via a jointly-defined likelihood, as reconstructing variables individually can lead to further difficulties (Juggins, 2013).

To transform our 28 dimensional pollen into 3 dimensional climate, we use a large set of modern pollen and climate data (detailed in Haslett et al., 2006). These data have been collected from around the world; each data point consists of the modern climate and a set of pollen counts taken from the surface sediment at that location. The climates are obtained from weighted averaging of local weather station data over periods of approximately 30 years. We treat these climates as being known precisely, their uncertainties are likely to be orders of magnitude smaller than the uncertainties we obtain for ancient climate. The surface pollen counts are obtained similarly to the counting of fossil pollen. Indeed many of these surface samples are simply the top layer of a core extracted for palaeoclimate reconstruction. The modern data are available as part of the supplementary material to this paper.

Statistical models for modern pollen/climate data sets have been built previously (Haslett et al., 2006; Salter-Townshend and Haslett, 2012; Sweeney, 2012) and involve finding climates that pollen varieties particularly favour, though these too ignore uncertainty in the modern climate data. The models use spatial processes, though it is climate rather than physical space that is the spatial variable. We can use these climate-space processes to 'look up' the climate that is favoured by any particular 28-vector of ancient pollen. This modern data set contributes another statistical model in our framework; we term it the modern analogue data set.

We do not consider the problem of spatio-temporal or multi-proxy inference on palaeoclimate in this paper. Whilst data are available for such a task (e.g. from

the websites given above) and there is some sophisticated statistical modelling in this area (e.g. Li et al., 2010, though this was based on simulated data), we feel that a proper understanding of the processes and models required for palaeoclimate inference at a single site using a single proxy have not yet been fully developed. We hope that this paper provides a way forward for those interested in such extensions.

The paper is arranged as follows. In Section 2 we provide details of our hierarchical time series approach and show how this may be modularised to produce three sub-models. In Section 3 we outline how our reconstruction model can be fitted using a novel MCMC algorithm. In Section 4 we apply our model to our case study site in Northern Ireland. We conclude in Section 5. Our paper contains three technical appendices; the first deals with the MCMC algorithm, the second with model validation, and the third with instructions for the associated R package `Bclim`.

2. Bayesian calibrated hierarchical time series models

We structure our model similarly to that defined by Berliner (1996), though similar concepts are found in state space models (e.g. Cressie and Wikle, 2011) and dynamic linear models (West and Harrison, 1999). We separate the model into four layers: the observation layer, the calibration layer, the process layer, and the parameter layer. Each layer consists of multiple parts: the ancient pollen and the radiocarbon dates form the observation layer, the modern analogue data are the calibration layer, whilst the climate and sedimentation process form the process layer.

We start by outlining our notation:

- y are the observed ancient pollen data from the core, y_i is a 28-vector of pollen for layer i in the core, $i = 1, \ldots, n$.
- x are the observed radiocarbon dates in the core. x_k is the kth radiocarbon date, $k = 1, \ldots, r$. Usually $r \ll n$. Since radiocarbon forms in the upper atmosphere at a variable rate, the radiocarbon age of an object is not the same as its calendar age. The radiocarbon calibration curve (Reimer et al., 2013) provides a method for transferring radiocarbon ages into calendar ages.
- d are the observed depths in the core. d_i is the depth associated with layer i.
- c are ancient climate variables. c_i is the 3-vector of climates associated with layer i. These are our main inference target.

- t are variables representing ages of the ancient pollen data. t_i is the age (given in calendar years before present; BP) of layer i.
- y^m is the observed modern pollen data. y^m_j is a 28-vector of modern pollen for observation $j = 1, \ldots, s$ where j indexes each modern sample site. (These are surface samples, so depth is not relevant for the modern analogue data.)
- c^m is the observed modern climate data. c^m_j is a 3-vector of modern climates for observation j.
- θ are a set of parameters governing the relationship between pollen and climate.
- ψ are a set of parameters governing the sedimentation process (i.e. linking age and depth).
- v are a set of parameters governing the climate process. As these parameters deal with the dynamics of climate change they are also of key interest. We use v rather than a Greek character because we use these to measure stochastic volatility. We set v_i to be a 3-vector of volatility parameters associated with time increment (t_i, t_{i+1}).

From the above we create a posterior distribution of our parameters given data:

$$\begin{aligned} p(c, t, \theta, \psi, v | y, x, d, y^m, c^m) \quad \propto \quad & \underbrace{p(y|c,\theta)\; p(x|t,c,\psi)}_{\text{observation layer}} \times \underbrace{p(y^m|c^m,c,\theta)}_{\text{calibration layer}} \\ & \times \underbrace{p(c|t,v)\; p(t|\psi,d)}_{\text{process layer}} \times \underbrace{p(\theta)\; p(\psi)\; p(v)}_{\text{parameter layer}} \end{aligned}$$

Before proceeding further we make some simplifying assumptions. We assume that pollen at layer i is conditionally independent of other layers given climate and the parameters θ (for both ancient and modern data). Furthermore, we assume that radiocarbon dates are conditionally independent given the calendar age at that layer. These assumptions are well-established in the literature (e.g. Haslett et al., 2006; Haslett and Parnell, 2008; Tingley et al., 2012) and have rarely been challenged. Strictly speaking the conditional independence assumption for pollen layers only holds when all possible climate variables are observed and when pollen counts react instantly to changing climate. In reality we can only infer a finite number of climate variables, and plants will react at differing speeds to changing climates. In this paper we move from two to three climate dimensions (compared with Haslett et al., 2006) by including a moisture component of climate (AET/PET) as well as the two temperature components. However, the problem of differing rates of plant response to changing climate has only been dealt with through far more advanced deterministic models (see,

e.g. Garreta et al., 2009).

We make two further assumptions that require slightly more discussion. First, we assume that the modern analogue data set dominates the ancient data set to the extent that we can write $p(y^m|c^m, c, \theta) \approx p(y^m|c^m, \theta)$. Such an assumption was presented as uncontroversial in Haslett et al. (2006) though may present problems where ancient climate lies beyond the range of the modern analogue data. However, no models which account for ancient climate have been proposed in the literature. Second, we make the assumption that climate plays no role in the sedimentation process, so that $p(x|t, c, \psi) \approx p(x|t, \psi)$. A simple argument against this assumption may be that warmer climates yield more pollen and so faster sedimentation rates. However, the relationship is likely to be far more complex, and again no such models have yet been created which relate stochastic sedimentation to climate.

Following these assumptions, we obtain:

$$\begin{aligned} p(c, t, \theta, \psi, v|y, x, d, y^m, c^m) \quad \propto \quad & \prod_{i=1}^{n} p(y_i|c_i, \theta) \times \prod_{j=1}^{s} p(y_j^m|c^m, \theta) \times \prod_{k=1}^{r} p(x_k|t_k, \psi) \\ & \times p(c|t, v) \times p(t|\psi, d) \times p(\theta) \times p(\psi) \times p(v) \end{aligned}$$

As presently stated, the model requires a posterior of dimension $dim(c)+dim(t)+dim(\theta)+dim(\psi)+dim(v)$ where c is of dimension $n \times 3$ and t of dimension n. The parameters θ, ψ and v all turn out to be similarly high dimensional (see later sections for elaboration), so we make two final assumptions that are purely to reduce the complexity of the model fitting, and result in breaking the overall model into three separate modules. Such approximations have been previously suggested by Liu et al. (2009), although they occur naturally in many settings where data are pre-processed before analysis. First, we remove the influence of the fossil pollen y on the parameters θ. Second we remove the influence of t on the climate process for c. The effect of this modularisation can be seen most clearly by looking at the complete conditional distributions for these parameters:

$$\begin{aligned} p(\theta|...) \quad & \propto \quad \underline{\prod_{i=1}^{n} p(y_i|c_i, \theta)} \prod_{j=1}^{s} p(y_j^m|c_j^m, \theta)\, p(\theta) \\ p(t|...) \quad & \propto \quad \prod_{k=1}^{r} p(x_k|t_k, \psi)\, p(t|\psi, d)\, \underline{p(c|t, \psi)} \end{aligned}$$

In each case we remove the underlined terms from the updates, thereby creating three separate models that no longer need to be fitted simultaneously. The

parameters θ are now learnt solely from the modern data, so that:

$$p(\theta|y^m, c^m) \quad \propto \quad \prod_{j=1}^{s} p(y_j^m|c_j^m, \theta)\ p(\theta) \tag{1}$$

The ages t and parameters ψ are learnt solely from the radiocarbon dates and depths:

$$p(t, \psi|x, d) \propto \prod_{k=1}^{r} p(x_k|t_k, \psi) \prod_{i=1}^{n} p(t_i|\psi, d)\ p(\psi) \tag{2}$$

Finally, the posterior distribution for ancient climate involves the posterior distributions above, the ancient pollen data and the climate process:

$$\begin{aligned} p(c, t, \theta, \psi, v|y, x, d, y^m, c^m) \quad \propto \quad & \prod_{i=1}^{n} p(y_i|c_i, \theta)\ p(c|t, v)\ p(t, \psi|x, d) \\ & \times p(\theta|y^m, c^m)\ p(v) \end{aligned} \tag{3}$$

We call Equations 1, 2 and 3 the modern analogue module, the chronology module, and the reconstruction module respectively. This modularisation (the act of removing the underlined terms from the updates) is a conservative assumption as it reduces the precision in the parameters because we are removing terms that are multiplicative to the complete conditional. The three modules can be seen most clearly in a Directed Acyclic Graph (DAG), Figure 2. In the subsequent sections we discuss the modelling choices for each of the modules in more detail.

2.1. Modern analogue module

The modern analogue data set we use contains 7742 modern surface samples of 28-vectors y^m of modern pollen and 3-vectors c^m of modern climate. In this module we aim to estimate parameters θ governing the relationship between pollen and climate using the model specified in Equation 1. The pdf $p(y_j^m|c_j^m, \theta)$ used in the likelihood here is sometimes known as a forward model as it provides a data generating mechanism from which pollen can be simulated given climate. Numerous methods for creating a forward model between climate and pollen proxy data have been suggested; see Ohlwein and Wahl (2012) for a full review.

A first attempt at a Bayesian modern analogue forward model was given by Haslett et al. (2006) where the likelihood distribution was Dirichlet-Multinomial to explicitly model over-dispersion in the modern analogue data:

$$y_{j,1}^m, \ldots, y_{j,28}^m|c_j^m \sim DirMult\left(K_j, \{\theta_1(c_j^m), \ldots, \theta_{28}(c_j^m)\}\right)$$

where K_j is the total number of pollen grains for sample j (a palynologist will often stop after counting 400 grains), and $\{\theta_1(c_j^m), \ldots, \theta_{28}(c_j^m)\}$ are a set of

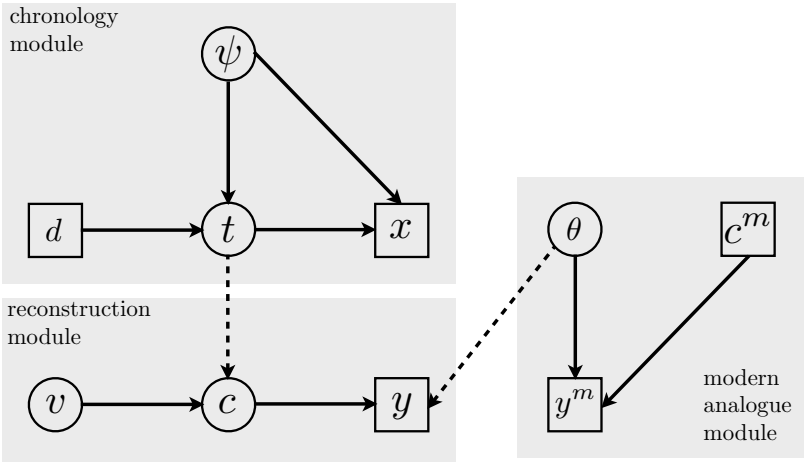

Fig. 2. A Directed Acyclic Graph (DAG) of our palaeoclimate model with different modules indicated in grey boxes. The notation is provided in Section 2. Circles indicate parameters/latent random variables whilst boxes indicate data. The solid lines indicate the direction of information flow, whilst the dashed lines indicate relationships where modularisation occurs.

parameters governing the likelihood of that particular variety of pollen being present in the sample, given the climate that is associated with it. The parameters θ_j were given a Gaussian Markov Random Field (GMRF; Rue et al., 2009) prior with two-dimensional climate as the spatial variable. The net effect is to produce 28 surfaces (known as response surfaces) which govern how that particular pollen variety responds to climate.

The Haslett et al. (2006) model was extended by Salter-Townshend and Haslett (2012) to account explicitly for zero inflation (rather than just over-dispersion) and to allow for a richer covariance structure amongst the multinomial proportions. This new covariance structure uses a nested multinomial distribution where the nesting structure was created from an expertly elicited highly informative prior distribution. The model retains the GMRFs (in two climate dimensions) and so the Integrated Nested Laplace Approximation (INLA; Rue et al., 2009) can be used to bypass Monte Carlo fitting techniques and provide extremely fast posterior inference. We use the Salter-Townshend and Haslett (2012) model in this paper to learn about the modern analogue data, though we extend their

model slightly to account for our three climate dimensions rather than the two originally used.

In Figure 3, we show a schematic plot of how a pollen variety may respond to different climates. The modern data points (denoted as Y here) provide the means to fit a non-parametric curve with climate as the explanatory variable and pollen count as the response. When presented with ancient pollen (in a later module) we can invert these surfaces to get an estimate of the pdf of ancient climate. The graph shows two example ancient pollen counts, denoted 1 and 2. The second of these leads to a naturally multi-modal climate pdf.

2.2. Chronology module

The chronology module is concerned with estimating the ages t of the ancient pollen in the core. These ages will necessarily be uncertain, since the radiocarbon dates are observed with uncertainty, and the interpolation required to infer ages at all depths will add further uncertainty. A useful constraint is that age must increase with depth (older sediments lie deeper in the core) so a monotonic stochastic process is used. A number of statistical age-depth models have been proposed (e.g. Bronk Ramsey, 2008; Haslett and Parnell, 2008; Blaauw and Christen, 2011); see Parnell et al. (2011) for a review. We use the Bchron model of Haslett and Parnell (2008) since it has been specifically developed for inclusion in palaeoclimate reconstruction and allows full access to all posterior quantities.

Expanding on Equation 2, we treat the radiocarbon dates x as normally distributed (a common assumption in radiocarbon dating) around a known function of the calendar ages t (via the radiocarbon calibration curve; Reimer et al., 2013). More importantly, the sedimentation process $p(t_i|\psi, d)$ is governed by a compound Poisson-Gamma process on the time increments:

$$t_i - t_{i-1}|\psi, d = \sum_{i=1}^{N(d_i - d_{i-1})+1} g_i(\psi)$$

where $N(d_i - d_{i-1})$ follows a Poisson distribution with a rate that depends on the depth increment, and g_i is a gamma distributed random variable parameterised by ψ. Further complications exist in the form of outlying radiocarbon determinations which may break the monotonic structure of the process. However, we do not discuss these further here; see Christen and Perez (2009) for a discussion of outliers in radiocarbon dating. Figure 4 shows the estimated ages (with uncertainties) for our case study site.

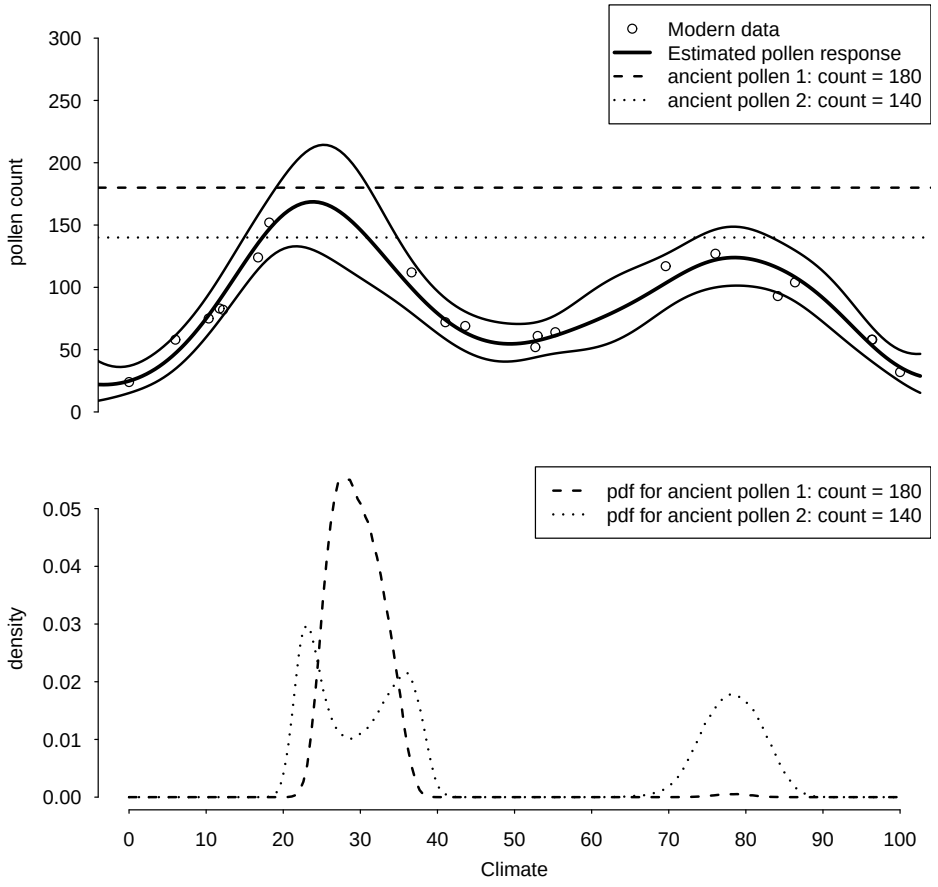

Fig. 3. Schematic of the modern pollen/climate model for the modern analogue data. The upper panel shows example modern pollen and climate data for a single pollen count and climate dimension. This pollen variety seems to prefer values of the climate variable to be around 30, for which we would expect around 180 grains to be counted in a sample layer. When ancient pollen 1 (with a count of around 180) is introduced we obtain a climate pdf (lower panel) strongly focussed around climate 30. When a lower count of ancient pollen is found (at around 140) we obtain a bi-modal climate pdf focussed away from climate value 30, with a further possible mode at climate value 80.

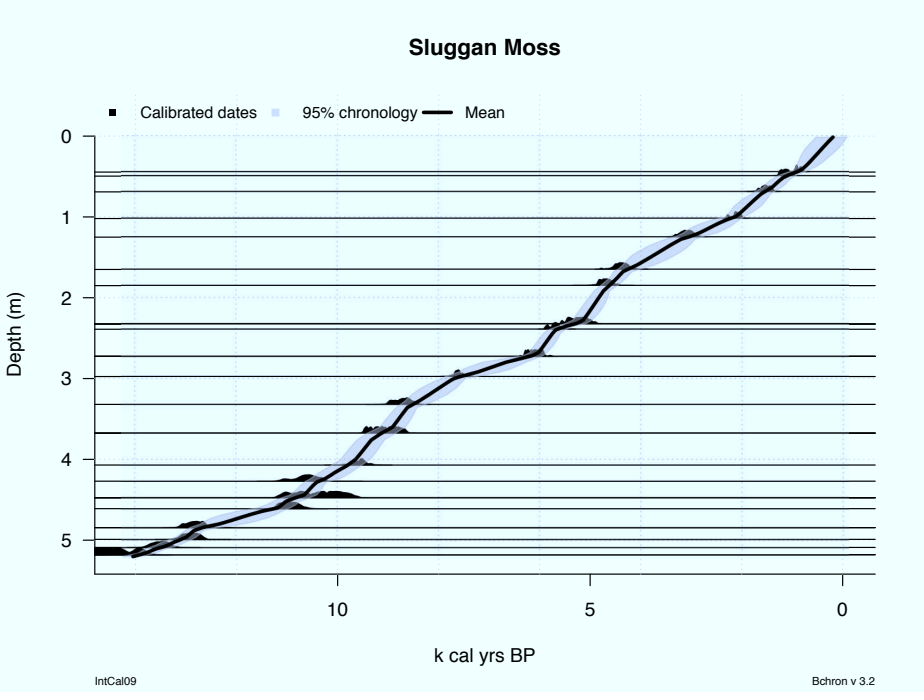

Fig. 4. The output from a chronology model run on our case study site at Sluggan Moss. Each of the horizontal lines represents a radiocarbon date taken from the fossil pollen core. Its associated pdf is shown in black. The chronology model run provides age estimates at the dates at which pollen is counted in the core, represented here by the shaded 95% point-wise credible intervals. The chronology model allows us to work on a calendar timescale, albeit with uncertainty.

2.3. Reconstruction module

Our final module creates our quantity of interest; the posterior distribution of three-dimensional climate. Returning to Equation 3, we need to define the quantities $p(y_i|c_i, \theta)$, $p(c_i|t_i, v)$ and $p(v)$. For the first of these we use the same nested multinomial likelihood as in the modern analogue module. The modularisation allows us to marginalise over θ to give $p(y_i|c_i) = \int p(y_i|c_i, \theta)\ p(\theta|y^m, c^m)\ d\theta$ since $p(\theta|y^m, c^m)$ uses exclusively the modern analogue data. We use this expression to our advantage in the model fitting section below. An alternative would be to fix θ at its posterior mode though we have found that this tends to underestimate the uncertainty in the resulting ancient climate estimates.

The key modelling choice for this module concerns the climate process $c|t, v$. We require a time series process in continuous time that can appropriately capture climate dynamics like those seen in the Younger Dryas period. Previous work in this area has included traditional ARIMA time series models (Tingley and Huybers, 2010), models with covariates (Li et al., 2010) and models based on Brownian motion (Haslett et al., 2006). For our purposes we use a simple stochastic volatility model based on the Normal-Inverse Gaussian distribution which allows us to focus on both climate and climate volatility. We write (for an individual climate dimension):

$$c_i - c_{i-1}|t_i - t_{i-1} = h \sim N(0, v(h)), \quad v(h) \sim IG(\phi_1 h, \phi_2 h)$$

where IG is an Inverse Gaussian distribution (Betrò and Rotondi, 1991) and ϕ_1 and ϕ_2 are given informative prior distributions (see next paragraph). When marginalised over the squared volatilities v, we obtain a Normal-Inverse Gaussian (NIG) process on c (Barndorff-Nielsen, 1997). This is long-tailed, has explicit pdf and is closed under addition. Bayesian inference for the NIG process has been discussed by Karlis and Lillestol (2004). The NIG process is extremely simple to work with, and provides many of the features we might expect to appear in a dynamic and volatile system such as climate.

There are various choices as to how we use the NIG process in our final model with respect to the multivariate nature of climate. The simplest version is perhaps to use the NIG on each dimension independently with a single volatility process shared between the climate dimensions. Since it is likely that changes in temperature and moisture occur at different rates we rejected this model in favour of something more flexible and so allow for independent volatilities in each climate dimension. We do, however, specify common prior distributions for the hyper-parameters ϕ_1, ϕ_2 (discussed below). An even richer model would involve a multivariate NIG model (Barndorff-Nielsen, 1997) which would explicitly model correlation between the climate dimensions. However such correlation is partly induced through the likelihood and the multivariate version would require substantially more coding. We leave a more detailed study of the choice of prior

distribution for the climate process to another paper.

We obtain informative prior distributions for the IG parameters ϕ_1 and ϕ_2 by fitting the NIG process to the last 14k years of a similarly irregularly-spaced ice core data set from Greenland (Stuiver, 2000). The data here are precise measurements of δ^{18}O, a chemical proxy that approximately measures the temperature of rainfall. Once fitted, the posterior distributions are well approximated by log-normal distributions, so that we obtain $\phi_1 \sim LN(1.28, 0.08)$ and $\phi_2 \sim LN(4.23, 0.27)$. Henceforth, our focus is on a posterior distribution for ancient climate c, ancient volatilities $\sqrt{v}$ and hyper-parameters ϕ_1, ϕ_2.

3. Fitting the reconstruction module

Having covered the modelling choices for each of our modules, we now outline a novel Markov chain Monte Carlo (MCMC) fitting algorithm for the reconstruction module. The fitting algorithm for both the modern analogue and chronology modules have been discussed elsewhere (Salter-Townshend, 2009; Haslett and Parnell, 2008) so we do not cover them. However, our algorithm makes use of the fact that we can simulate parameters θ from the posterior distribution of the modern analogue module, and simulate ages t from the posterior distribution of the chronology module.

Of course MCMC is not the only means by which such models can be fitted. For similar models there are numerous algorithms based on Sequential Monte Carlo (see, e.g. Carvalho et al., 2010, for a review). These proceed (in our notation) in a 'forward' or filtering stage by simulating from $p(c_1)$ and then forming $p(c_i|y_1, \ldots, y_i)$ sequentially for $i = 2, \ldots, n$. Filtering densities are created of the form $p(c_i|y_1, \ldots, y_i)$, though their creation requires both the use of the observation and process layer for every time point. Below, we show that in our situation it is possible to produce a valid joint posterior without such restrictions, i.e. which does not require full calculation of the likelihood or process distributions at the forward stage.

We introduce our algorithm by first remarking that it is feasible to calculate, for a single layer of ancient pollen, a posterior distribution of ancient climate for that layer only, written as $p(c_i|y_i) \propto p(y_i|c_i)p(c_i)$, where $p(y_i|c_i)$ is calculated using the integral $p(y_i|c_i) = \int p(y|c_i, \theta)\ p(\theta|y_i^m, c_i^m)\ d\theta$ and $p(c_i)$ is flat (note that our prior on c_i in Section 2.3 is intrinsic, i.e. we model the changes in c but make no a-priori statement about the marginal values of c). This likelihood is slow to calculate as the integration may involve a high dimensional grid, but can be done in parallel for multiple layers simultaneously. The resulting sets of $p(c_i|y_i)$ for $i = 1, \ldots, n$ we term *marginal data posteriors* (MDPs), as they contain the posterior information on the ancient climate given pollen at only

that layer. Clearly, they are strongly related to the filtering densities outlined above. In fact they are far easier to store, being of dimension $n \times 3$ rather than $n \times 28$. Once obtained, we need no further calls to the expensive likelihood terms $p(y_i|c_i)$. The use of MDPs is somewhat equivalent to being given climate as 'data' (with provided uncertainties); a common occurrence in many fields where data are adjusted prior to analysis. Here that adjustment is done explicitly with full respect to the uncertainty.

The benefit of using MDPs is not immediately obvious so it is helpful to consider a simplified version where $y_i|c_i \sim N(c_i, 1)$; the MDP is trivially $c_i|y_i \sim N(y_i, 1)$. The complete conditional distribution of the parameters of interest c and v is:

$$p(c, v|\ldots) \propto \prod_{i=1}^{n} p(y_i|c_i) \prod_{i=2}^{n} p(c_i|c_{i-1}, v(t_{i-1}, t_i)) \prod_{i=2}^{n} p(v(t_{i-1}, t_i)|\phi_1, \phi_2).$$

All the terms involving c are now Gaussian and so c can be analytically integrated out of the model. The same holds were the MDP used in place of $p(y_i|c_i)$ without any recourse to approximation. This means that we can focus inference on v and create c at a second stage from $c|y, v \sim N(Vy, V)$ where $V = (I+W)^{-1}$. Here W is a singular tridiagonal matrix containing the volatilities which can be written as $W = \sum_i v_i^{-1} B_i B_i^T$ where B_i is the ith row of difference matrix B, an $(n-1) \times n$ differencing matrix with the first row structured as $(-1, 1, 0, \ldots, 0)$, and subsequent rows structured similarly.

In our situation the marginal data posteriors are not Gaussian, though it is relatively simple to approximate them accurately using Gaussian mixtures so that:

$$p(c_i|y_i) \approx \sum_{g=1}^{G} \pi_{ig}(y_i) N(\mu_{ig}(y_i), \Sigma_{ig}(y_i))$$

where the mixture component probabilities π_{ig}, the component means μ_{ig} and covariance matrices Σ_{ig} are explicitly written here to show their dependence on the observed data y_i. The size of the approximation can be arbitrarily reduced by simply increasing the number of mixture components G. By conditioning on a mixture component (with probability proportional to π_{ig}), the factorisation in the previous paragraph occurs again and we have the same shortcut to creating a posterior distribution with minimal approximation error. Gaussian mixtures can be created using the MClust package of Fraley and Raftery (2002). For our algorithm we find it sufficient to set $G = 10$ and force all covariance matrices Σ to be diagonal.

The above allows us to create posterior samples of climate and climate volatility from ancient pollen data. Our final step is to interpolate on to a regular grid so that we can obtain, e.g. climate estimates at a centurial level. We can interpolate

the squared volatilities v using the Inverse Gaussian bridge of Ribeiro (2003). Climates can then be created from a standard Brownian bridge conditional on the volatilities. More technical detail on the fitting algorithm is given in Appendix A. We consider the robustness of our model to mis-specification in Appendix B.

4. Case study: Sluggan Moss, County Antrim, Northern Ireland

We now apply our three modules to a site from Northern Ireland, previously published in Smith and Goddard (1991). Our goal is to create a posterior distribution of the three climate dimensions GDD5 (warmth of growing season), MTCO (harshness of winter) and AET/PET (availability of moisture) and their associated volatilities over the previous 14,000 years. A plot of a subset of the data at this site (pollen, depths and radiocarbon dates) is given in Figure 1.

The modern analogue module relies only on the modern data so the creation of the posterior $p(\theta|y^m, c^m)$ is a once-only exercise. This posterior distribution can be re-used for other ancient pollen cores. For further computing details we refer the reader back to Salter-Townshend and Haslett (2012). The chronology module is also independent of the fossil pollen and can also be created at an offline stage, though it is only relevant to one particular core. As stated earlier we use the Bchron (Haslett and Parnell, 2008) R package to create an age-depth model and thus posterior distributions of the age of each ancient pollen layer. The output from the chronology module is shown in Figure 4.

For the reconstruction module, the creation of MDPs and their approximation as mixtures is a relatively fast step taking less than 5 minutes on a modern PC with several CPUs, though the former is strongly dependent on the number of layers n. For our core we have $n = 115$ layers which is fairly typical, though other cores may have many more. The MCMC stage to create posterior volatilities was run for 100,000 iterations with a burn-in period of 20,000 and thinning by 40. The resulting 2,000 iterations were checked for convergence using the R package boa (Smith, 2005). Posterior creation of climates, and their subsequent interpolation, are of negligible computational impact. A full run of the modern analogue, chronology and reconstruction modules for this core took less than 10 minutes on an Intel Core-i7 2.6GHz processor with 8 CPUs and 16Gb of RAM.

Figure 5 shows GDD5, MTCO and AET/PET posterior distributions for Sluggan interpolated via bridging on to a regular centennial grid from 0.2 to 13.8 ka years BP (approximately the age range of the Sluggan Moss core). We show point-wise summaries of the climate sample paths, though other summaries (e.g. first differences) are available just as simply. A Younger Dryas type event is clearly visibly in MTCO, and there appear to be contemporary changes in both

GDD5 and AET/PET. Unsurprisingly, the last 10k years BP are reasonably constant, much like comparable ice core data (Stuiver, 2000). Figure 6 shows the posterior distributions of interpolated volatilities derived via the Inverse Gaussian bridge. Given the extra uncertainty in the volatility, there is less signal here and we only show the plot for MTCO. There is some evidence of increase in volatility at around the Younger Dryas period; all of the highest mean volatilities occur before 10k years BP. Further precision could be attained by reducing chronological uncertainties or incorporating multiple sites in a spatial model.

Figure 7 shows the prior and posterior distributions for the Inverse Gaussian parameters ϕ_1 and ϕ_2. These control the mean and the variance of the volatility process such that the mean of the squared volatility per unit time is ϕ_1 with variance ϕ_1^2/ϕ_2. The model indicates that the prior and posterior of ϕ_1 are broadly comparable, but the posterior of ϕ_2 is shifted lower than the prior, corresponding to an increase in volatility variance. This may be the result of a climatological phenomenon, for example cores in Europe may exhibit more variability in volatility, or may be a result of the increased data uncertainty when compared to that of an ice core.

These results from Sluggan Moss have a number of implications from a palaeoclimate perspective. It appears from Figure 5 that the uncertainty in the reconstructed palaeoclimate values is greater for all three variables before ca. 10.5 ka BP; mean volatility is also generally higher during this period (Figure 6). The very high uncertainty is principally a consequence of the 'multiple analogues' problem discussed by Haslett et al. (2006); this likely also underlies at least in part the generally high volatility seen during this interval. The first implication is thus that pollen data alone provide an inadequate basis for a reliable palaeoclimate reconstruction at this site for the period before ca. 10.5 ka BP, although this limitation could probably be overcome if the choice of analogues could be constrained using data from other proxies (Huntley, 1993, 1994).

After ca. 10.5 ka BP, throughout most of the Holocene, the palaeoclimate has been relatively stable, with some reduced uncertainty in the reconstructed values for all three variables. Of the three, MTCO (harshness of the winter) has shown least change and also has relatively limited uncertainty, the 50% range of the joint posteriors generally being ca. 3°C and even the 95% range rarely exceeding 8°C. Whilst such uncertainties exceed those typically quoted by transfer function studies (Brooks and Birks, 2001), it is important to have more complete and realistic estimates of uncertainty if robust comparisons are to be made between such reconstructed values and either those reconstructed from other proxies or those derived from climate models. The median GDD5 (growing season warmth) is higher before ca. 6 ka BP, falling thereafter, albeit that the change is of much smaller magnitude than the uncertainty. Whilst the fall in the median value is consistent with a wide range of other evidence that indicates that the early

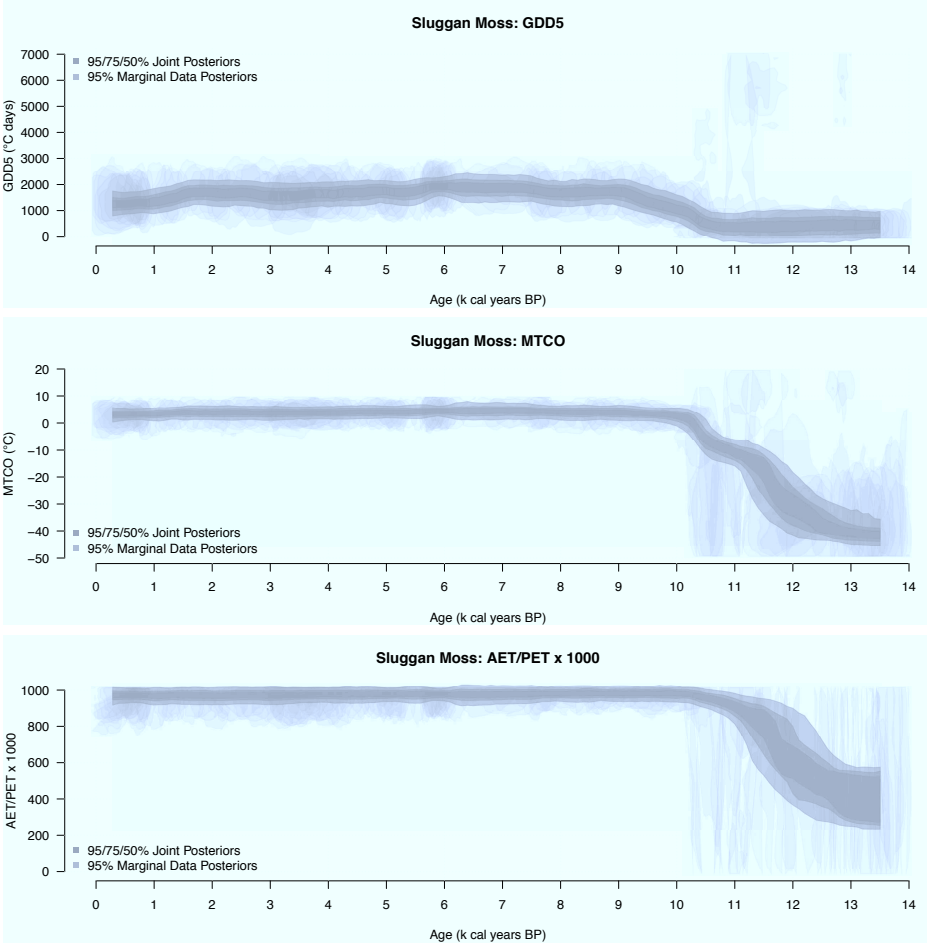

Fig. 5. A plot of the centennial interpolated GDD5 (growing season warmth), MTCO (harshness of winter) and AET/PET (available moisture; scaled up to (0,1000)) over the period 0 to 14ka BP. The blue 'blobs' represent the marginal data posteriors whereas the red bands represent summarised posterior stochastic interpolations of climates c from our interpolated stochastic volatility model.

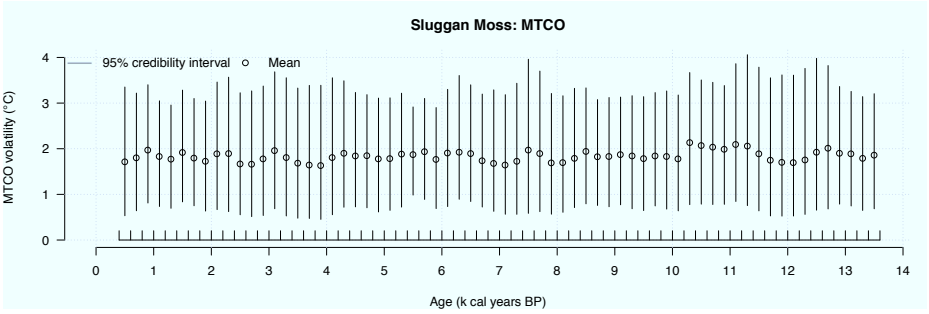

Fig. 6. A plot of the posterior stochastically interpolated volatilities (in 200-year time windows) for the Mean Temperature of the Coldest Month for Sluggan Moss. The vertical lines represent 95% credibility intervals for the centennial volatility, whilst the circles represent the mean.

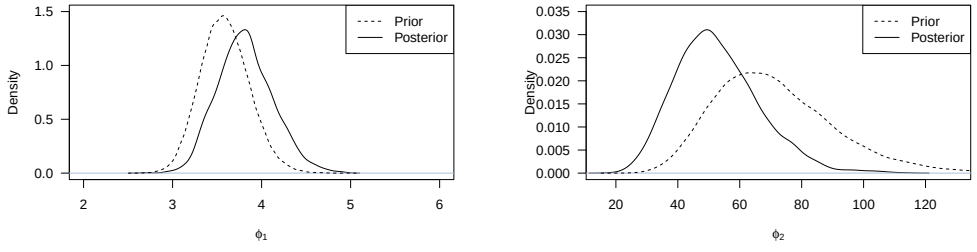

Fig. 7. Plot of prior and posterior distributions of the volatility parameters ϕ_1 and ϕ_2.

millennia of the Holocene were generally warmer in summer in Europe (e.g. treelines extending to higher latitudes and altitudes, as summarised by Grove, 2004), the range of the uncertainty in the reconstructed values would urge caution if considering this record in isolation.

In the case of AET/PET (moisture availability) the median value is generally lower after ca. 6 ka BP, but also has greater uncertainty. The higher uncertainties after ca. 6 ka BP urge caution in interpreting the reconstructed values; indeed, it is likely that the increased abundance of Gramineae (grasses) is principally driving the reconstructed changes, whereas this increase mostly reflects human forest clearance as agriculture developed. Furthermore, there is a discrepancy in the change in the median reconstructed value when compared to other lines of evidence (e.g. the widespread development of blanket bogs in the uplands of the British Isles after the mid-Holocene, see Birks, 1988) that indicate generally greater moisture availability in north-west Europe during the second half of the Holocene.

5. Discussion

The model we have presented performs inference on palaeoclimate whilst quantifying uncertainties in a more detailed and thorough fashion than previously possible. The foundation of the model is a Bayesian hierarchical time series which explicitly separates out the dynamical systems (climate; sedimentation) from the observation model (the link between climate and proxy pollen data; the formation of radiocarbon dates). This idea, proposed originally in Haslett et al. (2006), had also been suggested by Tingley et al. (2012). We have implemented and considerably expanded this approach and developed a modular algorithm which can perform inference on both climate and climate volatility through the use of mixtures of marginal data posteriors. The Normal-Inverse Gaussian process we apply to imitate the dynamic nature of changing climate allows us to focus inference on questions that could not previously be answered using existing modelling approaches.

The modularity invoked by following our modelling assumptions seems appropriate for use in future extensions. This modularity enables various steps to be run in parallel, and also allows us to change modules as required. For example, to produce interpolations using a different chronology model as that of Figure 4, the creation of MDPs and their approximation as mixtures does not need to be re-run, being entirely independent of any chronological uncertainty. Thus only the other steps were required. This has larger implications for future modelling as, for example, a new forward model can be used in place of the one we use with no other changes required to any of the other steps. The same applies for the chronology model, the Mclust mixture algorithm, and the climate process itself

(though it would still need to be intrinsic). The price we pay for such flexibility is increased uncertainty in the posterior distributions of climate.

There are several other potential drawbacks to the model as proposed. First, it is conceivable that the mixture formulation does not properly cover the marginal data posterior to sufficiently learn the climate volatility parameters. Such a problem will increase with n and G (the number of mixture components used). However, our model validation runs (Appendix B) show that this is almost never the case and coverage properties, even when G is under-estimated, still seem adequate. Another potential disadvantage is the modularisation assumption, both between the likelihood parameters θ and the rest of the model, and also between the chronology model and the climate process. The former seems most reasonable, as new cores are unlikely to impact much on the climate process given the strength of modern analogue data available. The latter, however, poses an interesting challenge, as if the sedimentation process is also posed as an intrinsic prior it is feasible for inclusion in our MDP-style inferential approach.

Some enhancements which follow immediately and to which our algorithm may still apply include:

- A multi-proxy analysis of palaeoclimate. This would require multiple forward models describing the relationship between climate and the various proxies. Our modularisation approach would not be hindered by such an extension, as we could simply create MDPs for the different proxies and include them as standard, so that the product MDP now becomes, e.g. $\pi_{\text{MDP}}^{\text{proxy1}}(c|y_1) \times \pi_{\text{MDP}}^{\text{proxy2}}(c|y_2)$. However, care needs to be taken in selecting the aspects of climate to which the different proxies are responsive. If these are substantially different, it may be that an extra process layer is required to match the different climate variables appropriately.

- The development of probabilistic forward models. These describe the causal chain from climate to proxy data. The forward model we use is relatively simplistic in its description of the mechanics of climate/pollen interaction, though it is far more sophisticated in its description of the uncertainty relating to the counting of pollen data and the relationships between pollen varieties. We encourage the development of physical forward models provided they retain suitable stochastic elements. A recent example of such thinking is Tolwinski-Ward et al. (2011).

- Richer climate process models. We might extend our time series approach into the spatial domain to give:

$$\begin{aligned} y(s_i, t_i)|c(s_i, t_i) &\sim f_\theta(c(s_i, t_i)),\ i = 1 \ldots, n \\ c(s_i, t_i)|\left\{c(s_1, t_1), \ldots, c(s_{i-1}, t_{i-1})\right\} = c &\sim \zeta_\kappa(c),\ i = 2, \ldots, n \end{aligned}$$

where now both pollen and climate are indexed by space s and time t and the prior distribution ζ is applied to climate change, parameterised by κ. We might assume that this would use all observations from previous time points $t_1, \ldots, t_{i-1}$ so that a particle algorithm might now become more appropriate. The prior ζ might be a simple stochastic climate model, or a richer version of our independent increments process including covariates and a spatial process. It is immediately obvious that c will no longer factorise out of the posterior, yet if ζ remains intrinsic a Laplace approximation might still allow our algorithm to proceed, though with caveats as to the size of the approximation error. Finally, even in situations where the prior is not intrinsic, it may be that other non-Gaussian mixture arrangements will yield simple tractable forms.

Performing inference on palaeoclimate over multiple sites may be possible by following the proposed methodology of Lindgren et al. (2011). In fact, the borrowing of strength from nearby cores may overcome one of our main issues: that of temporal uncertainty. It is certainly feasible that the constrained correlation of neighbouring sites will reduce temporal variability and thus provide more precise estimates of climate and possibly its associated volatility. Following this approach seems most promising in producing a pan-European map of palaeoclimate and its uncertainty.

6. Acknowledgements

We would like to thank the following for fruitful discussions as part of the SUPRAnet network: Jonty Rougier, Caitlin Buck, Tamsin Edwards, and Michel Crucifix. We would also like to thank T Brendan Murphy for his assistance with the mclust algorithm and Richard Chandler for comments on an earlier draft.

References

Barndorff-Nielsen, O. E. (1997). Normal Inverse Gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of Statistics 24*(1), 1–13.

Berliner, L. (1996). Hierarchical Bayesian time series models. In K. Hanson and R. Silver (Eds.), *Maximum Entropy and Bayesian Methods*, pp. 15–22. Dordrecht: Kluwer Academic Publishers.

Betrò, B. and R. Rotondi (1991). On Bayesian inference for the Inverse Gaussian distribution. *Statistics & Probability Letters 11*(3), 219–224.

Birks, H. (1988). Long-term ecological change in the British uplands. In M. B. Usher and D. B. A. Thompson (Eds.), *Ecological change in the uplands*, pp. 37–56. Oxford: Blackwell.

Blaauw, M. and J. A. Christen (2011). Flexible paleoclimate age-depth models using an autoregressive gamma process. *Bayesian Analysis 6*(3), 457–474.

Bronk Ramsey, C. (2008). Deposition models for chronological records. *Quaternary Science Reviews 27*(1-2), 42–60.

Brooks, S. and H. Birks (2001). Chironomid-inferred air temperatures from Lateglacial and Holocene sites in north-west Europe: progress and problems. *Quaternary Science Reviews 20*(16-17), 1723–1741.

Carvalho, C. M., M. S. Johannes, H. F. Lopes, and N. G. Polson (2010). Particle learning and smoothing. *Statistical Science 25*(1), 88–106.

Christen, J. A. and E. Perez (2009). A New Robust Statistical Model for Radiocarbon Data. *Radiocarbon 51*(3), 1047–1059.

Cressie, N. and C. K. Wikle (2011). *Statistics for Spatio-Temporal Data (Wiley Desktop Editions)*. Wiley-Blackwell.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*(458), 611–631.

Garreta, V., P. A. Miller, J. Guiot, C. Hély, S. Brewer, M. T. Sykes, and T. Litt (2009). A method for climate and vegetation reconstruction through the inversion of a dynamic vegetation model. *Climate Dynamics 35*(2-3), 371–389.

Grove, J. (2004). *Little Ice Ages: Ancient and Modern*. London: Routledge.

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

Haslett, J., M. Whiley, S. Bhattacharya, F. J. G. Mitchell, J. R. M. Allen, B. Huntley, S. P. Wilson, and M. Salter-Townshend (2006). Bayesian palaeoclimate reconstruction. *Journal of the Royal Statistical Society, Series A 169*, 395–438.

Huntley, B. (1993). The use of climate response surfaces to reconstruct palaeoclimate from Quaternary pollen and plant macrofossil data. *Philosophical Transactions of the Royal Society of London, Series B - Biological sciences. 341*, 215–223.

Huntley, B. (1994). Late Devensian and Holocene paleoecology and paleoenvironments of the Morrone Birkwoods, Aberdeenshire, Scotland. *Journal of Quaternary Science 9*(4), 311–336.

Huntley, B. (2012). Reconstructing palaeoclimates from biological proxies: some often overlooked sources of uncertainty. *Quaternary Science Reviews 31*, 1–16.

Jansen, E., J. Overpeck, K. R. Briffa, J.-C. Duplessy, F. Joos, V. Masson-Delmotte, D. Olago, B. Otto-Bliesner, W. R. Peltier, S. Rahmstorf, R. Ramesh, D. Raynaud, O. Rind, R. Solomina, V. R., and D. Zhang (2007). Palaeoclimate. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, T. M., and H. L. Miller (Eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Juggins, S. (2013, March). Quantitative reconstructions in palaeolimnology: new paradigm or sick science? *Quaternary Science Reviews 64*(null), 20–32.

Karlis, D. and J. Lillestol (2004). Bayesian estimation of NIG models via Markov chain Monte Carlo methods. *Applied Stochastic Models in Business and Industry 20*(4), 323–338.

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

Lindgren, F., H. v. Rue, and J. Lindström (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society - Series B: Statistical Methodology 73*(4), 423–498.

Liu, F., M. J. Bayarri, and J. O. Berger (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis 4*(1), 119–150.

Mann, M. E., R. S. Bradley, and M. K. Hughes (1998). Global-scale temperature patterns and climate forcing over the past six centuries. *Nature 392*(6678), 779–787.

Mann, M. E., R. S. Bradley, and M. K. Hughes (1999). Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. *Geophysical Research Letters 26*(6), 759.

McShane, B. B. and A. J. Wyner (2011). Discussion of: A statistical analysis of multiple temperature proxies: Are reconstructions of surface temperatures over the last 1000 years reliable? *The Annals of Applied Statistics 5*(1), 1–45.

Ohlwein, C. and E. R. Wahl (2012). Review of probabilistic pollen-climate transfer methods. *Quaternary Science Reviews 31*, 17–29.

Parnell, A. C., C. E. Buck, and T. K. Doan (2011). A review of statistical chronology models for high-resolution, proxy-based Holocene palaeoenvironmental reconstruction. *Quaternary Science Reviews 30*(21-22), 2948–2960.

Prentice, I. C., M. T. Sykes, and W. Cramer (1993). A Simulation-Model for the Transient Effects of Climate Change on Forest Landscapes. *Ecological Modelling 65*(1-2), 51–70.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (2002). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press.

R Foundation For Statistical Computing Austria, V. (2011). R Development Core Team, R: a language and environment for statistical computing.

Reimer, P. J., E. Bard, A. Bayliss, J. W. Beck, P. G. Blackwell, C. Bronk Ramsey, P. M. Grootes, T. P. Guilderson, H. Haflidason, I. Hajdas, C. Hatté, T. J. Heaton, D. L. Hoffmann, A. G. Hogg, K. A. Hughen, K. F. Kaiser, B. Kromer, S. W. Manning, M. Niu, R. W. Reimer, D. A. Richards, E. M. Scott, J. R. Southon, R. A. Staff, C. S. M. Turney, and J. van der Plicht (2013). IntCal13 and Marine13 Radiocarbon Age Calibration Curves 0–50,000 Years cal BP. *Radiocarbon 55*(4), 1869–1887.

Ribeiro, C. (2003). A Monte Carlo method for the Normal Inverse Gaussian option valuation model using an Inverse Gaussian bridge. *Business* (02), 1–15.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B 71*, 1–35.

Salter-Townshend, M. (2009). *Fast Approximate Inverse Bayesian Inference in non-parametric Multivariate Regression (with application to palaeoclimate reconstruction)*. Ph. D. thesis, Trinity College Dublin.

Salter-Townshend, M. and J. Haslett (2012). Fast inversion of a flexible regression model for multivariate pollen counts data. *Environmetrics 23*(7), 595–605.

Scott, E. M., G. T. Cook, and P. Naysmith (2010). A report on phase 2 of the fifth international radiocarbon inter- comparison (viri). *Radiocarbon 52*(3), 846–858.

Smith, A. G. and I. C. Goddard (1991). A 12,500 year record of vegetational history at Sluggan Bog, Co. Antrim, N. Ireland (incorporating a pollen zone scheme for the non-specialist). *New Phytologist 118*, 167–187.

Smith, B. J. (2005). {B}ayesian {O}utput {A}nalysis Program ({BOA}), Version 1.1.5.

Stuiver, M. (2000). GISP2 oxygen isotope ratios. *Quaternary Research 53*(3), 277–284.

Sweeney, J. (2012). *Advances in Bayesian Model Development and Inversion in Multivariate Inverse Inference Problems with application to palaeoclimate reconstruction*. Ph. D. thesis, Trinity College Dublin.

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

Tingley, M. P. and P. Huybers (2010). A Bayesian algorithm for reconstructing climate Anomalies in space and time. Part I: development and applications to paleoclimate reconstruction problems. *Journal of Climate 23*(10), 2759–2781.

Tolwinski-Ward, S., M. Evans, M. Hughes, and K. Anchukaitis (2011). An efficient forward model of the climate controls on interannual variation in tree-ring width. *Climate Dynamics 36*(11), 2419–2439.

West, M. and J. Harrison (1999). *Bayesian Forecasting and Dynamic Models (Springer Series in Statistics)*. Springer.

A. Technical details of the model fitting approach

For the following derivations we assume univariate climate so that c_i is actually a scalar. Everything that follows is extendable to multivariate climate with only minor notational changes, as discussed in the last section of this appendix.

First, we re-express the mixture component of the marginal data posteriors which were previously $p(c_i|y_i) = \sum_{g=1}^{G} \pi_{ig} N(\mu_{ig}, \Sigma_{ig})$. We introduce indicator variables z_{ig} that are 1 if observation i is in group g and zero otherwise, and vectorised as z_i. Thus we have $p(c_i|y_i) = \int p(z_i|\pi_i) \prod_{g=1}^{G} N(\mu_{ig}, \Sigma_{ig})^{z_i} \, \partial z_i$ where $z_i|\pi_i$ is a multinomial $(1, \pi_i)$ distribution where π_i are the known mixture weights for layer i. This re-expression has the advantage that, conditional on z, $\prod_{i=1}^{n} p(c_i|y_i, z_i)$ is multivariate normal. However, this introduces extra parameters z which must be estimated.

Following this re-arrangement, we require the posterior distribution:

$$\begin{aligned} p(c, v, z, t, \phi_1, \phi_2|\text{data}) \quad \propto \quad & \left[\prod_{i=1}^{n} \prod_{g=1}^{G} p(c_i|\mu_{ig}, \Sigma_{ig})^{z_{ig}}\right] \times \left[\prod_{i=2}^{n} p(c_i|c_{i-1}, t_i, t_{i-1}, v_{i-1})\right] \\ & \times \left[\prod_{i=1}^{n} p(z_i|\pi_i)\right] \times \left[\prod_{i=1}^{n-1} p(v_i|t_i, t_{i+1})\right] \times p(t|x, d) \times p(\phi_1) \times p(\phi_2) \end{aligned}$$

where all the distributions on the right hand side are known. $p(c_i|\mu_{ig}, \Sigma_{ig})$ is a univariate normal distribution with fixed mean μ_{ig} and variance Σ_{ig}, $p(c_i|c_{i-1}, v_{i-1})$ is Gaussian with given mean and variance, $p(z_i|\pi_i)$ is multinomial $(1, \pi_i)$ with $\pi_i = [\pi_{i1}, \ldots, \pi_{iG}]$ the vector of known mixture proportions for layer i. $p(\phi_1)$ and $p(\phi_2)$ are log-normal distributions with informative hyper-parameters set as described in the main text.

All the distributions involving c above are Gaussian so c can be analytically integrated out of the posterior. We now let $\prod_{i=1}^{n} \prod_{g=1}^{G} p(c_i|\mu_{ig}, \Sigma_{ig})^{z_{ig}} = MVN(M_z, D_z^{-1})$ where M_z is an n-vector of mixture means defined by the allocations in z and D_z is a diagonal matrix of mixture precisions, again defined by the allocations in z. Furthermore $\prod_{i=2}^{n} p(c_i|c_{i-1}, v_{i-1})$ can be written as $\left[\prod_{i=1}^{n-1} v_i^{-1/2}\right] e^{-\frac{1}{2} c^T W c}$ where $W = W(v)$ is the singular random walk precision matrix introduced in Section 3.

Setting $A = \prod_{i=1}^{n-1} v_i^{-1/2} \times \prod_{i=1}^{n} p(z_i|\pi_i) \times \prod_{i=1}^{n-1} p(v_i|t_i, t_{i+1}) \times p(t|x, d) \times p(\phi_1) \times p(\phi_2)$ and $Q = Q_z(v) = (D_z + W)^{-1}$, and focussing on the Gaussian part of the

full posterior, the above can be re-arranged to give:

$$\begin{aligned}
p(c, v, z, t, \phi_1, \phi_2|\text{data}) &\propto A \times |D_z|^{1/2} \exp\left[-\frac{1}{2}(c - M_z)^T D_z (c - M_z)\right] \times \exp\left[-\frac{1}{2} c^T W c\right] \\
&= A \times |D_z|^{1/2} \times \exp\left[-\frac{1}{2}(c - Q^{-1} D_z M_z)^T Q (c - Q^{-1} D_z M_z)\right] \\
&\quad \times \exp\left[-\frac{1}{2} M_z^T (D_z - D_z Q^{-1} D_z) M_z\right] \\
&= \frac{|D_z|^{1/2}}{|Q|^{1/2}} |Q|^{1/2} \exp\left[-\frac{1}{2}(c - Q^{-1} D_z M_z)^T Q (c - Q^{-1} D_z M_z)\right] \\
&\quad \times \exp\left[-\frac{1}{2} M_z^T (D_z - D_z Q^{-1} D_z) M_z\right]
\end{aligned}$$

Thus the posterior can be marginalised over the distribution $c|y, v, z \sim N(Q^{-1} D_z M_z, Q^{-1})$, removing climate c from this stage of the inference. It can subsequently be simulated from this multivariate normal given the posterior of $v, z, \phi_1, \phi_2|y$. It is this latter posterior on which we now focus.

The marginalised posterior is now:

$$p(v, z, t, \phi_1, \phi_2|\text{data}) \propto A \times \frac{|D_z|^{1/2}}{|Q|^{1/2}} \times \exp\left[-\frac{1}{2} M_z^T (D_z - D_z Q^{-1} D_z) M_z\right]$$

where all terms are now trivial to compute, as Q is a simple tri-diagonal matrix so its inverse requires only $O(n)$ steps. The full conditionals for the remaining parameters (treated individually as v_i and z_i) now become:

$$\begin{aligned}
p(v_i|\ldots) &\propto \frac{v_i^{-1/2}}{|Q|^{1/2}} \exp\left[\frac{1}{2} M_z^T D_z R_z\right] \times p(v_i|t_i, t_{i+1}) \\
p(z_i|\ldots) &\propto \frac{|D_z|^{1/2}}{|Q|^{1/2}} \times \exp\left[-\frac{1}{2} M_z^T D_z M_z + \frac{1}{2} M_z^T D_z R_z\right] \times p(z_i|\pi_i)
\end{aligned}$$

where R_z is the solution to $(D_z + W) R_z = D_z M_z$. These parameters can thus be updated extremely fast using simple Metropolis-Hastings. ϕ_1 and ϕ_2 can be similarly updated using only the Inverse Gaussian and log-Normal distributions upon which they depend.

A notable increase in speed can be obtained for parameter v using the Woodbury formula (e.g. Press et al., 2002) since when proposing a new v_i as, say v^* we can create $Q^* = Q + ((v^*)^{-1} - v_i^{-1}) B_i B_i^T$ with B_i as described in Section 3. The ratio of determinants now simplifies as :

$$\begin{aligned}
\frac{|Q|}{|Q^*|} &= \frac{|Q|}{|Q + (t_{i+1} - t_i)^{-1} B_i ((v^*)^{-1} - v_i^{-1}) B_i^T|} \\
&= \frac{|Q|}{|Q||1 + ((v^*)^{-1} - v_i^{-1}) B_i^T S_i|} \\
&= \left(1 + \left[(v_i^*)^{-1} - v_i^{-1}\right] B_i^T S_i\right)^{-1}
\end{aligned}$$

with S_i the solution to $(D_z + W)S_i = B_i$.

A.1. Multiple climate dimensions

For multiple climate dimensions $j = 1, \ldots, m$ we now have climates c_{ij} and increment variances v_{ij} parameterised by ϕ_{1j} and ϕ_{2j}. The mixture means μ and variances Σ are further parameterised by j and the joint posterior is:

$$\begin{aligned} p(c, v, z, t, \phi_1, \phi_2|\text{data}) \quad \propto \quad & \left[\prod_{j=1}^{m}\prod_{i=1}^{n}\prod_{g=1}^{G} p(c_{ig}|\mu_{gij}, \Sigma_{gij})^{z_{ig}}\right] \times \left[\prod_{j=1}^{m} p(c_{1j}) \prod_{i=2}^{n} p(c_{ij}|c_{i-1,j}, t_i, t_{i-1}, v_{i-1,j})\right] \\ & \times \left[\prod_{i=1}^{n} p(z_i|\pi_i)\right] \times \left[\prod_{j=1}^{m}\prod_{i=1}^{n-1} p(v_{ij}|t_i, t_{i+1})\right] \times p(t|x, d) \times \prod_{j=1}^{m} p(\phi_{1j}) \times p(\phi_{2j}) \end{aligned}$$

Unsurprisingly, the same marginalization over c occurs as before, and we obtain:

$$p(v, z, t, \phi_1, \phi_2|\text{data}) \quad \propto \quad A \times \prod_{j=1}^{m} \frac{|D_{zj}|^{1/2}}{|Q_j|^{1/2}} \times \exp\left[-\frac{1}{2} M_{zj}^T (D_{zj} - D_{zj} Q_j^{-1} D_{zj}) M_{zj}\right]$$

where A is now $\prod_{j=1}^{m}\prod_{i=1}^{n-1} v_{ij}^{-1/2} \times \prod_{i=1}^{n} p(z_i|\pi_i) \times \prod_{j=1}^{m}\prod_{i=1}^{n-1} p(v_{ij}|t_i, t_{i+1}) \times p(t|x, d) \times \prod_{j=1}^{m} p(\phi_{1j})p(\phi_{2j})$. The updates for v and ϕ_1, ϕ_2 are unaffected as they factorise across climate dimensions. The update for z is now:

$$p(z_i|\ldots) \quad \propto \quad \prod_{j=1}^{m} \left\{ \frac{|D_{zj}|^{1/2}}{|Q_j|^{1/2}} \times \exp\left[-\frac{1}{2} M_{zj}^T D_{zj} M_{zj} + \frac{1}{2} M_{zj}^T D_{zj} R_{zj}\right] \right\} \times p(z_i|\pi_i)$$

B. Model validation

In this section we determine the properties of our model fitting algorithm using simulated data under some idealised and non-idealised circumstances. To improve the speed of our tests we simplify the likelihood somewhat. Similarly we simulate data only observed on fixed, unit time. We consider 5 different scenarios:

(a) A simple Gaussian test that the parameters are identifiable when simulated from the model. We set $n = 100$ and $m = 3$. For $j = 1, \ldots, m$ we first simulate $\phi_{1j} \sim U(0.1, 10)$ and $\phi_{2j} \sim U(0.1, 10)$. For $i = 1, \ldots, n-1$ we then create $v_{ij} \sim IG(\phi_{1j}, \phi_{2j})$ and, for $i = 1, \ldots, n$, we create $c_{ij} - c_{i-1,j} \sim N(0, v_{ij})$. Finally we create $\delta_i \sim U(0.02, 2)$ and simulate pseudo pollen $y_{ij} \sim N(c_{ij}, \sqrt{\delta_i^{-1}})$. From the pseudo pollen data and the Gaussian likelihood we obtain Gaussian MDPs (with no simulation or mixture approximation required) which are passed, with the values of ϕ_j and η_j, to our MCMC functions to provide posterior distributions of 3-dimensional climate and volatility.

(b) A zero-inflated Poisson likelihood with 3 pollen taxa. The IG parameters, volatilities and climates are simulated as above, but we create pseudo-pollen via $y_{i1} \sim ZIP(p_1, \sqrt{a_1 c_1^2 + a_2 c_2^2})$, $y_{i2} \sim ZIP(p_2, \sqrt{a_1 c_1^2 + a_3 c_3^2})$ and $y_{i3} \sim ZIP(p_3, \sqrt{a_1 c_1^2 + a_2 c_2^2 + a_3 c_3^2})$. Here $ZIP(p, r)$ is a zero-inflated Poisson distribution with zero inflation parameter p and rate r. We set p_1, p_2, p_3 respectively as $p_j \sim U(0, 0.2)$ and a_1, a_2, a_3 as Poisson rate parameters simulated as the modulus of a normal distribution: $a_j \sim |N(0, 1)|$. The pseudo-pollen data are turned into MDPs via importance sampling. The ZIP model specified above gives MDPs that are quite often multimodal. The MDPs are then approximated as mixtures using $G = 5$ mixture components. These mixture components are then passed to our MCMC algorithm to estimate climates and volatilities.

(c) Exactly as (b) but using only 2 mixture components. In many situations this will be a poor representation of the MDP and thus may bias estimates of climate or volatility.

(d) Split into two parts:

  (i) Exactly as (b) but with the Inverse Gaussian parameters ϕ_{1j} and ϕ_{2j} given an underestimating multiplicative bias value simulated from $U(0.5, 1)$.

  (ii) Exactly as (b) but with the Inverse Gaussian parameters ϕ_{1j} and ϕ_{2j} given an overestimating multiplicative bias value simulated from $U(1, 5)$.

Table 1.
Performance of the different model validation scenarios

Scenario	Detail	Proportion inside 90% CI	Proportion inside 50% CI
1	Gaussian likelihood	90.7%	50.8%
2	ZIP likelihood	90.8%	47.7%
3	ZIP likelihood (too few mixture components)	90.1%	44.5%
4a	ZIP likelihood (under-estimated IG parameters)	91.6%	46.5%
4b	ZIP likelihood (over-estimated IG parameters)	94.0%	51.1%

We run each of the above 1000 times, and check the coverage properties of the climate posterior to see whether they lie within the 90% and 50% credibility intervals. Table 1 shows the results. Under each scenario the model seems to perform extremely well.

C. R Package Bclim

Bclim is available as part of the open source, free, statistical software R (R Foundation For Statisti
R is available to download from `www.r-project.org`. To install the package Bclim simply type `install.packages("Bclim")` at the R prompt, followed by `library(Bclim)`. The Bclim package is made up of four main functions (covering the creation of MDPs, mixture approximation, MCMC, and interpolation), 2 plotting functions (for climate and climate volatility), and a function which runs all necessary steps in sequence.

Example data to run the function can be downloaded from `http://mathsci.ucd.ie/~parnell_a` or, if this is not available, as part of the supplementary material to this paper. To run the Sluggan example shown in Section 4, the files should be downloaded via the commands:

```
# Download and load in the response surfaces:
url1 <- 'http://mathsci.ucd.ie/~parnell_a/media/requireddata3D.RData'
download.file(url1,'required.data3D.RData')

# and now the pollen
url2 <- 'http://mathsci.ucd.ie/~parnell_a/media/SlugganPollen.txt'
download.file(url2,'SlugganPollen.txt')

# and finally the chronologies
url3 <- 'http://mathsci.ucd.ie/~parnell_a/media/Sluggan_2chrons.txt'
download.file(url3,'Slugganchrons.txt')
```

The response surfaces in the first command above are the pre-calibrated forward model parameters θ. The subsequent functions use the locations of the pollen and chronology file rather than loading them into RAM:

```
# Create variables which state the locations of the pollen and chronologies
pollen.loc <- paste(getwd(),'/SlugganPollen.txt',sep='')
chron.loc <- paste(getwd(),'/Slugganchrons.txt',sep='')

# Load in the response surfaces
load('required.data3D.RData')
```

The functions now proceed as `BclimLayer` which produces the marginal data posteriors, `BclimMixPar` or `BclimMixSer` which approximate the MDPs as mixtures (either in parallel or serial respectively), `BclimMCMC` which produces posterior chains of volatilities and climates, and `BclimInterp` which uses the Inverse Gaussian and Brownian bridges to interpolate climate. Finally `BclimCompile` produces a list object which can be passed to `plotBclim` or `plotBclimVol` for plotting:

```
step1 <- BclimLayer(pollen.loc,required.data3D=required.data3D)
step2 <- BclimMixSer(step1)
step3 <- BclimMCMC(step2,chron.loc)
step4 <- BclimInterp(step2,step3)
results <- BclimCompile(step1,step2,step3,step4,core.name="Sluggan")

# Create a plot of MTCO (dim=2)
plotBclim(results,dim=2)

# Create a volatility plot
plotBclimVol(results,dim=2)
```

Each of the above functions has an associated help file which provides further information and options.