Geophysical Journal International

Geophys. J. Int. (2018) **215,** 2008–2034
Advance Access publication 2018 September 18
GJI Geomagnetism, rock magnetism and paleomagnetism

doi: 10.1093/gji/ggy383

Transdimensional inference of archeomagnetic intensity change

Philip W. Livermore [1], Alexandre Fournier [2], Yves Gallet[2] and Thomas Bodin [3]

[1] *School of Earth & Environment, University of Leeds, Leeds, UK. E-mail: p.w.livermore@leeds.ac.uk*
[2] *Institut de Physique du Globe de Paris, Sorbonne Paris Cité, Université Paris Diderot, CNRS, Paris, France*
[3] *Univ Lyon, Université Lyon* 1, *Ens de Lyon, CNRS, UMR* 5276 *LGL-TPE, F-*69622, *Villeurbanne, France*

Accepted 2018 September 14. Received 2018 September 3; in original form 2018 March 27

SUMMARY

One of the main goals of archeomagnetism is to document the secular changes of Earth's magnetic field by laboratory analysis of the magnetization carried by archeological artefacts. Typical techniques for creating a time-dependent model assume a prescribed temporal discretization which, when coupled with sparse data coverage, require strong regularization generally applied over the entire time-series in order to ensure smoothness. Such techniques make it difficult to characterize uncertainty and frequency content, and robustly detect rapid changes. Key to proper modelling (and physical understanding) is a method that places a minimum level of regularization on any fit to the data. Here we apply a transdimensional Bayesian technique based on piecewise linear interpolation to sparse archeointensity data sets, in which the temporal complexity of the model is not set *a priori*, but is self-selected by the data. The method produces two key outputs: (i) a posterior distribution of intensity as a function of time, a useful tool for archeomagnetic dating, whose statistics are smooth but formally unregularized and (ii) by including the data ages in the model of unknown parameters, the method also produces posterior age statistics of each individual contributing datum. We test the technique using synthetic data sets and confirm agreement of our method with an integrated likelihood approach. We then apply the method to three archeomagnetic data sets all reduced to a single location: one temporally well-sampled within 700 km from Paris (here referred to as Paris700), one that is temporally sparse centred on Hawaii, and a third (from Lübeck, Germany and Paris700) that has additional ordering constraints on age from stratification. Compared with other methods, our average posterior distributions largely agree, however our credible intervals appear to much better reflect the uncertainty during periods of sparse data coverage. Because each ensemble member of the posterior distribution is piecewise linear, we only fit oscillations when required by the data. As an example, we show that an oscillatory signal, associated with temporally localized intensity maxima reported for a sparse Hawaiian data set, is not required by the data. However, we do recover the previously reported oscillation of period 260 yr for the Paris700 data set and compute the probability distribution of the period of oscillation. We further demonstrate that such an oscillation is unresolved when accounting for age uncertainty by using a fixed age and with an artificially inflated error budget on intensity.

Key words: Archaeomagnetism; Magnetic field variations through time; Inverse theory; Statistical methods; Time-series analysis.

1 INTRODUCTION

Archeomagnetism, in a broad sense, is the study of Earth's magnetic field over the last ∼10 000 yr, conducted by analysis of the magnetization carried by archeological artefacts, lake sediments and volcanic records. Archeomagnetic reconstructions allow insight into how the Earth's magnetic field has altered over time (e.g. Korte *et al.* 2011; Licht *et al.* 2013; Nilsson *et al.* 2014; Pavón-Carrasco *et al.* 2014), useful not only for supplying constraints on the process by which the field is generated in the fluid outer core, but also for archeomagnetic dating of artefacts (Le Goff *et al.* 2002; Pavón-Carrasco *et al.* 2011).

One of the principal challenges of reconstructing the archeomagnetic field is data sparsity, in both space and time. Archeological features such as ancient kilns and pottery that record the magnetic field from when they last cooled are relatively rare, and localized to those ancient human settlements where remains were unearthed. The data set also may be very sparse or particularly poorly dated over

certain periods (sometimes referred to as dark ages), a reflection of either lack of data collection or simply that there are few remaining artefacts suitable for analysis. Scarcity is a particular problem for analysis of localized transient phenomena, which appear to punctuate the quasi-steady ancient magnetic field, such as *archeomagnetic jerks*: sharp changes in the local direction of the field over timescales of about one century associated with an intensity peak (Gallet *et al.* 2003), and *geomagnetic spikes*: large, rapid changes in the local intensity over a few decades (e.g. Shaar *et al.* 2011). In rare cases, some data sets offer a well-dated stratified sequence (Schnepp *et al.* 2009; Shaar *et al.* 2011; Kostadinova-Avramova *et al.* 2014) which add significant constraints to the behaviour of the archeomagnetic field through time.

Models of the archeomagnetic field may be global, regional or reduced to a single location (in the latter case, by using the virtual axial dipole approximation). All studies to date have used some form of prescribed smooth dependence in time to accommodate temporal changes of the field. The most widely used framework adopts two means of smoothing: first by using a temporal expansion in terms of cubic spline functions centred on a set of knot points, and secondly by reducing the temporal complexity of the model by a penalization method (Korte & Constable 2003; Lanos 2004; Thébault & Gallet 2010; Hervé & Lanos 2017; Tema *et al.* 2017). A prescription of the knot points in advance places an *a priori* constraint on the solution, for dynamics on timescales more rapid than the closest spacing cannot be accommodated. Furthermore, the penalization of temporal complexity (often the second time derivative of the radial field at the core–mantle boundary) is applied uniformly over the model, which means that some periods will be oversmoothed and some undersmoothed. A sliding-window with linear regression (Lanos *et al.* 2005) is an alternative to these methods, where the window duration may be a function of the temporal distribution of data (Le Goff *et al.* 2002). It is possible to remove temporal smoothing all together by adopting a stochastic process representation of the geomagnetic field (Hellio *et al.* 2014), although this procedure is still subjective as the parameters of this process (the covariances) need to be prescribed.

Transdimensional Bayesian methods offer an attractive alternative, in which the models are not regularized but allow the data to self-select model complexity, here the number and position of internal knot points (also known as internal vertices or change points). Despite the freedom of this procedure to choose an unlimited number of degrees of freedom leading to overfitting and highly complex models, in fact owing to the natural parsimony of the Bayesian framework the models are inherently smooth and large scale, and admit fine-structure only where required by the data (Sambridge *et al.* 2013). Rather than resulting in a single time-dependent solution, the Bayesian modelling produces a posterior distribution, numerically approximated by an ensemble of models sampled using a Markov chain, whose shape and variance gives important information about the behaviour of the solution and confidence that should be placed in it. Such transdimensional methods have been widely applied across the geosciences, in applications such as Earthquake rupture (Dettmer *et al.* 2014), geomagnetic reversals (Ingham *et al.* 2014), seismic imaging (Bodin *et al.* 2012), climate studies (Hopcroft *et al.* 2007), 2-D seafloor resistivity (Ray *et al.* 2014), inversion for gravitational anomalies (Luo 2010) and inference of abrupt changes in geochemical records (Gallagher *et al.* 2011).

An additional benefit of using a Bayesian procedure is that poorly known quantities (on which the posterior distribution depends), that would otherwise need to be fixed at some arbitrary level in a traditional inversion approach, can be co-estimated (Malinverno & Briggs 2004; Bodin *et al.* 2012; Hervé & Lanos 2017). This has a particular benefit for archeomagnetism for which data ages can be difficult to assess. Archeological data are typically assumed uniformly distributed between two dates determined by typology or historical reference; in contrast, volcanic data are dated using radiocarbon techniques and in this work their errors are assumed normally distributed with zero mean and a given standard deviation, conforming to the databases from which they are extracted. By including the data ages as additional or hyperparameters, a Bayesian formulation of the problem allows us to sample both the data ages and the model description in the same way, producing marginal posterior distributions for the data ages themselves (Lanos 2004; Hellio *et al.* 2014; Schnepp *et al.* 2015; Hervé & Lanos 2017) alongside that of the archeomagnetic intensity model. In this way, data with different intensities but the same mean estimated age may be judged by the posterior distribution to be associated with distinct ages. Such a scheme differs fundamentally from a typical global modelling approach in which apparently contemporaneous data, rather than being treated as distinct, are averaged to a single value (Korte & Constable 2018). Furthermore, in terms of archeomagnetic dating, this fully integrated approach that we adopt here should be contrasted with typical methods of dating in which the magnetic signature of a datum is matched against a separately produced time-dependent model of the geomagnetic field (e.g. Pavón-Carrasco *et al.* 2011; Hervé & Lanos 2017; Lanos & Philippe 2017): here the age distributions are a component of the model output.

In this paper, we apply the transdimensional Bayesian framework to three time-series of archeomagnetic intensity, chosen to represent very different situations regarding the nature and quality of the archeomagnetic data sets available at a regional scale (in particular, their age distribution and uncertainties, consistency between the data, number of data), which are described in Section 2. In Section 3, we summarize the Bayesian method; more detail is included in the Appendices. Section 4 describes important benchmarks of the method against synthetic data sets. Sections 5–7 show the results of our method applied to the three archeomagnetic data sets, whose results are further discussed in Section 8 with reference to results from other methods.

2 THREE EXEMPLAR DATA SETS

Archeomagnetic data includes measurements of either (or both of) intensity (i.e. magnitude) and direction (inclination, declination). For this initial exploration of the method, we will focus only on intensity variations. Below, we give an overview of the three chosen data sets; more details (e.g. on the specifics of data selection) can be found in Appendix A. These data sets are shown in Fig. 1, which shows not only the individual data (blue points) but the temporal data density (green histogram). For each data set, the magnetic intensity of each datum is assumed to be normally distributed, with a mean and standard deviation estimated from laboratory analysis as shown by the vertical error bars; the interpretation of the horizontal error bars is described in turn for each data set.

(1) **Paris700**: The first data set, *Paris700*, relies on stringent selection criteria (see Genevey *et al.* 2013), has quasi-uniform temporal resolution from about 1000 BC to 2000 AD and comprises data collected from within 700 km from Paris. Its 154 entries are all reduced to the latitude of Paris (48.9°N) using the virtual axial dipole moment (VADM) approximation. The data age errors are assumed uniformly distributed within given ranges (shown by the horizontal error bars in Fig. 1(top panel).

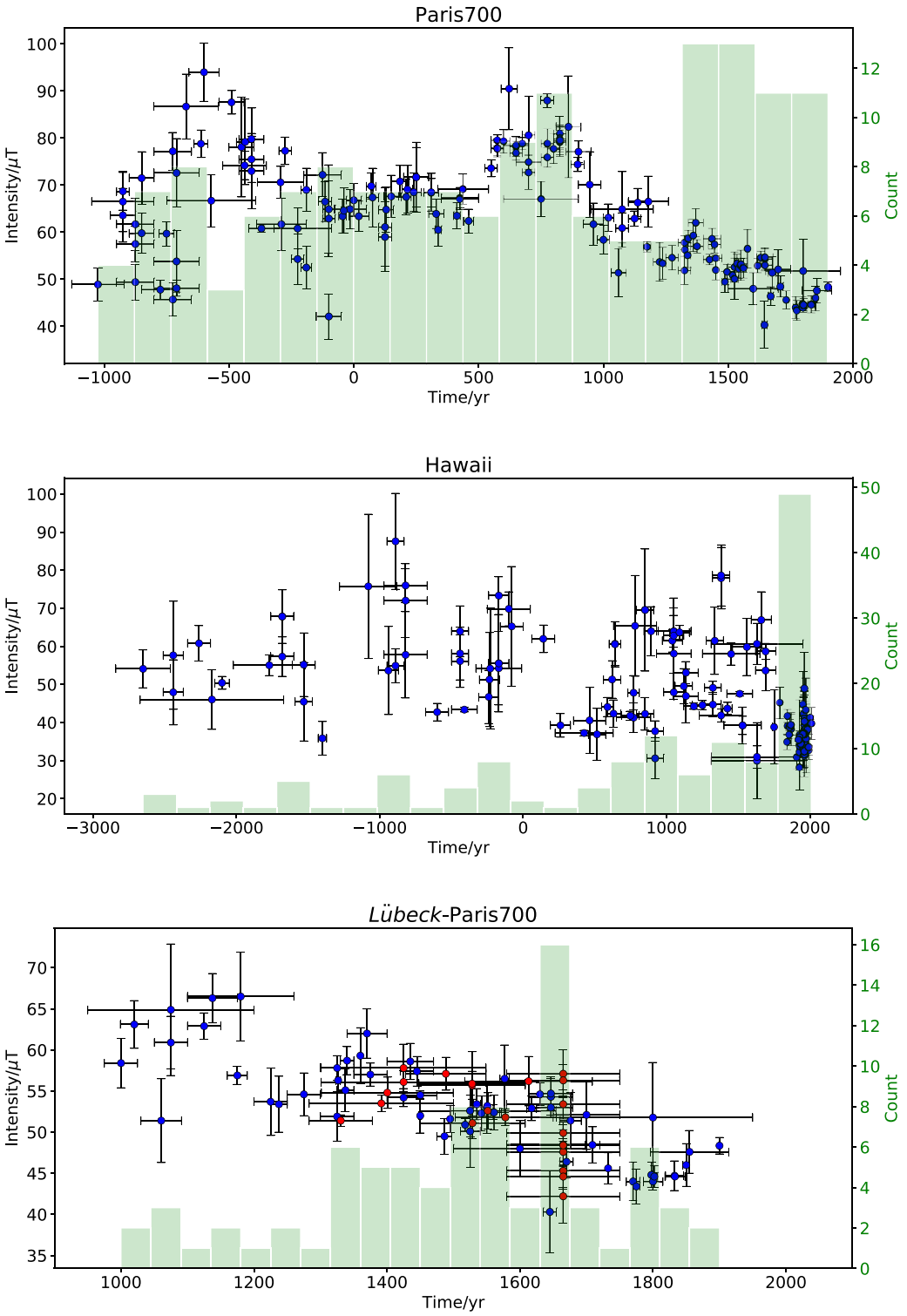

Figure 1. The three exemplar data sets that will be analysed by the Bayesian method in this paper: Paris700 (top panel); Hawaii (middle panel) and Lübeck-Paris700 (bottom panel). The data points are shown as blue circles with error bars indicating the uncertainty as described in the text. The red data points obey a strict ordering in age. The data-density is shown by the green histograms using 20 bins of equal width for each data set.

(2) **Hawaii**: To examine the effect of data sparsity and non-uniform sampling on our method, we consider a second data set of Hawaiian intensity measurements taken from the Geomagia.v3 database (Brown *et al.* 2015) over the past 4500 yr. Here, we adopted minimalist (i.e. loose) selection criteria and this case therefore corresponds to extracting a local data set from the global database in a relatively blind manner. All 134 retained entries are reduced to the latitude of Kilauea volcano (19.42°N). The data are shown in

Fig. 1(middle panel), where the horizontal error bars indicate one standard deviation in radiocarbon-determined age. Note that recent 'historical' lava flows are assigned a normally distributed age uncertainty of 0.5 yr. The density of data is heavily weighted towards the most recent times.

(3) **Lübeck-Paris700**: Finally, some archeomagnetic data sets contain stratified data whose ages obey a strict ordering in time. We investigate a third data set, comprising the (unstratified) Paris700 data (restricted between 900 and 2000 AD) with 22 stratified data with otherwise poorly constrained ages from Lübeck, Germany (Schnepp *et al.* 2009), reduced to Paris using the VADM approximation, for the period ∼1300 AD to ∼1750 AD. The data (unstratified: blue; stratified: red) are shown in Fig. 1(bottom panel) where all ages are uniformly distributed in time with additional strict ordering constraints where relevant.

3 THE TRANSDIMENSIONAL BAYESIAN ALGORITHM

3.1 Bayesian inference

Bayesian inference is a procedure by which the knowledge of model parameters is improved from the *prior* information, expressed in terms of probability distributions, by the introduction of data. The result is not simply a single optimized model that fits the data, but rather the *posterior distribution*, characterized by an ensemble of models, which describes the probability of the unknown model parameters given the observed data. In this way, the posterior distribution not only describes the most likely value (i.e. the mode of the distribution), but also diagnostics such as credible intervals (the analogue of confidence intervals). The analysis rests upon Bayes' theorem:

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})\, p(\mathbf{m})}{p(\mathbf{d})}, \tag{1}$$

where $\mathbf{m}$ is the model and $\mathbf{d}$ is the collection of data. Expressed in words, the posterior (on the left-hand side) for the model *given* the data, is proportional to the product of the likelihood (the first term in the numerator) and the prior; the term in the denominator is termed the evidence (which does not depend on the model $\mathbf{m}$) and although useful for choosing between physical models, will not enter into our analysis. The posterior distribution then depends on several key elements: the model parametrization, the likelihood and the prior. We will describe each in turn.

3.2 The model

The principal part of any individual model is the representation of intensity change with time, which we parametrize using continuous piecewise-linear functions defined over a period $[t_{\rm start}, t_{\rm end}]$. The reason for our choice of linear, rather than higher order temporal dependence (e.g. cubic splines) is that we wish to assert minimal smoothness on the posterior. Any curvature within the posterior must be required by the data and cannot be an artefact of our parametrization. The piecewise linear regression function, g, is interpolated from the value of the intensity at the endpoints $F_{\rm start}$ and $F_{\rm end}$, along with the age t_j and intensity F_j with $1 \leq j \leq k$ of any of the k internal change points. The number of internal change points, $k \geq 0$, is a free parameter of this transdimensional model. We note that although every model individually is piecewise linear, taking ensemble statistics yields a smoothed evolution that gives a good description of magnetic field variability.

In this study, we will assume that the errors on both the data intensities and ages are known. However, because outliers are common in archeomagnetic data sets, a fuller treatment could include these errors as hyperparameters (Malinverno & Briggs 2004). Key to our algorithm is an appropriate handling of the data ages, on which the model depends. There are multiple possibilities of how we can treat them. First, we could treat the ages as known and error-free, and simply increase the error budget in the intensity uncertainty as a proxy for a combined error estimate (e.g. Ingham *et al.* 2014). Second, we could use a bootstrap method to repeatedly draw possible ages from their given distributions (e.g. Thébault & Gallet 2010). Third, we could incorporate the unknown data ages into an 'integrated likelihood', in which the data ages are entirely absent from the model vector (Sambridge 2016). Fourth, we can include the data ages as hyperparameters into our model vector. In this work, we adopt the last approach in our *age-hyperparameter* (or AH) method: not only does this handle the age uncertainties in an identical manner to the uncertainty in the other model parameters, but after recovery of our posterior distribution, the marginal distributions of the data ages can be readily computed. For the sake of completeness, however, we will confirm some of our calculations using an integrated likelihood approach (IL; see Section 4.3).

It is worth remarking that our incorporation of data ages as hyperparameters conflicts with the usual division between model-space and data. To an archeomagnetist, the data *are* the set of measurements of intensity along with the data ages. Yet, in isolation, the data ages are not informative about the distribution of the temporal variation of intensity. Therefore, we partition the intensity and ages—absorbing the ages into the unknown model vector and treating the 'data' ($\mathbf{d}$ of eq. 1) as simply the set of N_{data} intensity values, denoted here as $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_{N_{data}}$.

We write the model vector as

$$\mathbf{m} = [\mathbf{f}, k, \mathbf{a}]^T,$$

where T indicates transpose, and $\mathbf{m}$ is a column vector of size $2k+3+N_{\rm data}$, where $\mathbf{f} = (F_{\rm start}, F_{\rm end}, F_1, F_2, \ldots, F_k, t_1, t_2, \ldots, t_k)$ describes the $2k+2$ vertices of the piecewise linear time-dependence, and $\mathbf{a} = (a_1, a_2, \ldots, a_{N_{\rm data}})$ are the ages of the $N_{\rm data}$ data. When adopting the integrated likelihood, the model is simply $\mathbf{m} = [\mathbf{f}, k]^T$. Fig. 2 shows a cartoon of a low-dimensional model with two data points and three internal vertices.

Two aspects of notation are worthy of comment. The first is that although we have used both t and a to denote an age, this allows us to distinguish between the ages of the change points t_j and the data a_j. Second is that we have used the calligraphic symbol $\mathcal{F}_j$ to denote the data (laboratory determined intensity) and F to denote the intensity value of a model vertex.

3.3 The likelihood

Any specific realization of the model $\mathbf{m}$ prescribes the data ages $\mathbf{a}$ along with the set of internal vertices required to define the piecewise-linear regression function, g, that describes intensity variation with time. To define a likelihood, for the jth datum we compute the difference between its intensity $\mathcal{F}_j$ and the value of the regression function $g(a_j)$ evaluated at its age. Because of the assumed normally distributed errors on the data $\mathcal{F}_j$, and because we assume errors independent between measurements, the likelihood of all the

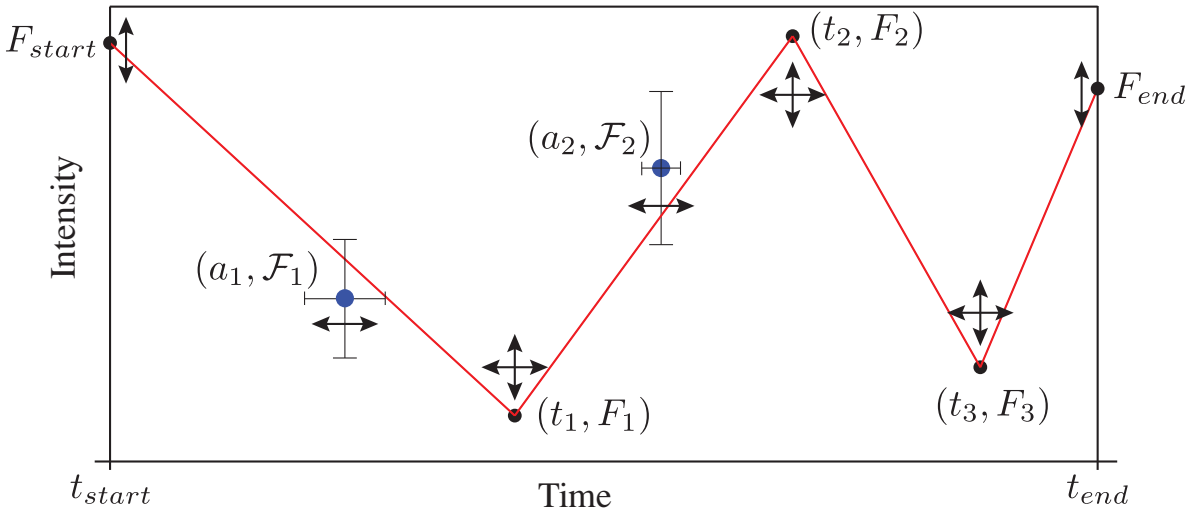

Figure 2. Cartoon of a low-dimensional model over a fixed period [$t_{\rm start}$, $t_{\rm end}$] constrained by two archeomagnetic data, with given intensities $\mathcal{F}_1, \mathcal{F}_2$ and unknown ages a_1 and a_2. The intensity variation is described by a piecewise-linear interpolation between the three internal vertices (t_i, F_i) and the two endpoint intensities $F_{\rm start}$ and $F_{\rm end}$. The arrows indicate how the Monte Carlo method allows each parameter to change.

data intensities being realized is proportional to $e^{-\phi}$, where

$$\phi = \frac{1}{2}\sum_{j=1}^{N_{data}} \left[g(a_j) - \mathcal{F}_j\right]^2/\sigma_j^2, \tag{2}$$

where σ_j is the given estimate of the standard deviation. We note that, because only the ratio of likelihoods is relevant, the constant of proportionality never enters the analysis. In Sections 5 and 6, for the Paris700 and Hawaii data sets, the distribution of residuals will confirm the assumption of normally distributed intensity errors.

The integrated likelihood approach of Sambridge (2016) is a generalization of the formula above to the case where the data ages are formally unknown but have a known distribution. By integrating (2) over all its possible values, it is possible to calculate a likelihood weighted by the age distribution; see Section 4.3 for further details.

3.4 The choice of prior distributions

A keystone of Bayesian methods is the description of the prior information as a probability distribution that characterizes what is known or supposed about the model before the introduction of data. The data ages are independent of the rest of the model, and by conditioning on the value of k, we can write the prior as

$$p(\mathbf{m}) = p(\mathbf{a})\, p(\mathbf{f}\,|k)\, p(k).$$

We assume that $p(k)$ is a uniform distribution over the interval [0, $k_{\rm max}$], where $k_{\rm max}$ is prescribed (we typically take $k_{\rm max} = 50$), that is

$$p(k) = \begin{cases} (k_{\rm max}+1)^{-1}, & 0 \le k \le k_{\rm max}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Given a value of k, the intensity at each vertex F_j and its corresponding age t_j is assumed mutually independent, thus

$$p(\mathbf{f}|k) = p(\mathbf{t}|k)\prod_{j=1}^{k+2} p(F_j|k),$$

where $j = 1, 2, \ldots k$ indexes the internal intensity coefficients, and we have adopted the notational convenience $F_{k+1} = F_{\rm start}$ and $F_{k+2} = F_{\rm end}$. We assume that the prior distribution of each intensity value F_j is uniform:

$$p(F_j|k) = \begin{cases} (\tilde{F}_{\rm max} - \tilde{F}_{\rm min})^{-1}, & \tilde{F}_{\rm min} \le F_j \le \tilde{F}_{\rm max}, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $\tilde{F}_{\rm max}$ and $\tilde{F}_{\rm min}$ are prescribed values (typically taken to be 100 μT and 30 μT).

For the purposes of model development, it is expedient to assume that the vertex ages are distributed uniformly from a choice of N equally spaced ages between $t_{\rm start}$ and $t_{\rm end}$ (Bodin & Sambridge 2009). This means that the joint probability distribution of the vertex ages (in which their order is irrelevant) is

$$p((t_1, t_2, \ldots, t_k)|k) = \left[\frac{N!}{k!(N-k)!}\right]^{-1}.$$

However, for infinitely large N (that will describe our final model, see Appendix A), we convert from a discrete to a continuous variable; the prior distribution of each vertex age, t_j, becomes independent, uniformly and identically distributed:

$$p(t_j) = \begin{cases} (t_{\rm end} - t_{\rm start})^{-1}, & t_{\rm start} \le t_j \le t_{\rm end}, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

For unstratified data, each datum age is supposed independent so that

$$p(\mathbf{a}) = \prod_{j=1}^{N_{data}} p(a_j). \tag{6}$$

For a datum of index j that is archeologically dated, its age is uniformly distributed between given dates $[a_j^{\rm min}, a_j^{\rm max}]$ and has a prior distribution of

$$p(a_j) = \begin{cases} (a_j^{\rm max} - a_j^{\rm min})^{-1}, & a_j^{\rm min} \le a_j \le a_j^{\rm max}, \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

whereas for normally distributed ages (e.g. from radiocarbon dating) the prior distribution is

$$a_j \sim \mathcal{N}(\mu_j, \lambda_j^2) \tag{8}$$

where μ_j and λ_j are the given mean and standard deviation. For a subset of data that is stratified, the corresponding set of ages are not independent and must obey a constraint of the form $a_j <$

$a_{j+1} < \cdots$. Formally, such a constraint will alter the mathematical structure of the prior distribution as given above, but because of the methodology by which we sample the posterior distribution this structure never actually enters our analysis. For the data ages, we only actually need to be able to draw from their prior distribution; practically, after drawing ages according to (7) or (8), we simply discard any set of ages that violates any required age ordering, and redraw. It is further noteworthy that although we will focus on data whose ages are solely either uniformly or normally distributed, it is straightforward to consider a mixture of age distributions.

We note that despite specifying each model vertex intensity F_j to have a uniform distribution, in general the associated prior for the linearly interpolated intensity $g(t)$ (for given arbitrary t) is not uniformly distributed. The reason why the uniform characteristic does not carry over to $g(t)$ is because it only attains values close to its bounds of $\tilde{F}_{\max}$ and $\tilde{F}_{\min}$ when both of the vertices used in determining $g(t)$ by linear interpolation also have extreme intensity, a situation which has small probability compared to the probability of a single vertex being extreme. This can be seen more formally in the simple case of $k = 0$ (no interior vertices). The distribution of intensity at the middle of the age range $t_{\text{mid}} = (t_{\text{end}} + t_{\text{start}})/2$ is simply the average of the two endpoint intensity distributions, that is, one half of the sum of two identically uniformly distributed random variables, which takes the form of a triangular distribution that has a peak value of $(\tilde{F}_{\max} + \tilde{F}_{\min})/2$. Thus $g(t_{\text{mid}})$ is not uniform, in contrast to both $g(t_{\text{start}})$ and $g(t_{\text{end}})$, which are uniform. For arbitrary k this picture is more complicated, but a similar reasoning applies. In this way, our mean, median, mode, and credible intervals in the figures below are not exactly the mean, and median of the posterior model $p(\mathbf{m}|d)$, but rather of the posterior projected in the space of $g(t)$.

Finally, it is of note that several authors (e.g. Green 1995; Hopcroft *et al.* 2007) use order statistics in their prior, in order that the change points are spread out. Here, we do not want to assume this, as in fact we are interested in features for which there is rapid time-dependence and therefore we want to allow clustering of vertices. Many of our prior distributions are uniform and may not in fact encode zero information despite assigning all values within their range the same probability (Jaynes 2003).

3.5 The AH-RJMCMC sampling algorithm

Having defined the prior and likelihood, all that remains is to characterize the posterior distribution, which we undertake numerically by drawing a sufficiently large set of samples using the reverse-jump Monte Carlo Markov Chain algorithm (RJMCMC; Green 1995). The resulting Markov chain of models, in which each model depends only on its predecessor, represents a random walk through the space of all permissible models; its distribution converges to the posterior distribution. The chain is built iteratively by either adding a duplicate of the current last model $\mathbf{m}$ or adding a proposed model $\mathbf{m}'$, a perturbation to $\mathbf{m}$, which is drawn from a distribution of alternative models $q(\mathbf{m}'|\mathbf{m})$. The decision of whether to add $\mathbf{m}$ or $\mathbf{m}'$ to the chain is based on an acceptance test. The underlying methodology does not require any particular choice of q, but rather different choices of q will simply alter the speed of convergence to the posterior distribution (Green 1995).

In our algorithm, which is based on that of Bodin & Sambridge (2009) and Gallagher *et al.* (2011), the model $\mathbf{m}'$ differs from $\mathbf{m}$ by one of several possible perturbations, as depicted in Fig. 3 (see also Fig. 2) grouped by type. These perturbations depend on a set of user-specified parameters: σ_{change}, σ_{birth}, σ_{move} and β that are discussed later.

For perturbations of type 1, the intensity value of a randomly chosen internal vertex is perturbed from its current value by a random amount distributed as $\mathcal{N}(0, \sigma^2_{\text{change}})$. Perturbations of type 2 alter the temporal arrangement of the vertices. For a perturbation 2a, the age of a randomly chosen vertex is altered by the addition of a normally distributed perturbation, $\mathcal{N}(0, \sigma^2_{\text{move}})$. For a perturbation 2b (vertex birth), a new vertex is proposed randomly (according to a uniform distribution) within the temporal limits of the model, with an intensity that is distributed $\mathcal{N}(0, \sigma^2_{\text{birth}})$ relative to its linearly interpolated value based on the current vertex distribution. Perturbation 2c describes vertex death, where a vertex is removed from the model. Lastly, perturbation 3 describes the resampling of $\lfloor N_{data}/\beta \rfloor$ ages within their given prior distributions where $\lfloor x \rfloor$ denotes the floor (i.e. the integer part of) x. The specific details of each move type and their associated acceptance criteria mirror those of Bodin & Sambridge (2009) and are described in Appendix B.

Each run is initialized with a model with a random number of vertices, described by coefficients $\mathbf{f}$ and $\mathbf{a}$ randomly chosen from their given prior distributions. Diagnostic statistics of the distribution of the posterior intensity and the marginal distribution of the data ages are computed 'on-the-fly', usually adopting thinning (e.g. analysing only every 100th chain member in order to minimize the effect of any temporary localized confinement in model space of the chain). The algorithm is run until the posterior distribution (as indicated by the computed diagnostics) converges, which typically occurs after about $1–5 \times 10^6$ model proposal iterations; the compute time for this is a matter of a few seconds on a single-core desktop computer. In accord with standard practice, we discard the first 50 000 iterations as 'burn-in' in order to remove dependency on the initial model. We note that our method, in which the model vector contains the data ages, is considerably (i.e. thousands of times) faster than an implementation (which we tried in a development phase) of the two-stage approach of Hellio *et al.* (2014): in which the ages were drawn in an outer loop by a Monte Carlo method, and the posterior distribution of intensity was then determined within an inner loop assuming fixed ages.

The rate of convergence of the Markov chain (and hence that of the posterior distribution) is affected by the choice of model proposal. Slow convergence will occur if either the proposed models are too different from any current model (they are unlikely to be accepted and the Markov chain will change only infrequently), or if they are too close to the current model (the chain will never effectively explore distant parts of model space). The proposal q depends on σ_{change}, σ_{move}, σ_{birth}, β, which are adjusted to attain suitable acceptance ratios of about 15–50 per cent (Roberts 1996). Typical values are, respectively, 20 μT, 200 yr, 8 μT, 20, which give acceptance ratios (for model perturbation types of 1, 2a, 2b, 2c, 3) of 19, 11, 14, 14, 21 per cent for the Paris700 data set reported in Section 5.

In addition to reporting diagnostics such as histograms of vertex position, we also compute the relative entropy, or Kullback–Leibler divergence (e.g. Press *et al.* 2001), a quantification of the difference between the posterior $p(g(t)|\mathbf{d})$ and prior distributions $p(g(t))$ of intensity variation (or in other words, a measure of information gain from the data):

$$D_{KL}(t) = \int p(g(t)|\mathbf{d}) \ln \left(\frac{p(g(t)|\mathbf{d})}{p(g(t))} \right) dg(t).$$

As we remarked earlier, the prior on the intensity evolution is not immediate from the prior information on the internal vertices; here,

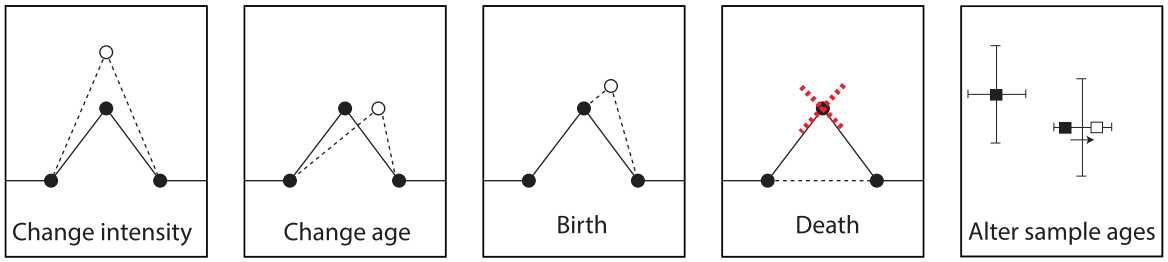

Figure 3. Cartoon showing the five possible perturbations (unfilled symbols) to the current model (black symbols). In type (1), the intensity at a vertex is changed; in (2a) the age of a vertex is changed; in (2b) a vertex is born; in (2c) a vertex is removed; in (3) a subset of the data ages are resampled. The equal spacing of the black symbols is for graphical purposes only: in general they are spaced unequally.

we compute it by running the AH-RJMCMC algorithm with the likelihood set to 1. As a side note, this procedure of amending the likelihood should (and does) recover the prior distribution on each of the variables that we have specified.

If D_{KL} is close to zero, it means that the posterior and prior distributions are very similar, so that the data add little information. Conversely, a high value of D_{KL} means that the prior and posterior are significantly different. The relevance of this quantity here is that all Bayesian inferences are dependent on the choice of prior. If the KL divergence is low, then we might have little confidence in the robustness of the posterior distribution for it would likely alter if we changed the prior. On the other hand, a high value of D_{KL} would render the posterior largely invariant of the choice of prior. As shown in Sections 5–7, high values of D_{KL} occur when the posterior distribution is tightly focused, signifying strong constraints from the data relative to the assumed broad uniform prior.

4 SYNTHETIC-DATA TESTS

4.1 Recovery of known underlying behaviour

In order to test the AH-RJMCMC methodology, we created synthetic versions of the Paris700 and the Hawaii data sets, termed Synt-Paris700 and Synt-Hawaii, both based on the smooth intensity variation of the coupled-Earth (CE) dynamo model (Aubert *et al.* 2013). The particular time window in the CE model was chosen such that the magnitude of variation of intensity was comparable to the observations, and is time-shifted in order that it is defined over the same age ranges as the observational data sets. It is worth remarking that the CE model time is scaled to terrestrial ages through considerations of the secular variation (Lhuillier *et al.* 2011), and any resemblance to existing intensity changes at either location is fortuitous. The CE model was sampled at the appropriate latitude/longitude location and, crucially, according to the same age distributions as the Paris700 and Hawaii data sets themselves. Pseudorandom noise is added in both intensity and age (according to their assumed forms) to mimic errors in the real data sets. We are now in a position to test the recovery of intensity evolution from noisy data, from either well-sampled or sparsely sampled data sets.

Fig. 4 shows diagnostics of the AH-RJMCMC method applied to the Synt-Paris700 data set. The top row shows the posterior distribution of intensity characterized by its (time-dependent) average, median and modal values, alongside the 95 per cent credible intervals in filled orange; in the middle row is a density plot of the intensity distribution. Of primary importance is that for almost all the time window (except a small deviation around 1000 AD) the 'true' evolution (shown in black) largely agrees with the mode, median and average of the posterior distribution. Indeed, the true evolution is contained within the 95 per cent credible intervals (and the regions of highest intensity density), which gives us great confidence that the method can recover the underlying evolution. Fig. 4(bottom panel) shows the vertex position histogram with the KL divergence overlaid in red. The histogram shows strong evidence for a change in linear slope around 1400 AD and 1700 AD which occurs during a period of densely sampled data. There are several other ages which favour a change in slope, although their probability is smeared out and consequently not so high due to lower confidence from more sparsely sampled data. High values of the KL divergence in 500–1800 AD correspond to periods when the posterior distribution differs markedly from the prior and here takes the form of a tightly focused intensity density. It is worth noting, however, that high KL does not necessarily mean that the posterior better describes the true evolution, for in this example the deviation of 1000 AD is contained within this high KL period.

Fig. 5 shows similar diagnostics but for the Synt-Hawaii data set. It is notable that despite the poor data coverage at early times, the method generally returns a good recovery of the intensity evolution. The asymmetric distribution of data (sparse at earlier times, dense at later times) means that the posterior distribution (shown for example either by the 95 per cent credible intervals or the intensity density) becomes very focused post 500 AD and has a high KL divergence. As with the previous synthetic data set, the posterior distribution does not follow the true evolution exactly: there are some discrepancies during short-period oscillations of the true evolution. Here, these are at 1400 BC, 800 AD and 1700 AD, and are caused by either poor sampling or because there is insufficient evidence to exclude a linear fit. Also of note is that, for the discrepancy around 1400 BC, the uncertainty bounds are relatively narrow compared with the Hawaiian data set (Fig. 13) despite being sampled at the same times and with the same assumed errors on age and intensity. This is because the synthetic data happen to be consistent with a single linear segment (which in our framework will always be preferred if the data allow), whereas for the Hawaiian data set any linear evolution fits poorly. Although the absence of the true solution feature within the 95 per cent credible intervals may be viewed as a limitation of the method, it is really a reflection of the fact that the non-linear

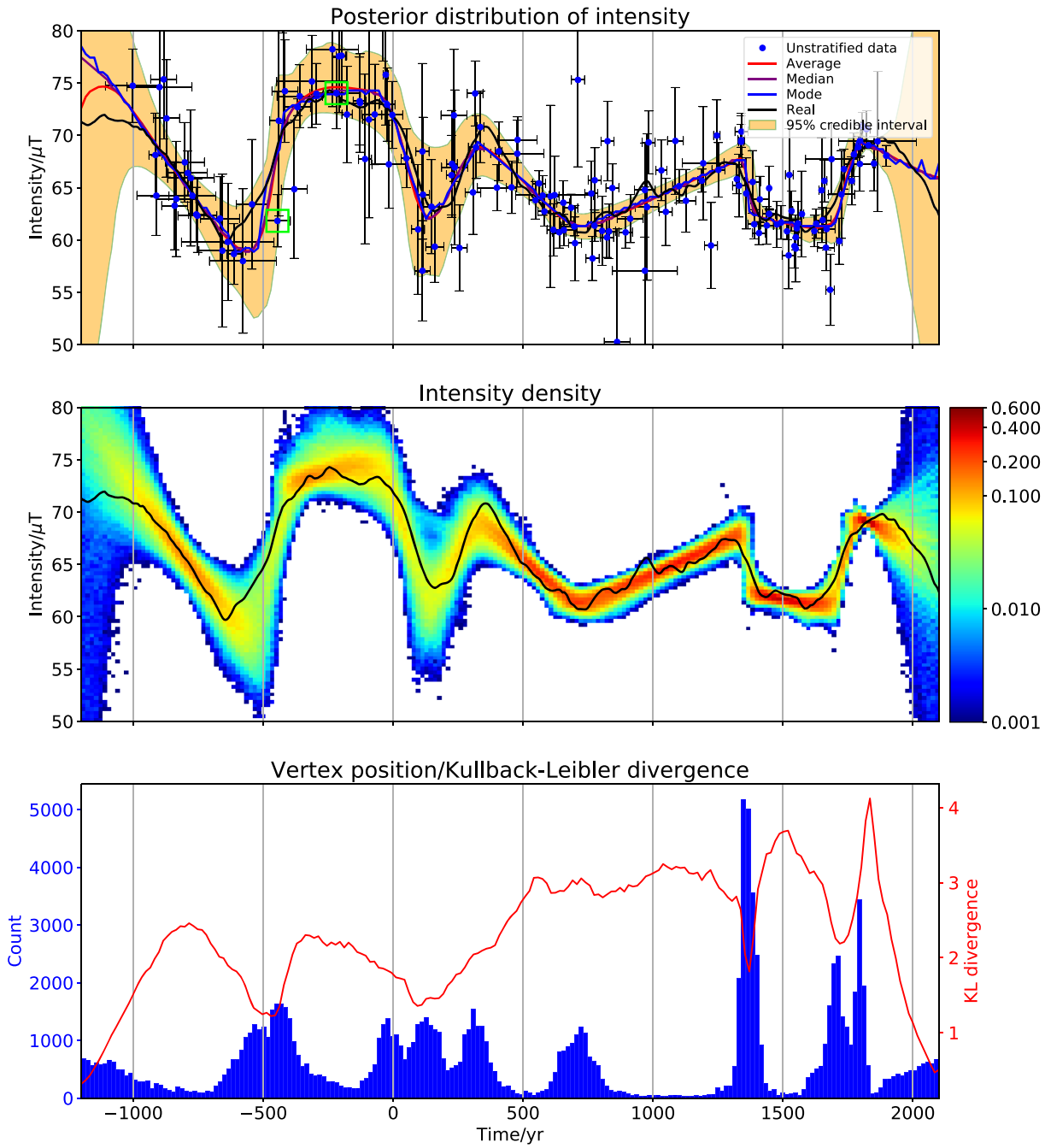

Figure 4. Recovery of the intensity variation for the Synt-Paris700 data set. Top panel: the posterior distribution depicted by the average, median, and modal curves with 95 per cent credible intervals shown in filled orange; the true (synthetic) variation is shown in black. The data highlighted in green boxes are referenced in Fig. 6. Middle panel: density plot of the posterior distribution, overlaid by the true variation (thin black line). Bottom panel: combined plot of a histogram of the vertex position showing the most likely ages for change in linear slope, with the Kullback–Leibler divergence in red (right-axis).

feature is not sampled at all so it is unsurprising that it is not present in the posterior.

Overall, for both data sets, the method recovers the intensity evolution of the CE-model very well. It is important to note that, by design, the method fits a posterior distribution of minimum curvature (being based on piecewise-linear segments). Thus any oscillatory behaviour in the posterior necessarily must be required from the data, and cannot be an artefact of the model. Conversely, minor and/or rapid oscillations of the true evolution which may either be poorly temporally sampled or lie within the error tolerance of a best-fitting linear-interpolation will not be present in the AH-RJMCMC posterior.

4.2 Marginal age distributions

Since the AH-RJMCMC method samples the joint distribution of the model space, we can derive the (joint) posterior distribution of

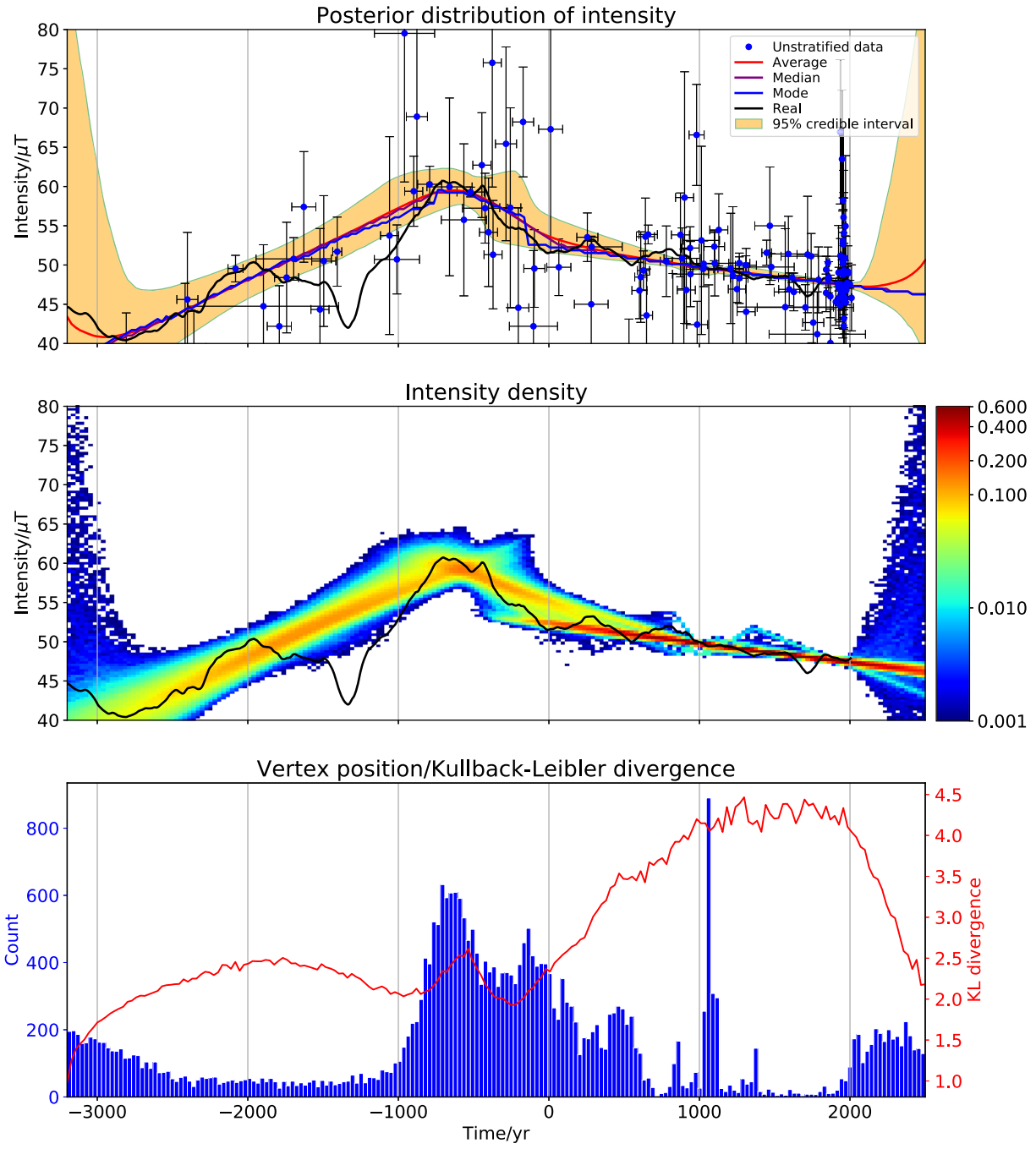

Figure 5. Recovery of the intensity variation for the Synt-Hawaii data set. Top panel: the posterior distribution depicted by the average, median, and modal curves with 95 per cent credible intervals shown in filled orange; the true (synthetic) variation is shown in black. Middle panel: density plot of the posterior distribution, overlaid by the true variation (thin black line). Bottom panel: combined plot of a histogram of the vertex position showing the most likely ages for change in linear slope, with the Kullback–Leibler divergence in red (right-axis).

any single variable, or subset of variables, by marginalization. Of particular interest is to examine the joint posterior distribution of the age and intensity of any given datum, which is simply achieved by collecting 'on the fly' the age a_j (specified within each model) and its associated intensity, $g(a_j)$, interpolated from the vertex information of the model.

Fig. 6 shows two such joint distributions for the two data 119 (left) and 154 (right) of the Synt-Paris700 data set that were chosen to represent two extremes; these data are highlighted with green boxes in Fig. 4. The joint distribution is shown as green hexagons [dark(light) shading means higher(lower) values]. The marginal distributions of both age and intensity for each datum are shown on the top and right axes; for reference, the uniform prior distribution on age, and the assumed normally distribution of intensity, are shown in transparent orange. The noisy "observed" datapoint is shown in purple, while the true values are shown by the red triangle.

On the left the marginal posterior age distribution is comparable to the prior, so the Synt-Paris700 data offers no new information about the datum's age beyond what is already assumed in the prior.

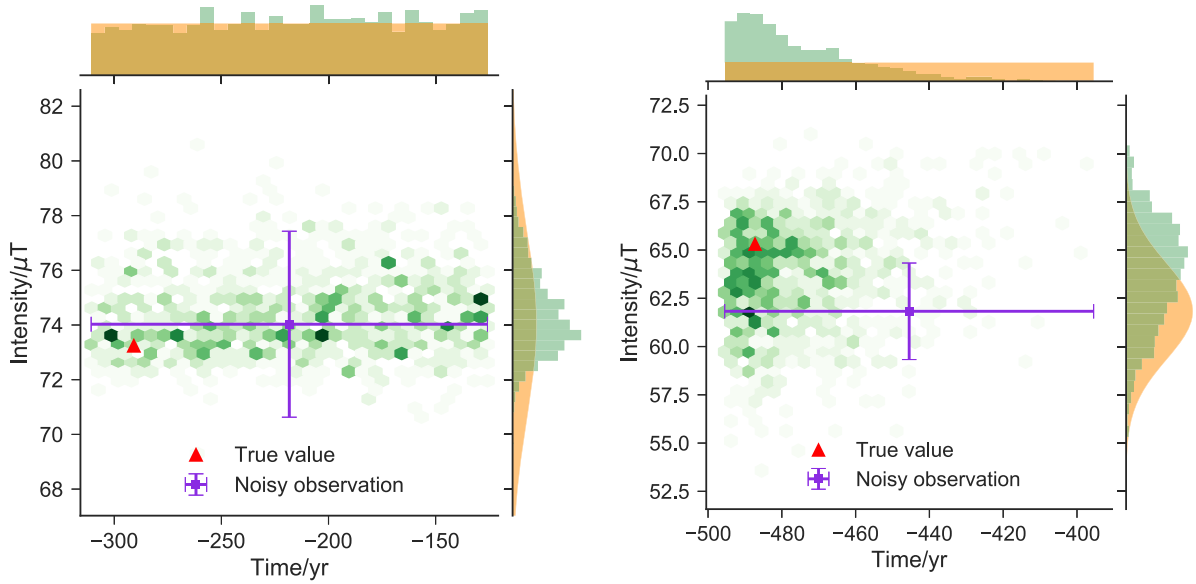

Figure 6. Joint posterior probability distribution of the age and intensity for the two data of index 119 and 154 of the Synt-Paris700 data set, as highlighted with green boxes in Fig. 4. Green hexagonal bins show the posterior joint probability distribution, while the marginal distributions for age and intensity are shown above and to the right (in green); semi-transparent orange shows the prior on the data age, and also the assumed normal distribution of intensity. The true age and intensity of the data is shown in red while the synthetic 'observation' is shown by the purple triangles (with error bars).

Although the posterior distribution of intensity is centred on approximately the same value as the datum, it has a much smaller standard deviation.

By contrast, on the right the marginal posterior age distribution is heavily skewed, the datum being much more likely to have an age at the earlier end of the range [400, 500] BC. Indeed, the true value falls close to the peak of the joint distribution, at around 500 BC. Thus other data within the data set add significant constraints on the likely age of this particular datum because we fit the data set as a whole. In this example, the distribution of posterior intensity is comparable to that of the observed datum but shifted upwards in value.

4.3 Model consistency

In this section, we briefly address two aspects of model consistency. First is to check that our diagnostics of the posterior distribution (average, mode, median) do actually reflect the properties of the posterior distribution itself. Fig. 7 shows 1000 individual models that are spaced equally along the Markov chain (red lines) for the Synt-Paris700 example, plotted alongside the average model (black line) that is comparable to the other diagnostics (Fig. 4). It is apparent that the average variation does indeed follow the evolution of the ensemble.

The second issue is more subtle. In our method, we have accommodated the imprecise knowledge of the archeomagnetic data ages by including them as hyper-parameters, $\mathbf{a}$, in our model vector $\mathbf{m} = [\mathbf{f}, k, \mathbf{a}]^T$. For our focus on archeomagnetism, this is a natural way to proceed because the marginalized posterior distributions of the data ages (which are of significant scientific value) are straightforward to compute. However, it is worth noting that this is not the only way of setting up the model. For example, rather than including the data ages in the model vector and using a likelihood based on a weighted misfit in intensity, we could swap these around: an alternative would have been to adopt the model vector $\tilde{\mathbf{m}} = [\mathbf{f}, k, F]^T$ where $F = (\mathcal{F}_1, \mathcal{F}_2, \dots)$ are the data intensities, and use a likelihood based on the misfit of the uniformly distributed data ages. The fact that these two model setups are different is a manifestation of the asymmetry of the way in which the data is handled. Ideally, the ages and intensities of the archeomagnetic data should be handled in the same way (i.e. symmetrically), reflecting the fact that there is no objective reason to treat errors in age any differently than those in intensity. Of course, because the uncertainties in the ages and intensities do not necessarily obey the same distribution type, we cannot expect the implementation of the algorithm to be symmetric.

Sambridge (2016) describes an 'integrated likelihood' method, handling 2-D data with known error distributions symmetrically by a piecewise-linear MCMC algorithm, in which the uncertainties in age are integrated out. He considers only the case where both variables are described by a joint normal distribution; in Appendix C we include details of the mathematical extension of his methodology to our focus on uniformly distributed ages and normally distributed intensities.

Using the integrated likelihood method has the advantages that (i) it treats uncertainty in both intensity and age in a symmetric way, and (ii) the model vector $\mathbf{m}_{\text{IL}} = [\mathbf{f}, k]^T$ does not contain the data ages: therefore the model space that must be sampled has many fewer dimensions and convergence (in terms of the length of the Markov chain) may be more rapid although each forward model evaluation may be slower because of the additional marginalization over age uncertainty.

Fig. 8 compares the converged posterior distributions from the age-hyperparameter and integrated-likelihood methods, using Markov chains of length 10^6 and 10^5, respectively beyond their burn-in period of 5×10^4. The fact that the methods closely agree gives us significant confidence in our methodology; it is apparent that the asymmetry in our method does not affect the final result. Computationally, for a given chain length, the integrated likelihood method is about a factor of 10 slower than the age-hyperparameter model. However, in this example this is balanced by requiring a Markov chain about 10 times shorter and so both examples shown

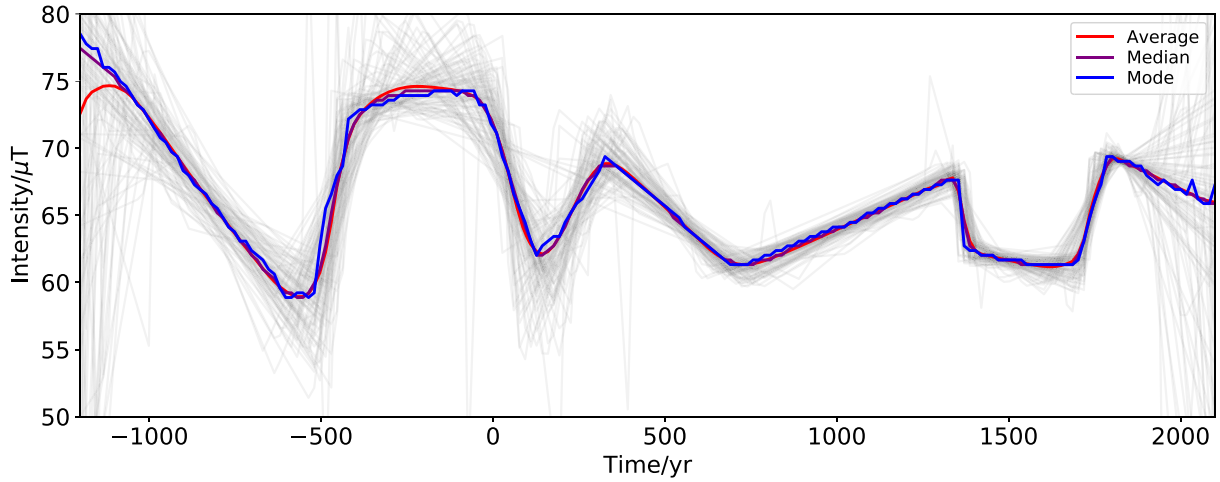

Figure 7. An ensemble of 1000 individual models (equally spaced along the Markov chain) in grey, compared to the ensemble statistics, for the AH-RJMCMC method applied to the Synt-Paris700 data set.

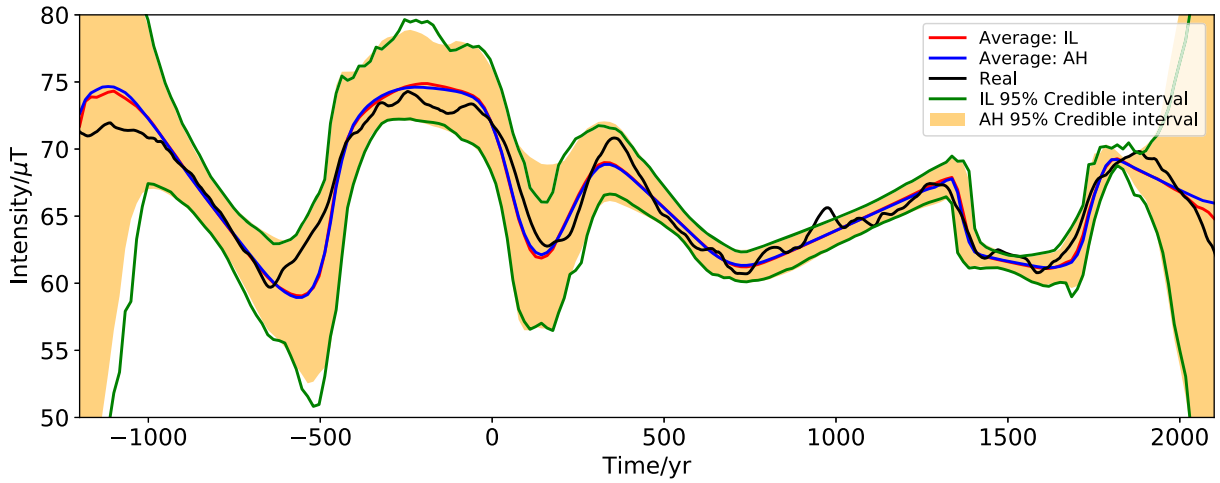

Figure 8. A comparison of the posterior distribution of intensity for the integrated likelihood IL-RJMCMC and the age hyper-parameter AH-RJMCMC methods based on the Synt-Paris700 data set. The 95 per cent credible intervals for the AH-RJMCMC method are shown in filled orange, while the green lines show the same credible interval for IL; the ensemble averages are shown in red/blue. The underlying real (synthetic) variation is shown in black.

above take about the same amount of time to run: approximately 5 seconds on a modern desktop computer.

5 APPLICATION TO THE PARIS700 DATA SET

We now apply the age-hyperparameter methodology to the Paris700 data set; the key results are shown in Fig. 9. As for synt-Paris700, the average, median and mode follow each other closely, and post 0 AD have a focused intensity density giving confidence in the posterior. Other diagnostic plots are given in Appendix D, which includes confirmation of the assumption of normally distributed intensity errors. The time variation after 500 AD is broadly similar to Genevey *et al.* (2016), which is not surprising given the dense-sampling of the secular variation and the small scatter.

There are several data points which do not fit the general trend, which include a low intensity value (about 40 μT) around 100 BC. To gauge the effect of outlying results, Fig. 10 shows the effect of removing data (including the result around 100 BC) whose mean intensity lies outside the 95 per cent credible interval (evaluated at each datum's mid-point age) and re-running the AH-RJMCMC algorithm. With the outliers removed, at ages 200–300 BC, the posterior distribution differs markedly: the temporary dip in the intensity is now removed and values extend no lower than about 50 μT during this period.

One important feature of the posterior shown in Fig. 9 is a periodic signal contained within the envelope of relatively narrow 95 per cent credible intervals, most evident post 500 AD. Importantly, as noted in Section 4, because our model favours minimal curvature, the existence of such a signal cannot be attributed to anything other than the influence of the data themselves. This signal was identified in a similar data set by Genevey *et al.* (2016) who showed, by analysing the power spectral density (PSD) of their best-fitting curve, maximum power at approximately a 256-yr period. Here, we apply a similar procedure but to the entire posterior distribution to produce a probabilistic assessment of frequency content.

From the ensemble of 10^6 models that we generate from the AH-RJMCMC method, we take 1000 models (equally spaced along the Markov chain) and for each we perform the following analysis in order to determine the PSD as a function of period:

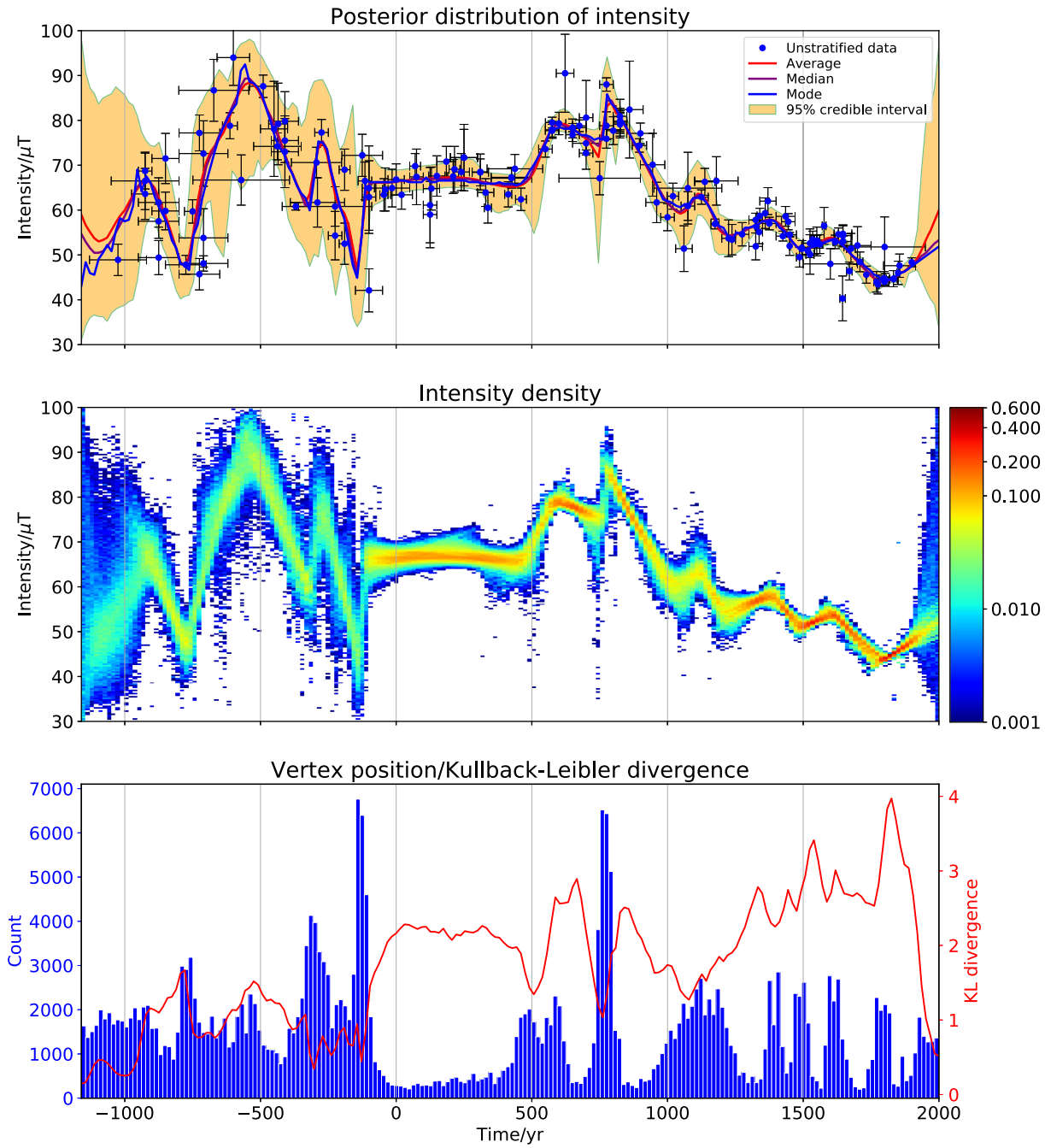

Figure 9. The time-dependence of the posterior intensity distribution for the Paris700 data set as determined by the AH-RJMCMC method. Top panel: the average, median, and modal curves with 95 per cent credible intervals shown in filled orange. Middle panel: density plot of the posterior distribution. Bottom panel: combined plot of a histogram of the vertex position showing the most likely ages for change in linear slope, with the KL divergence in red (right-axis).

(i) Isolation of the time series for the window 500–1900 AD (note that the models themselves are defined over a broader range, but this age range is strongly constrained by the data).

(ii) Removal of the linear trend.

(iii) Application of a forward-backward Butterworth filter using cut-off frequencies of 400 yr^{-1} and 40 yr^{-1} to isolate the periods of interest.

(iv) Computation of the PSD using the method of Welch (1967).

Fig. 11(left-hand panel) shows the PSD as a function of period for each ensemble member (thin black line), with the PSD for the median, modal and average model shown in colour. The average of the set of 1000 individual PSD curves is shown in orange. We note that the average PSD agrees well with the PSD of the average (blue), a result that is not expected to be true in general unless all curves share a similar frequency content. There is a strong indication of a dominant period of 260–280 yr. Fig. 11(right-hand panel) shows a normalized histogram of the period corresponding to the maximum PSD for each of the 1000 representative models, with the best-fitting normal distribution shown in black (mean 269 yr, standard deviation 9 yr) that fits the general behaviour very well. We note that our preferred period of 269 yr is within 1.5 standard deviations (and thereby consistent with) the approximate 256 yr signal found

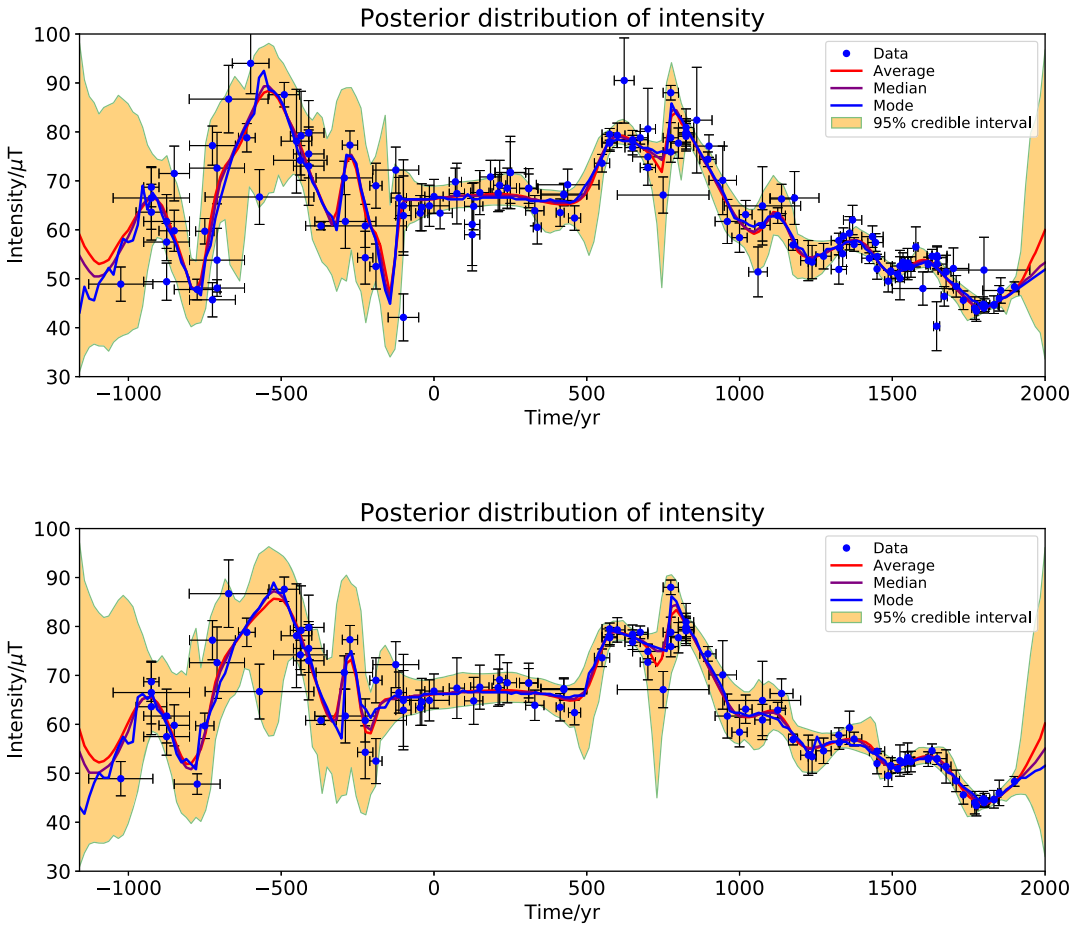

Figure 10. Comparison of the posterior distribution of intensity variation obtained with the Paris700 data set (top panel) and a refined data set with outliers removed (bottom panel).

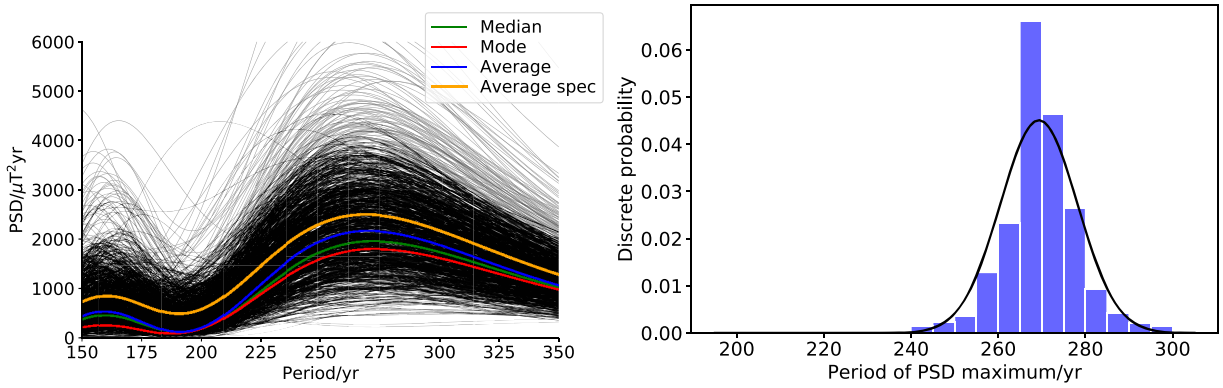

Figure 11. Power spectral density as a function of period for the AH-RJMCMC method applied to the Paris700 data set, restricted to the period 500–1900 AD. Left-hand panel: shown in thin black is the PSD for each of the 1000 representative ensemble members; the PSD for the median, mode and average are shown as green, red and blue, respectively, while orange shows the average of the 1000 individual PSD curves. Right-hand panel: a normalized histogram of the period corresponding to the maximum PSD, with the best-fitting normal distribution (using only periods 200–300 yr) shown in black (mean 269 yr, standard deviation 9 yr).

by Genevey *et al.* (2016). It is worth mentioning that, for their computations, Genevey *et al.* (2016) subtracted a non-linear trend (that differs from our use of a linear trend), which may explain the small difference in the period. We note that the shape of the histogram is converged using 1000 models, and that the plot of PSD is another example of projecting the model **m** onto a different space.

It is also of interest to compare our method of handling errors both in intensity and age with more common methods of treating the ages as known and increasing the error budget for the intensity. Fig. 12 shows our AH-RJMCMC method applied to the Paris700 data set in the top panel, compared with the same method applied to three other variants of the same data set: assumed age errors of zero (with the age being defined by the mid-point age of the interval), age errors

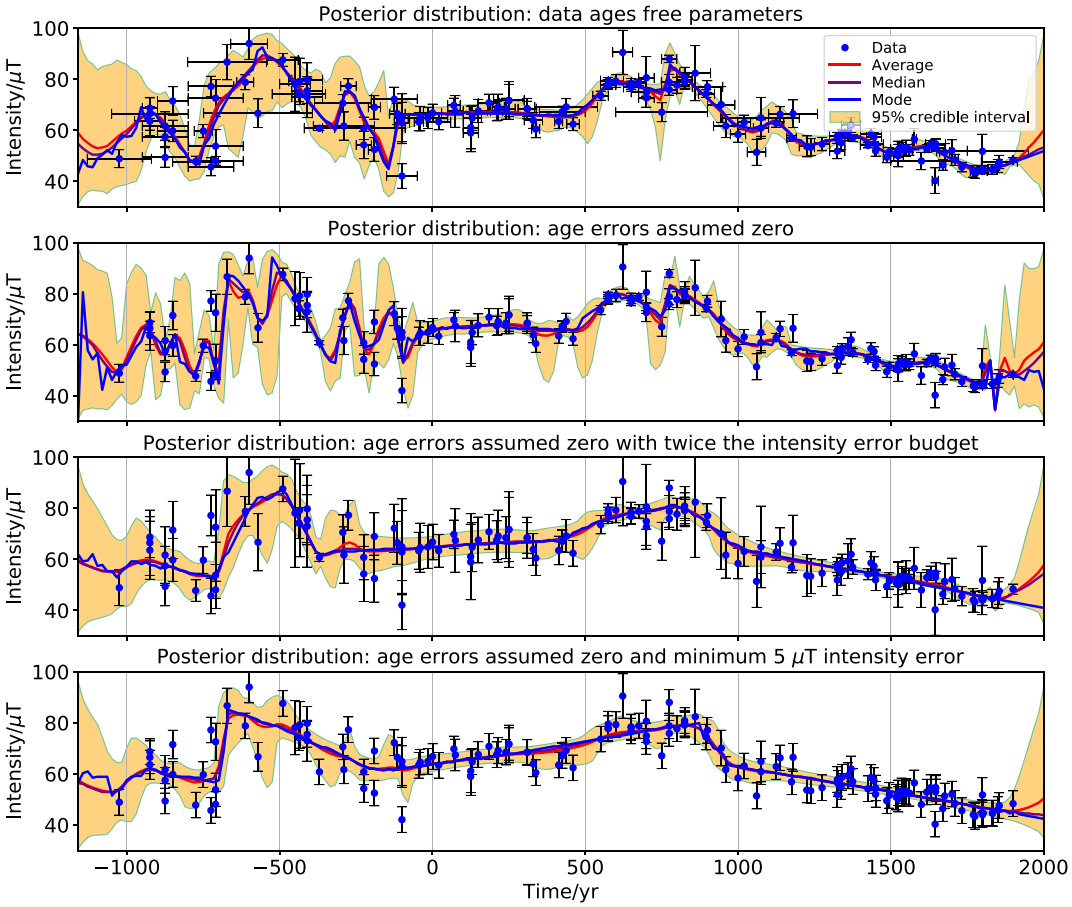

Figure 12. A comparison of posterior distributions of intensity using variants of the Paris700 database. From top to bottom: a reproduction of our AH-RJMCMC method applied to Paris700 (Section 5); the same as above but with assuming the ages are known to be exactly their midpoint values; as above with twice the intensity error budget; assuming known ages with a minimum intensity error of 5 μT.

of zero with twice the error budget on intensities, and age errors of zero with a minimum intensity error of 5 μT (as assigned by Korte & Constable (2011) to all intensity data for the construction of global field models). There are two key aspects of note. First, comparing the top two panels of the figure shows that assuming fixed ages causes rapid changes in the posterior intensity (which are absent when the ages are allowed to be free parameters), particularly for relatively sparsely sampled epochs (here clearly identified pre 0 AD). Second, either in doubling the intensity error budget or setting a minimum threshold of 5 μT results in the smoothing out of most fluctuations: of importance is that the periodic signal previously identified post 500 AD is absent in the bottom two panels. A similar result would likely hold were we to draw fixed ages from the given uniform distribution as in a bootstrap method, rather than using their midpoint values. This corroborates other studies (e.g. Genevey *et al.* 2016; Hellio & Gillet 2018) that have highlighted the mismatch in time variability between regional and global models, because of the necessity to include additional smoothing for the global case.

6 APPLICATION TO THE HAWAII DATA SET

We now turn our attention to the data set of the Hawaiian area extracted from Geomagia.v3 (Brown *et al.* 2015) that is much more sparse in time, and which has data ages and intensities both normally distributed. We adopt an extended prior for the intensity to be uniform in the range 10–100 μT because of the greater range of intensities in the data, and also increased the maximum number of internal vertices to 100. The increased scatter in the data (compared with Paris700) means that a converged posterior requires a longer Markov-chain. The results are summarized in Fig. 13, with a chain length (before thinning) of 5×10^6 after burn-in. Other diagnostic plots are given in Appendix E, which includes confirmation of the assumption of normally distributed intensity errors.

As with previous models, the average, median and mode have a very similar time dependence. However, as expected, the more scattered data set gives less confidence in the posterior distribution, reflected in its much broader credible intervals and less focused density. Where the data is particularly sparse or scattered, the 95 per cent credible interval range is about 30–80 μT (for example, at very early times), similar to the range of the prior on the intensity of the model vertices. Apart from for the very recent data dated to the 19th and 20th centuries thanks to direct observations, the width of the credible interval range does not extend much below 20 μT. This should be contrasted with the posterior from the Paris700 data set, which has a typical credible interval width of just 5 μT. In terms of KL divergence, typical values for the Hawaiian data set are between 0 and 1 for much of the age window, which is much less than the comparable diagnostic of the Paris700 data set (of about 2, for the period post 0 AD). Thus compared with Paris700, the Hawaiian data do not add as much information to the prior.

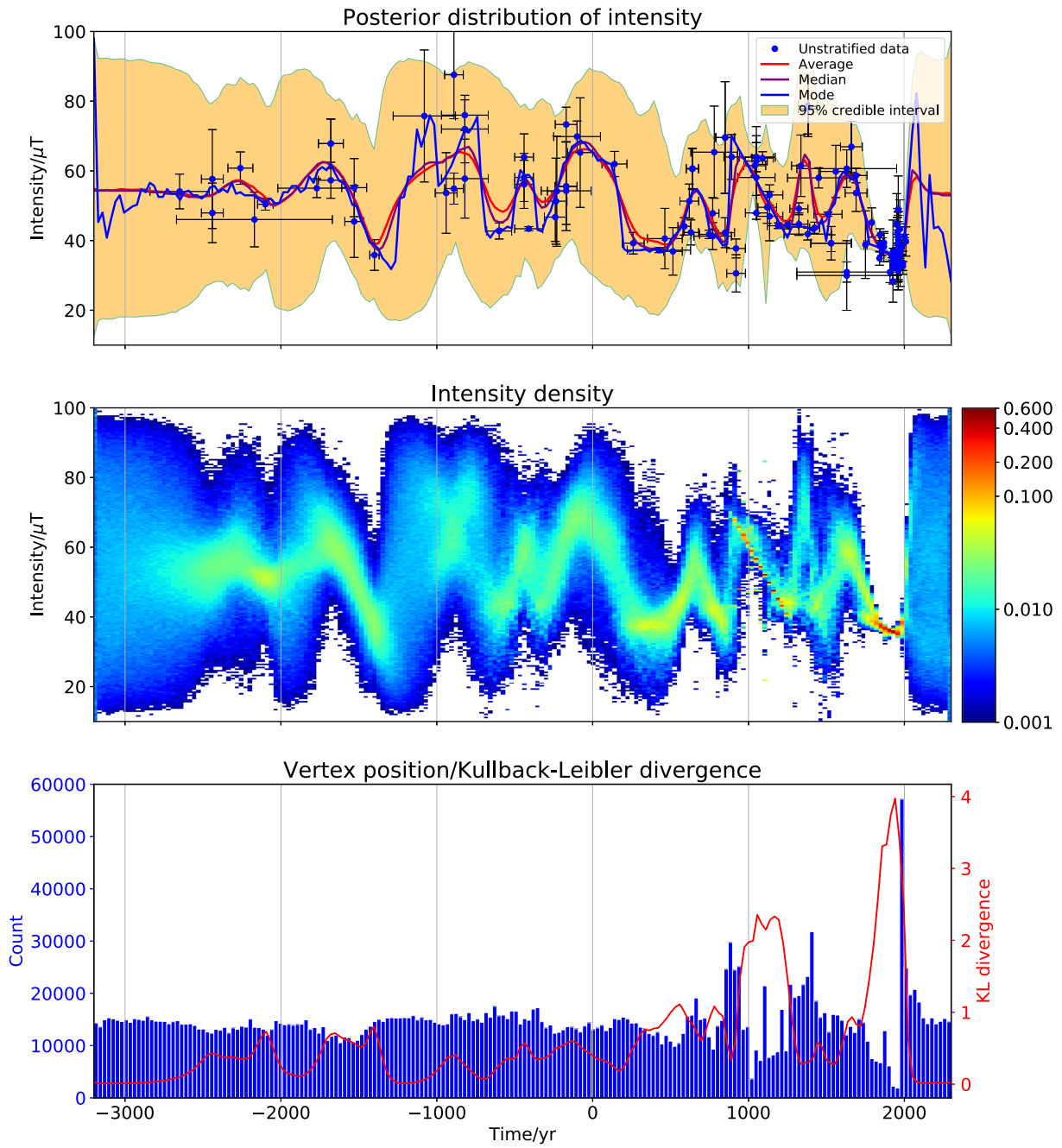

Figure 13. The posterior intensity variation for the Hawaii data set. Top panel: the average, median, and modal curves with 95 per cent credible intervals shown in orange. Middle panel: density plot of the posterior distribution. Bottom panel: combined plot of a histogram of the vertex position showing the most likely ages for change in linear slope, with the KL divergence in red (right-axis).

One aspect of interest is the quasi-periodic behaviour identifiable in the top panel of Fig. 13 post 500 AD. Fig. 14 shows the same power spectral density analysis of Section 5 applied here, for the period 500–2000 AD. The scatter of the lines in the left-hand plot is a reflection of the poorly constrained ensemble of model curves (owing to the sparse data distribution), and in this case the average of the spectra is not similar to the spectra of the average model. Nevertheless, the spectra of the average, modal and median diagnostics for the posterior all have a maximum at around 325 yr. The right plot shows a histogram of the periods of the PSD maxima, with a very broad but bimodal structure with peaks loosely defined at around 210 and 330 yr. Thus the sparsity of the data set has an associated very broad spread in the possible models that fit the data, and this analysis does not identify any specific period. Furthermore, we note that the Hawaiian curves established by de Groot *et al.* (2013) and Tema *et al.* (2017) using different and more severe selection criteria do not show the same quasi-periodic features (see discussion).

7 APPLICATION TO THE LUBECK-PARIS700 DATA SET

Our final application of the AH-RJMCMC methodology is to the Lübeck-Paris700 data set that contains data with ordered age constraints. Fig. 15 shows a summary of the posterior intensity varia-

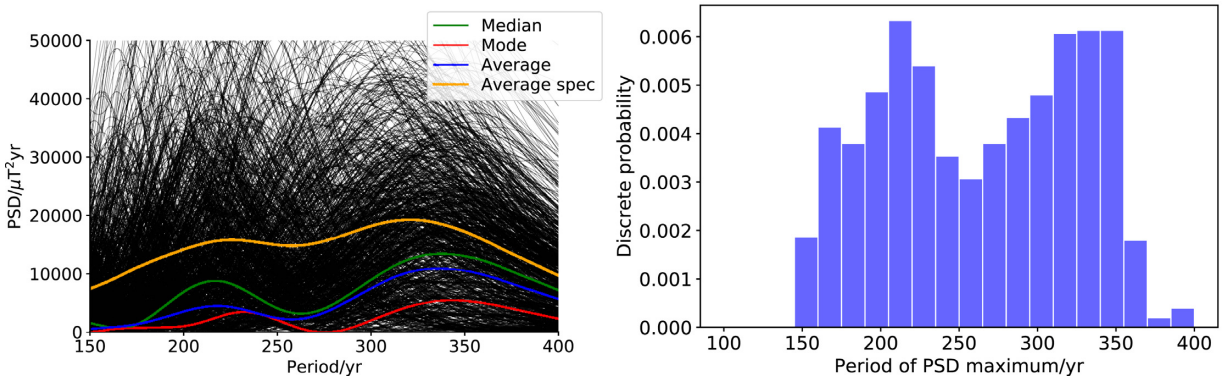

Figure 14. Power spectral density as a function of period for the AH-RJMCMC method applied to the Hawaii data set, restricted to the period 500–2000 AD. Left-hand panel: the PSD for each of the 1000 representative ensemble members are shown by thin black lines; the PSD for the median, mode and average are shown as green, red and blue respectively, while orange shows the average of the 1000 individual PSD curves. Right-hand panel: a normalized histogram of the period corresponding to the maximum PSD.

tion with time. For the majority of the time window, the posterior is highly focused and well constrained by the data (with a KL divergence of at least 2).

Of primary interest is the inclusion of stratified data, in particular, the 10 archeomagnetic data all of age 1665 ± 85 AD but which obey strict ordering constraints. Although each of these data are consistent with (within error tolerance) the quasi-periodic fit to the unstratified data (Fig. 16), the posterior age distributions will be highly skewed as, for example, the datum of least intensity is most likely to have an age at the extreme upper end of the interval of [1580,1750]. Fig. 16(a) shows the stratified sequence of posterior ages for the 10 data each coloured differently, with the nominal midpoint age of 1665 AD marked as the vertical dashed line. The posterior ages differ markedly from the uniform prior (shown by the flat coloured rectangles). The posterior estimates of the age and intensity of each datum (given by the average of the relevant marginal posterior distributions) are shown in Fig. 16(b) using the same colour scheme. Each of the original data are marked at their midpoint age of 1665 AD (with error bars) with the posterior estimate marked as a solid square of the same colour. Thus, for example, the topmost datum (shown as blue) when treated as part of the entire data set, has an age which is most likely to be shifted earlier in time than 1665 AD and to a smaller value of intensity than its original estimated value. Such shifts are also shown in Fig. 17 that shows the joint age-intensity posterior probability distribution for the two end-members (red, sky-blue) of Fig. 16(a). On the left we see that the posterior age is on the lower-most extreme of the prior age interval, and has an associated posterior intensity which is more focused than the distribution describing the intensity error but centred on a similar value. On the right, the posterior age is at the extreme upper end of the prior interval, and has an intensity distribution which has a greater mean and a much smaller spread than the laboratory-determined intensity error.

It is worth remarking that, although all the data have a highest probability that is shifted relative to their nominal original age and intensity estimates, the stratified data are not passive in the process of model determination, but rather themselves contribute to the overall posterior distribution. This process therefore differs fundamentally from using a given secular variation curve as a means to estimate age (Le Goff *et al.* 2002; Pavón-Carrasco *et al.* 2011). Although applied here to stratified archeological data, the same technique could also be used to improve knowledge of the dating of sedimentary sequences, probably with interesting consequences for global field modelling.

8 DISCUSSION AND COMPARISON TO OTHER METHODS

In this paper, we have presented a transdimensional Bayesian method that can infer the time dependence of intensity evolution; it has four key elements. First, we co-estimate marginal distributions for data ages as well by treating these as unknown model parameters. Second, by averaging over a large ensemble of model realizations with differing parameter values and complexity, diagnostics (such as the average and credible intervals) of the posterior distribution are temporally smooth, yet formally unregularized. Third, because our method is based on linear interpolation, oscillations in the posterior distribution are only present when absolutely required by the data, and thus our method presents a powerful tool to probe oscillatory behaviour. Finally, our method is very fast (taking a matter of seconds) and the code is publicly available.

With these points in mind, it is now instructive to compare the results of our method to those obtained from two well-known and independently developed approaches of determining archeomagnetic intensity evolution. The first we consider is the sliding window method of Le Goff *et al.* (2002), in which the intensity evolution is defined within windows of variable duration (M. Le Goff, personal communication). The second is the Bayesian scheme developed by Lanos (Hervé & Lanos 2017; Hervé *et al.* 2017; Lanos & Philippe 2017). Like our scheme, it co-estimates data ages using a similar Monte-Carlo approach, but assumes a cubic spline (rather than linear) dependence of intensity with time, and achieves a smooth evolution by regularizing using a 'shrinkage' parameter (Tema *et al.* 2017) rather than (transdimensional) model averaging. The differences between the results from all these methods reflect not only the different methodologies, but also specific user choices. For Bayesian methods (here, our AH-RJMCMC method and that of Lanos and colleagues) the posterior depends on the chosen prior distributions, although for our model this dependence can be quantified (see the Kullback Leibler divergence described in Section 3.5). For the sliding window method, a choice needs to be made on the factors controlling the duration of the window.

Fig. 18(a) shows our AH-RJMCMC method compared to the sliding window scheme for the Paris700 data set. Although the average curves largely agree, there are distinct differences, particularly pre 0 AD and around 750 AD where either the data are relatively scattered or are outlying to the general trend. In particular, the single data point of 40 μT around 200–300 BC causes a large oscillation in our AH-RJMCMC average, yet the sliding window results are

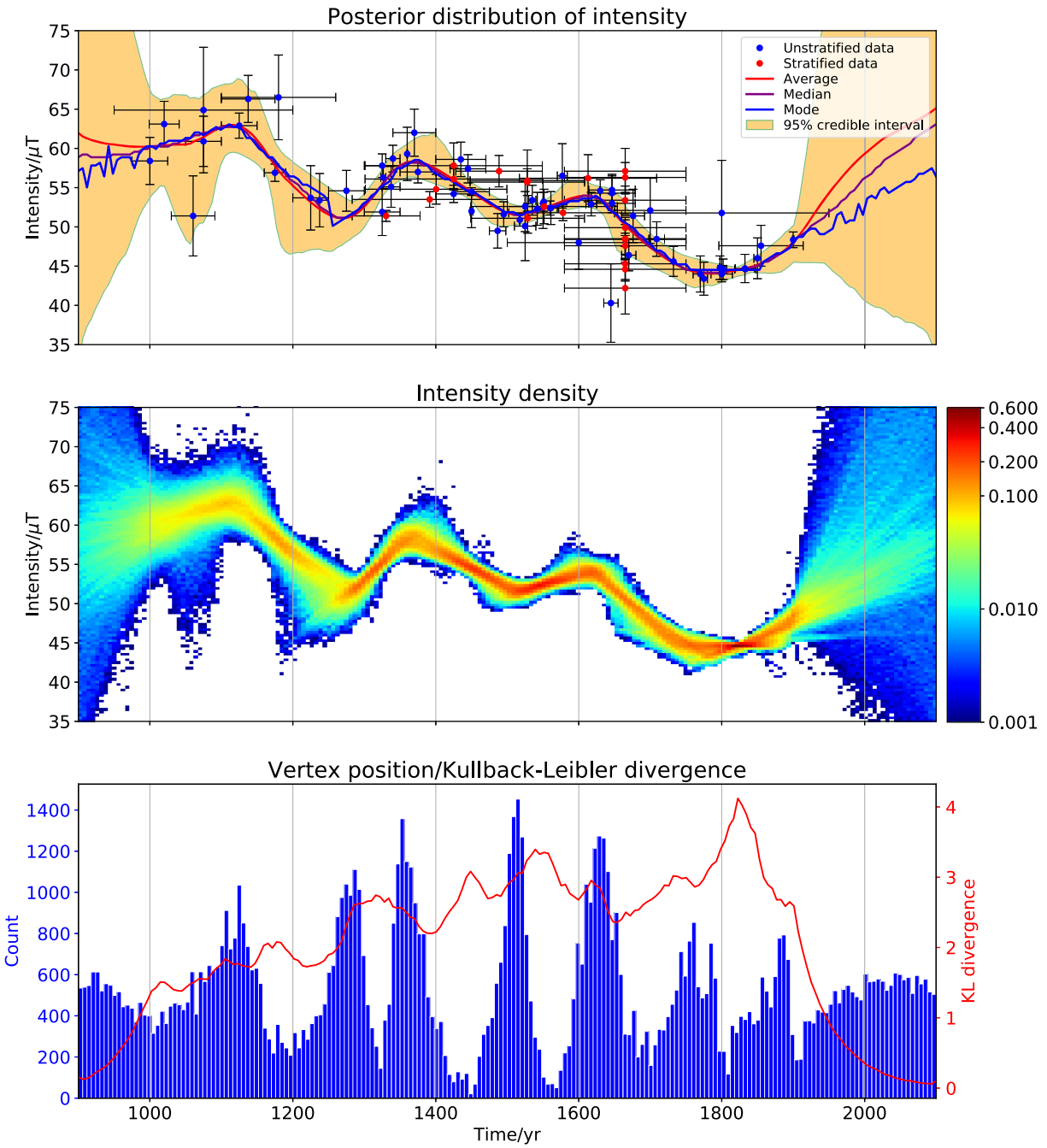

Figure 15. The posterior intensity variation for the Lübeck-Paris700 data set. Top panel: the average, median and modal curves with 95 per cent credible intervals shown in orange. Middle panel: density plot of the posterior distribution. Bottom panel: combined plot of a histogram of the vertex position showing the most likely ages for change in linear slope, with the KL divergence in red (right-axis).

more similar to our AH-RJMCMC method where we have rejected outliers (see Fig. 10) where the oscillation is absent. In both methods, the uncertainties are mainly smaller during times of dense data, but larger during times of data sparsity.

Fig. 18(b) shows our AH-RJMCMC method compared to the Bayesian method of Lanos using a French data set described in Hervé *et al.* (2017) (which is almost identical to our Paris700 data set within the common time interval). During times of dense data coverage (50 BC onwards) the two methods agree in both their average and credible intervals. However, pre 50 BC there are some significant differences in the average models, in particular at 600 BC, 300 BC and 100 BC, probably due to the rejection of the nearby data points in the method of Lanos and colleagues. Furthermore, pre 50 BC the 95 per cent credible intervals are not in agreement. In our method, periods of large uncertainty generally correspond to periods of sparse data coverage. In contrast, the curve of Hervé *et al.* (2017) appears to have a credible interval width that varies only relatively slightly in time, even during periods of sparse data (see, in particular, the period close to 1000 BC). Thus their credible interval width does not appear to have a simple link with data sparsity. Similar conclusions can be drawn when comparing the two Bayesian methods but on the Hawaiian data set of Tema *et al.*

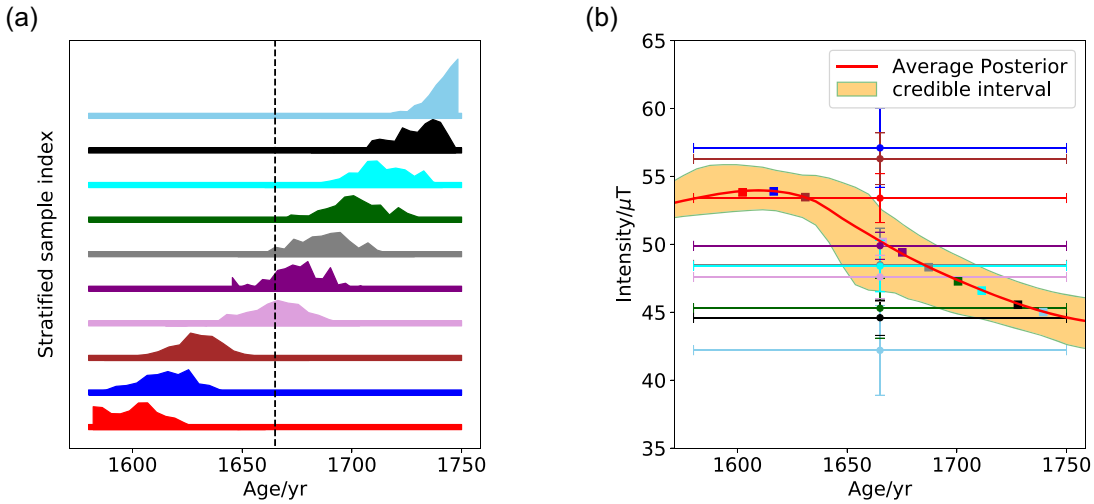

Figure 16. (a) Posterior age distribution for the 10 data with identical midpoint age but which obey stratification constraints; the associated prior distributions are shown as the coloured rectangles. (b) For each of these 10 data, a comparison of the midpoint age estimate (coloured circle with error bars) with their corresponding posterior (average) estimates (coloured squares). Both plots use the same colour scheme.

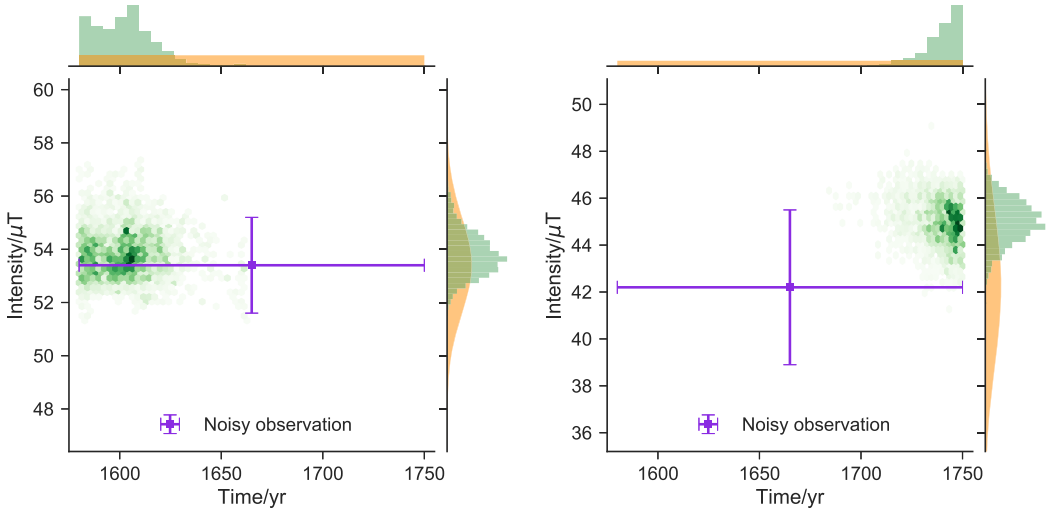

Figure 17. Joint posterior probability distribution with marginals of the data age and intensity for the two end-member cases of the 10 stratified data all with midpoint age 1665 AD. The left-hand plot shows the distribution of the datum plotted in red in Fig. 16, while the right-hand plot shows the datum plotted in sky blue in Fig. 16. Green denotes the posterior, orange the prior distribution on age and the normally distributed likelihood for intensity.

(2017), shown in Fig. 19(a). In this case, comparable data gaps of 1500 yr between [4500,6000] BC and [1000, 2500] BC have very different credible interval widths: respectively, enormous (off scale) and narrow. For the same periods, our AH-RJMCMC method gives broad credible intervals whose widths are comparable with each other.

An important issue concerns treatment of data points that do not fit with the general trend: of particular note in the Paris700 data set (e.g. Fig. 18a) are those defining an intensity peak that may have occurred around 250 BC, and the single point defining an intensity low around 100 BC. Although it is not possible to know definitively whether such data represent the true geomagnetic field evolution or simply have a greater error than assumed, there are two methods by which such data can be handled. First, we can define and remove outlying data; we followed this approach in Section 5, where we pragmatically defined an outlier by assessing if the single point defined by the datum's midpoint age and intensity lay outside the 95 per cent credible interval of the posterior distribution calculated using the full data set. According to this definition, the single point having a low intensity value around 100 BC is outlying, whereas the data cluster around 250 BC are not. We remark that a small difference in age would place the datum around 100 BC inside the 95 per cent credible interval. Indeed, perhaps a more robust definition of outliers might be to proceed probabilistically, by calculating the probability, p, of any given datum (with its assumed distribution of errors in both age and intensity) falling outside the 95 per cent posterior credible interval bounds. Data for which p is greater than some given threshold value (say) 95 per cent (so that there would be only a 5 per cent probability that the true intensity and age values of the datum came from within the 95 per cent credible interval) could be labelled as outliers; such a procedure would likely result in only a few, if any, data being marked as outlying in the Paris700 data set. A second approach that proceeds within the Bayesian framework is to keep all the data but allow the data intensity errors to

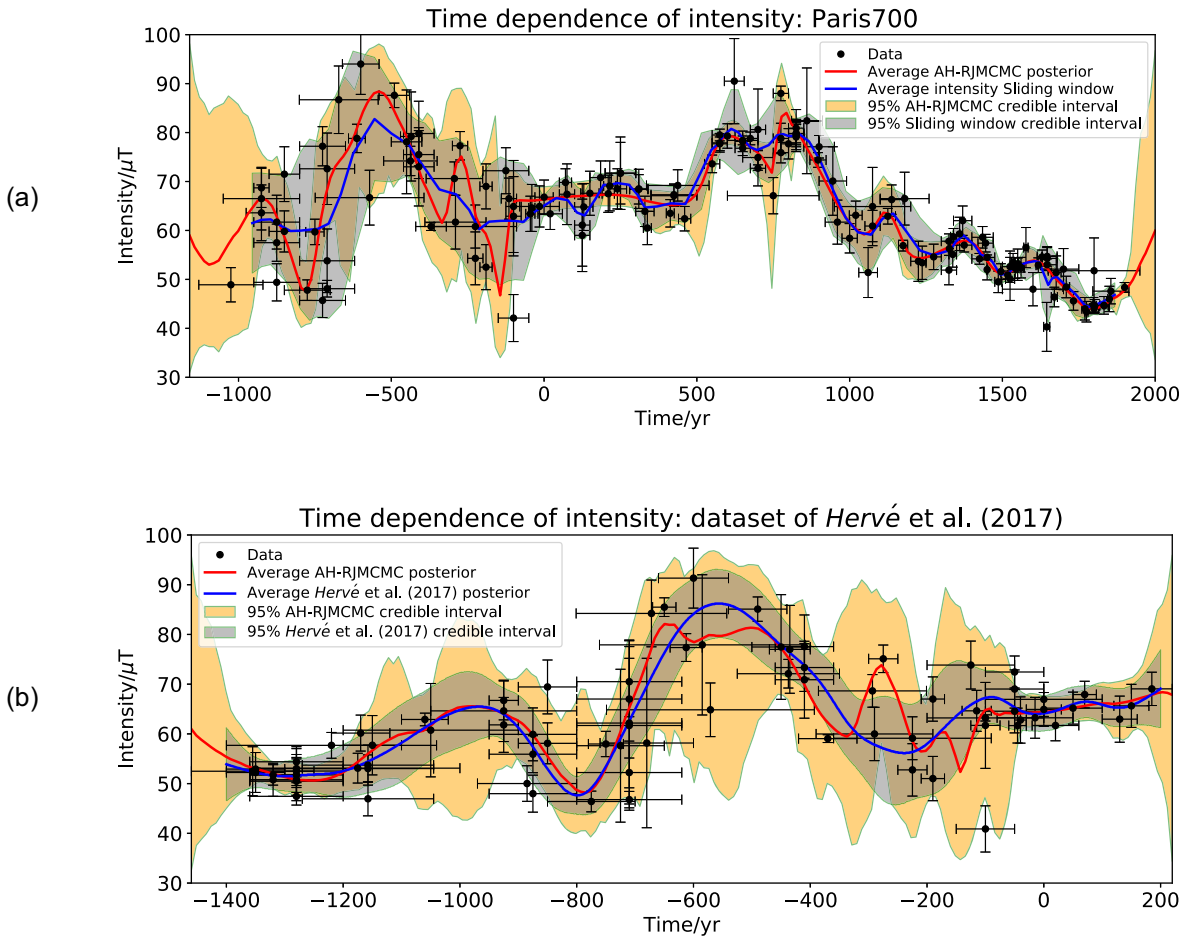

Figure 18. (a) Comparison of methods on the Paris700 data set: AH-RJMCMC compared with the sliding window method of Le Goff *et al.* (2002). (b) Comparison of our AH-RJMCMC method and the Bayesian method of Lanos and colleagues, using a French data set (Hervé *et al.* 2017).

be hyperparameters, whose posterior distribution is an output of the algorithm (e.g. Malinverno & Briggs 2004). Data that do not follow the general trend will then be associated with larger values of intensity error and therefore given less weight (e.g. Thébault & Gallet 2010; Licht *et al.* 2013).

The cluster of one or two data points around 250 BC in our study gives a local maximum in the posterior intensity. Interestingly, in both the methods of Lanos and Le Goff, this peak is totally missing, likely the consequence of inherent smoothing since the timescale of this feature is roughly the same as the one characterizing the different peaks over the past 1500 yr (see Fig. 18a and Genevey *et al.* 2013, 2016). Although more data are obviously required to better constrain its existence, there is no reason to believe that such a feature could not have occurred around 250 BC. In passing, we note that our approach, which relies on a piecewise-linear temporal interpolation, is well suited to detect rapid fluctuations of the geomagnetic field, such as archeomagnetic jerks (Gallet *et al.* 2003) or geomagnetic spikes (Shaar *et al.* 2011).

Next, we comment on the power of our method to assess periodic behaviour, which is only permitted as part of the solution where required by the data. For the Paris700 data set, the fact that our method shows a dominant periodic signal (see Section 5) of about 270 yr is therefore a strong conclusion. Indeed, it is notable that this period is in agreement with the period of about 260 yr deduced by Genevey *et al.* (2016), despite the differences between approaches (in particular, in the long-term trend assumed), adding weight to the robustness of this observation.

Conversely, our method can also be used to assess whether a data set is compatible with the absence of a periodic signal. We explore one such example of the sparse Hawaiian data set (Tema *et al.* 2017) that is similar to that presented in this paper but with more stringent selection criteria. It is worth recalling that our own data set can be viewed as a blind (though reasonable) use of a global database or a data compilation. Tema *et al.* (2017) identified four consecutive peaks in the intensity evolution from 2000 BC to 2000 AD. Fig. 19(a) compares results from our method with that of Lanos (see Fig. 6c of Tema *et al.* 2017). Although our method fits a quasi-oscillatory signal post 0 AD, we do not find evidence for the final oscillation around 1500 AD: this is because the data are actually consistent with linear behaviour and do not require an oscillation. In this respect, the curve derived from the AH-RJMCMC method appears quite similar to that proposed by de Groot *et al.* (2013) constructed using another Hawaiian data selection. Following an application of our method, Fig. 19(b) shows the mean posterior value of each data age and corresponding intensity, and shows that a linear dependence is compatible by shifting the data within their error bounds. In a parallel study, a similar assessment of testing whether data can be shifted (within error bounds) to fit an overall trend was recently performed by Korte & Constable (2018) to assess geomagnetic spikes, although in a more *ad hoc* and non-probabilistic fashion. Overall, we find that this Hawaiian data set is entirely consistent

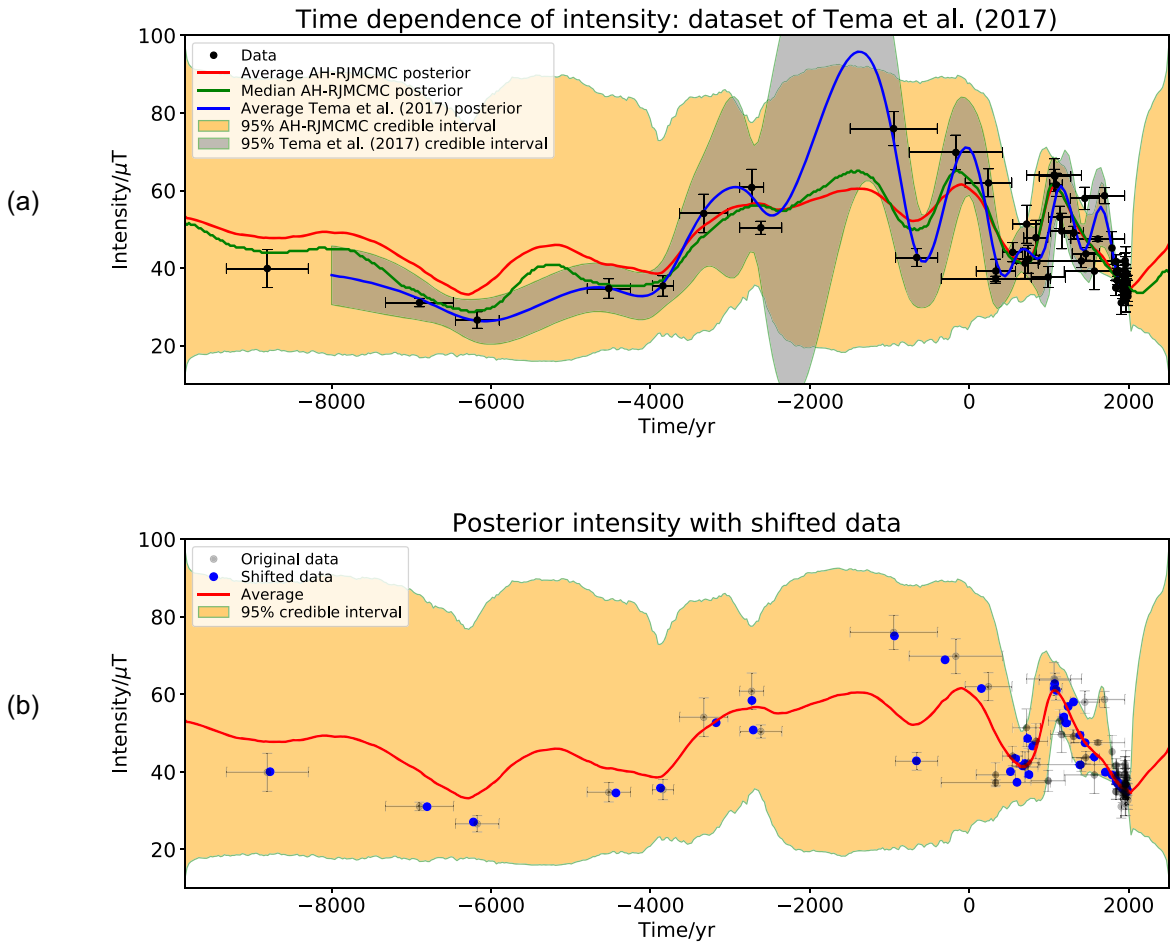

Figure 19. Comparison of Bayesian methods on the Hawaiian data set of Tema *et al.* (2017). (a) results from our AH-RJMCMC method compared with the Bayesian method of Lanos and colleagues (Tema *et al.* 2017); (b) the shift in age and intensity from their original mean values (semi-transparent black) to the mean posterior values (blue).

with a much less oscillatory signal with no obvious period (our red curve).

We end by commenting briefly on the use of our AH-RJMCMC tool for archeomagnetic dating of a datum of some given intensity. In our approach, we would add this datum (with broad uniform uncertainty on age) to the data set, and run the AH-RJMCMC model. The posterior marginal age distribution is then part of the model output. This is to be contrasted with a more standard approach, in which the datum is compared to a reference curve. Both the AH-RJMCMC method described here, and the method of Lanos and colleagues, can perform either calculation. Although our approach of 'all data together' differs from the standard method of distinguishing a reference data set from a comparative datum, it has two advantages. First is that, within a Bayesian formulation, noisy data should not be removed from the problem but instead accounted for with their large uncertainties. In our method, this is implemented by using a very broad prior distribution on the datum age. Second is that separating the problem into two steps requires an arbitrary choice. How do we decide what is too noisy to be accounted as 'reference data'? Taking everything together is an integrated approach that combines all pieces of information, and that avoids this arbitrary distinction.

ACKNOWLEDGEMENTS

PWL would like to thank Université Paris Diderot and the Institut de Physique du Globe de Paris for funding two research visits in 2015 and 2016, during which much of this work was begun. PWL was partially supported by the Natural Environment Research Council grant NE/G0140431. TB is supported by the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 716542. We acknowledge support from the INSU (PNP) programme. We are very grateful to Maxime Le Goff for running his sliding window algorithm on our data, and for helpful comments. We also thank Gabrielle Hellio and an anonymous reviewer for constructive comments on the manuscript. All figures were created in Python using Matplotlib (Hunter 2007). All authors contributed to the ideas behind the approach taken in the manuscript; YG provided the data sets and background on archeomagnetic secular variation curves; AF, TB and PL developed the methodology and wrote the code; PL generated the results and wrote the paper. All authors provided comments on the manuscript. This is IPGP Contribution no. 3967. *Code availability:* The code and data sets are available at https://github.com/plivermore/AH-RJMCMC, which has DOI:10.5281/zenodo.1442345.

REFERENCES

Aitken, M., Allsop, A., Bussell, G. & Winter, M., 1988. Determination of the intensity of the Earth's magnetic field during archaeological times: reliability of the thellier technique, *Rev. Geophys.,* **26**(1), 3–12.

Aubert, J., Finlay, C.C. & Fournier, A., 2013. Bottom-up control of geomagnetic secular variation by the Earth's inner core, *Nature,* **502,** 219–223.

Bodin, T. & Sambridge, M., 2009. Seismic tomography with the reversible jump algorithm, *Geophys. J. Int.,* **178**(3), 1411–1436.

Bodin, T., Sambridge, M., Rawlinson, N. & Arroucau, P., 2012. Transdimensional tomography with unknown data noise, *Geophys. J. Int.,* **189**(3), 1536–1556.

Brown, M.C., Donadini, F., Korte, M., Nilsson, A., Korhonen, K., Lodge, A., Lengyel, S.N. & Constable, C.G., 2015. GEOMAGIA50. v3: 1. General structure and modifications to the archeological and volcanic database, *Earth, Planets Space,* **67**(1), 83.

Chauvin, A., Garcia, Y., Lanos, P. & Laubenheimer, F., 2000. Paleointensity of the geomagnetic field recovered on archaeomagnetic sites from France, *Phys. Earth planet. Inter.,* **120**(1), 111–136.

Coe, R.S., 1967. Paleo-intensities of the Earth's magnetic field determined from tertiary and quaternary rocks, *J. geophys. Res.,* **72**(12), 3247–3262.

Cromwell, G., Tauxe, L., Staudigel, H. & Ron, H., 2015. Paleointensity estimates from historic and modern Hawaiian lava flows using glassy basalt as a primary source material, *Phys. Earth planet. Inter.,* **241,** 44–56.

de Groot, L., Biggin, A., Dekkers, M., Langereis, C. & Herrero-Bervera, E., 2013. Rapid regional perturbations to the recent global geomagnetic decay revealed by a new Hawaiian record, *Nat. Commun.,* **4,** 2727.

Dettmer, J., Benavente, R., Cummins, P.R. & Sambridge, M., 2014. Trans-dimensional finite-fault inversion, *Geophys. J. Int.,* **199**(2), 735–751.

Gallagher, K., Bodin, T., Sambridge, M., Weiss, D., Kylander, M. & Large, D., 2011. Inference of abrupt changes in noisy geochemical records using transdimensional changepoint models, *Earth planet. Sci. Lett.,* **311**(1), 182–194.

Gallet, Y., Genevey, A. & Courtillot, V., 2003. On the possible occurrence of 'archaeomagnetic jerks' in the geomagnetic field over the past three millennia, *Earth Planet Sc. Lett.,* **214**(1), 237–242.

Genevey, A., Gallet, Y., Constable, C.G., Korte, M. & Hulot, G., 2008. Archeoint: an upgraded compilation of geomagnetic field intensity data for the past ten millennia and its application to the recovery of the past dipole moment, *Geochem. Geophys. Geosyst.,* **9**(4), doi:10.1029/2007GC001881.

Genevey, A., Gallet, Y., Thébault, E., Jesset, S. & Le Goff, M., 2013. Geomagnetic field intensity variations in Western Europe over the past 1100 years, *Geochem. Geophys. Geosyst.,* **14**(8), 2858–2872.

Genevey, A., Gallet, Y., Jesset, S., Thebault, E., Bouillon, J., Lefèvre, A. & Le Goff, M., 2016. New archeointensity data from French Early Medieval pottery production (6th–10th century AD). Tracing 1500 years of geomagnetic field intensity variations in Western Europe, *Phys. Earth planet. Int.,* **257,** 205–219.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika,* **82**(4), 711–732.

Hammond, M., Lanos, P., Hill, M. & Colleoni, F., 2017. An archaeomagnetic study of a Roman Bath in Southern France, *Archaeometry,* **59**(2), 356–372.

Hellio, G. & Gillet, N., 2018. Time-correlation-based regression of the geomagnetic field from archeological and sediment records, *Geophys. J. Int.,* **214**(3), 1585–1607.

Hellio, G., Gillet, N., Bouligand, C. & Jault, D., 2014. Stochastic modelling of regional archaeomagnetic series, *Geophys. J. Int.,* **199**(2), 931–943.

Hervé, G. & Lanos, P., 2017. Improvements in Archaeomagnetic Dating in Western Europe from the Late Bronze to the Late Iron Ages: an Alternative to the Problem of the Hallstattian Radiocarbon Plateau, *Archaeometry,* **51**(1), 123–14.

Hervé, G., Schnepp, E., Chauvin, A., Lanos, P. & Nowaczyk, N., 2011. Archaeomagnetic results on three Early Iron Age salt-kilns from Moyenvic (France), *J. geophys. Int.,* **185**(1), 144–156.

Hervé, G., Chauvin, A. & Lanos, P., 2013. Geomagnetic field variations in Western Europe from 1500BC to 200AD. Part II: new intensity secular variation curve, *Phys. Earth planet. Inter.,* **218,** 51–65.

Hervé, G., Chauvin, A., Milcent, P.-Y. & Tramon, A., 2016. Archaeointensity study of five Late Bronze Age fireplaces from Corent (Auvergne, France), *J. Archaeol. Sci.: Rep.,* **7,** 414–419.

Hervé, G. *et al.*, 2017. Fast geomagnetic field intensity variations between 1400 and 400 BCE: new archaeointensity data from Germany, *Phys. Earth planet. Inter.,* **270,** 143–156.

Hopcroft, P.O., Gallagher, K. & Pain, C.C., 2007. Inference of past climate from borehole temperature data using Bayesian Reversible Jump Markov chain Monte Carlo, *Geophys. J. Int.,* **171**(3), 1430–1439.

Hunter, J.D., 2007. Matplotlib: a 2D graphics environment, *Comput. Sci. Eng.,* **9**(3), 90–95.

Ingham, E., Heslop, D., Roberts, A.P., Hawkins, R. & Sambridge, M., 2014. Is there a link between geomagnetic reversal frequency and paleointensity? A Bayesian approach, *J. geophys. Res.,* **119**(7), 5290–5304.

Jaynes, E.T., 2003. *Probability Theory: The Logic of Science,* Cambridge University Press.

Kapper, K.L., Donadini, F. & Hirt, A.M., 2015. Holocene archeointensities from mid European ceramics, slags, burned sediments and cherts, *Phys. Earth planet. Inter.,* **241,** 21–36.

Korte, M. & Constable, C., 2003. Continuous global geomagnetic field models for the past 3000 years, *Phys. Earth planet. Inter.,* **140,** 73–89.

Korte, M. & Constable, C., 2011. Improving geomagnetic field reconstructions for 0-3ka, *Phys. Earth planet. Inter.,* **188**(3-4), 247–259.

Korte, M. & Constable, C.G., 2018. Archeomagnetic intensity spikes: global or regional geomagnetic field features? *Front. Earth Sci.,* **6,** 2160–15.

Korte, M., Constable, C., Donadini, F. & Holme, R., 2011. Reconstructing the Holocene geomagnetic field, *Earth planet. Sci. Lett.,* **312,** 497–505.

Kostadinova-Avramova, M., Kovacheva, M. & Boyadzhiev, Y., 2014. Contribution of stratigraphic constraints of Bulgarian prehistoric multilevel tells and a comparison with archaeomagnetic observations, *J. Archaeol. Sci.* **43**(C), 227–238.

Kovacheva, M., Boyadziev, Y., Kostadinova-Avramova, M., Jordanova, N. & Donadini, F., 2009. Updated archeomagnetic data set of the past 8 millennia from the Sofia laboratory, Bulgaria, *Geochem. Geophys. Geosyst.,* **10**(5), doi:10.1029/2008GC002347.

Lanos, P., 2004. Bayesian inference of calibration curves: application to archaeomagnetism, in *Tools for Constructing Chronologies,* pp. 43–82, Springer.

Lanos, P. & Philippe, A., 2017. Hierarchical Bayesian modeling for combining dates in archeological context, *J. Soc. Francaise Statistique,* **158**(2), 72–88.

Lanos, P., Le Goff, M., Kovacheva, M. & Schnepp, E., 2005. Hierarchical modelling of archaeomagnetic data and curve estimation by moving average technique, *Geophys. J. Int.,* **160**(2), 440–476.

Le Goff, M., Gallet, Y., Genevey, A. & Warmé, N., 2002. On archeomagnetic secular variation curves and archeomagnetic dating, *Phys. Earth planet. Inter.,* **134**(3), 203–211.

Lhuillier, F., Fournier, A., Hulot, G. & Aubert, J., 2011. The geomagnetic secular-variation timescale in observations and numerical dynamo models, *Geophys. Res. Lett.,* **38,** L09306.

Licht, A., Hulot, G., Gallet, Y. & Thébault, E., 2013. Ensembles of low degree archeomagnetic field models for the past three millennia, *Phys. Earth planet. Inter.,* **224,** 38–67.

Luo, X., 2010. Constraining the shape of a gravity anomalous body using reversible jump Markov chain Monte Carlo, *Geophys. J. Int.,* **180**(3), 1067–1079.

Malinverno, A. & Briggs, V.A., 2004. Expanded uncertainty quantification in inverse problems: hierarchical Bayes and empirical Bayes, *Geophysics,* **69**(4), 1005–1016.

Nilsson, A., Holme, R., Korte, M., Suttie, N. & Hill, M., 2014. Reconstructing Holocene geomagnetic field variation: new methods, models and implications, *Geophys. J. Int.,* **198**(1), 229–248.

Pavón-Carrasco, F.J., Rodríguez-González, J., Osete, M.L. & Torta, J.M., 2011. A Matlab tool for archaeomagnetic dating, *J. Archaeol. Sci.,* **38**(2), 408–419.

Pavón-Carrasco, F.J., Osete, M.L., Torta, J.M. & De Santis, A., 2014. A geomagnetic field model for the Holocene based on archaeomagnetic and lava flow data, *Earth planet. Sci. Lett.,* **388**(C), 98–109.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P., 2001. *Numerical Recipes in Fortran 77,* 2nd ed., CUP.

Ray, A., Key, K., Bodin, T., Myer, D. & Constable, S., 2014. Bayesian inversion of marine CSEM data from the Scarborough gas field using a transdimensional 2-D parametrization, *Geophys. J. Int.,* **199**(3), 1847–1860.

Roberts, G.O., 1996. Markov chain concepts related to sampling algorithms, in *Markov chain Monte Carlo in Practice,* pp. 45–57.

Sambridge, M., 2016. Reconstructing time series and their uncertainty from observations with universal noise, *J. geophys. Res,* **121**(7), 4990–5012.

Sambridge, M., Bodin, T., Gallagher, K. & Tkalčić, H., 2013. Transdimensional inference in the geosciences, *Phil. Trans. R. Soc. A* **371**(1984), 20110547.

Schnepp, E., Lanos, P. & Chauvin, A., 2009. Geomagnetic paleointensity between 1300 and 1750 A.D. derived from a bread oven floor sequence in Lübeck, Germany, *Geochem. Geophys. Geosyst.,* **10**(8), doi:10.1029/2009GC002470.

Schnepp, E., Obenaus, M. & Lanos, P., 2015. Posterior archaeomagnetic dating: an example from the Early Medieval site Thunau am Kamp, Austria, *J. Archaeol. Sci.: Rep.,* **2,** 688–698.

Shaar, R., Ben-Yosef, E., Ron, H., Tauxe, L., Agnon, A. & Kessel, R., 2011. Geomagnetic field intensity: how high can it get? How fast can it change? Constraints from iron age copper slag, *Earth planet. Sci. Lett.,* **301,** 297–306.

Tema, E., Herrero-Bervera, E. & Lanos, P., 2017. Geomagnetic field secular variation in pacific ocean: a bayesian reference curve based on holocene hawaiian lava flows, *Earth planet. Sci. Lett.,* **478,** 58–65.

Thébault, E. & Gallet, Y., 2010. A bootstrap algorithm for deriving the archeomagnetic field intensity variation curve in the Middle East over the past 4 millennia BC, *Geophys. Res. Lett.,* **37**(22), doi:10.1029/2010GL044788.

Thellier, E., 1959. Sur l'intensité du champ magnétique terrestre dans le passé historique et géologique, *Ann. Geophys.,* **15,** 285–376.

Welch, P., 1967. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms, *IEEE Trans. Audio Electroacoust.,* **15**(2), 70–73.

Yu, Y., Tauxe, L. & Genevey, A., 2004. Toward an optimal geomagnetic field intensity determination technique, *Geochem. Geophys. Geosyst.,* **5**(2), doi:10.1029/2003GC000630.

APPENDIX A: DATA

A1 Paris700

We first consider the application to a well-studied data set from a region, of radius 700 km, centred on Paris, France, in order to compare our results with previous work. The data were selected using the same quality criteria as those previously considered by Genevey *et al.* (2013, 2016). Our compilation focuses on the past three millennia and includes all the data selected by Genevey *et al.* (2016; and references therein), to which we added data obtained by Chauvin *et al.* (2000), Hammond *et al.* (2017), Hervé *et al.* (2011, 2013, 2016), Kapper *et al.* (2015) and Kovacheva *et al.* (2009). It comprises 154 entries, which were all reduced to the latitude of Paris (48.9°N) using the virtual axial dipole moment (VADM) approximation. The magnetic intensity of each data point is assumed to be normally distributed, with a mean and standard deviation estimated from laboratory analysis. The age of each datum is unknown but assumed uniformly distributed within a given historically dated interval.

All data points were acquired using the experimental method developed by Thellier (1959) or using one of its variants (Coe 1967; Aitken *et al.* 1988; Yu *et al.* 2004). At least 2 partial-thermoremanent magnetization (p-TRM) checks were carried out during the experiments to assess the absence of alteration of the magnetic mineralogy during the thermal treatment. Each archeointensity result relies on the average of at least three intensity values regardless of whether this average was performed at the fragment level or at the fragment group level. The standard deviation of the averaged intensity values does not exceed 15 per cent of the corresponding means. A correction of the anisotropy effect on TRM acquisition was made for the archeological artefacts most prone to this effect (i.e. pottery and tiles; Genevey *et al.* 2008). Furthermore, when the effect of the cooling rate on TRM acquisition was not analysed, we chose to apply a correction of 5 per cent-decrease to the archeointensity data (see discussion in Genevey *et al.* 2008). We did not consider a selection criterion on age uncertainties.

Finally, to remain consistent with the data selection criteria in Hervé *et al.* (2017) we omit the datum 412.5 ± 125 BC reported in Hervé *et al.* (2013).

A2 Hawaii

To examine the effect of data sparsity on our method, we consider a data set of Hawaiian intensity measurements taken from the Geomagia.v3 database (Brown *et al.* 2015). Each data point has both an intensity and age that are assumed normally distributed; the age estimate comes from radiocarbon analysis except for the most recent data whose ages are constrained by historical observations. For the latter data, we arbitrarily fixed the age standard deviation $\sigma = 0.5$ yr. Our approach here relied on much less selective criteria, which nevertheless led to a data set much more sparse than the Paris700 compilation. We retained the results whose quality is supported by an alteration test, regardless of the nature of this test, and which possess uncertainties in both intensity and in age. Between about 9000 BC and 4500 BC most available data lack age uncertainties, and we have therefore focused on the past 4500 yr. We complemented the compilation with the data sets recently obtained by de Groot *et al.* (2013) and Cromwell *et al.* (2015). All 134 data points are reduced to the latitude of Kilauea volcano (19.42°N).

A3 Lübeck-Paris700: an example of stratification

We combined the Paris700 data set (between 900 AD and 2000 AD) with the stratified data set from Lübeck, Germany (Schnepp *et al.* 2009), whose data ages are not independent and must obey a strict ordering in time. These archeointensity data were obtained from a long series of superimposed bread oven floors in a bakery that was in activity from ∼1300 AD to ∼1750 AD. Although Lübeck is 800 km (marginally more than 700 km from Paris), we reduce these measurements to the latitude of Paris by assuming that the non-dipole field effect was negligible between the two regions.

APPENDIX B: THE AH-RJMCMC ALGORITHM

B1 Model overview

Our aim is to sample the joint posterior probability distribution of the model (the intensity as a function of time, alongside the data ages), given the data (the intensity values), which is described in Section 3 in the main text. Here we present in detail the method we use to sample from (and therefore characterize) the posterior

distribution. For completeness, however, we first briefly summarize the modelling framework.

The model is defined within the age interval $[t_{\text{start}}, t_{\text{end}}]$, which needs to be large enough to span the temporal distributions (with uncertainty) of all data; otherwise some realizations of the data ages will be outside the model space. In the case of uniformly distributed ages, this is straightforward; for normally distributed ages, we ensure that the model boundaries lie at least 2 standard deviations from each datum mean-age.

The posterior is parametrized by the model vector $\mathbf{m} = [\mathbf{f}, k, \mathbf{a}]^T$, where $\mathbf{f}$ is the vector containing the endpoint intensities and the k internal vertex intensities and ages:

$$\mathbf{f} = (F_{\text{start}}, F_{\text{end}}, F_1, F_2, \ldots, F_k, t_1, t_2, \ldots, t_k),$$

and $\mathbf{a}$ are the data ages. The posterior requires a prescription of the likelihood (assumed normal, see Section 3.3) and prior distributions.

The prior distributions on the vertex intensities and number of vertices k are assumed independent and uniform, $F_j \sim \text{U}(\tilde{F}_{\min}, \tilde{F}_{\max})$, $k \sim \text{U}(0, k_{\max})$. Without stratigraphy constraints, the ages are assumed independent and either uniformly or normally distributed $a_j \sim \text{U}(a_j^{\min}, a_j^{\max})$, or $a_j \sim \mathcal{N}(\mu_j, \lambda_j^2)$, where μ_j and λ_j are the given mean and standard deviation. If a subset of the ages have stratigraphy constraints, then they are assumed to be independently distributed yet ordered: $a_j < a_{j+1} < \ldots$. In practice, this is handled by drawing the ages independently and discarding a set of draws if the stratification constraints are violated. For ease of model development, the vertex ages are assumed confined to an equally-spaced discrete set of N values between the temporal limits of the model $[t_{\text{start}}, t_{\text{end}}]$; the joint prior distribution of the vertex ages (in which their order is irrelevant) is then

$$p((t_1, t_2, \ldots, t_k)|k) = \left[\frac{N!}{k!(N-k)!}\right]^{-1}.$$

Our algorithm is to assemble a Markov chain, whose distribution converges to the posterior distribution we seek. To add a new model to the Markov chain, we propose a new model $\mathbf{m}'$ which differs from the current model $\mathbf{m}$ by virtue of one of several possible moves with given probability:

(1) Change the value of the intensity (F_j) of a randomly selected vertex j (prob 1/3)

(2) Alter the age-distribution of vertices (prob 1/3) by either

(a) Move in age: change the age (t_j) of a randomly selected vertex j (prob 1/9)

(b) Birth: create a new vertex (prob 1/9)

(c) Death: remove a vertex (prob 1/9)

(3) Resample a randomly chosen subset of ages in $\mathbf{a}$ (prob 1/3)

B2 Proposals

Each of these proposals is given in more detail below.

B2.1 Change the value of intensity

For move type (1), we perturb

$$F_j' = F_j + z,$$

where $z \sim \mathcal{N}(0, \sigma_{\text{change}}^2)$, where σ_{change} is a user-specified standard deviation. Note that

$$p(F_k'|F_k) = \frac{1}{(2\pi)^{1/2}\sigma_{\text{change}}} \exp\left[-\frac{(F_k' - F_k)^2}{2\sigma_{\text{change}}^2}\right]$$

which is equal to $p(F_k|F_k')$. Thus this move is symmetric.

B2.2 Move in age

For move type (2a), we alter the age t_j of vertex j where j is randomly selected

$$t_j' = t_j + y$$

where y represents a temporal shift to one of the unoccupied discrete set of ages (of which there are $N - k$). However, in the limit ($N \to \infty$) that we will eventually take, all positions are unoccupied with probability 1 and so we may model age as a continuous variable and draw $y \sim \mathcal{N}(0, \sigma_{move}^2)$ where σ_{move}^2 is user-specified. Again, because of the symmetry of the distribution,

$$p(t_j'|t_j) = p(t_j|t_j')$$

and the move is symmetric.

B2.3 Birth

For move (2b), we give birth to a new vertex by randomly selecting an unoccupied epoch and assigning an intensity value; we also increase k by 1. In the discrete case, there are $N - k$ available choices of age, which in the continuous limit becomes a uniform distribution. The value of intensity we give the point is not arbitrary, but a perturbation away from its interpolated value from the current model. Suppose we choose t_j' which lies between t_j and t_{j+1}. Then we set

$$F_j' = F_j^* + r,$$

where $r \sim \mathcal{N}(0, \sigma_{\text{birth}}^2)$, and F_j^* is the value found using a linear interpolant between t_j and t_{j+1}. We then need to augment the model vector by F_j' and t_j'. The proposed model vector differs then from the current model by just two additional parameters: all other values (aside from k) remain the same.

B2.4 Death

For move (2c), we remove at random one of the vertices (with its associated intensity value) from the current model, and reduce k by 1. The proposed model therefore has two fewer parameters than the current model—all remaining parameters (aside from k) are identical.

B2.5 Resample ages

A subset of the ages in $\mathbf{a}$ of size $\lfloor N_{\text{data}}/\beta \rfloor$ (with β a user-specified parameter) is resampled according to the prior distributions (uniform or normal). This move is symmetric.

B3 Acceptance probabilities

Having defined a new proposal $\mathbf{m}'$, we need to decide on whether to accept it or not. For fixed-dimension MCMC modelling (i.e. k constant) we define

$$\alpha = \min\left(1, \frac{p(\mathbf{d}|\mathbf{m}')\, p(\mathbf{m}')\, q(\mathbf{m}|\mathbf{m}')}{p(\mathbf{d}|\mathbf{m})\, p(\mathbf{m})\, q(\mathbf{m}'|\mathbf{m})}\right)$$

that is the ratio of the products of the likelihood, prior, and proposal probability. This ratio is then compared with a randomly drawn

uniform variable $r \in U[0, 1]$. If $\alpha > r$ then the model is accepted, if not it is rejected.

For the general transdimensional case, Green (1995) showed that the proposal ratio can be written as

$$\frac{q(\mathbf{m}|\mathbf{m}')}{q(\mathbf{m}'|\mathbf{m})} = \frac{j(\mathbf{m}|\mathbf{m}')q_{\mathbf{m}'\to\mathbf{m}}(u')}{j(\mathbf{m}'|\mathbf{m})q_{\mathbf{m}\to\mathbf{m}'}(u)}|J|.$$

Here, $j(\mathbf{m}|\mathbf{m}')$ is the jump probability: the probability that the algorithm chooses to select proposal $\mathbf{m}$ from model $\mathbf{m}'$. The pdfs $q_{\mathbf{m}'\to\mathbf{m}}(u')$ and $q_{\mathbf{m}\to\mathbf{m}'}(u)$ give the probability of the random variables needed to make the proposal; note that, for example, $q_{\mathbf{m}'\to\mathbf{m}}(u')$ can depend only on the current state $\mathbf{m}'$ of which u' is a perturbation. Finally, J is the Jacobian of transformation. When making a proposal, if any of the random variables drawn lie outside their prior distribution then $p(\mathbf{m}')$ vanishes and the acceptance ratio is 0: the proposal is never accepted.

B3.1 Moves of fixed dimension

If k is unaltered, then since both moves (1), (2a) and (3) are symmetric i.e. $q(\mathbf{m}|\mathbf{m}') = q(\mathbf{m}'|\mathbf{m})$, the prior distributions cancel out and α is simply the minimum of 1 and the ratio of the likelihoods: $e^{-(\phi'-\phi)}$, where ϕ is defined in eq. (2).

B3.2 Birth

For a birth move, suppose that model $\mathbf{m}$ has dimension k and $\mathbf{m}'$ dimension $k+1$. Then the ratio of priors is

$$\begin{aligned}\frac{p(\mathbf{m}')}{p(\mathbf{m})} &= \frac{p(k+1)\; p(t_1, t_2, \ldots, t_{k+1})\, p(\mathbf{a}) \prod_{j=1}^{k+1} p(F_j)}{p(k)\; p(t_1, t_2, \ldots, t_k)\, p(\mathbf{a}) \prod_{j=1}^{k} p(F_j)} \\ &= \frac{\left[\frac{N!}{(k+1)!(N-[k+1])!}\right]^{-1}}{\left[\frac{N!}{k!(N-k)!}\right]^{-1}} (\tilde{F}_{\max} - \tilde{F}_{\min})^{-1} \\ &= \frac{k+1}{N-k}\,(\tilde{F}_{\max} - \tilde{F}_{\min})^{-1}. \end{aligned} \tag{B1}$$

We now need to consider the ratio of the proposal probabilities. For birth, the jump probability is 1/3, and we further need to draw a specific additional age from the unoccupied ages (probability $(N-k)^{-1}$) and then draw an intensity from its assumed normal distribution centred on the interpolated value with probability $q(F_j')$. The model vector is simply augmented with the new age and intensity, and the transformation (described by the Jacobian J) is simply a relabelling of indices; therefore $J = 1$. The ratio of transition probabilities is then

$$\frac{q(\mathbf{m}|\mathbf{m}')}{q(\mathbf{m}'|\mathbf{m})} = \frac{\frac{1}{3}(k+1)^{-1}}{\frac{1}{3}(N-k)^{-1} q(F_j')}, \tag{B2}$$

where

$$q(F_j') = \frac{1}{\sqrt{2\pi\sigma_{\text{birth}}^2}} e^{-(F_j' - F_j^*)^2/2\sigma_{\text{birth}}^2}. \tag{B3}$$

The acceptance ratio is then

$$\begin{aligned}\alpha &= \min\left[1, e^{-(\phi'-\phi)}\,\frac{k+1}{N-k}\,(\tilde{F}_{\max} - \tilde{F}_{\min})^{-1}\,\frac{\frac{1}{3}(k+1)^{-1}}{\frac{1}{3}(N-k)^{-1} q(F_j')}\right] \\ &= \min\left[1, e^{-(\phi'-\phi)}\,(\tilde{F}_{\max} - \tilde{F}_{\min})^{-1}\, q(F_j')^{-1}\right].\end{aligned}$$

Note that if instead of drawing the new value from a Gaussian distribution $q(F_j')$ we proposed a new value from the uniform prior distribution $U[\tilde{F}_{\min}, \tilde{F}_{\max}]$, we would have terms that would cancel out and α would then depend only on the ratio of likelihoods.

B3.3 Death

For the acceptance probability of the death of vertex j, suppose that model $\mathbf{m}$ has dimension k and $\mathbf{m}'$ dimension $k-1$. The jump probability is 1/3 and the doomed vertex (from k equally probable choices) needs to be removed; there is no requirement to generate any new variables. Each of the ratio of the priors and the proposals are simply the reciprocal of those for birth with k replaced by $k-1$.

$$\alpha = \min\left[1,\; e^{-(\phi'-\phi)}\,(\tilde{F}_{\max} - \tilde{F}_{\min}) q(F_j)\right] \tag{B4}$$

where $q(F_j)$ is given by (B3). We finally note that none of the acceptance ratios depend on N, or on the choice of prior distribution for the data ages (which cancels out in all cases). We therefore take N to be infinitely large, converting each vertex age into an independent, continuous, uniformly distributed random variable.

APPENDIX C: AN INTEGRATED LIKELIHOOD FUNCTION

At the heart of the Bayesian method is a likelihood function, which gives the probability of an observed intensity value $\mathcal{F}$ given an assumed underlying model. If the datum's age is assumed error-free and the true intensity value is y, the likelihood is

$$p(\mathcal{F}|y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\mathcal{F}-y)^2}{2\sigma^2}\right]. \tag{C1}$$

However, because the datum's age itself is not known precisely, we need a method of incorporating its uncertainty into the modelling process. In the principal method we consider in this paper, these effects are accommodated by including the data ages into the model vector $\mathbf{m} = [\mathbf{f}, k, \mathbf{a}]^T$, and then successively resampling $\mathbf{a}$ according to the specified prior distribution. For our problem, this is a natural way to proceed, because the marginalized posterior distributions of the data ages (which are of significant scientific value) are trivial to compute. However, it is worth noting that this is not the only way of setting up the model: rather than including the data ages in the model vector and using the intensities in the likelihood, we could swap these around. Thus in an alternative model, we might use the model vector $\mathbf{m} = [\mathbf{f}, k, F]^T$ where $F = (\mathcal{F}_1, \mathcal{F}_2, \ldots)$ are the data intensities, and use for the likelihood the uniform distribution on data age. The fact that these two model setups are different is a manifestation of the asymmetry of the way in which the data is handled. Ideally, the data ages and intensities, both accompanied with a specific error distribution, should be treated symmetrically as they enter on the same footing.

Sambridge (2016) describes a method in which 2-D data with known error distributions are treated symmetrically by a piecewise linear MCMC algorithm. He considers only the case where both variables are described by a joint normal distribution: here, we extend his methodology to our focus on uniformly distributed ages and normally distributed intensities.

The keystone of Sambridge (2016) is a definition of the likelihood that expresses the probability of realizing a particular datum $\mathbf{d} = (\mathcal{F}, a)$, where $\mathcal{F}$ is its intensity and a is its age, given a model curve

c (here, the piecewise linear description of intensity with time). This can be written as an integral over the whole curve in intensity-time space, considering a general *source* point q on the curve c parametrized by arc length:

$$p(\mathbf{d}|c) = \int_c p[\mathbf{d}|q(s)]p[q(s)]\mathrm{d}s. \tag{C2}$$

Each choice of q represents a different and equally probable [and thus $p(q(s))$ is uniform] realization of the datum's true intensity and age. It is worth noting that the choices of (i) parametrising c in terms of arclength, rather than age or intensity (as we could have done in our case) and (ii) assuming that all values of q are equally probable, produces a method which is symmetric in the variables.

There are two ingredients required in the above. First, because each choice of q is assumed equally probable, we can calculate $p[q(s)] = C$ by integrating over the whole curve:

$$\int_c p[q(s)]\mathrm{d}s = C\,L = 1,$$

where L is the total arclength of the curve of interest; hence $C = 1/L$.

Second, we need to be able to evaluate the likelihood of realizing the data $\mathbf{d}$ assuming the true value is at $q(s)$. We assume a normally distributed intensity $\mathcal{F} \sim \mathcal{N}(F_q, \sigma^2)$, where F_q is the intensity at q, and uniform in age $a \sim U[T_0, T_1]$. We will always consider a model curve c that includes the entire age interval $[T_0, T_1]$, and because the datum age probability is zero outside this interval, in (C2) we can restrict attention to ages within $[T_0, T_1]$. Suppose there are j line segments that we need to consider, each with a start and terminal age. We will consider each in isolation, and add their contributions together.

We can parametrize the jth linear segment of the model by a new variable $0 \leq \theta \leq 1$, where $s = l_j\theta$ and l_j is the arclength of the segment. Hence $\mathrm{d}s = l_j\,\mathrm{d}\theta$ and the contribution to (C2) is

$$\frac{l_j}{L}\int_0^1 p(\mathbf{d}|q(\theta))\mathrm{d}\theta = \int_0^1 \frac{l_j}{\sqrt{2\pi}\,\sigma\,L\,(T_1 - T_0)} \times \exp\left[-\frac{(\mathcal{F} - F_q(\theta))^2}{2\sigma^2}\right]\mathrm{d}\theta,$$

where $F_q(\theta)$ is the intensity at $q(s(\theta))$. Now,

$$\mathcal{F} - F_q(\theta) = \mathcal{F} - [(1-\theta)F_0 + \theta F_1] = a - \theta b,$$

where F_0 and F_1 are the start and end values of the intensity in the segment, $a = \mathcal{F} - F_0$ and $b = F_1 - F_0$. Hence the contribution can be written

$$\frac{l_j}{\sqrt{2\pi}\,\sigma\,L\,(T_1 - T_0)}\int_0^1 \exp\left[-\frac{(\theta - \bar{\theta})^2}{2\sigma_\theta^2}\right]\mathrm{d}\theta, \tag{C3}$$

where using the notation of Sambridge (2016), $\bar{\theta} = a/b$ and $\sigma_\theta = \sigma\,b^{-1}$. Note that in Sambridge (2016), his eq. (A9) the definition of σ_θ has a typographic error.

The contribution is thus

$$\frac{l_j\sqrt{2}\sigma_\theta}{\sqrt{2\pi}\,\sigma\,L\,(T_1 - T_0)}\frac{\sqrt{\pi}}{2}\left[\mathrm{erf}(\mathrm{t}_2) - \mathrm{erf}(\mathrm{t}_1)\right]$$

where $t_2 = (1-\bar{\theta})(\sqrt{2}\sigma_\theta)^{-1}$, and $t_1 = -\bar{\theta}(\sqrt{2}\sigma_\theta)^{-1}$.

Summing up over all the line segments,

$$p(\mathbf{d}|c) = \frac{1}{2\,L\,(T_1 - T_0)}\sum_j \frac{l_j}{b_j}\left[\mathrm{erf}(\mathrm{t}_{2,\mathrm{j}}) - \mathrm{erf}(\mathrm{t}_{1,\mathrm{j}})\right].$$

Assuming that the data (indexed by $i = 1, 2, \ldots N_{\mathrm{data}}$) are independent, the combined likelihood is then

$$\prod_{i=1}^{N_{data}} \frac{1}{2\,L\,(T_{1,i} - T_{0,i})}\sum_j \frac{l_j}{b_{j,i}}\left[\mathrm{erf}(\mathrm{t}_{2,\mathrm{j,i}}) - \mathrm{erf}(\mathrm{t}_{1,\mathrm{j,i}})\right]. \tag{C4}$$

It is interesting to consider the limit of small error in age in eq. (C3), corresponding to the case of $F_0 = F_1 = F_q(\theta)$, in which there is only a single line segment (as the age interval is too small to contain more than one). In this limit the likelihood collapses to

$$\frac{1}{\sqrt{2\pi}\,\sigma\,(T_1 - T_0)}\exp\left[-(\mathcal{F} - F_0)^2/2\sigma^2\right],$$

which recovers the usual normal likelihood (C1), albeit normalized by the (infinitesimal) time duration.

Finally, we note that numerically, it is easiest to use the log-likelihood, rather than the likelihood itself. In cases where t_1 and t_2 are large (where the model is a very poor fit to the data), the log likelihood is numerically challenging to compute since both error functions are -1 to machine precision with a difference of zero. In this case, we can set the likelihood (for this datum) to be a prescribed small number (for example, 10^{-50}). Note that this case only arises when the model is a very poor fit to the data, which can occur on initializing the algorithm with a randomized model, but practically will never enter the MCMC sampling beyond 'burn-in'.

APPENDIX D: SUPPLEMENTARY PLOTS FOR THE PARIS700 DATA SET

Fig. D1 shows other diagnostics of the AH-RJMCMC method applied to the Paris700 data set that supplement those given in Section 5. The top row shows the misfit against iteration (i.e. the index of the Markov-chain) and a normalized histogram of the number of internal change points. The burn-in period (marked by the red bar) is sufficiently long for the misfit to drop to close to its lowest value, and the range of change points (fixed between 1 and 50) is large enough to capture the distribution of vertices, which peaks at around 40. Finally, the bottom panel of Fig. D1 shows a normalized histogram of the difference between the intensity values from Paris700 and the average posterior evaluated at the data mid-point ages weighted by the data intensity errors: $\Delta F/\sigma = \left(g(a_j) - \mathcal{F}_j\right)/\sigma_j$ for each datum j (see Section 3.3). The close similarity of the histogram and the standard $\mathcal{N}(0, 1)$ distribution gives us confidence in the assumed normal distribution of intensity error.

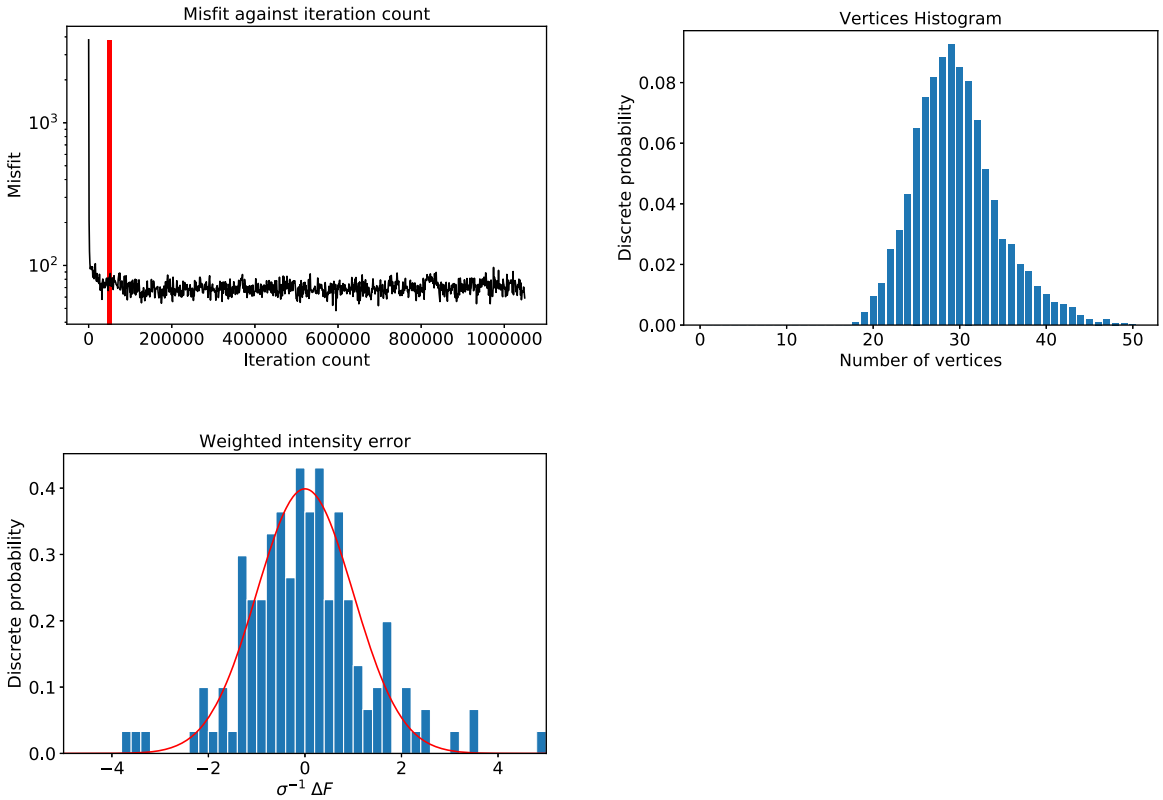

Figure D1. Supplementary diagnostics for the AH-RJMCMC method applied to the Paris700 data set. Top left-hand panel: misfit against Markov-chain index where the red bar indicates the end of the burn-in period. Top right-hand panel: normalized histogram of number of internal change points. Bottom left-hand panel: histogram of weighted residuals of the data set and the average posterior compared with the normal distribution $\mathcal{N}(0, 1)$ in red.

APPENDIX E: SUPPLEMENTARY PLOTS FOR THE HAWAII DATA SET

Fig. E1 shows other diagnostics of the AH-RJMCMC method applied to the Hawaii data set that supplement those given in Section 6. The top row shows the misfit against iteration (i.e. the index of the Markov-chain) and a normalized histogram of the number of internal change points. The burn-in period (marked by the red bar) is sufficiently long for the misfit to drop to close to its lowest value, and the range of change points (fixed between 1 and 100) is large enough to capture the distribution of vertices, which peaks at around 50. Finally, the bottom figure shows a normalized histogram of the difference between the intensity values from Hawaii and the average posterior evaluated at the mid-point ages weighted by the data set intensity errors: $\Delta F/\sigma = \left(g(a_j) - \mathcal{F}_j\right)/\sigma_j$ for each datum j (see Section 3.3). The close similarity of the histogram and the standard $\mathcal{N}(0, 1)$ distribution gives us confidence in the assumed normal distribution of intensity error.

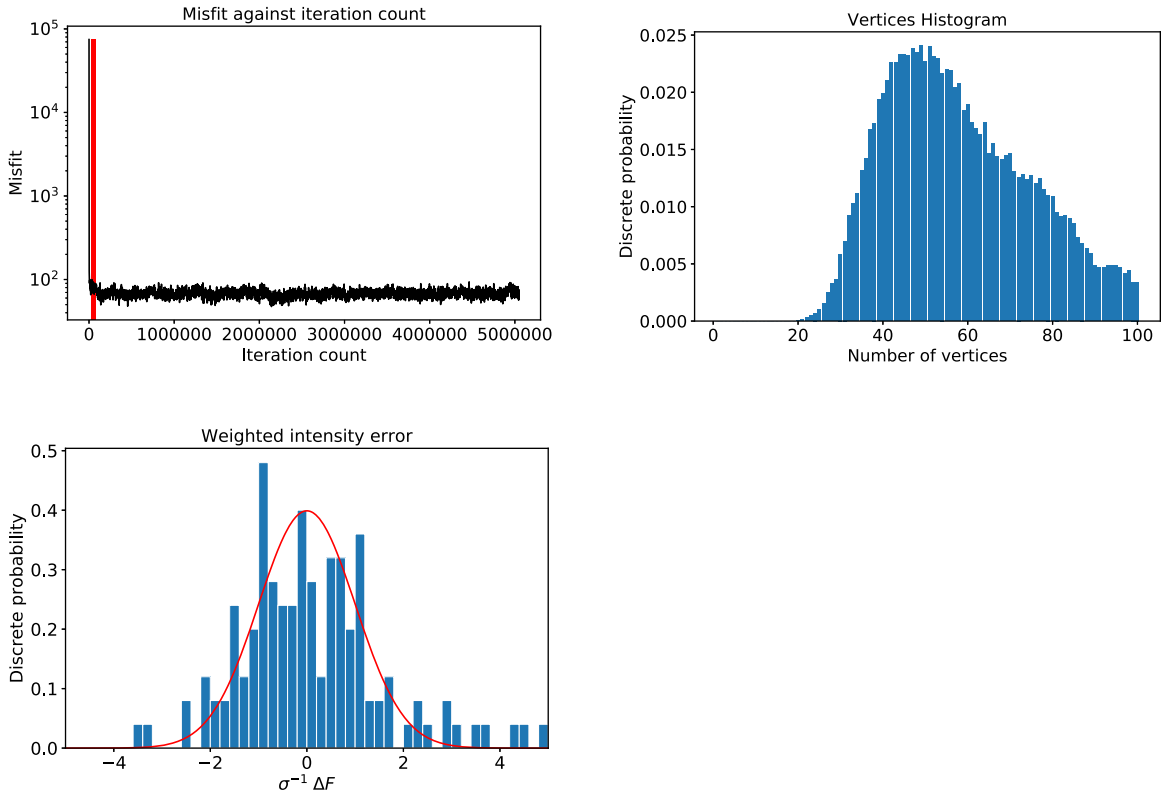

Figure E1. Supplementary diagnostics for the AH-RJMCMC method applied to the Hawaii data set. Top left-hand panel: misfit against Markov-chain index where the red bar indicates the end of the burn-in period. Top right-hand panel: normalized histogram of number of internal change points. Bottom left-hand panel: histogram of weighted residuals of the data set and the average posterior compared with the normal distribution $\mathcal{N}(0, 1)$ in red.