https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

Overcoming multimodalities in age–depth model posteriors using On–the–fly Probability Enhanced Sampling and Parallel Tempering

Hanna Kjellson[1], Andreas Nilsson[1], and Anders Irbäck[1]

[1]Department of Earth and Environmental Science, Lund University, Sölvegatan 12, 223 62, Lund, Sweden

Correspondence: Hanna Kjellson (hanna.kjellson@mgeo.lu.se)

Abstract. Bayesian age–depth models are central to paleoclimate research, linking depth in natural archives to calendar age. When synchronizing variable data from different sites, inference in these models requires sampling from high-dimensional, often multimodal posteriors. Consequently, standard samplers become trapped in local optima, miss plausible chronologies, and bias downstream analyses. Here, we investigate two enhanced sampling approaches, On–the–fly Probability Enhanced Sampling (OPES) and Parallel Tempering (PT), and find that tempering only the likelihood components that drive multimodality can improve exploration. Our results suggest that OPES and PT combined with targeted tempering provide robust age–depth models with implications for a wide range of applications in Earth sciences, including paleoclimate and paleomagnetic reconstructions.

1 Introduction

Understanding how Earth's climate may evolve over the coming decades remains one of the most pressing scientific challenges of our time. Physical modeling and climate projections are cornerstones of climate science, while paleoclimate reconstructions offer a unique opportunity to investigate underlying mechanisms and test projections against evidence from the past. These paleoclimate reconstructions are based on stratigraphic archives such as deep–sea sediment cores, polar ice cores, and cave deposits that have slowly accumulated material over time. The stratigraphic constraints (oldest deposits at the bottom and youngest at the top) provide a timeline, but to use the data for climate reconstruction or most other purposes, we also need a timescale.

For ice cores, laminated sediment cores, and stalagmites, ages can sometimes be inferred by counting layers. However, in the absence of clear layering, age must be modeled, typically using an age–depth model such as BChron (Haslett and Parnell, 2008), Oxcal (Ramsey, 2008), or Bacon (Blaauw and Christen, 2011). Age information is given either by direct dates from radionuclides, such as ^{14}C and U–Th, or by indirect dates obtained by matching data, such as δ^{18}O or magnetic data, across sites, assuming that they vary according to a common process (e.g. Edmonsond and Dyer, 2025; Lee et al., 2023; Nilsson et al., 2022; Parnell et al., 2015). Both direct and indirect dates carry uncertainties that are typically accounted for within a Bayesian framework, where inferring an age–depth model amounts to sampling from a posterior distribution.

With a few data points of direct dates, this posterior can often be effectively sampled using, for instance, Hamiltonian Monte Carlo (HMC; Duane et al., 1987). However, when the common signal is highly variable, incorporating indirect data can

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

introduce additional complexity into the posterior distribution, often leading to multiple modes separated by low–probability transition barriers. Sampling from such a distribution cannot be done efficiently with standard HMC and instead requires enhanced sampling techniques.

Here, we investigate On–the–fly Probability Enhanced Sampling (OPES; Invernizzi et al., 2020) and Parallel Tempering (PT; Swendsen and Wang, 1986), previously used by Eichenseer et al. (2025), and in particular propose a targeted tempering strategy that acts only on the parts of the likelihood that induce multimodality. We explore the utility of the tempering strategy using four semi-synthetic datasets derived from three real datasets, the Dayu Cave (Tan et al., 2009) and Wah Shikhar Cave (Sinha et al., 2011) stalagmite records, as well as a sediment core from Lake Biwa (Ali et al., 1999). The targeted tempering approach performs well, clearly outperforming HMC across all datasets, and full–likelihood tempering for the Wah Shikhar Cave dataset and a slightly modified version of the Dayu Cave dataset.

2 Age–depth model and data

To arrive at the sampling techniques, we first introduce our simplified age–depth model, which is based on Blaauw and Christen (2011). Assuming constant inverse sedimentation rates or, equivalently, sedimentation times per unit depth, $x_1, x_2, \ldots, x_K$, between discrete depths, $d_0, d_1 = d_0 + \Delta d, \ldots, d_K = d_0 + K\Delta d$, the age at depth d is given by:

$$A(d) = A(d_0) - \Delta d \sum_{k=1}^{i} x_k - (d - d_i)x_{i+1}, \text{ where } d_i < d < d_{i+1}. \tag{1}$$

$A(d_0)$ is assumed to be known, and the sedimentation times are distributed as:

$$x_k \sim \text{Lognormal}(\mu, \sigma), \tag{2}$$

where μ and σ are chosen to yield an approximately correct mean accumulation (based on the upper- and lowermost dates), and a variance corresponding to that of a gamma process with shape parameter 1.5 as recommended in Blaauw et al. (2026). Since our goal is to study sampling in challenging posteriors, we omit the auto–regressive "memory" term of Blaauw and Christen (2011) which can sometimes simplify sampling. As described in Trachsel and Telford (2016), there is no single best way to model age vs. depth, so this choice does not make the model less realistic.

With the prior in Eq. 2, denoted $\pi(\mathbf{x})$, in place, we form the posterior by combining it with a likelihood $L(\text{data}|\mathbf{x})$:

$$p(\mathbf{x}|\text{data}) \propto \pi(\mathbf{x}) L(\text{data}|\mathbf{x}). \tag{3}$$

The likelihood can incorporate direct dates with uncertainty, as well as indirect data that only become informative when compared to a reference function. This could, for instance, be a known local or global signal of δ^{18}O, as in Axelsson (2024) or a modeled local signal of the geomagnetic field based on data from multiple records at different locations, as in Nilsson and Suttie (2021).

Here, we consider four datasets (see Table 1 and Fig. 1), all consisting of both direct and indirect data, similar to those investigated by Nilsson and Suttie (2021) and Axelsson (2024). The datasets consist of one synthetic sediment core with dated

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

Table 1. Summary of datasets used.

Abbreviation	**Biw**	**Day**	**Daye**	**Wah**
Dataset	Lake Biwa[a]	Dayu Cave[b]	Dayu Cave[b]	Wah Shikhar Cave[c]
Type of record	sediment	stalagmite	stalagmite	stalagmite
Direct dates	tephra layer	^{230}Th	modified ^{230}Th	^{230}Th
Indirect synthetic data	declination	δ^{18}O	δ^{18}O	δ^{18}O
Reference function	Midpath Geodynamo[d]	ECHAM5–wiso[e]	ECHAM5–wiso[e]	ECHAM5–wiso[e]

[a] Ali et al. (1999). [b] Tan et al. (2009). [c] Sinha et al. (2011). [d] Aubert et al. (2017). [e] Sjolte et al. (2018).

tephra layers and declination data generated based on a Lake Biwa sediment core (Ali et al., 1999), two synthetic speleothem records with ^{230}Th and δ^{18}O data based on a Dayu Cave stalagmite (Tan et al., 2009), and one synthetic speleothem record with ^{230}Th and δ^{18}O data based on a Wah Shikar Cave stalagmite Sinha et al. (2011).

For the Lake Biwa dataset (biw), one of the Dayu Cave datasets (day), and the Wah Shikhar Cave dataset (wah), the direct dates are the true measured values, with ^{230}Th dates for the stalagmites and dated tephra layers for the sediment core. As the reported standard deviations of around 1 yr for the ^{230}Th ages are likely too optimistic for the Dayu Cave dataset, we also consider a modified dataset (daye) with larger standard deviations of 20 yr for the ^{230}Th dates, which we believe are more realistic in a general setting (Schorndorf et al., 2026). The set of direct dates for a record will be referred to as y_n, with standard deviation σ_n at depth d_n for $n = 1, \ldots, N$.

The slowly varying Midpath Geodynamo simulation from Aubert et al. (2017), previously used by Nilsson et al. (2026), is used as a reference function for the sediment core declination data, while a low–pass–filtered version of the ECHAM5–wiso simulation of δ^{18}O from Sjolte et al. (2018) is used for the δ^{18}O data in the three stalagmite datasets. We denote both types of reference functions with f. Since the original measurements of the indirect data do not come from the reference functions considered, we generate synthetic δ^{18}O and declination data for each dataset, with value z_m and standard deviation σ_m at depth d_m for $m = 1, \ldots, M$. Here, σ_m and d_m correspond to, respectively, the uncertainties and depths of the original data, whereas the z_m's are synthetic data created in a two-step procedure. First, we transfer the reference function from age to depth, using an HMC–generated random age–depth curve sample, $A_{dataset}(d)$, from a posterior defined by direct dates only. Second, for each depth d_m of the original data, we generate synthetic measurements by corrupting the reference value $f(d_m)$ with Gaussian noise, with uncertainty equal to that of the original data.

With one set of direct dates and one set of indirect data, all datasets have two types of factors for the likelihood:

$$\begin{aligned} y_n|\mathbf{x} &\sim \mathcal{N}(A(d_n), \sigma_n), \\ z_m|\mathbf{x} &\sim \mathcal{N}(f(A(d_m)), \sigma_m), \end{aligned} \tag{4}$$

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

which, combined with the prior, yield the posterior distribution:

$$p(\mathbf{x}|\mathbf{y},\mathbf{z}) \propto \overbrace{\prod_{k}^{K} \pi(x_k)}^{\pi(\mathbf{x})} \overbrace{\prod_{n}^{N} p(y_n|\mathbf{x}) \prod_{m}^{M} p(z_m|\mathbf{x})}^{L(\text{data}|\mathbf{x})} =: \tilde{p}(\mathbf{x}). \tag{5}$$

This has high probability for age–depth models that both align with the direct dates and fit the indirect data to the reference function. If the direct dates are not that restrictive or the reference function varies rapidly, the posterior can be multimodal, corresponding to several ways of fitting the indirect data to the reference function. However, if the direct dates are restrictive and the reference function varies slowly, there is often essentially only one way to fit the indirect data to the reference function, yielding a unimodal, or at least simple, posterior that can be sampled using standard HMC.

In the previous studies of Nilsson and Suttie (2021) and Axelsson (2024), this is indeed the case. Nilsson and Suttie (2021) include both inclination and declination data and account for the fact that geomagnetic field directions are gradually locked–in in the sediment. This makes the common signal more slowly varying and yields fewer potential fits of the indirect data. In comparison, our model only includes declination data and uses a rapidly varying reference function. This allows for more potential fits of the indirect data, and while it is a simplification, similar data could occur in other geomagnetic field applications.

For δ^{18}O, Axelsson (2024) also uses a more slowly varying reference function than we do. This accounts for the percolation of water down to the cave before deposition (Parrenin et al., 2024). However, δ^{18}O can be more variable than what is assumed in Axelsson (2024), and the data and reference functions considered for the stalagmites here vary on more realistic time–scales, (e.g. Sinha et al., 2011). By using a faster varying reference function and excluding a priori correlations between sedimentation times, we allow for more potential fits of the indirect data while modelling more realistic δ^{18}O variations.

Thus, all our datasets consist of synthetic, but realistic, data that promote multimodal posteriors. In addition, we choose $K = 50$ segments for the discretization in Eq. 1 for all datasets, between depths $d_0 = 0\,\text{cm}$ and $d_K = 1000\,\text{cm}$ for the sediment core, and $d_0 = 0\,\text{mm}$ and $d_K = 100\,\text{mm}$ for the three stalagmites. Using $K = 50$ variables of sedimentation time, the posterior is, in addition to multimodal, also high–dimensional.

3 Sampling from the posterior

Finding regions of high probability for such posteriors can be challenging, especially as probable regions tend to be separated by highly unlikely regions. The standard approach for sampling from an unnormalized distribution $\tilde{p}(\mathbf{x})$ is Markov Chain Monte Carlo (MCMC), such as Gibbs or HMC sampling. In HMC, one defines the potential energy $u(\mathbf{x}) = -\log(\tilde{p}(\mathbf{x}))$ and generates proposals by running Hamiltonian dynamics over this energy landscape, yielding draws from a Markov chain with the normalized probability $p(\mathbf{x})$ as stationary distribution. This works well when there is one dominant global mode. However, in multimodal landscapes, HMC can become stuck around a local mode and sample only from that region. To address this issue, various approaches have been developed. Here we focus on two tempering methods, PT and a multi–thermal variant of OPES.

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

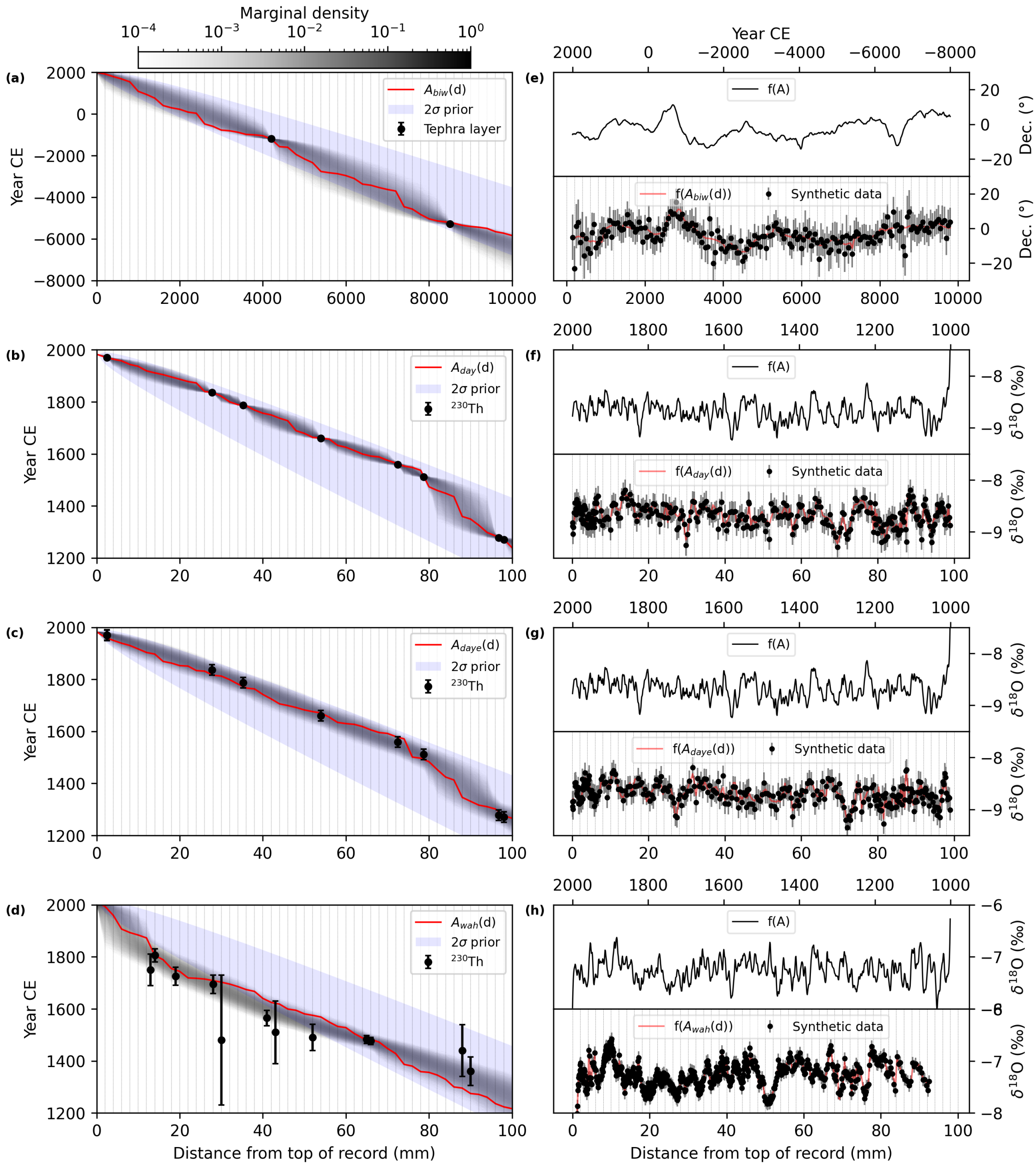

Figure 1. In (a)–(d) we present direct dates (^{230}Th and tephra layers) with two standard deviations, the prior with two standard deviations and the marginal density of age–depth models based on only direct dates along every 0.1 mm depth for the stalagmites and every 1 cm for the sediment core. From (a) to (d), the order of datasets is: the Lake Biwa sediment core, the Dayu Cave stalagmite (original and with inflated ^{230}Th uncertainties), and finally the Wah Shikhar Cave stalagmite. For each dataset, we also present a randomly chosen "true" age–depth curve, $A_{dataset}(d)$, which we use to generate the data in (e)–(h). The latter illustrate, in the same order as (a)–(d), the reference functions vs time, $f(A)$ (top panel), and the synthetic declination or δ^{18}O data vs depth (bottom panel).

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

3.1 Tempering methods

Introducing $\beta_s = 1/T_s$, with the "temperatures" T_s for $s = 1,2,\ldots,S$, and splitting the posterior distribution, $\tilde{p}(x)$, into one part that is easy to sample and another that is hard to sample, $\tilde{p}(\mathbf{x}) = \tilde{p}_e(\mathbf{x})\tilde{p}_h(\mathbf{x})$, we consider the tempered distributions:

$$p_s(\mathbf{x}) = \frac{\tilde{p}_s(\mathbf{x})}{Z_s} = \frac{\tilde{p}_e(\mathbf{x})\tilde{p}_h(\mathbf{x})^{\beta_s}}{Z_s}, \quad s = 1,2,\ldots,S, \tag{6}$$

where:

$$Z_s = \int \tilde{p}_e(\mathbf{x})\tilde{p}_h(\mathbf{x})^{\beta_s} d\mathbf{x}, \tag{7}$$

are unknown normalization constants. Note that $\beta_1 = 1$ corresponds to the true posterior, and the smaller the β_s, the flatter the distribution. This means better exploration for HMC but also less detailed sampling of the high–probability regions. The idea behind tempering methods is to combine more detailed sampling for β_s close to one with more global exploration for β_s less than one. To choose which temperatures to consider, advanced schemes such as the one proposed in Katzgraber et al. (2006) are available, but here we use a simple geometrical distribution:

$$T_s = T_1 \left(\frac{T_S}{T_1}\right)^{\frac{s-1}{S-1}}, \quad s = 1,\ldots,S, \tag{8}$$

which works well for our examples.

3.1.1 Simulated Tempering (ST)

One way to combine the tempered distributions, called simulated tempering (ST; Lyubartsev et al., 1992; Marinari and Parisi, 1992), is to create a Markov chain in the extended space $\bar{\mathbf{x}} = (\mathbf{x}, s) \in \mathbf{R}^K \times \{1,\ldots,S\}$ that samples from the joint stationary distribution:

$$p_{joint}(\bar{\mathbf{x}}) := w_s p_s(\mathbf{x}). \tag{9}$$

The w_s parameters are tunable and set the relative weights of the different temperatures (choosing $w_s = 1/S$ corresponds to having equal weight for all temperatures), and $p_s(\mathbf{x})$ contains the unknown normalization constants Z_s in Eq. 6 and 7.

Given w_s and estimates $\hat{Z}_s$ of Z_s, one can create a chain sampling $p_{joint}(\bar{\mathbf{x}})$ by using ordinary fixed–temperature Monte Carlo updates in $\mathbf{x}$, interrupted at regular intervals by Metropolis–controlled updates of the temperature index s. In the latter updates, one typically proposes a change to either $s-1$ or $s+1$ with equal probability (unless we are at a border temperature). The new temperature state s' is accepted with probability:

$$\alpha(s \to s') = \min\left[1, \frac{w_{s'}\tilde{p}_{s'}(\mathbf{x})\hat{Z}_{s'}^{-1}}{w_s\tilde{p}_s(\mathbf{x})\hat{Z}_s^{-1}}\right]. \tag{10}$$

Samples of the desired posterior are obtained by considering samples with $s = 1$.

Many variants of the original ST scheme have been developed. One line of development is to remove the dynamical variable s and instead let the system evolve in a single effective potential, as in the STeWiE method (Nymeyer, 2010). Rather than

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

sampling the distribution in Eq. 9, this method samples the corresponding marginal distribution, which, in what follows, will be referred to as the target distribution and is given by:

$$p_{tg}(\mathbf{x}) := \sum_{s=1}^{S} w_s p_s(\mathbf{x}). \tag{11}$$

Another major development line concerns the determination of the $\hat{Z}_s$ estimates, as finding reasonable values is crucial to ensure temperature mobility. Originally, this was achieved through an iterative procedure involving potentially costly simulations. In more recent work, it was realized that this preparatory work could be avoided by using an adaptive Monte Carlo approach (Salakhutdinov, 2010), inspired by the Wang–Landau method (Wang and Landau, 2001).

3.1.2 On–the–fly Probability Enhanced Sampling (OPES)

OPES (Invernizzi et al., 2020) combines the recent advances in ST and samples from the marginal distribution in Eq. 11, while on–the–fly estimating the (relative) normalization constants. The method was primarily developed to enable the combination of tempering and local biasing methods, such as umbrella sampling (Torrie and Valleau, 1977). Here, since there are no obvious collective variables, we focus on the tempering aspect and, following Invernizzi et al. (2020), choose $w_s = 1/S$ to sample with equal weight across all temperatures. This corresponds to considering the target distribution:

$$p_{tg}(\mathbf{x}) = \frac{1}{S} \sum_{s=1}^{S} p_s(\mathbf{x}). \tag{12}$$

When sampling, we only need to define the target distribution up to a proportionality constant, so we can also consider an unnormalized version:

$$\tilde{p}_{tg}(\mathbf{x}) = \frac{1}{S} \sum_{s=1}^{S} v_s \tilde{p}_s(\mathbf{x}), \tag{13}$$

where $v_s = Z_1/Z_s$ and the proportionality constant has been chosen as Z_1 to convert the problem of estimating normalization constants to the problem of estimating relative normalization constants.

To sample from this, the idea of Invernizzi et al. (2020) is to introduce a bias or, equivalently, a scaling function $f_v(\mathbf{x}) = \tilde{p}_{tg}(\mathbf{x})/\tilde{p}(\mathbf{x})$. By sampling from the modified density $\tilde{p}(\mathbf{x})f_v(\mathbf{x})$, for instance using Langevin dynamics, as considered in Invernizzi et al. (2020), or using HMC, as considered here, we obtain samples from the target distribution.

The key point is that the scaling function f_v and the relative normalization constants v_s are coupled. In particular, using Eq. 6 and 13, the scaling function can be written as:

$$f_v(\mathbf{x}) = \frac{1}{S} \sum_{s=1}^{S} v_s \tilde{p}_h(\mathbf{x})^{\beta_s - 1}, \tag{14}$$

with:

$$v_s = \frac{Z_1}{Z_s} = \frac{\int [\tilde{p}(\mathbf{x})/\tilde{p}_{tg}(\mathbf{x})]\tilde{p}_{tg}(\mathbf{x})d\mathbf{x}}{\int [\tilde{p}_s(\mathbf{x})/\tilde{p}_{tg}(\mathbf{x})]\tilde{p}_{tg}(\mathbf{x})d\mathbf{x}} = \frac{\mathbf{E}[f_v(\mathbf{x})^{-1}]_{p_{tg}}}{\mathbf{E}[f_v(\mathbf{x})^{-1}\tilde{p}_h(\mathbf{x})^{\beta_s-1}]_{p_{tg}}}. \tag{15}$$

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

This allows for an iterative scheme. Starting from an initial guess, $v_{s,0}$, for the relative normalization constants, v_s, we in iteration q use an estimate, $f_{v,q}(\mathbf{x})$, of the scaling function, $f_v(\mathbf{x})$, defined as:

$$f_{v,q}(\mathbf{x}) = \frac{1}{S}\sum_{s=1}^{S} v_{s,q-1}\tilde{p}_h(\mathbf{x})^{\beta_s - 1}. \tag{16}$$

Before the next iteration, we recompute estimates, $v_{s,q}$, of the relative normalization constants, v_s, using:

$$v_{s,q} = \frac{\sum_{k=1}^{q} f_{v,k}(\mathbf{x}_k)^{-1}}{\sum_{k=1}^{q} f_{v,k}(\mathbf{x}_k)^{-1}\tilde{p}_h(\mathbf{x_k})^{\beta_s-1}}, \tag{17}$$

where $\mathbf{x}_k$ denotes the sample at iteration k and not the kth sedimentation time x_k.

As the iterative scheme can be fragile in early iterations, we use a slightly different initialization than the original article. Instead of starting with $f_{v,0} = 1$ and initializing the run in one of the high–probability regions (Invernizzi et al., 2020), we start with an initial guess for $v_{s,0}$ and allow the sampler to start from anywhere, partly to avoid the practical difficulty of finding high–probability regions but also to ensure that the results are not affected by the starting point. Details are given in Appendix B1.

Once samples have been generated from the biased distribution, we reweight them to obtain draws from the posterior. In particular, the expected value of an observable $O(\mathbf{x})$ under the posterior can be approximated as:

$$\mathbf{E}[O(\mathbf{x})]_p \approx \frac{\sum_{q=1}^{n_s} O(\mathbf{x}_q) f_{v,q}(\mathbf{x}_q)^{-1}}{\sum_{q=1}^{n_s} f_{v,q}(\mathbf{x}_q)^{-1}}, \tag{18}$$

where n_s denotes the number of samples. Therefore, sample q for $q = 1, \ldots, n_s$ is assigned weight $\xi_q = 1/f_{v,q}(\mathbf{x}_q)$ and is resampled with replacement and probability $p_q = \xi_q / \sum_q \xi_q$. Uncertainties are estimated using the jackknife method (Miller, 1974).

3.1.3 Parallel Tempering (PT)

Like ST, PT samples from the S probability distributions $p_s(\mathbf{x})$ in Eq. 6, associated with different predetermined “temperatures” T_s. However, PT samples from these distributions in parallel. Thus, we define the Markov chain on the state space $\bar{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_S) \in \mathbf{R}^{K \times S}$, which consists of one state, $\mathbf{x}_s$, for each temperature. The chain is constructed to converge to the joint stationary distribution:

$$p_{joint}(\bar{\mathbf{x}}) = \prod_{s=1}^{S} p_s(\mathbf{x}_s), \tag{19}$$

which assigns equal weight to all temperatures. Although this can be done by simply running parallel Markov chains for the different temperatures, using, e.g., HMC, PT adds a second type of move. This is a Metropolis–controlled exchange of states between a pair of temperatures T_s and $T_{s'}$ that allows the state of one chain to jump to the state of the other in a single step, potentially leading to longer moves and faster convergence.

The swap is accepted with probability:

$$\alpha(\mathbf{x}_s \leftrightarrow \mathbf{x}_{s'}) = \min\left[1, \frac{\tilde{p}_s(\mathbf{x}_{s'})\tilde{p}_{s'}(\mathbf{x}_s)}{\tilde{p}_s(\mathbf{x}_s)\tilde{p}_{s'}(\mathbf{x}_{s'})}\right]. \tag{20}$$

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

To have a reasonable acceptance rate, the proposed swaps are typically between neighboring temperatures and in our computations, we follow the swap scheme of Syed et al. (2022), in which all temperature pairs $s, s+1$ are considered with even (odd) s in even (odd) iterations.

PT differs from ST and OPES in that it does not require an estimate of the normalization constants Z_s. Samples from the true posterior can be obtained without reweighting by considering samples with $s = 1$, which correspond to the lowest temperature. To account for autocorrelation and compute statistical errors, standard block analysis can be used. This is the same as for ST, but allows for a simpler post–processing scheme compared to OPES.

3.2 Partial tempering for age–depth modeling

Having introduced the tempering methods in the previous section, an important aspect to note in Eq. 6 is that we do not have to temper the full distribution. For example, one can choose $p_e(\mathbf{x})$ as the prior distribution (e.g. Eichenseer et al., 2025; Minson et al., 2013; Mosegaard and Tarantola, 1995; Sambridge, 2014), which, compared to tempering the full distribution, proves computationally necessary for our setup. Taking this one step further, in age–depth modeling, the part of the likelihood that contains direct dates is often unimodal and can be sampled using standard HMC (see Fig. 1), whereas the indirect data typically induce multimodality. While tempering the full likelihood can improve sampling, the high temperatures needed to escape the narrow valleys of the indirect likelihood tend to weaken constraints from absolute ages in the direct likelihood, causing many high–temperature samples to be spent in low–probability regions of the broad prior. Thus, introducing the direct likelihood $\tilde{p}_d(\mathbf{x}) = \prod_n^N p(y_n|\mathbf{x})$ and the indirect likelihood $\tilde{p}_i(\mathbf{x}) = \prod_m^M p(z_m|\mathbf{x})$, we suggest choosing:

$$\begin{aligned} \tilde{p}_e(\mathbf{x}) &= \pi(\mathbf{x})\tilde{p}_d(\mathbf{x}), \\ \tilde{p}_h(\mathbf{x}) &= \tilde{p}_i(\mathbf{x}). \end{aligned} \tag{21}$$

4 Results and discussion

4.1 Resulting age–depth models

Our implementations of OPES and PT are available in a C-Python framework at Kjellson (2026). Starting with random draws from the prior, both OPES and PT with partial tempering converge for all datasets after around 10,000 iterations, and running 100,000 iterations takes around an hour on a standard desktop computer. Parameters and performance measures are presented in Table B1–B4 in Appendix B2, and choosing the samples from chain $s = 1$ for PT, we obtain the age–depth models and the corresponding marginal densities presented in Fig. 2. As OPES and PT yield virtually identical plots, the OPES results, based on reweighted samples from all chains, are deferred to Appendix A. Note that for both PT and OPES, the Wah Shikar Cave age–depth models are presented for temperature $T_{22} \approx 94$ to emphasize that there are multiple modes, even though the samples for $s = 1$ align almost perfectly with the original age–depth curve. This is due to the large number of low–uncertainty δ^{18}O data.

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

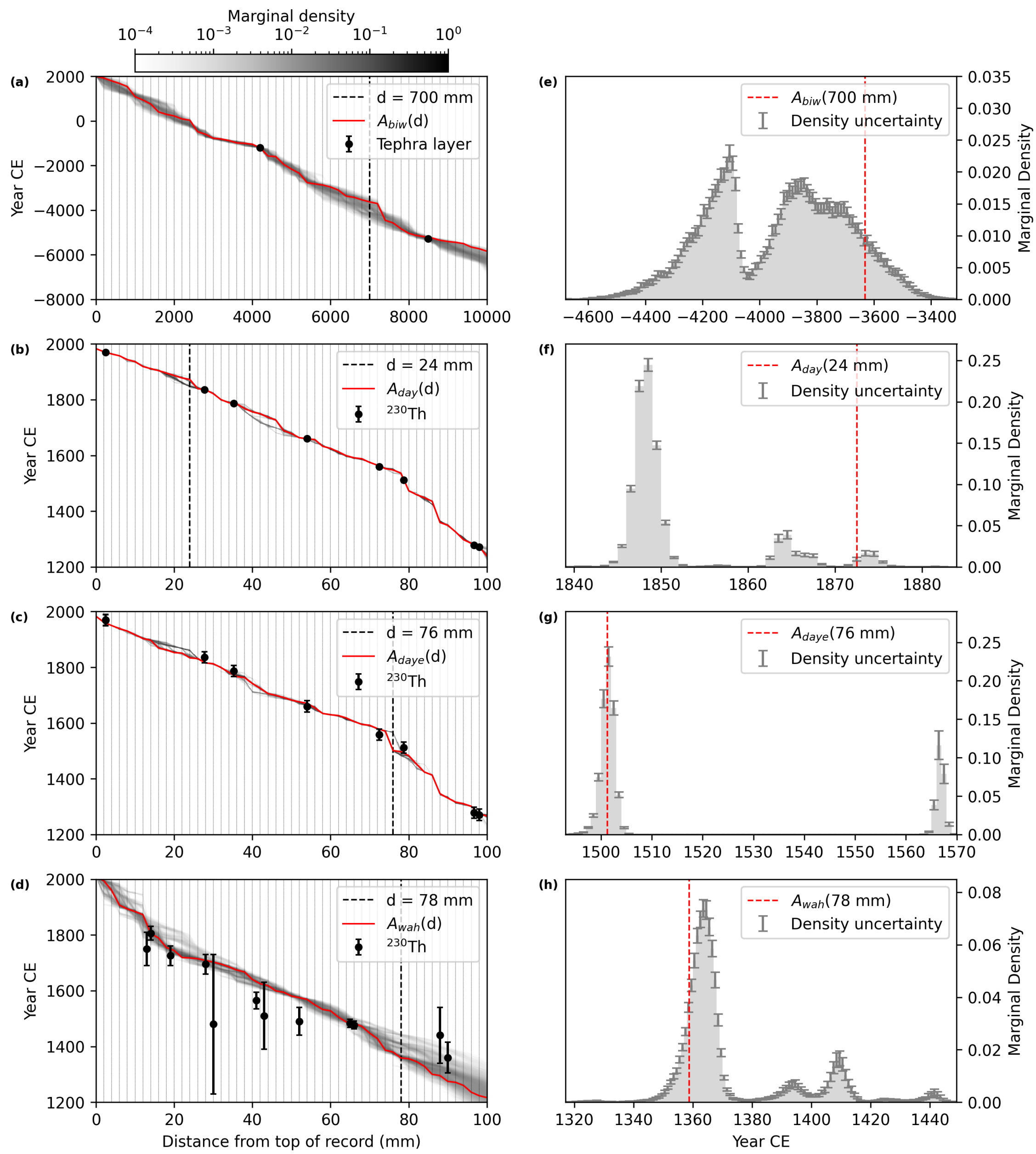

Figure 2. In (a)–(c) we present age–depth models corresponding to the $s = 1$ PT samples for the Lake Biwa, Dayu Cave, and modified Dayu Cave datasets. In (d), we present the age–depth models corresponding to the $s = 22$ PT samples for the Wah Shikar Cave dataset. The age–depth models are plotted as histograms for every $0.1\,\text{mm}$, with bins of width $1\,\text{yr}$, for the two caves, and for every $1\,\text{cm}$, with bins of width $10\,\text{yr}$, for the Lake Biwa sediment core. The colorbar, which is in log–scale to better show less likely solutions, represents the marginal density. In (e)–(h), this is specifically presented along depths where there appeared to be multimodality. Again, in the order of the Lake Biwa sediment core, the Dayu Cave and the Wah Shikhar Cave datasets. In all panels, $A_{dataset}(d)$ is the true age–depth model.

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

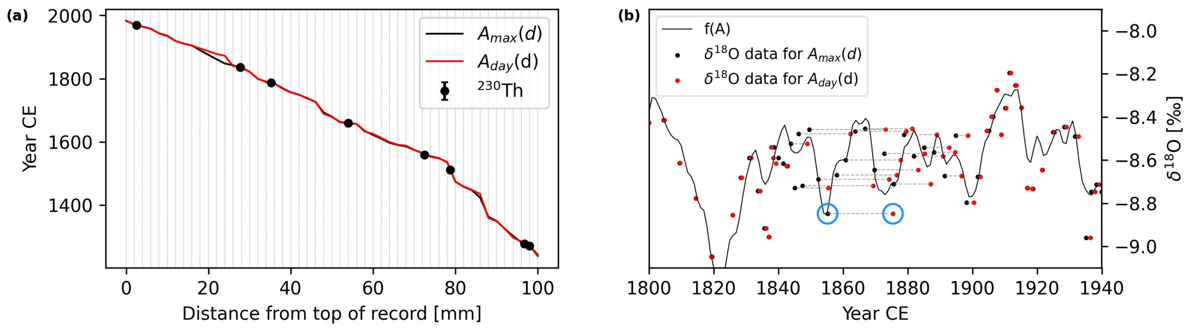

Figure 3. In (a) we show the most likely solution found during sampling, $A_{max}(d)$, and the "true" age–depth model, $A_{day}(d)$. The Dayu Cave data, based on the most likely solution $A_{max}(d)$ and the true age–depth model $A_{day}(d)$, are shown in (b). With some imagination, it can be seen that the $A_{max}(d)$ data is closer to the low-pass filtered ECHAM5-wiso simulation, $f(A)$, than the $A_{day}(d)$ data. An interesting point where the two age–depth models place one datapoint at two completely different ages is marked with blue circles.

Given the resulting age–depth models, we would first like to focus on the Dayu Cave age–depth model with the original ^{230}Th dates in Fig. 2b and 2f. An important observation, particularly visible in Fig. 2f, is that the true age–depth curve, $A_{day}(d)$, from which the δ^{18}O data is created, does not coincide with the most likely solution. Figure 3 illustrates this, and shows how the data, when mapped through the most likely solution, $A_{max}(d)$, align more closely with the reference function than when mapped through $A_{day}(d)$. For instance, the data point at around 1875 CE has been perturbed from the reference function to the extent that it fits better at around 1855 CE. As the true solution hides in a quite unlikely mode, this highlights the importance of finding all solutions rather than focusing only on the dominant mode.

4.2 Comparison of methods

To emphasize the need for partial tempering to find these modes, we compare HMC, PT, and OPES across all datasets, applying both partial and full tempering of the likelihood to PT and OPES. The results are presented in Fig. 4, which illustrates the convergence of the mean and 95% quantile potential energy, $u(\mathbf{x}) = -\log(\tilde{p}(\mathbf{x}))$, as a function of iteration for the different samplers. The energy for the OPES runs is computed using Eq. 18, without jackknife analysis for simplicity.

Starting with the Lake Biwa sediment core, all samplers except for HMC successfully sample the posterior in Fig. 4a. The constant energy offset of the HMC sampler comes from one of the chains diverging to a local minimum and all other chains converging to the same global minimum as the tempered samplers. So, while there are issues with HMC for the Lake Biwa sediment core, they are not as problematic as for the other datasets. For the Dayu Cave dataset in Fig. 4b, all HMC chains diverge to different local minima, but both full and partial tempering are sufficient for the samplers to converge properly. It is first for the Dayu Cave dataset with inflated ^{230}Th uncertainties and the Wah Shikhar Cave dataset in Fig. 4c-d, that partial tempering clearly exceeds both HMC and full–likelihood tempering.

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

Hence, sufficient sampling is heavily dependent on the dataset. The large uncertainties of the declination data, together with the relatively slow variations in the reference function for the Lake Biwa sediment core, yield a broad range of possible age–depth models and a smooth posterior landscape. Fig 2e shows that the posterior has multimodalities. Still, the high–likelihood regions are separated by quite likely regions, making it possible for the HMC sampler to traverse between modes. The Dayu Cave dataset presents greater sampling challenges, as the δ^{18}O data have lower uncertainties and follow the fast variations of the reference function, leading to a posterior distribution with modes separated by highly unlikely regions. Although tempering is required for the samplers to converge, the strong ^{230}Th dates constrain the high–temperature samples to the ^{230}Th likelihood, even with full tempering, so these samplers perform as well as those with partial tempering.

The situation is different for the edited Dayu Cave dataset. With larger and more realistic uncertainties for the ^{230}Th dates, the high–temperature samples of the full–likelihood samplers are drawn from irrelevant regions of the prior. This is because the high temperatures required to escape the narrow δ^{18}O modes effectively remove the information from the ^{230}Th dates. As the high–probability regions are narrow and difficult to find, partial tempering proves necessary in this setup. With large uncertainties for the ^{230}Th dates and dense low–uncertainty δ^{18}O data, the situation is similar for the Wah Shikar Cave dataset. In fact, the δ^{18}O data is even denser than in the Dayu Cave dataset, and the situation for the full–likelihood tempered samplers is further worsened by the mismatch between prior and posterior mass (see Fig. 1d), making the high–temperature samples completely irrelevant for the very deep highest–likelihood mode. While one can question the choice of prior and potentially promote a prior that better aligns with the data, we believe that this is a quite realistic choice if the goal is to keep the prior unbiased with respect to the data.

One could imagine similar issues with partial tempering, in particular if the direct part of the likelihood does not "contain" the indirect part. While such a mismatch between data types points towards errors in uncertainty estimation, there could be situations where the indirect data are most likely in the tails of the direct part of the likelihood. To resolve this issue, one practical solution, apart from revising the data uncertainties, could be to use a scaled version of the direct part of the likelihood in $\tilde{p}_e(\mathbf{x})$ and account for this in $\tilde{p}_h(\mathbf{x})$.

In our examples, direct and indirect data align, and convergence depends more on parameter tuning. While OPES requires additional and less straightforward tuning of $v_{s,0}$, the PT setup is relatively straightforward, and once the HMC settings and the temperature ladder are chosen, the method behaves robustly. On the other hand, OPES has the advantage that it can be run with fewer chains than there are temperatures. Many chains are needed to obtain a stable estimate of v_s in our runs, but once it converges (after ~10,000 samples), it can be fixed, allowing for multiple independent runs for any number of chains. In general, both methods have advantages and disadvantages, but the main point is that they converge to the same posterior distributions and do so more quickly when partial tempering is applied.

4.3 Other methods

In addition to OPES and PT, we briefly mention two other enhanced sampling methods: Sequential Monte Carlo (SMC; PyMC Contributors, 2026), a tempering method related to Annealed Importance Sampling (Neal, 2001), and Nested Sampling (NS; Albert, 2020, 2024; Koposov et al., 2025; Skilling, 2006; Speagle, 2020), which does not use tempering but still follows a

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

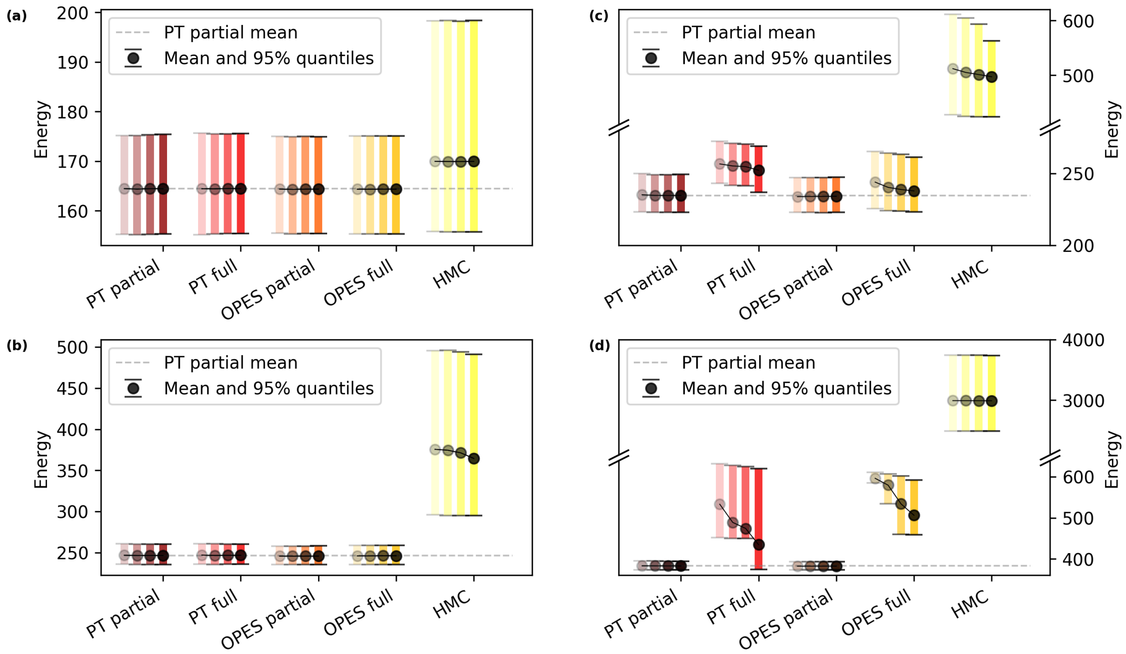

Figure 4. Comparison of HMC, OPES, and PT (the latter two with and without partial tempering) for the Lake Biwa sediment core in (a), the Dayu Cave stalagmite in (b), the Dayu Cave stalagmite with inflated ^{230}Th uncertainties in (c), and the Wah Shikhar Cave stalagmite in (d). The plots show the mean potential energy ($u(\mathbf{x}) = -\log(\tilde{p}(\mathbf{x}))$ as dots and the 95% quantiles as colored error bars, excluding a 10,000-sample cutout. The mean is computed over all chains for HMC and all temperatures but reweighted for OPES. For each method, the four error bars represent the mean and quantiles, using the first 10%, 25%, 50%, and 100% of the data, from left to right and from shaded to full color. For Fig. (c)–(d), we cut the y–axis because the HMC sample energies are far too high.

procedure similar to SMC. In our examples, when applied with standard settings and recommended tuning (Ashton et al., 2022; PyMC Contributors, 2026), neither approach reliably found the relevant modes and weighted them properly. One likely reason is that the considered variants do not exploit gradients to the same extent as HMC samplers, and the SMC version we used does not allow for straightforward partial tempering. In addition, NS is more sensitive to dimensionality than MCMC samplers and typically benefits from a good alignment between prior and posterior mass (Ashton et al., 2022; Koposov et al., 2026). To improve this alignment, an approximate distribution can be fitted to the samples obtained using only direct dates and then used as a prior when incorporating indirect data, similarly to Ashton et al. (2022). Applied to our datasets, this helped identify plausible modes, but their relative weights remained inaccurate.

As our main focus here is OPES and PT, we believe that there is still room for further investigation of NS, SMC, and other methods. However, we would say that OPES and PT with partial tempering constitute two good basic methods for this application. They can both be run in parallel, and while they require tuning, it is not too extensive, particularly for PT. In the future, our goal is to implement PT and/or OPES with partial tempering in a more user–friendly framework. With this, we hope

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

to enable simpler access to good age–depth modeling in a multitude of applications, ultimately leading to improved dating and better synchronization of records locally and globally.

5 Conclusions

This article shows how to apply the methods of OPES and PT to the high–dimensional multimodal posteriors that arise in the field of age–depth modeling. Once v_s has converged, OPES enables running multiple independent chains simultaneously, each providing samples from the required distribution. On the other hand, it demands more tuning, and for difficult problems, a simple PT scheme is often the more practical choice. An important finding is that, for challenging problems, it is important to temper only the parts of the likelihood that create multimodality, as this improves exploration and reduces the time spent in low–likelihood regions of the prior. With more robust age–depth modeling, we hope to open opportunities across applications such as paleoclimate and paleomagnetic field reconstructions, where variable stratigraphic data is used.

Code and data availability. Code and data are available at Kjellson (2026). So far, the code is written specifically for these datasets. A more general framework will hopefully be available soon.

Appendix A: OPES results

We present the resulting age–depth models for the OPES runs in Fig. A1. The samples are based on all chains and have been reweighted to the lowest temperature $T_1 = 1$ for the Lake Biwa, Dayu Cave, and edited Dayu Cave datasets, while they have been reweighted to $T_{22} \approx 94$ for the Wah Shikhar Cave dataset.

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

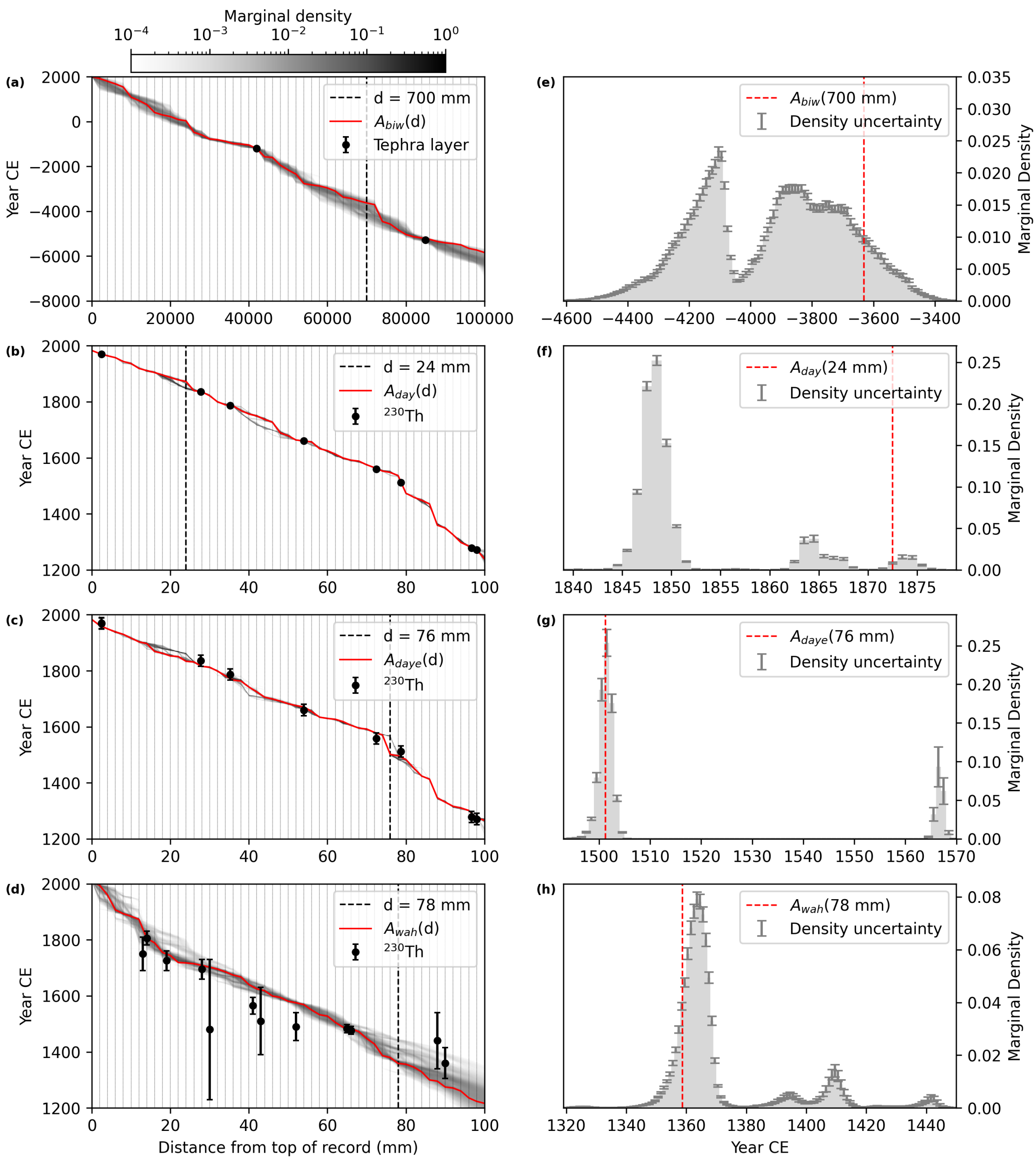

Figure A1. In (a)–(c) we present age–depth models corresponding to the $s = 1$ OPES samples for the Lake Biwa, Dayu Cave, and modified Dayu Cave datasets. In (d), we present the age depth models corresponding to the $s = 22$ OPES samples for the Wah Shikar Cave dataset. The age–depth models are plotted as histograms for every $0.1\,\mathrm{mm}$, with bins of width $1\,\mathrm{yr}$, for the two caves, and for every $1\,\mathrm{cm}$, with bins of width $10\,\mathrm{yr}$, for the Lake Biwa sediment core. The colorbar, which is in log–scale to better show less likely solutions, represents the marginal density. In (e)–(h), this is specifically presented along depths where there appeared to be multimodality. Again, in the order of the Lake Biwa sediment core, the Dayu Cave and the Wah Shikhar Cave datasets. In all panels, $A_{dataset}(d)$ is the true age–depth model.

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

Appendix B: Computational details

B1 Initializing OPES

Invernizzi et al. (2020) advises to start the OPES simulations in a high–probability region with $f_{v,0}(\mathbf{x}) = 1$. As we want to ensure that starting from one basin does not bias the results, and since it is not an easy task to find these regions, particularly for the Wah Shikar Cave stalagmite, we instead start with an initial estimate of $v_{s,0}$:

$$v_{s,0} = \exp\left(\frac{n_r}{2}\left(\beta_s - 1\right)\right). \tag{B1}$$

This estimate follows from Eq. 17 with $f_v(\mathbf{x}) = 1$ and the assumption that the lowest–temperature samples correspond to a timescale where the indirect data are roughly one standard deviation away from the reference function, yielding the approximation $\tilde{p}_h(\mathbf{x}) \approx \exp\left(-\frac{n_r}{2}\right)$, where n_r is the number of indirect data measurements.

Furthermore, since the iterative update of $v_{s,q}$ tends to diverge at the start, particularly if high–probability samples are not found fast enough, we, in practice, use the update:

$$v_{s,q} = \frac{n_{init} + \sum_{k=k_0}^{q} f_{v,k}(\mathbf{x}_k)^{-1}}{n_{init} v_{s,0} + \sum_{k=k_0}^{q} f_{v,k}(\mathbf{x}_k)^{-1} \tilde{p}_h(\mathbf{x_k})^{\beta_s - 1}}, \tag{B2}$$

instead of Eq. 17. Denoting the number of samples with n_s, we set $k_0 = \min[1, q - n_s/100]$. This makes $v_{s,q}$ more variable and adaptive. To allow for convergence after the initial samples, we then fix $k_0 = \frac{n_{ch}}{10} q_{ta} - n_s/100$ after $\frac{n_{ch}}{10} q_{ta}$ samples. This is a valid rule since $\frac{n_{ch}}{10} \geq 1$ in all our examples. If this is not the case, some other rule should be applied. q_{ta} is the sample index at which $v_{s,q}$ first starts to increase, n_{ch} is the number of independent chains contributing to $v_{s,q}$, and we choose $n_{init} = 0.0005\, n_s n_{ch}$ empirically. n_{init} can be interpreted as an initial number of samples with $f_{v,q} = 1$ and $\tilde{p}_h(\mathbf{x_k})^{\beta_s - 1} = v_{s,0}$. While this term biases $v_{s,q}$, it stabilizes the early evolution and has little effect on the converged value.

B2 Parameters and performance measures

In Table B1–B4, we present some important parameters and performance measures for the HMC and partial tempering OPES, and PT runs. n_t denotes the number of time steps in each HMC iteration, Δt the length of each time step, and n_c the number of cutout samples due to initial transients (this is always chosen as $n_c \geq$10,000). The number of chains for OPES, n_{ch}, is not presented as we choose it as $n_{ch} = S$, and the time step and number of time steps are adapted to achieve an acceptance rate of around 0.65–0.95. This is not always possible for HMC, as the chains diverge to different local modes. Then, the acceptance rate is set too high rather than too low. For the performance measures, $\hat{r}$ denotes the Gelman–Rubin convergence statistic, with values of $\hat{r} > 1$ indicating non-convergence, ESS_ρ, the standard effective sample size that accounts for autocorrelation, computed using Martin et al. (2026), and ESS_w, the effective sample size given in Invernizzi et al. (2020), which accounts for magnitude differences in the weights. Both ESS_w and ESS_ρ are always less than or equal to n_s, but should be as close as possible to n_s for good performance. Note that for OPES, the actual effective sample size is lower than ESS_ρ and ESS_w separately, because they account for different effects.

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

Table B1. Discretization for the Lake Biwa, Dayu Cave (unedited and edited), and Wah Shikhar Cave datasets

	Biw	**Day**	**Daye**	**Wah**
K	50	50	50	50
d_K	$1{,}000\,\mathrm{cm}$	$100\,\mathrm{mm}$	$100\,\mathrm{mm}$	$100\,\mathrm{mm}$

Table B2. HMC parameters and performance measures for the Lake Biwa, Dayu Cave (unedited and edited), and Wah Shikhar Cave datasets

	Biw	**Day**	**Daye**	**Wah**
n_t	700	700	700	700
Δt	0.0005	0.0001	0.001	0.0001
n_s	100,000	100,000	100,000	100,000
n_c	10,000	10,000	10,000	10,000
max $\hat{r}$	1.44	3.18	3.18	3.68
min ESS_ρ $\forall$ chains	11	6	6	5

Table B3. OPES and PT parameters for the Lake Biwa, Dayu Cave (unedited and edited), and Wah Shikhar Cave datasets

	Biw	**Day**	**Daye**	**Wah**
T_S	20	20	20	400
S	20	20	20	30
Temperature ladder	geometric	geometric	geometric	geometric
n_t	700	700	700	700
Δt	0.001	0.001	0.001	0.0003
n_s	100,000	100,000	100,000	100,000
n_c	10,000	10,000	10,000	10,000

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

Table B4. OPES and PT performance measures for the Lake Biwa, Dayu Cave (unedited and edited), and Wah Shikhar Cave datasets

	Biw	**Day**	**Daye**	**Wah**
OPES				
max $\hat{r}$	1.00	1.00	1.01	1.08
min ESS_ρ $\forall$ chains	147,801	6,603	2,593	328
ESS_w $\forall$ chains	305,259	182,615	193,887	165,462
PT				
Mean # of round trips	267	44	23	2
Swap rejection rate	0.05	0.2	0.2	0.3
min ESS_ρ for chain $s = 1$	13,671	1,371	374	3,009

These statistics show that HMC does not work for any dataset. For OPES, ESS_ρ and $\hat{r}$ are computed on the samples before reweighting. While the Wah Shikhar Cave dataset has $\hat{r} = 1.08$, suggesting slightly non-converging chains, the same value after reweighting is 1.0. Thus, the mixing is not optimal between higher temperatures, but the relevant samples are captured in all chains. Here, it could be beneficial to use an improved HMC sampler with a No-U-Turn Sampler (NUTS; Hoffman and Gelman, 2014) and adaptive step size to improve exploration. Similarly, for PT, the time step is constant across all temperatures, and to improve performance, one could make it temperature-dependent. Also, the rejection rate for PT is lower than suggested, for instance, in Syed et al. (2022), and performance could improve further with higher values. Finally, the ESS_ρ for the edited Dayu Cave dataset is notably low, and we could potentially benefit from slightly increasing the maximum temperature. However, while there is room for more tuning, our main point is to show the benefits of partial tempering for both OPES and PT. As our next step is to implement this in a more user–friendly setting, we leave tuning and optimization for later.

Author contributions. Andreas Nilsson, Anders Irbäck, and Hanna Kjellson designed the original project idea. Hanna Kjellson developed the code, performed the simulations, and prepared the manuscript, with continuous input from Andreas Nilsson and Anders Irbäck, both for writing and interpreting the results.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors would like to thank Max Arthus Schanner for insightful discussions and suggestions. The research was funded by the European Union (ERC, PALEOCORE, 101125394). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Albert, J. G.: JAXNS: a high-performance nested sampling package based on JAX, arXiv [preprint], https://doi.org/10.48550/arXiv.2012.15286, 2020.

Albert, J. G.: JAXNS: Probabilistic Programming and Nested sampling in JAX, GitHub [code], https://github.com/Joshuaalbert/jaxns, (last access: 2026-05-27), 2024.

Ali, M., Oda, H., Hayashida, A., Takemura, K., and Torii, M.: Holocene palaeomagnetic secular variation at Lake Biwa, central Japan, Geophysical Journal International, 136, 218–228, https://doi.org/10.1046/J.1365-246X.1999.00718.X, 1999.

Ashton, G., Bernstein, N., Buchner, J., et al.: Nested sampling for physical scientists, Nat. Rev. Methods Primers, 2, 39, https://doi.org/10.1038/s43586-022-00121-x, 2022.

Aubert, J., Gastine, T., and Fournier, A.: Spherical convective dynamos in the rapidly rotating asymptotic regime, Journal of Fluid Mechanics, 813, 558–593, https://doi.org/10.1017/JFM.2016.789, 2017.

Axelsson, J.: Interglacial climates in proxies and models: Utilizing sampled oxygen isotopes and model simulations to understand past Indian summer monsoon variability, Ph.D. thesis, Stockholm University, https://doi.org/10.13140/RG.2.2.16845.86242, 2024.

Blaauw, M. and Christen, J. A.: Flexible paleoclimate age–depth models using an autoregressive gamma process, Bayesian Analysis, 6, 457–474, https://doi.org/10.1214/11-BA618, 2011.

Blaauw, M., Christen, J. A., Lopez, M. A. A., Vazquez, J. E., V., O. M. G., Belding, T., Theiler, J., Gough, B., and Karney, C.: Package 'rbacon', https://CRAN.R-project.org/package=rbacon, 2026.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D.: Hybrid Monte Carlo, Physics Letters B, 195, 216–222, https://doi.org/10.1016/0370-2693(87)91197-X, 1987.

Edmonsond, S. and Dyer, B.: A Bayesian framework for inferring regional and global change from stratigraphic proxy records (StratMC v1.0), Geoscientific Model Development, 18, 4759–4788, https://doi.org/10.5194/gmd-18-4759-2025, 2025.

Eichenseer, K., Sinnesael, M., Smith, M. R., and Millard, A. R.: StratoBayes: a Bayesian method for automated stratigraphic correlation and age modelling, Geochronology, 7, 545–570, https://doi.org/10.5194/gchron-7-545-2025, 2025.

Haslett, J. and Parnell, A. C.: A simple monotone process with application to radiocarbon-dated depth chronologies, Journal of the Royal Statistical Society: Series C (Applied Statistics), 57, 399–418, https://doi.org/10.1111/j.1467-9876.2008.00623.x, 2008.

Hoffman, M. D. and Gelman, A.: The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, Journal of Machine Learning Research, 15, 1593–1623, http://jmlr.org/papers/v15/hoffman14a.html, 2014.

Invernizzi, M., Piaggi, P. M., and Parrinello, M.: Unified Approach to Enhanced Sampling, Phys. Rev. X, 10, 041 034, https://doi.org/10.1103/PhysRevX.10.041034, 2020.

Katzgraber, H. G., Trebst, S., Huse, D. A., and Troyer, M.: Feedback-optimized parallel tempering Monte Carlo, Journal of Statistical Mechanics: Theory and Experiment, 2006, P03 018–P03 018, https://doi.org/10.1088/1742-5468/2006/03/p03018, 2006.

Kjellson, H.: hannakjellson/sample-age-depth-models: Implementation of OPES and PT for the preliminary article "Overcoming multi-modalities in age–depth model posteriors using On–the–fly Probability Enhanced Sampling and Parallel Tempering", Zenodo [code], https://doi.org/10.5281/zenodo.20507158, 2026.

Koposov, S., Speagle, J., Barbary, K., Ashton, G., Bennett, E., Buchner, J., Scheffler, C., Talbot, C., Cook, B., Guillochon, J., Cubillos, P., Ramos, A. A., Dartiailh, M., Ilya, Tollerud, E., Lang, D., Johnson, B., jtmendel, Higson, E., Vandal, T., Daylan, T., Angus,

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

R., patelR, Cargile, P., Sheehan, P., Pitkin, M., Kirk, M., Xu, L., Leja, J., and joezuntz: joshspeagle/dynesty: v3.0.0, Zenodo [code], https://doi.org/10.5281/zenodo.17268284, 2025.

Koposov, S., Speagle, J. S., et al.: dynesty v3.0.0 documentation: User guide and recommendations, https://dynesty.readthedocs.io/en/stable/, (last access: 2026-04-09), 2026.

Lee, T., Rand, D., Lisiecki, L. E., Gebbie, G., and Lawrence, C.: Bayesian age models and stacks: combining age inferences from radiocarbon and benthic δ^{18}O stratigraphic alignment, Climate of the Past, 19, 1993–2012, https://doi.org/10.5194/cp-19-1993-2023, 2023.

Lyubartsev, A. P., Martsinovski, A. A., Shevkunov, S. V., and Vorontsov-Velyaminov, P. N.: New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles, The Journal of Chemical Physics, 96, 1776–1783, https://doi.org/10.1063/1.462133, 1992.

Marinari, E. and Parisi, G.: Simulated tempering: a new Monte Carlo scheme, EPL, 19, 451, https://doi.org/10.1209/0295-5075/19/6/002, 1992.

Martin, O. A., Abril-Pla, O., Deklerk, J., Axen, S. D., Carroll, C., Hartikainen, A., and Vehtari, A.: ArviZ: a modular and flexible library for exploratory analysis of Bayesian models, Journal of Open Source Software, 11, 9889, https://doi.org/10.21105/joss.09889, 2026.

Miller, R. G.: The jackknife-a review, Biometrika, 61, 1–15, https://doi.org/10.1093/biomet/61.1.1, 1974.

Minson, S. E., Simons, M., and Beck, J. L.: Bayesian inversion for finite fault earthquake source models I—-theory and algorithm, Geophysical Journal International, 194, 1701–1726, https://doi.org/10.1093/gji/ggt180, 2013.

Mosegaard, K. and Tarantola, A.: Monte Carlo sampling of solutions to inverse problems, Journal of Geophysical Research: Solid Earth, 100, 12 431–12 447, https://doi.org/https://doi.org/10.1029/94JB03097, 1995.

Neal, R. M.: Annealed importance sampling, Statistics and Computing, 11, 125–139, https://doi.org/10.1023/A:1008923215028, 2001.

Nilsson, A. and Suttie, N.: Probabilistic approach to geomagnetic field modelling of data with age uncertainties and post–depositional magnetisations, Physics of the Earth and Planetary Interiors, 317, 106 737, https://doi.org/https://doi.org/10.1016/j.pepi.2021.106737, 2021.

Nilsson, A., Suttie, N., Stoner, J., and Muscheler, R.: Recurrent ancient geomagnetic field anomalies shed light on future evolution of the South Atlantic Anomaly, Proceedings of the National Academy of Sciences, 119, e2200749 119, https://doi.org/10.1073/pnas.2200749119, 2022.

Nilsson, A., Suttie, N., Troyano, M., Gillet, N., Aubert, J., and Irbäck, A.: Joint Bayesian inference of Earth's magnetic field and core surface flow on millennial timescales, arXiv [preprint], https://doi.org/10.48550/ARXIV.2601.10344, 2026.

Nymeyer, H.: Serial tempering without exchange, The Journal of Chemical Physics, 133, 114 113, https://doi.org/10.1063/1.3480013, 2010.

Parnell, A. C., Sweeney, J., Doan, T. K., Salter-Townshend, M., Allen, J. R. M., Huntley, B., and Haslett, J.: Bayesian inference for palaeoclimate with time uncertainty and stochastic volatility, Journal of the Royal Statistical Society Series C: Applied Statistics, 64, 115–138, https://doi.org/10.1111/rssc.12065, 2015.

Parrenin, F., Bouchet, M., Buizert, C., Capron, E., Corrick, E., Drysdale, R., Kawamura, K., Landais, A., Mulvaney, R., Oyabu, I., and Rasmussen, S. O.: The Paleochrono-1.1 probabilistic model to derive a common age model for several paleoclimatic sites using absolute and relative dating constraints, Geoscientific Model Development, 17, 8735–8750, https://doi.org/10.5194/gmd-17-8735-2024, 2024.

PyMC Contributors: Sequential Monte Carlo, Zenodo [code], https://doi.org/10.5281/zenodo.5654871, 2026.

Ramsey, C. B.: Deposition models for chronological records, Quaternary Science Reviews, 27, 42–60, https://doi.org/https://doi.org/10.1016/j.quascirev.2007.01.019, 2008.

Salakhutdinov, R.: Learning deep boltzmann machines using adaptive MCMC, in: Proceedings of the 27th International Conference on International Conference on Machine Learning, pp. 943—-950, Madison, WI, USA, ISBN 9781605589077, 2010.

https://doi.org/10.5194/egusphere-2026-3766
Preprint. Discussion started: 15 July 2026

Sambridge, M.: A Parallel Tempering algorithm for probabilistic sampling and multimodal optimization, Geophysical Journal International, 196, 357–374, https://doi.org/10.1093/gji/ggt342, 2014.

Schorndorf, N., Warken, S. F., Eichstädter, R., Mielke, A. S., Avilés Olguín, J., Keppler, F., Hennhöfer, D., Tec Pool, F., Evia, C., Gómez, M. J., Stinnesbeck, W., and Frank, N.: Challenges of initial Thorium and approaches to robust speleothem age models: A case study from the Yucatán peninsula, Mexico, EGUsphere, 2026, 1–31, https://doi.org/10.5194/egusphere-2026-2284, 2026.

Sinha, A., Stott, L., Mudelsee, M., Cheng, H., and Biswas, J.: The leading mode of Indian Summer Monsoon precipitation variability during the last millennium, Geophysical Research Letters - GEOPHYS RES LETT, 38, https://doi.org/10.1029/2011GL047713, 2011.

Sjolte, J., Sturm, C., Adolphi, F., Vinther, B. M., Werner, M., Lohmann, G., and Muscheler, R.: Solar and volcanic forcing of North Atlantic climate inferred from a process-based reconstruction, Climate of the Past, 14, 1179 – 1194, https://doi.org/10.5194/cp-14-1179-2018, 2018.

Skilling, J.: Nested sampling for general Bayesian computation, Bayesian Analysis, 1, 833–859, https://doi.org/10.1214/06-BA127, 2006.

Speagle, J. S.: DYNESTY: a dynamic nested sampling package for estimating Bayesian posteriors and evidences, Monthly Notices of the Royal Astronomical Society, 493, 3132–3158, https://doi.org/10.1093/mnras/staa278, 2020.

Swendsen, R. H. and Wang, J.-S.: Replica Monte Carlo simulation of spin-glasses, Phys. Rev. Lett., 57, 2607–2609, https://doi.org/10.1103/PhysRevLett.57.2607, 1986.

Syed, S., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A.: Non-reversible Parallel Tempering: A scalable highly parallel MCMC scheme, Journal of the Royal Statistical Society Series B: Statistical Methodology, 84, 321–350, https://doi.org/10.1111/rssb.12464, 2022.

Tan, L., Cai, Y., Cheng, H., An, Z., and Edwards, R. L.: Summer monsoon precipitation variations in central China over the past 750 years derived from a high–resolution absolute–dated stalagmite, Palaeogeography, Palaeoclimatology, Palaeoecology, 280, 432–439, https://doi.org/10.1016/j.palaeo.2009.06.030, 2009.

Torrie, G. M. and Valleau, J. P.: Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella Sampling, Journal of Computational Physics, 23, 187 – 199, https://doi.org/10.1016/0021-9991(77)90121-8, 1977.

Trachsel, M. and Telford, R.: All age-depth models are wrong, but are getting better, The Holocene, 27, https://doi.org/10.1177/0959683616675939, 2016.

Wang, F. and Landau, D.: Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States, Physical Review Letters, 86, 2050 – 2053, https://doi.org/10.1103/physrevlett.86.2050, 2001.