

Supplement to "Synchronization of layer-counted paleoclimatic proxy archives using a Bayesian regression modeling framework"

Eirik Myrsvoll-Nilsen ^{*,†} 

Keno Riechers ^{†,‡}  and Niklas Boers ^{‡,†,§} 

This supplementary material contains the following:

- Section A: Shows how to obtain the cumulative age distribution from Gaussian layer increment models.
- Section B: Explains how Gaussian age distributions can be synchronized to tie-points, assuming no unknown biases.
- Section C: Demonstrates how the layer increment and age discrepancy models can be fitted simultaneously using an MCMC approach on simulated data.

A Joint age distribution of Gaussian models

Assume that the underlying layer-increment model is a Gaussian process defined by some fixed effects β and some additional hyperparameters θ , and that there is no reason to suspect a bias in the layer-counted data $\mathbf{y}$. Then it is possible to define a computationally efficient framework for incorporating tie-points to constrain the age uncertainty expressed by the posterior predictive distribution

$$\pi(\mathbf{y}^* | \Delta \mathbf{y}) = \pi(\mathbf{y}^* | \beta, \theta) \pi(\beta, \theta | \Delta \mathbf{y}). \quad (1)$$

In order to achieve this we require an analytical expression for $\pi(\mathbf{y} | \beta, \theta)$. By defining $A_k = y_0 + \sum_{i=1}^k a(w_i | \beta)$ the likelihood of $\mathbf{y}$ given (β, θ) can be expressed as

$$\mathbf{y} | \beta, \theta \sim \mathcal{N}(\boldsymbol{\mu}_y, \mathbf{Q}_y^{-1}) \quad (2)$$

with mean vector $\boldsymbol{\mu}_y = (A_1, \dots, A_n)^\top$ and some precision matrix $\mathbf{Q}_y$.

The precision matrix $\mathbf{Q}_y$ can be derived analytically without inverting $\boldsymbol{\Sigma}_y$ by expanding the joint age distribution $\pi(\mathbf{y} | \beta, \theta) = \pi(y_1 | \beta, \theta) \cdots \pi(y_n | y_1, \dots, y_{n-1}, \beta, \theta)$, and taking advantage of any conditional independence to simplify the expression. We

^{*}Department of Mathematics and Statistics, UiT –The Arctic University of Norway, Tromsø, Norway, eirik.myrsvoll-nilsen@uit.no

[†]Complexity Science, Potsdam Institute for Climate Impact Research, Potsdam, Germany

[‡]School of Engineering & Design, Technical University of Munich, Munich, Germany

[§]Global Systems Institute and Department of Mathematics, University of Exeter, Exeter, UK

demonstrate this procedure for the case where ε follows an AR(1) process. This yields a sparse precision matrix with bandwidth $b_y = 2$:

$$\mathbf{Q}_y = \frac{1}{\sigma_\varepsilon^2} \frac{1}{1 - \phi^2} \begin{pmatrix} 2 + 2\phi + \phi^2 & -(1 + \phi)^2 & \phi & & & \\ -(1 + \phi)^2 & 2(1 + \phi + \phi^2) & -(1 + \phi)^2 & \phi & & \\ \phi & -(1 + \phi)^2 & 2(1 + \phi + \phi^2) & -(1 + \phi)^2 & \phi & \\ & \ddots & \ddots & \ddots & \ddots & \\ & \phi & -(1 + \phi)^2 & 2(1 + \phi + \phi^2) & -(1 + \phi)^2 & \phi \\ & & \phi & -(1 + \phi)^2 & 2 + 2\phi + \phi^2 & -(1 + \phi) \\ & & & \phi & -(1 + \phi) & 1 \end{pmatrix}. \quad (3)$$

If the precision matrix for the layer increments, $\mathbf{Q}_{\Delta y}(\boldsymbol{\theta})$, is known, then the corresponding $\mathbf{Q}_y(\boldsymbol{\theta})$ can be computed for general Gaussian random fields as follows. For clarity, we assume $y_0 = 0$, and simplify the notation $\boldsymbol{\mu} := \boldsymbol{\mu}(\boldsymbol{\beta})$ and $\mathbf{Q} := \mathbf{Q}(\boldsymbol{\theta})$. The partial sum $y_k = \sum_{i=1}^k \Delta y_i$ can be expressed as the matrix product $\mathbf{y} = \mathbf{R}\Delta\mathbf{y}$, where the lower triangular matrix $\mathbf{R}$ and its inverse $\mathbf{S} = \mathbf{R}^{-1}$ are given by

$$\mathbf{R} = \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & \vdots & \ddots & \\ 1 & 1 & \dots & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{S} = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix}.$$

Finally, the change-of-variables formula implies

$$\pi(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = \mathcal{N}_n(\boldsymbol{\mu}_y, \mathbf{Q}_y) = \mathcal{N}_n(\mathbf{R}\boldsymbol{\mu}_{\Delta y}, \mathbf{S}^\top \mathbf{Q}_{\Delta y} \mathbf{S}). \quad (4)$$

which constitutes the desired analytical expression for the likelihood of the ages. We can then sample from the posterior predictive distribution for the ages

$$\pi(\mathbf{y}^* \mid \Delta\mathbf{y}) = \int \pi(\mathbf{y}^* \mid \boldsymbol{\beta}, \boldsymbol{\theta}) \pi(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \Delta\mathbf{y}) d\boldsymbol{\beta} d\boldsymbol{\theta} \quad (5)$$

using a Monte Carlo sampling scheme.

It can be shown that the precision matrix $\mathbf{Q}_y = \mathbf{S}^\top \mathbf{Q}_{\Delta y} \mathbf{S}$ will be a band matrix of bandwidth $b_y = b_{\Delta y} + 1$. Thus, if $\Delta\mathbf{y}$ can be described using a Gaussian Markov random field (GMRF), where the precision matrix is a band matrix with bandwidth $b_{\Delta y} \ll n$, then the ages $\mathbf{y}$ will also follow a GMRF with a banded precision matrix with equally $b_y \ll n$. This ensures that the log-likelihood can be computed efficiently (Rue, 2001) and that simulations can be obtained in linear time.

B Naïve constraining of Gaussian age uncertainties

Consider a proxy record together with the corresponding age–depth model of the form of Equation (4). Assume that for a set of depths $\mathbf{z}_b = (z_{b_1}, \dots, z_{b_m})^\top$ the exact ages $\mathbf{y}_b =$

$(y_{b_1}, \dots, y_{b_m})^\top$ are known from some external source of information. We call the set of tuples $\{(z_{b_k}, y_{b_k})\}_k$ tie-points. Assuming that tie-points can be considered as realizations of the posterior predictive distribution, then the synchronous age uncertainties are given by

$$\pi(\mathbf{y}_a \mid \mathbf{y}_b, \boldsymbol{\beta}, \boldsymbol{\theta}). \quad (6)$$

This assumes a partitioning of the joint vector age variables into $\mathbf{y}^\top = (\mathbf{y}_a^\top, \mathbf{y}_b^\top)$, where $\mathbf{y}_b$ represents the subset of the m fixed age variables and $\mathbf{y}_a = (y_{a_1}, \dots, y_{a_{n-m}})^\top$ represents the remaining $n - m$ free variables, with $n \gg m$. A similar partitioning is consequentially imposed on the mean vector and the precision matrix

$$\boldsymbol{\mu}_y(\boldsymbol{\beta}) = \begin{pmatrix} \boldsymbol{\mu}_a(\boldsymbol{\beta}) \\ \boldsymbol{\mu}_b(\boldsymbol{\beta}) \end{pmatrix}, \quad \mathbf{Q}_y(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{Q}_a(\boldsymbol{\theta}) & \mathbf{Q}_{ab}(\boldsymbol{\theta}) \\ \mathbf{Q}_{ab}(\boldsymbol{\theta})^\top & \mathbf{Q}_b(\boldsymbol{\theta}) \end{pmatrix}. \quad (7)$$

Since the joint distribution of $\mathbf{y}$ is Gaussian, so is the conditional distribution

$$\pi(\mathbf{y}_a \mid \mathbf{y}_b, \boldsymbol{\beta}, \boldsymbol{\theta}) = \mathcal{N}_{n-m} \left(\boldsymbol{\mu}_{a|b}(\boldsymbol{\beta}, \boldsymbol{\theta}), \mathbf{Q}_{a|b}^{-1}(\boldsymbol{\theta}) \right). \quad (8)$$

The conditional mean vector is given by (Rue and Held, 2005)

$$\boldsymbol{\mu}_{a|b}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{\mu}_a(\boldsymbol{\beta}) - \mathbf{Q}_a^{-1}(\boldsymbol{\theta}) \mathbf{Q}_{ab}(\boldsymbol{\theta}) (\mathbf{y}_b - \boldsymbol{\mu}_b(\boldsymbol{\beta})), \quad (9)$$

and the conditional precision matrix is simply

$$\mathbf{Q}_{a|b}(\boldsymbol{\theta}) = \mathbf{Q}_a(\boldsymbol{\theta}). \quad (10)$$

Sampling from $\pi(\mathbf{y}_a \mid \mathbf{y}_b, \boldsymbol{\beta}, \boldsymbol{\theta})$ can be performed in linear time provided $\mathbf{Q}_y(\boldsymbol{\theta})$ is sparse.

In any realistic application, tie-points will be associated with a non-negligible amount of uncertainty. Let the joint uncertainty be described by some distribution $\pi(\mathbf{y}_b \mid \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ denotes a vector of known parameters or conditions that affect the joint tie-point distribution. A set of uncertain tie-points is then comprised of a vector of depths and a (joint) probability distribution of associated ages $(\mathbf{z}_b, \pi(\mathbf{y}_b \mid \boldsymbol{\xi}))$.

The synchronized dating uncertainty can then be expressed using the law of total probability

$$\pi(\mathbf{y}_a \mid \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \int \pi(\mathbf{y}_a \mid \mathbf{y}_b, \boldsymbol{\beta}, \boldsymbol{\theta}) \pi(\mathbf{y}_b \mid \boldsymbol{\xi}) d\mathbf{y}_b, \quad (11)$$

where we have assumed conditional independence $\pi(\mathbf{y}_a \mid \mathbf{y}_b, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\xi}) = \pi(\mathbf{y}_a \mid \mathbf{y}_b, \boldsymbol{\beta}, \boldsymbol{\theta})$. The integral in Eq. (11) is often impossible to solve analytically. But given expression (8) for $\pi(\mathbf{y}_a \mid \mathbf{y}_b, \boldsymbol{\beta}, \boldsymbol{\theta})$, sampling from $\pi(\mathbf{y}_a \mid \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\theta})$ is simple and can be achieved using a two-step Monte Carlo sampling procedure. First, sampling from $\mathbf{y}_b \mid \boldsymbol{\xi}$ yields predictive posterior distributions according to the fixed tie-point synchronization for each sampled set of tie-points. Second, from each fixed tie-point synchronization chronologies can be sampled. This two-stage sampling procedure suitably reflects all remaining uncertainty. A diagram illustrating our methodology, from model specification and fitting to synchronization and sampling, is provided in Fig. 1.

If the joint tie-point uncertainty is expressed by a Gaussian distribution,

$$\mathbf{y}_b \mid \boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}_\xi, \mathbf{Q}_\xi^{-1}) \quad (12)$$

then Eq. (11) can be computed analytically. We then find for synchronized ages $\mathbf{y}_a$

$$\mathbf{y}_a \mid \boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}_{a|\xi}, \mathbf{Q}_{a|\xi}^{-1}), \quad (13)$$

where

$$\boldsymbol{\mu}_{a|\xi} = \mathbf{Q}_{a|\xi}^{-1} \left(\mathbf{b}_a - \mathbf{Q}_{ab}(\mathbf{Q}_\xi + \mathbf{Q}_{ab}^\top \mathbf{Q}_a^{-1} \mathbf{Q}_{ab})^{-1} \mathbf{b}_b \right), \quad (14)$$

and

$$\mathbf{Q}_{a|\xi} = \mathbf{Q}_a - \mathbf{Q}_{ab} \left(\mathbf{Q}_\xi + \mathbf{Q}_{ab}^\top \mathbf{Q}_a^{-1} \mathbf{Q}_{ab} \right)^{-1} \mathbf{Q}_{ab}^\top, \quad (15)$$

with

$$\mathbf{b}_a = \mathbf{Q}_a \boldsymbol{\mu}_a + \mathbf{Q}_{ab} \boldsymbol{\mu}_b, \quad \mathbf{b}_b = \mathbf{Q}_\xi \boldsymbol{\mu}_\xi + \mathbf{Q}_{ab}^\top \mathbf{Q}_a^{-1} \mathbf{Q}_{ab} \boldsymbol{\mu}_b. \quad (16)$$

This implies that the resulting synchronous joint dating uncertainty is Gaussian and retains the Markov property.

This approach assumes that tie-points can be considered to be realizations of the posterior predictive distribution of $\mathbf{y}$. It does not perform Bayesian updating on any model parameters, nor does it take into account any potential biases in the counting process that could cause a discrepancy between the layer-counted chronology and the real time scale represented by the tie-points. This approach is therefore only useful in incorporating tie-points that can be assumed to follow the same distribution as the layer-counted data, to constrain dating uncertainties expressed by the posterior predictive distribution. It should not be used if there is suspicion of a non-negligible bias in the layer-counted data.

C Joint fitting with Markov chain Monte Carlo simulations

To demonstrate how the layer-increment model and the age-discrepancy model can be fitted simultaneously in a single step, we will present a simulated example. While this approach is more computationally demanding and may require additional strategies to ensure numerical stability, it provides a more realistic demonstration of joint inference.

In this example, we assume that the tie-points are fixed, meaning they have no inherent uncertainty to account for. We first simulate $n = 100$ layer increments, $\Delta \mathbf{y}_1$, using piecewise defined trends between pre-determined change-points. Synthetic proxy measurements are sampled from an AR(1) process and combined linearly using coefficients $\boldsymbol{\beta}$. Additional variability is added by sampling from an AR(1) model with parameters $\boldsymbol{\theta}$. Next, tie-points are constructed by taking the cumulative sums of the synthetic layer increments, $\mathbf{y}_1$, and adding a component following a sine wave, representing the systematic bias $\boldsymbol{\psi} = \mathbf{y}_2 - \mathbf{y}_1$. Tie-points are then extracted at locations 1, 25, 50, 75 and 90, while the remaining values are left unobserved, yielding $\mathbf{y}_2 = (\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{missing}})$.

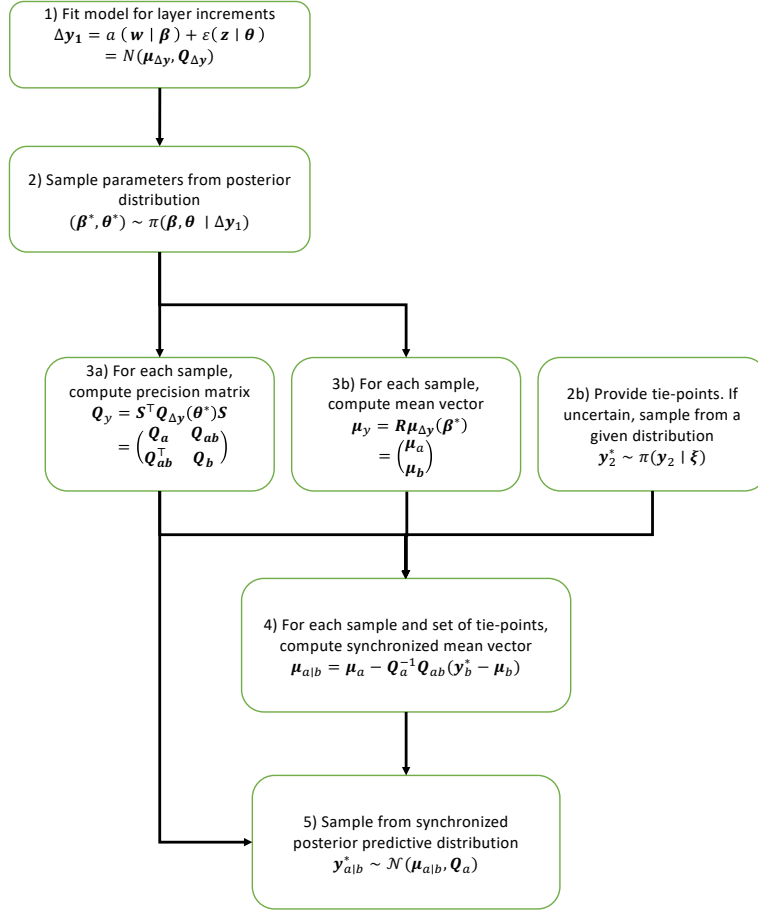


Figure 1: Diagram describing the process of fitting the age increment model to observed layer increments and, subsequently, simulate chronologies that have been constrained using tie-points, under the assumption of a Gaussian layer-increment model and no bias in the layer-counting process. Steps 3–6 are repeated for every chronology in the ensemble, while steps 1 and 2 are performed only once.

The objective is to fit the layer-increment model to $\Delta \mathbf{y}_1$ and simultaneously fit an appropriate model to the age discrepancies $\boldsymbol{\psi}_{\text{obs}}$, and then infer the unknown ages $\mathbf{y}_{\text{missing}}$. To complement our examples in the main text, we demonstrate here an alternative age-discrepancy model: the ARIMA(1,1,0) process. The model is fitted using a no U-turn sampler (NUTS) MCMC approach using a Stan implementation (Carpenter et al., 2017; Stan Development Team, 2024). The results are presented in Fig. 2, which depicts the synthetic layer increments, tie-points, and their respective model fits. The figure illustrates that the fitted ARIMA(1,1,0) model is able to mimic the underlying true sine function, despite only having 5 out of 100 observations. When the model

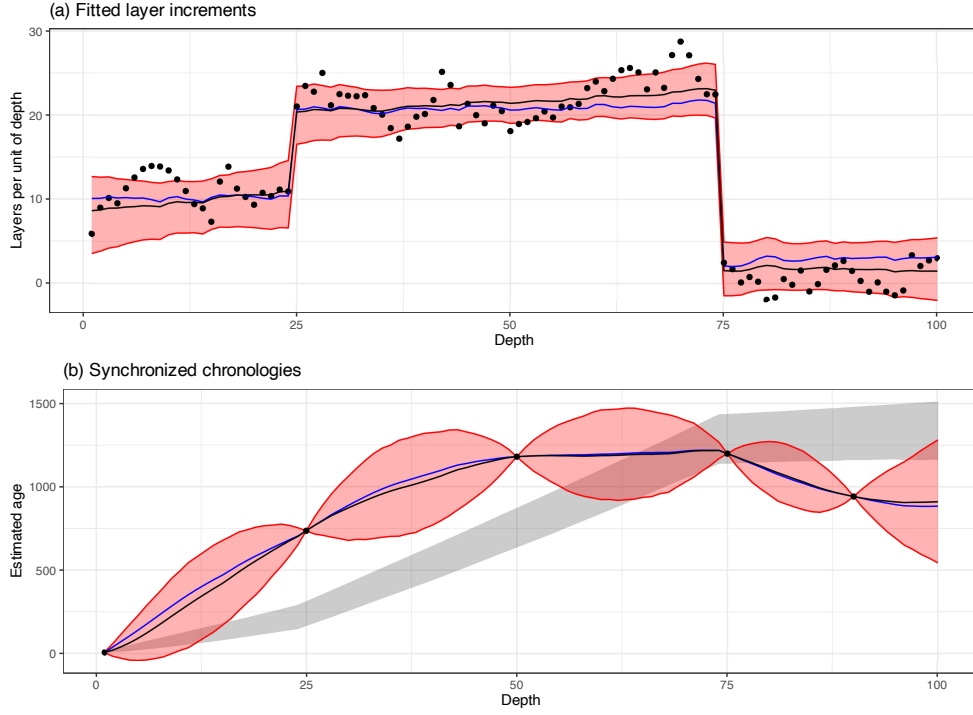


Figure 2: An AR(1) model is fitted to simulated layer increments (panel (a)) and an ARIMA(1,1,0) model representing potential structural biases is fitted to the discrepancy between the cumulative sum of the layer increments and five fixed tie-points (panel (b)). The models are combined and fitted to the extended dataset simultaneously. The black points represent the observed data, the blue lines represent the true layer-increment trend term (panel (a)) and the true bias from which the tie-points are taken (panel (b)). The black lines represent the posterior marginal means and the red shaded area represent the 95% credible intervals. The gray shaded area in panel (b) shows the 95% credible intervals of the unsynchronized age, obtained by taking the cumulative sums of the layer increments without incorporating tie-points.

accounts for an additional age-discrepancy component, the inferred uncertainties increase. This contrasts with the unsynchronized chronologies, where no such component is incorporated.

References

- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). “Stan: A probabilistic programming language.” *Journal of Statistical Software*, 76(1): 1–32. [5](#)

Rue, H. (2001). “Fast sampling of Gaussian Markov random fields.” *Journal of the Royal Statistical Society, Series B*, 63(2): 325–338. [2](#)

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory And Applications (Monographs on Statistics and Applied Probability)*. Chapman & Hall/CRC, New York, NY, USA, 1 edition. [3](#)

Stan Development Team (2024). “RStan: the R interface to Stan.” R package version 2.32.6.

URL <https://mc-stan.org/> [5](#)