*Submitted to the Annals of Applied Statistics*

# PHENOTYPIC EVOLUTION STUDIED BY LAYERED STOCHASTIC DIFFERENTIAL EQUATIONS - SUPPLEMENTARY MATERIAL

BY TROND REITAN[*], TORE SCHWEDER[†] AND JORIJNTJE HENDERIKS[‡*]

**1. Introduction.** There were some technical issues that had to be resolved in order to describe in more detail the framework described in section 2 of the main text and turn it into an inferential tool that could be used for instance for the coccolith application. The analytical expressions for expectation and covariance, had to be derived and studied. In order to speed up the calculation of the likelihood, a Kalman filter was used instead of calculating the likelihood directly from the correlation matrix. Restrictions on the model complexity were necessary in order to get results in a reasonable amount of time. Also, numerical methods had to be used for analyzing the likelihood for each model.

In this supplementary, we will describe the Kalman filtering approach, the restrictions imposed and the numerical methods in more detail than in the main text. But we will also describe some details about the modeling framework. The program code and the data file are also supplemented.

## 2. Some properties of vector linear stochastic differential equations (SDEs).

2.1. *Itô representation and moments of a linear SDE.* General results concerning linear SDEs as well as some special cases of vectorial linear SDEs are described here. We refer to the main article, section 2, for the description of such processes, but will remind the reader of the following general representation of a linear SDE:

$$dX(t) = (m(t) + AX(t))dt + \Sigma(t)dW(t), \tag{1}$$

where the state vector $X(t)$ is a $p$-dimensional function of time $t$ and $W(t)$ are random walk stochastic contributions. We write it a little more general here than

[*]CEES, Dept. of Biology, University of Oslo, P. O. Box 1053 Blindern, N-0316 Oslo, Norway. E-mail: trondr@bio.uio.no

[†]Dept. of Economics, University of Oslo, P. O. Box 1095 Blindern, N-0316 Oslo, Norway. E-mail: tore.schweder@econ.uio.no

[‡]Dept. of Earth Sciences, Uppsala University, Villavägen 16, SE-75 236 Uppsala, Sweden. E-mail: jorijntje.henderiks@geo.uu.se. Funding for this research was provided by the Swedish National Research Council and Knut and Alice Wallenberg Foundation

*AMS 2000 subject classifications:* Time series, Latent processes

*Keywords and phrases:* Ornstein-Uhlenbeck process, Time series, Latent processes, Fossil data, Coccolith

in the main paper, as we initially do not assume the diffusion matrix $\Sigma(t)$ to be constant in time, though it should not depend on $X(t)$ or any other stochastic time series.

Here $m(t)$ is a $p$ dimensional expectation term which for instance can consist of a fixed expectation function and/or contributions from exogenous (externally specified) processes to any part of the state process. The pull matrix, $A$, which we will assume is constant in time, is a $p \times p$ matrix describing how the $p$ processes in the state vector reacts to themselves and each other.

Let the dimensionality of the stochastic contributions, $dW(t)$, be $q$, where $W$ is a vector of standard Wiener process for each component, with no dependency between components. Thus, the diffusion matrix has $dim(\Sigma(t)) = p \times q$.

The Itô representation of a $p$-dimensional stochastic process $X(t)$ governed by Eq. (1) is

(2)
$$X(t) = B(t-t_0)^{-1}\left\{X(t_0) + \int_{t_0}^{t} B(u-t_0)m(u)du + \int_{t_0}^{t} B(u-t_0)\Sigma(u)dW(u)\right\}$$

where $B(t) = \mathrm{e}^{-At} \equiv \sum_{n=0}^{\infty}(-A)^n t^n/n!$ and $X(t_0)$ is the state at a previous time $t_0$.

When the pull matrix $A$ is diagonalizable, the exponential matrix function, $\mathrm{e}^{-At}$, has a convenient expression in terms of the matrix of eigenvectors $V$ (which then is of full rank) and the eigenvalues $\lambda_j$ in appropriate multiplicity in the diagonal matrix $\Lambda$ of the pull matrix:

$$VA = \Lambda V,$$

(3)
$$B(t) = \sum_{n=0}^{\infty} V^{-1}\Lambda^n(-t)^n V/n! = V^{-1}\mathrm{e}^{-\Lambda t}V.$$

We will here concentrate on models with diagonalizable pull matrices.

Here $\mathrm{e}^{-\Lambda t}$ is a diagonal matrix with elements $\mathrm{e}^{-\lambda_j t}$. Thus, Eq. (2) can be expressed as

$$X(t) = V^{-1}\mathrm{e}^{\Lambda(t-t_0)}VX(t_0) + V^{-1}\int_{t_0}^{t}\mathrm{e}^{\Lambda(t-u)}Vm(u)du + V^{-1}\int_{t_0}^{t}\mathrm{e}^{\Lambda(t-u)}V\Sigma(u)dW(u)$$

This representation can for instance be found in [1]. It is parallel to that for ordinary linear differential equations obtained by the eigenvalue method (see for instance [12], p. 269).

The first and second order moments of the state process, given $X(t_0)$, are

$$EX(t) = V^{-1}e^{\Lambda(t-t_0)}VX(t_0) + V^{-1}\int_{t_0}^{t} e^{\Lambda(t-u)}Vm(u)du$$

$$cov(X(v), X(t)) = V^{-1}\left[\int_{t_0}^{v} e^{\Lambda(v-u)}V\Sigma(u)\Sigma(u)'V'e^{\Lambda(t-u)}du\right](V^{-1})', \tag{4}$$

for $v \leq t$. Being a linear transform of the Wiener process $W$, the state process $X$ is Gaussian. Thus the two first moments are sufficient to determine the entire state process.

When the diffusion matrix is constant in time, Eq. (4) can be expressed as

$$cov(X(v), X(t)) = V^{-1}\Xi(t, v, t_0)(V^{-1})' \tag{5}$$

where $\Xi(t, v, t_0)_{i,j} = -\frac{e^{\lambda_j(t-v)} - e^{\lambda_i(t-t_0+v-t_0)}}{\lambda_i+\lambda_j}\Omega_{i,j}$, $\lambda_i = \Lambda_{i,i}$ is the i'th eigenvalue of $A$ and $\Omega = V\Sigma\Sigma'V'$.

When the expectation contribution $m(t) = m_0$ is constant, the first part of Eq. (4) can be expressed as:

$$EX(t) = V^{-1}e^{\Lambda(t-t_0)}VX(t_0) - V^{-1}\Lambda^{-1}(1 - e^{\Lambda(t-t_0)})Vm_0. \tag{6}$$

These expressions can be used in a Kalman filter setting, for updating between the process state from one time point to the next, $X(t_{k-1}) \to X(t_k)$, as well as in a direct calculation of the likelihood by the multinormal distribution.

Note also that $EX(t)$ and $cov(X(t), X(t)) = var(X(t))$ converge as $t - t_0 \to \infty$, provided all eigenvalues are real and negative. Then the process is stationary provided $X(t_0)$ has the limiting normal distribution.

2.2. *Hierarchical models and layers.* A model is said to be hierarchical if the variables can be grouped in linearly ordered layers with directed causal flow up the chain. Each layer can consist of several processes, for instance representing the geographical sites in the coccolith application. See Fig. 1 for an illustration of the layered modeling framework.

Causal flow is here understood as Granger causality [3, 4, 11], and causality is in SDE models equivalent to local dependency. We have that $X_i$ is locally independent of $X_j$ when $A_{ij} = 0$ and is locally dependent on $X_j$ when $A_{ij} \neq 0$, see [10]. Note that $X_i$ is the set of processes belonging to layer $i$ and $A_{ij}$ refers to a sub-matrix of $A$ where the rows belong to layer $i$ and the columns belong to layer $j$. Keep in mind that all layer processes $X_i$ can be multi-dimensional. The total state will be $X = (X_1', \ldots, X_l')'$ with $l$ being the number of layers. With dimensionality $k$ in each layer, the total state dimensionality will be $p = kl$.

If $X_j$ is locally independent of $X_i$ but not vice versa, then we represent this graphically as $X_j \to X_i$.

For a hierarchical model with $l$ layers, the pull matrix has upper triangular block form. If the causal flow is from a layer to the layer just above, the pull matrix will also have band structure, so that each layer reacts only to the one immediately below it, in an OU-like fashion:

$$A = \begin{pmatrix} -A_1 & A_1 & \tilde{0} & & \\ \tilde{0} & -A_2 & A_2 & \ddots & \\ & \ddots & \ddots & \ddots & \ddots \\ & & \ddots & -A_{l-1} & A_{l-1} \\ & & & \tilde{0} & -A_l \end{pmatrix}. \tag{7}$$

Here $\tilde{0}$ are $k \times k$ sub-matrices having all elements equal to zero. The $k \times k$ sub-matrices $A_i$ are diagonal, typically having strictly positive diagonal elements, to achieve stationarity. The lowest layer could be allowed to have zeros on the diagonal, in which case it would not be stationary but rather represent a random walk. Diagonal elements equal to zero in layers above that would mean that the layer process is not affected by the layers below, and that these layers then could be removed.

Denoting the layers from top to bottom by $X_1, \ldots, X_l$, with $X_i$ belonging to layer $i$, the causal structure described in Eq. (7) is $X_l \to X_{l-1} \to \cdots \to X_1$. In addition to this endogenous flow of causality, there might be exogenous forcings in any layer.

A layer could react to a layer below with a different scaling factor than it reacts to itself, but with that freedom and the freedom to specify the mean and variance of the layer below, the system would be non-identifiable.

We will call a hierarchical model, with all elements in Eq. (1) having a decomposition into layers and where the pull matrix has the form of Eq. (7), a strict layered process. The form of other elements of Eq. (1) in a strict layered model, is described in the text below.

In order for a system to have distinct layers, the diffusion matrix would also need to have a band structure, but this time only with elements on the diagonal band. This can perhaps most easily be described by the covariance matrix of the

stochastic contributions,

$$\Sigma\,\Sigma' = \begin{pmatrix} \Sigma_1\Sigma_1' & \tilde{0} & & & \\ \tilde{0} & \Sigma_2\Sigma_2' & \tilde{0} & \ddots & \\ & \ddots & \ddots & \ddots & \ddots \\ & & \ddots & \Sigma_{l-1}\Sigma_{l-1}' & \tilde{0} \\ & & & \tilde{0} & \Sigma_l\Sigma_l' \end{pmatrix}. \tag{8}$$

Even in a strict layered model, the square sub-matrices $\Sigma_i\,\Sigma_i'$ are not required to be diagonal. We want to allow for the possibility of having instantaneous stochastic contributions being shared among the individual processes in a layer.

The expectation contribution can, for strict layered model purposes, be decomposed into $m(t) = (0\ \ldots\ 0\ A_l\mu_0(t))' + A\sum_{m=1}^{s}\beta_m T_m(t)$. The first term containing $\mu_0(t)$ describes the expectation contribution from the lowest layer, which if it is constant, will be the expectation vector of layer processes in the stationary case. The second term denotes $s$ regression terms on the set of exogenous time series $T_m(t)$ with $m \in \{1,\ldots,s\}$. In a strict layered model, we would want to latch each regressor $m$ into a single layer, denoted $L(m)$, so that the vector $\beta_m$ only has non-zero elements for processes belonging to that layer.

With the model being strictly layered, each layer $i$ can then be described by the following OU-like equation:

$$dX_i(t) = -A_i\left(X_i(t) - X_{i+1}(t) - \sum_{m|L(m)=i}\beta_{m,i}T_i(t)\right)dt + \Sigma_i dW_i(t) \tag{9}$$

where $\beta_{m,i}$ are those elements in the $\beta_m$ vector that belongs to layer $i$ and where we define $X_{l+1} = \mu_0(t)$, so that the lowest layer responds to the expectation vector $\mu_0(t)$.

An apparatus for handling more than one type of datasets in the analysis can be made by using strictly layered processes for representing each dataset. Causal flow between specific layers belong to each of two dataset could then be allowed. This would be most natural to do if the sites correspond to the different datasets or some of the datasets are deemed global in nature.

2.3. *Relating two examples to Eq. (1).* We will cover two examples here. First is a situation with two layers and only one site. The second will be our initial model, having three layers and $k$ sites, where we allow for correlations between sites in a special manner.

The two-layered examples described in the main manuscript, can be put in the form of Eq. (1) by setting

$$A = \begin{pmatrix} -\alpha_1 & \alpha_1 \\ 0 & \alpha_2 \end{pmatrix},$$
$$\Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$$

and

$$m(t) = \begin{pmatrix} 0 \\ \alpha_2 \mu_0 \end{pmatrix}.$$

This will result in the two processes being described as

$$\begin{aligned} dX_1(t) &= -\alpha_1(X_1(t) - X_2(t))dt + \sigma_1 dW_1(t) \\ dX_2(t) &= -\alpha_2(X_2(t) - \mu_0)dt + \sigma_2 dW_2(t). \end{aligned} \tag{10}$$

The three-layered model described in the manuscript called the initial model, consists of three layers, with a middle layer that receives global stochastic contributions while the two other receive strictly local contributions. The state will have dimension $3k$ and can be de-composed into three layers, which we will label as $X_{1,j}$ in the top layer, $X_{2,j}$ in the middle layer, and $X_{3,j}$ at the bottom. The site index $j$ runs over the number of sites, $k$. The flow of causality is $X_{3,j} \to X_{2,j} \to X_{1,j}$, and for each $j$, $\alpha_3$ is the pull in the bottom Ornstein-Uhlenbeck (OU) processes $X_{3,j}$, $\alpha_2$ is the pull for $X_{2,j}$, and $\alpha_1$ for $X_{1,j}$. The instantaneous standard deviations are also identical within layers, and are $\sigma_i$ at layer $i$. We let the middle layer be affected by an external time series $T(t)$, as well as a common Wiener process $W_2$. Otherwise there are no instantaneous cross covariances. The equations for this process are

$$\begin{aligned} dX_{1,j}(t) &= -\alpha_1(X_{1,j}(t) - X_{2,j}(t))dt + \sigma_1 dW_{1,j}(t) \\ dX_{2,j}(t) &= -\alpha_2(X_{2,j}(t) - X_{3,j}(t) - \beta T(t))dt + \sigma_2 dW_2(t) \\ dX_{3,j}(t) &= -\alpha_3(X_{3,j}(t) - \mu_0)dt + \sigma_3 dW_{3,j}(t), \end{aligned} \tag{11}$$

where $j = 1, \ldots, k$.

This can be described using Eq. (1) as

$$A = \begin{pmatrix} -\alpha_1\tilde{1} & \alpha_1\tilde{1} & \tilde{0} \\ \tilde{0} & -\alpha_2\tilde{1} & \alpha_2\tilde{1} \\ \tilde{0} & \tilde{0} & -\alpha_3\tilde{1} \end{pmatrix},$$
$$\Sigma = \begin{pmatrix} \sigma_1\tilde{1} & \underline{0} & \tilde{0} \\ \tilde{0} & \sigma_2\underline{1} & \tilde{0} \\ \tilde{0} & \underline{0} & \sigma_3\tilde{1} \end{pmatrix}$$

and

$$m(t) = \begin{pmatrix} \underline{0} \\ \alpha_2 \beta T(t) \\ \alpha_3 \mu_0 \underline{1} \end{pmatrix}$$

where $\tilde{1}$ is the $k \times k$ identity matrix, $\tilde{0}$ is an $k \times k$ of only zeros and $\underline{1}$ and $\underline{0}$ are $k$ dimensional column vectors consisting only of ones and zeros respectively.

Note that $\Sigma$ is a $3k \times (2k+1)$ matrix, rather than $3k \times 3k$. The covariance matrix then becomes the following $3k \times 3k$ matrix:

$$\Sigma\,\Sigma' = \begin{pmatrix} \sigma_1^2 \tilde{1} & \tilde{0} & \tilde{0} \\ \tilde{0} & \sigma_2^2 \underline{1}\underline{1}' & \tilde{0} \\ \tilde{0} & \tilde{0} & \sigma_3^2 \tilde{1} \end{pmatrix}.$$

2.4. *Covariance calculations for examples.* In all the examples, we assume the process to be stationary, so that the initial state is irrelevant.

We first look at an OU process, $dX(t) = -\alpha(X(t) - \mu_0)dt + \sigma\, dB(t)$, such that $\alpha > 0$. The stationary mean and covariance, using Eq. (4) is respectively $\mu_0$ and $cov(X(0), X(t)) = \mathrm{e}^{-\alpha t}\sigma^2/2\alpha$.

The two-layered process in Eq. (10) will also typically have stationary covariance at the upper layer, which then is a linear combination of exponential functions. The stationary mean for the topmost layer of a two-layered system will be $\mu_0$, while the covariance is, using Eq. (4),

$$cov\left(X_1(0), X_1(t)\right) = \frac{\sigma_1^2}{2\alpha_1}\mathrm{e}^{-\alpha_1 t} + \frac{\sigma_2^2 \alpha_1^2}{\alpha_1^2 - \alpha_2^2}\left(\frac{1}{2\alpha_2}\mathrm{e}^{-\alpha_2 t} - \frac{1}{2\alpha_1}\mathrm{e}^{-\alpha_1 t}\right), \tag{12}$$

provided $\alpha_1 \neq \alpha_2$ are both positive.

For a three-layered model with six sites in each layer, the eigenvector matrix is

$$V = \begin{pmatrix} \tilde{1} & \frac{\alpha_1}{\alpha_2 - \alpha_1}\tilde{1} & \frac{\alpha_2 \alpha_1}{(\alpha_2 - \alpha_1)(\alpha_3 - \alpha_1)}\tilde{1} \\ 0 & \tilde{1} & \frac{\alpha_2}{\alpha_3 - \alpha_2}\tilde{1} \\ 0 & 0 & \tilde{1} \end{pmatrix}, V^{-1} = \begin{pmatrix} \tilde{1} & -\frac{\alpha_1}{\alpha_2 - \alpha_1}\tilde{1} & \frac{\alpha_2 \alpha_1}{(\alpha_3 - \alpha_1)(\alpha_3 - \alpha_2)}\tilde{1} \\ 0 & \tilde{1} & -\frac{\alpha_2}{\alpha_3 - \alpha_2}\tilde{1} \\ 0 & 0 & \tilde{1} \end{pmatrix}$$

and

$$\Lambda = \begin{pmatrix} -\alpha_1\tilde{1} & 0 & 0 \\ 0 & -\alpha_2\tilde{1} & 0 \\ 0 & 0 & -\alpha_3\tilde{1} \end{pmatrix}$$ where $\tilde{1}$ is the unit matrix.

Using Eq. (4), the following formula for the stationary covariance of the top

layer emerges, say when $\alpha_1 > \alpha_2 > \alpha_3 > 0$:

$$cov(X_{1,i}(t), X_{1,j}(u)) = \delta_{ij}\frac{\sigma_1^2 e^{-\alpha_1(t-u)}}{2\alpha_1} + \frac{\sigma_2^2\alpha_1^2}{\alpha_1^2-\alpha_2^2}\left(\frac{e^{-\alpha_2(t-u)}}{2\alpha_2} - \frac{e^{-\alpha_1(t-u)}}{2\alpha_1}\right) + \left(\frac{\alpha_1^2\alpha_2^2 e^{-\alpha_3(t-u)}}{\alpha_3(\alpha_1^2-\alpha_3^2)(\alpha_2^2-\alpha_3^2)} - \frac{\alpha_1^2\alpha_2 e^{-\alpha_2(t-u)}}{(\alpha_1^2-\alpha_2^2)(\alpha_2^2-\alpha_3^2)} + \frac{\alpha_1\alpha_2^2 e^{-\alpha_1(t-u)}}{(\alpha_1^2-\alpha_3^2)(\alpha_1^2-\alpha_2^2)}\right) \times \frac{\delta_{ij}\sigma_3^2}{2} \quad (13)$$

where $\delta_{ij}$ is the Kronecker delta.

In the application to coccolith size evolution, there are hidden layers below the top layer. The top layer represents the six site-specific population mean coccolith sizes. Underneath this layer is a hidden layer of site-specific fitness optima. These are possibly influenced by a global temperature indicator series $T(t)$, which is assumed known over the 57 million years of data, and also by unobservable variables, one for each site, in a bottom hidden layer. The model is thus a process of the form of Eq. (11) with $l = 3$ layers and with six parallel trackers. The variables in the lowest hidden layers might be regarded as confounders.

2.5. *Identification in hierarchical models with flat hidden layers - swapping layers.* Identifiability issues arise when there are multiple ways of defining a hierarchical model such that the mean and correlation structure is the same at the topmost, measured layer. Since the mean and correlation structure specifies the likelihood, two models that are able to replicate each other's mean and covariance, will not be identifiable.

For a model with two layers, the process of the lower layer $X_2$ will be an OU process. The stationary covariance for the top layer $X_1$, is given in Eq. (12). An identical covariance is only possible in models with $\alpha_1$ and $\alpha_2$ as the eigenvalues of the pull matrix. The only possibility is thus to interchange the two values.

**Lemma 1** *The stationary covariance of the top layer process in a two-layered process is the same as the stationary covariance in a two-layered process with the pull coefficients being switched ($\alpha_1 \to \alpha_2$, $\alpha_2 \to \alpha_1$) provided that $\sigma_2^2 \to \frac{\sigma_2^2\alpha_1^2 - \sigma_1^2(\alpha_1^2-\alpha_2^2)}{\alpha_2^2} \geq 0$. When $\alpha_1 > \alpha_2 > 0$ it is impossible to switch the layers and re-scale when $\sigma_2^2 < \sigma_1^2\left(1-\alpha_2^2/\alpha_1^2\right)$. When $0 < \alpha_1 < \alpha_2$, it is always possible.*

This can be seen by studying the process where $\tilde{\alpha}_2 = \alpha_1$ now belongs to the lower layer, while $\tilde{\alpha}_1 = \alpha_2$ belongs to the upper. The old two-layered system is described by $dX_1(t) = -\alpha_1(X_1(t)-X_2(t))dt+\sigma_1 dB_1(t)$ and $dX_2(t) = -\alpha_2(X_2(t)-\mu)dt+\sigma_2 dB_2(t)$. The new two-layered system is written as $dY_1(t) = -\tilde{\alpha}_1(Y_1(t)-Y_2(t))dt + \tilde{\sigma}_1 dB_3(t)$ and $dY_2(t) = -\tilde{\alpha}_2(Y_2(t)-\tilde{\mu})dt + \tilde{\sigma}_2 dB_4(t)$. The processes

will have the same expectation, if $\tilde{\mu} = \mu$. The covariance of the top (observable) layer will be the same if $cov\left(X_1(0), X_1(t)\right) = cov\left(Y_1(0), Y_1(t)\right)$ for all $t$. Using the expression for two-layered covariance in Eq. (12) for both processes, this means that $\frac{\sigma_1^2 e^{-\alpha_1 t}}{2\alpha_1} + \frac{\sigma_2^2\alpha_1^2}{\alpha_1^2-\alpha_2^2}\left(\frac{e^{-\alpha_2 t}}{2\alpha_2} - \frac{e^{-\alpha_1 t}}{2\alpha_1}\right) = \frac{\tilde{\sigma}_1^2 e^{-\tilde{\alpha}_1 t}}{2\tilde{\alpha}_1} + \frac{\tilde{\sigma}_2^2\tilde{\alpha}_1^2}{\tilde{\alpha}_1^2-\tilde{\alpha}_2^2}\left(\frac{e^{-\tilde{\alpha}_2 t}}{2\tilde{\alpha}_2} - \frac{e^{-\tilde{\alpha}_1 t}}{2\tilde{\alpha}_1}\right)$ for all $t$. Putting $\tilde{\alpha}_2 = \alpha_1$ and $\tilde{\alpha}_1 = \alpha_2$, this yields that $\frac{\sigma_1^2}{2\alpha_1} - \frac{\sigma_2^2\alpha_1^2}{\alpha_1^2-\alpha_2^2}\frac{1}{2\alpha_1} = -\frac{\tilde{\sigma}_2^2\alpha_2^2}{\alpha_1^2-\alpha_2^2}\frac{1}{2\alpha_1}$ and $\frac{\sigma_2^2\alpha_1^2}{\alpha_1^2-\alpha_2^2}\frac{1}{2\alpha_2} = \frac{\tilde{\sigma}_1^2}{2\alpha_2} + \frac{\tilde{\sigma}_2^2\alpha_2^2}{\alpha_1^2-\alpha_2^2}\frac{1}{2\alpha_2}$. Solving for $\tilde{\sigma}_1^2$ and $\tilde{\sigma}_2^2$ yields $\tilde{\sigma}_1^2 = \sigma_1^2$ and $\tilde{\sigma}_2^2 = \sigma_2^2\left(\frac{\alpha_1}{\alpha_2}\right)^2 - \sigma_1^2\left(\left(\frac{\alpha_1}{\alpha_2}\right)^2 - 1\right)$. With $\alpha_1 > \alpha_2 > 0$, $\tilde{\sigma}_2^2$ is only positive when $\sigma_2^2 > \sigma_1^2\left(1 - \alpha_2^2/\alpha_1^2\right)$.

**Lemma 2** *The stationary covariance of the top layer process of a two-layered process having $\alpha_1 > \alpha_2$ is the same as the stationary covariance of a sum of two particular OU processes, when $\sigma_2^2 < \sigma_1^2\left(1 - \alpha_2^2/\alpha_1^2\right)$.*

A sum of two OU processes, $X_s = Y_1 + Y_2$ specified as $dY_1(t) = -\tilde{\alpha}_1(Y_1(t) - \mu/2)dt + \tilde{\sigma}_1 dB_3(t)$ and $dY_2(t) = -\tilde{\alpha}_2(Y_2(t) - \mu/2)dt + \tilde{\sigma}_2 dB_4(t)$ will have expectation $\mu$. One could get the same expectation with $X_1$ having expectation $\tilde{\mu}_1$ and $X_2$ having expectation $\tilde{\mu}_2$, where $\mu = \tilde{\mu}_1 + \tilde{\mu}_2$. For identification purposes, each can be given half the total expectation. The covariance is $cov\left(Y(0), Y(t)\right) = \frac{\tilde{\sigma}_1^2}{2\tilde{\alpha}_1}e^{-\tilde{\alpha}_1 t} + \frac{\tilde{\sigma}_2^2}{2\tilde{\alpha}_2}e^{-\tilde{\alpha}_2 t}$. Equating this with the expression for two layers, where $\alpha_1 > \alpha_2$ will only be possible if either $\tilde{\alpha}_1 = \alpha_1$ and $\tilde{\alpha}_2 = \alpha_2$ or $\tilde{\alpha}_1 = \alpha_2$ and $\tilde{\alpha}_2 = \alpha_1$. As there is an identification problem between $Y_1$ and $Y_2$ when all that is observed is $Y = Y_1 + Y_2$, we impose $\tilde{\alpha}_1 = \alpha_1$ and $\tilde{\alpha}_2 = \alpha_2$. We then get that $\frac{\sigma_2^2\alpha_1^2}{\alpha_1^2-\alpha_2^2}\frac{1}{2\alpha_2} = \frac{\tilde{\sigma}_2^2}{2\alpha_2}$ and $\frac{\sigma_1^2}{2\alpha_1} - \frac{\sigma_2^2\alpha_1^2}{\alpha_1^2-\alpha_2^2}\frac{1}{2\alpha_1} = \frac{\tilde{\sigma}_1^2}{2\alpha_1}$. Solving for $\tilde{\sigma}_1^2$ and $\tilde{\sigma}_2^2$ yields, $\tilde{\sigma}_1^2 = \sigma_1^2 - \frac{\sigma_2^2\alpha_1^2}{\alpha_1^2-\alpha_2^2}$ and $\tilde{\sigma}_2^2 = \frac{\sigma_2^2\alpha_1^2}{\alpha_1^2-\alpha_2^2}$. As long as $\alpha_1 > \alpha_2$, $\tilde{\sigma}_2^2$ will be positive. However, $\tilde{\sigma}_1^2$ will only be positive, as long as $\sigma_2^2 < \sigma_1^2\left(1 - \alpha_2^2/\alpha_1^2\right)$.

Note that the sum of two OU processes will be within the framework of multi-dimensional linear SDEs. By setting $X_s = Y_1 + Y_2$ as in lemma 2, one gets $dX_s(t) = (-\alpha_1(Y_1(t) - \mu/2) - \alpha_2(Y_2(t) - \mu/2))dt + (\tilde{\sigma}_1 dB_3(t) + \tilde{\sigma}_2 dB_4(t)$. This can be made into the form of Eq. (1) by setting the total process as the three dimensional vector $X = (X_s\ Y_1\ Y_2)'$. Then

$$A = \begin{pmatrix} 0 & -\alpha_1 & -\alpha_2 \\ 0 & -\alpha_1 & 0 \\ 0 & 0 & -\alpha_2 \end{pmatrix}, m = \begin{pmatrix} (\alpha_1+\alpha_2)\mu/2 \\ \alpha_1\mu/2 \\ \alpha_2\mu_2 \end{pmatrix} \text{ and}$$

$$\Sigma = \begin{pmatrix} \tilde{\sigma}_1 & \tilde{\sigma}_2 \\ \tilde{\sigma}_1 & 0 \\ 0 & \tilde{\sigma}_2 \end{pmatrix}, \tag{14}$$

will yield that $X_s$ is a sum of two OU processes.

The two-layered case having $\alpha_1 > \alpha_2$ can be called a fast tracking of a slow moving process (a). The two-layered case having $\alpha_1 < \alpha_2$ can be called a slow tracking of a fast moving process (b), while the sum of two OU processes will be called (c).

There is a degenerate situation for (c) when $\alpha_1 = \alpha_2$ or when $\sigma_2^2 = \sigma_1^2\left(1 - \alpha_2^2/\alpha_1^2\right)$. The degenerate situation $\sigma_2^2 = \sigma_1^2\left(1 - \alpha_2^2/\alpha_1^2\right)$ yields the covariance form of a single OU process, which can be covered as a special case of both (a), (b) and (c), which would be pointless to present as a two-layered process. The degenerate situation $\alpha_1 = \alpha_2$ yields a non-diagonizable $B(t)$ in Eq. (2), and can not be expressed in the form of Eq. (3) (though an analytical expression can still be found for it). In a Bayesian inference with continuous prior distribution on the parameters, both cases will have zero measure and can safely be ignored.

What has been shown here is that except for the degenerate situation, cases (b) and (c) are mutually exclusive, while case (a) is able to replicate the covariance of both these two other cases (though no case outside of them). Thus we prefer to impose the restriction $\alpha_1 > \alpha_2$ instead of contemplating $\alpha_1 < \alpha_2$ or the model described in Eq. (14). Such a restriction on the pull parameters will not be a restriction on the top layer of a two-layered process itself.

For an $l$ layer process $X_l \to \cdots \to X_1$, the suggested restriction is

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_l > 0. \tag{15}$$

The identifying restriction in Eq. (15) is argued successively, from top to bottom. If $\alpha_1 \geq \cdots \geq \alpha_j < \alpha_{j+1}$ then the two layers $j$ and $j+1$ could be switched without altering the model. Starting from the top again, the first violation of the restriction allows a switch of layers without any change to the covariance at the top layer. Such layer switching is repeated until Eq. (15) is satisfied.

Some other identification issues can also be spotted. Consider models with only one variable at each layer, $X_l \to X_{l-1} \to \cdots \to X_1$, but without the restrictions in Eq. (7). The governing equations will have the form $dX_j(t) = \left(a_{j,j}X_j(t) + a_{j,j+1}X_{j+1}(t)\right)dt + \sigma_j dW_j(t)$. With $-\alpha_j \equiv a_{jj}$, and by successive re-scaling of the variables, this is a hierarchical system of layered processes with pull coefficients $-\alpha_j$ taking the form of Eq. (7). Assuming $\alpha_j > 0$, the process is stationary.

Also note that if the pull matrix of a layer has diagonal components $-\infty$, the processes belonging to that layer track precisely the processes of the underlying layer. The layer is thus of no consequence, and a model with this layer removed will yield the same likelihood. Such a layer should thus be dropped for identification purposes.

Imposing the restriction in Eq. (15) will make sense in a single site setting. However, for multi site settings as in our application, with the possibility of site-

specific pulls, whether this restriction should be applied can be debated. If the top layer in a two-layered model has regional pulls, then the restriction in Eq. (15) can still be applied for a single site. If that restriction makes us switch layers, the lower layer can now be given regional pulls, so that each pull in the new second layer is equal to that of the previous first layer and the new first layer now has the pull of the previous second layer. However, the pulls in the other layer may not be subject to Eq. (15), so there is no way to impose increasing pull as you go from low to high layers and still be able to get the same covariance structure for all sites. One could still impose the restriction to a single site, but with no site being deemed more fundamental than another, this would seem unnatural.

One could also impose Eq. (15) on all sites, but with site-specific pull parameters, this will give a harder restriction on the possible covariances than what is strictly necessary. Still, we did apply this restriction as an option, in our analysis. We tried this because it wasn't clear whether a model with site-specific pulls would be preferred, inference would be easier without having to deal with the multimodality introduced by non-identifiability and a model with strictly ordered pulls could be deemed easier to interpret than one without this quality.

Note also that imposing this pull identification restriction created problems for the specification of the Bayesian prior distribution for the pull parameters, see section 3.2.

## 3. Prior distributions.

3.1. *General prior policy.* For all parameters, the distributional family was set by first transforming the parameters so that each re-parametrized version was allowed to range through all real numbers. The transformed parameter was then assigned a normal distribution.

Since we did not have any prior knowledge about dependency between parameters, the joint prior density was set as the product of those for the individual parameters.

For the mean $\mu$ and the regression parameter $\beta$, no re-parametrization was necessary. For the diffusion and pull (or characteristic time) parameters, the logarithm can be allowed to roam freely. The situation for the inter-regional correlations $\rho$, was a bit more complicated, as it can roam from $-1/(k-1)$ to 1, where $k$ is the dimensionality of each layer, i.e. the number of sites. In our case, where $k = 6$, correlations below $-0.2$ will not yield positively definite diffusion matrices. We here used a logistic transformation, from $[-0.2, 1]$ to the real line and let that be normally distributed. The mean and standard deviation of each parameter was set so as to yield the 95% credibility intervals listed in Table 2 of the main article.

3.2. *Extra pull parameter policies when the identification restriction is enforced.* An identification restriction on the pull parameters, so that the pulls are decreasing down the layers, can be imposed in any multi-layered model. However, since this is imposed before the data, the specification of the prior distribution of the pulls or characteristic times must take this restriction into account. At first, we tried starting with the normal distribution on the log-transformed characteristic times with a fixed 95% credibility band (typically stretching from $1ky$ to $1Gy$), and then simply impose the pull identification restriction. One needs to correct for the probability mass removed, by multiplying the restricted distribution with the number of non-identifiable ways to order the pull parameters. However when doing so, the marginal distribution of each pull parameter will be more narrow than it was initially. This will create an unfair advantage for models with a high number of layers, as long as the parameter estimates are still within the core range of the marginal distributions. This problem was discovered when the simulation study was performed.

We created and studied several methods for avoiding or at least decreasing this problem. All methods started with a basic characteristic time distribution as described above (namely log-normal and typically with a 95% credibility interval stretching from $1ky$ to $1Gy$), which we will call the base distribution.

One way of doing this is to start with the upper layer and give the characteristic time of that layer the base distribution. Conditioned on the characteristic time of the upper layer, the difference in characteristic times between the second and the first layer could also be given this base distribution and the same for the difference between the characteristic time of the third and second layer, conditioned on the characteristic time of the second layer. With this Markov chain prior, one has an expression for the joint distribution of the characteristic times: $\pi(\Delta t_1, \Delta t_2, \ldots, \Delta t_l) = \pi(\Delta t_1)\pi(\Delta t_2|\Delta t_1)\ldots\pi(\Delta t_l|\Delta t_{l-1}) = f_b(\Delta t_1)\, f_b(\Delta t_2 - \Delta t_1)\ldots f_b(\Delta t_l - \Delta t_{l-1})$. where $\pi(.)$ denotes the prior distribution density function of a specified parameter and $f_b$ denotes the base distribution density function.

An alternative to this approach is to start from the lowest layer, and give the characteristic time of that layer the base distribution. Then $\Delta t_l - \Delta t_{l-1}$ conditioned on $\Delta t_l$ is given the base distribution also. One can thus proceed to the upper layer. Then the joint distribution becomes $\pi(\Delta t_1, \Delta t_2, \ldots, \Delta t_l) = \pi(\Delta t_l)\pi(\Delta t_{l-1}|\Delta t_l)\ldots\pi(\Delta t_1|\Delta t_2) = f_b(\Delta t_l)\, f_b(\Delta t_l - \Delta t_{l-1})\ldots f_b(\Delta t_2 - \Delta t_1)$.

One could also use a mixture of these two priors, so that there is a prior probability for starting either at the lower or upper layer. One could even consider a mixture of starting at any given layer. This may be an unnecessary complication, though.

If one starts with the upper layer, it should be noted that while the width of the characteristic time of a layer conditioned on the layer above it has the same width

and form as the base distribution, the marginal distribution of the characteristic time of that layer will be wider, since the the characteristic time of the layer above will vary. This will be similar if one starts with the lowest layer instead. Thus these two prior strategies could give an unfair advantage to models with too few layers.

It is possible to make other characteristic time joint prior distributions by the same strategies as above, namely by starting off with either the upper or lower layer and develop a joint distribution for the prior according to a Markov chain. One such alternative would be to again start at the upper layer, giving the characteristic time of that layer the base distribution, but then let the characteristic time on the second layer conditioned on the first be the base distribution conditioned on the characteristic time of the second layer being larger than the characteristic time of the first layer. In mathematical terms, $\pi(\Delta t_2|\Delta t_1) = f_b(\Delta t_2|\Delta t_2 > \Delta t_1) = f_b(\Delta t_2)/(1 - F_b(\Delta t_1))$ where $F_b$ is the cumulative distribution function of the base distribution. Similarly one could start from the lowest layer and then derive the joint prior distribution by setting $\pi(\Delta t_{l-1}|\Delta t_l) = f_b(\Delta t_{l-1}|\Delta t_l > \Delta t_{l-1}) = f_b(\Delta t_{l-1})/F_b(\Delta t_l)$. Note however that while these prior strategies also only use the base distribution, the marginal of one characteristic time conditioned on another will be more narrow than the base distribution. This may also hold for the marginal distribution, except for the layer (upper or lower) where one starts the Markov chain.

What all these strategies have in common is that the marginal distribution of the characteristic time of one specific layer (the upper or lower layer in the cases mentioned) will have the same distribution for all models, no matter the number of layers each model has. Thus the problem of giving sharper prior distribution to the characteristic times in all layers for a model with more layers, is avoided. The distribution of other characteristic times may be more or less narrow than this base distribution.

It was deemed unfeasible to let all characteristic times have the same marginal distribution except for a location shift, and still impose the pull identification restriction on the joint distribution.

In our current application, we used the strategy of starting at the lowest layer and progressing upwards by conditioning the base distribution on the characteristic time of the layer above being less than that of the layer below. Thus the joint distribution of the characteristic time prior becomes:

$$\pi(\Delta t_1, \Delta t_2, \ldots, \Delta t_l) = f_b(\Delta t_l)\frac{f_b(\Delta t_{l-1})}{F_b(\Delta t_l)} \cdots \frac{f_b(\Delta t_1)}{F_b(\Delta t_2)}.$$

## 4. Kalman filtering, transition matrices, Kalman smoothing.

4.1. *Why Kalman filtering.* For Bayesian and classic maximum likelihood (ML) inference on model structure and parameters from a time discrete set of partially

observed states at the top level, the likelihood function is required. Assuming stationarity of the process, except possibly for deterministic trends, and assuming independent and normally distributed measurement errors, the data have a multinormal distribution.

The state process is Markovian. The covariance function and the likelihood can therefore be calculated recursively by the Kalman filter. The conditional mean and variance of the next observation, given the state of the full process at the time of the current observation, is found from the Itô representation in Eq. (2).

The Kalman filter can therefore be applied for calculating the likelihood for data modeled using linear SDE systems, as described in Eq. (1). In addition, the Kalman smoother provides predictive inference on the evolution of hidden processes in the model, and of top layer processes at time points without observations.

4.2. *The Kalman filter.* The Kalman filter rests on two foundations. First, there is a Markovian system equation for an unobserved state,

$$X_k = F_k X_{k-1} + u_k + w_k, \tag{16}$$

for $k \in \{1, \ldots, n\}$, $X_k$ is p-dimensional and where $w_k \sim N_p(\underline{0}, Q_k)$. Secondly, there is an equation for the observations,

$$Z_k = H_k X_k + v_k \tag{17}$$

where $v_k \sim N_{q_k}(\underline{0}_{q_k}, R_K)$ describes the observational noise. The time indexes $k$ can be associated with the index of the ordered observational times, $t_1, \ldots, t_{k-1}$, $t_k, t_{k+1}, \ldots, t_n$ and $n$ is the number of observations so that $1 \leq k \leq n$.

The number of observations at a single time, $k$, can be different for different times, thus $dim(v_k) \equiv q_k$ will vary and so will the dimensions of the matrices $R_k$ and $H_k$. With a continuous time scale, we assume that there is only one observation at a time, so that $dim(v_k) = 1$, $dim(R_k) = 1 \times 1$ and $dim(H_k) = 1 \times p$. The actual data has a finite time resolution, so that there were originally some time points with observations at multiple sites. For these cases, the sampling times were adjusted with steps of about 100 years in order to avoid the problem of dealing with multiple observations at the same time. These adjustments are far smaller than the time resolution, so they should not significantly affect the results.

Both noise contributions, $w_k$ and the states $v_k$ are normally distributed and independent. This is encouraging, since the linear SDE framework in this work also has states that are normally distributed. The system state is Markovian and with linear updates, see again Eq. (2). Furthermore, the observational noise is assumed normal and independent, as in the Kalman filter.

The SDE framework described here works in continuous time, but the transition matrix, $F_k$, the covariance matrix, $Q_k$, and the additive change, $u_k$, can be found

using the SDE framework described in the main text. Thus, starting from Eq. (4), the contents of the state in Eq. (16), becomes:

$$F_k = V^{-1}\mathrm{e}^{\Lambda(t_k - t_{k-1})}V$$

$$u_k = V^{-1}\int_{t_{k-1}}^{t_k} \mathrm{e}^{\Lambda(t_k-u)}Vm(u)du \tag{18}$$

$$Q_k = V^{-1}\Xi(t_k, t_k, t_{k-1})(V^{-1})' \tag{19}$$

where $\Xi(t, u, \upsilon)$ and $\Omega$ are defined as in Eq. (5).

The integral in Eq. (18) can be solved analytically if $m(t)$ is a constant, see Eq. (6). This will be the case with a fixed layer below the lowest stochastic layer (i.e. an OU process) and no external time series. With an external time series, one needs to perform a numerical integration. With these expressions, the state equation of the Kalman filter is available.

The observational formula described in Eq. (17) is simpler, as the observations are one-dimensional. Element number $j$ of $H_k$ at time $k$ will be $H_{k,(1,j)} = I(\text{observation } k \text{ is performed at location } j)$. Thus $R_k^2 = s_{t_k,j}^2/n_{t,j}$, where $s_{t_k,j}$ is the sample variance at time $t_k$ and site $j$ and $n_{t_k,j}$ is the corresponding number of samples.

With both the state and observational equations available, the Kalman filter can be used. For a given time index $k$, the Kalman filter gives inference on the state, $X_k$, can be made conditioned on the previous observations, $Z_1, \ldots, Z_{k-1}$ as well as one the observation up to and including the present, $Z_1, \ldots, Z_k$. Defining $\hat{X}_{k|l} = E(X_k|Z_1, ..., Z_l)$, $P_{k|l} = Var(X_k|Z_1, ..., Z_l)$, $\hat{Z}_k = E(Z_k|Z_1, ..., Z_{k-1})$ and $S_K = Var(Z_k|Z_1, ..., Z_{k-1})$, one gets a prediction step

$$\begin{aligned} \hat{X}_{k|k-1} &= F_k\hat{X}_{k-1|k-1} + u_k \\ P_{k|k-1} &= F_k P_{k-1|k-1} + Q_k \end{aligned} \tag{20}$$

and an updating step

$$\begin{aligned} \hat{Z}_k &= H_k\hat{X}_{k|k-1} \\ S_k &= H_k P_{k|k-1}H_k' + R_k \\ K_k &= P_{k|k-1}H_k'S_K^{-1} \\ \hat{X}_{k|k} &= \hat{X}_{k|k-1} + K_k(Z_k - \hat{Z}_k) \\ P_{k|k} &= (\tilde{1} - K_k H_k)P_{k|k-1}. \end{aligned} \tag{21}$$

In this application, it is the first two lines of the updating step that are of importance. It gives the probability density of a new data point conditioned on the previous data, which is what is needed for calculating the likelihood, $L(\theta) =$

$\prod_{k=1}^{n} f(Z_k|Z_1,\ldots,Z_{k-1}) = \prod_{k=1}^{n} N(Z_k|\hat{Z}_k, S_k)$. Here $N(x|\mu,\sigma^2)$ denotes the probability density function of the normal distribution with argument $x$, expectation $\mu$ and variance $\sigma^2$.

With the number of data $n$ large compared to the number of latent processes, $p$, and preferably with sparse matrices, this method will provide a faster algorithm for calculating the likelihood than using the covariance matrix of the measurements. It must be mentioned that while calculating the likelihood using Kalman filtering is efficient when considering a high number of data, $n$, it can be inefficient when considering a high number of process states, $p$. This is because one needs to multiply state-related matrices when using this method. Thus the likelihood calculation can be of computation cost of order $O(np^3)$, while using a precision matrix method, the cost is of order $O(n^3)$ (provided analytical expressions for the covariance has been derived). As long as $p << n$, this means a Kalman filter will tend to be more efficient at calculating a likelihood than the precision matrix method. For special cases, where $p$ is comparable to $n$, the computational cost of the Kalman filter could be worse than $O(n^3)$. If the matrices are sparse, as they are in our application, using this fact can help alleviate this problem.

4.3. *The Kalman smoother.* While the Kalman filter deals with updating the inference on states and observations conditioned on previous observations, the Kalman smoother gives the inference on states conditioned on all observations.

The Kalman smoother operates by the following updating rules, where one starts at the last time index, $n$:

$$\begin{aligned} C_k &\equiv P_{k|k} F_{k+1} P_{k+1}^{-1} \\ \hat{X}_k^{(s)} &= C_k(\hat{X}_{k+1}^{(s)} - \hat{X}_{k+1|k}) \\ P_k^{(s)} &= C_k(P_{k+1}^{(s)} - P_{k+1|k})C_k' \end{aligned} \tag{22}$$

where the notation $(s)$ stands for Kalman smoother results (conditioned on all observations). See [9] for a description of the Kalman smoother, using much of the same notation as here.

Note that time points with no observations can be introduced here, in order to get inference about the process states for other time points than those in the dataset. A time point, $k$, with no observations, will be updated as $\hat{X}_{k|k} = \hat{X}_{k|k-1}$ and $P_{k|k} = P_{k|k-1}$. Having a continuous time stochastic model means that one gets transition and covariance matrices for such inserted time points. Thus one can get a picture of the mean and uncertainty of the process states at any given time point.

## 5. Practical restrictions.

5.1. *Number of layers.* While the number of layers described in section 2.2 of the main text is not restricted, for practical purposes such a restriction is needed. Certainly, a finite number must be used in order to have the analysis run in finite time. Also, an eigenvalue decomposition is needed in order to assess correlations and do Kalman filtering. Finding an analytical rather than numerical solution for the eigenvector matrix, avoids the computation of the eigen-decomposition each time a new parameter value is studied, but can restrict the number of layers one is willing to contemplate.

In the analysis, we looked at a maximum of three layers. Since the best model among the models studied turned out to be a three-layered model, it may be wise in future analysis to go even further to a four-layer model, in order to study if even more layers may be necessary. Still, the three-layered models were only marginally better than the two-layered models according to the analysis. Thus a four-layered model can be expected to perform worse.

In order to keep the dimensionality of the analysis constant for different models, all models were initially formed as three-layered models. If a two-layered model was wanted, this was achieved by ”collapsing” one layer. Collapsing a layer is done by setting the diffusion to zero ($\Sigma_{layer} = \tilde{0}$) and by setting the pull very high ($\alpha_{layer} = (4 - layer) \times 10^8 My^{-1}$, which corresponds to the characteristic time being less than a year). With such rapid dampening of previous states, the processing belonging to a site in this layer will track the corresponding site process of the lower layer almost exactly. Thus, the layer specified will be removed from the dynamics of the state, while retaining the dimensionality. Going to one layer (or even zero) can similarly be achieved by collapsing two (or three) layers.

5.2. *Regionality.* One or several parameters (expectation, layer-specific pull or diffusion) may be regional. A regional parameter means that it will be different for different sites. However, in order to keep the number of models and the model complexity under control, only one parameter at a time was allowed to be regional. Making a parameter regional means increasing the number of parameters by five, for our six site dataset. For a three-layered model, the possibility of having one regional parameter means that seven different regionalizations must be studied, in addition to models with no regional parameters. If two regional parameters were possible, this would give a total of $28 = 7 + 7 \times 6/2$ new models to study for each model with no regionalization. For three regional parameters, the situation grows even worse. Thus restricting the number of regional parameters to only one, made sense. However, the possibility of even more regional parameters could be explored by starting with the best model with only one regional parameter and then increasing the number of regional parameters in a step-wise fashion.

## 6. Numerical methods.

6.1. *Finding the maximum likelihood.* Various ways of doing hill-climbing constitute a fairly simple set of methods for optimization and have been implemented in many programming libraries (including the GNU Scientific Library, GSL, which was used for the analysis in this paper). However, the problem with hill-climbing is that only one local optimum is found. The starting point in parameter space will determine which of the local optima are found. If there is only one optimum, then of course all starting points will result in that same optimum. However, for most models, the optima found could vary with varying starting points, meaning that more than one local optimum existed. Thus a hill-climbing method on its own could not find the global optimum required in an ML approach when the likelihood is multimodal.

There are more sophisticated optimization methods, like genetic algorithms and simulated annealing that could possibly find the global optimum relatively robustly, but such methods tend to be difficult to implement and take a lot of computer resources to run. A simulated annealing method was tried, but for the simulation parameter used, it did not converge for execution times comparable to the simpler methods described here.

Instead, we opted for using a shotgun approach, where we started from 50 different starting points and found the largest optimum among the 50 runs. With 50 starting points and using the initial model, the algorithm seemed to converge to the same optimum all times, indicating that this could very well be the global optimum. Note that while this approach may work for a moderate number of parameters, it is not expected to work so well for a high-dimensional model.

In Figure 3a, we have shown the outcome of ML optimization from 1000 different starting points drawn from a wide distribution for the original model. The optimized diffusion and characteristic time of the first layer is shown as a scatter plot, indicating that while some results are more probable than others, there is great variance in the optimization outcome. Some ridges with more outcomes than others can be identified, indicating that the likelihood surface could be relatively constant at these ridges. One of these ridges could correspond to the global maximum, as seen in Figure 3b. Here the likelihood is plotted against the characteristic time of layer 1. As can be seen, there is a set of values for the characteristic time that yields approximately the same likelihood, which can indicate that one is dealing with a single peak, but with great uncertainty about the actual value for the parameter. This is further supported by the Bayesian analysis.

The absolute maximum of these 1000 runs had a value for the logarithmic likelihood of about 211.76. About 11% of the runs ended with a logarithmic likelihood higher than 211.00. Thus more than 25 runs are needed in order to ensure that the probability for not finding this optimum is less than 5%. For 60 runs, the probability for not finding this peak drops to less than 0.1%. However, for different models,

this situation gets even worse. For the best model according to the Bayesian analysis, the ML approach ended around the best optimum only in about 1% of the runs.

Another variant of the shotgun approach, which didn't use a vague prior notion of reasonable parameter values, but rather the posterior samples from an Markov chain Monte Carlo (MCMC) algorithm, seemed to be much more efficient. A low number of hill-climbing iterations from the posterior samples, yielded stable results. This efficiency is of course conditioned on the choice of doing MCMC sampling in the first place. Bayesian analysis yields a competing inference, making the ML analysis optional. Still, if classic information criteria like Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) are wanted, as well as Bayesian analysis of each model, then such an approach may be justified.

6.2. *Bayesian MCMC analysis.* For the analysis of a single model in the Bayesian setting, an MCMC method was used. The proposal distribution for the Metropolis-Hastings algorithm was chosen to be as simple as possible, namely a random walk Metropolis algorithm [6]. The reason for choosing such a simple MCMC method aim to explore many different models using the same algorithm. The scale of the random walk for each parameter was chosen by an adaptive term in the burn-in phase, adjusting the random walk scale to get an acceptance rate of about 1/3.

A parallel tempering [2] scheme was implemented in order to deal with the multimodality of the likelihood found during ML optimization. The tempering 'temperature' strategy was chosen as $T_i = 2^{i-1}$ with $i \in \{1, .., m\}$ where $m$ is the number of tempering chains. Thus, the number of tempering chains determines the whole tempering algorithm. Only two chains were used in the large model comparison analysis, so $T_1 = 1$ and $t_2 = 2$.

The total input needed to run the algorithm was the number of iterations in the burn-in phase $n_{burn-in}$, the number of samples wanted $N$, the number of MCMC steps in between those fetched for the analysis $n_{indep}$ and the number of tempering chains, $m$. These algorithmic specifications were set using initial runs in order to check how much burn-in was needed, how many MCMC samplings were needed in order to get approximate independence between samples and how many samples were necessary in order to get roughly stable results. In the multi model analysis, these values were chosen to be $N = 2000$, $n_{indep} = 10$, $n_{burn-in} = 3000$ and $m = 2$. Note that these algorithmic parameters were used for all models, so for some models, the samples may not have been as close to independent as one would have wanted. This combined with the method described in the next sub-section was however sufficient to get stable model selection. More robust parameter estimates could be gotten with $N = 10000$ and possibly with $n_{indep} > 10$.

6.3. *Bayesian model likelihood (BML) calculation.* The numerical method we use is the importance sampler described in [7] and [8], where the MCMC samples are utilized indirectly. The reason is that the harmonic mean method, which utilizes the MCMC samples directly, is rather unstable. The mean vector and estimated covariance matrix of the samples, $m_{MCMC}$ and $V_{MCMC}$ respectively, are used for making a multi-normal proposal distribution for the importance sampler, $q(\theta) \sim N(m_{MCMC}, V_{MCMC})$. With $N$ samples from this distribution, $\theta_1, \ldots, \theta_N$, the BML is estimated as

$$\hat{BML} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(D|\theta_i)\pi(\theta_i)}{q(\theta_i)} \tag{23}$$

where $D$ is the data, $f()$ is the likelihood, $\pi()$ is the prior and $q()$ is the proposal distribution.

The efficiency of an importance sampler increases the more the proposal distribution approaches the contents of the integral (up to a normalization constant). The closer to normality the posterior distribution is, the more efficient this method is expected to be. The method works comparatively well in our case, yielding stable BML estimates ($sd(\log(\hat{BML})) < 0.1$ for our application) without taking more time than the MCMC sampling itself. This methodology is similar to, but simpler than the bridge sampling procedure of [5].

**7. Implementation issues.** The implementation of inference was initially done through ML optimization using the tool ADMB, which utilizes automatic derivation to perform hill-climbing (see section 6.1). In ADMB, the user makes a configuration script containing parameter specifications, starting points and C code for the likelihood. This script is then transformed into an executable program that delivers output files. The likelihood was computed using the covariance structure rather than the Kalman filter. Because multiple starting points were needed, see section 6.1, some scripting on top of this was needed. Using scripting, the system could be run several times and the maximal likelihood from this combined run could be found. As debugging and top-level scripting was labor-intensive in this setting and a more code-intensive Kalman filter combined with multiple model specifications was needed, it was decided that a single C++ framework would be better at performing the inference than an ADMB framework. Speed tests indicated that C++ using hill-climbing methods found in the GSL, performed only marginally slower than our ADMB implementation when doing multiple starting point hill-climbing.

Thus a C++ code base using GSL and the LGPL library Hydrasub (http://folk.uio.no/trondr/hydrasub) was used for later work. This also made it easier to consider a Bayesian approach using the same framework and the same code for the Kalman filter for calculating the likelihood. Model choices were represented

in the code using global variables, used in order to interpret the array coming in to the likelihood function. The random generator of GSL was used together with the built-in C routine *drand48()* for sampling from a uniform distribution.

## 8. Data issues.

8.1. *Normality.* The model framework proposed in this study uses the assumption of normality. The observed distribution (Fig. 2a) has however slightly heavier tails. This might be due to transients in the process, some non-normality in the log size distribution within samples (Fig. 2b), or other slight departures from the model. The mean log sizes are however correlated, and Fig. 2a should be interpreted with care.

Some of the samples exhibit a bi-modal distribution. Whether the coccolith size data reflect speciation (and extinction) events within the *Coccolithus* lineage, is an interesting question that will be studied elsewhere. Since size is strictly larger than 0 and we wish to use normal models, we did a log-transform of the sizes (which was originally measured in $\mu m$). Most of the samples are large enough to allow mean log size to be approximately normally distributed and we proceed under this assumption.

Bimodality due to speciations and exctinctions could introduce some extra complexity into the process that breaks the linearity of the system as well as introduce non-normality in some of the samples. If these events happen with different rates at different sites, this could explain the regionality of the models that were deemed best in the analysis. There is thus scope for further study of this dataset. But for modeling phenotypic means this is deemed an unnecessary complication, as there are in average 97 measurements per sample and as the central limit theorem tells us that the sample means should be close to normally distributed.

8.2. *Population variance.* Variability within samples is mainly due to variation in the respective populations. There might be a slight error present in the individual size measurements, but this is assumed to be much smaller than that due to the population variability. Population variability varies over time and between sites. This variability is smoothed out by a General Additive Model, with gamma distributed stochastic elements, in order to reduce sampling variability in the standard errors of population mean log body size. With $\widehat{s_t}$ being the GAM smoothed standard deviation in a sample of size $n_t$ at time $t$, the standard error of the mean $\widehat{s_t}/\sqrt{n_t}$ is used for the standard deviation of the error in measuring the population mean. Since typical samples have almost one hundred measurements and since we have performed the GAM smoothing, the sampling variability in the standard errors is here ignored. Fig. 4 shows sampling standard deviations together with GAM smoothings for Site 752.

## 9. More on simulation studies.

9.1. *Simulation studies when using the pull identification restriction.* Initially we did the simulation study using the pull identification restriction. For a three-layered model, we used the best model for the original dataset according to BML, see Table 2, with ML parameters estimates. Two-layered and one-layered models were then created simply by dropping the top layer or the top and middle layer respectively. We then simulated twenty datasets for each of these three cases. The same amount of data was produced in each dataset and the samples within each dataset were situated at the same sample times and having the same sample uncertainty as the real data. The state processes themselves were however simulated using these models, which was what generated the new datasets.

The results are shown in Table 1. From this modeling, it can be seen that when the pull identification restriction is being used, the information criteria AIC, AICc (corrected AIC) and BIC behaves approximately equally well. The MCMC-based deviance information criterion (DIC) criterion did not function satisfactory, however. For this specific case, an information criterion (except DIC) seems likely to support the right type of model for one and two-layered models, while it will wrongly conclude with two layers if a three-layered model produced the artificial data. The Bayesian model likelihood (BML) seems likely to identify one-layered and three-layered models, but tends to conclude with three layers also when a two-layered model has produced the data. All in all, the AIC seems slightly better than the rest for this case, but it does not support a property inference like BML does. Both methods seem capable of distinguishing between one-layered and multi-layered cases. Thus we will use both in an analysis.

When initially tested, this analysis suggested an error in our treatment of the combination of prior distribution and pull identification restriction, which resulted in support for over-complex models also for data produced by a one-layered model. See section 3.2 for how the prior was corrected.

An alternative approach is simply to disregard the pull identification restriction for BML purposes, which we will also test for the real data. Table 1 also shows the result for this approach, where both AIC, BIC and BML correctly supports a one and two-layered model most of the time, but incorrectly supports a two-layered model when a three-layered model produced the artificial data. When using Fisher's exact test for contingency tables, no difference between AIC, BIC and BML was found.

The overall results suggested to us that we could expect to be able to distinguish between a single and multi-layered case, but that we could not expect to infer three layers even when such a system had produced the data. However, note that the conclusions regarding inference strength is conditioned on the specific parameter

values used here. The results also seem to suggest that AIC and BML can be expected to agree when the pull identification restriction is not enforced.

9.2. *Description of the best one-, two- and three-layered models used in the simulation section of the main paper.* In the main paper, a simulations study was implemented using the best one-, two- and three-layered model according to BML and without imposing the pull identification restriction. These three models were as follows:

1. The best one-layered model was stationary, regional diffusion parameters and no inter-regional correlations. The ML estimates for the parameter set was as follows: $e^{\mu} = 7.34\mu m$, $\Delta t_1 = 0.96My$, $\sigma_{1,612} = 0.156My^{-1/2}$, $\sigma_{1,516} = 0.035My^{-1/2}$, $\sigma_{1,752} = 0.215My^{-1/2}$, $\sigma_{1,806} = 0.115My^{-1/2}$, $\sigma_{1,525} = 0.106My^{-1/2}$, $\sigma_{1,982} = 0.219My^{-1/2}$.

2. The best two-layered model was also stationary. It had regional pull and inter-regional correlations on the second layer. The ML estimates for the parameter set was as follows: $e^{\mu} = 7.42\mu m$, $\Delta t_1 = 18ky$, $\sigma_1 = 0.407My^{-1/2}$, $\Delta t_{2,525} = 0.97My$, $\Delta t_{2,612} = 5.4ky$, $\Delta t_{2,516} = 0.92My$, $\Delta t_{2,752} = 1.03My$, $\Delta t_{2,806} = 1.0ky$, $\Delta t_{2,982} = 2.6My$, $\sigma_2 = 0.140My^{-1/2}$, $\rho_2 = 0.556$.

3. The best three-layered model was stationary and had inter-regional correlations in the third layer. The second layer was deterministic and had regional pulls. The ML estimates for the parameter set was as follows: $e^{\mu} = 7.42\mu m$, $\Delta t_1 = 11ky$, $\sigma_1 = 0.482My^{-1/2}$, $\Delta t_{2,525} = 133ky$, $\Delta t_{2,612} = 215My$, $\Delta t_{2,516} = 0.41ky$, $\Delta t_{2,752} = 0.95My$, $\Delta t_{2,806} = 3.7Gy$, $\Delta t_{2,982} = 0.38ky$, $\sigma_3 = 0.166My^{-1/2}$, $\rho_2 = 0.664$.

The two- and three-layered models are also described Table 4, but not the one-layered model, as it was found far down in the list (at the 66th place) of best models according to BML (and also according to AIC).

## 10. More on the results.

10.1. *Top 5 models according to AIC and BML with and without using the pull identification restriction.* In our study, we used both AIC and BML as alternatives for doing model selection. Furthermore, we formulated a pull identification restriction, which may be wise to use in single-site cases or when not contemplating models with regional (i.e. site-specific) pulls and thus characteristic times, but which may be overly restrictive when it comes to models with regional pulls. Thus we did the analysis both with and without this restriction. In Tables 2-5 we present the top five model using all these four combinations of model selection.

In Table 2, it can be seen that the ranking is tight, as reflected by the BML of the highest and lowest ranking models, which yield a Bayes factor of only 1.67 in favor of model number one versus model number five. It also turns out that the five

models are very similar in structure and in parameter estimates for the common parameters. Model 3 is two-layered rather than three-layered and from the estimates it can be seen that the characteristic time and the diffusion in layer one of this two-layered model can be identified with the second layer in the remaining four models with three layers. Thus it is the existence of the top layer in the three-layered models that can be contested. Except for that, the parameters are fairly much the same as for the other models. The model variants seem to come only from the correlation structure of the fast-moving top and middle layers, where correlations will very soon be washed out anyway, and from whether or not the top layer is deterministic. Since the top layer is very fast-moving, stochastic contributions in general will very soon be washed out, so that it is difficult to say whether there are stochastic contributions in that layer or not. Thus the structural differences in the top models make very little difference for the description of the process itself. It's also worth noting that sites 612 and 806 have much shorter characteristic times at the lowest layer, than the rest. This will in practice make this layer flat for these two sites, and thus make them two or one-layered processes (for three- or two-layered models respectively) unconnected to the other sites. Also note that the first layer seems to have a characteristic time smaller than the smallest time intervals in the data (10ky), so that the only role of this layer is to add noise and otherwise smooth the process of the second layer.

Table 3 shows that AIC model selection yields different results than BML in the case of using the pull identification restriction. The top three models are two-layered while number four and five are three-layered. It seems that it is the existence of the second layer in a three-layered model that is in doubt. Other than this, the parameter estimates are quite the same over all five models. The model variation stems from uncertainty about the correlation structure and stochasticity of the upper layer, just as for the top BML models. While for the top BML models it was the characteristic times that were regional at the lowest layer, for the top AIC models when using the pull identification restriction, it is the diffusions of the lower layer that are regional. Site 612 and 806 have very low diffusion, thus flattening out the processes for these two sites. Again, these two sites have one layer less than the rest and since the only certain inter-regional correlation is found in this layer, they are in practice disconnected from the other layers (and from each other), just as for the models in Table 2.

When not using the pull identification restriction, the results change again. However, it is worth noting from Tables 4 and 5 that the best model according to BML is the second best according to AIC and the best model according to AIC is the second best according to BML. Thus AIC and BML no longer differ so much when it comes to model selection. Also, the parameter estimates are found to violate the pull identification restriction, thus yielding correlation structures which are differ-

ent from what can be achieved if the pull identification restriction was being used.

When interpreting the common two top models for AIC and BML, it is worth noting that with a deterministic second layer and with small characteristic times, a second layer process will just be a copy of the third layer. Thus the middle layer will be of no importance in such cases, namely for sites 516 and 982, one is left with a two-layered process. The lower layer has inter-regional correlations, so these sites will still be connected to the other sites. If the characteristic time is very high, on the other hand, the second layer process will in practice flatten out the third layer, and one is left with only one dynamic layer of any importance, namely the top layer. Thus sites 612 and 806 are again singled out, though in this case they loose two rather than one layer. As previously though, one of the layers they loose contain the only strong inter-regional correlation. Correlations in the upper layer are of little importance since the short characteristic times will wash out the correlated contributions fairly fast.

According to AIC, the top models are all three-layered with structural variation only in the correlation structure and the stochasticity of the first and second layer, see Table 5. Only model number five has parameter estimates somewhat different from the other models, and the difference is not too great.

BML ranks three two-layered models as the third to fifth best models. The parameter estimates belonging to the regional characteristic times seem to be very different from those of model one and two. However for sites 525 and especially site 752 (the two sites that did not in practice loose one or two layers according to the best AIC models), the parameter estimates are not so different. Site 612 and 806 now have very low characteristic times. Since there now is no layer below the second one, this however again means that the second layer will in practice be flat. Sites 516 and 982 now have moderate estimates for the characteristic times, comparable to those of sites 525 and 752. Thus while these three models may look structurally different from the two best according to BML, in practice they are not so different. The main difference seems to be that the number of layers in sites 525 and 752 are reduced from three to two. It should be noted that the fourth model has perfect correlation in the second layer. However, since the characteristic times are different, the layer two processes will be different for different sites. The other structural differences have to do with the correlation structure of the fast-moving upper layer and are thus of little importance.

These four model comparisons seem to agree on the following, namely that there are more than one layer (at least for some sites), that there are regional differences in the dynamic parameters, that there are inter-regional correlations and that one does not have a random walk in the lowest layer. The level parameter $e^{\mu_0}$ was consistently estimated to about $7.4\mu m$ and was deemed global. There seems also to be agreement in that sites 612 and 806 have less detectable layers and are discon-

nected (in practice no effect of inter-regional correlations) from the other sites and each other. Apart from that, there is disagreement between these four model selection methods in whether it is the diffusions or the characteristic times that are site-specific and if these regional parameters are found in the second or third layer. However, internally within each model selection, the models that were among the top five turned out to be very similar. Mostly, the uncertainty internally within each model selection was concerning the stochasticity and correlation structure of the top layer or the first two layers.

Since there seem to always be inter-regional correlations in the lowest layer, this indicates that each regional process in this layer is a sum of a global process and independent regional components, all of which are OU processes.

Tables 6 and 7 show the property analysis done both with and without using the pull identification restriction. In both cases, there seems to be good correspondence between the properties of the top five BML models and the properties inferred by this analysis. However, since there is some difference between the best model according to AIC and BML when using the pull identification restrictions, the match between property analysis and best AIC model is slightly worse. Still, the main features of multiple layers, the existence of regional parameters, inter-regional correlations and stationarity are supported by both.

## 11. Stability, robustness and data contributions.

11.1. *The effect of the parameter prior distributions.* Bayesian model comparison can be hampered by sensitivity to the prior probabilities. By a mistake, we did a full analysis with a much wider prior distribution for parameter $\mu_0$ than what we intended. The top 5 models were still retained, with approximately the same parameter estimates, indicating little sensitivity at least to this aspect of the prior. Some other parameter categories could be more sensitive.

However we did test for what we believed are the most sensitive parameters, that of the pull parameters and the diffusions. A prior with 95% of the probability mass within $1y$ to $300My$, as compared to within $1ky$ to $1Gy$, was used for the characteristic times (which determine the pulls). Also the 95% credibility interval going from 0.001 to 7 rather than from 0.02 to 3.5 was used for the diffusions. As the pull identification restriction seemed to be less motivated when there was regional pulls, the analysis was done without using it.

If ML optimization could be performed perfectly, then AIC should not be at all sensitive to the choice of a Bayesian prior distribution. However, our practical implementation of ML optimization starts from a finite set of posterior samples (this turned out to be far more efficient than a shotgun approach). If there are too few starting points used, then the best ML estimates after optimization (hill-climbing) may converge to a local rather than global optimum. Thus if one tries two different

prior distribution (or even if one did not change the prior but only did two different analysis), one may end up with different estimates. The probability of reach a certain convergent point may also depend on the prior, which can add extra variability into the ML optimization. Thus a comparison of the best models according to AIC may serve as a test of the stability of the ML estimates this method yields. The result of this test was that models one to three were in agreement, while model four and five were switched. The difference in the maximum likelihood between the two models was estimated to be less than 0.1 according to the previous analysis, so since the numerical optimization will not converge exactly to the optimum, these small differences are to be expected. All in all, it the ML estimates seem to be stable.

According to BML, the same model as before was the best and the runner-up model was identical also. The next three two-layered models were replaced with three-layered models having much the same structure as model one and two. The third model was the same as one of the top five models according to the previous AIC analysis (except for not having inter-regional correlations on the upper layer, which does not make a practical difference). The fourth model was the same as the fourth model according to AIC (except for the lack of upper layer correlations). The fifth model was the same as the fourth except that it had inter-regional correlations in the second layer.

The over-all impression is that little was changed when it comes to uncertain properties (like that of having two or three layers), even though there may be slight changes in how much support these properties get. A property analysis showed this to be the case. While there was probability weight 32% and 60% for two and three layers respectively, after changing the prior the weights changed to 26% and 73%, respectively. The support for three layers is not great and the robustness analysis suggest that under variations in the prior (while keeping the credibility bands within reasonable boundries), but it anyway seem to suggest slight support for three layers rather than two. And even more so than previously, one layer seem to be ruled out, now having a weight of only 1.3%.

Regionality in pull is a little more doubtful now, having a weight of 70% rather than 88%, but again there is still support for the same property. The third layer is given a weight of 62% rather than 69% of having regional parameters. Inter-regional layers with non-perfect correlation is given a weight of 83% while before it was 85%. The existence of deterministic layers were found to have a posterior weight of 91% before while it is now 81%. And lastly, non-stationarity (random walk) on the lowest layer was found to have a weight of 98% now while 99% before.

For all these properties, it seems that the shifts in the property weights are too small to change which property is favored by the weights. Thus we can feel rela-

tively secure about the properties in this case. The agreement concerning the top models between BML and the classical measure of AIC, which does not use priors, should also suggest robustness. There was less agreement if the pull identification restriction was used, which may indicate that using this restriction may cause some instability.

11.2. *The effect of portions of the data.* The effect of portions of the data on the outcome of the analysis can be studied by removing such portions and redoing the analysis. When the credibility bands are compared, this can show how a certain subset of the data affects the inference on the state of the parameters, by looking at how narrow the band is before and after including the data. Also, if there are parts of the data that pull in different directions, this can suggest that a more complicated model may be needed.

In Table 8, the result of such an operation is shown. Various sites have been removed and compared to when all sites are included. Also, the data was subdivided into $age < 10My$ and $age > 10My$. The regional parameters for sites that have been removed, are of course not available and are marked NA. The same is the case for parameters belonging to sites with little or no data younger than 10My.

As for the effect of data from Sites 525 and 982, these sites seem to agree well with the rest so that removing these sites does not affect much the credibility interval of the inter-regional correlation in the lowest layer. Site 752 seems to contribute so much to this parameter, that the credibility interval encompasses zero without data from this site. Site 516 on the other hand, seems to yield a stronger result for the inter-regional correlation when this data is removed than when it is included, suggesting that this site is correlated to the rest but having a correlation which is less strong. These observations, together with the disconnectedness of sites 612 and 806 through their small pulls in the second layer, suggests that a model with unequal correlation between sites, may be justified. However, that will increase the number of correlation-parameters from 1 to 15 for each layer with inter-regional correlations (which in this case is just the lower layer).

As for subdividing the data in time, it's reassuring to see that the expectancy seems to be approximately the same for the two time periods, though the uncertainty of the expectancy is much larger in each subdivision than in the total dataset.

No signs conflicting signals were found. Removing some portions of the data did not yield drastically different estimates, but seemed rather to just increase the uncertainty. The only possible exception to that might be the impact of site 516 data on the inter-regional correlation, though it should be noted that while the credibility bands are shifted when removing site 516 data, these bands are overlapping before and after this procedure. Thus there are no strong suggestions of inconsistencies between data and model.

**12. Source code.** Further developments of the source codes for our analysis programs, can be found on the web page http://folk.uio.no/trondr/layered.

TABLE 1

*Results for the analysis of artificial data. The table shows the number of times the correct model is inferred using different techniques, in a total of 20 artificial datasets per model. A parenthesis shows which model, represented by the number of layers, is preferred most often if the number of correct model inferences drops below* 50%.

| Actual model | AIC | AICc | BIC | BML | DIC |
|---|---|---|---|---|---|
| When using the pull identification restriction | | | | | |
| One layer | 20 | 20 | 20 | 15 | 7 (3) |
| Two layers | 20 | 20 | 19 | 3 (3) | 0 (3) |
| Three layers | 0 (2) | 0 (2) | 0 (2) | 18 | 20 |
| When not using the pull identification restriction | | | | | |
| One layer | 20 | 20 | 20 | 18 | 6 (3) |
| Two layers | 19 | 19 | 19 | 17 | 2 (3) |
| Three layers | 0 (2) | 0 (2) | 0 (2) | 0 (2) | 16 |

## References.


[1] FEDCHENKO, N. V. and PRIGARIN, S. M. (2004). Boundary Value Problems for Linear Stochastic Differential Equations. *Russian Journal of Numerical Analysis and Mathematical Modelling*, **19(6)**, 507–525.

[2] GEYER, C. J. (1991). Markov chain Monte Carlo Maximum Likelihood. in E. Keramigas (ed.) *Computing Science and Statistics: The 23rd symposium on the interface*, Interface Foundation, Fairfax, 156-163.

[3] GRANGER, C. W. J. (1969). Investigating Causal Relations by Econometric Models and cross-spectral methods. *Econometrica* **37(3)**, 424–438.

[4] GRANGER, C. W. J. (1981). Some Properties of Time Series Data and Their Use in Econometric Model Specification. *J. Econometrics* **16**, 121–130.

[5] MENG, X. and WONG, W. (1996). Simulating Ratios of Normalizing Constants via a Simple Identity. *Stat. Sinica*, **6**, 831-860.

[6] METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. and TELLER, A. H. (1953). Equation of State Calculations by Fast Computing machines. *Journal of Chemical Physics*, **21**, 1087-1092.

[7] REITAN, T. and PETERSEN-ØVERLEIR, A. (2009). Bayesian methods for estimating multi-segment discharge rating curves. *Stochastic Environmental Research and Risk Assessment* **23(5)**, 627–642.

[8] REITAN, T. and AAS, K. (2011). A New Robust Importance Sampling Method for Measuring VaR and ES Allocations for Credit Portfolios. *Journal of Credit Risk* **6(4)**.

[9] SÄRKKÄ, S., VEHTARI, A., and LAMPINEN, J. (2004). Time Series Prediction by Kalman Smoother with Cross-Validated Noise Density. *IEEE International Joint Conference on Neural Networks*, **1-4**, 1615-1619.

[10] SCHWEDER, T. (1970). Decomposable Markov Processes. *J. Applied Prob.* **7**, 400–410.

[11] SCHWEDER, T. (2011). Causality and Markov completeness. Ms in prep., to be submitted to *Scandinavian Journal of Statistics*.

[12] SYDSÆTER, K., HAMMOND, P., SEIERSTAD, A. and STRØM, A. (2005). *Further mathematics for economic analysis*, Prentice Hall.

TABLE 2

*ML estimates for the* 5 *top ranked models according to BML among the* 710 *models under comparison with the pull identification restriction. Bayesian 95% credibility bands are shown in parenthesis. The characteristic times reported for the upper and middle layer ($\Delta t_1$ and $\Delta t_2$ ) is in units of $ky$, while those for the lower layer are in units of $My$. The expectancy term was transformed to the original scale, so $e^{\mu_0}$ has units $\mu m$. The second layer in a two-layered model and the third layer in a three-layered model will be labeled the lower layer. The middel layer will only be used for three-layered models. In a two-layered model, all middle layer parameters will be marked with the text NA.*

| Parameter | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| log(BML) | 201.63 | 201.37 | 201.36 | 201.30 | 201.12 |
| #layers | 3 | 3 | 2 | 3 | 3 |
| #parameters | 13 | 14 | 11 | 12 | 14 |
| $e^{\mu_0}$ | 7.42 (7.31-7.54) | 7.42 (7.31-7.56) | 7.42 (7.32-7.54) | 7.42 (7.31-7.56) | 7.42 (7.31-7,58) |
| $\Delta t_{upper}$ | 1.5 (0.02-13) | 0.7 (0.03-19) | 8.4 (0.1-40) | 2.1 (0.02-17) | 0.1 (0.04-90) |
| $\Delta t_{middle}$ | 3.1 (0.2-30) | 1.4 (0.2-6) | NA | 11 (0.6-17) | 2.6 (0.2-230) |
| $\Delta t_{lower,525}$ | 0.9 (0.6-3.3) | 1.0 (0.62-4.1) | 1.0 (0.6-3.5) | 1.0 (0.6-3.3) | 1.0 (0.59-4.1) |
| $\Delta t_{lower,612}$ | 0.003 (0.001-0.5) | 0.006 (0.002-0.6) | 0.01 (0.0009-0.47) | 0.02 (0.002-0.59) | 0.005 (0.002-0.94) |
| $\Delta t_{lower,516}$ | 0.9 (0.54-5.7) | 0.9 (0.50-3.8) | 0.9 (0.54-5.1) | 0.9 (0.52-6.9) | 0.9 (0.51-6.0) |
| $\Delta t_{lower,752}$ | 1.1 (0.5-14) | 1.0 (0.52-12) | 1.0 (0.50-9.6) | 1.1 (0.53-12) | 1.0 (0.52-11) |
| $\Delta t_{lower,806}$ | 0.01 (0.001-0.2) | 0.006 (0.002-0.27) | 0.01 (0.0006-0.14) | 0.01 (0.002-0.14) | 0.003 (0.001-0.53) |
| $\Delta t_{lower,982}$ | 2.6 (1.5-88) | 2.6 (1.4-75) | 2.5 (1.4-76) | 2.6 (1.5-76) | 2.6 (1.4-73) |
| $\sigma_{upper}$ | 1.2 (0.02-3.0) | 1.8 (0.03-4.6) | 0.57 (0.28-5.1) | 0 (exact) | 0.39 (0.03-5.2) |
| $\sigma_{middle}$ | 0.54 (0.04-4.1) | 0.47 (0.04-2.1) | NA | 0.58 (0.18-1.8) | 1.0 (0.02-2.4) |
| $\sigma_{lower}$ | 0.14 (0.09-0.18) | 0.14 (0.08-0.18) | 0.14 (0.09-0.18) | 0.14 (0.08-0.18) | 0.14 (0.08-0.18) |
| $\rho_{upper}$ | 0 (exact) | -0.16 (-0.19 - 0.97) | 0 (exact) | 0 (exact) | 0 (exact) |
| $\rho_{middle}$ | 0 (exact) | 0 (exact) | NA | 0 (exact) | -0.17 (0.19 - 0.95) |
| $\rho_{lower}$ | 0.54 (0.07-0.89) | 0.54 (0.09-0.92) | 0.53 (0.07 - 0.90) | 0.54 (0.08-0.90) | 0.55 (0.08-0.95) |

TABLE 3

*ML estimates for the* 5 *top ranked models according to AIC among the* 710 *models under comparison with the pull identification restriction. Bayesian 95% credibility bands are shown in parenthesis. The characteristic times reported for the upper and middle layer ($\Delta t_1$ and $\Delta t_2$) is in units of $ky$, while those for the lower layer are in units of $My$. The expectancy term was transformed to the original scale, so $e^{\mu_0}$ has units $\mu m$. The second layer in a two-layered model and the third layer in a three-layered model will be labeled the lower layer. The middel layer will only be used for three-layered models. In a two-layered model, all middle layer parameters will be marked with the text NA.*

| Parameter | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| AIC | -434.80 | -434.76 | -434.57 | -433.10 | -433.09 |
| #layers | 2 | 2 | 2 | 3 | 3 |
| #parameters | 12 | 11 | 11 | 13 | 12 |
| $e^{\mu_0}$ | 7.41 (7.24-7.64) | 7.42 (7.26-7.63) | 7.42 (7.19-7.57) | 7.41 (7.26-7.63) | 7.42 (7.24-7,63) |
| $\Delta t_{upper}$ | 17 (0.6-101) | 10 (0.6-162) | 0.02 (0.02-17) | 8.0 (0.02-17) | 8.7 (0.03-18) |
| $\Delta t_{middle}$ | NA | NA | NA | 13 (0.5-80) | 9.7 (0.9-78) |
| $\Delta t_{lower}$ | 1.6 (0.7-3.2) | 1.5 (0.73-3.6) | 1.5 (0.62-2.01) | 1.6 (0.7-3.3) | 1.5 (0.7-3.4) |
| $\sigma_{upper}$ | 0.4 (0.19-1.8) | 0.50 (0.18-1.8) | 12.5 (0.02-6.9) | 0 (exact) | 0 (exact) |
| $\sigma_{middle}$ | NA | NA | NA | 0.58 (0.18-1.8) | 0.72 (0.19-1.7) |
| $\sigma_{lower,525}$ | 0.12 (0.08-0.19) | 0.12 (0.08-0.18) | 0.12 (0.10-0.21) | 0.12 (0.08-0.19) | 0.12 (0.08-0.19) |
| $\sigma_{lower,612}$ | 0.01 (0.006-0.09) | 0.01 (0.007-0.09) | 0.01 (0.01-0.09) | 0.01 (0.006-0.9) | 0.01 (0.007-0.09) |
| $\sigma_{lower,516}$ | 0.16 (0.10-0.24) | 0.16 (0.09-0.24) | 0.16 (0.15-0.28) | 0.16 (0.11-0.24) | 0.16 (0.11-0.25) |
| $\sigma_{lower,752}$ | 0.08 (0.05-0.15) | 0.08 (0.05-0.15) | 0.08 (0.08-0.17) | 0.08 (0.06-0.15) | 0.08 (0.06-0.15) |
| $\sigma_{lower,806}$ | 0.001 (0.004-0.13) | 0.0005 (0.003-0.12) | 0.001 (0.01-0.18) | 0.001 (0.003-0.12) | 0.001 (0.003-0.11) |
| $\sigma_{lower,982}$ | 0.17 (0.12-0.27) | 0.17 (0.12-0.27) | 0.17 (0.15-0.30) | 0.17 (0.12-0.27) | 0.17 (0.12-0.28) |
| $\rho_{upper}$ | -0.20 (-0.20 - 0.62) | 0 (exact) | 1 (exact) | 0 (exact) | 0 (exact) |
| $\rho_{middle}$ | NA | NA | NA | -0.19 (-0.19-0.57) | 0 (exact) |
| $\rho_{lower}$ | 0.55 (0.03-0.81) | 0.54 (0.06-0.81) | 0.53 (-0.07 - 0.53) | 0.55 (0.04-0.79) | 0.54 (0.04-0.78) |

TABLE 4

*ML estimates for the* 5 *top ranked models according to BML among the* 710 *models under comparison without using the pull identification restriction. Bayesian 95% credibility bands are shown in parenthesis. The characteristic times reported for layer one ($\Delta t_1$ ) is in units of $ky$, while those for layer two and three are in units of $My$. The expectancy term was transformed to the original scale, so $e^{\mu_0}$ has units $\mu m$. The third layer in a three-layered model will be labeled the lower layer. The middel layer will be defined as the second layer whether this is in a two or three-layered model. In a two-layered model, all lower layer parameters will be marked with the text NA.*

| Parameter | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| log(BML) | 203.77 | 203.13 | 201.11 | 200.61 | 200.53 |
| #layers | 3 | 3 | 2 | 2 | 2 |
| #parameters | 12 | 13 | 11 | 10 | 12 |
| $e^{\mu_0}$ | 7.42 (7.15-7.69) | 7.42 (7.18-7.67) | 7.42 (7.31-7.60) | 7.55 (7.30-7.70) | 7.43 (7.29-7,59) |
| $\Delta t_{upper}$ | 11 (0.3-80) | 14 (0.9-78) | 18 (0.9-276) | 180 (16-400) | 72 (2.1-220) |
| $\Delta t_{middle,525}$ | 0.13 (0.001-0.52) | 0.13 (0.002-0.52) | 1.0 (0.6-3.5) | 1.5 (0.8-5.1) | 1.5 (0.68-4.6) |
| $\Delta t_{middle,612}$ | 215 (0.57-3000) | 238 (1.3-10000) | 0.005 (0.0009-0.47) | 0.05 (0.0004-3.3) | 0.01 (0.0001-0.95) |
| $\Delta t_{middle,516}$ | 0.0004 (0.0001-0.15) | 0.002 (0.0001-0.17) | 0.9 (0.54-5.1) | 1.1 (0.04-11) | 1.2 (0.51-7.9) |
| $\Delta t_{middle,752}$ | 1.0 (0.24-2.9) | 1.1 (0.21-3.2) | 1.0 (0.50-9.6) | 2.4 (1.2-62) | 2.0 (0.58-23) |
| $\Delta t_{middle,806}$ | 4000 (2.0-10000) | 22000 (3.9-9000) | 0.001 (0.0006-0.14) | 0.001 (0.0002-0.49) | 0.0002 (0.0001-0.21) |
| $\Delta t_{middle,982}$ | 0.0004 (0.0001-0.14) | 0.006 (0.0001-0.17) | 2.6 (1.4-76) | 6.8 (3.2-300) | 5.2 (1.6-160) |
| $\Delta t_{lower}$ | 1.2 (0.64-3.7) | 1.3 (0.66-3.6) | NA | NA | NA |
| $\sigma_{upper}$ | 0.48 (0.22-2.8) | 0.43 (0.22-1.7) | 0.41 (0.18-1.8) | 0.21 (0.17-0.70) | 0.27 (0.19-1.1) |
| $\sigma_{middle}$ | 0 (exact) | 0 (exact) | 0.14 (0.07-0.17) | 0.10 (0.05-0.15) | 0.11 (0.07-0.18) |
| $\sigma_{lower}$ | 0.17 (0.11-0.24) | 0.17 (0.11-0.24) | NA | NA | NA |
| $\rho_{upper}$ | 0 (exact) | -0.19 (-0.19 - 0.28) | 0 (exact) | 0 (exact) | -0.19 (-0.20 - 0.98) |
| $\rho_{middle}$ | 0 (exact) | 0 (exact) | 0.56 (0.19-0.97) | 1 (exact) | 0.83 (0.20 - 0.98) |
| $\rho_{lower}$ | 0.66 (0.29-0.85) | 0.67 (0.32-0.88) | NA | NA | NA |

TABLE 5

*ML estimates for the* 5 *top ranked models according to AIC among the* 710 *models under comparison without using the pull identification restriction. Bayesian 95% credibility bands are shown in parenthesis. The characteristic times reported for layer one ($\Delta t_1$ ) is in units of $ky$, while those for layer two and three are in units of $My$. The expectancy term was transformed to the original scale, so $e^{\mu_0}$ has units $\mu m$. The third layer in a three-layered model will be labeled the lower layer, while the middel layer will be defined as the second layer.*

| Parameter | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| AIC | -437.14 | -436.93 | -436.75 | -435.25 | -435.05 |
| #layers | 3 | 3 | 3 | 3 | 3 |
| #parameters | 13 | 12 | 12 | 14 | 14 |
| $e^{\mu_0}$ | 7.42 (7.18-7.67) | 7.42 (7.15-7.69) | 7.42 (7.21-7.70) | 7.41 (6.94-7.84) | 7.42 (7.06-7,76) |
| $\Delta t_{upper}$ | 14 (0.9-78) | 11 (0.3-80) | 0.1 (0.01-10) | 14 (0.9-97) | 14 (3.3-380) |
| $\Delta t_{middle,525}$ | 0.13 (0.002-0.52) | 0.13 (0.001-0.52) | 0.13 (0.001-0.4) | 0.13 (0.001-0.7) | 0.13 (0.009-4.1) |
| $\Delta t_{middle,612}$ | 238 (1.3-10000) | 215 (0.58-3000) | 324 (1.2-5000) | 530 (0.01-540) | 270 (0.0007-270) |
| $\Delta t_{middle,516}$ | 0.002 (0.0001-0.17) | 0.0004 (0.0001-0.15) | 0.002 (0.0001-0.10) | 0.001 (0.0001-0.54) | 0.002 (0.0003-8.6) |
| $\Delta t_{middle,752}$ | 1.1 (0.21-3.2) | 1.0 (0.25-2.9) | 0.9 (0.10-2.6) | 1.0 (0.03-3.4) | 1.0 (0.33-24) |
| $\Delta t_{middle,806}$ | 22000 (3.9-9000) | 4000 (2.0-10000) | 2000 (1.0-10000) | 14000 (0.08-520) | 3000 (0.0004-170) |
| $\Delta t_{middle,982}$ | 0.006 (0.0001-0.17) | 0.0004 (0.0001-0.14) | 0.001 (0.0001-0.12) | 0.002 (0.002-1.9) | 0.004 (0.0004-0.15) |
| $\Delta t_{lower}$ | 1.3 (0.66-3.6) | 1.2 (0.64-3.7) | 1.2 (0.7-3.5) | 1.3 (0.63-5.4) | 1.2 (0.0003-3.7) |
| $\sigma_{upper}$ | 0.43 (0.22-1.7) | 0.48 (0.22-2.8) | 14 (0.46-14) | 0.43 (0.21-1.4) | 0.42 (0.17-0.84) |
| $\sigma_{middle}$ | 0 (exact) | 0 (exact) | 0 (exact) | 0.00001 (0.003-0.18) | 0.00005 (0.004-0.15) |
| $\sigma_{lower}$ | 0.17 (0.11-0.24) | 0.17 (0.11-0.24) | 0.17 (0.11-0.23) | 0.16 (0.06-0.24) | 0.17 (0.014-1.2) |
| $\rho_{upper}$ | -0.19 (-0.19 - 0.28) | 0 (exact) | 1 (exact) | -0.20 (-0.19 - 0.51) | -0.20 (-0.19 - 0.25) |
| $\rho_{middle}$ | 0 (exact) | 0 (exact) | 0 (exact) | 0 (exact) | 1 (exact) |
| $\rho_{lower}$ | 0.67 (0.32-0.88) | 0.66 (0.29-0.85) | 0.66 (0.27-0.85) | 0.57 (0.23-0.89) | 0.66 (-0.15 - 0.97 |

TABLE 6

*Posterior weights for different properties when the pull identification restriction is and is not used to the right and left of the forward slash, respectively. Number of models in parenthesis.*

| Property | Options | | | |
|---|---|---|---|---|
| Number of Layers: | 1 | 2 | 3 | |
| | 7.4% / 7.5 % (18) | 44.7% / 32.4% (114) | 47.9% / 60.1% (578) | |
| Regionality in: | none or $\mu_0$ | pull | diffusion | |
| | 0.4% / 0.2% (177) | 78.7% / 88.3% (259) | 20.9% / 11.5% (274) | |
| Regional parameters in: | no layer | upper layer | middle layer | lower layer |
| | 0.2% / 0.2% (177) | 20.9% / 13.1% (205) | 27.6% / 69.1% (196) | 51.3% / 17.5% (132) |
| Inter-regional | none | intermediate (6D) | perfect (1D) | |
| correlations: | 20.2% / 4.9% (50) | 66.1% / 85.2% (212) | 13.7% / 9.9%(448) | |
| No diffusion in | no layer | layer 1 | layer 2 | both layer 1 and 2 |
| | 39.1% / 7.7% (486) | 25.5% / 1.2% (132) | 35.1% / 91.1% (72) | 0.4% / 0.007% (20) |
| Random walk in | no | yes | | |
| lowest layer: | 98.1% / 99.1% (414) | 1.9% / 0.9% (296) | | |

TABLE 7

*Posterior weights of correlation structure properties, when the pull identification restriction is and is not used to the right and left of the forward slash, respectively. Only three-layered models with non-degenerate pull at the bottom layer (333 in total) were examined.*

| Inter-regional correlation at | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Top layer | N | Y | N | Y | N | Y | N | Y |
| Intermediate layer | N | N | Y | Y | N | N | Y | Y |
| Bottom layer | N | N | N | N | Y | Y | Y | Y |
| Number of models | 15 | 16 | 30 | 32 | 56 | 60 | 60 | 64 |
| Posterior probability (%) | 11.1 / 1.6 | 10.0 / 1.1 | 9.2 / 4.8 | 7.3 / 2.5 | 23.6 / 49.9 | 16.5 / 25.2 | 12.7 / 10.1 | 9.5 / 4.9 |

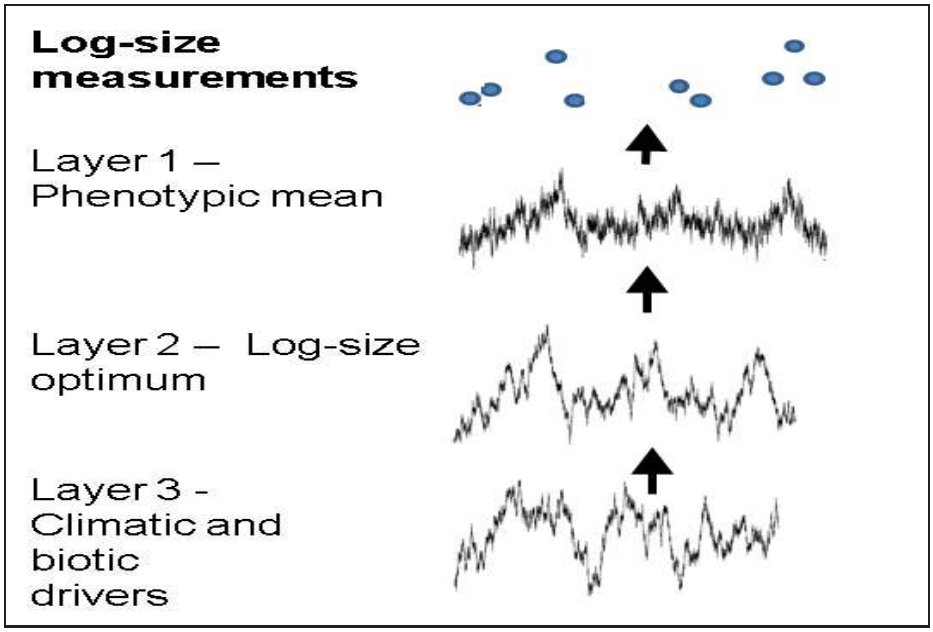



FIG 1. *A graphical representation of the modeling framework, having three labeled stochastic layers. The thick arrows represent how one series affect the other (causal flow).*

TABLE 8

*Credibility bands (95%) for parameters belonging to the best model according to Bayesian analysis, with some data retracted and without using the pull identification restriction. In general, credibility bands seem to get a little bit wider when portions of the data are excluded, which is to be expected. Parameters belonging to sites that have been removed or which do not have data in the specified time intervals, are marked with NA. Note that site 612 and 806 seem to have little impact on the uncertainty of the inter-regional correlations in the lowest layer. Since these layers are in practice disconnected from this layer, this is to be expected. Removing site 516 data on the other hand seems to enhance the estimate of the inter-regional correlation, suggesting that this site may be connected to the rest, but with less strong correlation. It's also worth noticing that the inter-regional correlation seems insignificant in each of the time periods considered, while in the full data, it reaches significance. Thus large time periods are needed in order to discover this feature. Almost all parameters seem much more uncertain with the age restrictions than without.*

| Parameter | Full | - Site 525 | - Site 612 | - Site 516 | - Site 752 |
|---|---|---|---|---|---|
| $e^{\mu}$ | 7.15$\mu m$-7.69$\mu m$ | 7.13$\mu m$-7.78$\mu m$ | 7.09$\mu m$-7.88$\mu m$ | 7.06$\mu m$-7.83$\mu m$ | 7.12$\mu m$-7.67$\mu m$ |
| $\Delta t_1$ | 0.3ky-80ky | 1.2ky-79ky | 0.9ky-150ky | 0.8ky-130ky | 0.7ky-93ky |
| $\sigma_1$ | 0.22-2.8 | 0.22-1.6 | 0.22-1.8 | 0.14-1.6 | 0.22-2.0 |
| $\Delta t_{2,525}$ | 1ky-0.5My | NA | 0.6ky-0.6My | 1ky-0.4My | 0.7ky-0.7My |
| $\Delta t_{2,612}$ | 0.6My-3.2Gy | 0.97My-6.4Gy | NA | 1.5My-6.3Gy | 1.2My-6.3Gy |
| $\Delta t_{2,516}$ | 90y-150ky | 89y-170ky | 180y-240ky | NA | 0.1ky-210ky |
| $\Delta t_{2,752}$ | 0.25My-2.9My | 0.33My-3.3My | 210ky-3.1My | 130ky-2.0My | NA |
| $\Delta t_{2,806}$ | 2.0My-10Gy | 3.7My-7.7Gy | 2.2My-7.8Gy | 2.6My-9Gy | 0.9My-6.9Gy |
| $\Delta t_{2,982}$ | 140y-44ky | 170y-0.30My | 150y-250ky | 90y-84ky | 150y-270ky |
| $\Delta t_3$ | 0.6My-3.5My | 0.6My-5.1My | 0.8My-5.0My | 0.8My-6.4My | 0.7My-3.5My |
| $\sigma_3$ | 0.11-0.23 | 0.11-0.24 | 0.09-0.21 | 0.10-0.24 | 0.11-0.23 |
| $\rho_3$ | 0.29-0.85 | 0.17-0.95 | 0.34-0.90 | 0.44-0.94 | -0.07 - 0.79 |
| Parameter | Full | - Site 806 | - Site 982 | $age < 10My$ | $age > 10My$ |
| $e^{\mu}$ | 7.15$\mu m$-7.79$\mu m$ | 7.02$\mu m$-7.88$\mu m$ | 7.26$\mu m$-7.64$\mu m$ | 5.75$\mu m$-10.1$\mu m$ | 7.00$\mu m$-7.61$\mu m$ |
| $\Delta t_1$ | 0.3ky-80ky | 2ky-100ky | 1ky-120ky | 2ky-9.7My | 1ky-230ky |
| $\sigma_1$ | 0.22-2.8 | 0.23-1.2 | 0.18-1.6 | 0.11-1.13 | 0.20-1.8 |
| $\Delta t_{2,525}$ | 1ky-0.5My | 0.5ky-0.6My | 0.5ky-0.7My | 0.4ky-40My | 0.4ky-3.4My |
| $\Delta t_{2,612}$ | 0.6My-3.2Gy | 0.9My-6.5Gy | 0.8My-4.1Gy | NA | 62ky-4.5Gy |
| $\Delta t_{2,516}$ | 90y-150ky | 120y-220ky | 110y-0.3My | NA | 190y-3.0My |
| $\Delta t_{2,752}$ | 250ky-2.9My | 170ky-3.3My | 190ky-2.7My | 10ky-41My | 72ky-0.9Gy |
| $\Delta t_{2,806}$ | 2.0My-10Gy | NA | 2.2My-8.6Gy | 20ky-37My | 2ky-2.1Gy |
| $\Delta t_{2,982}$ | 140y-44ky | 110y-230ky | NA | 180y-19My | 0.4ky-5.0My |
| $\Delta t_3$ | 0.6My-3.5My | 0.7My-4.8My | 220ky-1.9My | 190ky-37My | 170ky-15My |
| $\sigma_3$ | 0.11-0.23 | 0.09-0.21 | 0.11-0.50 | 0.06-1.1 | 0.03-0.61 |
| $\rho_3$ | 0.29-0.85 | 0.31-0.88 | 0.17-0.94 | -0.05 - 0.97 | -0.10 - 0.97 |

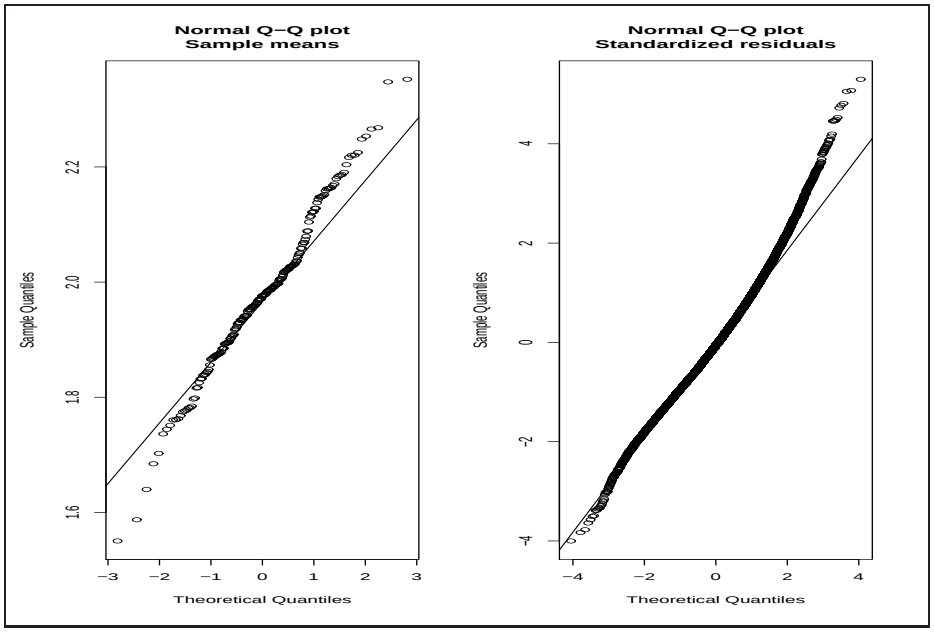
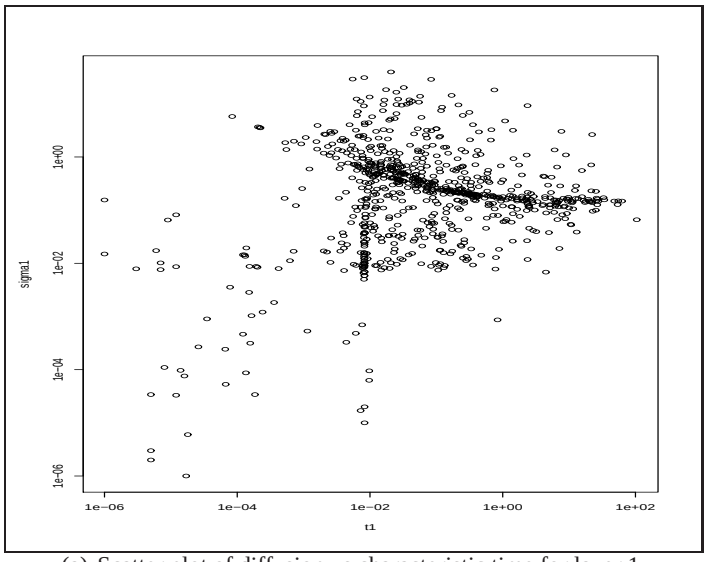
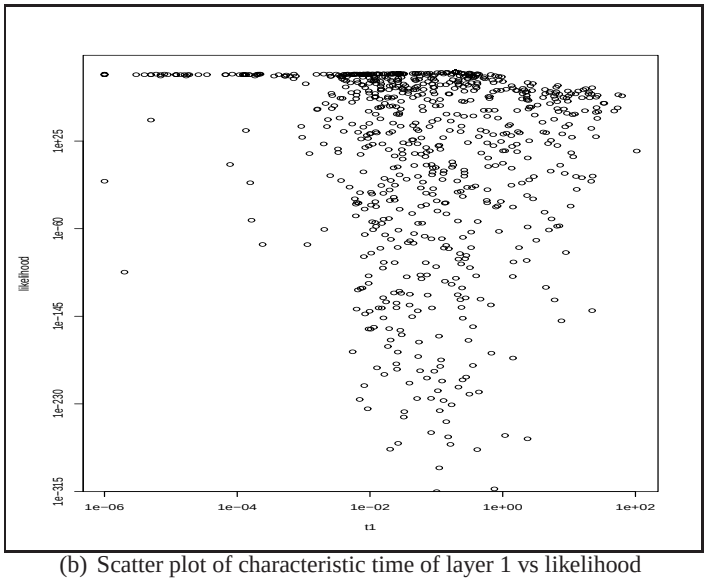



FIG 2. *Sample means for log size (a) and residuals (measurement minus sample mean) for all individual coccolith measurements belonging to a sample of 10 measurements or more, standardized with regards to the sample standard deviation (b).*




(a) Scatter plot of diffusion vs characteristic time for layer 1




(b) Scatter plot of characteristic time of layer 1 vs likelihood

FIG 3. *ML optimization outcomes for the original model, using a shotgun hill-climbing approach.*

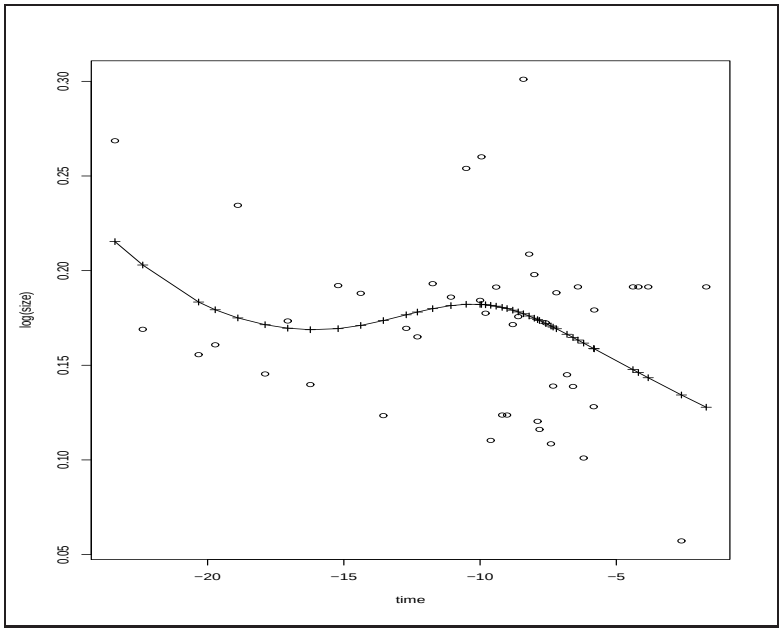



FIG 4. *Sampling standard deviations for Site 752 as open symbols, with GAM smoothed standard deviations as crosses with lines.*