#review

A Bayesian framework for inferring regional and global change from stratigraphic proxy records (StratMC v1.0)

criteria

Principal criteria	Excellent (1)	Good (2)	Fair (3)	Poor (4)
Scientific significance: Does the manuscript represent a substantial contribution to modelling science within the scope of Geoscientific Model Development (substantial new concepts, ideas, or methods)?		2		
Scientific quality: Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)? Do the models, technical advances, and/or experiments described have the potential to perform calculations leading to significant scientific results?		2		
Scientific reproducibility: To what extent is the modelling science reproducible? Is the description sufficiently complete and precise to allow reproduction of the science by fellow scientists (traceability of results)?	1			
Presentation quality: Are the methods, results, and conclusions presented in a clear, concise, and well-structured way (number and quality of figures/tables, appropriate use of English language)?	1			

General Comments

The authors provide a novel Bayesian approach for simultaneous absolute age modeling and correlation of stratigraphic sections with multiple proxies. This approach is based on the inference of a "common signal" shared by proxies (on a proxy-by-proxy level), which may be biased in individual sections. The model is clearly presented, and the code implementing the model is well-documented and easy enough to run. The authors validate the model with

synthetic examples primarily focusing on d13C in carbonates, for which they explore various scenarios of signal recovery via their model. While I find that the correlation aspect of their model is appropriate for the applications they envision, my main feedback is that the age modeling (upon which correlation is dependent, as they are simultaneously modeled) is ad-hoc and requires a stronger basis in the statistics of the distribution of time in stratigraphy. Similarly, the experiments exploring "temporal noise" require firmer grounding in theory. See subsequent specific comments for more details. I believe that this concern can be addressed with deeper interrogation of the age modeling priors and/or reformulation of the priors to more deliberately incorporate our knowledge of the temporal statistics of stratigraphy. The authors also need to better build the intuition for how the age model priors in conjunction with age constraints affect the output of the Bayesian model. My other comments address more minor issues, including more thorough comparison with the Bayesian model of Lee et al. (2023) and the utilization of reliability diagrams as an additional, more nuanced evaluation of their probabilistic model performance. Overall, with some revisions along the lines elaborated on below, I think this manuscript provides an important contribution as a modeling framework for correlation and age modeling, especially for the types of applications that the authors have highlighted.

Specific Comments

Age modeling and correlation

The authors clearly lay out various difficulties in both age modeling and stratigraphic correlation, and they also provide an intuition for how correlation can help with age modeling by constraining likely synchronous levels within various sections. The authors, however, fail to clearly establish *how* the information that goes into the age modeling propagates to the posterior inference for the common proxy, and how this information is tied to the fairly well-established statistics of the distribution of time in stratigraphy.

Imagine a common proxy signal that is simply a sine wave with a single period, which is perfectly recorded in several sections. This signal can be exactly correlated. Now imagine that the sections only contain precise age constraints at their tops and bottoms. In this case, even though the signals are trivially correlated, the location of the peak and trough of the sine in absolute time will be highly uncertain. With extremely ignorant priors on the age modeling, it's conceivable that the posterior distribution for the common proxy signal in absolute time may even have flat contours that completely encapsulate the amplitude of the sine. Ideally, the priors on the age modeling would ensure that the temporal structure of the posterior appropriately encapsulates the true structure. However, the figures that the authors present demonstrate that this is not the case for their model. For example, in Figure 8, the authors show that the posterior model nicely recovers the overall shape of true common proxy signal. However, the locations of peaks and troughs in absolute time seem (subjectively speaking) too tightly constrained, such that the posterior model significantly deviates from the true model at the locations of most major peaks and troughs. This result seems to be due entirely to the age modeling, for which the

priors appear to be *too* informative. This subjectively described behavior can be quantified with a reliability diagram (see a subsequent comment).
All this discussion brings me to the main point, which is that the authors need to more critically consider the prior age modeling. The authors state that they (line 138) "construct prior age models with the goal of imposing no limits on sedimentation rate between age constraints," and yet in Figure 3c they show the distribution resulting from their prior modeling approach. What is this distribution? Sadler (who the authors cite) demonstrated that the concept of sedimentation rate is only relevant at a timescale of interest, since the Sadler Effect shows that sedimentation rates decrease as a power law with averaging timescale. At a particular timescale of interest (within an order of magnitude or so), Sadler also demonstrated that sedimentation rates follow a log normal distribution. Figure 3c does not appear to be log normal. The ad-hoc approach that the authors have taken with the shift and scale parameters was probably motivated by modeling convenience, but it muddies the waters in terms of incorporating empirical statistical information about sedimentation rates. I recommend that the authors reformulate their priors for the age modeling. Specifically, the authors should explicitly consider how the sedimentation rates they model probabilistically are tied to a timescale (as they must be), which itself might be a random variable in their model. If the authors think that proxy sample spacing may span timescales over multiple orders of magnitude, they should grapple with the implications that makes for prior sedimentation rate modeling throughout a sampled section. I recommend abandoning the ad-hoc approach, which imposes a poorly interpretable prior. Finally, the prior that is ultimately chosen should result in age models that yield *reliable* posterior models (see subsequent point).

> As an aside, I could imagine that this modeling framework might be able to introduce an intermediate hidden variable that captures the well-constrained, correlated component of the common proxy signal, which exists along a coordinate that has a monotonic relationship with both absolute age *and* stratigraphic height in each section. This internal representation separates the correlation from the age modeling problem. The task would then be to evaluate the likelihood of the monotonic maps between age and height for each section, mediated via this hidden variable. These mappings would be informed by the prior assumptions about sedimentation rate as well as potentially any stratigraphic information indicating disconformity, etc.

Quantifying model performance with reliability diagrams

I appreciated the authors' mean signal likelihood metric, which captures the overall performance of their modeling approach in a single number. However, given that they are evaluating the performance of a probabilistic inference with respect to the truth, they should also utilize reliability diagrams, which show how the predicted distribution of values corresponds to the actually observed distribution. Bröcker and Smith (2007) (https://doi.org/10.1175/WAF993.1) provide a useful reference for constructing reliability diagrams with bootstrapped confidence intervals. I suspect the authors will find that the current

formulation of their model underestimates the true signal at both the lower and upper prediction quantiles due to the afore-mentioned over-confidence in the absolute time location of the common proxy signal. That is, I expect that the reliability diagrams for the current modeling approach would have low slopes falling off of the 1:1 line for a perfectly reliable model, in which case hopefully a modified age modeling prior would improve reliability.
In figure 9c, reliability diagrams may also reveal another slightly troubling result. The authors nicely show how the incorporation of multiple proxies significantly increases the synchronicity of the posterior with each true common proxy. However, the confidence intervals, especially at the low and high tails, do not seem to sufficiently collapse to reflect the improved modeling. Why might the model be overestimating uncertainty in the tails for inferences with more proxy systems?

Comparison with Lee et al. (2023)

While the authors do mention Lee at al. (2023) in the introduction, I think more can be done to compare the two modeling approaches. To my knowledge, Lee et al. (2023) provide the sole other Bayesian approach to simultaneous age modeling and (single) proxy correlation. Given that the authors are presenting exactly the same sort of model, they should be more explicit in acknowledging the similarities between the models and then highlighting the contributions they have made to this type of modeling, namely:

1. prior age modeling that is not strictly tied to assumptions about deep sea sedimentation rates (although as previously mentioned, this approach needs to be better grounded in theory) and
2. multi proxy correlation (which is mentioned in section 4.2, but with insufficient context)

Section 4.2 needs to reference Lee et al. (2023), and there could be a couple more sentences highlighting the similarities between the models either in the introduction or methodology. For instance, Lee et al. (2023) also permit inference of section-by-section offsets and variance scaling with respect to the inferred common proxy signal (albeit for a single proxy).
I think the authors should consider applying their model to the same d18O and radiocarbon dataset modeled by Lee et al. (2023) (perhaps just the Deep North Atlantic dataset, for example) as an application with real-world data and an opportunity for direct inter-model comparison. Several if not all of the cores utilized by Lee et al. (2023) in that stack have other proxy measurements (d13C, elemental concentrations, etc.) that could be utilized by the authors' new approach.

"Non-uniform" depositional histories

The treatment of "episodic" sedimentation could also be better grounded in the theory of time's distribution within stratigraphy. The Sadler Effect arises due to the power law distribution of hiatus within stratigraphy; the distribution of hiatus therefore dictates apparent sedimentation

rates. Hiatuses result from the dynamics (autogenics) of sedimentation as well as processes such as sea level, tectonics, etc. A critical timescale is the compensation timescale: below this timescale, stratigraphy is incomplete, which approximately corresponds to the "episodic" realm described by the authors. Beyond this timescale, stratigraphy is complete (the "continuous" regime), up until hiatuses resulting from longer timescale processes such as tectonic modifications of basin accommodation. What the authors refer to as "temporal noise" and "episodic sedimentation" are in fact the statistical structure of hiatus in stratigraphy, which results in stratigraphic incompleteness at short and long timescales.

This background brings me to the main point of this comment, which is that the authors need to take care that they are realistically modeling stratigraphy as best as we understand it when constructing their synthetic examples. For instance, by modeling the elapsed time between approximately evenly-spaced (in space) samples as a gamma distribution, do the resulting height increments ('the devil's staircase' shown in the right panel of Figure 6b) have a truncated/exponentially tempered power law distribution for small values of k, as we expect (Ganti et al. 2011)? How does the truncation of the power law (i.e., the compensation timescale) depend on the value of k? The authors should establish the theoretical connections between their current stratigraphic synthesis protocol and the relevant quantities in our current understanding of time's distribution in stratigraphy (such as the compensation timescale, the power law distribution of hiatus in stratigraphy, which truncates at intermediate (post compensational) timescales). Paola et al. (2019) provide a great review of these concepts (https://doi.org/10.1146/annurev-earth-082517-010129).

Alternatively, the authors could reformulate how they generate their synthetic stratigraphies. For example, rather than sampling time increments according to a gamma distribution, which appears to be an arbitrary decision unmotivated by theory, they could instead simulate stratigraphies with varying compensation timescales and stratigraphic completeness (i.e., hiatus power law exponents), which are then sampled regularly (or not) in space. The results of Section 3.3.2 would then be much easier to interpret with respect to the theoretical framework that exists for stratigraphy; perhaps the authors could modify Figure 13 to reflect various values of the hiatus power law exponent (completeness) and compensation timescale.

Technical Comments

- In Equation 4, the notation seems imprecise. Is it not in fact the evaluation of the posterior over θ at the true proxy value? $P_{\theta_{f(t_n)}}(g(t_n))$.
- Fig 1
  - would be nice to annotate the relevant parts with the notation introduced in Equation 1
  - empty box under model seems to serve no purpose
- fig 5
  - would be easier to compare panels a and b if the axes in panel a were flipped

- line 363: might be worth clarifying that white noise is independent, identically distributed, zero mean
- line 484-485: This sentence minimizes the importance of the age modeling procedure for the construction of the common proxy signal...it's not really like the age modeling is a byproduct of the authors' model. It's an integral part of the inference.
- line 567, 596: siliclastic -> siliciclastic