Dear Melanie Hopkins, Dear PCI Paleontology Managing Board

Thank you for agreeing to be the recommender of our manuscript, and for reaching out to three reviewers for their comments on it. We appreciated their thoughts on the manuscript, as they reflect the perspectives of different scientific communities on the problem we address in this manuscript.

Below we provide our responses to the comments made by the reviewers. They are structured as follows: general responses to broader points made by multiple reviewers, and specific responses to comments made by individual reviewers. All line numbers mentioned refer to the manuscript deposited on bioRxiv on the 24th March 2024 (https://www.biorxiv.org/content/10.1101/2023.12.18.572098v2).

General Responses

Test performance

Multiple reviewers were surprised by the poor performance of the model selection procedure used. This was attributed to different causes (e.g. the continuous time expansion, or our implementation of the mode of evolution). Melanie Hopkins wrote R code to examine this effect. This code uses the simulation procedures provided by the paleoTS package to simulate the trait evolution, and performs model selection on the resulting time series. Adjusting the code to match the parameters used in our simulation study, we found that the qualitative behavior of the AICc weights persists. This shows that the poor test performance is not an issue with our implementations, but the methodology of the paleoTS package (code available under https://doi.org/10.5281/zenodo.10843692 or https://github.com/MindTheGap-ERC/paleoTS_test). We did not include this comparison into the revised version of our manuscript for two reasons: (1) As we argue below, our implementations of trait evolution are identical in all important aspects to those in the paleoTS package (2) the paleoTS performance should have been tested independently, and we do not think it is our role to perform a comprehensive performance test in this study. However, the revised manuscript now refers to the parameter study using paleoTS internal simulation procedures (line 654).

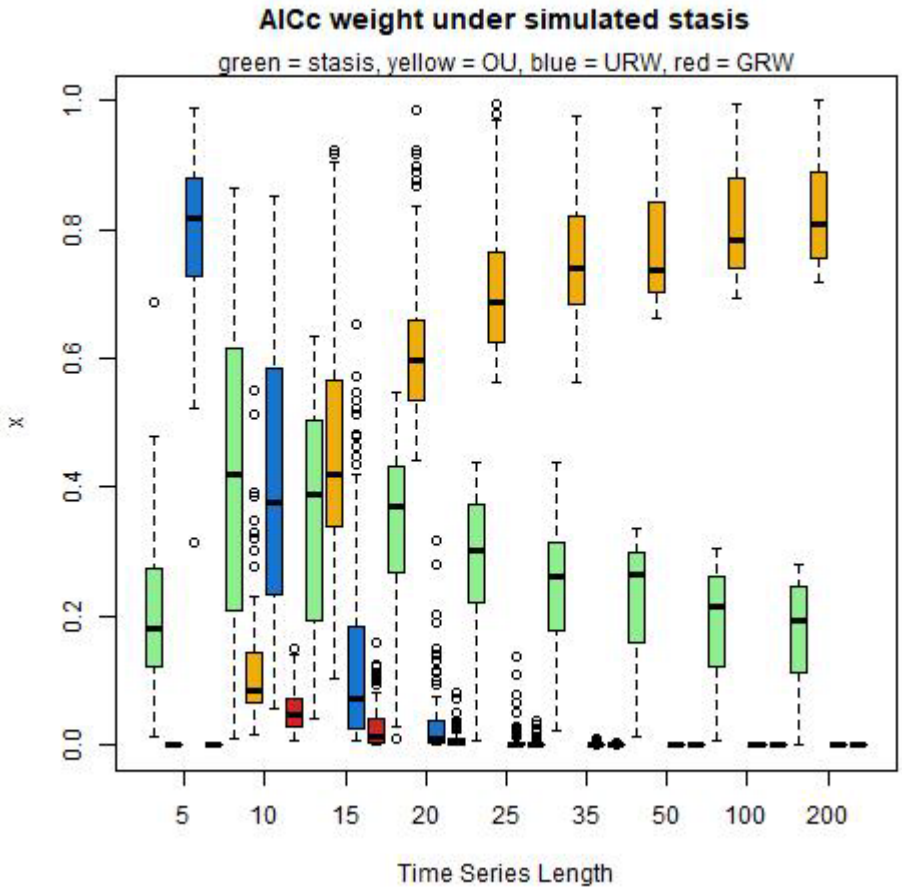

Figure 1: AICc weights of models under evolutionary stasis simulated by the paleoTS package.

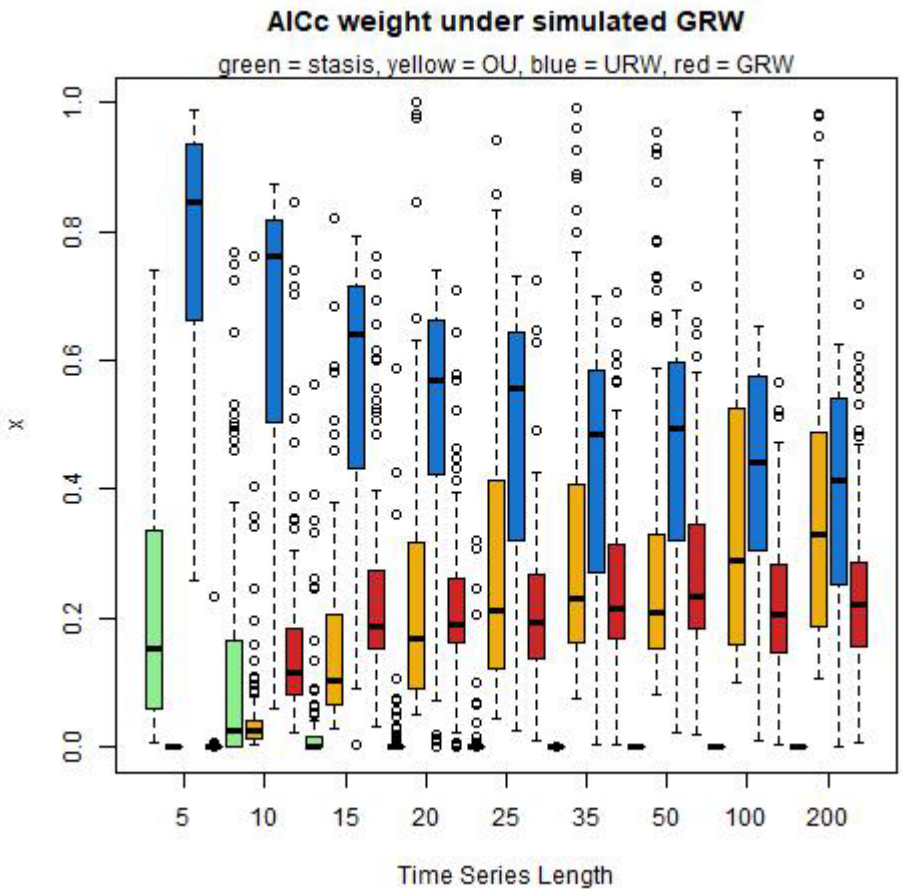

Figure 2: AICc weights of models under the GRW model with mean 0 (URW model) simulated by the paleoTS package.

Continuous time expansion

Both Melanie Hopkins and Bjarte Hannisdal pointed out that our choice of a continuous-time expansion of traditional models of phenotypic evolution is unusual, and might be the cause of the observed poor performance of the model selection. The justification for the choice of a continuous time extension was expanded in the manuscript. We discuss this point for the Brownian motion (BM) and Brownian drift (BD) – it is not relevant for stasis, which is simulated as independent, identically distributed trait values, and is independent of any choice of time step.

It is correct that the time step of the CarboCAT simulation is 1 kyr, and simulation evolution on this timescale would be an intuitive choice. We did opt not to do this for the following reasons:

1. We have a prescribed sampling strategy (1 sample per m) in the stratigraphic domain to emulate a best-case scenario for paleontological sampling. Because sediment accumulation is highly irregular, the difference in time between two sample varies by three orders of magnitude as is mentioned in the text. As a result, the evolutionary history is in general not sampled at the same temporal resolution or time increments of the basin simulation. This could be resolved by simulating evolution at the timescale of the carbonate platform model, and interpolate trait values between the tie points in time. However, this generates dependencies between successive trait values when they fall within the same time bin, and makes trait values a combination of values at both tie points. This dependency would contradict the assumption of the random walk model, which is that incremental changes of traits are independent of each other.

2. We compare the effect of time series length while keeping the timespan of observation constant. Subdividing this interval into time intervals of equal size could be solved via interpolation, which

would introduce the same methodological problems mentioned above.
3. Random walk models based on discrete time steps have a non-obvious scaling behavior when the time steps are altered, making their results harder to generalize across timescales.

In total, we think that using continuous-time expansions of models of trait evolution are methodologically superior: They are more consistent with our study design, which heavily relies on irregular sampling, reduce statistical artifacts due to interpolation, and reduce scaling problems when comparing results with other studies. In the section below we show that for equidistant time steps, the continuous-time expansion matches the discrete time implementation.

Differences between simulations and tests

The reviewers have raised the concern that some of the poor test performance is due to a discrepancy between our (time-continuous) implementation of the modes of evolution and the model implemented in paleoTS.

Comparing our implementation of BD and BM (supplementary code, code/utils.R, function "myBD") with the implementation in the paleoTS package (source code is available by executing "sim.GRW" in the console after loading the package), we find the following:
1. Both implementations generate mean trait values by summing up normally distributed trait increments. In our implementation, the standard deviation increments are dependent of the time difference between successive sampling times according to the defining properties of the Wiener process (https://en.wikipedia.org/wiki/Wiener_process) and the derived Brownian drift model. This is not the case for the paleoTS implementation, as it is based on fixed time steps and the effect of time step length is implicitly contained in the variance parameters. When our implementation is applied to equidistant time steps, the simulation of trait values is identical to the approach used in paleoTS.
2. After mean trait values are simulated, paleoTS adds a normally distributed random variable with mean 0 and standard deviation sqrt(vp/nn) to them to incorporate the effects of a finite number of specimens on the observed mean trait values. Here, vp is the variance, and nn is the sample size. This step is not incorporated in our simulation study, as we assumed that mean trait values are purely based on the underlying simulated mode of evolution. This means that for equidistant time series, the implementations differ by a normally distributed random variable with mean zero and a standard deviation of sqrt(0.1/100) = 0.0316, resulting in a mean absolute difference of 0.025. This is two to three orders of magnitude smaller than the difference in traits observable at the examined timespan, and is not biasing trait values into a specific direction.

The differences between the stasis implementations are similar, they differ by a normally distributed random variable with mean 0 and standard deviation of 0.0316.

With these small differences, we believe our implementations if the models match the paleoTS implementation in all important aspects, and their difference is not sufficient to explain the poor test performance. This is demonstrated by the fact that, qualitatively, test performance does not change when trait evolution is simulated with the procedures provided by the paleoTS package.

Ornstein Uhlenbeck Process

Bjarte Hannisdal and the anonymous reviewer both had question about our choice not to simulate Ornstein-Uhlenbeck processes, but still include them into our test case.

Ornstein-Uhlenbeck processes are a class of mean-reverting Gaussian processes. Their sample paths are solutions to a stochastic differential equation, which is commonly solved using the Euler-Maruyama method (https://en.wikipedia.org/wiki/Euler%E2%80%93Maruyama_method) using

equidistant time steps. Based on the choices to use time-continuous models to get exact trait values at specific points in time, this method was not suitable for our study design. Ornstein-Uhlenbeck processes can also be simulated using the fact that they are Gaussian processes by drawing from a high-dimensional Normal distribution. The dimension of this distribution is equal to one less than the time series length. Drawing from a 200 dimensional normal distribution is computationally challenging, and we decided not to include it.

We decided to include OU into the set of modes we test for, as each tested mode corresponds to distinct evolutionary dynamics – evolutionary stasis, random walk, drift, and evolution towards an adaptive optimum. Each of these four has very different biological meaning, and we wanted to see whether stratigraphic effects favor the recognition of specific modes of evolution, whether they are the "true" mode of evolution in the time domain or not. Not including OU into the modes tested for would have narrowed down the biological implications of our study. This decision is now stated explicitly in the text.

Usage and interpretation of AIC

Multiple reviewers have raised concerns regarding the usage of AIC vs. AICc and the usage of a stringent criterion for what qualifies as an identified model.

Regarding AIC vs. AICc: As was pointed out by Melanie Hopkins, the paleoTS package uses AICc and not AIC (as previously stated in the manuscript). This was a mistake and is now corrected.

Both the anonymous reviewer, Bjarte Hannisdal, and Melanie Hopkins have pointed out that our AICc threshold of 0.9 is high. We define a mode of evolution as correctly identified if its AICc weight is larger than 0.9. We chose this high value in our initial study because sampling conditions in our study design are exceptional – the majority of time series was longer than empirical fossil time series, the number of specimens per time point was high (100 specimens), and intrapopulation variance was small compared to the overall change in trait values observed (1 to 2 orders of magnitude smaller). We think this optimistic sampling scenario justifies using a rigorous threshold for AICc weights.In addition, AICc values below 0.9 are documented in the main text as well, so they can be evaluated by anyone interested in applying a lower threshold.

We do nowhere argue that if that 0.9 threshold is not met, there are no supported models as phrased by Bjarte Hannisdal. We would like to repeat the points made in the discussion – we think the discussion about the AICc cutoff threshold used should not be the focus, but rather the qualitatively unintuitive behavior of the AICc weights.
In the revised manuscript, more emphasis was put on the raw AICc values, which can be directly interpreted as support for a model (given the data and the other models) rather than the cut-off value of 0.9. Where we discuss the failure of the test to identify the correct mode of evolution, the cut-off value is highlighted more.

Response to specific comments

Review by Bjarte Hannisdal

Comments: *"When simulating the three canonical 'modes' of trait evolution (stasis, unbiased random walk, and biased random walk), the authors […] use continuous models (Brownian motion/drift, in their terminology; aka Wiener process) that can be sampled at arbitrary points in time. That's fine, but it seems like an unnecessary complication"*
Response: See the section above on continuous expansion on a justification why we chose this

Comments: *"The authors then state that the reason why they excluded Ornstein-Uhlenbeck models from their simulations is because they couldn't generate samples unequally spaced in time."*
Response: See section on Ornstein-Uhlenbeck process under "General responses" above.

Comment: "It is unclear why the the DAIME package was used. […] The use of some formal notation in this paragraph (e.g. a morphism) hints at a theoretical framework underpinning this software"
Response: The theoretical framework of the DAIME package is described in the supplementary material in Hohmann (2021). This package was used to enable us to sample the depth & time domains at arbitrary points. This was clarified in the text.

Comment: *"The authors simulate somewhat extreme versions of the different modes of trait evolution. […]. By restricting within-sample trait variance and setting a fixed sample size, the authors minimize the effects of sampling error on the trait mean, which would render directional patterns more random, and random patterns more static. This effect would shift the distribution of observed evolutionary modes towards stasis"*
Response: This is correct. We aimed to isolate the effects of the heterogeneous distribution of time in the stratigraphic record on the recognition of the mode of evolution in this study, which is why we assumed an optimistic sampling strategy. We agree with the reviewer, but this is intentional, to make the study easier to follow.

Comment: *"In their description of this time domain analysis (p. 18, lines 366-374) it is not clear how this time domain sampling is performed, and I initially thought the sampled points were evenly distributed in time."*
Response: This is correct. As mentioned, sampling was equidistant in time. This was clarified in the text (line 394).

Comment: *"I may misremember, but my understanding was that if the AIC weight is >0.9, then one is justified in identifying a single best model, and otherwise one should present and discuss the relative support for multiple models".*
Response: See section on AICc under "General Responses" above

Comment: *"Arguably, the most striking finding is that the results for the stratigraphic domain analyses and the time domain analyses are so similar, which could mean two things: (1) The time domain data are also stratigraphically distorted to some extent because the temporal sampling is so highly irregular, which implies that their analysis is not well designed to test for the effect of stratigraphic biases per se. (2) The paleoTS analysis is actually very robust to the simulated stratigraphic distortions, which would be in sharp contrast to the authors' conclusions."*
Response: See the comment on the time domain analysis. Sampling in the time domain is equidistant, so there is no stratigraphic distortion whatsoever. Time domain analysis was performed to establish a baseline for the paleoTS test performance. Our results for this case match those from simulating lineages with the paleoTS package (see general comments), indicating there is indeed a problem with model selection in paleoTS under best-case conditions.

Comment: *"paleoTS analysis clearly favors the wrong model (OU). […] the authors may want to consult the literature on phylogenetic comparative methods that discuss issues surrounding bias towards OU models in AIC model selection (https://doi.org/10.1111/2041-210X.12285)."*
Response: Thank you for pointing this out, this was added to the manuscript

Comment: *"The sensitivity of the time domain analysis to time series length (Fig. 10) is counterintuitive. […] ? I wonder if this might be an expression of the kind of stratigraphic distortion the authors are seeking to investigate"*

Response: We agree that this is counterintuitive. As mentioned above, there are no stratigraphic distortions in the time domain. Even if all model assumptions are met, and traits are simulated with the paleoTS package, results remain similar (see general comments), which makes them even more puzzling.

Regarding additional comments: We are familiar with your previous work on this topic, and it served as a motivation for this study. Thank you for sharing the dissertation, this is an invaluable resource.

Review by anonymous reviewer 1

Main comment

Comment: *"It seems a bit worrying that the correct (simulated) evolutionary model is not recovered under excellent sample conditions in the absence of stratigraphic biases. […]"*.
Response: We agree, it is worrying that a well-established method performs poorly under perfect conditions. In the above section, we have addressed all concerns regarding the used methodologies, and found identical effects for simulations performed with the paleoTS package (code available under https://doi.org/10.5281/zenodo.10843692 or https://github.com/MindTheGap-ERC/paleoTS_test), demonstrating that it is not an implementation issue. This fact was highlighted in the text (line 654). As pointed out, we do not know how exactly this unexpected behavior originates, and we provide a list of potential causes in the discussion. We think it is important that the community is made aware of these unexpected results, and further investigations into the reliability of the methods should follow.

Minor comments

Comment: *"Figure 1: This figure is important as it describes the study design. I would have appreciated a more detailed figure caption to make it easier to understand the different steps in the study."*
Response: caption was adjusted.

Comment: *"Why include a sample variance? Including this will introduce noise into the data, and this is not one of the aspects under investigation."*
Response: Sample variance was included to match the time series format required by the paleoTS package, which was used for model selection. It was specifically chosen to be small compared to the mean change in traits so it does not obscure evolutionary trends (line 359).

Comment: *"The rationale for including the OU model as one of the candidate models when investigating relative model fit is unclear. […]"*
Response: See section on Ornstein-Uhlenbeck processes under "General Responses".

Comment: *"Wouldn't it make more sense to assess how a model of punctuated evolution performs?"*
Response: The punctuated evolution model is mentioned in the discussion, and this section was expanded in the revised version of the manuscript. We decided not to test for punctuated modes because of the poor test performance without stratigraphic biases: Complex modes of evolution are fit by combining the standard modes of evolution (stasis, random walk, etc. – the ones used in our study) with a break point (Hunt et al. 2015, https://doi.org/10.1073/pnas.1403662111). Our results show that the performance of model fitting without a breakpoint and for adequate models is poor. We decided that adding another step of methodological complexity would not add anything to the present study.

Comment: *"Using a criterion of 0.9 for AIC weights is quite stringent"*
Response: See section on AIC under "General Responses" above.

Comment: *"Is there a specific reason why you have favored AIC over AICc?"*
Response: See section on AIC under "General Responses" above.

Review by Katharine Loughney

Comment: *"I suggest finding an alternative phrase to refer to the control of stratigraphy on the fossil record, as "bias" has negative connotations that may serve to justify the perceived shortcomings of paleontological investigations."*
Response: This was adjusted in the revised version

Comment: *"I take it that an assumption of the model is that all lineages are assumed to have an equal chance of being sampled across facies (or environments, as they are represented in the model). The authors should clarify this point in the Methods […]"*
Response: Added to methods section (line 352).

Comment: *"The model constructs lineage patterns based on sampling one synthetic column at a time. I am curious whether the reconstruction of the evolutionary modes improves from tracking lineages from composite columns, similar to how graphical correlation integrates stratigraphic or biostratigraphic data from multiple locations"*
Response: We agree that this is an interesting question, but beyond the scope of this manuscript. We aim to explore this further of follow-up studies, but the current study design is already stratigraphically complex enough given that the target audience are evolutionary biologists.

Comment: *"The majority of the Results and Discussion focuses on general trends from scenarios A and B, and the figures almost exclusively show output from scenario A. […] I also think there is a missed opportunity to not only emphasize the relevance of the model findings to the real-world record, but also to say something about reconstructions of trait evolution from the real record"*
Response: Scenario A was chosen because the effects discussed (spatial heterogeneity, spatial completeness, differential effects) are most clearly displayed within them. We kept figures of scenario B in the supplementary materials to keep the main message clear and concise, and not overwhelm the reader with too many figures. In the revised manuscript, more focus was now put on the implications of scenario B in discussion and results.

Comment: *"Because the SL curves in each scenario impart different frequencies and durations of hiatuses, I think it is worth adding more emphasis of the importance of this to the real-world record. When the importance of hiatus frequency and duration is discussed in the Results (section beginning on line 485), the differences between scenarios are hinted at but not explicitly stated. If the real-world SL curve imparts many short hiatuses (and a more continuous age-depth model), then the potential to measure real modes of trait evolution is perhaps not as bad as we tend to fear because the record is "incomplete.""*
Response: The difference is now more explicitly stated. This was moved to earlier in the text.

Tables and figures:

Comment: *"In Figure 2, it would be helpful to have a color key for the different facies depicted in the simulated shelves."*
Response: This was adjusted for figure 2 and 3.

Comment: *"It is unclear why a column from 2 km in scenario A is compared to a column from 6 km scenario B."*
Response: This was clarified

Comment: *"Figures 8 and 9, 10: Captions need explanation of the abbreviated tested modes in the legends."*
Response: This was adjusted

Specific comments:

Comment: *"Line 19: the clause […]"*
Response: Thank you for pointing this out. This was corrected.

Comment: *"What is meant by "adequate model"?"*
Response: In contrast to many technical statistical terms, there is no general definition of model adequacy available. Even the cited references on model adequacy (Voje 2018, https://doi.org/10.1111/2041-210X.13083, Voje et al 2018, https://doi.org/10.1086/696265) do not provide a rigorous mathematical definition, but rather define different tests that examine "adequacy". Here we use adequacy in the sense that the model is an appropriate description of the data, which is the case if the model was used to simulate the data (as we did). We added a mention of this in the introduction.

Comment: *"Line 71: This phrasing makes it sound like sedimentology and stratigraphy are the only disciplines that are jargon laden."*
Response: This was corrected, the focus is now on the utility of these disciplines for evolutionary biology.

Comments regarding line 94, 203, 289, 295, 691, 716, 723: Thank you for pointing this out, this was adjusted.

Comment 658-660: *"Yes, it is important to examine multiple columns along dip for interpreting the patterns. The authors may want to acknowledge here that this is a bit of an oversimplification that may not directly apply to carbonate platforms."*
Response: This was added.

Comment: *"Line 778: I'm confused by the use of "ground truthing" here or by the wording of this sentence. Are the authors suggesting that the models offer ground truth? Ground truth can only be gotten in the field."*
Response: In this sentence, ground-truthing specifically refers to the methods used: As written "use of stratigraphic forward modeling to *ground-truth the methodologies* serving this palaeobiological research program." Their performance needs to be demonstrated in a simulation setting before they should be applied to empirical data. This is clearly shown by our study design, where tests failed when all model assumptions are met, which made it impossible to test one of our hypotheses.

Comments: *"Lines 783–784: see also strataR in Holland (2022)"*
Response: Thanks for mentioning this, was included.