Community Comment

I very much enjoyed reading the paper by Lin et al. This is an extremely impressive and thorough piece of work. The associated Python framework is extremely well done, and I was very impressed to see so many options given for how the models are fitted and plotted. My only real concern about this work is the functionality for those who are not experts in Python (or Pyro) as the vast majority of users would be. There are very thorough notebooks in the tutorial section of the Github repo but these are not really helpful for those who want to do a quick straightforward model. My guess is that most users want to fit a GP model using the default values for time uncertainty, error variance, kernel choice, etc, and would like a simple guide for how to get their data from the Excel spreadsheet through the PalaeoSTEHM pipeline. Going even further I would strongly encourage the authors to create a proper Python package to simplify the instructions and coding.

We thank the reviewer for recognizing the importance and thoroughness of this study, as well as for acknowledging the effort invested in developing the PaleoSTeHM framework. We agree that simplifying the modeling process for non-expert users is crucial for broader accessibility. In response to this insightful feedback, we have now included a new PaleoSTeHM UI section in our Github page (https://github.com/radical-collaboration/PaleoSTeHM/blob/main/PaleoSTeHM_UI/Holocene_Spatiotemporal_analysis/Holocene_SP_anlysis.ipynb) that allows users to automatically conduct temporal and spatiotemporal GP implementation, optimization, and plotting with minimal input. This feature is designed to streamline the process, enabling users to fit GP models using default implementations of model structure and to easily transition data from common formats such as Excel into the PaleoSTeHM pipeline (https://github.com/radical-collaboration/PaleoSTeHM/blob/main/PaleoSTeHM_UI/Holocene_Spatiotemporal_analysis/Holocene_SP_anlysis.ipynb). We hope this addition will greatly enhance usability for all users. Now, PaleoSTeHM is available as a PyPI package (https://pypi.org/project/PaleoSTeHM/).

One notable thing I couldn't see in the notebooks I ran (or in the paper) was convergence checking for the model. I would say this is absolutely vital for having any faith in the results. Models of this complexity can be extremely difficult to obtain convergence on, and there is a whole range of summary stats available for this in Pyro. It should be part of the default workflow everywhere. Perhaps it is for some scripts and I've missed it, but it certainly isn't discussed in the paper. On a similar vein I'd like to know if the model is calibrated, via scoring rules or even just some posterior predictive distributions (though these can be tricky with bivariate uncertainties), and some kind of out-of-sample performance metrics as would be common in standard ML pipelines.

We thank the reviewer for their valuable feedback. In response, we have included a dedicated Section 3.6 for model validation, which comprehensively addresses convergence checking and validation processes. Additionally, for each illustrative model presented, we now provide detailed validation metrics, including residual plot checks, prior and posterior predictive checks, cross-validation, MCMC convergence diagnostics, and optimization trace plots. These additions ensure a thorough evaluation of model reliability and performance, addressing the concerns raised.

Otherwise I really enjoyed the paper and I'm super excited to see how this develops. There were a number of poor sentences which I've highlighted below but I don't think I've got all of them. It just needs another language check.

We thank the reviewer for pointing out some language problems, which significantly improve the readability of this paper. The points below are now modified as suggested.

L32: Change-points

We thank to review to raise this point, but we suggest to keep the current term to follow the notation in previous paleoenvironmental paper like Ashe et al., 2019 and Caesar et al., 2021. We've modified change "change-point" to "change-point models" to keep overall consistency.

L35: I'd put the reference to the GitHub repo here so people can start coding without needing to read the whole paper.

Modified as suggested.

Table 1: I'd just review some of these definitions. The conditional probability one about conditioning on an unknown quantity doesn't read quite right. I also think you: should include one for parameter itself; adjust the line spacing or add horizontal lines to separate the entries better; re-write the likelihood one; and change the uncertainty one which seems to be a frequentist definition.

Table 1 revised according to feedbacks.

Fig 1: External

Fixed.

Fig 1 caption: Platforms

Adjusted.

L174: I assume the number of change points is fixed and not learnt?

Correct, a learning change-points functionality is not include in this paper, while we discussed the potential development of this method (i.e., transdimensional method in the discussion section).

L176: delete 'and'

Deleted.

L187: Mu(t) or mu(X) (as used in Eq 10)?

Updated.

L201: will be shown in Section 2
Refined.

Table 2: I got confused by what the sampling covariance is and how it is calculated for deterministic models. Please expand in the text

More information is now included.

Eq 15 (and perhaps others). The usual way to present normal distributions is mean and variance, not sd.

Revised.

Fig 5 (bottom right) and others. It always bugs me slightly that the uncertainty in the rate for the present is more unknown when it is the period when we have the most data. Is there a way to solve this with these models? It strikes me that we should be using temporally non-stationary models that allow for far reduced variance (and hence variance on the derivative) to capture the rate of the most recent periods.

We appreciate the reviewer's insightful observation regarding the challenge of capturing reduced uncertainty for recent periods where more data are available. With insights from previous studies (e.g., Fig. 1 of Kopp et al. 2016), we note that this issue can be addressed without the need to introduce temporal non-stationarity. Instead, the solution lies in allowing higher-frequency temporal variability in the modeling process. This approach leverages the dense data available for recent periods to resolve such variability, while in sparse-data regions, this variability effectively manifests as additional white noise.