Paper Review: PaleoSTeHM v1.0-rc: a modern, scalable spatio-temporal hierarchical modeling framework for paleo-environmental data

November 2024

Summary of content

The authors present a novel spatio-temporal hierarchical modeling framework designed for examining paleo-environmental data. It provides an in-depth discussion of the underlying architecture of the PaleoSTeHM software and showcases its capabilities through several case studies focused on paleo sea-level data.

Comment to the Author

This paper showcases the PaleoSTeHM software, which represents a significant and valuable contribution to the field. The integration of machine learning techniques with a variety of Bayesian inference methods marks a notable advancement. However, the paper's structure, terminology and layout would benefit from further refinement. It assumes considerable prior knowledge, which may pose challenges for readers. I have outlined several questions and observations regarding the explanations, along with substantial content-related feedback (refer to Main and Minor comments).

Main Comments

- The target audience for this paper is somewhat unclear. While the introductory section provides an overview of statistical methods, the machine learning component and the software structure is not explained in sufficient detail. Additionally, the introduction feels incomplete and the paper's layout inconsistently transitions between examples, such as those involving paleo-sea level data. Furthermore, the paper assumes prior familiarity with Ashe et al. (2019), particularly regarding the authors' conventions for distinguishing between analytical choices (e.g., Variational Bayes) and modeling choices(e.g. GP). This assumption could make it challenging for readers unfamiliar with the referenced work to fully understand these distinctions. I would recommend keeping consistency with terminology throughout the paper. I would ensure the paper focuses on the capabilities of the software.

- The paper lacks a dedicated section on model validation, and I could not find any methods addressing this in the GitHub tutorial. How can users assess and ensure model convergence? Were prior posterior predictive checks performed or was a simulation analysis conducted to evaluate the techniques? Additionally, were cross-validation methods employed, with residual analysis and true versus predicted plots? It is also unclear how users should compare models—for example, using metrics like RMSE, MAE or empirical convergence. I strongly recommend including a comprehensive section on model validation in the paper, along with links to resources or tutorials demonstrating how these assessments can be performed using the software.

- The literature review on statistical models lacks references that directly link to the introductory equations, which undermines the connection between the theoretical framework and existing research. Additionally, the paper provides extensive details on a wide range of topics, from physical models to various proxy data sources. Please ensure that appropriate references are included for all these topics. For instance, I came across a recent paper by Upton et al. (2024) that used Bayesian spatio-temporal generalized additive models and an R package available on Cran called reslr, which appear to adopt a similar approach to modeling sea-level changes. A comparison of the methods employed in the current work and those in reslr and similar approaches would be particularly valuable for readers.

Minor Comments

- Title: "for paleo-environmental data". The paper primarily focuses on paleo-sea level data, and the tutorials associated with the software exclusively use this dataset. It would be helpful if the authors included an example in the GitHub repository demonstrating how the methodology could be adapted for other types of proxy data, such as temperature. Alternatively, if the scope of the paper is limited to sea-level data, the title might be revised to reflect this focus more accurately. If I have overlooked another example where an alternative climate proxy is examined, it would be beneficial for the authors to highlight it more clearly.
- Update ", and " through out the paper to remove the comma before the "and".
- Abstract:
  - "Geological records of past environmental change provide crucial information for assessing long-term climate variability, non-stationarity, and nonlinearities" in climate? This sentences seems unfinished.
  - "This framework enables the implementation of flexible statistical models that rigorously quantify spatial and temporal variability from geological data with clear distinguishing between measurement and inferential uncertainty from process variability".. An improvement: "clearly distinguishing measurement and inferential". Additionally, some sentences in the paper could be shortened to improve readability and clarity.
- Introduction:
  - Line 11: "As humans push the planet's climate and biosphere increasingly far outside the range of our species' experience, the geological record provides critical out-of-sample data against which to test the models used to project future environmental change". This sentence is misleading as the paper does not address projections.
  - " Yet, as an environmental record, the geological data is quite sparse and often noisy and indirect." Revision: "However, as an environmental record, geological data is sparse, often noisy, and indirect." Also, there is minimal discussion on what the geological data is. Include more information about the data or use a reference to direct readers to e.g. Shennan et al 2015 for paleo-sea level data.
  - Line 16: "an analytical perspective" keep the naming convention consistent. In Table 1 you use analysis choice and other locations interchange these terms.
  - Line 20: Include a definition of a geological proxy and references for examples of "temperature and precipitation".
  - Line 23 and Line 24: the acronym GMSL needs to be defined.
  - Line 27: "(Tan et al., ...)" include "e.g." at the beginning of this list as it is not the complete list of papers.

- Hierarchical statistical Modeling:
  - Providing a definition of hierarchical modeling at the start of this section would help readers understand the topic more effectively.
  - I would begin by explaining Bayesian statistics and then link it to the example you will describe in the next section. Include references to the original mathematical papers for the Bayesian and conditional probability definitions.
  - Line 40: " data (y) can be inverted". I would use a different verb then inverted as this is misleading.
  - Table 1: I recognize that the table is to reflect the table in Ashe et al. 2019, yet, there are a number of relevant terms used extensively in the paper which would be beneficial if they were included in this table, e.g. Variational Bayes, Bayesian statistics, empirical Bayes, full Bayes, machine learning terms. Highlight where these terms are used throughout the paper, i.e. refer to Table 1.
  - Line 53: " A basic hierarchical statistical model for paleo sea level distinguishes the fundamental RSL change from both its inherent variability and the observational noise." I would place this sentence after equation 3 and clearly state that is the your example case. This is an example where the layout of the paper is blending the example with definitions as mentioned in the first Main Comment. Also, RSL acronym has not be defined. Explain to the reader what is relative sea level.
  - Line 61: Missing references for "geochronology techniques", for example Wright et al 2017.
  - Line 67: Include GIA acronym with the definition of glacial isostatic adjustment. Also, Table 1 has been referenced here but does not correspond to any of the terms in the table.

- Line 72 and 73 are very important as they define the authors' conventional terminology i.e. modeling choices and analysis choices. As mentioned in the Main comments section, this is where the authors need to clarify their meaning of analysis choice and model choice before describing how this applies to the new software.
- Line 74 and 75 highlights how prior knowledge has been assumed as mentioned in the first Main comment. The authors need to include brief explanations for deterministic methods and the difference between those methods and probabilistic methods.
- "Several factors, including the complexity of the problem, the size and resolution of the data available, the computational resources at hand, and the extent of prior knowledge applicable to the modeling effort, should guide the selection of modeling and analytical choices." Shorten this sentence to improve readability. This is a key sentence for future software users, again relates to the first Main comment.

- Model Description: Would this benefit with an update of name from model description to Software Architecture? This is the key section of this paper.
  - Section 3.1 needs to be restructured. The authors use L3, L2, and then L1 without defining these modules. I would recommend including lines 95 to 100 earlier in the section for clarity.
  - Line 89 and 90 "L2 employs Python as the user interface language and utilizes a high-performance machine learning platform as the execution back-end". This sentences needs more explanation.
  - Line 96: "including auto-differentiation, GPU acceleration, and modern optimization algorithms." Include definitions and references for each of these topics.
  - Line 100: " multiple methods to consider temporal uncertainty". Include how the uncertainty in the response is included. Also, address how each model fits into Figure 1. Instead of placing Figure 1 in brackets at the end of a sentence, consider rephrasing to begin the sentence with a reference to the figure. For example: "As shown in Figure 1, L3..." This approach integrates the figure more seamlessly into the narrative and emphasizes its relevance to the discussion.
  - Line 104: Should this not be "(L4, Figure 1)"? Or is this section describing how L3 interacts with L4?
  - Figure 1: The caption does not include the term module which has be extensively used in the section above. Include a sentence on how these modules or layers interact. There should be references used for Pyro and PyTorch in the caption. The discussion about L1 in both the caption and the corresponding section is very limited. Consider adding an additional sentence to the text to provide more context or explanation about L1, ensuring its significance is adequately addressed.
  - Section 3.1: "experiment architecture" does this relate to the result section or is this section address how the user should use the software?
  - Line 112: "Training" this term is confusing. Will the user be repeating these steps " sequential selection steps" or is it that they are identifying the best option for their data?
  - Line 113: This relates back to the first Main comment. Instead I recommend that the authors start with: " In Figure 2 we define the 5 steps of the PaleoSTeHM software focusing on L3 from Figure 1. These steps are...
  - Line 116: " These five steps reflect core functionalities developed within three PaleoSTeHM modules, shown in Figure 1." Should this not be " the core functionalities of Layer L3 from Figure 1. This needs clarity.
  - Line 117: "To support the effective selection of modeling and analytical choices provided by PaleoSTeHM for various paleo-environmental applications." This sentence is unfinished.
  - Line 118: "modeling option" update to " modeling choices" for consistency.
  - Figure 2 is a very important figure and the caption needs to discuss how it relates to Figure 1. Again, "experiment architecture" is a confusing term. I would number the 5 steps, instead of using colors and boxes to define the steps as this is not inclusive for all readers. "temporal uncertainty treatment" is not defined anywhere, should this reference the EIV method (Dey et al 2000) and the NI method (and McHutchon and Rasmussen, 2011). Include e.g before "temporally linear and Gaussian Process"
  - Line 122: "commonly used temporal or spatio-temporal modeling choices used". Update this sentence.
  - Line 125: "While we do not include a specific section for parameter-level modeling, leveraging the ecosystem of Pyro and Pytorch enables users to easily define prior probabilities for data and process-level model parameters using most of the commonly used probability distributions". Include references for Pyro and Pytorch. Additionally, improve the readability of this sentence. Does this mean that the software allows users to define priors?

- "e.g., radiocarbon Reimer". should this include "radiocarbon dating, Reimer.."
- Could equation 4 not be discussed in section called hierarchical statistical modeling?
- Line 145 and 146: Explain why strong covariance could be an issue and how adapting the likelihood structure could improve this.
- Lines 150 and 151: The term "likelihood sampling code" has not been defined, which may confuse readers. Please provide a clear explanation of what this means and how it is used in the context of the model. Link this back to the discussion in Line 146. Additionally, Pyro is mentioned without a reference—please include the relevant citation.
- The "Process Level Modeling" section should be restructured as a new subsection, with the subsequent sections organized as subtopics within it, i.e. temporal linear models to GP Kernal module.
- Line 160: "qualitatively assessed". How was this carried out? Was a residual analysis undertaken? Include a reference with this statement.
- Equation 8 and the other equations in this section should be kept general to apply to various paleo-environmental changes. For example, β could represent the rate of change in the paleo-environmental variable. Additionally, the authors should reference Ashe et al 2019 for the upcoming sections.
- Include reference for change point Carlin et al 1992.
- Line 176: define non-parametric and parametric in the table.
- Line 196:" Yet, its justification can sometimes be complex (Stein, 2012)." Expand.
- Line 203 and 204: Should include a reference.
- Line 211 missing key reference to Peltier's multiple GIA models.
- Line 213: "spatial teleconnections" update this term as it is misleading.
- Line 217: " PaleoSTeHM to probabilistically..." expand on this feature. This is very important and a novel component of the software.
- Line 221: given a definition for "sampling covariance functions".
- Line 225: Figure 2 instead of Figure 1?
- Line 226 to 229: The list requires a reference to the original statistical paper where the technique is defined, along with the examples of where it is used in the paleo-sea level field.
- Line 235 Include reference.
- Line 237-238: "temporal and spatial white noise kernels". Need to expand on this and include references to equation 4.
- Line 243: "The spatial correlation is computed for spatial kernels based on the 1-dimensional geographical radial distance between data points." Expand on this.
- Table 2: Please include references to the statistical papers where these equations were defined.
- Analysis Choice and Modeling Choice. It is confusing when the analysis choice which represents least square or Bayesian approaches keeps changing its name. Similarly modeling choice or model characteristic keeps changing terminology.
- Line 249: "deterministic methods (e.g..." This has not been explained previously or in Table 1. Also "implemented in other packages", reference the other packages?
- Line 260-265: "sampling process", "autocorrelation", "effective sample size per iteration and..", "path length" and "step size". All require definitions.
- Line 283: "details below": Include the specific section.
- Line 284-287: Could you provide a more detailed explanation? The current text assumes prior knowledge of many machine learning techniques and terms.
- Line 296: Need to include a reference for variational bayes.
- Equation 12 needs more explanation and Kullback-Leibler divergence needs a reference.
- Line 302-303: "..scales linearly". Is there a reference for this?
- Line 308: "Cahill.." missing reference to "Dey et al, 2000:.

- Results: Consider renaming this section to Case Studies. Additionally, the layout should more closely align with the steps outlined in Figure 2 and Section 3.2.
  - Line 319 include link to codes.

- Line 320: Update the sentences into bullet points of the different test cases. References for the data sources are missing. Also, "analysis choice modeling techniques" blending the two separate modules, update this.
- Line 325: should there be more references include e.g. Walker 2021?
- Line 330: Have a sentence at the start stating what is being discussed, i.e. "in this section we will address the impact of data level". Also, there is minimal discussion on what the input data is.
- Equation 14 and 15: The treatment of the white noise component in both equations is unclear. Is the white noise modeled as a random variable drawn from a specified distribution, or is it treated as a deterministic fixed parameter? This distinction needs to be clarified. Additionally, the notation in Equation 15 is non-standard, as the inclusion of the square root means that the second parameter no longer represents the variance. Please clarify whether the square root is intentional and, if so, provide an explanation for this modeling choice.
- Figure 3: The caption for Figure 3 requires more detail. Please reference the data source used in the figure. Clarify what the additional noise component represents, is it the standard deviation of the white noise or the residual standard deviation of the model? Also, explain why 90% credible intervals are used and why uncertainty boxes are set to 2 σ? Was a 2 *sigma* uncertainty applied in the models, or was it 1 σ? Add labels (a), (b), etc., to each panel for clarity. The legend is missing for the RSL vs. Age plot. For readers unfamiliar with RSL, include a brief explanation of why RSL values are negative.
- Line 347: How does the user determine whether to use a normal or a uniform distribution? The results show a difference between the two, but is there a preferred distribution, or should the choice depend on the specific characteristics of the data?
- Line 355: "3 change points". Why 3 change points? Should you reference the original paper? Also, "RBF" acronym not defined.
- Figure 4: Update RSL with Relative Sea Level. Include legend explaining the red boxes and in caption describe how you model the midpoint of the boxes with 1 σ error. Include reference for the data. Label each row of plots (a). Why was variational Bayes not used for GP in time?
- Line 370: Why not use the same data location throughout the case study in order for the reader to clear see the impacts of the model and analytic choices on the same dataset? "use a subset of original data" is this to speed up the model run times. How many data points were used?
- Line 372-375: Include the run times in this paragraph as stated in Figure 5. Is 2200 posterior samples and 500 iterations sufficient, or does this seem on the lower end as the default is 5000 for other software? Additionally, could the model convergence checks for all approaches be included in the appendix for consistency?
- Figure 5: Include reference for data. Update RSL with relative sea level and include legends explaining the red boxes. Label each row for easier explanations.
- Section 4.2: Spatio-Temporal Analysis: This section requires further refinement to enhance readability and clarity. It lacks sufficient references to the original data collection process and previous model results, which are crucial for context. Additionally, the model definitions assume a level of prior knowledge that may not be accessible to all readers. The discussion of Figure 6 is also lacking, as the individual panels have not been thoroughly addressed. Each panel should be analyzed in detail or with a brief summary, comparing the results from the different model choices after discussing them independently.
- Line 390: This section examines the spatio-temporal model used to examine the different RSL drivers
- Line 392: Include link to the sea-level proxy database.
- Line 395 - 400: Present this paragraph as a table and name the models instead of using (i). With the corresponding equations. Describe the purpose of each of these models.
- Equations 16 - 21: The paper encourages the reader to refer to Ashe et al 2019, however, the notation does not correspond completely. For example, "g(t)" is not described as the "global component" in the text and Line 416 " g(t) as we assume..." expand on this. Also, all the Ks should be defined or reference to Table 2.
- Lines 423-430: The terms "sampling covariance function" and "physical ensemble m" require further explanation and appropriate references. This section needs be improved. For example: "Lin et al. (2023b) described a method for incorporating physics-based GIA models using an ensemble approach. In our software, the ensemble method is implemented as...". Please clarify these concepts and provide the necessary references to improve the readability and context of this section.
- Line 431: Provide an explanation for " weighted mean of different physical models" including a reference. Similarly Line 435 - 440 assume prior knowledge and do not include references, this needs to be altered.

- Line 440 - 454: Update " Figure 6 demonstrates the results from our 5 spatio-temporal process level models defined in Table (3)". I would recommend reviewing the layout of this paragraph to improve readability. For example: "regional common kernal, g(t), in equation 17" has not been clearly defined. Include a recap of what model (i) is examining. Check references, for example Cahill et al., 2015 did not examine data in Florida. Describe each panel of Figure 6 in a specific order, for example "Figure 6i, geological sea-level..", is this (i) representing the model or is it representing a different panel plot that is highlighting a specific process?
- Line 455: "GIA model" missing a reference.
- Line 464: "teleconnections"? What is this referring to?
- Figure 6: Include legend to show what the red box is. Include labels for each row explaining what the panels represent. Update model i - iv with the name of the models used. Caption: "Process level models impact.." instead " The impact of process level modeling choices for..". Include reference to the data. The labels a - l are hard to read. Also we now have model i and panel i. Reference for ICE7G is required. Why examine the year -5500CE? Explain the standard deviation of RSL prediction more clearly.
- Line 471: "is equivalent to a linear combination of physical models according to data-model misfits". Clarify this sentence.
- Line 477- 478: Either include a reference or explain what is meant by " direct constraints on ice history, the ..."

- Discussion:
  - Line 481 - 485: Improve the readability of this sentence. A paragraph should contain a least 3 sentences and these sentences are long.
  - Line 486: "Because of" avoid using because at the start of any sentence.
  - Line 488: " process level models introduced this paper" .. "in this paper".
  - Line 490: "to describe latent some space-time..". Improve this sentence.
  - Lines 495-498: The term "principal component..." requires either references to alternative methods or a clear definition of what these methods are. Please provide the necessary context for clarity.
  - Line 500: Upton 2023 used generalized additive models for RSL changes. This should be included.
  - Line 506-510: "subject to complex likelihoods". This requires a reference.
  - Line 520: " Another outstanding issue for GP based process level models is scalability, the standard GP models included in PaleoSTeHM v1.0 cannot scale well to large data sets (>10 thousands data points)." Has there been test done to examine ¿10 thousands or 10 thousand data points? More common to describe the computational requirement of a Gaussian Process being of O(n^3) where n is the number of data points.
  - Line 530: "and (Lin et al., 2023b) developed" remove the brackets.

- Conclusions:
  - Line 545: "though the limited availability of user-friendly software often hinders it." There are other packages available for the paleo-sea level community which have not been referenced in this paper.
  - Appendix, Table A1: Include more information in the Appendix regarding its relevance to the paper. Why was 90% credible interval used in this paper? Convention is 95%, is there a reason why 90% was used instead? Should parameter level title include "Priors for Parameters". Need to define GBR in caption for table. Should you include reference to where the data is sourced? Also a reference to where the models have been used in the past?
  - Appendix Figure A1: Update "Model ii" with the name of the modeling option and analysis choice which has been used. Improve text in caption "prediction on - 5500CE RSL", for example " Model predictions from Model X for relative sea level at time point -5500CE...". Include more information about how this plot relates to the paper and its relevance. Included a reference to highlight the data source. Linked in the paper on line 455, however it requires more discussion e.g. (as shown in Figure A1). The caption needs more information regarding the axis. RSL should be Relative Sea Level. The reference for the ICE model and VM model should be included in the caption. Could the letters a,b,... be placed outside the plots in black to make it easier to read.

- I commend the author's Github repository which contains many tutorials that demonstrate to the user how to implement the PaleoSTeHM. I recommend reviewing the documentation to improve readability, there is a number of spelling mistakes and some of the sentences are long and misleading. Additionally, the software requires a google drive connection is there another option as some university do not allow Google accounts? The 2 hour tutorial videos are very useful however, is there any possibility to split them into smaller segments in order to be used in future lectures.