1 General Overview

This paper should be viewed as a review and tutorial for the Bayesian analysis of computer models aimed at paleo climate and ice-sheet modellers, although the general framework transcends many scientific disciplines. The authors have a clear aim which is to encourage the paleo modelling community to employ a full Bayesian uncertainty quantification framework in the analysis of their (coupled systems of) computer models, and most importantly making inferences for the real world physical system by jointly incorporating all sources of uncertainty, including structural model discrepancy. Moreover, the paper sets out the importance of ensuring that all aspects of a modeller's statistical framework are transparent, defensible and explicitly outlined within papers; as is good practice for peer-reviewed scientific publication and for obtaining reproducible results. Other positives from this manuscript include: the discussion of the limitations of Multi-Model Ensembles (MMEs) which are prevalent across paleo modelling, as well as in present-day and future modelling; and the call for a high-quality paleo constraint database, including the associated uncertainties, thus greatly decreasing the time spent devising suitable constraints whilst enabling consistency between studies in terms of the data with which model output is compared.

The framework presented within sections 2 and 3 does not constitutes novel methodology, instead forming an amalgamation of the Bayesian analysis of computer models literature, as well as some applications to other scientific disciplines. However, the authors attempt to present this as a complete Bayesian framework for paleo modellers, including numerous practical suggestions for its implementation. The content is therefore of greatest use to those less familiar with such methods, although could also be used as a reference by those more experienced in statistical uncertainty quantification.

2 Major Concerns

Below I outline my major concerns regarding this manuscript.

2.1 Manuscript Length

Firstly the length of the manuscript is excessive and would require a committed and motivated reader to finish the paper. In addition, the abstract, introduction and conclusion do not necessarily provide an adequate summary of the main methods and the consequences of (none) implementation. The authors are ambitious in the breadth of topics from the Bayesian analysis of computer experiments literature that they include resulting in this manuscript having the feel of the notes from a lecture course. For any reader with even a moderate understanding or familiarity of Bayesian statistics, this treatment seems unnecessary. For example, sections 2.1 and 2.2 could be much briefer with the reader referred to an introduction to probability theory undergraduate textbook or online resource where necessary, rather than repeating the material within this manuscript. The discussion of MCMC in section 2.7 is not the main purpose of this work, hence it could also be condensed without detriment to this manuscript by referring the reader to other papers, textbooks and resources where desired.

2.2 Structure and Examples

Throughout the manuscript the authors incorporate numerous examples, both toy models, and from across paleo climate and ice-sheet modelling, but also in relation to present and future modelling efforts. Whilst these provide clarity to the meaning of the statistical framework, in particular, relating this to the analysis of paleo computer models, it also seems like an attempt to conform to the scope of Climate of the Past. In many instances these examples seem unnecessarily long and are often vague in their description without any actual results, for example, lines 399-429, in the discussion of internal model discrepancy. Consequently, these examples greatly add to the length of the manuscript without adding much to the contents. Another concern is the examples interrupt the flow in the presentation of the Bayesian framework that the authors are keen for the reader to adopt. Greater clarity of exposition could be achieved by taking inspiration from Vernon et al. (2018) and restructuring the manuscript, only using short examples where absolutely necessary. Firstly, I would suggest combining sections 2 and 3 to describe the framework including: prior and model specification; emulation (see comments in section 2.3); uncertainty quantification; history matching and the need for simulations to bound reality; and uncertainty quantification in making predictions/retrodictions. This should be followed by a separate section demonstrating how (the majority of) this framework is then applied to a paleo climate and/or ice-sheet model. The paper should finish by briefly highlighting what further steps are required by the paleo modelling community, rather than the extensive section 4 which seems to repeat many of the points found in sections 2 and 3.

2.3 Statistical Emulation

Statistical emulators form a vital tool in the analysis of computationally expensive simulators such as paleo climate and ice-sheet models. The authors refer to emulators at multiple points throughout the main body of the paper, however they are only first discussed in any detail at the end of section 2.7, lines 590-596. Given their importance within the presented Bayesian framework as a fast statistical approximation to the simulator's output(s) for as yet unevaluated paper settings along with a corresponding statement of the uncertainty, I believe that emulators warrant a more detailed exposition within the main body of paper, rather than leaving it to Appendix A6. This should be introduced early on within the manuscript, as discussed in section 2.2.

It is not guaranteed that the reader will know what is an emulator, or more specifically, a Regression Stochastic Process Emulator (RSPE) as it is termed in this manuscript. It is therefore necessary to expand on the discussion currently provided in Appendix A6 with further mathematical details of their formulation, as in Vernon et al. (2010), highlighting the range of possible choices to encapsulate the possible model output behaviour, for example, see Rasmussen et al. (2006) for an in-depth discussion of Gaussian Processes (GPs).

Within this manuscript there is a focus on emulating outputs one-by-one and treating them independently which for many applications is a suitable and justifiable approach. I would recommend that the authors also reference that there exists numerous multivariate emulation techniques such as: separable GP emulation (Conti et al. (2010)); the outer-product emulator which extends the separability assumption onto the regression components (Rougier (2008) and Rougier et al. (2009)); non-separable emulators (Fricker et al. (2013)); parallel partial GPs (Gu et al. (2016)); and through the use of basis representations of multivariate outputs (Higdon et al. (2008) and Salter et al. (2019, 2022)). It is unnecessary to provide further methodological details in this manuscript. A further minor point regarding the approach to emulation described in lines 500-502; it would help to provide the name for this technique as multilevel, multiscale or multi-fidelity emulation, in order to aid the reader in identifying other literature or code implementations.

Practical implementation of the described framework is important and the authors therefore signpost the reader to the recently released `hmer` R package. It would also be useful to mention the accompanying website, `https://hmer-package.github.io/website/index.html`, which provides detailed tutorials, as well as links to published research using the package. For GP emulation, see the RobustGaSP R package, `https://cran.r-project.org/web/packages/RobustGaSP/index.html`, and the associated papers (Gu et al. (2016, 2019)). In addition, many within the paleo modelling community use Python as their main programming language, hence it would be of use to suggest similar Python modules such as GPy for GP emulation, `https://sheffieldml.github.io/GPy/`, which also includes tutorials.

2.4 Uncertainty Quantification

In section 2.4, lines 270-278, the authors introduce what is commonly referred to as an additive error structure. It would aid the reader to state this. Moreover, this seems like the default choice with only a brief mention given to a multiplicative error structure in lines 293-295, whilst it should also be noted that it is not strictly necessary to log-transform the output (both the simulation and observation/reconstruction data) to simply return to the more common additive error structure. It would also add to the content of section 2.4 to comment on biases in observational errors and structural model discrepancy and how these should be accounted for within the error model. A comment should also be included regarding any sources of uncertainty exhibiting a parameter dependency, with this being particularly relevant for (internal) structural model discrepancy. In equation 14, the authors should comment on the implications of whether the model M is deterministic or stochastic. This is not immediately clear when going from the first to the second line of this equation.

The authors include the following overly strong statement about structural model discrepancy: “Within most (if not all) scientific disciplines, the tendency to date has been to effectively ignore this source of uncertainty” (lines 340-341). Whilst it is true that model discrepancy is often overlooked (or given minimal treatment) within the (paleo) climate and ice-sheet modelling literature, there exist examples across a range of applications where model discrepancy is explicitly assessed, for example: in cosmology (Vernon et al. (2010)); epidemiology (Andrianakis et al. (2015, 2017)); and in climate modelling (Edwards et al. (2019)).

2.5 A Bayesian Framework for the Analysis of Paleo Computer Models

Appendix A1 serves as a useful recipe for performing a Bayesian analysis of paleo computer models. Given that the main aim of this manuscript is to promote this framework to the reader for their analyses, would it be best to include a more compact version of this recipe within the main body, for example, at the start of section 2, with cross-referencing to the relevant subsection for each numbered step? A more discursive example could then be left to the appendices.

Throughout appendix A1, the authors provide guidance on the number of simulations required for each step of the framework. However, the exact number of simulations really depends on numerous factors including: the computational cost of running the model; whether multiple models of differing complexities within a hierarchy are being used; what are the available computing resources; whether an iterative analysis such as history matching being performed, or a single stage analysis such as a full Bayesian calibration; the smoothness in the behaviour of the simulator outputs of interest with respect to changes in the parameter settings; prior beliefs about the simulator(s) combined with information gained from previous studies; and the type or form of the statistical model to be fitted. Without explicitly referring to such factors in the authors’ previous analysis or analyses, the quoted number of simulations holds limited value. The authors should therefore provide more details of the configuration of previous study or studies to which they are referring, noting that this also links to my above concerns about the lack of a concrete and numerical example raised in section 2.2. In addition, the manuscript would be further strengthened by referencing guidance on the practical choice of the number of simulations detailed in Loeppky et al. (2009).

3 Technical Corrections

The following is a list of the technical corrections I have spotted during my review of the manuscript. This is by no means complete and should serve as examples of the types of errors found. I would advise that the authors take greater care to check their manuscript before submission as the many errors or vague choice of language made the paper difficult to follow.

- Line 48 – Remove repeated fullstop.
- Tables 1 & 2 – Improve the presentation of these tables to aid the reader by adding horizontal and vertical lines to better delineate rows and columns.
- Table 3 – In the row for the "Mathematically limited paleo researcher", there is a "" in the second column. Evidently something is missing here.
- Section 2 title – Capitalise "t" in "the Bayesian framework".
- Line 162 – Model configuration is denoted by the parameter vector C_M. To adhere to general conventions regarding scalars and vectors, this would be better denoted in bold font as $\mathbf{C}_M$.
- Line 173 – There is a rogue closing bracket at the end of the line after "multiplication rule".
- Lines 171-175 – Combine equations 7 and 8 as the repetition is unnecessary.
- Footnote 2 – This point seems more important to the understanding of the reader and as such should be elevated to the main body of text.
- Equation 12 (page 9) – The denominator, $P(D)$, is computed using the law of total probability which is not introduced until equation 19 (page 22). For those readers less familiar with probability, it would be useful to provide this formula before (or within) equation 12.
- Lines 195-202 – The translation of the hypothesis based science to Bayes rule would be clearer as a numbered list.
- Line 302 – Unnecessary opening bracket before $R - M(C_M)$.
- Line 328 – Remove the second "this".
- Equations 15-17 – There is a change in notation from capital "M" to lowercase "m" when describing the model. Is this deliberate? If so, please explain why. Otherwise, change to M for consistency with the rest of the paper.
- Figure 2 – Add light blue confidence interval for the structural error corrected model to demonstrate that this overlaps the data.
- Line 501 – Correct the formatting for the citation for "Cumming and Goldstein, 2009" which runs into the margin.
- Footnote 11 – Add a fullstop to sentence.
- Section 3.1 – The notation *cm* appears to be used to denote a parameter vector. I assume this is supposed to be C_M for consistency with section 2.
- Line 906 – Remove the word "to" before "towards".
- Section 4.7 title – Capitalise the first word, "a".

4 Conclusion

I am supportive of the Bayesian uncertainty analysis framework presented in this manuscript and strongly agree with the need for such reproducible and defensible analyses, whilst admiring the authors' ambition to unify the paleo modelling community in their approach to uncertainties. At this stage I recommend the authors consider rewriting this manuscript due to the numerous points mentioned in this review. In particular, it is necessary to greatly condense the description of the framework to the core elements, with additional information provided for those interested through supplementary material and via references to other resources that already provide an appropriate coverage of the facet. The exposition would be greatly helped by including a substantive example applying this framework to paleo models, along with appropriate code, which will also motivate and demonstrate to the reader how such an analysis can be practically implemented.

References

Andrianakis, Ioannis, Ian R. Vernon, Nicky McCreesh, Trevelyan J. McKinley, Jeremy E. Oakley, Rebecca N. Nsubuga, Michael Goldstein, and Richard G. White (2015). “Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda”. In: *PLOS Computational Biology* 11.1. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003968.

— (2017). “History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance emulation”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 66.4, pp. 717–740. ISSN: 0035-9254. DOI: 10.1111/rssc.12198.

Conti, Stefano and Anthony O’Hagan (2010). “Bayesian emulation of complex multi-output and dynamic computer models”. In: *Journal of Statistical Planning and Inference* 140 (3), pp. 640–651. DOI: 0.1016/j.jspi.2009.08.006.

Edwards, Tamsin L., Mark A. Brandon, Gael Durand, Neil R. Edwards, Nicholas R. Golledge, Philip B. Holden, Osabel J. Nias, Antony J. Payne, Catherine Ritz, and Andreas Wernecke (2019). “Revisiting Antarctic ice loss due to marine ice-cliff instability”. In: *Nature* 566, pp. 58–64. DOI: 10.1038/s41586-019-0901-4.

Fricker, Thomas E., Jeremy E. Oakley, and Nathan M. Urban (2013). “Multivariate Gaussian Process Emulators With Nonseparable Covariance Structures”. In: *Technometrics* 55.1, pp. 47–56. DOI: 0.1080/00401706.2012.715835.

Gu, Mengyang and James O. Berger (2016). “Parallel Partial Gaussian Process Emulation for Computer Models with Massive Output”. In: *The Annals of Applied Statistics* 10.3, pp. 1317–1347. ISSN: 1932-6157. DOI: 10.1214/16-AOAS934.

Gu, Mengyang, Jesus Palomo, and James O. Berger (2019). “RobustGaSP: Robust Gaussian Stochastic Process Emulation in R”. In: *The R Journal* 11.1, pp. 112–136. DOI: 10.32614/RJ-2019-011.

Higdon, Dave, James Gattiker, Biran Williams, and Maria Rightley (2008). “Computer Model Calibration Using High-Dimensional Output”. In: *Journal of the American Statistical Association* 103.482, pp. 570–583. DOI: 10.1198/016214507000000888.

Loeppky, Jason L., Jerome Sacks, and William J. Welch (2009). “Choosing the Sample Size of a Computer Experiment: A Practical Guide”. In: *Technometrics* 51.4, pp. 366–376. DOI: 10.1198/TECH.2009.08040.

Rasmussen, Carl Edward and Christopher K. I. Williams (2006). *Gaussian Processes for Machine Learning*. The MIT Press. ISBN: 0-262-18253-X. URL: https://gaussianprocess.org/gpml/chapters/RW.pdf.

Rougier, Jonathan (2008). "Efficient Emulators for Multivariate Deterministic Functions". In: *Journal of Computational and Graphical Statistics* 17.4, pp. 827–843. DOI: 10.1198/106186008X384032.

Rougier, Jonathan, Serge Guillas, Astrid Maute, and Arthur D. Richmond (2009). "Expert Knowledge and Multivariate Emulation: The Thermosphere–Ionosphere Electrodynamics General Circulation Model (TIE-GCM)". In: *Technometrics* 51.4, pp. 414–424. DOI: 10.1198/TECH.2009.07123.

Salter, James M., Daniel B. Williamson, John Scinocca, and Viatcheslav Kharin (2019). "Uncertainty Quantification for Computer Models With Spatial Output Using Calibration-Optimal Bases". In: *Journal of the American Statistical Association* 114.528, pp. 1800–1814. DOI: 10.1080/01621459.2018.1514306.

Salter, James M. and Daniel B. Williamson (2022). "Efficient Calibration for High-Dimensional Computer Model Output Using Basis Methods". In: *International Journal for Uncertainty Quantification* 12.6, pp. 47–69. ISSN: 2152-5099. DOI: 10.1615/Int.J.UncertaintyQuantification.2022039747.

Vernon, Ian, Michael Goldstein, and Richard G. Bower (2010). "Galaxy Formation: a Bayesian Uncertainty Analysis". In: *Bayesian Analysis* 5.4, pp. 619–670.

Vernon, Ian, Junli Liu, Michael Goldstein, James Rowe, Jen Topping, and Keith Lindsey (2018). "Bayesian uncertainty analysis for complex systems biology models: emulation, global parameter searches and evaluation of gene functions". In: *BMC systems biology* 12 (1). ISSN: 1752-0509. DOI: 10.1186/s12918-017-0484-3.