Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-136-RC1, 2020

Geoscientific Model Development Discussions

Interactive comment on "Constraining stochastic 3-D structural geological models with topology information using Approximate Bayesian Computation using GemPy 2.1" *by* Alexander Schaaf et al.

Ashton Krajnovich (Referee)

akrajnov@mymail.mines.edu

Received and published: 17 October 2020

General comments:

The article "Constraining stochastic 3-D structural geological models with topology information using Approximate Bayesian Computation using GemPy 2.1" provides a practical and easy to understand method for the use of topological analysis in probabilistic geomodeling as part of a likelihood-free Bayesian inference scheme. The article moves research in probabilistic geomodeling in a meaningful direction towards the in-

corporation of geologic knowledge (or "knowledge-based inputs"), which is a valuable contribution as the uncertainty of geological knowledge is a traditionally underrepresented aspect of geologic modeling. The article does so by building upon recent works in topological analysis of 3D geologic models, demonstrating the application of geologic knowledge in the form of topology graphs describing the known distribution and relationships of normally-faulted stratigraphic units. The method put forth in the article presents practical advancements to enable the broader use of knowledge-based inputs in probabilistic geomodeling through the combined use of likelihood-free Bayesian inference (via Approximate Bayesian Computation (ABC)) and advanced sampling schemes (Sequential Monte Carlo (SMC)). The positive implications of the use of these tools in the research are clearly stated: circumventing the intractability of defining mathematical likelihood functions for abstract geologic knowledge and demonstrating processing performance improvements through a brief discussion of simulation efficiency. The authors give clear credit to related works both in the realm of geologic modeling as well as the broader fields of topological analysis and Bayesian statistics. The work fits well into the current state of probabilistic geomodeling research, with the research objectives achieved coinciding with recommendations made in recent works. The work provides necessary codes, algorithm descriptions and parameter files for reproduction of the research results both in the text and as supplemental material (Zenodo DOI). The article title effectively communicates the contents of the paper.

The work utilizes the developing, open source GemPy geologic modeling environment effectively, particularly highlighting the strengths provided from GemPy's ability to efficiently integrate stochastic simulation, topological analysis and 3-D geomodeling in a single platform. This integration is a critical component necessary to improve geomodel processing efficiency through iterative sampling schemes such as SMC. The discussion of the processing efficiency improvements however is brief, and as a significant contribution of this work, should be addressed in more detail (see specific comments for suggestions).

A key assumption that the demonstrated method operates under is that the observed topology graph of the geologic model being analyzed is known without uncertainty. This assumption needs to be stated more clearly to the reader, as the current presentation is confusing (e.g., the train of thought from line 21-24 is not in line with the proposed method in line 64). This is important as the convergence to a single topology graph plays a role in the significant reduction in uncertainty seen across the final probabilistic geomodel ensembles, so the discussion of this reduction should be clearly stated in light of the use of a known subsurface topology. Clarifying this assumption could also help clarify confusing mathematical notations and technical terminology used when describing summary statistics in Section 2.3.2.

Another aspect of the paper that is not sufficiently addressed is the selection and definition of the input parameter prior uncertainty distributions. It is unclear whether they were defined as broad, non-informative priors, based on empirical analyses, drawn from previous works, or assumed by the modeler. While the focus of the paper is on the use of ABC to incorporate geologic knowledge in the form of topology information, the core methodology is based in input-based, probabilistic geomodeling, and as such the discussion of input parameter prior uncertainty distribution selection and characterization needs to be discussed in some more detail.

The structure, language and mathematical notation of the paper could use some improvement. Figure callouts are often out of order and separated from their referenced figures by up to a page or more, hurting the paper's readability. Some figures might also be combined for ease of reference, e.g., Figures 6 and 7 and Figures 10 and 11. The synthetic and realistic geologic models case studies share many similarities (differing mainly in size and the presence of an overlying unconformity), leading to some avoidable repetition of information in the description of methodology and results between these two models. The language used leans towards a somewhat casual style, exhibiting some repetitive sentence structures and, in some cases, run-on sentences and other grammar related readability issues (see technical corrections for edits and

suggestions!). The mathematical notation is unclear in some places (e.g., Paragraph at line 161) and should be reviewed to be consistent and in line with the general statistical literature (rather than just from a specific cited work). Some technical terms lack definitions before their introduction (line 21) and in a few cases are provided with confusing definitions (e.g., Paragraph at line 161). These issues do not significantly impede the quality of the research, but will require minor revisions.

Overall, this research is a valuable and fitting contribution to GMD. Minor revisions are suggested regarding structure, figures, language, mathematical notation and definition of technical terms. More importantly, additional clarification and discussion is necessary regarding: (i) the reasoning behind and implications of key assumptions used in the work, namely the use of a known topology graph, and (ii) on the description of input parameter prior uncertainty distributions (and their impact on potentially low model acceptance rates). Following these revisions and clarifications, I would recommend this paper for publication in GMD.

Specific Comments:

Title: Consider rephrasing to avoid the repetitive use of the word "using". I would suggest: "Constraining stochastic 3-D structural geological models with topology information using Approximate Bayesian Computation in GemPy 2.1"

Abstract: As the research is built in the GemPy environment, it would be beneficial to highlight it's usage in the abstract (perhaps at Line 13).

Line 129: Sentence requires revision to be accurate about what the likelihood function represents in Bayes' theorem. I suggest: "This updating process relies on the use of a likelihood function p(y|theta), representing the conditional probability of the observed data y given the prior probability of the underlying parameter theta and the theoretical connection to the occurring event."

Line 144: You have reversed the conditional probability described by the likelihood

function, which is: the likelihood for observing the data y, given the model based on uncertain parameters theta.

Line 147: This is unclear, as likelihood functions are inherently encoding information regarding not just the parameters theta, but also the observations y and the assumed theoretical relationship between theta and y. Consider removing or revising.

Section 2.3.2: This section requires additional clarification between "observed data" and "simulated data". Refer to the treatment of ABC in Gelman et al., 2004 where y is the observed data (observed "summary statistic" in ABC) and y-rep is the simulated data (simulated "summary statistic" in ABC). The use of y-hat to represent the observed summary statistic and y to represent the simulated summary statistic creates additional confusion (as the observed data introduced in Bayes' theorem were defined as y, not y-hat).

Line 156-157: Perhaps add a reference to (Wood and Curtis, 2004)? (Geological prior information, and its applications to geoscientific problems)

Line 160: Please add an additional clarifying sentence on what the summary statistic is in this work rather than the short parenthetical (to avoid confusion with typical summary statistics like mean, mode, median etc.). Also, a comment: In the proposed (approximate) inference scheme, the new evidence y (or data) is the "summary statistic". So, while the definition of the additional term "summary statistic" to describe "y" is useful for highlighting the approximate nature of ABC, the equivalency of these two terms should be clarified for the reader.

Line 162-163: Clarify the 2nd part of the sentence to illustrate that the "observed summary statistic y-hat" is static for the entire geomodel ensemble (i.e., the known, observed topology graph), while "the summary statistic y" is tied to each individual geomodel realization (i.e., a simulated topology graph).

Line 165: Theta-prime has not been introduced. What does it refer to as opposed

to theta? I assume you are referring to a single draw from the parameter distribution theta, but please clarify. When relying on mathematical notations from another work (the ones in question here seem to be borrowed from Sadegh and Vrugt, 2014), make sure notations are introduced properly. It also helps to also have a "sanity check" to make sure that the notation used is not confusing with respect to the broader statistical literature (e.g., where the observed data in Bayes theorem are typically represented without a ˆ or ')

Section 2.5: Section could be made much more concise to avoid excessive overlap with existing works (seeing as the major contributions of the paper are not focused on novel applications of Shannon entropy).

Line 227: How and why were the prior uncertainty ranges chosen? Were they considered to be broad, non-informative priors, derived emprically, based on background information or simply assumed by the modeler for the sake of simulation? Same question should be addressed more directly for the Gulfalks case study as well (Line 249-251), where the uncertainties appear to be derived from the referenced work though this is not stated definitively. Also, just a comment: I am quite interested to see how incorporating structural uncertainty (by way of the methods put forth by Pakyuz-Charrier et al., 2018a,b, Roberts et al., 2019 or Krajnovich et al., 2020) would influence the geomodel topology.. Intuitively, there is a high potential for confounding effects on the range of possible geomodel topologies when interface location and interface/fault orientation are varied together!

Line 246: How was the interface uncertainty applied to the surface points? Independently at each node, or generally to the set of surface points (so as to retain surface shape). From reading into the supplemental codes, it appears that the uncertainty was applied to the group of surface points – but this information needs to also be included in the text for the typical reader. This also applies to the synthetic model, which appears (from the code provided) to have been modeled from similar groups of surface points, though this is not clarified in the text.

Line 251: Tying back to the earlier comments on how prior uncertainty ranges were chosen, I believe that “ease of implementation” is somewhat of an inconclusive reasoning. The rest of the sentence provides more meaningful perspective but still could be expanded upon (e.g., what is “simplified uncertainty modeling” in this context?). Please add some more detail.

Line 268: A figure representing this most frequent topology graph from simulation (or other selected simulated topology graphs) would be quite insightful, especially if accompanied by a discussion of their geologic significance (e.g., tying back to points made during the introduction (Line 51), did any simulated topology graphs represent a compressional rather than extensional tectonic regime?). If length permits of course - perhaps if some figures are combined or suggested section lengths reduced, this could be added.

Line 287: Since the Jaccard Index used could allow for multiple topologies to be present in the final model ensemble (depending on the rejection threshold used), it would be beneficial to see some exploration of what these possible model topologies looked like (how geologically unrealistic do they get? Are all 675 unused topologies absolutely unrealistic?). Including a discussion of this sort would help guide future works investigating uncertainty of the applied topology information itself (without requiring reproducing the results to show geomodel uncertainty when multiple simulated topologies were present in the final ensemble). See also Comment for Line 268.

Line 295: In line with the missing clarification regarding the assumption of the observed topology graph being known without uncertainty, add some clarification behind the reasoning for setting the rejection threshold such that only the applied initial topology remains in the probabilistic geomodel ensemble. Was the goal of empirical testing of thresholds to find the largest threshold which resulted in only a single model topology remaining across the probabilistic geomodel ensemble?

Line 297: How does simulation time for ABC-REJ compare to simulation time for the

standard MC approach?

Line 298: This is a significant improvement in efficiency! Perhaps include a description of acceptance rates from each epoch of SMC, or at least a comparison of the final acceptance rate at the threshold value of 0.025 in SMC for comparison with the rate given for REJ. This information might fit naturally in Figure 12.

Line 324: If the information applied were non-meaningful (e.g., an incorrect topology graph), the geomodel ensemble would likely still exhibit a reduction in entropy due simply to the convergence of the model realizations towards the single model topology applied. That is, the reduction in uncertainty is arising from the reduction of possible model topologies, not necessarily the meaningfulness of the model topology used in the ABC algorithm.

Line 334: It appears that expanding the ABC approach proposed here to incorporate multiple observed topology graphs would not be a matter of "easily scaling". Revise to clarify that the general ABC framework would definitely allow for this, although it would require reparameterizing the current summary statistic and discrepancy measure (distance function), and also possibly changing the simulation method (as mentioned in Line 357-359).

Line 335: This would be a good place to bring up again the implications of using the demonstrated ABC approach if there were uncertainty about the observed topology graph.

Line 345: “. . .reducing the parameter dimensionality” – how so? The number of input parameter probability distributions is the same in standard MC or in ABC-REJ/SMC. The computation efficiency improvements arrive from reducing the number of input parameter draws that are run through uncertainty propagation to the 3D geologic model space, which in SMC also allows for reducing the size of the uncertainty space (note, not the parameter dimensionality) iteratively.

Line 370: This was not discussed earlier in Section 3.2 when the acceptance rate was initially 0.0059 (0.59%). Does that low acceptance rate warrant reassessing the prior input uncertainties used in the probabilistic geomodeling? Should be discussed to better frame the current work and guide future work.

Figure 4: Consider replacing X & Y with N & E to be more intuitive for geoscientists. Applies to all figures of geomodels with labeled axes.

Figure 6: In my opinion, the XZ difference section (and possibly then also XY and YZ sections) from Figure 7 could be appended onto Figure 6 for ease of reference. Also, what do the overlain crosshairs show?

Figure 8: Figure does not show (a), (b), (c). . . tags. Also, as mentioned in the comment for Line 279, the figure does not show histograms.

Figure 10: The significant reduction in model entropy indicates the strong dependence on the initial topology used - this potential source of bias should be addressed. Please discuss the implications of using a rejection threshold which only allows one model topology across the entire final ensemble of geomodel realizations. Since the authors are operating under this (valid) assumption, it needs to be clearly stated earlier that the initial geomodel topology is "known" and treated without uncertainty. See also comments regarding Lines 324, 295, 287 and 268. Also, I believe this figure could be merged with Figure 11.

Figure 12: Figure needs correction to show Y-axes labels. Perhaps acceptance rates per epoch would be useful to add as well, as they are tied to the processing efficiency improvement of 10.1x (see comment regarding Line 298).

Technical Corrections:

See the annotated PDF provided as supplementary material.

Please also note the supplement to this comment:

https://gmd.copernicus.org/preprints/gmd-2020-136/gmd-2020-136-RC1-supplement.pdf

---

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-136, 2020.