Monday, 29 March 2021 3:32 PM

From <https://gmd.copernicus.org/preprints/gmd-2020-400/gmd-2020-400.pdf>

This article is in an important area for geology - building models of the subsurface quickly, using all or most available data and highlighting uncertainties. The paper is an important contribution in terms of the software that it is offering for the community and demonstrating this software with a series of datasets in a case study. However, the article suffers from a number of issues that mean that I would advise the authors reorganising the paper, running some additional benchmarking and documenting how efficient the code is as well as how it works in terms of datatypes and inputs and outputs. The paper would also benefit from a better review of the literature in places, especially the introduction and discussion. The comments below are as detailed as I could be to improve the paper and should not require a lot of additional scientific work, mainly rewriting and organising the text.

General Comments

a. The article is poorly organised and confusing in its current state, mixing a several terms, data types, layer types, categories, inputs and outputs without clearly structuring their relationship. Parts of the writing leave a lot to be desired - there are a lot of unscientific terms and quite a lot of referencing missing. New paragraphs are not always given a line in between (or alternatively indented), making the article hard to read at points. The use of caplock text is annoying (e.g. heading 2.1, line 121)

b. A development and technical paper "usually include[s] a significant amount of evaluation against standard benchmarks, observations, and/or other model output as appropriate." I would not consider the example dataset given a significant amount of evaluation - other datasets such as NSW etc. are mentioned and it is also said that the time to build such models is very small ("The example map and associated data used in this paper took just over 3 minutes to deconstruct with map2loop and a further 4-15 minutes to build with the three target modelling engines, running on a standard laptop computer." - lines 456-7) so I'm not sure why more examples are not given. In addition, the inclusion of a very imprecise single runtime should not be considered a benchmark. The architecture of the machine needs to be mentioned as well ("standard laptop" does not suffice).

c. I suggest that the examples of data given in 5,7,9 and 10 are actual results of the model run against a particular dataset rather than just examples of inputs/outputs. Therefore, I would split the results section as a case study and describe the model separately. Even better, the results from this dataset could be used as an example and several case studies used to show different run-times etc. for benchmarking the results and the difficulties/comparisons drawn between these models discussed.

1. Introduction

a. The introduction reads more like an abstract (and indeed is rather similar to most of the abstract) rather than a detailed review of the literature in this area. There are not enough papers cited and the introduction seems to define new concepts rather than base those concepts clearly on previous work. The second half of the introduction (Lines 75-104) reads more like it belongs in the methodology section.

b. Line 53-54: "Unfortunately, away from basins and mines, drill-holes are often too shallow to provide constraints at the regional scale, and also often lack stratigraphic information." It would be great to specify in which context your proposed solution is targeted at and what other solutions for other regions would work (e.g. in basins). This might be in the introduction or in the discussion or both.

c. Lines 52-62: This paragraph is difficult to read and no references are given as examples of the types of data or on alternative methods of validation mentioned, e.g. via gravity and magnetic data (line 62).

2. Input Data

a. The introduction to this section does not let the reader know that the following 6 subsections correspond to the six input data types on Figure 1. The lack of consistency between headings of the subsections and the titles of the datatypes from Figure 1 makes this also difficult to understand. The text should additional specify what constraints each of the input data types have - it mentions which datasets were used to derive them for this example but not what other similar datasets (e.g. are there easily available WAROX or ASUD equivalent databases in other parts of the world? If so give us some examples)

b. More details could be provided about how they were generated for this case (e.g. data type 5). Why are some of the polygons mentioned in input 1 missing from input 5 (e.g. Hardey Formation)? How is the operation coded and where is that handled? Is it a preprocessing step or is it included in the current code? If it is preprocessing, how is this calculated? (e.g. array sorting?).

c. It is difficult for the reader to understand the difference between input data types and how they relate to the three categories of data mentioned in the introduction (position, gradient and topological) as well as how they map to the subsections in Section 2 on input data (e.g. 2.1 Chronostratigraphic layer)

d. Programming terms related to classes in the code (such as POLYGON) should be introduced in some form in the introduction and demarcated from the regular word polygon (which might suffice in any case because I think your programming classes are also regular shapes with 3 or more sides. Multipolygon/Polyline/Multipolyline could just be defined in the intro as well. Again, the headings of 2.2 seem to indicate layers contain a single type of shape (e.g. polyline) that is appropriate to a single datatype (e.g. a fault). It would be good to indicate in a table of some kind which types can be associate with which layers/datatypes etc.

3. Augmented Data Outputs

a. Each of the subsections of this section are poorly written, full of jargon and hard to understand what is being discussed or why. For example "For regional geological models[,] a high resolution topography model is usually not needed" - under what conditions is "usually" applied? The subsections miss introductory sentences that explain how this datatype fits into the broad categories of data outlined in the intro to section 3 (positional/gradient/topological) or why it exists and the intro to section three does not let us know why these particular datatypes exist either. This section also includes quite a bit on methodologies used to create the outputs - so perhaps this could be a methodology section? Or a separate section on methodology and then a section on outputs showing how generated data could be outputted from the code. Also could be clearer which datatype corresponds to map2model / map2loop by organsing them under those headings or mentioning this in the introduction (e.g. map2model can generate x,y,z while map2loop can generate x,a,b,c).

b. Lines 216-218*:* "This paper focuses on two libraries, map2model (C++), and map2loop (Python).

map2model performs a spatial and temporal topological analysis of the geological map, and map2loopfurther refines this analysis by including information from non-map sources, such as stratigraphic databases, acts as a wrapper for map2model and performs all other calculations."

c. I'm not sure if this belongs at the end of this section or was cut and paste from the introduction at some point. It seems out of place here and I'm not sure how it relates to either the preceding or following paragraphs.

4. 3D Modelling of map2loop outputs

a. Confusingly the title suggests that only map2loop should be included here, but line 439 immediately mentions map2model. This section is very brief and needs to be expanded or integrated into the discussion section or a results section created.

b. Line 447: "Of course the whole system relies on the quality of the input data" - why "of course" and how is this handled? Remind us of any quality checks and how the quality of the input data might affect results (or papers that show how this can be handled). Why is the integration of geophysics beyond the scope of the paper (see my comments in the introduction about basins - this should be tightly linked here)? Immediately, it seems that DEM might be considered a geophysical dataset already and if certain shapes can be extracted from geophysical data (polygons/polylines) then why can't they be included?

5. Discussion

a. This section reads like an uncertainty section. However, a lot of the discussion fails to discuss uncertainties very scientifically, with terms like "reasonably robust" (line 464) or "'reasonable' results" (line 469). The discussion fails to identify ways to handle the uncertainties in a well-defined way. Since there is "no unique solution" (line 470) and many choices are "arbitrary" (line 505), it seems impossible to quantify model errors or uncertainties. However, the workflow nature of the model generation and mapping of particular choices should make error estimation possible. I would have liked to have seen a more robust discussion of the literature in this regard. Section 5.3 mentions wrapping the workflow in a Bayesian network but it would be good to have this discussion linked back to the previous sections on different kinds/sources of uncertainty in the models and what the implications for these different kinds/sources are on the Bayesian framework.

b. The sensistivity analysis is mentioned to be essentially trivial (line 565), so why not perform it for the current dataset?

6. Conclusions

I'm not sure that all the conclusions are carefully based on the results of the paper. In particular:

a. It has not been proven (only claimed) that the model reduction is significantly reduced.
b. Priors (a Bayesian concept) is first mentioned in the conclusion
c. And see my remarks about uncertainty above, that this provides a homogeneous pathway to sensitivity analysis etc - since these are only claimed to be in development.

Minor Issues:

1. Line 52: New paragraph or not?
2. Line 56: In a 3D workflow; (, should replace ;)
3. Line 56-59: Difficult to understand what is meant here:
4. *In a 3D workflow; spatial and temporal topological data, such as the age relationships between*

faults and stratigraphic units, we combine all of these direct observations with conceptual information, based on our understanding of the tectonic history of the region, including assumptions regarding the subsurface geometry of faults and plutons, to provide sufficient constraints to build a 3D geological model.

5. Lines 206-208*:* These are the same types as input data. Why not refer readers back to the definition of these types in the introduction.
6. *These outputs are grouped by type: positional outputs, which provide information on the location and shape of features; gradient outputs, which provide information on the orientation of features; and, topological outputs that provide information on the spatial and temporal relationships between features.*
7. Line 248 and Line 258: are essentially the same: "Processing of X geometries consists essentially of extracting the x,y location…."
8. POLYGON vs polygon (etc.): An alternative is just to use "Polygon" after defining it. However, the word polygon with lower case could probably suffice in most cases. In some cases, the capslock is not used consistently, e.g. line 172 (POINTS) and 173 (point) or in heading 2.1 ("Chronostratigraphic POLYGON and MULTIPOLYGON layer") and Figure 1 ("1. chronostratigraphic polygons"). Consistency is a problem, e.g. Line 288: Is the series of x,y,z points - are these POINTS or points? Also, for example, Figure 3 references POLYGONS etc. in caption by "polylines" in the figure insert.

From <https://gmd.copernicus.org/preprints/gmd-2020-400/gmd-2020-400.pdf>