Review of, “Mapping Antarctic Geothermal Heat Flow with Deep Neural Networks optimized by Particle Swarm Optimization Algorithm”, by Liu et al.

Review by Michael Wolovick

Summary

In this manuscript, Liu and coauthors infer Antarctic Geothermal Heat Flow (GHF) by training a neural network on a global database of heat flow measurements and a variety of geological or geophysical predictor variables. They train their model to predict spatially binned global observational GHF data using the input variables, without any knowledge of spatial structure (other than the spatial structure implicit in the predictor variables). They use a technique called particle swarm optimization to select the optimal hyperparameters in their deep learning model. They try four different configurations of this technique to ensure that their results are robust, and they also test the robustness of their model by leaving out variable amounts of input data in a densely sampled region of Europe. Their analysis reveals significant advantages of deep learning techniques as compared to linear regression. Their final predicted Antarctic GHF map is broadly reasonable from a geophysical perspective, with higher heat flow in West Antarctica and the Antarctic Peninsula compared with East Antarctica, although they do find some notable local maxima of GHF in East Antarctica under the Gamburtsev Subglacial Mountains and Vostok Subglacial Highlands. They conclude with a comparison of their model against four other published estimates, and against the limited in situ observational data available.

Major Comments

My expertise is in glaciology, numerical modeling, and geophysics, so I cannot truly evaluate the deep learning techniques that the authors have used. As far as I can tell, they appear to have been quite thorough from a machine learning perspective, with a lot of effort having been spent to ensure that their model fits the data well with the right set of hyperparameters. However, I do know something about Antarctic heat flow, and I do understand the importance of having good input datasets and predictor variables for a project like this, especially since their deep learning models have no internal knowledge of spatial relationships: their model treats each grid cell as an independent data point, and therefore any spatial structure in the output must come from spatial structure in the inputs. I also understand how vital it is that the input datasets contain realistic and accurate data underneath the Antarctic Ice Sheet; no matter how good the neural network is, if it is supplied with bad inputs, then it will produce bad outputs. It is therefore vital that the authors manuscript shows readers maps of all of the predictor variables, both globally and with a south polar view for Antarctica.

Unfortunately, this manuscript does not show those maps. It is therefore impossible for readers to evaluate how much the authors’ results can be trusted. These figures need not all be in the main text; with 16 variables (14 after removal of two highly colinear variables) there is a lot of information to display, so supplemental figures or an appendix would be fine here. However, it is impossible to properly evaluate the authors’ model without seeing those inputs.

This is an especially important issue given the fact that at least two of the input variables are problematic in Antarctica. The “rock type” variable (which is one of only 3 that the authors show, in Fig 2), classifies the entire Antarctic Ice Sheet under the category of “ice”. This of course makes sense for a dataset which represents the surficial rock type of a region, but we are interested in the heat flow *underneath* the ice. The subglacial rock type is generally unknown for most of Antarctica, so this dataset is worthless. Additionally, sediment thickness is another variable they use that is poorly constrained underneath most of the Antarctic Ice Sheet. While it is true that active-source seismic surveys have constrained sedimentary basin depths in a handful of locations, for the most part we simply have no idea how thick the sediment is underneath the ice, so this dataset is also worthless. The authors made decisions about which variables to keep and which to exclude on the basis of colinearity (sensibly choosing to discard redundant variables that were highly correlated with other variables), but they seemed not to have considered physical plausibility or under-ice uncertainty in their decision-making. Good input datasets for a project such as this one must be datasets for which the structure in Antarctica is well-constrained. For at least two of the inputs, that is categorically not the case. The authors need to look over their input datasets and remove those that are not well-constrained underneath the Antarctic Ice Sheet. Most obviously this includes rock type and sediment thickness, but they should double-check all of their inputs to ensure that they have realistic spatial structure in Antarctica. This will require that the authors retrain their deep learning models using only the datasets that are reliable in Antarctica. Unfortunately, this may degrade the quality of the fit and reduce predictive capacity in the rest of the world. However, that is simply the nature of the problem we are trying to solve. If the goal is to infer GHF in Antarctica, then there is no point in using datasets that are unconstrained there.

In addition to the requirement that the input datasets be well-constrained in Antarctica, it is also important that they be free of spatial artifacts, since the authors' deep learning model has no internal knowledge of spatial relationships; it treats every grid cell as an independent data point, and thus it relies on the input datasets to produce spatial structure. Unfortunately, the authors' output model (Fig 6) contains pronounced meridional stripes radiating out from the South Pole. This is likely a result of the fact that the authors interpolated all of their inputs onto a latitude-longitude grid with constant grid spacing. Constant grid spacing in lat-lon space works fine in the mid-latitudes, but it can produce artifacts near the poles, and the authors' result clearly has such artifacts. Since the authors' deep learning model treats every grid cell as an independent data point, it follows that these meridional stripe artifacts in the output are a result of similar stripe artifacts in at least one of the inputs.

There are three main methods that they could use to fix this: 1) they could use projected x/y coordinates for their Antarctic prediction while keeping lat-lon coordinates for the rest of the globe, although this potentially introduces problems in applying a deep learning model trained on lat-lon data to a new set of x/y data if the statistical distributions of the two datasets are different; 2) they could use variable grid spacing in longitude, with more grid points in each row near the equator and fewer grid points in each row near the poles, a method that looks especially attractive given that their deep learning models treat the data as a list of independent points rather than a structured grid anyway; 3) they could keep their regular lat-lon grid but apply latitude-dependent smoothing in the longitude dimension in order to ensure that their input datasets have constant spatial resolution even as the grid converges near the poles[1]. The exact method is up to the authors' choice, and they are of course free to choose a different method from the three that I propose here, but it is important that they appropriately pre-process their predictor variables to remove artifacts in polar regions, because their deep learning algorithm is not going to be capable of removing those artifacts on its own. And, of course, it is vital that the authors *show* us these predictor variables, so that we can verify for ourselves that they are indeed artifact-free.

My overall recommendation is that this paper needs major revisions. I chose major revisions rather than minor mostly because I am recommending that the authors retrain their models after removing datasets that are unconstrained in Antarctica and pre-processing to remove meridional artifacts. The manuscript itself might not need a great many changes. The additional figures I requested showing the input datasets can be placed in a supplement or appendix rather than the main text, and most of the main text can probably be kept without too much change. It might very well be that the new model has a broadly similar distribution of GHF, just without the artifacts. However, I want to see the authors' models retrained after the changes to the inputs that I described above, and since I am recommending that the authors redo their main modeling work, I classify this as a major revision.

---

1 In other words, they can apply a separate low-pass filter to every longitude row of their 0.5°x0.5° gridded datasets, with a wavelength in degrees equal to 0.5/cos(lat) and periodic boundary conditions in longitude. This would ensure that their grids have the same effective spatial resolution at high latitudes as they do at the equator.

Minor Comments

L29-30: "As an important heat source beneath the Antarctic ice sheet, GHF directly affects the hydrological system under the ice sheet (Kang et al., 2022)."

While I appreciate the reference to a paper I am coauthor on, there is probably a better reference to use here. We didn't talk much about hydrology in that paper, although we did show basal melt rates.

L33-34: "In addition, the complex interaction between GHF and climate results in a significant degree of variation in Antarctic ice mass distribution."

I'm not sure what exactly you mean here. How does GHF interact with climate? This sentence needs to be reworded or clarified.

L37-38: "...lays a significant factor for understanding the feedback mechanisms produced by Antarctic ice mass loss and predicting sea-level change"

This sentence also needs to be clarified.

L40-41: "However, the sparse and uneven distribution of in situ borehole data for GHF, coupled with the severe climatic challenges of direct measurements in the Antarctic continental interior, presents significant challenges for data acquisition (Fisher et al., 2015).

This sentence should be rephrased. How does the sparse distribution of borehole data present a challenge to data acquisition? It would be more accurate to say that the challenges of data acquisition result in a sparse distribution of data. Perhaps rephrase as, "Unfortunately, the severe logistical difficulties involved in collecting direct measurements in the Antarctic continental interior ensure that the distribution of in situ borehole data for GHF is sparse and uneven (Fischer et al., 2015)."

L42-48: " Conventional approaches fall into two categories: one based on the derivation of geothermal processes, such as decreasing west-to-east heat flow derived from some assumptions of geological conditions (Pollard et al., 2005), crustal and upper-mantle heat flow inferred from seismic models (Shapiro & Ritzwoller, 2004; Shen et al., 2020; Hazzard & Richard, 2024), and Curie temperature depths estimated using satellite magnetometry and thermal models (Maule et al., 2005; Martos et al., 2017). The other was from statistical methods such as multivariate similarity analysis (Stål et al., 2021), Bayesian inversion of multiple datasets (Lösing et al., 2020) and machine learning (Lösing & Ebbing, 2021)."

These sentences need to be reworked as well. It is wrong to describe the first set of sources as "deriv[ing] geothermal processes". "Geothermal processes" is an ambiguous phrase that could be misinterpreted as referring to hydrothermal circulation, which none of these sources represent. In addition, many of the sources in the first category are also engages in some form of statistical modeling, not process modeling. Shaprio and Ritzwoller, for example, use a similarity function to relate seismic structure in Antarctica to seismic structure elsewhere in the world, where GHF observations are available. They don't perform any thermal modeling. It would be better to say that the first group use one type of data (usually seismic tomography or magnetic anomalies), which the second group use multiple types of data. In addition, there are some references missing here.

Perhaps this section could be rephrased as: "Conventional approaches fall into two categories: on the one hand are those which use a single type of observation to infer GHF, most commonly seismic tomography (Shapiro & Ritzwoller, 2004; An et al., 2015; Lucazeau, 2019; Shen et al., 2020; Haeger et al., 2022; Hazzard & Richard, 2024) or magnetic anomalies (Maule et al., 2005; Purucker et al., 2012; Martos et al., 2017), although broad tectonic reconstructions have been used as well (Pollard et al., 2005). On the other hand, there are a newer set of statistical methods which integrate multiple types of observational constraints to infer GHF using multivariate similarity analysis (Stål et al., 2021), Bayesian inversion, (Lösing et al., 2020) or machine learning (Lösing & Ebbing, 2021).”

L60: “...deep learning algorithms … due to its high accuracy…”
Should be “due to their high accuracy”.

L91-92: “ Subsequently, these filtered, high-quality point measurements were aggregated by calculating the mean value within a 0.5° × 0.5° latitude-longitude grid. “
See my major comments about the problems with using a regular lat-lon grid when studying the polar regions.

Figure 1
Would it be good to include a couple sentences talking about the overall geographic distribution of the global data used to constrain the model? By eye, these data seem to be heavily biased towards wealthy countries, with much lower data density in Africa, South America, and the Middle East.
In addition, the color scale should be changed. Blue-white-red is appropriate for data that represent anomalies with respect to a mean or zero value. The GHF measurements being shown here are all positive, however, so a different color scale should be used.

Table 1
As I discussed in the major comments, **all** of these geophysical features need to be shown to the reader, both in global view and in south polar view. These figures can be placed in a supplement or appendix if necessary.
In addition, some of these data inputs have two sources listed. What does it mean when two sources are listed? Does that mean that the dataset is the mean of both sources? Or is it the case that one source is a publication and the other is a link to the actual dataset?

L115-116: “Sedimentary layers, due to their low thermal conductivity, act as an insulating blanket, significantly influencing the dissipation of deep-seated heat”
That may be true, but unfortunately, we have no meaningful constraint on sediment thickness underneath the ice sheet, at least not on a large scale. That is the challenge for a project like this: useful datasets are not merely those that have a meaningful physical relationship with heat flow, but those that have a meaningful relationship with heat flow **and** which are well-constrained in Antarctica. Excluding sedimentary thickness will, no doubt, reduce the quality of the global fit. However, the challenge of a project like this is to generate a model that can explain global heat flow *using only variables that are known and well-constrained in Antarctica*. Any predictive power added by sedimentary thickness will be of no help in Antarctica.

L123-125: “The Global Lithological Map (GLiM) database (Hartmann & Moosdorf, 2012) provides surface rock type data, explaining spatial variations in thermal conductivity.”
Same concern as above. Their map (at least as shown in your Fig 2) lists the entire Antarctic Ice Sheet as the “ice” rock type, which is useless for inferring subglacial heat flow.

L127-128: “To ensure dataset consistency, all predictor variables were resampled to a uniform 0.5° × 0.5° grid using Ordinary Kriging.”
As I discussed in my major comment, a uniform lat/lon grid can produce meridional stripe artifacts near the poles. Potential solutions include: 1) using projected x/y coordinates in Antarcitca; 2) using uneven grid spacing in longitude; 3) using latitude-dependent smoothing in the longitude dimension. Or perhaps a different solution that I haven’t thought of. But regardless, something has to be done to help this uniform lat-lon grid perform better near the South Pole.

L157: “the Adam optimizer”

Does this need a reference?

Equation 3

I thought that R^2 was the squared correlation coefficient? The formula for that would be:

$$R^2 = \left(\frac{\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2} \right)^2$$

Am I wrong about that? Is this a different definition of R^2?

Figure 4

There is not much range on the y-axis here. Does that mean that all four of the configurations tested here have roughly the same performance? Or that the final result is relatively insensitive to the hyperparameters? In any case, the text should probably discuss the narrow range at some point.

Figure 5

Why does the circle enclosing your test region include parts of the Black ans Aegian Seas? You have excluded marine observations from your dataset, so it seems like you could make a better dense test region by shifting the circle to only cover terrestrial parts of Europe.

In addition, why is R^2 negative for the linear regression model? Is this a function of the fact that you have defined R^2 differently than normal?

L318-322: “However, in the Gamburtsev Subglacial Mountains, Vostok Subglacial Highlands, and the area around Subglacial Lake Vostok, there is an increasing trend of heat flow values, which shows that these regions may have been affected by deep tectonic activity or localized heat sources (Artemieva, 2022).”

My own inversion for GHF (Wolovick et al, 2021) also showed a local maximum of GHF in the Gamburtsev Mountains which is necessary to fit observations of subglacial water networks there.

Figure 6

The meridional stripe-artifacts are quite prominent in the final result and uncertainty estimate here. In addition, it would be nice if the uncertainty estimate made some attempt to account for the uncertainty in the input datasets, which have uneven spatial resolution in Antarctica even for variables that are relatively well-constrained like seismic velocity or Curie Depth.

Figure 7

It would be better to show signed difference rather than absolute difference here. It is important to know which estimate is hotter! This would be a good place to use the blue-white-red color scale from figure 1.

In addition, there are quite a few additional published estimates that you could compare your model against. Additional comparison datasets include: Shapiro and Ritzwoller (2004); Maule et al., (2005); Purucker et al., (2012); An et al., (2015); Lucazeau, (2019), Haeger et al., (2022); Hazzard and Richards, (2024).

L356: “In instance…”

Should be, “For instance…”

Figure 8

It appears that many of the observations that you use to validate your model are actually located on the seafloor around Antarctica. While it certainly makes sense to include these data points when so few in situ observations are available, does it really make sense to compare your model against these data when you excluded marine observations from your training data?

L389-392: “This discrepancy may result from the heterogeneity of local geologic features, differences in raw data processing methods, or the influence of complex processes such as shallow water circulation and unsteady convection in the lithosphere, and further studies are needed to elucidate the underlying mechanisms.”

In addition, the discrepancy between your results and those of Shroeder et al. (2014) could be the result of model assumptions made by Schroeder et al. They made very specific and potentially limiting assumptions about the form of the subglacial hydrological system when constructing their inverse model, and those assumptions could potentially introduce errors into their result.

Section 6 Data Availability

This section should be after the Conclusions section, not before it.

L426: Zenodo link

It would be nice if this link also contained the processed and gridded datasets used as input to your model. While it is true that these datasets are all available at their original sources, it would be nice if it were possible for interested users to access the gridded inputs that you created for your model at one place.

L437: “...which is consistent with the active geological structures.”

Rephrase this, this sounds awkward. Perhaps try: “...which is consistent with the locations of present-day tectonic and volcanic activity.”

References:

The references should be in alphabetical order, not in citation order.

Review References

An, M., Wiens, D. A., Zhao, Y., Feng, M., Nyblade, A., Kanao, M., Li, Y., Maggi, A., and Lévêque, J.-J.: Temperature, lithosphere-asthenosphere boundary, and heat flux beneath the Antarctic Plate inferred from seismic velocities, Journal of Geophysical Research: Solid Earth, 120, 8720–8742, https://doi.org/10.1002/2015JB011917, 2015.

Haeger, C., Petrunin, A. G., and Kaban, M. K.: Geothermal Heat Flow and Thermal Structure of the Antarctic Lithosphere, Geochemistry, Geophysics, Geosystems, 23, e2022GC010501, https://doi.org/10.1029/2022GC010501, 2022.

Lucazeau, F.: Analysis and Mapping of an Updated Terrestrial Heat Flow Data Set, Geochemistry, Geophysics, Geosystems, 20, 4001–4024, https://doi.org/10.1029/2019GC008389, 2019.

Purucker, M. E.: Geothermal heat flux data set based on low resolution observations collected by the CHAMP satellite between 2000 and 2010, and produced from the MF-6 model following the technique described in Fox Maule et al. (2005), 2012.

Wolovick, M. J., Moore, J. C., and Zhao, L.: Joint Inversion for Surface Accumulation Rate and Geothermal Heat Flow From Ice-Penetrating Radar Observations at Dome A, East Antarctica. Part I: Model Description, Data Constraints, and Inversion Results, Journal of Geophysical Research: Earth Surface, 126, e2020JF005937, https://doi.org/10.1029/2020JF005937, 2021.