Referee's comments are in blue, our reply in black, quotes in the revised manuscript in red.

In this study, Wang et al. compare modelled velocities derived from a full-Stokes thermomechanical model (from a previous study, Huang et al., 2024) with surface velocity observations from MEaSUREs. I do not think this study is a novel contribution to the field, and found the manuscript confusing, so I do not recommend it for publication.

Major comments

In a preceding study, Huang et al. (2024) conduct full-Stokes thermomechanical simulation forced by 8 different geothermal heat flux models. The strength of this study is the comparison of the model results with an independent constraint – the radar specularity content of the bed. To make this comparison, Huang et al. perform an inversion for the basal parameters, based on modelled vs. surface velocities. So in this study, Wang et al. are calculating metrics based on the residuals of the modelled vs surface velocities, and are thus evaluating the performance of the Huang et al. inversions. While it seems reasonable that better performing inversions reflect more accurate GHF map, a range of other factors can be at play. For instance, the authors note the possible influence of anisotropic viscosity. Poorly-performing inversions could also be the result of uncertainties in the form of the basal sliding law, or in the basal topography model. Wang et al. present a ranking of GHF maps that is similar but not identical to that of Huang et al. It is unclear why Wang et al. think that this ranking is more accurate, especially given the circularity of the velocity residual argument mentioned above. It is also unclear that the Wang et al study presents results which are novel, relative to the Huang et al. study.

Reply: Thanks for the referee's comments. There is a misunderstanding from the referee regarding our method. In Huang et al. (2024), modelled surface velocity is compared with the observed over the whole domain in the inversion for basal parameters for each GHF data. The referee said "So in this study, Wang et al. are calculating metrics based on the residuals of the modelled vs surface velocities, and are thus evaluating the performance of the Huang et al. inversions." While it is true that our overheating metrics are based on (a subset of) the velocity residuals from the inversion, our overcooling metrics are based on the basal sliding velocity from the inversion, which does not enter into the residual of a mechanical inversion at all, since the observational constraints for a mechanical inversion are surface velocities. Moreover, a mechanical inversion does not take into account the physical plausibility of the sliding result it produces. It is not circular reasoning to compare two different parts of a model to each other; rather, it is a check of internal consistency, or lack thereof. A mechanical inversion may well be able to fit the surface velocity observations equally well when forced with many different models of the ice sheet thermal structure and rheology; however, if some of those models require high sliding

velocities in cold-based regions, then they should be downweighted in comparison to models that show a good agreement between basal temperature and velocity.

We understand how this can appear to be getting something from nothing, so to speak, since our method does not require any additional observations beyond the surface velocities used in the mechanical inversion. However, it is wrong to say that we do not have any "independent constraints" in this method. The "independent constraint" that we are using here is not an observation, but rather the a priori physical understandings that: 1) rapid sliding requires warm basal temperatures and subglacial water, and, 2) reducing the basal slip coefficient cannot prevent the ice from flowing by internal shear deformation. The inconsistency metrics that we developed in this paper are an attempt to quantify and rank the extent to which these basic (and uncontroversial) physical understandings are violated.

Of course, the reviewer is correct that other factors beyond GHF can cause inconsistencies between model velocities and temperatures. We discuss some of those factors in the discussion section of the manuscript. However, we used the same forcing fields for all of our models except for GHF, so it is fair to attribute differences in the inconsistency metrics to the only factor that varied between the models, which was GHF.

As for the referee's comment, "Wang et al. present a ranking of GHF maps that is similar but not identical to that of Huang et al. It is unclear why Wang et al. think that this ranking is more accurate,", we do not claim that our ranking is more accurate than theirs. The purpose of demonstrating this method in the same domain as used by Huang et al is so that we can validate the method against their results based on radar specularity measurements, which is useful before applying our method to other domains where such measurements are not available. However, it would be unlikely for both methods to produce exactly the same ranking. The two methods both have their own sources of uncertainty and their own sets of assumptions and arbitrary parameter choices. We are not claiming that either ranking is better in any sort of absolute sense, only that they both seem to be measuring roughly the same thing.

Wang et al. justify their study by saying that they are introducing a new method – this is misleading. They are instead introducing new terminology for "metrics" based on velocity residuals, which is a common practice in the field of glaciological inversion. One metric masks the thawed bed, another metric masks the frozen bed, but both metrics are used in the final evaluation – so it is unclear why the masking was necessary to start with. I don't understand why you need "bidirectional constraints" when you could take the root-mean-square error. Even if they were presenting a new method, they do not provide any tests to show that this method improves upon existing ones. Finally, the terminology of these metrics is confusing, and I found did not provide me with a better physical understanding of why some models performed

better than others (although this last point can be rectified with clearer language).
Reply: As we discussed in our reply above, there seems to be a confusion here: the overheating metrics use the surface velocities, and can thus be thought of as a subset of the inversion residual, but the overcooling metrics use the basal velocities, and thus they are not in any sense a subset of the inversion residual. The "new method" that we are introducing is the practice of quantifying the internal (in)consistency between a sliding inversion and the rheology structure used to force that inversion. The specific metrics that we use to quantify this inconsistency could be changed, for example by using a squared error term as the reviewer suggests instead of the linear error terms that we used, but the general practice of caring about and quantifying the inconsistency between a sliding inversion and the temperature/rheology field used as an input to that inversion is new, and that is what we mean when we describe our contribution as a new method.

We were concerned about the need for "bidirectional constraints" in the manuscript because we wanted to avoid a situation in which we only penalized models for being too cold or too warm, and thus ended up with a ranking that favored the most extreme GHF maps rather than those which accurately balanced warm and cold conditions. The reviewer does not seem to understand why we were concerned about this, because they assumed that all of our metrics were computed from the surface velocities, and thus that they could be easily combined into a single RMS metric just as in the cost function for the sliding inversion. However, our overcooling metrics are computed by comparing *basal* velocity with basal temperature, and basal velocity does not appear in the inversion cost function. This is the metric that is easiest to explain and understand: areas that are cold-based should not be sliding. But if we only used overcooling metrics, then our final ranking would be biased towards the hottest GHF maps, regardless of whether those were the most accurate. This is why we also needed overheating metrics to obtain an unbiased result. As we explained in the methods section, our overheating metrics are built around the physical understanding that a sliding inversion cannot prevent the ice from flowing by internal shear deformation. The inversion cannot reduce the sliding velocity below zero, and since it does not adjust the near-basal rheology of the ice, it is powerless to reduce the residual with the observations if that residual is caused by internal deformation. Thus if the basal ice is warmer and softer than it should be, we should expect the inversion to be biased towards positive residuals (meaning the model surface velocity is faster than the observed) in areas of warm bed. This is why we use a subset of the inversion residual for our overheating metrics.

We acknowledge that it might be preferable to use one metric or constraint over the whole domain. But we, nor anyone else to our knowledge, has proposed and applied the same metric on both warm bed and cold bed. For instance, we can use the temperature difference between ice temperature and pressure melting point in the metric for the cold bed. But we cannot use it on the warm bed, as the ice temperature at the warm bed equals the pressure melting point.

Finally, we do not claim that our method improves upon existing ones, which we assume you refer to Huang et al. (2024). As we discussed above, Huang et al. (2024) used radar specularity observations to evaluate the model result. But we do not use any englacial or subglacial data in this study. The significance of this work is that the metrics we proposed provide an effective approach to evaluate modelling results and identify problem areas without using any englacial or subglacial measurement data.

Minor comments

In Figure 3 & 4, I think it would be clearer to change the term "warm bed" to "thawed bed". And in Figure 5, "cold bed" to "frozen bed".

Reply: Okay, we change "warm bed" to "thawed bed", and "cold bed" to "frozen bed" in Fig. 3, 4 and 5.

The authors should note where they obtain their velocity observations within the main text, and maybe plot them.

Reply: Thanks for your suggestions. In the main text, we write

Huang et al. (2024) used the present-day surface ice temperature (Le Brocq et al., 2010), observed surface velocity from MEaSUREs InSAR-Based Antarctic Ice Velocity Map, version 2 (Rignot et al., 2017) and ice sheet topography data from BedMachine Antarctica, version 2 (Morlighem et al., 2020).

We also add a plot of surface velocity observation in Figure 1.

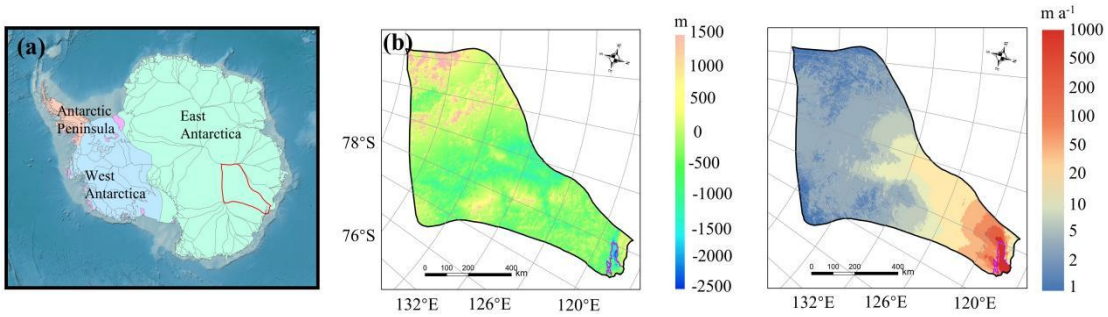

Figure 1. (a) Geographic location of Totten Glacier (red outline) in Antarctica; (b) bed elevation of Totten Glacier, the purple curve represents the grounding line; (c) observed surface velocity.

I think the authors should add a section describing the model methodology, so that the reader can understand this paper without having to read Huang et al.

Reply: Thanks for your suggestion. We add the following subsection in the revision:

2.2 Methodology in Huang et al. (2024)

Huang et al. (2024) employed thermo-mechanical coupled simulations using eight GHF datasets to investigate the steady-state thermal regime of Totten Glacier. The methodology involved two interconnected modeling components:

1. Forward Modeling: An enhanced shallow-ice approximation model integrated with a subglacial hydrology module was utilized to simulate englacial temperature profiles.
2. Inverse Problem: A full-Stokes ice flow model was applied to resolve basal friction coefficients through inverse analysis, to minimize the misfit between simulated and observed velocities while simultaneously generating velocity predictions.

A feedback loop was established: the velocity outputs from the inverse model were used to refine key parameters in the forward model - specifically constraining the basal slip ratio, rheological properties, and shape functions. This bidirectional coupling process underwent multiple iterations to achieve convergent steady-state solutions. Huang et al. (2024) further used specularity content as a critical constraint to differentiate localized wet versus dry basal conditions. They compared modeled basal thermal states with different GHFs and observational returns, and thus evaluate the reliability of GHF datasets.