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

The paper takes a thoughtful approach to assess a collection of Geothermal Heat Flux (GHF) using two sets of radar sounding observations: the detection of subglacial lakes and bed echo specularity content. The authors rightly point out that "one-sided" tests using either observable will result in the selection of the highest GHF values. The authors are thoughtful about where they do and do not apply the specularity content in terms of upstream and downstream portions of the catchment. The authors are also thoughtful about the difference between how the lakes and specularity observations are used in terms of "one-sided" vs. "two-sided" constraint.

However, the authors seem to view the choice as either a "one-sided" approach that only evaluates if the lakes/high-specularity correspond to thawed areas or a "two-sided" approach in which that comparison is combined with evaluation of "no lake"/low-specularity correspond to cold areas. However, it seems like there's another option. To "reward" the match between lakes/high-specularity and thawed areas as in the "one-sided" and then to "penalize" a mismatch between lakes/high-specularity and cold areas. This seems like more than the "one-sided" approach and like it might not be vulnerable to the same preference for the highest meld as the pure "one-sided" approach. However, it also seems like it has the benefit of applying to both lakes and specularity. It has the additional benefit of allowing the specularity to be used in all regions of the catchment. This seems additionally important because, just as the authors describe for lakes, it's not the case that low-specularity areas have to correspond to cold areas, it can correspond to thawed areas where water is not pooled in sufficient quantities to be specular and/or form a lake. As a result, this intermediate between "one-sided" and "two-sided" metrics could apply to both observables.

Reply: We tried the option that the referee suggested, i.e., to "reward" the match between lakes/high-specularity and thawed areas as in the "one-sided" and then to "penalize" a mismatch between lakes/high-specularity and cold areas.

The reward of the match between lakes/high-specularity and thawed areas is the ***warm hit rate*** we defined as the ratio of the number of grid cells with modelled warm bed that have $specularity_5 > 0.4$ to the total number of grids with $specularity_5 > 0.4$.

We name the penalize of mismatch between lakes/high-specularity and cold areas as ***cold miss-fit rate***, which can be defined as the ratio of the number of grid cells with modelled cold bed that have $specularity_5 > 0.4$ to the total number of grids with $specularity_5 > 0.4$.

We call the new metrics that the referee suggested ***Total rate 2***, which is

$$\textbf{\textit{Total rate 2}} = warm\ hit\ rate - cold\ miss\text{-}hit\ rate.$$

For comparison, we call the "two-sided" metrics in the manuscript as ***Total rate 1***,

which is the *average* of warm hit rate and cold hit rate minus the abs(*imbalance*), where *imbalance* reflects the difference between warm hit rate and cold hit rate.

We show ***Total rate 1*** and ***Total rate 2*** in the table below using the 5 GHF datasets.

Table 1. *Warm hit rate*, *cold hit rate*, *cold miss-fit rate*, *Total rate 1* and *Total rate 2* for the modelled results with 5 GHFs. The threshold of $specularity_5$ is taken as 0.4 for warm hit rate, and 0.2 for cold hit rate.

GHF	*Warm hit rate*	*Cold hit rate*	*Cold miss-hit rate*	***Total rate1*** *(original result)*	***Total rate 2***
Martos et al., 2017	0.9750	0.1042	0.019	-0.27	0.956
Shen et al., 2020	0.7250	0.5682	0.256	0.53	0.469
An et al., 2015	0.5563	0.6591	0.425	0.52	0.1313
Shapiro and Ritzwoller, 2004	0.5750	0.6951	0.406	0.54	0.169
Purucker, 2012	0.5375	0.7254	0.444	0.48	0.0935

We can see the ranking of ***Total rate 2*** is the same as that using *Warm hit rate.* Especially for the highest GHF, Martos et al. (2017), the area of modelled warm bed is much larger than modelled cold bed, hence the value of *warm hit rate* is much larger than *cold miss-fit rate,* and plays the dominant role in ***Total rate 2.***

Therefore, the *cold miss-fit rate* did little to "penalize" the mismatch between lakes/high specificity and cold areas. It cannot be serve as an intermediate between "one-sided" and "two-sided" metrics. We prefer to stick on the "two-sided" metrics used in the manuscript.