We thank M. Hoelzle for his constructive comments.

Obvious grammatical or syntax errors have been changed without further comment. Otherwise, all the comments by the referee were considered, and the text in the publication has been adapted accordingly. We have emphasised that the model results are of course the result of assumptions and abstractions, and that their use in, for example, selecting potential measurement sites, must be seen within this context. Further, we have added model runs showing a higher, cold location, where uncertainty with depth varies more as a result of phase changes (i.e. we are near the melting point). And finally, we have made clearer the assumptions and parameters selected in our model.

In the following, answers to the specific comments of M. Hoelzle are given.

Comments M. Hoelzle:

Page 802: The authors choose several ground types for their modeling exercise including clay, sand, silt, peat, gravel and rock writing that liquid water influences the thermal conductivity. I fully agree with this statement. However, dealing with alpine permafrost equally important is the air filled part in the ground materials with high porosity such as coarse blocks, which are covering large parts of alpine permafrost areas.

We fully agree with that statement. Representing coarse blocks such as typically found on rock glaciers is important for modeling permafrost in the Alps. In this setting, we parameterize them with the hydrologic conductivity of gravel and a high porosity. This allows a free drainage of the pore space and the corresponding air content is accounted for in the calculation of ground thermal conductivity that constitutes one element of the importance of coarse blocks for permafrost (Gruber & Hoelzle 2008). The advection of air in blocky surfaces, however, is not included in the model, which is a complex problem that we are not yet in the position to address.

Page 811/812: I strongly disagree with some statements in chapter 5.1. The authors view is, in my opinion, strongly biased by a selective modelers perspective. If the authors are tuning their model to the measurements, the output variability will be reduced automatically and implicitly they assume that the model represent the real physical processes, which is in their case definitively not true. The authors state that 'These findings can in turn inform future measurement campaigns by quantifying the benefit of an individual measurement'. This statement is of major impact and I find it very dangerous because based on this statement the authors claim that they were able to model all important processes. This is certainly not the case here as the authors probably know by themselves neglecting e.g. important processes in ground materials with high porosity containing large amounts of air with several different processes such as air circulation, long-wave radiative heat transfer between blocks etc. not included in their model. As long as a model study is not able to take all processes into account, I find it somehow foolhardy to make such a statement in a paper. I recommend that the authors reformulate this part of the paper taking into account the suggestion above.

We fully agree that exaggerated trust in models for the planning of measurements is dangerous and it has not been our intention to insinuate this. There is however some merit to our argument: Let us consider two situations in which we wish to make measurements to learn about a system. In the first, we have no knowledge of the system investigated and will employ some sort of random sampling. In the second, we know something about the system and tailor our measurements accordingly. Whether or not this process of measurement design is aided by a model (the formulation of our knowledge in computer code) does not make it more or less subjective or dangerous but it may make it more reproducible. To make the possible benefits and caveats clearer we have extended the discussion to: „These findings can inform future measurement campaigns. Model uncertainty (for a given location, time and variable) can be interpreted as one metric for the potential benefit of an individual measurement. It does however not provide information on the correspondence of model results with reality and should therefore be treated with care and as one of several metrics to inform the design of measurement campaigns.“

Specific comments:

1. Page 792, line 9: Why only hydraulic properties? In coarse debris, it is a remarkable percentage of air within the pore spaces. It has to be taken into account!

This was maybe not specified clearly enough in the paper. The pore space includes water and/or air, and is filled according to the saturation of the soil. We added: "… characterized by porosity and hydraulic properties."

2. Page 792, line 13-15: These results are already, since decades known, so this is not a really exciting result of your study. The result of your study is the careful and really well done evaluation of errors and uncertainties evaluated by your model. Therefore, you should concentrate in the abstract on these results!

We agree. We deleted these sentences, and shortened the abstract.

3. Page 793, line 21: Before the authors start writing about model evaluation, it has to be considered that a model is a) a strong abstraction of the real system, b) it is always a simplification and c) it is influenced strongly by the modeler's perception of the system, which is often strongly biased. These effects have to be taken into account when writing about model evaluation, because uncertainties are often influenced by the biased perception of the modeler.

Yes, we agree. We changed the sentence to: „Since every model is an abstraction and simplification of reality, and since model outputs are thus strongly dependent on the modeler's perception of the system, any model must in a first step be evaluated for its fit to an intended purpose \citep{Rykiel1996}“.

4. Page 800, line 3: Please give some references for your applied setting of the thresholds for rain and snow.

Done. We included a reference by Kienzle, 2008.

5. Page 801, line 18: What is the basis of this precipitation correction factor? Please describe this more detailed.

Done. We wrote: „To deal with this systematic measurement error, which has great effects on snow accumulation and soil moisture, GEOtop considers a~\textit{precipitation correction factor}. Hence, all precipitation measurements used as input to the model are multiplied by the correction factor. The value of the correction factor must be assigned before running the model, and may be used for tuning.“

6. Page 801, line 19 to 24: Probably the determination of the fetch distance would be necessary. However, as the mountain topography as well as local surface condition is influencing the local turbulent heat fluxes considerable also the heights of 0.5 to 16 m seems to me somewhat arbitrary chosen. The authors can also select a range between 0.01 and 800 m, which would probably better reflect the range of the real system based on knowledge from balloon soundings in alpine valleys.

Yes, the values of 0.5 and 16 m have been chosen somewhat arbitrarily as to our knowledge, there is no practicable solution for doing so otherwise. To make this clear, we have reformulated: "The height of the sensor at which a temperature or wind speed are measured influences the calculation of the turbulent fluxes. While the exact height of the meteorological station can be measured precisely, the topography of the station in mountain regions may influence the equivalent height with respect to an infinite planar surface (Fig. 3). As a consequence, its determination is partly arbitrary and in this study, the height was varied between 0.5 and 16m."

7. Page 803, line 3: When the authors define ground as a volume below earth surface, then they have to include the part, which is filled with air too. Otherwise their simulations cannot really be compared with alpine permafrost conditions (see general comment above)

Answer: see answer to general comment 1.

8. Page 803, line 17: What do the authors mean with estimated?

We changed "estimated" to „calculated" to make the sentence more clear.

9. Page 804, line 8: What the authors mean with plausible parameter values. Please be more specific.

We changed the sentence to: „The simulations are calibrated with the observations to obtain parameter values that minimize the difference between model outputs and observations."

10. Page 804, line 17: Again, what is a plausible range? Please be more specific.

We wrote: „and varying x_j within values that are physically plausible. The ranges of the parameters are determined based on literature review and/or expert opinion. However, it must be kept in mind, that even though intended to be as objective as possible, the selection of a parameter range is always partly subjective, which influences the results and conclusions that are obtained from the analysis."

11. Page 804, line 18: Here one of the fundamental problems in the study is directly introduced -> parameters are determined based on literature review and/or expert opinion -> this approach will automatically lead to a strongly biased model setting. How can the authors prevent this? I do not have a solution but I would expect that the authors discuss this aspect to show that they are aware of this serious problem.

Yes, please see answer 10.

12. Page 804, line 19: What is a local sensitivity or better what is the contrary of a local sensitivity -> a general sensitivity? Please define this better.

We added a sentence: „Local sensitivities are obtained when each parameter is varied separately and all other are kept fixed. This procedure contrasts to global sensitivities where all parameters are changed simultaneously \citep[e.g.,][]{Saltelli2004, Saltelli2008}, "

13. Page 805, line 3: The authors write that 'all parameters are assumed independent'. I do not really understand this assumption, because the authors are of course aware that this is certainly never true! May they can support this assumption by a more fundamental explanation, because writing that the study setting is 'synthetic, spatial autocorrelation of the parameters are therefore not taken into account' is not enough to justify this approach. Especially not if the authors at the end of the paper *want to justify that they model study can be used for an improved measurement concept in the field.*

We assume that the uncertainties in the parameters do not correlate with each other. Spatial autocorrelations of the parameter uncertainties are not taken into account since, as we mentioned, the setting is synthetic and run at the point scale, and not on a real DHM. Taking spatial autocorrelation into account would not modify the results in such a setting. However, if run in a three dimensional mode (solving a three dimensional heat condution, autocorrelations of individual parameters would have to be assessed preliminarily based on literature. Otherwise, the uncertainty estimations would lead to erroneous results. However, since here our point are all independent, this does not affect our results.

14. Page 811, line 24: I strongly disagree with this statement (see general comment above)

Please refer to answer to general comment.

15. Page 813, line 19: I strongly agree with the statement that it is important to evaluate individual processes separately if used in impact models. I would like to remember that the authors would have had the opportunity to do such independent evaluations, because at their research site 16 years of energy balance measurements exist to perform such an independent evaluation! Such evaluations were already performed in the past by Stocker-Mittaz et al. (2002), which is maybe worth to mention here.

Yes, thank you for the idea. We include this reference in the paper.

16. Page 814, line 10 to 18: Yes, again I strongly agree but see comments above.

See answers to general comments.

17. Page 814, line 23&24: The authors state ‘ Parametric uncertainty of MAGT at different depth is almost constant’. Is this result not simply caused by the assumptions made by the authors for the model definition?

The reason for that is that modeled MAGT is determined by the ground surface temperature, and hence uncertainty in GST propagates down into the ground. However, we added a new Figure 11 for a northfacing point at 3500m, where we can see that temperatures (and their uncertainties) at greater depth are influenced by cold conditions and e.g. the phase change. At temperatures close to phase change, the uncertainty between different depths varies strongly. We have also added a statement to this effect in the paper.

18. Page 815, line 9 to 11: In general, I agree with this statement. However, if you miss some processes in the model, the model is producing nonsense although with the help of your measurements and the corresponding tuning of the model produces good results!

Yes, you are right. But as you also stated, any model is wrong (since it is a simplification of reality), so we think we have to keep this restriction in mind, and still try to learn most of what we can do at the moment.

19. Page 815, line 25: see comment 15.

See answer 15.

20. Page 816, line 10: I would add to this sentence the following ‘after an in-depth evaluation of all *processes in the field'.*

Done, thank you.

Technical comments:

1. Page 792, line 7: uncertainties instead of uncertainty

Changed.

2. Page 796, line 22: in general the use of expression is global radiation or total shortwave radiation instead of global shortwave radiation.

Changed.

3. Page 833, Fig. 5, Page 835, Fig. 7, Page 836, Fig. 8: The labeling of the axis in these Figures are too small.

The labeling was enlarged.

Gruber, S. & Hoelzle, M. (2008): The cooling effect of coarse blocks revisited, *Proceedings of the 9th International Conference on Permafrost 2008*, Fairbanks, Alaska, USA, 557–561.