Reviewer 1:

We thank the reviewer for her/his comments and advice which we have implemented in a revised version of the manuscript. In the following, we give a point-by point reply to the points raised (in bold):

This article is clearly presented and well written. The modelling technique appears to be effective at capturing small-scale variability in ground temperatures. I would recommend it for publication in this journal, however, I would suggest that significant improvements should be made beforehand.

For the introduction and methods, I have only minor comments:

Line 20-21 on page 3909, "which is of similar order of magnitude as the" should be "which is of a similar order of magnitude to the".

done

Line 8 on page 3910 "improved predictions on" -> "improved predictions of"

done

In Section 3.1 (first part) you should mention that only vertical heat fluxes are simulated (if this is indeed the case?). It would also be useful to know the temporal resolution of driving data that CryoGrid2 requires.

CyroGrid 2 is a 1D-model and indeed only simulates vertical heat fluxes, as mentioned in 3.1. We have inserted a clarifying sentence under 3. Driving data sets: As driving data sets for CryoGrid 2, we use gridded data sets of daily average air temperature and snow depth which are obtained from a downscaling scheme and a snow redistribution model (Sects. 3.4, 3.5).

Later in Section 3.1 (under "ground properties"), page 3918 line 12 says saturated conditions are assumed "except for fell", but in Table 1 it seems saturated conditions are assumed for all including fell?

The modeled active layer is significantly shallower than 3m, below which saturated conditions are assumed. We have clarified this sentence accordingly: "…below the current active layer for all classes, except for fell for which no in-situ data are available and saturated conditions are assumed below a depth of 3m."

In Section 3.3, the final paragraph discusses how the snow modelling was validated. However there are no numbers or plots. I would strongly recommend giving more quantitative information here, ideally a plot. (For example, Figure 2 supports the nfactor scheme well.)

We have inserted the required quantitative information in the revised version, including a graph showing measured vs. modeled melt-out dates: "The modeled snow depths and the timing of snowmelt were validated against automated snow depth measurements acquired near ZeroCalm 1, i.e. close to the ZEROline. Linear regression analyses showed that the modeled snow depth represented 77-97% of the variability in the observed snow depth in five of the seven hydrological years and approximately 47% in two years (2004-2005 and 2008-2009). The latter offsets were caused by earlier modeled snowfall the valley in the autumn than in reality due to monthly applied lapse rates in SnowModel/MicroMet during these autumn

months. Additionally, SnowModel/ MicroMet represented the timing of snowmelt with on average ±4 days, while the maximum deviation was 8 days (Fig. 4). Furthermore, a visual comparison between modeled and manually observed snow depth collected in 2005-2010 in two areas of 100 x 100m along ZERO-line (ZeroCalm 1 and 2) showed correspondence in the spatial snow depth distributions. Finally, the modeled melt-out dates were validated for ZERO-line by comparing to orthorectified images taken by an automatic camera system at 400ma.s.l. at the slope a mountain overlooking ZERO-line (Hinkler et al., 2002) for the years 2006 to 2009 at a resolution of 5m. From grayscale images, the presence or absence of snow was determined using a simple threshold filter which was adapted for each year. In case of missing images due to clouds in front of the camera, the date of the snow melt was set to the midpoint between the last snow-covered and the first snow-free date. The results confirm the results from the comparison to point observations: in 2006, the deviation of the melt-out dates between measurements and SnowModel/MicroMet results was 0.0±8.6 days, -1.8±5.6 days in 2007, 0.7±8.2 days in 2008 and 5.4±6.0 days in 2009. The melt-out date is therefore represented within one week for most grid cells, but larger deviations can occur for a number of grid cells. Note that cloudy periods with no images of up to four days lead to an uncertainty of several days in the determination of the snowmelt date for some years and pixels. Furthermore, Hinkler et al. (2002) suggest an absolute referencing error of about 10m for each pixel which also contributes to a reduced match between images and model results."

In Section 3.4, most of the steps seem to refer only to air temperature? Is the same procedure used for snow depth? Please clarify this. It would also be helpful to clarify the random sampling procedure. It is not clear whether the offsets are taken from a whole year (from 2003-2010) for each past/future year, or the year from which offsets are taken changes every month. It also also not clear whether offsets are used only from one time in 2003-2010 for each correction factor, or whether a mean offset for several (randomly selected) years is calculated?
We have significantly expanded this section and added clarifications in the revised version.

In Section 3.5, page 3923 line 12, "date" -> "data". Also line 21, remove "be".

Sect. 3.5 has been removed.

For the remaining Sections (results, discussion, conclusions) more significant improvements are recommended.

For Figure 5, it would be useful to show the range of values for each vegetation class, similar to Figures 3/4/6. In fact you could even combine Figures 3 and 5, showing the observations on only the early part of the graph, and with a shaded band showing the range of simulated values. Combining the graphs may not work, since they are discussed in different places in the text. Please consider this but use your own judgement.

We have added the modeled ranges to Fig. 5. Since the CALM sites comprise a lower number of grid cells compared to ZERO-line and the NDVI values are fairly homogeneous (and also the snow cover in most years), the spatial variability is not large (which is the reason this was left out in the previous version). However, we agree that the figures are more consistent this way.

In Section 4.1 (under Active layer thickness) you state that "CryoGrid2 can capture the significant differences between the three sediment classes Dryas, Cassiope and wetland caused by the different soil moisture contents". Is there no other difference between these classes in the model? Such as the NDVI factors, mineral soil composition? The statement that it is due only to the moisture contents does not appear to be fully justified.

We agree, this statement is misleading. We have replaced it by "…caused by different ground and surface properties".

Page 3925, "the biological activity ... are" -> "the biological activity ... is"

corrected

The list of uncertainties (Section 5.2) is a nice, concise list. However, it would be helpful to make it clearer which is likely to be the greatest uncertainty. Perhaps this could be achieved by considering order of magnitude estimates from the previous studies that you cite.

This is a tricky point which would in principle require a sensitivity study for all factors (including the GCM/RCM results) which in turn would require knowing plausible ranges for the different parameters or driving data sets. In the revised version, we have conducted a sensitivity analysis for snow depth and ground and surface properties, which are both determined by the NDVI (see Sect. 4.2 in the revised version, and below for more details). This shows that ground temperatures are most sensitive to snow depth, while thaw depths are most sensitive to NDVI (i.e. ground and surface properties). To back up these results, we cite the results of another sensitivity study by Langer et al. (2013), with similar findings for a site in Siberia: they showed that the snow properties give rise the largest uncertainty concerning ground temperatures, while the ground properties related to the ground stratigraphy are the crucial source for thaw depth.

In the discussion and conclusions some claims are not fully supported by the work. Namely the final paragraph where you claim that GCM's "are not capable of correctly predicting the onset of permafrost thaw". Your study has shown that your technique can predict the onset of permafrost thaw, but it has not shown anything about any other model. It may be that large-scale models can predict it via some parametrization or grid-box mean - at least you have not made any justification that they cannot. In 5.3 you mention "a simple increase of the spatial resolution seems a prerequisite to resolve such shortcomings", but there is no discussion of how this could be achieved on a large scale. I suggest that you include more discussion of the possible applications of your techniques and emphasise how they can be useful for large-scale modelling,if you wish to address shortcomings of large-scale models. It may be that your highresolution work could inform parameterizations of sub-grid-scale processes in GCM's?

The crucial point of the two sections is to highlight the significance of spatial variability at scales smaller than typical model grid cells (both for large-scale GCM modeling and for more dedicated permafrost modeling). In 5.3 we do not make any reference to GCM modeling, only dedicated permafrost modeling studies are cited. We have added two sentences on how statistical methods may be used to reach a representation of subgrid variability and thus a refined "effective grid size": "For modeling of large spatial domains, a grid cell size of 10m is generally not feasible due to computation power. Statistical representations of small-scale variability are a promising approach to overcome this problem, as recently explored by Fiddes et al. 2013 and Gisnås et al. 2014."

Since this work is not primarily related to permafrost representation in GCMs, we have removed the potentially misleading remark "as they are typical for the land-surface schemes of atmospheric modes" from the final part of the Conclusion.

Finally a major improvement that I suggest for this work is some more analysis of the results. You have included spatial variability of several factors (listed nicely in the beginning of Section 3), but there is no analysis of which of these factors is the most important to take into account -

to which are the soil temperatures most sensitive? You claim briefly in Section 5.3 that the spatial variability of ground temperatures is caused mainly by the snow depth, but there is no evidence of this presented in the paper. It would be very interesting if there were some plots actually along the transect. If, for example, you plotted mean snow depths and soil temperatures along the transect (and other variables too for comparison), it may be clear that the snow has the biggest influence. Or you could compute the strength of correlation with soil temperature and snow depth. I would leave it up to you how you show the influence of each factor, but I must stress that it would add a lot of value to the paper from the point of view of a modeller, who may wish to know what the key aspects are that they should first consider.

To provide a methodologically sound assessment of the different factors, we have conducted a sensitivity analysis by running the simulations with constant ground and surface properties (i.e. for each class with associated ground stratigraphy and typical NDVI value). In Sect. 4.2, the following text has been inserted: "In order to investigate the sources for this spatial variability, a sensitivity analysis was performed by running CryoGrid 2 for ZERO-line with a uniform ground stratigraphy and associated characteristic NDVI value (Sect. 3.1) for each of the four stratigraphic classes Fell, Dryas, Cassiope and wetland. This analysis suggests that snow depth has the largest effect on 1m ground temperatures, with a variability 3-5 times larger than the variability caused by ground and surface properties. On the other hand, modeled maximum thaw depths are much more influenced by ground and surface properties than by snow depths which only lead to differences on the order of 0.1 to 0.2m, compared to differences of more than 0.5m for different stratigraphic classes/NDVI values. A statistically significant correlation between NDVI values (and thus stratigraphic classes) and snow depths modeled by SnowModel/MicroMet does not exist in the employed data set."

I hope these comments have been useful, please let me know if it doesn't make sense or if you disagree. I enjoyed reading your paper and look forward to reading an improved version.

Best wishes.

Thank you very much!

On behalf of the authors,

Sebastian Westermann