We are grateful for this detailed review and the thoughtful suggestions for improvements. In the following, a point-by-point reply to all comments and questions is given. Our answers are marked in bold.

On behalf of all authors,

Sebastian Westermann

Anonymous Referee #2

Westermann et al. present a modelling scheme based on MODIS acquired LST and ERA-Interim reanalysis products in order to derive MAGST at 1 km at a continental scale. Subgrid variability is addressed by computing a distribution of MAGST's using a simple equilibrium model (CryoGrid 1) for each grid-cell. The approach is applied to approximately 5 million km2 in the North-Atlantic region. The approach is evaluated against a network of 143 boreholes (IPA 2010) and suggests a model accuracy > 2.5degC. The probabilistic approach allows a classification of each grid cell as continuous, discontinuous and sporadic permafrost.

In general I found the manuscript to be an interesting approach (and enjoyable read) that could be an important contribution to large area permafrost modelling and would recommend publishing after consideration of the following suggestions and comments.

MAIN COMMENTS

1. TOPOGRAPHY AND SURFACE COVER

The scheme accounts for subgrid variability by computing a range of physically plausible nf (roughly the offset between GST and surface T generally caused by snowpack) and rk (roughly the offset between GST and temperature at depth) parameter values. You mention rk to be primarily dependent upon water content of the soil but also of significance is the thermal properties of the surface cover itself – blocky, vegetation, fine grained material etc. Is this somehow considered in computing variability?

We agree, this would be important. At this stage, however, it can in practice not be included in such modeling approaches, since reliable global maps/ classifications of such phenomena are lacking. Therefore, we restricted ourselves to only two simple classes.

We have inserted a paragraph under Sect. 2.7, which clarifies such restrictions: "The rk values are chosen close to 1 for both classes, which implies that large changes in the thermal conductivities do not occur. Areas with a potentially large thermal offsets (i.e. low rk values), such as wetlands or extensive block field areas (Gruber and Haeberli, 2007), are clearly not represented by this choice. However, with the currently available land cover products, it is not possible to reliably detect such areas (see Sect. 2.7), so that they cannot be accounted for by the employed scheme.

Another very important source of subgrid variability, even at the 1 km scale is topography, particularly elevation and aspect (although reduced influence in northern climates). It would be nice if this could somehow be considered in some kind of ruggedness index as obviously this effect on uncertainty is far greater in a steeper region of coastal Greenland compared to flat regions of the Russian Arctic.

We also agree with this comment, but there are is a conceptual problem with including topography and exposition: The applied FDD and TDD is a composite of MODIS LST, i.e. LST values reflecting the true thermal surface temperature conditions of a certain pixel, and downscaled ERA reanalysis. For the latter, including topographical variations would be straight-forward, as the employed DEM provides a standard deviation of altitudes for each pixel. For the LST measurements, however, there is no accepted procedure to obtain the variation within a pixel. The signal received at the sensor is an average one, which is a result of the distributions of exposition, altitude, shaded areas, surface albedo, etc. within the footprint area. Simply applying an atmospheric lapse rate to estimate the spread would not do justice to these many factors, especially considering that the measurements are taken during clear-sky conditions. Moreover, in really steep topography, the satellite footprints can vary strongly from scene to scene depending on the respective satellite orientation and view angle, so that MODIS LST is of limited use in high mountain areas, such as the European Alps.

Accounting for the topographic variations only for the ERA reanalysis-derived values is conceptually not satisfactory, since their fraction is highly variable throughout the study area (Fig. 1), and the "weight" of the topographic variations would basically depend on the abundance of clear-sky conditions.

We have therefore chosen not to include topographic variations as a factor in the model approach and rather clearly state this as a limitation. We believe that the presented modeling cannot be more than a first-order approximation of the permafrost distribution in high mountain areas with strong topographic variations and that its main potential lies in areas with rather gentle topography. However, this is the case in the large part of the study area. Of all grid cells with modeled negative mean ground temperatures, only about 2-3% feature a standard deviation of the altitudes larger than 75m (according to the employed DEM), and even in countries like Norway, the majority of the permafrost occurs in high-lying plains rather than in jagged mountain areas.

We have included the following statement within Sect. 2.7: "Furthermore, variations of altitude and exposition within grid cells are an additional source of spatial variability of ground temperatures which is not accounted for in the presented model scheme. In mountain areas with strong topographic variations, spatial variability of ground temperatures within a grid cell is most likely underestimated."

Furthermore, we added a statement to the discussion in Sect. 4.3: "Furthermore, in mountain areas with strong topographic variations, satellite-based LST measurements at 1 km scale cannot sufficiently capture variations of altitude and exposition, so that

the modeling scheme at best can be expected to deliver a first-order approximation of the permafrost distribution."

2. UNCERTAINTY OF ERA-INTERIM AIR TEMPERATURE FIELD

You address some of the downscaling issues with using a coarse scale product such as ERA-Interim and I agree with the simple approach for your application. However, you don't discuss the spatial variability of bias in a reanalysis product such as ERA-Interim which rely on stations / upper air measurements to constrain the weather model. Such variability could lead to strong differences in regional patterns of bias. Some references (if available) on the performance of ERA-Interim in the North could be useful. You mention the finding that model levels below grid level do not yield as good results and in general this was found to be related to poor representation of the surface boundary layer. One significant effect of this that is worth mentioning is that valley inversions are not captured in the temperature field.

In the revised version, we refer to two studies, which have independently validated the performance of the ERA-interim analysis for areas adjacent to our study region. However, these are only unsystematic snapshots, and we are not aware of a conclusive study on regional biases of near-surface air temperatures in the study region. Furthermore, we have added a statement on the limitations of the downscaling procedure, and possible ways to overcome such differences: "We emphasize that the procedure cannot account for many regional and local climate and weather conditions, such as temperature inversions in valley systems. Downscaling of reanalysis data using e.g. the Weather Research and Forecasting (WRF) Model, though computationally expensive, may be a way to overcome such difficulties (Aas et al., 2015)."

Ref: Schanke Aas, K., Berntsen, T., Boike, J., Etzelmüller, B., Kristjánsson, J., Maturilli, M., Schuler, T., Stordal, F., Westermann, S.: A comparison between simulated and observed surface energy balance at the Svalbard archipelago, Journal of Applied Meteorology and Climatology, in print, doi: http://dx.doi.org/10.1175/JAMC-D-14-0080.1, 2015.

3. UNCERTAINTY OF ERA-INTERIM SNOW DEPTH

There are well documented biases in reanalysis precipitation fields (eg Schmidli et al 2006) and such coarse resolution of ERA-Interim will never capture the variability of snowdepth in complex topography which in turn has a large effect on spatial patterns of the ground thermal regime. I think these possible sources of bias should be mentioned in the scheme evaluation. Also, how are snowdepths at each grid cell used to determine the range of nf factors? I don't see this described anywhere.

We have included the reference in 2.5, and mentioned documented biases. As detailed in 2.5, the aim of using ERA snowfall is to capture large-scale precipitation patterns, and adjust the ranges of nf-factors, so that the true range is contained within. This means that in an area with high snowfall, such as the W coast of Norway, the minimum nf should be significantly lower compared to an area with low snowfall, such as N

Greenland. In the revised version, this is explained in more detail in Sect. 2.7, where we describe how nf is determined from ERA snowfall.

4. UNCERTAINTY OF MODIS LST

You mention in Section 2.2 the seasonal average cold biases of up to 3K and therefore question the reliability of MODIS. To overcome this problem you say you have a composite product of reanalysis and MODIS. However, the reanalysis is used to gap fill cloudy days. Does this mean the 3K bias is due to cloudiness or some other effects? It would be good to be clear how the composite directly addresses the bias in MODIS and is not just a gap filling strategy.

Yes, the bias is mainly due to cloudiness, as shown in Westermann et al., (2012) and Østby et al. (2014). We have modified the respective passage to better explain this: "This is in particular attributed to prolonged cloudy periods with systematically different average surface temperatures that are not captured by the satellite measurements. Furthermore, detection of clouds is imperfect in particular during polar night conditions (Liu et al., 2004), so that cloud top temperatures are contained in the MODIS LST time series. Validation studies have shown that these effects can lead to a systematic cold-bias of up to 3 K in seasonal averages (Westermann et al., 2012; Østby et al., 2014).

6. CLOUDINESS

How is cloudiness detected and subsequently MODIS scenes rejected? Some details on methods, thresholds used would be useful.

Cloudy regions are automatically masked out by the MODIS cloud detection, so that such data are not contained in the employed MODIS LST products. An explanatory sentence has been inserted: "Cloudy regions are automatically detected and removed by the MODIS cloud mask (Frey et al., 2008)".

7. BIAS

There appears to be a cold bias in "North America" (Figure 3) compared to "Nordic" and "Russia". Do you have any suggestions why?

It is already stated in the manuscript that there is a cold-bias for the boreholes in N America. Possible reasons are: 1) the comparatively small number of boreholes located only at a few different sites, which makes it an unrepresentative selection; this would at least leave the possibility open that the bias in reality does not exist; 2) all boreholes are located in a high-Arctic setting, where the performance of the employed reanalysis may be less good – a similar bias may well occur e.g. in N Greenland, where there are no boreholes; 3) the employed parameterizations in particular for nf may be biased for the high-Arctic setting.

Since it is not entirely clear if there really is such a regional bias, we do not wish to speculate about the potential reasons. We have inserted a sentence in Sect. 3.1: "…,

while there is a slight cold-bias for the available boreholes in N America. However, the latter cannot be fully secured due to the comparatively weak data basis of only 12 boreholes, all of which are assigned large model standard deviations of 1.9 to 2.7 degree C (see Supplement) which indicate a potentially large spatial variability of ground temperatures."

8. VISUALIZATION OF UNCERTAINTY

Linking back to comment 1, among other factors – key sources of uncertainty would be related to ruggedness of topography in a grid cell. Would it be possible to devise an uncertainty map based around a topographic index?

As outlined above, this cannot be done in a satisfactory way since the footprint of a MODIS satellite pixel is to some extent variable, which causes problems exactly in areas with high roughness. The uncertainty will depend on the exact nature of this footprint area.

Another figure I would like to see would be a spatial representation of the variability of modelled values for each grid-cell. This would again give some insight into regional patterns of uncertainty at least with respect to the perturbed model variables. You could also overlay bias point values on this map from the borehole-modelled results.

A new Figure 5 visualizing the standard deviation of all model realizations has been inserted.

TECHNICAL COMMENTS

1. *P754 l6: on continental scale > at continental scales.*

Done

2. *P755 l16:18: reads a bit like this sentence has just been thrown in there without too much context.*

Changed to "In order to improve such shortcomings, Gruber (2012) recently derived a global high-resolution data set…"

3. *P755 l25:26: which SWE product do you refer too and why do you not use instead of ERA-Interimsnow-depth.*

In the Langer et al.-study, GlobSnow SWE at a spatial scale of 25km is used, which has a significantly higher resolution compared to ERA-interim snow depth. We have inserted more information in the text.

4. *P756 l5 Schmid et al. 2012 could be a good reference here too.*

Reference inserted

5. *P756 l18: on continental scale > at continental scales.*

Done

6. *P756 l18: factor > variable?*

We agree, done

7. *Section 2.3: why was ERA-Interim chosen?*

Our analysis was performed with ERA-interim, and we have not tested any other reanalysis data set, which we consider beyond the scope of the presented study. We have added an additional sentence on ERA-interim to Sect. 2.3, and paragraphs discussing the uncertainty of ERA-interim to Sect. 4.1.2.

8. *You mention SST and upper-air soundings but surface stations are a significant source of data.*

Inserted

9. *P762 l6 round > ground*

Done, thanks!

10. *P762 l19 remove 'a'*

Done, thanks!

11. *P765 l10 "regime on <a> continental scale"*

done

12. *P769 l8 a proxy of temperature is measured and converted to LST via an algorithm*

13. *P769 l7:8 You mention that the strength of LST is to provide an actual measurement – perhaps its worth moderating that statement as there is a lot of algorithm going into deriving the final LST value. In that sense you could argue MODIS LST is closer to a modelled value. This separation is of course not black and white - but I think you would not classify MODIS LST similarly to a sensor directly at the ground surface.*

We agree, this separation is not black and white. However, we argue that satellite LST measurements are significantly closer to surface observations than to a coarse-scale atmospheric model (even when downscaled). The main difference is that reanalyses use the implemented model physics to compute e.g. skin or air temperatures from sparse observational data sets that are generally acquired FAR from the point of interest. However, the MODIS LST algorithm relies on an observational data set (i.e. received radiances in a whole range of spectral bands) DIRECTLY from the point of interest. This implies that only information taken at or at least close to the actual spatial and temporal resolution is employed to derive the LST value. So, if there is e.g. a forest fire or a volcanic eruption, the satellite measurement will see the increased temperature, but

the model will not, since neither the fire nor the volcano are part of its model physics. If a large Arctic river adverts heat and warms a narrow area adjacent to its banks, the satellite measurements will detect it, but the model will not. If there is a temperature inversion in a clear-sky night in a 2km wide valley system, it is clearly visible in satellite product, but not in a coarse-scale model. We agree that the underlying set of algorithms is complicated and prone to error, but many state-of-the-art measurement systems and devices (eddy covariance systems, for instance) share this characteristics.

We have reformulated the statement to be more neutral: "This approach combines the strengths of both products - the capacity of satellite sensors to provide actual measurements of the footprint area at a high spatial resolution, and the dense, gap-free time series of the reanalysis product."

REFERENCES

Schmidli, Jürg, Christoph Frei, and Pier Luigi Vidale. "Downscaling from GCM Precipitation: A Benchmark for Dynamical and Statistical Downscaling Methods." International Journal of Climatology 26, no. 5 (April 1, 2006): 679–89. doi:10.1002/joc.1287.

Schmid, M.-O., S. Gubler, J. Fiddes, and S. Gruber. "Inferring Snowpack Ripening and Melt-out from Distributed Measurements of near-Surface Ground Temperatures." The Cryosphere 6, no. 5 (October 2012): 1127–39. doi:10.5194/tc-6-1127-2012.

Additional corrections:

There has been an error in the derivation of the TTOP equation in the Appendix. The way it was defined, it is incorrect to set MAGT=T_{TOP}, since the latter refers to the winter temperature at the bottom of the active layer. MAGT is the weighted average of winter and summer temperatures at the bottom of the active layer. This has been corrected by modifying Eq. A1, and adding Eq. A2.