Overview

We thank the reviewers for their constructive comments and suggestions. We have revised the manuscript throughout in response to the reviewers' comments. For reference, our response to comment "m" by reviewer "n" is labeled "R[n]C[m]").

Throughout the discussion below, the text is colored as follows:

Black: Reviewer comment
Blue: Author response
Red: Text newly inserted into the revised manuscript
Green: Text already inserted into the manuscript during the previous round of revisions but missed by the reviewers.

Reviewer #1

The problem with the not yet published study by Chen et al. (always cited as Chen et al. 2018 instead of in review) remains. Large parts of the manuscript are based on it. The Cryosphere journal rule is ccording to the webpage 'Works cited in a manuscript should be accepted for publication or published already.' Results etc. referring to it should be removed.

R1C1: The study by Chen et al. (2019) has been accepted and is in press. We have updated the citation to this paper.

Other
Table 4 - state source of observation in caption

R1C2: We added the citation to the observation (Brown et al., 2002).

> "Table 4 – Evaluation results for simulated permafrost extent against the permafrost map by Brown et al. (2002)." (Page 45, line 1 in the revised manuscript)

Figure 8 - properly cite Brown et al.

R1C3: We have properly cited this reference.

> "The red boundary outlines the continuous and discontinuous permafrost regions according to Brown et al. (2002)." (Page 53, line 4 – 5 in the revised manuscript)

Reviewer #2

This study uses point measurement and airborn data in combination with the results of global model driven my MERRA-2 reanalysis modeling data to analyze present permafrost conditions and extent. Authors compare datasets from different scales to study the match between them. The main problem is how to compare in-situ data with averaged to 20x60 m2 grid cell data and then averaged to 81 km2 grid cell. Then authors touch on the problem on why global model unable to model permafrost in the Western Russia and Eastern Canada. Global model fail to model permafrost in those regional because those area represent ecosystem protected permafrost zones (Shur et al., 2007). This means that thick organic layer, most importantly including moss layer, protect permafrost below from warm air temperatures. To achieve this increasing the amount of the organic layer as was also done for example global models like CLM and SiBCASA (Nicolsky et al., 2007; Jafarov and Schaefer 2016) is simply not enough. It is important to drive those regions with cold initial temperatures with enough moss-organic insulation on top. In addition deep soil column should allow keeping permafrost in those regions.

Overall, the paper indicates some important and interesting analysis, including the effect of soil moisture on the ground temperature and ALT. However, current version of the paper need some major clean ups to improve clarity. I suggest cutting the number of Figures, removing discussion from the conclusion and making results and discussion section, since results already have a lot of discussion. Keep the conclusion straight to the point, do not summarize your work in the conclusion. Instead suggest what improvement can be made to improve discrepancies in the ALT simulation in Mongolia, Russian etc. and how the permafrost extent can be better modeled on the global scale.

We sincerely thank the reviewer for the helpful suggestions. We should mention that the reviewer's comments are based on the original submission, rather than the revised manuscript resulting from the previous round of reviews. The revised manuscript included major changes from the original submission, including some changes that were requested by Reviewer #2 here. Below we address the reviewer's comments that still apply to the first revised version of the manuscript and point out below where the reviewer's comments had already been addressed in that revision.

Abstract
L27 …some permafrost areas… Be specific, spell out those areas.

R2C1: The text in question (concerning the trend analysis) was removed following the suggestion by Reviewer #3. See R3C3 and R3C15.

Introduction
P3. L26. I suggest acknowledging all the work done ALT measurement using GPR as a part of the pre-ABoVE campaign. Chen et al., (2016) documented extensive GPR ALT data collection near Toolik Lake, Alaska. Jafarov et al., (2018) documented extensive GPR ALT data collection near Barrow, Alaska. These datasets a unique because they represent spatial ALT collection in oppose to point measurements by CALM. Both dataset available for download from ABoVE website.

These datasets can be extremely useful in this study because they give a better idea on spatial variability of the ALT on meter scale. The standard deviation from those works can be used to better constrain the uncertainty in measured ALT at a finer spatial scale. In addition, I highly suggest checking the most recent and the most complete work on the nearsurface permafrost data in Alaska (Wang et al., 2018). The data collected in that dataset provides a wider coverage for Alaska and can be extremely useful for this study.

R2C2: We appreciate the reviewer for pointing out these datasets. We now point to other remote sensing-based ALT retrievals, including the GPR ALT data:

> "Due to the sparsity of in-situ measurements at the regional to global scale, evaluating the spatial pattern of ALT produced by any such simulation remains challenging. Indeed, it is difficult to compare the simulated values at model resolutions with in-situ observations taken at the point scale unless the measurement point is uniformly representative of the area covered by the model grid cell or the representation errors associated with the point-to-grid comparison are well defined. Remotely sensed permafrost products, which provide a unique source of spatially distributed ALT at the landscape-scale, may provide help in this regard. Existing remote sensing ALT products have been retrieved from ground-based Ground Penetrating Radar (GPR) (Chen et al., 2016a; Jafarov et al., 2017), airborne polarimetric Synthetic Aperture Radar (SAR), and spaceborne interferometric SAR (Liu et al., 2012; Li et al., 2015; Schaefer et al., 2015). These ALT products are available at the landscape-scale and can complement our modelling analysis. In this study, we use remote sensing information from the NASA Airborne Microwave Observatory of Subcanopy and Subsurface (AirMOSS) mission. In 2015, AirMOSS acquired P-band (420-440 MHz) SAR observations over portions of northern Alaska from which Chen et al. (2019) retrieved regional estimates of ALT and soil layer dielectric properties that are related to soil moisture and freeze/thaw states. In their study, Chen et al. (2019) mainly focus on the development and improvement of the ALT retrieval algorithm, whereas the present study uses the ALT retrievals in combination with in-situ measurements to aid in assessing the (fully independent) ALT simulations." (Page 4, line 21 – 25, and page 5, line 1 -10 in the revised manuscript)

Given the length of the paper and the fact that the reviewer also suggested cutting back the number of figures and the discussion, we decided to leave the use of the GPR data for future work. In fact, we are already working on just such a study.

Regarding the observations reported by (Wang et al., 2018), we used the permafrost sites from the GI-UAF in our previous study (Tao et al., 2017). We will assemble more measurements from USGS and NPS sites as included in Wang et al. (2018), as well as other remotely sensed ALT retrievals the reviewer pointed out in our future study to further evaluate and analyze model uncertainty.

P4. L 22-30. Do this freeze-thaw formulation allows multiple thaw zones? E.g. talik and seasonal frost above with the existing permafrost at a deeper depth.

R2C3: The equation implicitly allows for multiple thawing zones in the vertical dimension. However, considering the size of the model grid cell (81 km^2), it is very unlikely the talik zone can be represented. The seasonal frost above the existing permafrost at a deeper depth, however, can be well represented in the model. In fact, seasonal frost is taken into account when determining ALT.

We added one sentence here to make this point.

> “We search for the deepest l if multiple thawed-to-frozen transitions are present (e.g., if a seasonal frost at the surface is separated from the permafrost below by a thawed soil layer). The annual ALT for a given year, then, is defined as the deepest depth at which a thawed-to-frozen transition occurs within that year. Note that the calculation of equation (1) is made at the scale of a model grid cell, and thus features such as talik are not represented if they occur at sub-grid cell scale.” (Page 7, line 11 – 15 in the revised manuscript)

P5.L12 Not sure why the model was spun up for 180 years? Typically spin up means total equilibrium.

R2C4: To reach equilibrium states for deep soils (down to 13 m), the 180 years’ spin-up procedure is needed (Sapriza-Azuri et al., 2018; Paquin and Sushama, 2015).

Methods section needs some better organization. For example,

1. In-situ to AirMoss comparison
2. In-situ to CLSM comparison

R2C5: This comment no longer applies. We had already reorganized the structure of the manuscript during the previous round of revisions.

P7-8. L30-12. The main point of those two paragraphs is the difference. I suggest plotting the difference between AirMoss and CLSM with 81 km2 resolution, just one Figure instead of ABC. Then it will be clear when they do not match and then discussion can be more focused on the why they do not match.

R2C6: We appreciate the reviewer’s suggestion. We now added an additional panel to show the difference between AirMOSS and CLSM results at 81 km^2 resolution. We also removed the subpanel a), but still kept b) and c) to compare the spatial patterns of AirMOSS retrievals and model results.

P8. Paragraphs 3 and 4. Similarly don’t need Figure 4 AB. In-situ data has smaller uncertainty and variability, when scaled up we average the variability into a one grid cell. The question is what is the uncertainty for CLSM should be, which was answered later in the manuscript by analyzing the effect of different factors (snow, organic layer, soil moisture). If you plot the CLSM uncertainty bars and they intercept with the solid lines then this makes the overall results much better.

R2C7: The revised manuscript from the previous round of revisions is already very different from the original manuscript to which the reviewer's comment refers. For this round of revisions, we deleted Figure 4a and modified Figure 4b and 4c. Indeed, we removed the comparison results for AirMOSS retrievals at the in-situ site scale, since Chen et al. (2019) already evaluated the retrievals at the site scale.

Regarding the CLSM uncertainty, we had already added some relevant discussion on it during the previous round of revisions, i.e., during the open discussion period. We copied some relevant information from the open discussion below. Please refer to the online discussion response for detailed analysis and supplementary figures (https://www.the-cryosphere-discuss.net/tc-2018-119/tc-2018-119-AC2-supplement.pdf and https://www.the-cryosphere-discuss.net/tc-2018-119/tc-2018-119-AC3-supplement.pdf).

> "Test simulations (not shown) with alternative model configurations indicate that increasing the number of soil layers may act to decrease somewhat the simulated ALT, suggesting that our values may be a little overestimated; however, based on results from a new study by Sapriza-Azuri et al.(2018), our use of a no-heat-flux condition at the bottom boundary rather than a dynamic geothermal flux may lead to underestimates of ALT. Such uncertainties should naturally be kept in mind when interpreting our results. Our supplemental simulations (not shown) also suggest that increasing the total modelled soil depth has only a small impact on simulated ALT. Uncertainty in our description of soil organic carbon, i.e., both soil carbon content and vertical carbon distribution, leads to corresponding uncertainty in our ALT simulations. We indeed find a significant improvement in simulated ALT at several Mongolian sites when we arbitrarily impose less total soil carbon content and concentrate less soil carbon in top layers (not shown). Besides the vertical distribution of soil carbon, the vertical variation in other soil hydrological properties (e.g. soil texture and porosity) should also play a significant role since they all affect soil thermal conductivity and heat capacity. In addition, the lack of a necessary organic layer on top of soil column and the related thermal processes is also a major deficiency for the model especially in ecosystem-protected performant regions.
>
> Another issue affecting our ALT comparisons is the climatological representation of vegetation parameters such as LAI used in CLSM. An additional investigation (not shown) revealed large differences between the LAI climatology used in CLSM and more realistic, time-varying, satellite-based LAI products at several Mongolian sites. In addition, while we did exclude from our analyses any measurements that were affected by notable disturbance (e.g., wildfire), the impacts of other potential land changes on ALT, including overgrazing in Mongolia (Sharkhuu and Sharkhuu, 2012; Liu et al., 2013), were not explicitly treated in the model. The model also lacks the vertical advective transport of heat in the subsurface due to downward flowing liquid water, which can significantly affect permafrost thawing (Kane et al., 2001; Rowland et al., 2011; Kurylyk et al., 2014)." (Page 29, line 2 – 21 in the revised manuscript)

P9. L16-30. It mainly depends on the pixel size (grid cell) of the modeled ALT. The authors should think how they can address the overall uncertainty in the global model, and how that uncertainty would change when they compare it with in-situ or AirMoss data.

R2C8: Please see our response above in R2C7.

P14. L6-20. Cite Shur et al., (2007) draw the discussion from that work. Refer to my main comment.

R2C9: We added relevant sentences as shown below relying on the reviewer's comments. (Note that we had already edited this paragraph during the previous round of revisions.)

> "One possible reason is that the permafrost in western Siberia is characterized as an ecosystem-protected permafrost zone (Shur and Jorgenson, 2007) where a thick moss-organic layer (i.e., moss-dominated mires (Anisimov and Reneva, 2006; Anisimov, 2007; Peregon et al., 2009)) protects the permafrost below from thawing under a warm air temperature. This is mainly attributed to the low thermal conductivity of the organic layer in summer, which strongly insulates the permafrost from the warm atmosphere, and the high thermal conductivity of the frozen organic layer in winter, which allows cold temperature penetration from above, provided the snowpack is not too thick (Nicolsky et al., 2007a; Jafarov and Schaefer, 2016). This mechanism is lacking in the current version of CLSM (Tao et al., 2017). Thus, improving the model through a better representation of thermal processes in an organic layer above the soil column in combination with initializing the simulation with a sufficiently cold soil temperature should improve the simulation results. This work is reserved for a future study." (Page 24, line 5 – 14 in the revised manuscript)

We also mentioned this in the abstract.

> "The resulting simulated permafrost distribution across the Northern Hemisphere mostly captures the observed extent of continuous and discontinuous permafrost but misses the ecosystem-protected permafrost zones in western Siberia. " (Page 1, line 16 – 16 in the revised manuscript)

P14. L31. There are many CALM sites within a CLSM grid cell. The variation in CALM sites is a standard deviation (std). Again this deviation is from hand full of sites where the GPR measurement provides a wider range of the possible (std) in Barrow and Toolik Lake regions.

R2C10: We agree with the reviewer about the limited information about CALM sites. We hope to include GPR data in the first author's next manuscript.

P15. L1-3. The soil characteristic in Mongolia might include rocky type environment. In mountain areas the ALT along the south face slopes might be quite deep. I wonder if that might explain the deep ALT in those regions.

R2C11: We thank the reviewer for pointing out this. We checked the soil information at Mongolian sites and did find some useful information, and we accordingly added several sentences to the manuscript:

> "An additional reason for the underestimation of ALT in Mongolia might be a mismatch between the land surface parameter values used in the model and the actual conditions at each site. For instance, detailed soil information (https://www2.gwu.edu/~calm/data/webforms/mg_f.html) indicate that some Mongolian sites have special "rocky" soil types including limestones (e.g., M04), slatestones (e.g., M05), gravelly sand (e.g., M06 and M08), etc. that are not well represented in the model. As another example, sites on south-facing slopes presumably have much deeper ALT than those on slopes with less exposure to the sun, which is not captured by CLSM. This large representative errors of Mongolian sites are clearly illustrated by the standard deviation (although computed only with 3 to 5 measurements) as shown by the error bars in Figure 11a." (Page 26, line 21 – 26, and page 7, line 1 – 5 in the revised manuscript)

P15. L30. Do you think if you drive the model with different reanalysis data (ERA-Interim or similar) it might give you better results?

R2C12: This is possible. Please note that we've deleted the relevant sentence the Reviewer pointed.

Regarding the impacts of using different reanalysis data, we had already added relevant discussion following the previous round of reviews (copied below).

> "Note that some other global models, such as CLM3 and the Community Climate System Model version 3 (CCSM3) as reported in Lawrence et al. (2012), also missed this area of permafrost and that updated versions of these models (i.e., CLM4 and CCSM4) showed improved performance in this regard (Lawrence et al., 2012). Guo et al. (2017) reported underestimated permafrost extent simulated in western Siberia using CLM4.5 driven by three different reanalysis forcings (i.e., CFSR, ERA-I and MERRA), and they showed an improved simulation of permafrost extent in this area when using another reanalysis forcing, the CRUNCEP (Climatic Research Unit - NCEP) (Guo and Wang, 2017). Guimberteau et al. (2018) found similar improvements stemming from the use of CRUNCEP forcing. We leave for further study whether the MERRA-2 forcing data is responsible for the western Siberia deficiency seen in our own results." (Page 24, line 17 – 25 in the revised manuscript)

P16. L19. I would drop unnecessary words phrases like at least to some extent from the text.

R2C13: We removed this section and also the original Figure 13 and 14. See R3C3 and R3C15.

Reviewer #3

The authors run a series of simulations of permafrost dynamics using the MERRA2 reanalysis and the MERRA land model. They compare the results to remotely sensed active layer thickness (ALT) and in situ measurements of ALT. The paper has the potential to be a solid model-data comparison of simulated permafrost dynamics. I recommend acceptance after major revisions.

We sincerely thank the reviewer for the careful review and constructive suggestions. We now have revised our manuscript throughout according to the reviewer's comments.

I have three major comments:

1) The authors need to account for measurement uncertainty when comparing to observations and refine their statistical comparison techniques. I found a number of errors in the statistical comparisons that I identify below.

R3C1: We now provided a supplementary file to explain the measurement uncertainty in detail, and we added more discussion on measurement uncertainty of radar retrievals and representative error of in-situ measurements to better interpret the comparison results throughout the manuscript. See R3C18, R3C21, R3C27, R3C31, R3C33, R3C34, R3C48, and R3C49 for details.

In the comments below, we did not see further, more specific information about "a number of errors in the statistical comparisons." We do address all of the issues that are raised.

2) The authors need to clarify the role of the remote sensing data in this analysis. They spend as much space comparing the remotely sensed ALT with in situ data as with the model. Is the paper a means to validate the model or the remote sensing data?

R3C2: The unique advantage of remotely sensed technique is that it provides spatially distributed ALT at the landscape-scale, well complementing the point-/plot-scale in-situ measurements. The present study is meant, in large part, to compare model-simulated, remotely sensed, and observed in-situ ALT to help understand the limitations in both remote sensing techniques and modeling mechanisms, and explore how remotely sensed ALT could be used to improve modeling skills.

We now deleted all the site-scale comparison between AirMOSS retrievals and in-situ observations since Chen et al. (2019) had already discussed the quality and uncertainty of the retrievals in detail.

3) The authors need to change their spinup procedure or drop the trend analysis. Repeating the full time period for spinup introduces a dynamic response that produces false trends aliased on top of real trends. This pretty much invalidates the trend analysis.

R3C3: The reviewer is correct. Given such considerations, we have removed the trend analysis, including Figures 13 and 14, from the manuscript. It was fortunately not a central part of the study.

I have the following specific comments:

P2L17-20: Reword. This is a runon sentence with two, double nested parenthetical clauses, making it very difficult to understand.

R3C4: We deleted the relevant sentence.

P2L25: State how models are useful. This paragraph emphasizes resolution as a weakness of models.

R3C5: We augmented the sentence with more information. The revised sentence is copied below.

> "Physically-based numerical model simulations are potentially useful for quantifying and understanding these dynamics at large spatial scales; they can also provide insights into associated impacts on the global carbon cycle." (Page 3, line 13 – 14 in the revised manuscript)

High-resolution simulation results of permafrost states at the global scale are necessary but are still lacking. We have modified the relevant sentences in the manuscript regarding resolution to elaborate our focus better:

> "Permafrost dynamics can be modelled, for example, by driving a land surface model (LSM) offline (i.e., uncoupled from an atmospheric model) with meteorological forcing data (including air temperature, radiation, precipitation, etc.) from some credible source. LSMs that have been used to quantify large-scale permafrost patterns (i.e., distributions and thermal states) and their interactions with a warming climate include, for example, the Joint UK Land Environment Simulator (JULES, Dankers et al., 2011), the ORganizing Carbon and Hydrology in Dynamic EcosystEms (ORCHIDEE) - aMeliorated Interactions between Carbon and Temperature (ORCHIDEE-MICT, Guimberteau et al., 2018), the Catchment Land Surface Model (CLSM, Tao et al., 2017), and the Community Land Model (Alexeev et al., 2007; Nicolsky et al., 2007a; Yi et al., 2007; Lawrence and Slater, 2008; Lawrence et al., 2008; Lawrence et al., 2012; Koven et al., 2013; Chadburn et al., 2017; Guo and Wang, 2017). Most of these land models were run at coarse spatial resolutions, e.g., ranging from 0.5° × 0.5° to 1.8° × 3.6° for LSMs participating in the Permafrost Carbon Network (PCN) (Wang et al., 2016a) and from 0.188° × 0.188° to 4.10° × 5° for the models participating in the Coupled Model Intercomparison Project phase 5 (CMIP5) (Koven et al., 2013; https://portal.enes.org/data/enes-model-data/cmip5/resolution).
>
> Differences in the permafrost behaviour simulated with these models reflect model-specific process representations as well as biases associated with different meteorological forcing datasets (Barman and Jain, 2016; Wang et al., 2016a; Wang et al., 2016b; Guo et al., 2017; Guimberteau et al., 2018). Such forcing biases are difficult to avoid given the sparsity of direct observations of meteorological variables in most parts of the high latitudes. Even

> reanalyses, which assimilate a variety of global observations, inevitably have biases in high latitudes due to observation sparsity in cold regions combined with the many challenges of physical process modelling. Nevertheless, despite these issues, permafrost behaviour simulated with LSMs driven offline by reanalysis forcing fields can still be useful for understanding the impacts of climate variability on permafrost. The present paper utilizes this approach. Specifically, we generate here a dataset of Northern Hemisphere permafrost conditions by driving an updated version of NASA's Catchment Land Surface Model (CLSM) with Modern-Era Retrospective Analysis for Research and Applications-2 (MERRA-2; Gelaro et al., 2017) surface meteorological forcing fields for the middle-to-high latitudes across the Northern Hemisphere over the period 1980-2017. We perform the simulations at 81 km2 resolution encompassing permafrost areas in the middle-to-high latitudes of the Northern Hemisphere. This resolution is high relative to most existing modelling studies at the global scale; published simulations at higher resolution are limited to plot scales (e.g., CALM-site scale in Shiklomanov et al. (2010)), landscape scales (e.g., polygonal tundra landscape scale in Kumar et al. (2016)), or regional scales (e.g., 4 km2 in Jafarov et al. (2012) covering Alaska; 1 km2 in Gisnas et al. (2013) covering Norway)." (Page 3, line 14 – 25, and page 4, line 1 - 19 in the revised manuscript)

P3L2-3: This is not difficult. One must account for representation error when comparing a point measurement to the area average of a model pixel.

R3C6: We modified the relevant sentences as copied below. (Note the sentences were moved to another paragraph.)

> "Due to the sparsity of in-situ measurements at the regional to global scale, evaluating the spatial pattern of ALT produced by any such simulation remains challenging. Indeed, it is difficult to compare the simulated values at model resolutions with in-situ observations taken at the point scale unless the measurement point is uniformly representative of the area covered by the model grid cell or the representation errors associated with the point-to-grid comparison are well defined. Remotely sensed permafrost products, which provide a unique source of spatially distributed ALT at the landscape-scale, may provide help in this regard." (Page 4, line 21 – 25, and page 5, line 1 in the revised manuscript)

P3L7: The resolution of these simulations is essentially the same as for many published simulations, so I am not sure this is the best claim to make.

R3C7: We deleted this sentence and modified this paragraph. The relevant sentences are copied above in R3C5.

P4L3: State or described the improved performance.

R3C8: We now deleted all the discussion comparing MERRA with MERRA-2, since we did not conduct simulations forced by MERRA and the sentence here is not very relevant to the scope of this study.

P4L8: State or describe exactly what is inferior.

R3C9: Again, we have deleted the relevant discussion comparing MERRA and MERRA-2.

P4L10-12: Delete. Each reanalysis has strengths and weaknesses and I find it very difficult to believe that one version of MERRA is truly superior to another, especially considering the scarcity of measurements in the Arctic. I have no objection to using MERRA-2, of course, but claiming superiority is not warranted and best deleted from the manuscript.

R3C10: We have deleted this sentence and have reorganized this paragraph.

P4L18-20: Here the authors state they will use the remotely sensed ALT to validate the model, but later they actually validate the remotely sensed ALT against ground observations. This makes the actual purpose of including remotely sensed data unclear in this paper.

R3C11: We have modified relevant sentences and have also added sentences before this paragraph to better articulate our purpose of using remotely sensed ALT in this study. Note we also already deleted the discussion on site-scale comparison between AirMOSS retrievals and in-situ observations which was already included in Chen et al. (2019).

> “Due to the sparsity of in-situ measurements at the regional to global scale, evaluating the spatial pattern of ALT produced by any such simulation remains challenging. Indeed, it is difficult to compare the simulated values at model resolutions with in-situ observations taken at the point scale unless the measurement point is uniformly representative of the area covered by the model grid cell or the representation errors associated with the point-to-grid comparison are well defined. Remotely sensed permafrost products, which provide a unique source of spatially distributed ALT at the landscape-scale, may provide help in this regard. Existing remote sensing ALT products have been retrieved from ground-based Ground Penetrating Radar (GPR) (Chen et al., 2016a; Jafarov et al., 2017), airborne polarimetric Synthetic Aperture Radar (SAR), and spaceborne interferometric SAR (Liu et al., 2012; Li et al., 2015; Schaefer et al., 2015). These ALT products are available at the landscape-scale and can complement our modelling analysis. In this study, we use remote sensing information from the NASA Airborne Microwave Observatory of Subcanopy and Subsurface (AirMOSS) mission. In 2015, AirMOSS acquired P-band (420-440 MHz) SAR observations over portions of northern Alaska from which Chen et al. (2019) retrieved regional estimates of ALT and soil layer dielectric properties that are related to soil moisture and freeze/thaw states. In their study, Chen et al. (2019) mainly focus on the development and improvement of the ALT retrieval algorithm, whereas the present study uses the ALT retrievals in combination with in-situ measurements to aid in assessing the

> (fully independent) ALT simulations." (Page 4, line 21 – 25, and page 5, line 1 -10 in the revised manuscript)
>
> We also now provide an additional reason to include the retrieval results in the introduction:
>
> "As a side benefit, the side-by-side comparison of modelled and remotely sensed ALT estimates is an important first step toward combining this information effectively in future model-data fusion efforts." (Page 5, line 17 – 18 in the revised manuscript)

P5L10: The vertical resolution seems too coarse to simulate ALT. The total depth is fine, but other models typically use much higher resolution to simulate ALT. The authors need to explain why this resolution will work.

R3C12: While our vertical resolution is coarse, we emphasize that our ALT values are not strictly tied to this resolution; we can compute any value of ALT depending on the simulated ice fractions of the model layers. This is clarified by some added text:

> "Equation (1) then expresses that the thawed-to-frozen depth is equal to the bottom depth of the layer l but adjusted upward according to the ice fraction within the partially thawed layer l. The annual ALT for a given year, then, is defined as the maximum thawed-to-frozen depth within that year. This upward adjustment, by the way, allows the thawed-to-frozen depth to be a continuous variable; it is not quantized to the imposed layer depths." (Page 7, line 9 – 11 in the revised manuscript)

Also note that we did test some simulations with increased vertical resolution and total depth during the previous round of revision. Our general conclusions are consistent with other studies in terms of how soil configuration affects permafrost modeling (e.g., Alexeev et al., 2007; Lawrence et al., 2008; Sapriza-Azuri et al., 2018; Nicolsky et al., 2007; Dankers et al., 2011). Please refer to the online discussion response for detailed analysis and supplementary figures (https://www.the-cryosphere-discuss.net/tc-2018-119/tc-2018-119-AC3-supplement.pdf).

We did include some appropriate discussion about this issue in the conclusion section which are copied below.

> "Test simulations (not shown) with alternative model configurations indicate that increasing the number of soil layers may act to decrease somewhat the simulated ALT, suggesting that our values may be a little overestimated; however, based on results from a new study by Sapriza-Azuri et al.(2018), our use of a no-heat-flux condition at the bottom boundary rather than a dynamic geothermal flux may lead to underestimates of ALT. Such uncertainties should naturally be kept in mind when interpreting our results. Our supplemental simulations also suggest that increasing the total modelled soil depth has only a small impact on simulated ALT." (Page 29, line 2 – 7 in the revised manuscript)

In response to reviewer's specific comment about the rationale of this vertical resolution, we added one sentence here as shown below.

> "The soil layer thicknesses increase with depth following a geometric series for consistency with the linear heat diffusion calculation (Koster et al., 2000). A no-heat-flux condition is employed at 13m depth." (Page 6, line 14 – 16 in the revised manuscript)

P5L23: This is a good formulation. Models often use it, but rarely document it.

R3C13: Thanks. (No response required.)

P6L17: A 180 year spinup is adequate for stabilizing soil temperatures, but not for soil carbon. Does this model include dynamic soil carbon pools? If yes, then a spinup of 1000-5000 years is more appropriate.

R3C14: We do not have the dynamic soil carbon pools in this study, as now explicitly stated:

> "This version of CLSM, however, does not include dynamic soil carbon pools." (Page 6, line 24 - 25 in the revised manuscript)

P6L20: The chosen spinup technique pretty much invalidates the trend analysis. The typical response time for soil temperature in a model such as this is 20-30 years, exactly matching the length of the MERRA forcing data. If they had spun up using only 1980-85 MERRA data, then the trend analysis makes sense. I suggest either changing the spinup or dropping the trend analysis.

R3C15: We have followed the reviewer's suggestion and have removed all discussion regarding the trend analysis, which was not central to the paper.

P7L29-30: This means one can use the radar data only where one expects the alt to be less than 60 cm. If the radar cannot penetrate below 60 cm, I question the utility of using it for validation. The authors need to supply a rationale for including it in the study.

R3C16: We understand the reviewer's concern about the sensing depth limitation with the remotely sensed ALT retrievals. Even so, the radar retrievals are practically useful for shallow permafrost (for instance in the North Slope of Alaska). As section 4.1 demonstrated, the AirMOSS ALT retrievals show good performance when compared with in-situ observations excluding the sites with ALT measurements deeper than 60 cm. The ALT retrieval algorithm as documented by Chen et al. (2019) represents the current status of P-band radar technique for monitoring permafrost, and will be gradually improved in the future. Combining remote sensing techniques and land models together is a way to better monitor permafrost dynamics. As the first attempt here, we believe the intercomparison of ALT among model results, remotely sensed retrievals and in-situ observations could provide useful insights to the research community.

The manuscript also discussed the potential use of ALT retrievals in the "5 Conclusion and Discussion" section (as copied below).

> "The remote sensing approach is still relatively new, with many aspects still requiring development. It is important, though, to begin considering the modeling and remote sensing approaches side by side, as both should play important roles in permafrost quantification in the years to come. Indeed, once the science fully develops, joint use of modeling and remote sensing (e.g., through the application of downscaling methods) should allow the generation of more accurate permafrost products at higher resolution." (Page 27, line 22 – 25, and page 28, line 1 - 2 in the revised manuscript)

In addition, as noted above, we include in the introduction this additional side benefit of our inclusion of the retrieval results:

> "As a side benefit, the side-by-side comparison of modelled and remotely sensed ALT estimates is an important first step toward combining this information effectively in future model-data fusion efforts." (Page 5, line 17 – 18 in the revised manuscript)

P8L11: Please identify which site got covered with lava. This is so unusual that you have to tell the reader.

R3C17: The documentation (meta-data) for two sites (R30A and R30B in Kamchatka) specifically notes that they "were covered by lava after Nov 2012 eruption". We have added the two sites into the manuscript.

> "We did not use measurements that were flagged as having been taken too early in the season or under unusual conditions (e.g., after the site was burned or covered with lava, which occurred at sites R30A and R30B in Kamchatka)." (Page 10, line 14 – 16 in the revised manuscript)

P8L28: The section on comparison with the radar ALT must include uncertainty. The best that a model can do is match the observations within uncertainty.

R3C18: We provided a supplementary to explain the measurement uncertainty of AirMOSS retrievals. We then added relevant sentences discussing the radar ALT uncertainty and how the model results compare with the radar retrievals.

> "We employed the strategy of Schaefer et al. (2015) to handle the uncertainty propagation, i.e., adding in quadrature the uncertainty components from each scale/level involved (see the supplementary file for a detailed description). For AirMOSS retrievals, the sampling uncertainty of mean ALT at the 81 km2 model grid-cell scale is negligible given the large sampling size and the fact that the retrieval uncertainty dominates the overall uncertainty (see supplementary file). Here, we use a nominal estimate of 0.15 m to represent the AirMOSS uncertainty (i.e., the average of the lower and upper bound of the actual retrieval uncertainty for individual radar pixels as discussed by Chen et al. (2019))." (Page 12, line 13 – 18 in the revised manuscript)

> “Figures 4c further shows that the CLSM-simulated ALT agrees well with the AirMOSS ALT retrievals to within the measurement uncertainty of 0.15 m at all the site-located model grid cells. Indeed as Figure 3c illustrated, the differences between simulated ALT and the AirMOSS retrievals over all the transects examined here are generally below the measurement uncertainty of 0.15 m.” (Page 18, line 19 – 22 in the revised manuscript)

P8L28: The authors should include a description of the statistical comparison itself. There are many ways to do this, ranging from a cost function to a regression.

R3C19: We now explicitly discuss the specific metrics used for the comparison (see also R3C34).

> “We rely on several metrics to evaluate the model and radar-retrieval performance, including bias, root mean square error (RMSE), and correlation coefficient (R).” (Page 11, line 13 – 14 in the revised manuscript)

P8L28: The authors need to change the section title. The title covers only comparison with the radar data, but the text covers comparison with CALM data.

R3C20: We changed the title from “Comparison With AirMOSS ALT Retrievals” to “Comparing ALT from In-situ Observations, AirMOSS Retrievals, and CLSM Results in Alaska”.

P9L7-16: The comparison of a point, in situ measurement to a model or remote sensing pixel must account for representation error. Representation error is the uncertainty when a point measurement represents an average. The standard deviation of the CALM grid measurements is a good estimate of representation error.

R3C21: We did include the standard deviation of ALT measurements in the analysis (also explicitly stated in the caption of Figure 4). We now include an additional paragraph to provide more information regarding the representation error.

> “When comparing in-situ measurements with model results at the 81 km2 scale (i.e., a point-to-grid comparison), the ultimate measurement uncertainty propagated from the point-scale measurements to the 81 km2 scale is, for all intents and purposes, unknown due to a lack of sufficient measurements over the 81 km2 scale to compute upscaling errors (see supplementary file). We thus show instead the standard deviation of CALM measurements to illustrate, in a highly approximate way, the spatial representativeness error of the in-situ measurements – a small (large) standard deviation represents a homogeneous (heterogeneous) area in terms of ALT, meaning that the in-situ mean likely can (cannot) represent an average over a larger scale, assuming the site-scale heterogeneity is somewhat transferable to the larger scale. Such transferability might only apply to the largest in-situ site scales (e.g., 1000 m × 1000 m) to the model grid-scale (81 km2) and is thus, in general, questionable. We thus make no claim here that the standard deviations shown represent true uncertainty levels. ” (Page 12, line 20 – 25, and page 13, line 1 – 3 in the revised manuscript)

We also obtained the standard deviation of measurements at all the CALM sites that provide such information, and we now include the standard deviation as an estimate of representative error in the comparison against modeled ALT all over the simulation domain in the Northern Hemisphere (new Figure 11). We discussed the CALM representation error later in the result section. Please see our response in R3C27.

P10L25-30: The authors should state this is a standard degree day model and find some references.

R3C22: Indeed, this is different from "a standard degree day model" for predicting ALT based on relevant variables as reported in many studies (e.g., Shiklomanov et al., 2010). Here, we are simply trying to quantify how much of the temporal variability in ALT can be explained by air temperature and the snow mass. Please see further discussion in the following response (R3C23).

P11L5-8: This is a standard degree-day model for ALT with a snow adjustment. There are hundreds of variants of this model in the literature derived from the original thermodynamics equation, models, or empirically from in situ observations. The authors need some references here and text explaining that this is a degree day model.

R3C23: Actually this is not "a standard degree-day model for ALT with a snow adjustment". We are aware of many studies using the standard degree-day model to predict ALT, and we have added these references into the manuscript (see the new text below). However, our purpose here is not to derive a standard degree-day model to predict ALT. Instead, we simply construct a multiple linear regression relationship between ALT time series and two important variables, i.e., the accumulated positive air temperature and maximum snow mass, in an attempt to determine how much of the temporal variability of ALT is explained by air temperature and snow mass. We added relevant discussion here to clarify the difference between our multiple linear regression analysis and the conventional degree day models used by other studies to predict ALT.

> "The correlation coefficient relating ALT to $\sqrt{T_{cum}}$ and $\mathrm{SWE}_{\mathrm{max}}$ is the square root of the coefficient of multiple determination (R^2) obtained through fitting Equation (4). This equation is similar in form to the common degree-day model for predicting ALT from accumulated degree days of thaw based on the Stefan solution (e.g., Shiklomanov and Nelson, 2002; Zhang et al., 2005; Riseborough et al., 2008; Shiklomanov et al., 2010). Here, however, we constructed equation (4) for a different purpose: to explore how much of the temporal variability of ALT can be jointly explained by snow mass and above-freezing air temperature. Before calculating these correlation coefficients, we removed the linear trend within ALT, T_{cum}, and $\mathrm{SWE}_{\mathrm{max}}$ to avoid potentially exaggerating the correlation due to an underlying trend." (Page 15, line 19 – 25, and page 16, line 1 -2 in the revised manuscript)

P11L8: The authors should explain why they included the a0 term. The a0 term is not often seen in a degree day model because one typically assumes the soil starts frozen (a0=0).

R3C24: Again, this is not a degree day model for predicting ALT. The a0 is necessary here to ensure a well-defined fitting to the multiple linear regression relationship in order to obtain statistically meaningful coefficients of determination and correlation. We have elaborated more about our purpose for performing the regression and have clarified the differences between equation (4) and the standard degree day model. Please see the above response (R3C23).

P11L8: The authors should explain why they chose Tcum rather than the square root of Tcum. One can derive the sqrt(Tcum) relationship directly from the original thermodynamics equation and the relationship appears many times in analyses of in situ measurements. Because of the strong theoretical basis of sqrt(Tcum), using plain Tcum is rare, so the authors need to justify its use.

R3C25: Although Equation (4) was not intended to be a standard degree day model as explained above (R3C23), we have followed the reviewer's suggestion and have replaced Tcum with sqrt(Tcum). The resulting correlation coefficient map is very similar to that using Tcum (as shown below). Note, however, we now detrend the data before calculating the correlation coefficient and the figures shown below are not included in the revised manuscript. Please see our response in R3C43 for updated figures.

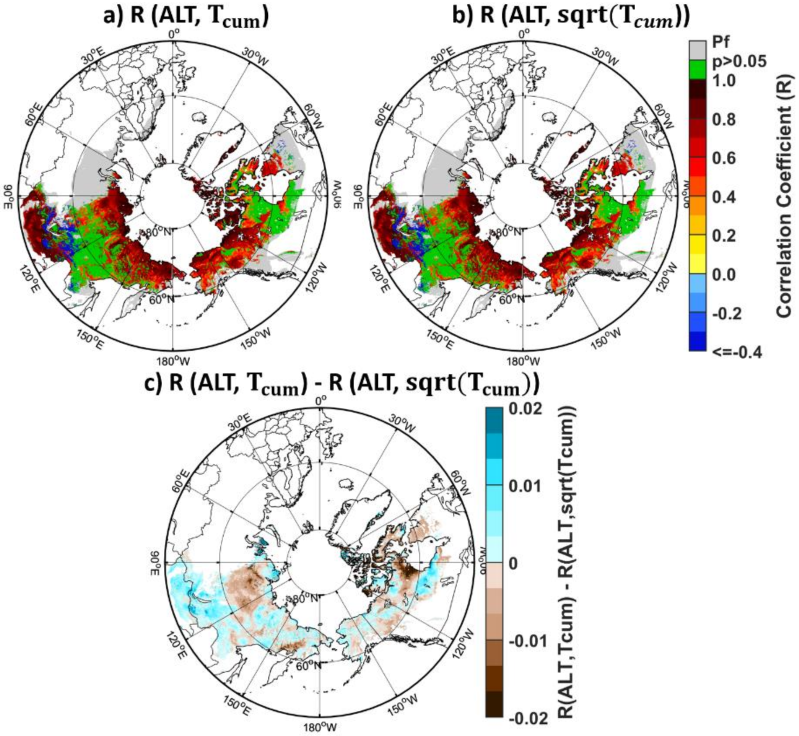


Figure R1: Correlation coefficient between a) ALT and effective accumulated air temperature (T_{cum}) and b) ALT and sqrt(T_{cum}). c) the difference between a) and b).

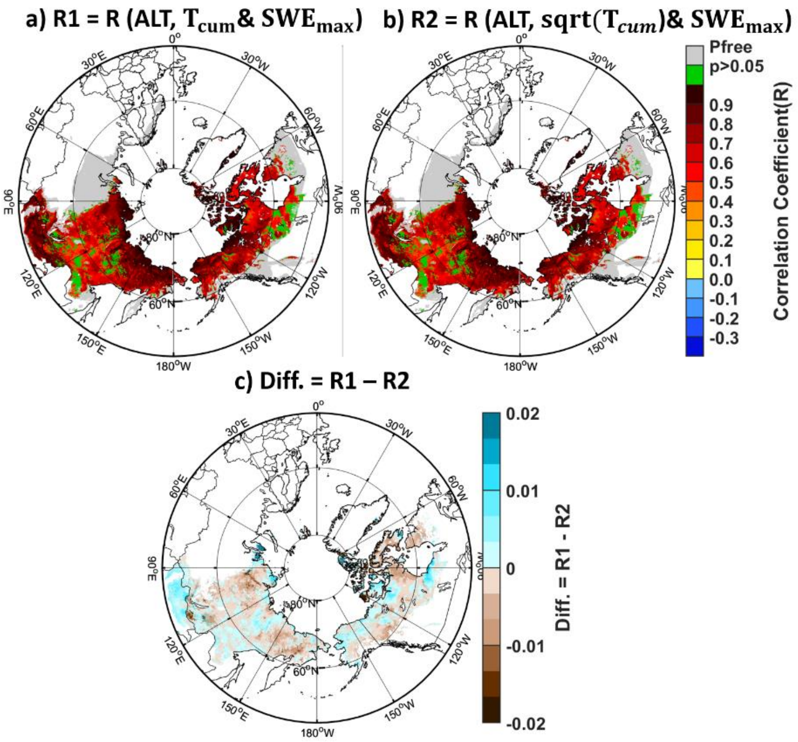


Figure R2: Multi-variable coefficient of correlation for a fitted multiple linear regression model a) between ALT and T_{cum} and SWE_{max} and b) between ALT and sqrt(T_{cum}) and SWE_{max}. c) the difference between a) and b).

P11L12-16: This description of comparing to CALM is out of place and should be moved to section 3.1.

R3C26: We consider the context here to be distinct from that of section 3.1, which only compares an ALT snapshot in the thawing season of 2015 in Alaska (when and where AirMOSS flights were conducted). Here we focus, as explained in the manuscript, on the comparison of multi-year ALTs and the climatological mean between model results and CALM observations all over the simulation domain in the Northern Hemisphere.

P11L12-16: The authors need to account for uncertainty in the CALM measurements when comparing to the model output.

R3C27: We have managed to obtain the standard deviation of measurements at all the CALM sites that provide such information, and we now include the standard deviation as an estimate of representative error in the comparisons against modeled ALT. Please see our new Figure 11.

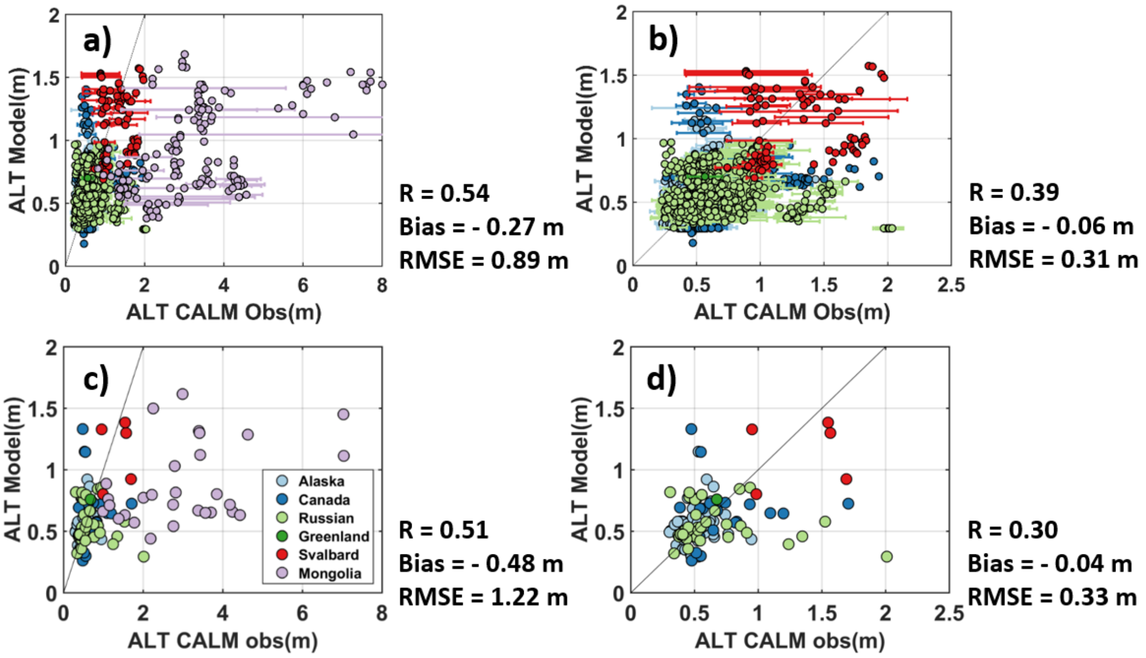


Figure R3 (Figure 11 in the manuscript): a) Annual ALT from CLSM simulation vs. CALM observations with horizontal error bars indicating standard deviations of measurements within the model grid cell. Error bar is absent if the number of measurements within a 81 km^2 grid cell is less than three. b) As in a) but excluding the Mongolia sites. c) 38-yr average ALT for the period 1980-2017 from CLSM simulation vs. CALM observations. d) As in c) but without the Mongolia sites. The correlation coefficient (R), bias, and root mean squared error (RMSE) are provided next to each subplot.

However, as explained by the supplementary file, it is impossible to estimate the CALM measurement uncertainty at the model grid scale due to the lack of upscaling errors. But by showing standard deviation here, we could explain how heterogeneous the CALM site sampling area is and in general how reasonable the point-to-grid comparison results are. We added relevant sentences as below.

> “As noted in section 3.1 and in the supplementary file, the uncertainty of the CALM ALT measurements in the context of evaluating grid cell-scale model results theoretically involves uncertainty derived from probing point measurement uncertainty, site-scale mean uncertainty, and upscaling errors in going from the site-scale to the model-scale. This latter uncertainty in particular is unknown. In our figures (in section 4.4) we show the standard deviation of the observed ALT as a very crude surrogate for the spatial representativeness error associated with the point-to-grid comparison. As before, we make no claim here that the standard deviations shown represent the relevant statistical uncertainty.” (Page 16, line 18 – 25 in the revised manuscript)

> “As another example, sites on south-facing slopes presumably have much deeper ALT than those on slopes with less exposure to the sun, which is not captured by CLSM. Th large representative errors of Mongolian sites are clearly illustrated by the standard deviation

> (although computed only with 3 to 5 measurements) as shown by the error bars in Figure 11a." (Page 27, line 2 – 4 in the revised manuscript)

Please also see discussion in R3C21.

P11L18-24: The spinup technique invalidates the trend analysis. Either drop the trend analysis or modify the spinup.

R3C28: We have removed the trend analysis.

P12L6-8: Delete. Unneeded.

R3C29: We modified this paragraph and moved it to Section 3.1, since it fits better there.

P12L10-12: Move to methods.

R3C30: Done as suggested.

P12L14: Figure 3 shows the difference between the model and observations, but is this difference within the uncertainty? If yes, then the two are statistically identical and thus a match. If no, then there is a statistically significant mismatch. The magnitude of the difference is unimportant if the difference is less than uncertainty. The authors need to account for uncertainty in this comparison.

R3C31: We used Figure 3 to compare the spatial patterns of remotely sensing data and model results, and we discuss the differences next in Figure 4. We added relevant sentences discussing the radar ALT uncertainty and how the model results compare with the radar retrievals.

> "We employed the strategy of Schaefer et al. (2015) to handle the uncertainty propagation, i.e., adding in quadrature the uncertainty components from each scale/level involved (see the supplementary file for a detailed description). For AirMOSS retrievals, the sampling uncertainty of mean ALT at the 81 km2 model grid-cell scale is negligible given the large sampling size and the fact that the retrieval uncertainty dominates the overall uncertainty (see supplementary file). Here, we use a nominal estimate of 0.15 m to represent the AirMOSS uncertainty (i.e., the average of the lower and upper bound of the actual retrieval uncertainty for individual radar pixels as discussed by Chen et al. (2019))." (Page 12, line 13 – 18 in the revised manuscript)
>
> "Figures 4c further shows that the CLSM-simulated ALT agrees well with the AirMOSS ALT retrievals to within the measurement uncertainty of 0.15 m at all the site-located model grid cells. Indeed as Figure 3c illustrated, the differences between simulated ALT and the AirMOSS retrievals over all the transects examined here are generally below the measurement uncertainty of 0.15 m." (Page 18, line 19 – 22 in the revised manuscript)

P12L21: Explain here why soil type influences the result. The reader should not have to flip forward in the paper to get this answer.

R3C32: We have added relevant discussion at this point in the text:

> "Variations of the simulated ALT within a single transect (Figure 3a) are predominantly induced by changes in soil type (indicated in Figure 2c and 2d). In essence, the higher the organic carbon content within the soil, the smaller the simulated ALT due to slower heat transfer associated with lower thermal conductivity, higher porosity, heat capacity, etc. (Tao et al., 2017). See also section 4.2 for a discussion of the influence of soil texture on the spatial pattern of ALT." (Page 17, line 15 – 19 in the revised manuscript)

P12L25: Agreement 'to first order' is too vague and carries no meaning. The authors need to quantify the agreement accounting for uncertainty.

R3C33: We deleted "to first order" here and modified the sentence as below.

> "Figures 4a and 4b show that the CLSM-simulated ALTs agree with the in-situ observations with an overall mean bias of -0.05 m and a RMSE of 0.17 m." (Page 17, line 22 – 23 in the revised manuscript)

Given the large scale gap between the in-situ site and the model grid cell, it is impossible to estimate the measurement uncertainty to be comparable with model results. However, we do show the standard deviation as a representative error of the in-situ measurements. We added relevant sentences as below.

> "When comparing in-situ measurements with model results at the 81 km2 scale (i.e., a point-to-grid comparison), the ultimate measurement uncertainty propagated from the point-scale measurements to the 81 km2 scale is, for all intents and purposes, unknown due to a lack of sufficient measurements over the 81 km2 scale to compute upscaling errors (see supplementary file). We thus show instead the standard deviation of CALM measurements to illustrate, in a highly approximate way, the spatial representativeness error of the in-situ measurements – a small (large) standard deviation represents a homogeneous (heterogeneous) area in terms of ALT, meaning that the in-situ mean likely can (cannot) represent an average over a larger scale, assuming the site-scale heterogeneity is somewhat transferable to the larger scale. Such transferability might only apply to the largest in-situ site scales (e.g., 1000 m × 1000 m) to the model grid-scale (81 km2) and is thus, in general, questionable. We thus make no claim here that the standard deviations shown represent true uncertainty levels. " (Page 12, line 20 – 25, and page 13, line 1 – 3 in the revised manuscript)

P13L1-8: The authors need to expand their statistical analysis of the model-data comparison. All they have is correlation, which says nothing about magnitude. They should expand the residual analysis to include bias (mean residual), root mean square error (residual standard deviation), and chi-squared (standard deviation of residuals normalized by uncertainty).

R3C34: In addition to correlation coefficient, we already included bias and RMSE as discussed in section 4.1 and the associated Table 3 (and also in section 4.3), and the Figure 11 and the associated discussion in the manuscript, which the reviewer might have missed. Also, since it is impossible to estimate the uncertainty of in-situ measurements at the model grid scale as explained in detail in the supplementary file, we are not able to compute the chi-square.

We also provided a supplementary file and added more description to explain our model-data comparison strategy. Please also see our discussion in R3C18, R3C21, R3C27 and R3C31.

P13L9: 'Broadly consistent' is too vague and carries no meaning. The authors need to quantify the agreement.

R3C35: We deleted "broadly" here and quantified the agreement using the following text

> "Although the AirMOSS ALT retrievals generally underestimate the in-situ ALT measurements (as shown in Figure 4a), the retrievals tend to be more consistent with the observations when the in-situ measurements are within the ~60 cm sensing depth of the P-band radar data, as indicated in Table 3. Specifically, excluding the sites with in-situ ALT measurements that exceed the AirMOSS sensing depth of ~60 cm, the overall mean bias for the AirMOSS retrievals at the 81 km2 scale drops to -0.01 m, and the correlation coefficient increases to 0.64. In contrast, the CLSM simulation results show a bias of 0.01 m and a zero correlation coefficient at these sites." (Page 18, line 7 – 13 in the revised manuscript)

P13L20: 'In general' is too vague and has no meaning. Comparing modeled and observed trend with latitude is perfectly valid here. The limited number of in situ measurements will simply result in higher uncertainty.

R3C36: We deleted "In general" here.

P13L23: Again, 'generally' is too vague.

R3C37: We deleted "generally" here.

P13L23-33: Figure 5 is not the correct format to show the relationships described here. The reader cannot visualize the relationships and correlations from the simple time series plots in Figure 5. The authors should replace the time series plots with three plots to illustrate the relationships: ALT vs. latitude, ALT vs. organic matter content, and ALT vs. air temperature.

R3C38: We feel the reviewer did not fully understand this figure. This is not a time series plot, but is rather a spatial-series plot. We do, however, appreciate the reviewer's point and have thus added scatter plots (e.g., ALT vs. air temperature, ALT vs. organic matter content, and ALT vs. maximum snow depth) next to the spatial-series plots. Please see our new Figure 5.

P14L20-23: Perhaps, but the author's argument is not convincing. Shading associated with higher LAI represents an equally valid explanation. Higher water content associated with higher organic matter content could also explain the difference. The authors have the full suite of model output on hand. They should do a statistical analysis of available output to track down exactly what explains the difference.

R3C39: We agree with the reviewer about potential shading effects; however, such effects are not resolved in the current model. The statement "Higher water content associated with higher organic matter content could also explain the difference" is also true but not valid here since the differences between HomF&Veg vs. HomF relate only to the vegetation differences.

The argument about the snow impacts caused by changing vegetation is based on our previous investigation in (Tao et al., 2017). We have accordingly modified the relevant sentences and cited our previous study here.

> "This is because the generally lower albedo of the taller and leafier trees (representative of the IVO transect) during the snow season resulted in increased snowmelt and thus reduced snowpack during the snow season (compare the green and red curves in Figure 6c), thereby reducing the thermal insulation of the wintertime ground. With reduced insulation, cold season ground temperatures dropped, making it more difficult for temperatures to recover during summer(Tao et al., 2017). " (Page 20, line 14 – 17 in the revised manuscript)

P14L31: The authors need to identify exactly what soil parameters changed. Porosity? Thermal conductivity? Volumetric water content? Also, the authors need to explain how they specify soil properties in the model. A sharp change as seen here is common when specifying properties by soil type, such as sandy loam defined in the USGS soil triangle. A sharp change would be unusual when specifying soil properties by maps of soil texture (sand, silt, and clay fraction).

R3C40: This information is mentioned in the methods section (section 3.2) and in Table 2 when we describe the idealized experiments. We now clarify the approach as follows:

> "Owing to the strong control of soil type-related parameters (see section 3.2 and Table 2) on soil moisture, spatial variability in soil moisture remains high in HomF and HomF&Veg and is only eliminated once these soil type-related parameters are homogenized (Figure 6d), which explains the abrupt changes shown in Figure 3c as mentioned in section 3.1. " (Page 20, line 21 – 23 in the revised manuscript)

Note also that, as indicated in the caption of Figure 2, we effectively use a very highly resolved version of a soil texture triangle having 253 subtypes, not the more standard USDA triangle with ~10 types.

P15L33: The reason simulated ALT is deeper is the same reason identified later in the manuscript: the model either has permafrost or it does not because it cannot represent sub-grid scale processes. When the model does simulate permafrost in sporadic regions, it is always greater than observed because it represents an area average of permafrost and non-permafrost areas.

R3C41: We agree with the reviewer about the model deficiency on handling sub-grid variability as we also discussed in the manuscript (Section 4.4, Page 18, line 21 -26)). What we meant here is that the simulated ALT in discontinuous and sporadic permafrost regions is deeper than that in continuous regions. We modified the sentence as below for clarity.

> "The largest ALT standard deviations (red color in Figure 8b) are found mainly in discontinuous and sporadic permafrost regions (see Figure 1b) where ALTs are deeper on average than that in continuous permafrost region." (Page 22, line 5 – 6 in the revised manuscript)

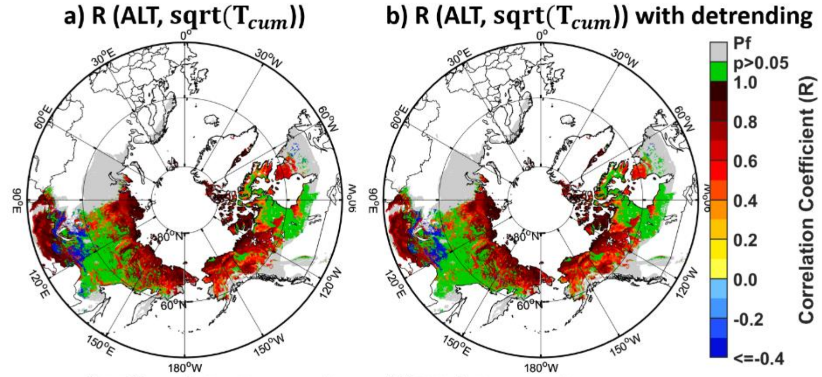
P16L9-10: The authors need to either perform the analysis with air temperature or at least summarize and reference the results of other studies that did perform the analysis.

R3C42: We have modified this sentence and added relevant references here.

> "The relationship between the spatiotemporal characteristics of simulated ALT and air temperature forcing has been investigated before in many studies at the site to landscape scale (e.g., Klene et al., 2001; Shiklomanov and Nelson, 2002; Zhang et al., 2005; Juliussen and Humlum, 2007) and at the regional scale (e.g., Anisimov et al., 2007). Here we simply analyze the correlation coefficient between ALT and two variables: the proxy of total energy input into the ground (i.e., $\sqrt{T_{cum}}$, see section 3.3) and the maximum SWE. Our goal is to explore how much of the spatiotemporal variability of ALT across the globe can be jointly explained by these two variables." (Page 22, line 16 – 22 in the revised manuscript)

P16L12: The authors need to remove the trends in ALT, Tcum, and SWEmax before calculating the correlation coefficients. We see nice strong correlations because all three variables show strong trends over the time period of the simulation. Removing the trends will significantly change Figure 9 and its interpretation. If the authors want to isolate the effects of trends on the ALT, then they should include an analysis using the congruent trend fraction.

R3C43: As requested by the reviewer, we now detrend ALT, Tcum, and SWEmax before calculating the correlation coefficients between ALT and sqrt(Tcum), between ALT and SWEmax, and between ALT and the two factors (sqrt(Tcum) and SWEmax). The resulting correlation coefficient maps differ very much from the results without detrending mostly over the areas with large p (> 0.05) as shown below. But the general pattern of the correlation coefficients do not change very much, and our core conclusion still holds. Please see our new Figure 9 in the revised manuscript. Please also see R3C23 and R3C25 for relevant description and discussion.



a) R (ALT, $\mathrm{sqrt}(\mathrm{T}_{cum})$)

b) R (ALT, $\mathrm{sqrt}(\mathrm{T}_{cum})$) with detrending

c) Diff. = a) No Detrending – b) With Detrending

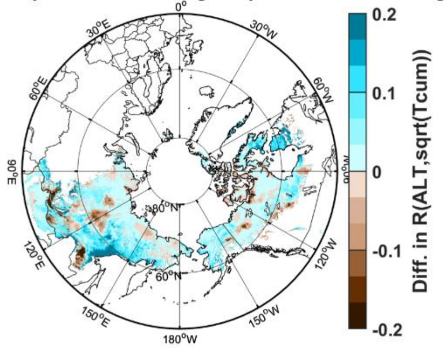


Figure R4: Correlation Coefficient between ALT and sqrt(T_{cum}) a) without and b) with detrending. c) the difference between a) and b).

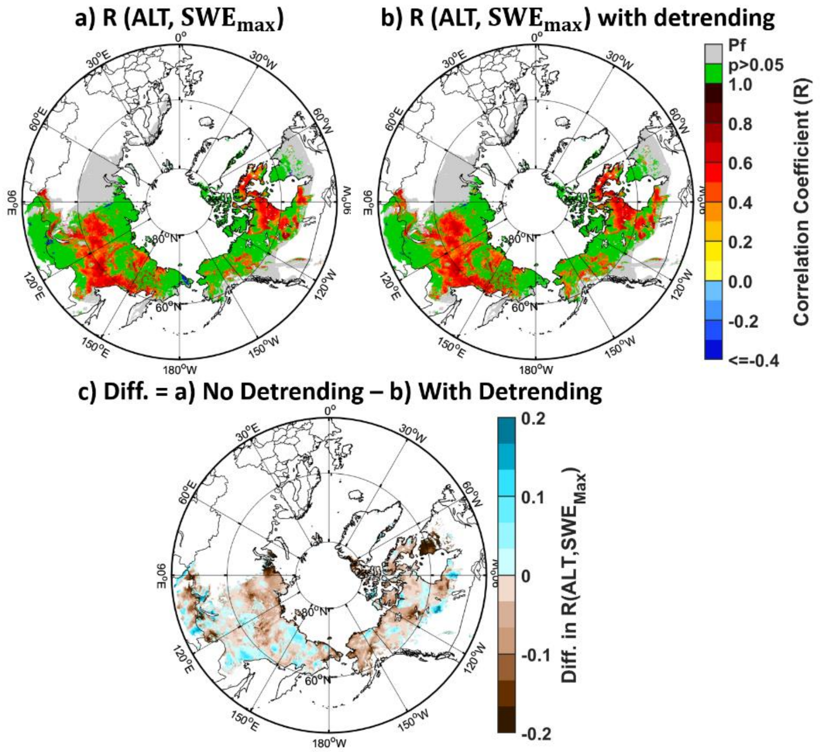


Figure R5: Correlation Coefficient between ALT and SWE_{max} a) without and b) with detrending. c) the difference between a) and b).

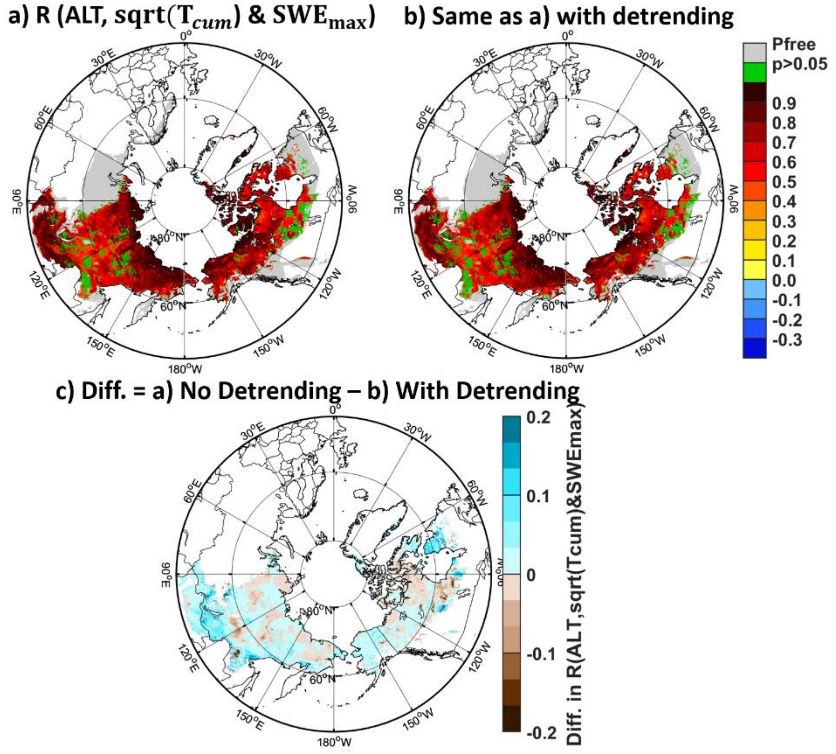


Figure R6: Multi-variable coefficient of correlation for a fitted multiple linear regression model between ALT and sqrt(T_{cum}) and SWE_{max} a) without and b) with detrending the data. c) the difference between a) and b).

P17L1: The regions identified on the maps do not correspond to high mountains. Please clarify.

R3C44: We now explicitly identify these mountainous regions:

> "Even so, a few limited areas still exhibit low correlations ($p>0.05$, green colour in Figure 9c). Some of these areas are in mountainous regions, for instance, the Eastern Siberian (Ostsibirisches) Bergland, where more complex environmental controls might be playing a dominant role. In addition, MERRA-2 snow forcing might be severely erroneous in these regions." (Page 23, line 20 – 22 in the revised manuscript)

P17L5-6: Delete.

R3C45: Done.

P17L22: 'Geographically thin' is too vague. Please reword.

R3C46: We have deleted 'geographically thin' and reworded this sentence.

> "The disagreements between the simulated and observed permafrost extents (covering about a few degrees latitude) toward the south in Figure 10a (green and blue areas at the southern edge of permafrost regions) are less of a concern, since the comparison in such areas is muddied by the interpretation of "isolated" permafrost in the observational map (Figure 1b)." (Page 25, line 2 – 4 in the revised manuscript)

P18L20: The authors cannot make this claim without an actual comparison with other models. Drop the statement.

R3C47: Indeed, we made an extensive effort to compare our results with existing results in literature during the previous round of revisions. See the online discussion from the open review for details (https://www.the-cryosphere-discuss.net/tc-2018-119/tc-2018-119-AC3-supplement.pdf). We now add "perhaps" here to tune down our tone a little bit.

P18L14 and P18L34: This is the first mention of representation error. The authors need to estimate the representation error of the in situ measurements and include this in the comparison with the modeled ALT. There are several ways to do this and I leave it to the authors to determine the most appropriate method for this paper.

R3C48: We now provide the standard deviation of CALM measurements to represent spatial representative error in the comparison with modeled annual ALT. Please also see our new analysis and discussion in R3C21, R3C27 and R3C33.

P19L1-2: This statement is not true. A point measurement can represent an area average if one includes representation error in the point measurement.

R3C49: We now include the standard deviation for Mongolian sites if more than two sites present within the same model grid cell. Please see our new Figure 11. We also change the relevant sentence as below:

> "An additional reason for the underestimation of ALT in Mongolia might be a mismatch between the land surface parameter values used in the model and the actual conditions at each site. For instance, detailed soil information (https://www2.gwu.edu/~calm/data/webforms/mg_f.html) indicate that some Mongolian sites have special "rocky" soil types including limestones (e.g., M04), slatestones (e.g., M05), gravelly sand (e.g., M06 and M08), etc. that are not well represented in the model. As another example, sites on south-facing slopes presumably have much deeper ALT than those on slopes with less exposure to the sun, which is not captured by CLSM. This large representative errors of Mongolian sites are clearly illustrated by the standard deviation (although computed only with 3 to 5 measurements) as shown by the error bars in Figure 11a." (Page 26, line 21 – 25 in the revised manuscript)

P19L4: Either change the spinup or drop the trend analysis.

R3C50: We have removed the trend analysis from the manuscript.

P20L15-16: Again I am confused about the motivation of including the remotely sensed ALT in this paper. This paper compares the model to the RS data, but also compares the RS data to the in situ measurements. Do the authors want to validate the model or the RS data?

R3C51: We now deleted all the site-scale comparison between AirMOSS retrievals and in-situ observations since Chen et al. (2019) had already discussed the quality and uncertainty of the retrievals in detail. Please also see R3C2 and R3C11.

P20-P22: Please reduce the summary to one page or less. The current summary is way too long and simply repeats material from the results section. What is the primary, take-away points the authors want to convey? What are their most important or most interesting results? What are the broader implications of their results?

R3C52: As suggested, we have shortened the summary, which now focuses on our primary and take-away points. We kept some essential discussion on model uncertainty and deficiency, though. Please see our new section 5.

Reference

Alexeev, V. A., Nicolsky, D. J., Romanovsky, V. E., and Lawrence, D. M.: An evaluation of deep soil configurations in the CLM3 for improved representation of permafrost, Geophys Res Lett, 34, 10.1029/2007gl029536, 2007.

Anisimov, O. A., Lobanov, V. A., Reneva, S. A., Shiklomanov, N. I., Zhang, T., and Nelson, F. E.: Uncertainties in gridded air temperature fields and effects on predictive active layer modeling, Journal of Geophysical Research: Earth Surface, 112, 2007.

Brown, J., Ferrians, O., Heginbottom, J. A., and Melnikov, E.: Circum-Arctic Map of Permafrost and Ground-Ice Conditions, Version 2. [Permafrost Extent], NSIDC: National Snow and Ice Data Center, Boulder, Colorado USA, https://nsidc.org/data/ggd318, 2002.

Chen, R. H., Tabatabaeenejad, A., and Moghaddam, M.: Retrieval of permafrost active layer properties using time-series P-band radar observations, IEEE Transactions on Geoscience and Remote Sensing, 10.1109/TGRS.2019.2903935, 2019.

Dankers, R., Burke, E. J., and Price, J.: Simulation of permafrost and seasonal thaw depth in the JULES land surface scheme, Cryosphere, 5, 773-790, 2011.

Juliussen, H., and Humlum, O.: Towards a TTOP ground temperature model for mountainous terrain in central-eastern Norway, Permafrost Periglac, 18, 161-184, 2007.

Klene, A. E., Nelson, F. E., Shiklomanov, N. I., and Hinkel, K. M.: The n-factor in natural landscapes: variability of air and soil-surface temperatures, Kuparuk River Basin, Alaska, USA, Arctic, Antarctic, and Alpine Research, 33, 140-148, 2001.

Lawrence, D. M., Slater, A. G., Romanovsky, V. E., and Nicolsky, D. J.: Sensitivity of a model projection of near-surface permafrost degradation to soil column depth and representation of soil organic matter, Journal of Geophysical Research-Earth Surface, 113, 10.1029/2007jf000883, 2008.

Nicolsky, D. J., Romanovsky, V. E., Alexeev, V. A., and Lawrence, D. M.: Improved modeling of permafrost dynamics in a GCM land-surface scheme, Geophys Res Lett, 34, 2007.

Paquin, J. P., and Sushama, L.: On the Arctic near-surface permafrost and climate sensitivities to soil and snow model formulations in climate models, Clim Dynam, 44, 203-228, 2015.

Riseborough, D., Shiklomanov, N., Etzelmuller, B., Gruber, S., and Marchenko, S.: Recent advances in permafrost modelling, Permafrost Periglac, 19, 137-156, 10.1002/ppp.615, 2008.

Sapriza-Azuri, G., Gamazo, P., Razavi, S., and Wheater, H. S.: On the appropriate definition of soil profile configuration and initial conditions for land surface-hydrology models in cold regions, Hydrol Earth Syst Sc, 22, 3295-3309, 10.5194/hess-22-3295-2018, 2018.

Shiklomanov, N. I., and Nelson, F. E.: Active-layer mapping at regional scales: A 13-year spatial time series for the Kuparuk region, north-central Alaska, Permafrost Periglac, 13, 219-230, 2002.

Shiklomanov, N. I., Streletskiy, D. A., Nelson, F. E., Hollister, R. D., Romanovsky, V. E., Tweedie, C. E., Bockheim, J. G., and Brown, J.: Decadal variations of active-layer thickness in moisture-controlled landscapes, Barrow, Alaska, Journal of Geophysical Research: Biogeosciences, 115, 2010.

Tao, J., Reichle, R. H., Koster, R. D., Forman, B. A., and Xue, Y.: Evaluation and Enhancement of Permafrost Modeling With the NASA Catchment Land Surface Model, J Adv Model Earth Sy, 9, 2771-2795, 10.1002/2017MS001019, 2017.

Wang, K., Jafarov, E., Overeem, I., Romanovsky, V., Schaefer, K., Clow, G., Urban, F., Cable, W., Piper, M., and Schwalm, C.: A synthesis dataset of permafrost-affected soil thermal conditions for Alaska, USA, Earth Syst Sci Data, 10, 2311-2328, 2018.

Zhang, T., Frauenfeld, O. W., Serreze, M. C., Etringer, A., Oelke, C., McCreight, J., Barry, R. G., Gilichinsky, D., Yang, D., and Ye, H.: Spatial and temporal variability in active layer thickness over the Russian Arctic drainage basin, Journal of Geophysical Research: Atmospheres, 110, 2005.