**Response to comments of reviewer #2:**

We thank Reviewer 2 for providing valuable comments on our paper that helped improving our manuscript. Please find **the reviewers comments in bold**, our point-by-point answers without formatting, and *changes to the initial manuscript in Italics*.

**Comments:**

**1. P5 L22-24: Most of the analysis is based on using the meteostation data across the Russian territory from 1980-1990. Unfortunately, there is no in-depth analysis of the observations and why these data can be used. Observations at many meteostations are performed at the disturbed conditions and thus only a small subset of this data could be used in the direct comparison. The authors simply assume that just all data could be used. The meaningful comparison of the model to observations at the disturbed sites could be done if the model also simulates the disturbed conditions. One of the major problems in this paper is that many observations were taken at the disturbed sites near meteostations, while models simulate typical ground conditions. To do the comparison correctly it is either necessary to model disturbed conditions in the models or remove 'corrupted' observations from the analysis.**

This data set is quality checked and officially released by the All-Russian Research Institute of Hydrometeorological Information-World Data Centre (RIHMI-WDC; <http://meteo.ru/>). They provide quality controlled soil temperature at depths to 320 cm from meteorological stations of the Russian Federation. The data was subject to quality control by using four methods of statistical control (histograms, standard deviation check, check for connectedness of values that are neighboring in time).

Meteorological sites are located in an open and typical place within the surrounding terrain. They are located far from the major obstacles and water bodies that may have a direct effect on the quality of measurements. To keep the surface of the meteorological site in its natural state, it is permitted to walk only on the specially laid tracks, a width not exceeding 40 cm, within the site territory. In the warm season grass on the site is regularly mowed and trimmed. The height of the grass on the site is not more than 20 cm. In the winter, the natural state of the snow cover is not broken. Observing conditions at the Russian stations in all meteorological elements correspond with WMO standards. The observations presented have been included in data sets, such as GSOD, HadSRUT4 etc. and are widely used in climate research.

Soil temperature measurements are carried out simultaneously with the measurements of the whole complex of meteorological observations (temperature, characteristics and dynamics of snow cover and precipitation and so on). The soil temperature observations are under the original surface (from draw-out thermometer data). All meteorological observations, including soil temperature, are produced exclusively within the same site area (26 by 26 meters in size), and under constant careful control of workers who are caring for a site. This prevents any accidental violations of the integrity of the site and guarantees the quality of observations. A detailed history of the development of methodology for soil temperature measurements was provided by Bykhovets et al. (2007). A detailed description of dataset preparation is provided in Sherstiukov (2007). The archived soil temperature dataset was run through four independent methods of quality control (Sherstiukov, 2012).

Thus, the data of meteorological observations at the Russian stations for soil temperature can be effectively used in the evaluation of thermal changes in the upper layers of permafrost zone, as well as in the analysis of processes of thermal interaction between the atmosphere and soil. Similar conclusions were obtained by leading Russian scientists in this area (Anisimov and Sherstiukov, 2016; Pavlov and Malkova, 2009) and internationally (Park et al., 2014; Brun et al., 2013; Decharme et al. 2016; PaiMazumder et al., 2008).

However, as Park et al. (2014) pointed out, the observations at some locations could have been unavoidably disturbed by grass cutting during the warm season and the removal of organic materials, mainly at agricultural sites. These disturbances may cause increased warming of the soil over time. Therefore, long-term soil temperature trends in could potentially include this non-climatic component (Frauenfeld et al 2004).

Thus, we agree that we have to be careful which and how we use the observation data. But we can argue that our study results are solid. First, and most important is that we do not present either direct point-by-point comparison with station observations nor trends, but we investigate variables relationships. We evaluate the models functional behaviors (inter-variable scatter plots, Figs. 1, 4, 5 and PDFs, Fig. 2). The inter-variable scatter plots (which show medians and the 25th and 75th percentiles) as well as the PDFs present the overall functional behavior of snow insulation effects, and not the individual station’s representation. Therefore, this kind of benchmark to evaluate the models skill is a solid approach using the station data, which include uncertainties (by showing the percentiles). This approach, as the other reviewer points out, reveals some structural issues of the models in simulating snow depths and its insulation effects on soil temperature. Secondly as we are only looking at upper soil temperatures, any change due to changes in soil column temperature gradient resulting from grass removal are negligible. Thirdly, the agriculture sites concerns are not an issue for our permafrost area focused study.

According to the reviewer’s comment, we include in section 2.2 two references for more details of the data set: “*A detailed description of dataset preparation is provided in Sherstiukov (2012a). Observing conditions at the Russian stations in all meteorological elements correspond with WMO standards. The observations presented have been included in data sets, such as GSOD, HadSRUT4 etc. and are widely used in climate research (e.g. Anisimov and Sherstiukov, 2016; Decharme et al. 2016; Park et al., 2014; Brun et al., 2013; Pavlov and Malkova, 2009; PaiMazumder et al., 2008).The soil temperature dataset was run through four independent methods of quality control (Sherstiukov, 2012b).*”

Further, we emphazise the point that possible disturbances do not affect our results:

“*However, some soil temperature observations could be disturbed by grass cutting during the warm season and the removal of organic materials, mainly at agricultural site, which may affect the trend in warm season (Park et al., 2014), but this does not affect our results about the air- upper soil temperature relationship in winter*.”

Finally, we cannot compare the model results and observation over only bare ground (or short grass) in the PCN simulations. Only the grid-cell average results were kept (and not the value for each vegetation and bare ground tile).

References:

Anisimov, O.A., Sherstiukov A.B. Evaluating the effect of environmental factors on permafrost factors in Russia, Earth’s Cryosphere, XX(2), 90-99, 2016.

Brun, E., Vionnet, V., Boone, A., Decharme, B., Peings, Y., Valette, R., Karbou, F. and Morin, S.: Simulation of northern Eurasian local snow depth, mass and density using a detailed snowpack model and meteorological reanalysis, J. Hydrometeorol., 14, 203–214, doi:10.1175/jhm-d-12-012.1, 2013.

Bykhovets, S. S., Sorokovikov, V. A., Martuganov, R. A., Mamykin, V. G., and Gilichinsky, D. A. History of soil temperature measurements at the network of meteorological stations in Russia, Earth’s Cryosphere, XI(1), 7-20, 2007.

Decharme, B., Brun, E., Boone, A., Delire, C., Le Moigne, P. and Morin, S.: Impacts of snow and organic soils parameterization on North-Eurasian soil temperature profiles simulated by the ISBA land surface model, The Cryosphere, 10, 853–877, doi: 10.5194/tc-10-853-2016, 2016.

Frauenfeld O W, Zhang T, Barry R G and Gilichinsky D.: Interdecadal changes in seasonal freeze and thaw depths in Russi,a J. Geophys. Res. 109 D05101, 2004

Park, H., Sherstiukov, A.B., Fedorov, A.N., Polyakov, I. V., Walsh, J.E.: An observation-based assessment of the influences of air temperature and snow depth on soil temperature in Russia, Environmental Research Letters, Vol. 9, 2014, http://iopscience.iop.org/1748-9326/9/6/064026

PaiMazumder, D., Miller, J., Li, Z., Walsh, J. E., Etringer, A., McCreight, J., Zhang, T., Mölders, N. Evaluation of Community Climate System Model soil temperatures using observations from Russia, Theoretical and Applied Climatology , 94(3),187-213, 2008.

Pavlov, A.V., Malkova, G.V. Small-scale mapping of trends of the contemporary ground temperature changes in the Russian North, Earth’s Cryosphere, XIII(4), 32-39, 2009.

Sherstiukov, A. Dataset of daily soil temperature up to 320 cm depth based on meteorological stations of Russian Federation, RIHMI-WDC, 176, 224-232, 2012a.

Sherstiukov, A. Statistical quality control of soil temperature dataset, RIHMI-WDC, 176, 224-232, 2012b.

**2. P5 L29: “comprehensive Russian station data set”. Please describe this comprehensive dataset. Provide a reference and in-depth discussion about the site conditions, disturbances of the ground cover, what stations are not qualified for the comparison.**

This comment is related to the comment above. Please see our answer above.

**3. P7 L26: “Snow depth was then calculated from SWE using a snow density of 250 kg m-3.” What was the rational to use 250 kg/m3. Please provide references. How would the results change is 300 kg/m3 is used? Why not to use the SWE instead of the snow depth in all other further comparisons. It looks like converting to the snow height can add additional uncertainties to the consecutive analysis.**

The stations only give the snow depth (and not SWE). In addition to the station’s ground snow observations we also use the gridded SWE data from the GlobSnow-2 product to support the station data results. Thus, to compare the GlobSnow data with the station data, we must convert one data set. No way has a preference. We decided to convert SWE from GlobSnow to snow depth, and for this we need an assumption about snow density. We use 250 kg m-3 because it is a mean observed value. Zhong et al. (2013) report snow density values of 180-250 kg m-3 for tundra/taiga and 156-193 kg m-3 for alpine snow classes in winter. Woo et al. (1983) report snow density values of 250-400 kg m-3 for various terrain types. Thus we use the mid-value of 250 kg m-3. We added a paragraph in section 2.2: “*Snow depth was then calculated from SWE using a snow density of 250 kg m-3, which is a median observed value in winter. Zhong et al. (2013) report snow density values of 180-250 kg m-3 for tundra/taiga and 156-193 kg m-3 for alpine snow classes. Woo et al. (1983) report snow density values of 250-400 kg m-3 for various terrain types. Choice of density does not materially affect the results.”*

To show you that this parameter choice does not affect our results, Fig. X1 is the comparison with snow depth from GlobSnow using a density of 300 kg m-3. This figure confirms that the pattern does not change and shows that the differences in snow depth are small (less than 10 cm).

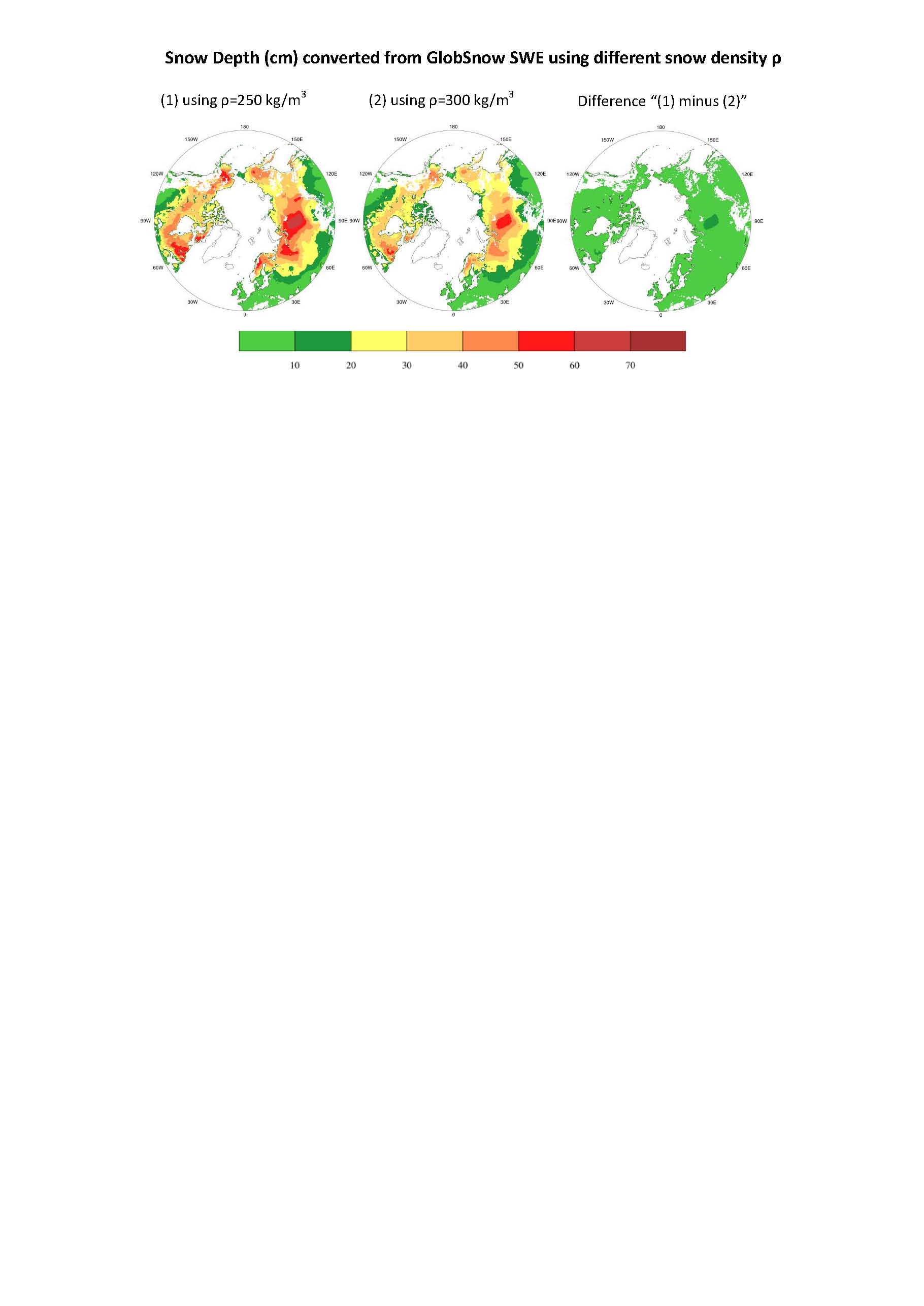


Figure X1: Snow depth (cm) derived from GlobSnow SWE using different assumptions of snow density, winter 1980-2000.

Further, our calculations show that the model biases in snow depth are quite similar using either the station or the GlobSnow data (see earlier SI Tab.3, now moved to Table 3). Also, if we re-calculate the model biases with respect to GlobSnow derived with a density of 300 kg m-3, the biases change only slightly and the model ranking is unaltered. For your convenience, we list here the comparison of the RMSE for snow depth (cm):

|  |  |  |  |
| --- | --- | --- | --- |
| RMSE of snow depth (cm) | | | |
|  | GlobSnow | | Stations |
| Density=250 kg/m3 | Density=300 kg/m3 |
| CLM4.5 | 18.1 | 20.5 | 18.1 |
| CoLM | 22.1 | 24.9 | 21.4 |
| ISBA | 19.8 | 22.9 | 18.8 |
| JULES | 12.8 | 11.5 | 14.1 |
| LPJ-GUESS | 16.0 | 13.7 | 17.3 |
| MIROC-ESM | 14.0 | 13.7 | 17.9 |
| ORCHIDEE | 15.3 | 12.4 | 16.5 |
| UVic | 16.8 | 16.5 | 18.9 |
| UW-VIC | 20.0 | 22.6 | 19.8 |

**4. P7 L29: “the GlobSnow product can show regional differences (of ca. 0.5-5 cm) with biases increasing with increasing SWE” Is it a systematic bias in conversion of the SWE to snow height? If so, why not to take it into the account. Please specify where the regional differences are the largest, what areas have smallest. How many stations are located in the areas where differences are largest. Also, please list differences in % not the absolute values, since in some locations 5cm could be 25% of error vs 10% of error.**

We do not aim to evaluate the GlobSnow data. There are quite a lot of papers about details of GlobSnow validation and uncertainties. The given sentences are from the references. The paper is already rather long and we do not feel a digression into this data set is of interest to most readers. Again we emphasize that all the relationship analysis we do is based on the station data (Figs. 1-5). We use GlobSnow only for the evaluation of the simulated spatial maps of snow depth (Fig.6). This has been done to support the station data results, and to arrive at a more solid comparison with the models. And indeed, we show that the calculated model biases in snow depth are quite similar using either the station or the GlobSnow data (see earlier SI Tab.3; now *Table 3*). To make this clearer in the text, we now directly cite the sentence from the references and also include a reference related to our considered Russian Arctic or Eurasia region. We improved the according part in section 2.2: *“…GlobSnow-2 product (http://www.globsnow.info/swe/), which has been produced using a combination of passive microwave radiometer and ground-based weather station data (Takala et al., 2011). Orographic complexity, vegetation cover, and snow state (e.g. wet snow) affect the accuracy of this product. When compared with ground measurements in Eurasia, the GlobSnow product can show root-mean-square error* (*RMSE) values of 30 to 40 mm for SWE values below 150 mm, with retrieval uncertainty increases when SWE is above this threshold (e.g., Takala et al., 2011; Muskett, 2012; Khelemet et al., 2013*).”

**5. P8 L11-13: “We assume that there is relatively little impact due to soil moisture and texture between surface and 20 cm depth in winter.” This is probably true for the cold climate conditions, while in the warm climate the surface might freeze, but 0.20m could stay thawed.**

Yes, we consider here cold climate winter conditions. We follow the other reviewer’s suggestion to formulate it “*In winter, the effects of other factors (e.g. soil moisture, texture) on ∆T are much smaller than that of snow*.”

**6. P8 L13-14: “Although we recognize the difference between ground surface and 20 cm soil temperatures and that soil organic layer could play a role in certain locations”. What certain locations? In warm climate? Please be specific.**

We do agree that the snow effect is the key in winter cold climate. You (and the other reviewer) also emphasize, other factors like soil moisture, texture, organic matter are of secondary importance, and we agree on this. Therefore, we deleted this sentence to avoid confusion.

**7. P8 L17: “However, we find that the results do not significantly change when the model simulated temperature differences between ground surface temperature (GST) and near-surface air temperature (Tair) are used instead of between 20 cm soil (T20cm) and near-surface air temperatures.” Again please be specific. How much is not significant?**

This sentence is to confirm that our results of the functional behavior (Figs. 1-4) do not qualitatively change if we use either the ΔT=Tsoil-Tair or ΔT=GST-Tair. It is to tell the reader that we checked this. To make it clear, we do not have GST observations! Therefore, we have to use Tsoil. The question is how sensitive the results would be when using GST instead of upper soil-layer temperature (Tsoil at 20 cm depth). This would give an indication if/how soil characteristics are important for our presented relationships. The only thing we can do for this is to look into the model results. For your convenience, we show here one example plot (Fig. X2). The comparison of the red (ΔT=Tsoil-Tair) and green (ΔT=GST-Tair) clearly show that the regression does not materially change. “*To test how sensitive are results using 20 cm temperatures instead of ground surface, we also analysed model simulated temperature differences between ground surface and Tair, and found not qualitative differences, hence justifying use of 20 cm observations.”*

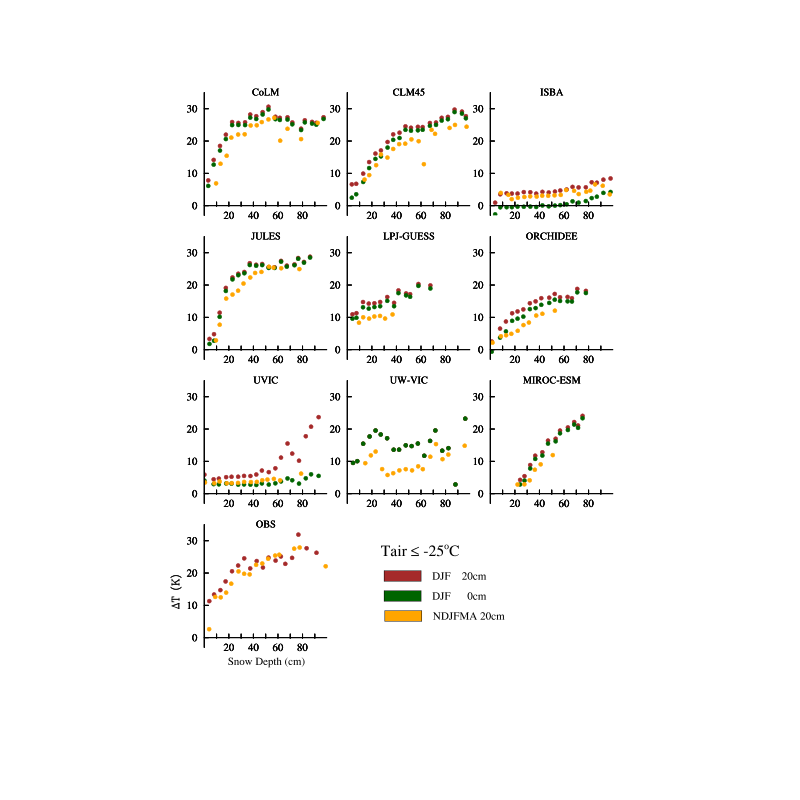
****

Figure X2: Variation of ΔT (K) with snow depth (cm) under cold conditions (Tair ≤ -25°C) for winter 1980-2000. The dots represent the medians of 5 cm snow depth bins, calculated from all Russian station grid points (n=268) and 21 individual winters. Red: ΔT is the difference between soil temperature at 20 cm depth and Tair, winter (DJF); Green: ΔT is the difference between ground surface temperature and Tair, winter (DJF); Orange: ΔT is the difference between soil temperature at 20 cm depth and Tair, winter (NDJFMA).

**8. P8 L4: “we checked that a different winter definition (NDJFMA) does not substantially change the results” What is substantial, please be specific.**

This sentence is to confirm that our results of the functional behavior (Figs. 1-4) do not change if we use another definition for the winter season. It is to tell the reader that we checked this. The question is how sensitive the results would be when using a different winter definition, because we know that snow can begin in November and end at the beginning of May. For your convenience, we show here one example plot (Fig. X2). The comparison of the red (winter-DJF) and orange (winter-NDJFM) clearly show that the regression does not materially change. We deleted “substantially” and substitute “qualitatively”. “*Our analysis is focused on the common winter (DJF) condition, although snow can begin in November and end at the beginning of May, but we checked that a different winter definition (NDJFMA) does not qualitative change any of the inter-variables relationships found*.”

**P3 L5-6: delete “as expressed by simulated differences”**

Done.

**P4 L5: Please try to avoid parentheses, usually it is possible to write the manuscript without them.**

Done.

**P4 L7-8: delete “quality” and parentheses.**

Done.

**P4 L14, “soil temperature”, reference is missing**.

The reference is provided just in the former sentence.

**P4 L22-23: Please rewrite without parentheses.**

Done.

**P8 L21: “We use correlation analysis to investigate the co-variability…”. Please provide formulae or references to this analysis**

Done. “*We use the Pearson product-moment correlation coefficient* *and its significance (von Storch and Zwiers, 1999)…* “

**P8 L22: “The input consists of detrended time series of winter means at each grid point.” How did you compute detrended time series? You may add details to the appendix.**

Done. “*Before we compute the correlations we detrended the data by removing a least squares regression line*.“

**P8 L26: “Student t-test” Reference?**

Done. “… *coefficients is estimated by the Student’s t-test (von Storch and Zwiers, 1999).*”

**P8 L28: “To further examine the functional behavior between different variables, we present relationship…”. What variables? Please be specific.**

Done. We introduce the key variables in the paragraph above. “… *4 key variables: near-surface air temperature (Tair), near-surface soil temperature (soil temperature at 20 cm depth; Tsoil), snow depth (dsnow), and the difference between Tsoil and Tair.”*

**P8 L31: “We illustrate the dependence on air temperatures by evaluating”. Dependence of what? What variable? Please be specific**

Done. “*We illustrate the dependence of ΔT vs. dsnow and Tsoil vs. dsnow relations for three Tair ranges*.”

**P9 L2-3: “The principal motivation for such classifications is to distinguish dry snow pack regimes from those where sporadic melt may occur even in winter.” Usually, the motivation or idea to do something goes first. The next sentence describes how this motivation is implemented. Please re-write this paragraph such that the motivation is at the beginning.**

Done. We revised this according to your suggestion. We start with “*To distinguish dry snow pack regimes from those where sporadic melt may occur even in winter, we split Tair into 3 regimes: the coldest conditions…* “

**P9 L11: “observations and all models produce a clear relationship” I do not see this "clear relationship" at the first glance in Figure 1. Please be specific. Describe what the reader may find on this figure and then state the relationship.**

Done. Fig. 1 clearly indicates that the observations as well as all models show an increase of ΔT with increasing dsnow. We improved this paragraph. First we start with the observations “*The air-soil temperature difference (ΔT) - snow depth (dsnow) relationship in winter (Fig. 1) shows in the Russian station observations an increase of ΔT with increasing dsnow. The data exhibit a linear relation between ΔT and dsnow at relatively shallow snow depths with a trend towards asymptotic behavior at thicker snow, which is in agreement with earlier findings (Zhang, 2005; Ge and Gong, 2010; Morse et al., 2011)…*“. Then we continue with the models behavior “*All models reproduce the observed relationship, i.e. increasing ΔT with increasing dsnow. However, Fig. 1 also shows a wide across-model spread in the simulatedrelationships, and that some of the models are not consistent with the behavior in the observations…”*

**P9 L14: “that some of the models are not consistent with the behavior in the observations. There is also significant scatter in the observation-based relationship, the inter-quartile range…”. Observations are just plotted on the last panel in figure 1. It takes time to find the observations on figure 1. Please try to re-design Figure 1 and all other figures such that observations stand out and could be easily noted. And please point out/circle this scatter on the figure.**

Sorry, we do not agree that it is hard to recognize the observations in the figures, though obviously this a personal issue. However, we consistently show observations in the bottom subpanel of our figures. To highlight the scatter in addition to the median (which is presented by the dots) we plot therefore upper and lower bars on these dots, indicating the 25th and 75th percentiles.

**P9 L21-24: “The Russian station data and some model results exhibit a linear relation...larger snow depths (Fig. 1), which is in agreement with earlier findings (Zhang, 2005; Ge and Gong, 2010; Morse et al., 2011).” Move this sentence up to the beginning of the above paragraph.**

Done. We largely revised the discussion in Section 3.1, such that we discuss observations first, and then followed by the model representation.

**P9 L31-32: “Figure 2 views the…relationship in the complementary form of the PDFs …different snow depth and air temperature regimes”. Before showing the PDF, please explain what you plan to analyze and how the PDF can help you to achieve this goal. It is really confusing for a reader to understand why PDF are now considered and plotted.**

Done. We revised the introduction part of this analysis: “*Figure 2 views the ΔT vs. dsnow relationship in a complementary form using the PDFs of ΔT for different snow depth regimes. This analysis allows a detailed evaluation of the snow regime-dependent ΔT separation by quantifying and comparing the modal value and width of the different conditional PDFs*.”

**P10 L1: “we divide the data into “shallow” (dsnow≤20cm) and “thick” (dsnow≥45cm) regimes”. Why do you choose these thresholds to characterize shallow and deep snow?**

Done. We explain that the Russian snow depth data show a non-normal distribution with a median value of 30 cm (SI Fig. 1). Therefore, we have chosen two classes, one with a threshold below (dsnow≤20cm) and one with a threshold above (dsnow≥45cm) of this median value. “*Since the Russian snow depths are clearly non-Normal in distribution (SI Fig. 1, with a median dsnow of 30 cm), we divide the data into "shallow" (dsnow ≤ 20 cm) and "thick" (dsnow ≥ 45cm) regimes to separate two snow depth regimes*.”

**P10 L3: “Based on the ∆T PDFs”. There are lots of line of figure 2. Maybe a revision to figure 2 is necessary. Could you please be specific how these five models separate the regimes, while others do not?**

Done. We agree that Fig. 2 was too busy. We improved the readability of Fig.2 by splitting it into two figures *Fig.2a and Fig.2b*. Fig.2a shows the PDFs for snow depth classes, while Fig.2b shows the PDFs for different air temperature regimes. We also use now the same colors for the different snow classes as in Fig. 4 to be consistent. The color of the different Tair classes is consistent in all relevant figures (Figs. 1, 2, 5). We improved the discussion of the separation of the PDFs: “*Based on the ΔT PDFs, five models (CoLM, CLM4.5, JULES, ORCHIDEE, MIROC-ESM) successfully separate the ΔT regimes under different snow depth conditions. Their simulated ΔT PDFs have a smaller modal value for thin snow than for thick snow, as in the observations. The other models clearly fail in separating the ΔT PDFs for the two different snow depth regimes.”*

**P10** **L9-17: This is mainly a description of observations. I would suggest to fold in this paragraph into the text right after the introduction of figure 1.**

Thanks, but we disagree because: First we discuss the general functional behavior of ΔT vs. dsnow relationship, i.e. increasing ΔT with increasing dsnow (Fig.1). Here we start with observations, followed by model evaluation. This is then further supported and discussed by Fig.2a. Next we discuss then the impact of Tair on the ΔT vs. dsnow relationship (different colored curves in Fig. 1 and Fig.2b). Therefore, we keep this paragraph there at its place, because this is the introduction/motivating paragraph, why we expect and look at the impact of Tair on the the ΔT vs. dsnow relationship. To make this clearer, we add at the end of this paragraph “*Therefore, we can expect that the same thickness of snow in colder climates will provide greater insulation than it would in warmer climates.*”

**P10** **L19-20: “The observations in Figs. 1 and 2 indicate that snow under colder climates have greater insulation than under warmer climates.” Please add an opening sentence to this paragraph. It would be great to give a reader a small hint about what is going to follow in the text below. This seems to be very typical for this manuscript, starting from the details and then reveal a motivation behind all of this.**

Actually, we do have an opening paragraph above these lines, which gives the motivation looking at the impact of Tair on the the ΔT vs. dsnow relationship. There we explain the hypothesis that snow under colder climates have greater insulation than under warmer conditions, based on known facts: ”*Both Figs. 1 and 2 further indicate that air-soil temperature differences are related to air temperature conditions. This is expected due to snow pack properties, particularly its density and moisture content, that affect the thermal conductivity of the snow. For example, the density of fresh fallen snow tends to be much lower under cold air temperatures than warm (Anderson, 1976), leading to increased insulation (larger ΔT). Snow densification is also a function of air temperature, for example, depth hoar metamorphosis of the snow pack, which produces more insulation (loosely packed depth-hoar crystals have very low thermal conductivity), is promoted by strong thermal gradients in the snow pack, and is typical of continental climates (e.g., Zhang et al., 1996).* *Therefore, we can expect that the same thickness of snow in colder climates will provide greater insulation than it would in warmer climates*.” This paragraph gives the motivation and hypothesis. Then, with the next paragraph we discuss that indeed our observational analysis confirm this expected impact of Tair on the ΔT vs. dsnow relationship. According to this comments, we improved the connection between the two paragraphs. The first paragraph is followed by “*Indeed, our observational analysis (Figs. 1 and 2) confirm this. This is shown by a larger ΔT for colder Tair than for warmer Tair (for a certain snow depth) and a greater sensitivity of ΔT to changes in dsnow (Fig. 1), and by the larger modal value of the ΔT PDF for colder Tair than for warmer Tair (21 K for Tair ≤ -25°C and 9 K for -15°C < Tair ≤ -5°C; Fig. 2b). This is consistent with colder climates having lower density snow packs, and the differences are in line with measurements of snow density variability (Zhong et al., 2013).”*

**P10 L32: “∆T/ dsnow”. On page 9, Line 25 there was a different notation for this relationship. I would suggest to make uniform notations. The sign "/" means division and maybe there is a better choice, e.g. \Delta{T}(d\_snow) P12 L16: “Tsoil /Tair” Maybe it is possible to find a better notation. The sign / means division. Could you use T\_obs(T\_air)**

Done. We agree with you. We changed all the “/” with “vs.” throughout the paper.

**P11 L5-7: “Our analysis (Fig. 1) indicates that some models (CLM4.5, CoLM, JULES) are better able to…”. Please move this section of the text close to figure 1, the reader now needs to go back in text. Everything related to figure 1 needs to stay close to figure 1.**

Thanks for the suggestion, but the structure as it stands seems clearer to us. First we discuss the general functional behavior of ΔT vs. dsnow relationship, i.e. increasing ΔT with increasing dsnow (Fig.1). Here we start with observations, followed by model evaluation. This is then further supported and discussed by Fig.2a. Next we discuss the impact of Tair on the ΔT vs. dsnow relationship (different colored curves in Fig. 1 and Fig.2b). Again, we start with observations, followed by model evaluation. Here we discuss in detail that part of Fig.1 which shows the snow-dependence on ΔT vs. dsnow relation (colored curves in Fig.1). To make this structure even more clear, we add one opening sentence before this model evaluation part starts. “*If we evaluate the models with respect to this observed impact of Tair to the ΔT vs. dsnow relationship, we demonstrate that some models (CLM4.5, CoLM, JULES) are better able to replicate the effect than others (LPJ-GUESS, MIROC-ESM, ORCHIDEE, UW-VIC) (Fig. 1)…”*

**P11 L22: “strong”. I also see lots of green and blue. I guess this is an indicator of weak correlation.**

Sorry, we think this is a misunderstanding; we were not clear enough. Here “strong spatial variability” refers to the pronounced spatial variability in the correlation. We see in Fig.3 in some regions redish color (high correlation) or in other regions greenish color (low correlation). We revised this sentence “*The maps of the ΔT vs dsnow correlations in winter (Fig. 3) demonstrate a pronounced spatial variability in the ΔT vs dsnow relationship*.” And, we re-ordered some text such that this sentence is then directly followed by the description of the details of regional variation in the correlation.

**P11 L23: “but indicate that most models agree on the general large-scale pattern.” Please describe what you see on the generated maps and only then state your conclusion. The first sentence is to describe what is going to happen in the rest of paragraph. It appears that the conclusions/details are always are beginning of many paragraphs. of course, there is a style to present a material this way, but it is only good for press-conferences when it is not enough time to describe assumptions, hypothesis, etc.**

Done. We improved this paragraph to make it clearer. We start with the most obvious finding that the maps of the ΔT vs. dsnow correlations show a pronounced spatial variability in this correlation. Then, we describe the details of the observed spatial variation. This is then followed by the model evaluation. Thus, we do have a clear and consistent presentation structure.

**P12 L9: delete “Previous authors” and parentheses.**

Done. We re-arranged this sentence. “*Previous studies have noted that the strength of relationship between Tsoil and Tair is modulated by dsnow and the snow insulation effect increases only up to a limiting depth beyond which extra snow makes little difference to soil temperatures (Smith and Riseborough, 2002; Sokratov and Barry, 2002; Zhang, 2005; Lawrence and Slater, 201*0).”

**P13 L18: “SI Fig.3”. L24: “SI Fig.4”. Bring this figure from the appendix to the text and describe how this figure is obtained. P15 L32: “SI Table 3”. There is no reason to move table and figures to SI when they are cited so often. Ideally, it should be no mentioning of figure from the SI.**

We moved SI Table 3 to the main text; it is Table 3 now. Both figures (SI Fig.3, SI Fig.4) are cited only once now and are supplementary information.

**P13 L20-21: “Obvious outliers in the Tsoil /Tair correlation maps (SI Fig. 3) are ISBA and UVic, which strongly overestimate the correlation (r > 0.9)”. What does it physically mean?**

Done. We improved this discussion. “*Obvious outliers in the Tsoil vs Tair correlation maps are ISBA and UVic, which strongly overestimate the correlation (r > 0.9) over most of the Arctic. This indicates an underestimated snow insulation effect, and confirms the weak insulation in both models, which has been initially discussed based on the underestimated ΔT (Fig. 1) and weak correlation between ∆T and dsnow (Fig. 3).”*

**P14 L10: “the biases range from -0.8 K to -4.7 K (SI Table 2)”. Please state that this is for winter months in the text.**

Done. “*The biases of winter mean air temperature ranges…”*

**P14 L25-26: “underestimate” and “overestimate”. Does it depend on the values of 250 kg/m3 used to convert the satellite measured SWE to snow depth.**

No, this is not the case. Please see our answer to the related comment #3. Furthermore, we show in Table 3 (earlier SI Tab.3) that the model biases are consistent regardless of which observations are used for the model evaluation. The model biases with respect to the station data are consistent with those with respect to the GlobSnow data; the model ranking is not affected.

**P14 L28-29: “The evaluation of the model performance for SWE compared to GlobSnow indicates the same bias characteristics as described here for snow depth (not shown).” This is really confusing... please re-write**

Done. We deleted this sentence.

**P15 L5-9: “Across-model differences in the interannual variability of winter precipitation do not translate simply to corresponding differences in the interannual dsnow variability (not shown).” This is another conclusion before the supporting statement. “For example, UVic calculates the (unrealistically) largest interannual dsnow variability in the boreal Europe permafrost region which is not reflected in the precipitation variability.” This is the supporting statement.**

This is a subjective criticism of style rather than substance. There are countless examples in the scientific literature of similar styles as we adopt here, that is a general point followed by a specific example. We do think this paragraph is clear.

**P15 L13: “We have shown that the across-model spread in the representation of snow insulation effects”. How did you show that? I think I lost something while reading Sections 3.1 and 3.2**

This done in Figs.1-5 as well as in Table2.

**P15 L15-16: “By considering the relationship plots and the conditional PDFs (Figs. 1, 2, 4, and 5)”. Add figure #s here after “plots” and not all the figures in parentheses are related to PDFs.**

Done. “*By considering the relationship plots (Figs. 1, 4 and 5), and the conditional PDFs (Fig. 2) we were able to classify the models*…”

**P15 L17: “sort the models in terms of their snow insulation performance.” Here “sort” means “classify”?**

Done. Yes, this is what was meant. For clarity, we replace “sort” by “classify”.

**P15 L20-29: It is better to move this paragraph to the introduction, where the employed models are described.**

Sorry, we do not agree with you. Here we describe those specific model characteristics and processes which explain why these models show a better performance than the others. It makes no sense to list which models have performed better in the introduction before the analysis is done.

**P17 L1-2: “The results are further improved by updating the snow albedo and snow densification parameterization”. References?**

Done. “*Decharme et al. (2016) still showed that the ISBA results are further improved by updating the snow albedo and snow densification parameterization.”*

**P17 L20-32: Move this paragraph into the above sections.**

Sorry, we think this is a misunderstanding. This paragraph clearly belongs to the summary and conclusion. In this paragraph, we summarize the main findings.

**P18 L1: “The primary aim of this study was to …”. I thought that "the aim" is stated on line 14, page 17. Please clarify.**

Done. We revised this sentence to “*This study uses the ensemble of models to document model performance with respect to Tsoil versus Tair relationships*...”

**P18 L9: “Those models which show limited skill in snow insulation representation …have some deficiencies”. Maybe over simplifications?**

Done. We revised this to “… *have some deficiencies or oversimplifications*...”

**P18 L24-30: The manuscript has not discussed the modeled area of permafrost prior to this paragraph. The conclusions are to summarize primary findings. Please move this paragraph into the main part of this manuscript.**

Done. We agree with you and the other reviewer and we added an additional section “*5. Permafrost area*” in the results section in the paper with the additional new *Tab.4*, which presents the simulated permafrost area. “*Snow cover plays an important role in modulating the variations of soil thermodynamics, and hence near-surface permafrost extent (e.g., Park et al., 2015). Here we evaluate if there is a simple relationship between the simulated Northern hemisphere permafrost area and the sophistication and ability of the snow insulation component in the LSM to match observed snow packs. The simulated near-surface permafrost area varies greatly across the nine models in the hindcast simulation (1960-2009; Table 4). Some of the better performing snow insulation effect models (CLM4.5, JULES) simulate a near-surface permafrost area of 13.19 to 15.77 million km2, which is comparable with the IPA map estimate (16.2 million km2) (Brown et al., 1997; Slater and Lawrence, 2013). CoLM and ORCHIDEE, identified as reasonable models with respect to snow insulation, simulate much lower (7.62 million km2) and higher (20.01 million km2) areas, respectively. The main deficiency of CoLM is its too small soil depth (3.4 m) compared with CLM4.5 (45.1 m) despite having very similar snow modules (Table 1). However, ISBA, one of the two models that showed rather limited skill in representing snow insulation effects, also simulates the highest permafrost area (20.86 million km2). This is inconsistent with previous studies (e.g., Vavrus, 2007; Koven et al., 2013) which concluded that the**first-order control on modelled near-surface permafrost distribution is the representation of the air-to-surface soil temperature difference. Table 4 shows that the situation is more complex and that snow insulation simulation is not the dominant factor in a good permafrost extent simulation. When the land surface models having poor snow models are eliminated, the remaining models’ simulated permafrost area show little or no relationship with the performance of the snow insulation component, because several other factors such as differences in the treatment of soil organic matter, soil hydrology, surface energy calculations, model soil depth, and vegetation also provide important controls on simulated permafrost distribution (e.g., Marchenko and Etzelmüller, 2013).****”***

Accordingly, we shortened the permafrost part in the “6. Summary and conclusions” section. The according paragraph reads: “*Snow and its insulation effects are critical for accurately simulating soil temperature and permafrost in high latitudes. The simulated near-surface permafrost area varies greatly across the nine models (from 7.62 to 20.86 million km2). However, it is hard to find a clear relationship between the performance of the snow insulation in the models and the simulated area of permafrost, because several other factors e.g. related to soil depth and properties and vegetation cover also provide important controls on simulated permafrost distribution.”*

**P26 L3: “Table 1. PCN snow model details.” Please specify how the upper 20cm of the soil column are resolved in each model. How is the heat capacity is calculated, may add another column.**

According to our analysis, and as also noted by Referee #1, it is the snow layer in winter that is the key here, not the soil conditions (moisture, organic layer). Therefore, we think the current table contains enough information for the discussion in this paper, with its emphasis on snow above the soil. The general structure and some key parameterizations of snow processes can explain the main deficiencies in the simulated results. The interested reader may find such soil details in 3 papers we cite in the introduction on the model general characteristics.

**P28 L4: “The dots represent the medians of…”. Please use different symbols. If this is printed black&white or someone is color blind, it might be hard to differentiate the results. P29 L3: “Conditional probability density functions…”. Why are they conditional? “color”. Same comments as for the above figure. Please think how to make this plots available to color-blind people. P30 L3: “Spatial maps of the correlation…” Same comment as above P31 L4: “The dots represent the medians…” Same comment as above. Maybe use different symbols for dots. P32 L4: “The dots represent the medians…”**

We do indeed sympathize with colour-blind readers. Personally journal should do more to encourage such consideration in their publications, but this is clearly an issue at the journal publisher level, not at the author or editor level. Our figure presentation is consistent with what is common in TC and other journals.

“Conditional probability” is a completely standard term in statistics. The “conditional” means the ∆T PDFs are PDFs for specific different conditions, namely for different conditions of air temperatures and snow depths.