Reply to Anonymous Referee #1

The manuscript describes a dataset of thermal conditions for permafrost-affected soils in Alaska. This complements an earlier published dataset providing deep ground temperature that was described by Clow (2013, ESSD). The compilation could be described as a value-added dataset which might be a more preferable term than synthesis dataset. The authors have gone beyond providing simply a compilation of raw data acquired from many sites and have calculated a number of key parameters and statistics. This value-added compilation will be useful to permafrost scientists, ecologists, hydrologists, engineers and practitioners as well as the modelling community. The manuscript provides a detailed description of the development of the dataset including the various processing steps and techniques used for quality control. It provides useful information that might serve as guidance for others that are compiling similar types of data for public dissemination. For these reasons, this manuscript should be published. The manuscript however, requires a bit of work before it is acceptable for publication. There are a few places in the manuscript where more explanation would be helpful. For example, some of the parameters utilized such as effective snow depth and SHTM require further explanation (see specific comments). Other sections that require further explanation are outlined in the specific comments. For the most part the manuscript is well written but some editing is required to improve language and increase clarity. Some suggestions for editorial revisions have been provided but the authors should thoroughly proofread the revised manuscript before submission. Although I have made several comments on the manuscript that I hope the authors will find helpful, dealing with them should not take much time. I expect that a revised manuscript that is acceptable for publication can be prepared within a reasonable time. I look forward to reading the published paper.

Specific Comments (keyed to page and line numbers)

P1,L1 – This is not a conclusion of this paper so it could be deleted.
Response: DONE.

P1,L15 – It is better to use "increasing" rather than "warming" when referring to temperature. Suggested revision "Continuous increases in near-surface air temperatures. . ." or alternatively you could say "Continuous warming at the ground surface. . .."
Response: DONE.

P2,L1 – Are you placing a dollar value on ecosystems?

Response: We mention dollar amount to explain the relevance of permafrost change, previous research provides an estimates on the potential damage (e.g. Melvin et al., 2017). We revised the wording to make clear that ecological impacts, while important were not included in the dollar estimate:

> Continuous increases in near-surface air temperature over the Alaskan Arctic causes warming and thawing of permafrost in Alaska, which is expected to continue throughout the 21st century with significant impacts on ecosystems, and estimated multi-billion dollar economic loss due to infrastructural damage.

P2,L13-17 – Are these really permafrost datasets or is soil temperature (or shallow ground temperature) dataset a better description given that the measurements may not necessarily be in permafrost.

Response: We agree not all of those dataset are necessary permafrost datasets. Brown et al. (1998) does not have ground temperatures, whereas the Russian and China soil datasets do not solely measure in permafrost-affected soils. We add the following clarifications:

> The permafrost extent map by Brown et al. (1998) is one of the most frequently and widely used metrics for comparing permafrost model results against ground-based data (Koven et al., 2015; McGuire et al., 2016). Another widely used dataset in model-data validation is the Russian Soil Temperature dataset of daily ground temperature measurements at different depths ranging from 0 to 3.2 m for 51 years (Sherstiukov, 2012). An additional ground temperature dataset includes daily-mean ground temperatures at various depths from 0 to 3.2 m at more than 800 stations in China which in selected locations dates back to the 1950s (Wang et al., 2015).

P2,L18-26 - There are other permafrost monitoring sites in Alaska and perhaps these should be mentioned. There are the measurements to about 20m that UAF collects and also the deeper temperatures collected by the USGS which have been published (see Clow 2013, ESSD). These could also be mentioned either here or in previous paragraph.

Response: We agree and added Clow (2013). Our dataset does include permafrost monitoring wells, but only those that are shallow, up to 3 m, not the data from deep boreholes. There are several publications on the temperatures from deep borehole datasets (e.g. Also Romanovsky et al. 2010; 2015). To clarify we added the following text:

> In addition to shallow borehole ground temperatures data (i.e. up to 3 meters) there are datasets that archive temperatures from the deeper (generally >5 m) boreholes (Clow, 2013; Biskaborn et al., 2015). Here we consolidated data from shallow borehole ground monitoring stations across Alaska from multiple government agencies. The importance of the shallow borehole data is that it can provide an immediate response to the changing climate in oppose to deep ground temperatures which would take more time to respond.

P2, L27 – "near-surface ground temperatures" or "shallow ground temperatures" might be better terminology.
Response: Done.

P2,L28 – revision suggested ". . ..from the three most reliable monitoring networks over the past several decades:. . ."
Response: Done.

P2,L30 – indicate at what depth the ground temperatures are measured, i.e. ". . .ground temperatures to x depth). . ."
Response: Done.

P2,L31 – revision suggested ". . ..for 72 stations. . .."
Response: Done.

P2,L31-34 – Consider reducing the use of first person. Eg. "Detailed information and meta-data are provided for the dataset. . ." "Futhermore, two types of data . . .. . .were implemented: (i) testing for inconsistencies. . ...; and (ii). . ..use of the snow. . ..."
Response: Done.

P3,L6-9 – I don't think you need to give the description of the CALM network as these data are not compiled in the dataset that is the subject of this paper. I suggest that this section be deleted. You can mention in the Introduction that the dataset you have compiled complements other permafrost relevant datasets compiled for AK such as CALM and USGS (see above) datasets. The focus in this section should only be a description of the sources for your data compilation.

Response: We moved the description of CALM data to Introduction section (Raw page 2, L12-17). The revised paragraph reads as follow:

> The permafrost extent map by Brown et al. (1998) is one of the most widely used metrics for comparing permafrost model results against ground-based data (Koven et al., 2015; McGuire et al., 2016). Another widely used dataset in model-data validation is the Russian Soil Temperature dataset of daily ground temperature measurements at different depths ranging from 0 to 3.2 m for 51 years (Sherstiukov, 2012). An additional ground temperature dataset includes daily-mean ground temperatures at various depths from 0 to 3.2 m at more than 800 stations in China, which for selected locations dates back to the 1950s (Wang et al., 2015). In addition to shallow borehole ground temperatures data (i.e. up to 3 meters) there are datasets that archive temperatures from the deeper (generally >5 m) boreholes (Clow, 2013; Biskaborn et al., 2015). Moreover, the Circumpolar Active Layer Monitoring (CALM) monitoring network measures active layer thickness (ALT) - the maximum soil depth above permafrost that thaws every summer and refreezes in the winter (Shiklomanov et al., 2008). Here, we consolidated data from shallow borehole ground monitoring stations across Alaska from multiple government agencies. The importance of the shallow borehole data is that it records the more immediate response to the changing environmental conditions, whereas deep ground temperatures take extensive time to respond."

P3, L13 – In figure reference (here and elsewhere) you can remove the symbol as this information should be in the figure caption or legend.
Response: Done.

P3, L14 – revision ". . ..USGS installed stations to monitor permafrost. . ."
Response: Done.

P3, L16 – revision ". . ..the USGS has maintained 17 automated. . ."
Response: Done.

P3, L17 – is "NPS has monitored ground temperatures since 2004" more appropriate?
Response: Done.

P3, L26 – P4,L6 – There is some repetition in this section and it is a bit confusing. You could say that thermistors are utilized to measure temperatures to depths of 1.5m and that these are embedded in a rod, anchored in a single hole or inside a fluid-filled hole. You could then describe the calibration procedure and give the accuracy (should also give precision). The details of the systems used in the 3 networks including thermistor type and temperature range, measurement depths and installation method could then be

summarized in a table along with any relevant publications for the particular network. The data acquisition system (datalogger) should also be mentioned as well as frequency of site visits for downloads.

Response: To clarify, we added a new table summarizing the thermistor information and revised the description in the manuscript accordingly.

Table 1. Summary of the ground temperature instruments used for ground temperature monitoring in Alaska, USA.

Network	**Temperature sensor**	**Datalogger**	**Measurement Depths (m)**	**Temperature Range (°C)**	**Accuracy (°C)**	**Maintenance visits**
USGS	MRC thermistor	CR10X or CR1000	Surface, 0.10, 0.20, 0.25, 0.30, 0.45, 0.70, 0.95, and 1.20 m (except for Lake145Shore, where was only 0.25 m)	-30 to 75	0.01	July, August
GI-UAF	Campbell Scientific 107	CR10x or CR1000	Surface to >1 m, but various in stations	-35 to 50	0.02	July, August
	MRC thermistor	CR10x or CR1000	Surface to >1 m, but various in stations	-30 to 75	0.01	July, August
NPS	Campbell Scientific 107	CR-1000 XT	Surface, 0.10, 0.20, 0.50, 0.75, and 1.00, but various in stations	-35 to 50	0.02	July, August

P4,L4-6 – It would seem that you know that the probes are not well anchored in permafrost and the change in the "stickup" is due to heave rather than settlement of the ground in response to permafrost thaw (which might be the case if your probes extended to greater depths in the permafrost). Heave of the probe would take place over the winter as the freezing occurs and I assume you make the correction in the summer (although details are not provided). One might question how reliable your winter temperatures are in terms of the depth of measurement. More detail should probably be provided with respect to the amount of heave that occurs annually as well as how the temperatures are corrected.
Response: Correct; sensors mainly heave over winter time. Observations are corrected for sensor heave by the data contributors.

> "The USGS and NPS teams estimate frost heave by using ground temperature data from the topmost thermistor (at 5 or 10 cm depth). If temperature of the top thermistor during thaw-period exceeds air temperature then the sensor is considered exposed or partly exposed to solar radiation. The GI-UAF team measures frost heave at every site and then subtract heave depth from known sensors depth to correct for heaving Romanovsky et al.,

> (2008). Each team corrects for heaving every summer, and corrections are applied before releasing data."

P5, L11-12 – Do you correct for the vegetation effect? Trim the vegetation?
Response: Typically, vegetation remains undisturbed, as trimming vegetation would alter local conditions and add a bias to the measurements.

P5,L21 – revision suggested ". . .compile the dataset."
Response: Done.

P5,L23 – Does this mean that you might lose the 1m depth at sites where there has been significant heave of the probe? Minor revision suggested ". . .beyond the maximum observed depth. . ."
Response: Heaving of the ground depends on soil composition. In most cases ground does not heave much. However there are locations where heave is significant so that the shift of the initial location of sensors results in loosing measurements at 1 m depth. Minor revision was done.

P5,L25-26 – revision suggested ". . .models are monthly, the monthly means were calculated for all variables, including air and. . ..."
Response: Done.

P5, L26-27 – "Annual means were also calculated to allow. . .." Do you mean relationship between air and ground temperatures?
Response: We clarify:

> "In addition to monthly data, annual means were calculated to allow evaluation of the relationship between air and ground temperatures."

P5, L31 – Is the frost number determined for only the ground surface temperature or at each depth? Also, do you include the freezing and thawing degree day indices in the dataset as these are useful for models etc.
Response: We calculated the frost number indices for air and all available ground temperatures (as shown in Fig. 5 in the original manuscript). These results are attached as a value-added data in the revised dataset.

P6, Eq 1-3 – For DDF are you using a complete winter/freezing season (e.g. Oct – May). You should probably provide a bit more explanation.
Response: We calculate DDF following Nelson's method (1987), which uses monthly mean temperatures. Site-dependent freezing/thawing periods are then determined from the temperature curve. Comparison between results based on daily versus monthly temperatures showed that they are in good agreement ($R^2 > 0.99$) (Nelson, 1987). We added more explanation on this calculation:

> "Freezing (DDF) and thawing (DDT) indices were derived from monthly temperature curve. Nelson et al. (1987) and Zhang et al.(1996) have demonstrated from datasets with daily time resolution that results using monthly data correspond closely to results from daily data analysis."

P6, L6 – revision suggested "Data records from many sites have gaps. . ." Also, equipment malfunction is another problem that may result in data gaps. One thing that is not mentioned is the frequency of site visits.

Response: We added the timing of annual visits in a new table (Table 1 here, or Table 3 in the revised manuscript), visits were at least once per year in summer for all sites.

Table 1. Summary of ground temperature instruments from USGS, GI-UAF, and NPS networks of Alaska, USA.

Network	Temperature sensor	Datalogger	Measurement Depths (m)	Temperature Range (°C)	Accuracy (°C)	Maintenance visits
USGS	MRC thermistor	CR10X or CR1000	Surface, 0.10, 0.20, 0.25, 0.30, 0.45, 0.70, 0.95, and 1.20	-30 to 75	0.01	July, August
GI-UAF	Campbell Scientific 107	CR10x or CR1000	Surface to >1 m, but various in stations	-35 to 50	0.02	July, August
	MRC thermistor	CR10x or CR1000	Surface to >1 m, but various in stations	-30 to 75	0.01	July, August
NPS	Campbell Scientific 107	CR-1000 XT	Surface, 0.10, 0.20, 0.50, 0.75, and 1.00, but various in stations	-35 to 50	0.02	July, August

P6, L7-15 – For the missing data allowance, was there any consideration given to varying this according to the particular variable and its short-term variability. The deeper ground temperature would exhibit less variable so perhaps there could be allowance for more missing data than air temperature for example.

Response: We have indeed considered whether we set different criteria for missing data for different depths, particularly for deeper layers. However, it is difficult to determine a suitable threshold allowance for each site and depth. Ground temperature variability generally declines with depth, but the dampening rate is different between sites. Thus, for dataset consistency we applied the same threshold for missing data for all depths.

P6,L17 – P7,L7 – This section could probably be simplified and shortened. Maybe you could say that a unique name is assigned to each site. You could briefly mention how you deal with replacement sites.
Response: The revised section:

> "A unique name is assigned to each site. The Deadhorse monitoring site maintained by GI-UAF, and the Awuna site maintained by USGS have new monitoring stations, and the

old ones have been decommissioned. The new and retired systems simultaneously ran for a few months to evaluate the data consistency."

P7,L11 – Effects of snow on ground thermal state – is this validation or analysis?
Response: Good point, it is a validation. In our consideration, the effects of snow on ground surface temperature are basically to examine the physical mechanism among air temperature, snow cover and ground thermal states, i.e., seasonal snow cover ought to act to keep the ground warm.

P7,L21-22 – revision suggested ". . .keep the ground warm by reducing cooling (or heat loss) during the winter"
Response: Done.

P7,L23 – revision " . . .snow depth and soil thermal properties."
Response: Done.

P7,L24 – There is no snow cover outside of Oct-Mar for even more northerly locations?
Response: While there is some snow cover outside of the period Oct. through March, we selected this period to be consistent with an existing metric on snow cover effect by Slater et al. (2017).

P7,L24-30 – This section is a bit confusing and more information/explanation should probably be provided especially since the parameters mentioned are specific to Slater et al. (2017) and may not be familiar to many readers (i.e. use parameters like nfactors, offsets to describe effect of snow etc.). How is SNDeff determined. Is it represented by one of the terms in Eq 4? Is SHTM equivalent to deltaAmpnorm? Is Ampgrnd referring to ground surface temperature, since snow cover will influence surface temperature, whereas the damping effect at depth will be more dependent on ground thermal properties.
Response: SNDeff is determined by the average snow depth during Oct. through March, which is not represented in Eq.4. Eq.4 -6 are used to estimate the normalized amplitude difference between air and ground surface temperature. SHTM is the correlation between ΔAmpnorm and SNDeff. We added more explanation on these definitions:

> "We averaged the snow depth measurements over the period from October through March to obtain the effective snow depth (SND eff) (Slater et al., 2017). The amplitudes of air temperature (Ampair) and ground surface temperature (Ampgnd) were calculated following Slater et al. (2017), for those stations with available snow depth data. The snow and heat transfer metric (SHTM) captures the correlation between the normalized temperature amplitude difference (ΔAmpnorm) (i.e., Eq.4-6) and SND eff. As a physical mechanism examination, SHTM and ΔAmpnorm were only referring to ground surface temperature, because damping effect with depth (i.e., the difference between air temperature and ground temperature at deeper soil) is more dependent on ground thermal properties."

P8, L11 – "spatially variable" better than "spatially complex"

Response: Done.

P8,L12 – delete last part of sentence "according to the synthesis dataset" – not necessary as it is shown in the figure that is derived from your dataset.
Response: Done.

P8,L13 – You could just say "located near the glacier"
Response: Done.

P8,L16 – revise "The other two sites,. . ..."
Response: Done.

P8,L17-18 – Did you mean to include this last part of the sentence? You could make a comment that the thin snowcover is due to wind exposure.
Response: Done.

P8,L19 – This is the Frost Number calculated by Eq (1)? I don't see this value in table 3 only the freezing and thawing degree day indices.
Response: Yes, the Frost Number is calculated by Eq.1. We added the values in the revised dataset.

P9,L1-14 –Wouldn't the comparison of trends for ground temperature at various depths be the most important thing to check for sensor drift etc. (i.e. ignore any snow effects and focus on propagation of temperature wave with depth).
Response: Yes, it is possible to detect whether there was a systematic issue, such as sensor moving, we slightly revised Fig. 7 to allow more reliable comparison of trends.

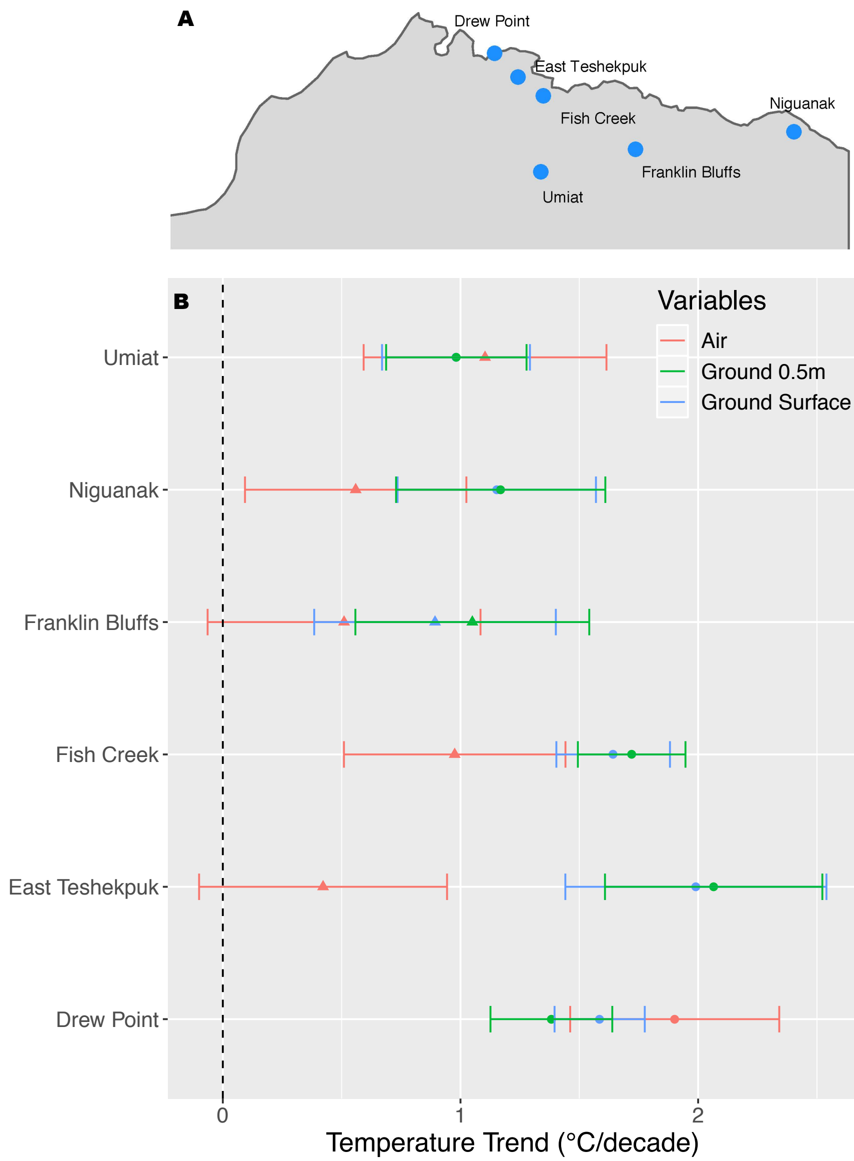

Figure 7. (A) Stations with at least ten years of identical period of air, ground surface and ground temperature at 0.5 m. (B) Trend comparison of air temperature, ground surface temperature, and ground temperature at 1 m over 1997-2016.Trends were estimated only for those stations comprising at least ten years of data. Error bars represent standard errors from the linear regression analysis. Circles indicate trends with p value <=0.05, triangles indicate trends with p value >0.05.

P10,L13-15 – See earlier comment regarding more explanation required for these parameters (SHTM, effective snow depth).

Response: We added more explanation on this:

> "Across these stations, effective snow depth was generally less than 0.4 m. The normalized temperature amplitude difference (ΔAmpnorm) that calculates the

> temperature difference between air and ground surface shows a positive linear relationship with effective snow depth. This correlation, so-called SHTM (Slater et al., 2017), implies snow insulation effects increase with effective snow depth, which is consistent with previous studies (Burn and Smith, 1988; Demezhko and Shchapov, 2001; Zhang, 2005; Morse et al., 2012; Slater et al., 2017). In addition, while snow is considered an important factor on winter ground temperature, vegetation can also effect the amplitude through their influence on summer temperature."

P11 Figure 6 – Labels on Y axis overlap between graphs. The trend requires correct units, degC/year, m/year. Are you showing standard error of the estimate also on the graph (should mention in caption)

Response:. Fig.6 is corrected. The shadow shows the standard error of the linear regression estimates.

P12, Figure 7 – Error bars represent standard error from the regression analysis?
Response: Yes, we now explain this in the Figure caption.

> Figure 7. (A) Stations with at least ten years of identical period of air, ground surface and ground temperature at 0.5 m. (B) Trend comparison of air temperature, ground surface temperature, and ground temperature at 0.5 m over 1997-2016.Trends were estimated only for those stations comprising at least ten years of data. Error bars represent standard errors from the linear regression analysis. Circles indicate trends with p value <= 0.05, triangles indicate trends with p value >0.05.

P13,L1-4 – While snow is an important factor and influences the winter ground temperature (and therefore the amplitude), vegetation and ground cover can also effect the amplitude through their influence on summer temperature. Is this part of the reason for the considerable scatter in your graph?

Response: Thank you for your suggestion. We revised this paragraph as shown earlier to include a statement about the effect of vegetation.

P13,L5-6 – Delete this – repetitive.
Response: Done.

P13,L8 – It is more correct to say "Changes in near-surface ground temperatures over time are important indicators of a changing climate" The direction of the change in ground temperature will indicate whether there is warming or cooling.
Response: Done.

P13, L8-18 – Will the database be periodically updated as new data are collected? You mention it is worth maintaining but you could say more regarding plans for updates.

Response: Our goal was an outline all important details about the current dataset complication, in-time updates of the dataset will mainly depend on the funding of the monitoring stations as well as our own funding.

Table 1 – In section 2.2, interpolation to determine ground temperature for 4 target depths is mentioned. In the table, reference is made only to 1m. Are they any statistics calculated for the other depths? It isn't clear from the table or section 2.2.

Response: An important synthesis objective was to make this data useful for the modeling community as a benchmark dataset. As we mention, data had been collected from different depths and these intervals were not consistent across the stations. Now all data are consistently presented for 4 depths.

Reply to Anonymous Referee #2

The dataset described in this manuscript is certainly useful, the manuscript, however, in its current form is not. There is too much ambiguity and missing information about the dataset, the language and organization are confusing in many places, and the presentation is a bit sloppy (especially the figures). This makes the utility of the dataset difficult to assess. The organization of the results section is strange and the paper leaves it unclear what all variables are actually included in this dataset and which ones are just presented for some type of qualitative validation.

Major points:
1. It is not clear what the dataset actually is. The introduction (p. 2 lines 30-31) says it is *measured* air and ground temperatures, snow depth, and soil volumetric water content. But then later other variables like frost numbers, thawing index, and freezing index are mentioned. It is also later written that the data are provided as interpolated values (p. 5 lines 22-23), which is very different from measured values. At the very least, readers should come away from this paper knowing clearly what the dataset actually is.

Response: We understand reviewer's frustration and made the corresponding changes to better address this issue in the paper. One main goal of the paper is to present a datasets useful for the modeling community. The air temperature, snow depth, soil water content, and ground surface temperature are all indeed measured values. Other presented variables are 'value-added' data which serve as additional metrics for the modelers to do model validation. Most of the time is not possible to compare data to model outputs directly (due to difference in resolution and physical parameterization). Instead it is useful for model validation to use the additional derived metrics.

We interpolate ground temperature data with depth, because in-situ observations of ground temperatures from the Alaskan Arctic region have been dispersed over different monitoring efforts, and there are no standard protocols for measuring depths. A revised, more detailed description on the interpolation has been added in the revised manuscript. Considering the unclear information, in this revision, we added flags to the variables. We believe this version should be more clear to what the dataset is.

2. A shortcoming of the dataset is the monthly timescale. While this may be OK for model validation, this is a very limited audience and most non-modelers would probably prefer the daily data. From a practical standpoint, you are disincentivizing users to turn to your synthesis product, given that the daily data are already readily available from the original UAF, USGS, and NPS sources. For example, based on the data you are providing I could not use it to quantify many processes that occur on shorter timescales, such as the onset of thaw or freeze, snowmelt timing, etc.

Response: Absolutely, higher temporal resolution captures much more seasonally or even hourly information. We aim to synthesize measurements from different observation teams, and use statistical methods to create a relatively uniform and easy-to-use database of monthly mean values of soil-related variables. This product can then be used to evaluate land surface model simulations or to investigate inter-annual variability and trends. Whereas daily or hourly resolution may ultimately be preferable, it would require massive additional statistical analysis to cope with missing data values at such increased resolutions. We advocate that ground temperature data at a monthly scale dataset is a proper start for data-model comparisons. And yes, whereas the original data are available from individual sources, it would still be complicated to make inter-comparisons since each group has their own protocols, station design, and data achieving methods, our work allows users to not have to do a uniformization themselves. Our synthesis data efforts not only assist with model evaluation, it may also benefit the discussion for better standards for data collection.

3. The dataset is supposed to be a synthesis of near-subsurface ground temperature data (p. 2 line 27). However, the "Overview of this dataset" section focuses on volumetric water content, snow depth, and frost number. It is very strange that you are providing an overview of some of the peripheral and derived variables, but not of the primary variables that make up the dataset. In fact, the entire final paragraph of section 3.1 should be deleted, because it suddenly presents research results, as opposed to describing and showcasing the dataset.

Response: We added an overview of primary variables, i.e., ground temperatures, including one new figure (Figure 4 in the revised manuscript or the following figure) and an explanatory paragraph in section 3.1. Please note that our title is ‘thermal conditions’, thus our focus is not solely on temperature, but instead present several permafrost related variables.

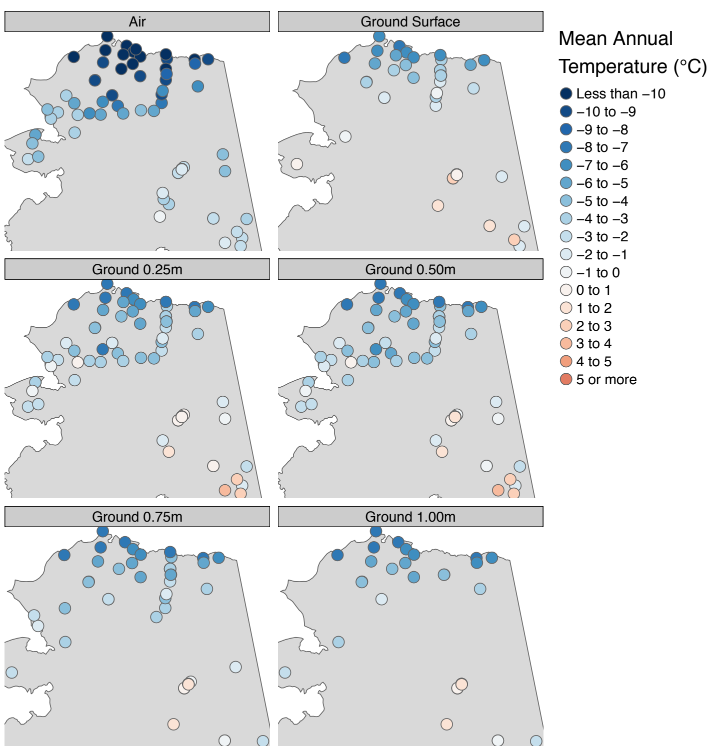

Figure X. Overview of spatial distribution of mean annual air temperature, ground surface temperature, and ground temperatures at 0.25 m, 0.50 m, 0.75 m, and 1.00 m.

> “Overall, mean annual air temperature is less than -10 C in the Alaskan Arctic while close to freezing point (-0.5 C at RUGA2 site) in the southern regions. Mean annual ground surface temperature for 46 available sites ranges from -7.6 C through 2.5 C,

which is higher than air temperature. Many sites comprise measurements of ground temperature at 0.25 and 0.50m, 69 and 67 sites respectively. The range of ground temperature at 0.25m and 0.50 m are roughly from -7.8 to 3.3 C. Mean annual ground temperature at 0.75 m varies from -7.5 to 1.2 C over 49 available sites. Our data comprises only 32 sites with ground temperature at 1 m, preferably located in the southern region of the Alaskan Arctic (~ 62N), and the range of mean annual ground temperature at 1 m is -7.8 to 1.2 C."

4. A crucial missing piece of this dataset is metadata information about the soil itself (soil type, density). This information is already available because, as stated (p. 5 line 4-5), at least the GI-UAF sensor installation was dependent on the soil profile and texture. Given the heterogeneity of soils and therefore its thermal conditions, adding this to the dataset would make it vastly more useful.

Response: This is a valid point. The main focus of this paper is description of the geophysical data related to the permafrost measured at the permafrost monitoring stations. The geological data is harder to obtain. Typically it is not available online and not always well documents due of the mixing of the soil layers. These complications makes it hard to add that data (like soil bulk density) in the current datasets. Currently, we added sentences and a new table (Table 2 in revised manuscript) to summarize available meta-information. Due to the different field work design of various teams, the soil and vegetation description may not fully comparable and not available at each sites.

Name	Vegetation	Soil Type
Drew Point	Moist-meadow, tussock-tundra complex	Silt
Fish Creek	Moist-meadow, tussock-tundra complex	Silt
Inigok	Moist-meadow, tussock-tundra complex	Silt
Tunalik	Moist-meadow, tussock-tundra complex	Silty Sand
Umiat	Moist-tussock tundra	Silt
Barrow 2	Graminoid-moss tundra (wet and moist acidic)	Typic Histoturbel, Typic Aquiturbel
Boza Creek 1	Open black spruce forest	Pergelic Cryaquepts
Boza Creek 2		-
Chandalar Shelf	Alpine meadow with low shrubs	Ruptic-Histic Aquiturbel
Deadhorse	Graminoid-moss tundra and graminoid, prostrate-dwarf-shrub, moss tundra (wet and moist nonacidic)	Terric Aquiturbel
Franklin Bluffs	Graminoid-moss tundra and graminoid, prostrate-dwarf-shrub, moss tundra	Ruptic-Histic Aquorthel
Franklin Bluffs Wet	Graminoid-moss tundra and graminoid, prostrate-dwarf-shrub, moss tundra	-
Galbraith Lake	Graminoid-moss tundra and graminoid, prostrate-dwarf-shrub, moss tundra (wet and moist nonacidic)	Ruptic-Histic Aquiturbel
Happy Valley	Tussock-graminoid, dwarf- shrub tundra and low-shrub tundra (moist acidic)	Ruptic-Histic Aquiturbel
Imnaviat	Tussock-graminoid, dwarf- shrub tundra and low-shrub tundra (moist acidic)	Typic Histoturbel, Typic Aquorthel
Ivotuk 3	Horsetail-rich variation of nonacidic tundra	-
Ivotuk 4	Moss dominated	-
Sag1 MNT (Moist Non-Acidic Tundra)	Moist nonacidic tundra	Pergelic Cryaquolls (43%), P. Cryaquepts (18%), P. Cry-oborolls (14%), others (25%)
Sag2 MAT (Moist Acidic Tundra)	Moist acidic tundra	Pergelic Cryaquepts (79%), Histic Pergelic Cryaquepts (21%)
Selawik Village	Upland DwarfBirch-Tussock Shrub	-
Smith Lake 1	White Spruce forest with high canopy	-
Smith Lake 2	Dense diminutive Black Spruce forest	-
Smith Lake 3	Forest surrounded by Black Spruce trees and tussock-shrubs	-
Smith Lake 4	Hummocks of sedges (tussocks) and shrubby vegetation with sparse Black Spruce	-
West Dock	Moist to wet tundra	Typic Aquahaplel
ASIA2	Dryas octapetala	Lithic Haplogelept
DVLA2	Arctagrostic latifolia, Petasites frigidus, Carex bigelowii, Empetrum hermaphroditum, Ledum palustre, Vaccinium uliginosum, Arctous alpina, Hylocomium splendens, Lupinus arcticus, Salix pulchra	Aquic Molliturbel
ELLA2	Umbilicaria, Alectoria migricans, Carex	Typic Haploturbel
HOWA2	Dryas octopetala, Salix phlebophylla	Typic Gelorthent
IMYA2	Dryas octopetala, Hierchloe alpine, Salix phle-bophylla	Typic Gelorthent
KAUA2	Dryas octopetala, Vaccinium uliginosum	Typic Gelorthent
KUGA2	Betula, Empetrum hermaphroditum, Ledium palustre, Vaccinium vitis-ideae	Typic Gelorthent
MNOA2	Dryas integrifolia, Potentilla biflora	Typic Haploturbel
SRTA2	Betula, Ledum palustre, Loiseleuria pro-dumbens, Stereocaulon, Flavocetraria cucul-lata, Vaccinium uliginosum	Typic Haplogelept
SRWA2	Betula, Dryas octopetala	Typic Gelorthent
SSIA2	Dryas octapetala, Arctous alpinus, Lupinus arcticus, Rhytidium rugosum	Typic Haplorthel
TAHA2	Betula, Dryas octapetala, Vaccinium uligi-nosum, Salix phlebophylla	Typic Gelorthent
UPRA2	Betula, Empetrum hermaphroditum, Ledum palustre, Picea glauca	Typic Dystrogelept

5. I question the linear interpolation method employed. Other studies, for example, Sherstiukov (2009) and Streletskiy et al. (2008, 2015) found that a polynomial fit better captures the exponential attenuation of temperature with depth. Regardless, this interpolation must be described with much greater detail if the entire dataset is based on this interpolation. How many soil depth observations were required for the interpolation? Did you only interpolate between 'adjacent' soil depths? Or, for example, if a USGS location only had a 5 cm and a 1.2 m observation, did you still interpolate and provide temperature at 0.25, 0.5, 0.75, and 1 m? Was the interpolation done on the raw hourly or daily data, or on the final monthly data? Does the final, published dataset only contain these interpolated values, or also the original ones? Do you distinguish between observed and interpolated data in the official dataset? These are all crucial details that cannot be omitted.

Response: We addressed this valuable comment in the following paragraph, and believe it allowed for an important change in clarity:

Previous studies, such as Sherstiukov (2009), Streletskiy et al. (2008), and Streletskiy et al. (2015), used non-linear interpolation methods to determine temperature-depth profiles by interpolation of typically 3-4 measurements of soil temperatures over a relatively large range of soil column (up to 3.20 m below the ground surface). Non-linear methods possibly have its advantages in these deeper subsurface cases, but Sherstiukov (2009) and Streletskiy et al. (2015) also suggested that a polynomial interpolation could provide a better fit.

In this study, we focus on the uppermost 1m which is a relatively shallow depth interval. We implemented any interpolation requiring measurements for at least 4 depths, which assures a relatively small interval around the target depths. In addition, soil temperatures were not extrapolated. In other words, ground surface temperature is only calculated when supporting measurements are indeed available.. Then, the calculated soil temperature at a specific depth depends on linear slope between just the observations at adjacent depths. Therefore, using a linear interpolation method does not result in a linear prediction from ground surface to 1 m, even if that probably is rather reliable fit. Here, we show an example at Drew Point from USGS network (Fig.1).

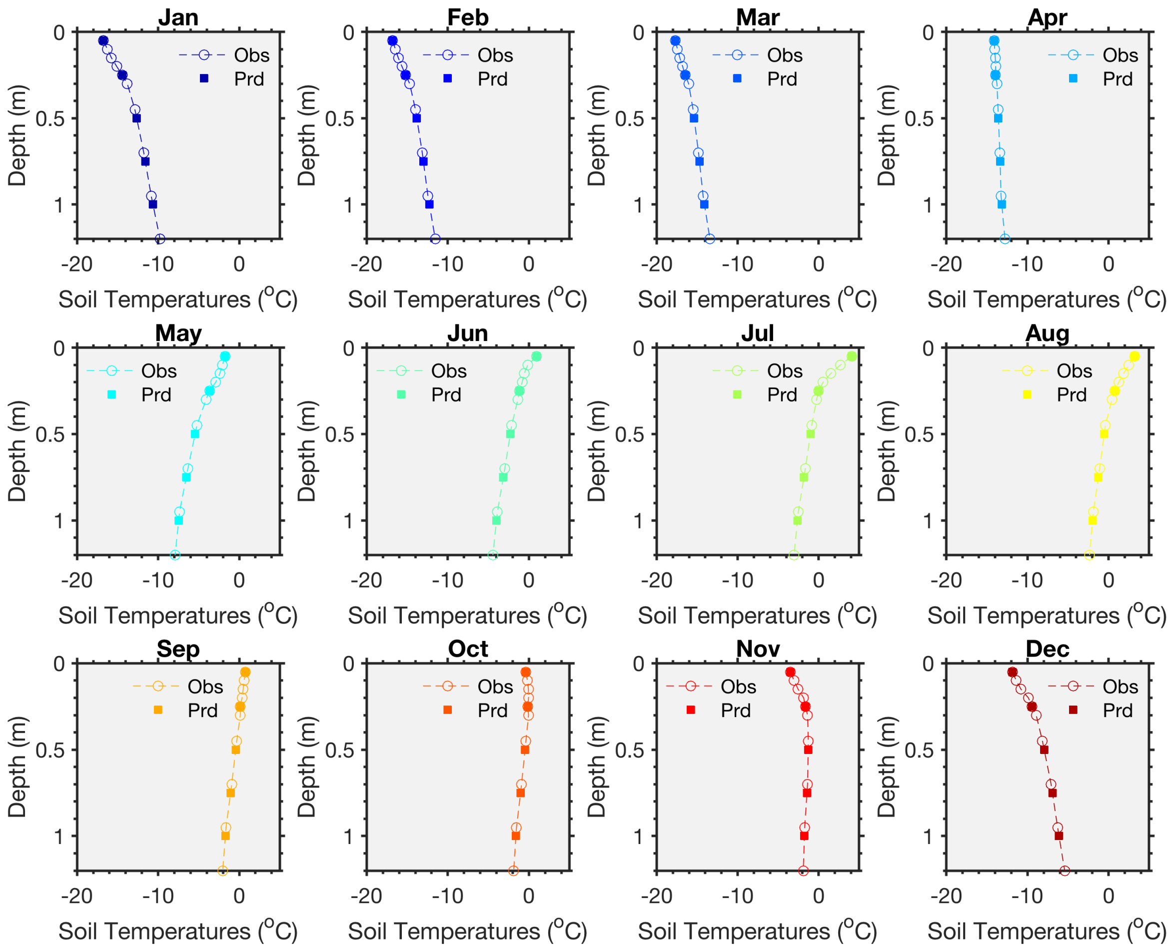

Fig.1 Monthly temperature observations (open circles) at the USGS Drew Point site for 2014 are displayed to illustrate the interpolation technique employed to arrive at data set values for specific depths; ground surface, 0.25 m, 0.50 m, 0.75 m and 1.0 m (solid circles).

Furthermore, we examined the uncertainty resulting from our linear interpolation method for the most data sparse case, i.e. when we only have observations at four depths. To do so we selected the entire year of data without any missing values and depths and used linear interpolation to predict temperatures at five depths (similar to Fig.1). Then we randomly selected only four depths, and interpolated again by using these four depths. While missing depths would reduce the number of available interpolation results (here is ~34% missing interpolated results), the influence from missing depths is limited (for this specific case is ~0.1 C). This analysis makes us more confident that for our shallow profiles linear interpolation between the adjacent depth observations is satisfactory.

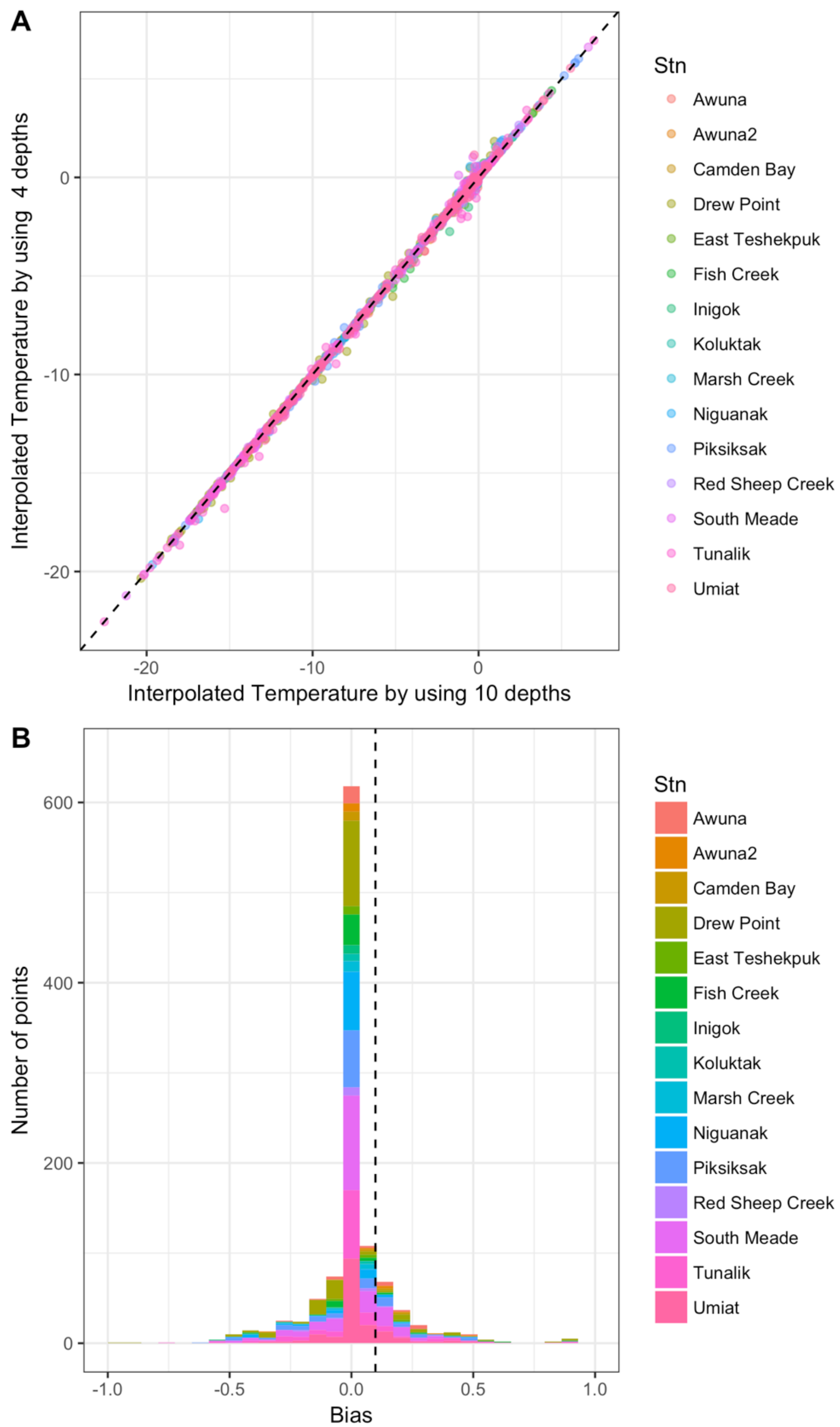

Fig.2 Proof of concept analysis to investigate the influence of missing depths on interpolated predicted temperatures at 4 distinct depths. The entire sample consist of 1725 data points in case we use all 10 depths measurements, while interpolations with missing depths consist of 1123 data points (or 65% of original sample). (A) Pairwise comparison between interpolated soil temperatures using all 10 depths measurements versus using randomly selected 4 depths. (B) Statistical distribution of bias from panel A. The black dashed vertical line represents the mean absolute bias (~0.1 C).

Finally, we added more sentences to clarify the interpolation process in this study:

> We implemented any interpolation requiring measurements for at least four depths, which assures a relatively small interval around the target depths. In addition, soil temperatures were not extrapolated. In other words, ground surface temperature is only calculated when supporting measurements are indeed available. Then, the calculated soil temperature at a specific depth depends on linear slope between just the observations at adjacent depths. Therefore, using a linear interpolation method does not result in a linear prediction from ground surface to 1 m, even if that probably is rather reliable fit. Furthermore, we examined the uncertainty resulting from our linear interpolation method for the most data sparse case, i.e. when we only have observations at four depths. To do so we selected the entire year of data without any missing values and depths and used linear interpolation to predict temperatures at five depths. Then we randomly selected only four depths, and interpolated again by using these four depths. While missing depths would reduce the number of available interpolation results, the influence from missing depths is limited.

6. Because this dataset is comprised of interpolated data (p. 5 line 22-23) anyway, why not also interpolate to fill in missing observations (or did you)? In cases where there are only one or two days of missing soil temperatures at a certain depth, you could relatively reliably fill in that gap based on the average of the previous and next day's observations. And if a certain soil depth has a missing observation but the immediately adjacent upper and lower depth observation is available, the missing depth could also be interpolated.

Response: Thank you for your comments and suggestions. We only interpolated soil temperatures at certain depths (i.e., 0.25, 0.50, 0.75, and 1.00 m) by using adjacent upper and lower depth measurements, while we did not fill in missing data in the time series. In many previous studies, the method mentioned has been applied to fill small gaps over time. We considered doing this, but refrained from interpolation in time. Our main reason is that all observations were obtained from automatic data logging stations, missing data generally indicate sensor damage. There were very few data gaps that spanned only one or two days. In the figure below we illustrate our point for a few sites (Fig.3). For any of such cases, data filling would not result in a significant improvement.

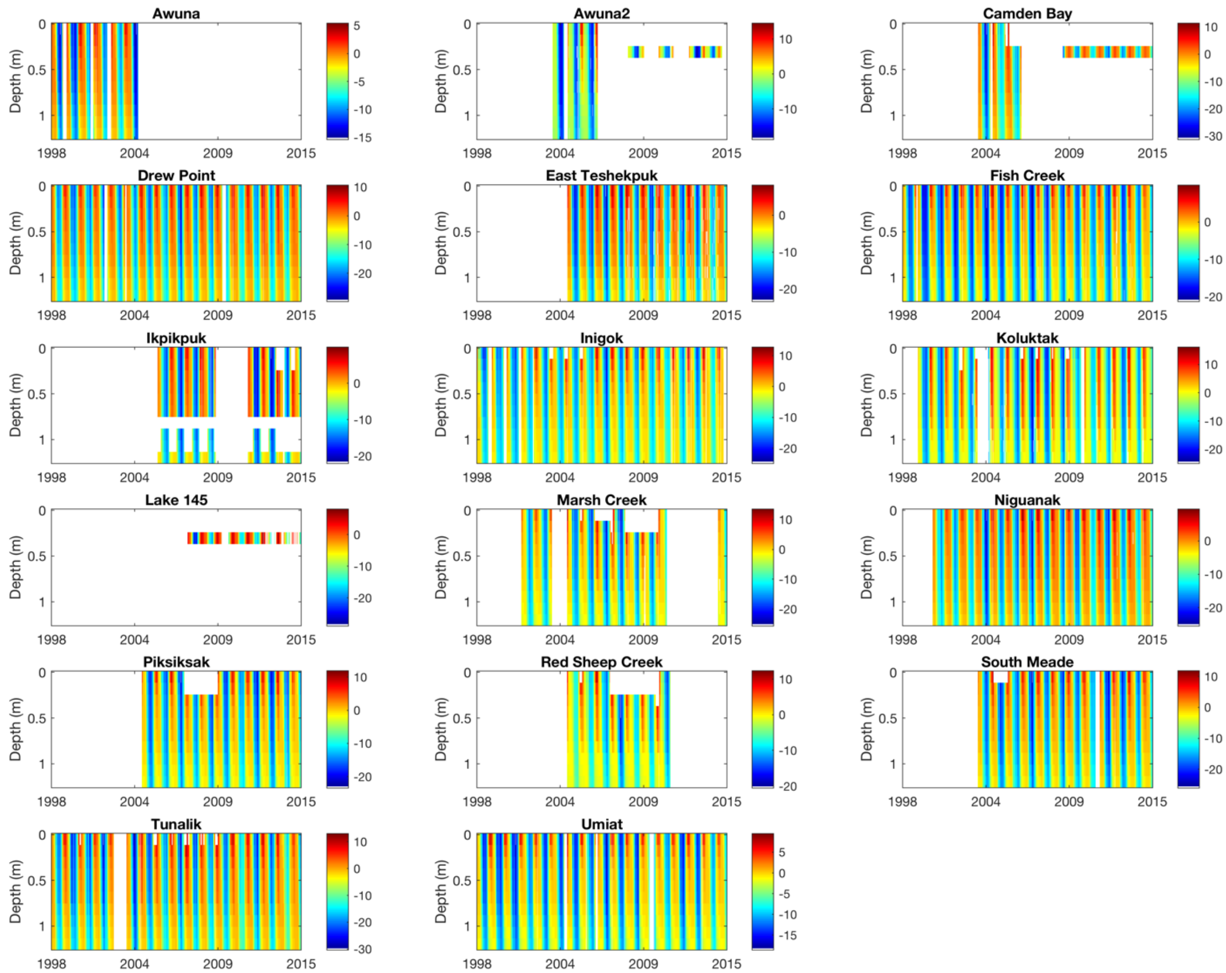

Fig.3 Illustration of period of missing data in soil temperatures (unit: degree C). Colormap shows temperatures.

7. It is peculiar that a linear regression trend analysis was chosen as a core validation technique, given that you acknowledge how there are probably not enough data available to reliably do so. Using trend analysis and other derived variables like frost numbers and effective snow depth to then qualitatively 'eyeball' whether things generally look right are not robust validation techniques. This may uncover glaring issues, but not the subtle non-climatic artifacts and discontinuities that commonly plague climate data.

Response: Thanks. We agree your concerns, data series are short, this is typical for Arctic meteorological data, and is worse for our ground temperature data set. Still, this is invaluable data as compared to no data and the topic of the paper is not to present results on trend analysis, rather to see whether the data is internally consistent. We acknowledge that any subtle issues (e.g., artifacts) in climate data are really difficult to evaluate. However, we choose to do some of the validation based on air temperature, which is one of the more robust, fundamental variables – and then can compare derived data to expert knowledge maps, such as Frost number. We may not able to address more detailed data problems in this study, but in the future, we would like to design some experiments to investigate trend analysis for sparse data series as you proposed.

Specific points:

Can you quantify or estimate the thermal disturbance caused by drilling the holes at the soil temperature measurement sites?

Response: All observational programs have made a point of minimizing disturbance, and first-year data are indistinguishable for later year data, but based on available data, we cannot evaluate quantitatively the thermal disturbance due to the drilling.

p. 1 line 9: why report that the dataset consists of 41,667 monthly values (to make it seem really big)? This is not a useful statistic but instead you should provide a percentage of how complete or how much missing data there are based on the overall date range.

Response: We present the total number of measurements for readers to have a sense of how many data points are available. Since we present data from different instruments (e.g., some stations do not have soil moisture data) and different observational periods for each sensor (e.g., air temperature data was installed often earlier than other variables), we provided a total available data (i.e. 41,667 data points). Based on the reviewers recommendation, we added another statistic in this revision:

> "Our data comprises 41,667 data points in total. There is significant missing data, e.g. some stations do not have soil moisture sensors installed, and there are different observational periods for each sensor (e.g., air temperature data was installed often earlier than other variables, sensor failure). Excluding the missing timeseries when certain instruments not installed, the percentage of complete data is about 77%."

p. 3 line 24-25: is this a personal hunch, or can you provide a reference for this statement?

Response: Removed.

p. 3 lines 28-29: please clarify how or why the thermistors are designed for a low temperature of -30 C, yet they record down to -35 C?

Response: They are two types of sensors, which we described in the previous sentence: "the networks use either a probe with several **thermistors** embedded into a single rod, typically 1.0 to 1.5 m long, or several individual Campbell Scientific **107 thermistors** anchored at specified depths within a single hole."

p. 5 lines 16-20: you explicitly mention the temporal availability of the USGS and GIUAF data, but why not for NPS?

Response: Thank you, we now make a more precise statement. USGS comprises data through July, 2015, whereas the GI-UAF and NPS datasets run up to December 2017.

p. 6 lines 12-13 and 14-15: how did you choose those thresholds (20 days and 90%)? Based on those cited references, or did you perform your own cost-benefit analysis to determine how many missing observations you can allow while still obtaining the most continuous monthly/annual dataset?

Response: These thresholds are selected based on careful evaluation of the different approaches mentioned in the references, and consider the fact that sparse data is typical in the Arctic.

p. 7 lines 5-6: what is this statement based on? Is this a visual assessment or did you perform a statistical analysis? Was this the only site that experienced a station move or other non-climatic change? Is there a list of all station moves, instrument changes, and other events that could affect the data quality?
Response: Yes, it was based on comparison. Because we didn't merge these two sites together. This sentence has been removed in the revised manuscript.

p. 7 line 25: what is effective snow depth and how was it calculated?
Response: This sentence has been revised:

> "We averaged the snow depth measurements over the period from October through March to obtain the effective snow depth (SND eff) (Slater et al., 2017)."

p. 8 lines 21-24: you are reporting frost numbers without even explaining what those 0.5 versus 0.6 values mean. How is frost number used to indicate permafrost occurrence? What does a 0.5-magnitude frost number indicate?
Response: The frost number was described in Eq. 1.

p. 9 lines 5-7: how or why were those stations chosen? Are they the only ones with 10 or more years of data?
Response: Indeed.

p. 10: is Smith Lake the only instance of multiple sites for the same (general) location? How or why was this Smith Lake example chosen for discussion?
Response: There are some other stations, e.g., stations around Franklin. But the stations at Franklin are limited to air temperature observations, and therefore we deemed them less suitable for an example. The selected stations (Boza Creek 1 and 2) are nearby, and thus the air temperatures were collected over similar monitoring periods and displays similar trends (as shown in Fig. 7). Thus, we selected the stations around Smith Lake because there are several different periods, and they illustrate well problems with differences in trends between relocated but nearby stations.

Figure 5: there is enough room on these figures to write out the actual variable names in the title.
Response: Thank you. Fig.5 is revised.

Figure 6: primary y-axis labels overlap (why not write "Air (C)," "Surface (C)," "Ground(C)" and elaborate in the caption that the top three rows show temperature?); I am unsure about the secondary y-axis labels, are they necessary? What does A, B, C, D refer to? What does the asterisk mean? What does the grey shading mean?
Response: Fig.6 is revised. The shadow shows the standard error of the linear regression.

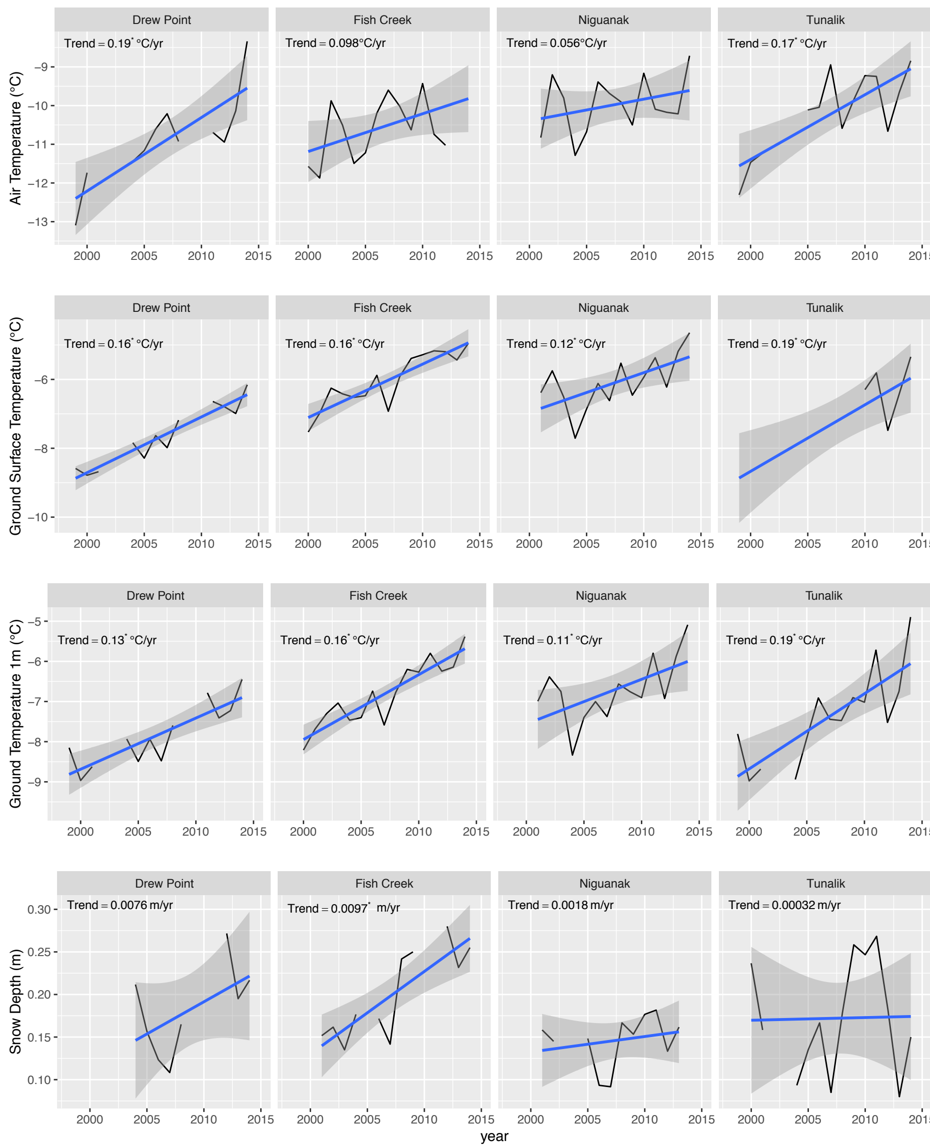

Figure 6. Examples of time-series in mean annual air, ground surface, ground temperature at 1 m below ground surface, and snow depth. Black line shows observations, the blue line is estimated linear trend. The shadow shows standard error of the linear regression estimates, asterisks mark trends with a p value < 0.05.

Figure 7: what is being shown here on these modified box and whisker plots? What does the circle versus the range indicate? Are all the trends plotted for the identical time period (the caption implies not)? If not, they are not comparable and this plot is misleading. Even if a location has more than 5 years of data, you cannot compare a 6-year trend to a 10-year trend if they have different beginning and end points. Are all the trends significant? What does A, B, C in the legend mean?

Response: Considering your comments, we revised this part to a more strictly comparison (P9 L10-14), which we believe addresses any confusion.

> "There are six stations with relatively long records (>=10 years) of air, ground surface, and ground temperature at 0.5 m for the same period. Fig. 7 shows that air temperature,

ground surface, and ground temperature at 1m have generally consistent trends. Furthermore, the trends in ground surface temperature and at 0.5 m depth were generally close."

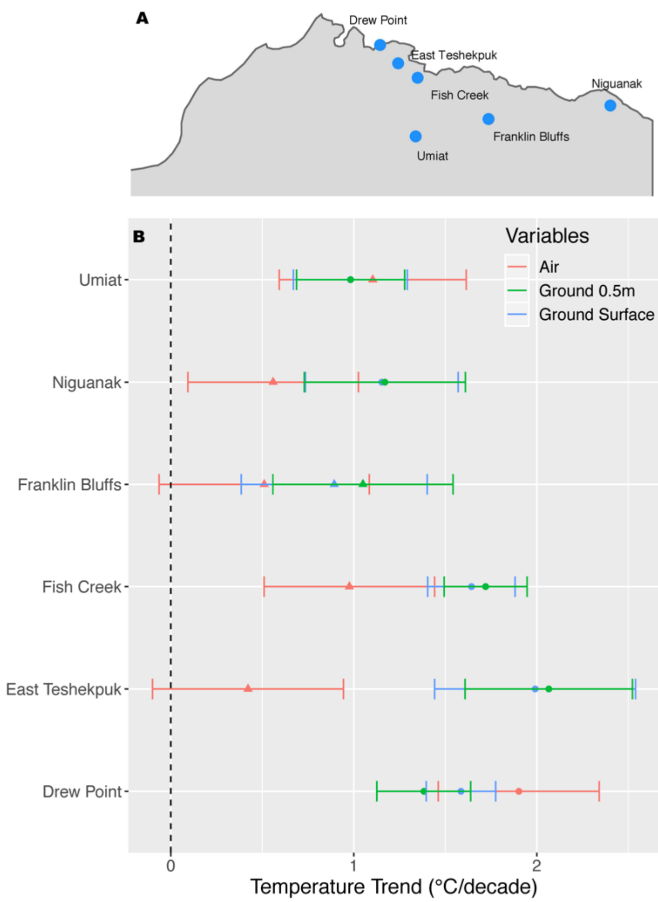

Figure 7. (A) Stations with at least ten years of identical period of air, ground surface and ground temperature at 0.5 m. (B) Trend comparison of air temperature, ground surface temperature, and ground temperature at 0.5 m. All trends were estimated only for those stations with at least ten years of data coverage. Air, ground surface, and ground temperature at 0.5 m were used in the same period. Error bars represent standard errors from the linear regression analysis. Closed cycle indicates trends with p value <=0.05 while triangle indicates trends with p value >0.05.

Figure 8: what is the grey shading?

Response: We revised the caption:

"Figure 8. Comparison between A) trends calculated using measured data at SL1, 2,and 3; B) merged data series and corrected trends at SL3 site. The shadow shows standard error of the linear regression estimates."

Figure 9: needs actual x and y axis titles (instead of acronyms), and units.

Response: Fixed.

The entire manuscript needs to be carefully edited. I am sure one of the 15 authors is a native English speaker who could do this?

Response: We apologize for the issues in manuscript and tried to correct them during the revision.

Reply to Anonymous Referee #3

This analysis provides an assessment of soil temperature, soil moisture, air temperature and snow depth data collected across Alaska, at depths up to 1 meter and over a time span of 1997 to 2016. The manuscript describes the processes used to harmonize the data; the harmonized data are presented in a resulting dataset posted through the Arctic Data Center. The dataset provides a useful contribution to the Arctic community and is especially relevant for model development. The manuscript itself could use more refinement prior to publication.

Response: Thank you for your comments concerning our manuscript. Those comments are valuable and helpful for revising and improving our paper.

Although the discussion appears focused on trends, the manuscript would benefit from some analysis and discussion of interannual variability observed in the data. Is it possible to calculate the start and end of the annual frozen period (where soil temperature is less than 0 degree C) from the compiled dataset? If so, please include this in the results and discussion, in addition to the freezing and thawing index. A brief description of vegetation and soil type should be included for all sites, as well as mean annual thaw depth.

Response: We deem the request of the reviewer for additional data analysis to be in conflict with the journal's policy (unknown to the reviewer, we had submitted an earlier manuscript version and were advised by the editor to keep analysis to a bare minimum and focus on presentation of the data process). However, we did add concise descriptions and meta-data of vegetation and soil information in the revised data set (see Table 2 in the revised manuscript). The onset and end time of the annual frozen period were also derived from our compiled dataset, which were included in the revised dataset. All 'derived' or 'value added' variables; such as freezing onset, freezing index, thawing index, were flagged to make a clear distinction with primary observed data.

Specific comments:
Abstract.
Line 2. Some of these temperatures are at or above 0 degrees C and near-surface soil temperature and soil moisture are also included in the compiled dataset. Perhaps refer to these data as representing active layer and permafrost.

Response: We corrected:

> "A comprehensive near-surface permafrost and active layer dataset is critical to better understand climate impacts and to constrain permafrost thermal conditions and spatial distribution in land system models."

Line 6. Add a comma to 1327 meters for consistency.
Response: Done.

Line 8. I don't think it necessary to have the paragraph mentioning missing data here, in the abstract. This is more a point to be made in the discussion and conclusion section.

Response: Thanks. We removed this here.

Introduction.
Line 26. Remove "of" prior to "allow".
Response: Done.

Line 35. Change "These technical validation would be useful for proving data harmonization and reusing these data" to "These technical validations are useful for data harmonization and future re-analysis of these data"
Response: Done.

Section 2, Page 3.
Line 10. "Hydra" probe instead of "Hydro" probe?
Response: Hydra Probe is a product of Stevens company.

Section 2, Page 5.
Line 11. Use "it is" instead of "it's"
Response: Done.

Section 2.2, Page 5.
Line 14. You define QC here but do not consistently abbreviate quality control past this point. Do so for consistency, or just use "quality-control" instead of QC.
Response: Done.

Line 22. Was bias introduced when applying linear interpolation to ground temperatures? The use of linear interpolation needs to be justified.
Response: Thank you, we addressed this. See extensive response on this same issue to previous reviewer (See P15-18, Comment 5 from Reviewer #2).

Line 32. Add another sentence or two to describe the Frost index in more detail. Provide an example of how the resulting index values might be interpreted.

Response: We added sentences to describe in more detail.

> "A Frost Number index of 0.5 implies equal freezing and thawing index. When the Frost Number index is > 0.5, it indicates the annual period of freezing is dominates thaw, implying climate condition that promote permafrost."

Section 3.1, Page 8.
Line 12. Southeast boreal or southeast mountain tundra?
Response: Corrected. "southeast mountain tundra".

Line 13. In which year was the 1.5 m depth recorded?
Response: Since this site is located close to a glacier, snow depth is generally thick. Our dataset showed that deep snow (> 1.5 m) occurred in 2008, 2009, 2011, 2012, and 2013.

Line 17. I do not think the note in parentheses is necessary.
Response: Fixed.

Section 3.2, Page 10.
Line 1. Change "while" to "that". Add "was" before "mainly".
Response: Done.

Lines 10 & 11. Change Smith Lake to SL2 and SL3.
Response: Done.

Section 3.2, Page 13.
Lines 5-6. These sentences are not necessary.
Response: removed.

Figure 1.
Move the legend for the pan-Arctic permafrost inset to the right-hand side of the inset map. Increase the font size – otherwise some will not be able to read this. Increase the font size for the main legend.
Response: Fixed. New Figure 1 is:

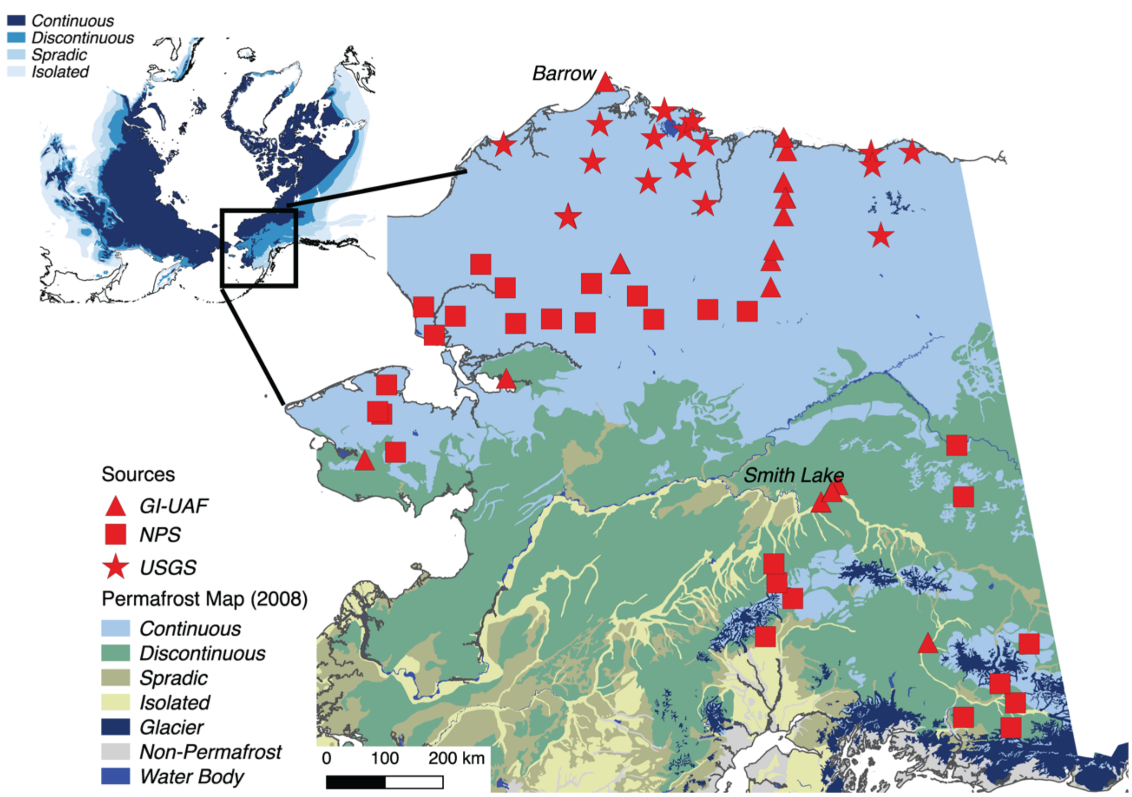

Figure 2. Why show only snow depth? Why not also show soil moisture, air and ground temperature? Indicate spatial locations where the "trend" analysis shows significant change or no significant change (I realize that locations having >= 10 years of data

may be limited, but it is still helpful to see these on a map). Color code by p value?

Response: We added an overview of primary variables, i.e., ground temperatures, including one new figure and paragraph. See response to previous reviewer (Page 13, Comment 3 from Reviewer #2).