Reviewer comments in bold, responses in plain text.

General comments: In this paper the authors develop a relatively straightforward diagnostic metric (SHTM - Snow Heat Transfer Metric) for establishing whether the heat transfer through the soil-snow layer is realistically simulated by a climate model. The diagnostic is based on the amplitude equation for the conductive heat flow which is integrated over the period when air temperature are below freezing to obtain the difference in the seasonal temperature amplitudes at some depth in the soil, and the effective snow depth which describes the insulating effect the snow layer over the accumulation season. The authors use observed air temperature, snow depth and soil temperatures at 20 cm from climate stations in Russia, Canada and the USA to obtain an estimate of the curve relating the effective snow depth to the normalized difference in temperature amplitude (Figure 3). There is considerable scatter around this curve which the authors describe as "noise" or "error". However, I suspect that the results shown in Figure 3 represent a number of different curves that reflect different snow-climate regions (e.g. Sturm et al. 1995) and soil properties (e.g. organic soils).

We agree and we have updated the text to make it clear that by noise / error, we are referring to observational error as well as the range of curves that arise due to different snow / soil regimes. We have rewritten to "The observations show the expected exponential shape and fit the underlying theory (5) well despite significant scatter in the data (Figure 3). The scatter is likely due to several things including measurement error, the range of conditions and snow regimes [Sturm et al. 1995], that occur across the landscape, including the timing of snowfall, the pattern of snow metamorphism, the properties and moisture content of the soil as well as uncertainty about the measurement locations of the observed data."

The authors then compare the ability of 13 CMIP5 climate models to replicate the observed relationship derived from the surface observations using all land grid points north of 55 deg N (Figure 4). The results show major differences between models but one question that crops up at this point is whether the somewhat limited spatial sample of observations (Fig. 2) influences this comparison. Repeating the analysis for grid points nearest the observations would answer this question. The large difference in SHTM between models is worrisome but we don't get any sense from the paper of the climatic consequences of a poor fit to the observed heat transfer relationship and how much of the poor fit is coming from representation of snowpack versus the specification of soil thermal properties.

The curves from the models are robust whether or not we sample at the same locations as the obs or globally, though obviously there is more scatter when sampling fewer points.

It is beyond the scope of this study to assess the climatic consequences of a poor representation of snow heat transfer. The one thing that is clear is that permafrost simulations will suffer if snow heat transfer is not represented accurately. To determine broader climatic consequences would require additional climate model sensitivity runs with a range of snow heat transfer representations / parameters.

Because the observations of soil temperature are not at the surface, it isn't possible to fully distinguish where in the snow/upper soil system a disagreement between models and observations might arise. However, the offset in Anorm at zero effective snow depth is an indication of the impact of soil heat transfer between the soil surface and 20cm depth. In obs, this value ranges from about 0.05 to 0.3. Many models exhibit a lower normalized temperature difference value at zero effective snow depth compared to obs, which means that they transfer heat through the soil too efficiently. The sources of a soil heat transfer bias are myriad and could be due to biases/errors in soil texture including organic matter, soil moisture and soil moisture phase, and soil thickness. Many to most of these models do not represent soil organic matter (which is highly insulative), so this is a potential source of bias in these models. We have added a paragraph to discuss this point.

Presumably one would not use a model with a poor SHTM metric for studies of the soil thermal regime or permafrost, but apart from that I'm not quite sure what the metric tell us. The metric would certainly have value in evaluating the performance of different versions of climate models and land surface schemes. One aspect of the paper that could be expanded on (topic for follow-on paper?) is the spatial variability in SHTM in observations and models. In conclusion this paper is a useful addition to the literature and a testament to Drew's ability to derive practical applications from complex processes.

The metric tells us that models with a poor SHTM metric are not correctly modeling the thermal insulation of snow and it should be used to assess the quality of different models and potentially identify what models are fit for purpose (e.g., as reviewer notes, a model with poor SHTM should likely not be utilized in permafrost studies). See below for more comments on spatial variability of SHTM, but basically we are not convinced that it is appropriate, without a lot more observations, to study the spatial variability of SHTM. Our goal here is to generate a constraint on the representation of snow heat transfer in models that can be applied globally.

Detailed comments:
- Page 1, line 30: Mudryk et al (2016) would be a useful reference to cite in this context as it specifically addresses the uncertainty issue in observational SWE datasets
Mudryk, L.R., C. Derksen, P.J. Kushner and R. Brown, 2015. Characterization of Northern Hemisphere snow water equivalent datasets, 1981–2010. Journal of Climate, 28:8037-8051.

Good suggestion. We have added the reference.

- Page 5: Observed data. The authors have a rather limited sample for characterizing the NH land area average amplitude used in eqn. (7). It would be instructive to provide the readers with some idea of the variability in Fig. 3 for a sample of the major snow climate (e.g. Sturm et al. 1995) and ecoclimatic regions.

It would certainly be good to be able to see how the curve differs for different snow climate regimes, but we are limited by the availability of collocated snow, air, and soil temperature data. As noted in the text, at least some of the significant scatter likely arises from the different snow climates that are sampled which lead to different snowpack densities across ecoclimatic regions. The best we can do is to note that some of the scatter is likely attributable to these factors and to further note that snowpacks with seasonal snowpack dynamics and average densities that lie outside our sampling could generate different curves.

- Figure 4: I suggest you use "scatter" rather than the statistical term "error"

Good point. We have modified to using the term 'scatter' rather than 'error' throughout the paper.

- Figure 5: the derivation of Figure 5 is not provided in the paper and there is no discussion of this Figure. This shows the CMIP5 ensemble close to the observations but this is a potentially misleading message.

We aren't clear what the reviewer thinks is missing. The derivation of the figure is in the figure caption and is pretty straightforward. It's true that the CMIP5 ensemble resembles the observations in terms of this diagnostic, which is only indicating that shallow snow depths are more common in both observations and models than deep ones. This seems uncontroversial. Not sure what additional discussion would be helpful.

- Figure 6: What about spatial variability in SHTM? Is this important? How does this vary between models? To what extent do the different geophysical fields used in models contribute to this variability i.e. how much of a model's behaviour in SHTM is related to representation of

the snowpack versus specification of soil thermal properties?

There may be some spatial variability in the SHTM, but that is not really the point. In Figure 6, we are emphasizing that the metric is relatively insensitive to climate forcing since the values remain constant through time and with climate change. One needs quite a bit of data to create the functional relationship curves so at best one could potentially create a map of very large regions of SHTM scores, but since the underlying data generating the observed curve is quite sparse, it doesn't really make sense to make a map of SHTM.

It is not possible to identify from these standard CMIP5 model runs where the source of discrepancy between model behavior and the obs comes from. That said, one can infer that the offset of approximately 0.05-0.3 in the observed normalized temperature difference in Figure 4 at zero effective snow depth, reflects the impact of the soil. Models that have a low normalized temperature difference value at zero effective snow depth compared to obs likely transfer heat through the soil too efficiently. The sources of a soil heat transfer bias are myriad and could be due to biases/errors in soil texture including organic matter, soil moisture and soil moisture phase, and soil thickness.