Reviewer comments in bold, responses in plain text.

This is an interesting paper with a data-driven and process-focussed approach that typifies the recently deceased first author's work. I expect that it will prove to be useful for model evaluation; indeed, this method has already been suggested for inclusion in the methods for evaluating models in the upcoming Earth System Model – Snow Model Intercomparison Project (http://www.climatecryosphere.org/activities/targeted/esm-snowmip). The observed relations ship between "temperature amplitude difference" and "effective snow depth" shown in Figure 3 has a great deal of scatter. A lot of this scatter will come from genuine physical processes. It would be useful to have some discussion of the influences of soil texture, soil moisture and freezing on the results. Without separating out these influences, it doesn't appear that this method could provide very strong constraints on models, but it is likely to still be useful because current models, as shown in Figure 4, have an even larger range.

We agree that the scatter in the observations can arise from multiple sources and in the revised version of the paper we better acknowledge this at the start of the Results and Discussion section: "The observations show the expected exponential shape and fit the underlying theory (5) well despite significant scatter in the data (Figure 3). The scatter likely arises from several sources including (1) the range of climate conditions and snow regimes [Sturm et al. 1995] that occur across the landscape, including the timing of snowfall and the pattern of snow metamorphism, (2) the properties and moisture content of the soil, and (3) uncertainties in the measurements themselves and the measurement locations of the observed data. It is not possible to distinguish which of these sources of uncertainty dominates the scatter seen in Figure 3."

Additionally, we note that we have already included discussion of the limitations of the SHTM metric and noted that small differences in a score are not necessarily indicative of a significant improvement. And, as the reviewer notes, current generation models show a wide range of performance for this metric. Our perspective is that even a weak constraint is better than no constraint and that if all models could be updated so that they lie somewhere near or within the gray shading in Figures 3/4, that would constitute a big improvement from an Earth System modeling perspective.

It would be interesting to know if the results of this paper can be related to the performances of the same models in simulating permafrost extent, as discussed by Koven et al. (2013). The Hadley Centre model is identified as one in which soil temperatures under snow track air temperatures too closely because of the simplicity of the snow model used. The developers of this model are well aware of this limitation and have implemented a multi-layer snow model to address it; the model is described by Best et al. (2011) and its impacts on permafrost simulations by Chadburn et al. (2015).

We have added the following statement to acknowledge that this issue has been resolved in the Hadley Center model. "Note that the Hadley Centre model developers have addressed this limitation by implementing a multi-layer snow model [Best et al., 2011; Chadburn et al. 2015]."

The definition of effective snow depth in Equation (6) is curious and requires explanation. Why is it chosen so as to give an effective depth that is greater than the average depth for any month for the green line in Figure 1?

We have updated the equation as per the suggestion of reviewer 1 and redrawn the figure.

page 2, line 31 "the period over which the forcing is applied" is ambiguous. Something like "the frequency of the forcing" would be better.

To improve clarity, we modify the sentence to "The value of d is a function of the thermal diffusivity of the medium and the length of time that the forcing is applied."

page 4, line 3 The R parameter is an effective damping depth, not an effective thermal diffusivity.

Corrected.

Best, MJ, and 16 others, 2011. The Joint UK Land Environment Simulator (JULES), model description. Part 1: Energy and water fluxes. Geoscientific Model Development, 4, 677–699.
Chadburn, SE, EJ Burke, RLH Essery, J Boike, M Langer, M Heikenfeld, PM Cox and P Friedlingstein, 2015. Impact of model developments on present and future simulations of permafrost in a global land-surface model. The Cryosphere, 9, 1505–1521.