Reviewer #2:

This paper describes an excessively detailed assessment of how to model a set of temperature measurements done at different depths in an Arctic landscape.

The topic area of this study and model development is important. However, this paper seems to overshoot the goal of providing a straightforward and useable modeling approach for these systems.

We thank the reviewer for their time and effort on this review and specifically agree that this is a timely topic that deserves much attention.

Our primary goal for this manuscript, as stated in the title, is to demonstrate how field observations can be incorporated into the development of process-rich models, both in terms of building confidence in the process representations and in systematically inferring model parameters that are not directly measurable. As noted by Kurylyk and Watanabe (2013) a need for better representations of permafrost environments in a warming climate has motivated the development of fine-scale thermal models. These emerging models are complements to the reduced complexity models used at regional and pan-Arctic scales and consider the wide range of coupled processes that are needed to model the permafrost environment at the level of detailed required for comparing to observations at their native spatial and temporal scales. Our manuscript fills a gap between these new process models that have been evaluated against laboratory data and regional models with coarse spatial resolution, which are poorly constrained by direct observation. By iterative calibration in what is termed the 'ModEx' cycle we are able to evaluate competing representations for processes governing ALT, calibrate the most successful representations, and then incorporate those process representations in the model development. This paper is an important step in the longer-term goal of refining and building confidence in land-surface models of permafrost affected regions and indeed far surpasses the objective, 'to model a set of temperature measurements'. Moreover, the iterative calibration and model refinement process documented in our manuscript has broader applicability to the development of environmental systems models in that a detailed guide for developing process rich models with available field data is presented and is of interest to readers of Geoscientific Model Development.

Given that the reviewer had missed the primary goals of this work, we have revised the manuscript to be more specific in the abstract, introduction and conclusions so that readers can clearly identify the purpose of the paper. First in the abstract we now state the goals of the this work and have added the reworded to now say, "A recently developed surface/subsurface model for permafrost thermal hydrology, the Advanced Terrestrial Simulator (ATS), is used in combination with field measurements to achieve the goals of constructing a process rich model based on plausible parameters and to identify fine scale controls of ALT in ice wedge polygon tundra in Barrow, Alaska." Then in the introduction we now say, "We use repeated calibration of model parameters against site-specific field measurements and iterative model adjustments of the model structure to reduce mismatch between model predictions and measurements in order to attain a viable model of thermal hydrological conditions." And we conclude the introduction section with a summary of our approach and have reworded part of it to say, "In this paper we summarize our ModEx experience involving the detailed use of subsurface *temperature and snow cover field data to develop and test process-rich simulations of ALT dynamics, such that observational data and necessary physical dynamics are incorporated into the model." Finally in the conclusion section we now restate how our ModEx approach achieved this works objectives with, "The particular variant of the ModEx approach combined calibration with iterative refinement of the model structure; parameter feasibility and model-observation mismatch were used as metrics to achieve the objective of model development and identification of viable representations of key thermal hydrological process." We also end the paper with a discussion regarding our approach to merging observations and model development, and how similar approaches may be useful in other applications.*

It is unclear how this kind of model simulations could be used to inform models at a regional scale.

There is significant interest in using fine-scale models to challenge and improve the coarse parameterizations used in regional and global land surface models. In particular, fine-scale models can represent processes and heterogeneities in greater detail and at the native spatial scales of field observations, and can thus bridge spatial scales and generally build confidence in coarse-scale models. This work is not focused on regional scale models, but by combining fine-scale models and observations to identify appropriate representations of key processes and appropriate parameterizations of those processes, the work indirectly informs regional scale models. Based on our results, we did suggest on page 3256 Lines 11-16 in the conclusion section that multiscale models that use overland flow to establish ponded depth in conjunction with subsurface thermal process models are a good approximation for simulating ALT at scale.

Furthermore, our work describes in detail what processes our model found to be important for representing ALT. By employing the iterative ModEx calibration process we found that 1) representing thermal conductivity as dependent on material properties and saturation states are necessary to propagate thermal changes in the subsurface (Section 3.2). 2) The dominant and transient saturation states are also necessary, especially considering how thermal conduction depends on both the phase and saturation state of the subsurface (Sections 3.4 and 4.3) 3) The representation of snow distribution, snow deformation, i.e ageing and depth hoar formation (Section 4.4). These are physical representations that may be important for large scale or multiscale models to consider, and in the case of subsurface thermal conductivity, may require extensive calibration that can be achieved using fine scale models. For greater clarity regarding how our work can be used to inform larger scale models, we also now describe in the introduction a general manor of how fine scale models such as the one presented here can be used to inform larger scale models, "Improved fine-scale simulation capabilities can inform the representation of soil thermal processes in regional to global scale models by identifying appropriate representations of key processes governing ALT, and by providing calibrated model parameterization."

Why spend so much effort on the detailed parameterisation of thermal properties if lateral heat flow might be important, which is then not included.

We acknowledge that lateral heat and water flow might have an influence on the system (Page 3251 Line 25) and future 3-D modeling and fieldwork is necessary to quantify what the influence of lateral heat flow might be (Page 3256 Lines 17-26).

However, 1D calibration and parameterization is beneficial in that the computational time to simulate a 1D problem allows for the many simulations necessary to sufficiently explore parameters space in order to identify what thermal properties are necessary to simulate ALT. Our work also shows that without a representation of lateral heat flow, we are able to match subsurface temperatures consistently for rims and centers, and with the exception of early snow melt and fall freeze, the simulated trough temperatures match the observed temperature with plausible subsurface properties.

It seems awkward to fix the lower temperature boundary, it is unclear what this is based upon.

The lower boundary temperature is a far field boundary condition that is within the average permafrost temperature for the North Slope Alaska (Romanovsky, et al., 2010). A fixed boundary condition is a reasonable approximation as seasonal temperature variations generally to not penetrate deeper the 10 to 16 meters and deep permafrost will see only negligible temperature increases over the course of a calibration that spans only a few years.

The paper is written very densely, but still does not contain enough information to fully appreciate what it is that has been carried out, and how. On the other hand this contains to much information without detailed description resulting in a difficult to read.

We believe that the additions to the manuscript, which now clearly state the objective of the work provide a bases for the level of detail in the manuscript. (See response to the reviewer number 2's first comment).

One is left with a feeling that the authors invested a lot of effort to develop a unusually detailed model but then fail to carry out a sensitivity or uncertainty study to evaluate the need for the complex model construction presented here. Could the same fit be obtained with a much simpler model too? In other words what is the sensitivity of the model fit to model complexity?

We acknowledge that a lot of effort was indeed invested to develop this model as is necessary for such complex and dynamic process representation. We further appreciate the question of needed complexity as distinguishing the governing processes from those that can be neglected is a central component to scaling the representation of thermal hydrology up to a larger scale model. Our manuscript documents how added process representation, i.e transient saturation, phase change, and the tightly coupled surface thermal conduction is needed to capture the subsurface thermal regime. In other words, our work started with a simple model, and as a result of incorporating observational data into the iterative calibration procedure, we were able improve model performance and identified the level of processes representation needed. For example, the addition of transient saturation in section 3.4 and 4.3, and refinement of the snow representation is section 4.4, demonstrate that without the additional process representation model fit and the plausibility the parameters used in the model would not be possible. Nevertheless, appropriate model complexity needs to be addressed, especially when attempting to find the appropriate level of process representation in larger scale models. We therefore added a comment about complexity in the conclusion section, "Further evaluations of the 1-D representations against 3-D model representations are needed to identify addition process representation and the appropriate level of model complexity to capture scale dependencies of thermal dynamics."

Other comments are provided in the attached annotated PDF.

Location in Original Text Page 3237, Line 16-17: "These local-to-intermediate scale processes are under-resolved or completely missing in ESMs. Therefore, improved fine-scale simulation capabilities can inform the representation of soil thermal processes in regional to global scale models."
Reviewer #2 Comment: This makes it sound like the upscaling of what you find at the small scale to the ESM scale is obvious. I am sure it isn't. The term 'inform' is somewhat ambiguous. How are the ESM's informed exactly?

> *We agree that this sentence is somewhat ambiguous and have therefore re-worded that sentence to be more specific and describe a general way in which fine-scale modeling efforts can inform larger-scale simulations. The sentence now states:* "Improved fine-scale simulation capabilities can inform the representation of soil thermal processes in regional to global scale models by identifying the appropriate representations of key processes governing ALT, and by providing calibrated model parameterization."

Location in Original Text Page 3239, Line 1-2: "Additionally, correct model structure representation is typically not known a priori."
Reviewer #2 Comment: how is this defined? what is correct?

> *We now define what is meant by correct model structural representation, and have changed the sentence to,* "Additionally, correct model structure representation, capable of representing the system based on known physical relationships while using plausible model parameters, is typically not known *a priori*."

Location in Original Text Page 3239, Line 5-7: Therefore, when dealing with a coupled system of complex processes, it is imperative that the conceptual model is refined during the calibration process to increase model structure adequacy (Gupta et al., 2012)."
Reviewer #2 Comment: Does this not all depend on the objective of the modeling study? Is the objective here to correctly estimate, forecast ALT?

> *The objective of this study is two fold, first to incorporate observational data into a process rich model representation of ALT dynamics, second is to identify the appropriate representations of governing processes that control the thermal hydrological dynamics that form ALT. Meeting these goals then creates a model that is capable of estimating and forecasting ALT.*
> *Many models may be capable of simulating ALT, however, without rigorous testing and comparison to observed variables (in this case, the plausibility of calibrated parameters) models may simulate the correct ALT for the wrong reasons. A model that is calibrated using the wrong structure, i.e., conceptual model, can result in erroneous forecasts, especially if conditions change such as the case for climate change. The problem of over-fitting and relying on models that are 'calibrated', but do not use plausible parameters is discussed in the previous sentence. If however, the model is tested and refined to produce both an accurate ALT and plausible calibrated*

parameters, some of the structural error can be reduced and confidence in ALT projections is increased.

Location in Original Text Page 3240, Line 14–16: "Here the ModEx procedure moves beyond the standard calibration by assuming the model itself is unknown, but can be refined through successive comparison to observation (outer loop in Fig. 2)."
Reviewer #2 Comment: how can something unknown be refined?

We now clarify the intent of this sentence by, "Here the ModEx procedure moves beyond the standard calibration by assuming the model itself is uncertain, but can be further constrained through successive comparison to observation (outer loop in Figure 2)."

Location in Original Text Page 3240, Line 18: "These improved model runs then inform the observation process by specifying the data needs, either through informal numerical experimentation or through more formal data worth exercises."
Reviewer #2 Comment: What are these?

Changed sentence to be more specific, "These improved model runs then inform the observation process by specifying the data needs, either through further calibration or through informal numerical experimentation."

Location in Original Text Page 3240, Line 20-22: "We implement ModEx model refinement by focusing on the plausibility of calibrated parameters in addition to the mismatch between field measurements and simulated responses."
Reviewer #2 Comment: Is this a manual process?

Calibration is automatic as described in the manuscript. The model refinement was done manually. We evaluate the set of calibrated parameters against the range of appropriate parameter values compiled from literature (See Appendix C) or field measurements. In doing so, insight about model behavior is gained which can them be used to improve the model and reshape the calibration response surface. For better clarity we have replaced the work 'focusing' with 'evaluating'.

Location in Original Text Page 3241, Line 12: "However, in the case of a complex model with high dimensionality, multiple local minima may exist, which results in gradient-based calibrations finding many solutions to the problem (Beven, 2006)."
Reviewer #2 Comment: non-uniqueness this is often called.

Yes. For clarification we now say, "However, in the case of a complex model with high dimensionality, multiple local minima may exist, which causes gradient-based calibrations to find non-unique solutions."

Page 3241, Line 16-17: "It is important to extend calibration boundaries beyond the acceptable parameter range in order to both diagnose model inadequacy and avoid boundary effects caused by automated calibration algorithms."
Reviewer #2 Comment: It is unclear what this exactly means, at least to me.

We now re-worded the sentence for clarity, to say, "Model structure error can also cause the response surface to slope to a parameter boundary indicating that over-

fitting is necessary to calibrate to observed data (Beven, 2005). Therefore, it is important to extend calibration boundaries beyond the acceptable parameter range to allow the optimization algorithm to travel into the infeasible range when the response surface dictates an implausible combination of parameter values, indicating an inadequate model."

Location in Original Text Page 3242, Line 23: "The focus of the model development chronicled here is NGEE-Arctic site "Area C" (Fig. 1), which is characterized by ~ 50 cm deep troughs, rims and shallow low centers."
Reviewer #2 Comment: I have a sense that Fig 1 is first mentioned after Fig 2, but I did not check in depth.

Figure 1 is introduced first on page 3239 line 15, while Figure 2 is introduced on line 17, after figure 1.

Location in Original Text Page 3243, Line 9-10: "The bottom boundary condition was held constant at a temper1ature of -6 ℃."
Reviewer #2 Comment: Why was a constant temperature boundary chosen? Surely the temperature at this depth cannot be considered constant over a 20 year period during which long term GST changes occur.

The bottom boundary condition is set as a far field boundary condition at depth of 50 meters. The boundary condition temperature is within the range of permafrost temperatures of the North Slope, which has seen between 0 to 2 degrees warming between the years of 1975 and 2010 (Romanovsky et al., 2010) for the entire permafrost zone. Furthermore, seasonal temperature changes do not penetrate to deep permafrost (See figure 4 in Romanovsky et al., 2010). Therefore, a calibration exercise over the course of a few years, such as this one will see negligible deep (greater than 15 meters) permafrost warming.

We now clarify that this is based on an average far field boundary condition and have changed the manuscript to now say, "The underlying mineral soil was a silty loam to a total depth of 50 m. A far field bottom boundary condition was held constant at -6℃ to represent the average deep permafrost temperature in the North Slope of Alaska (Romanovsky, et al., 2010)."

Location in Original Text Page 3243, Line 23: "In this model liquid water can coexist with ice below 0 C, as observed (Watanabe and Wake, 2009), which is an important process to represent in soils with rapid freeze thaw cycles in order to prevent unrealistic rapid cooling of the subsurface (Romanovsky and Osterkamp, 2000; Nicolsky et al., 2007)."
Reviewer #2 Comment: what process? the coexisting of liquid water and ice below 0 C is not a process, that is a phenomenon as a result of a process. But the process remains unnamed here. Liquid and ice partitioning is the process? I am not sure that is a process either. What is causing this?

We agree that 'process' is a poor word choice here. None-the-less, some water in pore space remains as liquid below 0℃ due to surface forces and pore geometry (Dash et al., 1995). In a thermal model it is also important to accurately represent the phases *of water, which have different thermal conductivities. We therefore have re-worded the sentence to be more specific and now state, "In this model liquid water can coexist with ice below 0°C, as is well known (e.g. MILLER, 1980; Williams and Smith 1991), which occurs due to soil surface forces and pore geometry."*

Location in Original Text Page 3251, Line 25-28: "A possible reason for the underestimated soil moisture is that the 1-D surrogate model neglected lateral surface- and subsurface flow that could be flowing on to the column, especially for troughs that are connected to an extensive trough-network."
Reviewer #2 Comment: Indeed, so what is the point of all this detailed calibration for thermal properties?

The point of the detailed calibration of thermal properties is to identify dominant controls of ALT and to best represent those processes in models. If adjustment to the conceptual model is warranted in order to attain both good fit to calibration targets and plausible parameters we changed the model accordingly to build a process rich model and noted why and how we think the model improvement is necessary. If we thought that additional process consideration may only slightly improve model performance, we noted what that process might be and how it could improve the model. Documenting the model development process in this way is important to 1) demonstrate the thought necessary for process rich model development and 2) add to literature the reasons why some processes are included and others are not necessary.

Here our model representation of the system was found to be good with plausible parameters for most times and depths with the exception of spring and fall periods in the trough. Lateral flow could contribute to the mismatch between observations and simulation in the troughs. However, if representing lateral flow were to improve the simulation, it would only improve the trough representation and for a small percentages of time (Figure 9).