Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2022-144-RC2, 2022

Comment on essd-2022-144

Anonymous Referee #2

Referee comment on "High-resolution predictions of ground ice content for the Northern Hemisphere permafrost region" by Olli Karjalainen et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-144-RC2, 2022

The data set provides very important information about near-surface (upper 5m) ground ice content for the Northern hemisphere. Such data are today available only very course but would be urgently needed for climate change projections and climate change impact studies related to permafrost thaw. The paper is well written and presented.

My main critics is that the model developed to estimate ground ice content is purely statistical based on scattered field measurements, and does little involve the physical processes behind the ground ice content and its evolution over a range in time scales. Statistical models are weak (or invalid) in representing conditions that are not or little represented in the training data. It is unclear to me to what extent the ice content data are biased or clustered in various aspects (topography, climate, groundtype, ice formation history, etc.). The authors mention actually a bias to ice-rich conditions. Such biases are not included in the uncertainty quantification. Similarly, it is unclear to what extent and why the input variables used to drive the statistical models represent the processes associated with ground ice formation and content. Given these uncertainties, the formal uncertainties given will not reflect the full reliability of the new data set, and it becomes thus not clear for which applications it could be used, and for which rather not. I can imagine users might apply your data in a way not justified by their validity and accuracy.

I understand my comments are likely not easy to include in the study. A more careful description of the non-formalized sources of uncertainties would be needed and an attempt to quantify these. Should areas be excluded that extrapolate outside of the conditions covered by the training data the parameter space?

Minor comments:

Though well-expected for expert readers, the title (or at least the abstract) should clearly state the the paper talks about the upper 5 m.

Lines 19 and 336: I wouldn't say the data "show" that... It is a statistical model. The produced data contain ... or similar

Line 72: You refer to SROCC, right? This is "an" IPCC special report, there are several other ones.

Fig 3 a: black dots on dark pink ground difficult to recognize.

Fig C4: Possible to have the input data in the background for each panel? For instance as colourcoded historgramme (as in Fig 5, perhaps in greyscale)?