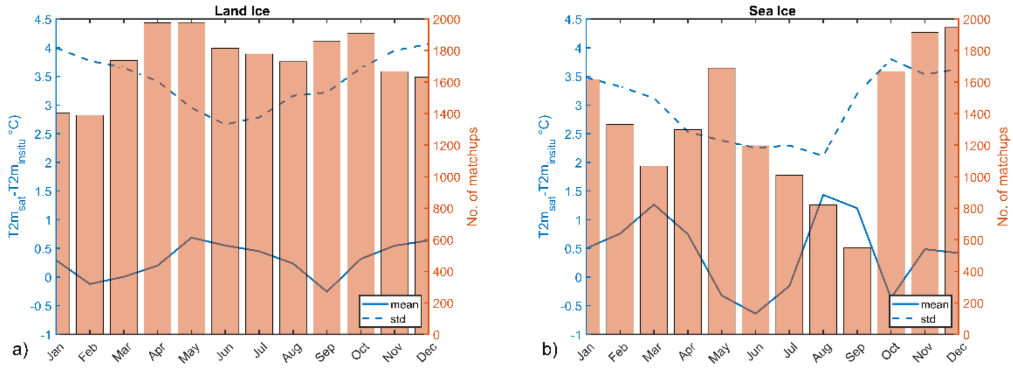
Referee #3:

The manuscript addresses estimation of 2-m air temperature (T2m) over the Greenland Ice sheet and the Arctic Ocean sea ice. The work is based on satellite remote sensing data on clear-sky ice/snow surface temperatures and their comparison with in-situ observations on T2m. I am impressed by the amount of data analysed, originating from different sources, and by the carefulness and detail of the analyses made. Among others, a lot of attention is paid on uncertainty analyses. The manuscript is well organized, the text is well written, and the illustrations are of high quality. The results are important, convincingly demonstrating that T2m estimates based on satellite remote sensing data are useful and reasonably accurate. Hence, they can supplement reanalysis products and rare in-situ observations on T2m. I suggest accepting the manuscript subject to minor revisions.

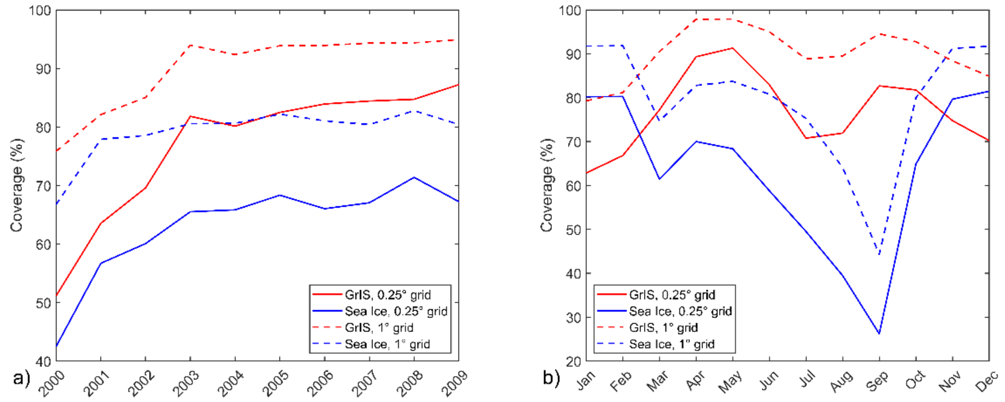
Specific comments

I wish to see more attention to the seasonal distribution and accuracy of the satellite-based T2m estimates. The quantitative results are given for the entire data set (e.g. page 22, lines 26-28), but it is well known that over the Arctic sea ice clouds are more common in summer than other seasons. Figure 8 demonstrates that there are less data from summers, but Figure 7 suggests that the estimates are better in summer than in winter, which must be related to smaller thermal differences between the ice and air in summer. Detailed comparison of skill scores (bias and rmse) between different seasons is, however, missing, and it remains unclear how useful the satellite-based T2m product is over sea ice in summer. This could be easily improved by presenting quantitative results for different seasons.

Thanks for pointing this out. We have added a figure (Fig 7; shown below) to the manuscript, which shows the seasonal variation in bias, standard deviation and the number of matchups. We agree that the thermal difference is smaller during summer, but this to first order, should be captured by the seasonal component of the regression model. The smaller stddev during summer is likely a combination of better performance due to improved cloud screening during summer and the smaller natural variability that is observed in summer conditions.



We have also added a figure (Fig. 8) to illustrate the coverage of the final $T2m_{sat}$ product in the end of the discussion section. As expected we see a minimum in coverage over sea ice in summer due to an increased cloud cover.



It would be good to make it clear already in the Abstract that the study addresses the Arctic Ocean and Greenland Ice Sheet, but not other land ice in the Arctic. Related to this, in Table 1 land snow must refer to seasonal snow. However, in the Discussion, Conclusions, and Abstract, there is nothing mentioned about results for seasonal snow. Were the results similar to those for land ice?

Yes, we agree on this. It has now been stated already in the abstract as well as in the introduction and conclusion. The study also includes the two ARM stations (BAR + ATQ) with seasonal snow cover. These are included as land ice stations in this study and only data from snow covered periods have been used. This is not very clear from the original manuscript, which has now been updated (see Sect 2.1). ATQ was used for training while BAR was used for validation. The RMS error at Barrows is shown in figure 12 (very first bar). From this figure, the results look similar to those for land ice. The $T2m_{sat}$ validation against Barrows gave a mean correlation of 95.8%, a bias of 2.49°C and std of 3.42°C. The correlation and std are similar to the average values over land ice while the bias is considerably larger. This is likely explained by physical differences between seasonal snow covered sites and GrIS sites, which are not fully captured by the regression model. This has now been stated in manuscript Sect. 4.2. In the discussion section, we have added a sentence on the possible extensions of this product to seasonal snow, which will require a dynamical surface mask and repeating the derivation of a regression model.

Page 1, last sentence: Could the information be updated (2010s vs. 2000s) and extended to also address the strong winter warming over the Arctic sea ice? The latter is evident on the basis of reanalyses.

Yes, the information has been updated according to the new historical climate data collection from DMI (Cappelen ed., 2021) and the winter warming over the Arctic has been addressed in the very first sentence.

Page 2, lines 7-8: "In particular" is questionable. It is even more difficult to achieve climate-quality precipitation and air humidity records from the region.

We agree on this and have replaced "In particular" with "Therefore".

Page 3, lines 4-5: Make it clear that the zones refer to the Greenland ice sheet.

Accepted

Table 1: Use of Polarstern data as a proxy for 2-m air temperature sounds surprising. According to my knowledge, in the Polarstern weather station the air temperature sensors are located 29 m

above the sea surface. In Figure 11, however, the rms-errors are small for Polarstern. Could it be because the data are mostly from summer, when the vertical temperature gradients in the boundary layer are mostly small?

You are right. The Polarstern air temperatures are measured 29 m above the sea surface. These measurements are not used in the derivation of the regression model, but only for validation. As noted by the reviewer the validation against Polarstern is relatively good. We agree that this is most likely explained by the fact that data are mainly from summer, when the vertical temperature gradients are mostly small, and the performance of the cloud screening algorithm reaches its maximum. This has been stated in manuscript.

Page 20, line 21: The temporal variability of air temperature is largest in winter because the meridional temperature gradient is largest, the atmospheric circulation is vigorous, and the cloud radiative effect is large (compared to near-zero in summer).

The explanation has been extended to include all of above.

Page 22, lines 1-10: As new results are presented here, consider moving this paragraph to the Results section.

The section has been moved as suggested. Also, the result section has now been divided into subsections to improve readability.

End of page 22 – beginning of page 23: I am surprised that the use of wind speed did not improve the results over sea ice (over Greenland Ice Sheet, the reanalysis errors for near-surface winds may be so large that it is not so surprising there). Also, it looks strange that the correlation coefficients and RMS-errors are exactly the same for the four regressions presented for sea ice in Table 3. I suggest double-checking the calculations.

The calculated values have been double-checked and the RMS-errors are exactly the same when using 2-digits, but this is not the case when using 3 digits. We found it also surprising that the inclusion of wind speed did not improve the results over sea ice. Possible explanations are provided in the discussion section and include limitations in the representation of surface roughness and the processes of snow-surface coupling, radiation and turbulent mixing partly due to limited resolution and varying relative importance of the processes with wind speed (Sterk et al., 2013). Better quality of wind speed (e.g. from regional high resolution reanalyses such as CARRA) will likely improve the performance when included in the regression models.

Page 23, line 30: In summer, the air temperature over sea ice in not so good indicator for Arctic climate change. Ice and snow are rapidly declining, but T2m remains the same, close to 0 deg C, as long as there is sea ice left. See e.g. Figure 2g in Vihma et al. (2008) (just for information, no reason to cite).

We agree on this and the sentence has been reformulated.

Page 24, last paragraph: The independence of the satellite-based T2m product and reanalyses is stressed here but, as mentioned earlier, reanalyses actively assimilate available in-situ observations. In particular, the ECMWF has had a pioneering role in assimilation of T2m data, which was done also in the Arctic already in the old ERA-40 reanalysis. However, I fully agree that the satellite-based T2m product is very valuable, and may indeed yield even better results than assimilation of (rare) observations into a reanalysis system. T2m is a diagnostic product in reanalyses, calculated on the basis of surface temperature and air temperature at the lowest model level, and liable to uncertainties in the parameterization of turbulence in the stably stratified atmospheric boundary layer. Note, however, that the NCEP-CFSR reanalysis, which is based on a coupled atmosphere – sea ice – ocean model, has performed better for near-surface atmospheric variables over sea ice.

We are not entirely sure what the reviewer is after in this comment. We agree with what the reviewer states about the assimilation of T2m into the models and it has now been stressed here as well for clarification. We see the derived product as a supplemental data set to the in situ observations in the data assimilation and this has now been clarified in the very last sentence. Moreover, we have added a comment on the better performance of NCEP-CFSR in Sect. 4.2.

Figure 6: In the legend, replace "air surface temperature" by "near-surface air temperature" or "2-m air temperature".

Accepted