Referee #1:

General Comments

Overall, this interesting and valuable paper is generally well written. There is some additional discussion needed in places and some tightening of prose, particularly in the second half of the introduction and in the discussion/conclusion. Yet the research itself is comprehensive and the science is good.

Specific Comments

- After Page 2 Line 30 the introduction needs a bit of attention to retain the good quality of previous paragraphs. It felt like the flow of the narrative was lost around here and there is a bit of repetition.
  The second part of the introduction has been rewritten to retain the quality of the first part.
- For the satellite data please give an explanation for the choice of a 0.25 by 0.25 degrees regular geographical grid. This seems quite a coarse resolution given the input data, is this to correspond to T2m datasets from other sources or model grids?
  The 0.25 degree grid was chosen within the EUSTACE project to ensure a common grid that could be used globally and that covers the surface skin temperature for all surfaces: ocean, land, lake and ice. The final products within EUSTACE was on this 0.25 deg grid. A sentence on this has been added to the manuscript.
- Page 7 Lines 5-7:
  - It took a couple of reads to figure out that the ISTskin_L3 is the daily version of the L3 not the 3 hourly (assuming I understand correctly). I think changing the text to something like "Here, we have aggregated the AASTI ISTskin_L2 observations into 3 hourly and daily, gridded 5 Level 3 (L3) averages of ISTskin_L2 on a fixed 0.25 by 0.25 degrees regular geographical grid. The daily gridded averages (ISTskin_L3) are calculated by averaging" might make this clearer. The text has been changed as suggested.
- Page 8 Line 26:
  - You mention the distances but not temporal matching. I guess from the data this is a daily average comparison, but this might be worth mentioning explicitly here.
    It has now been stated that is a daily average comparison.
- Page 14 28-29:
  - This sentence reads as though only the random uncertainties are provided for each pixel and the others are provided as one value for the whole dataset. Is this correct?
    The sentence has been reformulated. All uncertainty components are provided for each pixel as explained in Section 3.2. Thanks for pointing this out.
  - Surely given the inclusion of a sampling uncertainty in the synoptic component and separate equations for land ice and sea ice there should be a value for each pixel? This is the case for synoptic uncertainties for similar satellite products such as from e.g. Ghent et al (2017).
    Agree. See above.

- Page 15 Line 1-4:
    - This sentence makes it sound like only ice over sea is included; ice shelf (land ice) and sea ice. But Figure 6 and previous figures show data over the Greenland ice sheet. Is the Greenland ice sheet included in the dataset?
    Yes, the Greenland Ice Sheet is included. The sentence has been reformulated.
- Results:
    - I would like to see a few more citations and additional discussion of the results here as I was often left wondering how the results compared with previous research or observations that might back up your results. For example:
        - "The monthly mean air temperature typically reaches 20 a maximum of -4°C during July and a minimum of about -28°C during winter." Does this correspond to previous research and/or observations? Perhaps the GrIS in situ observations could be added to Figure 12 to show how close the T2m_sat is?
        The range in monthly mean air temperature is in agreement with those reported by van As et al. (2011) at a number of PROMICE AWSs. This has now been stated in the manuscript. In relation to Fig 12, we don't think it will be a good idea to include in situ observations here. Fig. 12 shows the average temperature of the entire GrIS, and we don't think the average of the sparse in situ observations can provide a spatially representative measure of the mean T2m of Greenland for the period.

        - It seems that the results from comparisons of ERA-Interim (and ERA-5) suggest that the long standing warm bias in these reanalyses for the Arctic still exists, which deserves some comment here and perhaps citation of previous studies on this. I also don't know if ERA-5 has been compared to in situ T2m over the Arctic in other research yet so this could be an interesting result given the warm bias may still be a feature of this dataset series. Again, if so this is worth noting.
        Thanks for pointing this out. Previous studies have been cited and two recent studies (Wang et al. 2019 and Graham et al. 2019) have evaluated both ERA-I and ERA5 for Arctic sea ice against buoy observations and N-ICE2015, respectively, and found that ERA5 also suffers from a warm bias. Similarly, recent studies found no significant improvement in performance over the GrIS for ERA5 compared to ERA-I (Delhasse et al., 2020; Zhang et al., 2021). This has now been stated in manuscript.

        - Page 17 Lines 5-9, I believe your statements about Arctic cloud cover are correct, but I think a citation would be useful to back this statement up. Citations have been added.
    - Discussion:

- This section was more like a list of additional things to note that did not fit in the rest of the paper rather than a coherent discussion. I felt it didn't really tell the story of the research in the way it deserves. The section has been restructured and rewritten to make it coherent and to include more references to the result section and previous research.
- I think this needs a bit of work to restructure and possibly a short summary of results relevant to each point made. For example:
    - Page 22 Lines 11-18 seem to refer to the fact that "The satellite 10 derived air temperatures are about 0.3°C warmer than measured in situ air temperature for both land ice and sea ice." which is probably due to the influence of the linear regression? It would be nice to include the context for these sorts of statements in the discussion.
      These lines actually refer to the cold sky bias observed when satellite L3 IST is compared with in situ observations (~-2°C as shown in table 2). The section has been rewritten and hopefully it is more clear now what we have done to remove this clear-sky bias in the final product (for which the bias is 0.3°C for both surfaces). As shown in Sect. 4.3, part of the remaining bias for the GrIS is likely explained by topographic effects.
- Will the surface temperature dataset be extended to seasonal snow and ice? It is possible to extend this work to seasonal snow. It requires a dynamical surface mask and repeating the derivation of a regression model. However, within EUSTACE similar efforts have already been made to cover seasonal snow (Morice et al., 2019; Rayner et al., 2020; Good 2015). This point has been added to the discussion section.

Technical Corrections

- Page 2 Line 19: Satellite not Satellites. Accepted
- Page 2 Line 23: Either un-capitalise The or remove. Removed
- Figure 3, 4: might be a big ask depending on the software used, but is there any chance of removing the sliver of white from the plots? The figures have been updated and the silver of white has been removed.
- Page 8 Line 22: I think a subheading for this validation section would be useful to the reader. Accepted
- Page 16 Line 8: there is not there are. Accepted
- Page 23 Line 17: MODIS not Modis (acronym of sensor). Accepted
- Page 15 Line 2: could you expand the acronym ETOPO1 or provide a very brief indication in text of what this is? Perhaps "ETOPO1 global relief model" or similar? Accepted
- Conclusion section: a bit long and should be distilled down to the major points. The section has been shortened a bit, and now excludes the listing of tested models that where not used.