Reply to Anonymous Referee #2

Authors demonstrate no competence for the topic of Antarctica. Product has no value.

We are well aware of the fact that this response will not change the reviewers opinion on our work. However, we want to act professionally, and will endevour to reply to comments provided that these were articulated appropriately.
We strongly refute the claim that we "demonstrate no competence for the topic of Antarctica": the team builds on a wealth of experience on Antarctic environments for the last 15 years, and has a combined field experience of approx. 15 seasons.

The opening statement - lines 13 and 14 of page 1 - is thoroughly, demonstrably and emphatically false. Antarctica, essentially 14 x 106 km2 of snow and ice (not counting winter sea ice), exists, for six months or more of each annual cycle, in a completely frozen state. Frozen = zero terrestrial ecology/biology/biodiversity. During the 'warm' season (SH summer), the minuscule areas (< 5 x 103 km2, 0.03% or less of the total area) of Antarctica not snow covered (for hydrological rather than temperature reasons), e.g. McMurdo dry valleys, support a tiny, desperate exotic (and fascinating) mini-ecosystem which has virtually no impact on hydrology or glaciology beyond its restricted boundaries. The biodiversity of Convey, which the authors like to cite, refers to sub-Antarctic islands wherever BAS operates long-term research bases; note that Pete's very good work rarely if ever refers to British Halley Station on the continent (adjoined ice shelf) itself. Biodiversity issues for Antarctica, including invasive species, habitat (sea ice) modification or reduction, competition with human predators (e.g. for krill) occur almost exclusively in the marine realm. Likewise for proposed protection areas. No liquid-water hydrology exists at the surface of Antarctica. Extensive glacier mass balances and motions have little to no dependence on surface air temperature. Snow surface halogen chemistry, particularly within regions exposed to wind-blown sea salt aerosols, does show temperature dependence, on reaction rates and - to less extent - on products, but the authors seem blind to that entire field.

The following paper demonstrates clearly that biodiversity is, by far, not limited to the McMurdo Dry Valleys.

- Wauchope, H., Shaw, J.D. & Terauds, A. A snapshot of biodiversity protection in Antarctica. *Nat Commun* **10,** 946 (2019). https://doi.org/10.1038/s41467-019-08915-6

There is also a wealth of literature (as cited below) supporting the importance of near-surface air temperature as a driver of *terrestrial biodiversity, and hydrological and glaciological processes.* For a revision, we will support our statement by further literature and we will slightly rephrase it to make clear that we're mainly talking of ice-free areas when it comes to terrestrial biodiversity:

"Near-surface air temperature in Antarctica is an important driver of terrestrial biodiversity in the ice-free areas(Convey&Smith, 2006; Convey, 2010; Hogg et al., 2011) and is decisive for hydrological (Herbei et al., 2016) and glaciological processes (Cook et al., 2005). "

- Hogg, I.D., Wall, D.H. Global change and Antarctic terrestrial biodiversity. *Polar Biol* **34,** 1625 (2011). https://doi.org/10.1007/s00300-011-1108-9
- Convey, P., Smith, R.I.L. Responses of Terrestrial Antarctic Ecosystems to Climate Change. *Plant Ecol* **182,** 1–10 (2006). https://doi.org/10.1007/s11258-005-9022-2

Do the authors not understand 24-hour polar night alternating with polar constant daylight? They provide a daytime nighttime data extraction routine (page 3 line 27) which, one can scarcely believe, apparently ignores the entire issue of seasonal light levels (complete light, complete dark).

We are fully aware of 24-hour polar night alternating with polar constant daylight. MODIS is a global dataset that provides 2 acquisitions a day, one called “day-time” the other being named “night-time”. Obviously in Antarctica the name of those products can be a little misleading.

Later (on page 5, paragraph starting at line 5) they describe use of solar angles to calculate hillshading as one of their predictor variables but they give no indication that they understand Antarctica; the description sounds more relevant to mid-latitude Germany.

Certainly there are hillshading effects in Antacrtica as well which affect the temperature. This is certainly nothing that applies to mid-latitude Germany only. Complete darkness (same as complete sunlight) has been considered since the hillshading is used here as a temporally dynamic variable as extensively described in the manuscript.

One appreciates mention and use of the RadarSAT DEM, but even 200 m resolution (which they interpolate to 1000 m) misses most relevant surface texture. Higher-resolution airborne radar surveys over large areas of the ice sheet show flat smooth areas of various extents (over basal lakes) amidst much rougher ridged and fractured ice, often (evidently) with substantial temporal evolution. Again, they apparently have no idea. Their predictors have no relevance.

We are working on 1 km resolution in agreement with the MODIS LST data. Though higher resolution DEMs exist, LST can, with this temporal resolution, only be provided at 1 km spatial resolution. We agree that 1 km misses a lot of the high resolution features, however, 1 km is still a huge improvement upon the existing temperature products available at continental scale.

Also the first reviewer was asking about the accuracy of the DEM and was suggesting using the DEMs from Bamber et al. (2009) or Slater et al. (2018), rather than RAMP DEM.

Although the RAMP DEM might have an accuracy not sufficient for e.g. glacier drainage basin delineation for mass balance analyses (Cook et al., 2012), however has been indicated to be suitable as a surface topography dataset (Cook et al., 2012), which is what we need it for. Please note that not all the characteristics of the DEM were used as predictor (see results after variable selection), so the only terrain related relevant information is daily maximum hillshade. We compared the results for daily maximum hillshade for the DEM of Bamber et al. with hillshade derived from RAMP and found that the differences are small (R^2 for the entire study area = 0.98 and with focus on the DryValleys where hillshading is most variable= 0.84). Therefore we are convinced that no change in the results must be expected.

We will justify our choice of the DEM in the manuscript: “*The Radarsat Antarctic Mapping Project (RAMP) digital elevation model (DEM) (Liu et al., 2015), version 2, was used, which has been indicated to be suitable as a surface topography dataset (Cook et al., 2012)*”

Their primary tool, MODIS LST, has demonstrated and much-argued weaknesses over snow and ice, both for cloud masks and surface temperature extractions. One might have hoped that Meyer et al (2016, the predecessor to this work and again much cited by these authors) might have addressed if not offered new resolution to some of those known issues but that paper

blithely accepts MODIS products (citing primarily mid-latitude terrestrial examples) as de facto valid despite a large, vociferous and continuing debate about applicability, suitability and errors over snow and ice. Until or unless these authors demonstrate and document new algorithms or techniques to improve performance of MODIS products over snow and ice they and we must regard this particular application as un-proven at best. A large literature, none of if cited here, debates these troublesome issues of single or multiple sensors and their individual or combined effectiveness at retrieving surface air temperatures over snow and ice. Again, the authors demonstrate no competence whatsoever in the use of MODIS LST.

Also following the recommendation of the first reviewer, we included a more detailed information on the MODIS LST product.

"The data are cloud-masked using the MODIS Cloud Mask algorithm (Ackerman et al., 1998) that applies typical thresholds in the visible and infrared channels. Though the MODIS LST product is cloud masked, the "white on white" and "cold on cold" effect is a challenge for cloud detection in Antarctica (Allen et al., 1990). This holds especially true for cirrus clouds that could in parts not reliably be detected in the used LST product. This is an ongoing challenge and further research effort on this will certainly improve the presented AntAir dataset in the future.
The MODIS LST data are reported with a quality of better than 1°C in the range from -10 to 50°C (Wan et al., 2004). However this did not involve an extensive validation for Antarctica. For the antarctic McMurdo Dry Valleys, Wan (2014) reported a mean error of 1K. Note that a general bias is not problematic for this study due to the applied machine learning based regression approach, but that robust relationships is relevant. Here, previous studies have indicated robust correlation between MODIS LST and measured air temperature, e.g. Wang et al., 2013 over the Lambert glacier drain or Li et al. (2016) between measured snow surface temperature and MODIS LST."

After casual application of four different machine learning techniques, the authors in the end rely on visual inspection!!! Their complete inability, despite multiple runs of multiple software tools, extensive spatial and leave-out cross-validation, to rely on any single outcome despite extensive statistical evaluations disqualifies the entire effort.

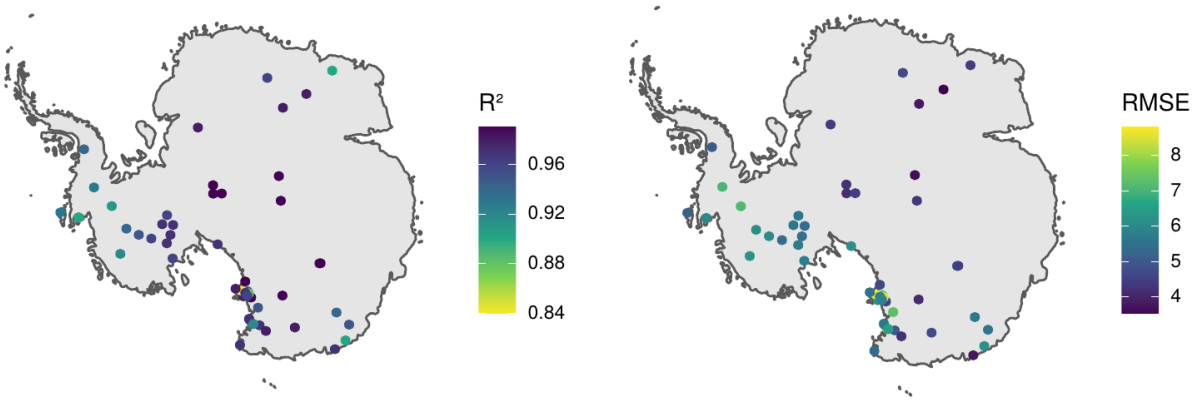
We disagree with the reviewer's view on the model validation strategy, and found the wording of a "casual application" surprising, as the procedure of extensively tuned and validated models go far beyond most commonly seen validation strategies.

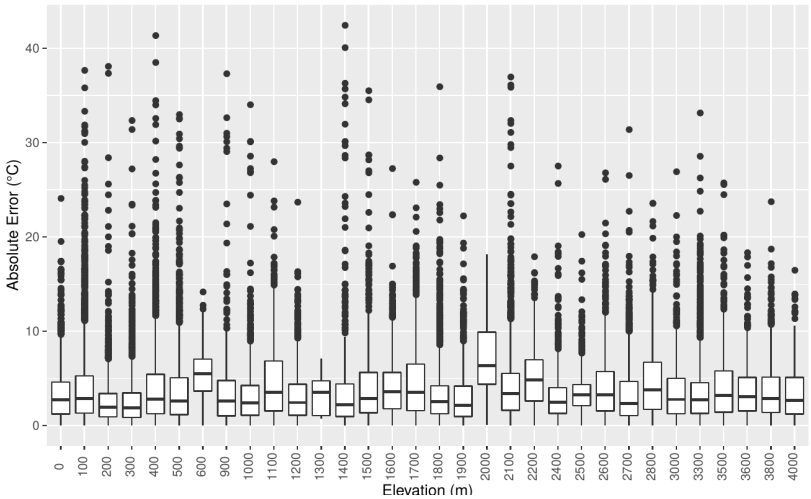
Moreover, in this paper we are not "relying" on visual inspection, rather, but we acknowledge visual inspection as an additional validation step, in addition to the outlined a) spatial and b) temporal validation by independent data. It is widely documented that blindly looking at statistical values is very dangerous when analysing what flexible algorithms actually learn (this is a large topic in machine learning, see e.g. Lapuschkin et al., 2020 or Patrick Schramowski et al., 2020). We ensured that the algorithm is learning meaningful relationships (meaningful in terms of spatial and temporal prediction for new unseen data) by spatial variable selection, spatial and temporal statistical validation, and a final expert inspection of the model predictions by visual analysis. It has been shown by many other studies that it is not sufficient to look at statistical values, and that expert knowledge should not be left out in a validation procedure.

- Lapuschkin, S., Wäldchen, S., Binder, A. *et al.* Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun* **10,** 1096 (2019). https://doi.org/10.1038/s41467-019-08987-4
- Schramowski et al. (2020): Right for the Wrong Scientific Reasons: Revising Deep Networks by Interacting with their Explanations. arXiv:2001.05371

This potential user might have asked for fine-scale validation in areas of (relatively dense) met measurements or perhaps RMSE sorted by elevation, but why bother?

Following the constructive suggestions of the first reviewer, we included a figure showing the RMSE/R^2 per station. We can also provide a figure for a revised manuscript, showing the RMSE sorted by elevation, although, no significant patterns are obvious.





They literally have nothing valid to show. To report absolute and RMS errors of 5K seems absurd. Who do they think might use such imprecise unreliable data? From any of several authors (try anything from Scambos, for example) these authors should know concern about long-term climate induced trends of 0.3 to 0.4 oC per decade over higher elevations of East Antarctica. Here they can't provide better than 5K? Over 15 years (assuming their time period of 2003 to 2018 (but apparently 2003 to 2016 according to page 11 line 6)), we might expect temperature change of perhaps 0.6 oC? Even by yearly averaging (1.73 oC, page 11 line 5), they fail to come close to necessary precision. They refer (e.g. page 8 line 31) to "RF being superior in the temporal prediction" but they fail entirely to demonstrate necessary temporal skill. Again, once senses that they fundamentally do not understand the system they attempt to model.

Indeed, we have higher errors than in other studies related to air temperature estimation from satellite, but we do not believe this is due to "going backward in precision", but rather to the very strict validation strategy that we used here. We validate our model using independent spatial

locations and years different from what has been used for model training, and argue that while the validation statistics are less impressive than in a lot of studies, it is also a lot more realistic. We refer to Roberts et al. (2017), Meyer et al. (2018), Valavi et al. (2018), Pohjankukka et al. (2017) to support our claim.

- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F. & Dormann, C. F. (2017), 'Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure', Ecography.
- Meyer, H., Reudenbach, C., Hengl, T., Katurji, M. & Nauss, T. (2018), 'Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation', Environmental Modelling & Software 101, 1– 9.
- Pohjankukka, J., Pahikkala, T., Nevalainen, P. & Heikkonen, J. (2017), 'Estimating the prediction performance of spatial models via spatial k-fold cross validation', International Journal of Geographical Information Science 31(10), 2001–2019.
- Valavi, R., Elith, J., Lahoz-Monfort, J. J. & Guillera-Arroita, G. (2018), 'blockcv: an r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models', Methods in Ecology and Evolution.

What, by the way, do bold values in Tables 1 or 2 indicate? Some kind of statistical certainty of statistical summaries? And what do the axis units in Figure 4 indicate?

The caption of the tables 1 and 2 point on that: "The best performances are highlighted". We will include the more detailed information that they are "highlighted in bold" for a revised version. We will also add the coordinate reference system to the units of Figure 4.

One gets the strong sense that we have gone substantially backward in precision, accuracy and reliability with this product.

See our comment above. Also, there is currently no spatio-temporal dataset we are aware of, that has a better performance in this spatio-temporal resolution.