Review of " An improved glacier parameterisation for the ecLand land-surface model: local, regional and global impact"
by
Gabriele Arduini, Christoph Rüdiger, and Gianpaolo Balsamo

The authors present results of evaluation tests of a new tile within the ecLAND module of the IFS-code of the ECMWF. I have read this paper with interest - the paper is generally well written - but needs improvement in the depth of the evaluation. The authors now convincingly show that the code update is an improvement. However, it would be good if the authors show also the performance against existing models, to see if the new tile/module performs similar as those models, or that there is still room for further improvement.

The essential improvements are those in the evaluation of the skin temperature, and especially of those of the subsurface temperature and to add an evaluation of in situ surface mass balance. I would genuinely like to see the technical improvements that I think are essential for better representation of the SEB and SMB without much coding - at the other hand I am afraid this request/wish will be in vain.

Comments
Introduction: glaciers and ice sheets are two different things, so by using "glaciers", ice sheets are sometimes forgotten. Therefore, I think it is better to replace
"glaciers" of lines 24, 32, 39, 106 to "glaciers and ice sheets",
"glaciers" of lines 30, 94, 97, 100, 107 to "glaciated surfaces".
L 108 is a special case, I leave it to the authors to make it more general. And check other instances when I have missed the word "glacier" in this list.

L 34: Greenland Ice Sheet - so with capitals.

L 38: Don't forget that for ice sheet, Earth System Models (like CESM2, UKESM) perform increasingly well (if the ESM developers try to model the SMB well), so regional climate models are not the only alternative of using observations only. These things are discussed later in the introduction too, so this formulation is already somewhat inconsequent.

L 43-54: I don't think this sidestep to ice dynamics is relevant for this manuscript nor introduction. For example, the initial-condition problem of ice sheet models is completely different to those of atmospheric models. I propose to remove this paragraph.
L 48: I think this reference if outdated, take a newer one if you decide to retain this paragraph.

L 65: A regional climate model is not an Earth System model. So, HARMONIE-AROME improvement are not an example of ESM improvements.

L 74-82: Given that the land-ice parameterization of ecLand is simple and uses very few layers compared to the advanced schemes in, particularly, polar adapted Regional

Climate Models, I would like that the authors formulate very specifically what their new parameterization should do well, and what not necessarily. For example, is that the SEB (including albedo evolution) for the typical range of a weather forecast (14-21 days), on seasonal timescales; does it include the surface mass balance (including refreezing effects)?

L 94: As far as I can recall, this simple snow scheme has been used also in earlier IFS cycle. If so, replace by "In CY49R1 (the NWP ....2024) and preceding versions since version <first version with this code>, glaciers and ice sheets are ...."

L 135: This method of confining the skin temperature to the melting point is already long in the IFS code even before Arduini 2019. Still, I remain to the opinion that this workaround is a poor solution, given that one can easily solve this issue in a mathematically sound and numerical simple way. We have used previous IFS versions (e.g. CY33R1) and in those versions considerable errors arose. Specifically, the time step after observing melt, the initial skin temperature guess was below the freezing point, the large conductivity number was not used, and the actual skin temperature became again well above the melting point. Even if this problem does not arise in this version, the method of Arduini leads to inconsistent skin temperatures if the uppermost snow or ice layer is not at the melting point.
Our solution was and is the following: When one observes that the skin temperature of a snow/ice/glacier tile is above the melting point, we keep the skin conductivity as is, but (simply) apply that one thus knows that the skin temperatures of that tile is at the melting point. If those tiles cover the whole grid box, one only need recalculate the fluxes. If these melting surface cover only a part of the grid box, we solve the SEB again, applying that we know the skin temperature of these melting snow/ice tiles, while the other tiles remain unknown as in the normal linearized SEB solve method. The authors can have more extensive documentation of our approach if they wish to have.

Section 2.2.2:
To me, this decision to not create separate prognostic variables for glaciated variables is a very, very poor compromise. For the glacial ice it is acceptable - sea ice is not active over land - but for snow I really do not see the urgency to mess up your snow physics to safe 4x4 prognostic variables (layer mass/thickness, layer density, layer temperature, layer water content). Snow surfaces, and particularly the snow albedo, is a classic example of a non-linear process, so the aggregate of two implies that now both are wrong.
We have no longer the computers of the 70s for which fast memory was a severe limiting factor and is the fancy type system of IFS not specifically set up to allow for adding new variables without having to adjust the code from top-to-bottom?
It sounds like that the authors were allowed to play around and improve the representation of glaciated surfaces; under the condition they won't bother the rest of the ECMWF-IFS community in any way. Don't get me wrong, I applaud the efforts of the authors to improve the representation of glaciated surfaces, but this compromise is very typical of the general and decades-long neglect of glaciers and ice sheet surfaces by the ECMWF.

Ideally, this poor compromise is rectified, and glacial snow is separated from land snow (and if you are doing that, please also separate snow over low vegetation and the snow below high vegetation into two independent snow layers sets, as that mix-up is equally bad). But I also do understand this strong suggestion (separate variables) is infeasible to effectuate (now), and I am aware this rant leads to zero change on the code or paper. However, hopefully it encourages the authors (or their successors) to fight even harder for a proper representation of glaciated surfaces in the IFC code. In all cases, I would like to see a longer motivation in the rebuttal why this poor compromise has been taken.

L 168: Please specify what Tsn is. It could be the temperature of the uppermost snow layer or the glaciated surface skin temperature - I don't know now.
I presume the authors are aware that this snow albedo (Eq. 3) includes two major simplifications. Firstly, the snow albedo is not a function of temperature (or density-per sé), but of snow grain size (and to some extend to the solar zenith angle). Secondly, the snow albedo is very strongly dependent on the wavelength, almost always 1 for UV to yellow; strongly varying for "red" and zero-ish for near-infra red - see e.g. Gardner and Sharp, GRL, 2010, or Van Dalum et al, The Cryosphere, 2020). So, using one albedo value number is similarly bad as running ecRAD with only one G-band. Again, I know it is not realistic to expect the authors to use a state-of-the-art grain-size based snow albedo scheme, but it is good to mention in the manuscript that this albedo is not regarded as state-of-the-art and specify its limitations with a reference to Gardner and Sharp, GRL, 2010 or a similar paper.

Section 2.4: It doesn't become completely clear to me how the experiments are carried out. This holds for both the point-scale as 2D global simulations. I understood from the description that the land model is rerun, but the atmospheric model not. Whether I got it right or not, explain in more detail which parts of the code have been rerun and which not. Furthermore, please specify which fields/fluxes are updated/adjusted in the experiments, and which fields/fluxes were kept constant. I conclude from the paper that the SEB has been recomputed, but I don't understand how that is done as that is far from trivial to do afterwards offline within the IFS framework (e.g. one needs the derivatives of fluxes to close the SEB and I would be seriously surprised as these derivatives are available from the ERA5 simulation.).

Figure 1 & 8: please use a map projection that shows Greenland with the right width to length ratio, so either Lambertian, Polar Stereographic or rotated lat-lon.

L 175: From the PROMICE dataset, the authors use the skin temperature. However, this temperature is not measured. I presume that the authors use the temperature derived from the upwelling longwave radiation. Please specify that explicitly here.

L 220: Make this description of the CLIM data more specific. Fgl is thus not 1 - otherwise that would have been stated - but still it would be sensible to use the tiled skin temperature for Figure 2-4. State which values have been used - grid-box average skin temperature or tile skin temperature. From line 275 I conclude that grid box averaged skin temperatures are used, it is better to adjust this - as the paper is about evaluating the glacier tile and the observations are on the ice sheet. So, why is not the tiled skin temperature analyzed? This parameter can be exported, so practical reasons are not impeding this.

I like the analysis in Figures 2-4 - it is a good method to show what the new module can and cannot. The drawback is that it is hard to compare how well this new module performs compared to existing glacier surface descriptions, as I haven't seen it in other papers. Therefore, assuming that the PROMICE skin temperature is derived from the upwelling longwave radiation, the golden standard is the modelled skin temperature with a SEB model (e.g. https://doi.org/10.1017/jog.2024.68). It would be good add the performance of such a dataset, to compare it against the CTL-OBS and GLA-OBS results. Similarly, the "E5" and "E5-CLIM" could be compared of the perfomance of a polar adapted RCM like MAR (https://doi.org/10.5194/tc-14-957-2020) or RACMO (https://doi.org/10.5194/tc-12-811-2018 - although this dataset is outdated). I am quite sure the required data for such an analysis is available for the authors to be used.

L 252-260: The interpretation of Figure 3 is complicated without analysing the SEB and the T2m. An T2m temperature analysis could indicate if the atmospheric conditions of the E5 and E5-CLIM simulations are colder/warmer than those in the observations. Similarly the SEB analysis (against the PROMICE data) could indicate why the conditions are colder than observed. These figures don't need to be added to the manuscript, but such an analysis add depth to paragraph, which is now not much more than "Hmmm, our model is warmer/colder than observed." In that respect, the 3K warm winter bias of both OBS is remarkable, as for this experiment the 2m is "correct". In short, figure out why the new and old model deviate, and report that in the revised manuscript.

Figure 4: add in the caption that summer months are evaluated.

Concerning Figure 3-4, to which extend are the difference due to elevation difference between the observational site and the height of the grid box? Elevation biases induces temperature biases which have nothing to do with poor functioning parameterizations. Remove the effect of an elevation bias (if present).

Conserning Figure 4, again I would like to challenge authors to dive a bit more deeper into the 'why' the model is deviating from the observations. Very little physical explanations are given, and the manuscript (and your understanding of the model performance) will be improved if this indepth analysis is made. Again, the avenue to get this insight is through analysing the SEB, and again these figures don't have to be added to the manuscript (possibly supplementary materials), but allows for a more physical explanation why the model is deviating. I would guess that the underestimated cycle for the high locations is due to too high effective thermal capacity of the snow layer (being 50 cm thick), while alternatively missed nighttime refreezing (normally dampening cooling of the surface during the night) is the cause of deviations for the lower and upper ablation sites.

Section 3.1.2: Snow temperature

Snow temperature is a very good indicator of the performance of the model, and it is a very good idea to analyse and discuss this here. However, measuring snow temperature is in some sense trivial but using it is very complicated. With thermistor strings you can easily measure the snow temperature on a give location of the snow, either below or initially above the surface, depending how the string is installed. Given this installation, the temperature sensor moves with the snow pack in which is is burried, or stays at a given height above or below the measurement frame (like an Automated Weather Station) it is attached to. In all cases, the actual snow surface is moving away and towards the sensor all the time, so a sensor is never all the time at, say, 1 m depth below the surface. Compared to the data available online at the PROMICE website, the shown observational curve equals to those of "sensor 1" - although the online dataset has considerable datagaps in the summers of 2016 and 2018. The Fausto paper indeed state that this sensor is/was at 1 meter depth, but winter accumulation and summer snow melt are both well over 1 meter at TAS-A, so it could be - well, has been - anything like 0 to 2 meters. Given that sensor 2 has positive values in the summer of 2016 (after which the snow temperature sensors are reinstalled, visible in the shift in all readings), that happened for sensor 1 as well.
I really would like to see that the authors can retain this analysis, but that does require that they reconstruct the actual depth below the surface of "sensor 1" using the observed surface height - and not only for TAS A but for all stations used in Figure 5. After reconstructing the actual depth of the sensors, the adjacent model temperature can be extracted from the model data and compared to the observations.
If this is not possible, another way to assess the subsurface temperature is to replicate the evalution as provided by e.g. https://tc.copernicus.org/articles/15/1823/2021/.

And if this analysis is retained and improved, then it would be great if the results could be compared with subsurface temperatures from what is considered advanced surface models for either (or both) SEB models and or polar regional climate models, similar as requested for Figure 2.

L 287 & Figure 5: State explicitly that this Taylor diagram and bias on annual data.

Section 3.1.3 Trend in melting occurrences. Personally, I am more interested in the model performance than in the trend, as the latter has been documented in many papers already. So focus more on the statistics - in how many cases are the melting days well predicted, missed and are there many false alarms (and how does this compare to other models) - and remove the trendline.

Figure 8: please replace panels b and d by panels with the biases compared to the observations. A reader should be able to eyeball the improvement compared to CTL, but I don't want to eyeball how far off the GLA is from the observations, which is the current situation. And please be aware of the issues with the MODIS albedo if the solar zenith angle is larger than 70 degrees - which also may explain the positive bias in North Greenland in panel 8c.

L 338: I'm fine if you retain the reference to Zorzetto, but please be aware such parametrizations are already used for over a decade in polar adapted RCM like RACMO and MAR, the latter through the snow model CROCUS. Acknowledge that with appropriate references.

Section 3.2.2. This analysis is useful, but a common practice in other papers is to compare modelled SMB against the in-situ observations compilation from Machguth (J Glac 62, 2016, currently being updated by DMI) like done in Noël et al, 2019, Sci Adv. (Supplementary figure 3) or a subset of that like van https://tc.copernicus.org/articles/15/1823/2021/.

L 367: rephrase, ice sheets are not glaciers.

L 373: Please explain briefly, in the methods section, the Kling-Gupta Efficiency. Now every reader not familiar to it, is forced to dive up Gupta et al.

Figure 11, panel a. Not the KGE is shown, but the difference or change in KGE. Adjust the caption accordingly.

Section 3.3: KGE and glaciers are not my expertise, but when KGE is a common measure to evaluate river discharge, there are other model evaluation studies that have given KGE scores - so cite a few other studies and their scores to give the reader a clue if the KGE scores GLA and CTL are good or poor.