Referee comment on the paper *Evaluation of MARv3.14 over the Greenland Ice Sheet* by Timmermans et al. submitted to the TC.

General comments

The authors are presenting a new MAR version, v3.14, and have gone through extensive and thorough model evaluation in the context of simulated surface mass balance (SMB) over the domain of Greenland Ice Sheet. MAR outputs are compared to a range of observations, including in situ measurements, AWS records, satellite-derived melt extent, and albedo products. For example, the discussion on ablation underestimate (Sect. 4.1) describes, in detail, comparison to SMB observations from station to station. They have also conducted a model resolution-dependency analysis.

The paper is generally well structured, but needs some technical editing. I would like to see a structured Conclusions section. There are a lot of comparisons made in this paper, and the Conclusions section is the perfect place to summarize the main points found in the analysis that are of interest for a MARv3.14 (data) user. Please provide concrete, specific outcomes rather than vague, generalized summary. The Results section needs language polishing / stylistic editing. Namely, the over use of parentheses and over-referencing should be reviewed, as it severely disrupts the reading. This is also true in the Discussion section. At times, the Results section paragraphs are hard to read, as there is no clear structure (e.g in 3.4 Albedo representation). I suggest starting with a strict lead sentence, which is the opening sentence of a paragraph that introduces its central point, and then following with supporting sentences with data. As this is a paper using a lot of different datasets to evaluate MAR, the results need to be presented clearly.

I have only some minor scientific comments, mostly asking to expand and elaborate. Overall, the evaluation is done properly and the results make sense. MAR is an important tool in climatological studies, and such a thorough and rigorous improvement and evaluation is appreciated.

Suggestions for improving the content are provided here. I recommend the paper to be published at the TC, after these minor, mostly technical, improvements.

Specific comments [scientific]

1. L31: *Several approaches have been developed to estimate the GrIS SMB including positive degree days, energy balance, regional (RCM) or global circulation models (GCM) (Fettweis et al., 2020).*
   An important takeaway of the paper of Fettweis et al. 2020 is not the existence of several approaches of SMB estimation, but that RCMs out-perform other methods when validated against satellite and in situ field observations. I think this is an important motivation of this paper of development of MAR and should be emphasised here.

2. L90: *Grid-cells with an ice fraction lower than 50 % are discarded to avoid unrealistically high ablation values near the ice sheet–tundra interface.*
   I would like a more thorough explanation on what unrealistically high ablation values are, so as not to disregard data semi-arbitrality from the analysis. Furthermore, is this hard-coded within the model (as in grid-cell fractionation is only from 50% ice fraction and above), or is it that you later only analysed GrIS pixels with grid-cell fractionation > 50% ice?

3. L147: *We present and discuss the results obtained using a melt rate–based criterion, whereby a grid-cell is classified as melting when the simulated melt exceeds a threshold of 0.2 mm w.e. day−1 , following Fettweis et al. (2011) and Puggaard et al. (2025).*
   Later, in Sect. 3.5, Sensitivity tests that derived the MARv3.14 threshold of 0.2 ± 0.1 mm w.e. day−1 is referred to Table S8. Puggard et al. 2025 determined 1 mm w.e. day−1 to be MARv3.12 threshold using ROC and PR curves. Fettweis et al. (2011) tested a range of values, but suggested 8.25 mm WE day−1. I do expect different thresholds in MARv3.14 and previous versions. Your methodology of minimizing the RMSE tests quite low values, 0.1-0.3 mm w.e. day-1 and does not test the values suggested by Fettweis et al. (2011) and Puggaard et al. (2025). Could you expand on the choice of low threshold values in the sensitivity tests? Why such a large difference between 1 (Puggard et al. 2025), 8.25 (Fettweis et al. 2011) and your 0.2 ± 0.1 mm w.e. day−1? Is it the methodology of determining the threshold, or is it changes between the MAR version (such as implementing meltwater freezing or ecRad?). Could you have hit a local minimum with 0.2 mm w.e. day−1? This is relevant also considering the results you obtained: *Figure 6b indicates that the use of a melt threshold of 0.2 mm w.e. day−1 results in an overestimation (> 30 days per year) of the mean annual number of melt days over southeastern Greenland.*

4. L215 and L313: *Note that our results show similar biases and correlation to those obtained from another RCM (i.e., RACMO2.3p2 at 5 km (Noël et al., 2019) using the same near-surface meteorological variables and radiative fluxes. Both MAR and RACMO2 tend to underestimate high wind speed.*
   Also in the Discussion section: that the biases are “of the same order of magnitude as another state-of-the-art RCM at 5.5 km“.
   Two points here. Firstly, please support the “similar biases and correlation” and “*tend to underestimate* ” claims by giving numbers, so the reader does not have to go to Noël et al., 2019 paper and compare these results side by side. The second point is technical: there is one bracket too much: (*i.e., RACMO2.3p2 at 5 km, Noël et al., 2019).*

5. L232 and L238, sentences: *…but also reflects a systematic negative bias (~ −0.10), likely originating from the inability of the snow albedo scheme of MAR to capture high fresh snow surface albedo (~ 0.87) in accumulation zones.*
   and: *We find that MAR overestimates low albedo values in ablation zones exposing bare ice in summer. In fact, MAR does not simulate values below 0.35 whereas values as low as 0.1 are measured by AWSs. This discrepancy stems from the MAR albedo scheme,*

which constrains the prescribed MODIS-derived ice albedo between 0.35 and 0.55 to prevent the development of unrealistically low albedo values over large areas (25 km2).
Please add extra information about the albedo scheme to Sect. 2.1. From the MAR improvements description, the available information there lead me to believe that bare ice albedo (prescribing MODIS observations instead of a surface meltwater dependent bare ice albedo ranging between 0.5 and 0.55 as before) and the possibility for the meltwater to freeze above bare ice, would allow for lower albedo values than 0.35. Furthermore, how are meltpond albedos taken into account? When it comes to fresh snow albedo, is there a maximum value?

6. Figure 5a.
   Where are the horizontal line features in the summer albedo on Fig. 5a coming from?

7. L265: *Moreover, MAR suggests the presence of aquifers in this area where melt occurs sometimes 10 m or 20 m below the surface.*
   Do you mean that meltwater percolates and is stored 10-20m below the surface in MAR's simulations? Could you expand a little on that? How surface melt water reaches the ocean, and how fast it does so, is poorly understood, but an important process for SMB. These depths, or even deeper, are realistic, eg. Humphrey et al. 2021 and Forster et al. 2013.

8. L340: *Additional biases, such as in precipitation, drifting snow erosion and deposition, sublimation, and measurement uncertainties, may also play a role.*
   I agree. But, is drifting snow erosion to some extent included in MARv3.14 snow scheme or none whatsoever? This sentence also raises points not discussed in the paper: drifting snow erosion, deposition and sublimation, while precipitation was evaluated between different MAR resolutions. These topics stay out of the scope of this paper and are difficult to evaluate. But, this sentence raises the question: what about them? I suggest spending a few more sentences explaining why these SMB components were not evaluated in the paper, and referencing MAR papers that have done so (e.g MAR (v3.11) drifting snow paper by Amoury et al. 2021) and what they found.

9. *L411: Therefore, future developments should prioritize improving the representation of clouds and associated precipitation, which directly affects both radiative fluxes and accumulation.*
   You are emphasising a conclusion that is not really backed up by this paper. While I am sure that there are biases coming from clouds and precipitation, and while these issues are coupled also to the surface conditions e.g via albedo, this paper does not evaluate clouds and precipitation. You should conclude what you learned from your evaluation, and if there is a suspicion of poor cloud and precipitation processes from the biases that cannot be explained with surface processes (e.g Sect 3.3), then we should conclude that clouds and precipitation needs further study. I would like to know, however, what future developments should prioritize, considering this evaluation.

Technical corrections

1. L25 (**PR ;** the sum of snowfall and rainfall) - unnecessary space
2. L36: *Since then, numerous developments have been implemented in versions 3.10–3.14, as well as an increase in spatial resolution over the Greenland domain, i.e., from 15 km (Fettweis et al., 2020) to 5 km in this study.*
   Increase of the HIGHEST spatial resolution? As one of the next paragraphs states, 30 km MAR simulations exist, so it would be fitting to show the true extent of the model resolution increase or explicitly state that the highest suggested resolution has increased.
3. Section 2. This is not a requirement, but I would think that a separate Section (2) should be reserved to MAR only. The data you use to evaluate MAR should have a Section of its own (Section 3). This would lead to a better overall structure of the paper.
4. L93: We use SMB measurements from the GrIS SUMup database (Vandecrux et al. (2024) ~~; purple dots in Figure 1a~~).
   L96: (Ohmura and Reeh (1991)~~; green squares in Figure 1a~~).
   Leave these for the caption, not useful in the main text.
10. L94: *Here, we compare modelled SMB with measurements derived from stakes, snow pits, firn and ice cores, snowfox records, mass balance profile estimates and δ 18 95 O dating.*
    Be more concise with words: either use “SnowFox” records, or better yet, “ground-based SnowFox cosmic ray neutron sensor snow measurement records”.
11. L157: *Applying topographic correction (Eq. (1)), SMB bias and RMSE decrease by 27 % and 20 % respectively, while correlation increases to 0.95 (Fig. 1c).*
    Double brackets *(Eq. (1))* → (Eq. 1), or [Eq. (1)].
12. L158: *The linear regression lines, which exhibit slopes lower than 1, indicate that MAR tends to underestimate both high ablation and high accumulation observed values.*
    It is noteworthy that elevation correction barely changes the regression.
13. L171: *Comparing MARv3.14 to airbone radar transects from Karlsson et al. (2016, 2020) leads to biases, RMSEs and correlations of -0.01 m w.e. and 0.02 m w.e. and 0.99, similar to those of MARv3.9.*
    →Comparing MARv3.14 to **airborne** radar transects from Karlsson et al. (2016, 2020) **reveals** biases, RMSEs and correlations of -0.01 m w.e. and 0.02 m w.e. and 0.99, similar to those of MARv3.9. (Biases are already in the model and in the radar measurements). Air-bone radar typo is also in Table S1.
14. L170: *We find that MARv3.14 (5 km) improves on MARv3.9 (15 km), while model differences remain statistically non-significant.* And elsewhere.
    A Results section should be readable and understandable on its own, even without looking at tables. Please back the claims of “improving” and “*remain statistically non-significant”*. No need to write out all Table S1 results, but please back up your claims either with an example or range of values. Similarly to paragraph expanding L174-L178.
15. Figure 1: At least inside one Figure, please use the same units, either mm w.e or m w.e
16. L193: *At coarser resolutions (Table S3), the agreement with observations decreases slightly in terms of RMSE, correlation and bias. However, the differences are small, and the RMSEs remain systematically lower than the corresponding standard deviations.*
    A Results section should be readable and understandable on its own, even without

looking at tables. Please back the claims of “decreases slightly” and “*differences are small”*.

17. Sect *3.3 Radiative fluxes modelling* but also everywhere else
    The frequent use of brackets to give information and side notes makes it very hard to read. If every sentence has a parenthetical aside, it disrupts reading flow. In some cases, one sentence has several. Rewording some sentences would make the language better. For example: “Except for the RMSEs which are slightly larger for shortwave fluxes (~12 W m−2), no significant differences are found between the annual and the summer (JJA) statistics (Table S4).” JJA is already defined as summer months previously in the text. Instead of referencing tables in brackets, you can use full sentences: “Table S4 presents … etc.” In cases of sentences such as these: “The annual (resp. JJA) standard deviations (in W m−2) are 129.95 (resp. 93.60) for SWD, 50.08 (resp. 35.98) for LWD, 99.01 (resp. 91.84) for SWU, and 38.98 (resp. 18.85) for LWU.” You could write two sentences, one for annual, one for JJA.
    Another example is from Sect. 3.5: *Figure 7a not only shows that MAR accurately reproduces the onset of the melt season (around early June) as well as its ending (midSeptember), but also that the timing of the peak melt extent is well captured, on average in mid-July.* → Figure 7a shows that MAR accurately reproduces the onset of the melt season around early June, its ending in mid September, as well as the timing of the peak melt extent in mid-July.
    Please cut down the use of parentheses. If the information is essential to the argument, put it in the main sentence.
18. Sect 3.3 *Radiative fluxes modelling* but also everywhere else
    This relates to the previous point. The referencing Table S4 is done in multiple places, and in brackets. The whole section is really a description of Table S4 and Figure 3, so you can lead with “Table S4 and Figure 3 present…”, and then you don’t need to refer to Table S4 separately any more. Same true everywhere else in the Results section.
19. L221: *Daily OLCI GEUS albedo values (Vandecrux et al., 2021) …* also in the other parts of Sect 3.
    Be mindful of over-referencing. You have introduced the OLCI GEUS *(Vandecrux et al., 2021)* dataset in Sect 2. Model and observational data. Later in the text, you can just call it OLCI GEUS without referencing. Same true for the SMB dataset of Vandecrux et al. (2024) in Sect. 3.1.. After establishing reference in Sect. 2, call it GrIS SUMup database. Same true for “AWS observations from the PROMICE and GC-NET networks (Fausto et al., 2021; How et al., 2022; Vandecrux et al., 2023)...” in Sect. 3.2.
20. L269: *Focussing on the melt extent, it appears that it is slightly (~ 2–5 % of the GrIS area) overestimated during the first half of the melt season and underestimated after the peak.*
    The extent of the overestimation is given, but not underestimation. Also, there is no space between “~” and the number.
21. Sect. *3.4 Albedo representation*
    The Section needs better structure. Use a lead sentence, followed by the supporting sentences for each paragraph. Start with an introductory sentence for the whole Sect. 3.4, as you did in Sect 3.1-3.

22. L299: *Figure 4c shows that the RMSE exceeds the OLCI-GEUS standard deviation over a large fraction of the GrIS (hatched)...*
    No need to add “(hatched)”, as the colour of the points, it belongs to caption, not text.
23. Figure 6 and Figure 7:
    This information belongs to Sect 2.6, not into the Captions: using a melt-based criterion of 0.2 mm w.e. Day−1 (although fine if you want to really emphasise this), satellite data are derived from the ascending and descending passes (combined with “OR” criterion) of the AMRS-derived product Picard (2025) from Picard and Fily (2006). The comparison is made over the satellite data grid using an IDW interpolation.
    Furthermore, I cannot see Picard and Fily (2006) referenced in Sect 2.6 and the “the AMRS-derived product Picard (2025) **from** Picard and Fily (2006)” is either typo or confusing sentence structure.
24. L311: *In the previous sections, we evaluated the latest MAR version (MARv3.14) at 5 km resolution against a set of observational data including in situ SMB (Fig. 1), meteorological (Fig. 2) and SEB (Fig. 3) records as well as albedo (Figs.4 and 5) and remotely sensed melt extent (Figs. 6 and 7).*
    No need to reference the Figures here again. A simpler sentence without all the parentheses would make it easier to read: *In the previous sections, we evaluated MARv3.14 at 5 km resolution against a set of observational data including in situ SMB, meteorological and SEB records as well as albedo and remotely sensed melt extent.*
25. *L321: To investigate the origin of the ablation biases at 5 km, we compare MAR to summer (JJA) SMB …*
    Summer is already established as JJA, so please pick one to limit usage of brackets.
26. L326: *Interestingly, the comparison with pressure transducer data (Figure S3a) yields similar results to those shown in Figure 1b, despite the smaller sample size (137 vs 3012 points in Fig. 1).*
    Please write out what comparison was shown in Figure 1b.
27. L338: *However, while this radiative bias should result in an ablation overestimation of only ~ 0.7 mm w.e., we observe a bias of ~ 2 mm w.e..*
    Double punctuation.
28. Sect. 4.1 Ablation underestimate
    Here, the first paragraphs with station comparisons have metrics to back up claims of biases. But, from KAN_L and KAN_M comparisons, not. Only words like “underestimation”, “overestimation”, “small” are used. You need to use key findings in numbers consistently to prove physical points, explain biases, or compare with literature.
29. Sect 5. Conclusions
    Please provide concrete, specific outcomes rather than vague, generalized summary. There were a lot of comparisons made, and a Conclusions section is the perfect place to summarize the main points found in the analysis that are of interest for a MARv3.14 (data) user.