We thank both reviewers for their useful comments and positive feedback. The suggested revisions substantially improved the manuscript, and we addressed all comments, point-by-point, in this document. The comments of the reviewers are shown in black and our replies in blue. We number reviewer comments for referencing purposes throughout the document (comment 1 = C1, etc.). The line, figure and table numbers are based on the updated manuscript.

Responses to Referee #1 (R1)

Summary

This manuscript presents an overview over the performance of ERA5, ERA5-LAND and dynamical downscaling with the WRF model (to dx~3km and dx~1km) for calculating the surface energy balance (SEB) of mountain glaciers over Western Canada. Four glaciers are chosen for evaluation, where observations of the relevant variables (e.g., turbulent heat fluxes, temperature, radiation, wind speeds, etc.) are available during the summer season. The authors derive the simulated variables for the SEB, after some corrections, from the model output and directly compare the results with the observed values. Furthermore, they run the WRF model in multiple configurations for parametrizations to find the "optimal" setup for a satisfactory calculation of the SEB. Results suggest that dynamical downscaling with WRF does not automatically outperform ERA5, except for the wind speed and direction - mostly due to the higher horizontal resolution. Generally speaking, both ERA5 and WRF are useful for calculating the SEB, while a correct simulation of the meteorological fields over the glaciers would require even higher horizontal resolution at the hectometric range.

The manuscript is extensive and has a valuable purpose in discussing the challenges of dynamical downscaling over glaciated environments and suggesting an "optimal" setup for future applications. However, in some sections, the authors need to argue in more detail on why they apply a new method; some reasonings are given in the discussion, while they would be already required in the methods section. The interpretation the WRF results is sometimes lacking an important factor - namely terrain resolution. Comments and suggestions are given in the list below.

We thank the reviewer for their useful comments and positive feedback that helped us improve the quality of the manuscript. By addressing these comments, we have been able to provide a clearer explanation of the need for a new method and enhance the discussion of key points.

Major comments

R1 C1: **Calculation of the surface fluxes from model output via the bulk method**. I agree that the modelled albedo values strongly differ from the observations; however, while only reading the methods it is difficult to follow the argumentation why the authors decide to calculate the turbulent fluxes with the observed albedo via the bulk method instead of directly using the values for sensible & latent heat fluxes from model output. Is this a common method to utilize output from an atmospheric model for glacier SEB modelling- was this approach also used in previous studies?

We thank the referee for this comment, and we realized that more clarification on this issue should be introduced from the start (in the “Data and Methods” section) rather than in the “Discussion” section. As we see it, there are two points raised by the referee: one is the use of observed albedo rather than derived albedo from ERA5 and WRF, and the other is the use of the commonly used bulk methods (in glacier studies) to derive turbulent heat fluxes rather than outputting these directly from ERA5 and WRF. (Please note that albedo does not feature in the calculation of turbulent heat fluxes, but in the calculation of net shortwave radiative fluxes – we guess that the referee points to the fact that albedo will be linked to roughness length, i.e. depending whether snow or ice albedo is on the surface the roughness lengths are determined.) Both albedo and turbulent fluxes, as daily timeseries, are known to be poorly simulated by ERA5 and WRF at relatively coarse spatial resolution (> 1 km grid spacing). We also show this poor simulation in the "Discussion" section. Prompted but the referee's comment, we now provide more rationale in the "Data and Methods" section to justify our use of the observed albedo and the bulk method (lines 291-308 in the updated manuscript). For accurately modeling turbulent fluxes, it is crucial that both ERA5 and WRF correctly represent the glacier surface roughness lengths, as well as the temperature, humidity, and wind speed in the surface boundary layer (SBL). Considering that ERA5 and WRF fail to do so at their respective grid spacing, their derived turbulent heat fluxes (regardless of the parametrization scheme used for the SBL) are far off from the observed values. We also added Figure 5 and Table 4 to the main text, representing the modeled vs. observed timeseries of daily albedo and seasonal roughness lengths, respectively.

Lines 291-308: The primary goal of the evaluation analysis is to assess the performance of the SEB model, forced with either ERA5 or WRF data, in simulating seasonal melt energy at our sites. To do so we evaluate the total simulated energy available for melt (Q_M ; Eq. 1), as well as the daily timeseries of Q_M , as calculated from the SEB model forced with the reanalyses (ERA5, ERA5-Land), as well as with the WRF output, against the reference calculations when the same SEB model is forced with the AWS data. Thus, the input for the SEB model, i.e. the atmospheric variables K_{in}, L_{in}, T, RH, and U, are taken from: (1) the AWS at each site, representing the reference or true values, (2) ERA5, (3) ERA5-Land, and (4) WRF at grid spacings of 3.3 and 1.1 km, using each of the three configurations (REF, min-NRMSE, and TOPSIS). For the reanalysis and WRF, only the data from the grid cell covering each study site is used. As we are interested in the evaluation of meteorological rather than surface variables (albedo and surface roughness), we use in-situ observations of daily surface albedo and seasonally-averaged roughness lengths in the SEB model. These surface variables could have been taken directly from the reanalysis and WRF; however, we found that these values can differ substantially from the observed ones throughout the observational periods (Figure 5 and Table 4). The discrepancy between WRF and observed albedo at glaciers, especially in the ablation zone, has also been noted previously in Eidhammer et al. (2021). Thus, to avoid any evaluation biases originating in poorly assigned surface variables, we stick to the choice of using observed surface variables in the SEB model. The observed daily surface albedo is calculated as the ratio of measured daily totals (in local daylight hours) of reflected and incoming shortwave radiation at each site. The incoming shortwave radiation at the surface is taken from these datasets without any further modifications (e.g. separation into direct and diffuse radiation).

Table 4. Mean seasonal roughness lengths [m] for momentum (z_{0v}) derived from the observations (AWS), ERA5 and WRF (1.1 km) at each study site.

Site	AWS	ERA5 (30 km)	WRF (1.1 km)
Castle Creek 2012	0.003	0.067	0.002
Nordic 2014	0.003	0.936	0.002
Conrad 2015	0.003	1.169	0.002
Conrad 2016 AWS_1	0.001	1.166	0.002
Conrad 2016 AWS_2	0.003	1.166	0.002
Kaskawulsh 2019	0.001	0.001	0.002

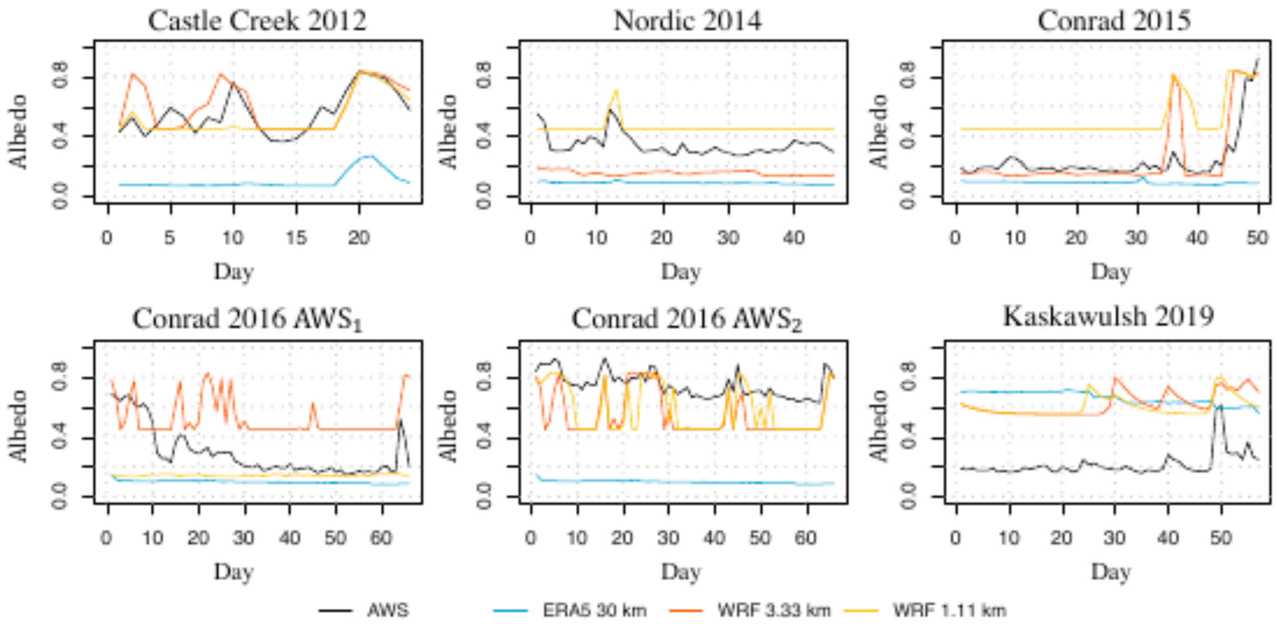

Figure 5. Modeled (ERA5, WRF at 3.3 km and WRF at 1.1 km) versus observed (AWS data) timeseries of daily albedo over the observational period (starting at Day 1) at each site. WRF is run with the REF configuration.

The authors mention in the discussion the unsatifactory performance from the turbulent fluxes from the direct model output (lines 497--508), but for the general understanding of the manuscript, it would make sense to add these paragraphs directly after they introduce the new method (ca. Line 385).

As requested, we now moved the discussion of the performance from the turbulent fluxes from the direct model output from the discussion to the “Results” section. We also provided more detailed rationale in the “Data and Methods” section (lines 291-308; see comment R1 C1).

Furthermore, changing one parameter to derive a quantity from the rest of the modelled output might lead to physical inconsistencies, because all the other variables used for the bulk method still indirectly depend on the "wrong" albedo. Did the authors calculate the SEB with directly modelled turbulent fluxes?

It is correct that we use the bias-corrected temperature, humidity, and wind speed directly from WRF/ERA5, while surface variables (albedo and roughness lengths) are derived from observations. The referee is correct that this leads to physical inconsistency in the use of modelled simulations; however the main objective of this study was to investigate how well an ‘off-grid’ SEB model (meaning that the SEB model is decoupled from WRF and ERA5 models) performs at a given location on a glacier, when forced with meteorological variables from WRF/ERA5 in comparison to when forced with AWS data. We now make this objective more clear and also mention the limitations associated with this approach (lines 297-317, see comment R1 C1).

Yes, we also evaluated the performance of directly modeled turbulent fluxes from WRF/ERA5 in the SEB model and this resulted in large biases relative to the SEB model when forced with AWS data. We had previously addressed this in the “Discussion” section, but to highlight this finding, we have moved it to the “Results” section (lines 403-413):

We also investigated the use of surface Q_H *and* Q_L *as outputted directly from the reanalysis and WRF into the SEB model, rather than calculating those fluxes with our bulk method. In WRF, these fluxes are derived through a local or non-local closure scheme in the planetary boundary and surface layer, depending on the parameterizations used (Skamarock and Klemp, 2008). When* Q_H *is directly taken from ERA5, the NRMSE of* Q_H *is 83 %, which is twice as large as the original error when* Q_H *is calculated with the bulk method. In WRF at 1.1 km, the error in* Q_H *is increased from 31 % when the bulk method is used to 60 %, while the error for* Q_L *is increased from 21 % to 54 %. For Kaskawulsh glacier, the largest glacier among our study sites, the performance of simulated* Q_H *and* Q_L *directly from ERA5 is similar or only slightly worse (few percent) than the performance based on the bulk method. However, looking across all the sites, taking* Q_H *and* Q_L *directly from ERA5 leads to an increased underestimation of mean* Q_M *from 6% in the original estimate to 72%. For WRF at 1.1km, the relative error in* Q_M *increased from 8% in the original estimate to 17%. These results justify our choice to assess the turbulent heat fluxes via the bulk method instead of taking them directly from the reanalyses and WRF.*

R1 C2: **Interpretation - terrain resolution.** The authors argue that the poor performance of wind speed and direction simulation yields from the inability to simulate the katabatic glacier wind. The authors could check whether the "bad" model performance only happens during the wind directions corresponding to the down-glacier wind - the model seems to perform better during synoptically-forced conditions. However, glacier winds are not the only meteorological phenomenon present over mountain glaciers; such as thermally-induced circulations, downslope windstorms, etc, which are all mostly governed by the topography (Goger et al, 2022). Therefore, well-resolved topography is essential for the correct simulation of the wind field - tis also explains the general bias reduction of wind speed & direction for small horizontal grid spacings (dx=1.1km and dx=370m). This is an important point which should be mentioned in the discussion and interpretation of the results. Publications from idealized simulations argue that at least 10 points across a valley are necessary to simulate the relevant processes well, and that the correct representation of topography is likely more important than the choice of parameterization schemes (Wagner et al, 2014).

We thank the reviewer for this comment. We incorporated a more extensive discussion on this topic in the paragraph on wind speed in the discussion (lines 586-594):

Apart from katabatic winds and synoptic storms, other meteorological phenomena mainly governed by topography, such as thermally-induced circulations and downslope windstorms, occur over mountain glaciers (Goger et al., 2022). Therefore, accurately representing the topography is crucial for correctly simulating the wind patterns. A better representation of topography explains the improved accuracy in wind speed and direction for smaller grid spacings in our simulations (1.1 km and 370 m). The finer grid spacing not only improves the elevation representation of the analyzed grid cell (Table S1), but also likely improves the elevation representation of the neighboring grid cells, leading to a more accurate representation of slopes and aspects of the terrain. According to Wagner et al. (2014), the correct representation of topography is likely more important for the simulation of local flow regimes and turbulent heat fluxes than the choice of physics parameterization schemes.

R1 C3: **TOPSIS and minRMSE configurations.** Maybe I missed it, but do the authors somewhere list the final WRF model setup of TOPSIS and minRMSE, like Table 2 for the REF run? This might be of use for future dynamical downscaling studies.

Yes, the final WRF setup for TOPSIS and minNRMSE was listed in the Supplementary Information. We now moved this table to the main text (Tables 2 and 3).

Minor comments

R1 C4: line 50: which simplified assumptions?

We revised the sentence (lines 52-54):

Nevertheless, as statistical downscaling relies on simplified assumptions (e.g., the existence of linear relationships between local and large-scale climate variables), the technique introduces another source of error or uncertainty into the model output (Marzeion et al., 2020).

R1 C5: line 57: make a new paragraph

Done.

R1 C5: line 83: An extensive analysis of real-case, high-resolution large-eddy simulations over a glacier is provided by Goger et al (2022), and Sauter & Galos (2016) performed semi-idealized LES over a glacier and evaluated the calculation of turbulent fluxes.

Thank you for pointing out these studies. We updated the references in the text (lines 88-90).

R1 C6: line 85: "Downscaling to several kilometers": Several kilometers might not be the optimal target for mountain glaciers embedded in highly complex terrain, which requires likely horzintal grid spacings of less than 1km.

We agree with the reviewer and revised the sentence (lines 90-92):

Therefore, when incorporating WRF into long-term glacier evolution modeling at regional scales, downscaling to a grid spacing of approximately one kilometer seems to be the computationally optimal target.

R1 C7: lines 134-203: I understand that it is important to mention the most commonly used parameterization schemes in WRF, but this is too lengthy for an introduction - perhaps it's enough to mention this configuration in the methods and finally say how it performs within the ensemble.

We have condensed the introduction to focus only on the challenge of identifying optimal physics parameterization schemes in WRF (lines 93-102). The detailed explanation of commonly used parameterization schemes in glacier studies has been moved to the "Data and Methods" section (209-221).

R1 C8: line 213: You can place the optimal configuration of parameterizations from the introduction here.
Done accordingly.

R1 C9: line 220: What do you mean exactly by "reflect different time windows during melt season"?

We randomly assigned the six-day periods for the sensitivity analysis to represent different time periods in the melt season (e.g. early, middle and late melt season). The sentence is revised accordingly (lines 229-230):

The six-day periods are selected randomly to represent different time windows throughout the early, middle, and late melt season.

R1 C10: line 398: " none of these altered WRF configurations yield a strong impact on the calculated Q_M from the SEB model": Did you reset the albedo for calculating the turbulent fluxes here as well? Because then this relative agreement is not very surprising.

The albedo and roughness lengths are set to the observed values and are kept the same across all three WRF configurations. Considering that the simulation of individual components of the SEB components, such as the radiative and turbulent fluxes, can vary substantially among the three WRF configurations (as observed in the sensitivity tests, Figure 14), the relative agreement in simulating QM is somewhat surprising. Despite the substantial differences in these individual SEB components due to different parameterization choices, the simulation of melt energy over the observational period remains relatively consistent for the three optimal configurations.

R1 C11: line 478: ...."do not distinguish between ice and snow categories": It's true that the land use category does not distinguish between snow and ice. However, after initialization, WRF indeed initializes snow cover on glacierized surfaces. The authors mention observed snow cover at one of the glaciers during the time window of interest - is this snow cover present in WRF as well? If yes, the snow cover indeed has an influence on the SEB in the model.

Here in the text, we were referring to the initial land category data. Yes, WRF does simulate the snowfall and therefore updates the snow cover in the simulations which is then reflected in the surface albedo for each day/hour. We now made this clear in the text (lines 534-546). We did compare the timeseries of daily albedo from WRF (and ERA5) with the observed albedo at our sites and found large discrepancies (lines 300-306, see R1 C1; Figure 5). This figure has now been moved to the main text to better motivate our choice of using the observed albedo in the SEB model.

Also, in response to comment R2 C1 from referee #2, we have included a comparison between modeled (WRF and ERA5) and observed daily precipitation at our research sites (lines 355-362 and 374-379, Figures S2 and S3). Although precipitation (in the form of rainfall) contributes relatively little to the SEB at seasonal scales, events of fresh snowfall can significantly impact albedo and consequently the melt energy, depending on the frequency of such events during a melt season.

Lines 534-546: The commonly used land cover categories used for initializing WRF, based on the default MODIS data (Friedl and Sulla-Menashe, 2004) or ESA CCI data as used in this study (ESA, 2017, Table 2), do not distinguish between ice and snow categories. This distinction is crucial for the performance of albedo on glacier surfaces, and, consequently, the net shortwave radiation. While WRF does simulate snowfall and therefore updates the surface albedo at each time step, the timeseries of modeled daily albedo can substantially differ from the in-situ observations (Figure 5), justifying our approach to use the observed daily albedo in the SEB model. Nevertheless, in the absence of observations, there are multiple albedo models of varying complexity (e.g., Oerlemans and Knap, 1998; Brock et al., 2000; Hirose and Marshall, 2013; Marshall and Miller, 2020) that could be incorporated in the SEB modeling, but this application is beyond the scope of our study. A promising result for these albedo models is that ERA5 and WRF timeseries of daily precipitation, including snowfall, are relatively well correlated with the observed timeseries (Figure S3). This correlation analysis, however, may not be robust due to the likely poor quality of in-situ precipitation measurements as highlighted before. More research is thus needed to adequately assess the performance of ERA5 and WRF in precipitation modeling at our sites.

R1 C12: Figures 8 and 10: Please add a background grid to the figure, this improves their readability.

Done.

References

Goger, B., Stiperski, I., Nicholson, L., and Sauter, T. (2022): Large-eddy simulations of the atmospheric boundary layer over an Alpine glacier: Impact of synoptic flow direction and governing processes, Q. J. R. Meteorol. Soc, 148, 1319–1343, https://doi.org/10.1002/qj.4263

Sauter, T. and Galos, S. P. (2016): Effects of local advection on the spatial sensible heat flux variation on a mountain glacier, The Cryosphere, 10, 2887–2905, https://doi.org/10.5194/tc-10-2887-2016

Wagner, J. S., A. Gohm, and M. W. Rotach (2014): The impact of horizontal model grid resolution on the boundary layer structure over an idealized valley. Mon. Wea. Rev., 142, 3446–3465, https://doi.org/10.1175/MWR-D-14-00002.1