Dear RC2,

Thank you for taking time to thoroughly review our manuscript and for your constructive comments. Here is a preliminary reply to your major concerns. A more detailed version will follow with the revised manuscript.

Major Concerns

1. *I don't think it is unreasonable to use the single NARR grid cell that coincides with this glacier, given the 32-km scale of NARR grid cells and the good correlation with available ground T, RH, and shortwave radiation data. I do wonder why NARR incoming longwave radiation data was not used in this study though, for consistency. Was this considered, or compared with the longwave radiation parameterization that is used? The parameterization requires cloud cover fraction that is taken from NARR, so why not just use incoming LW from NARR, which appropriately considers the 3-hourly vertical temperature, humidity, and cloud structure, vs. a parameterization that only uses near-surface values?*

We also think this choosing the closest NARR gridpoint would be the most obvious choice, but Rev1 had some reserves and made useful suggestions, which will be included in the revised MS. Regarding the incoming longwave: thank you for spotting this…The Konzelmann parameterization described in section 3.3.3 of the methods is not what was used. This was an earlier application of the model, and as you correctly spotted this is a crucial aspect that we in fact had carefully addressed! Here is what was done:

We did not use the raw NARR incoming longwave (LWin) because these would carry an elevation bias, since the NARR gridcell is 237 m higher than the AWS. NARR LWin will therefore be too low. We did not have observations of LWin at the AWS to correct this bias directly. Instead, LWin was calculated from the bias-corrected NARR air temperature (Ta) at the AWS location and the atmospheric emissivity (E) from NARR: LWin = theta*E*Ta^4, where theta is the Stefan Boltzmann constant. The result is a LWin that is almost equal to the raw NARR Lwin, but with a slight correction for the difference in air temperature between the NARR gridcell and the AWS. The key point is to use the NARR emissivity, which reflects the vertical humidity and cloud structure. I am glad you spotted this, we will correct that section in the revised MS. Accordingly, panel f of figure 3 showing the bias-corrected NARR forcing will be changed to replace the cloud cover (CC) with the atmospheric emissivity

2. *The authors note that ERA5 was not available at the time of the study, but it has been out for more than one year now - about 1.5 years I believe, and offers 1/4 degree resolution with hourly data, which can avoid some of the complexities and assumptions in mapping the 3-hour NARR data onto hourly estimates. ERA-land is even higher resolution. I would not insist on this, as it is a lot of additional work, but I think it would be valuable and would strengthen the manuscript a lot to explore model results with ERA5 and/or ERA-land as well - this could very much help to test the robustness of the results and conclusions, and lower the reliance on NARR's relatively unproven veracity in the complex mountain terrain. This could be follow-up work, but I also have the nagging sense that this study risks being already out of date. The argument that ERA5*

precipitation is questionable is not really valid, as all of the reanalysis output is bias-adjusted and NARR is weak in this respect.

As you say, comparing with ERA-5 would be interesting but would entail too much work at this point, so it will have to be left for follow-up work. However, it is worth mentioning that ERA5 has a 0.25 degrees resolution (~ 28 km) vs. 0.30 degrees for NARR (~32 km). Most importantly, all atmospheric forcings for the ERA5-**land** product (0.1 degree) are interpolated from ERA5, so there is no real gain in spatial resolution for atmospheric variables, but only so for land surface variables. We think that the methodological framework used in this study could very well be re-applied with ERA5 in follow-up work, and this will be emphasized.

3. *I agree with R1 on the confusion regarding terminology in the methods. As I read it, the NARR output is not downscaled to the AWS; this study uses bias-adjusted rather than downscaled NARR output, as I understand it. Bias-adjusted NARR fields are then distributed over the glacier from a reference site (the AWS) using locally-relevant lapse rates and sophisticated methods for the incoming SW radiation. As I understand it, the reference site for the precipitation differs (e.g., Parker Ridge), but it is a similar approach. I am happy to leave it to the authors to decide how they would like to refer to this process (extrapolated, downscaled, lapsed, or distributed over the glacier), as long as it is clearly defined and consistent in the manuscript. It is not what I think of or what is commonly referred to as downscaling though.*

Yes, we agree we need to clarify the terminology. We think that bias-correction should be used to described the correction of NARR forcings, and "distribution to the glacier grid" for the subsequent step.

4. *I admire the use of the regional network of permanent weather stations to develop temperature and precipitation lapse rates, but I worry about the relevance of these values to the glacier itself. These are all off-glacier sites with a maximum elevation of 2025 m, while Saskatchewan Glacier extends from about 1800 to 3300 m. Glacier near-surface temperatures (and the surface energy balance that influences these) are very specific, as is the snow accumulation regime on glaciers and in the unsampled elevation band from 2025-3300 m. I don't have great confidence in the applicability of the lapse rates as determined by the off-glacier climate station network. For temperature, why not use the average daily or monthly lapse rates as determined by the HOBO temperature transect? I realize that these are summer-only, and the data are limited, but this is what is used for the diurnal lapse rates so this would seem relevant and consistent. Winter temperature lapse rates are not important to the glacier melt, so could be assigned an average or May value. For the precipitation lapse rates over the glacier, is there a way to use available winter mass balance data (in situ and/or LIDAR-inferred) to look at this? The current precipitation lapse rate may be appropriate, but it would be helpful to constrain and evaluate this, as well as the assumption of a sustained (and strong) linear increase in precipitation across the icefield plateau from 2800 to 3300 m.*

This is a common issue, to rely on lapse rates calibrated with lower-elevation stations. Your preoccupation is warranted, though. We thus checked the mean on-glacier lapse rates derived from the Hobos on the ablation stakes for their available period from May to August 2015, which yields a value of -0.46°C/100 m, versus the average from the permanent network for the

same months: -0.49 °C/100. This rather similar value gives us confidence that the lapse rates extrapolate well to the AWS site. When considering the mean lapse rate from the permanent network for May to October, i.e. the months during which the toe of the glacier (1784 m) is above the zero-degree isotherm (Supplementary Figure S1a), then the mean lapse rate is -0.47 °C/100. Hence replacing the mean monthly lapse rates by the mean lapse rates from the Hobos would not significantly change the results. These considerations will be included in the revised MS. We also want to recall that diurnal anomalies in the temperature lapse rate derived from the Hobos were superimposed on the seasonal lapse rate cycle derived from the permanent network, to more realistically simulate diurnal changes in the lapse rates not captured by the permanent network.

As for the precipitation lapse rate: this is harder to constrain, as in other studies, as there are no high elevation precipitation gauges. The Columbia/Parker Ridge stations (2000 m) are actually pretty high already compared to typical stations in the area. The sensitivity of the modelled mass balance to this lapse rate was tested in Fig. 6c: the sensitivity is expectedly high, which highlighted the need to constrain this value as best as possible from ancillary observations. We found that the lapse rate constrained on the permanent network gave good results. The performance, as you suggest, can be specifically assessed against the end of winter mass balance data (bw), which was done in section 4.4 and figure 5a: the lapse rate used yielded satisfactory bw simulations up to the highest stake/snowpit, which gives confidence on the lapse rate used. We did discuss the limitation of using a time-invariant lapse rate, which gave poorer results in the dry 2016 year (Fig6a), but we will also better discuss the unknown biases for the highest reaches of the glacier not covered by snowpits/stakes. This area only represents 8.8% of the total glacier area.

5. *The precipitation lapse rate that is used is based on the reference climate station data from November to March (l.206). This does not coincide with the accumulation season on the glacier, which is more like September to May. Is this same precipitation lapse rate used for April to October, and is there objective support for that? This needs to be discussed and addressed, perhaps with an examination of the primary data or perhaps by bringing in the winter mass balance data from the glacier, if there is some from the 2014-2016 study. November to March is relevant for the lower-elevation snow season, but not that of the glacier, where autumn and spring often bring a lot of snow.*

Please refer to section 7 and figure S1 that discusses this extensively based on observations. The current model code only allowed using a single value for the precipitation lapse rate, which is a limitation. We sought the most representative winter value. The precipitation gradient is best defined (strongest relationship between precipitation and elevation) during November-March which is also the period when most snow accumulation occurs, i.e. when the whole glacier is above the zero-degree isotherm (Fig S1). The lapse rate value for this period (mean = 15.6 % per 100 m) yielded satisfactory simulations of the end of winter mass balance (Figure 5 a), but we do show that there is a high sensitivity of the simulated mass balance to the uncertainty in this lapse rates (± 4%/100 m: see figure 6c and discussion). In fact, this point was a strong one made in the paper, that it is crucial to constrain this lapse rate to obtain good mass-balance simulations, but we realize that we did not express it clearly enough, so that we will emphasize this better in the revised MS. I also agree that snow accumulation occurs from September to May, for the higher parts of the glacier, and that the lapse rate starts decreasing in these shoulder seasons (April-May and September-Oct.: Fig. S1). Yet the bulk of snow accumulation is

rather from Nov. to March and the ± 4% sensitivity used in Figure 6 captures the uncertainty of adding Sept-Oct and May in the mean value). Most importantly, the value used yielded good end of winter mass balance, which is dominated by accumulation.

6. *Wind speed results on ll.393-395. These are extremely high average wind speeds, an annual average of 16 m/s and up to 23 m/s in February. I appreciate this is likely a windy site, and there are katabatic winds here, but those are typically stronger in summer. Are the authors confident that these units are correct - is this perhaps km/hr, or are these maximum (vs. mean) wind speeds that are reported here and plotted in Figure 2? An average monthly wind speed of 23 m/s equates to 83 km/hr, which is not plausible. Values reported and used later in the manuscript (e.g., from NARR, means closer to 5 m/s) are more reasonable. I would also add that I have spent some time on this glacier, and there is a steady and reliable down-glacier wind, but not of the knock-you-over variety.*

Thanks for spotting this. Indeed, quite high… I will say I was nervous for a moment… The wind speeds on Figure 2d are in km/h, not m/s as labelled, but they were correctly converted to m/s before bias-correcting the NARR wind speeds. This is why the resulting NARR forcings (Fig.3c) are within expected values, as you spotted correctly. We will modify figure 2d to plot wind speeds in m/s rather than km/h.

7. *I am not sure what 'homogenized' means here in the context of the observational precipitation records that are spliced. Homogenized has a very specific meaning for meteorological data sets, involving corrections for discontinuities associated with station moves or changing conditions/instruments/methods at an observation site. The precipitation data also seem to have a lot of gaps, which makes me worry about the time series of mean annual values. It seems best from about 1972 to 1994, not for the full period plotted in Figure 2. What methods were used to gap-fill this data for missing months? Apologies if I missed this. My sense is that it would be best to use these data for long-term mean monthly values from 1979-1994, using all available monthly data over this period. This can then inform a bias-adjustment of NARR mean monthly values for the same period, 1979-1994. Then go with bias-adjusted monthly NARR (or ERA5) precipitation for the study. Just my surficial thoughts on inspection of the observational data in Figure 2.*

We will rephrase to ‘merged’ instead of ‘homogenized’. Yes, there are gaps, as discussed in the MS. Recall that we used the daily observations from the ‘merged’ Parker Ridge/Columbia station to bias-correct the daily NARR precipitation, using all common days with available data. So we did not gap-filled the observations, we used whatever was available to bias-correct the NARR data on a daily time scale. The numerous gaps were the prime reason to use the (more uncertain) NARR precipitation as forcings. I hope this is clear, we will revise the text to make this clearer to the reader.

8. *Perhaps my most significant concern: the sensible heat fluxes seem far too high for a mid-latitude continental glacier, and compared with other data from the region (Peyto, Haig Glaciers). Also, it is surprising and unusual that latent heat fluxes are positive. I don't trust either of these results. Are the erroneously high winds speeds (point 6) the reason for this? This could explain the high values of sensible heat flux, though it is still*

odd that latent heat flux is positive. What is the basis for determining the snow/ice surface temperature in these calculations? This is critical to the turbulent heat flux calculations, and I did not see a discussion of this in the paper - apologies if I missed it. Is a melting glacier surface assumed in the summer? What is assumed through the rest of the year? I wonder too if the snow roughness value is appropriate for winter conditions - 6 mm is high, perhaps more reflective of sun cups than the smooth winter and spring snow surface. Snow roughness values closer to 1 mm are commonly adopted in glacier modelling. The sensitivity to this variable could be more thoroughly explored, perhaps considering order of magnitude rather than \pm 1 mm variations.

First, the wind speeds are correct, as mentioned in point 7. We were also a little surprised at first by the large sensible heat fluxes and small but positive latent heat fluxes. We will double check again that everything is correct, but we think they are correct and that the climate of the Columbia icefield (as you may have experienced) is actually quite different than Peyto and the other glaciers you refer to, which have more continental climates. The icefield is higher and has its owns weather, more humid than Peyto. This is why the Columbia icefield is often wrapped in clouds while the valley is clear (and explains why one can get stuck on the icefield waiting for a helicopter while it is sunny below the icefield…). But you are correct that this needs to be better discussed, and thank you for suggesting references. Some are already quoted, but we will discuss this point in more details in the revised MS.

You are right, we did not mention how the surface temperature (Ts) is calculated. Ts calculation follows Hock and Holmgren (2005). Ts is assumed to be zero if the computed energy balance is positive. If the energy balance is negative, Ts is lowered iteratively by steps of 0.25 K until the energy balance for the time-step and the gridcell considered becomes zero.

Regarding the roughness values we used the closest analogs on Peyto (Munto, 1989) to guide theses parameters. The snow value is indeed on the high side but still closer to 'glacier snow' during the snowmelt period. Since we used a time-invariant value a value close to metamorphized snow is better than mid winter when ablation is limited. We could expand the range for sensitivity analyses (Fig 6) to enclose a broader range of published values.

9. *AWS snow accumulation is reported on l.485. How is this recorded? Isn't this just a tipping bucket rain gauge at this site? Or is this based on measurements from site visits? It would be helpful, per the comment above, to report winter snow accumulations on the glacier and use this data to help evaluate the precipitation modelling.*

Using an ultrasonic gauge on the AWS. I see it was not mentioned, we will include this when describing the station. End of winter mass-balance (bw) at stakes and snowpits/soundings between stakes are presented in Figure 5a, we will try to better emphasize this.

10. *Agree with R1 that it will be really helpful to use mm or m w.e. throughout for mass balance, rather than a mix of mm, cm and m.*

I agree, we will homogenize the units.

11. *l.494, the balance gradients. Please give in units of m or mm w.e. per m. This is an interesting result, though I worry that the unusually high value (steep gradient) on the upper glacier is in part due to the unconstrained precipitation/accumulation gradient on the glacier. Looking at the available data in Figure 5 from 2015 and 2016 (2014 data are not sufficient), it would be hard to justify a bi-linear vs. linear relationship for ba. This is purely a model result then, as I understand it - can it be explained via mass balance processes here, since the authors describe the balance ratio as an unusual result? Is it reflected in the geodetic mass balance profiles? This is a significant point, as the balance gradients (values, linear vs. bi-linear) are potentially significant for regional-scale mass balance modelling - I can imagine other authors using the values that are reported in this study. Perhaps it is early to think about this too much, as the results may change upon revisiting of the wind speeds, modelled surface temperatures, and modelled turbulent fluxes.*

It is true that the modelled gradient is not constrained by point mass balance beyond 2900 m. Geodetic gradients include ice flux, which would complicate deriving mass balance gradient from them, no? The area of the glacier above 2900 m represents only 8.8 % of the total area, so extrapolation errors in this unsampled area would have a small impact on glacier-wide balances. Yet I agree that we should perhaps not use the modelled gradient, at least not beyond the last validation point at 2900 m. Perhaps calculating the gradient from the ELA up to 2900 m would be acceptable? This slightly changes the gradient, from 0.31 cm w.e./m (whole accumulation zone) to 0.32 for the period of observation, and from 0.29 to 0.31 for the long-term (1979-2016) period. The revised balance ratio would be 3.10 in place of 3.34, still high compared eastern Rockies 'drier' glaciers, and closer to 'wetter' West coast glaciers. This point, along with the larger simulated sensible heat flux and positive latent heat flux, suggest that the weather of the Columbia icefield is indeed distinct from the more continental types glaciers of the Rockies. We see no reasons at this stage to question the validity of the model results, but we will certainly carefully review our analyses as per your comments before re-asserting this conclusion in the revised MS.