Response to Referee #2:

We thank Dr. Ruzica Dadic for her detailed review that helped us to significantly improve the manuscript. Our responses to her comments are given in bold font, while any quotes from the revised manuscript are copied in italics. Before we list those changes, we summarize the major changes in the revised manuscript that incorporate comments from both referees:

1) Restructured (streamlined) the methods and results sections. In particular, all evaluated bulk methods in the study are clustered into two main types: C-methods and methods based on katabatic models. C-methods consist of four subgroups: C_log, C_Rib, C_M-O and C_SR, whereas the second cluster consists of the C_kat method and K_Int method. All the bulk methods depend on mean meteorological variables only (i.e. there is no application of OPEC data in the bulk methods). All other variants of the bulk methods (i.e. those with OPEC-derived variables) are now presented as part of the sensitivity analysis;
2) Removed the section on newly derived stability functions (following the comments from Referee #1);
3) Improved the discussion on possible spurious self-correlations and removing the results where the self-correlation was present;
4) Introduced a proxy variable for the background temperature lapse rate by using the near-surface air temperature observations from the two nearby meteorological stations, at different altitudes and in the glacier vicinity, in addition to the temperature measurements from the glacier station. The usage of these data led us to identify a dependency of parameters in the K_Int method on the proxy variable (i.e. difference in off-glacier and on-glacier near-surface air temperature). This empirical relationship allowed us to assess the parameters purely from the mean meteorological variables, instead of using the OPEC-derived stability parameter (z/L) as was the case in the initial manuscript;
5) Quantified the errors in the modelled sensible heat fluxes resulting from the radiative overheating of the temperature sensor;
6) All results (sensible and latent heat fluxes) are now evaluated for the cases with 30-min wind speed exceeding 1 m/s (instead of wind speed exceeding 3 m/s as was the case in the initial manuscript); and
7) Revised several figures and added a few that explicitly show dependencies among variables: z/L, Bulk Richardson number, wind speed and air-surface temperature difference.

Responses to Referee #2's comments:

● Generally the paper is well written, but I did find it cumbersome to read, because it includes many different parameterisations that are not easily distinguishable in the text. So I suggest that the authors consider a restructuring of the methods to clarify the difference between the model runs they performed and maybe ”cluster” the methods that are similar.

A great suggestion made by the reviewer. In the revised manuscript, the methods have now been clustered into two main types: C-methods and methods that rely on a katabatic model. We significantly revised the structure of the manuscript to make it more streamlined and less 'cumbersome' to read. We now feel that that paper is easier to read following this clustering and general streamlining.

● My main concern with the paper is that it neglects the very stable conditions by only looking at conditions where wind speed is >3m/s or (the moisture/temperature gradients are large enough). I appreciate that the measurement of turbulent fluxes under very stable conditions are harder to obtain because the mean flow is non-stationary and characterised by brief episodes of intermittent turbulence Mahrt [1989]; Beljaars and Holtslag [1991]; Mahrt [1998]; Cheng et al. [2005]. Considering the significant amount of the periods where low wind speeds occur (Figure 4 in the submitted manuscript), those periods should not be neglected when trying to improve the turbulent fluxes parameterisations over glaciers. A number of studies have been dedicated to finding valid flux-profile relationships for very stable conditions, such as are often found over snow and ice surfaces [e.g. Webb,1970; Kondo et al., 1978; Lettau, 1979; Brutsaert, 1982; Holtslag and deBruin, 1988; Beljaars and Holtslag, 1991; Cheng and Brutsaert,2005; Grachev et al., 2007] and those studies have also been applied to snow and ice surfaces [Pomeroy et al., 1998; Jordan et al., 1999;Sharan, 2009; Dadic et al., 2011].

We only use the near-neutral stability criterion in calculating the roughness lengths, while for the comparison of measured versus modeled turvulent fluxes we include all stability conditions that satisfy -2 < z/L < 2. In the revised manuscript we now also include the conditions for which wind speed exceeds 1 m/s (instead of the original threshold of 3 m/s). We note, however, that the inclusion of these data did not change the results of the bulk method evaluation. One needs to be careful when assuming that the very stable conditions are present during low wind speeds only → sloped glacier surfaces can have very stable conditions (z/L > 1) present during high wind speeds (e.g. Uz > 5 m/s), as we now show in Figure 7 of the revised manuscript (see below).
We thank the referee for the list of references. We already have quite extensive list of references but we have added now a selection from this recommended list.

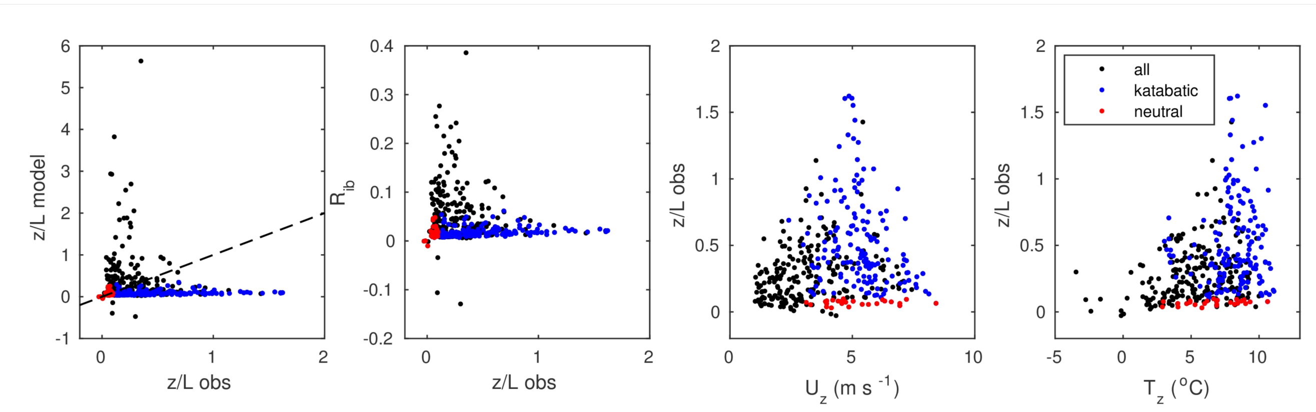

Figure 7: Modeled z/L with the fixed-point iterative scheme in the $\mathrm{C_{M-O}}$ method and the Bulk Richardson number (R_{ib}) against the OPEC-derived stability parameter (z/L obs). Dashed black line shows 1:1 line. Also shown is a dependency of the 30-min OPEC-derived stability parameter (z/L obs) on the wind speed (U_z) and the near-surface air temperature (T_z).

● All bulk methods assume a logarithmic profile, and they only differ in what stability correction they use. This should be clarified in the manuscript.

This is correct if the stability correction is negligible. However, introducing a stability function into the parametrizations for K (eddy viscosity) as K= ku*/phi(z/L) does change the logarithmic profile in the bulk method to log-linear profile (under stable conditions). This is why we differentiate between C_log (with logarithmic profile) and C_M-O (with log-linear profile). We prefer to keep this differentiation, and clarify this issue in the revised manuscript.

● Figures 1, 2, and some of Figure 4 (radiation, precipitation, wind direction) are not needed in this paper and can be removed.

We prefer to keep the figures showing the study area and the location of AWS, as well as the photos of the station setup on the glacier. We also prefer to keep radiation and wind direction since we used these variables to help us identify conditions with katabatic flow. Alternatively, we could place these figures in the Supplementary material and will consult ourselves with the editor about whether this is recommended.

● Figure 5: It is of no surprise that pretty much all 4 methods in this Figure have the same results, considering they all use he bulk method at almost neutral conditions. By neglecting the stable conditions, they don't have much reason not to vary. I am therefore not sure what the point of his comparison is.

As noted above, we use all the stability conditions (-2 < z/L < 2), not just the near-neutral ones (-0.1 < z/L < 0.1). It seems that this was not clearly communicated in the original paper, and therefore we clarified this issue better in the revised manuscript.
Page 9 in the revised manuscript (copied here in the Latex form):
To assure that the bulk method evaluation is performed on the high-quality measurements, all of the filters above are applied to the OPEC measured u_*, sensible heat (Q_{H}) and latent heat (Q_{E}), except the 'neutrality' and 'wind speed' filters. The latter two filters are modified so that all runs with $|\frac{z_{v,t}}{L}| < 2$ are included in the calculation of fluxes, as well as all runs with U_z > 1 m s^{1}. The threshold of $\frac{z_{v,t}}{L}=$ 2 is chosen because the universal stability functions for stable stratification are commonly defined up to $\frac{z}{L}=$ 2 \citep{Foken2008}, which represents strongly stratified stable regime.

● P18–19: It is not surprising that the "parameterisations" which use measured u_* as input lead to an increase in fit with the data. u_* goes into the Q_E equation by the power of 4, it's proportional to Q_H. It changes L with the power of 3, so will disproportionally decrease z/L. Some of this discussion (why u_* has more influence on the turbulent fluxes calculation than z/L) might be easier to understand by just looking at the equations and the relevance of the different parameters.

A good point. We now show the equations that relate the fluxes to u* in the bulk method (Equation 19 and 20 in the revised manuscript). Our goal here was to test how well this relation (between the fluxes and mean variables), expected by the theory, is supported by the data. We have now moved this analysis with measured u* and z/L into the sensitivity analysis section and have improved the discussion.
Page 9 in the revised manuscript (copied here in the Latex form):
In the gradient-flux relation, the eddy viscosity is parameterized as a function of z, u_* and M-O stability parameter ($\frac{z}{L}$). Because u_* and $\frac{z}{L}$, in the C-methods, are modeled rather than directly measured, any error in these modeled values can propagate into the flux estimates. Our goal in this section is to investigate the influence of the two variables, u_* and $\frac{z_{v,t}}{L}$, on the bulk method performance. To do so, we estimate the turbulent fluxes from each of the four bulk schemes using the OPEC-derived u_* and Obukhov length (L).

● Furthermore, I the observation on page 18 (L1–3) that the C_{log} and C_{SR} methods are not justified in table 3, where the difference between the u_* models in the correlation coefficient r is between 0.94 and 0.95 for Q_E and between 0.82 and 0.85 for Q_H, which is not exactly significant. I am not sure how to address this problem, but I'm sure the authors can come up with more robust conclusions than that.

Our discussion about the bulk method performance now reflects the results from all evaluation metrics (RMSE, MBE and correlation coefficient), not just the correlation coefficient. This section has also now been modified and moved to the sensitivity analysis section (Section 3.3.1) and the discussion has been improved, i.e. we first intercompare only the bulk methods with mean meterological variables, and later introduce the sensitivity tests when OPEC-derived u_* and z/L are used.

● p 29, L1-2: Considering that the authors have most SEB components to actually calculate the surface temperature, and that the surface temperature is an important feedback for the TF, the authors should consider calculating the surface temperature and including it in their calculations using the different parametrizations. It would be interesting what effect the different parameterizations have on surface temperature. I do not expect the authors to change all their results now, but maybe it's worth a discussion in the paper.

Using the SEB closure to derive surface temperature turned out to be unreliable, mainly because the measured radiative fluxes have large errors, in particular, the use of NR-Lite net radiometer sensor to estimate the outgoing longwave radiation from the measured net radiation and measured shortwave incoming and reflected radiation. We now explain this more clearly in the text. Also, we provide an error analysis in the calculated fluxes from the bulk methods assuming random errors in surface temperature.
Page 7 in the revised manuscript (copied here in the Latex form):
In the absence of direct measurements, the surface temperature (T_{0}) was assumed to be at melting point (0°C) and the surface vapor pressure at saturation (6.13 hPa). The assumption of consistent melting is corroborated with the sonic ranger measurements showing persistent surface lowering throughout the observational period. To assure that the assumption holds we use only the data for which $T_{z_t} > 1^{\circ}$. In general, assuming that $T_{0}=0^{\circ}$C works well on temperate glaciers during a melting season, and is more accurate than estimating the surface temperature from the longwave radiation measurements \citep{Fairall_etal1998} or from a SEB closure \citep{Hock2005}. Nevertheless, when the surface is not consistently melting, SEB closure can give much better results than the assumption of the melting surface (e.g. Conway and Cullen, 2013). Estimating T_{0} from our radiation data, proved to be unreliable because of the poor accuracy of NR-Lite net radiometer. As part of our uncertainty analysis, we will quantify errors in our results due to the assumed rather than measured surface conditions.

● p 30, L26-30: Considering that only near-neutral conditions are used for this study, I am not surprised that the stability corrections show very little difference when modelling the fluxes.

As already mentioned above, we use the stability conditions -2 < z/L < 2, not just the near-neutral conditions (-0.1 < z/L < 0.1). The main reason why the stability corrections did not significantly alter the fluxes is because the modelled z/L, calculated via the fixed-point iterative scheme, underestimates the OPEC-derived z/L.
Page 20 in the revised manuscript (copied here in the Latex form):
Intercomparion only across the C-methods (Fig.\ \ref{fig: scatter plots with basic bulk methods}) reveals that the performance of $\mathrm{C_{M-O}}$ method does not significantly differ from $\mathrm{C_{log}}$ method. This is because the M-O stability parameter ($\frac{z_{v,t}}{L}$), calculated with the fixed-point iterative scheme of \cite{Munro1989}, is uncorrelated with the OPEC-derived $\frac{z_{v,t}}{L}$, underestimates the stability during katabatic conditions and overestimates it during conditions with low speeds (Fig.\ \ref{fig: modeled versus observed stability}). The stability corrections that depend on calculated $\frac{z_{v,t}}{L}$, therefore have a small effect in modifying

the fluxes during the katabatic conditions, while during the non-katabatic conditions with the low wind speeds the fluxes are unnecessarily suppressed. Furthermore, we found no correlation between the 30-min OPEC-derived $\frac{z_{v,t}}{L}$ and any of the mean meteorological variables (e.g. temperature, wind speed; Fig.\ \ref{fig: modeled versus observed stability}), which explains the failure of the fixed-point iterative scheme that relies on these dependencies. The poor performance of the stability corrections in $\mathrm{C_{Rib}}$ method also follows from the lack of correlation between R_{ib} and the OPEC-derived $\frac{z_{v,t}}{L}$ (Fig.\ \ref{fig: modeled versus observed stability}).

● P31, L11-13: As far as I remember, the reason why the turbulent fluxes are suppressed in Conway and Cullen (2013) is that they assumed the log-linear relationship to be valid under very stable conditions. The log-linear relations, however, do not allow for significant fluxes to occur at very strong stability[Monin and Yaglom, 1971; Mahrt, 1998; Pleim, 2006] and underestimate the turbulent fluxes over these conditions [e.g. Deardorff, 1968; Webb, 1970; Kondo et al., 1978; Louis, 1979; Hogstrom, 1988; Launiainen, 1995; Mahrt, 1998; Jordan et al., 1999; Stossel et al., 2010].

Yes, this log-linear relationship explains part of the story in the findings from Conway and Cullen (2013). The other part is related to the presence of low wind maximum height. We have now incorporated this explanation more clearly in our discussion. We thank the referee for the provided references, a selection of which we included in the revised manuscript.

Page 22/23 in the revised manuscript (copied here in the Latex form):

While the original C-methods, however, overestimate Q_H during the katabatic conditions, the C-methods with stability corrections and measured u_* underestimate the fluxes (Table \ref{tab:evaluation results} and Fig.\ \ref{fig: scatter plot with sensitivity tests}). The overestimation of Q_H during katabatic flows has also been shown in \cite{Denby_Greuell2000} and explained by a failure of M-O theory in the presence of shallow katabatic wind speed maximum. At the wind speed maximum, measured u_* approaches zero, while the C-method assumes constant momentum flux in the surface layer and therefore overestimates u_*. The overestimation is less pronounced for Q_H than for u_* because the reduced turbulence at the wind speed maximum leads to an increase in the air-surface temperature difference, and subsequently an increase in the measured Q_H. However, when measured u_* is used in the $\mathrm{C_{M-O}}$ or $\mathrm{C_{Rib}}$ method, assuming that the eddy diffusivity is as effective as eddy viscosity (Pr=1), the C-method underestimates Q_H since the air-surface temperature difference alone can not compensate for the effect of reduced momentum flux. To correct for this bias in Q_H, Pr would need to decrease, i.e. the C-method would need to account for more effective eddy diffusivity than eddy viscosity at the given height. In the absence of wind profile measurements, we can only assume that these effects take place at our site, but we have no observational evidence for the presence of the wind speed maximum.