Response to RC2

We would like to thank the reviewer for their valuable comments, their thorough reading of the paper (and supplement), and their recommendation to accept it for publication. The reviewer raised important points and made smart observations that have substantially improved the paper. When we do not implement one of their recommendations, we provide a clear explanation of our reasoning.

More changes than listed have been implemented to accommodate the suggestions by the other three reviewers. Also, figure numbers are shifted by one due to the incorporation of a new figure depicting a synoptic overview (now Fig. 1).

Response to general comments:

- The science questions are good, but radiative effects that were a big part of the motivation for studying WCB microphysical properties aren't examined. Is there a reason for that? Connecting microphysical changes to radiative changes would be very insightful.

UTLS moisture is important for Earth's climate and we find large differences in the ice content and relative humidity in our study, which indicates that these results could also have an impact on the radiative budget of WCBs. Sadly, the original study by Oertel et al. (2025) did not focus on radiative effects and therefore did not write relevant variables to file. Nor are the simulations long enough to do a robust analysis of radiative impact. Therefore, we are not able to investigate what the impact of our findings are for the radiative effect of WCBs. However, it is important to point out that this could be a topic for future research, given the large differences we found in UTLS moisture conditions between PPE members. We therefore add an item to the final outlook bullet-point list following the point discussing the implications for geoengineering:
"Following this, future studies could investigate how the different UTLS moisture conditions produced by the PPE members influence the radiative effect of WCBs." (line 771)

- A few questions on sensitivities to the methodology:
There is a bit of justification provided for the range of some parameter values, but it isn't clear how all of the CAP, INP, CCN, SAT, and SST ranges are chosen for the PPE. Presumably these are constrained by possible real world values dictated by previous studies?

This is an important point and we add some additional information in the paper, based on the PPE design by Oertel et al. (2025), to make this more clear (Line 236).

- For the CCN scaling factor, values were chosen such that they approximately represent observed CCN variability ranges in the North Atlantic regions. We have added this information to the end of the paragraph on the CCN scaling factor:

  "*This variation results in CCN concentrations varying from approximately 100 cm−3 to 5,000 cm−3, which approximately represents the variability that has been observed for CCN concentrations (Hande et al., 2015, 2016; Genz et al., 2020; Wang et al., 2021).*"

- For the INP scaling factor, we now add the information that "*their concentrations in the atmosphere vary over several orders of magnitude*" (Line 205). Later in the paragraph, it is already noted by how much the concentrations approximately vary given the perturbations.
- For CAP, we already provide this information and cite relevant literature (Line 197):

  "*While theory predicts a normalized CAP value of 0.5 for perfectly spherical particles, realistic ice and snow hydrometeors often exhibit considerably different values (Westbrook et al., 2008; Chiruta and Wang, 2003). Consequently, in the PPE, CAP for ice and snow is varied (simultaneously) by a scaling factoring ranging from 0.2 to 1, which explores a range for CAP from 0.1 to 0.5.*"
- For SAT we add the following information (Line 249):

  "*For a factor f-SAT AD of 0.1, vertical velocities of 2 ms^{-1} can produce supersaturations of 20%, which aligns with theoretical examinations of supersaturation in liquid clouds (Korolev and Mazin, 2003; Morrison and Grabowski, 2008)*"
- For SST we believe that the paragraph is already sufficient to explain the +-2K range for variation.

With no test dataset for the RF model, couldn't that result in overfitting to this single case?

That is correct. However, this is not an issue in this case because of how we use and interpret the RF models. We use the RF models purely as a diagnostic tool—to rank parameter sensitivities and explore partial-dependence curves—not to make out-of-sample forecasts. If a model has a good fit — that is, if the root mean square error (NRMSE) is low — then it has determined which parameters and thresholds are important for "predicting" a target variable. We can then use these features to gain insight into what the model learned.
Even so, we guard against over-fitting by making sure the RF models did not have the capacity to memorize the data (number of trees, depth, etc.). Thus, the NRMSE is never small enough to indicate overfitting.

WCBs need some definition so thresholds are to be expected, but are results sensitive to the 600-hPa ascent depth and ascent rate threshold used?

The criterion that an air parcel must ascend at least 600 hPa in no more than 48 hours to be considered part of a WBC is a widely accepted threshold adopted in nearly every WBC study. A recent study by Heitmann et al. (2024) determined that relaxing this criterion results in only minor changes for various analyses (https://doi.org/10.5194/wcd-5-537-2024). Therefore, we are confident that our results are not substantially affected by this threshold.

For sensitivity to CCN, how much should we trust the snow and graupel changes in terms of applicability to the real world given arbitrary threshold conversions between these 2 categories as opposed to riming transitioning smoothly to produce a range of variably rimed precipitating ice?

This is a good point that was also pointed out be RC1 in a similar way. However, since this is a modeling study, we cannot make any definite statement on the effects in the real world without additional observational studies. Nevertheless, we have modified this paragraph to reflect that this process is continuous in reality (Lines 465-470) :
"*The mean of the maximum mixing ratio for snow (max(qs)) during ascent also increases with CCN, while for graupel (max(qg)) it decreases (Fig. 6 e). We explain this as follows: increased CCN concentrations lead to more numerous but smaller cloud droplets, which slows the formation of raindrops. This, in turn, reduces the collision efficiency between ice or snow particles and cloud droplets, slowing the (continuous) conversion to graupel. As a result, snow remains more abundant and graupel less abundant. However, this large effect of the CCN scaling factor*

on the hydrometeor population during the ascent is mostly lost by the end of the ascent, showing that once air parcels glaciate, CAP and INP dominate the ice-phase cloud microphysics"

Figure 6 also shows nonlinear relationships where CAP, INP, and CCN sensitivities are particularly low or high, so it should perhaps be noted that the distribution of values of these parameters in the real world is important for dictating whether overall sensitivities are large or small.

This is a great point which we now address in Line 494:
"*Notably, these relationships seem non-linear when CAP, INP scaling and CCN scaling are particularly low or high, indicating that the distribution of these parameter values in the real-world are important for determining whether the overall sensitivity of RHi to perturbations is large or small.*"

With the greater sensitivities in fast ascending trajectories, it seems like model representation of convective processes could be important and there could be some model resolution dependence there. Should that be mentioned?

For fast-ascending trajectories, only the sensitivities for the ice mixing ratio Ni and the ice radius ri are larger. The sensitivity for relative humidity over ice is smaller. We now point this out in Line 616:
"We also note that for fast ascending trajectories, Ni and ri are more sensitive to CAP and the INP scaling factor than for all trajectories, suggesting that overall sensitivities to these parameters might increase for higher resolution simulations, where a greater fraction of trajectories ascend quickly."
And in Line 720:
"*Ni and ri are more sensitive to CAP and the INP scaling factor for fast ascending trajectories than for all trajectories. This indicates that simulation scale might also influence Ni and ri, since Choudhary and Voigt (2022) found that a greater fraction of WCB trajectories ascend quickly when model resolution is increased.*"

The conclusions and discussion should include caveats. For example:

- This is a single case, and it isn't clear how representative it is of WCB events in general.
- PPEs sample the uncertain multi-parameter phase space but still have the weakness of assuming constant parameter values for some parameters that are not real world physical constants. Thus, sensitivities can be overestimated relative to a potentially more realistic stochastic framework in which constant parameters may be varying (e.g., Stanford et al. 2019).
- With only Hallett-Mossop rime splintering parameterized for secondary ice production, could mixed phase ice concentrations be biased low, potentially influencing the WCB outflow sensitivities? Recent studies by Alexei Korolev, Vaughan Phillips, and others have highlighted the potential importance of additional secondary ice mechanisms such as raindrop fragmentation upon freezing and ice collisional breakup.

We have included these (and other) caveats in Line 750: "*However, this study compromises of only one case study, and it isn't clear how representative our findings are of WCBs (or other model configurations and microphysics schemes) in general. Additionally, PPEs sample the uncertain multi-parameter phase space but assume constant parameter values throughout the simulation for some parameters that are not constant in reality. Therefore, sensitivities can be overestimated relative to a potentially more realistic stochastic framework in which parameters vary during the simulation (e.g Stanford et al. (2019)). Furthermore, the microphysics scheme employed in the ICON simulation only considers rime splintering; it does not account for raindrop fragmentation upon freezing or ice collisional breakup for secondary ice production. This could result in biases in mixed-phase ice concentrations. Finally, there are likely additional important parameters that influence the variables examined in this study, but that were not perturbed (see Hieronymus et al. (2025)).*"

- Could results be sensitive to the thresholds used to define the WCB (600-hPa depth) and their ascent rates?

Given our answer to this question above, we have not included this caveat. The ascent rates are discussed in the text.

Sentences beginning with "This" could be made clearer by stating the object that it is referring to (for example, "this difference…" instead of just "this…"). These are several instances:

Thank you for these suggestions which make the text more readable. We have implemented these changes and looked for other instances where we write "[…]. This is / makes / means / etc. […]" and have rewritten for clarity whenever it enhances readability.

Lines 67-68: How far south are the authors referring to? Could the authors be more specific?

We have rephrased to (Line 80): "[…] *when WCBs are usually located closer to the equator than during later stages, […]*". Also, we have added a synoptic overview of the case in the new Figure 1.

Lines 267-271: Are the authors saying that a high IBF score could be due to a parameter being highly correlated with another parameter instead of the parameter being "actually important"?

No, we mean that a high IBF score without a clear correlation with the target variable means that the variable might be important for the decision making structure of the RF model in *combination* with another parameter (PPE parameters do not correlate with each other, they are sampled according to the latin hypercube method). We try to clarify what we mean by rephrasing this sentence as follows (Line 295):
"*Note: a high IBF importance score for a PPE-parameter does not necessarily imply a clear or strong correlation with the output variable. Instead, it indicates that the PPE-parameter contributes strongly to the RF-model's predictions, possibly by affecting the RF-model's decision-making structure only in combination with other PPE-parameters, and is only a meaningful metric when the RF-prediction is good.* "

Lines 304-305: Since the authors are discussing temperatures in terms of Celsius, could the corresponding plots (e.g., Fig. 2) be modified to be in terms of Celsius? Using Celsius would make more intuitive sense in the framework of microphysics discussions including homogeneous freezing.

We modified all temperatures in the figures and in the text to degrees Celsius.

Line 325: For the first part of this sentence before the comma referring to all PPE members having a mean RHi > 100%, could the authors refer to Figure 2j?

Yes, we have added the reference to the figure.

Lines 325 to 327: What about the observation made is "particularly interesting"?

We find it interesting that changes to RHi are large eventhough changes in qv are small. RHi is largely dependent on vapor content, so if RHi changes by a lot, it is interesting that qv does not. We state this more clearly now (Line: 355)
"[…] which is particularly interesting ***because*** *the differences in qv are small."*

Lines 341 to 345: It seems like the argument here is that using the means for the RF model means that the spread of the distribution (5^{th} to 95^{th} percentiles) is not considered by the RF model. If so, the argument as stated appears a little convoluted and difficult to follow. Is what matters here the change in means between PPE members relative to the spread between the 5^{th} to 95^{th} percentiles (because all variables could be argued to have a large 5^{th} to 95^{th} percentile spread)?

In this paragraph, we aimed to point out that any changes in the mean or median of p or T should not be taken too seriously given the large spread of p and T values, which is much larger than any change in the mean or median of p or T. Furthermore, we state that the RF model does not identify meaningful relationships between PPE-parameter perturbations and these values anyway. We realize that this paragraph is somewhat redundant and have modified it to clarify this point (Line 362):

"As discussed above, T , p and qv show little variation between PPE members (Figure 3 b, d and f), with the variability within each PPE member being significantly larger than the mean differences between PPE members. The distributions for PPE members with the highest and smallest mean values for T , p and qv are also very similar (Figs. 3 a, c and e). The relatively large spread of the distribution (min/max shaded areas in Fig.3 a, c, and e) is a result of the differing number of WCB trajectories per PPE member (which is primarily controlled by SST, see Oertel et al. (2025)). The RF2
mean(p) and RF2 mean(T) model predictions (which we use to determine which parameters have the strongest influence on a variable, as long as the model prediction is strong) are also relatively weak, with mediocre R2 and large NRMSE-values (Tab. 2). Therefore, the RF model does not find meaningful changes in T_mean, p_mean and qv_mean with parameter perturbations. We interpret this inability of the parameters to change the pressure, temperature, and vapor conditions at the end of the WCB ascent as an indication that the thermodynamic conditions in the outflow of a WCB are largely constrained."

Lines 347: It is unclear how Figure 2c shows a correlation between T95 and qv95. Perhaps the authors meant to refer to Fig. A1c?

Yes, thank you.

Line 354: This statement ("The change is stronger for qv_95 than for T_95") presumably refers to Figs. 2f and 2b. It is unclear how this change is computed and how the 2 different variable changes can be fairly compared against each other. Perhaps the max 95^{th} value minus the min 95^{th} value divided by the 5^{th} to 95^{th} percentile spread to compare changes relative to the range of variable values?

We have removed this sentence because upon examining the figures and data again, we find that qv_95 does in fact NOT change more strongly with SST than T_95.

Line 359: It is clear visually that the highest and lowest qv value correlates strongly with the calculated saturation specific humidity. However, could the authors include a correlation coefficient to quantitatively support this claim?

We have added information on the spearman correlation coefficient of qv vs qv_sat for both groups (0.99) in the text.

Lines 357 to 361: Isn't Fig. A1a or something similar to it plotting qv as a function of temperature a more straightforward argument than Fig. A1b (qv vs. qv_sat) that qv is strongly constrained by temperature? Qv correlates strongly with qv_sat, but they are not 1:1, and it isn't clear from Fig. A1b alone how temperature vs. pressure modulate qv_sat to affect that relationship or how dynamics and microphysics affecting supersaturation.

We want to point out that qv at the end of the ascent is primarily constrained by the thermodynamic conditions at the end of the ascent for all simulations, and this includes the pressure as well as the temperature. Plotting qv vs temperature would only show how strongly qv correlates with temperature; using qv_sat instead also enocdes the dependence on pressure. Hence, we plot qv vs qv_sat for two simulations that have the most different 95th percentiles for qv, to show that for both simulations, qv at the end of the ascent correlates with qv_sat (and is therefore primarily constrained by the thermodynamic conditions).

Lines 376: "The spread… is unchanged…" The word unchanged seems a little too strong. Could the authors moderate it to "**mostly** unchanged"?

Yes.

Lines 377 to 378: "… reduces the spread". This reduction is not easily visible. Could the authors quantify this reduction?

We have added information on the spread of values to back up this claim (Line 404).

Line 380: The second mode in the distribution is not a "peak" since it is not a local maximum. It would be more accurate to describe this as a "second mode."

This is a more accurate description and we have made sure to call this second peak a "mode" throughout the text

Line 390: "…many small cloud droplets reach the homogeneous freezing level." How is "many" defined here? It is highly likely this is homogeneous freezing and glaciation temperatures in Fig. S11 provide some support, but could it be shown that this second mode is indeed due to homogeneous freezing, e.g., by examining drop concentrations at -38C or the change in ice concentration across that temperature level?

We meant to say the following, which we now write instead (Line 419):
"More homogeneously freezing cloud droplets when CAP is low explains why the distribution appears slightly bi-modal at low CAP values — indicating ***that a large proportion of trajectories contain small cloud droplets that reach the homogeneous-freezing level*** *— but becomes uni-modal at high CAP values."*

Concerning the number of cloud droplets reaching the homogeneous freezing level: due to vertical interpolation errors and time resolution of the trajectory data, it is not possible to examine drop concentrations just below -38°C accurately. Nevertheless, we are confident in our interpretation of the data, because the only physical process that can produce ice-crystal number mixing ratios in the order of $10^7 - 10^8$ kg^{-1} is homogeneous cloud freezing. Heterogeneously produced ice-crystal number mixing ratios cannot exceed the number of INPs, which is on the order of 10^4 kg^{-1}.

Figure S12 caption: "second peak" should be "second higher concentration mode."

We have implemented this change

Lines 423 to 424: "This clear dependency… parcel evolution (Fig. 5e)": In Figure 5e, the max_qc panel is dark red, and the max_qr panel is dark blue. Shouldn't these large magnitudes mean that CCN strongly modifies the liquid content rather than not strongly affecting it? Also, is “in early stages of parcel evolution” inferred from the liquid mass mixing ratios being maximum values?

Thank you very much for pointing this out; we are not sure why this false statement was made. CCN definitely effects the liquid water content during the ascent and we have modified the sentence to reflect this.
Regarding the second question: yes, the early stages are inferred from max_qc and max_qr being achieved during the ascent.

Line 424: “Therefore, cloud droplets are far smaller…” It is unclear how cloud droplets are proven to be smaller when qc strongly increases with the CCN scaling factor in Figure 5. Is there other evidence to support this assertion?

We have done some additional analysis reformulated these sentences to now write (Line 457):
“With an increase in CCN, cloud droplets during the ascent become smaller; for simulations with the largest and smallest CCN scaling factor, the mean cloud droplet mass is approximately $72 \cdot 10^{-9}$g and $9 \cdot 10^{-9}$g, respectively. Smaller
cloud droplets delay the formation of rain and enhance the residence time of liquid condensate. This delay […]“

Line 428 to 431: the word “presumably” is used twice in this sentence. Suggest replacing one of them with a synonym for improved readability.

We have changed this entire paragraph due to a more physically sound explanation suggested by RC1. We have summarized it about, in the question about the continuity in conversions from ice to graupel.

Lines 458 to 459: How robust are these red and cyan best fit lines? Could the authors include correlation values?

We have added spearman correlation coefficients for the red and cyan lines in the text.

Line 462: “… the spread of RHi values reduces strongly…” Is there a way to quantify this spread and its reduction?

We have added information in the text on the standard deviations of RHi values for simulations that have the maximum and minimum values for CAP, INP and CCN scaling.

Line 473: “The mean and median for the first group…” Could the authors clarify what “first group” and “second group” are referring to?

We have added clarification by adding “(CAP and INP scaling large)” and “(CAP and INP scaling small)” behind the respective groups

Line 501: Referencing specific visual cues would greatly help the readability of the sentence: “**as indicated by the red line**, CAP reduces…”
Line 509: “… below this threshold for the second group…” Could the authors clarify which line in the figure they are referring to?
Line 510: “…higher abundance of trajectories above the threshold also explains the long tail”. Could the authors clarify that they are referring to the lines in Fig. 8A?

We have modified this paragraph (Lines 445 to 455) to make all references to lines in plots, threshold values and trajectory "groups" more clear.

Line 553: The authors mention fast vs all trajectories, but the figures the authors are referring to within this sentence only contain fast trajectories. Could the authors reference back to the appropriate figure for comparison between the two trajectory groups? For example: "…for fast ascending trajectories than for all trajectories (compare xxx in Figures X and X)".

We reference appropriate figures and make more clear when we refer to differences in fast trajectories and all trajectories.

Lines 557 to 560: "This is because the peak t for fast trajectories **in Fig. 12b** is below 5*10^2…"

We have implemented this

Line 586: "The large Ni implies smaller Tsat…" Recommend referencing equation 2 here. Also, since decreasing ri has the opposite impact on Tsat, the authors should mention why the effect of Ni on Tsat is greater than ri.

We have implemented this

Line 605: Can the authors describe which behavior they are referring to?

We have modified the sentence (Line 658): "*which is probably a reflection of the* ***larger Ni*** *found at the end of the ascent*"

Line 606: Could Fig. 4e be referenced right after "bimodal distribution" for improved clarity?

We added the explanation that this is because of the increased number of high Ni values when CAP is small (while implementing this change we realized that we had falsely written "when CAP is large" and have changed this to "when CAP is small").

Line 643: "… a small shift in the lower Ni, which…". It appears that there is a word missing after "Ni".

We have modified the sentence to make it more understandable.

Line 697: Did the authors mean "Ni 5 hours after ascent"?

Yes, thank you

Throughout: When using terms like mean maximum variable, it would help to clarify what the mean and maximum correspond to, so it is clear how the variable is being computed.

We have changed these phrases to be phrased like: "*mean* ***of the*** *maximum mixing ratio […]*"

(Optional) The authors should consider re-spelling out the acronyms at key locations (e.g., figures, conclusion) to aid readers.

Figure 2: Make sure that to clearly state that largest and smallest "means" refer to the distributions of the variable for ensemble members with the smallest and largest mean value of that variable. The caption is a bit confusing as currently worded.

Figure 3: Could the authors mention in the caption that these distributions correspond to those at the "end of ascent"?

Figure 5e: (optional) Could the correlations be shown in a table format to add additional information such as the correlation coefficient value? Currently, max_qc and max_qs are very similar in shading.

Thank you for these suggestions. We have implemented most of them. The last comment we chose not to implement because the exact correlation coefficient is not important, only that the correlation is strong.