Response to reviewers for "Grid-Spacing Sensitivity of Rossby Wave Breaking to Mesoscale Diabatic Processes" by Rixen et al.

We thank the reviewers for their thorough reading and attempt to clarify some of the points that they found unclear or unconvincing, as discussed in the itemized response to the reviews below. We carefully revised the manuscript according to your comments and replied to them in blue font in the pages below, and the corresponding updates in the manuscript are highlighted in italic font.

Review 1

1 - Comment: The scientific focus in the list of main questions on page 4 could be improved. For example, the 2nd and 3rd questions are focused on which diabatic processes are better simulated and are these diabatic processes the main cause. However, the research, as shown in the title, is focused on Rossby wave breaking and the questions listed generally do not touch on the impacts of the changes in grid spacing on Rossby wave structure.

In lines 119-123, we re-framed the focus of the study by adapting the main research questions. We emphasized both the focus on the upstream moist mesoscale processes, as well as their impact on the upper-level waveguide and the subsequent downstream Rossby wave breaking. We believe these revised questions now frame the scope of the paper more appropriately and specifically also address the link between grid spacing and Rossby wave structure (lines 119-123):

"The main questions that we address in this study are:

1. Can the explicit simulation of upstream mesoscale processes at fine grid spacings in ICON mitigate large-scale forecast errors in IFS?

2. How does horizontal grid spacing modulate the intensity, spatial scale, and vertical structure of moist mesoscale processes?

3. To what extent do these processes modify the upper-level waveguide and how does this affect Rossby wave breaking?"

2 - Comment: A better understanding of why changes in the horizontal resolution matters:The manuscript clearly shows the impacts of grid-spacing and that higher resolution provides significant improvements. The findings state that the representation of mesoscale diabatic processes improve with higher resolution, which impacts the Rossby wave amplification. However, this finding is not sufficiently substantiated and a deeper exploration of why the resolution changes the accuracy of these simulations is needed. For example, the results should include showing how variations in the horizontal resolution impact Rossby waves through changing the height, continuity, location relative to the jet axis, and magnitude of the upper-level divergence from diabatic processes. While some insight is provided into these issues is included in the text, a deeper dive into how these characteristics (e.g., height, continuity, location, and magnitude of the upper-level divergence) vary with resolution is needed. An important issue is not only how does the

continuity and magnitude of the divergence change resolution but does the height, horizontal extent, and temporal continuity of the divergence also vary as one moves from deep convective parameterization, to shallow convective parameterization only to resolved convection. These issues are touched on in your manuscript and in the Wave Amplitude Error in figure A1, but the authors should include a brief and deeper dive (e.g., height, location relative to the jet stream, continuity, etc). Hence, these changes would lead the reader to be informed on poor forecasts exist.

We thank the reviewer for suggesting these analyses to improve the argument of why resolution matters and improves the representation of these processes. When looking at upper-level divergence, for example, the results mirror closely the ones from negative PV (from Figures 3b, 11a and A3). We show an example of such a plot in the following figure. We computed the upper-level divergence (at 250 hPa), and show the 95th percentile of divergence. As deep convection occurs, we can expect some strong upper level divergent flow. Though negative PV and upper level divergent outflow are two completely different metrics, they are both signatures of deep convection, and as we expected, this shows that negative PV is a good proxy for MCS strength. We decided to include a discussion of these results in the manuscript but not include the figure, as it would only make this manuscript longer (lines 260-262).

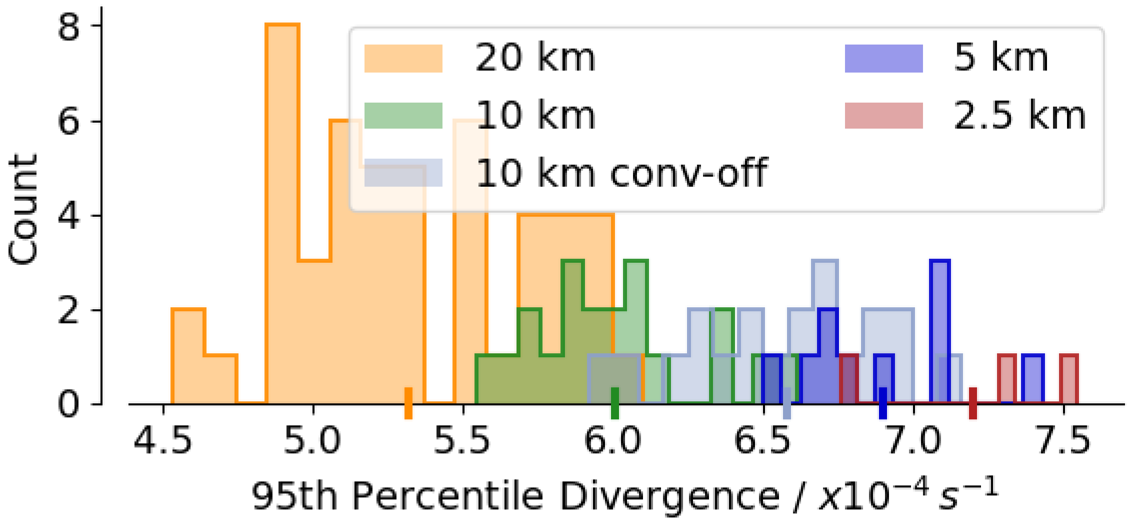

Figure 1: 95th percentile of upper-level (250 hPa) divergent outflow computed within the domain (105° to 70° W, 25° to 50° N).

Concerning your comment on the analysis of the location of these processes with respect to the jet axis: The second case study offers a great example to consider the distance of these processes with respect to the jet stream. We therefore performed such an analysis and added it in Figure A5 of the revised manuscript (see Figure 2 of this document). We discuss this in lines 368-371. Thank you very much for suggesting such an excellent idea.

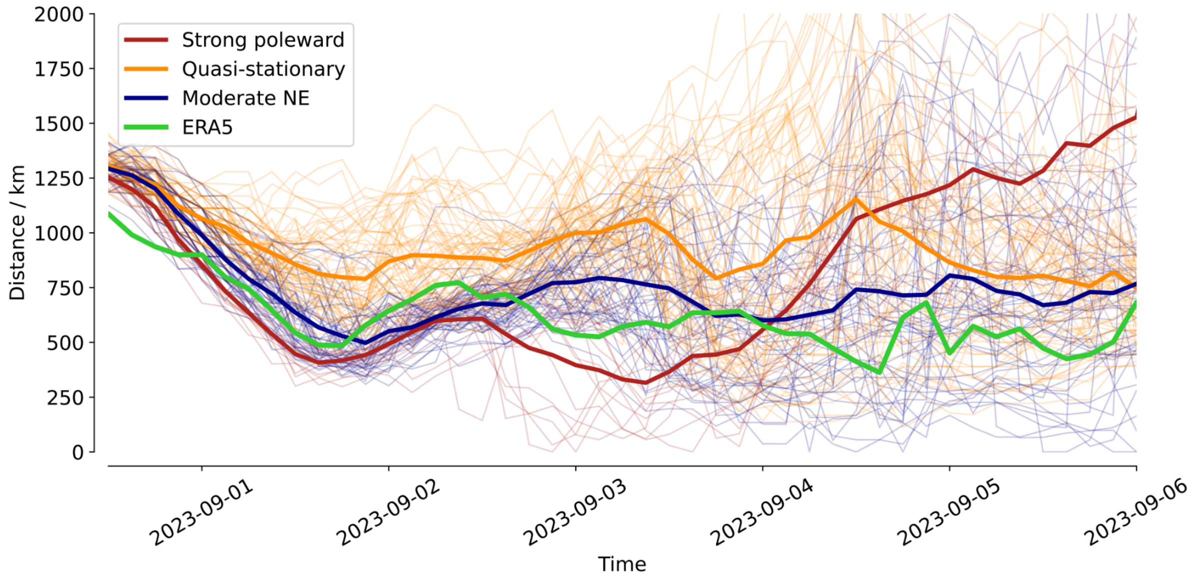

Figure 2: Distance between the extra-tropical cyclone and the jet stream for all ensemble members at all resolutions, colored depending on the cluster the belong to. The thick lines show clusters means, and the green line the corresponding ERA5 reference.

3 - Comment: How do the Rossby wave characteristics in these simulations vary with resolution? *Another issue to be considered is does the wave number for the Rossby waves change at higher resolution. The question arises since past research has shown that deep convection near the jet stream can generate high wave number Rossby waves with rapid propagation that can lead to Rossby wave breaking. These high wave number waves would likely require high resolution modeling.*

We thank the reviewer for pointing out the implications of deep convection on the formation of high-wave-number Rossby waves. We are currently working on generalizing these results (not only looking at case studies, but on the overall effect of deep convection leading to Rossby wave breaking). We believe that an analysis related to these high wave numbers would optimally be addressed in such a forthcoming study, rather than on the short timescales of these forecast busts. We suggest such directions for future research in the Conclusion of the paper (lines 542-543).

4 - Comment: Please clarify whether the issue of the relationship between the warm conveyor belt and the Rossby wave magnitude. Isn't the warm conveyor belt an inherent part of the Rossby wave dynamics? If not, clarify this point and substantiate with references that have made this finding. If the conveyor belt and Rossby waves are linked, then modify your discussion to increase that linkages between these two dynamical processes and clarify how the wave magnitude increases.

We clarified the link between WCB and Rossby waves in lines 55-61: "WCBs have also been shown to be closely linked to Rossby wave breaking (Madonna et al., 2014a), with most WCB outflows co-occurring with PV streamers (as proxies for Rossby wave breaking), although the converse relationship is less common, underscoring the importance of diabatic processes for downstream wave breaking. Building on this, Vishnupriya et al. (2025) documented different modes of WCB interaction with the waveguide and showed that WCBs characterized by stronger negative PV anomalies, deeper ascent, and stronger interaction with the waveguide can not only amplify ridges but also contribute to the formation of blocking events and cut-off lows. Together, these studies highlight the central role of WCBs in shaping Rossby wave breaking and downstream flow evolution." We believe that this underscores the linkage between Warm conveyor belts and Rossby wave breaking, highlighting more past research.

5 - Comment: Improve the linkage between your results and the past research on forecast busts. The discussion and conclusion section does not sufficiently link into the broad research discussed in the introduction. The text in the last section almost suggested that authors have more knowledge of the coarse grid model climate modeling than an insightful look at convection and Rossby wave dynamics. You should consider including a brief, but more in-depth linkage to the broad past research that is mentioned in the introduction. This issue is clearly important as your study is only looking at three cases so one cannot get a firm, trustworthy conclusion without working harder on placing your research into the broader prospective.

Thank you for pointing out the missing link to the broad state of research. We added some more literature discussion in the last section of the manuscript. (lines 515-525):
“Overall, the three cases consistently highlight the sensitivity of medium-range forecasts to mesoscale dynamical and diabatic processes, consistent with previous studies (Rodwell et al., 2013; Magnusson, 2017; Grams et al., 2018). All of the case studies examined feature some form of deep convection that generates negative PV anomalies and subsequently perturbs the waveguide, in agreement with earlier work (Oertel et al., 2020; Blanchard et al., 2021; Lojko et al., 2022, 2025a, b).[...] These results found in kilometer-scale simulations mirror some previous studies based on observations (Blanchard et al., 2020, 2021) and reanalysis (Madonna et al., 2014a; Oertel et al., 2023; Vishnupriya et al., 2025).”

6 - Comment: A possible point to consider: *The results show that coarse grid models have relatively poor accuracy in simulating these case studies. So to increase the impact of this research, the authors should consider including a sentence or two that has some speculation on how to improve convective parameterization in coarse models.*

In lines 505-514, we provide a discussion on scale-dependent parametrization schemes. Previous literature has shown that such schemes tend to improve the representation of convection in the gray zone, leading to generally improved precipitation, turbulence and dynamics compared to both fully parameterized and convection-off configurations. Going into more depth on parametrization schemes would be out of the scope of the paper, and there is already existing literature on this topic (see Zhang et al. 2021 as an example):

“To address such issues, scale-dependent deep convection schemes have been developed. These parameterization schemes tend to improve the representation of convection in the gray zone by providing a more consistent transition between resolved and parameterized processes, leading to generally improved precipitation, turbulence and dynamics compared to both fully parameterized and convection-off configurations (Tomassini et al., 2023; Park et al., 2024; Wang, 2022). However, such results have been shown to be flow dependent

(Tomassini et al., 2023), and sensitive to uncertain sub-grid parameterizations (Abbott et al., 2025).”

Review 2

Forecast error vs chaotic behaviour

The article would benefit from discussion on the interpretation of ensemble spread and low ACC even in those ensemble members that exhibit improved results. To what extent are we talking about forecast error and to what extent are we talking about inherent unpredictability of the system. The following are some comments relevant to this aspect:

Abstract: L18-19: 'ensemble spread and forecast error' Are these to forecast features at the same level, i.e. are both desirable or undesirable? Thank you for bringing up this point. The sentence could indeed benefit from some clarity. To emphasize how km-scale simulations allow for better resolution of the mesoscale diabatic processes and the following scale interaction, we decided to leave out the "ensemble spread" discussion for the main part of the manuscript. (eg. 3.1.2, 3.2.2, 3.3.2).

L106: An alternative is that forecast busts are related to inherent low predictability of the system. How can we distinguish between the two alternative possibilities? To some extent, forecast busts are associated with poor system predictability. However, if we observe an increase in forecast skill with decreasing grid spacing, it shows that the underlying issue comes from poor skill from coarse grid spacing simulations to represent key diabatic features and their interactions with the large-scale flow. We include such a discussion in lines 416-419:

"For comparison, the best and worst ICON ensemble members are shown in Fig. 10a. The former successfully captures the downstream ridge amplification, and the latter exhibits weak wave amplification and limited ridge development, highlighting the variability across ensemble members."

L473: Comment on situations where ensemble spread signals an inherently unpredictable situation. How do we distinguish between those and one in which the ensemble spread is erroneously large? Would it be useful to incorporate comparisons against the ensemble mean error in the analysis? Leveraging an ensemble forecasting framework for all three cases was the reason to distinguish the improved predictability from smaller changes in grid spacings compared to chaotic error growth. The improved ensemble forecasting skill shows that the improvement comes from the refined grid spacing rather than from the initial conditions.

L279-280: I agree that at 40km resolution the ACC is generally low but some ensemble members yield very high ACC. How do you interpret this? We thank the reviewer for highlighting these high-scoring ensemble members at 40 km resolution. These outliers achieve the right results, but not consistently, i.e., only a few members reach such high scores, which is not representative of the full distribution. A 40-km grid cannot explicitly resolve the structural evolution of an MCS (see Figure 3a and 3c for example). In a few ensemble members, specific initial condition perturbations happen to trigger parameterized convection in the right way and leading to a good forecast. However, because the physical mechanism (the MCS) is missing, this high ACC is driven by initial condition spread and compensatory errors, rather than an accurate physical representation of the scale interactions. We have updated the manuscript to clarify this distinction. (see lines 312-314):

“While a few coarse-grid-spacing simulations (e.g., 40 km or 20 km) also exhibit high ACC values, these cases do not reflect an improved representation of the underlying physical processes. They are rather outliers within the ensemble where specific initial-condition perturbations compensated for errors related to the coarse resolution, driving the correct synoptic development. Therefore, while the resulting forecast is accurate, it is achieved for the wrong physical reasons and lacks the consistent physical fidelity seen at km-scale grid spacings.”

If we take the third case as an example (and referring to Figure 11a): ensembles that (even at coarse grid spacing) perform well showcase an accurate trough development (close to the ERA5 line). This emphasizes that this trough is necessary, and finer grid spacings consistently resolve it better, yielding to higher ACC on days 6 of the forecast. (lines 268-270):

“The results from the negative PV area and precipitation distributions demonstrate that higher-resolution simulations better resolve the intense MCS convection in this region, ultimately leading to lower WAE. Figure A1 shows the percentage of ensemble members exceeding 80 mm of accumulated precipitation over the evaluation period across all model grid spacings (consistent with Fig. 3b).”

L291-292: I disagree. They are outliers when taken as part of this ensemble, but they are still able to capture the actual system's evolution. It is true that they do capture the system’s evolution, but there is no consistent (across members of the same grid spacing) accurate representation of the MCS for example, (c.f., last comment). However, we agree that the phrasing of this sentence is misleading, so we updated the manuscript accordingly. (lines 312-314)

L410: Are these two clusters pointing to inherent unpredictability in the system, which is also reflected in the ACC plot (Fig. 13). Thank you for bringing up this possible unpredictability in the system. It could be that the two clusters showcase this nicely, and that the ACC in IFS ensemble members exhibits this particular double cluster. We include a short discussion at the end of section 3.3.2, without entering in too many details as this would diverge from the main objective of the paper. (lines 472-476):

L458: Here or elsewhere, comment on the level of ACC reached as some of the cases do not reach particularly high levels. What is the mean ACC and how do the improvements compare to that? It is true that some cases do have low mean ACC even at fine grid spacings. As the reviewer mentioned, this could be related to low predictability of the system. We include some discussion on this topic as well as the median detached by the ensembles in the conclusion as well as in lines (381-391).

Results in the context of the operational IFS and ERA5
The article should include more comments on how the ICON simulations compare to the IFS and ERA5. For example, it should discuss what processes are expected to be resolved by ERA5 and what processes are not resolved by the reanalysis.

L349-351: In this case the 20km resolution is the one that performs the worst, similar to the IFS even though the IFS had at least double resolution…? It would be informative to comment on the results in the context of the operational forecasts and the reanalysis. Despite showing the lowest mean at 20km, ICON shows similar ACC for 40km and 20 km (the difference is not statistically significant). The operational IFS has a grid spacing of 9 km and shows a mean ACC that is similar to these results. One reason for such differences could be that IFS is a rather diffusive model and has a lower effective resolution compared to ICON (Zeman et al. 2021). We mentioned this past research in lines 528-530:

“In this context, a model's effective resolution could also be an important factor in the representation of such mesoscale processes. For example, the rather diffusive nature of IFS (Zeman et al., 2021) likely results in a lower effective resolution than ICON with a similar grid spacing. This difference could explain why ICON achieves similar ACC at half the grid spacing of the IFS“

In the next comments, we include more discussions on IFS and ERA5 ability to represent such features.

Figure 9: These simulations can be grouped as with and without parametrised convection. For both groups the ACC improves as resolution increases, but both reach similar levels at 10km the first and 2.5 km the second. It's interesting that some members at 10km seem to perform much better than all those at 2.5 km and that the ACC achieved in this case is generally lower than that in the first case. Following the reviewer’s comments, we include a short discussion on the grouping of these simulations with and without parametrization. We included a discussion on the overall low mean ACC (compared to the first case) at the end of the subsection. (lines 395-401):

“Categorizing the simulations into parameterized and explicit convection reveals that both groups show improved ACC with finer grid spacing. Nevertheless, both attain similar ACC performance at 10 km and 2.5 km resolution. With only three members, the ACC distribution of the 2.5 km ensemble is insufficiently sampled, and a larger ensemble would likely show higher-performing individual members. The overall low mean ACC, as well as little improvements with decreasing grid spacing, highlight the low predictability of the system. Ultimately, this case study illustrates the inherently low predictability of certain forecast busts, demonstrating that decreasing grid spacing alone is insufficient to fully overcome these initial-condition constraints.”

L426: Is this due to IFS and ICON having similar resolution at 10km? It could be caused by this. However, we do not perform a detailed and systematic analysis based on the IFS model, as this is out of the scope (technically and contents-wise) of the paper, so we do not enter this topic in detail. However, we include more details on the IFS simulations as well as the grid spacing and effective resolution in section 2.1:

"Since the selected cases are sampled from different years, different operational versions of the IFS with different grid spacings were used. For the first and last cases (in 2020), the ensemble product uses an horizontal grid spacing of approximately 18 km (Cycle 46r1) whereas it is 9 km in the second case study (Cycle 48r1)."

L455-456: Here or elsewhere, more discussion is needed on the ability of ERA5 to represent these features itself as its resolution is in the middle in comparison with those used here and it's low in comparison with that of the IFS. We include such a discussion in lines 133-135: *"Furthermore, although ERA5 does not explicitly resolve mesoscale dynamics (Li et al., 2024), its data assimilation enables it to capture their upscale impacts on the large-scale flow—which is not the case for IFS, which depends only on initial conditions."*

Physical phenomena v their model representation

L38-43: It would be useful to make a distinction between the physical phenomena and their representation in models. It is the latter that leads to subsequent forecast error. Tropical cyclones are sustained sources of diabatic heating. Are we systematically under- or over-estimating this heating in models? Thank you for this comment! The previous sentence did not make this distinction. Furthermore, coarse-resolution models systematically underestimate (and spatially smooth) the diabatic heating within organized convection and TCs. This underestimation of heating leads to insufficient upper-level PV destruction and an underestimation of downstream ridge building (see Willison et al,. 2013). We added a sentence on this topic in the introduction section:
"It has been shown (Willison et al., 2013) that coarse-resolution models systematically underestimate diabatic heating, thereby failing to capture the resulting downstream ridge amplification and flow evolution."

Physical phenomena as error vectors

L263-264: I'm not convinced this makes sense. A physical phenomenon, in this case a WCB, can't advect/propagate error. A more objective way of putting this would be to consider the initial discrepancy as a difference in initial conditions at the start of the WCB. We thank the reviewer for this important clarification. We agree that a physical phenomenon such as a WCB does not propagate error directly. Rather, the initial discrepancy in the phase and amplitude of the trough acts as a modified initial state. We have rephrased the text accordingly. (lines 283-284):
"This error, both in phase and amplitude of the trough, creates initial conditions differences in the formation and development of the WCB whose outflow contributes to the formation of the ridge on day 6 of the forecast (Fig. 1b)."

Technical comments

L100: Should it read 'investigating' or 'improving' rather than 'investing'? Has been corrected.

L114: 'Which diabatic processes are better simulated in fine grid spacings?' This will depend on the model and its associated parametrisation schemes therefore it would be helpful at this point to have a list of the processes that are being considered here. The questions have been refined according to the first Reviewer's comments including more details.

"The main questions that we address in this study are:

1. Can the explicit simulation of upstream mesoscale processes at fine grid spacings in ICON mitigate large-scale forecast errors in IFS?

2. How does horizontal grid spacing modulate the intensity, spatial scale, and vertical structure of moist mesoscale processes?

3. To what extent do these processes modify the upper-level waveguide and how does this affect Rossby Wave Breaking?"

L112-115: The way in which the questions are written here is an implicit distinction between large-scale forecast errors and forecast busts? Can this differentiation be made explicitly, e.g. are they different in terms of lead times? See previous answer.

L121: What is the IFS's native resolution? In general, the results from the ICON simulations should be put in context by considering the resolution of the IFS and ERA5. Since our case studies span multiple years, the IFS resolution varies. For the second case, it is 18km, and for the two others it is 9km. We added such details in section 2.1 (lines 138-140):

"Since the selected cases are sampled from different years, different operational versions of the IFS with different grid spacings were used. For the first and last cases (in 2020), the ensemble product uses an horizontal grid spacing of approximately 18 km (Cycle 46r1) whereas it is 9 km in the second case study (Cycle 48r1)."

L125: 'as it is based from data assimilation using IFS' I'm not clear on what the message here is. Do you mean it is based on the same forecasting system including data assimilation? Yes, exactly. ERA5 is produced using ECMWF's Integrated Forecasting System (IFS Cycle 41r2), which encompasses both the numerical forecast model and the 12-hour 4D-Var data assimilation scheme (Hersbach et al., 2020).

L126: Delete 'perturbed medium range'. Has been corrected.

L127: Why are these ensemble members characterised as perturbed? Are they not simply the IFS ensemble members, or are there additional perturbations applied to them? We did not apply any additional perturbation. The manuscript has been updated accordingly.

L130: It should read 'Multi-SatellitE'. Delete 'satellite' after '(IMERG)'. Has been corrected.

L135-136: Unless I'm misunderstanding the idea, '...perturbations taken from the 50 perturbed IFS ensemble.' should read '...given by the 50 IFS ensemble perturbations.' Has been corrected.

L136-138: Would it be better to use the SSTs that the operational IFS actually takes as boundary conditions? Otherwise, there are multiple factors that will make the results confusing. Thank you for raising this point. We agree that using IFS's SST as boundary conditions could have been possible and make it less confusing. However, we decided to stick with the method by *Prein et al.,2026*. Also, the main comparisons in the manuscript are between ICON simulations at different grid spacing, which use all the same SST boundary conditions and allows for a fair comparison.

L153: ‘In order to compare the same ensembles…’ I thought the whole idea was to generate new ensembles at different resolutions, or do you mean the same ensemble members? Even so, in what sense would they be the same? Thank you for raising this unclear point. Across different grid spacings, ensemble members use the same initial conditions (i.e., the first ensemble member initial conditions of all grid spacing are the same). We improved the phrasing and clarity the point in the manuscript. (lines 165-167):

“To maintain consistent initial conditions across different grid spacings, we matched ensemble members by index (i.e., corresponding ensemble members across all resolution runs share identical initial states). To ensure that these subsets are representative of the full ensemble distribution, we performed bootstrap resampling which confirmed that the selected members capture the overall distribution of forecast skill at day 6.”

Equation (1): Possibly not necessary but the equation can be simplified. More importantly, should the expression include a correction for latitude as grid points at higher latitudes represent smaller areas than those at lower latitudes? This same comment is valid for (2). We did correct for latitude, but did not mention it in the manuscript. This has been added in the updated version.

L183-184: ‘... ensuring that the deterministic forecast does not represent an outlier within the ensemble distribution.’ Why is this argument not deemed necessary for the first case? And how do you decide that the evolution of an ensemble member is an outlier and not a physically plausible realisation of the system? We thank the reviewer for this insightful comment. We appreciate the opportunity to clarify our selection strategy and refine our terminology. Our initial case study focused on a forecast bust where the deterministic run met the ACC and RMSE bust criteria. As we expanded our analysis to subsequent cases, we refined our methodology to additionally require that the ensemble mean also satisfied both criteria. This stricter condition allowed us to target broader, system-wide forecast failures across the ensemble rather than cases isolated to the single deterministic realization. We added a discussion following this comment in lines 192-198:

Regarding "Outliers" vs. Plausible Realizations: We agree that our original phrasing using the term "outlier" was imprecise. We do not evaluate individual ensemble trajectories as physically unplausible. Rather, requiring the ensemble mean to also exhibit a bust ensures that the poor forecast performance reflects a systematic shift or bias across the entire ensemble distribution, rather than a scenario where the deterministic run simply diverged along a less likely forecast. (lines 193-196):

“For the first case study, the selection criteria were applied to the deterministic IFS forecast to evaluate a classic deterministic forecast bust. For the two subsequent cases, the selection methodology was expanded to additionally require that the IFS ensemble mean satisfied both criteria. This ensures that the poor forecast performance reflects a systematic shift or bias across the entire ensemble distribution, rather than a scenario where the deterministic run simply diverged along a less likely forecast.”

L189-190: Move the references next to 'LAGRANTO', otherwise they appear to document the ICON simulated data. Has been updated.

L203: The subscript 'obs' is misleading as it actually labels reanalysis data. Has been updated.

L219: 'The storm produced the third-highest daily rainfall total in the UK since 1891.' Do you have a reference to back this claim? Has been added.

L221-222: An MCS would be associated with negative PV anomalies and therefore with the formation of a ridge rather than a trough. The development needs rewriting to make it clearer. We thank the reviewer for this comment. The phrasing was unclear and we rephrased it.

Throughout the text: Discuss the figures and panels in order. We corrected the text and references to the Figures to discuss them in order.

L229: '... highlighting differences in the WCB outflow.' How do you know these differences are caused by differences in WCB outflow? An analysis using the LAGRANTO algorithm showed that the outflow of WCB differs between the different simulations (and ERA5) so we attributed this shift in ridge formation to WCB outflow. We decided not to include the trajectories in the Figures as we already have 14 Figures in the paper. We nevertheless include a discussion of these results in the manuscript following the reviewer's comment. (lines 243-245):

"A trajectory analysis using LAGRANTO (not shown) confirms that the ERA5 WCB outflow was shifted relative to the IFS, directly linking these discrepancies in ridge positioning to differences in the WCB outflow."

L233: Repetition: Delete 'performed at horizontal grid spacing ranging from 40km to 2.5km'. Has been corrected.

L234: Repetition: Delete 'The simulations are used to analyse the sensitivity of forecast skill to grid spacing and to investigate key contributing processes'. Has been corrected.

L238: Here and other places where times are quoted. To improve the paper's readability, in addition to quoting the validation time (1200 UTC 10 February) indicate forecast lead time. Double check the journal's notation requirements. Has been added to the text.

L241-242: The standard deviation has grown but its link to MCS representation is not obvious. Could MCS variability be shown explicitly? In Figure 3 (of this document) we show the probability of event-accumulated precipitation exceeding 80mm for each grid spacing. We added this figure in the appendix and discuss it in lines 269-274:

"Figure A1 shows the percentage of ensemble members exceeding 80 mm of accumulated precipitation over the evaluation period across all model grid spacings (consistent with Fig. 3b). As grid spacing decreases, precipitation becomes less widespread and more spatially localized and intense, bringing the simulations into better agreement with IMERG satellite observations. These findings are not sensitive to the accumulated precipitation threshold and align with the conclusions of the preceding analysis."

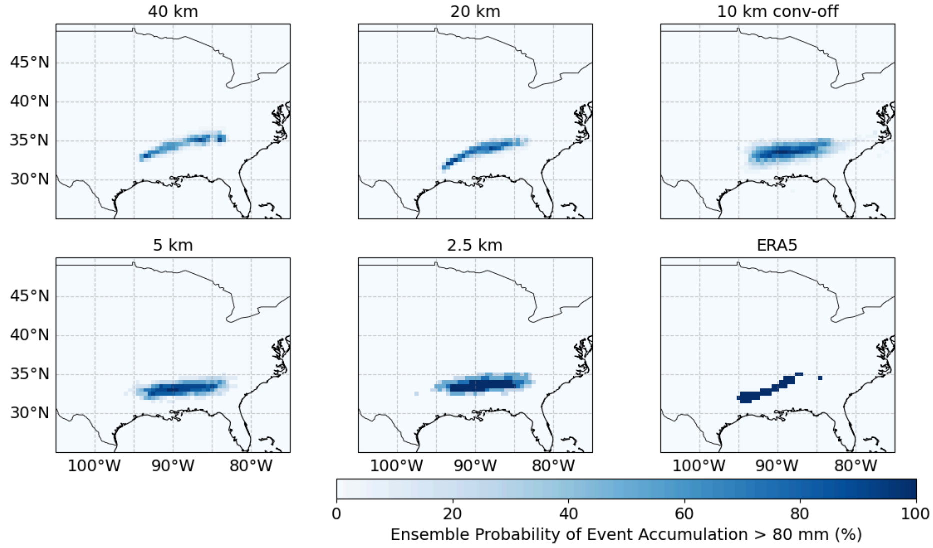

Figure 3: Probability of having an ensemble member exceeding 80 mm of precipitation over the entire event at each grid spacing.

Figure 1: The two panels could be merged into one. Has been updated.

L245: Leave conclusion to the end or indicate that this will be justified: Remove 'A clear sensitivity to horizontal grid spacing is observed'. Has been corrected.

L245-247: This statement is true but it could be affected by environmental factors such as vertical wind shear. Thank you for that comment. We also tried other proxies for MCS strength (such as vertical windspeeds) and they all led to the same conclusions so we decided to stick to NPV. We acknowledge the other factors that could affect this in the text.

L248: However, ERA5 native resolution is much lower than 2.5km grid spacing… How can this be reconciled? All of the data has been regridded on the ERA5 grid for meaningful comparison.

Figure 3 (and others): Improve figures so that legends do not obscure plots. Has been updated.

Figure 3 caption: Why do you use 3 PVU and not the standard 2 PVU contour which marks the position of the dynamical tropopause. At what level is the negative PV area calculated? We did use the standard 2 PVU contour, we apologize for this typo in the manuscript. It has been corrected in the new version. We computed the negative PV area at 250 hPa, and we added this in the caption of the Figure.

Figure 6: Language: An ensemble is a collection of ensemble members. I don't think this is the meaning intended in the legend of panel (a). Panel (b): Difference plots are easier to interpret if the reference (i.e. ERA5) is subtracted from the field you want to evaluate, i.e. IFS - ERA5. Yes, we meant ensemble member, and corrected the legend of Figure 6a. We also inverted the temperature difference so that it shows IFS - ERA5.

L311: 'positive temperature anomalies' I would have expected a negative anomaly as warm air fails to reach higher latitudes. Following the previous comment, we updated the Figure as well as the text, as they are now negative anomalies.

L336: Could you give more details on how this was calculated? How the jet stream wind speed is determined and then what does the 99th percentile represent? We first select the area shown in the box in Figure 7b (which includes the interaction between the cyclone and the jet stream). Then, we find the 99th percentile of wind speed at 250 hPa. We checked that these wind speed values are consistently associated with the jet stream location (clearly visible in the wind speed field), but decided to not show it in the paper for conciseness. In the following figure we show an example (for one simulation) of the cyclone's location as well as the location of the 99th percentile of highest wind speeds. The figure shows that this method captures the highest wind speeds in the jet stream. We confirmed the robustness of this method across timesteps and simulations.

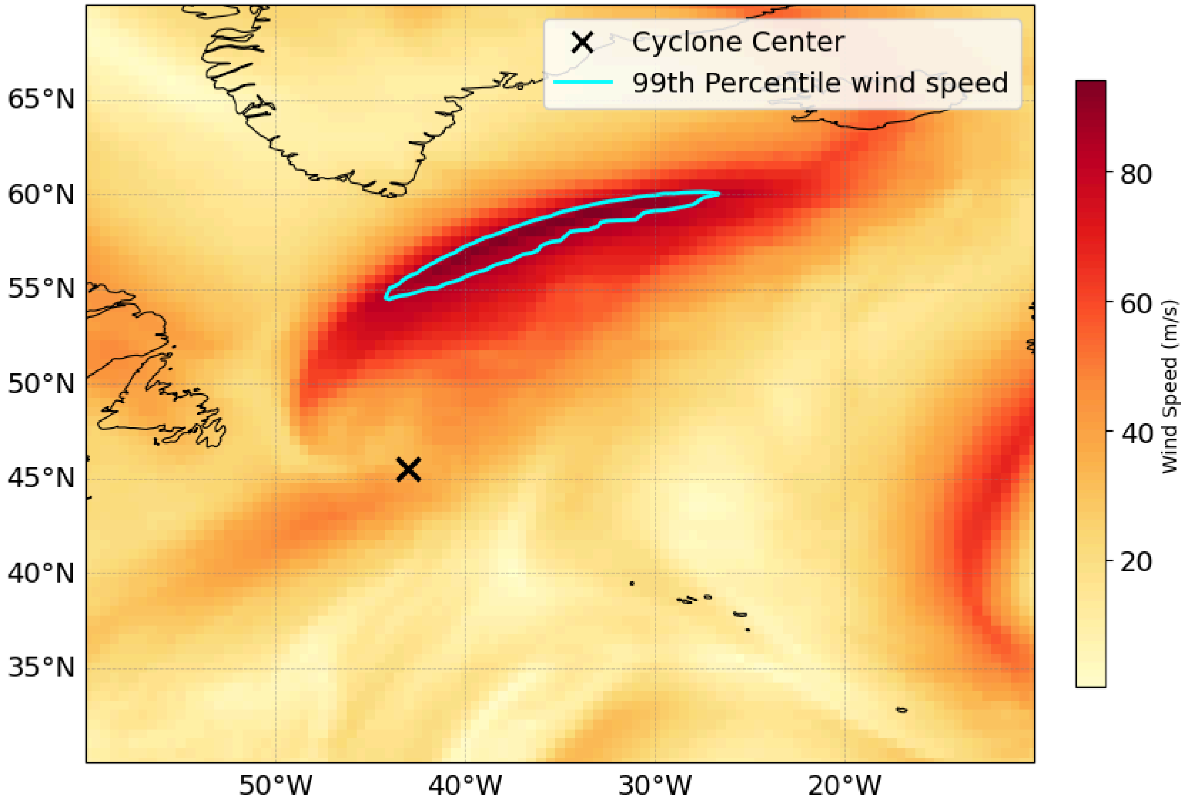

Figure 4: Example of a simulation: Windspeed (shading), position of the cyclone center (black cross) and 99th percentile of windspeed (cyan contour)

We thank the reviewer for pointing out that this was not clearly explained in the manuscript, we updated the description of the method as well as the caption of Figure 7 accordingly.

Figure 10(b) and Fig. 11(b): Are these panels discussed anywhere in the text? We corrected this and mentioned these panels in the text at the appropriate place.

L382: Delete 'that'. Has been corrected.

L383: Change 'exhibiting' to 'exhibits'. Has been corrected.

L292: 'wave activity energy' can be deleted as WAE has already been defined. Has been corrected.

L401: I assume it should read ‘Figure 10b’. Has been corrected.

L403: Change ‘ensembles’ for ‘ensemble members’. Has been corrected.

Figure 12 colorbar: Do you mean ‘percentage of ensembles’ or ‘percentage of ensemble members’? By showing only this variable it is difficult to assess whether the WCB outflow in ERA5 is actually being captured by any members in the various ensembles? Yes, the intended colorbar is “percentage of ensemble members”. We believe that this analysis shows a clear improvement of the location of the WCB outflow as it is directly comparable to ERA5. As an example, the *10km* or the *5km* simulations have an outflow that overlaps quite well with the ERA5 contour line compared to the *10km conv. off.* At 2.5km, it is hard to draw conclusions, but it is quite clear that it captures this “double cluster” outflow pattern. We could also show the Z500 patterns (like in Figure 10a), and this would also show that finer grid spacings display the ridge better than the coarse grid spacing. However such a plot would be hard to read considering the amount of simulations that we have done. Therefore, we believe that Figure 12 condenses the necessary information and conveys the right message.

L443-444: ‘divergent-flow anomalies’ What do you mean by anomalies in this context? What is the reference against which the anomalies are calculated? Do you actually mean forecast error? We thank the reviewer for pointing out that this discussion that was unclear. What Baumgart et al,. 2018 show is that if the model fails to accurately resolve MCS, it misses the divergent flow at upper levels, which is important for downstream amplification of Rossby wave dynamics. We corrected this in the manuscript. (lines 479-480)

L447: ‘...that resolving these upstream processes largely improves the forecast skill downstream’. You do show that the forecasts improve at shorter and longer lead times but I’m not convinced that the link to processes is fully shown in the paper. We thank the reviewer for this comment. We agree that the link between the improved representation of the upstream processes and the downstream forecast skill could have been made more explicit in the manuscript. Our interpretation is based on the process chain diagnosed for each of the three cases. For example, in Case 2, the improved representation of the cyclone trajectory is associated with stronger and more realistically represented jet-stream winds, which subsequently leads to an improved downstream circulation and higher ACC at day 6. Similar process-based links between improvements in the upstream evolution and downstream forecast skill are identified for Cases 1 and 3. We have revised the manuscript to clarify this connection and to avoid implying a stronger causal attribution than is supported by our analysis. (line 480-485):

“When models fail to accurately resolve this diabatic outflow, these errors project onto the tropopause and amplify downstream through nonlinear Rossby wave dynamics, ultimately degrading medium-range forecast skill. Our results extend these insights by showing that multiple diabatic sources—including MCSs, cyclones, and WCB ascent—can reinforce the jet stream and amplify ridge development through similar PV-mediated mechanisms, and that an improved representation of these upstream process chains is closely associated with enhanced forecast skill downstream.”

L453: It should read ‘Fine’ rather than ‘Fin’. Has been corrected.

References:

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803, 2020.

Li, Z., J. Peng, L. Zhang, and J. Guan, 2024: Exploring the Differences in Kinetic Energy Spectra between the NCEP FNL and ERA5 Datasets. J. Atmos. Sci., 81, 363–380, https://doi.org/10.1175/JAS-D-23-0043.1.

Prein, A. F., Pothapakula, P. K., Zeman, C., Lalonde, M., Rixen, M., Dipankar, A., Leclair, M., and Jocksch, A.: From single storms to large-scale waves: a multi-year kilometer-scale global simulation, Geoscientific Model Development, 19, 5277–5303, https://doi.org/10.5194/gmd-19-5277-2026, 2026

Willison, J., Robinson, W. A., and Lackmann, G. M.: The Importance of Resolving Mesoscale Latent Heating in the North Atlantic Storm Track, https://doi.org/10.1175/JAS-D-12-0226.1, 2013.

Zeman, C., Wedi, N. P., Düben, P. D., Ban, N., and Schär, C.: Model intercomparison of COSMO 5.0 and IFS 45r1 at kilometer-scale grid spacing, Geosci. Model Dev., 14, 4617–4639, https://doi.org/10.5194/gmd-14-4617-2021, 2021.