Response letter for *Advancing the Capabilities for Efficient Hurricane-Centric Simulations with the Atmospheric Model ICON*

Response to reviewer #1:

Review for "Advancing the Capabilities for Efficient Hurricane-Centric Simulations with the Atmospheric Model ICON" by Senf and Cremer for GMD

Summary

This study addresses the issue of performing high-resolution, global-storm resolving simulations while balancing computing resource and storage requirements. The authors propose a method to mimic Lagrangian-like simulations by shifting grids along the path of hurricanes, the phenomenon of interest in this study. The focus is on the efficiency gain as well as possible spin-up effects.

Dear Fabian and Roxana, it was a pleasure to read the manuscript. It is well structured, nicely written, and informative. The figures are clear and use appropriate colormaps. I believe the study should be published after considering my comments below, which are all of minor nature or editorial. Best, Nadja

Dear Nadja, we sincerely thank you for your thorough and insightful review. Your detailed comments have been invaluable in strengthening our manuscript and clarifying important technical aspects of our work. We appreciate your constructive feedback and the care you took in reading and reviewing our study. Best, Fabian and Roxana

Minor comments

- Line 138: I do not understand what you mean by "ten revolutions". Could you elaborate on that? That is mostly my ignorance and not being familiar with tobac. Also why, do you need to coarse-grain the vorticity field? What does the convolution filter exactly do?

  We were somewhat unclear here and have revised the wording accordingly. We mean that, for the field of coarse-grained relative vorticity, a threshold of $\zeta_c = 10\,\mathrm{day}^{-1}$ was chosen to identify the hurricane track. This threshold was determined manually and represents a good compromise to detect the investigated hurricane early and to obtain a continuous track of good quality.

  The vorticity field shows vortex structures on many different scales. To identify the hurricane track, it is necessary to distinguish the hurricane vortex structure from vortices on smaller scales. This is achieved by applying a running-average filter, which smooths the small-scale structures and thereby facilitates identification of the hurricane.

  The convolution filter is a mathematical operator that is applied to the vorticity field to perform the smoothing. It computes a weighted sum of values in the neighborhood of each point in the field, where the weights are all equal for a running-average filter. This leads to a reduction of high-frequency components in the field and facilitates identification of the hurricane. We made this more explicit in the revised manuscript.

- Line 224: In what way are these spin-up effects clearly visible? Did you perform a similar analysis as you did later for your inner nests?

  We removed the corresponding sentence from the revised manuscript, as it could be misleading. Spin-up effects in the base run are not particularly interesting for the current study, and any associated issues can be easily avoided by excluding the first full day of base run data.

- Line 228: What is the impact of having the daily-updated SST?

  We have not conducted extensive and systematic analyses to quantify the impact of daily SST updates. Therefore, we are hesitant to make definitive statements about the effects on our simulation results.

  For the base simulations, we find that the SST setup can significantly influence the hurricane track. In Fig. 1 we show Paulette's best track (from HURDAT, white) compared to simulated tracks from the coarser simulation (right, R2B9) and the finer simulation (left, R2B10 base run). For the coarser simulation, better agreement with the observed track is achieved when SST is held temporally constant. However, initializing the simulation at a later time (+ 3 days) has a larger impact on track forecast improvement than the SST setup does. For the finer simulation, track forecast deviations are smaller, and differences between SST setups are less pronounced. Nevertheless, the qualitative picture remains the same. Further studies on this topic would be very interesting but are currently beyond the scope of what our study offers.

Figure 1: Paulette best track (from HURDAT, white) compared to simulated Paulette tracks from (right) coarser simulation (on R2B9) and (left) finer simulation (base run on R2B10). Tracks are differently colored by sst setup, and init time

- Line 254: Do I understand correctly, that I as a user can specify the targeted grid resolutions? Or is it required to follow the halving of the resolution, as typical in an online nest? Could I perform a coarser simulation at 1.2 km and then immediately jump to 100 m, for example?

  Yes, this functionality is in principle possible. However, the current implementation misses the distinction between the grid of the base run and the grid from which refinement is performed. This means that currently the refinement is always performed from the base run grid, by halving

the grid spacing. The addition of a second grid for refinement can easily be implemented and left for future extensions.

We have added a sentence in the outlook part of the conclusions section to highlight that further research on this topic is needed.

- Line 255: What is your estimated effort in making more than three nests possible? What is technical difficulty behind this? Also, is there a check, that the widths of the grids are not so large, that the inner grid would overlap with the nudging zone of LBCs of the outer grid? In your case, it is clear that this is not the case, assuming you are using the standard configuration of 14 grid cells as a nudging zone.

  The real technical difficulties will probably only become apparent during implementation and testing of the functionality. We assume that implementing more than three nesting levels would mainly require changes to the runscript template. There, the corresponding required files for IC, BC, and extpar must be selected. Additionally, namelist parameters may need to be specified as a list that extends sufficiently long. All these aspects must be thoroughly reviewed and tested.

  Indeed, we have not implemented additional checks for grid properties. We assume that users will verify and adjust the grid settings themselves making the method not failsafe in that respect.

  And yes, it is certainly sensible to leave sufficient space to the edge.

- Line 279: Is it correct, that the base runs were performed with ICON v2.6.6, while the grid segments were done with v2025-04-2. Do you expect any discrepencies between the two model versions that are important to your hurricanes?

  We cannot say for sure whether significant differences arise from using different versions. We performed the base runs with version 2.6.6 because this was the current version at the time we started to work on the base runs. Since we were satisfied with the results of the runs after a series of adjustments and tests, we decided to use these simulations as a reference and did not repeat them with other ICON versions. When we developed the method for higher-resolution segments, ICON version 2025-04-2 had just become available.

  The inconsistency in the ICON versions we used is thus a reflection of the temporal development of our study. We believe that such development in the scientific environment is not atypical and hope that the differences between the versions are not so large that they substantially affect the results of our study in a qualitative manner.

- Line 282: What was the impact of adding graupel and ice number mixing ratios to the ICs? Did you see a reduction in spin-up?

  It would certainly be fine to have ICON's own diagnostics compute graupel, hail, and particle concentrations for the first calculated segment if they are not provided. Microphysical spin-up would lead to consistent cloud variables being established after some time (perhaps hours). However, if one considers the transitions to the next segments, the absence of information about graupel, hail, and particle concentrations would lead to abrupt changes in these fields.

Our method, however, places emphasis on the structurally consistent evolution of all dynamic, thermodynamic, and cloud variables over time, at least in the Lagrangian analysis region. This also makes it possible to consistently track existing smaller-scale convective elements across segment transitions. The inclusion of graupel, hail, and particle concentrations in the ICs is so essential for us that we have not performed any further tests without these quantities in the ICs.

We added a sentence to the revised manuscript to clarify this point.

- Figure 7: What is the temporal resolution of the output, as it seem quite noisy to me, especially in the maximum 10 m wind speed. Would a running mean help here? Can you also plot the maximum 10 m wind speed and minimum surface pressure of the base run, which is the white track?

  We output all 2D variables (surface, radiation and vertically-integrated cloud variables) on the regular longitude-latitude grid every 5 minutes. Convective activity generates particularly high gustiness of the 10 m wind, which explains the fluctuations on the very short time scales.

  We have implemented all suggested updates to Figure 7. Using a running mean with a width of 2 hours, we smoothed the curves for minimum surface pressure and maximum wind speed. The original curves are retained as light gray shading in the background. Additionally, we have added the base run data as a black curve for comparison. We have updated the figure caption accordingly.

- Figure 9: Could you be more specific regarding the discontinuous jumps in the subplots? I am not sure I fully see them, maybe highlighting them with some circles could help, or you point to them more explicitly in the figure caption.

  We recognize that our original description was unclear, and we have revised it in the updated manuscript for greater clarity.

  To explain: The discontinuous jumps arise from the distinction between Eulerian and Lagrangian reference frames. In the Eulerian perspective, the individual segment domains are fixed in space and time, while the Lagrangian analysis domain moves with the hurricane across these fixed domains. Conversely, in the Lagrangian perspective, the analysis domain remains stationary while the segment domains move beneath it in the opposite direction. At reinitialization times, the segments jump to new positions, causing the segment domain boundaries to shift discontinuously relative to the Lagrangian analysis domain. These discontinuous jumps, visible as abrupt spatial displacements of the segment domains in the figure, are what Figure 9 further illustrates.

- Line 323: Was the subtimestepping of the dynamical core changed in your grid setups? If so, why? Because of stability reasons?

  The number of dynamical substeps is set to the default value of 5 for our simulations. We carefully check all logfiles again and did not find any indication for adaptive changes in the number of substeps.

  We changed the formulation in the revised manuscript.

- Line 336: Is your time step at 2.5 km grid resolution 6 seconds? If so, is this is because of stability reasons? I would have expected a higher time step.

  6 seconds is applied at the 1.2 km grid resolution. We reduced the time step from 10 seconds - the limit imposed by the horizontal-flow CFL criterion - to 6 seconds for stability reasons. For comparison, the 5 km and 2.5 km setups use time steps of 30 and 15 seconds, respectively. The high-resolution setup employs 150 vertical levels, representing a relatively fine vertical discretization. Combined with the large vertical wind speeds generated by intense deep convection, this fine vertical resolution appears to necessitate the smaller time step.

  We mentioned 1.2 km grid spacing in the text body, but we will clarify this point in the revised manuscript.

- Line 343: I am convinced that your method is saving computing resources, and believing right now your numbers, this would lead to a substantial decrease in computing resources. However, I am wondering if this is true and not a slight overstatement, given that in your 3-nest setup + narrow width, the hurricane is cut off across all resolutions, but strongest in the case of 300 m naturally. My thinking here is, if your statement with a factor of up to 175 is a bit too strong, because coming from a scientific perspective, I am not sure how much we would gain from the innermost nest with the narrow width? Hence, I would be curious to hear "more realistic/applicable" numbers.

  We revisited the speedup part in depth and redesigned the calculations accordingly. The most important change is that we dramatically reduced the size of the considered reference case: the classical limited-area setup. In the earlier submitted manuscript version we had built a reference domain that covered the entire Atlantic, making it usable for many different studies. We have now changed that. In the current manuscript we only consider the Atlantic region that actually contains the hurricane (in our case, Paulette). We defined this region with a bounding box and we believe that users of high-resolution simulations might choose a similar approach to keep the final domain as small as possible. As a result of these changes, the speed-up factors are reduced by a factor of four.

  Another important change concerns the introduction of the tube as a reference case. This was suggested by Reviewer #2 to better assess the performance impact of grid segmentation. In our presentation we have now adopted the following strategy: first, we discuss the speed-up obtained by moving from a bounding-box domain to a tube domain, which yields a factor of roughly 2–8 depending on the tube's width and, of course, on the compactness of the hurricane tracks. Next, we introduce segmentation of the tube, which typically adds another factor of about 2–6. This way of presenting the results leads to more moderate speedup figures being shown, encouraging potential users to consider the limitations more carefully. Among those limitations is the fact that the narrow domain does not contain important parts of the organized cloud system and therefore sometimes functions more like a dynamic downscaling approach.

- Figure 10: Any idea what is happening to segment 02 and 03? Why are the behaving quite different to the other segments in your analysis? Why do you see for some segments and

variables the zigzag pattern? What is your output frequency that your analysis is based on?

In response to comments from reviewers #2 and #3, we have revised the spin-up analysis in Figures 10 and 11. New simulations were performed in which an additional three hours were computed for each segment, creating a three-hour overlap with the subsequent segment. This analysis is now substantially more robust and can be better utilized for quantifying spin-up effects.

With the new method, the zigzag pattern has disappeared. We have also modified the figure to display only the later segments, which clearly illustrate the spin-up effects in a systematic manner. The output frequency for the analysis is 5 minutes, which is the same as the output frequency of all 2d variables.

- Figure 11: So, blue represents the 10th percentile, while red the 90th? Could you specify this in the figure caption? Was the analysis here and Figure 10 also done for the higher resolutions? Are you changing vertical levels? This would also impact your spin-up. Have you looked at the cloud quantities vertically resolved before integrating them? I would be curious if you see some time dependence of the spin-up across the vertical dimension.

  The color coding in Figure 11 represents different segments, with blue indicating earlier segments and red indicating later segments. We have moved the color coding description to the beginning of the figure caption to make this point clear from early on. Due to the new method for analyzing spin-up effects, data are no longer available for performing this analysis at higher resolutions. This is because we needed to rerun the simulations with a three-hour overlap for each segment, which was computationally prohibitive for the higher-resolution setups.

  The number of vertical levels increases from 70 to 150 when transitioning from the base run to the higher-resolution segments. During further horizontal refinement, the number of levels remains constant at 150. We have not performed spin-up analyses for the vertical structure and therefore cannot make statements about spin-up effects in the vertical dimension. To further investigate these questions, three-dimensional data with high temporal resolution would need to be stored.

Editorial comments

- Line 8: You usually refer to your tropical cyclones, as hurricanes, so to stay consistent I would also call them in this instance hurricanes instead of tropical cyclones.

  done.

- Line 45: add the reference to Zängl et al., 2014 after introducing ICON

  done.

- Line 55: I am being very picky here, but what do you define as “good”? Realistic? Accurate? Precise? In relation to what?

  changed to ”accurate”

- Line 66/202: You introduced ICON already so I would omit the Earth system part / weather and climate model part.

  mentioned parts omitted.

- Line 107: Regarding the merging of ICs: Could you point to your extensive discussion later? I stumbled across this while reading and marked it hoping you would provide more details later, which you did. So, a simple reference to the section would already help.

  done.

- Figure 3: R02B10 should be explained in the figure caption, because this is only later used in the text. Same goes for Figure 4, and therein L70.

  done.

- Line 207: replace “times” by x and “km” should be km^2.

  done.

- Figure 5: The tiny world map in the upper right corner shows the extent of the shown reflect solar radiation flux, right? I was wondering, if you can add as well the nests described in Lines 207ff, such that it is clear how large these base runs are. Either for both R02B09 and R02B10, or at least for the latter.

  We implemented a yellow outline for R2B10 in the inset map of Figure 5. However, we decided not to include R2B09, as it would have made the figure too cluttered and less clear. We have added text to the figure caption to clarify this point.

- Line 307: there is a word missing in “must make between . . . ”.

  slightly rephrased the sentence

- Line 372: Because you showcase a variety of resolutions, it would help to reiterate which resolution you are talking about in the case of the “outer nest”.

  link to Tab. 1 added and domain configuration specified in text body.

Response letter for *Advancing the Capabilities for Efficient Hurricane-Centric Simulations with the Atmospheric Model ICON*

Response to reviewer #2:

We sincerely thank reviewer #2 for their thorough and constructive comments. The detailed and thoughtful feedback has helped us significantly improve the manuscript. We have carefully followed all suggestions, rewritten unclear passages, and included new analyses. These revisions provide greater clarity and rigor, particularly in the spin-up characterization and our discussion of the efficiency gains, limitations, and underlying assumptions of our method.

Peer-Review for "Advancing the Capabilities for Efficient Hurricane-Centric Simulations with the Atmospheric Model ICON" by Senf and Cremer

Summary

This study presents a workflow enabling efficient high-resolution simulations of tropical cyclones with the ICON atmospheric model. Rather than using a single large fixed limited-area domain spanning an ocean basin, the method generates a sequence of tube-shaped unstructured grid segments that follow the hurricane track, each covering 12 to 24 hours of the storm's movement. The workflow automates hurricane tracking, grid generation and initial condition merging across successive segments via distance-weighted blending. All simulations are performed in a pure limited-area model chain: ERA5 drives a fixed outer regional domain (5 km), which drives a fixed inner domain (2.5 km), which in turn provides initial and boundary conditions for the series of high-resolution segments. The method is demonstrated on Hurricane Paulette (2020), with efficiency gains of 13 to 175 times compared to a conventional fixed-domain approach are reported. Spin-up effects at segment transitions are argued to be manageable as well.

The manuscript is well-written and the figures are well designed. Figs. 8 and 9 are excellent, and give a compelling illustration of what the method actually produces at hectometer scales. I recommend acceptance after the authors address the comments below, notably in how they define the spurious transient response to the segmentation.

Main Comments

1. The spin-up characterization in Sect. 3.5 (Figs. 10 and 11) is the paper's central methodological claim, but this diagnostic could be made more convincing by comparing every segment with its neighbour, rather than with its own initial state at t = 0. As I understand, the author's diagnostic of what constitutes "spin-up" mixes two contributions that are hard to separate: the physical evolution of the hurricane and any spurious adjustment from the inconsistent initial conditions. Fig. 11 has a related issue: the PDFs are colored by segment index, and the interpretation is that a shift of the red bars indicates systematic spin-up. However, later segments correspond to a more intense and more variable storm, so how could their larger anomalies be attributed to restart quality alone? Therefore, the more informative diagnostic would be to overlay the same physical quantity from two consecutive overlapping segments k and $k+1$ during their shared time window. Then, any discontinuity at the restart time could be attributed to the initialization. This could be diagnosed quickly, as it requires no additional runs: the overlapping data exists by

construction.

In the original implementation of our workflow, the calculations were set up so that at time t_{k+1} the simulation moves directly from segment k to segment $k+1$. Consequently, in all of our production runs there are no data in which two consecutive segments overlap, which is why we could not apply the proposed analysis to our production runs.

Because Reviewer #3 also suggested running sensitivity experiments with a modified LBC flag and a diagnostic for pressure tendency, we decided to generate the proposed analyses from additional simulation runs. Our analysis of the spinup effects, now referred to as the re-initialization effect, follows your suggestion by using temporal overlap and additionally discusses the impact of the LBC flag. For greater clarity, we have placed this analysis in a separate section.

We hope that the new analysis and discussion of the re-initialization effect, which is now included in the revised manuscript, sufficiently addresses this major comment.

2. A simpler alternative to the restart chain that is not discussed anywhere in the paper is a single tube-shaped domain covering the full hurricane track, run as a continuous simulation using standard ICON checkpoint restarts on the same fixed grid. This would eliminate all IC-merging artifacts and spin-up at segment boundaries while still remaining more efficient than any rectangular fixed domain. For Paulette's relatively straight track, the cell-count penalty compared to the 12h segment approach would be modest: the along-track extension adds cells, but the tube geometry retains the full efficiency advantage over a rectangle. I get the argument in favor of short segments (that is cells far ahead of and behind the storm are computed at every timestep, far from the cyclone), but whether this is a significant cost depends on track geometry, storm speed, and the chosen domain width. To me it is not so straightforward for a case like Paulette. Perhaps the authors could provide a brief quantitative cell-count+timestep estimate comparing both approaches for this case, and discuss under what conditions the restart chain is comparable, if not preferable due to the absence of restarting artefacts?

   Following your suggestion, we have completely revised the reference case used to compute the speed-up of our method. First, we introduce a tube-shaped domain that covers the hurricane track with a chosen width. The compactness of the hurricane primarily determines the spin-up cost. Next, we address how much additional speed-up can be achieved by segmenting the tube.

   Another major change is the drastic reduction of the reference case—that is, the classic limited-area setup (also to address Reviewer #1's comments). In the originally submitted manuscript we constructed a reference domain that spanned the entire Atlantic, making it usable for many different studies. We have now replaced that with a much smaller domain. In the current manuscript we consider only the portion of the Atlantic that actually contains the hurricane (in our case, Paulette). This region is defined by a bounding box, and we demonstrate that users of high-resolution simulations can adopt a similar approach to keep the final domain as small as possible. As a result of these changes, the speed-up factors are reduced by roughly a factor of four.

Minor Comments

1. Line 138: I am not familiar with this criterion of ”ten daily revolutions” for identifying cyclonic features. What does this correspond to physically in terms of vorticity magnitude, and how sensitive is the tracking to this threshold?

   We were somewhat unclear here and have revised the wording accordingly. We mean that, for the field of coarse-grained relative vorticity, a threshold of $\zeta_c = 10\,\mathrm{day}^{-1}$ was chosen to identify the hurricane track. This threshold was determined manually and represents a good compromise to detect the investigated hurricane early and to obtain a continuous track of good quality.

2. Line 170: Is the distance-based weighting used to blend IC data from two sources at segment boundaries a linear function, or something more complex?

   The weights are calculated as linear function of distance to the interface or boundary between the two data sources. A scale factor can be applied such that at distances equal to or greater than the scale, weights for IC* from the reference are set to be zero. The implementation can be found in the source code at github, line 171.

   We slightly rephrased the description in the manuscript to clarify that the weighting is linear with distance.

3. Line 193: ”One of the longest tracks”, this makes this case particularly beneficial from the techniques, as opposed to a hurricane like Melissa that strenghtened and grew while staying relatively static.

   This is absolutely correct, and we also emphasize that Paulette is an ideal case for demonstrating the method.

   Of course, the domain extent also depends on which phase of hurricane development is investigated. In Fig. 1, we show the tube of Paulette (2020) and Melissa (2025), each with a radius of 500 km, compared with the bounding box. The ratio of tube area to bounding-box area is 0.4 and 0.36, respectively, so both are very similar. Katrina (2005), by contrast, is much more compact, with a ratio of 0.64 (not shown).

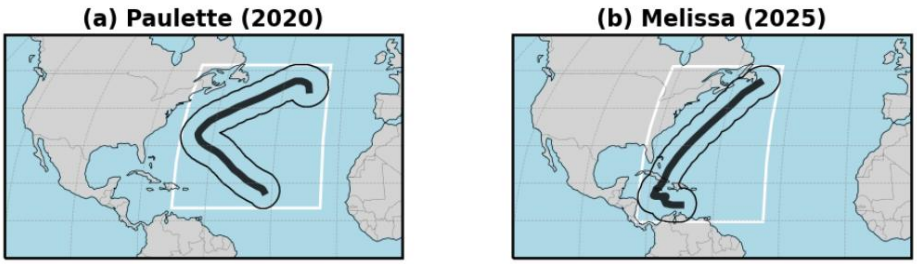

Figure 1: Comparison of the track (black thick line) tube-shaped domain (black, thin line) and bounding box (white) for (a) Hurricane Paulette (2020) and (b) Hurricane Melissa (2025). The tube across-track radius is 500 km. The ratio of tube area to bounding-box area is 0.4 and 0.36, respectively. Track data were taken from NOAA HURDAT2 database selecting time periods with status HU, extended by a maximum of 5 days prior and 2 days after the last HU flag.

4. Line 214: Is the same partial parameterization carried through to the finest nests, down to 300m-

resolution?

No, we switched off convection parameterization completely for the higher-resolution runs.

We indeed missed to specify the configuration of the hires setups. This has been corrected in the revised manuscript, where we added information in Section 3.3.

5. Line 225: I am not sure if this flag exists in the version of ICON the author used, but lsstice allows to prescribe hourly SST profiles. For a fast changing extreme event such as a cyclone, especially at 300m-resolution, this would be more accurate and could be mentioned for future implementation of this framework.

   We have not used the flag in question explicitly, to the best of our knowledge. Instead, for the mode `sst_ice_mode=6`, we expanded the capabilities and implemented this function for multiple online nests. We plan to extend the framework in the future to include hourly SST data for testing.

6. Line 280: As far as I understand, a warm start requires storing two timesteps for the time-marching scheme. Is this what is being used here, and blended in between segments using the distance-based weighting?

   No, we store and use only a single time step. Our original wording was misleading, and we have revised it in the updated manuscript for clarity.

   The blending is performed only with data from the re-initialization time step, not with the previous time step. During blending, two different data sources, data from the previous segment and data from the base run, are combined, with weights determined by the distance to the transition boundary.

7. Figure 7a: it is somewhat difficult to differentiate the two cases, but I guess that is the point? If not, perhaps one could be made with some opacity or use another colormap?

   Thank you for this suggestion. We indeed intend this figure to demonstrate that minimum surface pressure tracks at different resolutions remain very similar when the domain configuration is unchanged. We use the same color scheme with rainbow colors to represent different re-initialization times consistently throughout the manuscript. We prefer to maintain this color scheme in Figure 7a for consistency.

8. Figure 7b/c: I also have a difficult time sorting out the various overlapping resolutions from each case. The segment appear very clearly though.

   Following a suggestion by reviewer #1, we have added a running mean with a width of 2 hours to smooth the curves for minimum surface pressure and maximum wind speed in Fig. 7b and c. The original curves are retained as a light background (gray shadings) to maintain visibility. Additionally, we have added the base run data as a black curve for comparison. We have updated the figure caption accordingly.

9. Figure 8: This is really well-made figure! Do I see correctly that the eye of the storm is weaker, more diffused in the highest-resolution case? This is a bit counter-intuitive, so I presume this is caused by the proximity with the domain's boundaries?

   This is indeed a very interesting observation! The eye of the hurricane at the highest resolution

(300 m) does appear somewhat weaker and more diffuse at the selected time point. This can likely be attributed to the proximity to the domain boundaries, since the lateral boundary conditions constrain the dynamics in this region. We also use this figure to illustrate that domains chosen too small can lead to artifacts. However, we note that temporal variability in the eye structure is substantial, so the selected time point (synchronized with Fig. 5) may not be fully representative of what can be achieved at very high resolution.

In Fig. 2 you find the ice water path (IWP) for the same time point as in Fig. 8. In this field, the eye is clearly visible, and not as diffuse as it appears in the solar radiation flux. There are multiple aspects that must be considered for characterizing the hurricane eye, and the differences between these fields highlight the importance of examining various metrics when evaluating high-resolution simulations.

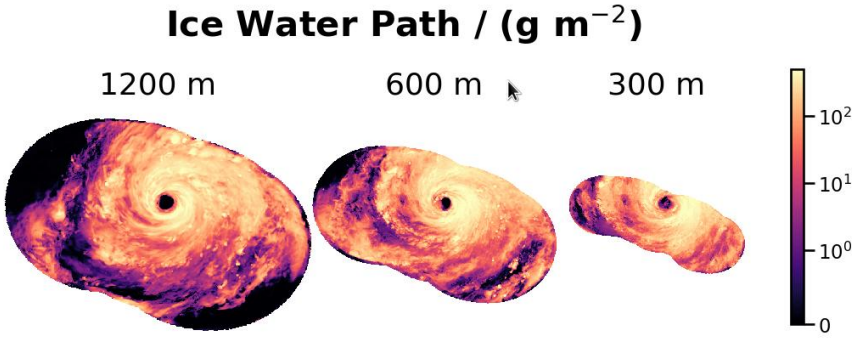

Figure 2: Ice water path (IWP) at 300 m resolution for Hurricane Paulette (2020) at the same time point as in Fig. 8. The IWP distribution shows a more diffuse structure in the eye region, likely due to proximity to domain boundaries.

We have also identified and corrected an image reference error in the text and hope that this aspect is now sufficiently clear.

10. Figure 9: Another beautiful figure, but I struggle to understand the author's point here: I cannot see any discontinuity with the background (opaque vs transparent contours), which looks in fact quite smooth.

    The same aspect was also raised by reviewer #1. We hope this answer clarifies the point:

    We recognize that our original description was unclear, and we have revised it in the updated manuscript for greater clarity.

    To explain: The discontinuous jumps arise from the distinction between Eulerian and Lagrangian reference frames. In the Eulerian perspective, the individual segment domains are fixed in space and time, while the Lagrangian analysis domain moves with the hurricane across these fixed domains. Conversely, in the Lagrangian perspective, the analysis domain remains stationary while the segment domains move beneath it in the opposite direction. At reinitialization times, the segments jump to new positions, causing the segment domain boundaries to shift discontinuously relative to the Lagrangian analysis domain. These discontinuous jumps, visible as abrupt spatial displacements of the segment domains in the figure, are what Figure 9 further illustrates.

11. Line 423: The amount of computational gain is real, especially for 12h-restarts. But eventually, as the cost diminishes due to further restricting the spatio-temporal bounds of the simulations, is the solution not too-strictly determined by the lateral/initial conditions? Can the 300m-resolution run sufficiently long to allow for the development of fine-scale physics, or has it become a sophisticated downscaling technique? I realise that there is no simple answer to this question, and that it is all a matter of trade-off, but it is worth discussing in the conclusion.

    As discussed in the answer to your main comment #1, we have recalculated the speed-up factors for a more compact reference domain, which reduces the speed-up factors by roughly a factor of four. This change is now reflected in the revised manuscript.

    We have also added a sentence in the conclusion section of the revised manuscript to address the limitations due to domain size. We emphasize that while our method allows for efficient high-resolution simulations, it is crucial to consider the trade-offs between computational efficiency and the ability to simulate the atmospheric phenomenon. The choice of segment length, the across-track width, and resolution should be guided by the specific research objectives and the physical processes of interest.

Response letter for *Advancing the Capabilities for Efficient Hurricane-Centric Simulations with the Atmospheric Model ICON*

Response to reviewer #3:

We sincerely thank the reviewer for this thorough and insightful assessment. Your detailed comments and thoughtful suggestions have been invaluable in improving our manuscript. The technical questions you raised have led us to refine our analysis and clarify important methodological aspects of our work. We particularly appreciate your constructive feedback on the spin-up analysis and the alternative initialization procedures - these have strengthened our analysis and outlook part significantly. We are grateful for the time and care you invested in reviewing our study.

Review of egusphere-2026-1412: Advancing the Capabilities for Efficient Hurricane-Centric Simulations with the Atmospheric Model ICON, Authors: Senf, F. and Cremer, R.

Summary

The authors present an automated workflow toolkit that enables the efficient simulation of tropical cyclones on a sub-kilometre scale. This is achieved by running the ICON model sequentially on multiple overlapping grid segments along the cyclone's expected path. Tailored to the ICON model, the toolkit automates several key tasks, including cyclone tracking, grid generation and the preparation of initial and boundary data, as well as conducting the high-resolution segment-based runs. The toolkit's key components rely on existing software packages, such as tobac for cyclone tracking, DWD ICON Tools for grid generation, and CDO for horizontal remapping.

The workflow toolkit is demonstrated using Hurricane Pauline (2020) with nested high-resolution simulations performed using grid sizes down to 300 m. The authors compare different configurations of the grid segments and conduct performance measurements on two different state-of-the-art HPC systems. They also discuss shortcomings of their approach, focusing on spin-up effects resulting from the remapping and merging of the initial conditions. Based on their findings the authors conclude that, compared to traditional approaches with one fixed limited-area domain, the efficiency gains are significant, reducing resource requirements by a factor of 13–175 depending on the configuration details. They also conclude that the observed spin-up effects are acceptable, given the substantial efficiency gains achieved.

General comments

The manuscript is well written and concise. It makes a valuable contribution to the scientific topic of how to conduct km-scale tropical cyclone simulations at a reasonable cost. The proposed workflow toolkit appears to be scientifically sound in most respects and is clearly described. The authors also effectively demonstrate the efficiency of their approach. The visual presentation of a Lagrangian cyclone analysis prototype is well designed, too.

However, I believe the manuscript requires a more detailed analysis of how the preparation of the initial conditions impacts simulation quality. I am not convinced that the analysis of the spin-up effects presented provides sufficient evidence to conclude that the negative impacts on simulation quality are sufficiently small. I have summarised potential pathways towards an extended model evaluation be-

low. In summary, I recommend that the manuscript be reconsidered for publication once the following major points have been addressed.

We would like to emphasize that we have made considerable efforts to better quantify the spin-up effects. The following points summarize our approach in this regard:

- We have improved the calculation of the spinup effect. It is now derived from temporally overlapping simulations. To achieve this, we created special sensitivity runs in which each segment was extended by three hours, and those additional three hours were compared with the first three hours of the subsequent segment. This adjustment follows a major comment from Reviewer #2.
- We added the mean pressure tendency as a diagnostic to our sensitivity simulations in order to provide a clearer picture of the dynamic adjustment processes. In this response we have inserted the temporal evolution of these diagnostics at the appropriate points.
- We carried out the sensitivity experiment with the proposed LBC flag and incorporated the results into the spin-up analyses in the revised manuscript. This now allows readers to see what changes and/or improvements can be expected from this setting.

Additionally, we treated the following points carefully, which we believe will further improve the manuscript:

- We have added a discussion of the Incremental Analysis Update (IAU) method as a potential alternative to our current initialization procedure. We have also highlighted the need for further research on this topic in the outlook section of the conclusions.
- We have also recomputed all speed-up calculations using a much smaller reference domain, addressing the key comments from Reviewers #1 and #2. This new reference domain is minimal rather than all-encompassing, which reduces the reported speed-up values by a factor of four. In addition, we introduced an unsegmented tube as another reference case.

Major points

Initialization of the high-resolution segment-based runs

The initial conditions (IC) for segment k are generated by merging the base run data with the output data from segment $k-1$. Section 3.5 examines the spin-up of several thermodynamic parameters resulting from that procedure during the first few hours of model integration. Spin-up effects are known to occur, if the driving model climate differs from that of the target model, or if the model climate changes significantly with grid resolution. Apart from spin-up effects, the proposed initialisation procedure may generate a significant amount of spurious sound and gravity waves ('noise') which might degrade the quality of the solution. However, the manuscript is lacking a quantitative analysis of the noise level.

We thank the reviewer for drawing attention to the important distinction between spurious sound and gravity waves, which are numerically generated artifacts, and the broader transient adjustment process following re-initialization. In response, we have rephrased the relevant sentence in the manuscript to explicitly read: "the merged ICs can excite spurious sound and gravity waves, and re-initialization effects occur at the beginning of a simulation in a new segment." Furthermore, we have

replaced the term “spin-up effects” with “re-initialization effects” throughout the manuscript wherever it refers to the adjustment following segment transitions in the high-resolution runs, reserving “spin-up” exclusively for the initial adjustment period of the base run.

As a quantitative measure of this noise, I recommend looking at time series of the domain-averaged surface pressure tendency (see the ICON model tutorial Section 2.2 for a description). Such time series can be extracted from the ICON log files. Specific questions:

- How does the initial noise level of the high-resolution runs compare to that of your base run? To make a fair comparison you might need to repeat the base run at the resolution of your coarsest segment, and to restrict the analysis to the area covered by the segment. However, a direct comparison of the existing runs may already be sufficient.

  Unfortunately, we set the message level for production runs to the default value, which means we cannot extract temporally high-resolution domain-averaged surface pressure tendency from the ICON log files for either the base run or the high-resolution simulations. To improve the spin-up analysis, an aspect raised by both you and the other reviewers, we reran experimental simulations of the hurricane-centric workflow with only a single domain (R2B11). As shown in the next response, an equilibrium level for the domain-averaged surface pressure tendency is reached within a short time, so we consider a comparison with the base run no longer necessary.

  We would also like to emphasize that the base run and the high-resolution segment runs are based on fundamentally different parameterization settings: NWP with partial convection parameterization versus LEM without convection but with 3D turbulence. This means that not only the spatial resolutions differ, but also the representation of physical processes considered in the simulations. Therefore, a direct comparison of the two simulations would convolve both changes, which complicates the interpretation of the results. Furthermore, we question whether a base run at high resolution is even meaningful, since the parameterization settings would not match the resolution. We also note that substantial computational resources would be required to conduct a high-resolution base run, which would undermine the purpose of efficiency gains achieved through segmentation. We have therefore concluded that conducting a base run equivalent at higher resolution for a direct comparison of the two simulation data is not productive.

- Does a reinitialisation interval of 12h or 24h provide sufficient time for the initial noise to decay? In other words, is the timeseries of the surface pressure tendency a flat curve by the end of the reinitialisation interval, or does it still slope? If it still slopes, your reinitialisation procedure may accumulate noise over successive segment runs, as your IC merging procedure will transfer spurious waves from one segment run to the next.

  In the figures 1–3, we present the time series of the domain-averaged surface pressure tendency for the high-resolution segment-based runs. The configuration is set to medium-sized width and 12h-reinit. The production setting with LBC flag OFF (thin lines) is compared to the experimental setting with LBC flag ON (thick lines). It shows that there is always significant noise at the beginning of each segment, which decays within a few hours. The time series of the surface pressure tendency is flat by the end of the reinitialization interval, indicating that the initial noise has decayed sufficiently. This suggests that the reinitialization procedure does not

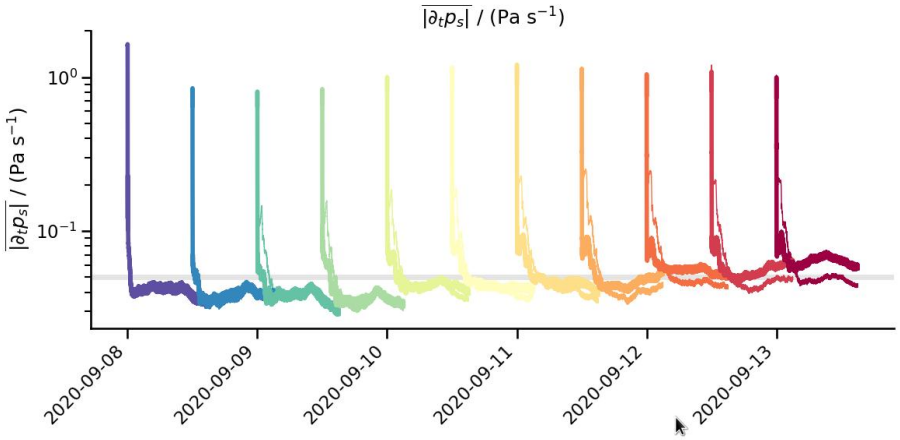

Figure 1: Full time series of the domain-averaged surface pressure tendency for the high-resolution segment-based runs. Configuration is set to medium-sized width and 12h-reinit. The production setting with LBC flag OFF (thin lines) is compared to the experimental setting with LBC flag ON (thick lines).

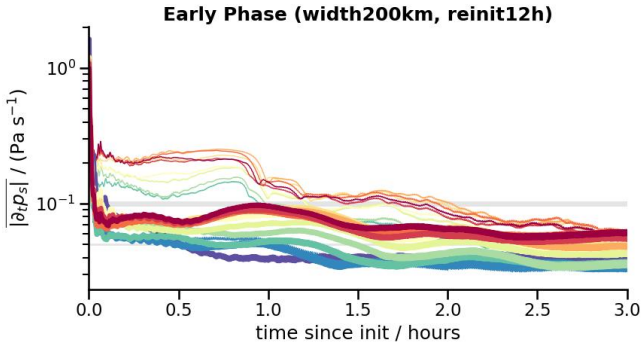

Figure 2: Early phase of the time series of the domain-averaged surface pressure tendency for the high-resolution segment-based runs. Configuration is set to medium-sized width and 12h-reinit. The production setting with LBC flag OFF (thin lines) is compared to the experimental setting with LBC flag ON (thick lines).

accumulate noise over successive segment runs.

Figure 2 shows the early phase of the time series, where the initial noise is most pronounced. It can be seen that the LBC flag has a significant impact on the noise level, with the experimental setting (LBC flag ON) showing a lower noise level than the production setting (LBC flag OFF). This suggests that the LBC flag is effective in reducing the initial noise in the high-resolution segment-based runs. This supports your recommendation to apply the LBC flag ON setting in our workflow, a recommendation we have now implemented in the revised manuscript.

Figure 3 shows the late phase of the time series, averaged over the last three hours, with error bars representing the standard deviation. The results indicate that LBC flag ON setting leads higher values of the domain-averaged surface pressure tendency, which suggests that the LBC flog ON setting also has some weaknesses. For this, we would derive that our production setting with LBC flag OFF still has some legitimation and further detailed research on this topic would be beneficial. However, we would like to emphasize much more detailed work on the impact of this LBC flag is outside our current capacity and would require a dedicated study.

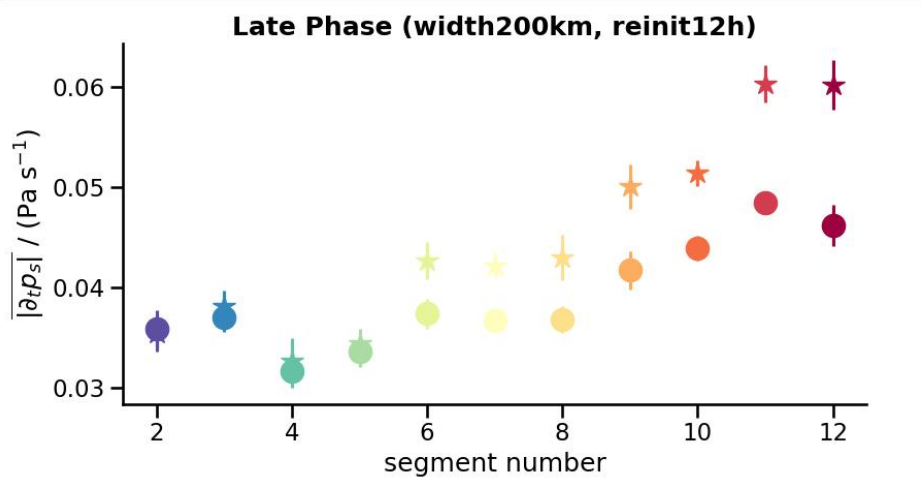

Figure 3: Late phase of the domain-averaged surface pressure tendency for the high-resolution segment-based runs averaged over the last three hours. Configuration is set to medium-sized width and 12h-reinit. The production setting with LBC flag OFF (filled circles) is compared to the experimental setting with LBC flag ON (stars). Errorbars represent the standard deviation.

Alternative initialization procedures

Have the authors explored alternative model initialisation procedures apart from IC merging? Might the Incremental Analysis Update (IAU) procedure be applicable? IAU is an effective filtering technique for spurious high-frequency sound and gravity waves (see, for example, the ICON model tutorial Section 11.3). In ICON, it is applied operationally to add analysis increments provided by the data analysis. If you define the difference between your base run fields and the fields from segment $k-1$ as the increments (or vice versa), the IAU procedure might be applicable to your problem.

We have indeed not explored alternatives to IC merging and have chosen to focus on the method we described, as it is very simple and efficient and, in our view, produces acceptable side effects. However, we were very pleased with the suggestion regarding the IAU method and will investigate this approach more thoroughly in future work.

For us, it is of paramount importance that the fields within the Lagrangian analysis region remain spatially and structurally consistent. Any discontinuities in these fields arising from alternative initialization methods could prevent us from tracking, for example, the eye of the hurricane or the development of deep convective clouds continuously over time. This possibly represents a significant difference compared to weather forecasting, where finding an optimal initial state is the primary objective. However, further insights can only be gained through a more detailed investigation of the IAU method in future work.

We have added a sentence to the outlook section of the conclusions to highlight that further research on this topic is needed.

Quantitative error analysis

The manuscript lacks a quantitative analysis of the errors introduced by the segment-based approach. Ideally, the results of the segment-based approach and the classic approach (single large LAM do-

main) should be evaluated against measurements. Only then, can the question be safely answered, as to whether the efficiency gains achieved justify the acceptance of potentially greater model errors. Admittedly, this is easier said than done.

There are many factors that affect the quality of a simulation, so agreement with observations depends on numerous factors. However, we want to emphasize that we are not proposing to use our method within an operational NWP framework, where observational agreement would become of primary concern. Instead, we believe our setup is particularly suited for basic-research questions and can be used to test very high spatial resolutions and their effects before the computational resources are available to conduct comparable studies with classical approaches.

A pragmatic alternative would be to compare the results of the segment-based approach with those of the classic approach at the same horizontal and vertical resolutions. Using the results of the classic approach as a reference would allow you to evaluate differences in noise levels, deviation of the model results, and the impact of the chosen across-track width.

We have significantly improved the analysis of errors introduced by our segmentation approach. To do this, we performed the sensitivity runs described earlier, using an overlap between two consecutive segments to more precisely characterize the behavior of the re-initialization effects. This technique is only reliable when only a short integration time has elapsed. As the simulation progresses, runs that were initially perturbed begin to diverge, and the divergence grows with time. Consequently, it quickly becomes unclear whether differences between two realizations of the same phenomenon are due to intrinsic divergence or to differences in the model settings. Our case study, the hurricane Paulette, is highly dynamic and strongly influenced by intense tropical convection. Point-wise comparisons between two realizations or difference maps are therefore inappropriate; only statistical analyses are meaningful. Robust statistics require a well-designed ensemble approach, which in turn demands a highly efficient simulation system.

Thus, as we argued earlier, a direct comparison with a single traditional limited - area domain would provide little insight, so we kindly ask to refrain from requesting such a comparison.

Lateral boundary conditions and external parameters

Lateral boundary conditions (LBCs) are an integral part of the presented workflow. However, Section 2.3 ‘Workflow Schemes and Implementation Details’ lacks a description of how LBCs and external parameter fields (e.g. topography) are treated. While the authors touch this topic in several sections of the manuscript, they provide no concise description. Specific questions:

- Are high-resolution segment-based runs at their boundaries always driven by hourly data taken from the base run, or is an additional post-processing step performed on LBCs similar to the merging step for initial conditions?

  Yes, hourly LBC are always taken from the base run and no special treatment is applied. We have added a paragraph about BC and EXTPAR handling at the end of section 2.3 and removed the discussion of BC and EXTPAR handling from section 3.3.

- Do you run the Extpar tool on the individual high-resolution segment grids, or do you interpolate the coarse external parameter fields from the base run onto the segment grids? One conse-

quence of the interpolation approach is that the high-resolution runs make use of an overly smooth topography. While I am aware that the focus of your simulations is on the ocean areas, the use of coarse external parameter fields is expected to notably impact hurricane simulations during landfall. Therefore, if this constraint exists in your toolkit, I recommend stating this clearly.

Extpar tool is always run on the individual high-resolution segment grids which yields the highest possible resolution for external parameters. We have made is aspect more clear now in the added paragraph about BC and EXTPAR handling at the end of section 2.3.

There might exist a few more shortcomings, which should be mentioned explicitly in the manuscript, if applicable:

- One possible shortcoming relates to the land and soil model. Initial condition fields containing tiled land surface information cannot be regridded to different resolutions, because the tile partitioning differs between the source and target grids. It is still possible to activate the tile approach for your segment runs, however a tile cold start will be required (i.e. initialisation from aggregated fields). The same restriction applies to the classic approach of a single large limited-area run such as your base run. However, the problem is amplified for your segment runs, as repeated initialisations are required. Hence, your segment-based approach might restrict the usage of the land and soil tile approach.

  Thank you for raising this interesting aspect. We did not apply our method for simulations over land in this study, as the focus is on hurricane-centric analysis over ocean areas. However, this is an important consideration for future applications. The limitations of re-initializing the land model at multiple segment boundaries need to be studied intensively to understand how repeated tile cold starts might accumulate errors or impact the representation of land-surface processes.

  We added a sentence in the outlook part of the conclusions section to highlight that further research on this topic is needed.

- Recently, the consequences of using inappropriate LBCs at the time of model initialization have gained interest within the ICON community. By 'inappropriate' I refer to LBCs that differ from the initial conditions read from the initial conditions file. In such cases, the model simulation is likely to suffer from model biases, proportional to the mismatch between the initial LBCs and the initial conditions. A detailed description of this problem is available on the ICON webpage `https://docs.icon-model.org/documentation/buildrun/buildrun_input_data.html`. It is likely that your model runs are affected by this problem. The proposed solution is to overwrite the initial LBCs with the initial conditions, by activating the namelist switch `init_latbc_from_fg=.TRUE.` (default FALSE). If this switch was deactivated in your model runs, I recommend repeating a subset of your runs to check whether this has any impact on the model results presented in your manuscript.

  Thank you for this interesting comment. It impacted our revision activities a lot. We have conducted a sensitivity experiment with the proposed LBC flag activated and incorporated the results into the spin-up analyses in the revised manuscript. This now allows readers to see what changes and/or improvements can be expected from this setting.

Please also find more answers to your questions regarding the LBC flag in the response to your major point on the "Initialization of the high-resolution segment-based runs".

Minor points

1. Line 44: '... and the influence of larger scale processes is not captured adequately.' I would merely argue that the influence of small- to regional-scale processes on large-scale processes is not adequately captured by design. Are the large-scale processes not accounted for by the lateral boundary conditions?

   We have rephrased the sentences accordingly.

2. Optional suggestion: Line 88: '... two one-way nested domains with grid spacings of 5 and 2.5 km are employed ...' The wording 'two one-way nested domains' might be a bit confusing for readers from the ICON community. If I understand correctly, your base setup consists of a regional domain and one nested domain (one-way nested). In particular, I propose to avoid using the word 'nest' for the base domain (DOM01), and to avoid counting it as a nest. The base domain can be a global or regional domain, which may be equipped with one or more nested domains. You might consider adapting the wording in the text (e.g. L153, L157, L205, etc.).

   We have adapted the wording at (former) lines 88 and 205 accordingly. These references apply to the base simulation.

   For the high-resolution segment runs, we have retained the original wording, as we consider the outer domain to be effectively an offline nest relative to the base run, and thus the terminology "nested domain" remains appropriate in this context.

3. Line 119: In your example the distance of the track segments A and D in terms of time is $\Delta t_{\mathrm{reinit}} + 2\Delta t_{\mathrm{add}} = 18\,\mathrm{h}$. However, in the text you state that it corresponds to the distance traveled by the hurricane, 'e.g. from 18 UTC the previous day to 15 UTC', which corresponds to a time interval of 21 h $\neq$ 18 h. Is there a typo in the text, or am I missing something?

   Thank you for pointing this out. We are lucky that the reviewer noticed this inconsistency. The interval of 18 hours is indeed correct, and right starting time would be 21 UTC.

   The text has been corrected to reflect this.

4. IC generation Line 159–166: This paragraph might benefit from some rephrasing, in order to point out more clearly that

   - CDO is used for horizontal regridding only (no vertical regridding)
   - vertical regridding (interpolation) is performed by the ICON test run, which reads the horizontally interpolated ICs, and outputs the vertically interpolated IC to file right before the model performs the first integration step (time step zero).

   Vertical regridding is mentioned again in Section 3.3 (L275–278). I suggest to discuss it only once.

   Thank you for the suggestion. We did the follwoing changes in the manuscript:

   - We have added "horizontal" to clarify that CDO is used for horizontal regridding only.

- We have removed the discussion of vertical regridding and only mention the number of vertical levels in the context of the ICON high resolution setup.
- We have added a paragraph about BC and EXTPAR handling at the end of section 2.3 and removed the discussion of BC and EXTPAR handling from section 3.3.

5. In Section 3.3 'Exploration of Different Refinement Strategies' the resolution of your grid segments (DOM01) is R2B11, which is exactly twice the resolution of DOM02 of your base run (R2B10). I assume that you are not constrained to a factor 2 in resolution jump and that it is possible to select arbitrary horizontal resolutions for DOM01 of the grid segments (within physically reasonable bounds)? In particular, I assume it is equally possible to choose R3B10 or R3B11 for DOM01 of your grid segments in this example? Is my assumption correct? Your statement 'In principle, it is also possible to start initial refinement from grids finer than the base grid' (L256) seems to point towards this direction, but it is not fully clear to me whether there exist technical limitations in your workflow or not.

   A similar question was asked by reviewer #1. We repeat our answer here:

   Yes, this functionality is in principle possible. However, the current implementation misses the distinction between the grid of the base run and the grid from which refinement is performed. This means that currently the refinement is always performed from the base run grid, by halving the grid spacing. The addition of a second grid for refinement can easily be implemented and left for future extensions.

   We have added a sentence in the outlook part of the conclusions section to highlight that further research on this topic is needed.

6. Figure 7b: Can you explain why the minimum surface pressure seems to undershoot when switching to the next grid segment? I would rather expect the minimum surface pressure to increase, as a consequence of regridding and weighted merging. Both, regridding and merging, will likely diffuse gradients in the prognostic fields which I expect to result in increased minimum surface pressure values.

   We can only provide a speculative explanation for the surface pressure undershoots. We suspect that the undershoots are caused by the re-initialization of segments with merged initial conditions. The temporal behavior of minimum surface pressure is also detailed in Figures 10 and 11. These figures show that the undershoots occur for each subsequent segment and are present across different domain configurations. This is very likely a dynamic adjustment process in which differences between the merged initial conditions are equilibrated through the radiation of atmospheric waves.

7. Line 327: '. . . the 12h-reinit setup is significantly more favorable in terms of resource requirements . . . '. Could you please explain why the 12h-reinit setup requires less computing time than the 24h-reinit setup (see Table 2)? Assuming an identical across-track width, switching from 24h to 12h will result in a domain size that is half the original size. However, the number of grid segments required will double. So, where does the observed speedup for the 12h reinitialisation interval come from?

This is likely a misunderstanding. We apologize for the confusion; the text in the manuscript may not have conveyed our intended meaning clearly. We added text to clarify this point.

Both setups, the 12h-reinit and 24h-reinit setup, do not differ in the number of time steps .They are identical for both configurations. However, because the 12h-reinit setup is always made off significantly smaller domains than the 24h-reinit setup, the computing time for the 12h-reinit setup is correspondingly smaller.

Editorial

1. Figure 1: Gray arrows are a bit hard to see in print form.

   We changed to a slightly darker gray tone for the arrows in Figure 1 to improve visibility in print form.

2. Table 1: Strictly speaking the column ‘reinit’ does not fit, as its content clashes with the labeling of the rows (similar for Table 2). It seems like a pragmatic approach to keep the table compact and, unfortunately, I do not have a better solution.

   Thank you for this feedback. We have attempted to improve the readability of the tables by adding vertical spacing and lines. We will work with the professional typesetting team to ensure that the table design is optimized during the preparation for publication.

3. Figure 7: ‘The minimum pressure . . . as in (b) and (c)’. ‘as in (b) and (c)’ does not make sense to me. Do you mean ‘as in (a)’?

   Yes, thanks! We have corrected the text to read "as in (a)".

4. There is some inconsistency in the way units are defined in the figures. Axis annotations sometimes use square brackets, while at other places a slash is used followed by units that are sometimes enclosed in round brackets and sometimes not. I suggest the consistent use of one style, such as square brackets for example.

   Thank you for this comment. Since only dimensionless quantities can be plotted as numbers, the physical quantity must first be divided by the corresponding unit. To follow this principle, we consistently use a slash in conjunction with the unit. We have additionally added round brackets around the unit everywhere to improve readability. We have applied this consistent formatting to all figures.

5. Line 313: ‘Most of the ice produced is located northwest of the hurricane center.’ Isn’t most of the ice located northeast rather than northwest in Figure 9?

   Thank you very much for catching this error! We have corrected the text to specify the correct direction.