Author comments in reply to anonymous referee #1

We thank anonymous referee #1 for their useful comments and the time invested in reviewing our manuscript. We very much appreciate the critical and thorough review of our work. We have addressed each of the referee comments as detailed point by point below, which we believe has significantly improved the quality of the manuscript. We hope our revised manuscript can be accepted for publication.

Adithya Vemuri, on behalf of all co-authors

Overarching comments

To fully address the major comment#2, we have added 2 additional extreme weather events to this sensitivity analysis. In the revised manuscript all extreme events observe fast changes in wind direction accompanied by severe yaw misalignment leading to potentially harmful conditions for wind farm operation. Furthermore, we have emphasized statistical significance of the differences between simulation setups and have therefore refocused the analysis on the quantitative MAE analysis of wind speed and wind direction. As a result, the revised manuscript observes significant changes to its structure, simulation setups, simulations pairs and the evaluation methodology.

Further, we mention that during the revision, we have checked and refined our post-processing tools, which led to some minor changes in metric values and time-series plots for the case of Storm Ciara.

Major comments

Comment 1: *A thorough technical English edit is required. There are numerous issues throughout the manuscript, most of which are specifically mentioned in the Minor Comments from the Abstract through Section 2 as examples of the issues that need fixing. I only mentioned a couple items from Sections 3–5.*

We thank the referee for pointing this out. We have spent significant attention to improving the overall quality of the English language with a native English speaker, both for the specific points raised by the referee, and for the manuscript in general. This is further detailed in our reply to the minor comments below.

Comment 2: *Lines 95–98 provide the key to defining the niche that this article aims to fill. I think you need to do a better job honing in and repeatedly showing how your work fills that niche. I also think that validating against data at only a single point for a single storm is inadequate to fill that niche. That is really the biggest fundamental issue I have with this manuscript, and why I gave it a Reject instead of a Major Revision. A single case study can have value if you do more than validate against observations only at a single point, while validating at a single point can have some value if you evaluate multiple cases. Are there any other buoys or towers that are available for offshore wind validation? I am aware of FINO1 in the North Sea region, but I believe it is outside your d04, unfortunately.*

We thank the referee for this valid comment. We address it in 2 ways as shown below. Modifications to the revised manuscript are highlighted in blue.

- Lines 95-98 indeed sparingly address one of the niche areas this paper aims to fill, on the applicability of Shin-Hong PBL scheme to coastal environments. Lines 38-45 establish the fundamental motivation and novelty behind this study, i.e., to perform a sensitivity analysis on

WRF physics parameterizations for the Belgian offshore wind farms from an operational SCADA point of view.

In the introduction of the revised manuscript, we emphasize the motivation in a better way:

Line 41: Sensitivity analyses are typically conducted to identify the optimal combination of physics schemes for a specific location (see, e.g., Efstathiou et al. 2013; Santos-Alamillos et al. 2013; Kala et al. 2015). This type of investigation has not been performed for the Belgian North Sea. Furthermore, to the authors' best knowledge, no previous studies have looked at potentially harmful EWE from a wind farm perspective as experienced by the machines themselves. Therefore, this sensitivity analysis aims to address this gap in research. The analysis presented in this paper assesses the impact of a wide range of physics parameterizations for PBL, cumulus and microphysics, and length of the update interval of LBC (lateral boundary conditions) on the simulated wind direction and speed.

This is again highlighted in the methodology section:

Line 170: This evaluation uses operational wind farm SCADA data for its quantitative analysis of wind direction and speed. Additionally, radar data from RMI-B allows for a qualitative perspective on precipitation. By combining these observational datasets, the premise of this study provides a unique opportunity to investigate EWE as experienced by an offshore wind farm to determine suitable WRF setups in the specific context of wind energy applications.

- As rightfully mentioned, d04 in our simulations does not encompass many public-domain observational datasets suitable for offshore wind validation. However, we agree that extending the analysis to multiple cases significantly enhances the quality of our manuscript. Therefore, to address the second part of this comment we include 2 additional extreme weather events that observe fast changes in wind direction accompanied by severe yaw misalignment. Furthermore, we have also opted for including statistical uncertainties in a more objective and quantitative analysis to assess which trends persist with statistical significance over different events. In this regard, we have refocused our analysis on quantitative comparison of the wind direction and speed MAE. We have removed quantitative discussions on precipitation (see also comment 3) and Kantorovich distances (for which we cannot define any error bars based on the available data). Furthermore, we have removed the somewhat subjective domain configuration sensitivity (see also comment 4d) in favor of an additional case with the 3D Zhang PBL scheme. As a result, we still define 12 runs per event (shown in table below), or a total of 36 WRF configurations for the 3 events combined.

<table>
<tr><th>Simulation run#</th><th>ERA5 LBC updates</th><th>PBL scheme</th><th>Cumulus scheme</th><th>Microphysics scheme</th><th>Update interval pairs</th><th>PBL pairs</th><th>Cumulus pairs</th><th>Microphysics pairs</th></tr>
<tr><td>1</td><td>3 h</td><td>MYNN</td><td>msKF</td><td>WSM5</td><td>A</td><td></td><td></td><td></td></tr>
<tr><td>2</td><td>3 h</td><td>Shin</td><td>KF</td><td>WSM5</td><td>B</td><td></td><td></td><td></td></tr>
<tr><td>3</td><td>1 h</td><td>MYN</td><td>KF</td><td>Thompson</td><td></td><td>C</td><td></td><td></td></tr>
<tr><td>4</td><td>1 h</td><td>MYNN</td><td>msKF</td><td>WSM5</td><td>A</td><td>D</td><td></td><td></td></tr>
<tr><td>5</td><td>1 h</td><td>Shin</td><td>KF</td><td>WSM5</td><td>B</td><td></td><td>F</td><td rowspan="2">I</td></tr>
<tr><td>6</td><td>1 h</td><td>Shin</td><td>KF</td><td>Thompson</td><td></td><td>C</td><td>G</td></tr>
<tr><td>7</td><td>1 h</td><td>Shin</td><td>msKF</td><td>WSM5</td><td></td><td>D</td><td>F</td><td rowspan="3">J</td></tr>
<tr><td>8</td><td>1 h</td><td>Shin</td><td>msKF</td><td>Thompson</td><td></td><td></td><td>G</td></tr>
<tr><td>9</td><td>1 h</td><td>Shin</td><td>msKF</td><td>Morrison</td><td></td><td></td><td rowspan="2">H</td></tr>
<tr><td>10</td><td>1 h</td><td>Shin</td><td>GD-3D</td><td>Morrison</td><td></td><td></td><td></td></tr>
</table>

11	1 h	Shin	GF	Morrison		E		
12	1 h	Zhang	GF	Morrison				
13	Ensemble average							

Comment 3: *Validating precipitation at only a single point is of limited use, even over many cases, when your goal is to determine which model configuration gave the most realistic simulation of precipitation. If you want to validate model precipitation or reflectivity, then you should leverage the land-based radar that you do have data from and do object-based validation with MODE for a more comprehensive validation.*

We thank the reviewer for this useful comment. We agree that single-point validation is of limited use, and that a MODE met-plus analysis tool would allow for a more comprehensive quantitative precipitation study. However, we are currently limited by data access to the land-based radar, with access to only the raw reflectivity radar data which, in itself, is not currently publicly available. We also lack the information required to post-process reflectivities into a form that could be quantitatively compared to WRF output, or to extract rainfall at ground level from this data as described, e.g., in Ref [1] below. Therefore, we omit any further quantitative evaluation of modeled precipitation in the revised manuscript. However, since precipitation plays a major role in wind energy and extreme weather events, we briefly mention the sensitivity and variability of precipitation rates produced by the model in the discussion section. Quantitative assessment of precipitation is mentioned as an area of future work in the conclusions section.

[1] Goudenhoofdt et al. 2016, Generation and Verification of Rainfall Estimates from 10-Yr Volumetric Weather Radar Measurements in: Journal of Hydrometeorology Volume 17 Issue 4 (2016)

The following lines have been included in the discussion section:

> Line 401: A qualitative perspective on precipitation indicates all WRF simulations to be highly sensitive to the combination of physics schemes and type of EWE. The qualitative analysis yielded little to no conclusions on the precipitation modelling fidelity of the considered WRF physics setups in this study. As an example, the results for the case of Storm Ciara are presented in Fig. 19 (in the revised manuscript). A direct quantitative comparison of simulated reflectivity and observed raw radar fields using, e.g., tools for comparing gridded observations such as MODE [2] is impeded by the lack of filtering and post-processing information on the latter raw data. Therefore, a quantitative assessment of precipitation modeling is out of scope of the current paper and left for future work.
>
> [2] Newman, K., J. Opatz, T., Jensen, J., Prestopnik, H., Soh, L., Goodrich, B., Brown, R. B., and Gotway, J. H.: MET-MODE, in: The MET Version 10.1.0 User's Guide, DTC, 2022

Furthermore, we explicitly mention the radar data is not publicly available in the description of the events section:

> Line 137: Three case studies are considered in this sensitivity analysis, namely, Storm Ciara on 10 February 2020, a low-pressure system on 24 December 2020, and a trough passage on 27 June 2020. The radar data presented therein is not publicly available, but was retrieved through a bilateral agreement with the Royal Meteorological Institute of Belgium (RMI-B). A brief synopsis of these events is presented in the following sub-sections.

Comment 4: *Updates on figures*

a) *Fig. 1 (several of these comments also apply to Figs. 6, 8, 10, and 11): It is customary to plot coastlines or national borders in black or gray on most maps. Using a color from your colorbar (red) is simply confusing. Also, your filled contour colors do not match the colorbar. The thin ribbons of darker colors around fields of pastel colors also make this figure difficult to interpret with any confidence. Additionally, in this figure the colorbar label says "Precipitation" but has units of mm/h. Precipitation would have units of mm, but mm/h are units for precipitation rate. The caption also states that the figure is depicting radar reflectivity, which again is not quite the same thing as precipitation rate, though they of course are related to one another. The caption also states that the observed radar reflectivity (or really, radar-derived precipitation rate) is valid at 04:00, but line 137 says it is valid at 04:40. Which is it?*

- The precipitation plots are changed to address these comments. An example plot format for the change in cumulus parametrizations is presented below. Thank you for pointing out discrepancies in the definition of precipitation units, we have updated the figure caption to reflect the correct unit of precipitation rate as mm/h. The time stamp for the case of Storm Ciara is at 04:40 UTC, also has been updated in the revised manuscript.

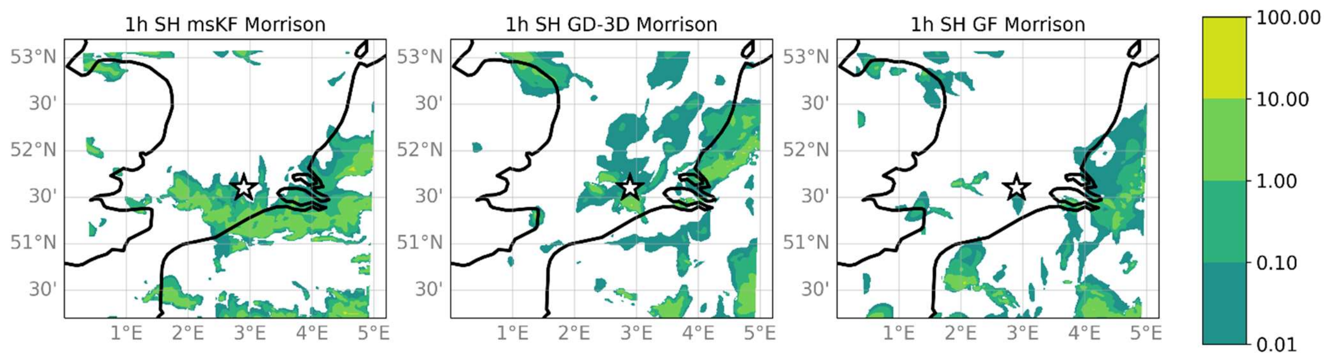
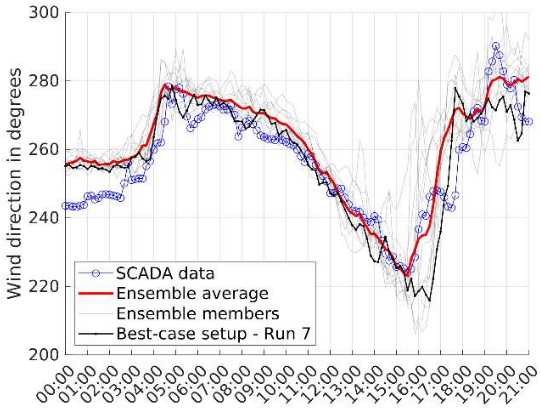


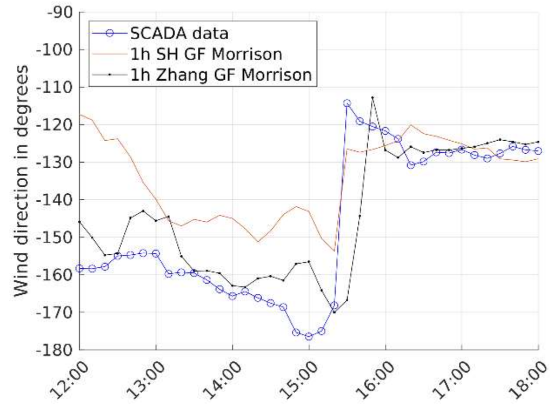
Fig. 19 caption: Contours of WRF precipitation rate in mm h^{-1} for the case of Storm Ciara on 10 February 2020 at 04:40 UTC. The plots are presented for cumulus simulation pair H for domain d04.

b) *Figs. 3 and 4: You should have a thin gray line in your legend if it is in your plot. In the x-axis label (also in Figs. 5, 7, and 9), also use a date format like 10 Feb 2020. 10/2/2020 will easily confuse American readers into thinking the date is 2 Oct 2020.*

c) *Fig. 5: The orange dashed line is quite faint and difficult to see.*

- We agree the date format can be confusing to readers of different region. Therefore, we have adopted the unambiguous date format as, e.g., 10 February 2020. The timeseries plots have also been updated to address the formatting comments while keeping in mind the color-blind readability. Example plots are presented below.





d) *Fig. 6 (most of these comments also apply to Figs. 8, 10, and 11): First, calling these Domain 1 and Domain 2 is misleading. These are really Domain Configurations 1 and 2; both these domain configurations have domains 1–4, so when you say domain 1 or domain 2, my mind automatically thinks of the outermost two WRF domains. Second, the radar contour lines all look the same color, which seems like a mistake. Third, restrict your WRF filled contour range to the equivalent of the radar derived precipitation rate, or at least restrict the lower bound cutoff to something like 0.01 mm/h. Is it really raining at lower values than that, anyway? Values below your lower bound should be transparent/not plotted. That will also solve the undesirable issue of the entire domain being filled with a dark blue that makes other features difficult to discern while also being meaningless. Fourth, state in the caption what the star is.*

- Although only a single domain configuration is considered in the revised manuscript, the comment on addressing the right terminology for domain configuration is still relevant to the revised manuscript. We update the captions of WRF domain configuration and nested domains throughout the paper.

  The following changes are made to the plot caption of the WRF domain configuration:

  Fig. 4: WRF and WRF Post-processing System (WPS) nested domain configuration (1-way nesting) considered common to all simulation runs in this study.

- In view of the reply to comment 3, we omit the radar contours from WRF precipitation rate plots, the updated plots are presented above (Fig. 19).

- Indeed, the precipitation rate from WRF simulations that was being plotted was lower than 0.01mm/h. Therefore, we update these plots considering precipitation rate values lower than 0.01mm/h to 0. Furthermore, we have amended the useful suggestions in the revised figures. These updated plots are shown above in the reply to Comment 4a.

Comment 5: *To be completely honest, changes in wind direction of 40° do not seem like a huge shift—it is not even half of a quadrant. If 40° is a hugely consequential shift that wind farm operators need to be quite concerned about, then it would be helpful to provide some justification.*

- The events featured in this study are selected based on specific alarms raised by individual turbines during fast changes in wind direction and severe yaw misalignment. However, bounds on such yaw misalignment and thresholds for raising alarms are specific to turbine models and wind farm operators. As this data is highly confidential, we cannot further elaborate in detail on this in the manuscript. However, we do understand the comment by the referee, and address this by elaborating on the methodology used for the identification of the extreme weather events.

  The following lines are added to the methodology section:

  Line 128: The selection of the events in this study is motivated by the occurrence of fast changes in wind direction accompanied by severe yaw misalignment leading to significant power loss as observed by a Belgian offshore wind farm in the North Sea. The methodology utilized to identify these events modifies the approach defined by Hannesdóttir and Kelly (2019) to include yaw misalignment. The wavelet analysis considers a minimum threshold to identify anomalous changes in wind direction accompanied by severe yaw misalignment experienced by several wind turbines. Severe yaw misalignment potentially has adverse effects on the operational lifetime and fatigue loading of a wind turbine (Wan et al., 2015; Bakhshi and Sandborn, 2016; Laino and Hansen, 1998; Damiani et al., 2018), highlighting its importance and relevance in this study. The SCADA analysis for the identification of these events includes confidential error codes and data that are protected under a non-disclosure agreement, therefore no further details can be provided herein.

Comment 6: *Table 2: I suggest either reordering pairs A–K based on the order they are discussed in Sections 4.1–4.4, reordering Sections 4.1–4.4, or both (my preference). It makes more sense to me to look first at the size of the domain and the lateral boundary condition temporal frequency, before then comparing different physics schemes. Also, in Table 2 ensure that cell borders are turned on to separate the different experiment pairs in the cumulus pair column. Additionally, in the section titles for Sections 4.1–4.4, it would be helpful to include the experiment pair letters.*

- We thank the referee for this insight and realize this could be confusing to the reader. The revised paper structure is re-organized to make the readability easier. The PDF was rendering at lower resolution, making the cell borders invisible in some places. This is improved in the revised manuscript, an example table for Storm Ciara is presented below (Table 3 in the revised manuscript).

Comment 7: *Table 7: In the Average NED column for rows 10 and 11, you have 1.10 in green, 1.111 in yellow, and 1.32 in red. The values 1.10 and 1.111 are so close that it is misleading to make them such different colors. Is the difference between 1.10 and 1.111 in this metric even meaningful? What would be a meaningful difference in NED or Kantorovich distance? In any case, in Tables 4–7, I really think you would be better off keeping the color scale and ranges from Table 3.*

- We agree with this comment and have categorized the values into 5 colors to avoid confusion.

  Following lines were added to the methodology section to highlight the same:

Line 229: Cells are colored based on a set of 5 categories between red and green. Categories are defined to cover 20% of the range between smallest and largest values for the considered metric. In this way, results are categorized into best (green, with errors in the 20% lowest range), good (light green), average (yellow), poor (light red), and worst (dark red).

The performance evaluation tables are updated to reflect the same, as an example the performance evaluation table (Table 3 in the revised manuscript) for the case of Storm Ciara is presented here.

Simulation run#	ERA LBC updates	PBL scheme	Cumulus scheme	Microphysics scheme	Wind direction MAE (degrees)	Wind speed MAE (ms^{-1})	NED (-)
1	3 h	MYNN	msKF	WSM5	10.46	3.88	2.08
2	3 h	Shin	KF	WSM5	8.48	2.57	1.51
3	1 h	MYNN	KF	Thompson	9.26	2.72	1.63
4	1 h	MYNN	msKF	WSM5	8.61	2.54	1.51
5	1 h	Shin	KF	WSM5	7.68	2.47	1.41
6	1 h	Shin	KF	Thompson	8.37	2.51	1.48
7	1 h	Shin	msKF	WSM5	6.59	1.78	1.11
8	1 h	Shin	msKF	Thompson	6.69	1.89	1.15
9	1 h	Shin	msKF	Morrison	7.17	1.89	1.20
10	1 h	Shin	GD-3D	Morrison	5.59	2.25	1.17
11	1 h	Shin	GF	Morrison	7.17	2.67	1.43
12	1 h	Zhang	GF	Morrison	8.69	1.84	1.34
13	Ensemble average				5.88	2.04	1.12

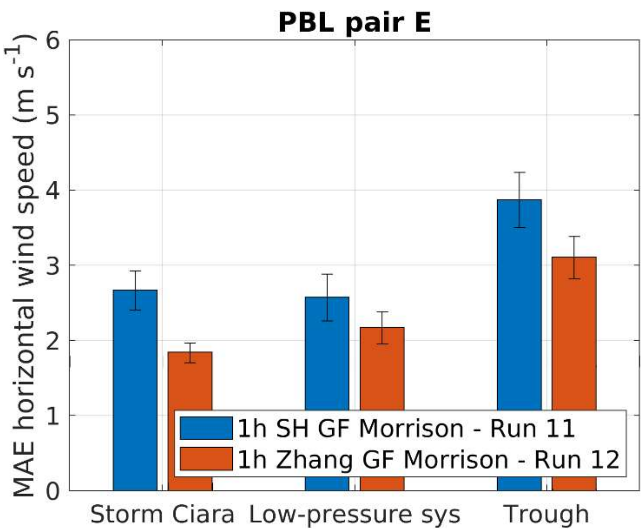
Comment 8: *Lines 325–327: First, change "ensemble" to "ensemble mean" or "ensemble average". Second (and more importantly), there are many papers and books that explain why the ensemble mean usually outperforms individual ensemble members (e.g., Wilks 2019, https://www.elsevier.com/books/statistical-methods-in-the-atmosphericsciences/ wilks/978-0-12-815823-4). It would be worthwhile to engage with some of that literature here, especially since your findings of the ensemble mean not being the best are contrary to what was expected. Do you have any insights as to why the ensemble mean performs comparatively poorly in the Kantorovich distance for wind speed and wind direction? This appears to be why the NED is not the best for the ensemble mean for the wind variables. Perhaps this is a side effect of the randomness introduced by having a sample size of only one event validated at only one point?*

- We thank the referee for pointing out a possible source of confusion regarding the ensemble average. We have updated the performance tables and timeseries plots (as depicted above) to state ensemble average instead of just ensemble. The definition of ensemble average in this study differs from a more traditional ensemble forecast, as defined in Wilks 2019, in that our ensemble average only includes the variability in physics scheme and the change in temporal resolution of boundary conditions, and hence does not reflect any uncertainties on initial conditions. We reflect on the literature and differentiate the ensemble average considered in this study in the methodology section.

  Following lines have been added to the methodology section to highlight the above point:

Line 217: This study also evaluates the performance of an ensemble average compared to single deterministic simulation runs. The ensemble average is defined as the mean of all simulation runs considered for a given case study. In this study, ensemble members are initialized with identical initial conditions from ERA5 reanalysis. Subsequently, variability in the ensemble average is only caused by the variation in update interval of LBC and physics parameterizations. Therein, the current definition of ensemble average differs from traditional ensemble forecasts, where variations in initial conditions are also considered, see, e.g., Wilks (2019).

- As mentioned above, with the addition of 2 more events to the revised manuscript, we have included a statistical comparison to quantify uncertainties on the MAE estimates and assess which trends persist with statistical significance over different events. An example plot (Figure 12b in the revised manuscript) is presented below indicating the change in performance evaluation for simulation pairs.



Minor comments

We thank the referee for their critical review on the manuscript's written English. The typos, comma style, abbreviation definitions and technical detail have been corrected in the revised manuscript. In addition to the specific points raised by the referee, we have thoroughly revised the use of English throughout the manuscript with the help of a native English speaker.

We have amended all the textual changes suggested by the referee. For brevity, we only mention the ones that require further explanation.

- *Throughout: Whether you use Oxford commas or not, the journal guidelines state that you need to be consistent, but there is not consistency of usage in your article.*

We have consistently made use of Oxford commas in the revised manuscript.

- *Throughout: Provide the time zone for all times in this article. I presume your times are in UTC, but it is never stated.*

Indeed, time zones are UTC, we mention this throughout the revised manuscript.

- *Line 7: It would be nice to mention in the abstract when Storm Ciara occurred.*

The dates of the events considered are now included in the abstract.

- *Line 30: Change "open-source" to "public domain". Also, since you use WRF v4.2, you should instead cite the WRF v4 technical note (Skamarock et al. 2019). Consider also citing Powers et al. 2017.*

We thank the referee for pointing this out and included both references.

- *Line 111: "concluded in no particular combination of WRF physics" — This is awkwardly worded. Please revise.*

The sentence has been revised to:

Line 114: In contrast, the study by Islam et al. (2015) for the Haiyan tropical cyclone over the west Pacific Ocean did not identify a suitable combination of WRF physics to best reproduce the extreme weather event.

- *Line 114: "time-lapse considered within the diurnal cycle" — This is awkwardly worded. Please revise.*

The sentence has been revised to:

Line 116: … analyses indicating a wide array of possible combinations of physics parameterizations depending on the type of weather phenomenon, the season and the time period considered for simulation within the diurnal cycle.

- *Line 124: "exposed" is an odd word choice here.*

The sentence has been revised to:

Line 125: Lastly, conclusions and future prospects are presented in Sect. 5.

- *Line 132: Please define what you mean by "over the local region". How local? D04 is not very big to begin with.*

Indeed, d04 is too small to capture all affected areas in Belgium by Storm Ciara. The areas affected as reported by RMI-B (RMI - Storm Ciara (meteo.be)) extend far in-land and offshore. The sentence is revised to:

Line 146: … the Royal Meteorological Institute - Belgium (RMI-B) reported wind gusts of up to 115 km h^{-1} in Ostend, located at the Belgian coast, with heavy precipitation accompanied by strong winds and thunderstorms.

- *All throughout Section 4: I think you need to consistently refer to domain configuration 1 and 2, as each domain configuration has domains 1, 2, 3, and 4. If you say domain 1, the reader will think of your outermost WRF domain, which is domain 1 (d01).*

We agree with this comment and have amended it (see also reply to comment 4d)

- *Line 322: This sentence is awkwardly worded.*

The sentence is revised to:

Line 401: A qualitative perspective on precipitation indicates all WRF simulations to be highly sensitive to the combination of physics schemes and type of EWE. The qualitative analysis yielded little to no conclusions on the precipitation modelling fidelity of the considered WRF physics setups in this study.