Review of WCD-2022-6, "The Cyclogenesis Butterfly: Uncertainty growth and forecast reliability during extratropical cyclogenesis" by Rodwell and Wernli

I thank the authors for their adjustments to the manuscript and responses to my recommendations concerning the initial submission of this work. I particularly appreciate the reduction in the number of sections in the manuscript, which has helped to improve its readability. A clear statement of the objectives of the work will further help to motivate the reader and will provide a useful "point of contact" between the otherwise disparate elements of the study.

The revised text still suffers from a lack of clear organization, which will make some of the most interesting results of the study difficult for future readers to access. Centralizing data and methodological descriptions in section 2 will avoid many of the current disruptions to the flow of the manuscript. It will also help to keep readers focused on the scientific contributions contained within the impressive amount of presented work.

I hope that these notes will provide some useful suggestions for this submission.

Recommendation: Major Revision

Reviewer: Ron McTaggart-Cowan

General Comments

1. It is unclear to me what was done to make the objectives of the study clearer in the introduction (response to General Comment #1 of the initial review). The closest thing that I can find to such a statement is the phrase that "This study focuses on uncertainty growth in the North American / North Atlantic / European region, and particularly the North Atlantic winter stormtrack (sic), with its embedded cyclogenesis events and other synoptic systems." However, this sentence does not explain what will be achieved by this "focus". Please include a clear thesis statement to help readers to understand what the intended outcome of the study is.
2. Although I appreciate the added discussion and authors' responses, I still think that invoking the "butterfly effect" is a misnomer. Based on the title, I would expect a paper about how small-scale and small-amplitude perturbations affect cyclogenesis. A title that is more descriptive – albeit less spectacular – would serve the content better. Maybe something like, "The impact of North Atlantic winter cyclones on uncertainty growth and forecast reliability in ensemble guidance".
3. I do not think that the strategy of "just in time" methodological description is effective or that it improves the readability of the text by motivating the reader. On the contrary, the decentralized methodology segments disrupt the flow of the text. Moreover, they are difficult to locate for readers that are not progressing linearly through the text and/or readers that wish to refer back to methodological descriptions at a later time. Please seriously consider introducing all relevant methods in section 2 of the manuscript.
4. The comparison of spread growth rates in select TIGGE models is interesting, particularly because of the wide range of patterns shown in Fig. 3. However, the follow-up on this analysis

lacks sufficient rigor to make it as useful as possible for future readers. It would be very interesting to know the growth rates of some systems differ systematically from others, for example. Imagine adapting the anomaly correlation score using the LGR from one model at a time as the "analysis anomaly" over the North Atlantic. For example, the LGR from each TIGGE model (i.e. the "forecast anomaly") could be compared to the ECMWF patterns: an ACC would be computed for JMA, NCEP and UKMO. Then each model could be compared to the UKMO patterns for another set of scores: JMA and NCEP (the ECMWF score already being known). Et cetera. In the end, symmetric matrix of ACC scores would be obtained, and could be presented as an effective synthesis for this component of the analysis. The 95th percentiles (or smaller, given the small number of cases) of the ACC scores could be used as a measure of the variability around the mean ACC score. Noting what the ACC score is for Fig. 3 would provide a quantification of the extent to which the case study aligns with the "typical" degree of agreement between LGR in the TIGGE systems.

5. Although much improved from the initial submission, the structure of the manuscript continues to present a challenge for readers. Aside from the need for a centralized methodology section (General Comment #3), a specific example arises at the end of section 3.3. The section was interesting, and ends with two interesting questions. If they are anything like me, the reader will be looking forward to diving into these questions. However, the section 4 introduction, and methodology introductions sections 4.1 and 4.2 mean that they will have to "hold that thought" for ~100 lines of text before they get to further discussions on these questions. By then, the reader will have forgotten the specifics of the questions or why they were interesting. If a review of reliability is required, it should appear either in the introduction or in section 2. Likewise, the complicated descriptions in sections 4.1 and 4.2 should appear in section 2. This reorganization will mean that the reader's momentum can be maintained as they progress through the results and synthesis.
6. How much of section 4.1 could be replaced by a reference to section 3 of Rodwell et al. (2015), but with "observation" (in that study) replaced by "analysis" (here)? The overlap is mentioned explicitly beginning on line 324, but a full replacement (and associated simplification of the current text) does not seem to have been considered: please consider it.
7. The term "cyclogenesis" appears to be used primarily to refer to the presence of a cyclone. This is important because the "cyclogenesis butterfly", based on a standard definition of cyclogenesis, implies uncertainty introduced by a cyclone is forming or deepening. However, the "cyclogenesis" clusters 1 and 2 (Fig. 7) only assess of the presence of a cyclone: they contain no direct information about whether the cyclone is intensifying or decaying (the westward tilt with height is not a guarantee of surface intensification). I understand that cyclones often deepen in this region; however, this makes the link to cyclogenesis anecdotal rather than data-driven. The "winding back" process (a term that should be clearly defined) appears to be an attempt to build in a cyclogenesis period. However, if I understand the procedure correctly then a cyclone moving into the defined area will be defined as "cyclogenesis", even if it has already reached its peak intensity. Alberta clippers, for example, reach peak intensity shortly after formation and slowly weaken thereafter as they move towards the region of interest for this study (Blaine and Martin 2007). Changing from "cyclogenesis" perspective to one that documents ensemble behaviour in the presence of a cyclone would not weaken the work, and

would better describe the analysis. The recommended title (General Comment #2) reflects this change in perspective.

8. Excessive spread in the storm track during cyclone passage is labelled as a "key conclusion of this study" (line 438). This conclusion appears to be based on Fig. 8o, which shows positive but non-significant differences between the composite residuals. If that is correct, then assertions of cyclone-related "over-spread" should be moderated in the text. Given the potential for type-I errors related to multiple-testing (Wilks 2016; BAMS) and an experimental design that does not sample interannual variability, the true significance of these differences is questionable.
9. Figure captions are not the appropriate place for methodological descriptions. Although figure-specific details might be provided in captions (specific threshold values for example), complete methodological descriptions should appear in the main body of the text where it can be easily found by future readers. Please move all methodological descriptions from captions to section 2 of the document.
10. Section 3.2 should be replaced with a brief description of the Lagrangian growth rate in section 2, including a reference to Rodwell et al. (2018). The derivation and extensive discussion of terms that will not be employed further in the analysis does a disservice to the current study by introducing unnecessary complexity. If the rhs of Eq. 3 will be useful in a future study, then it should be presented in the future study. The discussion section of this work could easily refer to a hypothetical expansion of the Lagrangian growth rate rather than specific equations that disrupt the flow of the text.
11. I understand that decisions related to writing style are typically left to the author; however, the over-use of em dashes disrupts the flow of the text and reduces its readability (there are seven in the introduction alone). Please consider rewriting the majority of phrases that currently use this form of subordination.
12. Single and double quotes are used liberally throughout the text; however it is unclear what they mean and how the authors choose between them in any given circumstance. Please consider removing the majority of these quotation symbols and/or provide a description of what they represent.

Specific Comments

1. [L19] Consider rewording split infinitive.
2. [L23] It isn't "NWP" itself that develops techniques, but researchers and system developers.
3. [L26] I believe that "leadtime" is usually written as "lead time".
4. [L28] I believe that "Stormtrack" is an application while, "storm track" is the usual term for the region discussed in this study.
5. [L32] Is "propone" the word that you mean to use here? Consider replacing with "prone" or "conducive".
6. [L35] Why is "blocking" (well-accepted terminology) enclosed in single quotes?
7. [L49] I think that a comma before the quoted question would be appropriate.
8. [L50] The term "reliability" has already been introduced with single quotes: consider removing them here for readability (the citations make it clear that this is a technical term).
9. [L53] Why does the bias problem apply only to short-range assessments of reliability as implied here?

10. [L58-59] This phrase suggests that improvements to the model and MU will not improve reliability in the presence of SV perturbations. It that guaranteed to be true? If the SV perturbations are scaled to become arbitrarily small, then they will presumably have a negligible impact on the forecast and model improvements will become dominant. This general statement might need either to be qualified or to be removed.
11. [L59] What does the term "the potential is raised" mean? Does this refer to an increase in potential, or to a subject that is raised later in the text. Please consider using clearer terminology.
12. [L63-65] This appears to be a run-on sentence: please rephrase.
13. [L69] Why is "Ensemble" capitalized here?
14. [L78] This is a highly condensed system description that is difficult to follow for those not already familiar with the ECMWF suite. Could a reference to a system description be added, either in the form of a peer-reviewed publication or an operational technical note?
15. [L90-92] Both SV and MU have already been defined. (I actually think that both acronyms should be replaced with complete terms throughout the text for readability.)
16. [L101] What does the "current EDA cycle" mean? Does that refer to the one that was operational when this paper was written? Please be more specific.
17. [L103-104] Does ERA5 use the same version and configuration of the EDA as described here? This is possibly important because a close connection might mean that systematic errors are common between the forecast and analysis.
18. [Section 2.2] The extremely brief introduction of non-ECMWF systems in section 2.2 stands in stark contrast to the preceding full page of detailed description about the ECMWF ensemble. Please provide at least a brief introduction for each system (beyond Table 1) along with relevant references.
19. [L111] For consistency with what?
20. [L121] Why is PV only conserved, "following the *horizontal* flow on an isentrope"? To my understanding the orientation of the isentrope doesn't matter for PV conservation (note that any flow across an isentropic surface is better expressed as "diabatic" rather than "vertical").
21. [L128] How is the "speed of cyclogenesis" defined? Do you mean "deepening rate" or "intensification rate"?
22. [L130] "Eastern North America" is located east of the Great Lakes. Does this mean that the cyclone initially tracked westward? I think that showing the track in Fig. 1 would be more effective than this text description.
23. [L136] Parcels with ascent midpoints at 25oN are unlikely to be ascending above the warm front in the comma cloud region. If these are not following typical WCB storm-relative trajectories, what is driving their ascent? Is this an anafront? Perhaps this is unimportant, but the WCB points are described in some detail here, as is the distribution of precipitation.
24. [L148-153] This is all standard Reynold's decomposition, is it not? If so, then that should be mentioned here. If not, then the differences should be explained and justified.
25. [L154] What is the advantage of the Eq. 2 form over that used by Baumgart and Riemer (2019)?
26. [L174] What does the "intrinsic context" mean?
27. [L176] I do not think that "ground-truth" is usually hyphenated or single-quoted.
28. [L187] How is a 24-h running mean taken for background forecasts with a range of only 12h? The preceeding methodological description should be expanded and moved to section 2.

29. [L188-193] A figure caption is not the appropriate place for methodological descriptions (the same applies for the WCB trajectory calculations described in the Fig. 1 caption). Please include this information in section 2. Lines 189-193 of the text contain the information that should appear in the Fig. 2 caption instead of the methodological description.
30. [L193-194] Please state explicitly how the location of large LGR is "consistent with Hoskins et al. (1985)", why this is important, and why further investigation would be useful (though not useful enough to be presented here).
31. [L197] Please provide a section reference rather than "above", particularly because the erosion of the trough has not been previously discussed.
32. [L198] Are the animations are for different initializing times for this case or for different cases? Please be specific about what these animations contain and why they are relevant.
33. [L199] What does it mean to "'shadow' the true synoptic evolution of the flow"? This term also appears on L226, although it remains unclear how the "true synoptic evolution" is defined, particularly given the similar amplitudes of analysis and short-range forecast uncertainty.
34. [L200] What are "large model growth rates"? Does this refer to large LGR values within model simulations? Please be specific about which synoptic features are associated with these growth rates, if they are important. If they are not, this sentence should be removed.
35. [L201-207] These events have already been listed in the introduction. Because their connection here is purely speculative (it is explicitly noted that they are "not investigated here"), these sentences should be removed. Any discussion to be retained should be included in section 6.
36. [L209-210] Rather than forcing the reader back to section 3.2 to identify the reasons, why not list them briefly here and provide a back-reference to section 3.2 for interested readers?
37. [L227-228] Does "DJF 2020/21" follow WCD date formatting conventions?
38. [L228] The phrase "the agreement can be better" is not specific enough for a scientific publication. Neither is the support of this statement with a new case study (not described in the text) sufficiently robust. Please refer to General Comment #4 for a recommended replacement.
39. [Section 4 introduction] This is a highly condensed description of reliability that is unlikely to describe the concept effectively to readers who are not already familiar with it. (I am reasonably familiar with it and have a very hard time following both this discussion and Fig. 5.) Consider moving this description to an appendix and focusing the in-text description of reliability on what it looks like to have a reliable system, or what problems are related to a lack of reliability. These concepts would be useful in the context of the current work and would help to motivate the subsequent analysis. This suggestion should be read in conjunction with General Comment #5.
40. [L243] The term "uni-modal" usually appears without a hyphen.
41. [L247-249] Has this notation not already been described in section 3.2? If so, it should not be repeated here because it appears to add complexity to this already complicated description of reliability.
42. [L255] Why is the operational status of the forecast important enough to be italicized here (or important at all for that matter)?
43. [L263] The phrase "for the interested reader" suggests that there is an alternative to reading sections 4.1 and 4.2 for the uninterested reader: is that true? If it is, then that alternative should be explicitly stated here.

44. [L271] The “as discussed above” phrase is not a useful introductory clause here: terms 1-6 of Eq. 4 have not been explicitly “discussed above”. Please either remove it or include it in the parenthetical statement at the end of the sentence.
45. [L287] The {} symbols should be referred to as braces rather than parentheses.
46. [L289] What “later” is being referred to here? Please be specific about where further discussion of this term appears.
47. [L315] Please be specific about where this “later” refers to in the text.
48. [L325-328] This discussion seems to be relevant only to the observation-based analysis undertaken in the Rodwell et al (2016) study. Please consider whether it is needed here, given that it seems to add little of direct relevance to the current work.
49. [L343] Please be specific about where this “later” refers to in the text.
50. [L343] How much is “a little”? Please provide quantification.
51. [L345] Please be specific about where this “later” refers to in the text.
52. [L346] Consider “suggests potential” rather than “reflects” because the compensation is not shown here.
53. [L347] What demonstrates the “recent deterioration in storm track reliability” claimed here?
54. [L347] It seems unlikely that the storm track itself has become unreliable. Please rephrase to make it clear that EDA reliability has recently deteriorated in the storm track region, if that is shown to be true.
55. [L358-360] How does one pick errors, spreads and reliability from different ensembles? My understanding is that Reliability is computed from the ensemble distribution, which involves both the 0^{th} and 1^{st} moments. As such, the Reliability is not an independent quantity that can simply be chosen from an arbitrary ensemble. From a more utilitarian perspective, how would picking the reliability of a given ensemble have an impact on guidance?
56. [L359] Suggest “day-2”.
57. [L362] What part of this analysis demonstrates that the JMA system has the slowest initial growth rates (the ensemble has the largest spread in the second column of Fig. 6)?
58. [L363-364] Which of the two questions posed at the end of section 3.4 is being answered here? The first one (over-spread during cyclogenesis) seems the most likely referent; however, the analysis in section 4.3 does not distinguish between cyclogenesis and no-cyclogenesis events. As a result, it cannot be asserted that the ECMWF ensemble is over-dispersive “in the vicinity of cyclogenesis”. It appears to be over-dispersive in the storm track, but no more detailed statement than that would seem to be appropriate here.
59. [L364-365] This is a statement rather than a question.
60. [L376-377] Why is the K-means algorithm any better able to “cluster on structures” than other clustering approaches? For example, EOFs could have been used and the clustering done with their PCs. Such an approach would arguably be even more structure-aware than one adopted. There is no clear need to change the clustering strategy; however, the rationale for the methodological selection should be defensible.
61. [L392-393] It is clear from the preceding paragraph that the LGR is not used as an input for the clustering algorithm. However, once the methodological description is moved to section 2 with the remainder of methodology information, this note will be relevant to remind readers of the independence of this field.
62. [L410] A scientific audience should not need to be old that 91:89 is “nearly 50:50”.

63. [L417] The spread maximum for the cyclone cases appears to occur in the middle of the North Atlantic storm track, or even at the eastern end of its highest track density, rather than over the "western part". See for example Fig. 7a of Hoskins and Hodges (2019; JCLIM).
64. [L426] Why is there a tilde before the figure reference?
65. [L433] If this region is described as the "western end of the North Atlantic winter storm track", then it would be useful to provide a graphical description of the storm track early in this study. Cyclone tracking studies [including the recent Hoskins and Hodges (2019)] find peak cyclone density near Newfoundland, placing the western end of the storm track along the eastern seaboard. If a different definition of the storm track is used in this study, it should be clearly described to make the associated discussions easier to follow.
66. [L436] There do not appear to be any significant differences in Residual (Fig. 8o) over Newfoundland. There is a small region of significant difference over eastern Quebec and the Gulf of St. Lawrence, but this is west of the coastal storm track. The small spatial scale and multiple testing make the significance of this region questionable (using a field significance test might help in this regard). This seems inconsistent with describing the red area in Fig. 8o as "particularly strong and significant".
67. [L437] What does it mean that the opposite-signed differences "might be associated with differences in downstream cyclogenesis"? Does this refer to different realizations of downstream cyclogenesis in different members, or to different forms of downstream cyclogenesis in reality, or something else entirely?
68. [L440] "Root-cause" is not usually hyphenated.
69. [L448-451] This does not appear to be a complete sentence.
70. [L457] Suggest "day-2".
71. [Fig. 10] What is the contour interval for MSLP?
72. [L462-463] Although the use of different colour bars allows different ranges of values to be shown, it is misleading in such a figure where the panels show the results of different sensitivity tests. Please consider using the same colour bars for all panels.
73. [L482-484 and L493-494] These discussions of changes to precipitation seem somewhat tangential to the main themes of the manuscript and could be removed.
74. [L500 and L547] Reword "2 d".
75. [L505] The phrase, "indicating that the conclusions drawn in this section are robust even with only two cases" does not seem logically correct. The fact that a second case shown a similar pattern gives adds to confidence about the conclusions; however, the similarity of two cases does not provide some sort of successfully conclusive evidence as implied by this statement.
76. [L515] Suggest replacing "these aspects might be developed" with "these techniques might be modified" for clarity.
77. [L554-555] Are any modern calibration techniques state *in*dependent as implied here for machine learning techniques?
78. [L560-562] It is unclear which results are being referred to here. Fig. 10 show that SV and MU have (by far) the leading impact on Z250 spread; DCP is a distant runner-up. However, this discussion seems to imply that DCP is dominant, with SV and MU also contributing. Although the results are more uniform between the three for 315K PV, this statement could easily lead future readers to think that deep convection has more of a relatively larger impact than it actually does in this case.

79. [L564-565] The wording of this sentence seems unnecessarily vague and complex.
80. [L566] Suggest removing hyphen in “model-uncertainty”.
81. [L571] This is a very abrupt ending to the manuscript. Consider adding a broader statement that is more directly related to the work undertaken in this investigation.