Dear Michael,

Thank you for your attention to this manuscript. We had not fully appreciated the problem some readers would have with the butterfly term. This has now been removed. The idea, that it is useful to evaluate uncertainty growth in a flow-dependent way, is hopefully still conveyed. We have spent not a small amount of time addressing the reviewer's concerns, but we greatly valued their diligence and insight. We hope that the revised manuscript has satisfied their concerns. Replies to your further comments are in blue text below.

Kind regards, Mark & Heini.

Dear Mark, dear Heini,

I have received the second round of reviews from two highly qualified and attentive reviewers. Both reviewers acknowledge the improvements made during the first round of revision. Both reviewers, however, have further major issues that need clarification before publication.

Both reviewers reinforce their issues with your use of the term "butterfly". I agree with the reviewers that re-defining the meaning of this term creates readers' confusion, without a benefit that a would see for the reader. Your reviewer Ron McTaggart-Cowan makes a constructive suggestion for an alternative title. (Shifting the focus on cyclones instead of cyclogenesis seems helpful to me also to avoid unnecessary confusion. I further agree with the reviewer that this shift would not at all diminish the significance of your results.)

A further issue that carries over from the first round of the reviews is the presentation of the material. There may be different opinions about how to best organize the material, and there may be different approaches that may yield satisfactory results. In its current version, however, the organization of the material affects the quality of the manuscript not to a small degree. To you as authors, this issue may not become so clear, because you are well aware of the storyline of your work and the major points that you would like to communicate. Switching between discussions of results, discussions of key concepts, and technical information on methods may not seem distractive to you. For your readers, however, that is very much different. Reviewer Ron McTaggart-Cowan's comments illustrate these distractions very well. I'd like to emphasize that this is not a critique of your writing style or your writing preferences; it's a matter of the functionality of the organization. I acknowledge that a solution to the issue will most likely be more complex than simply introducing a method section. When you introduce methods, you discuss conceptual aspects of these methods also. This conceptual guidance is highly appreciated. The guidance, however, is interspersed with technical information that distracts the reader from understanding the conceptual value of the respective methods. Combined with introducing methods during a discussion of results makes the current manuscript a difficult read. Helpful comments for re-structuring are found in the reviewer's comments. My own impression is that providing the conceptual guidance when you start discussing the results obtained by the method, while putting the more technical aspects of the method into a method section (or the appendix; section 4.1 seems to be a good candidate for that) will benefit

the reader. Both of you are highly experienced writers. I have no doubt that you will find a good solution to this issue once you "see the problem through the readers' eyes".

Noting these specific points, of course, does not imply that I mean to downplay any of the other points raised by the reviewers.

Below are a few minor points that I noted when I was having a look at your revised manuscript. I apologize if there is overlap with comments by the reviewers.

Kind regards, and I am looking forward to receiving your revised version.

Michael

- In the abstract, you refer explicitly to baroclinic and convective instability, just after noting the focus of your study. The role of these instabilities is hardly touched on in the manuscript. The explicit mention could thus raise readers' expectations that your manuscript will not meet. Do you see, for the reader, a clear benefit of referring explicitly to baroclinic and convective instability in the abstract? If not, consider omitting.
- We hope that the revised text makes the link more clearly
- The term overbar(v dot grad P) seems to be missing in the fourth line of Eq. 2. (I do not think that this derivation needs to be shown, though, at least not during discussion of results.)
- The term associated with overbar(dP/dt) in line three disappears when multiplied by P' and averaged over the ensemble. The final equation is shown in the main text and explained comprehensively in Sect. 2.6, and then discussed in the context of cyclogenesis in Sect. 3.1. The derivation is placed in Appendix B.
- L187: Can you clarify how a 24h running mean is applied to the 12h EDA forecast?
- Yes, we have clarified this in response to Ron's specific comment 29. The 24 h running mean is made after concatenation of hourly values from each 12 h assimilation cycle.
- - Acknowledgement: As Ron McTaggart-Cowan has revealed his identity you may want to consider referring to him by name.
- Yes we do acknowledge Ron by name.

wcd-2022-6

The Cyclogenesis Butterfly: Uncertainty growth and forecast reliability during extratropical cyclogenesis

by Mark J. Rodwell and Heini Wernli

Replies to the reviewer's comments (in blue text).

The authors would like to thank Ron McTaggart-Cowan and the other anonymous reviewer again for the time and care that they put into reviewing the revised version of this manuscript, and for their additional comments, which will be addressed in detail below. The main changes compared to the previous submission are the following:

- We implemented a more classical organization of the material, as requested by reviewer 1, and now present the many methodological aspects upfront in Sect. 2.
- We further clarified the objectives of the study.
- We completely removed discussion of butterflies.

Reviewer 1

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.

   The Abstract and Introduction have also been re-written, and we think that this makes the objective clearer.

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".

   Following the theme of our previous paper titled "Flow-Dependent Reliability: A Path to More Skillful Ensemble Forecasts", the aim of the current study was to continue (and promote) work that leads to flow-dependent improvements in ensemble reliability. The question, therefore, was "what flow aspects are most strongly associated with deteriorations in ensemble reliability at present?" Preliminary work here identifies flow aspects which have a strong impact on ensemble spread, and then we evaluate the maintenance of reliability. Hence, we believe that

the part of the current paper's title "Uncertainty growth and forecast reliability during extratropical cyclogenesis" is well justified (see response to general comment 7 about the use of the word "cyclogenesis"). The question remains over the "Cyclogenesis Butterfly" part. This was added because we foresee (or would like to see) research developing where other flow situations associated with strong uncertainty growth, other butterflies, are identified and evaluated for reliability. However, have not been able to convince the reviewers on this point, and appreciate their concerns over this phrase, so we have removed it from the title and paper.

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.

   There are pros and cons to both options, but we followed the reviewer's advice and centralized the methodological descriptions in Sect. 2. For the methodologies, we discuss the reliability before the growth rate, so that the text is less disjointed as we move next into the growth rate results.

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.

   This is an interesting idea. In view of the length of the current manuscript and the diversity of the employed methods and analyses, we suggest that a more systematic comparison of LRG be left for a subsequent study. In our view, several of the diagnostics used in this study are fairly novel and we regard it therefore as positive, if the results shown trigger additional and in parts more in-depth studies on certain aspects, as the one suggested here by the reviewer.

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.

We have re-ordered as requested. In response to this comment and specific comment 40, we now provide a more intuitive introduction to reliability up front, in Sect. 1.

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.

   The approach here is a little different, with departures relative to ensemble-mean rather than a single unperturbed observation. This study also goes further to evaluate the assumption of constant (flow-independent) bias. Nevertheless, we have followed the reviewer's advice. In Sect. 2.3, we appeal to the previous paper much more strongly, and simply discuss the differences. In order to discuss the impact of variations in forecast bias, the derivation and discussion of the residual term are retained, but moved to Appendix A.

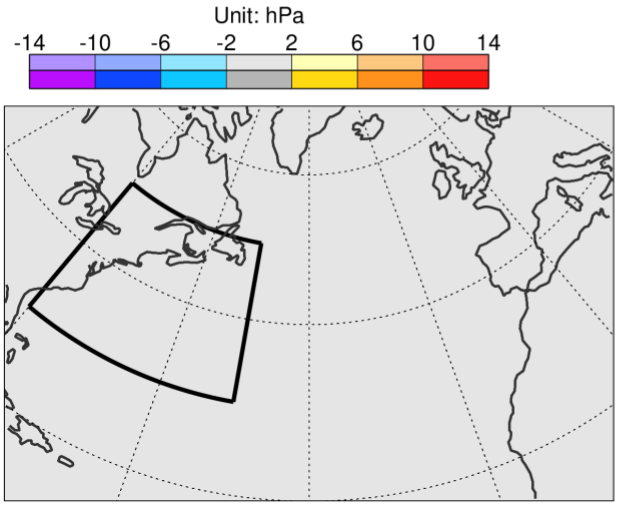
7. The term "cyclogenesis" appears to be used primarily to refer to the presence of a cyclone. This is important because the "cyclogenesis butterfly", based on a standard definition of cyclogenesis, implies uncertainty introduced by a cyclone is forming or deepening. However, the "cyclogenesis" clusters 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.

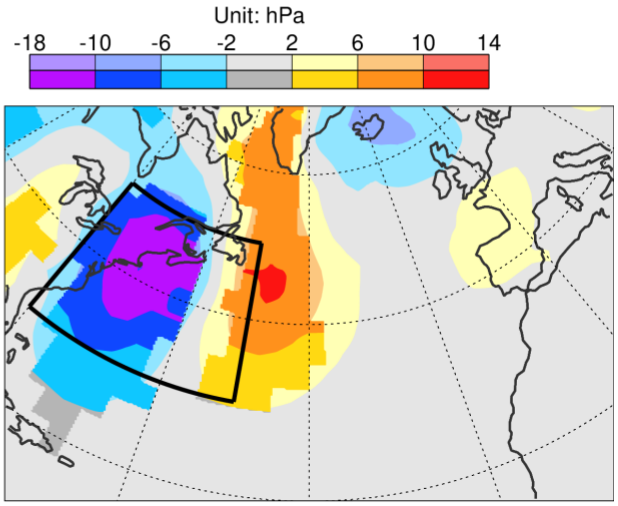
   The attached figures for the two clustering regions quantify the deepening of cyclones in this region. Despite being a somewhat diffuse average of events, for cluster region 1, the cluster-mean deepening for the "cyclogenesis cluster" attains 14 hPa over 2 days. For cluster region 2, it is 9 hPa. We believe that this justifies definitively our use of the word "cyclogenesis" in the sense that on average, there is strong intensification of the considered cyclones in the selected regions.

PMSL change D+0 to D+2

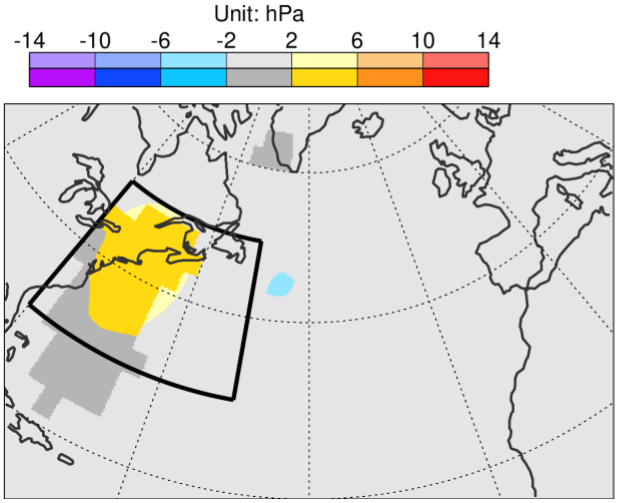
(a) All dates (size=180)

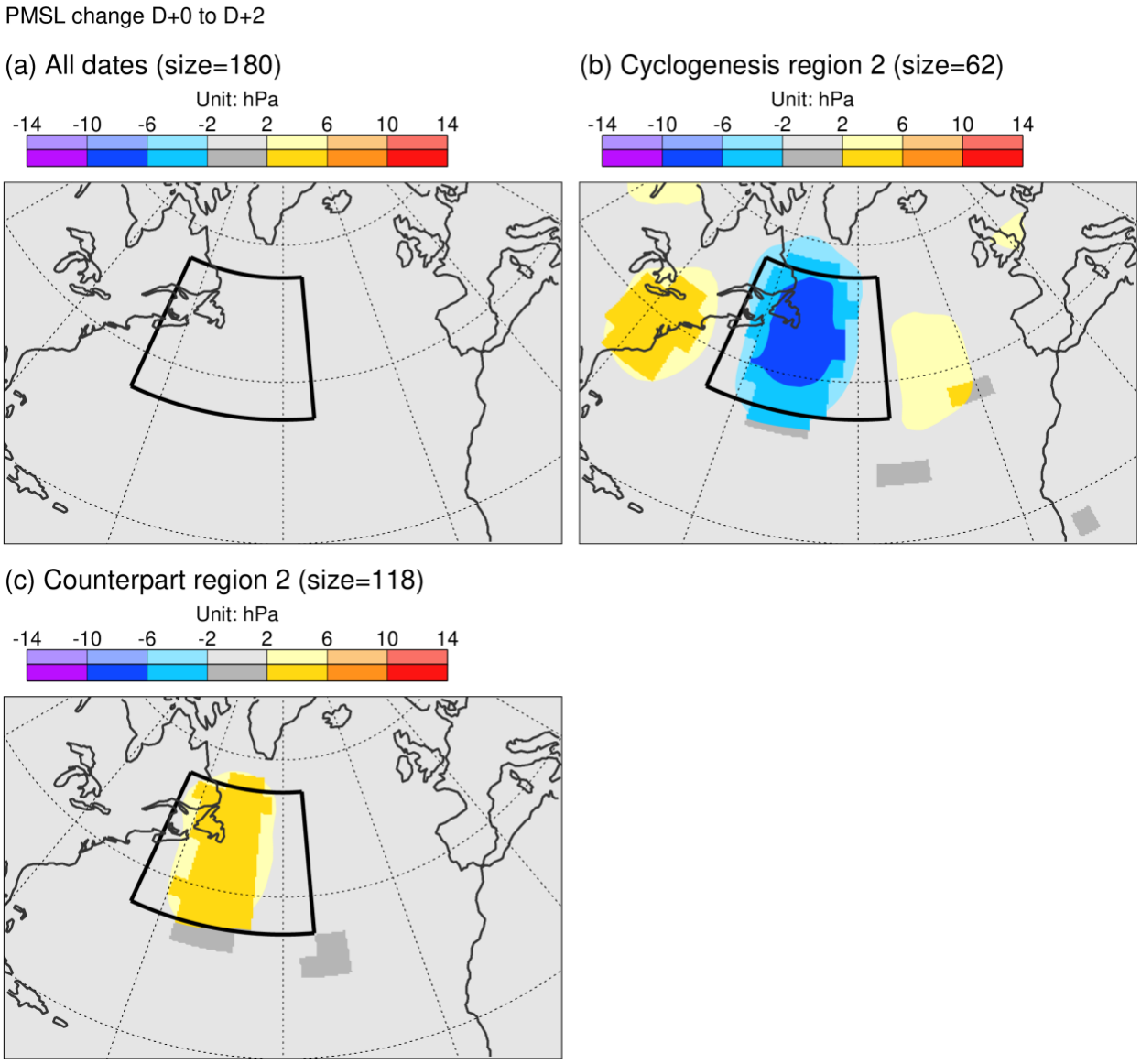


(b) Cyclogenesis region 1 (size=32)



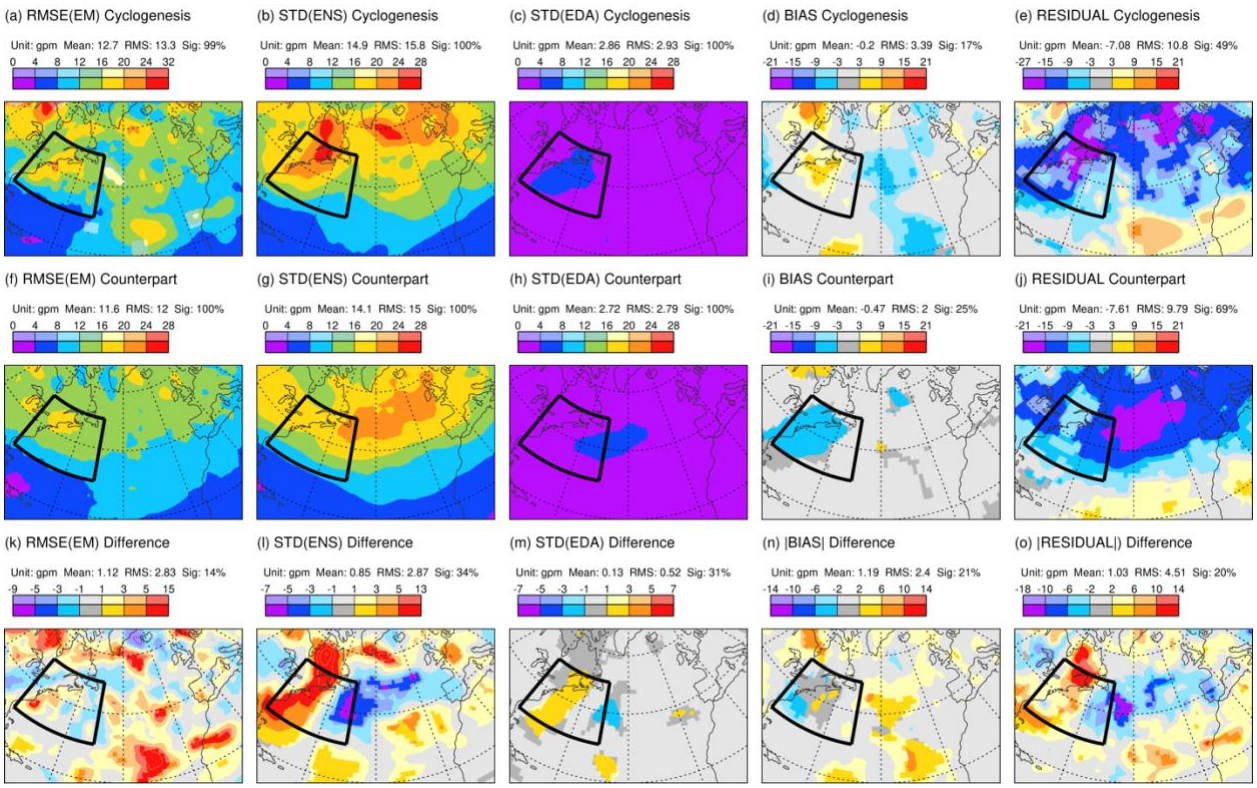
(c) Counterpart region 1 (size=148)





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.

   Our test is a strong one – 5% significance – and this is achieved in places in the Fig. 8o (new Fig. 7o due to re-organisation). We sample synoptic variability because that is what we are interested in. It is unclear how sampling interannual variability would help. A larger sample through adding years could improve significance, but then the model cycle will have changed. We maintain that new Fig. 7e, f, and o, together are sufficient to back-up our conclusion. Further justification, however, comes from the attached plot which shows the cyclogenesis/counterpart break-down when only the first region is considered. With better confinement of cases in a single region, we see stronger significance, even in the difference (panel o).



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.

   We have moved methodologies to Sect. 2.

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.

    Eq. 3 (including rhs) is central to the current study. We discuss the dynamical (rhs term 2) and diabatic (rhs term 1) aspects that could be associated with cyclogenesis in Sect. 3.1. It is also important to highlight what processes can be identified by the Lagrangian growth rate.

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.

    Our convention here is that the short hyphen is for use in hyphenated words like "co-ordinate", the longer en dash for numerical ranges like "2--3", and the longest em dash that splits text ---

particularly when a striking conclusion or important qualification follows. We have reduced usage of em dashes. There is only one em dash in the Introduction, for example.

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.

    There are differences in convention (UK versus US) for use of single and double quotes, and we did not fully adopt either. We have now removed most of the quotes, and follow the "single quote inside double quote" convention throughout.

Specific Comments

1. [L19] Consider rewording split infinitive.
   The re-wording avoids this issue.
2. [L23] It isn't "NWP" itself that develops techniques, but researchers and system developers.
   This is now avoided.
3. [L26] I believe that "leadtime" is usually written as "lead time".
4. This has been changed throughout.
5. [L28] I believe that "Stormtrack" is an application while, "storm track" is the usual term for the region discussed in this study.
   This has been changed throughout.
6. [L32] Is "propone" the word that you mean to use here? Consider replacing with "prone" or "conducive".
   Done.
7. [L35] Why is "blocking" (well-accepted terminology) enclosed in single quotes?
   Quotes removed.
8. [L49] I think that a comma before the quoted question would be appropriate.
   This has disappeared in the re-write.
9. [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).
   Done.
10. [L53] Why does the bias problem apply only to short-range assessments of reliability as implied here?
    For the forecast it doesn't, for the analysis it does – we have re-worded this in Sect. 2.3, L140-142, thanks.
11. [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.
    This now appears in the abstract and conclusions, and is re-worded.

12. [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.

    This has been re-worded in line with the previous comment.

13. [L63-65] This appears to be a run-on sentence: please rephrase.

    Done as part of the re-write of the Introduction.

14. [L69] Why is “Ensemble” capitalized here?

    Because it relates to the abbreviation which follows.

15. [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?

    We cited 5 papers on various aspects of the EDA. It is difficult to find publications specifying the configuration since this changes frequently. The best we can find is an ECMWF Newsletter article “A 50-member Ensemble of Data Assimilations” which is now also cited at L96-97.

16. [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.)

    We have made sure that both terms are only defined once, where they are first used. The acronyms are removed except in Sect. 5 when referring to the figures and experiment names.

17. [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.

    It refers to the EDA cycle at the specific time under consideration, so that a single EDA cycle can be run as an experiment. In operations, this is not possible, and the scaling is based on the previous cycle, 12 hours before. This is a very small point, which was originally relegated to the appendix but brought forward to the “Models, data and methods” Sect. 2 in response to reviewers’ previous comments. Getting the right balance between comprehensiveness and readability is difficult in complex diagnostic studies. We now say at L128 “singular vector perturbation scaling is based on the current EDA cycle rather than the cycle 12 h before”.

18. [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.

    ERA5 is based on an older model version and configuration. It is only used here for the plotting of Fig. 1.

19. [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.

    We have pointed the reader to the documentation available within the TIGGE archive at L137.

20. [L111] For consistency with what?

    For consistency of comparison, we need to use the same forecast start times. We now say (L137) “Here, comparisons are based on the common run times of 00 and 12 UTC”.

21. [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").
    The orientation of the isentrope does not matter for conservation of IPV, but the advection within the material derivative is based on the horizontal flow. If we omitted the word "horizontal" then we risk "flow on an isentrope" being interpreted as "flow along an isentrope".
22. [L128] How is the "speed of cyclogenesis" defined? Do you mean "deepening rate" or "intensification rate"?
    We now state L56-57: "However, the rate of deepening and the growth of uncertainty were not considered in the choice".
23. [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.
    Apologies for using wrong terms; we changed "Eastern North America" to "the Midwestern U.S." (the cyclone tracked eastward). The sentence then reads L42-43: "A little earlier than shown in Figure 1, on 26 November 2019, the cyclone had begun to develop over the Midwestern U.S. A day later, it had reached the Great Lakes …".
24. [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.
    Thanks for looking at this level of detail into this case study. Yes, to us, this looks like an anafront with slantwise ascent at the (extended) cold front (something we've seen already in very early WCB case studies with trajectories, e.g., Fig. 12 in Wernli 1997, QJ, 123, 1677-1706). For this study, however, we decided that adding such mesoscale information might be distracting.
25. [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.
    Yes, this is Reynold's decomposition into ensemble mean and deviations about the mean. Reynold's decomposition can also refer to (e.g.) a spatial mean and deviations about it (Reynold's stresses), and hence this could cause more confusion that help. We believe the new description (L213-214) goes some way to making things clearer.
26. [L154] What is the advantage of the Eq. 2 form over that used by Baumgart and Riemer (2019)?
    Eq. 2 (now Eq. 3) is written as the exponential growth rate (normalised by the spread). This is important, for example, when comparing TIGGE ensembles (as in Fig. 3) with different initial uncertainty. New Eq. 3 also extracts the strong advection of uncertainty by the ensemble mean, which we believe is useful for discussing the initial material growth rate when the ensemble members have similar wind fields. Baumgart and Riemer effectively extract the strong flux convergence term. We now discuss a little further these aspects in Sect. 2.6.
27. [L174] What does the "intrinsic context" mean?
    We have reduced discussion of intrinsic aspects in response to Reviewer 2's comments. We now only discuss once: L65-67: "Whether the answer hints at an intrinsic property of the atmosphere, or is dependent on the formulation of the forecast system, is explored by

comparing models within "The International Grand Global Ensemble" (TIGGE, Swinbank et al., 2016) archive".

28. [L176] I do not think that "ground-truth" is usually hyphenated or single-quoted.
    We no longer mention ground truth.
29. [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.
    We now mention the concatenation aspect ahead of the filter discussion, so it is clearer that a 24 h running mean is possible. Following the reviewer's advice, this text is now moved to Sect. 2.6
30. [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.
    This has been done.
31. [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).
    In Hoskins et al. (1985) Fig. 21, during cyclogenesis the equatorward flow anomaly at upper levels acts to enhance PV advection, which strengthens and slows the eastward progression of the trough. Uncertainties in this feedback process are represented in the second term on the right-hand side of Eq. 3. This is now discussed better in Sect. 3.1 para 2.
32. [L197] Please provide a section reference rather than "above", particularly because the erosion of the trough has not been previously discussed.
    The development of the LGR_P equation in the old Sect. 3.2 has been moved to the enlarged "Models, data and methods" section, to Sect. 2.6. The discussion of this equation in relation to cyclogenesis has been brought together into the Sect. 3.1 "Uncertainty growth in the EDA". This avoids the need for a backwards reference.
33. [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.
    There are EDA and TIGGE animations for the DJF 2020/21 season. There were also animations for the two original cases, but one of these cases has been dropped now and so it seems sensible to only make the full season animations available.
34. [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.
    We now state at L238-239: The resulting timeseries of fields can be used to produce animations of P315 which "shadow" (remain within the background uncertainty of) the true synoptic evolution of the flow.
35. [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.

Yes, this refers to LGR_P. We are now more specific about the features associated with these growth rates in Sect. 3.1.

36. [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.

    These events are no longer listed in the Introduction. They are left in Sect. 3.1, where the animations are discussed.

37. [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?

    The reference to intrinsic growth rates is no longer included here. The only discussion is in Sect. 3.1, L272-274.

38. [L227-228] Does "DJF 2020/21" follow WCD date formatting conventions?

    The first time the season is introduced (L163) we state "December–February 2020/21 season (DJF 2020/21)".

39. [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.

    Please see our reply to General Comment #4.

40. [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.

    We have followed this advice. A more intuitive introduction to the concept of reliability is now given in the first paragraph of the Introduction.

41. [L243] The term "uni-modal" usually appears without a hyphen.

    This has been changed.

42. [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.

    The main difference is that an overline in the reliability evaluation relates to a mean over forecasts, while an overline in the growth rate relates to a mean over ensemble members. We appreciate that this can be confusing. With some work within LaTeX, which we hope will be acceptable to the journal, we have managed to indicate the mean over forecasts with a thick overline.

43. The inclusion of both equations in the "Models, data and methods" section, with their description hopefully improves this.

    This has been done and, yes, it does.

44. [L255] Why is the operational status of the forecast important enough to be italicized here (or important at all for that matter)?

    We have removed the word operational where it is not required, and it is no longer italicized.

45. [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.

    The use of a long models, data and methods section means that this is no longer in the revised manuscript.

46. [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.

    This has been done.

47. [L287] The {} symbols should be referred to as braces rather than parentheses.

    This has been changed.

48. [L289] What "later" is being referred to here? Please be specific about where further discussion of this term appears.

    We now refer to the relevant Appendix section.

49. [L315] Please be specific about where this "later" refers to in the text.

    This has been removed

50. [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.

    Following the referral to this equation (following General comment 6), a discussion of the differences is probably more important than it originally was, and hence we retain this text.

51. [L343] Please be specific about where this "later" refers to in the text.

    We now refer to Sect. 4.3.

52. [L343] How much is "a little"? Please provide quantification.

    We leave "a little" here because it is now made clear that this is quantified in Sect. 4.3 (please see answer to previous point). This seems more informative than omitting "a little".

53. [L345] Please be specific about where this "later" refers to in the text.

    We again now refer to Sect. 4.3

54. [L346] Consider "suggests potential" rather than "reflects" because the compensation is not shown here.

    We now state L343-345 "As part of the current study, but not shown here, this reflects compensating deficiencies elsewhere, a recent deterioration in ensemble reliability in the storm track, and the importance of accounting for bias and analysis uncertainty".

55. [L347] What demonstrates the "recent deterioration in storm track reliability" claimed here?

    Please refer to the reply to the previous comment.

56. [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.

Please refer to the reply to specific comment 54.

57. [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?

    We were suggesting what might be fairly easily achievable. This is a small point and has been removed.

58. [L359] Suggest "day-2".

    We have improved the consistency of usage, using "day-2" and "2 day" throughout.

59. [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)?

    We now refer to new Fig. 3 and Fig. 4.

60. [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.

    We now state L363-365 "To answer the question in Sect. 3.2, whether the ECMWF growth rates are too strong in the vicinity of cyclogenesis, it needs to be determined whether the negative Residual in Fig. 5e is associated with a general level of over-spread or whether it can be linked to cyclogenesis events per se".

61. [L364-365] This is a statement rather than a question.

    This is removed by the change made in relation to specific point 60.

62. [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.

    We consider the K-means approach to offer a better chance of obtaining the required structures. EOFs do not necessarily represent physical structures, but can be used as a means of reducing dimensionality. See, e.g., Corti, S., Molteni, F. & Palmer, T. Signature of recent climate change in frequencies of natural atmospheric circulation regimes. Nature 398, 799–802 (1999). https://doi.org/10.1038/19745. For brevity, we have not included this discussion in the manuscript.

63. [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.

    Yes, this note is retained here.

64. [L410] A scientific audience should not need to be old that 91:89 is "nearly 50:50".

With “nearly 50:50” we are implying that DJF results lie nearly halfway between the values in the two clustered rows, and emphasizing that we have a sufficient sample on both sides.

65. [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).

    We have removed the words “storm track”. The point being made is that the spread maximum for the cyclogenesis cases is to the west, and the spread maximum for the counterpart is to the east. Close comparison with other studies would require a lot of understanding of the differences in methodologies.

66. [L426] Why is there a tilde before the figure reference?

    The figure panel indicates that there is variance in forecast bias. It may provide a rough estimate of this variance, although we have only separated into two flow regimes. There could be more variance in forecast bias associated with other flow regimes. We have changed the tilde to “implied in”.

67. [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.

    We refer to our reply to Specific comment 65, and have changed the text here to say L421-423 “Here it is evident for the ECMWF ensemble that most of the over-spread during DJF 2020/21 in the western North Atlantic region of focus (Fig. 5e), is associated with the cyclogenesis composite”.

68. [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”.

    We refer back to our reply, with attached figure, to General Comment 8.

69. [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?

    The cluster analysis was used to separate-off cyclogenesis events over the western North Atlantic. What is left, particularly associated with cluster 3 for region 1 (new Fig. 6c) and cluster 2 for region 2 (new Fig. 6e), contains increased cyclogenesis over the eastern North Atlantic. This might be expected from knowledge of the spatial and spatial-correlation scales of cyclogenesis. We now say L427-429 “Downstream, differences have the opposite sign — possibly because the occurrences of cyclogenesis events over the western North Atlantic are likely to be anticorrelated with the occurrences immediately downstream (as seen in cluster patterns Fig. 6c,d).

70. [L440] “Root-cause” is not usually hyphenated.

    We now avoid root-cause by saying L431-432 “This issue could be associated with several different aspects of the forecast system. Through sensitivity experiments, Sect. 5 explores some of the potential causes”.

71. [L448-451] This does not appear to be a complete sentence.

    Sorry, this was very garbled. We have changed the text to L440-444 “Firstly, singular vector perturbations to the initial conditions of the ENS are turned off globally (OP-SV) and then model uncertainty in the ENS is turned off globally (OP-SV-MU). From this point, the parametrization of deep convection in the ENS is turned off in a local box (OP-SV-MU-DCP) or the ENS model horizontal grid resolution is increased to ~ 4 km (OP-SV-MU+4km). Finally, the assimilation of observations in the EDA is turned off in a local box (OP-Obs) and the ENS is run again in the OP-SV-MU configuration”.

72. [L457] Suggest “day-2”.

    This has been changed. Please see response to specific point 58.

73. [Fig. 10] What is the contour interval for MSLP?

    10 hPa. This has been added to the new Fig. 9 caption.

74. [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.

    To allow the reader to see the structure of each sensitivity, it is necessary to vary the shading interval. We did state in the text L454-456 “Note that shading intervals vary over the panels shown in these two figures, so that the structures of all impacts can be seen”. We now also make this clear in the panel caption.

75. [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.

    We consider that these remarks are useful. Indeed, it is quite interesting that spread can change in response to DCP without any overall change in total precipitation.

76. [L500 and L547] Reword “2 d”.

    This is changed to day-2

77. [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.

    These are spread sensitivities based on 50-member ensembles, and do not relate to errors where only a single realisation of the truth is available. We now replace with “may be” and say L499-501 “Very similar results to those above were obtained for a second set of experiments initialised at 00 UTC on 17 January 2020 — indicating that the conclusions about ensemble spread sensitivity (based on 50 members) may be robust even with only two cases (the same could not be said for error sensitivity with just two cases).

78. [L515] Suggest replacing “these aspects might be developed” with “these techniques might be modified” for clarity.

    Done.

79. [L554-555] Are any modern calibration techniques state *in*dependent as implied here for machine learning techniques?

    We believe there is a lot of calibration done without knowledge of the history of the forecast. However, to avoid having to go into a lot of detail, this minor comment has been removed.

80. [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.

    We agree. In an attempt to highlight that the chaotic growth of EDA uncertainty is itself sensitive to deterministic model formulation, we added a phrase before the SV and MU impact discussion. This clearly led to the reviewer's interpretation. This phrase has now been removed at L551-554.

81. [L564-565] The wording of this sentence seems unnecessarily vague and complex.

    This has been re-worded in the re-write of the last paragraphs.

82. [L566] Suggest removing hyphen in "model-uncertainty".

    The hyphen is added as the text would otherwise include "model and model", which sounds a bit jarring.

83. [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.

    We have tried to broaden the statement in the last paragraph.

Reviewer 2

Review of "The Cyclogenesis Butterfly: Uncertainty growth and forecast reliability during extratropical cyclogenesis" by Mark John Rodwell and Heini Wernli

Dear authors,

Let me first apologize for being late with my review.

Thank you for carefully considering my comments! The paper has substantially been rewritten and in my opinion it has improved a lot. However, there are some crucial issues left which I suggest to reconsider.

Main points:

1. Butterfly and predictability

I am still not happy with the “butterfly” discussion. You did add some explanation what you mean with the term “Cyclogenesis butterfly”, but I still think this term might be confusing and also not really relevant for the paper and the issues it discusses:

We did not manage to convince either of the reviewers on this point, and appreciate their concerns. The broader aim with the term “cyclogenesis butterfly” was to promote the identification of other flow situations associated with large ensemble uncertainty growth, and to evaluate forecast reliability in these situations. We have dropped “cyclogenesis butterfly” from the title and paper. References to intrinsic predictability are also minimal now (only in Sect. 3.1). We suspect that many of the reviewer’s comments below are addressed by this change.

Since you are not investigating intrinsic predictability, I would suggest to not start the abstract with a quote from Lorenz that refers to intrinsic predictability.

This has been dropped from the abstract.

Further in the abstract and also in the introduction and conclusion you refer to decreased predictability and high sensitivity associated with the cyclogenesis. But can this really be concluded based on your investigation? Yes, the Lagrangian growth rates in the cyclone in your examples are high (mostly for ECMWF) and divers, but you show later that they are too high.

The differences with other models were/are highlighted in the abstract. The ECMWF over-spread is based on the model with SVs applied (which we consider to be the main culprit in the over-spread), while the large growth rates exist even without the SVs. The discussion is more nuanced now and, with the dropping of the cyclogenesis butterfly, we hope this satisfies the reviewer.

Second, looking at Figs. 8a) and 8f) the errors in your target area seem to be smaller in the cyclogenesis composite compared to the counterpart. Not much is said about this in the paper, but wouldn’t this indicate that cyclogenesis events are actually more predictable than the rest, at least on average in this period?

This is the error of the ensemble-mean, which is a different aspect. To clarify this, we have added the text L413-414 “Note that the stronger bias along the eastern coast of North America for the counterpart composite (cf. Fig. 7d,i) explains its larger ensemble-mean error (cf. Fig. 7a,f) in that area”.

So is there really a physical based high sensitivity or “butterfly”-Lorenz63 phenomenon present? Or is this just a “malfunction” sensitivity of the rather unphysical inflation methods used (SPPT, SV)?

The sensitivity studies at the end demonstrate that at least 50% of the uncertainty growth is associated with chaotic growth from initial (non-SV and non-SPPT) uncertainty.

2. Reliability

In Fig. 8e) you show a large residual in the target area and argue that the models uncertainty representation may have a problem with cyclogenesis. And I think that’s fair to say. But what about the even larger residual east of the target area in the counterpart? I think this should be discussed in somewhat more detail than in the current draft (only L436-437). Does the model may have even larger problems with other flow configurations? Is this related to ridge building, cyclone decay, secondary cyclogenesis? I think this should be at least roughly addressed, otherwise the statement “… flow-type clustering demonstrates that its over-spread in the stormtrack is indeed associated with cyclogenesis events” (L13-14, also L438) is not really justified in my opinion.

This downstream area is outside the compositing region. Increased downstream residuals in the counterpart composite can be due to cyclogenesis, which is likely to be more prevalent than in the composite with upstream cyclogenesis. We have modified the text to say L427-429 “Downstream, differences have the opposite sign — possibly because the occurrences of cyclogenesis events over the western North Atlantic are likely to be anticorrelated with the occurrences immediately downstream (as seen in cluster patterns Fig. 6c,d)”. Hence the downstream residuals may be explainable in exactly the same way.

Specific comments:

L42, 210: Intrinsic predictability is not really a sensitivity only to small-scale perturbations. It is rather characterized by a loss of sensitivity to the scale of the perturbation if their amplitude is sufficiently small (e.g. Sun and Zhang, 2016).

This has been removed, but there seem to be differing points of view. Reviewer 1 mentioned in their first review that “these are very big butterflies”.

L43 (Although the underlying processes…): I disagree with this statement. The recent study of Selz et al. 2022 (which you cite a few lines late) clearly showed how the error driving processes change when the amplitude of the initial condition uncertainty is reduced.

This is now acknowledged in Sect. 3.1

L45-46, L92: What kind of errors SPPT represent is not entirely clear, however, I think there is substantial evidence that it is not primarily missing interactions with unresolved scales: First, the recent success of SPP indicates that the parameters of the parameterizations (hence their assumptions and approximations) are associated with model uncertainty. Second, not all parameterizations account for unresolved motions (e.g. radiation, microphysics) but are also perturbed in SPPT and SPP. Third, a stochastic convection scheme that does account for unresolved motions only has virtually no impact on

error growth when the initial condition uncertainty is operational, see Selz et al. 2022. And don't you arrive at the same conclusion in L453-454?

Any model uncertainty representation is a pragmatic approach to improving ensemble reliability. One of the aspects which they purport to account for is missing interactions with unresolved processes. We have changed the text to L119-120 "A model uncertainty parametrization, which partly aims to represent scale interactions with (missing) sub-grid-scale variations".

L527-532: This paragraph confused me. First, yes, the initial growth rate in operational systems is much smaller than in intrinsic predictability experiments, but the latter are rather insensitive to the scale of the perturbations. Also small-amplitude large-scale perturbation lead to extreme growth rates on small scales. I think this was one of Durran's main points. Since the cyclogenesis growth rates are too high and associated with "unphysical" methods like SV and SPPT (see main point 1), I don't see how you can conclude that scale interactions and diabatic processes are important here. And what do you mean with "longer than expected intrinsic limit"?

This has all been removed. The "longer than expected intrinsic limit" aspect relates directly to the Palmer et al (2014) paper.

L533-534: No reason to speculate, this has now been done (Selz et al. 2022). The paper showed that on average error growth from operational uncertainties is mainly in the dry, balanced and larger-scale part of the flow.

A potential difference may be that here we consider growth rates in a very specific flow situation. In addition, Selz et al 2022 acknowledge that results can be sensitive to the model used in perfect model studies. Fig. 11d (new Fig. 10d) highlights this point from the perspective of diabatic processes.

L533-543: Again, is there really a problem with predictability in form of error growth or is there "only" a reliability problem, caused by the rather empirical model uncertainty representation methods?

The conclusions section has been re-worded. A conclusion of the study is that we would be better able to answer this point if SVs are removed from the operational forecasts. From an operational forecasting perspective, model uncertainty will likely remain a central aspect of the model, albeit empirical.

Minor comments, typos:

L131: 18UTC vs. 12UTC in the figure?

Thanks, should read 12 UTC.