wcd-2022-6

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

by Mark J. Rodwell and Heini Wernli

Replies to the reviewers’ comments

The authors would like to thank both reviewers for the time and care that they put into reviewing this manuscript, and for their insightful comments, which helped to improve the clarity of the presentation of our results. Before addressing the individual comments in detail below, the main changes compared to the original submission are briefly summarized:

- Large parts of the text have been rewritten to better explain the objectives of the study, to streamline the logic of the paper, and to increase the clarity of the analyses.
- Parts of the text have been shortened or omitted (e.g., discussion of link to Liouville equation).
- Only one case study is shown, as suggested by the reviewers.
- Reliability has been evaluated for all four TIGGE models.
- Two dates are now shown for the TIGGE models to give a better reflection of the range of results.
- The mathematical parts have been better integrated in the paper.

Reviewer 1

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

Overview

This manuscript addresses the very interesting problem of flow-dependent variability in ensemble reliability. Such an analysis is of significant practical utility because it gives ensemble designers important robust insights into their system's behaviour under identifiable meteorological conditions. Specifically for ensemble applications, such information is an essential replacement for the case study approach – although arguably conditional evaluations should also be preferred for deterministic systems.

It is clear from the breadth of the analysis that an impressive amount of work has gone into this investigation. However, the manuscripts suffers from the lack of a clearly stated objective for the complex diagnostics employed. As a result, the text gets mired in technical discussions rather than focusing on interpretations and discussions that support the objectives of the work and advance the main narrative of the manuscript. Similarly, many of the novel diagnostics themselves (for example Eqs. 1 and 2) seem to be overly complex for a study that arrives at relatively straight-forward – though very useful – conclusions regarding conditional overdispersion in the ECMWF ensemble.

As described in General Comments #1 and #2 below, I think that this work is interesting and important enough to be split into two separate manuscripts. The result will be two independent but complimentary studies that better motivate and demonstrate the utility of the proposed techniques. Such a reshaping of the investigation will also permit the introduction of more synthesis and interpretation of the results, resulting in a pair of papers that will have a larger impact on the field.

Recommendation: Resubmit after splitting the study into two separate manuscripts. Reviewer:

Ron McTaggart-Cowan

The authors would like to thank the reviewer for the time and care that they put into reviewing this manuscript, and for their insightful comments. We have tried to answer their comments below – with a better statement of the objective and less distractions. We will argue that both equations are important for this study. We also hope that this work motivates further flow-dependent evaluations, where these equations could be a useful reference. In our replies below, numbering for lines, figures, tables and equations refer to the revised manuscript unless otherwise stated.

General Comments

1. This manuscript presents a huge amount of material and it is clear that an awful lot of work has gone into this analysis. However, I think that the vast array of content actually reduces the potential impact of the study. Stronger curation of the information would focus the manuscript – and the reader – on the truly important elements of the work that lead directly to the conclusions. One way to start improving the focus of the study will be to identify and clearly state the objective of the work. That could effectively be done at the start of the last paragraph of the introduction. I

encourage the authors then to take a serious look at each element of content and decide whether or not it is essential to advancing the manuscript towards this objective. Components that do not fit into this focus should be removed and could probably form the basis for a separate submission.

Both reviewers made this point, so we have worked to make the statement of objectives clearer in the revised Introduction. We did give careful thought to the structure of the manuscript. In particular, we did try to cater for specialists and non-specialists by putting technical details in the appendices. This has also clearly failed, and we now have tried to improve the structure by using sign-posts in the main text. In the light of the reviewers' comments, we have removed discussion of the connection to the Liouville equation, and now only show one of the case studies. The appendices have also been reduced.

2. In the end, I think that this is really two papers. The first paper is about ensemble-estimated uncertainty growth rates and their relationship to cyclone intensification and/or trough amplification over the western North Atlantic. The second paper is about documenting and identifying the source of overdispersion in the ECMWF ensemble in the North Atlantic storm track. Although the second is clearly motivated by the first, these topics are separate enough that they would not even need to be a two-part submission: they could be treated entirely separately. Having two separate papers would allow for an expansion of discussions and dynamical interpretations, in addition to the introduction of important material into the main text that is currently relegated to the multiple appendices. I really think that the prodigious amount of effort that clearly went into this analysis would be much better served by two independent submissions.

Clearly there are several parts to this study. We have thought hard about splitting these up, but they really are strongly connected. Equation 3 (left hand side) is used to produce maps (and animations) of growth-rates. We focus on synoptic scales because these show the largest contribution to the overall variance growth over the first two days – as shown in Fig. B1. The maps demonstrate that growth-rates are concentrated into specific synoptic flow situations – identifying the extratropical flow-types others have linked to poor forecast skill. Hence, we suggest that it is a useful diagnostic for the study of flow-dependent predictability. This motivates the investigation into cyclogenesis, the title of the paper "the cyclogenesis butterfly", and the evaluation of flow-dependent reliability. Note that the growth involves diabatic effects and interactions between all scales – as demonstrated by Eq. 3 right-hand-side. We now make it clearer that this will depend on the scales of initial uncertainty, which will be different for intrinsic predictability studies, the representation of model uncertainty, and any use of initial SV perturbations. The question then arises as to whether other models display the same growth rates – we show that they are not that similar. To get a view of which might be best, we compare the models in terms of an extended spread-error equation. Following the other reviewer's advice, we now include all four models in this comparison. After showing that the ECMWF ensemble is over-spread in cyclogenesis events, we go on to investigate what might be done to improve this with a set of sensitivity experiments. These experiments are discussed in terms of Eq. 3 and suggest we could reduce the use of singular vectors. This would give us a more seamless system, and thus allow a better evaluation of model uncertainty and the key physical/dynamical processes driving the growth rates. We have tried to strengthen the motivation and links between the sections to justify keeping this as a single study.

3. [This comment is only directly relevant if the current submission is not split into two separate manuscripts.] Organizing the paper into 11 sections is highly unusual. Although I appreciate the use of sections and subsections as important tools for organizing content, I think that in this case there

are so many sections that readers will lose the "big picture" of the manuscript's organization. To a certain extent, the excessive number of sections appears to be a symptom of a stream-of-consciousness design. Rather than presenting the work in the order that it was executed, consider reorganizing it into larger logical chunks for the reader. For example, the extremely short Data section (2) should be augmented to include the methods currently described in sections 4 and 6, and part of section 8. It seems like sections 3, 5 and 10 would be more logically grouped as a single (case study) section with appropriate subsections. Sections 7, 8 and 9 should also be considered subsections of a "full-season" analysis section. The result would be a 5-section paper: (1) introduction, (2) data and methods, (3) case studies and sensitivity tests, (4) full-season analysis and model intercomparison and (5) conclusions. I believe that such a reorganization would really help to increase the potential impact of this study on the field.

We apologise if the study comes across as a stream of consciousness. We thought hard about the structure, and it is not presented in the order it was done. With several methodologies used in the study, we did not think it appropriate to place these all into a single section near the beginning, but rather discuss these only once they had been motivated. It is considered important to introduce the case studies at the beginning (now only one shown) because this gives a concrete example of what we go on to aggregate. However, it does not seem appropriate to place the sensitivity studies before the over-spread had been identified, because only then do solutions to the over-spread need to be considered. Furthermore, the sensitivity studies point to future avenues of research. We have tried again to improve the flow of the paper by grouping subsections, so that there are now only 6 sections, and including better motivation. We hope this is acceptable to the reviewer.

4. The two case studies appear to yield similar results. If the current document is to be revised as a single submission, one of the two case studies could be relegated to supplemental material. The main text could then claim demonstrable robustness with reference to the results shown in the supplement. If the material will be split into two independent studies (General Comment #2), then the two case studies could be retained in the first paper, along with augmented evaluation and interpretation.

   We have followed the first suggested course of action here, and think this does lighten the manuscript, thank you.

5. I think that a study of finite perturbation growth rates that cites the "butterfly effect" should mention Durran and Gingrich (2014), although I understand that the perturbation scales discussed here are much larger than the near-truncation scales found to be "unimportant" in the 2014 study (indeed, you mention this in your 2018 BAMS article). Perhaps this suggests that the "cyclogenesis butterfly" is a bit of a misnomer and (although catchy) might introduce some confusion: these are **very** big butterflies.

   Yes, these are big butterflies. As discussed above, our justification comes from the observation that growth rates in the first few days are largest at synoptic scales, and these growth rates appear to be orchestrated by specific synoptic flow-types. We are not discussing the Butterfly Effect (interpreted as an intrinsic predictability limit) and we make this clearer now. We did cite the Durran and Gingrich paper in the original manuscript, but this paper is cited more extensively now.

6. Based on the time periods discussed in the case studies, I think that "rapid cyclone deepening" would be a better description of the uncertainty precursor than "cyclogenesis". Both cyclones form 1-2 days before the period of interest, but intensify rapidly over the Gulf Stream. I think that the

distinction is important particularly in this region, where secondary cyclogenesis (i.e. the formation of a cyclonic circulation where none existed previously) is common and could easily be misunderstood to be the "butterfly". Clarifying the focus on rapid deepening of preexisting cyclones (if I am right about that) further emphasizes the fact that this study is looking at synoptic-scale uncertainty seeds, rather than the potentially mesoscale cyclone development precursors.

It seems difficult to define where a wave becomes a pre-existing system. In the composites, we identified cyclone structures, but the initial conditions were taken two days prior to these. The variance power-spectrum shown in the figure above indicated that the largest contribution to ensemble variance at initial time is from scales around 400 km (the diagonal lines indicate variance power per linear-unit distance on the x axis). It is likely that interactions between these mesoscale uncertainties, along with interactions at larger scales, play a role in the synoptic uncertainty that develops by day 2. We have added more discussion of this (see also Appendix B).

7. Although the breakdown of the Lagrangian growth rate into "non-conservative" and "advective" components (Eq. 1) is interesting, it does not seem to have any impact on this work. The analysis appears to proceed to look at only the Lagrangian growth rate itself, i.e. the l. h. s. of Eq. 1 rather than the forcing terms. If this is true, then the focus of the manuscript can be tightened by removing Eq. 1 and associated discussions, including most of appendix B (the remainder should be included in the augmented "Data and Methods" section, particularly if Z250 is adopted throughout as recommended in General Comment #13).

   We agree that the right-hand side of Eq. 3 was only briefly discussed before. We have added further discussion of the right-hand side of Eq. 3 – in particular, we emphasise how it represents multi-scale interactions and discuss the relationship to PV advection and generation within cyclogenesis events. The impact of multiscale interactions is evident in the blue shaded region of the figure shown above, where the representation of smaller scales in the 4 km experiments leads to increased ensemble variance even at scales already represented in the 18 km model.

8. The study references animations periodically. This means that readers will need to interrupt their progress to look at animations available in supplemental material. As far as I can tell, most of the relevant information could be presented as additional panels in the existing figures. For example, Figs. 2 and 3 are both single panel, but could be augmented to show other lead times to avoid the need for references to separate animations in the text.

   We have dropped the second case study but, for the TIGGE comparison, we show an additional date (Fig. 4) where the models' growth rates agree better. This allows for a more balanced discussion of their comparison which better represents the many cases in the animation.

9. Differences in the ensemble perturbation techniques between the different modelling systems investigated here seem potentially important, particularly given the short lead time. The use of SV perturbations in ECMWF ENS distinguishes it from most other systems in the TIGGE database, other than perhaps JMA. A discussion of these differences (or at least their itemization in an introductory table) would be very useful.

   We have included a table (Table 1) of these details.

10. This study looks at uncertainty (ensemble spread) growth rates from the perspective of synoptic cyclone dynamics. To make a convincing connection between the uncertainty growth and cyclone development it would be very useful to compare the former to something like the moist baroclinic growth rate (e.g. Booth et al. 2015; ASL). A high degree of correlation between the two would be

good evidence of the importance of rapid cyclone deepening to spread growth in the ensemble. Even something relatively simple like comparing the time series of area-averaged (over the Gulf Stream region) ensemble growth rates and moist baroclinic growth rates (with rapid deepening events identified) would provide a really nice dynamically based assessment of the importance of cyclone development to uncertainty.

This is an important idea. The authors are aiming to address this more fully in a future study focused on Eq. 3. We have discussed how this relates to cyclone development from the PV perspective in Sect. 3.2 and discuss the potential for future work in Sect. 6.

11. The maximum uncertainty growth region in Fig. 2 is upshear of the trough axis, where vorticity advection is negative aloft. Why is this? In both cases (Figs. 2 and 3) the cyclone is located between the dipole in growth rates, not at all within the peak growth rate south of the trough. This is not “ahead of the base of the upper-level trough” or “preceeding cyclogenesis” (line 142). I understand that some amount of spatial smearing arises from the use of 12-h differences to compute the growth rates, but the cyclones do not even appear to move through the maximum growth rate region. So then would it be more accurate to link large spread growth rates to amplifying upper-level troughs rather than cyclones per se? For example, perhaps uncertainties in the strength of the jet streak on the upshear flank of the trough (associated with its meridional extension) are more important than the lower-level cyclone itself.

    The ‘Lagrangian’ growth rate plotted does not include the advection of ensemble variance by the ensemble-mean flow. This advection is a major term and would immediately lead to Eulerian growth rates more aligned with the cyclone. We now make this clearer in the text. It is interesting that the Lagrangian growth rate highlights the upper-level trough region. Please see the previous response and new text in Sect. 3.2.

12. The bulk of discussions around the spread-error relationship appear to focus on the Spread and Residual terms of Eq. 2, leading to conclusions about overdispersion in the North Atlantic storm track. Is there no simpler way to arrive at the important conclusions of the study without going through this rather complicated derivation and analysis? The interesting flow-dependent aspect of the spread-error relationship is achieved through independent stratification (currently via cluster analysis), so I think the only thing that might be lost would be the conditional bias shown in Fig. 10i. However, this bias could be evaluated directly and shown to contribute significantly to the increased RMSE in the “counterpart” cluster without resorting to Eq. 2. The apparent ambiguity of the Residual term makes the discussions surrounding Eq. 2 quite difficult to follow and appears to make it difficult to make definitive statements about sources of problems within the ensemble. If the important message to be delivered by this work relates to the flow dependent overdispersion in the ensemble, then a simpler analysis (perhaps including regional and/or flow-stratified spread-reliability diagrams) might be a more effective vehicle. However, if the current investigation is just a showcase for the analytic technique itself then (a) that should be clarified and (b) the advantages of this technique over a simpler analysis should be emphasized.

    The intention is to showcase the technique and to consider carefully the assumptions made, so that it can be used in future (flow-dependent) evaluations of ensemble forecasts as well as here. Note that we are not just showing that there is a mismatch between MSE of the ensemble mean and variance of the ensemble; we are showing that the mean bias and analysis uncertainty cannot account for the mismatch. While the stratification would highlight the importance of bias in the non-cyclogenesis cluster, it is not immediately obvious to the reader that variance in bias would be important in the non-stratified budget. If we understand correctly, we would argue that spread-reliability diagrams make the same set of assumptions. What we have done is to highlight to the reader just before Sect. 4.1 that they do not need to go through the derivation, and that the essential summary of the equation is given in Sect. 4.2.

13. The lack of PV in the TIGGE database requires the use of Z250, which appears to produce similar results (Figs. 2-5). Although I can completely understand the appeal of starting with PV in this discussion, I think that for pragmatic reasons the entire study should focus on Z250. In the Data and Methods section the rationale for this can be very clearly explained. This would only really affect current sections 4 and 5. The PV 315 diagnostics in (current) section 10 could still be used because they are separate from the growth rate discussion.

    We agree that changing between PV315 and Z250 is not ideal. Nevertheless, we consider that it is important in order to discuss Eq. 3, where the right-hand-side requires PV. There is more extensive discussion of this equation now. We have simplified the case studies, which should help a little. Emphasizing the lack of PV in the TIGGE archive is also considered important to motivate its inclusion at some point.

14. Why was the clustering approach (current section 8) preferred over a much simpler cyclone identification approach? It seems as though clusters 2 and 3 for both domains are lumped into the "non-cyclogenesis" category when the results from the two domains were aggregated. As such, this seems like a very complicated way to identify dates with cyclones in the western North Atlantic.

    It is important to have an objective approach to classification, and we wanted the data to 'speak for itself' by showing the structures that emerged. In the end, the results are probably very similar to cyclone identification (as was alluded to in the manuscript). Many of the events in cluster 2 (44 out of 75) for region 1 do find their way into cluster 1 for region 2.

15. I am not sure grammatically why "growth-rate" is hyphenated throughout. This does not seem to be a common construction.

    We now use "growth rate".

16. I do not believe that forecast "lead-time" is usually hyphenated. More generally, there appears to be over-hyphenation throughout the text. Please limit the use of hyphens and ensure that they are represented using hyphen characters rather than the current em-dashes.

    We will consult on this.

17. Please confirm that date/time formatting conforms with WCD standards.

    Thanks, it now does.

Specific Comments

18. [L45] Distinguish between the true unstable modes of the flow and the computed singular vectors (optimal tangent linear growth with limited moist physics). The note about the "linear regime" points in this direction, but it would be useful to make this distinction right off the bat.

    We have removed this link to SVs in the introduction, and now distinguish SVs from intrinsic growth rates in Sect. 3.2.

19. [L48] It would be useful to itemize some of these approximations here because the difference between ensemble spread growth and error growth rate is fundamental to this study.

The differences are more clearly discussed in Sect. 3.2. Although not itemized in a list, they are separated into the two growth rate aspects (dynamical interactions and non-conservative aspects).

20. [L52-54] The punctuation of this sentence makes it difficult to follow: consider rewording.

This has been re-worded at L51-55, and is more clearly discussed now at the beginning of Sect. 4.

21. [L54] Remove hyphen from "ensemble-mean".

This has been removed throughout.

[L55-56] Replace "Jetstream" with "jet stream", "wave-guide" with "waveguide", and "down-stream" with "downstream".

We have done this.

22. [L57-72] This "outline" paragraph is overly long and complex because it strays into "abstract" territory by summarizing results. Consider shortening this paragraph by restricting its content to section descriptions only.

This has been shortened as suggested.

23. [L58] Provide a reference for TIGGE if it is to be mentioned here. Also confirm that this acronym can be used without definition in WCD, or define it.

The reference to the Swinbank et al. paper has been moved to L49 - the first place TIGGE is referred to (after the abstract).

24. [L73] Suggest dropping the first two sentences of this section and including all dataset descriptions here so that the flow of the remainder of the text is not interrupted by them. As noted in General Comment #1, this section should be rewritten to include information about the datasets and methods used throughout the study.

Section 2 on "Models, data sources and key parameters" has been extensively re-written to describe the datasets used and details of the TIGGE models. Both reviewers have mentioned including the methods here. The problem is that it would require so much motivation up-front. We hope that the revised approach, with better sub-sectioning later-on works better now.

25. [L73] I believe that "re-analysis" is more usually "reanalysis", including in Hersbach et al. (2020).

We have changed this.

26. [L75] The forecast range of the background does not seem to be identified here or in Appendix E. It seems to be 12 h (line 136), but that should be clarified here.

It is now clarified at L187 (12 h)

27. [L77] TIGGE stands for the "THORPEX Interactive Grand Global Ensemble".

The words that TIGGE stands for have changed over the years. The "I" now stands for "international" (please see: https://www.ecmwf.int/en/research/projects/tigge)

28. [L80-83] This information would probably be better displayed as a table for easier reference in later sections.

This has been done.

29. [L84] Suggest, "These data are used …".

The re-wording has removed this issue.

30. [Fig. 1] Are the trajectories that are used to identify the WCB region extending from -24h to +24h from the analysis valid time (i.e. these are the trajectory midpoints)? Suggest using the "red hatching" term consistently in the caption, rather than "shown in red".

The WCB region shown at a particular time *t** is based on all WCB trajectories, which, according to the WCB identification method by Madonna et al. (2014), are within the layer from 800 and 500 hPa at *t**. Since this method selects trajectories that ascend at least 600 hPa in 48 h, this means that the WCB region can be, in principle, based on trajectories calculated during all 48-h periods from [*t** - 48 h, *t**], [*t** - 42 h, *t** + 6 h] to [*t**, *t* + 48 h*], but given the fact that most trajectories ascend from about 900 to 300 hPa, there will be most likely no contributions from the very early and late of these periods, and as suggested by the reviewer, the bulk of trajectories shown at *t** are expected to be from the periods [*t** - 24 h, *t** + 24 h] and 6 h earlier and later, respectively. This explanation may sound complicated, but the interpretation of the WCB region shown in the figures is not: it indicates the region where WCB trajectories, irrespective of their exact start and end time, are ascending across the mid-tropospheric layer when they produce most of the latent heating and precipitation formation.

We have changed the text to "red hatching".

31. [Fig. 1] Should the mks form of PVU be provided in the caption?

    PV is now better introduced in Sect. 2.3.

32. [L104-106] Are these the forecast experiments discussed in section 10? If so, then this is additional motivation to move that section up as a "case study" subsection.

    We have argued, in response to point 3, that the case studies (now one) need to be up-front to give concrete examples, and then revisited once the result of the over-spread has been presented. Two examples are not enough to say anything definitive about reliability, but the similarity between the two cases suggests that two are sufficient to say something about the sensitivities of ensemble variance.

33. [L108] Suggest "… uncertainty grow-rate estimate …" because the ensemble provides only an estimate of the true forecast uncertainty.

    Agreed, we change to "can be estimated as".

34. [L109] What does the "1-dimensional" restriction mean here? Would this be better identified as "scalar", or can multiple state variables be included in a 1D state vector? This is obviously important because it reappears elsewhere in the text.

    This was a mistake and both reviewers suggested different solutions. We have changed "some 1–dimensional state–measure (of the atmosphere)" to "some atmospheric parameter field". Thank you.

35. [L114] The phrase "but with a different formulation" is too vague.

    We now make this clearer on L154.

36. [L118-124] This is a very complex sentence mixes conservative and non-conservative forcings in Eq. 1. It would be more useful to split this sentence to describe the physical relevance of the terms on the r.h.s of Eq. 1 individually.

    Since this relates to potential future work, it is discussed in Sect. 6 L533-542. By this stage, the conservative and non-conservative aspects have (now) been discussed extensively.

37. [L126] Should "Equation" be capitalized here? It wasn't in section 1. I do not think that the back-reference to section 1 is very useful here because the introduction did not go into much additional detail about the Liouville equation. A citation to relevant literature would be more useful here.

We have removed the link to the Liouville equation after considering the comments of both reviewers on the topic of inherent predictability limits.

38. [L125-130] I think that this discussion is fine, but it does not seem to advance the main thread of the study. It could be dropped to reduce the length of the manuscript.

    It has been dropped, as discussed above.

39. [L132-140] This information should be contained in the captions (most of it is) and/or left for supplemental material because it disrupts the flow of the main text.

    We have included as much information as we can in the captions and point to this at L188. The discussion in the text hopefully now keeps the reader engaged where details (e.g. of filtering) are required.

40. [Fig. 2] What is the contour interval for the contours showing extreme values?

    We stated that “Contours extend the shading scheme to the most extreme values, which are indicated at the ends of the colour bar”. We now clarify a little more with “Note that orange and blue contours extend the shading scheme, with the same interval. The most extreme values are indicated at the ends of the colour bar”.

41. [Fig. 3] Should this read “Case 2”?

    Yes, it should – thank you. Note that we are now only showing one case study in the main text.

42. [L145] Is this a third case study being introduced? I think that discussion of the full-season perspective should be left for the subsequent section (in the reorganized paper).

    This text referred to an example in the animation in the supplementary material, which was mentioned on old L139. We now do not mention the animation until later (L198-207) and it is not so necessary to the flow of the paper.

43. [L145-150] These seem like “future work” suggestions that would be better left for the concluding discussion.

    Yes, agreed, it is now in the conclusions Sect. 6.

44. [L152-155] The 12-h forecasts from the TIGGE database are for ENS rather than EDA, is that correct? If so, then is it true that Figs. 4a and 5a look different from Figs. 2 and 3 not only because the field is different but also because the perturbations are different? If I understand the ECMWF system correctly, SV perturbations are not added within the EDA cycle, but are added before ENS initialization. In that case, Figs. 4a and 5a have an additional source of optimized growth. That seems to make the comparison interesting, although it is complicated by the change in diagnostic field. Would it not be surprising if the SV perturbations have little impact on growth rates in these cases? Perhaps the Z250 growth rates could be shown for Figs. 2 and 3 to make this comparison possible.

    The reviewer is correct that the ENS is used from TIGGE and that, for ECMWF this includes SVs (and SPPT) while the EDA does not include SVs (but does include SPPT). We did discuss this in the context of improving seamlessness in order to be better able to diagnose the (2-day) growth rates of the model (and its uncertainty representation) from short-range forecasts. We make this distinction clearer (L219-220). Similarities between the EDA PV315 growth rates and the ENS Z250 growth rates suggest that there are strong growth rates even without the explicit inclusion of SVs, but we note (L220-222 and L470-471) that there is a slight westward shift of the maximum growth rates in the presence of SVs. While EDA Z250 growth rates would be interesting, we feel that they would complicate the manuscript unnecessarily.

45. [L155-156] So are these case studies (particularly Fig. 4) not representative of the general behaviour of these models? If so, perhaps another case study should be chosen for this comparison.

    The main point being made is that they disagree in terms of growth rates. We now also show a case (Fig. 4) where the agreement is better, to give a more balanced impression of the animation.

46. [L172-174] The source of Eq. 2 (appendix C) should be cited at the beginning of this discussion.

    We have brought this Appendix content into the main text (as discussed above), tried to make it more accessible, and sign-posted that the reader can avoid if desired.

47. [L181-182] I have a hard time understanding a lot of this discussion and how it relates to Fig. 6. It would be great to label the lines in Fig. 6 with the names of the terms in Eq. 2 that they relate to. The lines seem to be more directly related to the discussion in Appendix C, so perhaps Fig. 6 would be more appropriate in the appendix.

    We did experiment, but it is difficult to label all the lines in Fig. 6. For example, there are two lines which contribute to the forecast variance term in the final equation, and two which contribute to the analysis variance term. We have added a couple of lines to indicate the spread of the two reliable distributions. The reviewer is correct that the figure relates more directly with the variables in Appendix Eq. C1. We have moved the derivation to the main text, alongside the figure, and made it possible for readers to skip this if they choose.

48. [L192] Does this “main additional term in the Residual” refer to Eq. C5? If so, it would be useful to cite that equation here.

    Yes, it is the last term in the old Eq. C5. This has now been brought into the main text, as discussed above.

49. [Fig. 7] The change in colour scale range for panels (n) and (o) make comparison of the plots on the bottom row difficult. With the current plotting scheme, it looks like the difference in residual is almost entirely explicable by the difference in spread, but that is not really the case (is it)? The contour intervals for values beyond the standard colour bars should be noted in the caption.

    The reviewer is correct in their interpretation. In response to the other reviewer’s recommendation, this figure now includes the four TIGGE models investigated, with no differences plotted.

50. [L219-223] It is challenging to follow this discussion because of two forward-references to a description of the variance of forecast biases. It seems like that aspect of the discussion should be introduced before this text appears. In fact, it is not clear what discussion the forward references here are actually describing (the section 9 discussion seems to take an understanding of the forecast bias variance’s impact on the Residual for granted).

    The variance in forecast bias, and its relevance are now discussed in Sect. 4.1, L294 and L303-304 (thank you).

51. [L233] It was not obvious that this is a “key question”, so hopefully a clear statement of the study’s objective(s) in the introduction will help to make that link more direct.

    Hopefully so, motivation at the beginning seems to have been a key issue. We have tried to address this with changes to the Introduction.

52. [L233-235] Does this “either-or” statement arise from the form of the Residual term (Eq. C5)? If so, then it seems like it would be useful to put this equation in the main text, hopefully as part of a

discussion on the meaning of “variance in forecast bias”, which I think might be related to the “difficulties” proposed here (?).

The either-or statement is about whether the residual is associated with specific synoptic flow types (such as those with high growth-rates as in Figs. 2 and 3) or a more general issue. For example, associated with scale interactions with planetary wave uncertainties, or other scale interactions which might be more ubiquitous. We have tried to make this clearer from L364-367.

53. [L242-243] This region is quite complex: why would three clusters necessarily “provide sufficient degrees of freedom”? The optimal number of clusters is difficult to determine, but usually dropoffs in quantities like the AIC or BIC serve as some sort of semi-quantifiable justification for the number of clusters.

The aim was to balance realism of clusters with the need to obtain a sufficient sample size. This will be a compromise, but a broad ridge, a broad trough, and a tighter cyclogenesis flow-type seem to cover most eventualities. We try to make this balance of needs clearer now (L377-378). Note that later, on L402-404, we state that “visual inspection of plots similar to those for in Fig. 1 suggests that the objective clustering has been successful in partitioning the date/times into cyclogenesis and non–cyclogenesis flow types”. In addition, the clustering was successful-enough to provide statistically significant differences between the partitioned date/times.

54. [L256-258] This is the only discussion of the uncertainty growth rate in this section, and it does not seem to lead to any particular conclusion. Is there a good reason to include it here and in the Fig. 8 and 9 plots? (It does not seem to be discussed in the subsequent section either.)

This is useful for two reasons: firstly it again demonstrates that the cyclogenesis cluster is associated with the strongest growth rates (note that most of the date/times for cluster 1, area 2 come from the date/times in cluster 2, area 1) and hence consistent with Figs. 2 and 3, and secondly it provides the opportunity to note that these growth rates are not used in the clustering (that could potentially bias the reliability assessment). We change “since this is what will be evaluated” to “since this could potentially bias the reliability assessment”.

55. [L294] The phrase “almost the entire over-spread” seems like a bit of an overstatement. It is probably more defensible in terms of variance, but could perhaps be softened to “much of the overdispersion” or similar.

Since this is one of the major conclusions of the paper, we have followed the reviewer’s suggestion, changing the text to “The overall assessment of ensemble spread is seen in the residual terms (Fig. 8e and Fig. 8j). Here it is evident for the ECMWF ensemble that most of the over–spread in the region of focus, at the western end of the North Atlantic winter storm-track (Fig. 6e), is associated with the cyclogenesis composite — with statistically significant residuals in Fig. 8e and statistically insignificant residuals (indicated by the light blue and light grey colours) in Fig. 8j. Differences are shown in Fig. 8o. They are particularly strong and significant over Newfoundland. Downstream, differences have the opposite sign — possibly associated with differences in downstream cyclogenesis, and consistent with the increased spread noted above. Linking the day 2 stormtrack over–spread (in the region of focus) to cyclogenesis is a key conclusion of this study. It does appear, therefore, that ECMWF initial growth rates (Fig. 2, Fig. 3a) associated with cyclogenesis events are too strong. The next section explores the root–causes for this problem.”

56. [L297-301] I am afraid that I do not fully understand this discussion. How would the stratification of the groups (cyclone vs. non-cyclone) be done differently with multiple seasons or an independent assessment? Could this "regression to the mean" alternatively be considered a sampling bias?

    One approach could be to deduce a composite of the (local) initial conditions leading to cyclogenesis within one season, and to then pick date/times from the same season in a different year which project strongly onto the initial condition composite. Since this approach has not been tested, it is difficult to predict how successful it would be. Regression to the mean refers to the fact that, if one sampling of a random variable is extreme, the next sampling is likely to be less extreme. In the interests of brevity, we have removed this paragraph.

57. [L302.5] Consider simplifying the section title to "Sensitivity experiments to quantify uncertainty sources".

    We have changed to the suggested title – thank you.

58. [L315] I understand that resource constraints likely make additional tests difficult or impossible, but is it not conceivable that the ordering of MU and 4K is important? Systematic changes in the physics tendencies should be expected between 18 km and 4 km grid spacing (for example as more turbulent fluxes are represented by the dynamics), which will impact SPPT directly. This might mean that the impact of switching MU on and off at 4 km is different from what is observed in the 18 km configuration. I do not think that this is a big enough deal (or close enough to the focus of the paper) to justify additional simulations; however, you may want to put a bit more nuance in the wording of this statement.

    Thank you for pointing this out. We had been thinking more about SPPT's action on diabatic processes (there is a shift from 'convective' to 'large-scale' precipitation at 4 km, but the total precipitation – and thus diabatic tendency which SPPT works on – is largely unaffected). The power spectra in Fig. B1 do highlight more resolved variance in the 4 km experiments. We have change the wording to (L452ff) "The conclusions are not thought to be sensitive to the ordering of the various modifications. For example, it will be seen that the impacts on total precipitation of DCP and +4km are small, and hence these impacts should be little changed in the presence of the SPPT form of MU. However, parametrized turbulent fluxes might be weakened with +4km, and hence this impact could be a somewhat different in the presence of MU".

59. [L318] Why not show results from the 1200 UTC 27 November 2019 initialization so that the day-2 forecast aligns with the panels shown in Figs. 1, 2 and 4?

    The sensitivity plots show the impacts at day 2 of the growth rates that occurred previously – hence the difference in time.

60. [L326] Does the upshear maximum in the SV plot (Fig. 12b) really very well described as being in the "cold sector" of the cyclone? The cold sector is defined based on low-level airstreams but here the plot is showing spread differences in Z250. I think that this is much more related to the growth of perturbations in the jet streak on the upshear side of the trough, which is contributing to the "digging" of the trough / meridional amplification. Could the upper-level jet-front structure not an ideal place to have rapid SV growth (e.g. Hakim 2000; JAS)? By increasing vorticity at the base of the trough this feature will indirectly impact troposphere-deep cyclogenesis, but I think it is possible that the origins of the spread are more local. (The same is true for the second trough over the eastern North Atlantic that appears to be approximately equivalent barotropic.)

    Thank you for these interesting suggestions. See the reply below to your comment 61.

61. [L327-329] The spatial separation of the SV and MU contributions is beautiful. I think that it is very understandable based on the previous comment and the fact that model physics is largely inactive in upper-level jet-fronts, other than perhaps some turbulence. The MU is focusing on the regions where the physics is active (lower-level cyclone and WCB) while the SV is picking up dynamic growth along the jet streak on the waveguide. If you agree with this assessment, it could be a useful inference to add to the text.

Indeed, these are interesting inferences. They are difficult to prove, but we agree with the reviewer that we should include these considerations in the description of Fig. 10 b,c. We therefore have changed the paragraph (L464-478) to:

"Figure 10a shows the OP configuration with a well–developed surface low pressure system, as discussed in relation to Fig. 1. The warm conveyor belt (WCB) associated with this cyclone is seen to lead to the development of a prominent downstream upper–level ridge and a downstream trough west of Europe. As might be expected, the maximum Z250 spread is located downstream of the maximum 'Lagrangian' growth rates (cf. Fig. 3a).

The impact of the initial SV perturbations on Z250 spread (Fig. 10b) is particularly pronounced along the western flanks of the two prominent troughs over the western and eastern North Atlantic, respectively. This likely indicates the potential for dynamic growth along the intense jets in these regions, qualitatively in line with the idealized studies by Hakim (2000). This SV impact might help explain the apparent slight westward shift of the centre of maximum growth in the ENS (Fig. 3a) relative to the EDA (Fig. 2). There are places where the SV impact on spread is half the total (so that the fraction of variance explained reaches 25%). In contrast to the SV impact, the impact of the model uncertainty (MU) representation (Fig. 10c) is particularly pronounced in the cyclone centre and in the region of the WCB ahead of the surface low, i.e., in regions where cloud-related physical processes are particularly active. The large signal along the western flank of the ridge southwest of Greenland is consistent with the results of Joos and Forbes (2016), who found a large influence of cloud microphysical processes in the WCB on the tropopause structure in this part of the downstream ridge. MU also explains up to 25% of the total variance. The remaining variance must be associated with the (deterministic) growth of initial EDA analysis uncertainty".

Hakim, G. J., 2000. Role of nonmodal growth and nonlinearity in cyclogenesis initial-value problems. J. Atmos. Sci., 57, 2951-2967.

Joos, H., and R. Forbes, 2016. Impact of different IFS microphysics on a warm conveyor belt and the downstream flow evolution. Quart. J. Roy. Meteorol. Soc., 142, 2727-2739.

62. [L328] Missing closing parenthesis for figure reference.

This has been added, thanks.

63. [L334-336] Discussion of total precipitation seems tangential to this study (also L344-345).

The observation that the total precipitation stays the same, but the spread is altered seems interesting. As discussed above, old L344-345 is useful in the argument about the ordering of the experiments not being too important.

64. [L346-351] This is the first time that observation location is discussed. The Obs experiment seems largely unrelated to the other experiments and should be eliminated to focus the study on the "controllable" sources of spread quantified in the other experiments.

The reason that the observational experiments are included is that they help us quantify what extra predictability is currently achieved through the assimilation of local observations, such as cloud-affected radiances. While they may not directly impact reliability estimates, they allow us to gauge the relative importance of working towards reducing the use of SVs. The juxtaposition here is also useful because it motivates a future more comprehensive study of the impacts of assimilating observations in cyclogenesis flow situations.

65. [L343] Suggest changing to "… appears to yield a better depiction of uncertainty than that generated by …".

    In a combined response to both reviewers, we have changed the text L490-493 to: "The impact on P315 uncertainty of allowing the model to resolve more of the convection at the 4 km resolution (Fig. 11e) appears to be in closer agreement with the response to turning off the deep convection parametrisation (minus Fig. 11d), when the model is forced to represent the convection on the 16 km grid." Please note that the quoted ECMWF nominal resolution has been changed from 18km to 16km in line with the TIGGE documentation.

66. [L343] Remove extra "km".

    Thank you.

67. [L343] This seems like a really important statement because it suggests that the huge computational cost of a 4 km ensemble is not justifiable from this perspective.

    From this perspective, agreed. As noted, 16 km is not the scale to resolve convection.

68. [L374] Although they can likely be inferred, neither baroclinic nor convective instabilities were demonstrated in the analysis.

    Agreed, but we feel that the inference from baroclinic development and convection back to their respective instabilities can be assumed.

69. [L382] This conclusion does not seem as direct as it ought to be. Perhaps "could" should be replaced with "should"?

    We have changed to "should" at L568.

70. [L382-383] This seems like a fairly weak and somewhat confusing statement on which to end the manuscript. Moist singular vectors would be implemented in the TL/AD forms of the model, and as far as I know are quite independent of the SPPT-based model uncertainty estimate. Perhaps this discussion could instead be extended to consider the SPP-based uncertainty formulation as a look into the future ECMWF system.

    We have made a more direct statement now. The point we were trying to make was that it would be good to explore the idea of focussing model uncertainty on potential instabilities, rather than the effects of already-triggered instabilities. The former is more like what SVs are doing, although SVs would be costly to calculate at every timestep. It is possible that SPP could evolve into targeting these potential instabilities if the perturbed parameters were the triggering thresholds, for example. Since submitting the original manuscript, SPP has become more competitive with SPPT, and looks likely to be implemented at ECMWF in the near future. We have changed the wording to "It is possible that such a focus on instabilities (rather than the effects of already-triggered instabilities) might be better explored within the future 'stochastically perturbed parameter' (SPP) framework for model uncertainty — perturbing triggering thresholds for example."

71. [L392] Is a ^2 missing on the l.h.s of definition of the variance?

Yes, thank you for spotting this. The change is made at L151.

72. [L444] Why would the squared terms necessarily dominate, particularly if there are correlations between the constituents of the cross terms?

    In general, the terms are uncorrelated. However, we no longer state that we would expect the squared terms to dominate, but rather say that the cross terms are included in the epsilon term, and then point to the Appendix where they are quantified.

73. [L465-469] Providing a quantitative assessment of the relative size of each of these terms seems like it would be useful, particularly because the Residual is one of the (two) leading terms assessed in the text is key to conclusions regarding overdisperison.

    These terms were estimated in an earlier draft of this manuscript, but this was removed for brevity. A shortened version is now included in the appendix of the re-submitted manuscript.

74. [Appendix D] Why is a new field (500 hPa height) and season (JJA) introduced just for this appendix? I guess it might be to show the robustness of the analysis, but I think that the text on L228-232 distracts from the main message of the study. In a two-paper solution (General Comment #2), this figure and discussion could form the basis for a short subsection instead.

    The figures were meant for the supplementary material. We have removed them now, and simply state L356-558 “Note that conclusions drawn in this section appear to generalise to other parameters (such as geopotential heights and temperatures at 500 hPa), other seasons, other stormtracks, and continue until the most recent check for the March — May season 2022 (not shown).”

    Please note that, for old Fig. 7, a small bug was identified. The diurnal averaging aspect (discussed in the old figure caption) was implemented in a way that the error terms in the budget were decreased slightly (error of the diurnal average rather than diurnal average of the error). This has been corrected by removing the diurnal averaging. The impact is so small that it does not affect the conclusions.

    The authors would sincerely like to thank the reviewer for their insight and diligence in reviewing this manuscript. We feel that changes made have led to very useful improvements.

Reviewer 2

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

The paper investigates ensemble forecast reliability at 48h lead time, mainly in the ECMWF system with some comparison to other centers. To do so, a new spread-error budget is derived. It is found that the ECMWF ensemble is overspread in stormtracks in the winter season and it is argued that this is related to cyclogenesis events. In my opinion the core topic of this work is interesting and worth being published since it contradicts the intuitive expectation that cyclogenesis is associated with bad forecasts and low predictability. However, the paper requires a substantial revision. It is too long, difficult to read and not well structured. It spends a lot of time on tangent discussions and detailed case analyses, which I find distracting and misleading. Especially the discussion of the butterfly effect is imprecise, irrelevant and in part incorrect. On the other hand, topics more directly related to reliability and ensemble forecasting systems are very little discussed, if at all.

We appreciate the reviewer's comments and their time taken to review this manuscript. We attempt to address their specific comments below. In our replies below, numbering for lines, figures, tables and equations refer to the revised manuscript unless otherwise stated.

Major comments:

"The cyclogenesis butterfly"

This term, given already in the title, is never properly defined. It is introduced by phrases like "here we think of it as...". I am still not clear what is actually meant by this. The paper investigates reliability, which is an aspect of practical predictability, while the "real" butterfly effect refers to the existence of an intrinsic predictability limit caused by scale interaction in a multi-scale system (see Lorenz, 1969 and Palmer etal, 2014). Current weather prediction system are started from initial condition uncertainties that are much larger than butterflies and are on average far away from hitting the intrinsic limit (e.g. Zhang etal, 2019). The existence of singular vectors are not a manifestation of the butterfly effect, since they are still consistent with infinite predictability due to their constant growth rate. Judt, 2018 (Fig. 8b, day 0-2) for example has demonstrated the extreme increase in the error growth rate if the atmosphere really is perturbed with "butterflies" only. I am however not recommending to discuss the butterfly effect more precisely in this paper but rather to remove this discussion and to focus on much more relevant aspects with respect to (practical) reliability, like the long-standing underdispersive problem of ensemble forecasts and various methods that have been used to mitigate it (e.g. EDA, singular vectors, breading vectors, SPPT, SPP, etc.). It is probably a shortcoming in one or several of those methods that leads to the reliability problem that the paper investigates.

Sorry for the confusion here. We are not talking about the "real butterfly effect", intrinsic predictability limits, or Lorenz, 1969. To be fair, we did not mention the butterfly effect, but rather stated that our butterflies were defined as "local flow configurations where the chaotic and exponential growth–rate of uncertainty is particularly strong" (old L32). We also discussed SVs as indicating that divergence of trajectories within state–space is not uniform over the attractor (old L31), rather than as any manifestation of the butterfly effect. Our reference to Lorenz was to his 1963 paper – discussing sensitivity to initial conditions but not any intrinsic limit to predictability. Our use of the term "Cyclogenesis butterfly" is, rather, an attempt to encourage more flow-dependent thinking in the evaluation and development of forecast models. However, the potential for confusion is recognised, and we now ensure the reader does not get the wrong idea from the outset.

Figure B1 shows power spectra at day 0 and day 2 for the ECMWF ensemble. The maximum initial (EDA) variance contribution is from waves around 400 km, while the maximum D+2 variance contribution is from waves around 1500 km – i.e. synoptic scales. Notice that the initial variance is saturated at scales smaller than about 100 km. This plot motivates/justifies our interest in the growth of variance at the synoptic scales to D+2 – due to physics and dynamical interactions at all scales.

Incidentally, the Palmer et al. paper uses the current lead-author's previous example to speculate that intrinsic predictability limits might be longer due to the potential confinement of error-growth to intermittent flow regimes. While we did not discuss intrinsic predictability limits, the idea that certain flow-types can organise multi-scale interactions and focus error growth is very much in the spirit of the current study.

While lack of reliability might be explainable by shortcomings in the mitigation methods mentioned by the reviewer, this is not obviously the case. For example, sensitivity of parametrized convection to uncertainties in the resolved flow might be important.

To improve brevity, we have removed discussion of links to the Liouville equation. The thought was that the use of model uncertainty represents a route by which NWP could converge on the true dispersion rates on the real-world attractor (as initial uncertainty is reduced) but, reflecting on the reviewer's comments, it may be the case that NWP will never be able to say anything definitive about the real butterfly effect, as it pertains to the real world.

Reliability in a larger context

The specific overspread that is found over the Northern Atlantic stormtrack in winter (Fig. 7) has not been put into a wider context. If the system is reliable on average there must be compensating underspread somewhere, e.g. over the continents, outside the midlatitudes or in the other seasons. This should also be discussed, as well as the question if the system really is reliable on average at the considered 48h lead time. Some information can be found in appendix D, but I think this discussion should be central in the paper. Furthermore I am wondering what the downstream consequences of the stormtrack-overspread are. Does the overspread persist beyond the end of the stormtrack into the continent in a lagrangian sense, e.g. is the 5 day forecast for Europe in the winter season also overspread? Finally, what is the relation to forecast busts? According to Lillo and Parsons, 2017, East coast cyclogenesis has the potential to generate particularly bad forecasts over Europe. This kind of contradicts the (average) results from this paper. Possibly one season of data is not enough to investigate this but some discussion here would be helpful.

There is some compensation elsewhere. We now state L345-347 "Note that Rodwell et al. (2018) indicated better reliability for this model. Partly this reflects compensation in their annual and hemispheric means, partly it reflects the importance (here) of accounting for bias and analysis uncertainty, and partly it reflects a recent deterioration in stormtrack reliability".

At day 5, the over-spread is still evident but general interactions and the loss of any continued SV contribution make this less clear. The emphasis of this study is on the short timescales because these are the only timescales where agreement between ensemble members is sufficient to be able to make meaningful calculations of their dispersion rates. We do not see any contradiction with the Lillo and Parsons paper – cyclogenesis clearly results in large spread and deterministic forecast busts. The Lillo and Parsons paper is a major reason why we investigated cyclogenesis. We are demonstrating that we are over-spread, but we do not dispute that the spread should be large in these situations.

More use of TIGGE

While for the case studies 4 centers have been compared, the spread-error budget comparison is only done between the ECMWF and the UKMO system and finally the clustering analysis is only done for the ECMWF system. A reason for this is not given. I think the paper may miss an opportunity here to investigate possible reasons for the (ECMWF) overspread since the different centers use different methods to generate their ensemble. Hence I recommend to include more centers throughout the paper, particularly in the clustering analysis to see if the cyclogenesis overspread is a more general problem or specific to ECMWF.

We avoided a detailed 'beauty contest' because this can be problematic in manuscripts. For example, making sure that the details of the ensemble initialisation over the period of interest are documented correctly, and checking whether there are other factors to consider. Nevertheless, on the reviewer's advice, the variance budget has been calculated for the four models discussed, and this figure replaces the previous one. It suggests that the JMA model is more over-spread. While this model also uses SV perturbations, we prefer to focus on the SV aspect in the ECMWF model alone. Applying the clustering analysis to all the models would include a lot more plots and require even more discussion. The primary aim of the study was the evaluation and understanding of uncertainty growth rates in the ECMWF model; the TIGGE models were there primarily for context.

More focus

The paper spends a lot of time with a detailed discussion of two cases which in my point of view gives little insight. Furthermore, the paper oscillates between analyzing the cases and the entire winter and also between theta- and pressure-level analysis or squared and non-squared metrics, which I find confusing. With Eq. 1 the paper introduces a rather sophisticated diagnostic which later is not used at all. I think this is not a good use of the time of potential readers. The information given in this paper is distributed over 8 sections, 5 appendices, 17 figures and additional supplementary material. I suggest the authors should consider condensing the paper to the essential parts and keeping analysis and methods consistent across the paper.

Equation 3 is central to the study. Plotting it demonstrates that large growth rates are confined to particular flow features. The right hand-side shows that the growth rate (for the PV variable) represents the effects of uncertainties in diabatic processes and in non-linear dynamical scale-interactions. This is now discussed more thoroughly in Sect. 3.2 and revisited in the conclusions (thank you for pointing out this omission). It is hoped that Eq. 3 will feed into subsequent work examining these aspects in more detail. For the growth rate plots, only one case is now presented – with one PV growth rate example for ECMWF (which relates directly to the equation). Two dates are shown for the TIGGE models to give a better reflection of the range of results seen in the animations (which highlight similarities and differences). The sensitivity study (only one shown now) is considered important as it highlights the relative roles of model uncertainty and singular vectors, and points to possibilities for future development. This is now better discussed.

Specific comments:

L1:

This statement is incorrect (see major comment about the butterfly effect and comment below).

The abstract has been re-written to emphasise the focus on NWP. The differences with intrinsic predictability are made clear.

L21:

The Liouville equation as formulated in Ehrendorfer, 1994 assumes that the propagation operator is known and constant. Hence it cannot describe growth due to model uncertainty.

As discussed above, we have removed reference to the Liouville equation.

L26:

I would not say that EDA represents model uncertainty. The model uncertainty is rather part of the assimilation process to generate the initial condition ensemble.

Yes, we are saying that model uncertainty representation is included in the EDA system. The EDA and ECMWF system in general are now described in Sect. 2.1 in the “Models, data sources and key parameters” section.

L31:

This statement is incorrect. Lorenz-type butterflies, i.e. small-scale and small-amplitude perturbations limit predictability via scale interactions and not only due to chaos and strong sensitivity to initial conditions (see Lorenz, 1969 and especially Palmer etal., 2014). Hence the constantly growing singular vectors do not represent this "real" butterfly effect. Furthermore, current errors and uncertainties in forecasting system are neither small in scale nor small in amplitude and cannot be regarded as butterflies. If they were this would mean that current systems operate already now at the intrinsic limit, which is not true (e.g. Zhang etal. 2019). I agree that in some situations error growth in current forecasts is worse than average but if this might be related to the butterfly effect in rare cases is an open question.

As discussed above, we were not implying that SVs represented the Butterfly Effect. We were saying that they demonstrated that growth rates are not uniform. The introduction has been re-written to make the topic of this work clearer, and the role of SVs is discussed in Sect. 3.2.

L51:

I am not sure if I understand correctly what you mean with "cyclogenesis butterfly". The term is never clearly defined.

We now make this clearer in the abstract, at L39-47, and in Sect. 3.2.

L51:

"The key question is..."

This is a big gap in the line of argument and comes as a surprise to me. Please consider rewriting the introduction to focus on this question and the importance and flow-dependence of reliability and the need to extend the "spread-error" relationship.

The question is in the title. Clearly, we needed to work harder on motivating this in the introduction, and we have done this now.

L57:

The paper outline contains to many details in my opinion. Some should have been mentioned and discussed in the introduction, some are results.

We have shortened the outline to remove details and results.

Sec. 2, Data:

Since an entire section is dedicated to describe the data I would prefer that all the relevant details are given here rather than being distributed over the rest of the paper and the appendix. With respect to the other centers, only resolution and ensemble size are given but potentially interesting differences in the ensemble design are not mentioned and later not investigated (see major comment above).

We now increase the discussion of the ensembles from the other centers and include the initialisation details available within the TIGGE archive in a table. This section has been re-titled "Models, data sources and key parameters". We feel that it does not make sense to describe methodologies until the need for them has been motivated. Hopefully, with the better sub-sectioning throughout the manuscript, this is acceptable.

L74, caption Fig. 2:

What do you mean by "background ensemble/forecast"?

This is described L82-83 in Sect. 2.1 on the ECMWF forecast system (previously it was in Appendix E).

L90:

The arguments given about the case selection are very vague. How are they related to the key question? Are these cases in which the forecast was particularly unreliable? Or in which the cyclogenesis was very rapid?

This is a good point. We stated why the cases were chosen but did not make it clear that other attributes had not been a factor. The cases were not chosen because they were unreliable (difficult to establish for a single case) or that cyclogenesis was particularly rapid, or that uncertainty growth rates were unusually large. They were simply chosen to motivate to the reader the kinds of events we are considering, and for their suitability for the sensitivity experiments ("without being strongly affected by other flow perturbations in their environment"). We now make this clearer from L128.

Fig. 2-4:

It is unclear, which forecast lead time you show and why you chose to focus on this particular forecast lead time.

The information for these growth rate figures was given in Appendix B (Further details on the growth rate plots). We appreciate that this is not ideal and have now brought Appendix B partly into the main text in Sect. 3.3, and partly in the captions to Fig. 2 and 3.

To answer the reviewer's question, Fig. 2 is constructed using the 12 h background forecasts from the EDA (so no lead-times are greater than 12 h), started at 6 and 18 UTC. The fields shown are based on centred-means and differences between consecutive hourly lead times. The 24 h running-mean temporal filter then places the smoothed fields back on the whole hours. More specifically, Fig. 2 shows fields centred at 12 UTC on 29 November 2019. For the winds (including humidity fluxes), PV315 contour, the standard deviation in the growth rate parameter (PV315) and its advection, the lead times used are thus {28 Nov 18 UTC + 6,7,8,9,10,11,12 h}, {29 Nov 06 UTC + 0,1,2,3,4,5,6,7,8,9,10,11,12 h}, and {20 Nov 18 UTC + 0,1,2,3,4,5,6 h}. The ensemble-mean precipitation in a given hour and the time derivative in the standard deviation of PV315 are based on the differences (in precipitation accumulations and PV315 standard deviations) between these lead times {28 Nov 18 UTC + (7-6),…,(12-11) h}, {29 Nov 06 UTC + (1-0),…,(12-11) h}, and {20 Nov 18 UTC + (1-0),…,(6-5) h}.

For Fig. 3, again the lead times used are no greater than 12 h. The differences are that the forecasts are started at 00 and 12 UTC and data are only available at 6 h intervals. Hence for Fig. 3, the lead times used are thus {29 Nov 00 UTC + 0,6,12 h} and {29 Nov 12 UTC + 0,6,12 h} and the running mean has length 4 rather than 24.

We focus on these shortest lead times possible so that the ensemble members are as close as possible to each other, in particular representing the same synoptic systems, and hence the growth rates are the best flow-specific growth rates we can calculate. We now make this motivation more clearly. On L184-186 we now state “To understand how growth rates depend on the synoptic flow situation, it is useful to consider very short leadtimes when all ensemble members are representing essentially the same synoptic flow situation. A natural choice is to use the short background forecasts from ensemble data assimilation”.

L109:

I don't understand what you mean be 1-dimensional state-measure. Substitute 4-dim atmospheric field?

This was a mistake and both reviewers suggested different solutions. We have changed “some 1–dimensional state–measure (of the atmosphere)” to “some atmospheric parameter field”. Thank you.

Eq. 1:

sigma_hat is now the standard deviation of the PV, right? Use P_hat instead of sigma? The hat is not explained, same meaning as in eq. 2?

Sigma is an accepted way of representing the standard deviation, and the hat signifies that this is an estimator. Yes, it has the same meaning throughout. This is now all made clear in the text (L144-145).

I suggest to add an index i to P, P' and NC to indicate that these are quantities from individual members.

The derivation of Eq. 2, which was in Appendix A, has been brought into the main text. The notation has been explained better now. Subscripts are initially used (L149-151) and the meaning of the overline is more fully explained, including for non-linear terms (L150-152). We feel that the current approach of using an overline and no subscripts to signify a mean is neater and consistent with the standard approach for signifying the mean of a linear quantity. We use this approach throughout the paper. The use of subscripts in non-linear terms can indicate “Einstein notation”, when the summation over the subscript is implicit, and no overline is required. We do not want to cause this confusion for readers.

L120-L130:

Needs more introduction and explanation. However, this diagnostic is not used in the paper anyway. Consider removing it (see below).

Following the reviewer’s first major comment, we have dropped the discussion of the Liouville equation (old L125-130). We feel that old L120-124 are important in the discussion of the processes that can lead to the enhanced growth rates. This discussion has been improved with an explanation of the terms in

Sect. 3.2. These would be the (non-linear, scale interactive) processes that act on the initial uncertainty (including applied SVs) and the perturbations introduced by the model uncertainty. The more prospective aspects discussed at old L120-124 have been moved to the conclusions in Sect. 6, with a pointer from the end of Sect. 3.2.

L142:

"often preceding cyclogenesis", "occur within strongly precipitating WCBs":

These statements are rather vague and are either obvious or seem speculative. Do you mean that the growth rate is correlated with the amount of precipitation in the cyclone? And with cyclogenesis do you mean a depending of the trough where the growth rate peak occurs or does it lead to a cyclogenesis downstream?

This statement is an observation about the animation. We did not intend to make any link to a deepening of the trough or downstream cyclogenesis – simply the colocation of the growth rate with the base of the upper-level trough (we didn't anticipate the spatial interpretation of the word 'preceding'). The cluster mean results (Fig. 7) tend to confirm that there is enhanced growth at the base of the trough. We now make this link more clearly at L391-392 associated with the clustering. Hopefully the paragraph at L198-207 is better written now.

L145:

"Further investigation..."

I find the following statements distracting. But more importantly, the reader might have invested some time to understand eq. 1, the relevant papers and the appendix only to find out now that you are not investigating the right hand side of eq. 1 at all and leave this for future work. Hence I recommend to remove section 4 and appendix A and just state here that you are plotting a lagrangian growth rate.

We do now discuss the right-hand side of Eq. 3 more fully, as explained above. We direct the reader to more discussion about future work in the Conclusions and Discussion section (so that it does not distract so much here).

L152:

It is very confusing that you switch now to geopotential growth rates. I understand that PV is not in TIGGE. But what is the point of showing the PV growth rates first, especially since you did not explore the right-hand side of eq. 1, which for me is the main purpose of using PV? I suggest to stick with Z250 then and to omit the PV-plots.

We agree that changing between PV315 and Z250 is not ideal. However, we do not wish to leave the reader under the impression that it is all about SVs and model uncertainty – other modelling aspects will also be acting through the right-hand side of Eq. 3, in the PV context, to generate the spread. Emphasizing the lack of PV in the TIGGE archive is also considered important to motivate its inclusion at some point.

L157:

"... are very evident"

I actually was surprised to see how bad the agreement is. Not much is said however about where these discrepancies come from, L160 makes a very general statement.

We agree that the differences in growth rates are surprisingly large. Following a recommendation from the other reviewer, we now include a table summarising the initialisation procedures of the other models and show the extended spread error equation for these models. We have also included another example (Fig. 4) which shows a case of better agreement. This helps give a more balanced impression of the possibilities in the full DJF 2020/21 season.

L161:

"It would be useful..."

Please clarify what you mean. Uncertainty growth rates cannot be close to the truth since the truth is not uncertain.

Our wording was quite vague. We were referring to our prior expectation of the truth. The revised text L232-236 avoids discussion of the truth here. In particular, we have changed the wording from “which is closest to the truth” to “While it is difficult to evaluate these growth rates per se, it is possible to assess how well each ensemble system maintains short–range statistical reliability within the North Atlantic stormtrack”.

L167:

You state the essential information as e.g. in brackets. I suggest to change that and maybe write down an equation. Also I suggest to be more precise what average means (case average, area average, ensemble average).

We have worked hard to improve this introduction to reliability (the beginning of Sect. 4), including showing the equation suggested.

L174 (also L168):

I suggest to replace "ensemble forecast start times" with "(large number of) cases". Also please explain the symbols first and discuss the visualization afterwards. The notation is inconsistent with eq. 1, maybe express the ensemble mean with <..>.

The key issue here seems to be the definition of ensemble members and ensemble forecasts. We have now made this much clearer at the beginning of Sect. 4 (L246-249). If one looks through the literature, an overbar is generally used to denote a mean, regardless of what it is a mean over. Conversely, “<..>” is often used to denote an inner product. Hence, we would prefer to use an overbar throughout, including in both Eq. 3 and Eq. 9.

L188:

Is mu_A equal to the truth? And is mu_F also equal to the truth? I suggest to not discuss this in the figure caption.

These parameters are displayed in the figure and are different from the truth. We hope that by bringing the derivation together with the figure into the main text, this will become a lot clearer – since old Eq. C1 was the equation which related directly to the figure.

L196:

I find this square-root operation just for "more understandable units" confusing, especially since it introduces the complication with the residual and you admit that small contributions look larger than they actually are. Moreover, in the supplementary figure you switch back to square units. I suggest to keep the squares in every plot.

Reliability is about bias as well as spread (and all other moments too of course). Hence having bias in its correct units is valuable. In addition, a panel showing spread in squared units seems pretty meaningless per se. At L321-322 we now say that "While smaller terms will look more important than they are in the squared budget, the residual still correctly indicates spread deficiencies".

Sec. 7:

I wonder why you switched from comparing 4 centers to now only 2. Is there a reason for that? And why did you choose to compare with the UKMO?

We were trying to avoid a beauty contest, together with the difficulties in getting all the historical initialisation details correct. Since both reviewers have asked, we do now include all 4 models in Fig. 6, and have discovered that the TIGGE archive provides the necessary initialisation information.

L233:

In Rodwell et al, 2018 you showed (Fig. 1) that the (traditional) spread-error relation is perfectly matched for the Northern Hemisphere at any forecast lead time over an entire year (2014). Hence (if this is still true, is it?) the overspread you now show for the stormtracks in the winter must be compensated by an underspread at some other location or some other season. It would be interesting to investigate this (see major comment above).

We have replied to the major comment above. We discuss this in the revised manuscript.

L238:

I suggest to explain the K-mean clustering method with a couple of sentences.

We include this explanation up-front, with the text L369-370 "which seeks to minimise the sum of squared deviations from the relevant cluster-mean".

L245:

Why do you weight with the root? Isn't the grid cell area scaling with cos(lat)?

Because the clustering method is on the variance, which re-instates the square.

L264:

Does this mean you are combining the cluster1 cases from both clustering areas? Could you further justify this approach? The shift of the region doesn't seem that large. Would one clustering analysis based on a combined region lead to similar results? What about separating clusters by the surface pressure tendency in the region? Wouldn't this be a simpler method more directly related to cyclogenesis?

Yes, we are combining cluster 1 cases from both clustering areas to produce a 'cyclogenesis composite'. We have made this clearer now. We used the clustering approach (and the smaller regions) to give some coherence in flow structures, and to be consistent with the fields displayed in Figs. 3 and 4 (and the animations). We state at L376-377 "It is the ability to cluster on structures which motivated the choice of the K–means approach". To a large extent, the results justify the approach – we see the structures in the clusters and obtain statistically significant differences between the two composites. Other approaches, such as the one suggested by the reviewer, could also be successful and might be amenable to the use of a single combined region (indeed the whole storm-track), but the structure aspect might not be so well constrained. We suggest that this could be tested in future studies.

L293:

To me this statement seems a bit exaggerated. I would say that the overspread is reduced in the cyclogenesis composite. Also if I read the colors correctly, the residual difference does not reach statistical significance. Why is the overspread enhanced in the counterpart over the central/east Atlantic. Is it because cyclogenesis is shifted downstream in the counterpart cases? Is this spread reduction in cyclogenesis events also happening at other centers (see major comment above)?

We have moderated and extended the text at the end of Sect. 4.5 to say "The overall assessment of ensemble spread is seen in the residual terms (Fig. 8e and Fig. 8j). Here it is evident for the ECMWF ensemble that most of the over–spread in the region of focus, at the western end of the North Atlantic winter storm-track (Fig. 6e), is associated with the cyclogenesis composite — with statistically significant residuals in Fig. 8e and statistically insignificant residuals (indicated by the light blue and light grey colours) in Fig. 8j. Differences are shown in Fig. 8o. They are particularly strong and significant over Newfoundland. Downstream, differences have the opposite sign — possibly associated with differences in downstream cyclogenesis, and consistent with the increased spread noted above. Linking the day 2 stormtrack over–spread (in the region of focus) to cyclogenesis is a key conclusion of this study. It does appear, therefore, that ECMWF initial growth rates (Fig. 2, Fig. 3a) associated with cyclogenesis events are too strong. The next section explores the root–causes for this problem."

We have not examined the other models with this composite approach. This would require more work and a lot more explanation.

L296:

I did not understand this paragraph.

This paragraph was a note about conditional sampling. In the interests of brevity, we have removed the paragraph. The key text remains at L392-393 "(Note that the growth rate is not used within the clustering algorithm since this could potentially bias the reliability assessment)".

L313:

I notice you leave the convection scheme on at 4km resolution. Could you explain why? Usually only shallow convection is used at such high resolutions (e.g. Judt, 2018).

The 4 km experiment with the parametrization of deep convection turned off was also run. Turning off this parametrization at 4 km leads to a further enhancement in D+2 uncertainty in PV315 along the southern extreme of the cold front. It was concluded that, even at 4 km, forcing the ECMWF model to resolve convection can be unrealistic. This is an area of active research by others. We now cite Wedi et al. (2020) in the Introduction Sect. 1.

L324:

I don't understand this sentence.

We have changed the sentence 'The Z250 spread represents the temporal integral of the local tendency, and hence the effects of "material" generation and advection." to "As might be expected, the maximum Z250 spread is located downstream of the maximum 'Lagrangian' growth rates (cf. Fig. 3a)."

L333:

"attempts to resolve". This is misleading since the resolution is still 18km, right?

There is an understandable misunderstanding in our use of “resolve”. We have changed this to “that would otherwise be created when the model is forced to represent this convection on its 16 km grid”. Thank you. (Note that the nominal resolution is quoted now as 16 km to be consistent with the other models in the TIGGE documentation).

L338:

I am not sure about the relevance of this increased spread. It could just be a consequence of slightly displaced and explicitly resolved updrafts. Would there also be any enhanced spread in e.g. the precipitation averaged over the front?

We agree with the reviewer, and have modified our tentative explanation from “The smaller scales associated with PV315, and its sensitivity to vertical gradients in diabatic heating in the upper troposphere might help explain this sensitivity to the increase in resolution” to L487-488 “At 4 km, the model attempts to resolve more of the convection. The resolved convection can be associated with stronger updrafts, which might perturb the tropopause more vigorously, where PV gradients are particularly strong.” There is likely to be enhanced spread in precipitation within the frontal region, although we have not looked at this. It is difficult to say whether the precipitation averaged over the front will be more or less uncertain.

L342:

Possibly there are now explicitly simulated updrafts which are slightly displaced among the ensemble members and generate grid-pointwise spread. Again I am not sure how relevant this is. I don't think this is the kind of uncertainty SPPT was designed to account for. So I am not surprised to see less effect from SPPT in this region.

This comment is very thought provoking. We have changed the text on “Forcing the model to explicitly resolve this convection (even if at the wrong ~18 km scale) appears to better locate the uncertainty with that generated by the ~4 km model” to L490-493 “The impact on P315 uncertainty of allowing the model to resolve more of the convection at the 4 km resolution (Fig. 11e) appears to be in closer agreement with the response to turning off the deep convection parametrisation (minus Fig. 11d), when the model is forced to represent the convection on the 16 km grid.”

L379:

If a reduction of singular vectors would make forecasts more consistent then why are they still used? I suspect they do show a benefit at a different location, flow regime, lead time, etc. I think you should discuss these aspects in more detail and also possible alternatives (e.g. inflating SPPT or EDA, using SPP, higher resolution, etc). See also major comment above.

The SVs are particularly efficient at generating spread (by day 2) at synoptic scales (see the power spectra in Fig. B1). Recent spread-error spectra produced by the lead author suggest that this spread is now somewhat too large. The MU also produces spread at synoptic and planetary scales – at the planetary scales the spread-error agreement is much better. Hence this is evidence that a reduction in the magnitude of the SVs would be useful. We do now discuss the advantages of the SPP framework in the modified final sentence L569-571 “It is possible that such a focus on instabilities (rather than the effects of already-triggered instabilities) might be better explored within the future ‘stochastically perturbed parameter' (SPP) framework for model uncertainty – perturbing triggering thresholds for example."

L381:

This would only make sense if the observed increased spread with resolved convection does not mainly result from rather small displacements of individual updrafts. But this has not been investigated (see related comments above).

Please see our reply to the reviewer's comment on L342.

Minor comments:

Fig. 1:

The figure is hard to read and evaluate. I suggest to use a color for the >2PVU regions and to omit the red hatching, since it is kind of obvious and does not add any extra information. As labels of the panels I suggest "Case 1", "Case 1, +24h" or something like this for easier identification.

We only show one case now. We consider the red hatching to be important as not all readers will consider the colocation with precipitation to be obvious. The key southern PV=2 PVU contour seems clear to us.

Fig. 2 and others:

There is a lot of doubling between the figure caption and the text. I suggest to not repeat details in the text that are already included in the caption.

We have tried to eliminate duplication – this is easier now because most of the explanations are brought from the appendices into the main text.

L132:

"Single frame of animation" is not a good description of the plot.

This has been changed, thanks.

Fig. 3 caption:

I think you meant case 2.

Yes thank you.

L175:

Better "sampling from a population"?

We prefer "underlying distribution", as that is what is drawn in the figure. However, they mean the same.

L178:

Any reason why you call this "departure"? It is just a difference, isn't it?

The word departure comes from the world of data assimilation. It is used instead of "error" because the analysis (or observation) does not represent the truth. All the lines are differences.

L180:

"number of forecasts" is ambiguous. I suggest "cases".

We have better defined what we mean by “number of forecasts” now.

L182:

I suggest to remove the 2-superscript (looks like a footnote). It is clear from the context that you are considering squared quantities.

These superscripts are used to distinguish with the rooted terms. WCD discourages footnotes.

L191:

Remove (.

Done, thank you

Fig. 7 (and others):

The color bars are misleading to me. First I would appreciate if the color bars in panels a-j and k-o were identical. The main point of the figure is the comparison and identical color bars will help with that. Also it is misleading to color small positive value with saturated dark colors (e.g. panel c, it looks like a massive AnUnc). I suggest to use reddish colors for positive values, blueish colors for negative values and gray/neutral around zero (e.g. like you did in panel d).

The magnitude of these terms depends on the reliability of the forecast and the lead time. In some configurations, the errors (departures) and spread might be an order of magnitude larger than the bias and residual but, of course, the important difference between the departures and the spread (the measure of unreliability) has the same order of magnitude as (at least one of) the bias and residual. Hence, we really need to be able to see these latter terms in some detail. Moreover, the bias and residual can be of either sign, while the departures and spread are always positive. Hence, there is no clear reason to plot these terms with the same contour intervals.

Fig. 8:

The green geopotential lines are very hard to see. Please make them more prominent.

We have thickened these.

L273:

"Head of the stormtrack". This term is not clear to me. From the context I guess you mean the start/beginning (west).

We have made this clearer with L410: “western end of the stormtrack”.

Fig. 11:

I find the arrows more confusing than helpful. Also: CP->DP

We will change to DCP! We have experimented and feel that the arrows help. They convey the sense (sign) of the difference. We now state L452: “Vertical arrows in Fig. 9 indicate the sign convention of the difference to be plotted”.

Fig. 12:

I suggest to also revise the color bar. Panel a) shows a positive variable and should use neutral to reddish colors. Panels b)-f) should have the same color bar since this makes it much easier to assess the individual contributions. I cannot really distinguish gray from black contours.

We make it clear throughout where the contour intervals change – we feel it is important to see where a particular change has an impact, even if it is small compared to that of another change.

L328:

) missing.

We have done this, thanks.

Please note that, for old Fig. 7, a small bug was identified. The diurnal averaging aspect (discussed in the old figure caption) was implemented in a way that the errors terms in the budget were decreased slightly (error of the diurnal average rather than diurnal average of the error). This has been corrected by removing the diurnal averaging. The impact is so small that it does not affect the conclusions.

The authors would very much like to thank the reviewer for the time and insight that they have given to this review process. We feel that changes made have led to very useful improvements.

References:

Edward N. Lorenz (1969) The predictability of a flow which possesses manyscales of motion, Tellus, 21:3, 289-307

T. Palmer et al, 2014: The real butterfly effect. Nonlinearity 27 R123

Judt, F. (2018). Insights into Atmospheric Predictability through Global Convection-Permitting Model Simulations, Journal of the Atmospheric Sciences, 75(5), 1477-1497.

Zhang, F. et al, 2019: What is the predictability limit of midlatitude weather?. Journal of the Atmospheric Sciences, 76(4), 1077-1091.