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 – in particular 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, all line numbers refer to the originally submitted manuscript

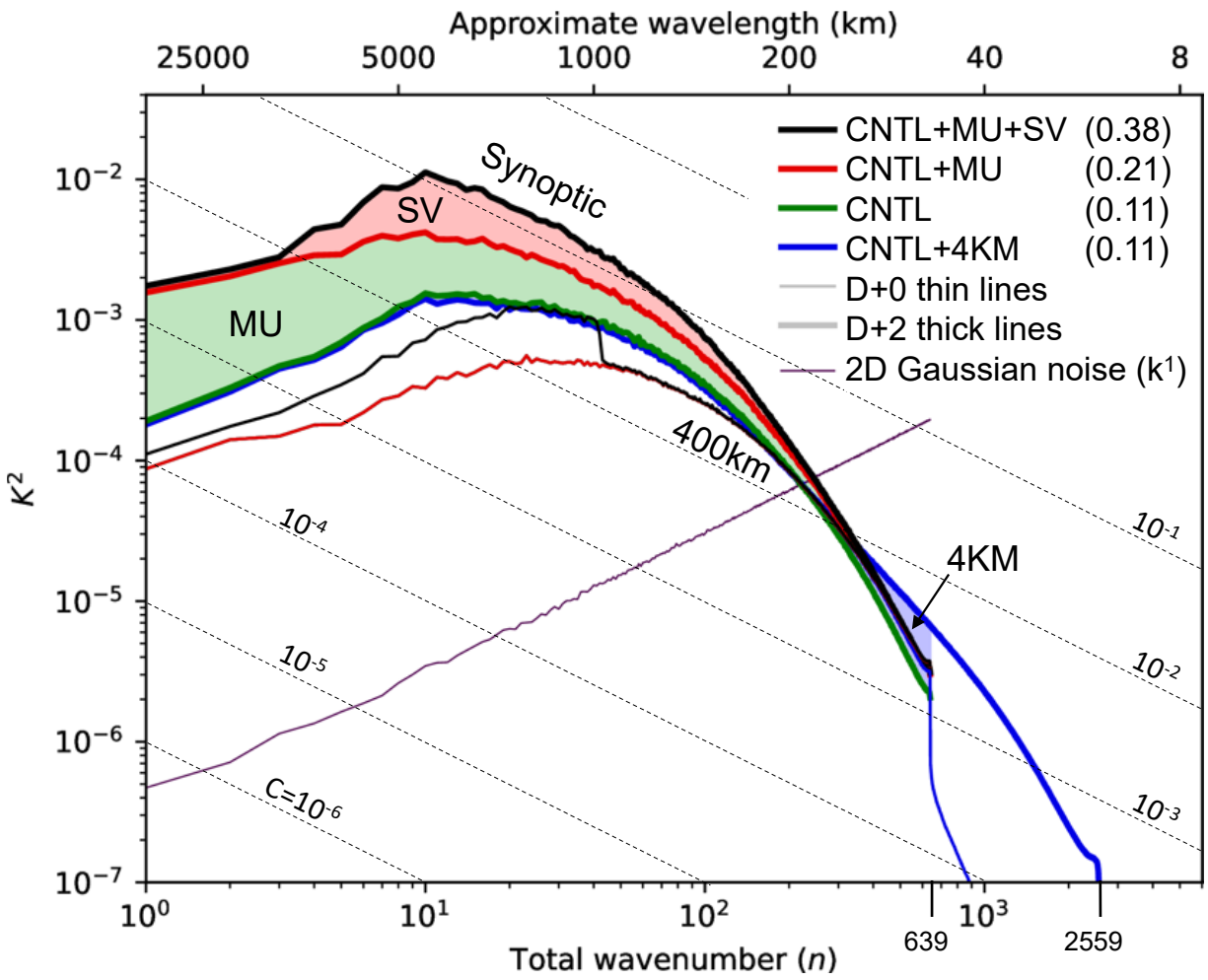
General Comments

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.

The broad objective of the work is in the title of the manuscript and this was outlined further on lines 49-52. However, both reviewers made this point, so we have worked to make the statement of objectives clearer. 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, but they really are strongly connected. Equation 1 (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 the attached variance power-spectrum plot. Note that this growth involves diabatic effects and interactions between all scales – as demonstrated by equation 1 right-hand-side. These maps demonstrate that growth-rates are concentrated into specific synoptic flow situations. This motivates the investigation into cyclogenesis and the title of the paper "the cyclogenesis butterfly". 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 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 these give concrete examples 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 and including better motivation, and 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/do cite the Durran and Gingrich paper.

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.

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. It is likely that interactions between these mesoscale uncertainties, along with interactions at larger scales, do play a role in the synoptic uncertainty that develops by day 2. We have added more discussion of this.

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 equation 1 was only briefly discussed before. We have added further discussion of the right-hand side of equation 1 – in particular we emphasise how it represents multi-scale interactions. 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 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 now only show one case study, but include an extra panel for a different lead time. This should avoid the need to refer to supplementary material.

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 will include these and cite the TIGGE archive documentation.

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 equation 1, as alluded to in lines 145-149 of the original manuscript.

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. We briefly discussed this in the manuscript. It is intended to be the focus of future work.

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.

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 equation 1, where the right-hand-side requires PV. 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 will consult on this.

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.

We will consult on this.

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 will do.

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.

We will do.

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

We will try to improve this.

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

The hyphen here has been fairly standard, but happy to go with the WCD style on this.

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

We will do.

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 Swinbank et al. paper was referenced at L77. We will bring this forward.

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.

Both reviewers have mentioned including the methods here. The problem is that it would require so much motivation up-front. We have emphasized this argument in the text, and signpost appropriately. We hope this satisfies both reviewers.

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

We will change 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 here (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.

We will do this.

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

Done thanks.

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 will change text to "red hatching"

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

We will change "2 PVU" to "2 PVU (=2 x 10^{-6} m^2 s^{-1} K kg^{-1})". It probably only makes sense to do this if we also define isentropic PV as $-g\zeta\frac{\partial\theta}{\partial p}$ in the main text, so we will do this as well.

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 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 say "Following, initially, the derivation of Baumgart and Riemer (2019), the growth rate of this measure can be related to sources of uncertainty growth via the equation".

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.

Without further study (which is anticipated) we cannot say much more about the relative importance of each of these aspects. We have given a citation for each aspect, but will also make the text more readable.

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.

Much of this detail is required to explain Figs. 2 and 3. This is because the figures show the growth rate after the spatio-temporal filter has been applied. We have placed the text in a methodology sub-section.

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 "Contours extend the shading scheme (with the same interval) to the most extreme values, which 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 refers to the animation in the supplementary material, which was mentioned on line 139. We now re-emphasise that this is in the supplementary material. We would be happy to share this animation with the reviewer.

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

Yes, agreed. This will be placed in the discussion, and we will simply say here that the growth rates could be model-dependent and then ask the final question "How well do growth rates agree amongst the TIGGE models?"

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 will make this distinction clearer. 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. 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 mentioned the slightly better agreement for some other cases for completeness. The full range of agreement/disagreement can be seen in the animations that we will provide with the revised paper.

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

We will bring this forward in the text.

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, and two which contribute to the analysis variance term. Lines are also conditional on the truth. The reviewer is correct that the figure relates more directly with the variables in Appendix equation (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 (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 forward reference is to the paragraph on lines 283-293 (in Section 9). In Section 9, the variance of the forecast bias (at least the "inter–cluster variability in forecast bias" – line 290) is explicitly represented (i.e. in the bias term, rather than in the residual). We have now changed the text of line 291 to say "explicitly represented in the bias terms shown in Fig. 10d and Fig. 10i". This is a subtle but important point in motivating the use of flow-dependent evaluations. We hope, with the derivation now in the main text, and further text refinements, that this will be clearer to the general reader.

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.

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.

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 clearer now. Note that later on, on line 265, we stated that "visual inspection of plots similar to those for the two case studies (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. We now make this point too.

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 "much of the over-spread in the region of focus, at the head of the North Atlantic winter storm-track (Fig. 7e), is associated with the cyclogenesis composite – with statistically insignificant residuals (largely light blue and light grey) in Fig. 10j and statistically significant residuals (largely dark blue and dark purple) in Fig. 10e. Differences are shown in Fig. 10o. Over Newfoundland in particular, they are comparable in magnitude to the full departures in Fig. 10a and statistically significant. Downstream, differences in the residual have the opposite sign – possibly associated with differences in downstream cyclogenesis, and consistent with the increased spread noted above."

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 the figure on page 3 of this document do highlight more resolved variance in the 4 km experiments. We will add the words "(the impact of increased resolution might be somewhat different in the presence of model uncertainty – because the parametrized turbulent fluxes could be weaker)".

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. 12 b,c. We therefore will change the paragraph (L322-330) as follows:

"Figure 12a 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 one might expect, the maximum spread is not co-located with the maximum 'Lagrangian' growth rates (cf. Fig. 4a). The impact of the initial SV perturbations on Z250 spread (Fig. 12b) is particularly pronounced along the western flanks of the 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). 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. 12c) 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 (L334-336) that the total precipitation stays the same but the spread is altered seems interesting. As discussed above, L344-345 is useful in the argument about the ordering of the experiments not being too important. We will add the words to the effect "and hence the impact of model uncertainty (in the form of SPPT) should be less sensitive to this aspect of increased resolution".

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 do 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 …".

We have changed to "The impact on PV315 uncertainty of allowing the model to resolve more of the convection at the 4km resolution (Fig. 13e) appears to be in closer agreement with the response to turning off the deep convection parametrization (minus Fig. 13d), when the model is forced to represent the convection on the 18km grid".

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, 18 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".

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 will change the wording to the effect "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.

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. We have reinstated our estimation of the cross-terms which might be non-zero (please see below).

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 will be 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 "Note that the over–spread is not specific to Z250 or the North Atlantic stormtrack (not shown)".

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 useful improvements.