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

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.
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.

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.
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.
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.
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.

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).
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.
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.
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.
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.
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.

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.
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.
15. I am not sure grammatically why "growth-rate" is hyphenated throughout. This does not seem to be a common construction.
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.
17. Please confirm that date/time formatting conforms with WCD standards.

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.
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.
20. [L52-54] The punctuation of this sentence makes it difficult to follow: consider rewording.
21. [L54] Remove hyphen from "ensemble-mean".
22. [L55-56] Replace "Jetstream" with "jet stream", "wave-guide" with "waveguide", and "down-stream" with "downstream".
23. [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.
24. [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.
25. [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.

26. [L73] I believe that “re-analysis” is more usually “reanalysis”, including in Hersbach et al. (2020).
27. [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.
28. [L77] TIGGE stands for the “THORPEX Interactive Grand Global Ensemble”.
29. [L80-83] This information would probably be better displayed as a table for easier reference in later sections.
30. [L84] Suggest, “These data are used …”.
31. [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”.
32. [Fig. 1] Should the mks form of PVU be provided in the caption?
33. [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.
34. [L108] Suggest “… uncertainty grow-rate estimate …” because the ensemble provides only an estimate of the true forecast uncertainty.
35. [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.
36. [L114] The phrase “but with a different formulation” is too vague.
37. [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.
38. [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.
39. [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.
40. [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.
41. [Fig. 2] What is the contour interval for the contours showing extreme values?
42. [Fig. 3] Should this read “Case 2”?
43. [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).
44. [L145-150] These seem like “future work” suggestions that would be better left for the concluding discussion.
45. [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.

46. [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.
47. [L172-174] The source of Eq. 2 (appendix C) should be cited at the beginning of this discussion.
48. [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.
49. [L192] Does this "main additional term in the Residual" refer to Eq. C5? If so, it would be useful to cite that equation here.
50. [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.
51. [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).
52. [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.
53. [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 (?).
54. [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 drop-offs in quantities like the AIC or BIC serve as some sort of semi-quantifiable justification for the number of clusters.
55. [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.)
56. [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.
57. [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?

58. [L302.5] Consider simplifying the section title to “Sensitivity experiments to quantify uncertainty sources”.
59. [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.
60. [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?
61. [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.)
62. [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.
63. [L328] Missing closing parenthesis for figure reference.
64. [L334-336] Discussion of total precipitation seems tangential to this study (also L344-345).
65. [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.
66. [L343] Suggest changing to “… appears to yield a better depiction of uncertainty than that generated by …”.
67. [L343] Remove extra “km”.
68. [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.
69. [L374] Although they can likely be inferred, neither baroclinic nor convective instabilities were demonstrated in the analysis.
70. [L382] This conclusion does not seem as direct as it ought to be. Perhaps “could” should be replaced with “should”?
71. [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.

72. [L392] Is a ^2 missing on the l.h.s of definition of the variance?
73. [L444] Why would the squared terms necessarily dominate, particularly if there are correlations between the constituents of the cross terms?
74. [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.
75. [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.