Comment on wcd-2022-6

Anonymous Referee #2

Referee comment on "The Cyclogenesis Butterfly: Uncertainty growth and forecast reliability during extratropical cyclogenesis" by Mark John Rodwell and Heini Wernli, Weather Clim. Dynam. Discuss., https://doi.org/10.5194/wcd-2022-6-RC2, 2022

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.

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.

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.

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.

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.

Specific comments:
L1:
This statement is incorrect (see major comment about the butterfly effect and comment below).

L21:
The Liouville equation as formulated in Ehrendorfer, 1994 assumes that the propagation operator is known and constant. Hence it can not describe growth due to model uncertainty.

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.

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.

L51:
I am not sure if I understand correctly what you mean with "cyclogenesis butterfly". The term is never clearly defined.

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.

L57:
The paper outline contains to many details in my opinion. Some should have been mentioned and discussed in the introduction, some are 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).

L74, caption Fig. 2:
What do you mean by "background ensemble/forecast"?

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?

Fig. 2-4:
It is unclear, which forecast lead time you show and why you chose to focus on this particular forecast lead time.

L109:
I don't understand what you mean be 1-dimensional state-measure. Substitute 4-dim atmospheric field?

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

L120-L130:
Needs more introduction and explanation. However, this diagnostic is not used in the paper anyway. Consider removing it (see below).

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 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?

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.

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.

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.

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.

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).

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 <..>.

L188:
Is μ_A equal to the truth? And is μ_F also equal to the truth? I suggest to not discuss this in the figure caption.

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.

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?

L233:
In Rodwell etal, 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).

L238:
I suggest to explain the K-mean clustering method with a couple of sentences.

L245:
Why do you weight with the root? Isn't the grid cell area scaling with cos(lat)?

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?

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)?

L296:
I did not understand this paragraph.

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).

L324:
I don't understand this sentence.

L333:
"attempts to resolve". This is misleading since the resolution is still 18km, right?

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?

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.

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.

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).

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.

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.

L132:
"Single frame of animation" is not a good description of the plot.

Fig. 3 caption:
I think you meant case 2.

L175:
Better "sampling from a population"?

L178:
Any reason why you call this "departure"? It is just a difference, isn't it?

L180:
"number of forecasts" is ambiguous. I suggest "cases".

L182:
I suggest to remove the 2-superscript (looks like a footnote). It is clear from the context that you are considering squared quantities.

L191:
Remove (.

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).

Fig. 8:
The green geopotential lines are very hard to see. Please make them more prominent.

L273:
"Head of the stormtrack". This term is not clear to me. From the context I guess you mean the start/beginning (west).

Fig. 11:
I find the arrows more confusing than helpful. Also: CP->DP

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.

L328:
) missing.

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.