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

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

We appreciate the reviewer's comments and their time taken to review this manuscript. We attempt to address their specific comments below. In our replies, all line numbers refer to the originally submitted manuscript

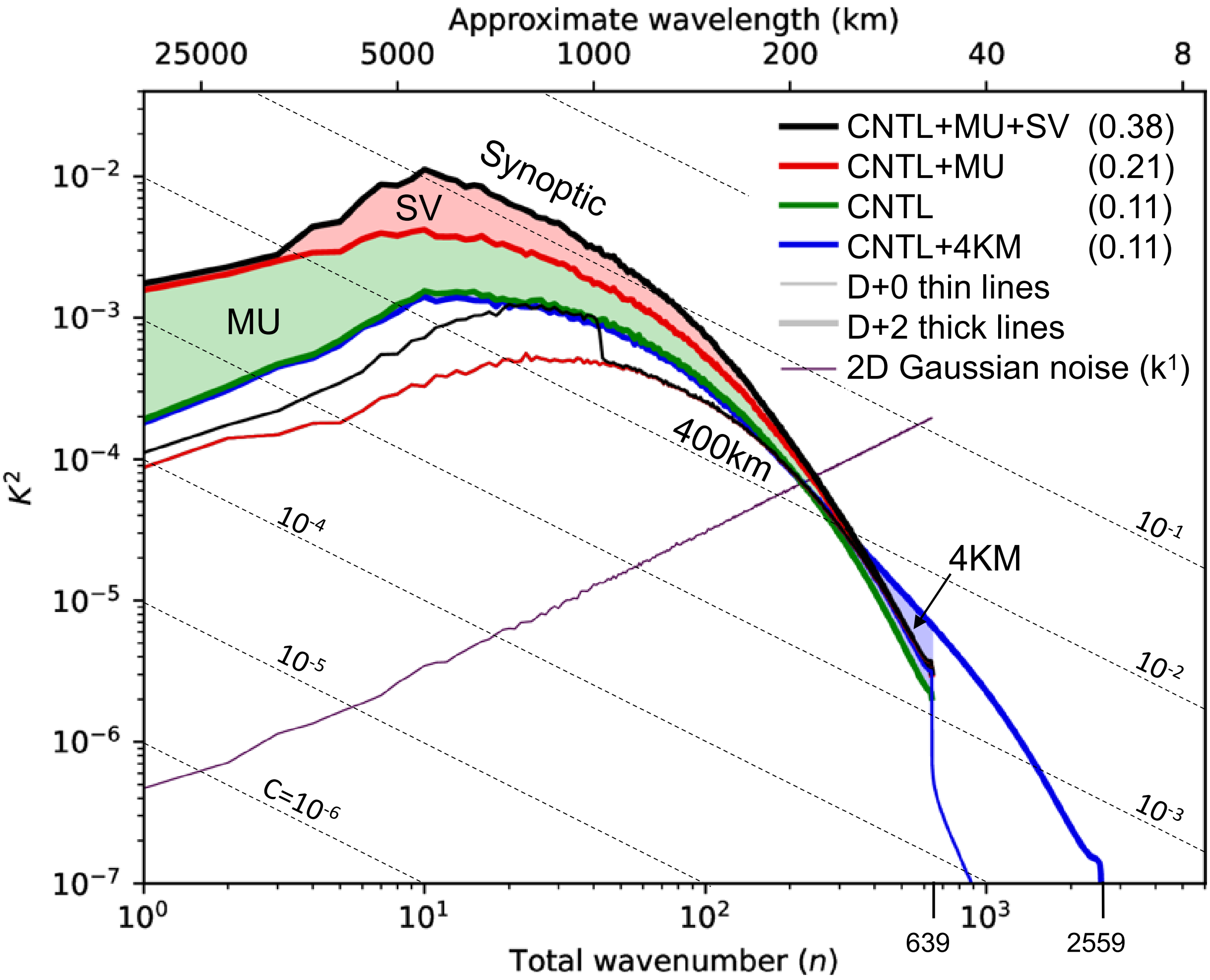
Major comments:

"The cyclogenesis butterfly"

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

Sorry for the confusion here. We are not talking about the "real butterfly effect", intrinsic predictability limits, or Lorenz, 1969. To be fair, we did not mention the butterfly effect, but rather stated that our butterflies were defined as "local flow configurations where the chaotic and exponential growth–rate of uncertainty is particularly strong" (line 32). We also discussed SVs as indicating that divergence of trajectories within state–space is not uniform over the attractor (line 31), rather than as any manifestation of the butterfly effect. Our reference to Lorenz was to his 1963 paper – discussing sensitivity to initial conditions but not any intrinsic limit to predictability. Our use of the term "Cyclogenesis butterfly" is, rather, an attempt to encourage more flow-dependent thinking in the evaluation and development of forecast models. However, the potential for confusion is recognised, and we now ensure the reader does not get the wrong idea from the outset.

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



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

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

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

Reliability in a larger context

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

There is some compensation elsewhere. The authors' previous paper on flow-dependent reliability shows good agreement in the annual-mean hemispheric-mean sense between RMSE and ensemble standard deviations. The current work (e.g. Fig. D1) demonstrates that this agreement may not have been so good (at day 2) if bias and analysis uncertainty had been factored-in. The lead author has also found that things have deteriorated (in the stormtracks) since then. For T850, some of the compensation in the operationally calculated scores actually occurs under Tibet, and this is screened-out in the diagnostics developed here. These details seemed to the authors to be of little long-term practical use to readers on this study. At day 5, the over-spread is still evident but general interactions and the loss of any continued SV contribution make this less clear. The emphasis of this study is on the short timescales because these are the only timescales where agreement between ensemble members is sufficient to be able to make meaningful calculations of their dispersion rates. We do not see any contradiction with the Lillo and Parsons paper – cyclogenesis clearly results in large spread and deterministic forecast busts. The Lillo and Parsons paper is a major reason why we investigated cyclogenesis.

More use of TIGGE

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

We avoided a detailed 'beauty contest' because this can be problematic in manuscripts. For example, making sure that the details of the ensemble initialisation over the period of interest are documented correctly, and checking whether there are other factors to consider. Nevertheless, on the reviewer's advice, the variance budget has been calculated for the four models discussed, and this figure will replace the previous one. Applying the clustering analysis to all the models would include a lot more plots and require even more discussion. The primary aim of the study was the evaluation and understanding of uncertainty growth rates in the ECMWF model.

More focus

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

Equation 1 is central to the study. Plotting it demonstrates that large growth rates are confined to particular flow features. The right hand-side shows that the growth rate (for the PV variable) represents the effects of uncertainties in diabatic processes and in non-linear dynamical scale-interactions. This is now revisited in the conclusions (thank you for pointing this omission). It is hoped that equation 1 will feed into subsequent work examining these aspects in more detail. For the growth rate plots, only one case is now presented – with one PV growth rate example for ECMWF (which relates directly to the equation) and one comparison of TIGGE model growth rates (which highlight similarities and differences). The sensitivity study is considered important as it highlights the relative roles of model uncertainty and singular vectors, and points to possibilities for future development. The lack of a clear signal associated with assimilation of local observations is consistent with the fact that the variance at small scales is largely saturated in the initial conditions – the grid-point skill is coming from the larger scales which are adequately constrained in the 4D Var without the local observations. This is now better discussed.

Specific comments:

L1:

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

We now state that "Global numerical weather prediction is often limited by particular flow features which are associated with pronounced uncertainty growth rates".

L21:

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

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

L26:

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

Yes, we are saying that model uncertainty representation is included in the EDA system. The missing piece of information seems to be that we do not mention that we are referring to the ECMWF EDA system. We now say "… in particular, at ECMWF, the ensemble data assimilation system (Isaksen et al., 2010) aims to represent flow–dependent error covariances in the background or "first–guess" (Bonavita et al., 2016) and in the observations (Geer et al., 2018), as well as model uncertainty (MU; Buizza et al., 1999) and a model grid's lack of representativity of point observations (Janjic´ et al., 2018)".

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.

We have changed the text “predictability is often limited by the Lorenz–type butterflies in the flow (Lorenz, 1972). Here we think of these as local” to “prediction is often limited by specific”.

L51:

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

We now make this clearer, as discussed above.

L51:

"The key question is..."

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

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

L57:

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

We have shortened the outline to remove details and results.

Sec. 2, Data:

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

We now increase the discussion of the ensembles from the other centers and include the initialisation details available within the TIGGE archive. This section has been re-titled “data sources”. We feel that it does not make sense to describe in detail how this data is used at this point (the parameters, lead-times etc.) – before the methodologies have been motivated. We have tried to indicate this structure better now, with better sub-sectioning throughout the manuscript.

L74, caption Fig. 2:

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

We now say “Depiction of cyclogenesis Case 1, as represented in the background forecasts of the EDA”. In response to the other reviewer’s comments, we have dropped most of the explicit references to the animations here. We still consider these animations to be very useful, and we would still like to provide them as supplementary material – we would be happy to share these with the reviewer if a means can be found (they are quite big files).

L90:

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

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

Fig. 2-4:

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

The information for these growth rate figures was given in Appendix B (Further details on the growth rate plots). We appreciate that this is not ideal and have now brought Appendix B into a subsection in the main text (with a note to say that this is for the interested reader and could be skipped).

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

For Figs. 4 and 5, again the lead times used are no greater than 12 h. The differences are that the forecasts are started at 00 and 12 UTC and data are only available at 6 h intervals. Hence for Fig. 4, the lead times used are thus {29 Nov 00 UTC + 0,6,12 h} and {29 Nov 12 UTC + 0,6,12 h} and the running mean has length 4 rather than 24.

We focus on these shortest lead times possible so that the ensemble members are as close as possible to each other, in particular representing the same synoptic systems, and hence the growth rates are the best flow-specific growth rates we can calculate. We now make this motivation more clearly. For example, at L138, we change "and highlight the features associated with enhanced growth rates" to ". Because ensemble members are synoptically very close to each other at short lead times, we can identify the synoptic features associated with enhanced uncertainty growth rates".

L109:

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

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

Eq. 1:

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

Sigma is the standard deviation, and the hat signifies that this is an estimator. This is now made clear in the text. We do not use P_hat since we use the same formulation of the left-hand side of (1) for other quantities (in particular, Z250). Yes it is the same meaning as in (2).

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

The derivation of (1), which was in Appendix A, has been brought into the main text. Here subscripts are initially used (e.g. L389-392). We feel that the current approach of using an overbar and no subscripts to signify a mean is neater and consistent with the standard approach for signifying the mean of a linear quantity. We use this approach throughout the paper. We appreciate that this can initially cause confusion and so where the variance of the P_i is defined on L392 (there is a superscript 2 missing here), we spell-out more clearly that the subscripts are dropped even for non-linear terms: “Note that we drop the subscripts below the overbar for non-linear as well as linear quantities so, e.g., $\overline{ab} \equiv \frac{1}{m}\sum_{i=1}^{m} a_i b_i$.

L120-L130:

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

Following the reviewer’s first major comment, we have dropped the discussion of the Liouville equation (L125-130). We feel that L120-124 are important in the discussion of the processes that can lead to the enhanced growth rates – these would be the (non-linear, scale interactive) processes that act on the initial uncertainty (including applied SVs) and the perturbations introduced by the model uncertainty. We did not discuss this fully enough previously, and have rectified this in the new Conclusions and Discussion section, with a pointer from L120-124 to this discussion.

L142:

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

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

This statement is an observation about the animation. We did not intend to make any link to a deepening of the trough or downstream cyclogenesis – simply the colocation of the growth rate with the base of the upper-level trough. The results (Fig. 8 and 9) tend to confirm that there is enhanced growth at the base of the trough. We now make this link more clearly at L257. We haven’t focussed on the WCBs so cannot confirm any correlation with precipitation amount – we change the text slightly to say “They also seem to occur within strongly precipitating WCBs”.

L145:

"Further investigation..."

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

We do now discuss the right-hand side of (1) more fully. We direct the reader to more discussion in the Conclusions and Discussion section (so that it does not distract so much here).

L152:

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

We agree that changing between PV315 and Z250 is not ideal. However, we do not wish to leave the reader under the impression that it is all about SVs and model uncertainty – both these aspects will be acting through the right-hand side of (1), in the PV context, to generate the spread. Emphasizing the lack of PV in the TIGGE archive is also considered important to motivate its inclusion at some point.

L157:

"... are very evident"

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

We agree that the differences in growth rates are surprisingly large. Following a recommendation from the other reviewer, we now include a table summarising the initialisation procedures of the other models and show the extended spread error equation for these models.

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.

We argue that there is a true growth rate associated with the real-world's attractor – this was why we discussed the Liouville equation. As discussed above, we agree that this might not be an attainable goal for NWP – or indeed if we can actually define such a goal. We have changed the wording from "which is closest to the truth" to "which system best maintains short-range reliability".

L167:

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

We have re-worded and re-arranged the text to say "when averaged over a sufficient number of ensemble forecasts, the average difference between the truth and the ensemble mean should match the average difference between an ensemble member and the ensemble mean (so, for example, the mean–squared–error of the ensemble–mean should match the mean ensemble variance). Adding an equation here would introduce extra notation that we do not consider helpful in the derivation of the extended equation.

L174 (also L168):

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

The key issue here seems to be the definition of ensemble members and ensemble forecasts. We have now made this much clearer, and now simply replace with “ensemble forecasts”. By bringing the derivation from the Appendix into the main text, the symbols are explained beforehand. If one looks through the literature, an overbar is generally used to denote a mean, regardless of what it is a mean over. Conversely, “<..>” is often used to denote an inner product. Hence, we would prefer to use an overbar in both (1) and (2), but we now make it clearer that the overbar in (2) is an average over ensemble forecasts.

L188:

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

These parameters are displayed in the figure and are different from the truth. We hope that by bringing the derivation together with the figure, this will become a lot clearer – since (C1) is the equation which relates directly to the figure.

L196:

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

Reliability is about bias as well as spread (and all other moments too of course). Hence having bias in its correct units is valuable. We have made this point now. We have also dropped the squared figures from the supplementary material.

Sec. 7:

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

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

L233:

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

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

L238:

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

We will include this explanation up-front. It was alluded to on L248, but not sufficiently.

L245:

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

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

L264:

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

Yes, we are combining cluster 1 cases from both clustering areas to produce a 'cyclogenesis composite'. We have made this clearer now. We used the clustering approach (and the smaller regions) to give some coherence in flow structures, and to be consistent with the fields displayed in Figs. 2 and 3 (and the animations). To a large extent, the results justify the approach – we see the structures in the clusters and obtain statistically significant differences between the two composites. Other approaches, such as the one suggested by the reviewer, could also be successful and might be amenable to the use of a single combined region (indeed the whole storm-track). We suggest that this could be tested in future studies.

L293:

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

We have moderated and extended the text to say "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 grey) in Fig. 9j and statistically significant residuals (largely dark blue and 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."

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

L296:

I did not understand this paragraph.

This paragraph was a note about conditional sampling. In the interests of brevity, we have removed the paragraph. The key text remains at L258: “(Note that the growth rate is not used within the clustering algorithm since this is what will be evaluated)”.

L313:

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

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

L324:

I don't understand this sentence.

We have changed the sentence ‘The Z250 spread represents the temporal integral of the local tendency, and hence the effects of “material” generation and advection.’ to “As one might expect, the maximum spread is not co-located with the maximum ‘Lagrangian’ growth rates (cf. Fig. 4a).”

L333:

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

There is an understandable misunderstanding in our use of “resolve”. We have changed this to “that would otherwise be created when the model is forced to represent this convection on its 18 km grid”. Thank you.

L338:

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

We agree with the reviewer here, and have modified our tentative explanation on L339-340 from “The smaller scales associated with PV315, and its sensitivity to vertical gradients in diabatic heating in the upper troposphere might help explain this sensitivity to the increase in resolution” to “At 4 km, the model attempts to resolve more of the convection. The resolved convection can be associated with stronger updrafts, which might perturb the tropopause more vigorously, where PV gradients are particularly strong”. There is likely to be enhanced spread in precipitation within the frontal region, although we have not looked at this. It is difficult to say whether the precipitation averaged over the front will be more or less uncertain.

L342:

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

This comment is very thought provoking. We have changed the text on L342-343 from “Forcing the model to explicitly resolve this convection (even if at the wrong ~18 km scale) appears to better locate the uncertainty with that generated by the ~4 km model” to “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 parametrisation (minus Fig. 13d), when the model is forced to represent the convection on the 18km grid".

L379:

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

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

L381:

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

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

Minor comments:

Fig. 1:

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

We only show one case now. We will experiment with the colouring suggestions.

Fig. 2 and others:

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

We will try to eliminate duplication – this might be easier now because some of the explanations are brought from the appendices into the main text.

L132:

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

This has been changed, thanks

Fig. 3 caption:

I think you meant case 2.

Yes thank you.

L175:

Better "sampling from a population"?

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

L178:

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

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

L180:

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

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

L182:

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

These superscripts are used to distinguish with the rooted terms.

L191:

Remove (.

Done, thank you

Fig. 7 (and others):

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

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

Fig. 8:

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

We will thicken these.

L273:

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

We will make this clearer.

Fig. 11:

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

We will change to DCP! We have experimented and feel that the arrows help.

Fig. 12:

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

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

L328:

) missing.

We will add this, thanks.

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

References:

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

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

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

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

Citation: https://doi.org/10.5194/wcd-2022-6-RC2