wcd-2022-6

Uncertainty growth and forecast reliability during extratropical cyclogenesis

by Mark J. Rodwell and Heini Wernli

Replies to the reviewers' comments

The authors would like to thank again both reviewers for the time and care that they put into reviewing the 2nd revised version of the manuscript, and for their comments, which helped to further improve the clarity of the text. Below we provide point-by-point responses to the individual comments; reviewers' comments are in black, and our replies in blue.

Reviewer 1 (Ron McTaggart-Cowan)

The authors have made significant changes to the submission that have improved the presentation of the work. I particularly appreciate the fact references to the "cyclogenesis butterfly" have been removed and that the number of sections has been reduced to six. The manuscript is certainly converging towards a publishable form.

Many thanks for this positive assessment of our efforts to improve the paper.

Recommendation: Minor Revisions

General Comments

1. Description of the plotting strategy within the main body of the text disrupts the flow and will distract future readers from the main points that the figures are meant to support. Leaving details of the plot description within the caption instead of the text is a stylistic decision; however, I believe that the paper would be more effective if this convention were followed throughout.

The caption is used to provide a complete and rigorous description of the figure. The main body text is useful for the reader to readily access the figure – for example "The black border indicates the union of the two clustering regions". Hence, there can sometimes be a small amount of duplication, but this seems necessary to us.

2. Starting on L204, the growth rate of spread is used as being analogous to forecast uncertainty. But one of the key findings of this paper is that the ensemble is conditionally over-dispersed. So then the events being studied are precisely those where the ensemble spread does not represent uncertainty well. It seems like a more careful use of the words "spread" and "uncertainty" is warranted throughout the text. This is particularly true in relation to the LGR, which I think is formulated to represent spread growth directly, and uncertainty growth only under the assumption that the two are interchangeable (which is shown not to be entirely true during cyclogenesis events).

This is an interesting point. We do view spread and uncertainty somewhat interchangeably (spread being one facet of uncertainty). This is because we believe that uncertainty can only be quantified in the context of the given observations and model (including model uncertainty representation); both of

which are likely to improve in the future. The text does make this clear. For example, in the previous version of the manuscript we state at L65 "The possibility that there is synoptic-scale coordination of uncertainty growth is interesting. Whether this hints at an intrinsic property of the atmosphere, or is dependent on the formulation of the forecast system, is explored by comparing models…" and at L70 "The variance of the ensemble, which generally grows with lead time, is then a measure of forecast uncertainty".

3. I still find the methods used in section 4 to be too complicated for the relatively simple outcome. If the goal is to demonstrate the conditional over-dispersion of the ECMWF ensemble, then please reconsider dramatically simplifying the analysis. If the goal is a demonstration that a complicated spread-error decomposition can be used, then please be sure to highlight the value added by the technique over a simpler analysis.

Although a simple comparison of spread and error also suggests over-spread, an analysis of the extended equation is useful. Partly this is to confirm that there are no strong (negative) covariance terms which invalidate the conclusion about over-spread. More importantly, the extended equation provides the better target for future system development. We mentioned the need to reduce day-2 ensemble standard deviation from the current 18 m to 13 m, rather than to the 15 m indicated by the simple spread-error relationship. We now increase the discussion in Section 6 by stating "We would argue that balance in the extended error-spread equation provides a superior reliability target for system development. For example, in the ECMWF ensemble, the standard deviation in day–2 Z250 over the east coast of North America during DJF 2020/21 was 18 m. The standard error-spread equation suggests a target of 15 m, while the extended error-spread equation suggests a target of 13.4 m. Including an estimate of the variance of forecast bias reduces this target further to 13.1 m. Moreover, at short forecast ranges, the extended equation can provide a better target for improvements in flow-dependent reliability".

4. The use of different shading ranges for different terms of the same equation (top two rows of Figs. 5 and 7) and equivalent plots for different sensitivity tests (Figs. 9 and 10) make these comparable panels very difficult to compare. I understand that not much will show up on the panels with small ranges, but that's useful information that the reader should be able to determine at a glance, not by looking simultaneously at the plotted structures and the colour bars simultaneously for two panels.

There is often a dilemma whether shading ranges should be such that they optimize the information content within a particular figure or across figures. We decided for the first option because we think that each figure per se also contains valuable structures that can be best seen with different shading ranges between figures. For example, the information in Fig. 9f shows the extent and magnitude of the influence of the local observations – this could be useful for future reference if (for example) SV perturbations were turned off, and model uncertainty scaled down. We have now added text in the caption to Fig. 5 stating "Note that the Bias and Residual, which can take positive and negative values, are shaded with a different interval to the other terms". The caption to Fig. 9 already stated "Note that the shading interval varies across the panels".

5. It appears that only the final paragraph of the conclusions (four lines) could really be classified as discussions or conclusions (the title of section 6). This means that ~90% of the section is actually dedicated to a thorough summary, including numerous figure and table references. As someone who read through the full manuscript, I find this redundancy a missed opportunity for opening up a broader

discussion of the implications of the work. Please consider either renaming section 5 to "summary and conclusion" or (better) redrafting section 5 to present a very short summary before taking a larger perspective on discussing the work.

As noted above in response to general comment 3, we now also discuss in this section the use of the extended error-spread equation as a target for system development. Nevertheless, we have re-named this section "Summary and Conclusions".

Specific Comments

1. [L40-58] I suggest redrafting this paragraph (and removing Fig. 1, with any needed panel combined with Fig. 2) for two reasons. One is that it is hard to imagine any reader of this work who wouldn't already been very familiar with cyclogenesis in the storm track. The other is that this study isn't really about cyclogenesis itself, but rather about the growth of uncertainties related to cyclogenesis. I understand that the reader needs to know what cyclogenesis is to appreciate how errors may grow, but I think that a well-crafted literature review would be more effective at relaying this than the current case study. Describing baroclinic instability and diabatic contributions to cyclone deepening could be done briefly with relevant citations. This would be followed by an introduction to error growth on the waveguide, for example citing the recent work of Baumgart et al. (2019).

This paper addresses at least two communities: scientists interested in ensemble forecasting in general (and we claim that not all of them are fully into the synoptic-scale dynamics of cyclogenesis) and colleagues familiar with cyclones and storm tracks (who need more background, e.g., about forecast reliability). We therefore think that it would be a pity to remove Fig. 1. Also, this figure is setting the scene for the later discussion of the same cyclogenesis event.

2. [L61] I think that the word "coordinate" implies too much intention here, and feels like an anthropomorphization as a result. Or maybe it's the word "act to": cyclogenesis doesn't really "act", it just happens. A similar construction appears in the subsequent sentence.

We like the word "coordinate", which links to the cited hypothesis of Palmer et al (2014). It is quite central for our results that there is a natural theoretically-based scale for baroclinic growth. However, we agree that "act" might be not the ideal word, we changed the sentence to "… to investigate whether cyclogenesis events can coordinate strong growth of forecast uncertainty …".

3. [L126-127] Many readers will probably know what it means to "warm start" VarBC and SPPT: please provide a brief explanation because this presumably impacts early spread growth in the forecast.

The sentence was providing technical information in an overly-complex way. The details actually ensure that the experiments are started as cleanly and smoothly as possible. We have removed this over-complexity by stating simply "Prior information for the experiments comes from the operational EDA".

4. [L137] Does the WCD style guide cover web references? If so, it will hopefully cover citation format and include information about access date.

This will be sorted out during the typesetting.

5. [L177] This introductory sentence is written as if clustering is the only way (or even the most obvious way) to accomplish the objective of identifying cyclones. Given that other methods for cyclone

identification have been used in the literature, please consider rewording this sentence to provide a stronger introduction to the need for a cluster analysis in this case.

We changed “clustering” to “focusing”. Further justification for the chosen method is given on lines 182-184.

6. [L210] The use of the thin overline here (ensemble mean) is distinct from the thick overline in Eq. 1 (time mean). I’m not sure that the typesetting is going to be clear enough to allow readers to distinguish between these two. Please consider using a different operator (for example <>) for the ensemble mean.

We have thought a lot about this, and experimented with different approaches. 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 overbars throughout. To help differentiate a little, we now used thick and thin overbars. However, thick overbars are defined in (and relate to) Section 2.3 while thin overbars are defined in (and relate to) Section 2.6, so we believe there is little risk of confusion.

7. [L248-252] This discussion makes it sound like Z250 is used throughout the remainder of the text, but the subsequent section moves back to P315. Please clarify here which sections are forced to use Z250 because of TIGGE database limitations.

Thanks, we have clarified this. We now state “Since the required fields are not available in TIGGE, when comparing with other models in Section 3.2, the pragmatic decision is made…”

8. [L259] “Cyclogenesis deepening” seems redundant given the adopted definition of cyclogenesis (i.e. cyclone deepening).

Agreed. We have changed “cyclogenesis deepening period” to simply “cyclogenesis”.

9. [L265-284] I don’t understand the “it is tempting to speculate” concept here, especially when the paragraph goes on to say that these hypotheses could be confirmed (or refuted) by investigating the terms on the r.h.s. of the LGR equation. If it is tempting and confirmable, then why isn’t it done? Then all of this theorizing could be replaced by a simple plot that shows what process is occurring. Between that simplification and the reduction in plot strategy description (General Comment #1), a solid analysis could be included without increase in manuscript length. This would also provide justification for the existence of the r.h.s. of Eq. 3, which is not otherwise used in the manuscript as noted by both reviewers in previous rounds of review.

It is not easy to verify these statements. They require work beyond quantifying the terms on the r.h.s. of the LGR equation. If we could have done it easily, we would have. The reviewer has already mentioned that this manuscript contains a lot of research, and we therefore prefer to keep our cautious formulation.

10. [L292-294] There is an odd asymmetry in this discussion. The “forecast bust” reference (behaviour of a forecasting system that is entirely a property of model space) seems out of place with this discussion of physical features and phenomena. Maybe pulling out the “forecast bust” phrase and making a separate statement would help, because then it is clear that the bust can have its origins in any of the listed features (or others).

We do not claim that the busts are entirely (or even partially) a property of the model space. Rather, we are relating the busts to the single, deterministic, sampling of the forecast. To make it clearer what is meant, we have reworded the sentence to: "All these situations can lead over Europe to extreme precipitation (Grams and Blumer, 2015), blocking events (Rodwell et al., 2013) and, in deterministic forecasts (which can be thought of a single sampling of an ensemble distribution), "busts" or "dropouts" (Lillo and Parsons, 2017)".

11. [L310-315] It is unfortunate that there is no quantification of this difference. It seems like there is enough information contained in the TIGGE database for a systematic assessment of LGR for Z250 during cyclogenesis cases from different models.

In the meantime, the TIGGE animations for the entire winter season have been uploaded to an open access repository: https://www.research-collection.ethz.ch/handle/20.500.11850/605102. We trust that the reviewer and future readers of the paper will appreciate the high variability of growth rates shown for the different events and between models.

12. [Sections 4.3 and 5] Does "day—2" refer to "day minus 2"? If so, please replace the em-dash with a space and a minus sign, to become "day -2" for readability and to distinguish it from a compound adjective (e.g. "the day-2 spread").

Apologies for this incorrect use of the Latex en-dashes. In the example, it refers to "day 2" rather than "day minus 2" (indeed it never refers to "day minus 2"). Throughout, we now retain "day—2" only when it is a compound adjective; otherwise we write, for example, "at day 2".

13. [L510-511] Is there really a need for further investigation to figure out how to reduce spread generated by SVs or SPPT? The options seem pretty obvious. Perhaps this would be better worded as how to reduce the associated spread growth during cyclogenesis without negatively affecting the overall well-balanced spread-error relationship.

The operational implementation of developments requires, in general, that they improve (or at least don't degrade) proper scores of the ensemble forecast. A well-balanced spread-error (or extended spread-error) relationship indicates reliability, which is one aspect assessed by a proper score, but refinement (or resolution) is another aspect. Hence, we rephrase this now slightly differently as "…it makes sense to investigate how these techniques might be modified to reduce the growth of spread during cyclogenesis without negatively impacting the overall performance of the ensemble."

14. [L564] I'm not really sure what this sentence means (even what "other" refers to or is distinct from) and how it follows logically from the current work.

The "other" flow types are distinct from the cyclogenesis situations discussed here. We have changed the last sentence to: "In addition to cyclogenesis situations, initial growth rates tend to support the idea that uncertainty can be concentrated in other flow situations, including those prone to mesoscale convection over North America (Palmer et al., 2014) and during the extratropical transition of tropical cyclones. Similar investigations of these initial growth rates, and how reliably the forecast predicts their evolution, could also lead to better flow-dependent reliability and improved overall forecast performance".

15. [Data availability statement] The “data and code available on request” doesn’t really live up to FAIR principles. Please consider at least uploading as much of the code used to generate the results shown here as possible to a public repository.

We also stated that the TIGGE and ERA data are freely available and have made the animations available under a doi. The details of the code are carefully laid-out in the manuscript. It is not practicable to provide the code without providing help and guidance, since there are numerous settings to configure and utilities (such as Metview and spectral transforms) required.

Reviewer 2

Thank you for again considering my previous comments and carefully rewriting the paper. I very much like how you restructured the manuscript. The objectives of the study come across much more clearly and it is much easier and less confusing to follow the paper and also more fun to read. In my opinion the paper can be published as it is, I see however still room for some improvements. I leave it to you and the editor to decide whether or not you want to incorporate these points.

Many thanks for this positive assessment of our efforts to improve the paper.

I give my main concerns below, starting with the most important one and then a few minor points in ascending order.

L427ff: Here I am coming back to a concern I already raised in the previous round which I think is still not really addressed convincingly: If I by eye average the residual over the plotted area in Fig. 7 of the cyclogenesis cluster and the counterpart, it seems that the counterpart is equal to even more over-spread, only the location of the over-spread is shifted downstream. It is fair to assume that, since cyclogenesis in the box is excluded in the counterpart cluster, it is more frequent in the downstream region. But since it is the main point of the paper to link cyclogenesis and over-spread I am still missing a quantitative analysis of this feature, e.g. by investigating surface pressure tendencies or even another cluster analysis based on the downstream region where the counterpart residual blob occurs.

We agree that we don’t discuss the structure of the residual outside of the clustering regions in detail. One reason is that, as also mentioned in the earlier reviews of the paper, the study is rather complex by combining different methods and sets of simulations. We therefore focus on the western North Atlantic, which is known from climatologies to be the hotspot of cyclone intensification and warm conveyor belt activity in the North Atlantic-European sector. A second reason is that towards the end of the storm track, processes get even more variable than in the entrance region. In the entrance region (which we focus on in our paper), days can meaningfully clustered into “cyclogenesis” and “non-cyclogenesis”, whereas in the eastern North Atlantic, there can be propagation of cyclones from upstream at different latitudes, or the formation of secondary cyclones along fronts, or the formation of cutoff lows, etc. This makes it more difficult to explain the main reasons leading to the residuum field in a clear and concise way. However, we manually analysed maps of the synoptic flow in this region downstream of the clustering regions, which revealed that indeed, at many time steps that belong to the non-cyclogenesis counterpart cluster, there is cyclogenesis in the region downstream of the clustering regions, but as mentioned above with a lot of variability between cases (with cyclones either close to Iceland, over the

UK or associated with upper-level cutoffs, further south near the Azores. Together, they most likely explain the over-spread mentioned by the reviewer.

We have added the text towards the end of Section 4.2 L392-394:

“Further east, towards the end of the storm track, processes get more variable. There can be propagation of cyclones from upstream at different latitudes, the formation of secondary cyclones along fronts, and the formation of cutoff lows, for example. This makes it more difficult to meaningfully cluster this region into cyclogenesis and non-cyclogenesis cases.”

To be more clear about the comparisons in Fig. 7, we also change, in Section 4.3, the text:

“Downstream, differences have the opposite sign — possibly because the occurrences of cyclogenesis events over the western North Atlantic are likely to be anticorrelated with the occurrences immediately downstream (as seen in cluster patterns Fig. 6c,d)”

to:

"Downstream, both composites individually display negative residuals. This is consistent with the above discussion since there has been no direct control for downstream cyclogenesis. Indirectly, it is likely that occurrences of cyclogenesis events over the western North Atlantic are anticorrelated with the occurrences immediately downstream (as seen in cluster patterns Fig. 6c,e) and this might explain the negative cluster differences in absolute residual for the downstream region (Fig. 7o)”.

L230ff: You did not state the motivation for using the EDA system instead of the forecast system to derive the Lagrangian growth rates at this point. I infer from later discussions that you wanted to exclude the singular vector perturbations? On the other hand, you have a forecast without singular vectors of this case in your sensitivity dataset. Also, for the TIGGE-analysis you (have to) include the singular vectors. I am still a bit confused here.

It is true that the case shown in the sensitivity experiments included the experiment without singular vectors. However, the growth-rates from the EDA are provided in the animation for the entire DJF 2020/21 season. It is important to show that the concentration of initial growth rates into particular flow situations is not purely due to the SVs (which we are suggesting could be reduced in amplitude). The most appropriate place to explain this seems to be in Sect. 6 where the effects of SVs are discussed. Following the text:

“Results highlight a few flow situations over the North American / North Atlantic / European region where ensemble variance growth at synoptic scales is particularly strong and concentrated”

to:

“The supplementary animation of LGRp for the DJF 2020/21 season highlights a few flow situations over the North American / North Atlantic / European region where ensemble variance growth at synoptic scales is particularly strong and concentrated. Note that this concentration of growth-rates is not dependent on singular vectors perturbations, since these perturbations are not included in the EDA.”

Appendix C: I could not find any reference to the variable and vertical level for which you show the spectrum. Is this for kinetic energy (i.e. variance of the wind)? Also, I don’t understand the meaning of the dotted lines.

Apologies, we have now added T500 in the main text and figure caption.

L269/L283, second vs. first source term: The text reads as if you are speculating about this. But wouldn't it be very easy to check if this is true by comparing the contributions from those terms?

It is not so easy to verify these statements. They require quite a lot of further work, and inevitably raise further questions. The paper is dense already, and we hope the reviewer will agree to this being left for a subsequent study.

L272ff: I find these references to intrinsic predictability misplaced here (maybe that's partly my fault) and they could maybe also removed. However, I would like to point out that first the term you are referring to can reflect scale-interactions (because it is non-linear), but it could also be dominated by interactions within similar (synoptic) scales. Which one it is is speculation at this point. Second, please note that in Selz et al 2022 also large total variances (as produced by the EDA) were considered. With similar methods also Baumgart et al. 2018 ("Potential Vorticity Dynamics of Forecast Errors: A Quantitative Case Study") considered (operational) large forecast uncertainties. Both studies showed that variance ("error") growth happens predominantly in the rotational component of the flow (e.g. figure 6 in Baumgart et at.), which in my opinion disfavours large contributions from scale interactions, since from the mesoscale downward the kinetic energy is roughly equally partitioned between the rotational and divergent component.

We are largely in agreement with the reviewer here; the only difference might be in terminology. As the reviewer says, it is up for speculation at present, but much of the growth could be "dominated by interactions within similar (synoptic) scales". We were very much including this within our reference to "scale interactions" within the continuum of scales between 2000 and 100 km. We agree that interactions can also be at the same scale. Maybe it is also possible that synoptic rotational anomalies can be driven by mesoscale interactions with divergent winds (as illustrated by the "Rossby Wave Source")? We have changed the text:

"Scale interactions embodied within this source term can produce initial uncertainty growth at synoptic scales because the EDA contains considerable variance power at spatial scales between about 100 and 2000 km (illustrative power spectra are presented in Fig. C1 in Appendix C). This may not be the case in predictability studies, where initial uncertainty is restricted to grid-point noise (Judt, 2018) or has small total variance (Selz et al., 2022). See also Durran and Gingrich (2014)."

to:

"Interactions within and between scales, represented in this non-linear source term, can produce initial uncertainty growth at synoptic scales because the EDA contains considerable variance power at spatial scales between about 100 and 2000 km (illustrative power spectra of T 500 are presented in Fig. C1 in Appendix C). This may not be the case in predictability studies, where initial uncertainty is restricted to grid-point noise (Judt, 2018). See also Durran and Gingrich (2014); Selz et al. (2022) for relevant discussions.

Minor points:

L146: "forecast time". I was confused at first. Maybe replace with "forecast init time" or "case" to distinguish it from "forecast lead time".

We have changed to “forecast initial time”.

L165: Should that be “...its correct form. For example, since R can be...”?

This was poorly worded. We have changed “This approach leaves the bias in its correct form, for example” to “Note that the square-root of Bias^2 is Bias (with its correct sign)”.

L250-251: I am confused. Are you using centred differences in time to calculate d/dt? So like f(12h)-f(0)/12h? But this then would be centred at 06 or 18 valid time (figure 3 however says 12Z)? Could you clarify?

For the TIGGE data, centred differences are valid at times 03, 09, 15, 21 UTC. The 24h running-mean then has the effect of putting the fields at 00, 06, 12, 18 UTC. For example, the 12 UTC field shown in Fig. 3 is the mean of centred differences at times 03, 09, 15, 21 UTC. Note that the animations of the TIGGE data include a frame at each hour - these being produced using linear interpolation between the 00, 06, 12, 18 UTC fields. We had omitted this detail. We now state:

Earlier, after “The filter also includes a 24 h running-mean“, “(the nominal validity time is at the centre of the running-mean window - placing the final fields back on the full hours)”

and, at the end of Sect. 2.6, we add “The resulting timeseries of fields can again be used to produce animations (hourly frames being derived using linear interpolation between the 6 hourly fields)”.

L264: Should that be 0.2 h^-1 (like in the colour bar of the figure)?

As per the caption for Fig. 2, the orange contours “extend the shading scheme, with the same interval. In the largest red blob, there are two orange contours which have values of 0.14 and 0.18 h^{-1}. This is why we state “in excess of 0.18 h^{-1}”. Note that the value at the end of the colour bar (0.26 h^{-1}) indicates that there is just a third contour at 0.22 h^{-1} but this must be so small that it is not plotted/visible.

L337: I suggest to first emphasise the discrepancy between the 18^2 and the 13.4^2 (the overspread of the ECMWF system, the main point of the paper) before going into the need to account for AnUnc and Bias.

The 13.4 m is the value which does take into account the analysis uncertainty and bias; it is 15 m when they are not accounted for. Nevertheless, we appreciate the spirit of the comment. We take this discussion to Sect. 6 because this is the natural place to put it and also because, by then, an estimate of the variance in forecast bias has been obtained (in Sect. 4.3), which lowers the ensemble variance target to 13.1 m. We now state in Sect. 6 “We would argue that balance in the extended error-spread equation provides a superior reliability target for system development. For example, in the ECMWF ensemble, the standard deviation in day–2 Z250 over the east coast of North America during DJF 2020/21 was 18 m. The standard error-spread equation suggests a target of 15 m, while the extended error-spread equation suggests a target of 13.4 m. Including an estimate of the variance of forecast bias reduces this target further to 13.1 m. Moreover, at short forecast ranges, the extended equation can provide a better target for improvements in flow-dependent reliability”.

L484: In my opinion this “lack of agreement” confirms the point I made in the last review that the model uncertainty representation (SPPT) is not accounting for the impact of missing sub-gridscale variability,

which is (better) resolved at 4 km, but mainly for “flow-dependent biases” in the parameterisation schemes on larger scales. Would you agree?

We say, “model uncertainty representation is thought to partly account for the impact of sub-grid-scale uncertainty”. We then highlight the discrepancy with the 4 km results. Yes, the reviewer’s hypothesis may well be correct. Probably few people think that the model uncertainty representation only accounts for missing sub-grid-scale uncertainty. In the end it is quite pragmatic - exemplified by the need at present at ECMWF to apply it at scales much larger than would be justified by the “missing sub-grid-scale uncertainty” argument. The modified text at the end of Sect. 5 now states “Since the motivation for the use of singular vector perturbations (Magnusson et al., 2009) and the initial reason for the development of model uncertainty representations (Buizza et al., 1999) was to increase ensemble spread, it makes sense to investigate how these techniques might be modified to reduce the growth of spread during cyclogenesis without negatively impacting the overall performance of the ensemble”.