Review of gmd-2017-260

Title: Methods of investigating forecast error sensitivity to ensemble size in a limited-area convection-permitting ensemble

Authors: Bannister et al.

Summary and general comments

This paper describes various ways forecast error sensitivity to ensemble size can be examined for high-resolution ensembles. These methods were applied to a single case for an ensemble over England with 1.5-km horizontal grid spacing.

In my opinion, there isn't any new science in this paper, because, as the authors note, one case is not enough to draw robust conclusions. It appears the authors believe the novelty of their work is in developing methods to analyze high-resolution ensembles with applications toward determining optimal ensemble sizes.

While I found some of these methods interesting, I think the paper lacks focus, and, in my opinion, a number of the methods do *not* represent new developments while some of the other methods seem overly complicated. It was difficult to tell whether the audience of this paper was data assimilation scientists or model evaluators, and, while the authors may have intended the paper to appeal to both, I don't think it was successful in that regard. Primarily because I believe this paper does not present enough novel material, in my opinion, I don't believe it is suitable for publication in *GMD*.

Specific comments

1. Content

This paper is not a review article, so I found quite a bit of material unnecessary. For example, section 1.1 seems irrelevant and can be omitted; the focus of this work is on high-resolution ensembles so you don't need to discuss these coarse-resolution studies (and, in fact, as you note, the coarse-resolution studies may not be relevant to higher-resolution applications). Similarly, lines 9-25 on page 7 (section 3.2) can be completely removed, as your ensemble generation method is unrelated to any of this material. It was also unclear why the completely different ETKF experiment was discussed and shown in Fig. 2; I think that just added confusion and should be removed.

Additionally, the introduction mixes data assimilation studies with work looking at forecast sensitivity to ensemble size. While there is certainly an overlap between the ensemble data assimilation and forecasting communities, the two groups do face different challenges, and I don't believe these differing challenges were laid out clearly enough. Overall, I think the paper would have been stronger had it focused on either data assimilation or ensemble forecasting; not both.

I also thought a few references were missing. Hagelin et al. (2017) and Raynaud and Bouttier (2017) should be discussed. These two very recent studies presented more

established methods of looking at forecast sensitivity to ensemble size [as do Clark et al. (2011) and Schwartz et al. (2014), which you cited]. Surcel et al. (2014) also has some interesting ways of looking at some of the topics you considered.

Finally, because you note that one case is insufficient to obtain robust results, there's no need to provide exhaustive discussion about your case-study results to place them in context of other work (because your findings cannot be generalized to other cases). For example, the text in lines 19-27 on page 18 is unneeded, and there are other areas of text that should also be omitted for similar reasons.

2. Appropriateness and novelty of evaluation methods

a. I didn't particularly find anything novel about your ensemble generation method, as downscaling from global to convection-allowing resolution has been used to generate ensembles for quite some time. Perhaps "negating existing perturbations" is somewhat novel, but, nonetheless, on its own, I don't think this part of the paper represents a new development. However, I think if you had more available resources, an interesting paper could focus on comparing your perturbation method to other cutting-edge perturbation methods (including those using ensemble data assimilation) for a large number of cases.

b. I thought section 3.3 was the most interesting and novel part of the paper. However, your method was somewhat complicated, and simplicity rules when proposing methods for model evaluators. Frankly, I don't see model evaluators adopting this technique (though maybe they should), and they will probably continue to use simpler, standard verification metrics to determine whether adding more members is beneficial.

c. In my opinion, the methods and analyses presented in sections 4.1, 4.2, 4.3, and 5 are not new, and we can't gain anything from the results because it's only a single case. Ensemble means, standard deviations, and probabilities have been computed for quite some time. Rank histograms are also standard (I do appreciate that you sampled from the observation error to construct your rank histograms).

d. Examining the kinetic energy (KE) spectrum (section 6) is also not new. Moreover, I don't understand why it's necessary to average KE spectra over all ensemble members. What information are you trying to obtain that isn't available from the single members? I believe looking at individual members should be sufficient for determining effective resolution as well as characterizing spectra.

e. I thought section 7 was somewhat interesting but only applicable to data assimilation scientists. Model evaluators would almost certainly never use these methods, as they are both complex and of questionable relevancy to those interested in determining how many ensemble members are needed for, say, 36-h forecasts. The idea of sampling error just isn't as important to folks using the ensemble forecasts compared to data assimilation researchers. Also, I believe some of this material was covered by Menetrier et al. (2014), which you cited, so I'm not completely sure of the novelty of at least portions of this section.

3. About your ensemble construction…

Fig. 8 points to problems with your ensemble construction. Namely, mismatches at the boundaries leads to spurious precipitation around the southern boundary at very light rainfall rates. There's probably not much you can do about this except to make your high-resolution domain bigger to move the area of interest further away from this noise.

4. References

Hagelin, S., J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant, 2017: The met office convective-scale ensemble MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.,* **143**, 2846–2861, doi: 10.1002/qj.3135.

Raynaud, L., and F. Bouttier, 2017: The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Quart. J. Roy. Meteor. Soc.*, doi: 10.1002/qj.3159.

Surcel, M., I. Zawadzki, and M. K. Yau, 2014: On the filtering properties of ensemble averaging for storm-scale precipitation forecasts. *Mon. Wea. Rev.*, **142**, 1093–1105, doi: 10.1175/MWR-D-13-00134.1.