Review on "Methods of investigating forecast error sensitivity to ensemble size in a limited-area convection-permitting ensemble" by R. Bannister, S. Migliorini, A. C. Rudd and L. H. Baker.

This paper examines the impact of ensemble size for the prediction of a particular rainy event over UK. In a first part some diagnostics on short-range forecast are examined for a small, an intermediate and a large ensemble. In a second part, the ability of these three ensembles to accurately estimate short-range forecast error variances and correlations is discussed.

Regarding the ensemble prediction part, this study, as mentioned by the authors, suffers from being based on a single case. Hence, neither objective scores nor robust conclusions can be derived, while we recall that a number of studies have already provided more comprehensive analyses of the ensemble size effect in convective-scale EPSs (Clark et al., 2011; Schwartz et al., 2017; Hagelin et al., 2017; Raynaud and Bouttier, 2017).

Regarding the estimation of error covariances, the authors mainly base their analysis on the sampling noise theory published in previous papers. Their additional developments are not clear to me and I suspect the methodology may be flawed to some extent. Overall, their results only confirm some well-known ideas. In particular, a similar comparison with a large 90-member ensemble is provided in Ménétrier et al. (2014).

In its present form, I thus consider the paper does not provide enough original and new results to warrant publication in GMD.

I provide hereafter a list of my major concerns.

Main comments

• 1. The generation of the large ensemble is not completely clear to me. Here are some points that require further details :

— Figure 2 suggests that you negate analysis (instead of forecast) perturbations from the 23-member MOGREPS-G, could you confirm ? Are these (positive + negative) perturbations added to the 4D-Var control to form the 46-member MOGREPS-G ? Same question for the passage from 46 to 92 perturbed members.

— The data assimilation step in MOGREPS-G adds ETKF perturbations around the control 4D-Var analysis. Is the ETKF size also increased for the DA cycling of the 46 and 92-member ensembles ?

• 2. Sections 7.2 and 7.3 : I don't fully agree with the application of the sampling noise theory to $\mathbf{g}^{(N)'}$ when $\mathbf{v}^{(\infty)} \approx \mathbf{v}^{(93)}$, and with the subsequent diagnostics.
In my opinion, one proper way to evaluate how suboptimal is a 93-member ensemble may be to derive an analytical expression for the covariance of the difference $\mathbf{v}^{(N)} - \mathbf{v}^{(93)}$ and to compare it to the 'true' sampling noise covariance. Then by varying the size of the large ensemble you could get an estimate of how large enough is the reference ensemble regarding different aspects of the covariance estimation. For instance, it is expected that an ensemble could be considered large enough for the variance estimate but not for the correlation.
Alternatively, note that the sampling noise theory indicates that the noise standard deviation of the 93-member ensemble is approximately $\sqrt{92/23} = 2$ times smaller than the noise standard deviation of the small 24-member ensemble.

• 3. Section 7 : Figures 11, 12, 15 and 16 do not provide any new results compared to previously published studies.

Minor comments

• Part 1.2 : in order to improve the understanding, please try to separate studies that concern ensemble forecasting from those that concern ensemble data assimilation since these are two different problems.

• How the four members of Figure 4 have been chosen ? Randomly ? A more relevant choice may be to show clusters from the large ensemble, since they may provide a better representation of the ensemble distribution.

• P11 L12 "This is consistent with the lack of sensitivity of ensemble mean forecast skill to ensemble size" : I don't agree with this sentence. Indeed, objective verification scores with convective-scale EPS clearly show an improvement of the ensemble mean skill when increasing the sample size, especially when starting from very small ensembles (e.g., 6 or 12 members). As you start from a larger ensemble this improvement is likely to be already close to saturation, which gives the feeling the ensemble mean skill is not sensitive to the sample size.

• Figures 4 and 5 : what is the rationale for showing the ETKF ensemble here ? It does not show the impact of ensemble size but rather the impact of data assimilation, which is not the subject of the paper.

• Figure 7 : could you explain how the frequency of each rank has been computed ? I don't understand why some ranks have a frequency equal to 1 while one would expect that all ranks frequencies sum to 1. In addition, it's quite unusual to overlay rank histograms of ensembles with different sizes (the x-axis should be different) and I find it very difficult to objectively compare the different ensembles on these figures. For instance, delta scores (distance to flatness) and number of outliers may provide a more understandable information.

• Section 7.5 : I don't understand why the correlation is estimated by dividing by $v_x^{(\infty)}$ and not $v_x^{(N)}$?

• P29 L30 "Not only does this show ... it is an interesting result in itself" : sampling properties of length-scale estimates have already been deeply documented in Pannekoucke et al. (2008); Raynaud and Pannekoucke (2012) for instance.

• At several places the authors say that it's difficult to judge if an ensemble is large enough to neglect sampling error. A possible useful tool is the signal-to-noise ratio (SNR), whose theoretical definition (e.g. Equation (17) of Ménétrier et al. (2015)) could be used to anticipate the SNR of the different components of the covariance matrix. In addition, estimating how large enough is the ensemble also depends on the subsequent analysis and forecasts scores.

• Final comment, you say that "the lack of sensitivity to other aspects, like rainfall properties and biases suggests that the quality of probabilistic forecasts would not be improved". This statement is in contradiction with objective scores from Hagelin et al. (2017), that indicate a clear improvement of precipitation forecasts when going from 12 to 24 members. In order to answer the question of the impact of ensemble size you need to look at relevant scores : ensemble mean, probabilities or rank histograms shown in this paper are clearly not sufficient and are known to saturate with 30-40 members. In order to better highlight the benefit of more members, it is recommended to look at scores that focus on the tails of the distribution, such as the weighted CRPS or quantile score for instance. In addition, the way the large ensemble is generated may also contribute to this lack of positive impact.

Specific comments

• P10 legend fig. 3 : to which pressure level does model levels 36 and 11 correspond ?

• P13 “The ensemble of rain rate .... a specified threshold rain rate" : please reformulate as this is not clear.

• P23 “The spectral weight of the variance sampling errors" : rather say this is the noise power spectrum.

References

Clark, A. J., J. S. Kain, D. J. Stensrud, M. Xue, F. Kong, M. C. Coniglio, K. W. Thomas, Y. Wang, K. Brewster, J. Gao, X. Wang, S. J. Weiss, and J. Du, 2011 : Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Monthly Weather Review*, **139**, 1410–1418.

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.

Ménétrier, B., T. Montmerle, L. Berre, and Y. Michel, 2014 : Estimation and diagnosis of heterogeneous flow-dependent background-error covariances at the convective scale using either large or small ensembles. *Quart. J. Roy. Meteor. Soc.*, **140**, 2050–2061.

Ménétrier, B., T. Montmerle, Y. Michel, and L. Berre, 2015 : Linear filtering of sample covariances for Ensemble-Based Data Assimilation. Part II : Application to a Convective-Scale NWP Model. *Monthly Weather Review*, **143**, 1644–1664.

Pannekoucke, O., L. Berre, and G. Desroziers, 2008 : Background-error correlation length-scale estimates and their sampling statistics. *Quart. J. Roy. Meteor. Soc.*, **134**, 497–508.

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

Raynaud, L., and O. Pannekoucke, 2012 : Sampling properties and spatial filtering of ensemble background-error length-scales. *Quart. J. Roy. Meteor. Soc.*

Schwartz, C. S., G. S. Romine, K. Fossell, R. Sobash, and M. Weisman, 2017 : Toward 1-km ensemble forecasts over large domains. *Monthly Weather Review*, **145**, 2943–2969.