Review of NPG-2018-21 : « Sensitivity of forecast skill to the parameterisation of moist convection in a limited-area ensemble forecast system »

General comments

The paper presents the impact of the use of two convection schemes (to represent uncertainties in the way convection is parameterized in numerical weather prediction models) in the COSMO system.
The manuscript is clear and well written but I think that some important points need to be clarified and reviewed before the paper can be accepted for publication.

Specific comments

When using a multi-physics approach to represent model error I think it is important to assess first each package of parametrization from a 'deterministic' point of view (using classical deterministic scores such as root mean square error, bias, heidke skill score, ...).
I suggest to perform two experiments (using unperturbed initial conditions), one using the Tiedke scheme, the other one using the Bechtold scheme, to compute some scores and to have a look at the distribution of precipitation. I think that the probabilistic evaluation that the authors use can not give informations about the behaviour of each of the schemes.

My main remark concerns the way 'outliers diagnostic' is computed and interpreted.
The 'outliers diagnostic' comes from the rank diagram score and represents the fraction of observations that lie outside the range of the ensemble.

My first comment concerns the way the rank diagram is constructed for precipitation forecast.
In an ensemble forecast of precipitation there are numerous cases in which one or several members forecast the same value. There also can be some cases in which all the members forecast a zero precipitation value.
How do the authors treat those cases ?
How do the authors treat the case in which the observed value and the minimum forecast value are equal ?
Do they add perturbations to the forecast values (as it is classically done in the literature) when one or several members forecast the same value ?

My second comment concerns the way the percentage of outliers is interpreted by the authors.
For a perfectly reliable ensemble of N members, the fraction of observations lying outside the range of the predicted values is 2/(N+1) (flat rank histogram).
Using a 10-member ensemble that is perfectly reliable, the fraction of outliers is 0.18 (0.095 for a 20-member ensemble).
Looking at figures 6 and 9 it can be seen that the percentage of outliers is below those theoritical values. In their comments of figures 6 and 9 the authors seem to consider that the lower the percentage of outliers, the better. This is wrong, especially if, as shown by the figures, the percentage of outliers is below the perfect theoritical value.

Looking at figures 6 and 9 we can only presume that the ensembles are over-dispersive (but this need to be confirmed by the rank histograms).
I think that using the percentage of outliers alone is not enough to properly evaluate the reliability of a forecast ensemble.
The authors should review all their comments of figures 6 and 9 and add, at least for one or two forecast ranges, rank histograms and comment their shapes (they can also use another score that measures the reliability of an ensemble).