I regret I cannot recommend acceptance of the paper. That it is because its content is too elementary for an international journal, and because there is a major misconception from the authors' part concerning the interpretation of some of their results.

The paper presents an assessment of the compared performance of ensemble predictions for precipitation performed with the mesoscale limited-area COSMO (COnsortium for Small-scaleMOdelling) model, using two different convection schemes. The first one, developed by Tiedtke, is classical and has been operational in the COSMO model for a number of years. The other one has been developed more recently by Bechtold at the European Centre for Medium-Range Weather Forecasts (ECMWF).

The performance of the two schemes is, as said by the authors (p. 6, l. 10), *roughly indistinguishable*, with perhaps a slight advantage to the Tiedtke scheme. This result is not by itself of sufficient interest for publication in *Nonlinear Processes in Geophysics*. In addition, the study presented in the paper is too superficial for an international journal. The comparison of the two schemes in ensemble prediction of precipitation should have been preceded by a description of the precipitation regime produced by those two schemes. For instance, does the distribution of precipitation in the two schemes differ from each other, and from the observed precipitation ? Simple histograms of the precipitation amount, in the two schemes and in the observations, would in that respect be very instructive, whether the histograms mutually agree of differ. I presume that type of study has been performed (I do not personally know well enough the literature on convection parameterisation), and that part of the paper could consist mostly of appropriate references, with description of the main conclusions of the referenced papers. But appropriate explanations, or at least an appropriate discussion, must be presented as to the links between the properties of the two convection schemes and the performance of the ensemble predictions.

In addition, there is a major misconception from the authors' part concerning the interpretation of their diagnostics of outliers (Figures 6 and 9 and corresponding text). The authors seem to consider that the presence of outliers is undesirable (*A better performance of Cleps-10T, which has a lower number of outliers than Cleps-10B, can be noticed*, … p. 5, ll. 26-27). Well, it would be absurd to assume that a 10- or 20-ensemble defines bounds between which the verifying observation must necessarily lies. Statistically, there must be outliers, and the question of their number has been abundantly discussed in the literature. The usual assumption made in the validation of ensemble prediction systems in that the predicted ensemble is a sample of independent realizations of a probability distribution of which the verifying observation must be an additional, independent realization. That property is called *reliability*. If it is verified, the verifying observation must be statistically indistinguishable from the N elements of the predicted ensemble. This means (among other things) that the probability for the observation to fall in any of the $N+1$ intervals defined by ranking the predicted ensemble values in increasing order must be the same for all intervals, and equal to $1/(N+1)$. The fraction of outliers must then be equal to $2/(N+1)$. That is 0.18 for $N = 10$, and 0.095 for $N = 20$. The fractions shown on the left panels of Figures 6 and 9 are smaller than that, especially at longer forecasts ranges. That means that the spread of the predicted ensembles are for both schemes significantly larger than the real uncertainty in the forecast. **That is in my mind a much more important conclusion than the minor differences observed between the two convection schemes**.

The right panels of figures 6 and 9 show the fraction of outliers lying above and below the upper and lower bounds of the ensembles. According to the same argument as above, that proportion must be equal to $1/(N+1)$, *i.e.* 0.09 and 0.048 for $N = 10$ and 20 respectively. Again, the fractions shown on the right panels of the figures are smaller than that, except for the 'max' curves of fig. 9, for which the observed fraction is correct.

The general conclusion is that both ensemble prediction schemes can mislead the user into expecting that the probability of occurrence of ‘extreme’ precipitation (either small or heavy) is larger than it really is, especially at longer forecast ranges (with as only exception for high precipitation in the right panel of fig. 9). Again, that is to me the main conclusion that must be drawn from the (limited amount of) results presented in the paper.

It may be of course that the authors do not consider that reliability of ensemble prediction systems (*i.e.*, overall statistical consistency between predicted probabilities of occurrence and observed frequencies of occurrence) is really important. But then they must explain clearly what makes in their minds that an ensemble prediction system is ‘good’, and why they consider that there must not be outliers.

I add two remarks.

- The right panels of Fig. 6 and 9 show a dissymmetry between low and high values of precipitation, without lower number of outliers below the minimum bound of the ensembles. This means that the ensembles, in addition to being overdispersed, are slightly biased towards low values of precipitation. These features may be visible (but not necessarily) on the global histograms I was mentioning above.
- The question of outliers, and more generally of the position of verifying observations with respect to predicted ensemble elements, has long been discussed, in particular through *rank histograms*. A rank histogram is, precisely, the histogram of the positions of verifying observations with respects with ensemble elements. The rank histogram of a reliable system must be flat. For a basic reference on rank histograms, see Hamill (2001).

Reference

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550-560