Weather Clim. Dynam. Discuss., referee comment RC2
https://doi.org/10.5194/wcd-2021-33-RC2, 2021

Comment on wcd-2021-33

Anonymous Referee #2

Referee comment on "Bimodality in ensemble forecasts of 2□m temperature: identification" by Cameron Bertossa et al., Weather Clim. Dynam. Discuss., https://doi.org/10.5194/wcd-2021-33-RC2, 2021

This manuscript deals with the identification of bimodality in the ensemble forecasts of 2-meter temperature produced by the ECMWF ensemble system. The ensemble forecasts are first dressed using a Kernel Dressing approach. The identification is done by means of a test developed in a simple framework of synthetic data generated by a Gaussian mixture model, optimized to a certain level of detection of false-positives (accepting bimodality while the sample is generated from a Gaussian distribution as far as I understood), fixed to 5%. The criteria used in the context of this test is (i) the presence of two modes, (ii) the mode probability ratio (ratio between the probability of one of the modes and the probability at the minimum between the two modes) must be larger than a minimum threshold and (iii) each mode must contain (through a form of clustering) at least 5 members. The results of the application of this test to the ensembles of the ECMWF showed that at several places all over the world, bimodality is detected up to 30% of the time, a level much larger than the false-positive rate. This appears mainly after the first week of the forecasts.

The question addressed in this paper is very interesting but there are several major points the authors should tackle before recommending publication. These are listed below by order of importance.

- The test is built in a very simple framework of synthetic data generated by a mixture of Gaussians. Data from the ECMWF are not purely random variables and one may wonder whether this type of approach is valid in a more realistic context. I am surprised that the test has not been validated on more sophisticated systems like the 3-variable Lorenz model (1963) displaying bimodality for certain variables, and also on spatially distributed systems like the Lorenz 95 system (1995). This would allow to evaluate the ability of the test to indeed detect bimodality, and to test the sampling issue in detecting false-positives. This should be done on these models or other similar types of models.
- Bimodality is detected in very active zones of the atmosphere such as for instance in

panel (d) of Figure 1. When I look at this figure, it seems to me that after 5 days there is no skill anymore, and after that period there is a random succession of unimodal, bimodal (and multi-modal) situations. At first sight I would say that this is a problem of sampling, and one can have on such a large range of possible values (from -15 to 5 °C) random clustering of ensemble members. In order to check whether it is not an artifact of sampling, one can pool several successive ensembles, for instance the ensembles started 12 hours before and after the nominal ensemble (or another grouping of ensembles). In that way one can get a set of 153 ensemble members and see if the bimodality detected is still present.

- The authors mention in the discussion (lines 325-330) the convergence toward the asymptotic distribution, and indicate that the bimodality is more a transient process. This should be checked and compared with the asymptotic climatological distribution. We should expect the convergence toward the climatology.
- In the discussion section (5), the authors discuss the possible origin of the bimodality, in particular related with the interaction with the ocean. Would the authors mean that the bimodality is related to a low-frequency variability which has skill at longer lead times? It would certainly be useful to investigate in greater details this in a specific region like the one presented in Figure 1d and see how is the convergence toward climatology through the use of skill scores, and see what is the difference between the events that are truly bimodal and the other ones. And also relate these to some specific processes (the presence of sea ice for instance).
- In the first part of the manuscript the authors address the problem of scoring and statistical postprocessing when the distributions are bimodal. In particular, the ability of scoring rules to detect bimodality. But at the end of section 2, the authors indicate the difficulty to detect bimodality based on the scoring rules, and they then move to the analysis of bimodality through the development of a simple statistical test. This section 2 looks unrelated with the rest of the manuscript and I think that this part can be considerable shortened and either placed in an Appendix, or as a few paragraphs in the introduction.

Minor points:

- In Figure 1, panel a, the authors show the observation in yellow. On what is this observation based on? And there is of course an observational error which is not taken into account in these plots and in the comments. The authors indicate under-dispersion of the ensemble, but is it really under-dispersed? This cannot be said from a single ensemble forecast.
- Lines 150-155. The authors mention the use of statistical postprocessing. It is not clear at all how it is done. This should be explained.
- Line 220. Please add that material in a supplementary. Also the material discussed at lines 170-175 if section 2.3 is kept somewhere (either in an Appendix or in a supplement).

References:

Lorenz, E N (1963). "Deterministic nonperiodic flow". *Journal of the Atmospheric Sciences*. **20** (2): 130–141.

Lorenz, E.N. (1995). "Predictability, a problem partly solved", ECMWF seminar on predictability.