Report on the manuscript:
Common metrics of calibration for continuous Gaussian data and exceedance probabilities
by Rita Glowienka-Hense, Andreas Hense, Thomas Spangehl, and Marc Schröder

The paper is concerned with the evaluation of ensemble forecasts for Gaussian data or categorical data. It appears that the authors are seeking to introduce univariate summary measures of calibration and sharpness of ensemble predictions that yield comparable values across Gaussian or categorical outcomes. I have several severe concerns with the paper which I am listing below.

Detailed comments:

1. The motivation, goal and results of the paper are unclear. Why is it important to compare calibration of predictions for Gaussian outcomes or for categorical outcomes on the same scale? Why can this not be achieved by using a proper scoring rule that applies to both, continuous and discrete outcomes such as the CRPS?

2. The paper suffers from a number of mathematical inconsistencies such as missing assumptions and definitions, or, overly simplistic and, therefore generally incorrect statements. Some examples are

   - Equation (8) suggests that rank histograms can either be flat, inverse U-shaped or U-shaped. However, there are many other shapes possible which are not just due to finite samples. It should be clearly stated and proved under which implications are intended in equation (8) and under which conditions they hold.
   - Equation (9) is introduced as a definition but then it is an inequality. This is particularly confusing in view of the statement on p.16,l.7-8. How is the RPC defined? What do you mean by "the optimal ESS is then equal to the optimal RPC"? Are they just both equal to one for calibrated predictions?
   - p.6,l.4-5: Starting with ensemble forecasts, there are many ways to derive predictive densities p. Similarly, the climate pdf can be estimated in numerous ways from the observations. What do you mean here?

3. The paradigm of Gneiting et al. (2007) to "increase sharpness subject to calibration" is not appropriately applied by the authors. Gneiting

et al. (2007) rigorously define both concepts and sharpness refers to the forecasts only, whereas calibration ensures statistical compatibility between forecasts and observations. This is contradictory to the statement on p.2,l.7–8: Calibration in the sense of Gneiting et al. (2007) is not a balance between sharpness and resolution. A calibrated prediction can be very sharp or not sharp at all.

4. In relation to my previous comments, I have severe reservation to speak of an "optimal" value of ESS being 1. It is not based on a proper scoring rule, and the authors do not give rigorous arguments of what is meant by "optimal" here an why (and under which conditions) this "optimum" is achieved *if and only if* $ESS = 1$.

5. In line with my last comment is the following issue: On p.16,l.12 the authors state that "the ESS of scaled variables contains the same information as the rank histogram". Firstly, if this is true then this is a strong argument against using the ESS to assess the quality of ensembles in terms of calibration *and* sharpness because the rank histogram does not assess sharpness. This can be shown rigorously and examples of this nature are provided in Gneiting et al. (2007). Secondly, in this broad generality, I believe that the statement is false, see my previous comment on equation (8) above.