Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-141-RC1, 2018

Geoscientific Model Development Discussions

Interactive comment on "Common metrics of calibration for continuous Gaussian data and exceedance probabilities" *by* Rita Glowienka-Hense et al.

Anonymous Referee #1

Received and published: 16 November 2018

General Comments

Probabilistic forecasts have recently received much attention across the disciplines. Calibration and sharpness are two common criteria in order to judge the performance of a probabilistic forecast. The present paper develops metrics for calibration and sharpness that (i) relate to information theory, and (ii) apply to continuous and discrete variables alike.

While I am sympathetic with the paper's topic and research question, the paper is not very accessible at present, and there is an important issue regarding the standardiza-

tion of forecasts (and observations) proposed in the paper. Some specific comments follow.

Specific Comments

- While I agree that the paper's objective is interesting, this objective should be better motivated in the introduction, e.g. by mentioning practical situations in which a joint evaluation of continuous and discrete variables is of interest.
- The ensemble-spread score (ESS) at Equation (1) is central to the analysis in Section 2. It seems important to highlight that the ESS is *not* part of the family of (strictly) proper scoring rules (see e.g. Bröcker, 2009, and the references therein). Proper scoring rules are decision-theoretically motivated tools which set the incentive for a forecaster to state what they think is the true distribution of the predictand. They have become a standard tool in recent years, and are probably the concept that many readers have in mind when reading about a 'score'. There are several differences between the ESS considered in the paper and proper scoring rules. First, the ESS is a tool for assessing the performance of a given forecast method, whereas scoring rules are used to compare two or more forecast methods. Second, adopting the paper's notation, the ESS is defined for an entire panel of forecasts $\{Y_{ij}\}$ (with $i = 1, \ldots, nrun$ denoting model runs and $j = 1, \ldots, J$ denoting time) and observations $\{X_j\}$. By contrast, proper scoring rules are defined for an ensemble/observation pair at a given date, such as $\{Y_{i1}\}/X_1$ corresponding to date $j = 1$. Third, the ESS attains its optimal value at one; smaller and larger values indicate a worse forecast, but it is not formally clear whether, say, a value of 0.8 is better or worse than a value of 1.2. By contrast, proper scoring rules attain their best value at zero, with larger values corresponding to worse forecasts. Finally, the ESS used in the paper is not to be confused

with the error-spread score of Christensen et al. (2015) which is a proper scoring rule based on the first three moments of a forecast distribution.

- The paper proposes to consider the ESS for *standardized* forecasts and realizations, as detailed at Equation (5). While I appreciate the simplicity of the characterizations that follow from standardization (see Equation 7), the comments on P5 highlight an important drawback of the methodology: a noninformative forecast ensemble that is drawn from some arbitrary distribution (such as a standard normal) which is the same at each date j will attain $ANOVA = 0$, as well as the best possible ESS value of one. Due to the standardization step, it is not even necessary for the ensemble to be correctly dispersed. Without the standardization step, the ESS could not be tricked so easily, in that it would at least be necessary for the mean model spread σ_e^2 to equal MSE, c.f. Equation (1). It would hence seem important to provide a more detailed motivation of the standardization step. (On P2, L27 the paper mentions that standardization is necessary in order to interpret ESS as a measure of calibration. This argument is not clear to me, and should be elaborated.)

- The presentation in Section 4 is unclear and should be improved.
  - In Equations (15) and (17), relevant notation (D_{KL}, $\hat{X}$, Y_k, etc.) is not defined.
  - The stated interpretation of Equation (19) does not become clear at present. To explain the equation, it is necessary to note that GCC^2/GAC^2 is the discrete analogue of $CORR^2/ANOVA$. The case $GCC^2/GAC^2 = 1$ then corresponds to $CORR = \sqrt{ANOVA}$, which corresponds to a flat rank histogram by Equation (8). Similar analogies apply when GCC^2/GAC^2 is either smaller or greater than one.
  - Relating to the previous comment, the format and wording for Equations (8) and (19) is inconsistent and should be streamlined.

Minor Comments

- There are some formal inconsistencies in the paper's citations, see e.g. "(von Storch and Zwiers, 2001)" versus "DelSole (2004)" on P6.

References

Bröcker, J. (2009): "Reliability, sufficiency, and the decomposition of proper scores," *Quarterly Journal of the Royal Meteorological Society*, 135, 1512-1519.

Christensen, H., I. Moroz, and T. Palmer (2015): "Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts,", *Quarterly Journal of the Royal Meteorological Society*, 141, 538-549.

Gneiting, T. and M. Katzfuss (2014): "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, 1, 125-151.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-141, 2018.