Thanks to the reviewer for reading our paper and providing some helpfull comments. Especially we will try to motivate the use of similar scoring rules for continuous and categorical data as well as the standardization of the ESS score and underline the different model problems for ESS below and above 1.

1 Replies to specific questions

1.1 Why is the evaluation of joint and discrete variables with the same score of interest?

Comparing continuous correlation of e.g. temperature between forecast and observation with transformed mutual information from exceedance probabilities of quantiles can reveal where the overall correlation comes from.

1.2 ESS is not part of the family of strictly proper scoring rules. It attains its optimal value at one. Differing from the ES

Assuming that the discussed time series have Gaussian distributions, i.e. the skewness terms are zero, the error spread score (ES), which is proper, contains the same 2 terms as the ESS.
Thus if the ESS is 1 the ES is zero.
The ESS is not a complete scoring rule in itself.
It is a measure of reliability. It measures whether the observations can be seen as ensemble members of the forecast.
The associated measure of resolution is the correlation and the anova with time as treatment is the associated measure of sharpness.

1.3 ESS attains optimal value one. Is 0.8 worse or better than 1.2?

If the ESS is equal to 1, then the ensemble spread indicates the model uncertainty.
Values of the ESS below and above one should be interpreted differently. Too low ensemble spread and thus too sharp forcast ensembles may well be a problem of model physics and only for very short term forecasts a problem of too small spread of the initial ensembles. ESS values above 1 indicate additional noise in the model. Thus theoretically values below 1 are more problematic than an ESS above one. We will underline this point in the paper.

1.4 ESS does not consist of ensemble observation pairs at given dates

The denominator of the ESS consists of the average over all forecasts of the pairs of ensemble mean at time j and the corresponding observations at time j with the respective distances between ensemble mean and ensemble members at a specific time step.
The ESS is created from the same two variables namely mean square error between ensemble mean and observation and ensemble spread as the error spread score ES which is a proper score. We will describe this point more clearly in the paper.

1.5 Why standardization

The ESS here has been performed with standardized variables. Thus the effects of too large/low model variance or bias - which could be remedied by post processing i.e marginal calibration - have been eliminated.
This helps to separate the marginal calibration from the reliability issue. This is in line with the argumentation of Bröcker (2009), who is dealing with calibration methods. He proposes to take ensembles as a source of information only. Moreover the standardization is a basic feature of regression analysis:

$\hat{Y} = \bar{Y} + \beta X$, where $\beta = \frac{\sigma_y}{\sigma_x} CORR(X - \hat{X})$
$\frac{(\hat{Y}-\bar{Y})}{\sigma_y} = CORR \frac{(X-\bar{X})}{\sigma_x}$
which is one way of taking the information from a forecast. Here X refers to the ensemble mean of the forecast and Y to the observation. CORR is thus the regression coefficient in case of marginal calibration. The inference of our calculations is that in the ideal case of an ESS=1 the ensemble spread is equal to one minus the squared correlation.

1.6 Uninformed forecast with ANOVA=0 has optimal ESS=1

We are aware of the fact and it is discussed in the text that an EPS can be perfectly reliable without being sharp. If a forecast is not sharp then every observation fits into the forecast ensemble. This is why there is a need for at least two of the variables - reliability, resolution and sharpness - to describe the performance of the forecast.

If the skewness is zero, the ES=0 in case the ESS=1. This is also true for ANOVA=0. It can be directly seen because the terms are the same.

2 Page 2 line 27 because only then the score is a measure of calibration as shown here..

Here probabilistic and exceedance calibration is meant and for claritiy will be replaced by the term reliability.

2.1 notation section 4

We will explain the notation in section 4. Further we will eliminate citation inconsistencies.