We thank the reviewer for discussing our manuscript. We hope that we can convince him of the benefit of our methods.

1 Replies to specific issues

1.1 The aim univariate summary measures..

- The *first* main result was to show the relation of the ESS to basic well established scores correlation and ANOVA with time as treatment.
  The standardization we introduced is also called marginal calibration and is a standard procedure in regression analysis. For standardized variables the correlation coefficient is the regression coefficient.
  Assuming Gaussian distribution of the variables the same terms are involved in the ES score (Christensen et al 2015) - to which we were made aware by the other reviewer - if used with standardized variables and assuming Gaussian distributions and thus zero skewness.
  This standardization leads to a reliability measure that is rid of marginal calibration errors.
  The result shows further that the optimal ensemble spread is equal to $1 - CORR^2$ and equivalently the RPC=1. The RPC is defined and discussed in Eade et al, 2014. We simply repeat this discussion.

- The *second* result was that we could show that the ANOVA ratio is very close to the mean *utility* defined by Kleeman (2002). This connects ANOVA analysis and relative entropy.
  The mutual information (MI), which is a special integrated relative entropy between the joint and the marginals of two variables, is directly related to correlation this comes from the literature. Together this shows that the classic tools used for the analysis of forecast ensembles are directly related to relative entropy.

- The *third* concern was to use a similar method for categorical forecasts.

1.2 Why not CRPS?

The CRPS has been shown to be beneficial for evaluating ensemble prediction of financial portfolios where always the complete pdf matters. In case of Gaussian distributed time series it can happen that an EPS with incorrect marginal calibration like too large variance but medium correlation has a larger CRPS than a second ensemble prediction system (EPS) with nearly zero correlation but well

adapted variance. Thus the uninformed second system wins in comparison to the informed first system.
If we perform marginal calibration then the only thing that remains from the orignal ensembles is the relative spread (1-ANOVA) and the ensemble mean.
The CRPS is important to evaluate large ensembles of financial portfolios because here always the complete pdf materializes. The CRPS is also well suited to compare two mean climate projections including their uncertainty.

1.3 The rank histogram can have different forms others than U-shaped and inverse U-shaped

The asymmetric forms of the rank histogram would be due to positive or negative mean biases of the EPS but in our case the data have been standardized (= marginal calibration see above) before the analysis. The rank histograms of these data will be generally U-shaped or inverse U-shaped. Overlays of these shapes are imagineable in case the ensemble sharpness varies with the initial state of the prediction. We guess that the time series of such systems will not have Gaussian pdf as is assumed here.

1.4 Under which conditions eq. 8 holds ...

We will further underline that also the relations to the rank histogram analysis in eq. 8 hold for standardized Gaussian variables. If the ESS for these scaled variables is equal to one or equally the ES be zero the distance of the ensemble members from the mean is the same as the mean distance between the observation and the ensemble mean. Thus the observation behaves like an additional ensemble member and therefore the rank histogram would be flat.
On the other hand if the ESS is less than one the ensemble members are generally closer to the ensemble mean than the observations. As we have scaled the variables to have zero overall mean, this means that the rank histogram is U-shaped. Analogously in case of an ESS greater than one the smaller distances of the observations from the ensemble means leads to an inverse U-shaped rank histogram.

1.5 How is the RPC defined?

The ratio is defined by Eade et al 2014 as a lower bound for the actual ratio of predictable components (RPC) which might be improved by future model developements. We took the definition directly from the paper. However the authors did not use the term ANOVA for the ratio of mean ensemble spread to total spread. They claim that the ratio $\frac{CORR}{\sqrt{ANOVA}}$ should ideally be equal to one without giving any proof.

1.6 Are the ESS and the RPC just equal to one for calibrated predictions?

The ESS can also be equal to one if the variables are not normalized. If this happens for model and the observational data with equal marginal calibration then the model ensemble is indeed reliable. On the other hand differences in the marginal calibration of observations and model can lead to an $ESS = 1$ but without having a reliable forecast ensemble.

Reliability of an EPS implies that the ensemble spread is equal to the mean square errror between observations and ensemble means and thus $ESS = 1$. In case of standardized variables this further implies that the correlation is equal to the square root of the ANOVA. This means that the claim of Eady et al. (2014) is equivalent to reliability of an EPS after standardization of the data.

1.7 Does the RPC need calibration

The ratio of correlation to the square root of ensemble mean to total variance depends on standardized variables. Thus the marginal calibration is inherent.

1.8 How have the predictive densities been derived?

Assuming that both forecasts and observations are Gaussian distributed an overall mean has been determined and the variance is an average variance with respect to that mean. The ensembles here do not show systematic differences. The pdf at a special forecast time is directly determined from the ensemble mean and the ensemble variance.

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

The sharpness measured here with ANOVA is an attribute of the forecasts only as is demanded by Gneiting et al (2007). It is calculated without any reference to observations, it can be generated right after the EPS prediction is available.

Measuring the reliability with standarized/marginally calibrated variables is intended to give an indication whether the sharpness - measured with ANOVA - is indeed associated with calibration (exceedance + probabilistic)/reliability.

1.10 A calibrated prediction can be very sharp or not sharp at all

The ESS analysis includes that forecasts can be reliable/calibrated (beyond marginal calibration) but not sharp at all. If the model sharpness/ANOVA is zero then the forecast is reliable or probabilistic and exceedance calibrated in any case (eq 8). For low correlation the model sharpness/anova should be correspondingly low then the forecast is also calibrated(exceedance + probabilistic)/reliable. Therefore as sharpness is increased the reliability/calibration(beyond marginal calibration) can only be hold ($ESS = 1$) if the resolution/correlation increases accordingly. Thus the reliability/calibration (beyond marginal calibration) is indeed a balance between sharpness and resolution.

1.11 Why is the optimal value of ESS=1

Optimal is meant in the sense that the ensemble spread is equal to the mean square error between observations and ensemble means. The same is demanded in the ES score of Christensen et al (2015) and citations therein in case of Gaussian distributions and thus zero skewness. Your co reviewer pointed us out to this article, we will cite it in our revised version. They have the same two terms - without standardizing the data - but take the squared difference. Thus the ES of standardized variables should be zero in case the ESS is one. The ES is a proper scoring rule. If you perform the same transformations the ES equally only depends on correlation and ANOVA. From the squared difference it can however no longer be determined whether the EPS is over- or underdispersive which we think is important.

1.12 Rank histogram does not assess sharpness

The rank histogram is as the ESS a measure of reliability of the forcast ensemble. In case one uses marginally calibrated data also for the rank histogram a U/inverse-U-shaped rank histogram is indicative of under/over-dispersion. This means on the one hand that the data must have also resolution because otherwise the rank histogram of a marginally calibrated data set of observations and prediction ensemble would be flat. On the other hand the underdispersion/overdispersion is indicative of too large/low sharpness of the forecasts compared to resolution. The rank histogram gives however no quantitative information for this missing balance and is no absolute measure of sharpness. The latter depends on the forecasts only. Such numbers are given by the triplet of ESS, correlation and anova. The relation between rank histogram and ESS only holds for Gaussian distributions and if standardization/marginal calibration is performed.