Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2017-366-RC3, 2017


Hydrology and Earth System Sciences
Discussions

# *Interactive comment on* "On the skill of raw and postprocessed ensemble seasonal meteorological forecasts in Denmark" *by* Diana Lucatero et al.

**Anonymous Referee #3**

Received and published: 18 September 2017

## 1 General comments

The paper by Lucatero et al. describes an assessment of uncorrected and post-processed GCM forecast over Denmark during the period 1990-2013. The study adresses a critical issue for GCM forecast users, especially in the field of hydrology where uncorrected forecast often lacks sufficient skill to be used as input for hydrological applications. Most methods applied in the paper are sound, for example the use of two well established post-processing techniques and a leave-out cross validation scheme. The paper is well written with a clear and concise structure. However, we believe that there is scope to improve its content before it can be published. Our major comments are listed below:



1. The approach that was adopted by the authors to downscale the ECMWF forecast is questionable. The authors applied an inverse weighting algorithm to convert the 70km resolution ECMWF grid to a 10 km resolution, and then used the downscaled data to post-process and analyse forecast performance. By doing this, they smooth ECMWF rainfall surfaces and break the conservation of mass, which artificially reduces the skill of uncorrected forecasts. To circumvent this problem, we recommend performing the analysis undertaken by the authors at the resolution of ECMWF forecasts (i.e. 70km), using a simple aggregation method for gridded observation data (see for example Schepen et al. (2014)). This alternative approach would eliminate the need for a downscaling algorithm, and provide a direct assessment of uncorrected ECMWF forecasts compared to post-processed forecasts.

   We understand the value of downscaling to work at a meaningful scale for hydrological applications. However, downscaling is a research topic in its own, and its impact should not mask the skill of the uncorrected forecasts. Without a proper assessment of uncorrected forecasts, it is difficult to select the appropriate downscaling model.

2. The quality and resolution of several figures is clearly below the standard of an international scientific journal. We strongly recommend redrawing figures 1, 4 and 6, increasing the resolution and/or converting them to a vector format. Unfortunately, with such low figure quality, it becomes difficult to check the comments made by the authors in reference to those figures. Additional comments on the figures are provided in the next section to improve their readability.

3. The analysis of forecast reliability (or statistical consistency as per the author's nomenclature) lacks important information to properly assess forecast performance:

   - It is not clear which variables are used to draw the PIT plots (figure 7). Such

plots require a single series of PIT values computed from matched pairs of observations and forecasts. The authors do not precise if the observations and forecasts are coming from a single grid cell, or from a spatial aggregation (e.g. the whole Denmark). This point is important to understand their difficulties in interpreting the PIT plot (see Line 6 page 9:“issues (...) are not remarkably clear”). We suggest drawing the PIT plots for a selected set of grid points and one lead time representing the main characteristics of the PITs across the study area.

- The analysis of reliability lacks a quantitative criterion similar to the skill scores. A standard approach is to compute the pvalue of a uniformity test such as the Kolmogorov-Smirnov test (Laio and Tamea, 2007). However, instead of the KS test, we recommend the Anderson-Darling test (Anderson and Darling, 1952), which exhibits a greater power (Noceti et al., 2003) and better ability to detect deviations from uniformity of the extremes. We strongly suggest computing the p-value of such tests for all grid points of the domain and all lead times, and then summarise the results by counting how many cells pass the test with a 5% threshold for a given lead time. This will provide a consistent and reproducible assessment of forecast reliability across the study area.

4. Temperature is treated differently than rainfall and PET both in the computation of the bias score and in the implementation of the LS post-processing model. This is quite confusing and should be harmonised. We recommend that all bias scores be computed in relative terms to facilitate comparison. The LS model should be applied to the three variables in two modes: multiplicative (which is the configuration used for P and PET) and additive (used for T). This approach would provide a clear advice on the choice of the LS configuration.

5. It is not clear how extrapolation is undertaken within the QM model. The authors only indicate that ‘’Extrapolation is then needed to map ensemble values and



percentiles that are outside the training range” (see Page 4, Line 21). This is a critical and frequent problem with the QM method which affects extremes, and is then of particular interest in hydrological forecasts. Please detail the extrapolation method.

Considering that the paper is covering an important topic for seasonal forecasting, but that the number of points to be improved is quite significant, we recommend the paper to be accepted with major revision. Detailed comments are provided in the following section.

## 2 Specific comments

1. Page 2, Line 26, “The most used methods are linear scaling and quantile mapping”: Zhao et al. (2017) provides an interesting perspective on the limitations of Quantile-Quantile mapping that could be cited here.

2. Page 3, Line 4, “A statistical consistent forecast system has low (or non-existent) bias in both mean and variance”: correct? ok.

3. Page 3, Line 14, “we make use of the Makkink equation (Hendriks, 2010) that takes as inputs temperature and incoming short-wave solar radiation from ECMWF System 4”: a mention on the quality of radiation forecast would be useful. A full analysis of radiation forecast performance is out of scope here, but perhaps some references can be cited to understand the impact of radiation inputs in ET0 calculation.

4. Page 3, Line 20, “The time and spatial variations of the variables can be seen in Fig. 1.”: Fig 1 is not readable, so the accuracy of this comment is hard to assess.

5. Page 3, Line 38, “once a scale factor has been applied”: In the case of T, the factor is not a scale, but a shift.

6. Page 6, Line 2, “a Wilcoxon-Mann-Whitney test was carried out.”: What was the data used to apply the WMW test? Was it applied to all grid cells for a single lead time? Please clarify.

7. Page 6, Line 39, “April shows an overestimation that might be due to the ‘drizzle effect’ in a month where dry days are more common.”: It is quite surprising that this “drizzle effect” is not affecting the bias in March and May where the bias shows an opposite trend compared to April. This statement requires more explanations. We suggest showing the number of dry days per month, to confirm that April has the highest proportion.

8. Page 11, Line 13, “One advantage ECMWF System 4 has over ensemble climatology is that forecasts are sharper”: The forecast is sharper but it is not clear if it is reliable. In this case, being sharp is not an advantage, but a problem.

9. Page 11, Line 24, “QM assumes that there is a linear relationship between ensemble mean and observations, assumption may not hold”: We believe that the authors mean LS instead of QM here. Otherwise we do not see why QM would assume such linear relationship.

10. Figure 5: The figure is trying to convey too many information at the same time with a confusing choice of symbol sizes (decreasing symbol size for better sharpness skill score) and color schemes. We suggest using the same presentation than Figure 3 and grouping the two figures into a single one that would provide an homogenous overview of forecast performance.

11. Figure 6: The scale of the color bar is different between the 3 plots. As a result, it is impossible to compare the skill between the three variables.



12. Figure 8: This plot is difficult to read because the legend does not explain that each location has a different symbol and all curves are shown with the same color. We suggest splitting each one of the three plots into 4 different subplots showing the results for one location only.

**References**

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics*, pages 193–212.

Laio, F. and Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11(4):1267–1277.

Noceti, P., Smith, J., and Hodges, S. (2003). An evaluation of tests of distributional forecasts. *Journal of Forecasting*, 22(6-7):447–455.

Schepen, A., Wang, Q., and Robertson, D. E. (2014). Seasonal forecasts of australian rainfall through calibration and bridging of coupled gcm outputs. *Monthly Weather Review*, 142(5):1758–1770.

Zhao, T., Bennett, J. C., Wang, Q., Schepen, A., Wood, A. W., Robertson, D. E., and Ramos, M.-H. (2017). How suitable is quantile mapping for postprocessing gcm precipitation forecasts? *Journal of Climate*, 30(9):3185–3196.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2017-366, 2017.