**To what extent does the diagnosis of multiple grid-box weather types add value in post-processing ensemble rainfall forecasts?**

Fatima Pillosu1,2, Tim Hewson2

1 University of Reading, Reading, UK

2 ECMWF, Reading, UK

**Abstract**

**Plain Language Summary**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Introduction

# Data

## Raw ECMWF ensemble

## ecPoint post-processing technique

## Rainfall observations

# Methods

Reliability (i.e., when x% probability is forecast is the event observed on x% of occasions?) and discrimination ability (i.e., the ability to distinguish between event and non-event) are desirable properties of ensemble forecasts (Wilks, 2019). While calibration deals with the meaning of probabilities, discrimination appraises the existence of a signal in the forecasts when an event materializes and its absence in the opposite situation (Ben Bouallègue & Richardson, 2022). To have value, post-processing techniques must improve both properties in the raw NWP forecasts. The performance of one year of retrospective forecasts from the three systems, the raw NWP, ecPoint-MultipleWT, and ecPoint-SingleWT, was compared using 12-hourly rainfall observations from the datasets described in section 2.3 as truth. Although raw NWP model output does not pertain to point values, it is very common to verify it using them, as in ECMWF’s two headline measures for precipitation (Haiden et al., 2023).

## Estimation of forecast reliability: Brier Score – Reliability Component

The reliability component of the Brier score is used to assess the reliability of the examined forecasts (Jolliffe & Stephenson, 2011).

## Estimation of “real” forecast resolution: ROC curves and area under the ROC curve computed with the trapezoidal approximation

The area under the relative operating characteristic (ROC) curve (AROC) is used as a summary measure of the forecasts’ discrimination ability (Jolliffe & Stephenson, 2011). The ROC plots the hit rate (HR) versus the false alarm rate (FAR) of an event for incremental decision thresholds. The ROC curve is defined by the line joining successive ROC points, where each point corresponds to results for increasing decision thresholds, from the top right to the bottom left corners of the plot.

The decision variable is the number of members exceeding the event threshold. Thus, the issued forecasts can take values in [0, 1/M, 2/M, …, 1] for an ensemble of size M. As a consequence, the resulting ROC curves are based on M+1 points (i.e., 52 for the raw ENS, and 100 for ecPoint-MultipleWT and ecPoint-SingleWT).

The area under the straight lines formed by connecting the M+1 points [including the (0,0) and (1,1) points] of the ROC plot correspond to the AROC with the so-called trapezoidal approximation (AROCT). This nomenclature come from the fact that the area is estimated considering straight lines between two consecutive points of the plot and so as a sum of trapeziums.

For rare events, there is a tendency for the points on the ROC to cluster toward the lower left corner of the graph (Casati et al., 2008). When computing AROCT, a straight line is drawn between the last meaningful point on the ROC curve and the top-right corner to close the ROC curve, giving the impression that part of the curve is missing. How much of the curve is missing depends on the lowest category, defined here by the ensemble size and the base rate of the event.

## Estimation of “potential” forecast resolution: Binormal approximation for the computation of ROC curves and area under the ROC curve

To draw a full ROC curve, one can apply the so-called binormal model (Harvey et al., 1992). The fitting of the ROC curve with the binormal model is based on the assumption that HR and FAR are integrations of normal distributions (i.e., a signal and a noise distribution), respectively. Harvey et al. (1992) provide a close-form for the computation of the AROC. The fitting of the HR and FAR requires a Z-transformation based on the unit normal distribution. For this reason, the result AROC is denoted as AROCZ.

When applied to ensemble-derived probability forecasts for rare events, this approach consists effectively in an extrapolation to a hypothetical continuous decision variable based on the limited set of decision thresholds materially assessable. Because such a decision variable may not be achievable in practice, AROCZ is sometimes considered as a measure of the potential discrimination that could be achieved for an unlimited ensemble size (Ben Bouallègue & Richardson, 2022; Bowler et al., 2006).

# Results

## Goodness fit test of the binormal model

**Figure 2** shows the results of the goodness fit test for the real ROC curves using the binormal model. The panels on the left (**a**, **c**, and **d**) show the z-score hit rate as a function of the z-score false alarm (red dots), and how they align with the prediction of the prediction of the best-fitting Gaussian signal detection model, represented straight line in z-score coordinates (shown in blue).

for ENS (left), ecPoint\_MultipleWT (centre), and ecPoint\_SingleWT (right).

The red dots, representing the z-score hit rate as a function of the z-score false alarm for ecPoint\_MultipleWT and ecPoint\_SingleWT follow the blue straight line, representing the prediction of the best-fitting Gaussian signal detection model (which in z-score coordinates is represented by a straight line).

**Figure 2a and Figure 2b** show a case of good fit. In **Figure 2a,** the red dots are aligned with the blue line, so that the ROCz represents a good continuation of the “real” ROC curve (**Figure 2b**).

## “Real” and “potential” forecast resolution

## Forecast reliability

Figure 2 displays values for the reliability component of the Brier Score (BSrel) for three verifying rainfall events (VRE) greater than 0.2 mm/12h (“dray or not”), 10 mm/12h (“wet conditions”), and 50 mm/12h (“extreme rainfall, with flash flood potential”).

# Case Study: Extreme rainfall and flash floods in China in July 2021

# Discussion

## Forecast reliability

## “Real” and “potential” forecast resolution

# Conclusions

# Figures

A diagram of a multi-wt system

Description automatically generated

**Figure 1 – Schematic representation of ecPoint’s single and multiple weather type approach.** Panel (a) shows the error formulation between forecasts and observations adopted for accumulated variables, called Forecast Error Ratio (FER). Panel (b) shows a visual representation of the error distribution, called Mapping Function (MF). The example pertains to the calibration of ENS 12-hourly rainfall forecasts for 47r3. Panel (c) shows the options of MFs adopted in ecPoint. If the MF for all data points, shown in panel (b), is split according to different grid-box Weather Types (WT) defined using predictors such as mainly large-scale or convective rainfall, rainfall totals, etc., each grid-box is post-processed according to its correspondent grid-box WT, and the post-processing approach is called ecPoint\_MultipleWT. The different grid-box WTs are represented using a decision tree (DT) representation (enclosed in the grey rectangle, DT partially shown). Different colours are assigned to leaves of the DT belonging to different predictors. If the MF for all data points is not split, all grid-boxes are post-processed using the same MF (enclosed in the black circle), and the post-processing approach is called ecPoint\_SingleWT (represented as a single leaf, as opposed to the tree-like representation of the ecPoint\_MultipleWT approach).

A collage of graphs

Description automatically generated

**Figure 2 – Goodness fit test for the fitting of the “real” ROC curves using the binormal model.** Panel (a) shows the z-score hit rate as a function of the z-score false alarm (red dots) for ENS (left), ecPoint\_MultipleWT (centre), and ecPoint\_SingleWT (right). The solid blue line represents the prediction of the best-fitting Gaussian signal detection model, which is a straight line in z-score coordinates. All three plots refer to the verifying rainfall event (VRE) >= 50 mm/12h and to the 12-accumulation period ending at t+126. Panel (b) shows the ROC curves for the same VRE and rainfall accumulation for ENS (turquoise), ecPoint\_MultipleWT (orange) and ecPoint\_SingleWT (grey). The “real” (ROC) and the “binormal” (ROCz) ROC curves are represented, respectively, with continuous and dashed lines.

A graph of different colored lines

Description automatically generated with medium confidence

**Figure 3 – “Trapezoidal” and “Binormal” Areas under the ROC curve (AROCt and AROCz, respectively) for lead times up to t+246 (i.e., day 10).** The panels from (a) to (d) show AROCt (continuous lines) and AROCz (dashed lines) values for verifying rainfall events (VRE) >= 0.2, 10, 25, and 50 mm/12h, respectively. The turquoise, orange and grey lines represent AROCt and AROCz values for ENS, ecPoint\_MultipleWT and ecPoint\_SingleWT, respectively. A 99% confidence level is applied to error bars represented by the shaded areas.

A graph of a function

Description automatically generated with medium confidence

**Figure 4 – ROC curves for the 12-hourly accumulation period ending at t+126 (i.e. day 5).** The panels from (a) to (d) show the “real” ROC curves (continuous lines) and the “binormal” ROC curves (dashed lines) for verifying rainfall events (VRE) >= 0.2, 10, 25, and 50 mm/12h, respectively. The turquoise, orange and grey lines represent the ROC curves for ENS, ecPoint\_MultipleWT and ecPoint\_SingleWT, respectively.

A screenshot of a graph

Description automatically generated

**Figure 5 – Reliability component of the Brier Score (BSrel) for lead times up to t+246 (i.e., day 10).** The panels from (a) to (d) show the BSrel values for verifying rainfall events (VRE) >= 0.2, 10, 25, and 50 mm/12h, respectively. The turquoise, orange and grey lines represent BSrel values for ENS, ecPoint\_MultipleWT and ecPoint\_SingleWT, respectively. A 99% confidence level is applied to error bars represented by the shaded areas.

# References

Ben Bouallègue, Z., & Richardson, D. S. (2022). On the ROC area of ensemble forecasts for rare events. *Weather and Forecasting*, *37*(5), 787–796. https://doi.org/10.1175/waf-d-21-0195.1

Bowler, N. E., Pierce, C. E., & Seed, A. W. (2006). STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Quarterly Journal of the Royal Meteorological Society*, *132*(620), 2127–2155. https://doi.org/10.1256/QJ.04.100

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E. E., Brown, B. G., & Mason, S. (2008). Forecast verification: Current status and future directions. *Meteorological Applications*, *15*(1), 3–18. https://doi.org/10.1002/met.52

Haiden, T., Janousek, M., Vitart, F., Ben-Bouallegue, Z., & Prates, F. (2023). Evaluation of ECMWF forecasts, including the 2023 upgrade. *ECMWF Technical Memoranda*, *911*, 1–60.

Harvey, L. O., Hammond, K. R., Lusk, C. M., & Mross, E. F. (1992). The Application of Signal Detection Theory to Weather Forecasting Behavior. *Monthly Weather Review*, *120*(5), 863–883. https://doi.org/10.1175/1520-0493(1992)120

Jolliffe, I. T., & Stephenson, D. B. (2011). Forecast Verification: A Practitioner’s Guide in Atmospheric Science. In *Forecast Verification* (2nd Editio). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119960003

Wilks, D. S. (2019). Statistical Methods in Atmospheric Sciences. In *Statistical Methods in the Atmospheric Sciences* (Fourth Edi). Elsevier Inc. https://doi.org/https://doi.org/10.1016/C2017-0-03921-6