**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

Weather forecasts nowadays rely heavily numerical weather prediction (NWP) models (Bauer et al., 2015), and commonly an ensemble of predictions is used to represent uncertainties (Buizza, 2019). NWP models do not output forecasts at specific sites (that most customers require) but instead output “average” values for grid boxes. If sub-grid variability (i.e., the variation seen amongst all point values observed within the same model gridbox) is low, then raw NWP forecasts are representative of point values. But if sub-grid variability is high raw NWP forecasts inevitably fail. The scale mismatch between in situ observations and gridded NWP forecasts is called representativeness error (Janjić et al., 2018). The two most common strategies to address sub-grid variability problems are using km-scale NWP models (e.g., ∼2 km or less), or using post-processing techniques to statistically convert the forecasts from grid box averages to point values. Km-scale resolution models show a much more realistic-looking spatial patterns and exhibit improvements in forecast skill where the more detailed representation, for example of complex orography, matters. Due to computational constraints, they have limited geographical coverage and forecasts’ lead times are typically limited to day 2-3. Although improvements for extreme events can still be done, post-processing techniques are a better prospect to provide rainfall forecasts at point scale over a continuous global domain.

ecPoint is a novel statistical post-processing technique that provides probabilistic point-rainfall forecasts at point scale for temperature and rainfall. In this study, only rainfall will be considered. Using the “remote calibration” approach defined in Hewson and Pillosu (2021), this post-processing technique has been able to post-process rainfall forecasts over a continuous global domain, also in places with no observations, and using as little as one year of calibration data (forecasts and observations). Sub-grid variability in rainfall is itself very variable (Fig. 1) and relates closely to the weather situation. There are clear-cut physical reasons for this. Dynamics-driven (large-scale) rainfall, often related to atmospheric fronts, arises from steady ascent of moist air across regions typically larger than GM scales (Fig. 1a). As rainfall rates mirror ascent rates, rainfall rate sub-grid variability tends to be small. Conversely instability-driven rainfall (i.e., showers/convection) arises from localised pockets of rapid ascent, which are typically hundreds of metres to kilometres across. During convection rainfall rate sub-grid variability, on GM scales, can be very large. The embodiment of ecPoint is that features of the NWP gridbox forecast output (and other global datasets) can tell us what degree of sub-grid variability to expect. For example, NWP output commonly subdivides rainfall into dynamics-driven and convective, and then for convective cases shower movement speed can be approximated by (e.g.) the 700 hPa wind speed. By using the convective rainfall fraction and the 700 hPa wind speed (two “governing variables”), “gridbox-weather-types” can be distinguished to anticipate a priori the expected sub-grid variability, and accordingly convert each forecast for each gridbox into a probabilistic point rainfall prediction (going from red to blue line in the PDFs in Figure 1). The logic outlined above could be successfully applied to a single deterministic forecast but in NWP ensembles furnish the most useful predictions2,15. So instead we apply separately to each ensemble member, creating an ensemble of probabilistic realisations (or “ensemble of ensembles”) that we merge to give the final probabilistic point forecast. Figure 1 shows only a sample of gridbox-weather-types. ECMWF’s current ecPoint-Rainfall system uses 300 G-WTs defined in decision tree form.

Knowledge this general approach, based on first principles of precipitation generation, has not been used before excepting in a limited way for nowcasting14 (Table 1, row 11). Generally, post-processing systems might not post-process rainfall differently according to different grid-box weather types. For example, Bouallegue (2021) advocates from a post-processing method that adds random noise to the rainfall forecasts to address the representative problem. While these systems are easier to calibrate than ecPoint, it is argued here that the cost of growing a decision tree for defining the G-WTs is worth it. This study compares rainfall forecasts from three forecasting systems: raw ECMWF ensemble forecasts (ENS), original ecPoint rainfall forecasts post-processed by a number of G-WTs (ecPoint\_MultipleWT), and ecPoint rainfall forecasts post-processed with the inclusion of a single WT, represented by no subdivision according to different predictors (ecPoint\_SingleWT). Results for a one-year verification period are presented. Results regards to the main two features of probabilistic forecasts (reliability and discrimination ability) are used to draw conclusions on the performance comparison for the three forecasts.

This study is structured as follows: section 2 describes the data (i.e., forecasts and observations) used in this verification study, section 3 described the methods used to compare the performance of the three considered forecasting systems, section 4 presents the results of the objective verification, section 5 show the results of a case-study-based subjective verification. Section 6 discusses the results obtained from the objective and subjective verification, and section 7 draw the final conclusions for the study.

# Data

## Raw ECMWF ensemble

## ecPoint post-processing technique

## Rainfall observations

# Methods

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

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 “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

## Forecast reliability: Brier score – reliability component

**Figure 2** shows the reliability component of the Brier Score (BSrel) for four verifying rainfall events (VRE): greater than 0.2 mm/12h (“dry or not”), 10 mm/12h, (“wet conditions”), 25 mm/12h and 50 mm/12h (the latter two representing “extreme rainfall, with flash flood potential” depending on local climatology conditions).

ecPoint\_MultipleWT’s reliability consistently outperforms the other systems, across all VREs and LTs. In contrast, ENS demonstrates the least reliability for VREs >= 0.2, 10, and 25 mm/12h. For the VRE >= 50 mm/12h, ecPoint-SingleWT shows a worse reliability than ENS up to t+120 (i.e., day 5). ENS shows again a worse reliability for longer lead times. For all three systems, the BSrel values diminish with increasing lead times, with a more rapid decrease at shorter lead times, especially for ENS. Out of the three forecasting systems, ecPoint\_MultipleWT (orange line) shows the most horizontal trend, meaning that the forecast reliability tends to not change significantly with lead time. In relative terms, the largest impact in the improvement of forecast reliability is observed for VRE >= 0.2 mm/12h and at shorter lead times for all VREs, where the relative distance between the green (ENS), grey (ecPoint\_SingleWT), and orange lines (ecPoint\_MultipleWT) is the greatest.

When examined as a function of lead time, the BSrel values for ENS (green line) exhibit a pronounced sinusoidal pattern. Notably, specific accumulation periods ending at t+30, t+54, t+78, ……, t+246 show enhanced reliability (lowest points on the curve, or throughs) compared to other periods ending at t+12, t+18, t+36, t+42, …..., t+228 and t+234 (highest points on the curve, or crests). Such pattern is particularly evident for small VREs (e.g., 0.2 and 10 mm/12h), and becomes noisier for large VREs (e.g., 25 and 50 mm/12h) likely due to the limited sample size associated with these larger events. Overall, the BSrel values for ecPoint\_SingleWT (grey line) and ecPoint\_MultipleWT (orange line) show a much less pronounced sinusoidal pattern than ENS. However, for larger VREs, the sinusoidal pattern becomes more pronounced for ecPoint\_SingleWT than for ecPoint\_MultipleWT.

The overall uncertainty in the estimation of the BSrel values is largest for ENS (i.e., large confidence intervals) and smallest for ecPoint\_MultipleWT (i.e., smaller confidence intervals). Yet, steps that mark the end of the 0-12 UTC accumulation period (i.e., t+12, t+36, t+60, …., t+228) in VRE >= 50 mm/12h, the uncertainty in the estimation of BSrel values peaks and become much higher than in the other forecasting systems. Up to VRE >= 10 mm/12h, the difference in reliability between ecPoint\_MultipleWT and ENS remain significantly different (i.e., their confidence intervals do not overlap) with a 99% confidence level, at all lead times. They remain significantly different also for VRE >= 25 mm/12h up to t+72 (i.e., day 3 forecast). The reliability difference between ENS and ecPoint\_SingleWT are shown to not be significantly different from VRE >= 10 mm/12h. The reliability difference between ecPoint\_SingleWT and ecPoint\_MultipleWT is not significant from VRE >= 25 mm/12h.

## Forecast discrimination ability: ROC curves and area under the ROC curve

## Goodness fit test of the binormal model

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

China's Henan Province experienced flooding between 17 and 31 July 2021 as a result of heavy rainfall. On July 20, Zhengzhou, the provincial capital, recorded 201.9 millimetres of rain between 4 and 5 pm local time (the highest ever figure recorded since measurements began in 1951), and between 00 an 12 UTC, 465.8 mm were observed (**Figure 6**, top right panel). Over the course of three days, between 20:00 on 17 July to 20:00 on 20 July, 617.1 mm of rain was recorded, nearing the usual average yearly precipitation. The floods caused the evacuation of 815,000 people and affected 14.5 million people around the province. The death toll reached the 398 deaths. **Figure 6** (top left panel) shows some of the flood impacts. Videos emerged, showing cars floating in streets and Zhengzhou Metro passengers waist-deep in water inside their carriage. Many cars on a road near the Danshi Subdistrict in Nanlong Lake were washed up by the rain. The Jingguang North Road Tunnel became flooded, trapping over 200 cars within.

The bottom panel in **Figure 6** compares the 12-hourly rainfall forecasts for ENS (first row), ecPoint\_MultipleWT (second row), and ecPoint\_SingleWT (third row), valid for the same period of the observations. The first three columns show the 99th percentile for day 5, 3, and 1 forecasts (from left to right). Focusing on the Zhengzhou area (within the black circle), the rainfall forecasts from ecPoint\_SingleWT overestimate significantly the observed rainfall totals (>700 mm/12h, when the highest record was 465.8 mm/12h). On the contrary, ecPoint\_MultipleWT appears to be more reliable providing a forecast that is of the same order of magnitude of the observations, even though forecasting a bit less rain than what observed (~ 400 mm/12h). ENS predicts well in advance (up to day 5) that the areas around Zhengzhou might receive some high rainfall totals. This shows that ENS is capable to identify areas at risk of flash floods. However, the absolute rainfall forecasts (up to 150 mm/12h) were underestimated by a factor of three. The fourth column of the bottom panel in **Figure 6** shows the probability of having less than 0.2 mm/12 (i.e., having no rain) for day 1 forecasts south-west of Zhengzhou (blue circle). ENS overestimates significantly the small-to-zero rainfall amounts, showing zero probability of having no rain. This is a typically observed condition during convective rain. ecPoint\_SingleWT shows much smaller probabilities than ecPoint\_MultipleWT of having no rain (between 20 and 25% instead of 50 to 80%). Therefore, the latter verifies better.

# Discussion

## Forecast reliability

The overall significant improvement in reliability by post-processed rainfall forecasts is consistent with results in previous studies (Saetra et al., 2004; Candille et al., 2008, Bouallegue et al. ,2020). The ensemble spread is increased in both post-processed forecasts. Therefore, the variability of point observations is better captured. However, the deterioration of ecPoint\_SingleWT’s reliability compared to ENS for very extreme rainfall events (VRE >= 50 mm/12h) shows that ecPoint\_SingleWT tends to overpredict the probabilities of extreme rainfall events. The ensemble spread (and, consequently, forecast error) is limited at short lead times. At these forecast ranges, the scale mismatch between model and observations plays a substantial role in the general impact of accounting for observations uncertainty. This is less the case at longer ranges when the ensemble spread (and forecast error) is larger. Therefore, post-processing shows it highest impact in short-range forecasts. This is consistent to results in Bouallegue et al. (2020).Furthermore, smaller BSrel values indicate an improvement in forecast reliability. However, the decrease of BSrel values with increasing lead times should not be read as the forecast reliability improves with increasing lead times because this contradicts forecasts’ expected error growth with increasing lead times. This can be potentially an artefact of the score used.

The observed sinusoidal pattern in the BSrel diagrams for ENS can be attributed to the current handling of the diurnal cycle in the ECMWF model (Bechtold et al., 2014). ENS convective rainfall have been found to perform in a more unrealistic manner during daytime (Section 9.6. Convective precipitation. Owens & Hewson, 2018). Although nighttime convective precipitation remains underestimated (Section 2.1.5.4. Convective precipitation. Owens & Hewson, 2018), having the crests (i.e., worse reliability) over daytime accumulation periods and the throughs (i.e., better reliability) over nighttime accumulation periods is found to be plausible. Addressing general representativeness errors in ecPoint\_SingleWT or using specific diurnal-cycle-related errors distributions in ecPoint\_MultipleWT (accounting for them using a predictor that represents daily accumulation of solar radiation) improves the shortcomings in ENS rainfall forecasts. The sinusoidal pattern is indeed significantly smoothed out for both post-processed rainfall forecasts. However, ecPoint\_MultipleWT remains smoother as the VREs increase thanks to a more target approach aimed to specifically tackle the diurnal cycle issues, excluding noise effects due to small sample sizes for very large VREs.

## Forecasts discrimination ability

# 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 screenshot of a graph

Description automatically generated

**Figure 2 – 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.

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 collage of graphs

Description automatically generated

**Figure 5 – 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.

Immagine che contiene testo, mappa, fiore, schermata

Descrizione generata automaticamente

**Figure 6 – Flash floods in Zhengzhou (Henah, China) on the 20th of July 2021.** The panel’s top-left side shows images of the impacts of the flash floods in Zhengzhou (credits to China Dialogue and CNN for top and bottom image, respectively). The panel’s top-right shows 12-hourly rainfall observations valid between 0 and 12 UTC on the 20th of July 2021. The bottom panel shows 12-hourly rainfall forecasts for ENS (first row), ecPoint\_MultipleWT (second row), and ecPoint\_SingleWT (third row) valid for the same period of the observations. The first three columns show the 99th percentile for day 5, 3, and 1 forecasts (from left to right). The fourth column shows the probability of having less than 0.2 mm/12 (i.e., having no rain) for a day 1 forecast.

# References

Bechtold, P., Semane, N., Lopez, P., Chaboureau, J. P., Beljaars, A., & Bormann, N. (2014). Representing equilibrium and nonequilibrium convection in large-scale models. *Journal of the Atmospheric Sciences*, *71*(2), 734–753. https://doi.org/10.1175/JAS-D-13-0163.1

Ben Bouallegue, Z., Haiden, T., Weber, N. J., Hamill, T. M., & Richardson, D. S. (2020). Accounting for Representativeness in the Verification of Ensemble Precipitation Forecasts. *Monthly Weather Review*, *148*(5), 2049–2062. https://doi.org/10.1175/mwr-d-19-0323.1

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

Owens, R. G., & Hewson, T. (2018). *ECMWF Forecast User Guide*. https://doi.org/10.21957/m1cs7h

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