**ECMWF TECHNICAL MEMORANDUM**

Does the diagnosis of multiple grid-box weather types add value when post-processing ensemble rainfall forecasts?

Fatima M. Pillosu1,2, Tim Hewson2, Christel Prudhomme2,3,4, Elisabeth Stephens1,5,6, Hannah L. Cloke1,5,7,8

1 Department of Geography and Environmental Science, University of Reading, Reading, UK

2 Forecast Department, European Centre for Medium-range Weather Forecasts, Reading, UK

3 Department of Geography and Environment, University of Loughborough, Loughborough, UK

4 UK Centre for Ecology and Hydrology, Wallingford, United Kingdom

5 Department of Meteorology, University of Reading, Reading, UK

6 Red Cross Red Crescent Climate Centre, The Hague, The Netherlands

7 Department of Earth Sciences, Air, Water and Landscape Science, Uppsala University, Sweden

8 Centre of Natural Hazards and Disaster Science, CNDS, Sweden

**Correspondence:** Fatima M. Pillosu(fatima.pillosu@ecmwf.int)

**Abstract.**

**Plain language summary.**

**Word count.** x words, excluding abstract, tables, captions, and references.

**Keywords.** Ensemble rainfall forecasts, ecPoint, Brier Score, ROC curves

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

# Introduction

Statistical post-processing based on weather scenarios remains relatively uncommon in operational environments (Roberts et al. 2023). Despite the argument that weather-scenario-based approaches should provide the best results in improving raw numerical weather prediction (NWP) outputs (Hewson and Pillosu 2021), no-weather-scenario-based approaches are still widely used because of their low calibration and production costs (Ben Bouallegue et al. 2020). This study compares the performance of forecasts post-processed with or without a weather-scenario-based approach to assess whether the former adds value to the latter.

Weather forecasts rely heavily on NWP models (Bauer et al. 2021), but their use is not always straightforward. Due to the scale mismatch between the site-specific predictions required by users and the NWP models' gridded predictions, the former can be largely inaccurate under certain conditions (Göber et al. 2008). This scale mismatch is called representativeness error (Janjić et al. 2018) and increases when the variation seen among observed point values within the model grid-box (i.e., sub-grid variability) is significant. Sub-grid variability relates closely to weather conditions. Dynamics-driven (large-scale) rainfall, often related to atmospheric fronts, arises from a steady ascent of moist air across regions typically larger than model grid-box scales. Thus, rainfall sub-grid variability tends to be small. Conversely, instability-driven rainfall (i.e., showers/convection) arises from localised pockets of rapid ascent hundreds of metres to kilometres across. Thus, rainfall sub-grid variability can be very large on model grid-box scales. Representativeness errors can be addressed by adopting ensembles (Buizza 2019) and increasing the spatial scale of NWP models (Cafaro et al. 2021). Km-scale models display realistic-looking spatial patterns (Roberts 2008) and improve forecast accuracy by better representing complex features such as orography (Casaretto et al. 2022). However, their geographical coverage is limited, with lead times rarely exceeding day 3. Although research in developing km-scale models is advancing rapidly (Zeman et al. 2021), statistical post-processing techniques still offer a cost-effective way to provide corrected, downscaled forecasts (Vannitsem et al. 2021).

The ecPoint statistical post-processing technique transforms global raw, gridded NWP model outputs into probabilistic predictions at point-scale, at a fraction of the cost of producing global km-scale forecasts (Hewson and Pillosu 2021). ecPoint's embodiment acknowledges that features of the NWP grid-box forecast output can tell what degree of sub-grid variability to expect for the considered variable. For example, "grid-box weather-types (G-WT)" can be defined using the convective rainfall fraction and the speed of steering winds at 700 hPa to anticipate the expected sub-grid variability in the case of mainly convective rainfall and low steering winds. The current ecPoint operational systems use more than 400 G-WTs and have been shown to improve the reliability and discrimination ability of raw rainfall and temperature forecasts, across different lead-time ranges, against point verification (Hewson and Pillosu 2021; Gascón et al. 2023; Hewson et al. 2023). However, is it worth bearing the cost of defining G-WTs, or would similar improvements be achievable with a post-processing system that does not differentiate between weather scenarios? To answer this question, this study compares rainfall forecasts' reliability and discrimination ability from the raw ECMWF ensemble (ENS), the original multiple G-WT ecPoint (Multiple-WT), and an experimental single G-WT ecPoint (Single-WT) represented by a no-weather-scenario-based approach. The study will focus on rainfall.

Section 2 describes the data used in the verification analysis, while section 3 describes the methods used to compute the forecasts' reliability and discrimination ability. Section 4 presents the results of an objective verification analysis, while section 5 shows the results of a case-study-based subjective verification analysis for an extreme rainfall event. Section 6 discusses the results, while section 7 draws the study's final remarks.

# Data

## Forecasts: ENS

The ECMWF ENS consists of one control run starting from the best possible representation of unperturbed initial conditions and 50 perturbed members starting from perturbed initial conditions (using singular vectors and a data assimilation ensemble) and stochastic model uncertainties (Buizza 2019). Up to day 15, ENS forecasts are stored in a native octahedral reduced-Gaussian grid with a resolution of ~18km at the equator (Owens and Hewson 2018). The forecasts for the considered verification period belong to the 47r3[[1]](#footnote-2) cycle.

## Forecasts: ecPoint

ecPoint is a statistical post-processing technique that transforms global gridded NWP outputs into probabilistic point-scale forecasts (Hewson and Pillosu 2021). The post-processing technique aims to provide forecasts that mirror observations from rain gauges by addressing the two main factors affecting the performance of global NWP model outputs against point verification: systematic biases (Lavers et al. 2021) and lack of representation of forecast sub-grid variability (Göber et al. 2008). The errors between forecasts and observations are computed for the calibration period. The Forecast Error Ratio (FER) is computed for accumulated variables such as rainfall, and its formulation is shown in **Figure 1a**. The graphical representation of the FER distribution is called Mapping Function (MF). The MF for all data points in the calibration period (**Figure 1b**) is differentiated according to weather scenarios defined at a grid-scale called grid-box weather types (G-WT). The derived MFs can be visualised with a decision-tree-like representation shown in **Figure 1c**. In the current operational system, each MF is sampled with 100 FER values, which are used as multiplying factors to post-process each raw ENS member. The (51 x 100) 5100 point-scale values are then distilled in 99 percentiles to reduce archive memory, and can be considered as the ensemble members of the ecPoint post-processed forecasts.

The MF shapes, linked to the expected degree of sub-grid variability, can change significantly according to the weather scenario. When on a grid-box, the raw ENS predicts high totals of mainly large-scale rainfall and strong steering wind speeds (case A in Figure 1c), and the MF takes a Gaussian-like form, meaning that the raw model output is representative of the point-scale rainfall totals. When the raw ENS predicts mainly convective rainfall with light steering wind speeds (case B in **Figure 1c**), the MF takes an exponential-like form, meaning that the expected degree of sub-grid variability is bigger than in case A. The diversity in the MFs shapes means that the corrections applied to the raw forecasts differ from grid-box to grid-box, and that the methodology can judge where to increase or decrease the expected point rainfall values according to the different expected degrees of sub-grid variability. Generally, ecPoint will show the tendency to increase the number of zeros (first row of **Figure 2**) to correct ENS’s tendency to overpredict small rainfall totals (Haiden et al. 2023), and to increase the amounts in the distribution’s wet tail (second row of **Figure 2**) to correct for ENS’s underestimation of high rainfall values (Haiden et al. 2023). However, Hewson and Pillosu (2021) showed that ecPoint can also change the location of the areas at risk of extreme localised rainfall as indicated in the raw ENS if the G-WT suggests that ENS might be overpredicting high rainfall totals.

The forecasts created with the multiple G-WTs shown within the grey box in **Figure 1c** will be called here Multiple-WT ecPoint. To answer the research question on the added value of the diagnosis of multiple G-WTs, an experimental ecPoint dataset was created using the MF in **Figure 1b**, which corresponds to a no-weather-scenario-based approach to process the raw ENS rainfall forecasts. This experimental dataset is referred to as Single-WT ecPoint.

## Observations: SYNOP and local high-density rain gauges

This study considered 12-hourly rainfall observations from two different resources stored internally at ECMWF: global surface synoptic observations (SYNOP) transmitted by the Global Telecommunication System, and high-density observations from local networks of rain gauges available, mainly for Europe (Haiden and Duffy 2016). Accumulation periods ending at 00, 06, 12 and 18 UTC, between the 1st of December 2021 and the 30th of November 2022, were considered. On average, there are 10,000 observations in each accumulation period. **Figure 3** shows a map with the location of the 12-hourly rainfall observations.

# Methods

Reliability and discrimination ability are desirable properties of ensemble forecasts (Wilks 2019). Both properties are defined against a rainfall threshold (e.g., 50 mm/12h), referred hereafter as a verifying rainfall threshold (VRT). Reliability measures whether the chosen VRT is predicted with a probability that equals the average frequency at which such an event is observed. Discrimination measures the ability of a forecasting system to distinguish situations that lead to events exceeding the VRT. While the property of reliability deals with the meaning of probabilities, the discrimination ability property appraises the existence of a signal in the forecasts when an event materializes (Ben Bouallègue and Richardson 2022). In order to say that post-processed forecast add value to their raw counterpart, both reliability and discrimination ability must be improved by the post-processing technique.

The performance of one-year retrospective Single-WT ecPoint, Multiple-WT ecPoint and ENS forecasts (i.e., 00 UTC runs from 01/12/2021 to 30/11/2022) are evaluated against global SYNOP rainfall observations and some local (mainly European), high-density rain gauge observations available internally at ECMWF (Haiden and Duffy 2016). It is worth noting that, although raw NWP model output does not pertain to point values, it is common to verify these gridded forecasts against point-rainfall observations (Haiden et al. 2023). Four VRTs were considered in the verification analysis: 0.2 mm/12h (i.e., “dry or not” condition), 10 mm/12h (“wet” condition), 25 mm/12h (“moderately severe rainfall, with some flash flood potential” condition), and 50 mm/12h (“severe rainfall, with flash flood potential” condition).

## Forecast reliability: reliability component of Brier Score (BSrel)

This study uses the reliability component of the Brier Score (BSrel) to measure the reliability of the three considered forecasting systems. BSrel is an integral measure of reliability across all issued probabilities (Ferro and Fricker 2012). Assume the occurrence of n events, and let x1, …, xn indicate whether the *i*th event occurs (i.e., xi = 1) or not (i.e., xi = 0). Suppose that each forecast can take one of the only K distinct values π1, …, πn. Let nk be the number of occasions on which πk is forecast. For those k for which nk > 0, define the conditional relative frequency to be the proportion of events that occur out of the nk occasions on which k is forecast:

|  |  |
| --- | --- |
|  | (1) |

BSrel is then defined as follows:

|  |  |
| --- | --- |
| BSrel = ( – ) | (2) |

Equation (2) shows that BSrel is a weighted average of the squared differences between the conditional relative frequencies and the corresponding forecasts. The reliability component of the Brier Score takes values in the interval [0,∞), with 0 being the best score, which is obtained when the conditional relative frequencies are equal to their corresponding forecasts.

## Forecast discrimination ability: Relative Operating Characteristic (ROC) Curves and Area Under the ROC curve (AROC)

ROC curves are built from a 2 × 2 contingency table that quantifies hits (H), misses (M), false alarms (FA), and correct negatives (CN) that occur when action is advised based on the VRT exceeding sampled probability thresholds (see **Table 1** for the definition of the constituting elements of the contingency table). Correspondent hit rates (HR) and false alarm rates (FAR) are computed, respectively, from equations (1) and (2):

|  |  |
| --- | --- |
| HR = H / (H+M) [values between 0 and 1] | (1) |
| FAR = FA / (FA+CN) [values between 0 and 1] | (2) |

HRs are mapped (on the Y-axis) against FARs (on the X-axis) within a unit square. The location of the ROC curve in the graph and the geometrical area under the ROC curve (AROC) determine the discrimination ability of the forecasting system. Perfect discrimination is obtained when only HRs grow and FARs remain equal to zero. This is represented by a ROC curve that rises from the bottom left corner (0,0) along the Y-axis to the top-left corner (0,1) and moves straight to the top-right corner (1,1). In this case, the AROC would be equal to 1. When the forecasting system has no discriminatory ability (i.e., it does not provide any additional information beyond climatological predictions), the HRs and FARs grow at the same rate. The ROC curve lies along the diagonal of the graph, and the AROC is equal to 0.5. How ROC curves are built and the AROC is computed has a significant impact on the interpretation of forecasts discrimination ability. This study compares forecasts discrimination ability estimated in two different ways, leading to the so called “real” and “potential” discrimination ability.

The “real” discrimination ability of a forecasting system measures the ability of a forecasting system to distinguish between events and non-events using the operational configuration of the forecasting system. It can be evaluated by building a ROC curve for specific incremental decision thresholds (Wilks 2019; Jolliffe and Stephenson 2011). To ensure that the ROC curves are as complete as possible, probability thresholds are determined using the full discretization available in the ensemble, rather than using fixed percentage bins (Ben Bouallègue and Richardson 2022). For an ensemble of size M, the maximum discretization is achieved by considering decision thresholds that correspond to the number of members exceeding the VRT. The ROC curve is therefore built by straight segments joining successive M+1 points (i.e., 0, 1/M, 2/M, …., M/M). The ROC curve for ENS is, therefore, built with 52 points, whereas the ROC curve for MultipleWT and Single-WT ecPoint is built with 100 points. The ROC curve is completed by joining the last meaningful point on the ROC curve and the top-right corner of the unit square with a straight line. The area under the ROC curve (AROC) is computed using the trapezoidal approximation. The nomenclature comes from the fact that the area is computed by simply adding the areas of single trapeziums formed by the straight lines between the ROC’s consecutive points. For rare events, the points of the ROC curve tend to cluster towards the lower left corner of the graph (Casati et al. 2008). Thus, completing the ROC curve with a straight line gives the impression that part of the curve is missing, and AROCs might be small. How much of the curve appears incomplete depends on the ensemble size and the base rate of the event. The AROC computed with the trapezoidal approximation is denoted as AROCT.

The ”potential” discrimination ability of a forecasting system measures the ability of a forecasting system to distinguish between events and non-events considering a hypothetical configuration of the forecasting system (e.g., what would it be the discrimination ability if a had an infinite number of ensemble members?). This approach aims at drawing a full ROC curve by fitting the real ROC curve, and for rare events, it consists effectively in an extrapolation to a hypothetical continuous decision variable based on the limited set of decision thresholds that are materially assessable from the real ensemble configuration. Since such decision thresholds may not be achievable in practice, the fitted ROC curves are considered as a measure of the potential discrimination that could be achieved with an unlimited ensemble size (Ben Bouallègue and Richardson 2022). Different fitting models are available in the literature (Harvey et al. 1992; Gneiting and Vogel 2022). Given its popularity, this study adopts the fitting of the ROC curve that uses the binormal model (Harvey et al. 1992). The binormal model assumes that HRs and FARs are integrations of normal distributions. Harvey et al. (1992) have also provided a closed-form for the AROC computation, denoted as AROCZ.

# Results

## Forecast reliability

Multiple-WT ecPoint shows the best forecast reliability for point verification across all lead times and all VRTs compared to ENS and Single-WT ecPoint. This can be seen by the relative distance between the orange line for multiple-WT ecPoint (lying at the bottom of the diagrams) and the green line for ENS and the grey line for the single-WT ecPoint (lying both above the orange one). For VRT = 0.2 mm/12h, the reliability for the single-WT ecPoint is similar to the reliability of the multiple-WT ecPoint. It worsens for the other VRTs, but it is still better than the reliability of ENS forecasts. For extreme rainfall (i.e., rainfall events exceeding the VRT = 50 mm/12h), the single-WT ecPoint shows the worst reliability out of the three forecasting systems.

In relative terms, the largest impact in the improvement of forecast reliability between raw ENS and post-processed forecasts (i.e., single- and multiple-WT ecPoint) can be seen for VRT = 0.2 mm/12h (the green line for ENS is far apart from the grey and orange lines, respectively for single- and multiple-WT ecPoint). For all three systems, BSrel decreases with increasing lead times. Out of the three systems, ENS and single-WT ecPoint show the biggest BSrel decrease, especially at shorter lead times, while multiple-WT ecPoint shows the most horizontal trend meaning that the reliability tends to not change significantly with lead time.

Out of all three systems, ENS shows the most sinusoidal pattern with the lead times at all VRTs (although with increasing noise for VRT = 25 and 50 mm/12h). This shows that reliability is different at different times of the day in the ENS forecasts. The post-processed forecasts improve on this, with the multiple-WT ecPoint showing the least sinusoidal pattern in all VRTs.

The uncertainty in the estimation of forecast reliability is greater for ENS and smaller for the single- and multiple-WT ecPoint. It is worth noting that multiple-WT ecPoint shows peaks of uncertainty in the reliability for VRT = 50 mm/12h in steps that mark the end of the valid period ending at 12 UTC. The differences between the raw and the post-processed forecasts are significantly different at all lead times with a confidence level of 99% for VRTs up to 25 mm/12h. The differences between the two post-processed forecasts are significant for the two smallest VRTs (i.e., 0,2 and 10 mm/12h), while for the two highest VRTs, the differences are significant only up to t+ 96 (i.e., day 4 for VRT = 25 mm/12h) and t+174 (i.e., day 7 for VRT = 50 mm/12h).

## Forecast discrimination ability

For point verification, multiple-WT and single-WT ecPoint show better AROCt values (orange and grey continuous lines in Figure 5, respectively) than ENS (green continuous line in Figure 5) across all lead times and VRTs. For VRT = 0.2 mm/12h (**Figure 5a**), the AROCt lines for all three forecasting systems almost overlap. As VRTs increase (**Figure 5b and c**), the distance between the AROCt lines for multiple-WT and single-WT ecPoint and the AROCt line for ENS increases, while the distance between the AROCt lines for multiple-WT and single-WT ecPoint remains not significantly different at the 99% confidence level.

The AROCz lines for all three forecasting systems (orange, grey and green dashed lines in **Figure 5**) always lie above the AROCt linesacross all lead times and all VRTs. The differences between the AROCz values for the three forecasting systems are not significantly different at the 99% confidence level. The uncertainty in the estimates of the AROCz values is shown to be similar for the three forecasting systems in all VRTs, except for the VRT = 50 mm/12h, where the uncertainty around the estimates of AROCz values for ENS is much bigger than those for the multiple-WT and single-WT ecPoint.

One aspect to highlight corresponds to the AROCt line for the multiple-WT ecPoint that remains above the AROCt line for the single-WT ecPoint for all VRTs, except for the VRT = 50 mm/12h. However, the relative position of the correspondent AROCz lines for VRT = 50 mm/12h is swapped. **Figure 6** shows the real and binormal ROC curves for the three forecasting systems and for VRT = 50 mm/12h. The example ROC curves are taken for the accumulation period ending at t+126 (i.e., day 5 forecast). **Figure 6** shows that the greater AROCt for the single-WT ecPoint compared to the multiple-WT ecPoint is due to the fact that the last meaningful point of the real ROC curve for single-WT ecPoint (point a) is located higher and to the right of the last meaningful point of the real ROC curve for multiple-WT ecPoint (point b). Although this make the AROCt for single-WT ecPoint (=0.82) bigger than for multiple-WT ecPoint (=0.79), this is also indicative of a higher rate of detection from Single-WT ecPoint than multiple-WT ecPoint, but at the cost of a higher rate of false alarms. The binormal approximation of the continuation of the ROC curves (grey and orange dashed lines for, respectively, Single-WT and Multiple-WT ecPoint) shows that because of the position of the last meaningful points on the real ROC curves (a and b for single-WT and multiple-WT ecPoint, respectively) the binormal ROC for single-WT ecPoint lies to the right of the binormal ROC for Multiple-WT ecPoint making the AROCz for single-WT ecPoint (=0.938) smaller than the AROCz for multiple-WT ecPoint (=0.946).

# 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 and 12 UTC, 465.8 mm were observed (**Figure 7**, 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 7** (top left panel) shows some of the flood impacts. Videos emerged, showing cars floating in streets and Zhengzhou Metro passenger’s 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 7** 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 significantly overestimate 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 7** 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.

# Discussions

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).

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.

# Conclusions

# Tables

**Table 1** - Definition of the four quadrants in a contingency table.

|  |  |  |
| --- | --- | --- |
| **FORECASTS (COLUMNS) /**  **OBSERVATIONS (ROWS)** | **YES** | **NO** |
| **YES** | QUADRANT I  Hits (H)  The event *was observed* when it *was predicted*. | QUADRANT II  False Alarms (FA)  The event *was not observed* when it *was predicted*. |
| **NO** | QUADRANT III  Misses (M)  The event *was observed* when it *was not predicted*. | QUADRANT IV  Correct Negatives (CN)  The event *was not observed* when it *was not predicted*. |

# 

# Figures

Immagine che contiene testo, schermata, diagramma, Carattere

Descrizione generata automaticamente

**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 weather map

Description automatically generated

**Figure 2** – Example of ENS (first column) and ecPoint (second column) rainfall forecasts. The panel compares the 50th percentile (first row) and the 99th percentiles (second row).

Immagine che contiene testo, diagramma, mappa

Descrizione generata automaticamente

**Figure 3** – Example of rain gauge locations for observations starting at 0, 6, 12, and 18 UTC.

A group of graphs showing different types of data

Description automatically generated

**Figure 4** - Reliability component of the Brier Score (BSrel) for lead times up to t+246 (i.e., day 10) and for verifying rainfall thresholds (VRT) >= 0.2 (a), 10 (b), 25 (c), and 50 (d) mm/12h. The turquoise, orange and grey lines represent BSrel values for ENS, ecPoint\_MultipleWT and ecPoint\_SingleWT, respectively. The shaded areas represent the correspondent confidence intervals at 99% confidence level.

A screenshot of a graph

Description automatically generated

**Figure 5** – Trapezoidal (continuous lines, AROCt) and Binormal (dashed lines, AROCz) areas under the ROC curve for lead times up to t+246 (i.e., day 10), and for verifying rainfall thresholds (VRT) >= 0.2 (a), 10 (b), 25 (c), and 50 (d) 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.

Immagine che contiene testo, linea, Diagramma, diagramma

Descrizione generata automaticamente

**Figure 6** – Real (continuous lines) and binormal (dashed lines) ROC curves for ENS (green lines), Multiple-WT ecPoint (in orange) and Single-WT ecPoint (in grey). The ROC curves are built for the 12-hourly accumulation period ending at t+126 (i.e. day 5 forecast), and for verifying rainfall thresholds (VRT) >= 50 (d) mm/12h.

Immagine che contiene testo, mappa, fiore, schermata

Descrizione generata automaticamente

**Figure 7** - 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.

**Acknowledgments.**

**Data and software availability.** The data is available under request to the correspondent author. The software is available in the following GitHub repository: *https://github.com/FatimaPillosu/Verif\_ecPoint\_SingleWT* .

**Author contributions.** FMP contributed to the design and the implementation of the research, and to the analysis of the results. HLC and CP supervised the project and helped built the manuscript structure. All authors contributed to the discussion of the results and the writing of the manuscript.

**Conflict of interest.** We declare that there are no competing interests.

**Funding statement.** The study was supported by the Copernicus Emergency Management Service.

# References

Bauer, P., P. D. Dueben, T. Hoefler, T. Quintino, T. C. Schulthess, and N. P. Wedi, 2021: The digital revolution of Earth-system science. *Nat Comput Sci*, **1**, 104–113, https://doi.org/10.1038/s43588-021-00023-0.

Bechtold, P., N. Semane, P. Lopez, J. P. Chaboureau, A. Beljaars, and N. Bormann, 2014: Representing equilibrium and nonequilibrium convection in large-scale models. *J Atmos Sci*, **71**, 734–753, https://doi.org/10.1175/JAS-D-13-0163.1.

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

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

Buizza, R., 2019: Introduction to the special issue on “25 years of ensemble forecasting.” *Quarterly Journal of the Royal Meteorological Society*, **145**, 1–11, https://doi.org/10.1002/qj.3370.

Cafaro, C., and Coauthors, 2021: Do convection-permitting ensembles lead to more skillful short-range probabilistic rainfall forecasts over tropical east africa? *Weather Forecast*, **36**, 697–716, https://doi.org/10.1175/WAF-D-20-0172.1.

Casaretto, G., M. E. Dillon, P. Salio, Y. G. Skaba, S. W. Nesbitt, R. S. Schumacher, C. M. García, and C. Catalini, 2022: High-Resolution NWP Forecast Precipitation Comparison over Complex Terrain of the Sierras de Córdoba during RELAMPAGO-CACTI. *Weather Forecast*, **37**, 241–266, https://doi.org/10.1175/WAF-D-21-0006.1.

Casati, B., and Coauthors, 2008: Forecast verification: Current status and future directions. *Meteorological Applications*, Vol. 15 of, 3–18.

Ferro, C. A. T., and T. E. Fricker, 2012: A bias-corrected decomposition of the Brier score. *Quarterly Journal of the Royal Meteorological Society*, **138**, 1954–1960, https://doi.org/10.1002/QJ.1924.

Gascón, E., A. Montani, and T. D. Hewson, 2023: Post-processing output from ensembles with and without parametrised convection, to create accurate, blended, high-fidelity rainfall forecasts. https://doi.org/10.48550/arxiv.2301.04485.

Gneiting, T., and P. Vogel, 2022: Receiver operating characteristic (ROC) curves: equivalences, beta model, and minimum distance estimation. *Mach Learn*, **111**, 2147–2159, https://doi.org/10.1007/S10994-021-06115-2/FIGURES/1.

Göber, M., E. Zsótér, and D. S. Richardson, 2008: Could a perfect model ever satisfy a naïve forecaster? On grid box mean versus point verification. *Meteorological Applications*, **15**, 359–365, https://doi.org/10.1002/met.78.

Haiden, T., and S. Duffy, 2016: Use of high-density observations in precipitation verification. *ECMWF Newsletter*, 20–25, https://doi.org/10.21957/hsacrdem.

——, M. Janousek, F. Vitart, Z. Ben-Bouallegue, and F. Prates, 2023: Evaluation of ECMWF forecasts, including the 2023 upgrade. *ECMWF Technical Memoranda*, **911**, 1–60.

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

Hewson, T., F. Pillosu, E. Gascòn, and M. Vučković, 2023: Post-processing ERA5 output with ecPoint. *ECMWF Newsletter*,.

Hewson, T. D., and F. M. Pillosu, 2021: A low-cost post-processing technique improves weather forecasts around the world. *Commun Earth Environ*, **2**, https://doi.org/10.1038/s43247-021-00185-9.

Janjić, T., and Coauthors, 2018: On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **144**, 1257–1278, https://doi.org/10.1002/qj.3130.

Jolliffe, I. T., and D. B. Stephenson, 2011: *Forecast Verification: A Practitioner’s Guide in Atmospheric Science*. 2nd Editio. John Wiley & Sons, Ltd,.

Lavers, D. A., S. Harrigan, and C. Prudhomme, 2021: Precipitation Biases in the ECMWF Integrated Forecasting System. *J Hydrometeorol*, https://doi.org/10.1175/jhm-d-20-0308.1.

Owens, R. G., and T. Hewson, 2018: *ECMWF Forecast User Guide*.

Roberts, N., 2008: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorological Applications*, Vol. 15 of, 163–169.

——, and Coauthors, 2023: IMPROVER: The New Probabilistic Postprocessing System at the Met Office. *Bull Am Meteorol Soc*, **104**, E680–E697, https://doi.org/10.1175/BAMS-D-21-0273.1.

Vannitsem, S., and Coauthors, 2021: Statistical postprocessing for weather forecasts review, challenges, and avenues in a big data world. *Bull Am Meteorol Soc*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

Wilks, D. S., 2019: *Statistical Methods in Atmoshperic Sciences*. Fourth Edi. Elsevier Inc.,.

Zeman, C., N. P. Wedi, P. D. Dueben, N. Ban, and C. Schär, 2021: Model intercomparison of COSMO 5.0 and IFS 45r1 at kilometer-scale grid spacing. *Geosci Model Dev*, **14**, 4617–4639, https://doi.org/10.5194/gmd-14-4617-2021.

1. https://confluence.ecmwf.int/display/FCST/Implementation+of+IFS+Cycle+47r3 [↑](#footnote-ref-2)