**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.** Statistical post-processing techniques allow to correct bias and anticipate representativeness errors in raw NWP model outputs. Due to their low calibration and production costs, no-weather-scenario-based approaches are widely used. However, this study argues that the use of a post-processing approach that differentiates its corrections according to different weather scenarios should provide better post-processed forecasts. This study investigates whether is worth bearing the cost of diagnosing multiple grid-box weather types to post-process NWP model outputs, or whether both approaches achieve similar improvements by calculating and comparing the reliability and the discrimination ability of forecasts from three systems: raw ENS and two post-processed forecasts based on ECMWF’s ecPoint technique, i.e., the currently operational multiple grid-box weather type approach (multiple-WT ecPoint) and an experimental system that does not diagnose weather types (single-WT ecPoint). Reliability is assessed using the reliability component of the Brier score. The “real” and “potential” discrimination ability is assessed by computing, respectively, the real and the binormal ROC curves and computing the correspondent areas under the ROC curve. This study focuses on rainfall. Results show that overall, multiple-WT ecPoint has better reliability and discrimination ability than single-WT ecPoint and ENS. While the area under the real ROC curve is bigger for single-WT than multiple-WT ecPoint for very extreme events, the area under the binormal ROC curve is smaller because the higher rate of detection in single-WT ecPoint comes with a higher rate of false alarms than in multiple-WT ecPoint. Moreover, the reliability for very extreme events is worse in single-WT ecPoint compared to multiple-WT and ENS, meaning that single-WT tends to overpredict the probabilities of occurrence of extreme events. This is also confirmed by the analysis of a case of severe rainfall and flooding in China. It is also shown that single-WT ecPoint reduces the chance of not having rain compared to multiple-WT. It is therefore concluded that it is worth bearing the cost adopting a weather-scenario-based post-processing approach as it would not reduce the rate of event detection but it would reduce the false alarms rates of the no-weather-scenario-based counterpart.

**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 the reliability and discrimination ability for forecasts from the raw ECMWF ensemble (ENS), the original multiple G-WT ecPoint (Multiple-WT), and an experimental single G-WT ecPoint (Single-WT) representing a no-weather-scenario-based post-processing approach.

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 to as 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 forecasts' ability 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 forecasts when an event materialises (Ben Bouallègue and Richardson 2022). Post-processing adds value to raw forecasts if both, reliability and discrimination ability, are improved.

The performance of one-year retrospective Single-WT ecPoint, Multiple-WT ecPoint and ENS forecasts (00 UTC runs, from 01/12/2021 to 30/11/2022) is evaluated against rain gauge observations. Although raw NWP model output does not pertain to point values, it is common practice to verify gridded forecasts against point-rainfall observations to assess forecasts' performance for site-specific predictions (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).

## Reliability: Reliability Component of the Brier Score (BSrel)

This study considers 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). Let us 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.

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

Relative Operating Characteristic (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). 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 (Y-axis) against FARs (X-axis) in a unit square. The location of the ROC curve in the graph and the geometrical area under the ROC curve (AROC) measure forecasts’ discrimination ability. Perfect discrimination is obtained when only HRs grow and FARs remain zero. This is represented by a ROC curve that rises along the Y-axis from the bottom left corner of the unit square to the top-left corner and moves straight to the top-right corner. In this case, the AROC is equal to 1. If HRs and FARs grow at the same rate, the forecasting system has no discriminatory ability (i.e., it does not provide additional information beyond climatological predictions). In this case, the ROC curve lies along the graph’s diagonal, and the AROC equals 0.5.

How ROC curves and AROCs are computed significantly impacts the interpretation of forecasts’ discrimination ability. ROC curves built for incremental decision thresholds materially assessable from the real ensemble configuration estimate the “real” forecasts’ discrimination ability (Wilks 2019). Probability thresholds are determined considering the full discretization available in the ensemble to ensure ROC curves are as complete as possible (Ben Bouallègue and Richardson 2022). These thresholds correspond to the number of members exceeding the VRT, so that for an ensemble of size M, maximum discretization is achieved by M+1 probability thresholds (i.e., 0, 1/M, 2/M, …., M/M). The ROC curve is built by straight segments joining successive points. It is then completed by joining the last meaningful point on the ROC curve with the top-right corner of the unit square with a straight line. The area under the ROC curve is computed using the trapezoidal approximation (AROCT), i.e. by adding the areas of the single trapeziums formed by the straight lines between ROC’s consecutive points. For rare events, the points of a ROC curve cluster in the graph’s bottom left corner, and completing the ROC curve with a straight line might give the impression that part of the curve is missing (Casati et al. 2008). How much of the curve appears incomplete depends on the ensemble size and the base rate of the event. ROC curves can also be built by fitting real ROC curves. For rare events, this method effectively consists of an extrapolation to a hypothetical continuous decision variable based on the limited set of probability thresholds materially assessable from the real ensemble configuration. Since such a configuration may not be achievable in practice, fitted ROC curves are considered to measure the “potential” discrimination that could be achieved with an unlimited ensemble size (Ben Bouallègue and Richardson 2022). Many fitting models are available in the literature (Harvey et al. 1992; Gneiting and Vogel 2022). This study employed the well-established binormal model, which assumes that HRs and FARs are integrations of normal distributions (Harvey et al. 1992). Harvey et al. (1992) also provided a closed form for AROC computation (AROCZ).

# Results

## Reliability

Across all lead times and all VRTs, multiple-WT ecPoint shows the highest forecast reliability for point verification. This can be seen from the orange line (Multiple-WT ecPoint) in **Figure 4** lying below the green (ENS) and grey (Single-WT ecPoint) lines. For VRT = 0.2 mm/12h, the reliability for single-WT ecPoint is similar to that for multiple-WT ecPoint. It gradually worsens with increasing VRTs until it shows the worst reliability out of the three forecasting systems for VRT = 50 mm/12h.

The reliability differences between the three forecasting systems and their statistical significance (at 99% confidence level) diminish with increasing VRTs. The uncertainty in the forecast reliability estimates also increases with increasing VRTs, but it is more prominent for ENS and single-WT ecPoint. It is also worth noting that, in VRT = 50 mm/12h, multiple-WT ecPoint shows uncertainty peaks in steps corresponding to accumulation periods ending at 12 UTC.

Forecasts' reliability as a function of lead time displays a sinusoidal pattern, especially for ENS and VRT = 0.2 mm/12h. The sinusoidal pattern indicates that reliability worsens for specific accumulation periods (ending at 12 and 18 UTC). Both post-processed forecasts show a much more reduced sinusoidal pattern, with multiple-WT ecPoint showing the most linear trend (although with increasing noise for VRT = 25 and 50 mm/12h). Multiple-WT ecPoint also exhibits the most horizontal trend out of the three systems, meaning that reliability does not change significantly with lead time.

## Discrimination ability

Across all lead times and VRTs, multiple-WT (orange continuous lines in **Figure 5**) and single-WT ecPoint (grey continuous lines) show larger AROCt values than ENS (green continuous lines). For VRT = 0.2 mm/12h (Figure 5a), the AROCt lines for all three forecasting systems overlap. As VRTs increase (**Figure 5b, c, d**), the difference between AROCt values for both post-processed forecasts and ENS increases and remains significant at the 99% confidence level. On the other hand, the difference between multiple-WT and single-WT ecPoint is much smaller and not significant.

For all lead times and VRTs, AROCz values for all three forecasting systems (dashed lines in **Figure 5**) are larger than AROCt. Their differences also appear to be small and not significant at the 99% confidence level. For VRT = 0.2 and 10 mm/12h, the uncertainty in the AROCz estimates is similar to that for AROCt but increases significantly for larger VRTs. In particular, the uncertainty in the estimates of AROCz for ENS for VRT = 50 mm/12h is the largest observed.

The AROCt line for the multiple-WT ecPoint remains above to that for the single-WT ecPoint for all VRTs, except for VRT = 50 mm/12h (respectively, orange and grey continuous lines in **Figure 5d**). However, the relative position of the correspondent AROCz lines is swapped (orange and grey dashed lines). **Figure 6** shows real (continuous lines) and binormal (dashed lines) ROC curves for the three forecasting systems, for VRT = 50 mm/12h and accumulation period ending at t+126 (i.e., day 5 forecast). **Figure 6** shows that the last meaningful point in the real ROC curve for single-WT ecPoint (point A) is located higher up and to the right of the last meaningful point of the real ROC curve for multiple-WT ecPoint (point B). Although this makes AROCt for single-WT ecPoint (=0.82) bigger than that for multiple-WT ecPoint (=0.79), this also shows that the higher rate of detection from single-WT ecPoint comes at the cost of a higher rate of false alarms. Due to the position of A and B, the binormal approximation of the continuation of the ROC curve for single-WT ecPoint (grey dashed line in **Figure 6**) lies to the right of the binormal ROC for Multiple-WT ecPoint (orange dashed line) so that AROCz for single-WT ecPoint (=0.938) results smaller than the AROCz for multiple-WT ecPoint (=0.946).

# Case study: extreme rainfall and flash floods in China in July 2021

In July 2021, the Henan Province in northeast China experienced extremely severe rainfall. Over three days, between July 17 and 20, 617.1 mm of rain were recorded in the province's capital, Zhengzhou, nearing the year's average precipitation. The most intense rainfall was observed on 20 July, when 201.9 millimetres of rain were recorded between 4 and 5 pm local time (the highest figure ever recorded since measurements began in 1951). The extreme rainfall generated severe, extensive flooding (**Figure 7a**), causing the evacuation of 815,000 people and affecting 14.5 million people around the province. The death toll reached 398 deaths.

**Figure 7b** shows rain gauge observations for 20 July between 00 and 12 UTC, where 465.8 mm of rain were observed in Zhengzhou. **Figure 7c** compares the 12-hourly rainfall forecasts for ENS (first row), multiple-WT ecPoint (second row), and single-WT ecPoint (third row), valid for the observations’ accumulation period. The first three columns in **Figure 7c** show the 99th percentile for day 5, 3, and 1 forecasts (from left to right). All forecasting systems, up to five days in advance, provided good guidance on which area was at higher risk of experiencing extreme rainfall, namely the area near Zhengzhou (highlighted by the small black circles). Closer to the event (i.e., day 1 forecasts), single-WT ecPoint predicted more than 700 mm/12h (zoomed in circle in the third row of **Figure 7c**), significantly overestimating the observed rainfall totals. Instead, ENS significantly underestimated the observed rainfall totals, predicting totals not higher than 150 mm/12h (zoomed in circle in the first row of **Figure 7c**). On the contrary, multiple-WT ecPoint predicted rainfall totals of the same order of magnitude as those observed, i.e., ~ 400 mm/12h (zoomed in circle in the second row of **Figure 7c**). The fourth column in **Figure 7c** shows the probability of having less than 0.2 mm/12h (i.e., having no rain) on day one forecasts for the area southwest of Zhengzhou (blue circles), where no rainfall was observed (blue circle in **Figure 7b**). ENS shows zero probability of having no rain. Single-WT ecPoint shows much smaller probabilities than multiple-WT ecPoint of having no rain, between 20 and 25% instead of 50 to 80%.

# Discussions

This study investigates whether is worth bearing the cost of defining the G-WTs that constitute the embodiment of the multi-WT ecPoint post-processing, or whether similar improvements to those shown in Hewson and Pillosu (2021) could be achieved by a post-processing system that does not differentiate between weather scenarios (represented by the single-WT ecPoint post-processing system). Calibration and forecast production costs would be indeed smaller for single-WT ecPoint than for multiple-WT ecPoint. To analyse the added value of both post-processed forecasts compared to the raw ENS, this study compares the reliability and discrimination ability for the forecasts from all the three systems. The reliability component of the Brier score is used to quantify forecasts’ reliability, and ROC curves and AROC are used to quantify forecasts’ discrimination ability for point verification. Although raw NWP model output does not pertain to point values, it is common practice to verify gridded forecasts against point-rainfall observations (Haiden et al. 2023).

* The more significant improvement in reliability at short lead times by the post-processed rainfall forecasts is consistent with results in previous studies, especially at short lead times (Saetra et al., 2004; Candille et al., 2008, Bouallegue et al. ,2020). Since the ensemble spread is limited at the beginning of the forecast range, the scale mismatch between model and observations plays a substantial role. This is less the case at longer lead times when the ensemble spread is larger.
* 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.
* Previous studies also found that much more significant improvements are obtained for small VRTs and persist at longer lead times. Moreover, in absolute terms, the reliability for ENS exhibits a decrease and a subsequent increase (i.e., a decrease in skill). This is not observed for the post-processed forecasts that exhibit a continuous decrease not showing a particular deterioration of the forecast reliability for increasing lead times.
* The differences in terms of reliability and discrimination ability as measured, respectively by BSrel and AROC can be explained by the large improvement of the post-processed when accounting for the expected sub-grid variability. The ensemble spread is typically (but not always) increased by the post-processing. AS a consequence, the post-processed forecasts are able to capture better the variability of point observations.
* Ben (2020) found that the lack of reliability for high rainfall thresholds, even after addressing the representativeness provides evidence that there are remaining systems problems that likely need to be addressed through prediction system improvements and/or post-processing. Here we argue that the needed improvements are applied by the multiple-WT ecPoint by differentiating the corrections to apply to the raw forecasts based on different weather scenarios.
* Regarding ROC curves, our results are comparable to those obtained in previous results. For small VRTs, the impact of the post-processing is minimal as the AROC lines are overlapping. This means that the information content of the post-processed forecasts is not modified for this type of event. However, when focusing on larger VRTs, the ROC curves clearly differ from each other. So the post-processing technique produces a shift in probability distribution that appears beneficial for users with small probability thresholds. The ability of the post-processed forecasts to capture the tail of the observations is rewarded in terms of forecasts discrimination ability.
* This analysis provides evidence that model-independent post-processing methods, that are based on independent observations only and so can be applied to forecasts from any model, simply adapting the parameters as a function of the model grid spacing, can improve forecasts’ reliability and discrimination ability. This is consistent to what has been found in previous studies Ben Boulange (2020). However, these type of approach can be seen as a way to account for model limitations linked to sub-grid scale uncertainty, but it cannot correct for model deficiencies. A model that accounts for weather-scenarios-dependent corrections, not only can applied better corrections linked to sub-gird scale uncertainty, but it can also correct for model biases which are typically linked to different weather scenarios as also suggested by Ben Boulange (2020).

# Conclusions

This study investigates whether is worth bearing the cost of implementing weather-scenario-based post-processing approaches or whether similar results can be achieved with less costly and more simple post-processing approaches that do not take into account weather-dependent corrections. The results suggest that accounting for weather scenarios when correcting for biases and anticipating for sub-grid variability in raw forecasts is crucial to improve forecast performance in terms of both, reliability and discrimination ability. This is especially true for high-impact events. Weather-dependant post-processing approaches not only are able to increase event’s detection rate, but it is also able to decrease the rate of false alarms.

# 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 map

Description automatically generated

**Figure 2** – Panel (a) displays the location of the rain gauges used in the objective verification. Panels (b), (c), and (d) display an example of forecast probabilities that can be obtained, respectively, from ENS, Multiple-WT ecPoint, and Single-WT ecPoint. The examples are shown for a day 2 forecast, issued on 9th December 2021 at 00 UTC, and valid between 10th December at 12 UTC and 11th December at 00 UTC. The probabilities shown are for rainfall exceeding 10 mm/12h.

A group of graphs showing different types of data

Description automatically generated

**Figure 3** – Panels (a), (b), and (c) display the reliability component of the Brier Score (BSrel) for VRT >= 0.2, 10, and 50 mm/12h, respectively, up to t+246 (i.e., day 10 forecast). Panels (d), (e), and (f) display the trapezoidal (continuous lines, AROCt) and Binormal (dashed lines, AROCz) areas under the ROC curve for VRT >= 0.2, 10, and 50 mm/12h, respectively, up to t+246 (i.e., day 10 forecast). The turquoise, orange and grey lines represent BSrel, AROCt, and AROCz values for ENS, ecPoint\_MultipleWT and ecPoint\_SingleWT, respectively. The shaded areas represent the correspondent confidence intervals at 99% confidence level.

A collage of graphs

Description automatically generated

**Figure 6** – Panels (a), (b), and (c) show the ROC curves the accumulatio period ending at t+24 (day 1 forecast) and for VRT >=0.2, 10, and 50 mm/12h, respectively. Real ROC curves are represented with continuous lines while binormal ROC curves are represented with dashed lines. Panels (d), (e), and (f) show the reliability diagrams for the same accumulation period and VRTs. The inserts represent the reliability diagrams for small forecast probabilities (at the bottom-left corner of the diagrams). Panels (g), (h), and (i) display the sharpness diagrams for the same accumulatio period and VRTs. The turquoise, orange and grey lines represent ENS, ecPoint\_MultipleWT and ecPoint\_SingleWT, respectively.

Immagine che contiene testo, mappa, fiore, schermata

Descrizione generata automaticamente

**Figure 7** - Flash floods in Zhengzhou (Henah, China) on the 20th of July 2021. Panel (a) shows images of the impacts of the flash floods in Zhengzhou (credits to China Dialogue and CNN for top and bottom image, respectively). Panel (b) shows 12-hourly rainfall observations valid for the 20th of July 2021 between 0 and 12 UTC. Panel (c) shows 12-hourly rainfall forecasts for ENS (first row), multiple-WT ecPoint (second row), and single-WT ecPoint (third row) valid for the observations’ accumulation period. 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.

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.

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.

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.

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.

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)