ECMWF TECHNICAL MEMORANDUM

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

Fatima M. Pillosu[1,2], Tim Hewson[2], Christel Prudhomme[2,3,4], Elisabeth Stephens[1,5,6], Hannah L. Cloke[1,5,7,8]

[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. This study compares 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). This study focuses on rainfall. Multiple-WT ecPoint is shown to have better reliability and discrimination ability than Single-WT ecPoint and ENS. This is also confirmed by the analysis of a case of severe rainfall and flooding in China. It is concluded that it is worth bearing the cost adopting a weather-scenario-based post-processing approach as, contrary to Single-WT ecPoint, Multiple-WT ecPoint would keep high events' detection rates while keeping low false alarms rates. This would contribute to enhance user's trust in the forecasts.

Plain language summary.

Word count. 5185 words, excluding abstract, tables, captions, and references.

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

---

1 Introduction

Statistical weather-scenario-based post-processing approaches remains relatively uncommon in operational environments (Roberts et al. 2023). Despite the argument that these would provide better results in improving raw numerical weather prediction (NWP) outputs (Hewson and Pillosu 2021), no-weather-scenarios-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 post-processed forecasts with or without weather-scenario-based post-processing approaches.

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). While ECMWF ensemble (ENS) forecasts are currently provided with 50 and 100 ensemble members[1], they might still not be able to reproduce all possible weather scenarios in convective situations. Km-scale models display realistic-looking spatial patterns

[1] https://confluence.ecmwf.int/display/FCST/Implementation+of+IFS+Cycle+48r1

(Roberts 2008) and improve forecast accuracy by better representing complex features such as orography and convection (Casaretto et al. 2022). However, due to computational costs, their geographical coverage is limited, and lead times rarely exceed day 3. Although research in developing probabilistic 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 weather scenarios at grid-box scale defined from NWP grid-box forecast output can anticipate the degree of sub-grid variability and estimate biases at grid scale for the considered variable. It has been shown that ecPoint improves the reliability and discrimination ability of raw rainfall and temperature forecasts against point verification across different lead-time ranges (Hewson and Pillosu 2021; Gascón et al. 2023; Hewson et al. 2023; Hemri et al. 2022). This study poses the following research question: is it worth bearing the cost of adopting weather-scenario-based post-processing approaches, or would similar improvements be achievable with no-weather-scenario-based approaches? To answer this question, the reliability and discrimination ability of forecasts from three forecasting systems are compared: the raw ENS, the original weather-scenario-based ecPoint, and an experimental no-weather-scenario-based ecPoint.

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 and draws the study's final remarks.

2 Data

2.1 Forecasts: ENS and ecPoint

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[2] cycle.

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 gridded rainfall forecasts and point rainfall observations (from rain gauges) are computed for the calibration period. For accumulated variables such as rainfall, the error computed is the Forecast Error Ratio (FER) whose formulation is shown in **Figure 1a.** The errors' distribution is called mapping function (MF) and represents the heart of the post-processing method. One could post-process the raw rainfall forecasts using the MF for all the data points in the calibration period (**Figure 1b**). However, the MF shapes, linked to the expected degree of sub-grid variability and biases at grid scale, can change significantly according to different weather scenarios at grid-scale (called grid-box weather types, G-WT). The variety of significantly different MFs obtained from defining a number of G-WTs can be visualised with a decision-tree-like representation (**Figure 1c).** 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 might take a Gaussian-like form, meaning that the raw model output is representative of the point-scale rainfall totals and that the expected degree of sub-grid variability is bigger than in case A. Post-processed forecasts created using the MF in **Figure 1b** will be called Single-WT ecPoint, because only one WT was used, the same for all data points. Post-processed forecasts created using the multiple G-WTs shown in **Figure 1c** will be called Multiple-WT ecPoint. In the Multiple-WT ecPoint, the corrections applied to raw forecasts differ from grid-box to grid-box, so that the methodology can judge where to increase or decrease point rainfall values according to the different G-WTs. This advantage would be lost in the Single-WT ecPoint because all grid-boxes would be post-processed in the same way. Generally, Multiple-WT ecPoint tends to increase the number of zeros 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 to correct for ENS's underestimation of high rainfall values (Haiden et al. 2023). Moreover, Hewson and Pillosu (2021) have shown that ecPoint can also change the location of areas at risk of extreme localised rainfall if the G-WT suggests that ENS might be overpredicting high rainfall totals. **Figure 2a-c** show how different can be the forecast probabilities (of exceeding 10 mm/12h) in the three forecasting systems. The probabilities for ENS (**Figure 2a**) are significantly reduced by Multiple-WT

[2] https://confluence.ecmwf.int/display/FCST/Implementation+of+IFS+Cycle+47r3

ecPoint (**Figure 2b**) and Single-WT ecPoint (**Figure 2c**). However, Multiple-WT ecPoint provides bigger reductions than Single-WT ecPoint at some locations, showing that corrections are not applied indifferently everywhere as done by the Single-WT ecPoint.

2.2 Observations: SYNOP and local high-density rain gauges

This study used 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 2d** shows a map with the location of the 12-hourly rainfall observations.

3 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. This study considers two types of scores to analyse reliability and discrimination ability. Summary scores are used to provide an overview of how reliability and discrimination ability compare between the three analysed forecasting systems. Subsequently, other scores will be used to provide a breakdown of the constituent parts of the summary scores. Both types of scores are described in section 3.1 for reliability and in section 3.2 for discrimination ability. The workflow is described in **Figure 3**.

The reliability and discrimination ability of one-year retrospective ENS, Multiple-WT ecPoint, and Single-WT ecPoint forecasts (from 00 UTC run, between 01/12/2021 to 30/11/2022, and up to day 10 lead time) are 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). Three VRTs were considered in the verification analysis: 0.2 mm/12h (i.e., “dry or not” condition), 10 mm/12h (“wet” condition), and 50 mm/12h (“severe rainfall, with flash flood potential”).

3.1 Forecast reliability: reliability component of the Brier Score (summary score) and reliability/sharpness diagrams (breakdown score)

The reliability component of the Brier score (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 $x_1, \ldots, x_n$ indicate whether the ith event occurs (i.e., $x_i = 1$) or not (i.e., $x_i = 0$). Suppose that each forecast can take one of the only K distinct values $\pi_1, \ldots, \pi_n$. Let n_k be the number of occasions on which π_k is forecast. For those k for which $n_k > 0$, define the conditional relative frequency $\overline{x_k}$ to be the proportion of events that occur out of the n_k occasions on which π_k is forecast:

$$\overline{\mathrm{x_k}} = \frac{1}{n_k} \sum_i x_i \tag{1}$$

BSrel is then defined as follows:

$$\mathrm{BSrel} = \sum_k \frac{n_k}{n} (\pi_k - \overline{\mathrm{x_k}}) \tag{2}$$

BSrel takes values in the interval $[0,\infty)$, with 0 being the best score obtained when the conditional relative frequencies are equal to their corresponding forecasts. The plot of BSrel values for different lead time allows us to compare how the discrimination ability in the three forecasting systems varies in time. (**Figure 3a**).

Reliability diagrams plot the relative forecast probability of an event against its correspondent relative observational frequency, indicating how often a forecast probability occurred in reality (**Figure 3b** - Reliability diagram). Perfect forecasting systems should have forecasts of x% being observed x% of the time. In this case, the reliability diagram lies on the diagram’s diagonal. If the reliability diagram is above the diagram’s diagonal for a specific forecast probability threshold, the forecasting system tends to underpredict the considered event with such forecast probability threshold. If it lies below the diagram’s diagonal, the forecasting system tends to overpredict it. When analysing reliability diagrams, it is important

to know the frequency distribution of forecasts issued with certain probabilities to indicate how forecasts tend to be distributed. For example, the small probability thresholds (within the red box in **Figure 3b** - Reliability diagram) are the most important when considering high VRTs because the sample of events exceeding the VRT with high probabilities is small. For this reason, reliability diagrams must be accompanied with sharpness diagrams, which plot the absolute frequency of forecasts issued with different probabilities (**Figure 3b** - Sharpness diagram).

3.2 Forecast discrimination ability: ROC curves (breakdown score) and area under the ROC curves (summary score)

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):

$$HR = H / (H+M) \text{ [values between 0 and 1]} \quad (1)$$

$$FAR = FA / (FA+CN) \text{ [values between 0 and 1]} \quad (2)$$

HRs are mapped (Y-axis) against FARs (X-axis) in a unit square. The location of the ROC curve in the graph (**Figure 3c**) breaks down the measure of HRs against FARs for each probability threshold. The values of the geometrical area under the ROC curve (AROC) provides a summary measure of the discrimination ability across all probability thresholds. The plot of AROC values for different lead times allows us to compare how the discrimination ability in the three forecasting systems varies in time (**Figure 3d**). 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. 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. The area under the ROC curve is computed using the trapezoidal approximation ($AROC_T$), i.e. by adding the areas of the single trapeziums formed by the straight lines between ROC's consecutive points. ROC curves can also be built by fitting the real ROC curve. 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 employs 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 ($AROC_Z$).

4 Results

4.1 Summary scores for reliability and discrimination ability: BSrel and AROC

Across all lead times and all VRTs (**Error! Reference source not found.a-c**), Multiple-WT ecPoint shows the best forecast reliability against point verification. This can be seen from the orange line (Multiple-WT ecPoint) in lying below the green (ENS) and grey (Single-WT ecPoint) lines. The best improvement in reliability compared to ENS, is obtained by both post-processed forecasts for VRT = 0.2 mm/12h (**Error! Reference source not found.a**). The distance between the orange and the grey line increases with increasing VRTs. In particular, the reliability for Single-WT ecPoint becomes increasingly closer to the reliability for ENS for VRT = 10 mm/12h (**Error! Reference source not found.b**) to finally show a worst reliability than the raw forecasts for VRT = 50 mm/12h (**Error! Reference source not found.c**). 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 more reduced sinusoidal pattern, with Multiple-WT ecPoint showing the most linear trend (although with increasing noise for VRT = 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. For example, ENS's reliability is highly lead time dependent, especially for VRT = 10 and 50 mm/12h. ENS gradually improves its reliability with lead time showing a downwards trend up to t+168 (i.e. day 7). Afterward, the reliability worsens progressively showing a n upwards trend. The significance (at 99% confidence level) of the reliability difference between the three forecasting systems 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. However, it is worth noting that, in VRT = 50 mm/12h, Multiple-WT ecPoint shows uncertainty peaks in steps corresponding to accumulation periods ending at 12 UTC.

Across all lead times and VRTs (**Figure 4d-f**), Multiple-WT (orange continuous lines) and Single-WT ecPoint (grey continuous lines) show larger AROCt values than ENS (green continuous lines). For VRT = 0.2 mm/12h (**Figure 4d**), the distance between the AROCt lines is the smallest. As VRTs increase (**Figure 4e-f**), the difference between AROCt 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 4d-f**) are larger than AROCt (continuous lines). The differences between the lines corresponding to AROCz for the three forecasting systems 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. However, it increases significantly for VRT = 50 mm/12h (**Figure 4f**), in particular for ENS. 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. However, the relative position of the correspondent AROCz lines is swapped for lead times up to t+144 (i.e. day 6).

4.2 Breakdown scores for reliability and discrimination ability: reliability/sharpness diagrams and ROC curves

Figure 5a-c show the reliability diagrams for the three forecasting systems, respectively, for VRT = 0.2, 10, and 50 mm/12h and the rainfall accumulation period ending at t+24 (i.e., day 1 forecast). Focused is given to small probabilities (between 1 and 10%). However, the reliability diagrams covering the full range of probabilities is shown in the figures' inserts. **Figure 5d-f** show the correspondent sharpness diagrams. Note that the extensive noise in the reliability diagram for VRT = 50 mm/12h (**Figure 5c**) for probabilities greater than 30% is due to the low number of occurrences of such extreme event with high probabilities (**Figure 5f**). For all VRTs, ENS (turquoise line) tends to significantly underestimate the frequency of events predicted with small probabilities. Forecasts tend to show probabilities that are half of those observed. When events are predicted with higher probabilities (see inserts), ENS tends instead to overpredict. Typically, events predicted with 80% or 90% of probability of occurrence are observed only 40% or 50% of the time. For VRT = 0.2 mm/12h, Single-WT ecPoint (grey line) shows the tendency to underestimate more than ENS the events with small probabilities (<7%). For the rest range of probabilities, the Single-WT ecPoint tends to overestimate the events' frequencies but less than ENS. For the small probabilities of events exceeding VRT = 10 mm/12h, Single-WT ecPoint tends to underestimate less than ENS the events' frequencies. Such underestimation gets amplified significantly for events with bigger probabilities. For example, events predicted with a 50% frequency are observed on average with a 75% frequency. Finally, Single-WT ecPoint significantly overpredicts events exceeding 50 mm/12h with small probabilities. On average, they are predicted with a frequency that it is observed only half of the times. Multiple-WT ecPoint (orange line) tends to underpredict slightly more than Single-WT ecPoint the events exceeding VRT = 0.2 mm/12h with small probabilities (<10%). For the rest of the range of probabilities, it still shows a tendency to underestimate the events' frequencies, but it has the closest reliability diagram to the diagonal out of the three systems. For small probabilities of events exceeding VRT = 10 mm/12H (<5%), Multiple-WT ecPoint shows perfect reliability (the reliability diagram lies over the diagram's diagonal). For bigger probabilities, it tends to underpredict the events' frequencies, but far less than the Single-WT ecPoint. Multiple-WT ecPoint shows again perfect reliability for events exceeding VRT = 50 mm/12h with small probabilities (<5%).

Figure 5g-i show the real (continuous lines) and binormal (dashed lines) ROC curves for the three forecasting systems, respectively, for VRT = 0.2, 10, and 50 mm/12h and the rainfall accumulation period ending at t+24 (i.e., day 1 forecast). For VRT = 0.2 mm/12h, the real and binormal ROC curves are almost overlapping, and both $AROC_t$ and $AROC_z$ are very similar. This means that both post-processed forecasts (Single- and Multiple-WT ecPoint) do not add any additional information that improves the discrimination ability of ENS in distinguishing between "dry" and "wet" conditions. For VRT = 10 mm/12h, the real and binormal ROC curves are again overlapping. The major difference compared to the previous case is the fact that the last meaningful point in the real ROC curve for ENS (point A, **Figure 5h**) is closer to the bottom left corner of the unit square than the last meaningful points in the real ROC curve for Single-WT ecPoint (point B) and Multiple-WT ecPoint (point C). This causes $AROC_t$ for both post-processed forecasts to be ~5% bigger than that for ENS, while $AROC_z$ remains similar for the three forecasting systems. The $AROC_t$ for Multiple-WT ecPoint (= 0.95) is slightly bigger than that for Single-WT ecPoint (=0.94). The real ROC curves for events exceeding a VRT = 50 mm/12h show a similar behaviour to the

one seen for VRT = 10 mm/12h. The distance between the last meaningful point in the real ROC curve for ENS (point A, **Figure 5i**), Single-WT ecPoint (point B) and Multiple-WT ecPoint (point C) is bigger, so that the $AROC_t$ for both post-processed forecasts is ~27% bigger than that for ENS. The $AROC_t$ for Single-WT ecPoint (= 0.86) is this time slightly bigger than that for Multiple-WT ecPoint (=0.85). This is due to the fact that last meaningful point in the real ROC curve for Single-WT ecPoint (point B) is located higher up and to the right of the last meaningful point of the real ROC curve for the Multiple-WT ecPoint (point C). Due to the position of B and C, the binormal approximation of the ROC curves for Single-WT ecPoint (grey dashed line, **Figure 5i**) lies to the right of the binormal ROC for the Multiple-WT ecPoint (orange dashed line) so that AROCz for the Single-WT ecPoint (=0.956) results smaller than the AROCz for the Multiple-WT ecPoint (=0.965).

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

The addition of a case-study-based subjective verification analysis is useful to understand the behaviour of the three considered forecasting systems during a well-documented extreme rainfall event.

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 6a**), causing the evacuation of 815,000 people and affecting 14.5 million people around the province. The death toll reached 398 deaths.

Figure 6b shows rain gauge observations for 20 July between 00 and 12 UTC, where 465.8 mm of rain were observed in Zhengzhou. **Figure 6c** 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 6c** 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 6c**), 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 6c**). 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 6c**). The fourth column in **Figure 6c** 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 6b**). 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 40% instead of 50 to 80%.

6 Discussion and conclusions

This study focused on evaluating whether the additional cost of implementing a weather-scenario-based post-processing approach (represented here by the Multiple-WT ecPoint) is justified compared to a simpler, no-weather-scenario-based system like Single-WT ecPoint. Calibration and forecast production costs would be indeed smaller for the latter post-processing system, and it is important to factor in such considerations when operationalizing post-processing systems.

Looking in conjunction at the results for reliability and discrimination ability provided by the summary scores (respectively, reliability component of the Brier score and area under the ROC curve), one can create a fairly complete comparative picture about the performance of the three forecasting systems. For events exceeding VRT=0.2 mm/12 (i.e., distinguishing between "dry" and "wet" conditions), the summary scores show that both post-processed forecasts have a similar performance when compared to ENS. They do not add any additional information that improves the discrimination ability (nor real nor potential) of the raw forecasts, but they both significantly improve their reliability, especially at shorter lead times. For events indicating "wet conditions" (i.e., VRT = 10 mm/12h), the potential discrimination ability for the three forecasting systems remain similar. In terms of the systems' real configuration, both post-processed forecasts add additional information that improves the discrimination ability of the raw forecasts, while remaining similar between themselves. However, while the improvements in reliability from Multiple-WT ecPoint compared to ENS remain significant, the improvements from Single-WT ecPoint are smaller, and in some cases, it worsens the reliability of the raw forecasts. The most interesting results are obtained for the "severe rainfall" condition (i.e., VRT = 50 mm/12h). Both post-processed forecasts improve the real and the potential discrimination ability of ENS forecasts. However, in the former case, the Single-WT ecPoint shows a better discrimination ability than the Multiple-WT ecPoint at all lead times, while in the latter case, the Multiple-WT ecPoint is better up to day 6 forecasts. However, while the reliability from Multiple-WT ecPoint remains better than ENS, the Single-WT ecPoint shows a reliability that is consistently worse than the one for ENS up to day 6. This shows that Single-WT ecPoint is likely providing too unrealistically high probabilities for extreme events compared to Multiple-WT

ecPoint. While this is rewarded by the measure used to estimate the discrimination ability of the real configuration of the forecasting systems, this is instead penalized by the measure used to estimate the discrimination ability of the potential configuration (i.e., if we had more ensemble members) and reliability.

The breakdown scores and the case study confirm the results displayed by the summary scores. The analysis of the extreme rainfall event in China highlighted that, while ENS provides good guidance several days on what is the area at risk of experiencing heavy rainfall, it tends to underpredict the actual rainfall totals. The case study also highlighted that ENS did not provide good guidance on which areas might not experience any rainfall by providing 0% change of having less than 0.2 mm/12h (i.e. "dry" conditions) where no rain was observed. The reliability diagrams showed that ENS tends to overpredict the small rainfall amounts and underpredict heavy rainfall. This is line with previous studies (Haiden et al. 2023). The two post-processing systems provide different degrees of improvements to these ENS shortcomings. Single-WT ecPoint overpredicts significantly less than ENS the small rainfall totals. The case study shows that Single-WT provides indeed a better guidance on which areas might not experience any rainfall, increasing the probabilities to 20-40% of having no rain in the dry areas. The reliability diagrams also show that Single-WT tends to overpredict heavy rainfall amounts. The case study confirms this outcome by showing that Single-WT ecPoint predicts rainfall totals that are ~40% more than those observed in the area affected by the heaviest rainfall. This would lead to a high number of false alarms, as confirmed by the ROC curves. This increased number of false alarms would inevitably have a deleterious effect in the trust users have in the post-processed forecast. Both reliability diagrams and case study show that Multiple-WT ecPoint performs best out of the three systems. It is indeed reliable at predicting the small rainfall totals and big rainfall. In the case study, Multiple-WT increases the probabilities to 50-80% of having no rain in the dry areas and provides good guidance on the peak rainfall by providing at day 1, a forecast of a similar order of magnitude of the observed amounts. These outcomes are also confirmed by the binormal ROC curves which show that, on equal hit rates, Multiple-WT shows smaller false alarm rates, increasing users' trust in the forecasts.

The results in this study show 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 particularly important for extreme events to provide timely and accurate forecasts for high-impact events. It is also important for small events (i.e., conditions of dry) to reduce false alarms that would inevitably reduce users' trust in the forecasts.

334 # Tables

335 **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*.

336

Figures

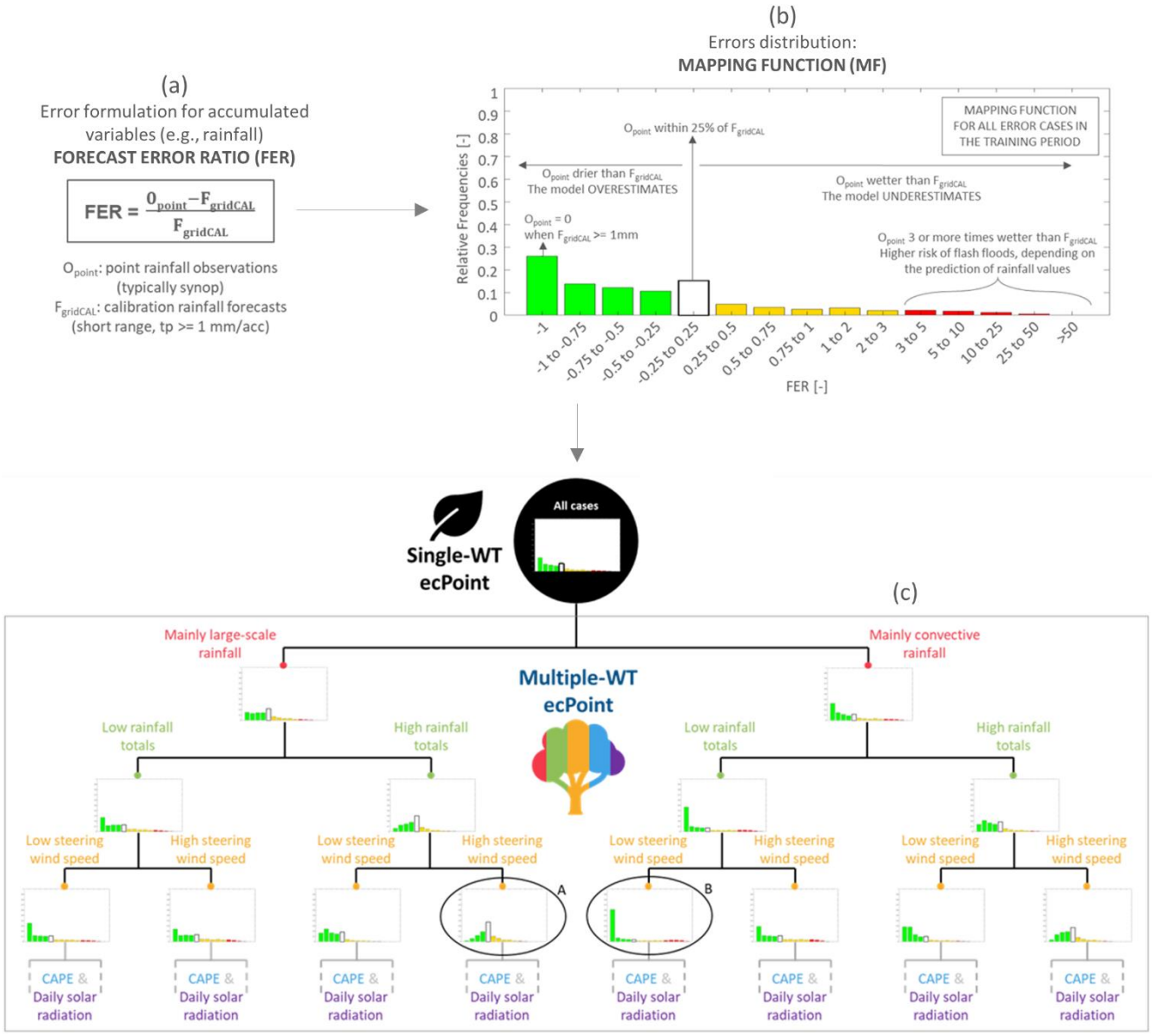

Figure 1 – Schematic representation of ecPoint's single and multiple weather type approach. Panel (a) shows the error formulation between forecasts and observations adopted for accumulated variables, called Forecast Error Ratio (FER). Panel (b) shows a visual representation of the error distribution, 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).

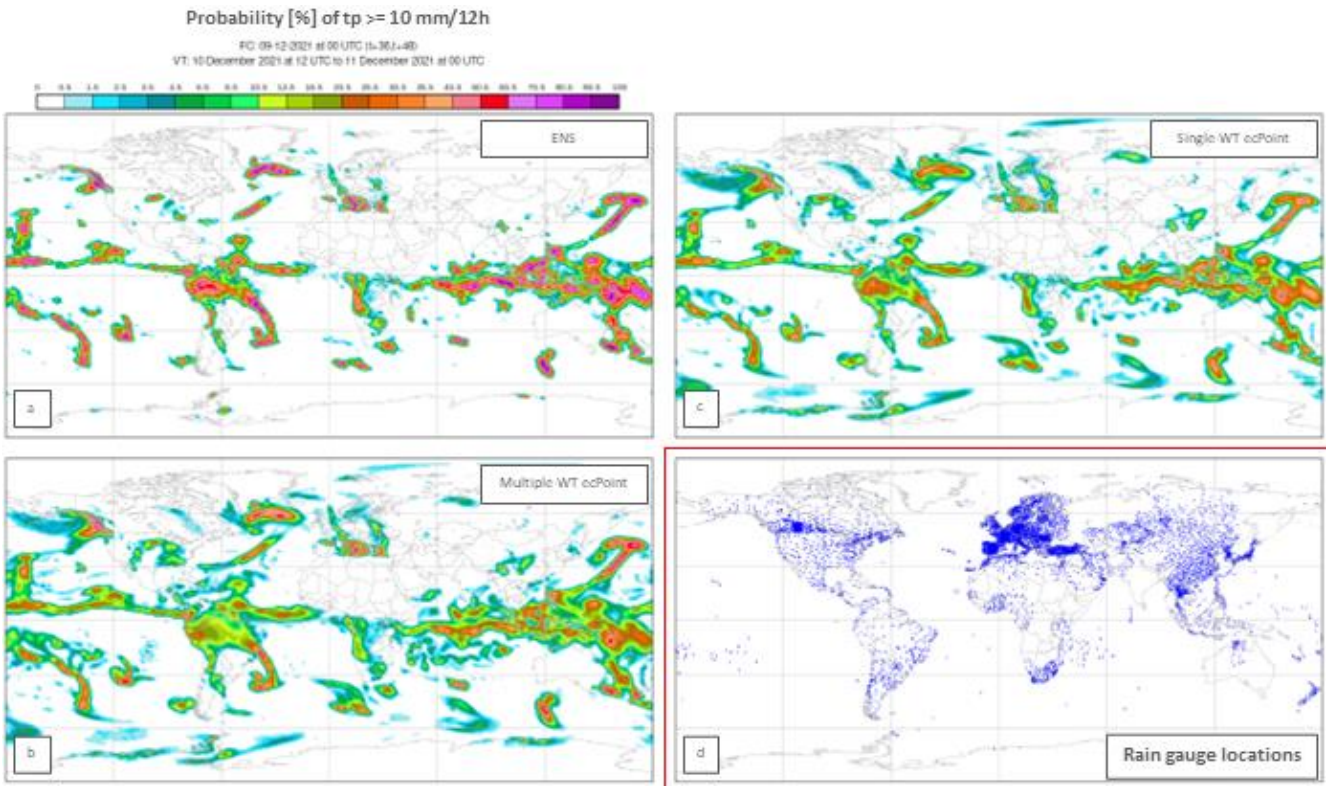

Figure 2 –Panels (a), (b), and (c) display an example of forecast probabilities exceeding 10 mm/12h that can be obtained, respectively, from ENS, Multiple-WT ecPoint, and Single-WT ecPoint. The examples are shown for a day 2 forecast, issued on the 9th of December 2021 at 00 UTC, and valid between 10th December at 12 UTC and 11th December at 00 UTC. Panel (d) displays the location of the rain gauges measuring 12-hourly rainfall for the accumulation periods used in the objective verification, i.e. ending at 0, 6, 12 and 18 UTC.

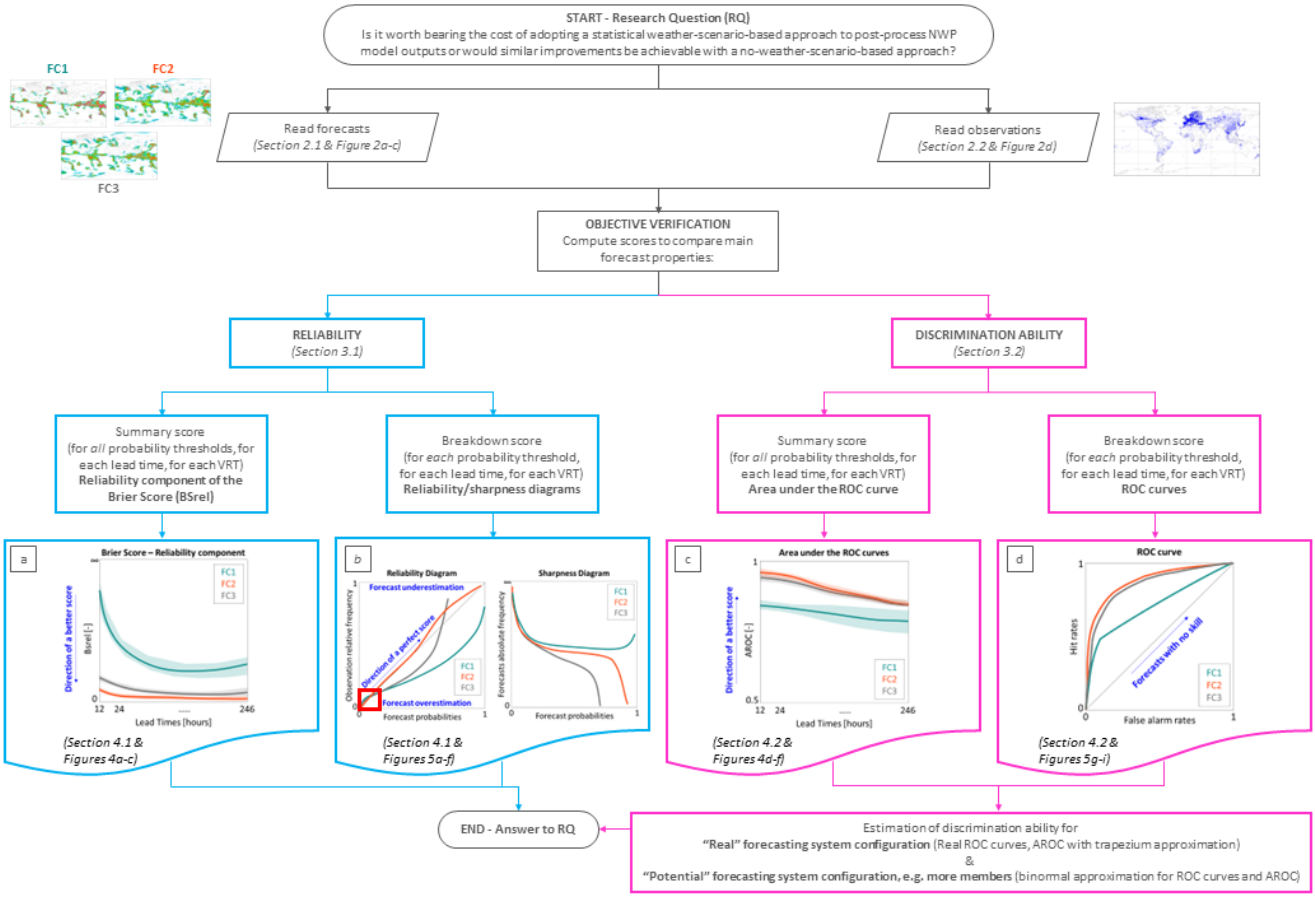

Figure 3 – Flowchart explaining the methodological approach to answer the study's research question (at the start of the flowchart) via objective verification. Two features of probabilistic forecast are compared: reliability (boxes in cyan) and discrimination ability (boxes in fuchsia). The sections where the methods are described in detail and where the correspondent results can be found are indicated in italics. For reliability, panels (a) and (b) show, respectively, examples of the diagrams for the reliability component of the Brier score and reliability/sharpness diagrams (the red box highlights the area representing small forecast probabilities). For discrimination ability, examples of the diagrams for the areas under the ROC curves and the ROC curves are shown, respectively, in panels (c) and (d).

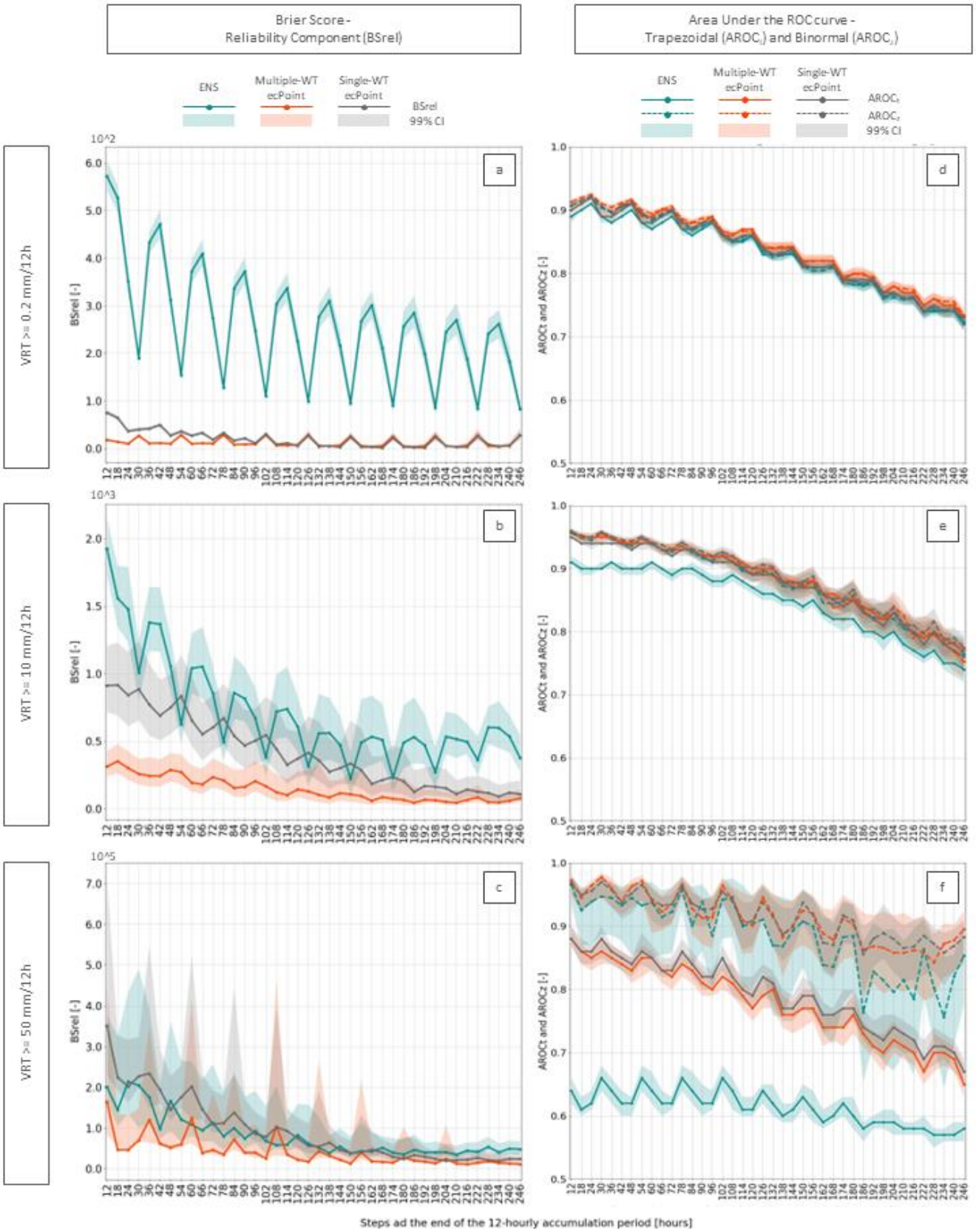

Figure 4 - Panels (a), (b), and (c) display the reliability component of the Brier Score (BS_{rel}) for VRT >= 0.2, 10, and 50 mm/12h, respectively, up to t+246 (i.e., day 10 forecast). Panels (d), (e), and (f) display the trapezoidal (continuous lines, $AROC_t$) and Binormal (dashed lines, $AROC_z$) 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 BS_{rel}, $AROC_t$, and $AROC_z$ values for ENS, ecPoint_MultipleWT and ecPoint_SingleWT, respectively. The shaded areas represent the correspondent confidence intervals at 99% confidence level.

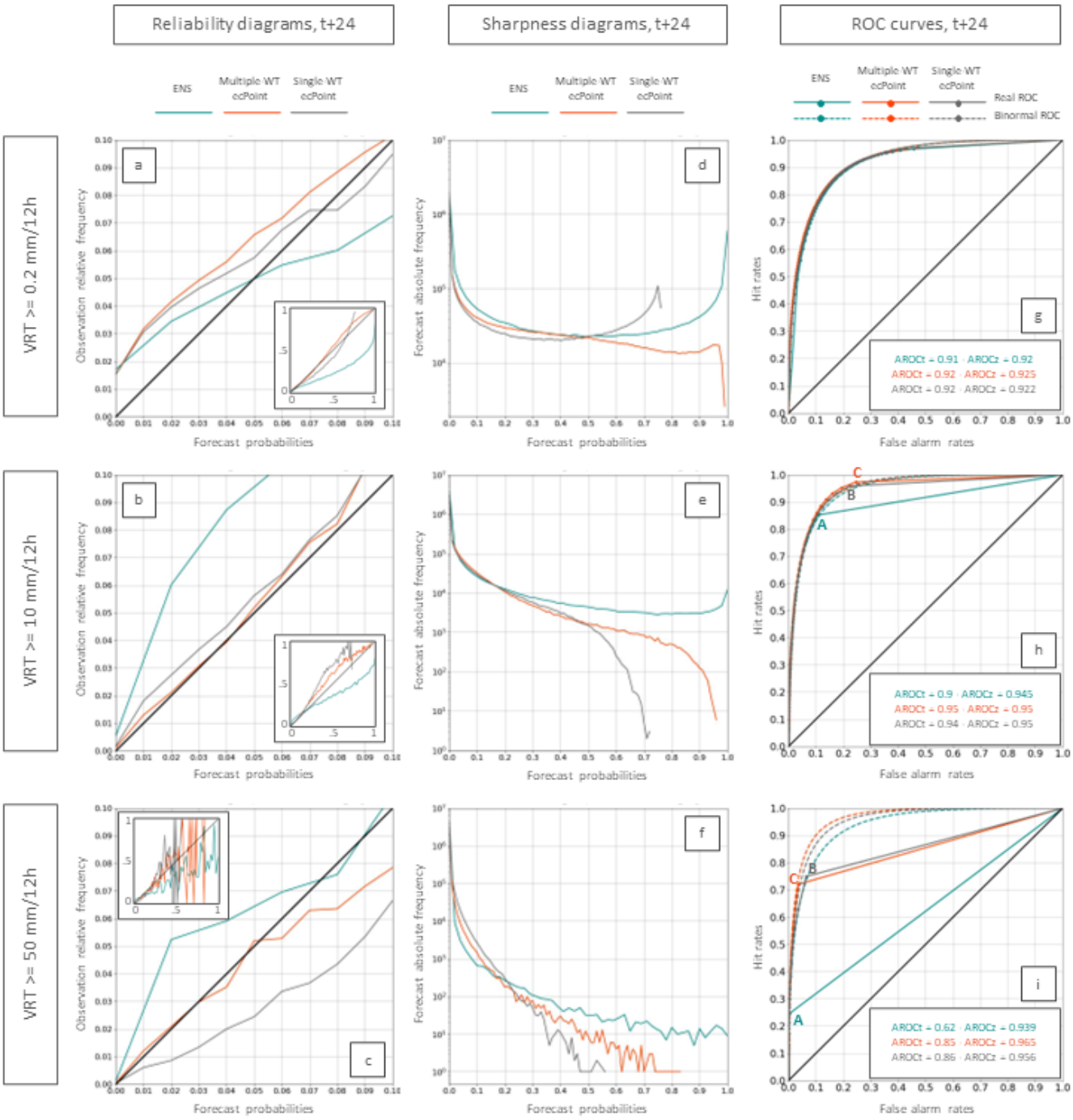

Figure 5 – Panels (a), (b), and (c) show the reliability diagrams for the accumulation period ending at t+24 (day 1 forecast) and for VRT >=0.2, 10, and 50 mm/12h, respectively. The reliability diagrams show forecasts probabilities up to 10%, while the full range of forecast probabilities is shown in the correspondent inserts. Panels (d), (e), and (f) display the sharpness diagrams for the same accumulation period and VRTs. Panels (g), (h), and (i) display the real (continuous lines) and binormal ROC curves for the same accumulation period and VRTs. Points (A), (B), and (C) in VRT = 10 and 50 mm/12h show the last meaningful points in the real ROC curves, respectively, for ENS, Single-WT ecPoint and Multiple-WT-ecPoint. The turquoise, orange and grey lines represent ENS, ecPoint_MultipleWT and ecPoint_SingleWT, respectively.

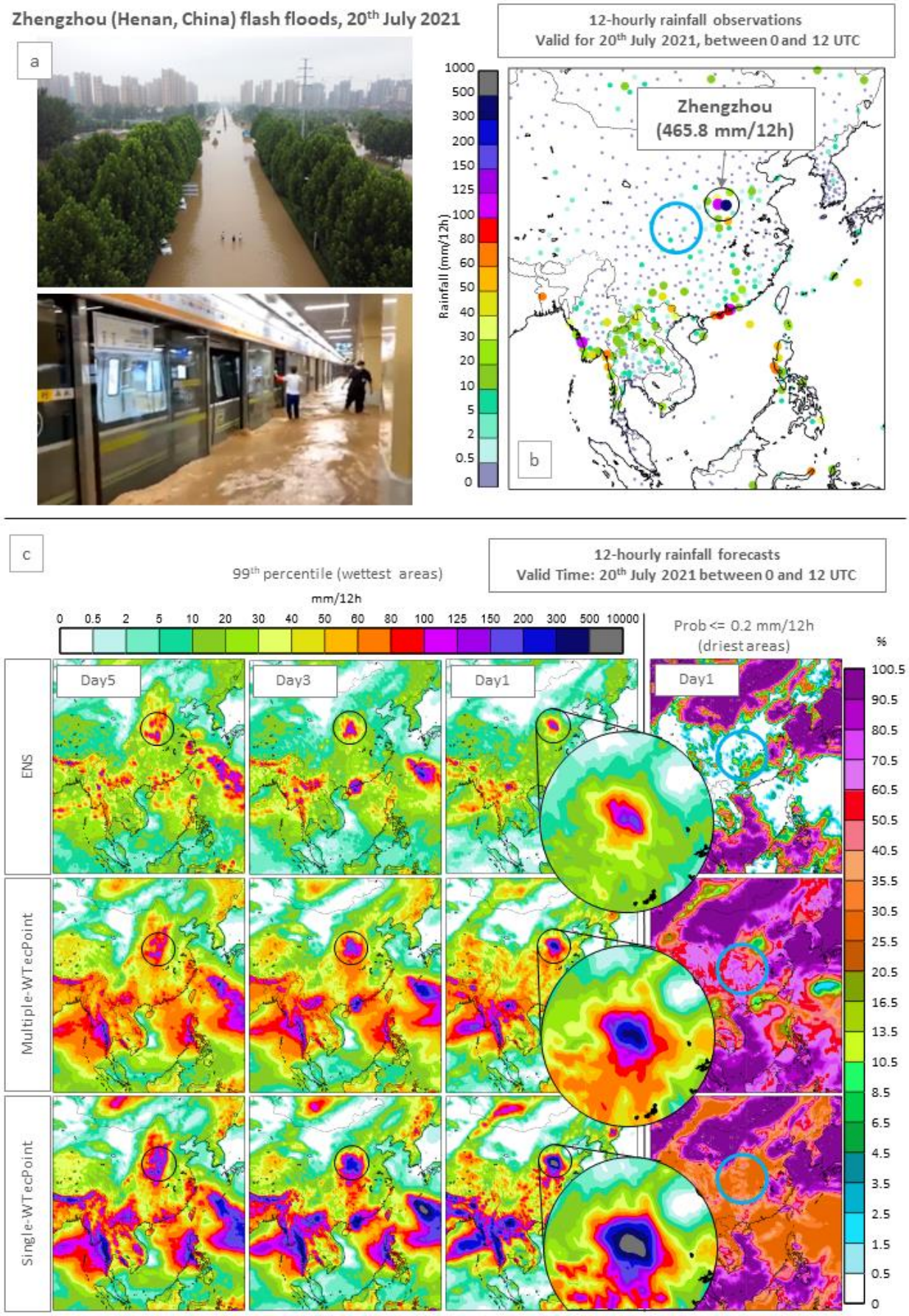

Figure 6 - Flash floods in Zhengzhou (Henah, China) on the 20th of July 2021. Panel (a) shows images of the impacts of the flash floods in Zhengzhou (credits to China Dialogue and CNN for top and bottom image, respectively). Panel (b) shows 12-hourly rainfall observations valid for the 20th of July 2021 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.

Hemri, S., T. Hewson, E. Gascón, J. Rajczak, J. Bhend, C. Spirig, L. Moret, and M. A. Liniger, 2022: How do ecPoint precipitation forecasts compare with postprocessed multi-model ensemble predictions over Switzerland? *ECMWF Technical Memoranda*, **901**, https://doi.org/10.21957/hy89j7svk.

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.