Technical Memo

ECMWF
European Centre for Medium-Range Weather Forecasts

924

Does a multivariate approach enhance univariate grid-to-point post-processed rainfall forecasts? A comparative analysis

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

[1] Department of Geography and Environmental Science, University of Reading, Reading, UK
[2] Forecasts and Services Department, European Centre for Medium-range Weather Forecasts, Reading, UK
[3] Department of Meteorology, University of Reading, Reading, UK
[4] Red Cross Red Crescent Climate Centre, The Hague, The Netherlands
[5] Department of Geography and Environment, University of Loughborough, Loughborough, UK
[6] UK Centre for Ecology and Hydrology, Wallingford, UK

December 2024

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/en/publications/

Contact: library@ecmwf.int

© Copyright 2025

European Centre for Medium Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, UK

Literary and scientific copyrights belong to ECMWF and are reserved in all countries. The content of this document is available for use under a Creative Commons Attribution 4.0 International Public License.
See the terms at *https://creativecommons.org/licenses/by/4.0/*.

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error or omission or for loss or damage arising from its use.

Abstract

Reliable and accurate point-scale rainfall forecasts are crucial for mitigating the impacts of localised extreme hydro-meteorological hazards. Nevertheless, biases and the low spatial resolution of global numerical weather prediction (NWP) models can hinder their utility for predictions at a local scale. Statistical post-processing is widely used to correct biases in raw global rainfall forecasts and can address representativeness errors. The most advanced post-processing techniques follow a multivariate approach, using relationships between rainfall observations and multiple predictors to capture physical processes that can influence the corrections applied to the target variable (i.e., raw rainfall forecasts). However, multivariate post-processing is expensive to run and maintain. Post-processing techniques can also follow a univariate approach, adjusting raw rainfall forecasts only through relationships between past forecasts and observed rainfall. Univariate approaches are widely used due to their simplicity and low calibration and production costs. They might also be the only viable approach when the required information to define predictors is unavailable. This study compares the reliability and discrimination ability of rainfall forecasts post-processed with a multivariate and univariate approach to assess whether the former improves performance. While not explicitly evaluated in this study, improved performance is a prerequisite to justifying the multivariate approach's higher maintenance, calibration, and production costs compared to its univariate counterpart. In this study, we considered the post-processing technique developed at the European Centre for Medium-range Weather Forecasts (ECMWF) to produce probabilistic rainfall prediction at point-scale called "ecPoint". One global forecasting system providing outputs in three different ways, two of which involved post-processing, were compared: the raw ECMWF ensemble (ENS), the original and currently operational multivariate ecPoint (M-ecPoint), and an experimental univariate ecPoint (U-ecPoint). Compared to U-ecPoint and ENS, M-ecPoint provides better reliability and discrimination ability. It also provides more timely and accurate predictions of "dry" and "extremely wet" conditions. Moreover, compared to U-ecPoint, M-ecPoint maintains high detection of low-probability, high-impact extreme rainfall events while reducing false alarms, delivering much higher user confidence.

Plain language summary

Accurate and reliable rainfall forecasts are essential for reducing the impacts of extreme weather at local scales. Global weather prediction models might struggle to provide accurate local forecasts due to biases and coarse resolution. Statistical post-processing can improve forecast performance. There are two main post-processing approaches: univariate and multivariate. Univariate approaches are simpler and are less expensive to run and maintain, relying only on the relationship between past forecasts and observed rainfall. Multivariate approaches are more advanced, using multiple predictors to capture physical processes influencing rainfall predictions. They are, however, more expensive and complex to run. This study investigates whether rainfall forecasts post-processed using a multivariate approach perform better than univariate ones, focusing on their ability to produce reliable and skilful rainfall forecasts. In this study, we considered the post-processing technique developed at the European Centre for Medium-range Weather Forecasts (ECMWF) to produce probabilistic rainfall prediction at point-scale called "ecPoint". We analysed one global forecasting system providing outputs in three different ways, two of which involved post-processing: raw ECMWF ensemble forecasts (ENS), the currently multivariate ecPoint (M-ecPoint), and an experimental univariate version of ecPoint (U-ecPoint). Compared to ENS and U-ecPoint, M-ecPoint provides more reliable and skilful rainfall predictions. It also provides more accurate and timely predictions of dry and extreme rainfall conditions, particularly for low-probability, high-impact events, increasing user confidence in the multivariate post-processed forecasts.

1 Introduction

Due to the scale mismatch between rain gauge rainfall totals and grid-scale outputs from Numerical Weather Prediction (NWP) models, NWP rainfall predictions can often be inaccurate, especially in global, coarser NWP models and under certain weather conditions, e.g., convective rainfall (Göber et al., 2008). While weather forecasts rely heavily on NWP models (Bauer et al., 2021; Bauer, 2024), using them is, therefore, not always straightforward. Such a scale mismatch is called representativeness error and increases when the variation seen among observed point values within the model grid-box (i.e., sub-grid variability) becomes large (Janjic et al., 2018). For example, a case of high rainfall sub-grid variability occurs when the model grid-box provides 10 mm of rainfall, but the distribution of point rainfall observations ranges from 0 mm (i.e., the model grid-box value overestimates the observed small rainfall totals) and 50 mm (i.e., the model grid-box value underestimates the observed high rainfall totals). The degree of observed sub-grid variability relates closely to the weather conditions. Dynamic-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 for model grid-boxes with a length scale of 10 km or more. Ensembles cannot address representativeness errors because they uniquely address uncertainties at grid-scale related to synoptic scale errors (Buizza, 2019). They can be addressed by increasing the spatial resolution of NWP model grid-boxes as done in regional, km-scale NWP models (Cafaro et al., 2021). Km-scale models can indeed display realistic-looking rainfall spatial patterns (Mamgain et al., 2024; Demuth et al., 2020; Roberts and Lean, 2008) and improve forecast accuracy by better representing complex features such as orography and convection (Casaretto et al., 2022). However, their geographical coverage is limited due to computational costs (Feijoó and Solman, 2022; Bauer, 2024) and lead times rarely exceed day 2 (Frogner et al., 2019; Keil et al., 2014).

Although research in developing km-scale models is advancing rapidly (Zeman et al., 2021; Bauer et al., 2024; Bauer, 2024), statistical post-processing still offers a cost-effective way to address biases in raw NWP forecasts and, sometimes, representativeness errors. Refer to Vannitsem et al. (2021) for a comprehensive literature review of state-of-the-art statistical post-processing techniques applied to NWP model outputs. Any statistical post-processing system aims to characterise errors in a database of past forecasts for a specific NWP model version and use such error characteristics to correct future forecasts produced by the considered NWP model. Some statistical post-processing systems apply corrections to raw forecasts according to information provided by historical observations (Klein et al., 1959). Other systems might use relations built between the observations and predictors that can be static, e.g. orography-related, or dynamic, e.g., generated by the NWP model itself (Glahn and Lowry, 1972). These two different approaches will be called univariate and multivariate in this study[1]. Univariate approaches are straightforward and computationally efficient. While simple, univariate approaches are still widely used (Bouallegue et al., 2020; Zorzetto et al., 2024), mainly due to their low calibration and production costs and because information from other sources (e.g., other NWP variables) might be unavailable. On the other hand, multivariate approaches can handle complex relationships and better capture the physical processes that might influence the target variable (Şan et al., 2023). One example of multivariate post-processing is "ecPoint", developed at the European Centre for Medium-range Weather

[1]The authors acknowledge that there is another, more widely adopted meaning for these two terms: "univariate" typically refers to an approach that post-processes a single target variable at a time, while "multivariate" typically refers to an approach that post-processes multiple target variables at the same time to capture intervariable, spatial and temporal dependence structures (Vannitsem et al., 2018). While recognising these definitions, the "univariate" and "multivariate" terms will be used in this study as indicated in the main text.

Forecasts (ECMWF) and operational since 2019 (Hewson and Pillosu, 2021). ecPoint is a statistical post-processing technique that transforms global raw, gridded NWP model outputs into probabilistic predictions at point-scale. ecPoint's formulation acknowledges that weather scenarios at grid-box scale can forewarn the degree of sub-grid variability for the variable under consideration and estimate biases at grid scale. ecPoint's multivariate approach differentiates its corrections according to weather scenarios defined using variables such as the fraction of convective rainfall, wind speed, CAPE, solar radiation, and standard deviation of sub-grid scale orography. For rainfall, ecPoint has improved the reliability and discrimination ability of global rainfall forecasts against point verification across different lead-time ranges (Gascón et al., 2024; Hemri et al., 2022; Hewson et al., 2023; Hewson and Pillosu, 2021). While ecPoint's multivariate approach has been shown to improve the quality of raw ENS rainfall forecasts, its training and forecast production are more costly than they would be for a univariate statistical post-processing system. The training of ecPoint consists of building a decision tree that increases substantially in size with the increase of the number of considered predictors. Since the creation of the decision tree is done in a semi-subjective fashion, increasing the number of predictors might require a person to work full-time on the training of ecPoint-based forecasts for a very variable period of time, from a few hours (e.g. the new predictors are at the top of the decision tree) to up to a week (e.g. the new predictors are at the bottom of the decision tree). On average, adding 50 leaves to the decision tree means adding approximately 1 minute to compute the post-processed forecasts (considering the currently available ensemble configuration and computing power at ECMWF). While this increase in computational time might seem small, it can be critical in operational settings, considering that adding only one predictor can increase the decision tree of up to 100 to 200 leaves. While not explicitly evaluated in this study, improved performance is a prerequisite to justifying the multivariate approach's higher maintenance, calibration, and production costs compared to its univariate counterpart. Hence, this study addressed the following research question: does a multivariate post-processing approach improve the reliability and discrimination ability attained through a univariate approach? To answer this question, one forecasting system providing outputs in three different ways, two of which involved post-processing, were compared: the raw ECMWF ensemble (ENS), the original and currently operational multivariate ecPoint (M-ecPoint), and an experimental univariate variant of ecPoint (U-ecPoint).

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

2 Data

2.1 Rainfall observations

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 mainly for Europe (Haiden and Duffy, 2016). Accumulation periods ending at 00, 06, 12 and 18 UTC were considered between the 1st of December 2021 and the 30th of November 2022. On average, there were 10,000 observations for each accumulation period.

2.2 Rainfall forecasts

2.2.1 ECMWF ENS

ECMWF ENS consists of one control run starting from the best possible representation of initial conditions and 50 perturbed members using singular vectors, a data assimilation ensemble and stochastic model uncertainties (Buizza, 2019). We used forecasts up to day 10 in their native octahedral reduced-Gaussian grid for the period between the 1st of December 2021 and the 30th of November 2022 (only 00 UTC runs were considered). Such forecasts belong to the 47r3[2] cycle, which had a spatial resolution of ˜18 km at the equator (Owens and Hewson, 2018).

2.2.2 Univariate and Multivariate 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 post-processed 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 sub-grid variability (Göber et al., 2008). The errors between global gridded rainfall forecasts (i.e., up to t+30, control run) and point-rainfall observations (i.e., rain gauges) are computed for a one-year calibration period. The error computed for accumulated variables, like rainfall, is the Forecast Error Ratio (FER), whose formulation is shown in Figure 1a). The error distribution is named Mapping Function (MF), and its shape is linked to the degree of sub-grid variability and biases at grid scale in the raw forecasts. The MF for all data points in the calibration period is also shown in Figure 1a, and it shows that ENS both overestimates (green bars) and underestimates (yellow and red bars) versus gage reports. The white bar indicates that only ˜15% of the point-rainfall observations were correctly predicted. The MF shape can significantly change according to different weather scenarios at grid-scale (called Grid-box Weather Type, G-WT). The multiple MFs can be visualised with a decision-tree-like representation, where each leaf of the decision tree corresponds to a G-WT and its corresponding MF (Figure 1b). When for a grid-box, the raw ENS predicts high totals of mainly large-scale rainfall and strong steering wind speeds (case A in Figure 1b), the MF takes a Gaussian-like form. This means the raw model output is relatively representative of the point-scale rainfall totals. When the raw ENS predicts mainly convective rainfall with light steering wind speeds (case B in Figure 1b), the MF might take an exponential-like form. This means that the raw model output is not representative of the point-scale rainfall totals and that the expected degree of sub-grid variability is bigger than in case A. Each raw forecast is converted into a distribution of N point-scale forecasts using the MFs (for example, operationally, for each raw ensemble member, N=100 point-scale forecasts are created).

Suppose all grid-boxes in the raw forecasts were post-processed by sampling only the MF in Figure 1a (i.e., MF within the black circle at the top of the decision tree in Figure 1b), the ecPoint post-processing would be following a univariate approach (U-ecPoint). This approach can be represented by a single-leaf decision tree. U-ecPoint generally increases the number of zeros in the distribution of point-rainfall forecasts to correct ENS's tendency to overpredict small rainfall totals (Haiden et al., 2023), as well as generally increasing the amounts in the wet tail of the rainfall distribution to correct for ENS underestimation of high rainfall values (Haiden et al., 2023).

When each grid-box in the raw forecast is post-processed differently by using the MF corresponding to

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

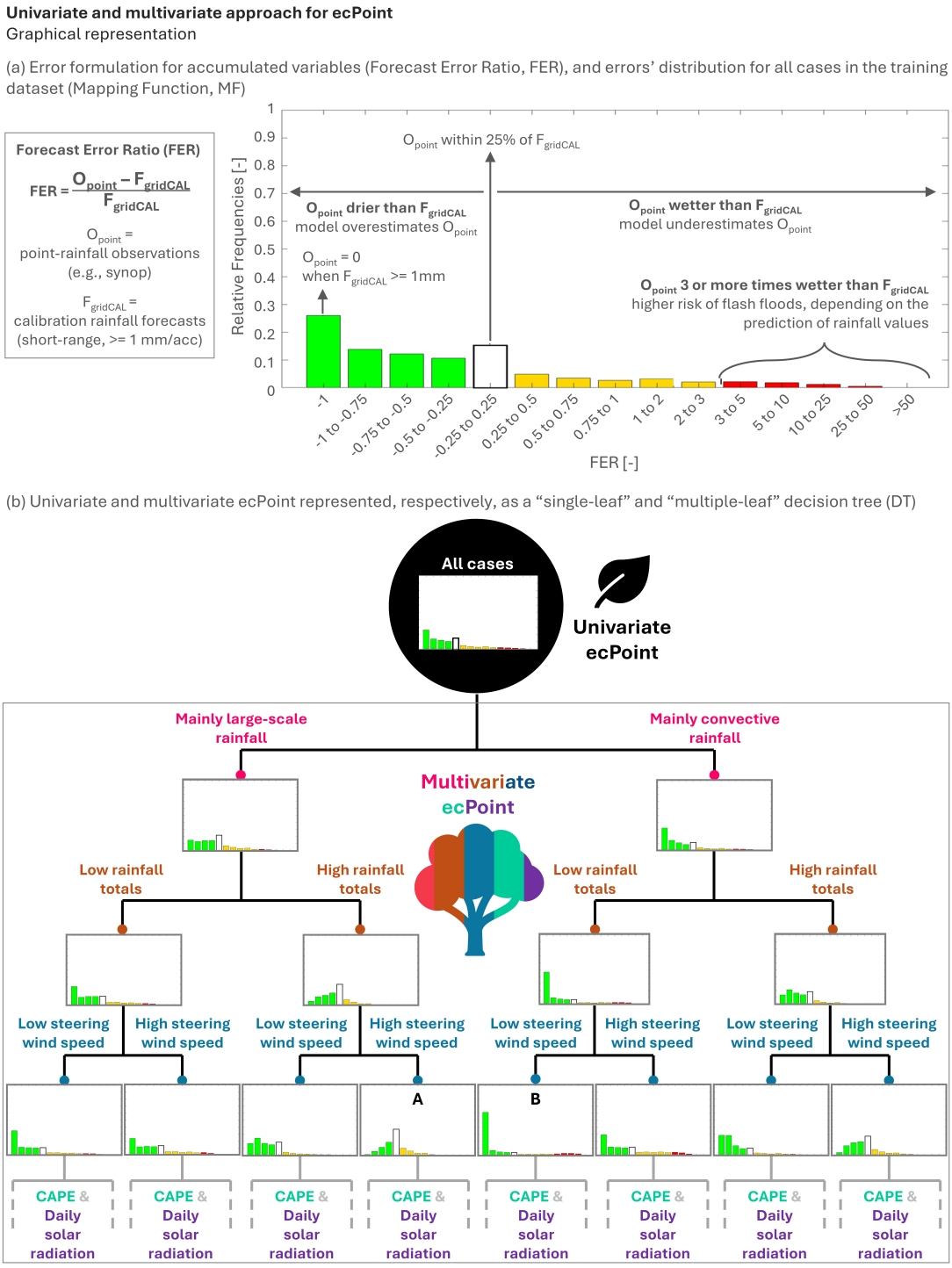

Figure 1: Graphical representation of univariate and multivariate approach for ecPoint. Panel (a) shows the error formulation for accumulated variables (Forecast Error Ratio, FER) and the error distribution for all cases in the training dataset (Mapping Function, MF). The example pertains to the calibration of 47r3 ECMWF ensemble forecasts for 12-hourly rainfall forecasts. Panel (b) shows the univariate approach for ecPoint (U-ecPoint) represented as a "single-leaf" decision tree (DT, within the black circle), while the multivariate approach (M-ecPoint) is represented as a "multiple-leaf" DT (within the grey square).

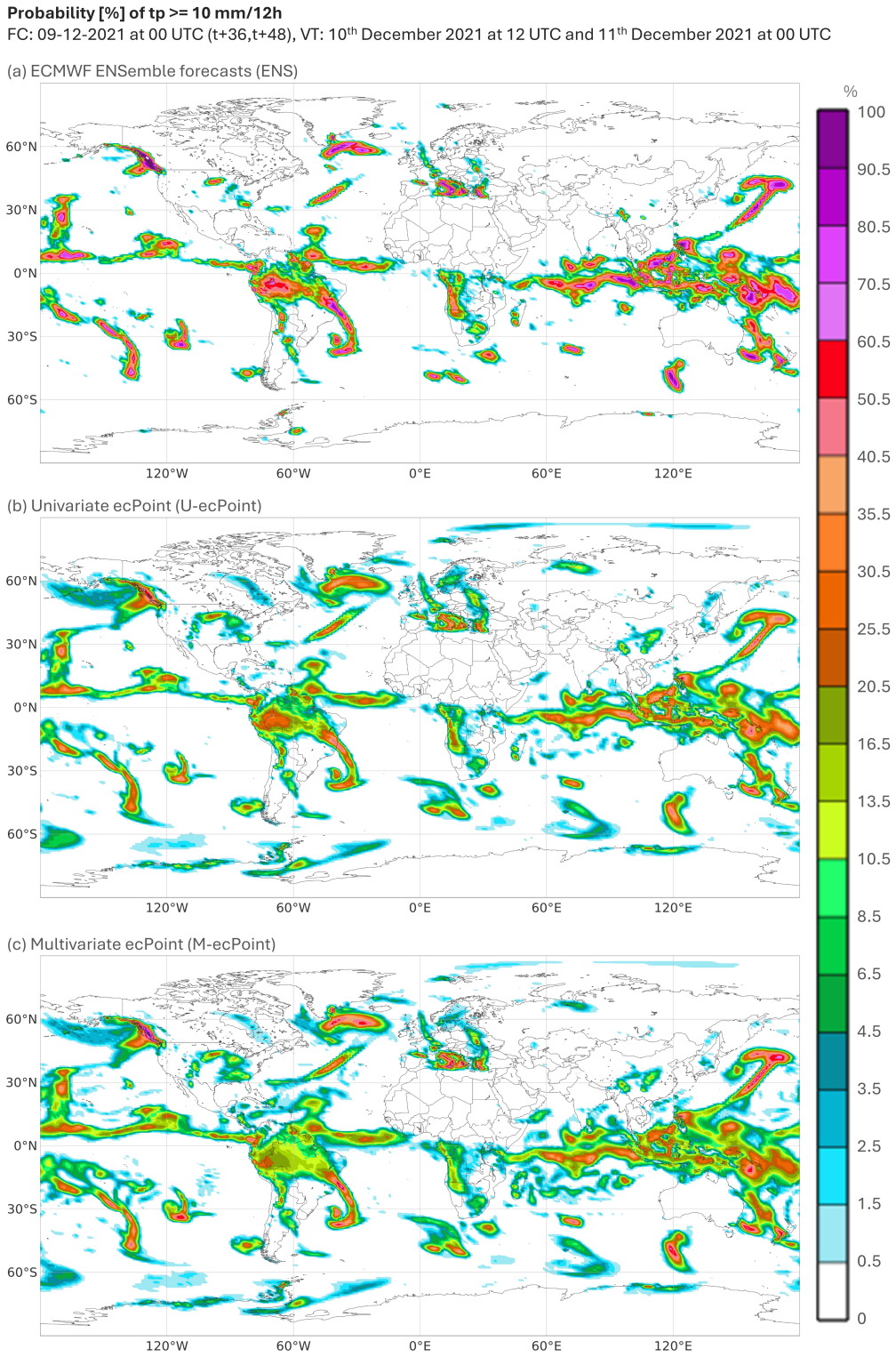

Figure 2: Panels (a), (b), and (c) show the probabilities (in %) of exceeding 10 mm/12h for, respectively, ECMWF ensemble (ENS), Univariate ecPoint (U-ecPoint), and Multivariate ecPoint (M-ecPoint). The examples are shown for a day 2 forecast, issued on the 9th of December 2021 at 00 UTC, and valid between 10th December at 12 UTC and 11th December at 00 UTC.

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*.

the matching G-WT in the decision tree (within the grey square in Figure 1b), the ecPoint post-processing would be following a multivariate approach (M-ecPoint). M-ecPoint can be seen as a multiple-leaf decision tree, and it corresponds to the original ecPoint system (Hewson and Pillosu, 2021). While M-ecPoint increases overall the frequency of small and large rainfall totals in the post-processed forecasts as U-ecPoint does, its adjustments are applied according to the different G-WTs. Hewson and Pillosu (2021) have shown that, due to the G-WT differentiation in the corrections, one of M-ecPoint's features is the ability to shift the location of areas at higher risk of extreme localised rainfall. This feature in U-ecPoint is lost as all grid-boxes are post-processed identically. Figures 2a-c show how different the forecast probabilities of exceeding 10 mm/12h can be across the three forecast outputs. Probabilities for ENS (Figure 2a) are mostly reduced by U-ecPoint (Figure 2b) and M-ecPoint (Figure 2c). However, M-ecPoint reduces the probabilities at certain locations more than U-ecPoint; this relates to the fact that corrections are applied differently across the ensemble members rather than uniformly, as done by U-ecPoint.

3 Methods

Reliability and discrimination ability are desirable properties of ensemble forecasts (Wilks, 2019). Both properties are defined against a threshold for rainfall (e.g., 50 mm/12h), referred hereafter to as Verifying Rainfall Threshold (VRT). Reliability measures whether the chosen VRT exceedance is predicted with probabilities that mirror the frequencies with which the event is observed. Discrimination measures forecasts' ability to distinguish situations that lead to events exceeding the VRT, appraising the existence of a signal in forecasts when an event materialises. Post-processing adds value to raw forecasts if reliability and discrimination ability are improved. This study considers two types of scores to analyse reliability and discrimination ability (Figure 3). Summary scores show how reliability and discrimination ability compare between the three analysed forecasts. Breakdown scores will provide detailed insights into how reliability and discrimination ability relate to the full distribution of probabilities. The scores used to assess reliability and discrimination ability are described in sections 3.1 and 3.2.

The reliability and discrimination ability of one-year retrospective ENS, M-ecPoint, and U-ecPoint forecasts were evaluated against rain gauge observations. Only 00 UTC run forecasts were considered between the 1st of December 2021 and the 30th of November 2022[3], up to day 10 lead time. Although raw NWP model outputs do not pertain to point values, verifying gridded forecasts against point-rainfall observations is a common practice to assess forecast performance for site-specific predictions (Haiden

[3]Both U-ecPoint and M-ecPoint use a calibration dataset created using 47r3 IFS forecasts computed for past dates in 2019.

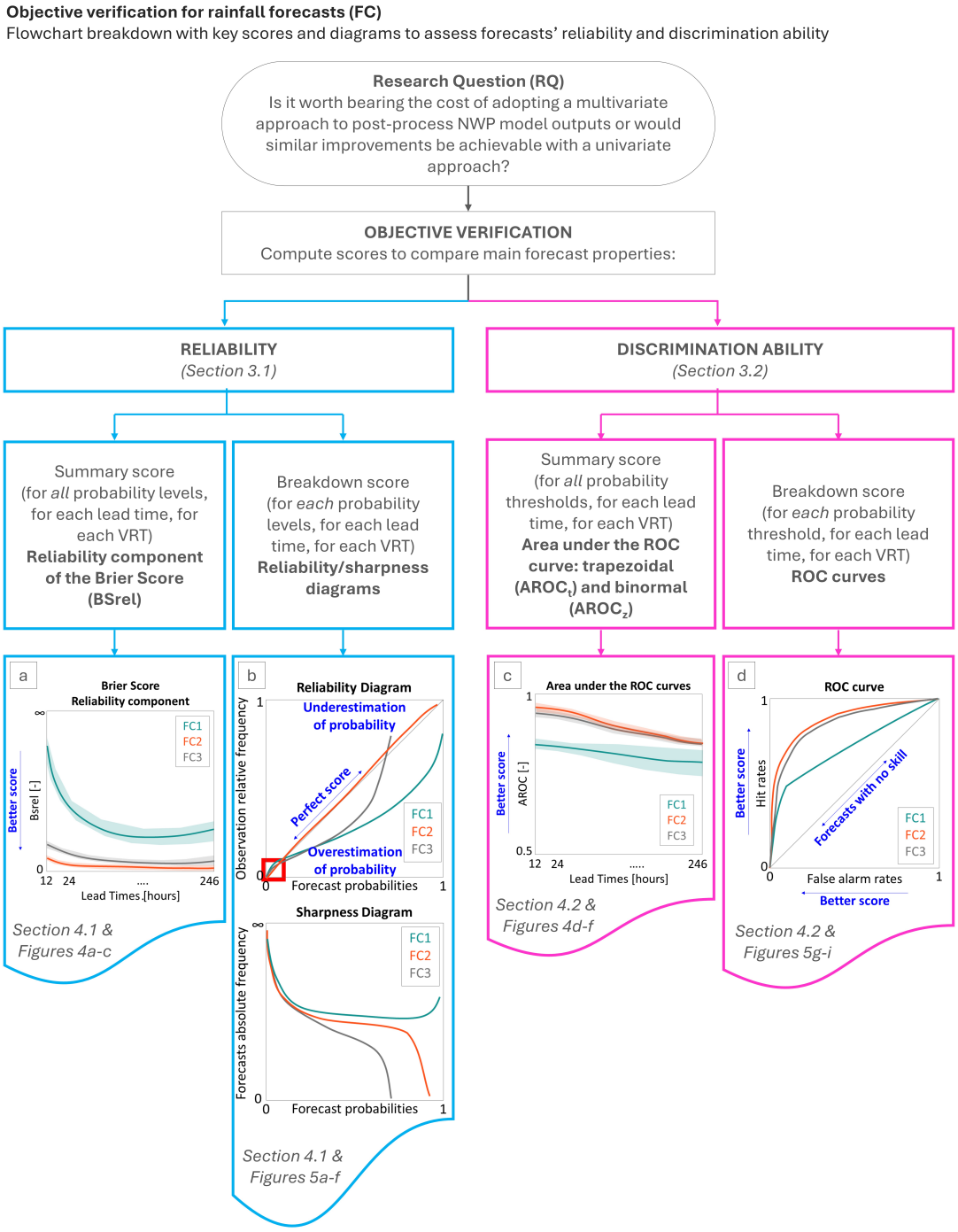

Figure 3: Objective verification for rainfall forecasts. The flowchart breaks down the key scores and diagrams to assess forecast reliability (boxes in cyan) and discrimination ability (boxes in fuchsia). The sections where the methods are described in detail and where the corresponding results can be 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, which is prominent in the results section). 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).

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"). Confidence intervals were added to the scores' estimates by applying the bootstrapping with replacement technique over 1000 repetitions.

3.1 Forecast reliability: reliability component of the Brier score (BS_{rel}, summary score) and reliability/sharpness diagrams (breakdown scores)

The reliability component of the Brier Score (BS_{rel}) quantifies how well the predicted probabilities align with the observed relative frequencies of an event across different forecast probability categories (Ferro and Fricker, 2012). BS_{rel} is defined by equations (1) and (2):

$$BS_{rel} = \sum_{k \in K_0} \frac{n_k}{n} (\pi_k - \bar{x}_k)^2 \tag{1}$$

with:

$$\bar{x}_k = \frac{1}{n_k} \sum_{i \in I_k} x_i \tag{2}$$

π_k is the forecast probability for category k, $\overline{\pi_k}$ is the observed relative frequency of the event for forecasts in that category, n_k is the number of forecasts in group k, and n is the total number of forecasts. I_k denotes the set of indices corresponding to forecasts that fall into category k, and x_i is the binary indicator of whether the event occurred ($x_i = 1$) or not ($x_i = 0$) for the i-th forecast. The term $(\pi_k - \bar{x}_k)^2$ measures the squared difference between the forecast probabilities and the observed outcomes, and $\frac{n_k}{n}$ weights this difference by the proportion of forecasts in each group. BS_{rel} takes values in the interval [0,∞), with 0 being the best score obtainable when all the conditional relative frequencies match their corresponding forecast probabilities. The plot of BS_{rel} values for different lead times (Figure 3a) allows us to compare how the reliability in the three forecast outputs varies in time.

Reliability diagrams plot the relative forecast probability of an event against its corresponding relative observational frequency, indicating how reliable the forecast probabilities are in different classes (Figure 3b - Reliability diagram). For perfect forecasts, when the forecasts show x% probability, observation should meet the criteria x% of the time, so that the reliability curve lies on the diagonal. If the reliability diagram is above the diagonal for a specific forecast probability, those forecasts are underpredicting the likelihood. If it lies below the diagonal, there is overprediction. When analysing reliability diagrams, it is also important to know the frequency distribution of forecasts issued with specific probabilities. 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 forecasts exceeding the VRT with high probabilities is small. For this reason, reliability diagrams should be accompanied by sharpness diagrams (Figure 3b - Sharpness diagram), which plot the absolute frequency of forecasts of different probabilities.

3.2 Forecast discrimination ability: receiver operating curve (ROC, breakdown score) and area under the ROC (AROC, summary score)

Relative Operating Characteristic (ROC) curves are built from 2×2 contingency tables (Table 1), quantifying hits (H), misses (M), false alarms (FA), and correct negatives (CN). Hit rates (HR) and false alarm rates (FAR) are computed, respectively, from equations (3) and (4):

$$\mathrm{HR} = \frac{H}{H+M} \quad [\text{values between 0 and 1}] \tag{3}$$

$$\mathrm{FAR} = \frac{\mathrm{FA}}{\mathrm{FA}+\mathrm{CN}} \quad [\text{values between 0 and 1}] \tag{4}$$

HRs are mapped (Y-axis) against FARs (X-axis) in a unit square. The form of the ROC curve (Figure 3d) shows how HRs vary with FARs as one systematically lowers the threshold probability at which it is assumed that an event has technically been forecast to happen (i.e. from 100% of probability the bottom left corner to 0% probability at the top right corner). The values of the geometrical area under the ROC curve (AROC) provide 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 forecasts varies in time (Figure 3c). Perfect discrimination is obtained when only HRs grow and FARs remain zero. It 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 forecasts have no discriminatory ability. In this case, the ROC curve lies along the diagonal, and the AROC equals 0.5.

How ROC curves and AROCs are computed can impact the interpretation of forecast discrimination ability. ROC curves built for incremental decision thresholds materially assessable from the real ensemble configuration estimate the "real" forecast discrimination ability (Wilks, 2019). Probability thresholds are determined by considering the full discretisation available in the ensemble to ensure ROC curves are as complete as possible (Bouallègue and Richardson, 2022). The number of thresholds corresponds to the number of members exceeding the VRT so that for an ensemble of size M, maximum discretisation 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 with a straight line in the top-right corner of the unit square. 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) by adding the areas of single trapeziums formed by the straight lines between consecutive points in the ROC curve. ROC curves can also be built by fitting a "real" ROC curve. This method effectively consists of an extrapolation to a continuous decision variable based on the limited set of probability thresholds materially assessable from the real ensemble configuration for rare events. Since such a configuration may not be achievable in practice, fitted ROC curves measure the "potential" discrimination that could be achieved with an unlimited ensemble size (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: BS_{rel} and AROC

Out of the three tested systems, M-ecPoint shows the best forecast reliability against point verification across all lead times and VRTs (Figures 4a-c). This is shown by the orange line (M-ecPoint) lying below the green (ENS) and grey (U-ecPoint) lines. Both U-ecPoint and M-ecPoint greatly improve the reliability for small rainfall totals (VRT = 0.2 mm/12, Figure 4a) compared to ENS, with M-ecPoint showing better reliability than U-ecPoint up to t+96 (i.e., day 4 forecast) and remaining comparable afterwards. The gap between the orange and grey line widens as VRT increases. Specifically, U-ecPoint reliability approaches that of ENS at VRT = 10 mm/12h (Figure 4b) and shows worse reliability than the raw forecasts for VRT = 50 mm/12h (Figure 4c) until about day 7. Forecast reliability as a function of lead time displays a sinusoidal pattern, especially for ENS and VRT = 0.2 mm/12h, suggesting a decline in reliability for certain accumulation periods (ending at 12 and 18 UTC). Both post-processed forecasts display a much less pronounced sinusoidal pattern, with M-ecPoint showing the flatter shape, though with increased noise for VRT = 50 mm/12h. M-ecPoint also maintains the most horizontal trend among the three systems, indicating minimal variation in reliability with lead time. In contrast, ENS reliability is highly dependent on lead time, particularly for VRT = 10 and 50 mm/12h, improving gradually until t+168 (i.e. day 7). The significance of the difference in reliability between the three forecasts diminishes with increasing VRTs. Likewise, uncertainty in the forecast reliability estimates also rises with increasing VRTs, but it is more notable for ENS and U-ecPoint. M-ecPoint and U-ecPoint show peaks in uncertainty at accumulation periods ending at 12 UTC for VRT = 50 mm/12h.

Across all lead times and VRTs, U-ecPoint and M-ecPoint show better forecast discrimination ability than ENS (Figures 4d-f). For VRT = 0.2 mm/12h (Figure 4d), the distance between the $AROC_t$ (continuous) lines is small and not significant at the 99% confidence level. Both post-processed forecasts improve upon the discrimination of raw ENS predictions up to t+96 (day 4 forecasts). Afterwards, U-ecPoint has a similar discrimination ability of ENS, while M-ecPoint still shows a better discrimination ability up to day 10 forecasts. The gaps between the $AROC_t$ lines for the post-processed forecasts and ENS widen significantly for increasing VRTs. The difference between the $AROC_t$ lines for U-ecPoint and M-ecPoint is much smaller and not significant at the 99% confidence level. M-ecPoint $AROC_t$ generally remains above U-ecPoint, except for VRT = 50 mm/12h, although the difference is not statistically significant. For all three forecasts, $AROC_z$ values (dashed lines in Figures 4d-f) are bigger than AROCt (continuous lines) at all lead times and VRTs. The differences between the $AROC_z$ lines for the three forecasts are very small and somewhat uncertain. For VRT = 0.2 and 10 mm/12h, uncertainty in the $AROC_z$ estimates seems to be similar to that for $AROC_t$, but increases substantially for VRT = 50 mm/12h (Figure 4f), particularly for ENS. At VRT = 50 mm/12h, the U-ecPoint $AROC_z$ line overlaps the one for M-ecPoint up to t+186 (i.e. day 7.75).

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

Figures 5a-c show the reliability diagrams, respectively, for VRT = 0.2, 10, and 50 mm/12h,), for the 12-hourly rainfall accumulation period ending at t+24 (day 1 forecast). ENS, U-ecPoint, and M-ecPoint are indicated, respectively, in turquoise, grey, and orange. Figures 5d-f display the corresponding sharpness diagrams. They show a steep reduction in the number of events predicted with probability above 10%, so the main reliability diagrams focus on small probabilities ($< 10\%$), while inserts show the

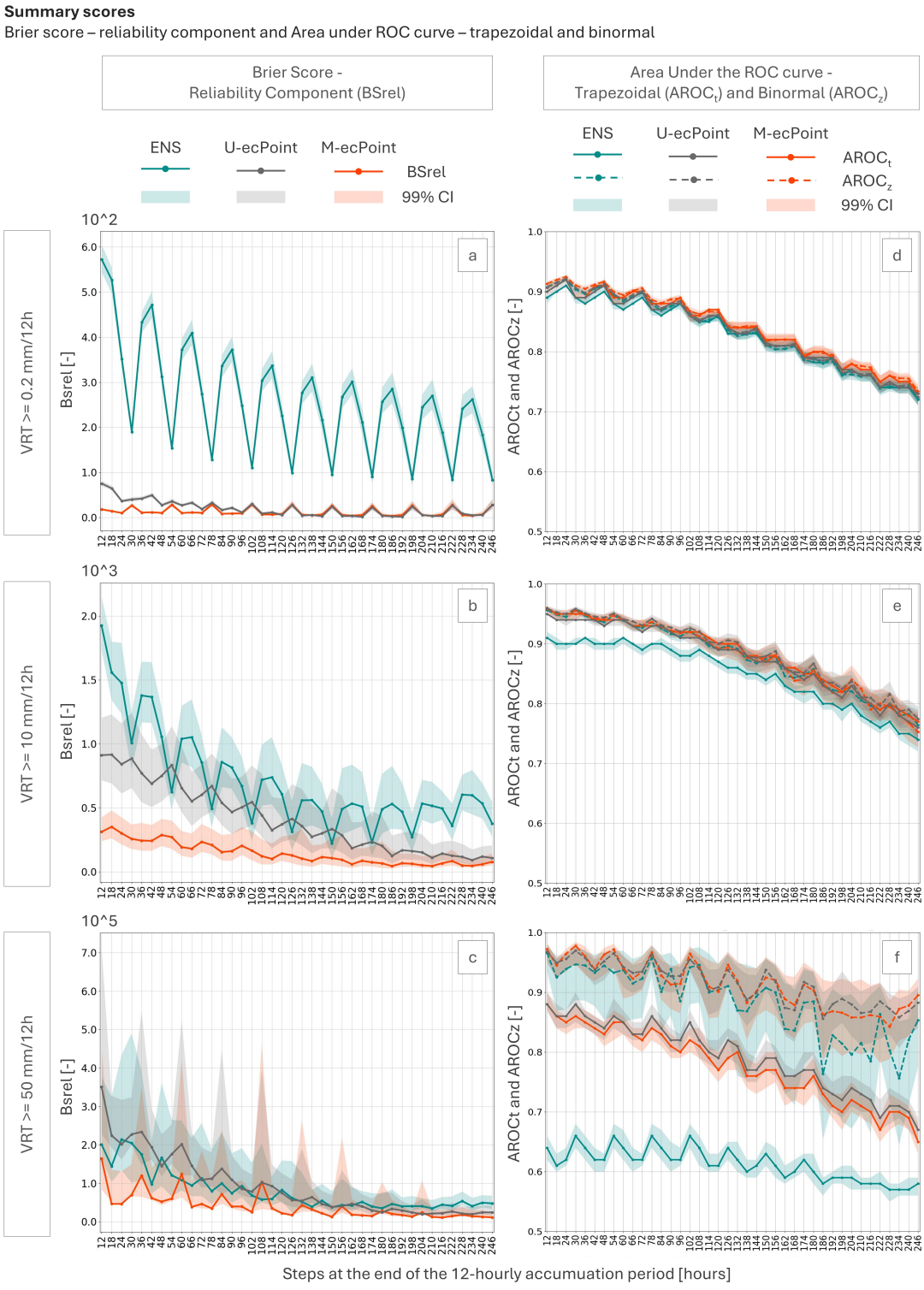

Figure 4: In all panels, ENS, U-ecPoint, and M-ecPoint are represented, respectively, by turquoise, grey, and orange lines. Panels (a), (b), and (c) display the reliability component of the Brier Score (BS_{rel}) for VRT >= 0.2, 10, and 50 mm/12h, respectively, and up to t+246 (i.e., day 10 forecast). Panels (d), (e), and (f) display the trapezoidal (continuous lines, AROC$_t$) and binormal (dashed lines, AROC$_z$) areas under the ROC curve also for VRT >= 0.2, 10, and 50 mm/12h, respectively, and up to t+246 (i.e., day 10 forecast). The shaded areas in all panels represent the confidence intervals at 99% confidence level.

Breakdown scores
Reliability and sharpness diagram and ROC curves – trapezoidal and binormal

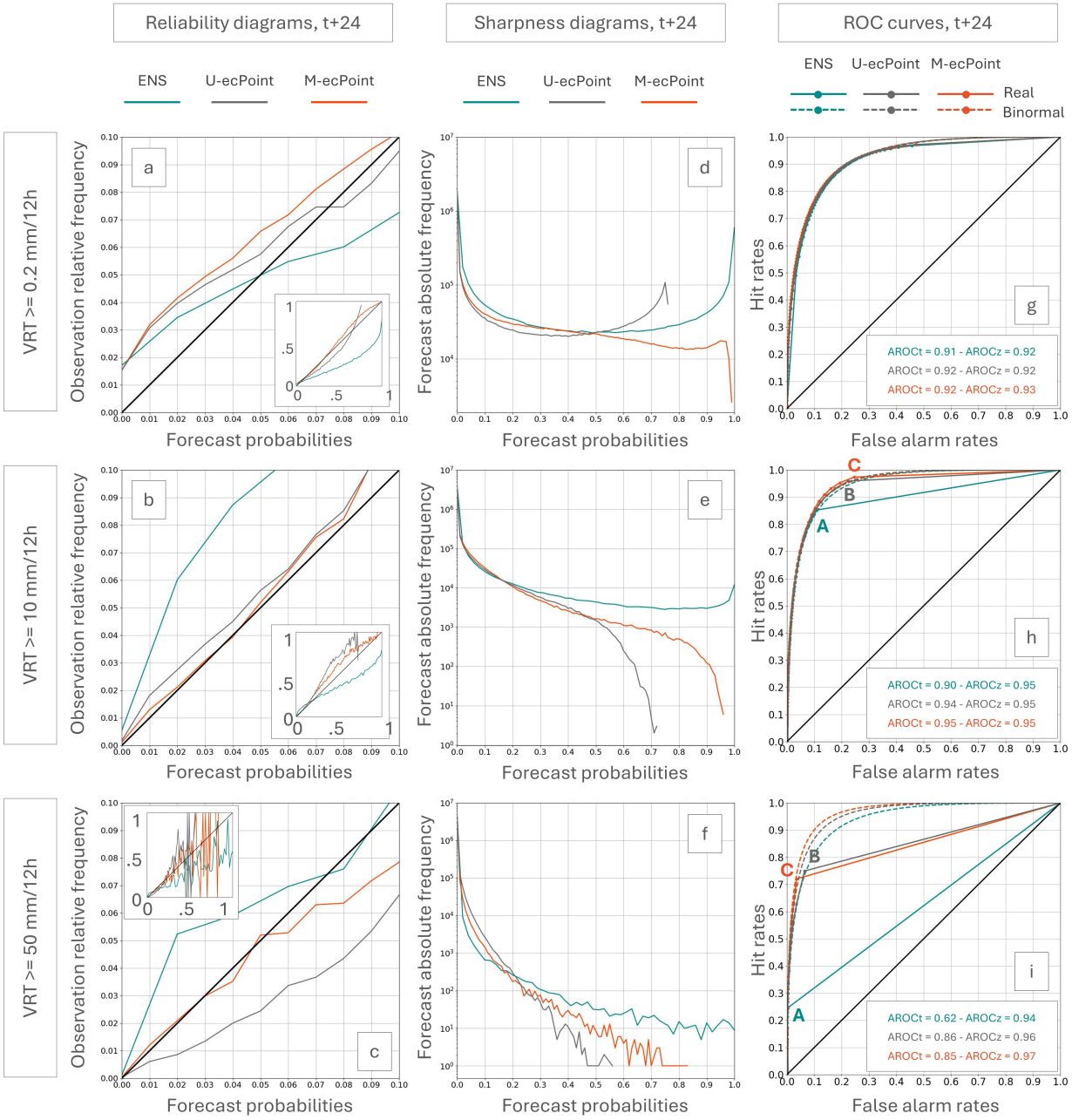

Figure 5: In all panels, ENS, U-ecPoint, and M-ecPoint are represented, respectively, by turquoise, grey, and orange lines. Panels (a), (b), and (c) show the reliability diagrams for the accumulation period ending at t+24 (day 1 forecast) and for VRT >= 0.2, 10, and 50 mm/12h, respectively. The main panels show forecast probabilities up to 10%, while the full range of forecast probabilities is shown in the corresponding inserts. Panels (d), (e), and (f) display the sharpness diagrams for the same accumulation period and VRTs. Note that there are half as many bins for ENS as for ecPoint, meaning the frequency in each bin will ordinarily be double in ENS than in ecPoint. Panels (g), (h), and (i) display the real (continuous lines) and binormal (dashed lines) ROC curves for the same accumulation period and VRTs. Points (A), (B), and (C) in panels (h) and (i) show the last meaningful points in the ”real” ROC curves for, respectively, ENS, U-ecPoint and M-ecPoint.

full probability distribution. For all VRTs, ENS underestimates the frequency of events whenever small probabilities are predicted, as indicated by the turquoise line above the diagonal. For example, when events exceeding VRT = 10 or 50 mm/12h (Figure 5b and Figure 5c) are predicted with a probability of occurrence of 2%, they are observed, respectively, with a 6% and 5% frequency. For higher probabilities, ENS overpredicts the event frequency. For example, when events exceeding VRT = 0.2 mm/12h (Figure 5a) are predicted with an 80% probability of occurrence, they occur only 50% of the time. For VRT = 0.2 mm/12h (Figure 5a), when small probabilities are foreseen, U-ecPoint underestimates more than ENS with the grey line lying above the turquoise line. U-ecPoint overestimates event frequency for higher probabilities but to a lesser extent than ENS. For occasions with small probabilities of exceeding VRT = 10 mm/12h (Figure 5b), U-ecPoint underestimates event frequency less than ENS (grey line below green line), and this underestimation increases for higher probabilities. For instance, when events are predicted with a 50% probability, they are observed with a 75% frequency. U-ecPoint also overpredicts events exceeding 50 mm/12h (Figure 5c), where they are predicted twice as often as observed. M-ecPoint generally underpredicts events exceeding VRT = 0.2 mm/12h (Figure 5a) with small probabilities slightly more than U-ecPoint. M-ecPoint reliability diagram is closest to the diagonal across the full distribution of event probabilities, indicating the best overall reliability out of the three systems. For small probabilities of events exceeding VRT = 10 mm/12H (< 5% in Figure 5b), M-ecPoint demonstrates near-perfect reliability. It underpredicts event frequencies for higher probabilities but much less so than U-ecPoint. M-ecPoint also shows near-perfect reliability for events exceeding VRT = 50 mm/12h when event probabilities are very small as they very often are (Figure 5f). Note that the extensive noise in the reliability diagram for VRT = 50 mm/12h (insert in Figure 5c) for probabilities greater than 30% is due to the low number of occurrences (<100 per bin) of such extreme event at high probabilities (Figure 5f). For VRT = 0.2 mm/12h (Figure 5g), the real (continuous lines) and binormal (dashed lines) ROC curves for all three forecasts nearly overlap, and $AROC_t$ and $AROC_z$ are almost identical (see insert in Figure 5g). This indicates that both post-processed forecasts (U-ecPoint and M-ecPoint) provide minimal additional discrimination ability to raw forecasts (ENS) to distinguish between "dry" and "wet" conditions. For VRT = 10 mm/12h (Figure 5h), the real and binormal ROC curves mostly overlap as well. The last meaningful point in the real ROC curve for ENS (point A) is closer to the bottom left corner of the unit square compared to those for U-ecPoint (point B) and M-ecPoint (point C). This results in $AROC_t$ for both post-processed forecasts to be ~5% larger than that for ENS, while $AROC_z$ remains similar for the three forecasts (see insert in Figure 5h). For VRT = 50 mm/12h (Figure 5i), the last meaningful point in the real ROC curve for ENS (point A) is much closer to the bottom-left corner of the diagram compared to that for U-ecPoint (point B) and M-ecPoint (point C), leading to an $AROC_t$ for U-ecPoint and M-ecPoint (= 0.86 and 0.85, respectively) that is ~27% larger than that for ENS (=0.62). Notably, $AROC_t$ for U-ecPoint is slightly bigger than that for M-ecPoint because the last meaningful point in the U-ecPoint real ROC curve (point B) is positioned higher and further right than the last meaningful point of M-ecPoint real ROC curve (point C). As a result, the binormal ROC curve for U-ecPoint lies to the right of that for M-ecPoint, making U-ecPoint $AROC_z$ (=0.96) slightly smaller than that for M-ecPoint (=0.97).

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

Considering a subjective verification analysis, based on a case study, is useful to understand the performance of the three considered forecasts during a well-documented low-probability, high-impact extreme rainfall event.

In July 2021, the Henan Province in north-east China experienced extremely severe rainfall. Over three

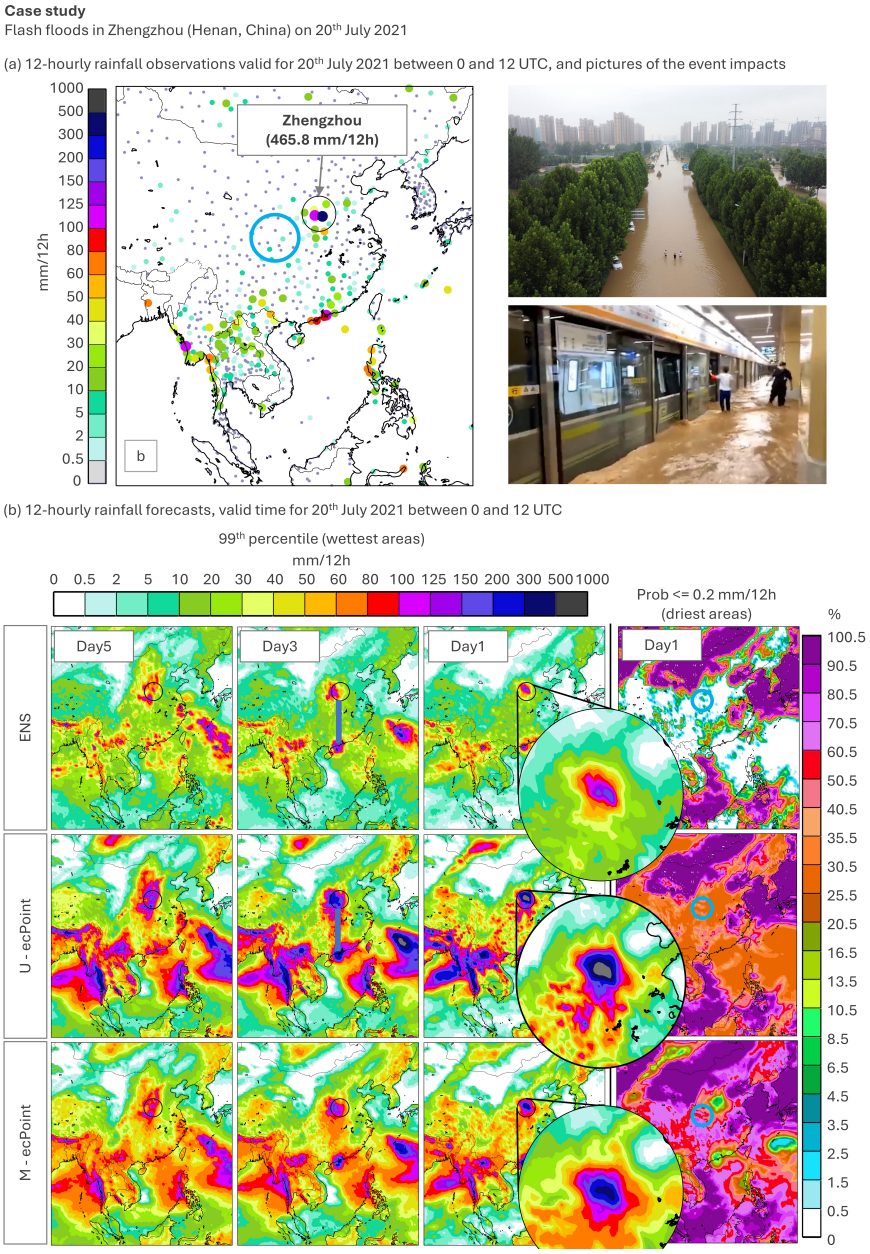

Figure 6: Flash floods in Zhengzhou (Henan, China) on the 20th of July 2021. Panel (a) shows 12-hourly rainfall observations valid for the 20th of July 2021 between 0 and 12 UTC and images of the impacts of the flash floods in Zhengzhou (credits to China Dialogue and CNN for top and bottom images, respectively). Panel (b) shows 12-hourly rainfall forecasts for ENS (first row), M-ecPoint (second row), and U-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), highlighting the rainfall peaks observed and predicted within the black circle. The fourth column shows the probability of having less than 0.2 mm/12h (i.e., having no rain) for a day 1 forecast, highlighting an area within the cyan circle where the rainfall totals were mostly zero. In panel (a), the rainfall totals between 0-0.5 mm/12h are indicated in light grey to allow contrast between the white background and the observations' symbol. In panel (b), the same small rainfall totals are indicated in white to avoid noise in the map plots.

days, between the 17th and the 20th of July, 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 the 20th of July when 201.9 millimetres of rain were recorded between 4 and 5 pm local time, the highest hourly figure ever recorded since measurements began in 1951. Figure 6a shows rain gauge observations for the 20th of July between 00 and 12 UTC, when 465.8 mm of rain were observed in Zhengzhou. Figure 6a also shows the extensive flooding caused by the extreme rainfall. 815000 people were evacuated, 14.5 million people were somehow affected around the province, and 398 people died.

Figure 6b compares the 12-hourly rainfall forecasts for ENS (first row), U-ecPoint (second row), and M-ecPoint (third row), valid for the same observation accumulation period in Figure 6a. The first three columns in Figure 6b show the 99th percentile for day 5, 3, and 1 forecasts (from left to right). For ENS, the 99th percentile corresponds to the wettest member. Due to the discretisation related to the number of raw ensemble members, the 99th percentile is identical to the 98th. For U-ecPoint and M-ecPoint, the 99th percentile represents the rainfall total that, on average, should be observed once at a point every 100 observations. All forecasts, 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). M-ecPoint also highlights many other areas across south-east Asia as being at risk (Figure 6b – day 3 forecasts). Closer to the event (day 1 forecasts), ENS greatly underestimated the observed rainfall totals, predicting rainfall values not higher than 200 mm/12h (zoomed in circle in the first row of Figure 6b). U-ecPoint predicted rainfall totals could reach 700 mm/12h (zoomed in circle in the second row of Figure 6b), while M-ecPoint predicted rainfall totals could reach 400 mm/12h (zoomed in circle in the third row of Figure 6b). The fourth column in Figure 6b shows day 1 forecast probabilities of having no measurable rain ($< 0.2mm/12h$) for the area south-west of Zhengzhou (blue circles), where ~75% of the sites showed no rain (blue circle in Figures 6a-b). ENS shows generally zero probability of having no rain. U-ecPoint shows much smaller probabilities of having no rain than M-ecPoint, between 20 and 40% instead of 50 to 80%.

6 Discussions and concluding remarks

Whatever the degree of sophistication of raw NWP models, statistical post-processing provides corrections that benefit end users who need very localised, yet seamless forecast information (Hemri et al., 2014). Hence, the statistical post-processing of raw NWP model outputs should be integral to any operational forecasting system (Roberts et al., 2023; Trotta et al., 2024). Statistical post-processing systems can follow a univariate or a multivariate approach, noting, however, that calibration and forecast production overheads will be greater for the latter. Hence, when operationalising a multivariate post-processing system, its greater costs must be justified by significantly better forecast performance in some net sense, based on reliability and discrimination ability metrics, compared to what could be provided by univariate approaches.

This study shows that both post-processed forecasts (i.e. U-ecPoint and M-ecPoint) greatly improve ENS reliability and discrimination ability. M-ecPoint, however, indicates overall better performance.

Previous studies (Haiden et al., 2023; Hemri et al., 2014) align with what was found in this study: ENS largely overestimates small point-rainfall totals (VRT = 0.2 mm/12 that distinguish between "dry" and "wet" conditions). In this case, U-ecPoint and M-ecPoint show slightly better (however, not significant) discrimination ability than ENS up to t+96 (i.e. day 4 forecasts) while remaining comparable afterwards. Both post-processed systems significantly improve ENS reliability. However, M-ecPoint shows better reliability than U-ecPoint up to t+96 (i.e., day 4 forecasts) and remains comparable thereafter. The

reliability diagrams also show that M-ecPoint has the best reliability out of the three forecasts, as it is the closest to the graph's diagonal. The real-world case shown in section 5 supports these findings. Of the three forecasts, M-ecPoint was the most reliable in separating between "wet" and "dry" conditions. In dry areas, M-ecPoint predicted ˜60-80% of probabilities of having less than 0.2 mm/12h, closely matching observations that show ˜75% probabilities of no rain. U-ecPoint and ENS largely underestimated the probabilities of having dry conditions, showing only ˜20-40% and 0%, respectively.

Both post-processed forecasts reduced ENS's sinusoidal pattern in the reliability component of the Brier score against lead time, for small and medium rainfall totals (VRT = 0.2 and 10 mm/12). The troughs consistently occur for 12-hourly accumulation periods ending at 06 UTC. Although the observational dataset is global, it is dominated by a greater number of observations over Europe. Therefore, it is not unreasonable to think that troughs reflect forecasts during nighttime conditions and that the sinusoidal pattern corresponds to a reliability degradation due to a mishandling of rainfall's diurnal cycle in ENS despite ongoing improvements (Bechtold et al., 2014). U-ecPoint and M-ecPoint [4] show a much reduced sinusoidal pattern, meaning that post-processing corrections of diurnal cycle mishandling are efficient without the need to include predictors targeting this specific issue. Out of the two post-processed forecasts, M-ecPoint shows a flatter pattern, and, therefore, including predictors targeting the misrepresentation of the diurnal cycle should allow M-ecPoint further to improve the prediction of nighttime rainfall for 12-hourly accumulations, as shown by Gascón et al. (2024) for 6-hourly rainfall.

Both post-processed systems improve ENS's ability to discriminate "wet" conditions (i.e. VRT = 10 mm/12h). While U-ecPoint and M-ecPoint show similar discrimination ability, M-ecPoint maintains significantly better reliability, as shown for VRT = 0.2 mm/12h.

For more extreme rainfall totals (VRT = 50 mm/12h), U-ecPoint shows marginally better discrimination ability than M-ecPoint according to $AROC_t$, caused by the last meaningful point in U-ecPoint "real" ROC curve being located above the last meaningful point for M-ecPoint. While this indicates that U-ecPoint provides more hits (i.e., identification of extreme rainfall events that did materialise), it does come at the cost of a higher number of false alarms (i.e., predictions of extreme rainfall events that did not materialise), which would inevitably have a deleterious effect on user trust due to the cry-wolf effect (LeClerc and Joslyn, 2015). Two factors suggest that U-ecPoint overestimates probabilities for extreme events compared to M-ecPoint. First, for low probability classes ($< 10\%$), where almost all the data resides, reliability diagrams show that U-ecPoint overforecasts observed rainfall probabilities by about a factor of two, corresponding to a clear false alarm signal. This means that 1% probability is more likely to be 0.5%, so the points on the ROC curve extend more towards the top right than they can for M-ecPoint, and ROC areas appear bigger than they might be. While this rewards $AROC_t$ (i.e., the measure used to estimate the discrimination ability of the "real" configuration of the forecasts), it penalises $AROC_z$ (i.e., the measure used to estimate the discrimination ability of the "potential" configuration with an infinite number of ensemble members), showing that, in reality, the discrimination ability of M-ecPoint is not worse than that for U-ecPoint. Additionally, U-ecPoint's overall reliability suffers to the point that its BS_{rel} score becomes worse than ENS. The case study also supports the overall U-ecPoint overprediction of extreme rainfall: compared to M-ecPoint, the 99^{th} percentile in U-ecPoint seems uniformly bigger across the domain. M-ecPoint provides, instead, a closer match to observations. These overall results should not surprise as U-ecPoint's single MF represents somewhat an average of the multiple MFs used in M-ecPoint. This suggests that U-ecPoint is applying multiplying factors that are too big when post-processing large rainfall totals (i.e., red bars in MFs).

Regarding the use of the binormal approximation in the computation of the ROC curves, there appear to

[4] M-ecPoint for 12-hourly rainfall does not contain predictors that aim at correction of the diurnal cycle mishandling by ENS.

be imperfections in the binormal approximation for extending the ROC curve for U-ecPoint: the dashed grey curve does not pass through point B as it probably should. This suggests $AROC_z$ for U-ecPoint is an overestimate. One should also question the validity of the binormal approximation for ENS for higher totals, given that ENS is not designed to capture higher totals attributable to sub-grid variability, no matter how many members there are. So, a proportion of hits at the top of the hit rate axis will only be attainable by chance. Because of this, if we had an infinite number of members, there would likely be a cusp in the curve, with the top right portion of the curve resembling a "no discrimination ability straight segment" rather than a curve. The binormal approximation does not capture the cusp and the subsequent straight segment, overestimating the area under the ROC curve. This is not an intrinsic limitation for ecPoint variants as they are designed to capture higher rainfall totals attributable to sub-grid variability.

The results of this study demonstrate that the multivariate variant of ecPoint (M-ecPoint) outperforms its univariate counterpart (U-ecPoint). M-ecPoint's enhanced performance is crucial for timely and accurate predictions of "dry" and "extremely wet" conditions, especially in advance of low-probability, high-impact rainfall events. It reduces false alarms, maintaining higher user confidence in the post-processed forecasts. While a formal cost-benefit analysis is beyond the scope of this study, our findings suggest that the improvements delivered by M-ecPoint justify its continued use, development, and long-term maintenance. However, further enhancements could streamline forecast production costs, balancing operational feasibility with accuracy and making the multivariate approach more attractive for long-term application. For example, future work should explore the optimization of M-ecPoint's complexity by investigating intermediate approaches between U-ecPoint and the current implementation, particularly examining the impact of using different numbers of predictors and leaves in the decision tree. Such analysis could help establish an optimal balance between model complexity and forecast skill improvement, potentially leading to a more computationally efficient implementation while maintaining forecast quality. Morever, the already streamlined calibration process could be further optimised by incorporating automatic modules to the currently semi-subjective approach, reducing required manpower when testing new predictors. The forecast production code could be optimised (e.g. through process parallelisation) to decrease runtime and facilitate the introduction of further refinements. Finally, a more streamlined verification framework could be integrated into the calibration process to conduct cost-benefit analyses and better assess whether new predictors or mapping functions provide sufficient improvement to justify their operational use.

Acknowledgments

The authors would like to thank Thomas Haiden for his insightful comments, which served as the catalyst for this study.

References

Bauer, P. (2024) What if? numerical weather prediction at the crossroads. *Journal of the European Meteorological Society*, **1**.

Bauer, P., Dueben, P. D., Hoefler, T., Quintino, T., Schulthess, T. C. and Wedi, N. P. (2021) The digital revolution of earth-system science. *Nature Computational Science*, **1**, 104–113.

Bauer, P., Hoefler, T., Stevens, B. and Hazeleger, W. (2024) Digital twins of earth and the computing challenge of human interaction. *Nature Computational Science*, **4**, 154–157.

Bechtold, P., Semane, N., Lopez, P., Chaboureau, J. P., Beljaars, A. and Bormann, N. (2014) Representing equilibrium and nonequilibrium convection in large-scale models. *Journal of the Atmospheric Sciences*, **71**, 734–753.

Bouallegue, Z. B., Haiden, T., Weber, N. J., Hamill, T. M. and Richardson, D. S. (2020) Accounting for representativeness in the verification of ensemble precipitation forecasts. *Monthly Weather Review*, **148**, 2049–2062.

Bouallègue, Z. B. and Richardson, D. S. (2022) On the roc area of ensemble forecasts for rare events. *Weather and Forecasting*, **37**, 787–796.

Buizza, R. (2019) Introduction to the special issue on "25 years of ensemble forecasting". *Quarterly Journal of the Royal Meteorological Society*, **145**, 1–11.

Cafaro, C., Woodhams, B. J., Stein, T. H., Birch, C. E., Webster, S., Bain, C. L., Hartley, A., Clarke, S., Ferrett, S. and Hill, P. (2021) Do convection-permitting ensembles lead to more skillful short-range probabilistic rainfall forecasts over tropical east africa? *Weather and Forecasting*, **36**, 697–716.

Casaretto, G., Dillon, M. E., Salio, P., Skaba, Y. G., Nesbitt, S. W., Schumacher, R. S., García, C. M. and Catalini, C. (2022) High-resolution nwp forecast precipitation comparison over complex terrain of the sierras de córdoba during relampago-cacti. *Weather and Forecasting*, **37**, 241–266.

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

Demuth, J. L., Morss, R. E., Jankov, I., Alcott, T. I., Alexander, C. R., Nietfeld, D., Jensen, T. L., Novak, D. R. and Benjamin, S. G. (2020) Recommendations for developing useful and usable convection-allowing model ensemble information for nws forecasters. *Weather and Forecasting*.

Feijoó, M. and Solman, S. (2022) Convection-permitting modelling strategies for simulating extreme rainfall events over southeastern south america. *Climate Dynamics*, **59**, 2549–2569.

Ferro, C. A. and Fricker, T. E. (2012) A bias-corrected decomposition of the brier score. *Quarterly Journal of the Royal Meteorological Society*, **138**, 1954–1960.

Frogner, I. L., Singleton, A. T., Køltzow, M. and Andrae, U. (2019) Convection-permitting ensembles: Challenges related to their design and use. *Quarterly Journal of the Royal Meteorological Society*, **145**, 90–106.

Gascón, E., Montani, A. and Hewson, T. D. (2024) Post-processing output from ensembles with and without parametrised convection, to create accurate, blended, high-fidelity rainfall forecasts. *Quarterly Journal of the Royal Meteorological Society*, **150**, 3117–3145.

Glahn, H. R. and Lowry, D. A. (1972) The use of model output statistics (mos) in objective weather forecasting. *Journal of applied meteorology and climatology*, **11**, 1203–1211.

Gneiting, T. and Vogel, P. (2022) Receiver operating characteristic (roc) curves: equivalences, beta model, and minimum distance estimation. *Machine Learning*, **111**, 2147–2159.

Göber, M., Zsótér, E. and Richardson, D. S. (2008) Could a perfect model ever satisfy a naïve forecaster? on grid box mean versus point verification. *Meteorological Applications*, **15**, 359–365.

Haiden, T. and Duffy, S. (2016) Use of high-density observations in precipitation verification. *ECMWF Newsletter*, 20–25.

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

Harvey, L. O., Hammond, K. R., Lusk, C. M. and Mross, E. F. (1992) The application of signal detection theory to weather forecasting behavior. *Monthly Weather Review*, **120**, 863–883.

Hemri, S., Hewson, T., Gascón, E., Rajczak, J., Bhend, J., Spirig, C., Moret, L. and Liniger, M. A. (2022) How do ecpoint precipitation forecasts compare with postprocessed multi-model ensemble predictions over switzerland? *ECMWF Technical Memoranda*, **901**.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K. and Haiden, T. (2014) Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, **41**, 9197–9205.

Hewson, T. and Pillosu, F. (2021) A low-cost post-processing technique improves weather forecasts around the world. *Communications Earth and Environment*, **2**.

Hewson, T., Pillosu, F., Gascòn, E. and Vučković, M. (2023) Post-processing era5 output with ecpoint. *ECMWF Newsletter*.

Janjic, T., Bormann, N., Bocquet, M., Carton, J. A., Cohn, S. E., Dance, S. L., Losa, S. N., Nichols, N. K., Potthast, R., Waller, J. A. and Weston, P. (2018) On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **144**, 1257–1278.

Keil, C., Heinlein, F. and Craig, G. C. (2014) The convective adjustment time-scale as indicator of predictability of convective precipitation. *Quarterly Journal of the Royal Meteorological Society*, **140**, 480–490.

Klein, W. H., Lewis, B. M. and Enger, I. (1959) Objective prediction of five-day mean temperatures during winter. *Journal of the Atmospheric Sciences*, **16**, 672–682.

Lavers, D. A., Harrigan, S. and Prudhomme, C. (2021) Precipitation biases in the ecmwf integrated forecasting system. *Journal of Hydrometeorology*, **22**, 1187–1198.

LeClerc, J. and Joslyn, S. (2015) The cry wolf effect and weather-related decision making. *Risk Analysis*, **35**, 385–395.

Mamgain, A., Prasad, S. K., Sarkar, A., Shanker, G., Dube, A. and Mitra, A. K. (2024) Evaluating short-range forecasts of a 12 km global ensemble prediction system and a 4 km convection-permitting regional ensemble prediction system. *Pure and Applied Geophysics*, **181**, 2217–2241.

Owens, R. and Hewson, T. (2018) Ecmwf forecast user guide. URL: https://www.ecmwf.int/node/16559.

Roberts, N., Ayliffe, B., Evans, G., Moseley, S., Rust, F., Sandford, C., Trzeciak, T., Abernethy, P., Beard, L., Crosswaite, N., Fitzpatrick, B., Flowerdew, J., Gale, T., Holly, L., Hopkinson, A., Hurst, K., Jackson, S., Jones, C., Mylne, K., Sampson, C., Sharpe, M., Wright, B., Backhouse, S., Baker, M., Brierley, D., Booton, A., Bysouth, C., Coulson, R., Coultas, S., Crocker, R., Harbord, R., Howard, K., Hughes, T., Mittermaier, M., Petch, J., Pillinger, T., Smart, V., Smith, E. and Worsfold, M. (2023) Improver: The new probabilistic postprocessing system at the met office. *Bulletin of the American Meteorological Society*, **104**, E680–E697.

Roberts, N. M. and Lean, H. W. (2008) Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, **136**, 78–97.

Trotta, B., Owen, B., Liu, J., Weymouth, G., Gale, T., Hume, T., Schubert, A., Canvin, J., Mentiplay, D., Whelan, J. and Johnson, R. (2024) Rainforests: a machine-learning approach to calibrating nwp precipitation forecasts. *Weather and Forecasting*.

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Bouallègue, Z. B., Bhend, J., Dabernig, M., de Cruz, L., Hieta, L., Mestre, O., Moret, L., Plenković, I. O., Schmeits, M., Taillardat, M., van den Bergh, J., van Schaeybroeck, B., Whan, K. and Ylhaisi, J. (2021) Statistical postprocessing for weather forecasts review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, **102**, E681–E699.

Vannitsem, S., Wilks, D. S. and Messner, J. W. (2018) *Statistical Postprocessing of Ensemble Forecasts*.

Wilks, D. S. (2019) *Statistical Methods in Atmoshperic Sciences*. Elsevier Inc., fourth edi edn.

Zeman, C., Wedi, N. P., Dueben, P. D., Ban, N. and Schär, C. (2021) Model intercomparison of cosmo 5.0 and ifs 45r1 at kilometer-scale grid spacing. *Geoscientific Model Development [preprint]*, **2021**, 1–35.

Zorzetto, E., Canale, A. and Marani, M. (2024) A bayesian non-asymptotic extreme value model for daily rainfall data. *Journal of Hydrology*, **628**.

Şan, M., Nacar, S., Kankal, M. and Bayram, A. (2023) Daily precipitation performances of regression-based statistical downscaling models in a basin with mountain and semi-arid climates. *Stochastic Environmental Research and Risk Assessment*, **37**, 1431–1455.