Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

On the ROC area of ensemble forecasts for rare events*

Zied Ben Bouallègue[a] and David S. Richardson[a,b]

[a] *European Centre for Medium-Range Weather Forecasts, Reading, UK*
[b] *Department of Geography and Environmental Science, University of Reading, Reading, UK*

November 25, 2021

Abstract: The relative operating characteristic (ROC) curve is a popular diagnostic tool in forecast verification, with the area under the ROC curve (AUC) used as a verification metric measuring the discrimination ability of a forecast. Along with calibration, discrimination is deemed as a fundamental probabilistic forecast attribute. In particular, in ensemble forecast verification, AUC provides a basis for the comparison of potential predictive skill of competing forecasts. While this approach is straightforward when dealing with forecasts of common events (e.g. probability of precipitation), the AUC interpretation can turn out to be oversimplistic or misleading when focusing on rare events (e.g. precipitation exceeding some warning criterion). How should we interpret AUC of ensemble forecasts when focusing on rare events? How can changes in the way probability forecasts are derived from the ensemble forecast affect AUC results? How can we detect a genuine improvement in terms of predictive skill? Based on verification experiments, a critical eye is cast on the AUC interpretation to answer these questions. As well as the traditional trapezoidal approximation and the well-known bi-normal fitting model, we discuss a new approach which embraces the concept of imprecise probabilities and relies on the subdivision of the lowest ensemble probability category.

Key words: ROC area, discrimination, verification artefacts, trapezoidal approximation, bi-normal model, imprecise probability.

1 Introduction

Over the past decades, the popularity of the relative operating characteristic (ROC) curve has steadily increased with applications in numerous fields (Gneiting and Vogel, 2018). In meteorology, verification of weather forecasts based on signal detection theory has been in usage since the seminal works of Mason (1982) and Harvey et al. (1992), and recommended as a standard verification tool by the Word Meteorological Organisation in Stanski et al. (1989). In the framework of probabilistic forecast verification, the area under the ROC curve (AUC) is often used as a summary measure of forecast discrimination. Discrimination is the ability to distinguish between event and non-event and, along with calibration, it is one of the key

*This work has been submitted to *Weather and Forecasting*. Copyright in this work may be transferred without further notice.

© 2021 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

attributes of a probabilistic forecast (Murphy, 1991). While calibration deals with the meaning of probabilities (its estimation is an attempt to measure whether taking a forecast at face value is an optimal strategy), discrimination appraises the existence of a signal in the forecast when an event materialises and its absence in the opposite situation.

Practically, the ROC plots the hit rate (HR) versus the false alarm rate (FAR) of an event for incremental decision thresholds. Examples are provided in Fig. 1 for probability forecasts derived from the 50-member ensemble run at the European Centre for Medium-Range Weather Forecasts (see more details about the data in Section 2). Corresponding probability fields for February 15, 2021 over the British Isles, are shown in Fig. 2. The targeted events correspond to precipitation exceeding the following thresholds: 1, 20, and 50 mm/24h. A ROC curve is defined by the line joining successive ROC points, where each point corresponds to results for increasing decision thresholds, from the top right to the bottom left corners of the plot. When the decision variable is the number of members exceeding the event-threshold (interpreted as a raw probability forecast), the issued forecast can take values in $[0, 1/M, 2/M, ..., 1]$ for an ensemble of size M. As a consequence, the resulting ROC curve is based on (up to) M+1 points.

The area under the straight lines formed by connecting the $M+1$ points (including the (0,0) and the (1,1) points) of the ROC plot correspond to the AUC with the so-called trapezoidal approximation (T-AUC). This nomenclature comes from the fact that the area is estimated considering straight lines between two consecutive points of the plot and so as a sum of trapeziums. Interestingly, T-AUC is equivalent to the result of a two alternative forced choice (2AFC) test for dichotomous events (Mason and Weigel, 2009). The 2AFC test consists in checking whether, for 2 different observations in a verification sample, one event and one non-event, the forecast associated with the former is larger than the forecast associated with the later (provided that the decision variable is oriented so that large implies more likely). If we denote x_1 a forecast when the event occurs and x_0 a forecast issued when the event does not materialise, the scoring function associated with the test is:

$$S(x_1, x_0) = \begin{cases} 0 & \text{if} \quad x_1 < x_0 \\ 1 & \text{if} \quad x_1 > x_0 \\ 0.5 & \text{if} \quad x_1 = x_0 \end{cases} \tag{1}$$

The result of the test is the average score over all event/non-event pairs in the verification sample. In Eq. (1), the test returns a value of 0.5 when the forecasts are indistinguishable, that is, in our examples, when two probability forecasts are identical. An average score of 0.5 is also the expected mean result for a random forecast.

For rare events, there is "a tendency for the points on the ROC to cluster toward the lower left corner of the graph" as noted by Casati et al. (2008) and illustrated in Fig. 1. When computing T-AUC, a straight line is drawn between the last meaningful point on the ROC curve and the top-right corner to close the ROC curve, giving the impression that part of the curve is missing. How much of the curve is "missing"

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

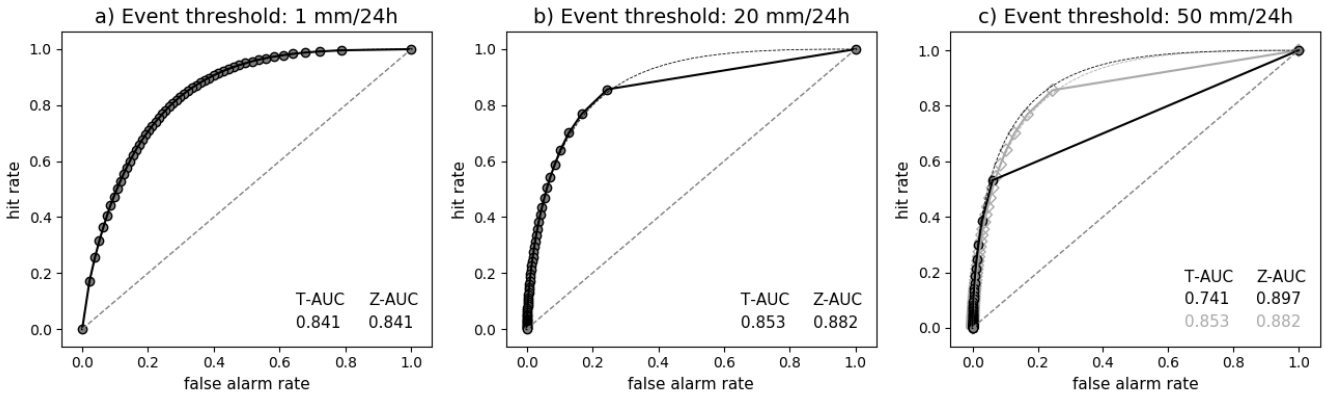

Figure 1: Examples of ROC curves for common and rare events. ROC curve and corresponding AUCs for precipitation forecasts with event-thresholds: a) 1mm/24h, b) 20 mm/24h, and c) 50 mm/24h. In grey, the results obtained when using the probability of exceeding 20 mm/24h for predicting the occurrence of precipitation exceeding 50 mm/24h.

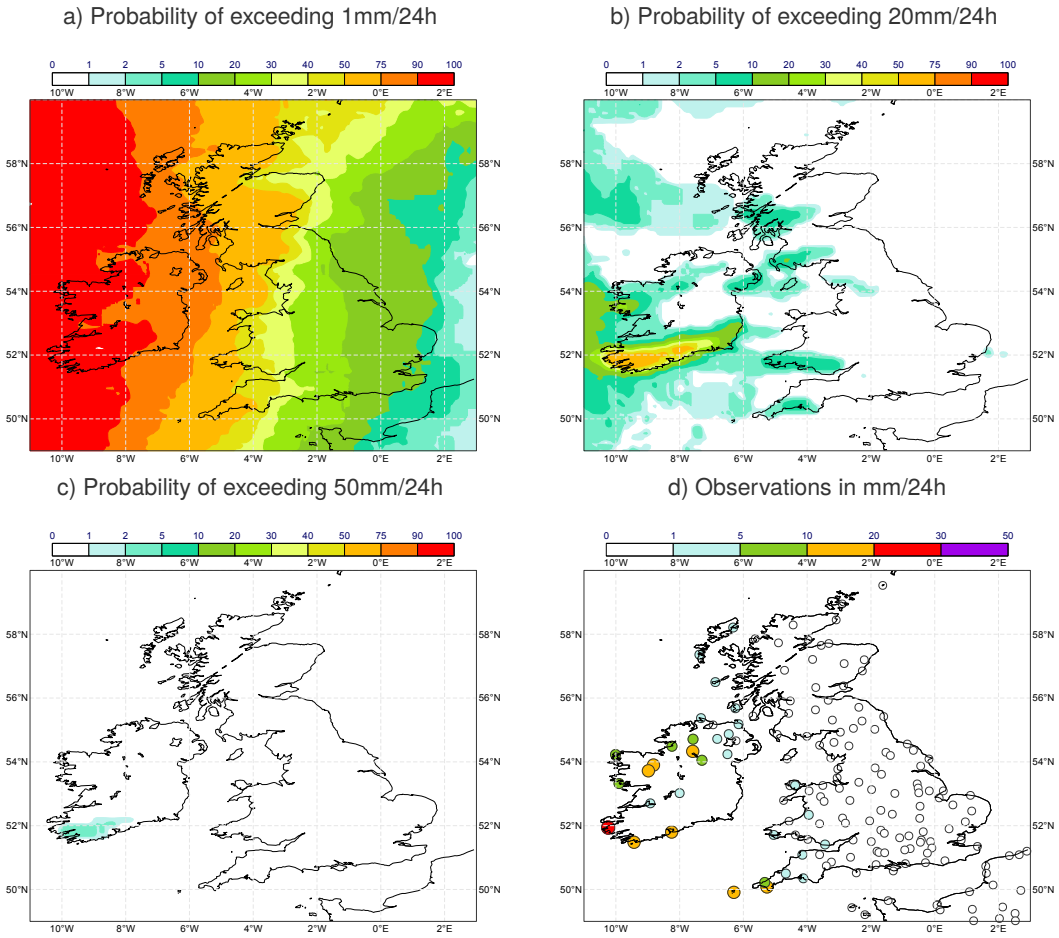

Figure 2: Storm Karim and probabilistic precipitation forecasts valid on 15 February 2021 over the British Isles. Probability of precipitation derived from a 50-member ensemble, 5 days in the lead time. Probability of exceeding a) 1 mm/24h, b) 20 mm/24h, and c) 50 mm /24h, and d) corresponding observations at synoptic stations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

depends on the lowest category[1], defined here by the ensemble size[2], and the base-rate of the event. As a rule of thumb, half of the ROC curve (the upper-right part) is missing when assessing the performance of an ensemble of size M focusing on an event with a base-rate $\alpha = \frac{1}{M}$. This rule is valid when the probability forecast is close to calibration, that is when the probability can be taken at face value for decision making. The rule is derived from the relationship between an optimal decision threshold, the slope of the line tangent to the ROC curve, and the event-base rate (see for example Eq. 25 in Ben Bouallègue et al. (2015)).

In order to draw a "full" ROC curve, one can apply the so-called bi-normal model (Harvey et al., 1992; Wilson, 2000; Atger, 2004)[3]. The fitting of the ROC curve with the bi-normal model is based on the assumption that HR and FAR are integrations of normal distributions, a signal and a noise distribution respectively. A closed-form for the computation of the AUC exists (see Eqs (2) and (3) in Harvey et al., 1992). The fitting of the HR and FAR requires a Z-transformation based on the unit normal distribution. For this reason, the resulting AUC is denoted here Z-AUC. When applied to ensemble-derived probability forecasts for rare events, this approach consists effectively in an extrapolation to a hypothetical continuous decision variable based on the limited set of decision thresholds materially assessable. Because such a decision variable may not be achievable in practice, Z-AUC is sometime considered as a measure of the potential discrimination that could be achieved, for example "for an unlimited ensemble size" (Bowler et al., 2006).

T-AUC and Z-AUC summary metrics can provide very different comparative results. The statistics reported in Fig. 1(c) are striking in that respect. In grey, we report the verification results obtained when using the probability of exceeding 20 mm/24h to predict the occurrence of precipitation exceeding 50 mm/24h. On the one hand, Z-AUC is (slightly) smaller than for the original forecast (in black): the interpretation is that the probability forecast for 50 mm/24h is potentially more informative than the probability forecast for 20 mm/24h when we focus on the higher event-threshold. On the other hand, T-AUC statistics point towards a larger predictive skill of the low event-threshold probability forecast practically users may benefit more from using the lower threshold unless additional post-processing is carried out to realise the potential additional benefit from using the higher threshold implied by the Z-AUC. As illustrated in Fig. 2, in many cases no ensemble member exceeds the high event-threshold. The small proportion of distinguishable forecasts explains the poor results of the original forecast in terms of T-AUC: the discrimination ability of two forecasts with the same value is equivalent to the discrimination ability of two random forecasts (see Eq. 1).

When verifying ensemble forecasts focusing on rare events, the AUC users face a dilemma: should they use T-AUC that relies on a clustering of points in the bottom left corner of the ROC plot, or should they use Z-AUC that extrapolates the results to compute scores based on a "full" ROC curve? The user's preference depends on the scientific question at hand, and in particular on whether the practical usefulness or the intrinsic information content of the ensemble forecast is the key aspect to be assessed. AUC assesses the discrimination ability of a decision variable, so special attention should be paid on how this decision variable

[1] more generally the conditional probability of the event given the lowest value of the discrimination variable

[2] the ensemble members are equally probable in our example

[3] more recently, Gneiting and Vogel (2018) introduce a flexible two-parameter beta family for fitting empirical ROC curves. This approach is not investigated or discussed further in our study.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

is derived from the ensemble forecast. A decision variable defined as the number of ensemble members exceeding a threshold can appear appropriate for common events but may be less useful when forecasting rare events as illustrated in Fig. 1.

Aiming at bridging the gap between T-AUC and Z-AUC results, we propose a new approach to enhance the probability forecast derived from an ensemble when focusing on rare events. Our approach is inspired by a suggestion in Casati et al. (2008): "one solution to this problem [the clustering of the ROC points] is to subdivide the lowest-valued forecast probability bins. The verification sample can usually support subdividing the lower-valued probability bins when fitting the ROC for low base-rates." In practice, the counting of the number of members exceeding an event-threshold provides a forecast interpreted as an imprecise probability (IP), a probability over an interval. A refinement of the raw probability forecast on that interval is conducted using additional information from the ensemble itself to better distinguish between different levels of low chances of event occurrence. In particular, we show how to use the ensemble mean (EM) as a "secondary" decision variable in this process. In the following, AUC estimated with this approach is referred to as IP/EM-AUC.

Having in mind the key question "How to interpret AUC results of ensemble forecasts when focusing on rare events?", we design a series of verification experiments in order to analyse T-AUC and Z-AUC in context. The verification experiments are chosen to show:

I. the loss of predictability with forecast lead time,

II. the impact of a post-processing step which accounts for subgrid-variability,

III. the impact of increasing the forecast probability categories,

IV. how to isolate the ensemble size effect with the help of a parametric model,

V. the impact of subdividing the lowest category with the help of an ensemble summary statistic.

The verification experiments and derived results are described and presented in Section 2 before drawing recommendations in Section 3.

2 Verification experiments

2.1 Verification setup

2.1.1 Dataset

Forecasts of daily precipitation are used in the following verification experiments, but similar qualitative results can be obtained with other accumulation periods or weather variables. The probability forecasts are derived from the ensemble prediction system run operationally at the ECMWF. The interpretation of the 50-member ensemble in terms of probability follows a simple (but common) approach. It consists in counting the number of members exceeding a threshold. Observation measurements at surface synoptic observation stations over the globe are compared with forecasts at the nearest grid-point over a verification

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

period running from 1 September 2019 to 31 August 2020.

Probability derived from ensemble forecasts are interpreted as imprecise probabilities as we are in a situation where the source of probabilistic information is incomplete and imperfect (Bradley, 2019). For example, when no member exceeds the event-thresholds of interest, the derived probability belongs to a probability interval close to 0. A ranking of the probability forecasts in that category can however be expressed with the help of an ensemble summary statistic such as the ensemble mean or an ensemble quantile. For illustration purposes, the ensemble mean is chosen as a secondary decision variable which is used to refine the ensemble interpretation for the lowest probability interval. Other choices can be valid as well and further research is encouraged in order to determine if an optimal summary statistic as a secondary decision variable exists in such a context.

Statistical post-processing of precipitation forecasts is also envisaged here and tested applying a parametric approach, that is relying on a pre-defined type of probability distribution. In the following, censored shifted gamma distributions are used to describe appropriately precipitation forecast distributions (a detailed description of the statistical method can be found in Ben Bouallègue et al., 2020). Post-processing aims here at correcting for the scale mismatch between forecasts (as model outputs on a grid) and observations (as point measurements at stations). We follow a so-called "perturbed ensemble approach" which consists in adding uncertainty to the forecast in order to represent the larger uncertainty at a finer scale. Practically, random perturbations drawn from a parametric distribution are added to each ensemble member.

Other ensemble post-processing techniques and their impact on the ROC are not investigated here. Techniques such as the neighbourhood method or the use of lagged ensemble lead to an increase of the effective ensemble size at much lower computational cost than running additional ensemble members (Ben Bouallègue et al., 2013). The availability of more members allows in any case a finer probability discretisation. In general terms, the impact of discretisation is discussed below along Experiment III.

2.1.2 Verification methodology

In our experiments, the central step of the verification process consists in populating contingency tables. The 2x2 tables are the raw material for generating:

- ROC curves (Mason, 1982),
- performance diagrams (Roebber, 2009),
- potential economic value (PEV) plots (Richardson, 2000, 2011).

Contingency tables are populated for incremental decision thresholds. In the traditional ensemble verification case, the decision variable is the number of members exceeding the event-threshold (e.g. 50mm/24h) and the number of decision thresholds is M+1 with M the ensemble size. In our experiment dealing with imprecise probabilities, the "zero category" is subdivided in 14 additional sub-categories using the ensemble mean as a secondary decision variable with the following (secondary) decision thresholds: [0.1, 0.2, 0.5, 1,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

2, 3, 4, 5., 6, 7, 8, 10, 15] in mm/24h.

So in practice, the following is performed:

1. choose an event-threshold,
2. count the number of members exceeding this event-threshold (this decision variable is called raw probability forecast),
3. if no members exceed the event-threshold, compute the ensemble mean (EM) and count the number k out of K cases for which EM is greater than each of the K secondary decision thresholds.
4. adjust the raw decision variable by considering a probability of $\frac{k}{M(K+1)}$ (rather than 0) for that forecast,
5. derive a contingency table for all distinct probability values in the forecast sample.

The trapezoidal approximation applied to the corresponding set of (HR,FAR) pairs correspond to IP/EM-AUC. This acronym reflects that the ensemble mean is used as a secondary decision variable.
Scores are aggregated over different domains, but mainly results for the globe are shown. Event thresholds are defined in absolute terms with a focus on an exceeding event threshold of 50 mm/24h. In the global verification sample, this event has a base-rate of around 1%, but with more occurrence in certain regions of the world than others. Hamill and Juras (2006) recommend the use of thresholds expressed in relative terms (quantile of a climatology) in order to avoid over-interpretation of the AUC results by mixing the forecast ability to distinguish between wet and dry regions and genuine predictive skill. Results for precipitation exceeding the 99^{th} percentile of the local climatology is also discussed as a final example.

2.2 Verification experiments results

Each of the following figures (Figs 3 to 7) comprises 3 plots: a performance diagram, a ROC plot, and a potential economic value plot (with a log scale on the x-axis). AUC estimates are indicated for both the trapezoidal approach (T-AUC) and the bi-normal method (Z-AUC). For each figure, two different sets of forecasts are compared. An asterisk indicates which of the 2 sets of forecasts (the red or the blue) has the best score in terms of T-AUC and Z-AUC. The relative superiority of the best forecast is indicated in percent. For all plots (except for experiment IV in Fig. 6), the results in red correspond to the raw ensemble-derived probability forecasts at day 5.

2.2.1 Experiment I: the impact of the forecast lead time

The first experiment compares the performance of forecasts at two different lead times. This experiment illustrates how different verification tools are impacted by a genuine change in forecast predictive skill. Fig. 3 illustrates the visual impact one can expect to see in such a situation where a difference in verification results can be explained only by a difference in predictive skill.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

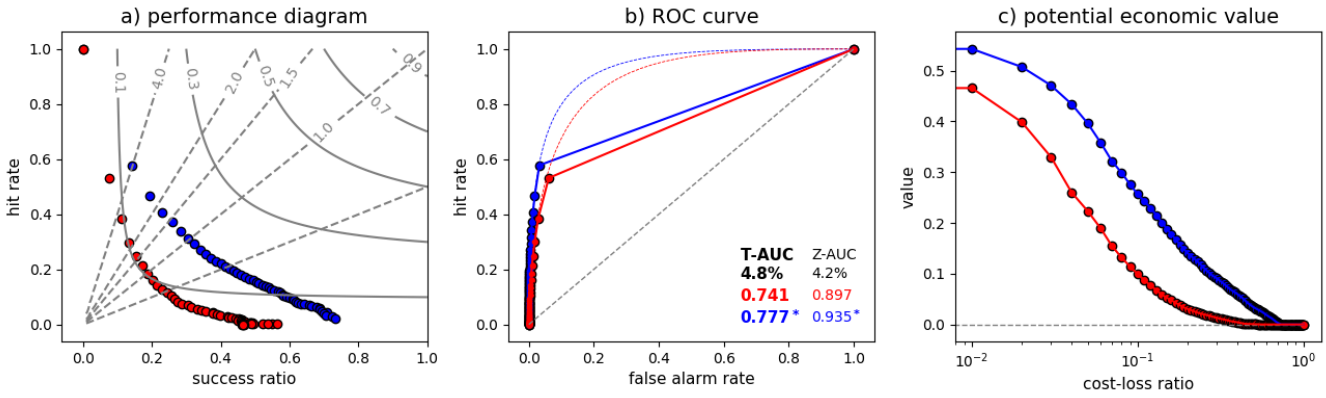

Figure 3: Results of **experiment I**, the benchmark experiment comparing forecast performance at day 1 (blue) and at day 5 (red).

In Fig. 3(a), the performance diagram shows the blue points (day 1 results) lie closer to the top right corner than the red ones (day 5 results). In Fig. 3(b), on the ROC plot, the blue points are distinct from the red ones and are closer to the top left corner. In Fig. 3(c), on the PEV plot, the blue curve lies above the red one, except for cost-loss ratios close to 1 for which both short and medium range forecasts have no value. This first experiment provides typical results expected from an increase in forecast predictive skill and, as such, serves as a benchmark for the following experiments. In terms of summary metrics, the relative skill improvement between day 5 and day 1 is of the same order of magnitude for T-AUC and Z-AUC, with 4.8% and 4.2% measured improvement, respectively.

2.2.2 Experiment II: the impact of a post-processing step

In this experiment, ensemble post-processing is applied in order to account for the scale mismatch between forecasts and observations. The method can be applied in a verification context to account for observation uncertainty, but also in a post-processing context to provide a forecast valid at any point within a model grid-box. The post-processed forecast is derived by adding a random perturbation independently to each member. The random perturbation is draw from a distribution described by the value of the forecast member (i.e. the predicted grid-scale precipitation) and with fixed parameters for all forecasts (as in Ben Bouallègue et al., 2020). In other words, a constant piece of information is added in the process: the expected sub-grid scale uncertainty expressed as a function of the grid-scale precipitation value. The main impact on the forecast is a significant increase of the ensemble spread with, in particular, larger distributional tails of the post-processed ensemble distribution compared with the original one.

All 3 plots in Fig. 4 show that the blue (post-processed ensemble results) and red points (original ensemble results) are overlapping with the exception of a couple of points in each plot. In the case of the performance diagram and ROC curve, blue and red points belong to the same underlying curves. In the PEV plot, we see larger value for users with a cost-loss ratio smaller than or equal to 2%. In terms of Z-AUC, the results are identical before and after post-processing up to the second decimal indicating that the underlying forecast information content is not altered by the process. The information has increased in the sense that

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

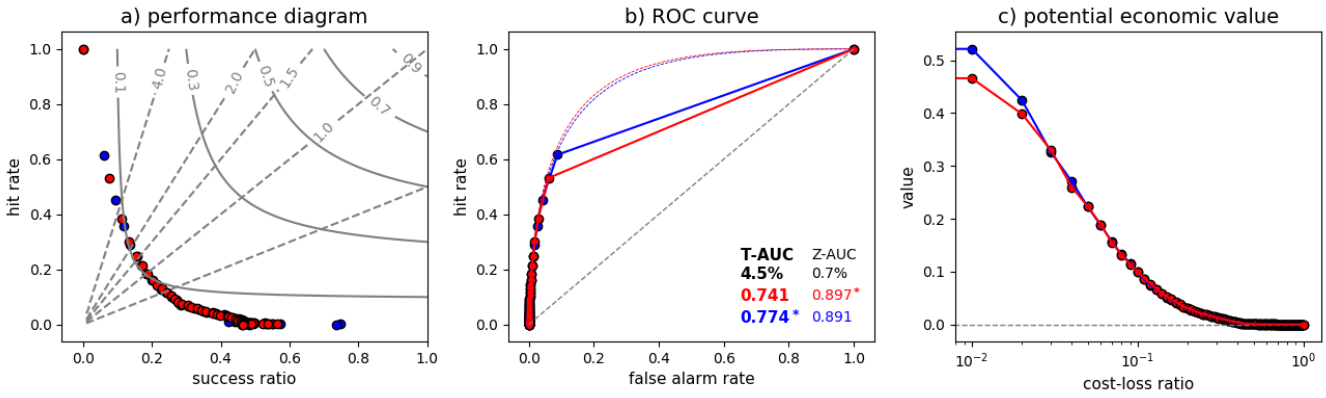

Figure 4: Results of **experiment II** illustrating the impact of post-processing. Results when accounting for representativeness (blue) compared with results for the corresponding raw day 5 forecasts (red).

information about the sub-grid uncertainty has been added to the forecast[4], but this bit of information is not different from one forecast to another. The improvement by post-processing as measured by T-AUC is 4.5%, close to the improvement measured between day 5 to day 1 discrimination ability (see Fig. 3(b)). While the latter is attributed to an improved predictability at shorter lead times, the former, the T-AUC improvement with post-processing, is attributed to a change in the frequency of forecast events: post-processed ensemble members exceed the event-threshold more often due to the larger spread (as in the example in Fig. 2(c)).

2.2.3 Experiment III: the impact of discretisation

The post-processing technique used in experiment II is based on a parametric approach. A random perturbation drawn from a parametric distribution is added to each member. The random draw is made from a distribution for which the parameters are known. Now, in experiment III, rather than a single perturbation for each member, we consider 2 random draws for each of the 50 raw ensemble members. So, the resulting post-processed ensemble has a size of 100. Let's recall that the ensemble perturbations are based on random draws from a distribution whose form depends only on the forecast value itself. The model for the precipitation sub-grid variability is a new but constant piece of information added to the forecast. Drawing additional random numbers from a parametric distribution better captures the form of the distribution but does not change the distribution itself or the quality of the underlying model.

Fig. 5 is similar to Fig. 4. The major difference between the 2 figures is the fact of a single point on each plot such as on the ROC plots with one point on the blue ROC curve closer to the top right corner when increasing the forecast discretisation. In terms of T-AUC, the change in the number of probability thresholds from 51 to 101 leads to a jump from 4.5% to 7.9% of improvement with respect to the original forecast. The larger the number of members describing the underlying forecast distribution, the better the decision variable defined as the number of members exceeding a threshold. In terms of Z-AUC, the doubling of the number of categories as part of the post-processing has no impact on performance. The underlying forecast distribution is the same, no additional information is provided that improves the forecast discrimination

[4]with a positive impact on the forecast calibration (not shonw).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

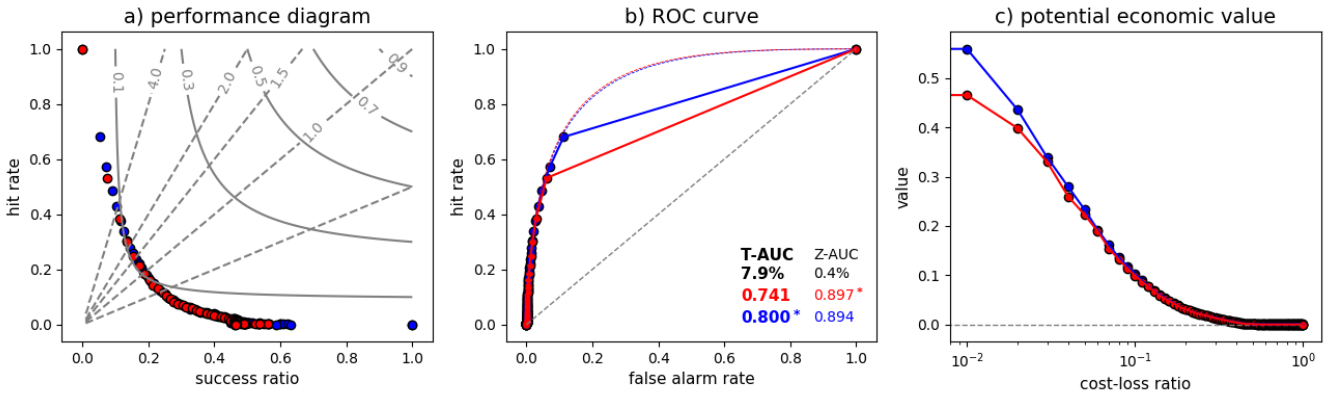

Figure 5: Results of **experiment III** illustrating the impact of discretisation. Results for 50 (red) and 100 (blue) post-processed members from a parametric model.

ability. With this experiment, we see that the choice regarding the categorisation of the probability forecast considerably influences the T-AUC results but not the Z-AUC ones.

2.2.4 Experiment IV: isolating the ensemble size effect

Here we again exploit the parametric nature of the post-processing approach we have followed. The goal here is to assess the ensemble-size effect on the forecast discrimination independently from the forecast discretisation effect. This experiment is designed to compare probability forecasts with the same discretisation (the same decision-thresholds) but derived from raw ensembles with different sizes. So, we distinguish raw ensemble size (the source of the forecast information) and the post-processed ensemble size which is an arbitrary choice in our setting. In this experiment, we compare raw ensembles of size 10 and 50, both post-processed in order to get 50 post-processed "members" (drawing 5 and 1 random perturbations for each member, respectively).

Fig. 6 shows the positive impact of increasing the raw ensemble size from 10 to 50 members. Derived probabilities from 50 post-processed forecasts (in both cases) exhibit better performance on the 3 plots. In particular, as expected, users with smaller cost-loss ratios) benefit most of the ensemble size increase. The discrimination ability as measured by T-AUC and Z-AUC shows an increase of 7.4% and 1.5%, respectively. These values can be compared with a 7.3% theoretical improvement of a proper score[5], the continuous ranked probability score, which encompasses both discrimination and calibration contributions.

2.2.5 Experiment V: subdividing the lowest category

In many cases, none of the ensemble members forecast a rare event, ie precipitation exceeding 50 mm/24h. The raw probability forecast is 0, but it is interpreted as an imprecise probability on an interval close to 0. Using a secondary decision variable, it is possible to distinguish between lower and higher chances within the pool of "0 probability" forecasts. In this experiment, we consider a forecaster that has access to the

[5]derived from Eq. 9 in Leutbecher (2019)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

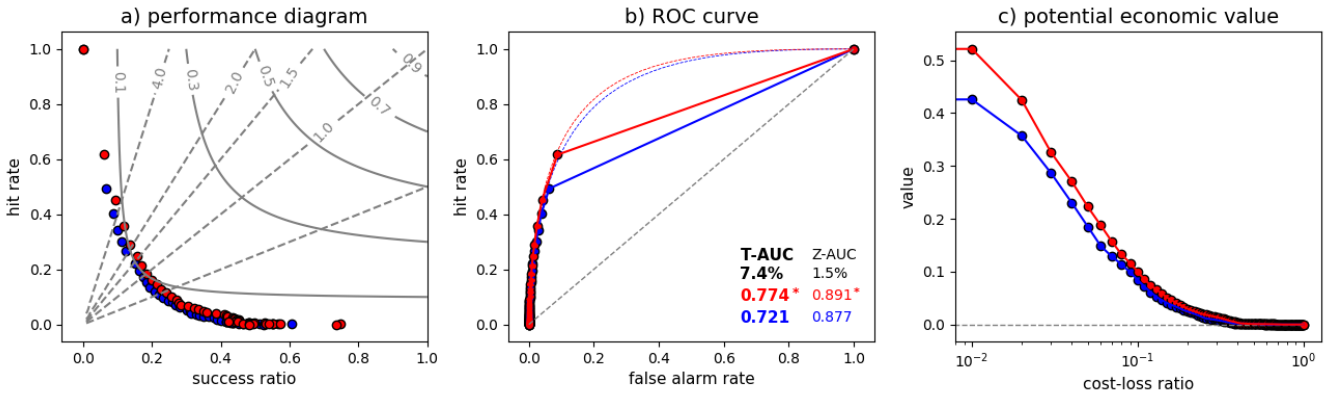

Figure 6: Results of **experiment IV** illustrating the expected impact of the ensemble size on verification results. Results for an ensemble of size 10 (red) and 50 (blue) post-processed to 50 members each. Note that the red points on this figure are the same as the blue points in Fig. 4.

ensemble mean in addition to the probability forecast itself. In the situation where a forecast probability of 0 is issued, the forecaster infers that the chance of occurrence of an event is higher when the ensemble mean is higher. Quantitatively, within an interval close to 0, a larger probability is assigned to a forecast with a larger ensemble mean (see methodology in Section 2.1.2). The verification results obtained with this ensemble interpretation are shown in Fig. 7 and compared with results obtained when using only the ensemble mean as a decision variable in Fig. 8.

Figs 7(a) and 7(b) show the continuity between the results based on the original data (red points) and the ones when subdividing the lowest probability category with the help of the ensemble mean (blue points). The so-called IP/EM-AUC corresponds to T-AUC computed using the blue points. In Fig. 7(c), PEV results diverge only for cost-loss ratio values smaller than 2%. These results share similarities with the ones obtained with post-processing in Fig. 4. When using the ensemble mean as a secondary decision variable, we obtain T-AUC and Z-AUC results that are (almost) identical, as a "full" ROC curve is now available based on the enhanced interpretation of the ensemble forecast. Interestingly, T-AUC with the lowest category subdivision is very close to the Z-AUC estimate for the original data. In other words, similar results can be obtained by extrapolation using Z-AUC or with T-AUC applied to an appropriate decision variable for rare events. On the plot in Figs 7 and 8, the blue points follow the red line derived with the binormal approximation, in one case they are above and in the other below. The results with the 2 approaches are consistent but subject to different level of uncertainty as discussed in Section 2.3.

The ensemble interpretation in two-steps with the help of a secondary decision variable refines the imprecise probabilities close to 0%. The ensemble mean forecast serves as a secondary decision variable in our example. This approach is different than using the ensemble mean as a unique decision variable as illustrated in Fig 8. Results here are produced using the 99% quantile of the local climatology as an event-threshold. This choice allows to better highlight the major benefit of integrating the ensemble distribution into the decision variable rather than using the ensemble mean only. Indeed, not only the ROC plot but also the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

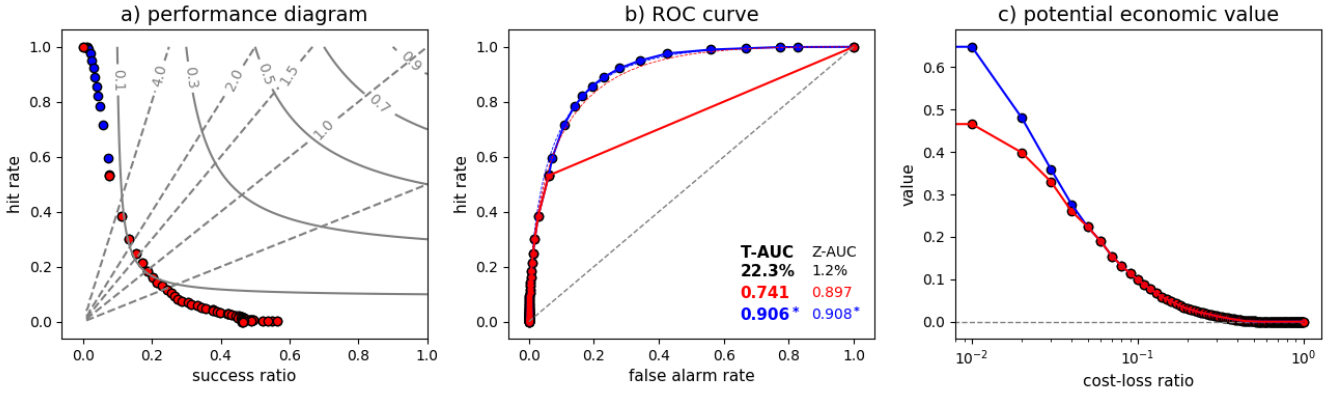

Figure 7: Results of **experiment V** showing the impact of subdividing the lowest category with the help of the ensemble mean (in blue) with respect to the original forecast results (in red).

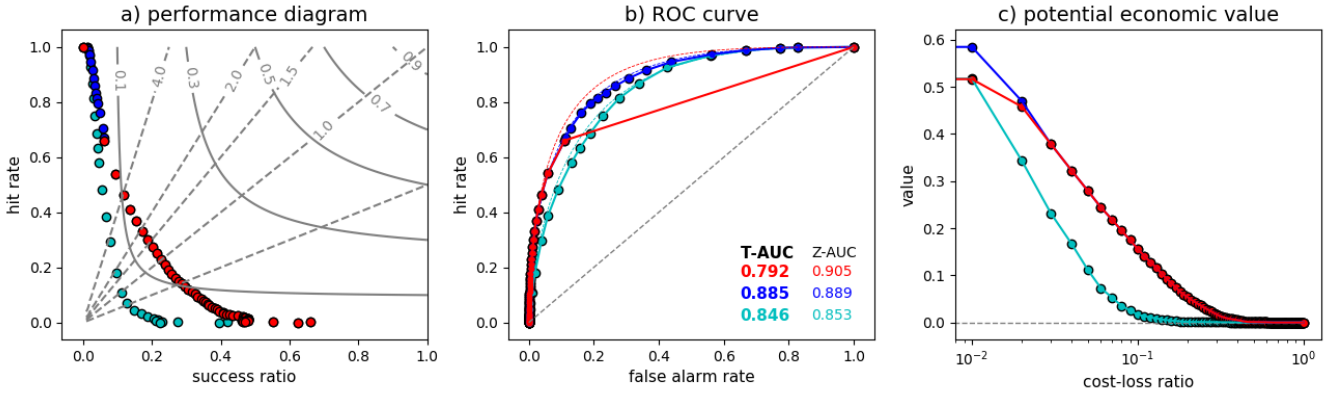

Figure 8: Results of **experiment V** as in Fig. 7, but using a threshold defined as the 99% of the local climatology. In addition to the results for raw probability (red) and when subdividing the lowest probability category with the ensemble mean (blue), we show results when using only the ensemble mean as a decision variable (cyan).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

performance diagram and potential economic value plot show the overall poor results of the ensemble mean (in cyan) compared with the probability forecasts (in red and blue).

Focusing on the T-AUC results in Fig. 8(b), we see that the ensemble mean appears as a better decision variable than the original ensemble interpretation as measured with the trapezoidal approximation. Because EM as a decision variable does not seems to performs nowhere better on the performance diagram (see Fig. 8(a)) or to benefit any user (see Fig. 8(c)), this result illustrates how misleading conclusions can be drawn when comparing T-AUC results from different "sources".

The ensemble size has an impact on both the raw probability and the ensemble mean estimates. The impact of the ensemble size on the discrimination ability as estimated when combining information from both ensemble aspects is not explored here.

2.3 Impact of the event base-rate

As a final investigation, we analyse AUC estimations and corresponding uncertainty as a function of the event base-rate. We consider 3 event-thresholds: 20, 40, and 50 mm /24 h, and build verification datasets for 4 different geographical domains (global, northern hemisphere, southern hemisphere, Europe) and 4 different seasons (autumn 2019, winter 2019, spring2020, and summer 2020). The event base-rate is different for each domain and threshold combination. The score uncertainty associated with each AUC estimation is derived as the inter-quantile range (5%-95%) of the score empirical bootstrapping distribution. Results are shown in Fig. 9 for T-AUC, Z-AUC, and IP/EM-AUC, the trapezoidal approximation when interpreting the ensemble in terms of imprecise probabilities and using the ensemble mean as secondary decision variable.

In Fig. 9(a), we observe a slight increase of the discrimination ability as a function of the rarity of the event, with the Z-AUC and IP/EM-AUC approaches displaying consistent estimates with respect to each other. However, a drop in discrimination is measured with T-AUC for event base-rates smaller than 3%. When applied to ensemble probability forecasts, T-AUC appears base-rate dependent. The drop is the result of the application of the trapezoidal approximation to raw probabilities derived from a non-infinite size ensemble: the AUC computation is confined to a smaller part of the full ROC curve as the event base-rate decreases. The drop in discrimination is also an indicator of when the ensemble interpretation into a decision variable by simply counting the number of members exceeding a threshold is no longer appropriate. The distinction between low probability of occurrence would require more ensemble members, or some sort of statistical post-processing, or simply to categorise low probability forecasts based on an additional ensemble summary statistic as for example the ensemble mean.

In Fig. 9(b), larger score uncertainty is seen for more extreme events. The increase in uncertainty is visible for all AUC methods, but T-AUC and Z-AUC estimates are more impacted than IP/EM-AUC ones. The bi-normal model exhibits the largest level of uncertainty for rare events because the original points are too close to get good estimates of the slope in the Z-transformed space, but the trapezoidal approximation is also subject to large variations. In practice, for event base-rates larger than 99%, the level of uncertainty appears

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

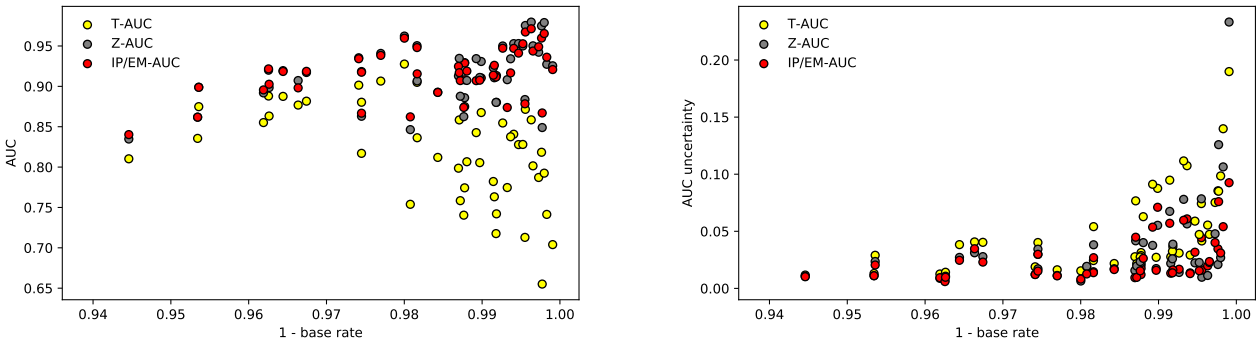

Figure 9: Area under the ROC curve (left panel) and corresponding uncertainty (right panel) estimated with the trapezoidal approximation (T-AUC, in yellow), a bi-normal model (Z-AUC, in grey), and interpreting the ensemble in terms of imprecise probabilities (IP/EM-AUC, in red) plotted as a function of the event base-rate. Each dot corresponds to the result for one domain, one season, and one threshold (see text).

too large to draw any useful conclusion with T-AUC or Z-AUC applied to raw probability forecast while the results for the enhanced probability using the ensemble mean as secondary decision variable appears more robust and reliable.

3 Recommendations

T-AUC and Z-AUC provide complementary information about the discrimination ability of a forecasting system. It is important to understand the differences between them and to use each appropriately depending on the question under investigation. Z-AUC measures the inherent discrimination ability of a forecasting system, while T-AUC measures how well this is achieved by a given implementation. Computing IP/EM-AUC shows a path to practically building a bridge between the two approaches.

Awareness of AUC properties is key in situations which combine 1) the assessment of probability forecasts derived from an ensemble and 2) a focus on rare events. Differences between T-AUC and Z-AUC are largest in such situations and it is especially important to interpret the results carefully. The recommendations below are targeted to this specific type of situation. In other cases (when focusing on more frequent events or when assessing probability forecasts derived from a parametric distribution for example), the different approaches for computing AUC converge to identical results as illustrated for instance in Fig. 1(a).

T-AUC, the AUC estimation with the trapezoidal approximation, is the traditional way to measure forecast discrimination. As illustrated, differences in T-AUC can be attributed to multiple sources, for example the level of forecast discretisation, the presence of ensemble forecast biases, or the way probabilities are derived from the ensemble forecast. So, when using the empirical ROC and summarising performance using the T-AUC, it is important to consider that:

- comparing T-AUC results should be done carefully. For example, a positive difference in T-AUC should be scrutinized before being interpreted as an improvement in intrinsic discrimination ability

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

(or as an improvement of the forecast performance in a broader sense),

- T-AUC is a measure of the performance of the forecasts using a particular discretisation of a chosen decision variable. Therefore, both the discretisation and the decision variable should be clearly described,
- the maximum available discretisation should be used (eg each member rather than fixed percentage bins), to ensure the ROC is as complete as possible,
- a comparison of the practical benefits of two competing forecast configurations should follow the above approach, describing the decision variable and discretisation used for each configuration. However this approach should not be used to draw conclusion about the underlying discrimination ability of the different systems.
- a fair comparison for the underlying discrimination ability of different systems would rely on using the same decision variable and the same discretisation. A sanity check for no significant differences in the underlying forecast bias is also important.
- comparing competing forecasts with T-AUC can lead to misleading conclusions if the above is not accounted for,
- to evaluate the impact of changing discretisation T-AUC should be used (Z-AUC will not be sensitive to this).

Z-AUC, the AUC estimate with a bi-normal model, is an alternative to the traditional trapezoidal approach. Z-AUC is a measure of "potential" discrimination ability of a system, in the sense that the extrapolation of the performance with the curve fitting has no practical meaning in terms of forecast usefulness. The bi-normal model is based on the assumption that HR and FAR follow the characteristics of normality distributed parameters. This assumption is not tested in our experiments, but our examples show good fits between model and data. As a summary:

- results with Z-AUC are not sensitive to forecast discretisation and to simple ensemble post-processing, in contrast to the T-AUC results,
- Z-AUC should be used to compare the potential discrimination ability of different forecasting systems. It does not indicate how this can be realised, but it does show which system has the better underlying performance,
- Z-AUC is useful to compare for example ensemble forecasts with different number of members it gives a better indication of the skill that could be achieved if sufficient discretisation is available,
- Z-AUC cannot distinguish whether a chosen discretisation is sufficient or can be improved; instead it shows what would be achieved if a sufficient discretisation was available,
- computing both T-AUC and Z-AUC allows useful comparison of potential and actual discrimination ability,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

- If T-AUC and Z-AUC are similar, the ensemble interpretation allows a discretisation of the derived decision variable which is sufficient and there is not much to be gained from post-processing to generate a better interpretation or a finer discretisation,
- If T-AUC is lower than Z-AUC there is the potential to improve the forecast performance by a better ensemble interpretation. This will be most beneficial to users with low C/L and most likely to happen for rare events (especially where low probability categories are not well resolved in the forecast).

We have demonstrated two methods to improve the ensemble interpretation and thus increase the forecast discretisation in such situations. Post-processing to account for sub-gridscale variability introduces a continuous distribution and allows arbitrary fine discretisation. The choice of discretisation should be sufficient to generate the full ROC (as well as plotting the ROC, a simple way to check this is to compare the T-AUC and Z-AUC).

IP/EM-AUC refers to the AUC estimated with a new approach which involves subdividing the lowest probability category by ranking the forecasts in this category according to an ensemble summary statistic, here the ensemble mean. We showed that this approach can provide sufficient discretisation to generate a full ROC curve based directly on the ensemble forecasts. The score estimations using this method have been shown to be more robust than with the bi-normal model. They also represent the real skill of the forecast system since users can act on each of the discretisation categories, rather than the potential skill that is shown by the extrapolation using a statistical model. Using this method, we have shown that while Z-AUC is strictly only a measure of potential discrimination skill, it may actually be straightforward to achieve in practice.

Other implications of this interpretation in the context of ensemble forecast verification and post-processing are topics for future research.

References

Atger, F., 2004: Estimation of the reliability of ensemble-based probabilistic forecasts. *Quart. J. Roy. Meteor. Soc.*, **130**, 627–646, doi:10.1256/qj.03.23.

Ben Bouallègue, Z., T. Haiden, N. Weber, T. Hamill, and D. Richardson, 2020: Accounting for representativeness in the verification of ensemble precipitation forecasts. *Mon. Wea. Rev.*, **148 (5)**, 2049–2062, doi:10.1175/MWR-D-19-0323.1.

Ben Bouallègue, Z., P. Pinson , and P. Friederichs, 2015: Quantile forecast discrimination ability and value. *Quart. J. Roy. Meteor. Soc.*, **141**, 3415–3424, doi:10.1002/qj.2624.

Ben Bouallègue, Z., S. E. Theis, and C. Gebhardt, 2013: Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorologische Zeitschrift*, **22**, 49–59, doi:10.1127/0941-2948/2013/0374.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0535.v1

Bowler, N. E., C. E. Pierce, and A. W. Seed, 2006: Steps: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled nwp. *Quart. J. Roy. Meteor. Soc.*, **132 (620)**, 2127–2155, doi:10.1256/qj.04.100.

Bradley, S., 2019: Imprecise probabilities. *The Stanford Encyclopedia of Philosophy (Spring 2019 Edition), Edward N. Zalta (ed.)*, URL ⟨https://plato.stanford.edu/archives/spr2019/entries/imprecise-probabilities/⟩.

Casati, B., and Coauthors, 2008: Forecast verification: current status and future directions. *Met. Apps*, **15 (1)**, 3–18, doi:10.1002/met.52.

Gneiting, T., and P. Vogel, 2018: Receiver operating characteristic (roc) curves. *arXiv:1809.04808*.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: is it real or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, doi:doi.org/10.1256/qj.06.25.

Harvey, L. O., J. K. Hammond, C. Lusk, and E. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120**, 863–883, doi:10.1175/1520-0493(1992)120⟨0863:TAOSDT⟩2.0.CO;2.

Leutbecher, M., 2019: Ensemble size: How suboptimal is less than infinity? *Q. J. R. Meteorolog. Soc.*, **145 (Suppl. 1)**, 107–128, doi:10.1002/qj.3387.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Met. Mag.*, **30**, 291–303.

Mason, S. J., and A. P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev.*, **137**, 331–349, doi:10.1175/2008MWR2553.1.

Murphy, A. H., 1991: Forecast verification: its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590–1601, doi:10.1175/1520-0493(1991)119⟨1590:FVICAD⟩2.0.CO;2.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **126**, 649–667, doi:10.1002/qj.49712656313.

Richardson, D. S., 2011: Economic value and skill. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. T. Jolliffe, and D. B. Stephenson, Eds., John Wiley and Sons, 167–184.

Roebber, P. J., 2009: Visualizing Multiple Measures of Forecast Quality. *Wea. Forecasting*, **24 (2)**, 601–608, doi:10.1175/2008WAF2222159.1.

Stanski, H., L. Wilson, and W. Burrows, 1989: *A Survey of Common Verification Methods in Meteorology*. World Weather Watch Tech. Rep., WMO, Geneva 8, TD No. 358, 114 pp.

Wilson, L. J., 2000: Comments on "Probabilistic Predictions of Precipitation Using the ECMWF Ensemble Prediction System". *Wea. Forecasting*, **15 (3)**, 361–364, doi:10.1175/1520-0434(2000)015⟨0361:COPPOP⟩2.0.CO;2.