Background Information: Coefficient of Predictive Ability (CPA)

Why a new forecast performance measure?

For illustration, we consider week-2 precipitation hindcasts over Addis Ababa by the ECMWF ensemble prediction system. Daily CHIRPS data aggregated to 7-day accumulations is used as a ground truth for verification.

The left panel of Fig. 1 shows ensemble mean forecasts $x_1, \ldots, x_n$ for all hindcasts with initialization dates in August during the years from 1999 to 2018 plotted against the verifying observations $y_1, \ldots, y_n$ derived from the CHIRPS data set. To check whether forecasts at this lead time have any skill we typically calculate anomalies[1] $\tilde{x}_1, \ldots, \tilde{x}_n$ from their climatological mean and calculate the correlation with the observed anomalies $\tilde{y}_1, \ldots, \tilde{y}_n$

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} \tilde{y}_i^2}} \tag{1}$$

as illustrated in the right panel of Fig. 1. The anomalies of both forecasts and observations are scattered rather symmetrically around zero, and thus measuring their association through the formula given in (1) makes intuitive sense. The Pearson correlation coefficient (PCC) obtained this way suggests a small amount of positive association between forecasts and observations, i.e. a small amount of forecast skill.

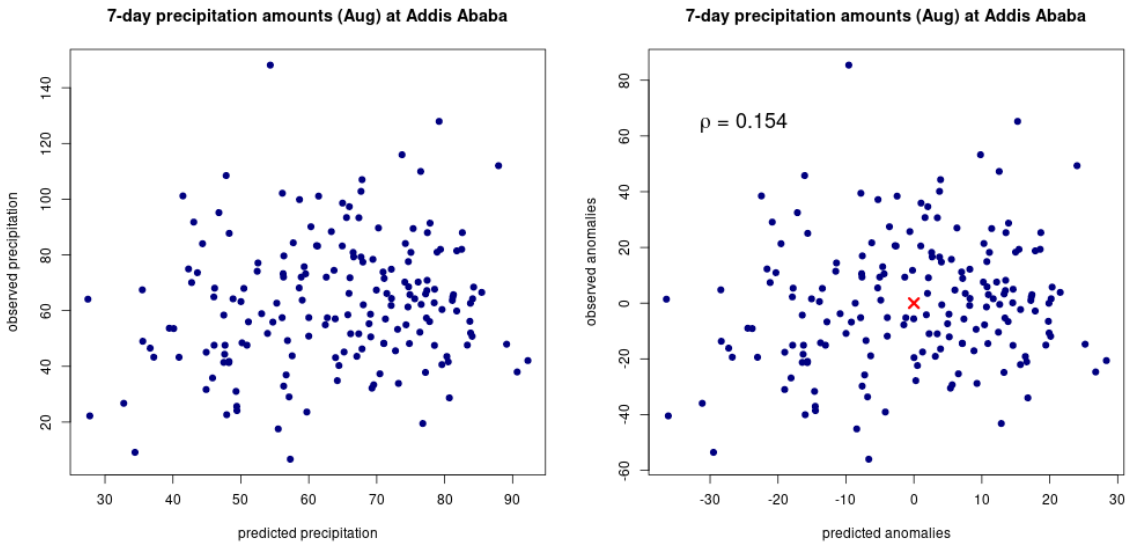

Figure 1: Ensemble mean forecasts of week-2 precipitation accumulations plotted against verifying observations (left panel), and the same plot for the respective forecast and observation anomalies (right panel).

Let's now consider the same example, but now with precipitation hindcasts initialized in October. The forecast valid times then fall into October and early November, and during this time of the year Addis Ababa receives much smaller amounts of precipitation. Many of the observed 7-day precipitation accumulations and even some of the predicted accumulations[2] are equal to zero, and the marginal distributions are much more left-skewed than in August, i.e. there are mostly small but also a few very large precipitation amounts. The strong asymmetry of the marginal distributions of forecast and observed anomalies around zero and the clustering of the zeros in the original data at some arbitrary negative value (see Fig. 2, right panel) raise some doubts as to whether the the formula given in (1) is still a good way to measure association of forecasts and observations.

This motivates the use of an alternative measure of forecast skill that can deal with quantities have duplicates and/or a skewed marginal distribution. The coefficient of predictive ability (CPA) is such a measure, that generalizes the area under the curve (AUC) metric used for binary outcomes to more general variables with real-valued outcomes. In the following we will briefly review the concept of receiver operating characteristic (ROC) curves and the associated AUC metric, and provide a definition and illustration of the CPA. For a more formal and detailed description we refer to [1].

[1] here by simply defining $\tilde{x}_j := x_j - \bar{x}, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

[2] which tend to get pulled way from zero due to the averaging over all ensemble members

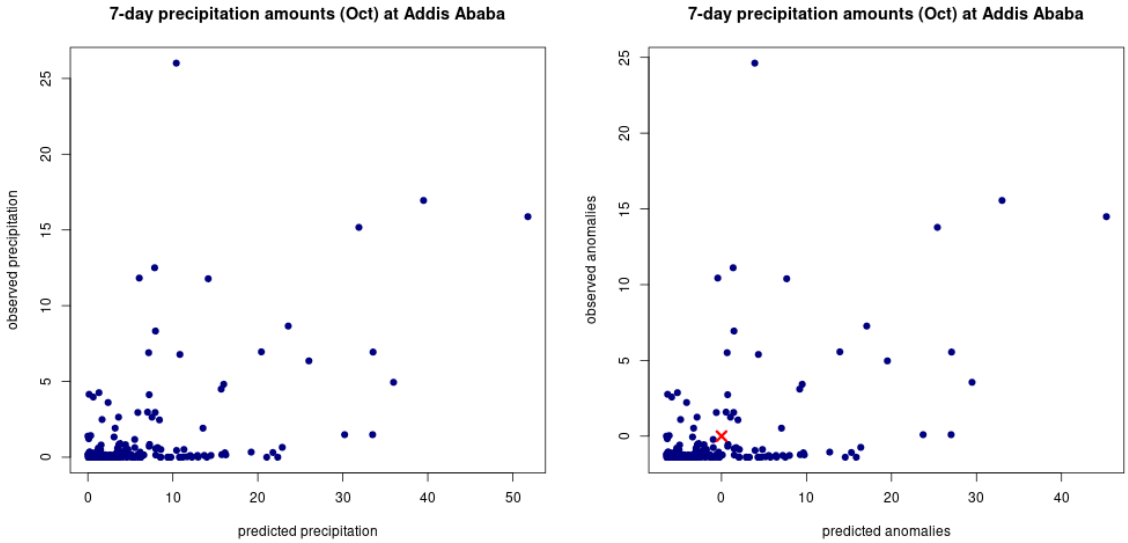

Figure 2: Same as Fig. 1 but now with forecasts initialized in October.

Receiver operating characteristic (ROC) curves

Let's take a step back and consider the case where instead of the exact 7-day precipitation amount we only want to predict whether this amount exceeds 1 mm ('rain' vs. 'no rain'). We still have the ensemble mean forecast, and based on that we can define a criterion for when we think the *observed amount* exceeds 1 mm. If we think that the forecast is unbiased and has the same distribution as the observation, then a natural choice would be to predict exceedance of 1 mm if the *forecast precipitation amount* exceeds 1 mm. But we are free to choose any other decision threshold if we think the forecast might be biased and a different threshold leads to better decisions. This is illustrated in the left panel of Fig. 3, where two overlapping histograms depict how much precipitation was forecast in the cases where the outcome was > 1 mm (≤ 1 mm).

We can evaluate our decision (i.e. forecast of 'rain' or 'no rain') by calculating

- **hit rate:** $\dfrac{\#\text{ cases where rain was predicted and occurred}}{\#\text{ cases where rain occurred}}$
- **false alarm rate:** $\dfrac{\#\text{ cases where rain was predicted but did not occur}}{\#\text{ cases where rain did not occur}}$

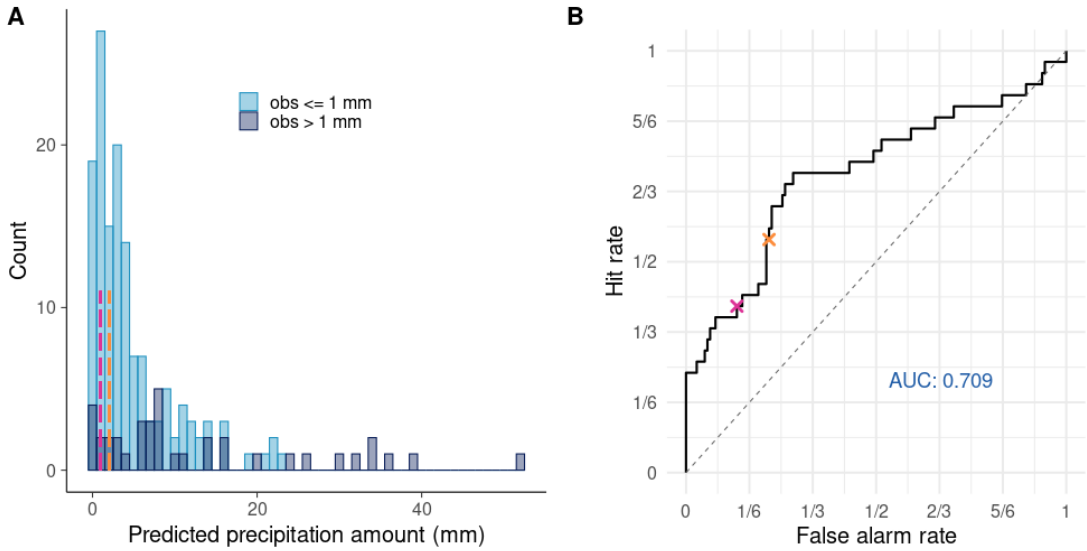

Figure 3: Histograms of predicted precipitation amounts when > 1 mm (≤ 1 mm) precipitation was observed (left panel) and corresponding ROC curve (right panel).

We would like to maximize our hit rate and minimize our false alarm rate, but in the presence of forecast uncertainty this is always going to be a trade-off. If in our example we make 1 mm our *forecast decision threshold* (magenta line in the histogram), we end up with only about 15% false alarms, but also with only about 40% hit rate (magenta cross in the right panel of Fig. 3). If we would like to increase our hit rate, we can increase the *forecast decision threshold* to 2 mm (orange line in the histogram), but this also entails an increase in the false alarm rate (orange cross in the right panel of Fig. 3). How exactly we trade off hit rate and false alarm rate is our choice, but as we adapt the *forecast decision threshold* to our personal preference we move through the black curve in right panel of Fig. 3, and this curve is called the receiver operating characteristic (ROC) curve.

The shape of this curve permits conclusions about the quality and utility of the forecast. For a completely random forecast the ROC curve would be on the diagonal. A highly skillful forecast has a ROC curve that is strongly bent upward, away from the diagonal. In that case we can find points on this curve (corresponding to particular decision thresholds in the left panel) with large hit rate and small false alarm rate. The area under the curve (AUC) is therefore a quantitative measure of forecast performance with a value of 0.5 corresponding to 'no skill' and a value of 1.0 corresponding to a perfect forecast.

The ROC curve is obtained as we move through all possible thresholds for our *forecast* in order to make a decision regarding the event of interest, here the exceedance of 1 mm precipitation. But since we know the actual observed precipitation amounts, we can consider other events as well, e.g. exceedance of 5 mm precipitation. This leads to a different ROC curve (see Fig. 4) and a different AUC that measures the forecast performance for this event.

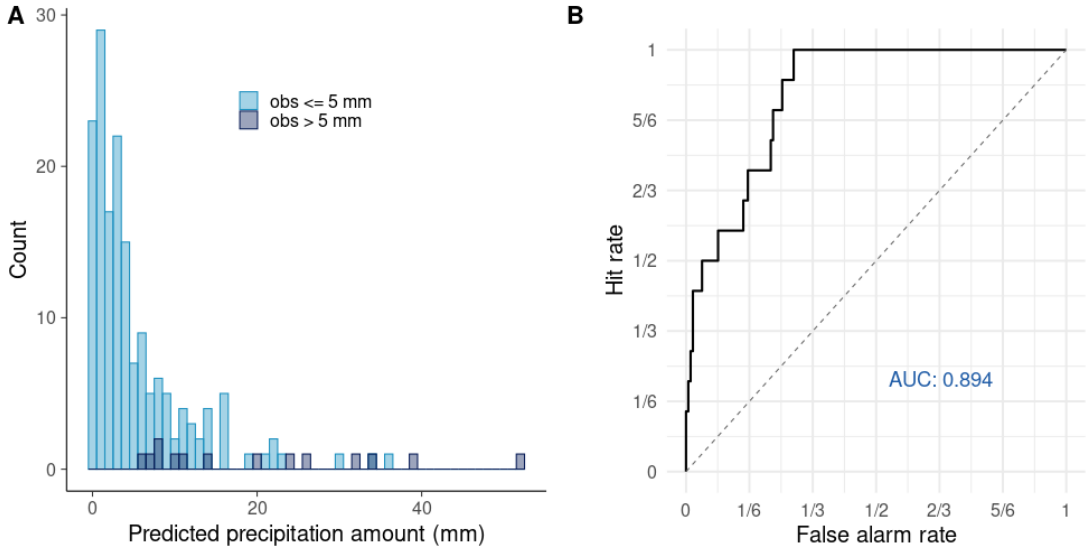

Figure 4: Same as Fig. 3 but now with an event threshold of 5 mm for the observed precipitation amounts.

Universal ROC curve and coefficient of predictive ability

Each threshold t of observed precipitation amounts defines a different precipitation event. Denoting the corresponding ROC curve by ROC_t we have a measure of forecast performance AUC_t for each such event. Is there a way to combine all these ROC curves into a single curve, and all AUC's into a single performance measure?

There is only a finite number of thresholds $t_1, \ldots, t_m$[3] that lead to distinct ROC curves, so a natural way of combining those ROC curves and the associated AUC is to consider a weighted average

$$\mathrm{UROC} = \sum_{j=1}^{m} w_j \mathrm{ROC}_{t_j} \quad \text{and} \quad \mathrm{CPA} = \sum_{j=1}^{m} w_j \mathrm{AUC}_{t_j} \ , \tag{2}$$

with weights $w_1, \ldots, w_m$ summing up to 1. In [1] a particular choice of weights is suggested[4], and the quantities in (2) calculated with these weights are called universal ROC curve and coefficient of predictive ability. The

[3] specifically, m is one less than the number of distinct values in the set of observations

[4] these weights are defined as a function of m and the number of duplicates for each unique value in the set of observations

suggested choice is motivated by the fact that in the case where all observations are distinct, CPA $= \frac{1}{2}(\rho_S + 1)$ where ρ_S is Spearman's rank correlation coefficient, which is defined in the same way as ρ in (1) but applied to the respective ranks of forecasts and observations instead of these data directly. The CPA can thus be viewed as a variant of ρ_S that deals with ties (e.g. zero precipitation amounts) among both forecasts and observations in a statistically rigorous way.

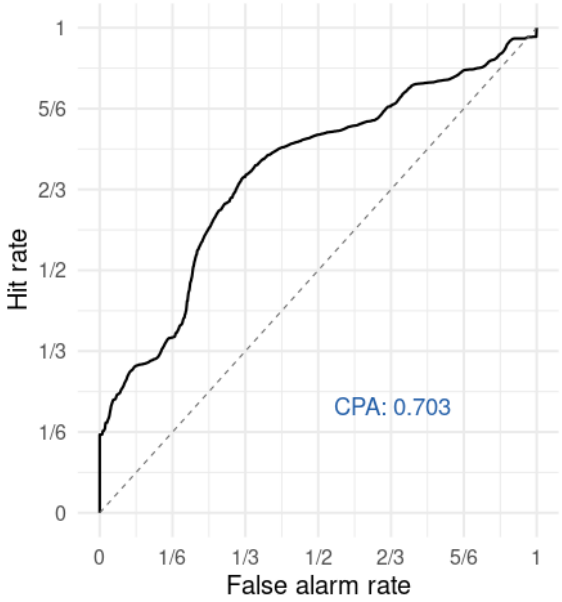

Figure 5: UROC curve for week-2 ensemble mean precipitation forecasts over Addis Ababa with initialization time in October and CPA associated with this curve.

Further reading and code

This document is essentially a brief and easy to read summary of [1], where further information and a rigorous mathematical derivation is provided. R code for the calculation of UROC curves and CPA is available on GitHub at `https://github.com/evwalz/uroc`. Figures 3, 4 and 5 above were created using this R code.

References

[1] Gneiting, T. and Walz, E.-M., 2021: Receiver operating characteristic (ROC) movies, universal ROC (UROC) curves, and coefficient of predictive ability. `https://arxiv.org/pdf/1912.01956.pdf`.