Probabilistic Skill Scores

or

The possibly eternal battle between probabilism and determinism

Let’s start this document with the root conceptualizations of both probabilism and determinism, which started in the philosophical sciences long ago:

Determinism – the philosophical position that for every event there exist conditions that could cause no other event // every event is the inevitable result of antecedent causes.

Probabilism is the philosophical position that, in the absence of certainty, probability is the best criterion.

By starting from these philosophical conceptualizations of determinism and probabilism, we can already get a clue of where the difference between probabilistic forecasting and deterministic forecasting lays, including when these are applied in weather and/or climate prediction.

Deterministic forecasting, which will not be deeply discussed in this document, aims at giving definite information on the occurrence or magnitude of a given event. In weather forecasting, deterministic forecasting means on getting a number – the number – that describes what is going to happen in the future. These forecasts rely on knowledge produced by the Mathematical, Physical and Computational Sciences, as the only way to quantitatively pin-point what is going to happen in the future is to understand all the chain of causality of different events, as well the initial conditions. In weather, this means not only understanding the equations that describe the atmosphere’s dynamics, it means also knowing the initial conditions of the atmosphere with sufficient precision, as well as knowing the initial conditions of other climate subsystems, their interactions (either lagged or unlagged, or both), and the overall simultaneous evolution of all these systems. This to get the exact number of, for instance, how much is going to rain 7 days ahead.

Of course, due to:

- present known or unknown limitations of our knowledge of the atmosphere, including chaotic behaviours;
- the modelling of it (computational power),
- the error of some observations (initial conditions, whose errors propagate over time)
- or unavailability of some observations, both in space and time.

Deterministic forecasts are rarely, if ever, completely and absolutely correct – of course, one can argue that if we use a deterministic model to forecast tomorrow’s weather and say, with 100% certainty, that tomorrow is going to rain, then that deterministic forecast is right, and this is a scenario were a probabilistic forecast would offer the same level of skill. But again, for both deterministic and probabilistic forecasts, there is always an extra step to take and, in this example of deterministic assessment, the extra step would be to forecast, again with 100% certainty, **how much** is going to rain – this is of course much harder than “just” predict that it is going to rain, which is already hard enough for most cases! This example also serves to illustrate that maybe, at least weather forecast wise, the level of advancement of deterministic predictions will only go as far as we need to. It is of Humanities’ interest to forecast how much is going to rain. It is not so much of the weather forecasters interest (or capacity?) to forecast with such precision where the extra information does not add benefit to the forecast (say, predicting rain with less than millimetre accuracy). In this regard, deterministic forecast systems might reach a stage where there is still room for improvement, but no benefits to improve them. But that is still not the case.

This is the part where deterministic forecasts have advanced a lot – and where they still need to advance a lot more: until we completely and perfectly understand and model the climate system and subsystems (which is *possible* that will never happen) chaos will be an inherent part of any weather prediction system, and therefore there will always be room for probabilistic forecasting. Probabilism, the root concept of probabilistic forecasting, rests on the fact that, while we either don't have or can't have all of the information to assess the specifics of an event, we have to assess and assign a probability of an event to happen. This is essentially a question of whether noise (stochastic behaviour) should be included or excluded from a forecast. In probabilistic forecasting, noise, or uncertainty, is included, as we will see further on the document. In deterministic forecasting, the inclusion of physically-motivated stochastic parametrization to represent uncertainty in the forecast systems is something that is still being discussed and worked upon. **[1]**

Because probabilism can be expressed in an infinite number of categories, the "initially limited" option of only assessing the probability of binary event to happen (or not to happen, a binary situation) can be expressed in a way where it branches out to give binary answers (forecasts) for different probability categories of a binary event (to rain or not rain, that is the question). Additionally, the same procedure can be applied to forecast sub-events (events inside an event). These options of probabilistic forecasting expand the domain of their useful application.

In this regard, it is only mathematically understandable why probabilistic forecasting rests on using conditional probabilities (**Bayes' Theorem**) or, in other words, why probabilistic forecasts typically present results as a the observed frequency of an event after it was forecasted to happen (or not to happen). Additionally, probabilistic forecasts can also express the likelihood of an event (sub-event) after knowing another event has happened or is going to happen (example: forecasting the likelihood of a severe precipitation with the previous knowledge that indeed there is going to rain). The extension of the application of conditional probabilities to forecasting events after the knowledge of the occurrence of another has given origin to what is called **Bayesian statistics and analysis**.

Understanding this is essential to understanding some skill scores, as these skill scores can be applied to both the forecasting of the event, and the forecasting of the sub-event (dichotomous vs polychotomous).

Introduction to skill scores

Some of the skills scores used in assessments of the quality of probability forecasts can and are also applied to determinist forecasts, like mentioned previously. It should not come as a surprise, then, that some of the formulations for these skill scores have the same general appearance to other scores, like the Mean Square Error.

Another thing that one should be aware when assessing probabilistic forecast's quality is that they are categorical forecasts, i.e., they only consider two probabilities – the event happening or the event not happening, even if those two categorized events have unequal chances of happening. This means that probabilistic forecasts consider pairs of observations of dichotomous events. Given that an event might happen or not (2 categories) independently if it was forecasted to happen or not (other 2 categories), al the possible outcomes of a dichotomous forecasting can be expressed in a 2 x 2 table, called a **contingency table**:

	Observed (x)		
Forecast (f)	Yes (1)	No (0)	Sum
Yes (1)	n11	n12	n1. = n11 +n 12
No (0)	n21	n22	n2. = n21+n22
Sum	n(.1) = n11 + n21	n(.2) = n12 + n22	n.. = N

Table 1: General Contingency Table for a dichotomous forecasting

Although on the most basic level probabilistic forecasts deal with dichotomous events, those can be built upon to get various “branches” of dichotomous events. In this case, the forecasts are called polychotomous forecasts (of dichotomous events), the contingency table has as many rows as the defined categories of probability, and the columns (still two) contain the joint distribution of observations and forecasts – one column is the frequency of the categorized event happening after it was forecasted (in that category), and the other column is the frequency of the categorized event not happening, again after it was forecasted it would happen. So, one category represents a right guess by the forecast, and the other a wrong guess by the forecast. This goes for every defined category of probability, and it is better understood with an example commonly mentioned in the literature – the probability of precipitation.

In this example, we can make a dichotomous forecast of precipitation, and 2 events are possible: either it rains or not. Extending this to the polychotomous forecast can leads us to defining more categories of probability, from forecasting zero percent of probability of precipitation, to forecasting 10 percent of probability of precipitation, and so on and on – until the forecast is saying 100 percent of precipitation. If we have only one forecast-observation event for each category (not desirable), then in each column of each category we would have a 1 in one column, and a 0 in the other. Which would mean the forecast got, for each category, a right/wrong guess 100 % of the times (depending on which column the 1 is).

In fact, the number of possible outcomes is the number of categories defined (*m*) times 2. So, for 4 defined categories of probability of precipitation, a forecaster will have 8 possible outcomes, and the contingency table is a 4 x 2 table (from *m* x 2).

If an event is well defined in the number of possible categories (*m*), it can be always expressed in a *m* x 2 table (where 2 is *k*). That being said, it is possible to define polychotomous categories in the observations as well. In this latter case, the number of possible outcomes is *m* x *k*.

All this was explained because probabilistic skill scores can be applied both to the overall forecast-observation pairs of forecasted and observed values (dichotomous events), as well as to the various categories defined which turn a dichotomous forecast in a polychotomous one. Therefore we can not only assess an overall skill of the forecast for a given event, we can also assess where (i.e. in which categories) that forecast model/system fails or gets the prediction right most of the times. This is extremely useful, as polychotomous forecasts are typically unequal in their performance of dichotomic forecasting in different categories of probability and, additionally, sometimes one forecaster is more interested in forecasting extreme categories than normal event categories and for that they need to assess the skill score of the dichotomous events in the various pre-defined categories.

The Generic Form of a Skill Score

Skill scores, or accuracy measures, are one of the ways to assess forecast accuracy in the atmospheric sciences. Therefore, forecast skill is usually presented as a skill score, which is interpreted as a percentage improvement over the reference forecasts. In the generic form, we can express a forecast's skill score this way:

$$SS_{ref} = \frac{A - A_{ref}}{A_{perf} - A_{ref}} \times 100\ [\%]$$

Where $\boldsymbol{A_{ref}}$ is the measure of accuracy of a set of reference forecasts (e.g. the climatology; persistence forecasts,…), ***A*** is the measure of accuracy of the forecast one is evaluating and $\boldsymbol{A_{perf}}$ is the value of the accuracy measure that would be achieved by *perfect* forecasts.

From the previous equation we can see the if the performed forecast as a perfect measure of accuracy (A=Aperf), then $\boldsymbol{SS_{ref}}$ would be 100%, meaning the forecast couldn't be better. If the performed forecast A doesn't improve over the reference $\mathbf{A_{ref}}$, then $\boldsymbol{SS_{ref}}$ = 0 %, meaning the forecast didn't improve the reference forecast. The performed forecast A can also be worse than the reference forecast $\mathbf{A_{ref}}$, and in this scenario, $\boldsymbol{SS_{ref}}$ would be lower than 0%, meaning that the performed forecast is worse than the reference forecast.

Brier Score (BS) and Brier Skill Score (BSS)

The Brier Score (BS) is a measure of the mean-square error of probability forecasts for a dichotomous event, i.e., and event that has two categories (e.g. rain or no rain).

$$BS = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2$$

yk and ok correspond to pairs of observations and forecasted probability of that event happening. These are called forecast-event pairs. ok represents the event, and I takes the value of either 1 or 0 – one it the event happened (it rained) and 0 if it didn't happen. Taking *n* as all the times an event was forecasted with a given probability (say, it was forecasted that it would rain with 60% of probability), the Brier score is the mean square error for every forecasted event, ranging from 0 – event was certain not to happen, to 1 – event was certain to happen.
Perfect forecasts represent a no difference between the forecasted and what happened. Therefore, a perfect forecast will have a BS = 0.

The Brier skill score **(BSS)** is the application of the generic skill score to Brier Scores of different forecasts:

$$BSS = \frac{BS - BS_{ref}}{0 - BS_{ref}} = 1 - \frac{BS}{BS_{ref}}$$

Like in the generic skill score form, a positive BSS is associated with a forecast better than the reference forecast, a BSS of 0 means no improvement and a negative BSS indicates a forecast worse than the reference forecast. The Brier Score can be decomposed in subsamples of forecasts, with the total number of forecast-event pairs being the sum of all the subsamples, or categories:

$$n = \sum_{i=1}^{I} N_i$$

and:

The Brier score can also be separated in three components that measure the Reliability, the Resolution and the Uncertainty of a probabilistic forecast.

$$BS = REL - RES + UNC$$

$$BS = \frac{1}{n}\sum_{i=1}^{I} N_i(y_i - \bar{o}_i)^2 \; - \; \frac{1}{n}\sum_{i=1}^{I} N_i(\bar{o}_i - \bar{o})^2 \; + \; \bar{o}(1-\bar{o})$$

These can be applied to the total set of observations-forecasts, subsampled by forecasted probabilities y_i. In this case, N_i is the number of times each forecast probability y_i is used for the collection of forecasts, or number of forecasts with the same probability, and o_i is the frequency of event occurring when forecasted with probability yi (oi number of times event occurred dividing by total number of cases in bin *i*).

The first term - the Reliability term - measures the closeness of forecast probabilities with real event probabilities, and in this previous equation will tend to 0 if the forecast is reliable, as it means the difference between forecasted and observed is minimal – a good forecast.

The second term - the Resolution term - expresses the deviation of observed frequency of each event (periods) from the average observed frequency of the event in analysis. It is a measure of the forecast’s ability to identify periods where observed frequencies of an event differ from the average observed frequencies of that event. The higher the Resolution term the better, as it means a given observed event’s frequency is different from the overall observed event frequency. This term is needed to illustrate the Brier Scores of some forecast systems/models, as some may have an artificially high Score due to the small differences between individual observed event’s frequency and the overall average of the observed events’ frequency.

The last term is the uncertainty term, which contains information about the inherent uncertainty of the forecasts, due to the stochastic nature of some natural systems in different time-scales, like the atmosphere. So essentially, this term is the probabilistic forecasts approach, in the Brier Score, of accounting for natural and yet-unknown chaotic behaviour.

This separation of the Brier Skill Score in 3 components is very useful to better describe if a given forecast model is performing well or not, and why it is performing in that manner.

Since these components of the Brier Score have different information about the forecast performance, they can also be applied individually, specially the first two terms, the Reliability (REL) and the Resolution (RES). Additionally, information about the terms can be seen in the Reliability Diagram, discussed below.

Reliability Diagram

Although the Brier score gives insights on the forecast performance, a comprehensive appreciation of the forecast quality in all the categorized events is better made through a joint graphical distribution of forecasts and observations. The Reliability Diagram, which shows this joint distribution, enables forecasters to see the scores distributed graphically, therefore making it easier to assess where the forecast is best or worst. The Reliability Diagram expresses the forecasted frequency on x axis, along with the observed frequency for each forecast (conditional probability p(observed | forecasted)), in the y axis. Taking as an example the following Reliability Diagram, from , which evaluated performance for wind speed forecasts, specifically, forecasting if the wind speed was above median.

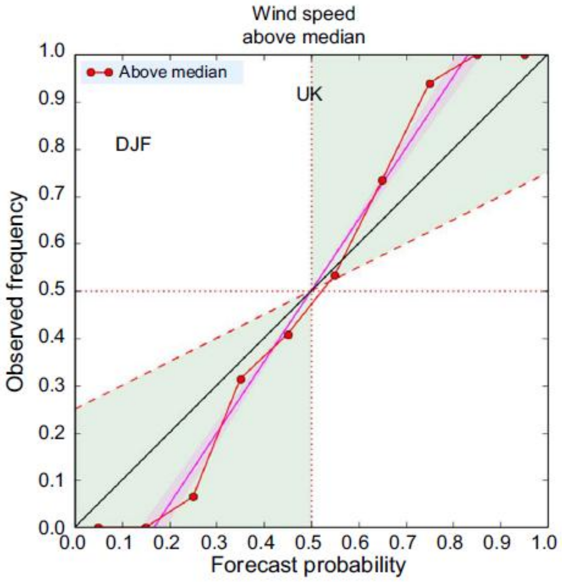

Figure 1: Reliability diagram for probabilistic forecasts of wind speed (above median), for UK's winter. From **[2]**.

Firstly, one can see that when the forecast said that those (above median) wind speeds were unlikely (f<0.3) to happen, the actual observed (above median) wind speeds didn't occur or occurred with less frequency. The same is true for the opposite situation: when the forecast issued likelihood of above median wind speeds to happen, they always happened with more frequency compared with what the forecast said. Both of these represent situations of *underconfidence*, and there are assessed in the Reliability Diagram with forecast lines of gradient steeper than 1.

The opposite situation, of *overconfidence*, can also happen in a forecast. In this case, the forecast over forecasts extreme high frequencies of events, and underforecasts extreme low frequencies of events. This is seen in the Diagram with forecast lines with a gradient shallower than 1.

In the previously presented Reliability Diagram, the forecast was underconfident. But apart from being underconfident or overconfident, probability forecasts can also have systematic bias where the forecasted frequencies are always above or below the observed frequencies – we can check these two situations in the central column plots of figure 2.

These types of diagrams are also sometimes presented with markers proportionally sized according to the sample size.

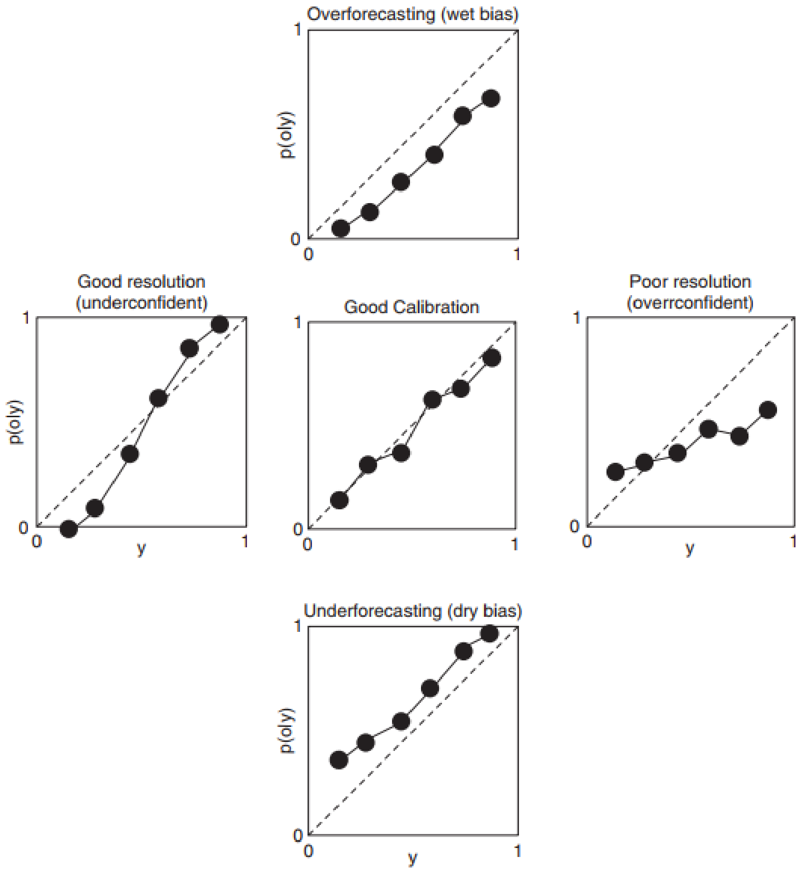

Figure 2: "Abstract" forecasts, showcasing different calibrations (Resolution and Bias). Center plot is a well-calibrated forecast, which is confident and shows no big bias over the whole range of forecasted probabilities. From **[3]**.

Sharpness Diagram

Another way of evaluating a forecast is to assess its sharpness (sometimes also called refinement). This evaluates the forecast probability alone, as in, it doesn't take into consideration the observed events that happened after each forecasted event. The sharpness diagram expresses the forecast occurances (absolute numbers) against the forecast probability for each forecasted frequency, or frequency bin. So, essentially, the Sharpness Diagram is a histogram of forecast probabilities. If a sufficient number of forecasting events has been performed; and if the forecast model is performing well, a sharpness diagram will tend to have a flat distribution, across all probability categories. Otherwise, it will tend to be like the plot shown in figure 3where forecasts near the climatological value dominate the number of forecast events (ocurrance of forecast), and the Sharpness Diagram will show decreased levels of deviation from the center:

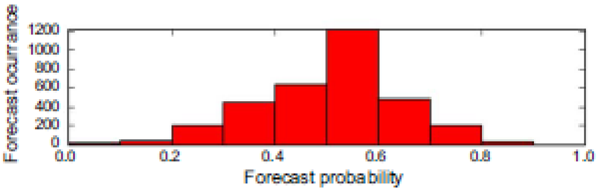

Figure 3: Sharpness diagram for the same dataset of figure 2. From **[2]**.

A sharpness diagram can be both an indication of both forecast sample size as well underperformances of the forecast model. A good forecast model, with a good sample size of observations and predictions/forecasts, will have a Sharpness Diagram that is not sharp as the one expressed in figure 3.

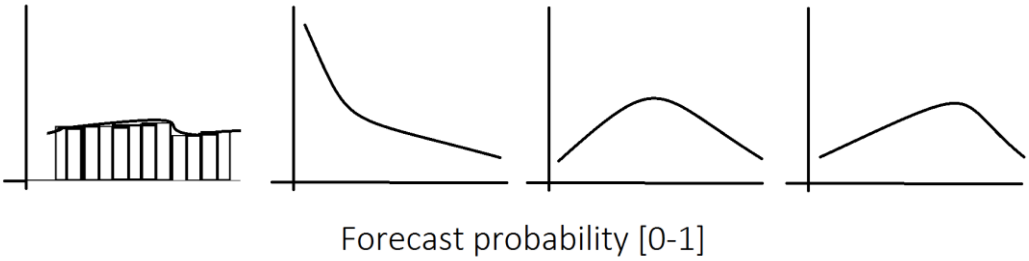

Figure 4: Conceptual sharpness diagrams. The left most one is the one relative to a forecast with highest skill, compared with the others.

ROC Diagram // ROC curves

The ROC (Relative Operating Characteristic, or, Receiver Operating Characteristic) diagram is another graphical way of verifying probability forecasts. Although it does not display full information regarding the quality of the forecast – unlike the Reliability Diagram – it presents information about a forecast in such a way that helps binary decision making, i.e., when a decision maker needs to choose between situation A and situation B. This decision will be rooted on the way a forecast performs for different forecast thresholds, which is expressed in the ROC curves.

The ROC curves are constructed using the Hit Rate (probability of detection) and False Alarm Rate (probability of false alarm) for every defined category of probability. This way the ROC curve contains information for every probability threshold issued and, depending on which category there is an issued forecast, the decision maker may check a ROC curve to see to how the forecast system/model performs in that category: does it forecast the happening of an event lots of times and it doesn't happen (high False Alarm Rate) or does it forecast the event happening with good accuracy (high Hit Rate).

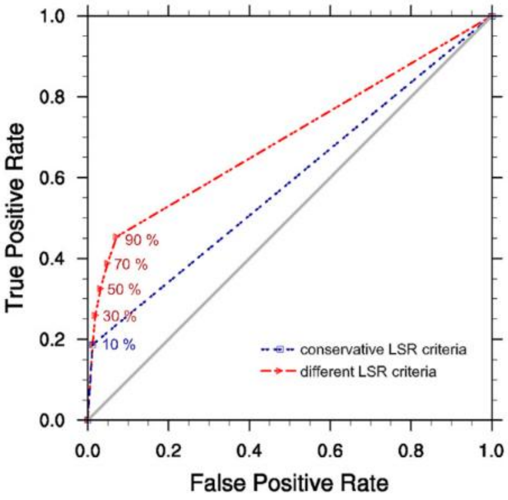

Figure 5: ROC curve depicting the false positive rate over the true positive rate for low stratus risk (risk of cloud formation, used for PV forecasting). The red line refers to a better forecast than the blue one. From **[4]**.

$$Hit\ Rate = True\ Positive\ Rate = \frac{\frac{Observed\ Yes}{Forecasted\ Yes}}{\frac{Observed\ Yes}{Forecasted\ Yes} + \frac{Observed\ Yes}{Forecasted\ No}}$$

$$False\ Alarm\ Rate = False\ Positive\ Rate = \frac{\frac{Observed\ NO}{Forecasted\ Yes}}{\frac{Observed\ No}{Forecasted\ Yes} + \frac{Observed\ No}{Forecasted\ No}}$$

The Hit Rate and False Alarm Rate are built from the 2x2 contingency tables (**Table 1**) for each category of probability of an event happening or not. The number of contingency tables available is the number of category of probabilities (or thresholds) defined, and the ROC curve is built from the Hit/False Alarm Rates calculated for each of those thresholds.

A perfect forecast is expressed in a ROC curve with two line segments, over the left and upper boundaries of the plot. This means that the area under the ROC curve of this perfect forecast is all the area of the plot, which is 1. A forecast that offers no added skill to random guessing (think of guessing which side of the coin will end up facing up) will have a 45º skewed line connecting the (0,0) and (1,1) points of the ROC – the area under the curve in this scenario is 0.5. So, the way to assess the skill score of a ROC curve of a given forecast is to introduce this information in the general skill score formula, which gives us the following equation:

$$SS_{ROC} = \frac{A - A_{random}}{A_{perfect} - A_{random}} = \frac{A - 0.5}{1 - 0.5} = 2A - 1$$

Where A is the area under under the ROC curve of the forecast under assessment. This skill score, expressed in a single scalar value, is useful to compare various forecast's ROC curves.

References:

[1] Arnold, H. M. (2013). *Should weather and climate prediction models be deterministic or stochastic?* http://onlinelibrary.wiley.com/doi/10.1002/wea.2151/full

[2] Clark, R. T. et al (2017). *Skilful seasonal predictions for the European energy industry*. http://iopscience.iop.org/article/10.1088/1748-9326/aa57ab

[3] Wilks, D.S. (2006). *Statistical Methods in the Atmospheric Sciences*.

[4] Kohler, C. et al (2017). Critical weather situations for renewable energies – Part B: Low stratus risk for solar power. http://www.sciencedirect.com/science/article/pii/S0960148116307844