Skill Scores for Probabilistic Climate and Weather Prediction

Trevor Maynard[1], Emma B. Suckling[1], Ed Wheatcroft[1] and Leonard A. Smith[1,2]

[1] Centre for the Analysis of Time Series, Dept. of Statistics, London School of Economics, U.K.
[2] Oxford Centre for Industrial and Applied Mathematics, Oxford University, U.K.

Introduction

A wide variety of skill scores are in use for evaluating probability forecasts. While the importance of using proper scores is well recognised [1], researchers often face requests to present results under a variety of scores. Is there any sense in which considering many "different" skill scores makes a case more (or less) persuasive? Which set of scores makes the most persuasive case?

Several skill scores for probability forecasts of continuous variables are considered, including the most commonly used metrics such as the proper linear (PL), continuous ranked probability (CRPS) and Ignorance (Ign) score, amongst others. Their strengths and weaknesses are contrasted under a variety of situations. The aim is to restrict the number of skill scores considered, reduce the use of misleading scores and identify independent evidence for the use of each score, based on a set of pre-defined "desirable" characteristics.

Desirable characteristics of a skill score

Propriety is a well known desirable characteristic of skill scores. Any score that is proper will always reward a forecast that reflects the forecasters true judgement. Locality, in which a forecast is rewarded solely on the probability assigned to the outcome, and is a property of the Ignorance score alone, is also debated as a desirable property in many cases. It is also proposed that any score considered to be a reliable measure of forecast quality, should also be robust against changes to the following forecast/verification properties:

- Sampling error in verification data (T1)
- Ensemble size (T2)
- Ensemble (forecast) sampling error (T3)
- Amount of verification data (T4)
- Underlying distribution shape (T5)

These properties become particularly important in the evaluation of weather and climate predictions, since small sample sizes and a limited forecast-verification archive often hinder the establishment of forecast skill.

[1] J. Bröcker and L. A. Smith, *Weather and Forecasting*, **22**:382-388 (2006).
[2] J. Bröcker and L. A. Smith, *Tellus A*, **60(4)**:663-678 (2007).

Experimental design

The performance and sensitivity of each skill score is tested, according to the set of 'desirable characteristics' outlined, in the following way. First, a 'true' underlying distribution is generated by kernel dressing [2] a set of initial points, drawn from either a) a Gaussian distribution (N(0,1)), or b) a forward projection of the Duffing map (which for different sets of initial conditions may produce distributions that are approximately Gaussian, bimodal, or something in between). Verification data are generated by sampling from the kernel dressed true distribution, f_u, defined as

$$f_u(t) = \frac{1}{N}\sum_{i=1}^{N} \tfrac{1}{\sigma_u}\phi\left(\frac{t-x_i}{\sigma_u}\right), \qquad (1)$$

with a fixed kernel bandwidth (σ_u), where

$$\phi(z) = \tfrac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \qquad (2)$$

Forecast distributions, $p(\sigma_m)$, are generated by dressing a separate sample of points, from the same initial distribution that generated the truth, using different kernel bandwidths ($\{\sigma_m\}_{i=1}^{k}$)

$$p(\sigma_m, t) = \frac{1}{M}\sum_{i=1}^{M} \tfrac{1}{\sigma_m}\phi\left(\frac{t-y_i}{\sigma_m}\right). \qquad (3)$$

The value of σ_m is found that minimises (i.e. finds the best) average score (across all verifications). A skill score is deemed to do well if the optimal σ_m is close to the true underlying value (σ_u).

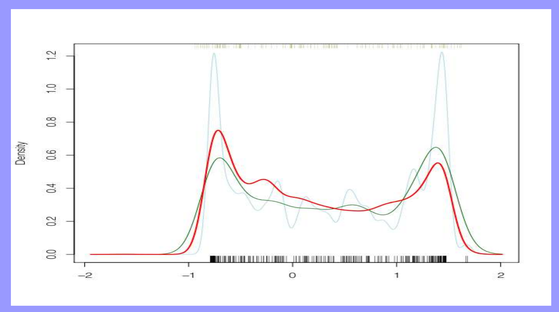

Fig 1: Illustration of an underlying distribution (red, where $\sigma_u = 0.1$), verification points (green rug at top) and the forecast distributions (blue, where $\sigma_m = 0.05$ and green, where $\sigma_m = 0.15$), both generated from the same seed set (black rug).

Figure 1 illustrates this procedure. The set of points that generates the forecast is often smaller than and different from those used to produce the true distribution. The less Gaussian the underlying distribution, the more this is likely to matter.

Results

Figure 2 illustrates each skill score as a function of kernel bandwidth for forecasts with corresponding verifications drawn from a Gaussian with $\sigma_u = 1.0$. It is clear that the proper scores all have optimal smoothing parameters that are relatively close to the true underlying value, whereas scores that are not proper (mean squared error and naive linear) do not find a minimum at all, preferring the smallest available value of σ_m, which would produce misleading results in forecast evaluation.

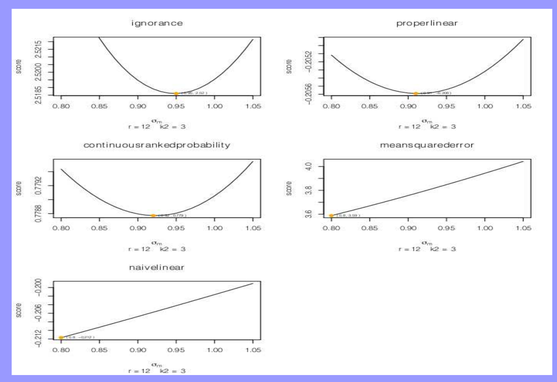

Fig 2: The optimal σ_m values for various scoring rules (for an underlying Gaussian distribution with $\sigma_u = 1.0$).

Tests are carried out for each of the proper skill scores in both the Gaussian and Duffing map scenarios to test the robustness of the score to the properties outlined earlier. Figure 3 illustrates how two scores are compared in each test case, by showing the optimal σ_m from one scoring rule against another. The numbered points indicate the forecast ensemble size, n, where n=2^n. In this example, it is clear that Ignorance performs better than CRPS (winning 6/7 and drawing once), since all points fall within the shaded area, indicating the points are closer to $x = \sigma_u$ than they are to $y = \sigma_u$). Results of all similar tests are summarised in table 1.

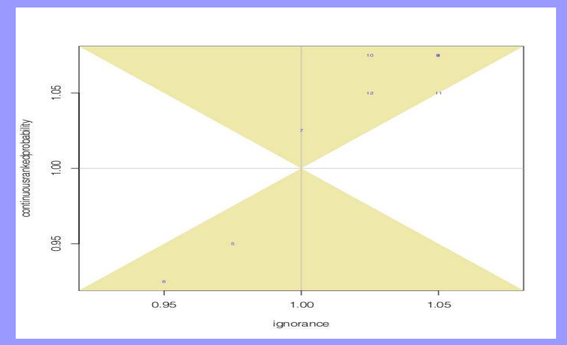

Fig 3: Ignorance against CRPS for forecasts based on a Gaussian distribution with $\sigma_u = 1.0$ and 2^8 verification points.

Test	Score 1 vs 2	1 wins	Tie	2 wins
T1:	Ign. vs CRPS	2	6	2
	Ign. vs PL	5	4	1
	CRPS vs PL	4	6	0
T2:	Ign. vs CRPS	31	18	31
	Ign. vs PL	36	8	36
	CRPS vs PL	37	7	36
T3:	Ign. vs CRPS	8	3	0
	Ign. vs PL	10	1	0
	CRPS vs PL	8	2	1
T4:	Ign. vs CRPS	8	1	0
	Ign. vs PL	8	1	0
	CRPS vs PL	6	2	1
T5:	Ign. vs CRPS	8	0	2
	Ign. vs PL	9	0	1
	CRPS vs PL	7	0	3

Table 1: Comparison of the peformance of each score under different scenarios.

Key Messages

1. A variety of common skill scores were compared and tested for robustness against the properties outlined;
2. Scores that are not proper fail to find forecast parameters close to the true underlying distribution;
3. For non-Gaussian distributions the proper linear score performs poorly compared to Ignorance or CRPS;
4. Ignorance tends to be more robust than CRPS to changes in the forecast/verification properties studied;
5. Robustness against changes in the properties studied here is a desirable characteristic of any skill score, since such scores are likely to provide less misleading estimates of skill in realistic situations, such as in weather or climate forecasting, in which sample sizes are often small and verification data is precious.