VERIFICATION OF EPS FORECASTS

CHARACTERISTICS OF SOME SCORES

Extended Abstract

Seijo Kruizinga
seijo@knmi.nl

Royal Netherlands Meteorological Institute
PO Box 201
3430 AE De Bilt
The Netherlands

Summary
In this paper some scores used in the verification of the output of Ensemble Prediction Systems (EPS) are studied with the use of simulated EPS output. The simulated output is disturbed with known errors and the response of the scores studied is assessed.

1. Introduction.

Ensemble Prediction is a special way of probabilistic forecasting. The goal of probabilistic forecasting is to specify the forecast in terms of a probability density function (PDF). In Ensemble Prediction such a PDF is represented by a random sample generated by the forecasting system. In general forecast PDF's are required:

- To be *statistically consistent* or *reliable* (Talagrand, 1997)
- To show *resolution* (narrower than the climate distribution) or alternatively to show *skill* (the mean of the forecast PDF is on average nearer to the observation than the climatic mean)

I order to verify these requirements a wide range of verification scores is applied in current daily practices. However, the interpretation of the results is often difficult because little is known about the response of these scores to errors in the forecast. In this paper we use a set of simulated forecasts to study the characteristics of some frequently used scores. The properties of the simulated forecasts are fully known. This makes it possible to disturb the forecast PDF with known errors and to study the response of the verification scores to these errors.

2. Simulation of forecast PDF's.

We assumed that the forecast PDF's had a Gaussian or normal distribution. This implies that only the mean and the standard deviation of each PDF had to be specified. Subsequently forecast PDF's were generated in the following way:

The means of the PDF's were drawn from a normal distribution with mean 0.0 and standard deviation 1.0

For the standard deviation we also used an independent draw from the same distribution but we applied the following transformation:

$$\mathrm{Std} = \alpha \exp(\beta x)$$

with α=0.8, β=0.3

The resulting PDF's show an overall skill and a variation in skill, which are comparable with day 6 forecasts of the ECMWF Ensemble Prediction System.

Verifying observations were generated by a single draw from the forecast PDF. The resulting distribution of the observations was nearly a Gaussian distribution with mean 0.0 and standard deviation 1.33. Ensemble Forecasts were constructed by taking several drawings from the forecast PDF.

3. Verification scores studied

Several widely used scores were applied to the simulated forecasts. Here only a list of scores is given. For the definition of the scores we refer to Wilks, 1995 or to the referred papers.

- Mean Error (ME) of the (Ensemble)Mean
- Root Mean Square Error (RMSE) of the (Ensemble) Mean, higher values indicate worse forecasts (0=perfect).
- Brier Skill Score (BSS), higher values indicate better forecasts (100%=perfect)
  - Bias of probabilities
  - Reliability of probabilities, lower values indicate high reliability (0=perfect)
- ROC-area, (Wilson,1995), higher values indicate improved forecasts (1=perfect)
- Ranked Probability Skill Score, RPSS, higher values indicate better forecasts.
- G^2, squared distance between average of the forecast probabilities and observed frequency, low values are to be preferred

- Continuous Ranked Probability Score (Unger, 1985), lower values are related with better forecasts
- Wilson's score, (Wilson, 1995), higher values indicate improvement of forecasts
- T^2, squared distance between average observed frequency in the consistency diagram and the expected frequency (Talagrand, 1997), lower values indicate better consistency.

4. Introducing errors in the forecast

As said earlier we studied the response of the verification scores to errors in the forecast PDF. Therefore we computed the verification results for the undisturbed forecast as well as for forecasts which were disturbed with some type of error. We distinguished the following cases:

UN: The undisturbed forecast.

CS: The standard deviation of the forecast PDF was assumed to be constant.

FM: A fixed error of 0.266 (20% of the total standard deviation of the observations) in the forecast mean.

FS: A fixed error in the forecast standard deviation. All standard deviations were multiplied by 0.7 representing forecasting systems that are too confident.

RM: A random error with mean 0.0 and standard deviation 0.266 was added to the forecast mean.

RS: A random error in the standard deviation. All forecast standard deviations were multiplied by a random factor varying between 0.45 and 2.20.

In order to introduce the same type of errors in the ensembles we first introduced the error in the PDF and drew the ensemble from the disturbed distribution subsequently.

5. Results.

We computed verification results based on our full knowledge of the forecast PDF as well as for ensembles with 10 members. The Brier Skill Scores were computed for two events:

- The 50% event, the probability of finding the observation above the climatic mean
- The 10% event, the probability of finding the observation in the upper 10% interval of the climate distribution.

For the assessment of the Ranked Probability Score we used 10 equally probable intervals.

In cases where we used our full knowledge of the PDF, probabilities were computed using normal probability theory. In the ensemble approach probabilities were estimated either by using sample statistics or by frequency counting and order statistics.

The ROC diagrams were either based on the probabilities associated with the 50 and 10% events described before. We used either 11 probability thresholds (0% (+10%) 100%) or the number of ensemble members for defining the ROC curve.

The results for the fully specified PDF are summarised in table 1. This table shows that the skill measures RMSE, Brier Skill Score (BSS) and Ranked Probability Skill Score (RPSS) respond only moderately to errors in the forecast. Wilson and ROC-area show the same moderate response. Furthermore it can be concluded that errors in the mean, either random or systematic, have the largest impact. Scores, which can be associated with the statistical consistency or reliability like the reliability component of the Brier Score and G^2, are strongly effected by all types of errors but most strongly by systematic errors.

Remarkable are the result for the ROC-area at "the fixed error in the mean case" and the result for the Wilson score at the "fixed error in the standard deviation case". Both indicate that the forecast improves when the indicated error is introduced into the forecast. Further research learned that with the ROC-area this is caused by the way it is computed. For the Wilson-score this effect is related to the way in which the score is defined.

The results for an Ensemble System with 10 members are summarised in table 2. In general the conclusions the same as in the previous case. The CRPS behaves similarly the BSS and the RPSS. Mark, however, the strong response of T^2 to systematic (fixed) errors either in the mean or the standard deviation. In figure 1 the consistency diagrams associated with these cases are shown. The strong response of T^2 can clearly be recognised in this figure as well. Furthermore it should be noted that the RMSE of the Ensemble Mean now indicates improvement when the forecasting system is overconfident.

The overall quality of an Ensemble Prediction System is expected to improve when number of members in the ensemble is increased. In table 3 it is demonstrated how this improvement is reflected in the verification scores. In this table the verification results are given for an undisturbed forecast but the number of ensemble members is varied from 5 to 50. In this table some additional scores are given as well. We added the Probability of Detection (POD) and the False Alarm Ratio (FAR) of an event with a 1% climatic probability, assuming that this event is forecast when one or more ensemble members contain the event. The relative frequency of warnings is given as well.

In table 3 it is shown that the improvement in skill of the ensemble is reflected in all skill measures. The results of G^2 and T^2 show that all ensembles are statistically consistent (the used definition of T^2 implies that it is proportional to the number of ensemble members). The reliability associated with the Brier score indicates that the reliability of the derived probabilities clearly improve with larger sample sizes.

6. Summary of conclusions.

- Skill measures (RMSE, Brier, RPSS etc) respond moderately to errors in the forecast mean of the PDF

- Skill measures respond only marginally to errors in the width of the forecast PDF

- Reliability(Brier), G^2, T^2 respond strongly to all type of errors and most strongly to systematic errors
- Wilson's score behaves irregular in case of overconfidence with fully specified PDF
- RMSE indicates improvement in the ensemble forecast in the case of overconfidence
- ROC-area should be used with care
- The consistency diagram and its associated T^2 is a very good score for testing statistical consistency of ensembles

7. References

Talagrand, O., R. Vautard and B. Strauss, 1997: Evaluation of probabilistic forecasting systems. ECMWF Workshop on Predictability. 20-22 October 1997, Reading

Unger, D.A.,1985: A method to estimate the continuous ranked probability score. Ninth conference on Probabilty and Statistics in Atmospheric Sciences, Virginia Beach, October 9-11, 1985.

Wilks, D.S., 1995: Statistical methods in atmospheric sciences. Academic Press, 467 pp.

Wilson, L.J. 1995: Verification of weather element forecasts from an ensemble prediction system. ECMWF Fifth Workshop on Meteorological Operational systems, 13-17 November 1995, Reading

Table1: Verification results for different error types
Fully specified PDF

Score	CASE					
	UN	CS	FM	FS	RM	RS
ME	0.00	0.00	0.27	0.00	0.00	0.00
RMSE	0.88	0.88	0.92	0.88	0.92	0.88
P(x>0.00), 50%						
Bias(%)	-0.2	-0.2	3.8	-0.2	-0.3	-0.2
BSS(%)	41.4	40.4	38.0	40.1	37.6	39.8
Reliability	5.6	25.9	190.0	301.0	54.6	60.8
ROC-Area	.867	.863	.864	.863	.851	.860
P(x>1.69), 10%						
Bias(%)	0.2	0.5	4.2	-2.3	0.7	1.0
BSS(%)	30.0	28.2	25.3	27.8	25.9	26.3
Reliability	8.7	13.2	371.3	130.5	60.0	59.5
ROC-Area	.865	.850	.889	.819	.853	.848
Overall scores						
Wilson(%)	18.0	16.4	17.4	20.8	17.4	17.7
RPSS(%)	39.0	37.7	35.4	37.4	35.3	36.9
G^2	7.1	10.5	829.1	443.0	17.1	35.4

Table 2: Verification results for different error types
Ensemble Prediction System (M=10)

	CASE					
Score	UN	CS	FM	FS	RM	RS
ME	0.00	0.00	0.27	0.00	0.00	0.00
RMSE	0.92	0.92	0.96	0.90	0.96	0.94
RPSS(%)	33.1	31.5	29.6	33.1	29.5	30.9
CRPS	5235	5367	5496	5226	5502	5474
G^2	11.7	16.8	830.5	442.1	21.4	43.5
T^2	9.8	118.9	1966.9	2831.6	86.2	146.6
Sample Statistics, P(x>0.00)						
Bias(%)	-0.2	-0.2	7.8	-0.2	-0.3	-0.2
BSS(%)	37.8	36.7	34.5	37.4	34.2	36.2
Reliability	54.9	28.1	737.3	331.3	160.4	68.5
ROC-Area	.851	.864	.849	.853	.837	.845
Order Statistics, P(x>0.00)						
Bias(%)	-0.2	-0.2	7.7	-0.2	-0.4	-0.2
BSS(%)	35.8	34.5	32.6	35.9	32.0	34.3
ROC-Area	.847	.840	.844	.854	.832	.841

Table 3: Verification results versus number of Ensemble Members

	5	10	20	50
ME	0.005	0.003	0.000	0.000
RMSE	0.96	0.92	0.90	0.89
RPSS(%)	27.3	33.1	35.8	37.7
CRPS	5683	5235	5015	4875
G^2	9.27	11.68	10,87	9.00
T^2	0.9	9.8	22.4	49.7
Sample Statistics, P(x>0.00)				
Bias(%)	-0.1	-0.2	-0.2	-0.2
BSS(%)	33.9	37.8	39.4	40.7
Reliability	193.6	54.9	17.8	6.7
ROC-Area	.836	.851	.858	.863
Order Statistics, P(x>0.00)				
Bias(%)	-0.2	-0.2	-0.2	-0.2
BSS(%)	30.2	35.9	38.3	40.2
ROC-Area	.825	.847	.857	.865
POD(1%)	37.6%	53.9%	67.5%	86.4%
FAR	3.0%	5.3%	8.1%	12.7%
Alarm	3.4%	5.7%	8.6%	13.4%

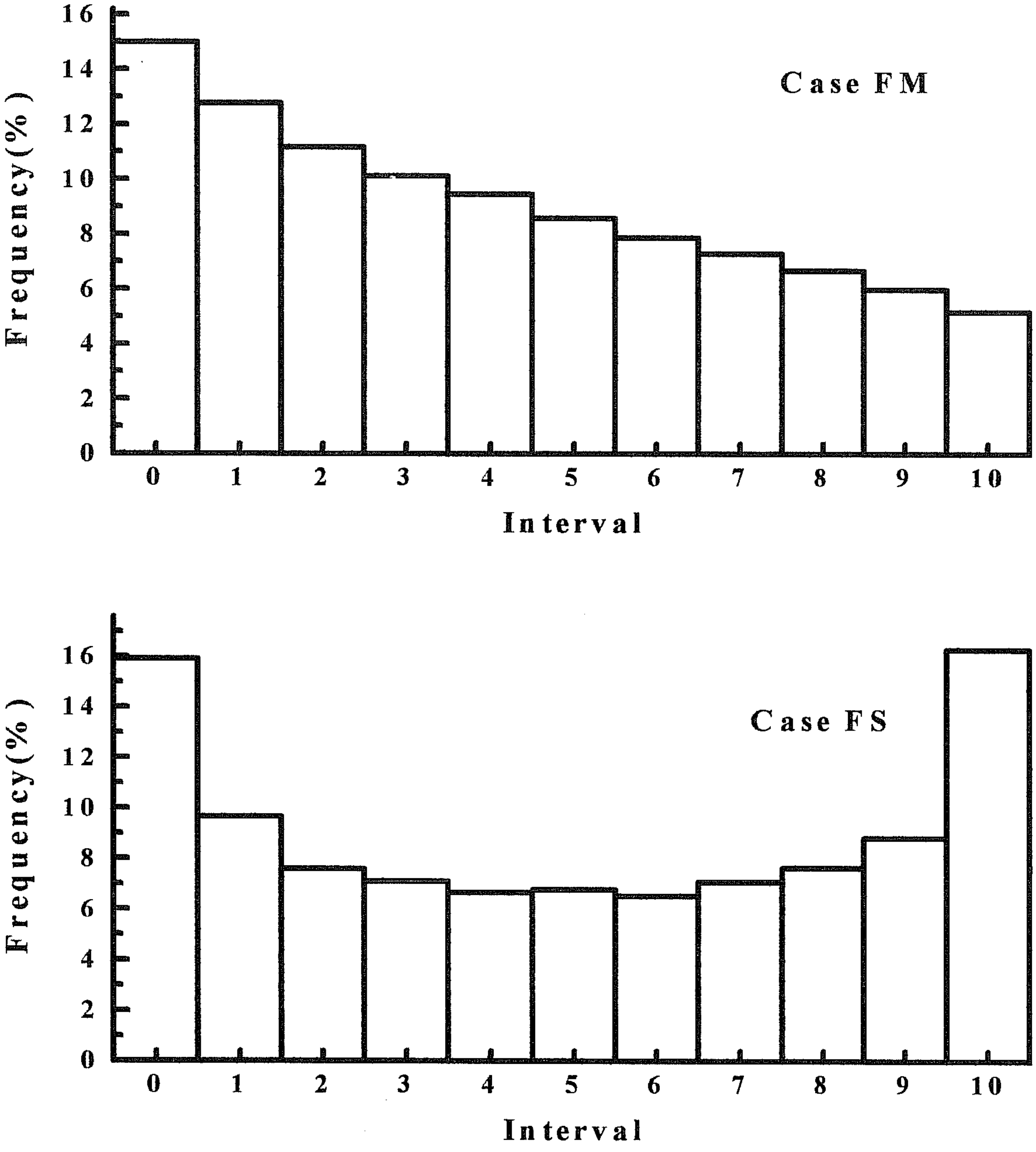

Figure 1: Consistency diagrams for the cases "Fixed error in the mean, FM" and "Fixed error in the standard deviation, FS".