

What looks bad might be good

On the interpretation of verification statistics

Anders Persson, SMHI

1. Introduction

Professor Tor Bergeron (1893-1977) used to say that the word “objective” was often misused by meteorologists. ”Objective” forecast methods had not descended from Heaven. They had been developed by humans and were selected by humans; there is always more than one “objective” method to choose from. The word “objective” only means that the numerical results are independent on the human being.

The same criticism is valid for the misinterpretation of “objective verification”. It only means that the numerical results are independent on any human who computes the verification. Whether the verification results are “good” or “bad” is a subjective matter. It depends on the subjective choice of verification method, it depends on what we subjectively are trying to achieve, it depends on the peculiarities of the period we subjectively have chosen etc.. This is nothing in particular for verifications, but for all statistics. If a country’s police one year caught 21.3% more drunken drivers than the year before, then 21.3 is an objective number. However, if this means that people had started to drink more or if the police had become more efficient, or both, is for further investigations to find out.

2. Standard verification scores

This article, the first of two, will try to address the problem of interpretation of verification statistics. It draws mainly from experiences at ECMWF, so some of it might not be applicable for short range forecasts, but most of it has, I think, general relevance.

To evaluate the performance of NWP the meteorological community has by tradition mainly used the Root Mean Square Error (RMSE), often complemented with the Anomaly Correlation Coefficient (ACC), Mean Absolute Error (MAE), Tendency Correlation (TC) and the mean error (ME). A reduced RMSE and ME, and an increased ACC have been deemed “good”, the opposite “bad”. These conclusions are often open for debate.

1. The sample size must be large enough to make the results significant. The daily values of ACC at D+10 can vary between 100 and -20%. An average over a period of 90 days would have an uncertainty of roughly $\pm 10\%$
2. If we consider the correlation between consecutive forecasts the problem of sufficiently large samples becomes increasingly true. To get uncorrelated forecasts of a set of 90 days we would have to pick every 5th forecast, which would reduce effectively the sample to just 18 forecasts.
3. Even if the sample size is sufficiently large the fact that the score is rising or falling is no certain evidence that something has become “better” or “worse” if for no other reason that “objective” scores often can give contradictory signals.
4. As we will see below, an improvement of the model’s ability to simulate the atmospheric motions, in particular the anomalous or extreme ones, can yield an increased RMSE.

5. Many verifications diagrams are made under the assumption that the error at initial time is zero. This might be true for a comparison with the analysis, but not against observations. The error at time=0 is the analysis error. By letting the verification start at the origin often leads to a curve with a “kink” at the first forecast interval.

Everything will follow from the mathematical equation $(a-b)^2=a^2+b^2-2ab$, where the left hand side will symbolize the RMSE, the three right hand side terms the characteristics of the NWP model, the behaviour of the atmosphere and forecast agreement respectively.

3. The RMSE

The standard expression of RMSE is

$$E_j = \sqrt{\frac{1}{T} \frac{1}{N} \sum_{n=1}^N \sum_{i=1}^T (f_{i,j} - a_{i+j})^2} \quad (1)$$

where E= the root mean square error, f=forecast, a=verifying analysis, i=forecast day, j=lead time, T= number of forecasts and N=number of grid points. For our pedagogical purposes it is convenient to simplify this expression into

$$E^2 = \overline{(f - a)^2} \quad (2)$$

That is, the mean squared error with the indices dropped and with an overbar which symbolizes the averaging over time and grid points. We can now re-formulate this equation by introducing c, the climatological average of the verifying day

$$E_j^2 = \overline{(f - a)^2} = \overline{(f - c + c - a)^2} \quad (3)$$

Using the powerful relation $(a-b)^2=a^2+b^2-2ab$ we have

$$E_j^2 = \overline{(f - c)^2} + \overline{(a - c)^2} - 2\overline{(f - c)(a - c)} \quad (4)$$

1. The first term, the variance of the forecasts around the climatological average is a measure of the realism of the model, its ability to simulate atmospheric features. It should have the same size as the second term, the variance around the climatological average.

2. It is only the last term, the covariance between forecast and analysed anomalies, which is related to the forecast skill in some sense. The higher the agreement between forecast and analyzed anomalies the higher its numerical value and it will contribute to a reduction of the RMSE.

3. The model variance is flow dependent, in particular with respect to the seasons. Since the model variance for a realistic model should vary with the atmospheric variance, the RMSE will tend to increase for dry parameters (wind, pressure, temperature, geopotential) during winter seasons in the extra-tropics and decrease in summer time.

4. The second term is beyond human intervention, but not the first term. If, intentionally or unintentionally, changes to the model properties are introduced which worsens the model's ability to forecast anomalous features, this will affect the RMSE. I case anomalies are over-forecast the RMSE

will increase, in case they are under-forecast, for example due to coarse grid or diffusive schemes, the RMSE might decrease which is a misleading impression of “improvement”.

5. Verification of model improvements should therefore not primarily be measured by the RMSE (or any other measure of forecast accuracy or skill) but by the variability. The variability displayed in (4) is not the only measure. But whatever measure, the variability of any parameter should remain on the same level through the forecast range and be at the same level of the variability of the analyses.

4. Important special cases

To understand the properties of the RMSE two special cases will be discussed. The first is to replace the forecast (f) with a trivial alternative, the climatological average (c). This reduces (2) to

$$E_c^2 = \overline{(a - c)^2} = A_a^2 \quad (5)$$

where A_a is the atmospheric variability. Another special case is when, with increasing lead time the covariance term in (3) will approach zero, i.e. no correlation between forecast and analysed anomalies.

$$E_{j \rightarrow \infty}^2 \Rightarrow \overline{(f - c)^2} + \overline{(a - c)^2} = A_f^2 + A_a^2 \quad (6)$$

where A_f is the model’s variability around climate.

If $A_f = A_a$ then

$$E_{\infty \rightarrow \infty}^2 \Rightarrow 2A^2 \quad (7a)$$

The RMSE will, on average, asymptotically converge towards the error saturation level (ESL).

$$E_{j \rightarrow \infty} \Rightarrow ESL = A\sqrt{2} \quad (7b)$$

The errors saturation level which is 41% above the climatological error level. This is also the ESL for persistence forecasts or pure guesses.

The high level of the ESL implies that at some early stage in the medium range forecast evolution, currently at +120 to +168 hours (depending on model), a NWP forecast model, literally interpreted, provides negative information. Medium range forecast models which after +120 hours are systematically unable to simulate intense cyclogenesis, vigorous blockings and other pronounced anomalies, might appear to provide lower RMSE than models which have this ability. Although this is not the case for short forecasts, it is important to be aware of the effect for at least two reasons:

1. With increasing resolution and ability to describe small scale features like squall lines, the verification becomes more sensitive to failures to forecast them correctly in time, place and intensity. Numerical or parameterization schemes which have a dampening effect might appear to have “improved” the model.
2. Any dampening of realistic, but less predictable features should not be achieved through introduction of more diffusion, but by some sort of ensemble averaging. This acts as a dynamical filter, first removing less predictable features and then re-introducing them as probabilities.

5. The size of ensemble perturbations

The perturbations in all current ensemble systems are scaled to have a size that agrees with the average analysis error. The previous discussion helps us to understand why the true errors of the perturbed analyses are $\sqrt{2}$ of the analysis errors. Let \mathbf{p} be the perturbed analysis, \mathbf{a} the imperfect analysis and \mathbf{t} the true analysis. The true RMS error of a perturbed analysis Δ is then

$$\Delta^2 = \overline{(p-t)^2} = \overline{(p-a)^2} + \overline{(t-a)^2} - 2\overline{(p-a)(t-a)}$$

Assuming that there is no correlation between the perturbations and analysis error

$$\Delta^2 = \overline{(p-a)^2} + \overline{(t-a)^2}$$

Since the size of the perturbations $|p-a|$ is the same as the analysis error $|t-a|$ we have

$$\Delta = \text{analysis error} \sqrt{2}$$

If the error growth during the first day or so can be regarded as linear, the perturbed forecast will on average be significantly worse than the non-perturbed Control. The remedy is to have many perturbed members so what they lack in individual skill they will compensate by their number.

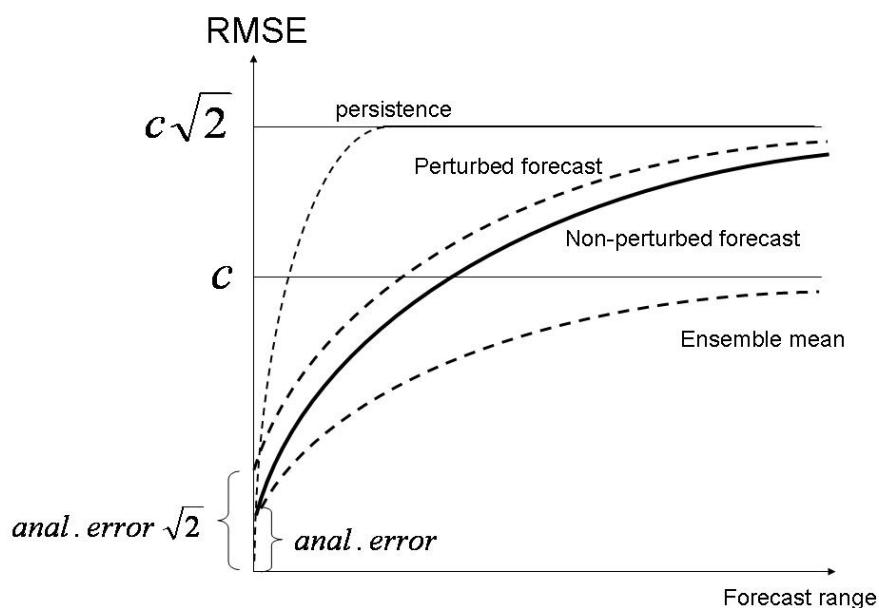


Figure 1: A schematic diagram of the RMS error growth for different forecast types. There are reasons to assume that the error level of a perturbed forecast is $\sqrt{2}$ of the RMS error or the ensemble mean.

6. Summary

The value of RMSE depends not only on the forecast “skill” but also on the atmospheric flow conditions and the ability of the NWP to simulate them. Model improvements should be evaluated by some variance measure instead of the RMSE.

7. The next part

Vector algebra will be introduced as a new way to understand, not only RMSE, but also ACC and more generally correlations. The vector approach will also help to illuminate the problems of verifying against analyses instead of observations, forecast “jumpiness” and its relation to forecast skill and the limitations of the lagged forecast in the short range.