The relation between forecast inconsistency and skill in deterministic and ensemble prediction

Anders Persson

European Centre for Medium Range Forecasts (ECMWF)
Reading, United Kingdom

1. Introduction

When medium range NWP forecasts (beyond 3 and up to 10 days) became more commonly used during the 1980's, the forecasters found that they could not apply the same forecast culture as with short range NWP forecast. Not only was it difficult to apply the traditional techniques to modify *a priori* a medium range NWP beyond D+3, it was also difficult to judge *a posteriori* the skill from the traditional verification scores.

1.1 A posteriori assessments

Verification scores like the Root Mean Square Error (RMSE) and Anomaly Correlation Coefficient (ACC) can not be trivially interpreted beyond D+3. Weather situations when the forecasted and/or verifying flow pattern is anomalous (blockings and cut-offs) has a tendency to yield favourable ACC scores and not so favourable RMSE, whereas the opposite is true for zonal flow patterns. If these statistical effects are not taken into account wrong conclusions may easily be drawn with regard to model changes or comparison between models.

1.2 A priori assessments

In the short range it is possible for a forecaster to perform an "update" by comparing a +12 h NWP forecast with more recent observations; phase or intensity differences can more or less linearly be extrapolated up to +24 or +36 h. However, beyond +48 h this linear approach breaks down. The difference between successive forecasts, the "forecast inconsistency", tends to be larger in the medium range, than in the short range. Unable to use their traditional tools to modify the medium range NWP the forecasters instead try the "forecast the forecast skill" using the " inconsistency" as an indication.

2. A vector interpretation of RMSE and ACC

It is essential to find if there is a reference level for "forecast inconsistency" which could define "acceptable" inconsistency and the statistical nature of this "consistency/skill" relation. In Persson (1996 a,b) derivations of certain statistical properties related to error and consistency can be found. Here an attempt will be made to illustrate the results using simple applications from vector algebra.

2.1 Definitions

By definition the correlation, r, between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ of length *a* and *b* and separated by an angle α is the normalized inner product of $\boldsymbol{a}$ and $\boldsymbol{b}$

$$r = corr(\boldsymbol{a},\boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|a\|\|b\|} = \cos\alpha \tag{1}$$

2.2 Error and model climate

Let ***a*** be the observed anomaly over a longer time period with size $\mathbf{A_a}$, and ***f*** the forecasted anomaly of length $\mathbf{A_f}$ as defined above. The two vectors are separated by an angle α at the origin **c**. We assume that there is no bias in the forecasts, i.e. they share the same climatology as the real atmosphere. The RMSE is then the length $\mathbf{E_j}$ of the vector difference **f-a** and the ACC the cosine of the angle (fig. 1a). For a model where $\mathbf{A_f > A_a}$ the "over-amplification" contributes to an increase of the RMSE (fig. 1b), whereas for $\mathbf{A_f < A_a}$ "the smoothing" might, for a certain range of angles α, act to decrease the RMSE (fig. 1c).

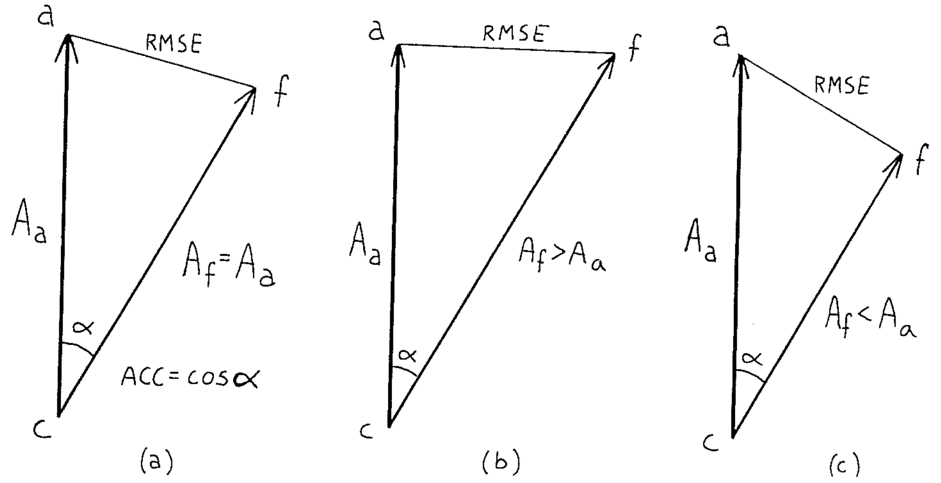

Fig.1 A vector representation of the average atmospheric and forecast state at some lead time. a) forecasted anomalies on average the mean the same as the observed; b) the forecast anomalies generally larger than the observed ("over-amplification") and c) the forecast anomalies lower than the observed ("smoothed" forecasts).

2.3 Error saturation level

When the forecast range and the averaging period increase, the forecast and the analyzed anomalies will tend to become uncorrelated. The vector $D+\infty$ represents a forecast with no skill and is thus orthogonal to the verifying vector. As discussed by Simmons et al (1995), when the skill vanishes $E_j \rightarrow A_a\sqrt{2}$ providing that $\mathbf{A_f = A_a}$ the model preserves the same variability as the analyses. A NWP which exhibits excessive dynamic activity will have $\mathbf{A_f > A_a}$ and an error saturation level > $A_a\sqrt{2}$.

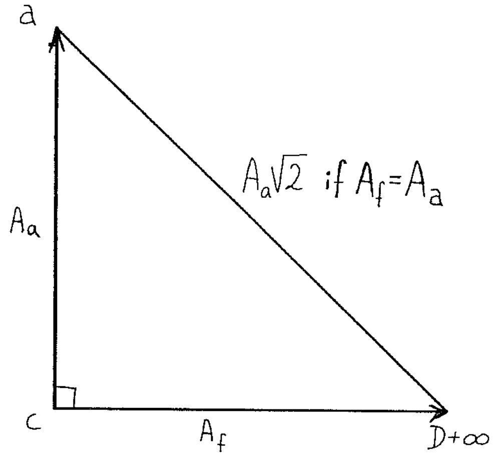

Fig.2 The observed and forecast state at a lead time when there is no skill in the forecast. In the case $\mathbf{A_j = A_a}$ then the maximum error level, the saturation level, is = $A_a\sqrt{2}$, else $\neq$ $A_a\sqrt{2}$.

2.4 Verification studies

A comparison between the operational and a smoothed version of the ECMWF 500 hPa forecasts was made for December 1994 - February 1995 over the Northern Hemisphere between 35-60 N, by weighting together the last three days' forecasts, the largest weight to the latest (Fig. 3). The weights were determined more or less ad hoc, e.g. 0.9, 0.1 and 0 for the smoothed D+1 forecast, 0.4, 0.3 and 0.3 for the smoothed D+8 forecast. An analysis of the of optimum weights, see Simmons (1996).

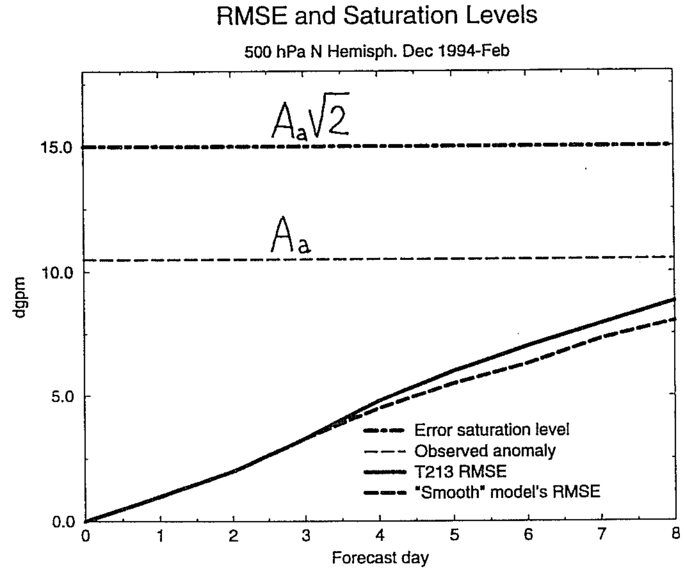

Fig.3a The RMSE evolution for one model with model climate, and one where smoothing has been applied. The RMSE of the model with stable climate approaches a higher level of error saturation than a smooth model.

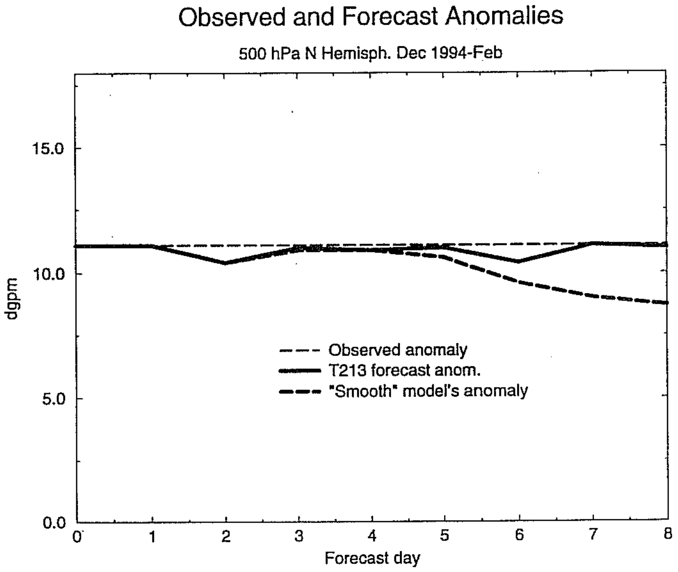

Fig.3b The observed and forecast levels of anomalies. Note that the smoothing has reduced the anomalies by 25-30%. If this smoothing is accomplished during the forecast computations it will be linked to severe model problems.

2.5 Stable model climate

A NWP model which due to model shortcomings (for example excessive diffusion) gradually through the forecast dampens its dynamic activity will experience difficulties forecasting flow anomalies like cut-offs and blockings in the medium range. In that case the climatological range of the forecasts will be lower than the observed and as a consequence $\mathbf{A}_j < \mathbf{A}_a$ and the forecasts tend to be close to the climatological mean then $\mathbf{E}_j \rightarrow \mathbf{A}_a$. The resulting lower RMSE level will give a unjustified favourable impression of the model's performance. Any under forecasting of the climatological variance, albeit its ability to reduce the RMSE, should be avoided by a NWP system:

1) the introduction or maintenance of "favourable" model deficiencies will make it more difficult to reduce unwanted and "unfavourable";

2) a model climate drift, in particular towards the climate state, makes the model almost inadequate both for long-range climate simulations and for ensemble forecast prediction since smoothing of small features also affects the ability to forecast blockings;

The development work with the ECMWF model, like most other global models, strive to reduced the RMSE under the condition of a stable model climatology; a 10-day forecast should look as realistic as the real atmosphere - even if it does not exhibit any operational forecast skill (Persson-Strauss, 1995). By computing the forecast anomalies or the typical 12- or 24-hour forecast tendency, any model drift can easily be detected by comparison with the analyzed values.

Any smoothing of small scale and less predictable features should be done in post-processing mode, by temporal or spatial filtering. The best and most consistent way of smoothing is achieved through the Ensemble Prediction System either with clustering or considering the Ensemble Mean(see below).

3. The forecast inconsistency

The latest update of a forecast should be regarded as the best and the previous forecast should be disregarded. Nevertheless, the quality of the output from a NWP system tends to be judged as much by its "consistency" with the previous output, as by its assumed skill. When, after a NWP model change, the level of "inconsistency" increases to a higher level it is taken as a negative development, even if the error level decreases. Analysing a relation between three or more vectors (one analysis and at least two forecasts) should of course be done in a multi-dimensional space, so some of the following two-dimensional figures should be seen as necessary simplifications.

3.1 Factors determining the inconsistency

The level of inconsistency depends, as the error, also on the seasonal anomalies and the realism of the model. Any model climate drift towards a climatological regimes, i.e. $\mathbf{A}_j > \mathbf{A}_{j+1}$ -> 0 will tend to decrease the inconsistency. Changes in the model climatology will affect the consistency level; a more active model will display a higher degree of inconsistency, a less active a lower level.

When a forecast system improves, in the sense that the RMSE decreases, the change in the level of inconsistency will not necessarily decrease. If the shorter forecasts have improved more than the longer (fig. 4a) this will be reflected in an increased inconsistency, which wrongly might be interpreted as the shorter forecast has become less reliable. If on the other hand the shorter forecasts have improved less, or not at all while the longer have improved, the consistency will appear to have improved and wrongly give an impression that the shorter forecasts have become more reliable (fig. 4b)

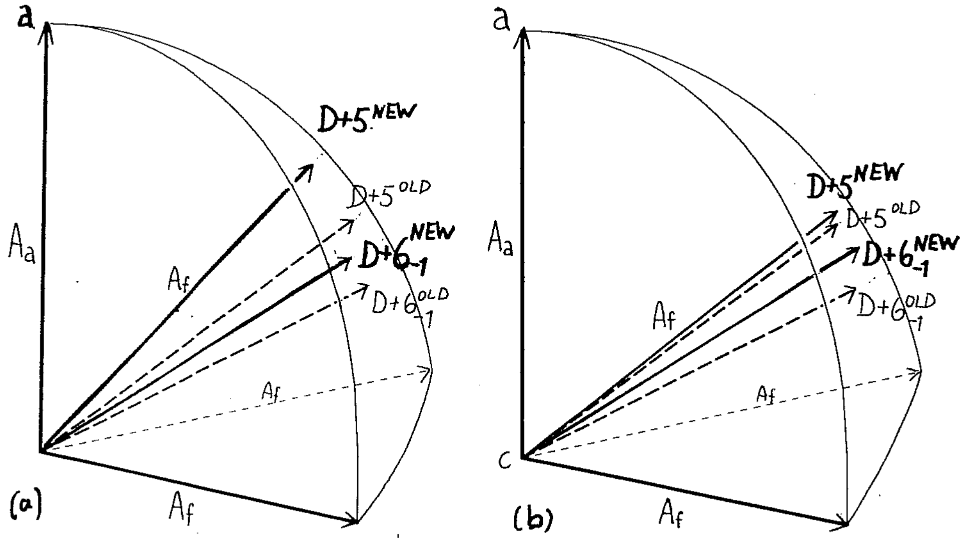

Fig. 4 Forecasts, verifying on the same day, from two consecutive runs. The dashed lines represent forecasts before an improvement, the straight line after. In a) the shorter forecasts have improved more than the longer and the inconsistency appears to have increased; in b) only the longer have improved more than the shorter and the inconsistency appears to have decreased.

3.2 Error correlations between different forecast runs

When meteorologists look for statistical relations they sometimes find that the errors from successive forecast runs from the same or different models verifying on the same time look similar. Correlations between the errors might also indicate high values around 0.5. But unfortunately these correlations are mainly a statistical artefact when the compared forecasts verify on the same analysis. It can easily be shown (fig. 5) that forecasts in the non-skilful forecast range have error correlations of 0.5.

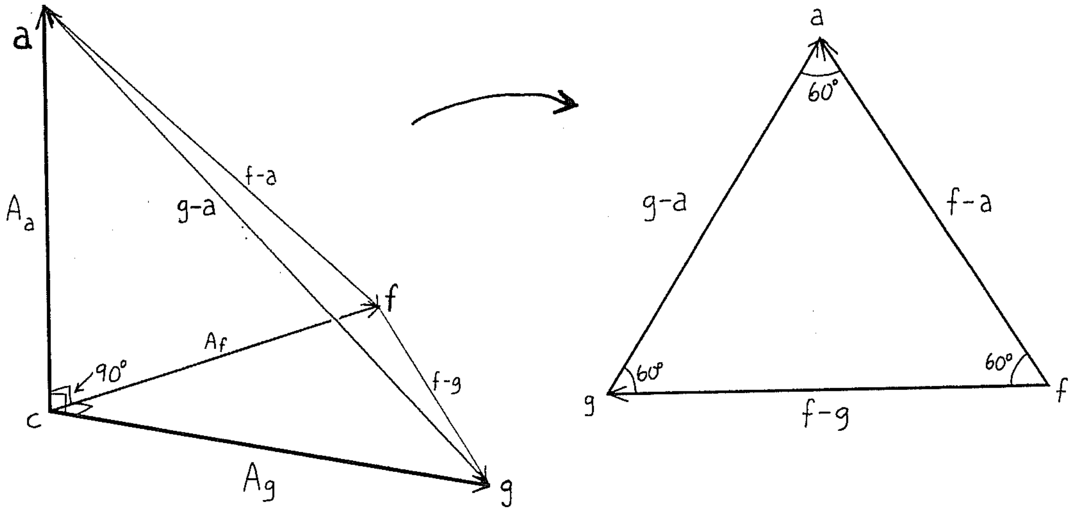

Fig. 5: a) The vectors **f** and **g** are two non-skilful and non-correlated forecasts verifying on the same analysis **a**. The error and forecast difference vectors are of equal size and together form an isosceles. b) This triangle yields a forecast-error and consistency-skill correlation of 0.5 (= cos 60 deg).

3.3 The correlation between inconsistency and error

Several verification studies in the early 1980's reported correlation around 0.3 between consistency and skill. These non-zero correlations gave hope of a "signal" that could be amplified applying more refined filtering or statistical stratifications. However, all attempts to distil any useful indicator came to nothing. In spite of these disappointing results the "consistency-skill" approach is still widely used, mainly because no other method to "forecast forecast skill" was available prior to the introduction of ensemble prediction.

Fig. 6 shows two forecast **f** and **g** verifying at the same time. They can be from the same model, but run on different initial times; they can be run at the same time but on different analyses (EPS) or they can be from two entirely different models. The both have the same and realistic model climate (**A_f=A_g=A_a**) and are correlated by **cosβ** where β is the angle between **f** and **g**.

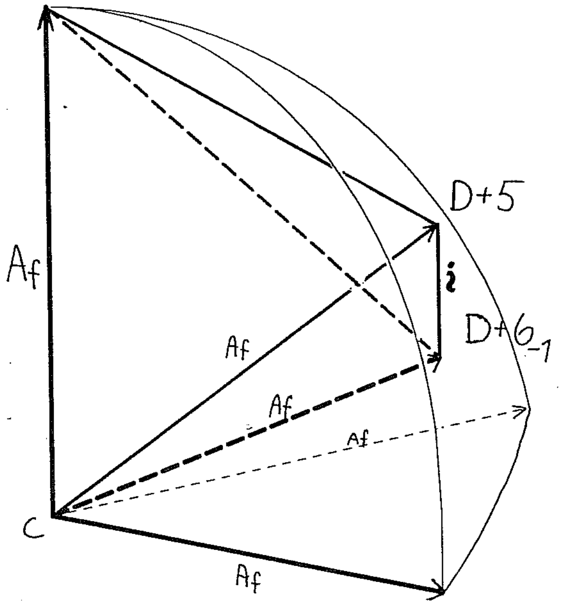

Fig. 6 A schematic vector representation of two forecasts **f** and **g** and their error vectors **f-a** and **g-a**. The forecast difference or "inconsistency" is the vector **f-g**.

The vector difference **f-g** is the forecast difference, inconsistency or "spread" and we want to know how it correlates with the error vector **f-a** ≤ **g-a**. The correlation is represented by the cosine of the angle α between vectors **f** and **g**.

If the difference in error between **f** and **g** is of the same magnitude as the inconsistency **f-g** it is obvious that α is close to 90 deg. and might even take values > 90 deg (fig.7a). This explains the low values of the "consistency-skill" correlations which have been found over the years.

The angle α can be decreased (and the consistency/skill correlation increased) in two ways: by making the two forecast of equal skill by improving **g** (fig.7b) and by making **f** and **g** less correlated (angleβ,in fig.6 increases) and in consequence the errors of **f** and **g** less correlated (fig. 7c).

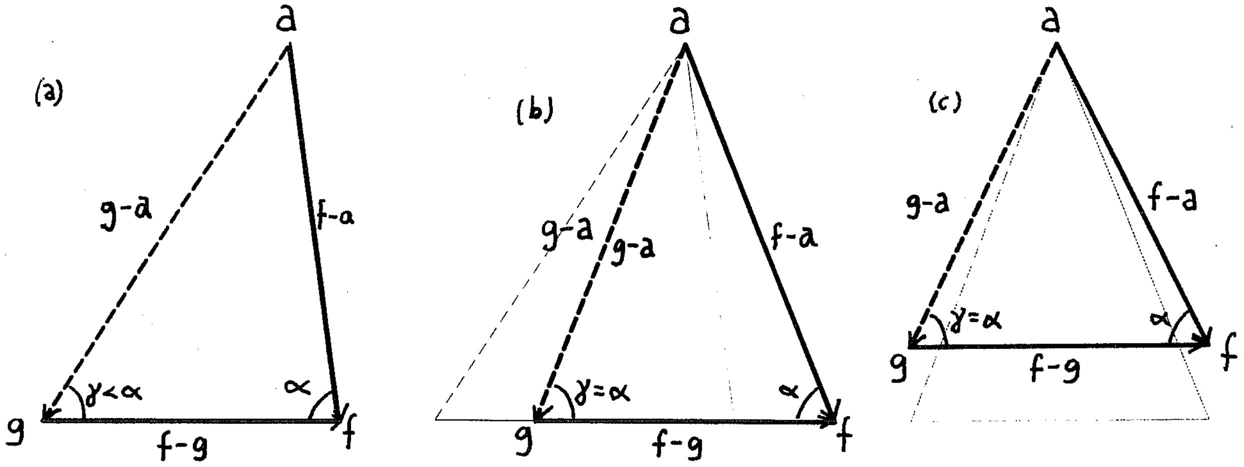

Fig.7: The three vectors **f-a**, **g-a** and **f-g** defining the consistency-skill relation in fig.6. Fig 7 a) represents the most common situation, b) when both forecasts are equally skilful and c) when the forecasts and their errors become less correlated.

3.4 The correlation between inconsistency and error

Some investigations in the 1980's found that the inconsistency **f-g** correlated strongly (around 0.7) with the error of the *previous* forecast **g-a**. There were even ideas that this could be used operationally until it was realized that having two consistent, but rather similar forecasts would not be of any help, even if one was probably better than usual. This correlation is defined by **cos** γ where γ is the angle between **g-a** and **f-g** (fig. 7). Since **g-a** > **f-a** it follows that α > γ and thus **cos** γ > **cos** α. It is easy to realize that the worse the forecasts **g** are, the smaller γ and the higher the correlation.

The same applies to two forecasts from different models: the errors of a generally worse model may be assessed from its similarity with the forecast from a better model, but not vice verse. Intuitively this can be compared with *calibration* of instruments: if a reading from crude instrument agrees with a more high-tech, it is rather the crude one that is performing better than normal.

3.5 The spread-skill approach in the EPS

This analysis can also be applied on the Ensemble Prediction System (EPS), where the spread of the ensemble is assumed to provide a measure of the probable error of the non-perturbed forecast. The EPS currently consists of one unperturbed analysis and 32 slightly perturbed in dynamically sensitive areas where shortage of observations may cause errors in the initial conditions. All these 33 analyses are then run on a T63L19 version of the ECMWF operational T213L31 global model. As with the inconsistency/error relation in the deterministic forecast system there exists in the EPS a "spread of the ensemble/error of Control forecast" relation. The conclusions drawn above for the deterministic case thus also applies for the EPS. If **f** is considered as the unperturbed "Control" forecast and **g** one of the EPS members, then an increase in the spread/skill relation would be achievedin two ways:

a) There should be no difference in overall skill between Control (**f**) and any of the members (**g**). With a more realistic model than the current T63L19, with more realistic perturbations and more advanced singular vector calculations, the overall forecasts error of an average ensemble member will decrease and the difference to Control will ultimately vanish.

b) Control and the ensemble member should become less correlated. Improvements in the model will work in this direction. The singular vectors are per definition orthogonal and thus uncorrelated, but since the perturbation patterns are computed from combining singular vectors the imposed perturbations will not be uncorrelated.

3.6 The consistency-skill equation

The discussion above can be nicely summarized using a theorem derived 1500 years ago by the Hindu astronomer and mathematician Aryabhata (476-550) and more commonly known as the Second Law of Cosine. For a triangle with **a**, **b** and **c** as sides and the angles opposite **a** being α then (fig.8)

$$a^2=b^2+c^2-2bc\cos\alpha$$

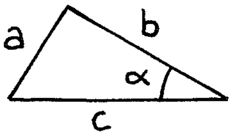

Fig. 8 The Second Law of Cosine

Applied on the triangle in fig.7 we get an expression for the correlation between **(f-a)** and **(f-g)**

$$\cos\alpha = \frac{(f-a)^2-(g-a)^2+(f-g)^2}{2(f-a)(f-g)}$$

from which follows that what will increase the spread/skill correlation is an improvement of **g** and/or an increased spread **(f-g)**.

4. Optimum use of medium range guidance

Weather forecasting, whether by subjective-empirical, statistical-climatological or dynamic-numerical methods is in essence a matter of estimating parameters, although they occur in the future. To the forecaster's disposal is a wide range of information sources, for the medium range in particular the ECMWF forecast products for the last days. A general rule in estimation techniques is that the best result is obtained when different sources of information are combined, instead of relaying only on the source which is attributed the highest quality.

This estimation technique, making the best estimate of the future state by a method of *combination*, instead of *comparison* where the approach is instead to try to determine, *which* of the available sources of information that at a specific time may provide the best estimate of the future state. Many forecasters have a feeling that they are able to tell which of several forecast alternatives is the best. But the "forecast-forecast skill" correlation is generally low, 0.3 at most. He might be able to "forecast the error", but on a higher error level. A better estimate of the future than the latest ECMWF forecast, the one that will provide the least error, can in a trivial way be achieved by weighting together the last 2-3 ECMWF forecasts; in an optimum way through the EPS.

6. Summary

This presentation has intended to illustrate that statistical results can not always be taken at face value and that simplistic interpretations can have negative consequences. "What looks good might be bad, what looks bad might be good" (Tim Palmer, personal communication):

-Comparing the scores of different NWP models using RMSE or ACC must be made on the condition that they have the same model climatology as the real atmosphere's.

-Inconsistency is an unavoidable consequence of a non-perfect forecast system. There is no straightforward relation between inconsistency and error, nor a level for "acceptable" inconsistency.

Considering the difficulties to interpret the verifications from the traditional deterministic forecast systems, one can only imagine the difficulties and potential pitfalls in relation to ensemble forecasting:

-An ideal EPS should be run on a model with a correct and stable climate, with the same overall skill of the perturbed and unperturbed analyses and forecasts, and with uncorrelated perturbations.

-If more than one forecast is available, either from previous runs, other models or from a EPS system, the optimum use is to weight together the information into one single forecast, or express the probability for some significant alternatives or intervals.

-The forecasters may display some skill in "forecasting the skill" or in "choosing the best forecast", but since the resulting forecast nevertheless does not provide the optimum forecast information, the benefit for the end-user is not optimal.

The introduction of medium range forecast guidance has transformed the rôle of the forecaster and confronted him with new challenges (Persson, 1993). Although the Ensemble Prediction System offers a very promising way ahead for the development of medium range weather forecasting, the scientific and operational complexities will take time to comprehend, further development and operational use will also demand a deeper understanding of the statistical nature of the forecast problem.

References:

Persson, A., On the operational use of ECMWF forecast products. In: Proceedings of the 1993 ECMWF Workshop on Meteorological Operational Systems, 22-26 November 1993, pp.116-123.

Persson, A. and Strauss, B., 1995: On the skill and consistency in medium range weather forecasting, ECMWF Newsletter, No 70, Summer 1995, pp. 12-15.

Persson, A., 1996a: On the consistency/skill and spread/skill relations in medium range weather forecasting. 13th Conference on Probability and Statistics in the Atmospheric Sciences, San Francisco 21-23 February 1996.

Persson, A.,1996b: On the statistical relationship between consistency and skill in medium range weather forecasting. ECMWF Operations Department Technical memorandum (under preparation).

Simmons, A.J., Mureau, R. and Petroliagis, T., 1995: Error growth estimates of predictability from the ECMWF forecasting system, Quarterly Journal of the Roy. Met. Soc. Oct. 1995 A, pp. 1739-1772.

Simmons, A.J. 1996: The skill of ECMWF 500 hPa height forecasts. In: Proceedings of the 1995 ECMWF Seminar on Predictability, in press.