FORECASTING FORECAST SKILL, PROBABILITY FORECASTING, AND THE PLAUSIBILITY OF MODEL PRODUCED FLOW

H.M. van den Dool
Climate Analysis Center/NMC
Washington, DC, USA

1. INTRODUCTION

We will discuss and review a few issues here. First, can we understand/interpret our attempts to forecast forecast skill as an attempt to measure the amount of low frequency variability present in the atmosphere at t=0? Second, perfect model experiments point to the fact that spread and skill are at best 0.7 correlated. We will offer an explanation as to why the correlation is less than unity in perfect model experiments. Finally, we will argue that probability forecasting is a more appropriate way of dealing with uncertainty than forecasting forecast skill.

2. FFS AND LOW FREQUENCY (LF) VARIABILITY

The question raised here is: Can we interpret forecasting forecast skill (FFS) signifying that the limit of predictability (LOP) varies from day-to-day as a function of the initial condition (IC)? It has been proposed in turbulence theory that the LOP is proportional to the time-scale of variation, or in other words systems (such as clouds, cyclones, weather regimes) are predictable over a time related to their life time (i.e. hours, days, and weeks respectively). While each phenomenon has its own predictability time-scale, the overall LOP (the time at which the RMS error reaches x%, x being for instance 90% of saturation) will be a function of time (function of IC) in the following way:

IC with strong LF components: long LOP, good FFS capability
IC with weak LF components : short LOP, poor case for FFS

We then have to ask the question: How does one measure the strength of the low frequency components of atmospheric flow at a given instant? Before discussing how FFS relates to this question, two points aside. Aside 1: it is a non-trivial problem to extract from data at one time level the power in the low (or any) frequencies. Aside 2: empirical long-range forecasting is often reduced to the question: which of the anomalies currently present in the atmosphere are likely to be long-lived? We tend to answer this question by constructing a time mean map (which is backward looking from t=0 by necessity) and by consulting a data library (spanning as many years as possible) to see in which areas time averaged anomalies tend to persist at this time of year.

Following is a list of ingredients commonly used in FFS. We then add comments as to how this ingredient relates to measuring low frequency variability at t=0.

1. Use a NWP model to calculate forecast persistence.
   This was proposed by *Chen* (1989) and has been used since in operational FFS schemes (see *Kalnay* (1991) and *Akesson* (1991), both in this volume). The idea is very simple. Let the NWP model determine whether persistence is high/low and expect the skill of the forecasts to be high/low.

2. Project the IC and/or forecast made by NWP onto EOF's (*Tribbia*, 1991, this volume). Although not by design, the lower order EOF represent the lower frequency components of atmospheric flow. The higher the projection onto the first EOFs the higher the skill is expected to be.

3. Along the same lines as 2) is the use of indices, such as PNA (*Palmer*, 1988; *Tracton et al.*, 1989) or the zonal index. Both indices characterize low frequency aspects of atmospheric flow.

4. Anomaly amplitude (*Branstator*, 1986). Simply because most of the variance in height resides at long time scales, we can only have large forecast anomalies if potent low frequency variability is in place.

5. "Spread" among different forecasts. It depends on how the ensemble is put together, whether or not spread measures low frequency variance.

All ingredients 1-5 appear to be measuring low frequency variability in one way or another. Items 3 and 5 do possibly more than that. The distinction between positive/negative values of the PNA index as a FFS tool goes beyond a simple statement of persistence of the initial flow. If an ensemble of IC's is constructed by using "optimal modes", we probably do not measure low frequency variability.

If one is willing to wait until months after t=0, one can do a temporal Fourier analysis to decide how strong the low frequency modes were at t=0. Following the method described by *Van den Dool and Saha* (1990) it was indeed shown that low frequencies have predictive skill over much longer times (10 days) than the high frequencies (4 days). There is little doubt that current FFS schemes, which generally use ingredients 1, 4 and 5 (plus other), measure low frequency variability. Assuming that non-changing flow is easier to forecast FFS schemes must have some success. The larger challenge is to specifically address the predictability of these IC's where large changes are imminent.

3. PERFECT MODELS AND AN UPPER LIMIT TO THE SKILL OF FFS

"Perfect" model experiments have been done to investigate how skillful our FFS schemes could be under the best of circumstances. This was done as follows (*Barker*, 1991). A very long control run was made with a GCM. From this run IC's were picked. Relative to each IC an ensemble of 8 was created by randomly perturbing the IC. Each member was integrated out to 60 days, using the same model. Other than the ensemble size, and the method of perturbing the IC, one has to accept this as a perfect setting. *Barker* (1991) found that even in this perfect setting spread, as measured by RMS differences among ensemble members, and skill, measured as the RMS difference of forecast and control, correlate only 0.6-0.7 at best. This is definitely a disappointment. Some of the operational FFS schemes have occasionally already approached this level of skill.

Another perfect "model" experiment was reported by *Van den Dool* (1989), using a limited area analogue approach. From his experiment it is clear, see his Fig. 9, that spread and skill (both measured by RMS) at 12 hrs are linearly related. But again the correlation was far from unity. The following is an attempt to explain why the correlation between spread and skill can never be perfect. In Fig. 1 we have sketched the initial cloud, and the cloud of solutions in phase space after 5 days. We assume a perfect model. Each member of the ensemble is equal in the sense that there is no a priori way of telling which is the best forecast. Arbitrarily we name one of the forecasts "Control", another "Verification". Generally the forecasts are denoted by F_j. It is easy to see that

$$\text{skill} = <(C - V)^2>^{1/2}$$

and

$$\text{spread} = <(F_i - F_j)^2>^{1/2}$$

are measures of the same quantity, namely the size of the cloud of solutions (the brackets refer to summing over space). This is, of course, the intuitive basis for FFS. Being a measure of the same thing does not mean that skill and spread are numerically equal. The problem is that skill is based on a sample size of one, and spread on a sample size N. In fact spread is a far more accurate measure of the size of the cloud than skill. With some irony one can say that there is not enough reality. (Combining many cases in bins,

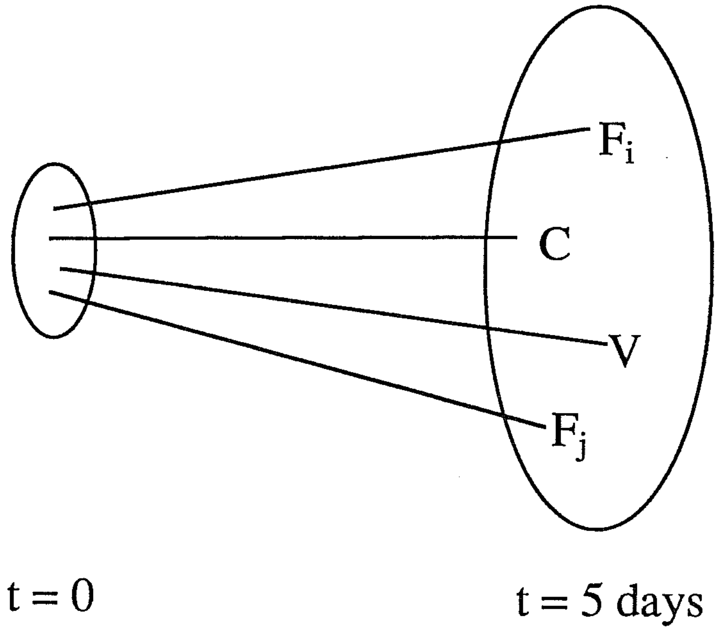

Fig. 1 A schematic of an initial cloud of possible states evolving to a cloud of solutions after some time. C and V stand for Control and Verification, F_j for a member of the ensemble.

Van den Dool (1989) did indeed find an almost perfect correlation between spread and the average skill of many forecasts.)

4. FFS OR PROBABILITY FORECASTING?

Let us consider an extremely simple hypothetical example of a probability forecast. A meteorological element is categorized into two classes, A(bove normal), and B(elow normal), with climatological probabilities for each class 50%. Following is a Table expressing 3 kinds of forecast.

B	A	Type of forecast
50%	50%	climatology
100%	0%	deterministic
0%	100%	forecasts
x%	100-x%	probabilistic

A deterministic forecaster will make categorical statements, either A or B will happen. A probability forecast assigns probabilities to each class, and with increasing abs(x-50) one can express increasing confidence that the more likely class will indeed happen.

There is a subtle difference between FFS and probability forecasting. If one recognizes that the forecast is not deterministic, there is no reason why we should want to forecast the skill of a single forecast. Rather one should go all the way by making an ensemble of forecasts spanning hopefully the whole space of possible solutions, and base a probability forecast on it. Having N "equal" forecasts in hand the question about the skill of forecasting forecast skill no longer arises. Suppose x=70% in the above Table. We could retrieve a FFS statement by noting that the chance that any of the individual forecasts (making up a large ensemble) landing in the B class is right equals 70%.

The difference between FFS and probability forecasting is more than semantics considering the operational practice we have had ever since NWP was introduced. At all major centers the practice has been to make a single run at the highest affordable resolution. In this framework it would make sense to forecast the skill of that single forecast. The alternative is to run the model many times at lower resolution (or less expensive physics) and give a firm base for probability forecasting.

Probability forecasts are very old and have been pioneered by E. Epstein, A. Murphy, and D. Gilman, to mention a few. In the past they were made by a single model run, augmented by the track record of the model in the past months or years. For instance if the model was right 80% of all cases when the single forecast landed in the A class, the probability forecast would be B: 20%, A: 80%. While this is better than a deterministic forecast (denying uncertainty), it is only a first approximation. Probabilities should not be based only on past performance, but also on the IC at hand, and that can only be done by multiple runs.

ACKNOWLEDGEMENT

I thank Steve Tracton for his input and for reading this manuscript.

REFERENCES

Akesson, O., 1991: A statistical technique for skill prediction (this volume).

Barker, T.W., 1991: The relationship between spread and forecast error in extended range forecasts. J.Climate, 4, 733-742.

Branstator, G., 1986: The variability in skill of 72-h global scale NMC forecasts. Mon.Wea.Rev., 114, 2628-2639.

Chen, W.Y., 1989: Another approach to forecasting forecast skill. Mon.Wea.Rev., 117, 427-435.

Kalnay, E., 1991: Forecast of the skill and extended-range prediction at NMC. (This volume)

Palmer, T.N., 1988: Medium and extended range predictability, stability of the PNA mode, and atmospheric response to sea surface temperature anomalies. Quart.J.Roy.Meteor.Soc., 114, 691-713.

Tracton, M.S., K. Mo, W. Chen, E. Kalnay, R. Kistler and G. White, 1989: Dynamical extended range forecasting (DERF) at the National Meteorological Center. Mon.Wea.Rev., 117, 1604-1635.

Tribbia, J. 1991: Monthly and seasonal prediction and predictability experiments using the NCAR climate model. (This volume)

Van den Dool, H.M., 1989: A new look at weather forecasting through analogues. Mon.Wea.Rev., 117, 2230-2247.

Van den Dool, H.M. and S. Saha, 1990: Frequency dependence in forecast skill. Mon.Wea.Rev., 118, 128-137.

Van den Dool, H.M. and Z. Toth, 1991: Why do forecasts for "near normal" often fail? Wea.Forecasting, 6, 76-85.