SIGNIFICANCE TESTING FOR FORECAST SKILLS

There is a recurrent need to compare the performance of forecast results from different analysis/forecast systems. First consider forecasts generated from a particular forecast system $\mathbf{F}$; most often, a series of forecasts $\mathbf{f_n}$ having a specified length (typically 5-days for GEOS) are produced from a set of initial (usually analyses) conditions. The forecasts are saved at discrete time intervals t_m, with the t=0 (initial condition) used as verification for each forecast. A sensitive measure of the skill of a forecast a time t_m is the anomaly correlation:

$$AC_n(t_m) = \frac{\langle (a_n va_n) \rangle}{\left[\langle {a_n}^2 \rangle \langle (va_n)^2 \rangle \right]^{1/2}} \quad ,$$

where, for a climatology C, the following are defined:

$$\begin{aligned} a_n &= f_n(t_m) - C \\ va_n &= f_n(t = 0) - C \end{aligned}$$

This process thus generates n m-element sequences of the spatial anomaly correlation statistics for the n forecasts from system F. These data are frequently displayed as "time-series plots" of the anomaly correlations. A robust estimate of the mean of the n-elements of the AC for a given m employs the Fisher Transform (here the "traditional" symbol for correlation ρ will be used for AC:

$$Z_{n,m} = \frac{1}{2} \frac{log(1 + \rho_{n,m})}{log(1 - \rho_{n,m})}$$

the mean of $Z_{n,m}$ is generated in the standard way, and the the inverse is performed to obtain ρ_m:

$$\begin{aligned} Z_m &= \frac{1}{N} \sum_{n=1}^{N} Z_{n,m} \\ \rho_m &= \frac{exp(2Z_m) - 1}{exp(2Z_m) + 1} \end{aligned}$$

The ρ_m are the standard "decay" curves generally shown for 500 hPa geopotential heights. Note, in practice a tiny non-zero term is added to the denominator of the above Fisher Transform in order to guard against any divisions by zero.

It is frequently the case that the forecast skills ρ_m from one system are to be compared with those from forecasts from other centers, or with forecasts from a modified version of that system. In this situation, the statistical machinery of hypothesis testing is invoked to test the null hypothesis that the mean ρ_m from the test system is statistically indistinguishable from the ρ_m from a different system. A further refinement to this testing process is based on the notion that all the forecast sequences in question are run from the same starting dates; thus the forecasts in question should contain the same underlying dynamics. This assumption allows for the use of *paired difference testing* (see von Storch and Zwiers. pp. 113-114). This approach tests on the null hypothesis that the difference of the means is zero.

For the purposes of the following discussion, consider two n-element ensemble forecast runs, with m saved forecast states (m_{tot}=11 for 5-day forecasts saved every 12 hours): $f^1_{n,m}$ and $f^2_{n,m}$. Define a Z-transform for the difference statistic:

$$\delta Z_{n,m} = \frac{1}{2} log \left(\frac{1 + \frac{1}{2}(\rho^{(1)}_{n,m} - \rho^{(2)}_{n,m})}{1 - \frac{1}{2}(\rho^{(1)}_{n,m} - \rho^{(2)}_{n,m})} \right) \quad .$$

Now generate the usual means and variances (over n) for $\delta Z_{n,m}$: μ_m and V_m. If the N members of the forecast ensembles are independent (likely a rash assumption), then the *degrees of freedom* or "dof" for this situation is N-1. The 90% two-sided t-distribution critical value for dof is obtained using GrADS functions "ASTUDT" and "ASTUDTOUT", and will be called "critval" here. The hypothesis test here then becomes:

$$\mu_m \leq critval \sqrt{\frac{V_m}{dof}} = \delta Z_c$$

If this inequality is met, then the difference mean is indistinguishable from zero to this level of confidence. For plotting purposes, these quantities are transformed back into "correlation space":

$$\begin{aligned}
\Delta\rho_m &= 2\frac{exp(2\mu_m) - 1}{exp(2\mu_m) + 1} \\
\rho^{upper}_{crit} &= 2\frac{exp(2\delta Z_c) - 1}{exp(2\delta Z_c) + 1} \\
\rho^{lower}_{crit} &= 2\frac{exp(-2\delta Z_c) - 1}{exp(-2\delta Z_c) + 1} \quad .
\end{aligned}$$

$\Delta\rho_m$ needs to be outside of the boxes defined by ρ^{upper}_{crit} and ρ^{lower}_{crit} for the anomaly correlation mean of $f^{(1)}_m$ to be considered significantly different from that from $f^{(2)}_m$.

REFERENCE

von Storch and Francis W. Zwiers, 1999, "Statistical Analysis in Climate Research", Cambridge University Press, 484pp.