errors of user-orientated forecasts *should not* exceed the $E_{climate}$ level whereas forecasts from good NWP models at some range *must*.

A-2.5 Error saturation level

Forecast errors do not grow indefinitely but asymptotically approach a maximum, the "Error Saturation Level" (ESL).

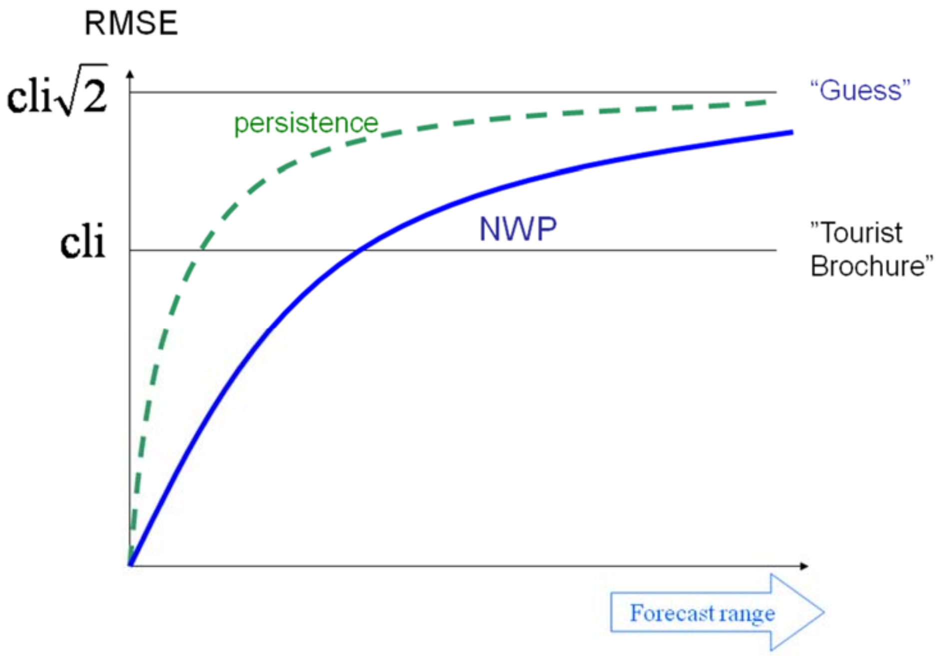

Figure 67: The error growth in a state-of-the-art NWP forecast system will at some stage display larger errors than a climatological average used as forecast and will, as do the errors of persistence forecasts and guesses, asymptotically approach an error level 41% above that of a forecast based on a climatological average

For extended forecast ranges, with decreasing correspondence between forecast and observed anomalies, the covariance term approaches zero. For A_f=A_a this yields an ESL at

RMSE = $E_{saturation}$= $A_a\sqrt{2}$

which is 41% larger than $E_{climate}$, the error when a climatological average is used as a forecast (see Figure 67). The value $A_a\sqrt{2}$ is also the ESL for persistence forecasts or guesses based on climatological distributions.

A-2.6 Measure of skill - the anomaly correlation coefficient

Another way to measure the quality of a forecast system is to calculate the correlation between forecasts and observations. However, correlating forecasts directly with observations or analyses may give misleadingly high values because of the seasonal variations. It is therefore established practice to subtract the climate average from both the forecast and the verification and to verify the forecast and observed *anomalies* according to the anomaly correlation coefficient (ACC), which in its most simple form can be written:

$$ACC = \frac{\overline{(f-c)(a-c)}}{\sqrt{\overline{(f-c)^2}\ \overline{(a-c)^2}}}$$

The WMO definition also takes any mean error into account:

$$ACC = \frac{\overline{\left[(f-c)-\overline{(f-c)}\right]\left[(a-c)-\overline{(a-c)}\right]}}{\sqrt{\overline{\left((f-c)-\overline{(f-c)}\right)^2}\;\overline{\left((a-c)-\overline{(a-c)}\right)^2}}}$$

The ACC can be regarded as a *skill score relative to the climate*. It is positively orientated, with increasing numerical values indicating increasing "success". It has been found empirically that ACC=60% corresponds to the range up to which there is synoptic skill for the largest scale weather patterns. ACC=50% corresponds to forecasts for which the error is the same as for a forecast based on a climatological average, i.e. RMSE = A_a. An ACC of about 80% would correspond to a range where there is still some skill in large-scale synoptic patterns.

A-3 Interpretation of verification statistics

The mathematics of statistics can be relatively simple but the results are often quite difficult to interpret, due to their counter-intuitive nature: what looks "good" might be "bad", what looks "bad" might be "good". As we have seen in A-1.3, seemingly systematic errors can have a non-systematic origin and forecasts verified against analyses can yield results different from those verified against observations. As we will see below, different verification scores can give divergent impressions of forecast quality and, perhaps most paradoxically, improving the realism of an NWP model might give rise to *increasing* errors.

A-3.1 Interpretation of RMSE and ACC

Both A_f and A_a and, consequently, the RMSE vary with geographical area and season. In the mid-latitudes they display a maximum in winter, when the atmospheric flow is dominated by large-scale and stronger amplitudes, and a minimum in summer, when the scales are smaller and the amplitudes weaker.

For a forecast system that realistically reflects atmospheric synoptic-dynamic activity $A_f = A_a$. If $A_f < A_a$ the forecasting system *underestimates* atmospheric variability, which will contribute to a decrease in the RMSE. As discussed in Chapter 4, this is "bad" if we are dealing with a NWP model but "good", if we are dealing with post-processed deterministic forecasts to end-users. On the other hand, if $A_f > A_a$ the model *overestimates* synoptic-dynamic activity, which will contribute to increasing the RMSE. This is normally "bad" for all applications.

Comparing RMSE verifications of different models or of different versions of the same model is most straightforward when $A_f = A_a$ and the models have the same general variability as the atmosphere.

A-3.2 Effect of flow dependency

Both RMSE and ACC are flow dependent, sometimes in a contradictory way. In non-anomalous conditions (e.g. zonal flow) the ACC can easily take low ("bad") values, while in anomalous regimes (e.g. blocking flow) it can take quite high ("good") values. The opposite is true for RMSE, which can easily take high ("bad") values in meridional or blocked flow regimes and low ("good") values in zonal regimes. Conflicting indications are yet another