6. Degrees of Freedom

How many degrees of freedom are evident in a physical process represented by f(s,t)? In some form questions about 'degrees of freedom' (dof) are common in mathematics, physics, statistics, and geophysics. This would mean, for instance, in how many independent directions a weight suspended from the ceiling could move. Dofs are important for three reasons that will become apparent in the remaining chapters. First, dofs are critically important in understanding why natural analogues can (or can not) be applied as a forecast method in a particular problem (Ch7). Secondly, understanding dofs leads ideas about truncating data sets efficiently, which is very important for just about any empirical prediction method (Ch 7 and 8). Lastly, the number of dofs retained is one aspect that has a bearing on how non-linear prediction methods can be (Ch10).

In view of chapter 5 one might think that the total number of orthogonal directions required to reproduce a data set is the dof. However, this is impractical as the dimension would increase (to infinity) with ever denser and slightly imperfect observations. Rather we need a measure that takes into account the amount of variance represented by each orthogonal direction, because some directions are more important than others. This allows truncation in EOF space without lowering the 'effective' dof very much.

We here think schematically of the total atmospheric or oceanic variance about the mean state as being made up by N ***equal*** additive variance processes. N can be thought of as the dimension of a phase space in which the atmospheric state at one moment in time is a point. This point moves around over time in the N dimensional phase space. The climatology is the origin of the phase space. The trajectory of a sequence of atmospheric states is thus a complicated 'Lissajous figure' in N dimensions, where, importantly, the range of the excursions in each of the N dimensions is the same in the long run. The phase space is a hypersphere with an equal probability radius in all N directions. A similar plot in M dimensions (M>>N), where each

direction represents one EOF, would have the largest excursions for the gravest EOFs, and tiny excursions in a very large number of directions representing the higher EOFs, with an obvious truncation problem, M going to infinity as resolution goes to infinity. In a sense, N results from folding the energy in modes > N back into modes < N, which serves to make N finite as well as creating a white spectrum in which the retained modes count as exactly one degree of freedom each. This makes the N processes somewhat hypothetical, a summary of the whole spectrum of variations in a single number: *the **effective** degrees of freedom*. As we shall see N, consistent with Toth(1995), is on the order of 30-50 for hemispheric daily instantaneous 500 mb height. Is 30 large or small? Thirty effective degrees of freedom is very large, as it renders the search for natural analogues virtually hopeless (van den Dool 1994). On the other hand 30 may seem very small compared to the nominal degrees of freedom (millions) used in NWP global models (the # of variables times the # of gridpoints) - certainly we have a lot more spatial and/or temporal coherence than random processes at each point in time and space would have.

6.1 Methods to estimate effective degrees of freedom N

Methods of estimating N have been reported infrequently in the literature. Moreover, the context and methods differ from paper to paper. Panofsky and Brier(1968) report that when correlating two time series of length N (drawn at random from the same parent distribution) the distribution of the correlation coefficient between the two time series is Gaussian with a zero mean and a standard deviation of $1/\sqrt{(N-2)}$, where N is the number of independent data in the time series (N is less than the length of the time series if autocorrelation were positive). Van den Dool(1981) extended this concept to maps (a string of data of length M) in order to test for statistical significance the spatial correlation between two anomaly maps (elements of the correlation rendition of Q^a). It was noted that N in this case is the number of processes going on independently in space, i.e. much less than the number of gridpoints M because of correlation in space. If one correlates anomaly maps that should not be related (by virtue of them being years

apart) one can retrieve N from an empirical pdf of correlations by inverting the expression $sd_{cor}=1/\sqrt{(N-2)}$. Following this approach we ask nature to conduct its own Monte Carlo experiment. Livezey and Chen(1983) had a similar purpose in mind when designing their famous field significance test. They integrated the spatial correlation (elements of Q) in space in order to find the integral space scale (S). S relates to N roughly as N≈domain size/S. This method is common in fluid dynamics and turbulence. Both Van den Dool (1981) and Livezey and Chen(1983) were concerned with the validity of claims of forecast skill by various methods. If someone's forecast anomaly map correlates 0.4 to the observed anomaly map, does that imply the forecast method is better than a random guess? Given N we at least can perform a statistical significance test.

Lorenz(1969) considered mean square differences (msd) between states in the atmosphere and, as explained in Toth(1995), a collection of msd should obey the chi squared distribution with N degrees of freedom. From an empirical collection of msd one thus needs to find in some way the chi-squared pdf that gives the best fit, then retrieve N by an 'inverse' method.

All these types of estimates of N (via msd, correlation), which work out numerically similar, but not identically the same, are closely related to classical notions of effective sample size, be it in time or space, or the de-correlation distance. Wang and Shen(1999) describe a more complete comparison of such methods, including theoretical background, and numerical estimates of N by each method on a given data set. Other aspects can be found in Fraedrich et al(1995) and Stephenson(1997).

6.2 Example

Here we report on a fresh example using the global Reanalysis data for daily Z500 data at 0Z on the domain 20 degrees latitude to the pole for both hemispheres, as well as for the tropics between 20N and 20S. These are large areas in terms of km^2. We consider each calendar month separately. We first take out a smooth daily Z500 climatology, forming anomalies. We correlate

all fields in, say, January, to all other fields in January but only in non-matching years. Therefore, the expected value of the covariance defined in Eq (2.14a) as

$$q^a_{ij} = \sum_s f(s, t_i)\, f(s, t_j) / n_s$$

is zero. q^a_{ij} are elements of the covariance matrix Q^a , but we consider only those elements that have a large separation in t_i and t_j . Due to sampling q^a_{ij} is not always exactly zero but has a spread measured by its standard deviation. In terms of correlation defined as per Eq (4.1) as:

$$\rho_{ij} = q^a_{ij} / \mathrm{sqrt}(q^a_{ii}\, q^a_{jj}) ,$$

the expected value is 0, and the spread of ρ_{ij} , as per Gaussian theory (Panofsky and Brier 1968), is 1/sqrt(N-2), where N is the effective degrees of freedom in the field f. If all M gridpoints had statistically independent equal variance processes going on, N would be M. But because of spatial correlation in the field f(s,t) the value of N is (much) less than M.

From the collection of the K correlations (K is very large, about a million for 30 years of data) we calculate the standard deviation of the empirical correlation distribution (ECD) as

$$sd_{cor} = \{ \Sigma\, \rho_{ij}^2 / K \}^{1/2} \quad (6.1)$$

where summation is over all admitted ij pairs and determine N as:

$$N = 1/sd_{cor}^2 + 2. \quad (6.2)$$

This procedure makes sense only for N > 3, and generally only for large N. For small N, the ECD would feel the boundaries at +1 and -1.

Fig.6.1 shows how N depends on season in the two hemispheres. Obviously N is more or less proportional to the size (in km^2) of the domain (20 to the pole), but because weather systems have different horizontal length scales in the two hemispheres and in winter versus summer there is clear variation in N. As shown by the lower curves in Fig. 6.1, N in the NH varies from about 30 in winter to 45-50 in summer. In the SH, N is relatively constant at about 30 with little or no seasonality (presumably because the oceans are a very large fraction of the SH). Except for

January the NH has always higher N than the SH. In summer weather systems in the NH are indeed smaller than in winter.

The upper set of curves in Fig. 6.1 are the radii of the N dimensional hypersphere given by

$$sd = (\sum_{s,t} f(s, t)\, f(s, t) / n_s * n_t)^{1/2} \qquad (6.3)$$

This radius or climatological standard deviation is over 100gpm in winter (in the respective hemispheres) but is as low as 65 in summer in the NH, (a lesser factor that also contributes to a high estimate of N given imperfect analyses. Summer has less of a signal.) The SH during summer maintains a higher standard deviation of at least 85gpm.

For the tropics (TR- not shown), the domain within 20 degrees from the equator N is around 30 also, with small irregular seasonality. So we have three domains (NH, SH, TR) with similar N, except the NH summer which has higher N. The sd according to (6.3) is only 20 gpm for TR.

An estimate of N for three dimension, all variables combined appears to be order 1000 (Toth 1995).

6.3 Link of degrees of freedom to EOF

Because EOFs do not have equal variance one may wonder how N relates to EOFs. It has been suggested recently that N can also be obtained through an integral involving the decrease of variance with EOF mode number (Bretherton et al 1999):

$$N = \frac{(\sum_k \lambda_k)^2}{\sum_k \lambda_k^2} \qquad (6.4)$$

where λ_k is the variance explained by the kth ordered EOF, or the kth eigenvalue of the covariance matrix. For most typical decays of eigenvalue with mode, the value of N equals

approximately the number of EOFs needed to describe 90% of the variance. So, as a rule of thumb, the value of N relates approximately to a tic mark on the cumulative variance versus EOF mode# graph (Huang et al 1996). For instance Fig 5.7 is such a graph and it would appear that seasonal mean Z500 in JFM (NH) have about 20 degrees of freedom. This number is lower than the N~30 in Fig 6.1 because of temporal averaging (3 month means), which filters out many smaller scale and high frequency weather systems and so decreases N.

For large N Eqs (6.4) and (6.2) give very similar numerical answers, with (6.2) being far easier to execute, so we used Eq (6.2) in the above. However, for interpretation's sake, Eq(6.4) is more helpful, especially it shows the relationship to EOFs. For instance, if there was only one EOF N should be 1 according to (6.4), a most reasonable result. If the atmospheric variability is represented by K equal variance EOFs, (6.4) reduces to N=K, which is exactly the interpretation of N we are seeking.

Note that (6.4) is more generally valid than (6.2) because it also applies for low values of N. Recently (6.4) has been independently proposed by Patel et al (2001) in the context of estimating the (sometimes low) dimension of an ensemble of forecasts - the name given by Patel et al being BV- or E-dimension.

Comparisons of N in atmospheric analyses and model generated data have been made since Van den Dool and Chervin (1986) - there are discrepancies, especially as the forecast lead increases, but there is broad agreement, even in the lower resolution models used in the 1980s. We made a recent calculation of N (for Z500 20-pole) from 5 years of 'reforecast' data (Schemm pers comm) in NH winter. Fig.6.2 shows that despite a few curious ups and downs, N is very nearly correct out to 30 days. This does not prove that each of the 30 or so independent processes is identically the same as in nature, although study of rotated EOFs appears to indicate mode-by-mode similarity for at least the first 6 EOFs (Peng 200.). In general, modern day atmospheric models appear to be close copies of nature on the resolved scales. In addition to N, the standard deviation around the mean, see Eq.(6.3), should also be maintained, see the dashed

line (labeled AMP) in Fig.6.2. Here too models are quite close to nature. Perhaps 30 days is not quite enough to arrive in the model's true climate, and the small ups and downs in Fig.6.2 are part of transitioning from nature to model climate.

6.4 Remaining questions

Some methods work with correlation in time, others with correlation in space. Possibly there are two values of N, depending on whether one has the regular or the reverse space time point of view (of the same data set), in the same way we have EOT and alternative EOT. The number of spatially independent processes does not(?) need to be the same as the number of temporally independent processes.

Another puzzling feature is that (6.2) is evaluated by correlating fields that, by construction, have an expected correlation 0 (by being far removed in time), while EOF calculations (followed by (6.4)) are done inclusive of correlation between neighbors in time. By including neighbors in time (today and tomorrow) the ECD would have a 2ndary maximum at +0.7 for daily 500mb height anomaly due to persistence, and N can no longer be retrieved as per (6.2) because the Gaussian assumption is violated. A complimentary thought is as to how EOFs would be calculated if one deleted the correlation to neighbors in time or space from the covariance matrix - with EOT this might be a doable exercise. This puzzle is a companion to the distinction between Wallace and Gutzler(1981) type teleconnection (high correlation at remote distance) as opposed to EOF and EOT which place a premium on explaining variance both nearby (the bulk) and far away.

A third issue is that in most modern texts correlations would be z-transformed, i.e. $z = 0.5\log[(1 + r)/(1 - r)]$, where r is sample correlation, so the ECD based on z is more strictly Gaussian. In that case $sd_z=1/\sqrt{(N-3)}$ or even $sd_z=1/\sqrt{(N-4)}$ (Wang and Shen 1999). We believe we avoided the need for this complication by setting up nature's Monte Carlo test, such that the expected correlation is zero, and the resulting ECD can be symmetric around zero. The Z

transform is needed when the expected value is non-zero (and the ECD asymmetric).

A final issue to be mentioned is that N as derived here is for a single data set. There are applications when one wants to know how many dof two data sets have in common. Some of the references (Bretherton et al 1999) describe some attempts, but much more work is needed.

6.5 Concluding comments

A value for N of about 30 implies that atmospheric states picked at random rarely (5% of the time) correlate more than +/-0.38. This presents a bleak prospect for finding natural analogues or anti-analogues deserving of that name. Chances are a little better in the SH than in the NH, and especially poor in the NH during its summer. In Chapter 7 we will inspect the wings of the ECD to make further statements about naturally occurring analogues .