7. Analogues

In 1999 the earth's atmosphere was gearing up for a special event. Towards the end of July, the 500mb flow in the extratropical SH started to look more and more like a flow pattern observed some twenty two years earlier in May 1977. Two trajectories in the N dimensional phase space, N as defined in chapter 6, were coming closer together. Fig.7.1 shows the two states at the moment of closest encounter, with the appropriate climatology subtracted. These two states are, for a domain of this size, the most similar looking patterns in recorded history[1]. But are these good analogues? They do look alike, nearly every anomaly center has its counterpart, but they are certainly not close enough to be indistinguishable within observational error, the anomaly correlation being only 0.81. The rms difference between the two states in Fig 7.1 is 71.6 gpm, far above observational error (<10gpm). The close encounter did not make it to the newspaper and, more telling, not even to a meteorological journal.

The idea of situations in geophysical flow that are analogues to each other has always had tremendous appeal, at least in meteorology. Even lay people may comment that the weather today or this season reminds them of the weather in some year past. The implications of true analogues would be enormous. If two states many years apart were nearly identical in all variables on the whole domain (of presumed relevance) , including boundary conditions, then their subsequent behavior should be similar for some time to come.[2] In fact one could make forecasts that way, if only it was easy to find analogues from a 'large enough' data set. The analogue method was fairly widely used for weather forecasting at one time (Schuurmans 1973) but currently is rarely used for forecasts persé (for all the reasons explained in section 7.1). Rather analogues are used to specify one field given another, a process called 'specification' or downscaling, or to learn about

---

[1]Slight exaggeration. We looked at data for 1968-2004 only.

[2] One should not expect exact analogues, because exact analogues would imply a periodic system, hence perfect predictability. Rather, we have states in mind that are close initially, with increasing differences as time progresses.

predictability (Lorenz 1969). In section 7.1 we review the idea and limitations of naturally occurring analogues, and explain why/when it is (un)likely to find analogues. Finding analogues, as such, is a diagnostic problem. If no analogues deserving of that name exist in a finite data set, the application towards forecasting does not even arise. Section 7.2 develops the idea of 'constructing' an analogue in the absence of any natural analogues (NA). The constructed analogue overcomes the main problem one has with natural analogues, although at some cost, and appears to have forecast applications (7.4). In the process of constructing an analogue, we make an empirical operator which allows us to address the calculation of unstable modes from just observations (7.6), study weakly non-linear processes, and the dispersion of initial sources (much like in Figs 3.1-3.2), but in a more realistic way than EWP.

7.1 Natural Analogues (NA)

The working definition for the existence of natural analogues is that two states in a dynamical system are so close they can be called each other's analogue. One qualifier is that these two states should be far apart in time, well beyond the de-correlation time. Two successive states could obviously be very close to each other, especially when observed at high temporal resolution, but the application to forecasting of such look-alike temporal neighbors would be meaningless as they stay temporal neighbors forever. We mean two states far apart in time, which by sheer co-incidence[3] happen to be close. We are interested in 'how close' two states can be on a given data set of size M, how long they will stay close, how long they have been close (i.e. the ramping up to the closest encounter; symmetry in time or not) etc. And if we can't find any worthwhile analogues, why not? We are also interested in the most dissimilar flow patterns, the highest -ve correlation, as it throws light on issues of linearity and symmetry. Importantly, we remove

---

[3]One may always wonder about co-incidence or chance, or the alternative that something forces the flow to be similar on specifically those dates. We remove the daily and annual cycle from the data and, in doing so, remove the effect of the main periodic forcings known to us.

periodic components from the system (daily and annual cycle) before searching for analogues on the anomalies.

The age old idea of a forecast based on analogues is displayed in Fig.7.2. Today's point in phase space (or alternatively: today's weather map) is called the 'base', and has time t=0 assigned to it. Then our task is to make a forecast for what will happen next. We look for analogues, i.e. cases in the past[4] that are very close to the base. These cases have to be around the same time of the year or at least subject to the same general climatic conditions. Having found the appropriate cases we line up the time axes and assign t=0 to these analogues also. The string of realizations, or the trajectory in phase space observed in the analogue year, following the analogue at t=0 is the forecast for the conditions that follow the base. If one compares this process to NWP, one might look upon the analogue as an analysis of the base, with an initial difference or error, and nature itself as the model that carries out the integration of the equations. The advantage of analogues, apart from simplicity, would be the use of a perfect model[5]. The problem with analogues is that the initial difference cannot be made small unless we either have an inordinate amount of data or very few degrees of freedom.

7.1.1 Similarity measures

The measure of similarity or analogy for two anomaly 'maps' observed at t_i and t_j includes any of the following three expressions:

a) RMSD, or root mean square difference,

$$RMSD = (\sum \{ f(s, t_i) - f(s, t_j) \}^2 / n_s)^{1/2} \tag{7.1}$$

---

[4]There is no difference in the role of base and analogue. If a state at time t1 is very close to a state at t2, the reverse is also true. It is only in real time that we cannot search for analogues in the future.

[5]Assuming the atmosphere-ocean-land system has not changed over time. The continual change in atmospheric composition, change in land use etc does violate this assumption, but we ignore this complication.

b) the covariance (from 2.14a)

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

or
c) the correlation from (4.1)

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

The RMSD makes a lot of sense, although it may favor states that have small amplitude, amplitude defined as

$$AMP(t_i) = (\sum_s \{f(s, \ t_i)\}^2 / \ n_s)^{\frac{1}{2}} \qquad (7.1a)$$

(The footnote#1 in Ch 2 about weighting applies to (2.14a), (4.1), (7.1), (7.1a))

The covariance measure, the elements of the, by now, familiar Q^a, is made to regress each state towards the other so as to minimize their RMSD. We will use here the closely related correlation, but we will not actually modify states by regression. Obviously, a near perfect analogue has low RMSD, and both high covariance and correlation, so the measures should agree in the limit of very good analogues. Many other slight variations in measures of (dis)similarity can be found in Stephenson(1997). In the past circulation types (on the order of thirty) were used to quantify similarity (Schuurmans 1973).

In practice we will search for the nearest neighbor in the N dimensional phase space, and delay answering the question as to whether these nearest neighbors are worthy of the name analogues. The area over which candidate analogues are sought depends on the application. If one wants to compete with global NWP, f(s,t) should include all variables, including boundary conditions and s runs over a global domain. However, in anticipation of the meager results (Lorenz 1969; Ruosteenoja 1988), we scale down expectations considerably, and search for analogues on the area from 20 degrees latitude to the pole, and one variable only, daily Z500 at

0Z.[6] This exercise is thus the exact continuation of Chapter 6, where the empirical correlation distribution (ECD) was already studied for the same variable on the same domain.

7.1.2 Search for 500 mb height analogues.

As an example what an analogues search yields, we inspect here the wings of the ECD for naturally occurring analogues on the domain from 20 degrees latitude to the pole, and using just one variable (Z500). Data treatment was already described in Ch 6.2, i.e. we first take out a smooth daily Z500 climatology, forming anomalies. First let's focus on just one calendar month, January and one hemisphere, NH. For each date t_i in January there is one other date t_j in January (in a non-matching year!) for which ρ_{ij} is the maximum over all j, denoted $\max(\rho_i)$. Averaged over all 31 days and all 37 years we determined that for January (Z500, NH, 0Z data) $\max(\rho_i)$ is typically around +0.54. Generally the nearest state is not very near. At 0.5 correlation, the nearest state, (and this would be after applying a regression) is at about the distance to the climatology. The record highest $\max(\rho_i)$ value in January, 0.71, is between January 25 1976 and January 10 1977. While this is better than 0.54 it is only for a single occurrence, and it still is not good enough to be considered very close.

The most dissimilar state with the highest negative correlation is typically -0.51 in January NH. There is a little asymmetry (present in all months and in both hemispheres) implying slightly better analogues than anti-analogues, like +0.54 versus -0.51 in January, NH. This is because high pressure and low pressure systems are not exactly each other's opposite (relative to the climatological mean). Indeed a small skew has been noted (White 1980). The record lowest $\min(\rho_i)$ value in January is between January 10 1974 and January 28 2000 at -0.69.

Fig.7.3 shows the mean value of $\max(\rho_i)$ and, with sign reversed, $\min(\rho_i)$ Along with the absolute highest correlations for the NH in all calendar months. Here we searched only in the

---

[6]If forecasts applications were appropriate here at all, a hemispheric barotropic model is about the most comparable technology.

same month. The quality of analogs as measured by $\max(\rho_i)$ (and anti-analogs as well) is better in winter than in summer. The average correlation one should expect for the best analogue varies from 0.55 in winter to 0.45 in summer, essentially following the inverse of N, compare Fig. 6.1. The record best pair in each month has a correlation of about 0.7 in winter (one case > 0.75 in November) to only 0.56 in summer. These 'best' lines are more noisy than the average lines because they are based on just one such occurrence in each month - they are the absolute extremes in the ECD. Generally the values for analogs are very slightly better than for anti-analogs, both in the mean and the extremes.

Year-round results for the SH, not shown, are similar to those shown in Fig.7.3 for the NH winter. This is because N in the SH is about thirty in any season, see Fig. 6.1, i.e. equal to N in the NH's winter. In seasons other than winter the SH is the better hemisphere for finding (dis)similar patterns because its N is much lower. Smaller N allows wider ECD and therefore slightly higher chance of finding analogues. I.e the average correlation in the SH is between 0.50 and 0.55 year round, and the record best pair by month correlates at better than 0.70 year round, with several record cases >=0.75.

In the above we used data windows of just one calendar month in order to make sure any analogues are during the same time of year, under the same general conditions in terms of solar radiation, SST etc. However, in the SH winter, both N and the standard deviation of height fields are nearly constant from May thru September which is core winter, see Fig.6.1. Using this five month window we redid the analog search for the SH. The number of paired comparisons increases with the square of the length of the data set, and 25 times more CPU is used to check all pairs. This extended search netted a handful of additional pairs correlating in excess of 0.75. We also found 20 pairs of anti-analogs that correlate -0.70 or better (but only one beyond -0.75). A search for Nov through March in the NH yielded only one more case >=0.75 and the two states in this pair are within a month (but not in the same named month).

While some positive research elements have been noted in the above, the main message is

the complete lack of any cases that are truly close, a conclusion reached before by Lorenz(1969), Ruosteenoja(1988), Van den Dool(1994) and many others. In some ways these findings have killed all applications of analogues. In synoptic meteorology, a day-by-day forecast map has been deemed useful as long as the correlation with reality is >=0.6. So to find a pair correlated at 0.80 is better than nothing in terms of pattern similarity and indeed the two maps looks alike noticeably, see Fig.7.1. But in our judgement even a 0.80 correlation is not enough for two states to be called each other's analogue. This is even clearer when we further note that the rms difference for this pair (71 gpm) is still 70% of the climatological standard deviation (100 gpm) in SH winter. In general the distance to the closest neighbor is barely smaller than the standard deviation. I.e. the nearest neighbor is, in general, no closer than the origin (climatological mean). The phase space of N dimensions dotted with realizations from 1968-2004 is incredibly empty due to undersampling. We have barely begun to observe the atmosphere. Any attempt to study clusters, preferred or non-preferred circulation patterns (on the domain) or analogues will run into serious data limitations unless we can lower N while retaining a meaningful physical system. If we had an analogue pair at the 0.99 correlation level on a large domain, we could veritably test how well NWP, as of today, makes forecasts compared to a perfect model, since they would be starting from similar initial error magnitude.

The situation with respect to finding a natural analogue at a correlation >0.80, > 0.90 etc on the NH/SH can be described as 'possible but highly improbable' - a matter of waiting very long. The problem is not the idea about analogues, the problem is a lack of data in view of the many degrees of freedom in the system.

Table 7.2 shows how the quality of analogues and antilogues improves when the search area is reduced progressively. As the area becomes smaller N decreases as well. For an area as small as 500 by 1000km the analogues are near perfect.

Table 7.2 The average value of max(ρ_i), min(ρ_i) and the value of N for January, 1968-

2004, for progressively smaller areas in the NH.

Ana	Anti	N	area	# of gridpoints
53.6	-51.4	29.6	20N-pole, 0-360E,	29X144 gridpoints
69.7	-68.2	16.3	45N-50N, 0-360E	3 X 144 gridpoints
80.2	-78.3	11.4	45N-50N, 0-180E	3 X 72 gridpoints
90.7	-88.9	7.0	45N-50N, 0-90E	3 X 36 gridpoints
95.3	-94.6	5.3	45N-50N, 0-45E	3 X 18 gridpoints
99.1	-99.0	3.7	45N-50N, 0-10E	3 X 5 gridpoints

7.1.3 How long do we have to wait?

As reasoned in Van den Dool (1994) a three-way relationship can be derived between the size (M years) of an historical data set or library, the distance ϵ between an arbitrarily picked state of the atmosphere and its nearest neighbor, and the size of the spatial domain, as measured by the effective number of spatial degrees of freedom (N). This heuristic derivation proceeds from the following steps and assumptions.

i) We assume N equal variance (carrying sd^2 variance each) processes are going on independently.

ii) The probability υ of two arbitrarily picked states to be within an acceptably small distance ϵ (ϵ = 15gpm would obviously be very good in view of Table 7.1), for a single one of the N processes can be found by integrating a standard normal distribution from $-\epsilon/(sd*\sqrt{2})$ to $+\epsilon/(sd*\sqrt{2})$. We found $\upsilon \approx 0.08$ for ϵ = 15gpm. An 8% chance is not bad as it practically guarantees an analogue if one has say 100 to 1000 independent realizations from the past.

iii) The probability of finding two arbitrarily picked states within tolerance ϵ for all N processes simultaneously is υ^N.

iv) The probability c of finding an analogue in an M year library is $c=1-(1-\upsilon^N)^{20*M}$, where the number 20 refers to the number of independent cases in say a two month window (and M is in units of years).

v) Demanding c>0.95 leads to

$$M > \ln(0.05)/(20 * \upsilon^N) \qquad (7.2)$$

Thus it would take a library of order $M=10^{30}$ years in an unchanging climate to regularly find 2 observed flows that match to within current observational error over a large area, such as the

Northern or Southern Hemisphere, for just one variable. Van den Dool(1994) gives a table for a range of values of N and ϵ. Obviously, with only 10-100 years of data, the probability of finding natural analogous over a large area (large N) is very very small. For constant M (~40years) and constant ϵ (or υ), the annual cycle in N basically explains the annual cycle in the width of the ECD and quality of analogue and anti-analogues as seen in Fig.7.3.

If the N dimensional phase space has a complicated structure the chance of finding an analogue may depend on where the base case is situated. Nicolis (1998) has shown this to be the case for what is perhaps the most researched dynamical system in history, Lorenz' three variable system (Lorenz 1963). For a system that simple (N is between 2 and 3) one can generate enough data to a) make analogue forecasting (using data alone, not the equations) a success, and b) to verify Eq 7.2 or more refined versions.

7.1.4 Application of Natural Analogues

Because natural analogues are hard to find for large N the applications in geophysics are at best limited. The main trick is to find meaningful physical problems, especially forecast problems, with no more than 2 to 3 degrees of freedom. This can be done by

a) limiting the search area for analogues to a circle of say 1000km radius (Van den Dool 1989). From the many good matches, see Table 7.2, one can create limited area forecasts which are valid at the center of the circle for short range (<=12 hours) forecasts. By moving around the circle, a laborious activity, one can make forecasts for large domains, then repeat the process for the next 6 or 12 hour timestep.

b) the same 1000km radius areas, or small enough otherwise, but now to downscale or 'specify' one field from another (no forward time stepping involved). For instance Kruizinga and Murphy(1983) and O'Lenic and Handel(2004) use limited area height analogues to translate NWP forecasts into surface weather elements. Hamill et al(2006) use limited area analogues on NWP model precipitation forecasts (of which they have a 22 year homogeneous reforecast data base) to replace historical analogues (to the current forecast) by their matching verification fields (which may be at much higher resolution). Such a practice has also long been in existence in aviation meteorology (Hansen 2000).

Another way of working with just 2 or 3 degrees of freedom would be

c) by retaining only 2 or 3 leading empirical orthogonal modes on a domain large in km2, but small in N. The question is whether this leaves any meaningful forecast problem intact. Given that ENSO (~ 1 d.o.f.) has such a big influence over at least 50% of the earth, small natural subspaces <<N, do suggest themselves.

Lowering N by time averaging helps, but not nearly enough. Generating data by very long GCM runs helps (Branstator and Berner 2005), but not nearly enough to make day by day short range forecasts on a global domain possible. (One may also question whether GCM data are a substitute for reality). Researchers in the UK have proposed to run global coupled ocean atmosphere models for a long time to generate the data from which to select analogue cases for forecasts of ENSO in the future. This could work only if ENSO has effectively only a few dof, which appears to be case for certain climate variables (Fraedrich 1986). Of course, if it is true that ENSO is contained in just a few dof one might as well use just the observations, see discussion Chapter 10.

Finally, one way out appears to be the construction of an analogue, see next section. Section 7.3-7.6 are all about applying the constructed analogue.

7.2 Constructed Analogues

7.2.1 The idea.

Because ***natural*** analogues are highly unlikely to occur in high degree-of-freedom processes, we may benefit from ***constructing an analogue*** having greater similarity than the best natural analogue. As described in Van den Dool (1994), the construction is a linear[7] combination of past observed anomaly patterns such that the combination is as close as desired to the initial state (or 'base'). We then carry forward in time persisting the weights assigned to each historical case. All one needs is a data set of modest affordable length.

---

[7]We do not rule out that non-linear combinations are possible, but here we report only on linear combinations.

Assume we have a data set f(s, j, m) of, for instance, monthly mean[8] data as a function of space (s), year (j = 1, M) and month (m). Given is an initial condition $f^{IC}(s, j_0, m)$, for example the most recent state (monthly mean map), where j_0 is outside the range j=1..M. A suitable monthly climatology is removed from the data - henceforth f shall be the anomaly. A (linear) constructed analogue is defined as:

$$f^{CA}(s, j_0, m) = \sum_{j=1}^{M} \alpha_j \, f(s, j, m) \tag{7.3}$$

where the coefficients α are to be determined so as to minimize the difference between $f^{CA}(s, j_0, m)$ and $f^{IC}(s, j_0, m)$. The technical solution to this problem is discussed below in sct 7.2.2 and involves manipulating the alternative covariance matrix Q^a.

Eq (7.3) is only a diagnostic statement, but since we know the time evolution of the f (we know the next value historically) we can make a forecast keeping the weights α_j =constant. More generally we seek a forecast of variable g (which could be f itself) as follows:

$$g^F(s, j_0, m+\tau) = \sum_{j=1}^{M} \alpha_j \, g(s, j, m+\tau) \tag{7.4}$$

For $\tau > 0$ we are dealing with a forecast, $\tau=0$ would be 'specification' or down- or up-scaling of g from f (the weights are based on f only!), and $\tau<0$ would be a backcast. The method is reversible in time. For g = f one can see that the time dependence of f is entirely in the time evolving non-orthogonal basis functions - this is the main trick of the CA forecast procedure and a significant departure from traditional spectral methods in which the basis functions are constant and the time dependence is in the coefficients α_j.

We will later refer to (7.3)-(7.4) in slightly rewritten form, the details depending on whether we use daily, monthly, seasonal or a sequence of seasonal data.

Eq (7.4) can also be written (for g=f and $\tau \neq 0$):

---

[8]One can construct analogues for monthly, seasonal or daily data. The procedure is the same. Here we start with monthly.

$$f^F(s, j_0, m+\tau) = f^{fC}(s, j_0, m) + \sum_{j=1}^{M} \alpha_j\,(f(s, j, m+\tau) - f(s, j, m)) \qquad (7.4')$$

In this form the equation looks like a forward time stepping procedure or the discretized version of the basic equation $\partial f/\partial t =$ linear and non-linear rhs terms. Note that on the rhs of (7.4') we make linear combinations of historically observed time tendencies.

Why should Eq (7.4) yield any forecast skill? The only circumstance where one can verify the concept is to imagine we have a natural analogue. That means α_j should be 1 for the natural analogue year and zero for all other years, and (7.4) simply states what we phrased already in section 7.1 and depicted in Fig 7.2, namely that two states that are close enough to be called each other's analogue will track each other for some time and are each other's forecast. Obviously, no construction is required if there was a natural analogue. But, as argued in the Appendix, in the absence of natural analogues a linear combination of observed states gives an exact solution for the time tendencies associated with linear processes. There is, however, an error introduced into the CA forecast by a linear combination of tendencies associated with purely non-linear components, and so a verification of CA is a statement as to how linear the problem is. Large scale wave propagation is linear, and once one linearizes wrt some climatological mean flow the linear part of the advection terms may be larger than the non-linear terms. This is different for each physical problem.

Is a constructed analogue linear? The definition in (7.4) is a linear combination of non-linearly evolving states observed in the past. So even in (7.4) itself there is empirical non-linearity. Moreover, in section 7.6, we will change the weights during the integration expressed in (7.4') - this will add more to non-linearity. (We are not reporting on any attempts to add quadratic terms in (7.4) - that would allow for more substantial non-linearity, but the procedure to follow is unclear).

7.2.2 The method of finding the weights α_j

We are first concerned with solving Eq(7.3). The problem is that the solution may not be unique, and the straightforward formulation given below leads to a (nearly) ill-posed problem. Classically we need to minimize U given by:

$$U = \sum_s \{f^{dC}(s, j_0, m) - \sum_{j=1}^{M} \alpha_j \, f(s, j, m)\}^2$$

Differentiation w.r.t. the α_j leads to the equation

$$Q^a \, \alpha = a \qquad (7.5)$$

This is the exact problem described in Eqs (5.1a) and (5.4a). Q^a is the alternative covariance matrix, α is the vector containing the α_j and the rhs is vector **a** containing elements a_j given by

$$a_j = \sum f^{dC}(s, j_0, m) \, f(s, j, m),$$

where the summation is over the spatial domain. Note that α_j is constant in space - we linearly combine whole maps so as to maintain spatial consistency. Even under circumstances where Eq(7.5) has an exact solution, the resulting α_j could be meaningless for further application, when the weights are too large, and ultra-sensitive to a slight change in formulating the problem.

A solution to this sensitivity, suggested by experience, consists of two steps:

1) truncate $f^{dC}(s, j_0, m)$ and all $f(s, j, m)$ to about M/2 EOFs. Calculate Q^a and rhs vector **a** from the truncated data. This reduces considerably the number of orthogonal directions without lowering the EV (or effective degrees of freedom N) very much.

2) enhance the diagonal elements of Q^a by a small positive amount (like 5% of the mean diagional elements), while leaving the off-diagonal elements unchanged. This procedure might be described as the controlled use of the noise that was truncated in step 1.

Increasing the diagonal elements of Q^a is a process called ridging. The purpose of ridge regression is to find a reasonable solution for an underdetermined system (Tikhonov 1977; Draper and Smith 1981). In the version of ridge regression used here the residual U is minimized but subject to minimizing $\sum \alpha_j^2$ as well. The latter constraint takes care of unreasonably large and unstable weights. One needs to keep the amount of ridging small. For the examples discussed below the amounts added to the diagonal elements of Q^a is continued until $\sum \alpha_j^2 < 0.5$.

We thus construct the analogue in an EOF truncated space. Strictly speaking we could find an exact solution to Eq 7.3, no ridging, once the fields are truncated to M/2 EOFs, with just any M/2 years chosen at random. But this procedure would be too sensitive. It is better to use all

13

years, and deal with the underdetermination by ridging.

Instead of using EOFs for truncation the alternative EOT suggest themselves, since we use Q^a. In essence we would rewrite (7.3) as

$$f^{CA}(s, j_0, m) = \sum_{j=1}^{M/2} \gamma_j \, e(s, j, m) \qquad (7.3a)$$

where the $e(s, j, m)$ are a set of alternative EOTs. One might look upon alternative EOT, orthogonal in space, as the most obvious way of removing co-linearity among the $f(s,t)$ so as to make the weights γ_j unique and their calculation easy (just projection). The α_j in (7.5) can be found from γ_j in (7.3a) by a recursive expression. The EOTs are linked to specific moments in time, see Chapter 5, such that the execution of Eq (7.4) is easy[9]. Note that in this variant the base functions are orthogonal initially, but turn non-orthogonal as the forecast proceeds.

One can raise the question about which years to pick, thus facing a near infinity of possibilities to chose from. Here we will use all years. No perfect approach can be claimed here, and the interested reader may invent something better. A large variety of details about ridging is being developed in various fields (Green and Silberman 1994; Chandrasekaran and Schubert 2005), see also appendix of chapter 8.

Note that the calculation of $\alpha(t)$ has nothing to do with Δt or future states of f or g, so the forecast method is intuitive, and not based on minimizing some rms error for lead Δt forecasts. There could not possibly be an overfit on the predictand.

7.2.3 Example of the weights

An example of the weights obtained may be illustrative. Table 7.3 shows the weights α_j for global SST (between 45S and 45N) in JFM 2000. We have solved (7.3), i.e. found the weights to be assigned to SSTA in JFM in the years 1956 through 1998 in order to reproduce the SST-anomaly field observed in JFM 2000, truncated to 20 EOFs, as a linear combination. For each

---

[9]Because EOTs are not unique it is tempting to use this freedom to think about tailoring EOT for the given base. I.e. find the one state among the f(s,t), say at t=t1, that explains the most of the base, then on to the 2nd tailored EOT etc etc as per Gram-Schmidt procedure. This yields the set of EOTs that is best suited to explain the base (better even than the EOFs of f(s,t)). However, the price to pay is that the historical data set will be truncated more by tailored EOT (than the EOT defined in Ch 5).

year we also give the inner product (ip) between the SST field in 2000 and the year in question, i.e. the rhs of (7.5). The ip gives the sign of the correlation between the two years. The sum of absolute values of ip is set to 1. Note the following:

i) there is no real large weight, 0.22 being the largest single value, indicating that no year is a natural analogue

ii) we allow both +ve and -ve weights, as the problem is cast in terms of anomalies. A high negative weight would point to an anti-analogue.

iii) the sum of the weights is unconstrained

iv) all years are used, even though years with small weights hardly participate.

v) the weights are somewhat similar to the ip's (have the same sign usually) but there are exceptions (1992 has zero ip but a high +ve weight). These exceptions are caused by the co-linearity as expressed by the off-diagonal elements of Q^a, i.e. the fact that years i and j have non-zero correlation.

vi) Since JFM2000 was a cold event in the Pacific, the reader may verify that previous cold events generally have +ve weight, especially 1989.

vii) The Pacific SST variability dominates variability globally during ENSO events, so the weights reflect ENSO. Nevertheless, one can also see a trend from mainly negative to mainly positive weights over the 40+ years shown. These trends can be even clearer when the tropical Pacific is quiet.

Table 7.3 Weights (*100.) assigned to past years (1956-1998) to reproduce the global SST in JFM 2000. The column ip refers to 'inner product', a type of weighting that ignores co-linearity. See also footnote # xx in sct 7.4 on the sequence of 4 seasons.

yr	ip	α_j	yr	ip	α_j	yr	ip	α_j	yr	ip	α_j	yr	ip	α_j
56	1	18	66	-5	-12	76	2	14	86	3	9	96	3	14
57	0	4	67	-5	-12	77	-3	-2	87	-1	-6	97	2	13
58	-4	-9	68	-2	-16	78	-4	-4	88	0	6	98	-1	3
59	-3	-7	69	-4	-8	79	-6	-7	89	4	22	99	NA	NA

60	-4	-2	70	-4	-6	80	-2	-7	90	5	14			
61	2	12	71	2	0	81	-2	-2	91	2	6			
62	-1	-1	72	1	-2	82	1	-3	92	0	18			
63	-1	2	73	-2	3	83	-1	-5	93	-1	-17			
64	-4	0	74	3	8	84	1	15	94	0	-6			
65	-1	-19	75	1	0	85	5	6	95	0	1			

Below we will discuss three applications to demonstrate how well constructed analogues (CA) work. The first example is specification of monthly mean surface weather from 500mb streamfunction. In section 7.4 we describe global SST forecasts - this has been the main application of CA so far. In section 7.5 we describe how CA works on daily 500mb height data, and how the dispersion of an isolated source is portrayed by CA, in comparison to EWP. This leads us into the possibility of calculating the fastest growing modes by empirical means in section 7.6

7.3 Specification or downscaling

Given a 500mb height map, what is the associated temperature at the same time near the surface? This question comes up when judging the implication of a 500mb forecast in terms of surface weather. In Fig. 7.4, upper left, we have the observed 500mb streamfunction (ψ)[10] anomaly map for February 1998. Using ψ data for the same domain in the years 1961-1990, and including neighboring months January and March (creating a data set of 90 cases), we first truncated to 50 EOFs (Fig 7.4 upper right), then constructed an analogue, as per Eq (7.3). The result is in the lower left. The EOF truncation is vitally important to find the solution, but as can be seen in the bottom right the error between the upper left and lower left panel is very small. Indeed 50 EOFs reproduces the original field, and the reader will be hardpressed to see any difference between panel (b) and (c), i.e. within the truncated world the CA is very accurate by

---

[10]We present new work and review older work using either 500mb height or 500 mb streamfunction. These two fields are closely related in mid-latitude, and nearly equivalent for specification purposes.

design. The error field (a) - (c) is given in panel (d). Making the CA for this application is no problem at all.

With the weights known as per Eq (7.3) we now execute Eq (7.4) for τ=0, and g is the 850mb temperature[11]. Note that no temperature fields were used in Eq (7.3). The linear combination of temp anomalies observed during 1961-1990 is in the lower left of Fig. 7.5 - and is to be verified against the observed in the upper left of Fig. 7.5. One can see that with one set of weights, independent of space, we largely reproduce the near-surface observed temperature anomaly from the streamfunction field aloft for most of the NH. In some small areas the error is considerable, and other factors may determine near surface temperature, but overall the error is 1.0C and the correlation 0.85. We considered many cases, and the example shown is typical in performance. In other years we find correlations ranging from 0.65 to 0.90. Specification of precipitation is harder, but still has appreciable skill (not shown).

To be sure: February 1998 is not an arbitrary month since it is during a strong ENSO winter, and the flow pattern in Fig.7.5 may be largely a response to tropical convection. Comparing to the 1st EOF of ψ in Fig.5.9, one may note high +ve projection for 1998. The surface weather in Fig.7.5 resembles regression or composites based on ENSO (Ch 8).

In Van den Dool(1994) this same procedure was applied to NH 500mb height, to specify US weather and using CA was shown to work much better than NA. Results over the US by CA appear consistent with far more laborious station by station regression equations developed by Klein(1985) for T and Klein and Bloom(1987) for P. Note that CA damps less than regression equations. Still CA damps some, and the error in Fig.7.5(d) (observed minus specified) often has the sign of the observed anomaly.

We thus find that CA works well on problems that are to some good approximation linear. The relationship between streamfunction (or height) and surface weather is fairly linear. As further proof that CA is a very good linear operator indeed we apply CA to the problem of specifying Z500 from ψ500. This problem presents itself in NWP when at the end of a numerical integration the height field (not a prognostic variable in modern models) has to be derived from ψ. The most

---

[11]850mb temperature is not literally a surface weather element. But due to difficulties in the Reanalysis/CDAS surface 2-meter temperature, 850mb is the closest proxy.

complex but still linear relation to obtain Z from ψ is given by the linear balance equation:

$$\nabla \cdot (f \, \nabla \, \psi) = \nabla^2 (g \, Z) \quad (7.6)$$

where f is the Coriolis parameter, and g the acceleration of gravity. ∇ is the horizontal gradient operator. If CA is a good linear operator it should score as high as Eq 7.6 on the task of calculating Z from ψ. We tested this for January mean 500mb data during 1991-2000, 10 cases in all. CA scores better than the linear balance equation in 8 out of 10 cases, their average scores are around 0.92 (Eq(7.6)) and 0.94 (CA) respectively. Both methods work well, and CA certainly succeeds in being as good as Eq (7.6). This proves we have correctly built a linear operator from data.

The reverse problem, calculating ψ from Z, is more difficult theoretically. But when we used the linear CA to calculate ψ from Z we still found a correlation in excess of 0.90 in all 10 cases.

7.4 Global Seasonal SST Forecasts

The best prognostic application for CA sofar has been the forecast of global SST. This has been done in real time at CPC since about 1993 (Barnston et al 1994). We use the (near) global SST that has been (is) used as lower boundary condition in the NCEP/NCAR Reanalysis (Kalnay et al 1996) (its continuation CDAS (Kistler et al 2001)), and form seasonal means. Thus, our data set f(s, j, m) is the seasonal mean SST over the period 1955-present, and m denotes season, m=1 for DJF etc. The most recently observed SST can be approximated by a constructed analogue as per (7.3) as

$$f^{CA}(s, j_0, m) = \sum_{j=1}^{M} \alpha_j \, f(s, j, m) \quad\quad\quad (7.3b)$$

where j=1 corresponds to 1956, and M is the last year. The weights are determined after truncating the global SST in EOF space. An example of the weights was shown in Table 7.3 when we used data through 1998, and M=43. As of this writing we use data through 2003 for the construction, and M=48. Given the weights the forecast is given by:

$$M$$

$$f^f(s, j_0, m+\tau) = \sum_{j=1} \alpha_j \, f(s, j, m+\tau) \qquad\qquad (7.4b)$$

which for $m+\tau > 12$ (NDJ) runs into the next year. In the example in Table 7.3 (initial state is JFM 2000), only data through JFM 1998 were used to determine the weights - in Eq (7.4b) the largest forecast lead τ used is 2 years. If one included 1999 in the construction τ could be only up to one year.

One of the most verified aspects of global SST forecasts is Nino34, an area between 170W and 120W and 5S to 5N. This index is thought to best describe ENSO (Barnston et al 1996). Fig 7.6 shows the skill of the CA forecasts for Nino34. The format of this graph is target season vs lead (in months). A zero lead forecast for season 1, DJF, is made at the end of November. The forecasts are automatically cross-validated because the year for which the analogue is constructed has to be left out in Eq(7.3b). For winter target seasons, the anomaly correlation is in excess of 0.9 out to a lead of 2 months, and in excess of 0.6 correlation until a lead of one year. One can clearly see the so-called 'spring barrier'. For instance, at a lead of 10 months, the correlation for June is < 0.4, but the correlation keeps increasing towards nearly 0.7 at lead 10 through fall and winter until about March, then suddenly drops back. Around the spring barrier the standard deviation is smallest, and most changes of sign take place in spring. It is clear that a relatively simple method can have a high level of skill. Earlier verifications (Barnston et al 1994; Landsea and Knaff 2000; Saha et al 2006) indicated that CA is among the leading forecast methods, both in real time and on retroactive forecasts. In fact CA stays better than random forecasts until a lead of more than 2 years. A problem in Nino34 forecasting, for all methods, is that there are just a few occasions of strong anomalies in the 50 year record and they make or break the overall verification scores. In between, and this can last for years, none of these forecast methods performs particularly well.

In Eq (7.3b) we construct an analogue to a recent initial condition. As an alternative we have constructed an analogue to a sequence of seasons, i.e. we construct an analogue that is similar to the development of global SST over the whole past year. This does help to boost the skill of the CA method for global SST. The justification for this may be as follows: We are forecasting a single variable from a single variable in a world where many variables are interrelated. An equation for a single variable would analytically be the result of eliminating all the

19

other variables, and the order of the final differential equation may be quite high. This justifies using more that a single initial condition. In effect we use initial conditions for the 1st, 2nd, 3rd derivative at recent times. To cut down on choices we have used either only a single season (the most recent season) or 4 non-overlapping seasons (one year)[12]. In EOF expenditure terms, these choices are the same, i.e. we either use a lot of precision in the latest season and none earlier, or we use moderate precision spread out over a whole year. The idea of using analogy over a longer period dates back to the days NA was used (Schuurmans 1973).

The use of four successive seasons also breaks the linearity (or symmetry) of ENSO warm and cold events somewhat. If warm and cold event evolve in different ways, an analogue constructed to a warm event for four successive times will not give high negative weights to a previous cold event.

The verification results shown in Fig. 7.6 are actually for an <u>ensemble mean</u> forecast. How is the ensemble made? One can vary the number of EOFs used, here 16, 21 or 26, the participating years as M+1 or M, and by spending all EOFs on the latest season's global SST, or spread the EOFs out across 4 non-overlapping seasons so as to mimic the evolution over a whole year. Any of these perturbations has the effect of changing the weights and creating a somewhat different initial CA. Following these options we obtained 12 members. An example is given in Fig. 7.7, which shows the ensemble issued in early July 2005. Half the members use 49(48) years, labeled l for late (e for early) in the 3rd index, and half the members spend the EOF on just the latest season (last 12 months), labeled 1 (12) in the first index. The middle index is 16, 21 or 26 for the number of EOFs used. The forecast by the late members using M+1 years stops after one year because the historical evolution following June 2004 was, at the time of forecast release, not known beyond June 2005.

Two members show an increase in Nino34, while 10 members show a tightly clustered decrease. Considerable spread is thus seen to be possible among the members. Traditional statistical methods minimize the rms error in the predictand (here Nino34) and all members would ultimately go to zero anomaly and collapse to zero spread. But even at very long lead the CA

---

[12]Note that Table 7.3 was constructed to the string of MJJ, JAS, OND1999 and JFM2000, each season five EOFs of global SST.

members do not agree. Clearly CA has features of error growth, divergence among states which reflects the fact CA has unstable modes. This will be worked out further in section 7.6.

7.5 Short range forecasts and dispersion experiments

Short range forecasting by CA is not a viable practical application at this time because superior methods have been available for decades. The verification presented below is to further understand the strengths and limitations of CA as a method. Especially the question as to why it would be useful in the long range and not in the short range. The question of utility for practical application involves comparison to other available tools.

Given is a space time data set f(s,t), in this case f is the instantaneous daily 500 mb geopotential height taken from NCEP-NCAR Reanalysis - s is a spatial coordinate (5° lat by 10° lon grid), and t is time. We form anomalies by subtracting a harmonically smoothed 1979-1995 daily climatology, appropriate for the time of year, produced by Schemm et al (1997). We consider the domain 20N to the North Pole. The data set f is processed as follows: In 1968 we take the fields for January, 1, 3, 5, ... 23 at 0Z, i.e. twelve fields in one year. Similarly for 1969 through 1992, for a total of 300 fields during 25 years. We now have f(s,t), where t=1, 300, representing a great diversity of NH January flows. The time t is a counter for both regular time and annual increments. f(s,t) represents the 'library' of historical states used for constructing analogues. A similar data set f(s,t+Δt) is used for making the forecast.

7.5.1 Short range forecasts

We now chose initial conditions every day from January 1-23 in the years 1993, 1994 through 2004, a total of 12 years, having no overlap with the library. Fields are truncated to 100 EOFs (determined from f(s,t)), and even though the CA is made in truncated space, the resulting CA field correlates 0.99 with the original. So we have:

$$f^{IC}(s, t_0) \approx f^{CA}(s) \equiv \sum_{t=1}^{300} \alpha(t) \, f(s,t) \qquad (7.3c)$$

And the forecast is given by:

$$f^F(s, t_0+\Delta t) \quad = \quad \sum_{t=1}^{300} \alpha(t)\, f(s, t +\Delta t) \qquad (7.4c)$$

We verified all 276 forecasts (12 years, 23 ICs) based on (7.3c) and (7.4c) for $\Delta t=1$ day on the domain 20N to the North Pole. The anomaly correlation is on average only about 0.82, ranging from 0.69 to 0.93 in individual cases. This has to be compared to the mean anomaly correlation of persistence on the same domain, 0.73, and ranging from 0.53 to 0.89 in individual cases.

Even though we have a CA match of 0.99 at the initial time, the correlation for CA drops like a rock to 0.82 in a single day - keep in mind that NWP stays above 0.9 for several days. Perhaps this can be explained away because we did not match the temperature field. But even a barotropic model has above 0.9 correlation at day 1 (Qin and van den Dool 1996). CA and a barotropic model only differ in the formulation of the anomaly vorticity advection by the anomaly wind. Apparently the linear combination of time tendencies associated with non-linear terms harms the CA forecast in the short range forecast. Still the total time tendency produced by CA has some skill, otherwise we could not have beaten persistence. It thus follows that even in this short range forecast problem the linear components of the time tendency are larger than the non-linear components. Nevertheless, the errors made in the latter put CA behind the competition. Looking back at the high scores for SST forecasts by the CA method months ahead of time (discussed in ch7.4), one must conclude that the SST prediction problem is a far more linear problem than the short range weather forecast problem.

Comparison of CA to NA and EWP can also be drawn. CA is better in forecast skill than NA because the initial match for CA is very good - the day 1 score (0.82) for CA is still much better than initial match for NA (which averages only 0.55 in winter, see Fig. 7.3). Even the day 2 score for CA, 0.60 correlation, is better than NA's initial match. The NA may be a perfect method, but its handicap (a bad initial state) makes it unusable - NA cannot even beat persistence averaged over all cases. As a short range forecast tool the skill of CA is not very different from EWP, compare to Table 3.7 for instance. Saying that the skill is similar does not mean that the forecasts are similar, as we will see in the next section 7.5.2. It may be a partial coincidence that CA and EWP have the same day 1 score. EWP is designed to do a simplified version of wave

propagation, an imperfect attempt to calculate the linear part of the time tendency. On the other hand CA is perfect at the linear part of the time tendency, but adds a presumably bad estimate of the non-linear term to it.

We also checked the results of CA forecasts for 20S to the South pole. Results in the SH are entirely consistent with the NH. The gain over persistence is much larger in the SH, roughly as much as it was for EWP in Table 3.7.

7.5.2 CA Dispersion Experiment

We are now in a position to redo the 'rock in the pond' experiments of chapter 3 to study CA behavior. We will start with the same 'round' 500 mb height anomaly disturbance at 45N used in Chapter 3. But what is different from the EWP dispersion is the CA dispersion depends on longitude. EWP as designed gives the same result, regardless of the longitudinal position of the initial disturbance. In contrast CA 'knows' about the stationary waves in the background field, and implicitly the underlying land-sea distribution. A technical issue is that for EWP we maintained wave amplitude harmonic by harmonic, while CA has an evolving amplitude. We undo this difference by restoring the amplitude of CA to its original value, where amplitude is as defined in Eq (7.1a). Still the spectral distribution may change somewhat.

Fig. 7.8 shows the CA dispersion from a source on the dateline and 45N. This has to be compared to Figs 3.1 and 3.2 for EWP1&2. The geography in the latter two is only for orientation while in Fig.7.8 the geography has real meaning. From the dateline and 45N, the dispersion by CA and EWP are broadly similar, with similar upstream and downstream developments. Mainly zonal dispersion can be seen in CA[13]. The fields after a few days are qualitatively similar, but the details of the far downstream traveling stormtracks show differences in slope, organization etc. This may be caused by the non-zonal background flow in which the CA dispersion takes place compared to the uniform background flow for EWP.

In Fig.7.9 we compare the 2-day forecasts by CA for four positions of an identical rock in the pond at 45N and the dateline, 90W, Greenwich and 90E respectively. The maps are rotated

---

[13]This could be in part because we use the domain 20N-pole for CA, while for EWP2 (spherical harmonics) we use the full globe.

such that initial source is always at the bottom. We now see that of the four starting longitudes, the dispersion from the dateline and Greenwich are the most similar to EWP. The other maps are somewhat similar, but not greatly so. In all cases one can see downstream development of opposite sign, but the intensity depends on longitude. Persistence of the original blob is strongest near Greenwich. Other features are more different. It does matter greatly at which longitude we place the rock, or rather relative to what background flow. Especially the position over Asia gives a different impression[14]. In any of the four plots there is some slight meridional propagation, the amplitudes being non-zero outside the latitudinal band of the original disturbance, but is not as clear as in Fig. 3.2. EWP works with a much more idealized zonally symmetric background flow. A plot similar to Fig.7.9 for the SH has simpler structures, looks more like EWP and shows little dependence on longitude.

---

[14]This difference may be slightly overstated because 100 EOFs explain less of the variance of the patch over Asia than elsewhere. I.e. the starting rock in the pond is not 100% identical at all positions once the EOF truncation is applied.