10. Conclusion

In this book we have reviewed empirical methods in short-term climate prediction. We devoted a whole chapter to the design of two of these methods, Empirical Wave Propagation (EWP, ch 3) and Constructed Analogue(CA ch 7). Other methods of empirical prediction were listed in Chapter 8, with brief descriptions and examples and references. One chapter is devoted to EOFs, as such a diagnostic topic, but widely used in both prediction and diagnostics, and thoroughly debated for a few decades. Two brief chapters are written in support of the subsequent chapter - Teleconnections (ch 4) should make the discussion on EOFs more interesting, and the topic of effective degrees of freedom (ch 6) is indispensable when one wants to understand why and when natural analogues would work (or not), or how an analogue is constructed, or how any method using truncation works.

Most chapters can be read largely in isolation, but connections can be made of course between chapters. EWP is claimed to be useful, if not essential, in understanding teleconnections. Dispersion experiments, featuring day-by-day time scales, link the CA and EWP methods. Examples of El Nino winter behavior can be found in a) the examples of EOFs on global SST and 500mb streamfunction (Ch 5), b) specification of surface weather from 500mb streamfunction (Ch 7), and c) the ENSO correlation and compositing approach (Ch 8). The noble pursuit of knowledge may have been as important in the choice of some material as any immediate prediction application. Chapter 9 is different, less research oriented, and more an eyewitness description of what goes on in the making of a seasonal prediction. This eyewitness account style spills over into chapter 8 here and there, because in order to understand why certain methods have survived to this day some practicalities have to be understood.

The closeness to real time prediction throughout the book creates a sense of application. However, the application in this book does not go beyond the making of the forecast itself - we completely shied away from such topics as a cost/benefit analysis or decision making process by, for example, a climate sensitive potato farmer or reservoir operator. Hartmann et al (2002) describe the CPC forecasts with an eye towards users.

Some special topics are left for this, the final, chapter to be emphasized, organized, discussed, maybe solved, or left for the interested reader to pursue further. In this order we will discuss 1) linearity, 2) why GCM's do not (yet handily) outperform empirical methods, 3) predictability and 4) the future of short-term climate prediction. The linearity is discussed first because it is very important for the 2nd topic.

10.1 Linearity

Although the equations of motion are non-linear, some aspects of behavior in the atmosphere are perhaps much more linear than expected. To avoid confusion let us mention that the words linear and linearity are used in somewhat different ways in various contexts.

a) The equations being non-linear means that they contains terms in which products of basic variables occur, the most obvious example being a momentum equation

$$\partial u/\partial t = - u \, \partial u/\partial x \text{ etc} \qquad (10.1)$$

where u is a wind component. In this case (non)linearity relates directly to time tendency on the left hand side.

b) The response of the atmosphere to El Nino and La Nina is said, by some, to be linear if the response to positive and negative SST anomalies is, except for the sign of the atmospheric anomaly, the same. This feature actually has more to do with the symmetry of the pdf with respect to the mean. Whether there is a relationship between type a dynamical and type b statistical linearity, we do not know, but see Burgers and Stephenson (1999) for a discussion.

c) A linear operator. Given the equation Ay=x, where A and x and known and y is solved for, the operator is said to be linear if $A(\alpha y)=(\alpha x)$ for any value of α. Type b linearity follows from type c. On occasion (Hoskins and Karoly 1981; Opsteegh and Van den Dool 1980), the atmosphere has been described as a linear steady state operator to describe teleconnections as a response to forcing. Doubling the forcing x, leads to doubling the response y.

d) An important feature of linearity in some contexts is that orthogonal modes do not exchange energy.

Obviously no one stops us from linearizing Eq (10.1), to the extent possible, i.e. substitute some mean state (U) plus a departure from it (u'). There are no definitive rules to choose U, so we leave in the middle exactly what U is (climatology, the mean state in the absence of transients, today's state in all its details, etc). The rhs of 10.1 can be written, $u\,\partial u/\partial x = U\,\partial u'/\partial x + u'\partial U/\partial x + U\,\partial U/\partial x + u'\,\partial u'/\partial x$. Assuming there is an equation for the mean state, $\partial U/\partial t = -U\,\partial U/\partial x + ... +$ known terms, we can thus write

$$\partial u'/\partial t = -U\,\partial u'/\partial x - u'\partial U/\partial x - u'\,\partial u'/\partial x \quad (10.2)$$

This recasts the forecast problem as one concerned with anomalies only (when U is climatology), a very familiar theme in this book. It is very possible that among the three terms on the rhs $u'\,\partial u'/\partial x$ is not dominant. The prototypical non-linear advection term can often be linearized to a considerable extent.

Instances in this book where linearity was noteworthy and possibly surprising are

1) EWP (Ch 3) appears to yield a linear wave dispersion relationship as if each wave, in an aggregate sense, can be treated without regard for the presence of other waves. In other areas of study it has been difficult to find averaging procedures that make the non-linearity small or vanishing. For instance, the mean state of the atmosphere can never be understood without including the mean effect of the transients (Peixoto and Oort 1992). The apparent linearity of EWP is related somewhat to using only short time separations in judging the phase speed. Presumable this is similar to the validity of tangent linear models for a duration of at most a few days. If non-linearity dominated in (10.2) (i.e. $u'\,\partial u'/\partial x \gg U\,\partial u'/\partial x + u'\partial U/\partial x$) none of this would be possible.[1]

2) CA (Ch 7) appears to duplicate much of EWP when applied to the same set of 1-day forecasts.

[1] In the less restrictive application of phase shifting (Cai and Van den Dool 1991) we found that sub-harmonics of the wave one follows survive the averaging, i.e. EWP is not entirely linear.

The trick of CA is to linearly combine states, and their subsequent evolution. But what exactly happens to the equations when two (only two to keep it simple) previously observed states u'_1 and u'_2 are averaged: $u'_3 = (u'_1 + u'_2)/2$. One must wonder whether the average of the observed $\partial u'_1/\partial t$ and $\partial u'_2/\partial t$ is a good forecast for $\partial u'_3/\partial t$. Again the success of CA, modest as it is, would be impossible unless the third term in Eq (10.2) is small, or at least not dominant. See Appendix in Ch7 for more details.

3) The success of CA in specification problems (Ch 7.3; no time derivative involved) is evidence of quasi-symmetric pdfs. It may be odd to say that rain and sunshine are each other's opposite, but when thought of as the pdf of a variable like vertical motion, there is indeed near symmetry relative to a climatology. Similarly, the search for natural analogues and anti-analogues (Ch 7.1) indicated only a very small difference in their quality, suggesting linearity of type b. Not shown is that the tendencies following (natural) anti-analogues, with sign reversed, are only slighty worse as a forecast than the tendencies following natural analogues (Van den Dool 1991), i.e. non-linearity of type a cannot be dominant.

4) There are plenty of linear correlations reported in this book. That the measurement of pressure at Darwin in Australia in SON (or the SST in a small area called Nino34) relates linearly to seasonal mean conditions half a world away for the next JFM should not stop to amaze us. How different is the situation in short range forecasting. In order to make a 5 day forecast we need a global model and accurate initial data over the whole globe. In fact one missing observation has sometimes been blamed for a failed forecast. How can things be so simple, type c linearity, for the seasonal forecast?

5) The linearity question has been raised with nearly all tools in Ch8, and addressed best in the literature in connection with the LIM. Why should Eq (8.8) $dx/dt = L\,x + R$ be approximately valid?. The answer offered by Winkler et al(2001) is that by taking a suitable (time) mean, like a weekly mean, the dynamical short time scale non-linearities of type a can be represented as a stochastic process acting upon the more slowly and linearly evolving time mean state. Whitaker

and Sardeshmukh (1998) even go as far as showing that the effect of stormtracks is linear in terms of anomalies in the time mean flow.

Consideration of linearity drive the answer in the next section.

10.2 Relative performance GCMs and empirical methods,

In 2005 NCEP gave some of its employees an award for developing a global land-ocean-atmosphere system called the Climate Forecast System. The citation included a sentence that stated that for the first time in history numerical seasonal predictions were on par with empirical methods. From a competing institution we got a publication entitled: Did the ECMWF seasonal forecast model outperform a statistical model over the last 15 years? (Van Oldenborgh et al 2005). How does one explain that after years of development, a costly GCM, absorbing enormous resources, does not handily outperform some simple empirical method. Why is it, in this day and age, that simple models, empirical methods etc are still of some value, and used at CPC and CDC in real time seasonal forecast operations? This may in fact be unanticipated, after the optimism expressed some 15 years ago about using models for this purpose (Palmer and Anderson 1994). The answer in our opinion has to do with linearity. Most empirical methods are linear or very close to being linear. The only fundamental advantage a GCM has over linear methods is that it can execute the non-linear terms. However, if for some magical reason the problem of seasonal forecasting is quite linear, a GCM cannot exploit its only advantage. Indeed following what we said in section 10.1, this may already be the case in the week2 forecast (Winkler et al 2001) where LIM and the NCEP global spectral model are not very different in skill. On may see in the same light the definition of the practical limit of predictability in Saha and Van den Dool(1988) - this limit is said to be reached if the continued model integration of an n-day forecast out to day n+1, is no better than persistence of the n-day forecast.

However, we do not need to go as far as requiring that the atmosphere is almost linear.

No matter how small, non-linearity is never trivial.[2] Instead, we argue below that as far as forecast skill in short-term climate prediction is concerned atmospheric (and oceanic) models are functionally linear. The lines of reasoning are as follows:

1) empirical methods are linear, or nearly so[3]

2) physical models have one clear advantage over empirical models: they can execute the non-linear terms

3) a model needs at least three degrees of freedom to be non-linear (Lorenz, 1960) so as to allow energy exchange among modes, although not any three dof will do.

4) we speculate that a non-linear model with nominally millions of degrees of freedom, but skill in only <= 3 dof is functionally linear in terms of the skill of its forecasts - and, to its detriment, the non-linear terms add random numbers to the time tendencies of the modes with prediction skill. It takes large ensembles to remove this noise.

5) empirical methods like analogues, see Ch 7.1, given ~50 years of data, can deal very well with about three effective degrees of freedom.

Therefore: Physical models need to have skill in, effectively, more than three dof before there is a scientific basis for expecting them to outperform linear empirical methods in a forecast setting[4]. The number three arises for two unrelated reasons. One needs at least three dofs for a non-linear model, and given ~50 years of observations empirical methods should be very good at problems with three or less edofs. (It follows incidentally that the NA method, while methodically non-linear, is also functionally linear as long as we cannot match more than 3 dofs.)

How many degrees of freedom (with forecast skill) do GCMs have in the seasonal forecast

---

[2]Here we are quoting H. Tennekes, who would say the same thing but dissipation instead of non-linearity.

[3]Some efforts to make non-linear empirical methods notwithstanding. Neural nets, and analogues are non-linear in principle.

[4]We emphasize forecast setting here - dynamical models may be far better than empirical methods in a simulation mode, but not in prediction mode.

problem? It would appear that currently the answer is only around one, maybe two. Quan and Hoerling (2005) have shown that the lion share of the GCM forecast skill can be duplicated by a linear regression on the first mode of tropical Pacific SST variability. That points to order one dof. Anderson et al(1999) have shown that models, vintage late 1990's, could not outperform CCA on identically the same task. Several authors (Straus and Shukla 2000) have found that the ensemble mean state of the model atmosphere in long AMIP runs has a very dominant first EOF, or, in our terms a very low N. That kind of simplicity is actually an argument against the application of GCMs to the forecast problem. That only one mode survives in the ensemble mean indicates that other modes, which do exist in individual members and in nature, do not correlate among members, and disappear upon taking the ensemble mean, a sure sign of low predictability. The forecast of a single EOF can be done very well by empirical means - no models needed. In fact Anderson(1999) found that CCA trained on model data could make an equal or better forecast of the next member in the ensemble than a model integration itself. Even under perfect model assumption the dofs with forecast skill thus appear very limited.[5]

Given that only ENSO, trends and soil moisture come to mind as factors in short term climate prediction, it is hard to imagine that the dof that can be skillfully predicted is very high.

A lack of non-linearity in the seasonal forecast problem can also be seen in the benefit of systematic error correction. Fig.10.1 shows along the y-axis the anomaly correlation of bias corrected ensemble mean Z500 forecasts[6] for AMIP runs by several models (over the period 1950-1994, see Peng et al(2002) for detail), as a function of the magnitude of the systematic error in Z500. GCMs have substantial systematic error. The standard deviation for seasonal mean Z500 in the NH is around 30gpm in DJF, so the four models shown Fig.10.1 have a systematic error ranging from one up to three times the natural variability. If the models were operating in a highly

[5]Characterizing the dof that two data sets (like forecasts and verification) have in common needs more work than what was presented in Ch6.

[6]These are not strictly forecasts, since perfect SST is provided.

non-linear environment the simulation of anomalies would be much less (more) skillful for models with the largest (smallest) systematic error. However, that dependence is only weakly present. Apparently even a huge systematic error can be removed and the anomalies salvaged, a sign of linearity which is unwelcome if GCMs are at some point in the future expected to exploit their ability to do non-linear calculations and beat linear methods. Extrapolation of the four entries in Fig.10.1 to zero systematic error may be dangerous, but does not point to anything higher than a 0.5 to 0.6 anomaly correlation for a perfect model.

Simply expecting that models will eventually handily outperform empirical methods, because this also happened in the short range, makes no sense. One may reason along the lines presented above, that in the short range forecast, GCMs, which have skill in very many dofs initially, should be much better than any linear (dynamical or empirical) method. This has indeed been found to be the case since about 1965. We have to identify the dofs we may be able to predict a season ahead of time under ideal circumstances, then rationally proceed by deciding which tools are the best approach in a real time forecast setting. This does not reduce the importance of models. A good simulation of atmosphere-ocean behavior on all time scales is very important and has many applications. For instance, it is unlikely we can estimate short-term climate predict***ability*** from anything but dynamical models.

At this point the reasoning along the lines of effective degrees of freedom and functional linearity is conjecture. We have obviously not proven that the above lines of reasoning, step 1 to 5, are correct.

10.3 Predictability

Predictability is thought of as the prediction skill one could achieve under ideal circumstances. Predictability is a ceiling for prediction skill. It helps to know predictability, so as to stay realistic, or to see how much improvement still awaits us. Below we give four approaches to determine prediction skill and predictability that have been in vogue over the last 40 years.

Hopefully this inspires the readers to some original work where it is needed most.

Approach 1: Evaluation of skill of real time prediction the old-fashioned way. Simply make forecasts and wait 25 or 50 years to see how well a method performs. Problems include a) small sample size, and b) a long waiting time (and funding agents are impatient).

Approach 2: Evaluation of skill of hindcasts. This removes the long waiting time aspect. Problems include a) small sample size, b) 'honesty' of hindcasts (overfit, tuning problems and fundamental problems with cross-validation) and c) hindcasts cannot be done for official forecasts, only strictly objective unambiguous methods.

Approach 3: Predictability of the 1^{st}kind (sensitivity due to uncertainty in initial conditions). Or, how long do two perturbed members in an ensemble stay closer together than randomly selected historical states. This method (Lorenz 1982) is very famous because of its clear connection to popular chaos theory. A model vs model verification amounts to a perfect model assumption. Problems include the choice of the size of the initial error and the nature of the error (growing, decaying..). Other problems arise if the spread of the ensemble members is low, or equivalently, the model's variability is lower than in nature. Studies of the predictability of the 1^{st} kind have lead to the insight that day-by-day weather in the mid-latitudes is predictable for at most 1 or 2 weeks, depending on criterion.

Approach 4: Predictability of the 2^{nd} kind due to variations in 'external' boundary conditions. This approach has come about mainly in the AMIP context (prescribed global SST variations) and goes by the names Potential Predictability, Reproducibility etc. Problems include unclarity about the lead of the forecast the predictability pertains to, and what is meant by external. Madden's (1976) approach based on data (the only empirical method to estimate predictability we are aware of) also fits in this category.

AMIP runs have indicated extremely high predictability of the 2^{nd} kind in the tropics (>50% explained variance), but more modest or sobering numbers in mid-latitude.

Approach 1 is out, because of impatience. With the development of 1-tier coupled atmosphere-ocean models, it appears that approach 4 has come to an end also. Only sensitivity to initial conditions (including erstwhile boundary conditions like SST) survives.[7] This leaves us only with evaluating hindcasts and predictability of the 1st kind as major tasks. Neither task is trivial.

There has been no developmental work after about 1985 on the method suggested by Madden(1976) who gave empirically based estimates of ~25% potential predictability in mid-latitudes for seasonal mean surface pressure. Except for CA (this book) and NA (Eshel 2006) empirical methods damp anomalies to zero as skill goes down with lead, and are thus unsuitable for predictability estimates. One needs diverging solutions to study predictability.

The definition of predictability, even the 1st kind, may need to be reworked. The older approaches were generally based on traditional verification, such as rms, anomaly correlation etc. But with the advent of creating full membered model ensembles new approaches using pdfs directly may have to be invented. Some recent work by DelSole (2004) points already in this direction.

Finally, the perfect model assumption needs work. While members of an ensemble obey exactly the same physics and numerics, why should the predictability estimate apply to the real world? In the end one must demand models to be good replicas of nature, including faithful simulation of say the Madden and Julian Oscillation and the QBO. The MJO is thought to be important for the shorter lead climate forecasts, but predictability estimates cannot be taken seriously when the models don't have an MJO, or have a weak MJO with erroneous phase speed.

We recently evaluated predictability of the 1st kind using the CFS model (Saha et al 2006), which is a state of the art 1-tier global ocean-atmosphere model. A single member was correlated against the average of 14 remaining members. The news is mixed. On the downside, the predictability of T and P over land in the NH does not exceed 0.4 correlation for any lead/verifying month. A more positive note is that the erstwhile boundary conditions (w and SST)

[7] Only CO2 increase, solar variability, atmospheric turbidity and the like come to mind as surviving external factors.

are not only predicted well but have even higher predictability. Over the oceans, modeling has apparently advanced to the point where we can beat the control forecast (persistence) for SST everywhere (not just in the equatorial Pacific). Over land, the high skill in w forecasts sis not as high as skill of persistence forecasts of w. This is indicative of problems with forecasting the sum of P minus evaporation minus runoff.

One must also understand that estimates of predictability look better, by definition, when ensemble members have low spread, low compared to the rmse of the control forecast. Others researchers appear to interpret low spread as motivation to add stochastic forcing to the models, so the increased spread and rmse are the same (and prediction skill and predictability become identically the same and all hope for improvement is gone).

10.4 The future of short-term climate prediction.

We end this book on prediction with speculation about the future. What will the state of short-term climate prediction be in the future?. And what are the priorities? A conservative approach to looking into the future is to extrapolate the advances of the last 10 years. Among the main advances:

1) The number of models or methods available in real time is increasing rapidly. Prior to 2000 organizations like CPC has only a handful of in-house tools. Technology and fast communication have allowed outsiders access to data they need to run their methods and make output available in a timely fashion. In spite of sobering estimates of prediction skill (and predictability) in mid-latitudes, the enthusiasm is enormous, both among modelers and empiricists. With so many new forecasts tools (hundreds of them) consolidation (Ch8) is an increasing priority.

2) Massive hindcast data sets. Each method should be accompanied by a hindcast data set, and increasingly this is what is happening. While the shining examples may have been for the seasonal prediction (Palmer et al 2004; Saha et al 2006), the generation of hindcasts is spilling over to the shorter forecast ranges as well. In order to run hindcasts, the reanalysis of land, atmosphere and

ocean (as far back as possible) is an increasing priority. Empirical studies are a major beneficiary of reanalyses also.

3) Probabilistic approach from beginning (perturbed IC) to end (application models). Although the seasonal prediction was one of the first to be expressed probabilistically, there has been a major push towards probability expression and verification. To a certain extent this may have been a spill over from the experience in ensemble forecasting in the medium range which gave new life to probability forecasting. Serious users are well served by reliable probability forecasts. The casual user may notice less benefit and feel excluded by the high abstraction level. Exactly how pdfs will be constructed from ~100 or 1000 members, each with a weight (based on skill and possibly co-linearity) remains a subject of study.

Two more topics were advances are required:

4) We need to come to grips with long term trends. Although trend tools, like OCN, are an ingredient in the seasonal prediction there is a dearth of methods contributing anything original about interdecadal variability. The emphasis has been on the interannual time-scale and ENSO. But now that the occurrence of the below normal temperature tercile has become a rare event we may need to adjust methods and presentation. A connection with longer term climate change research may be a natural avenue of progress with mutual benefits.

5) Since we work under a cloud of low predictability, we need to agree on how to define predictability, develop the notion, understand caveats and develop minimum requirements for any model to be used in a perfect model setting for a quasi-definitive predictability estimate.