Report of the Working Groups

1. PREDICTION OF FORECAST SKILL

Forecasts that extend beyond several days have an important and growing clientele, even though the average day-to-day skill of these forecasts can be characterized as modest to marginal. In order to make optimum use of predictions with such small overall skill, users need to be able to discriminate between those relatively few forecasts that are likely to be good from the larger numbers that are likely to be less successful.

Forecast skill variability includes the usual seasonal cycle but also encompasses large day-to-day and regime-to-regime fluctuations. Daily and regime changes in skill have been shown to have a non-random component and have been related to a variety of characteristics derived from the antecedent and forecast flow. Consequently, it should be possible to forecast forecast skill with some success, and attempts to do this are currently underway.

One line of effort emphasises the use of ensembles of forecasts and other more ambitious methodologies that will require considerable increases in operational resources. These approaches will be considered in sections 2 and 3. Here the emphasis will be on prediction of forecast skill in the context of present operational constraints. The subsequent discussion will include first a summary of potential predictors, their characteristics, relative merits, inter-relationships and potential usefulness, followed by considerations on how the predictors should be exploited (i.e. which methodology to use to produce which kind of products).

1.2 Predictors of forecast skill: nature, usefulness and intercorrelations

In a number of recent studies concerned with diagnostics or experimental predictions of forecast skill (see for example the contributions to this volume), the following predictors for the skill have been used or suggested:

1) Indices of the large-scale circulation (such as rotated principal components or teleconnection indices) deduced from forecast fields or from the initial analysis;

2) Spread (expressed in terms of RMS difference or anomaly correlation) between consecutive forecasts from the same NWP centre or between forecasts from different centres;

3) Day 1 error of the forecast or, possibly, 12-hour error deduced from a short-cut-off analysis;

4) Amplitude of the forecast anomaly;

5) Persistence of the forecast fields, or of the observed fields in the 2-3 days preceding the starting date of the forecast;

6) Energy budget terms deduced from forecast fields or from the initial analysis;

7) Regime-dependent instability diagnostics deduced from simple dynamical models.

All these predictors have advantages and disadvantages, and a lot of intercorrelation exists between them. Predictors 2 to 5 are very easy to compute, and can be deduced from a limited amount of data. Predictors 6 and 7 are much more expensive from a computational point of view, but they can provide a better comprehension of the underlying dynamical processes.

The circulation indices can give two kinds of information: they can show the dependence of the 'systematic' error of the model on the flow regime, or can be seen as statistical indicators of the instability of the flow itself, and consequently of the non-systematic component of the error. It is not clear if these two error components can be clearly separated. In any case, it seems that a better comprehension of this problem can be obtained if the indices explain some statistically relevant property of the variability of the real atmosphere.

When seen as indicators of quasi-systematic, regime dependent errors, the circulation indices can be complemented by the information about the energy budget. In fact, the systematic error of the ECMWF model is related to errors in the baroclinic conversion terms, and there are diagnostic indications about

the importance of baroclinic conversion by the ultra-long waves in the maintenance of the high-amplitude-planetary-wave regimes often misrepresented by most GCMs.

On the other hand, statistical indications about the dependence of error growth on the large-scale regimes provide a useful guidance to the instability studies that can be performed with simplified dynamical models. The results of these studies can be used to judge the dynamical significance of a statistical scheme for the prediction of forecast skill.

Another possible application of simple dynamical models is to provide a better use of the information about the very-short-range forecast error. A number of diagnostic studies have shown that downstream propagation and baroclinic amplification of analysis errors account for a large proportion of the short-range error. This will lead to additional complexity of the structure of correlation maps between, for example, day 1 and day 3 in addition to that provided by regime dependent space-time structure. Although some gross features of this structure can be revealed by purely statistical methods, experiments with simple models can provide independent guidance, given the limitations and the inhomogeneities of the currently available data sets.

1.3 Skill forecast schemes and products

The current experimental scheme for skill prediction in use at ECMWF assigns probabilities to five classes of skill. This probabilistic approach is probably best suited for this type of forecast. However, in view of the low correlations obtained, a three class approach may be preferred.

The regression technique used in the experimental scheme was based on a weighted average of 5x5 contingency tables. This scheme cannot cope with the unavoidable correlations between the predictors and it was therefore proposed to use well-known techniques like Regression Estimations of Event Probabilities (REEP) instead. Overfitting can be avoided through testing the significance of variance reduction in each step of the regression process and by extending the data set.

The data set used to develop the equations should be extended as much as possible. This is possible because many minor model changes are not

necessarily expected to have strong influence on these types of forecast equations. Additionally the scheme should be tested on an equally large dataset with independent data, perhaps through cross-validation techniques. The evaluation of the skill predictions in the independent sets needs special care because the standard evaluation methods are not very well suited for this type of forecast. Within the evaluation special attention should be given to periods when the numerical model is inconsistent, because skill prediction will be needed especially in these periods.

At present the predictions are issued for different preselected areas. However, there will also be a need for graphical output in the form of maps delineating areas with expected high and low skill. It is essential that this type of information be supported by maps with climatological information about the local model skill either in RMS or in ACC. For the time being both of these skill measures should be used in forecasting skill as well as in describing the climatology of the error field.

It is imperative that skill predictions are introduced in an early stage to the forecasters in order to give them the opportunity to gain experience and to provide feedback to the final product design. This introduction should be accompanied with some sort of educational program.

In the contribution Molteni and Palmer (this volume) it is shown that the skill of the skill prediction scheme improves when time averages of the skill are predicted. However, although the improvements are substantial, it was generally felt that the skill forecast should cover the same time window as the forecast. This means that in the short range, skill forecasts must relate to single forecast maps. In the medium and extended range time averages of skill forecasts will be more suitable.

1.4 Implications for MOS post-processing

At this moment no experience exists with Model Output Statistics (MOS) type applications using skill prediction forecasts in one way or another. Possible types of use are:

1) Adding the skill prediction to the MOS equation in such a way that it 'pushes' the resulting forecast to climatology in case of predicted low skill.

2) MOS-type equations for the prediction of credibility intervals associated with different forecast weather parameters like maximum and minimum temperatures.

3) Stratification of the dataset into high skill, normal skill and low skill subsets. Subsequently different MOS-equations can be developed for each of these subsets. This will lead to much sharper forecasts in the high skill subset.

These applications should be promoted by ECMWF and facilitated through cooperation with member states. Furthermore, for each of these applications, a substantial dataset will be needed to develop the MOS-equations. It is essential that these equations are developed with forecast values of skill and not observed values of skill. It is therefore important that the test data obtained during the final stage of the development of a skill prediction scheme are archived so that they can be used in the development of the above mentioned MOS-equations. Here again, cooperation between ECMWF and national services will accelerate useful application.

1.5 Recommendations

A. A full documentation of the temporal and spatial variability of the forecast error for the present ECMWF operational T106 model is required. Users should be made aware of the uncertainties in individual forecasts, especially those at the medium ranges. Both the systematic and random error should be documented as a function of season. Continual and updated documentation of the spatial distribution of the total error and systematic bias should be provided by ECMWF.

Data archives of the forecasts and analyses required to develop the climatological characteristics of the error fields of the T106 model should be made available to the research community. As a first step Lorenz tapes of the forecasts and the analyses of the 1000, 700, 500, and 200 mb heights should be developed followed later by surface variables such as temperature, humidity, wind and precipitation.

B. The ECMWF experimental forecast skill prediction scheme should be continued. Attempts to remove the covariability in the predictors as well as add other predictors should be considered. Studies into the

feasibility of providing probabilities for a smaller number of categories should be considered.

Although the present scheme is experimental, it would be appropriate at this time to consider what steps would be necessary to implement an improved scheme operationally. This would require informing the users of its possible availability, then creating the opportunity for a dialogue to determine their needs and to give them lead time for familiarization and development of applications.

C. Since this is a relatively new field of study, much research into the origin, variability and predictability of error characteristics are required. Characteristic circulation patterns associated with large and small errors need further study. Dynamical studies into the causes of this variability need to be expanded. Other studies such as the spread of forecasts inherent in lagged averages, Monte Carlo experiments and predictability experiments, should be considered further. Potential predictors such as forecast or observed persistence, and initial or forecast large amplitude anomalies should be studied, particularly their relationships to each other and different predictands. Energetics calculations should be continued. Simple models ranging from linear models to low resolution versions of the ECMWF operational model should be developed to study error characteristics as well as test potential empirical prediction schemes.

D. An important obstacle to a thorough description of model error climatology, optimization of the effectiveness of statistical schemes and understanding of relevant physical processes is the fact that homogeneous archived model forecast cases do not sufficiently represent the broad spectrum of common atmospheric flows, simply because of insufficient sampling periods. This situation is not likely to change if archives consist only of operationally produced cases. It is therefore recommended that a feasibility study be conducted to consider the augmentation of these archives with a reasonable number of forecasts run with the current version of the model from a selection of analyses from previous years. This selection would be based on a reasonable stratification of atmospheric flow and a comparison of the long-term distribution of flow types and their sample over the model period.

2. PROBABILISTIC WEATHER FORECASTING BY NUMERICAL METHODS

2.1 Introduction

Variability of forecast skill originates from a number of sources, e.g. the quality of the initial analysis, model deficiencies and intrinsic dynamical instabilities of the real (and model) atmosphere. All such possible causes are known to be strongly flow dependent. A probabilistic weather forecast provides the user with a variety of possible future atmospheric evolutions, together with estimates of their relative probability of occurrence and will thus give potentially very useful information about the relative forecast uncertainties.

Many recent observational, theoretical and numerical results seem to indicate that the Northern Hemisphere atmospheric circulation has a non unimodal distribution in some low-dimensional phase space of atmospheric states. The existence of such distinct "weather regimes" or "atmospheric modes" is supported by a wealth of synoptic experience, reflected in the concept of "großwetterlage", and of which blocking is one of the most important examples.

In this picture, both medium and (to an even larger extent) extended range weather forecast quality are heavily influenced by the model's ability to represent realistically transitions (or their absence) from one possible regime to another one. There is growing evidence that current numerical models show consistent deficiencies in forecasting transitions from one regime (e.g. zonal) to another (e.g. blocked). Such deficiencies can, again, be linked to the models' climate drift (or so called "systematic errors") but not exclusively to that.

Any probabilistic numerical weather forecast of some value should give indications about possible regime transitions and then, for each possible future atmospheric state, attempt to provide some more detailed information on smaller scale weather developments. This requires the model's statistical properties of regime transitions to be reasonably similar to the observed one. Although, for example, the onset of blocking beyond day 4 is currently poorly predicted (but in the range up to day 4 the statistics of models appear to be sufficiently close to the real atmosphere), it has been assumed , in the discussion that follows, that the situation will improve substantially in the

next few years, due to the improvements in the models' climatology that can reasonably be expected on the basis of current knowledge. This will make probabilistic weather forecasting by numerical methods a suitable tool to attack the problem of intrinsic deterministic unpredictability beyond a few days.

The most direct probabilistic weather forecasting system uses the Monte Carlo (MC) technique, in which many model simulations are repeated with slightly different (but similarly plausible) initial conditions. The results that can currently be obtained should, however, be interpreted carefully. There exists, for instance, no clear relationship between the spread of the ensemble and the skill. If both the model and the real atmosphere are in a stable regime, the spread of the forecast ensemble is likely to give good guidance on the probable forecast errors. If, however, the model is unable to represent a change of regime taking place several days into the forecast, the ensemble spread might again be comparatively small, although the ensemble forecast errors will be large, after the transition has taken place. In addition to using a model with the correct regime transition statistics, it is essential that the size of the ensemble is large enough to capture a realistic distribution of future atmospheric states and estimate quantitatively their relative probabilities. No precise evidence is currently available on the size of an ensemble adequate to do this (with purely random selection, it could for example be as high as 100; if a suitable selection strategy could be devised, though, a much smaller sample may be adequate). It is clear from these considerations that both the construction of the initial perturbations as well as the choice of the model are very important.

In section 2.2 various options for the construction of the cluster of initial states are discussed and in 2.3 the question of the appropriate resolution of the model to be used for MC forecasts is discussed. In section 2.4 the problem of postprocessing MC forecasts is examined and in 2.5 the possibilities for statistical dynamical models are explored.

2.2 Construction of the MC-initial perturbations

- Random perturbations

The easiest way of generating the initial perturbations is by adding random perturbations in grid-space or spectral space to the reference state. This

method poses serious problems in NWP models since the initialization in these models will dispose of all perturbations which project on the gravity modes, leaving an uncertain portion of the initial perturbation, little of which may readily excite unstable modes.

- Random analysis perturbations

In this method a small fraction of an arbitrary analysis (or short term error of an arbitrary forecast) randomly selected from the archives, (spatially weighted by a measure of analysis error) is added to the reference initial state. This procedure creates a largely balanced initial state and thus avoids the problems mentioned in the first method.

- Analyses from other weather centres

A simple possibility is to use the analysis of other NWP centres to create a small ensemble. This approach provides "perturbations" which are representative of analysis errors arising from data assimilation system differences and this can be interpreted as a lower bound for other methods of generating alternative initial conditions. All available experience on such a method should be examined, including the possibility of perturbing directly the observations.

- Analyses by different observing system scenarios

Although the most accurate analyses are generally achieved using all available observations, there are periods when the information from different observing systems are in conflict with each other. By withholding data from some observing systems in parallel assimilations, a priori equally plausible initial states can be produced for MC forecasting.

- Use of the knowledge about analysis errors

Perturbed initial states for Monte Carlo forecasting can be constructed from estimates of the accuracy of the initial state. Several different techniques can be used to define the statistical properties of the analysis errors.

The optimum interpolation method provides an estimate of the standard deviations of the analysis error and its spatial structure. This calculation assumes that the error statistics that enter the optimum interpolation scheme are perfectly known which makes the analysis error estimates too optimistic.

However, it gives a good measure of the data distribution and the uncertainties in the analyses as caused by data (un)availability. The usefulness of the operational analysis error estimation is also reduced by the simple error growth that is assumed. A more realistic error growth, by a Kalman filter, would produce more accurate analysis errors.

The temporal evolution of the coefficients of the dominant EOF's of the analyses generally does not exhibit rapid variations. Any high frequency variation in the coefficients can be regarded as (large-scale) noise or analysis errors. Discontinuities in the time evolution of the coefficients can therefore also be used to identify bad analyses or rapid atmospheric changes.

- Unstable modes of the atmosphere

In order to provide a strategy for reducing the potentially large number of perturbations it may prove valuable to calculate the dominant modes of instability of either the time mean or evolving forecast flow. The structure of these modes and their corresponding adjoints will determine possible "erogeneous" zones where the dynamically determined evolution of the atmosphere is most sensitive to perturbations. Perturbations which do not project significantly onto these regions of dynamical sensitivity may then be ignored in the initial MC ensemble.

- Time lagging

In this approach a sequence of recent analyses progressively lagging in time is used as the initial ensemble. Currently, this yields only a sample of small size, and a sample in which the perturbations are biased towards the short term systematic forecast errors and systematic observation errors. Furthermore, in data void regions effective perturbations would be small. With the development of 4-dimensional variational analysis methods, and an increasing availability of asynoptic data, it may become feasible to consider time lagged forecasts initiated more frequently than presently possible.

- Perturbation in the physics

An entirely different approach is to perturb the parameters in the physical parametrization of the model (radiation, boundary layer, convection). This would also make it possible to perturb the model continuously, such that

developing modes can be enhanced. Sensitivity studies are needed to help the choice of parameters to be perturbed.

2.3 The choice of resolution for Monte Carlo forecasting

For a given availability of computing resources, the number of forecasts that can be run in a Monte Carlo forecasting procedure will depend on the resolution chosen for the forecast model. The scope for a gain in the number of possible forecasts due to a reduction in vertical resolution is relatively modest, and experience of the sensitivity to changes in vertical resolution is relatively limited, so the discussion below is focused on the question of the appropriate horizontal resolution. We note, however, that if the Centre's aim of a 31-level T213 is realized operationally within a few years, then a model with around half this vertical resolution might be used for Monte Carlo forecasting. The possible need for parametrization schemes to work effectively across a range of vertical as well as horizontal resolutions should thus be kept in mind as the higher resolution model is developed.

Increased horizontal resolution has undoubtedly brought about significant gains in the accuracy of predicted synoptic patterns and of forecasts of local weather elements, and further improvements remain to be made. Despite this, there is evidence to suggest that horizontal resolutions lower than the current T106 resolution could provide useful information within a Monte Carlo forecasting system. This is likely to be particularly true for the larger scales of motion that will probably be the principal object of attention in a system aimed at the later medium range. However, the effects of model resolution on forecasts of blocking onset should be borne in mind.

In a paper presented to the Centre's 1986 Workshop on Predictability in the Medium and Extended Range, Jarraud compared a set of 24 cases for which forecasts had been carried out at resolutions T21, T42, T63 and T106. The T21 forecasts differed from the T42 forecasts (in the RMS error of the 500 hPa height) by rather more than the error of the T106 forecasts for most of the forecast range in the winter cases illustrated, whereas differences between T63 and T42, and between T63 and T106 were on average much smaller (though by no means negligible). Also, scatter diagrams indicated that differences between individual T21 and T42 forecasts could be almost as large as differences in skill from case to case, a result not found in the comparison

of higher resolutions. These results suggest that T21 is not an adequate resolution for Monte Carlo forecasts in the medium range, but that it may well be worthwhile to assess the performance of T42. It should be noted, however, that no attempt was made to optimize the physical parametrization on horizontal diffusion to T21 resolution.

Further evidence for the utility of resolutions below T106 is provided by the experience from running T63 forecasts from ECWMF and UK Meteorological Office analyses in cases where substantial differences had been observed in the operational forecasts of the two institutions. Use of T63 resolution was sufficient to demonstrate the predominant contribution of initial analysis differences to the subsequent forecast differences, although when running from the better of the two analyses (that is, when obtaining a broadly correct flow evolution) the extra detail added by T106 resolution brought the forecasts closer to reality. In the context of the present discussions, one set of experiments that could immediately be carried out would be to repeat the available cases with T42 resolution, to establish the extent to which this resolution too is capable of identifying the consequences of the analysis differences. Establishment of blocking featured in some of these cases.

Assuming satisfactory performance of T42, and the availability of a suitable method for the generation of a sizeable sample of initial states, the next appropriate step appears to be a Monte Carlo case study in which a large sample (at least several 10s) of forecasts is generated using T42, and compared with a smaller sample of T63 forecasts, and with one (or a few) T106 runs. The case should be one in which a marked transition of flow type occurs, and the initial date could be chosen as the earliest at which the transition is accurately captured by T106. Comparisons of conclusions from the full T42 ensemble with those drawn from subsets of different sizes should be made.

It would also be of interest to investigate a case of intense cyclogenesis, such as the major storm of October 1987. In this case it is unlikely that any of the lower resolution forecasts would simulate the intensity of the storm as captured in the best possible deterministic forecast, but useful indications of uncertainty of track and intensity might be gained.

2.4 Post processing of the Monte Carlo forecasts

Even with a modest sample size, some post processing of the ensembles will be necessary in order to compress the amount of information generated by the Monte Carlo technique. This is clearly vital for ensemble forecasting to be of practical use, but also must be considered carefully when analysing results in research mode.

Despite the fact that the RMS error is generally smaller for any ensemble mean forecast than for individual forecasts comprising the ensemble, and in the light of the considerations about atmospheric regimes outlined in the introduction, the straight ensemble mean forecast will not be a particularly suitable product (for the user who explicitly requires information about possible alternative scenarios). In this respect techniques for estimating, at each grid point, probabilities of occurrence of certain variables (e.g. 850 mb T) within predefined classes would be valuable.

From the research point of view in particular, it may be instructive to apply techniques of cluster analysis to identify groupings of trajectories of model evolution. In principle, estimates of the cluster mean, the probability of occurrence of each cluster, and standard deviation within each cluster, could be made. In practice, however, this may require large member ensembles to produce statistically significant results. Nevertheless, even if only a qualitative indication of clustering could be made, this would provide useful information to supplement the ensemble standard deviation diagnostic. For example, if the ensemble evolved into two widely separated clusters, and the intra-cluster standard deviation was small, the largeness of the total ensemble standard deviation would not necessarily imply total loss of predictability.

The final post processing of the Monte Carlo forecasts is highly dependent on the real need of the users. A market study to investigate such needs should therefore be initiated in the coming future.

2.5 Stochastic dynamic prediction

The problem we are addressing here is the explicit deterministic forecasting (via appropriate evolution equations) of the probability density function of the atmosphere. Due to the large number of degrees of freedom of current

operational models and to the non linearity of the equations, which leads to the need for a closure, one has to make simplifications to deal practically with the problem.

Lacarra and Talagrand (1988) have shown that the evolution in a shallow-water equations model of meteorologically realistic perturbations is well described for a period of something like two days by the tangent linear version of the model. Urban (1985) reached the same conclusion for a low truncation quasi-geostrophic model. Under that assumption, one can explicitly compute the covariance matrix of the forecast error, using a Kalman filter in forecasting mode. The initial covariance matrix of this forecast is provided by the use of the Kalman filter in assimilation mode.

This way of tackling the problem has also applications for data assimilation work i.e. structure functions in the optimum interpolation context, choice of the distance function and of the matrix in the variational approach.

It is possible to think about other approaches to the problem, e.g. higher order moment closure or band description of the covariance matrix. However one will always have to bear in mind that the real practical difficulty is the size of the problem.

2.6 Recommendations

A. Diagnose the practical predictability of regime transitions using available forecasts and analyses archives (e.g. ECMWF).

B. Assess the currently available best technique to generate Monte Carlo (MC) forecast ensembles.

C. Perform resolution studies in known cases of strong sensitivity to analysis details to assess lowest acceptable resolution for MC studies.

D. Perform 4-5 large ensembles with the best available technique to produce MC initial conditions to diagnose the areas of highest sensitivity to perturbations (that project readily on unstable modes) and compare with results from theoretical calculations of dynamical (baroclinic/barotropic) instability studies. Evaluate the relative

advantages of a large ensemble of lower resolution forecasts as opposed to a smaller ensemble of intermediate resolution forecasts.

E. Perform sensitivity studies to help the choice of physical parametrization parameters to be perturbed in MC-type integrations.

F. Encourage theoretical studies on the evaluation of the dimensionality of atmospheric attractors since it may have impact on the number of elements in a MC ensemble.

G. Initiate an investigation on the needs of users in the field of probabilistic-type weather forecasting products.

H. Given available resources, continue previous studies of the use of a poor man's ensemble taking contemporaneous forecasts from different centres.

References

Lacarra, J.-F. and O. Talagrand, 1988: Short-range evolution of small perturbations in a barotropic model. Tellus, Volume 40A, Number 2.

Urban, B., 1985: Error maximum growth in simple meteorological models (in French). (Météorologie Nationale Internal report).

3. EVALUATION OF EXTENDED RANGE DYNAMICAL FORECASTING

3.1 Current skill levels for prediction of the extended range

The current levels of skill for forecasts of the medium and extended range are evaluated using a variety of measures. Root mean square error and anomaly correlation applied to the 500 mb height field are by far the most common. For anomaly correlation, the result will depend on the "climatology" from which the anomalies are calculated. It is important to have a uniform and generally available climatology for this purpose as well as for general comparisons of model and observed climatologies.

The existence of useful skill in the day 1-30 time mean has been well established by numerous experiments. There is also considerable evidence of skill in 10 day time means of the forecasts centered at day 10 and for 10 to 40 day means of 850 mb temperature. The most common value used to evaluate this skill is the anomaly correlation for Northern Hemisphere 500 hPa height. Among the different models, generally values between .3 and .5 are found for 30 day means. A high percentage of skill is due to the first 8 days of the 30 day means and generally an average over the first 8 to 10 days gives the best estimates of 30 day means but there are spectacular exceptions.

Means over 10 days show similar values (.3-.5) at least for 4-13 day averages but on occasion the 10-20 day means have shown this value of skill as well. The 850 hPa temperature seems to have a better skill. This field moreover, is more directly connected to human activities. A point to be clarified is the relationship that exists between these fields and the surface temperature or the skill of a direct prediction of surface temperature.

The best seasons for the forecast are winter and spring and the main results presented were related to winter forecasts. The principal source of error in winter is the failure to predict the onset of blocking in the integration after 4 days since a good extended range forecast requires a good medium range forecast.

Systematic error is that part of the forecast error which survives ensemble averaging. It differs for different models although there are also common aspects. It is generally decreasing as models improve. Non-systematic error constitutes the largest source of error at all time scales but systematic

error may account for 10 to 50% of the total. The amount of systematic error exhibited by various models is not a clear function of resolutions since some low resolution models have smaller systematic error than higher resolution models. Different models exhibit systematic error for different flow components with error in the zonal components often but not always the largest contribution.

When comparing impacts from different resolutions on the skill of the forecasts there is a clear indication for the ECMWF model that a T21 resolution is too low but that T42 would generally be adequate. A further increase in resolution will improve the extended range skill further but only slightly. Other research has shown useable skill from low-resolution general circulation models which are tuned to minimize the systematic part of the error.

From short and medium range forecasts it has been found that forecast skills of different meteorological centres are highly correlated. Efforts should be made to find out if this is also true for extended range forecasts.

3.2 Theoretical limit of time-mean dynamical predictability

In view of the evidence of instantaneous limit of predictability being 10-14 days there remains a fundamental question regarding possible forecast skill at extended range. Estimates of predictive skill based on identical twin model experiments require : 1) that the models be capable of simulating a wide range of atmospheric variability and 2) that the ensembles of initial states be capable of exciting the wide range of atmospheric variability at realistic rates. Experimentation has shown that the time-mean forecast skill is considerably larger than the above mentioned estimates of instantaneous skill. For instance, identical twin predictability experiments show that the potential skill in a 30-day time mean is from 0.5 - 0.7 in the ACC for the 500 mb heights in the mid-latitudes. For some models with insufficient low frequency variance (compared to observations) potential predictability is not as high, whereas for other models with unrealistically large low frequency variance, potential predictability is far higher. For shorter time averaging, 10-day means centered on day 20 show potential predictability of 0.4 - 0.6. Evidence based on a large number of runs by operational and research Centres of exceptional forecast skill in the extended range is also found in many

actual forecasts up to 30 days. Most of these results are based on measurements of the 500 mb height skill, however, other variables show similar behaviour.

The above estimates of potential predictability are only measuring variability of internal dynamical processes. There is additional potential skill due to external forcing which has not been considered in these estimates. This variability and its potential predictability is especially large for the tropics. However, up to 30 days the ratio of external forcing to influence of the internal variability forcing is of the order of 0.1 - 0.2 for the mid-latitudes and therefore we feel that these represent a reasonable estimate of the theoretical predictability.

3.3 Diagnostic relationships concerning forecast skill

The current consensus is that evidence of useful correlation between ensemble spread and skill exists for the medium-range, say days 1-10 and possibly for days 6-15. Typical correlation values range from 0.3 to a maximum of 0.6 for regional verification. However the level of correlation decays quite rapidly with time, and is negligible at extended-ranges. Levels of correlation are somewhat higher when correlation-based spread/skill measures are considered, compared with rms statistics. In a synoptic sense there is a very close correspondence, through at least the medium range, between error fields and fields of forecast/forecast differences. However, based on present knowledge, the existence of small ensemble spread is a necessary but not sufficient condition for the occurrence of high forecast skill.

Ultimately the prospects for extracting reliable skill predictions from the spread/skill relationships depend on the ability of the models to simulate accurately the probability distribution of alternative evolutions consistent with the initial conditions. The comparison between the UK Met Office and ECMWF T63 model ensembles suggests tentatively that divergence between models can exceed the ensemble spread of each separate model in particular experiments. If different models contribute independent predictability information in such cases, there may be some benefit in constructing ensembles using a number of such models. Such an approach is already being pursued by NMC using the operational medium-range forecasts from NMC, ECMWF, and the UK Met Office.

A major contributing factor to the variability in skill is the present inability of models to capture blocking events beyond the first few days into the integrations. Additionally there appears to be a consensus that a strong relationship exists between the skill of predictions and the amplitude and sign of the PNA pattern. The NMC study strongly suggests that the relationships between skill and the PNA and blocking are indeed linked. Further evaluation and larger samples are necessary to confirm this link and to uncover any additional relationships that might exist between forecast skill and circulation regime.

3.4 Priorities for future research

There was a strong consensus from experimental work to date and predictability studies on dynamical extended range forecasts that research should continue in this area. We first recommend further experiments in potential predictability along the following lines: 1) It is suggested to look at the existing datasets of forecasts and calculate theoretical predictability from the forecast differences (e.g. the UK/EC intercomparison). 2) Further experimentation is necessary to evaluate sensitivity of identical twin experiments to differences in initial condition perturbation and to different models.

Since there appears to be substantial model sensitivity in the results, it is important that experimental programmes, and case studies, be coordinated as far as possible so that many cases can be run with different models from different analyses with different resolutions and with different physics packages. It is recommended that other centres try to repeat cases already used for the ECMWF mid-month couplets, for the UKMO/ECMWF seasonal ensembles and for the special study subsets of the NMC DERF experiments.

It is strongly recommended that ECMWF, UKMO and NMC produce and disseminate 10-year global/hemispheric climatologies for 1979-88 as soon as feasible.

There are important questions concerning the best strategy for future work. Current tentative plans indicate that the various Centres will use different experimental strategies. Some will make large ensembles of low-resolution runs, and others will make ensembles of high-resolution runs, and some a mixture of both.

There is no clear guidance available on the best method of generating the initial ensemble, when one can choose between the Lagged Average Forecast (LAF) approach, the Random Analysis Forecast (RAF) approach, or some other variant on the Monte Carlo (MC) approach. More experiments are needed in these areas. Experiments on these questions are also possible very cheaply on the medium range forecasts available on the GTS, where several analysis/forecast systems produce different forecasts from different subsets of the observational data. This body of data is the most complete Monte-Carlo experiment available.

The question of the optimal size of ensembles is still open; however evidence suggests that the size should be at least more than 4.

The work to date has indicated important dependencies of forecast skill on the flow. The relationship between forecast skill and the phase of the PNA has been documented at several centres, as has the relation between low forecast skill and the onset of blocking. Statistical investigations between the flow and scalar measures of skill are now being extended to relations between the flow field and the error field. These latter investigations are inherently more difficult than the simpler investigations. Given enough cases, they may offer the possibility of improved physical understanding through compositing techniques, estimates of preferred error tracks, and mechanistic experiments to test the resulting physical ideas.

Other physical measures of forecast performance, such as error budgets, instability studies, or energetics budgets may also provide clues to understanding the wide fluctuations in forecast skill. Error budget equations could help to distinguish between large scale barotropic error growth and shorter scale baroclinic error growth. They would thus provide quantitative tests for current interpretations of regime dependence of forecast skill.

Finally, results from several centres show that onset of blocking poses difficult forecast problems, already in the medium range, and therefore also in the extended range. Detailed case studies, coordinated between centres,

probably offer an important avenue for progress in understanding if these events are inherently unpredictable, or are merely not forecastable at present because of deficiencies in the analysis/forecast systems.