The EUROSIP system - a multi-model approach

T. N. Stockdale

ECMWF, Shinfield Park, Reading, UK
t.stockdale@ecmwf.int

1 Introduction

Seasonal prediction with numerical models is a challenging science on many levels. There are challenges to identify and understand the full range of processes that contribute to predictability, given the complexity of the physics, the non-linearity of the dynamics and especially the limited number of years that we can study observationally. There are challenges to initialize the various components of the climate system, especially given the general sparsity and non-stationarity of observational data. There are challenges to calibrate and interpret model output, again given a limited and non-stationary past, and to communicate information and its limits to a wide range of users without misleading or confusing. But the challenge I want to focus on here is the challenge of dealing with model error.

Model error is a fairly general concept, and I would like to be a little more precise in what follows. By **model error** I mean problems, inadequacies and imperfections with the model formulation and its numerical implementation. Note that model error does not mean coding mistakes, although of course any such mistakes might contribute to model error.

This model error causes integrations of the model to produce results which are unrealistic in various ways; e.g. the model climate (mean, variability, features) may be unrealistic. Often these unrealistic features are themselves referred to as model error; certainly it is common to talk of biases in model output as "systematic errors" of the model. This is a usage that I may slip into, since it is so common, but it is not the main focus of this presentation.

The imperfections in the model also contribute to errors in any seasonal forecast produced by the model. This contribution I define here as the **model forecast error**. We do not know its value in any particular case, but may try to estimate its statistical properties.

There are many examples of model errors and their effects, and of how once one aspect of a model's performance is wrong, errors in other aspects become inevitable. Examples that we have faced over the years which are relevant to seasonal prediction include: a biased mean ocean state in the equatorial ocean, which then affects SST variability - if the model thermocline is depressed too far, than SST variability will be damped; an incorrect position of the boundary between cold tongue and warm pool in

the western equatorial Pacific, such that the processes driving local SST variations become totally incorrect; an incorrect distribution of mean precipitation, such that shifts in precipitation inevitably give wrong anomalies - and also, given that tropical precipitation is a major driver of the whole atmospheric circulation, incorrect precipitation implies that both the mean and anomalous atmospheric circulation is wrong.

There are countless other model problems that could be described. Overall we believe that we have a broad spectrum of model errors. This means that when we improve a particular process in a model, the overall impact is almost as likely to be negative as positive. It is the fundamental reason why progress in reducing important categories of model error is very slow. For example, surface wind biases over the equatorial oceans are not really any better now than twenty years ago. To make matters worse, seasonal forecasts are highly sensitive to model errors. We are typically trying to calculate a relatively small shift in the centre of the probability distribution of an observed quantity; for temperature this might be a change of order of 0.5 deg C. If we integrate models forward in time for several months, it is easy to get much bigger errors than this. Of course, we use our standard bias correction technique to remove the effects of model bias to first order, but as just described, this is not enough - many model problems have strongly non-linear effects. In fact, model errors are believed to be the dominant cause of errors in the probabilities that we calculate with our seasonal forecasting systems.

So, if model errors are dominant and also painfully slow to reduce, needing perhaps many more decades of progress to reduce to acceptable levels, how can we hope to improve the reliability of our seasonal predictions?

2 The multi-model concept

The basic concept of a multi-model approach to seasonal prediction is very simple. Different coupled GCMs have different model errors. There may also be quite some errors in common, which the multi-model approach will not address, as we discuss later. We take an 'ensemble' of model forecasts. The mean of the ensemble should be better, because at least some of the model forecast errors will be averaged out. The 'spread' of the ensemble should also be better, since we are sampling some of the uncertainty. Thus by averaging over a number of imperfect models, we should get to a better forecast.

To consider this in more detail, note that we are dealing with an ensemble of forecast values, not an ensemble of models. The importance of this distinction will become apparent later.

Let us start by considering an "ideal" multi-model forecast system. We will assume a fairly large number of models (for example, 10 or more). We assume that the models have roughly equal levels of forecast error. We assume at this point that the model forecast errors are uncorrelated. And we assume that each model has its own mean bias removed.

Consider the case of a specific forecast. A priori, we consider each of the models' forecast pdfs equally likely, since they have equal levels of overall skill. Note this is a Bayesian sense of "equally likely" – in reality, all the model pdfs will be wrong, since all the models are imperfect.

A posteriori, this "equally likely" consideration is no longer valid: forecasts near the centre of the multi-model distribution have a higher likelihood. The situation is analogous to making repeated measurements of a length or mass: with a large number of measurements, the likely value of the measured quantity is known to be close to the centre of the distribution of measurements, and simple theory shows that the uncertainty is $\sigma/\sqrt{n}$, where σ is the standard error of the individual forecast model errors, and n the number of models averaged over.

Note that this idealized multi-model ensemble is very different to the more common ensemble forecast where the initial conditions are perturbed to represent uncertainty in the initial conditions, and then an ensemble is run with a single model. In such a perturbed initial condition situation, and assuming either that the model is run far enough forward to "scramble" the initial conditions or that the distribution of initial conditions are chosen to be equally likely, then the forecast values are considered to be equally likely and the uncertainty in the forecast would be σ, not $\sigma/\sqrt{n}$.

In practice, each model is run with a finite ensemble size, and so there will be some sampling uncertainty in the ensemble mean forecast of each model. We can take this into account, and then estimate the uncertainty in the multi-model ensemble mean, using a $\sqrt{n}$ factor for the number of models. We can then estimate a p.d.f. by, for example, assuming that the observed outcome will be normally distributed about its true expected value with an uncertainty given by the width of the distribution seen in a single model forecast distribution. The method by which a multi-model ensemble forecast is constructed and used to estimate the impact of model error on forecast uncertainty means that, even in an ideal case, the multi-model ensemble distribution is a distribution, **not** a p.d.f.

To move from this ideal case to a more realistic multi-model ensemble, we need to relax our assumptions and account for some other facts. Firstly, we cannot expect model forecast errors to be independent - we often see similar errors or features across a range of models. If we allow for the dependence of model forecast errors, this will reduce the degrees of freedom over which we sample, thus giving a smaller effective value of *n*, and increasing the uncertainty in our forecast. In some cases, the reduction in *n* could be drastic, and even a large number of models may only give a limited reduction in the errors in our forecast. Other assumptions (number of models, similarity of error statistics) are less fundamental, and can be addressed in theory by use of weighting and working to extend the ensemble.

To return to the key point of this section: a multi-model forecast ensemble is not a p.d.f., even in ideal circumstances. A single model forecast ensemble is not a p.d.f. either, but could be expected to approach a p.d.f. in a perfect model scenario. The general situation is that forecast ensembles need to be interpreted appropriately to

give a forecast p.d.f. It is this p.d.f. which should then be verified using probabilistic forecast statistics.

3 Multi-model results

The efficacy of a multi-model approach to seasonal prediction was well demonstrated by the DEMETER project, an EU funded project run during the years 2000-2003. This looked at multi-model seasonal prediction using seven coupled general circulation models. Some benefits of a multi-model forecast were overwhelming, but perhaps not that surprising: Figure 1 shows the Brier Skill score (BSS) for a single model versus a multi-model combination as a scatter diagram, where each point is a comparison of the score for a given region, lead-time, start date and model, calculated over a 43 year period (see Hagedorn et al, 2005). In more than 99% of cases, the multi-model beats a single model. Despite this overwhelming advantage, the comparison is not fair - the multi-model has a bigger ensemble size than the individual model. Since in this study the forecast distribution is equated to the forecast p.d.f., small ensemble sizes are penalised significantly.

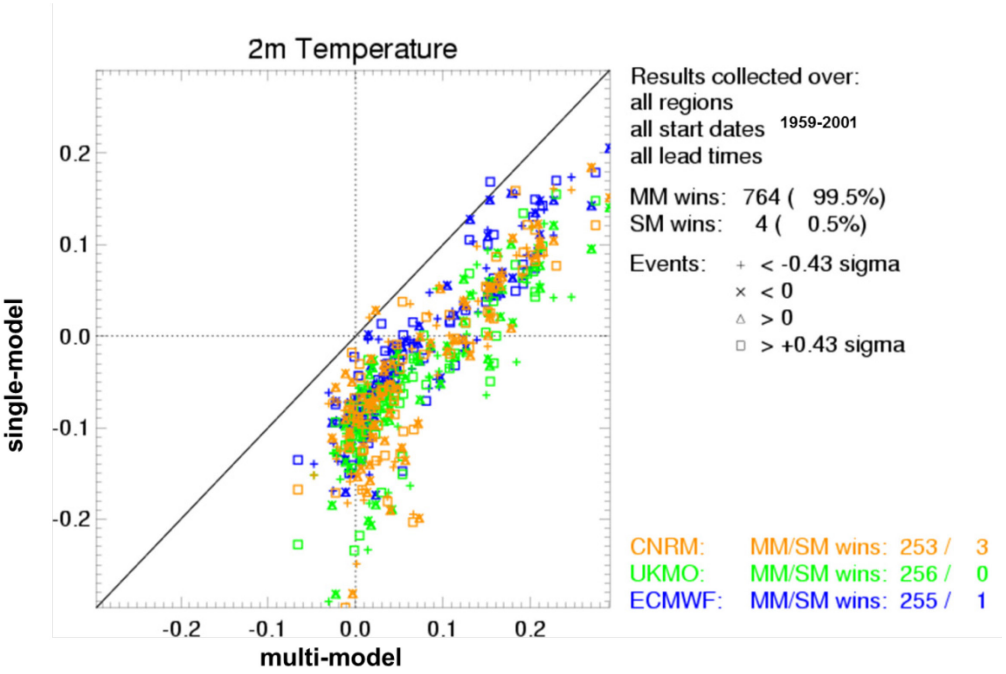

Figure 1: A scatter diagram showing single versus multi-model Brier Skill Scores, for forecasts covering a 43 year period. Each point compares the score for a given region, start season, lead time, and single model/multi-model combination. The multi-model beats the single model almost every time. From Hagedorn et al., 2005.

To enable fairer comparisons, one of the models was run using a large ensemble size, so single and multi-model forecasts could be compared with a constant ensemble size. This demonstrated that although ensemble size alone does help probabilistic skill scores, the multi-model combination gives a substantially larger benefit. A striking illustration is given in Figure 2, which shows the Rank Probability Skill Score for precipitation at grid points in the tropics. The blue lines shore scores for the ECMWF model, for increasing ensemble sizes from 9 (left) to 54 (right). There are multiple

lines because there are multiple ways to select e.g. 3 groups of 9 from the 54 members available. The coloured lines in the left column represent the scores from the 6 different models, each with 9 ensemble members. Note that the ECWMF model (in blue) is the highest scoring model for this statistic. We are going to compare a multi-model ensemble with a large ensemble of the highest scoring model! The second column shows, in red, the various possibilities for selecting two models from the available six. Several combinations beat the 18 member blue ensemble, although many combinations are still worse. With three models, the multi-model combination usually beats the best single model, and with four or more models it always does.

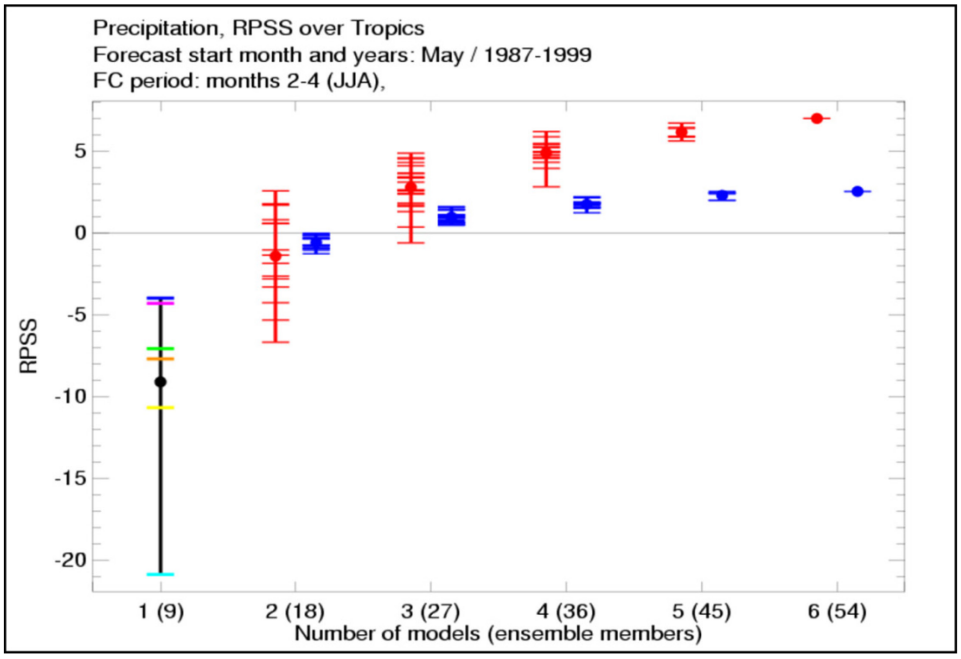

Figure 2: RPSS for precipitation in the tropics, for various multi-model combinations (in red) compared to an equivalent sized single model ensemble (in blue). See text for details. From Hagedorn et al., 2005.

4 EUROSIP

I want now to discuss EUROSIP, the operational multi-model seasonal forecasting system at ECMWF. This was developed following the encouraging results from DEMETER and other research projects, which established the scientific benefits of combining forecasts from several models. The first partners in the project were ECMWF, the Met Office and Météo-France. The initial design of the multi-model system had a number of features: a co-ordinated forecast strategy, a comprehensive and common data archive, and the production of real-time forecast products, although some aspects of this have evolved over time due to changes by the partners. The EUROSIP system became operational in 2005.

The data archive is an important part of the EUROSIP system. Individual model forecast data is archived in MARS, in a common data structure with the ECMWF seasonal forecasts. Both high frequency data (daily or more frequent) and monthly means are archived, although in recent years the Met Office have supplied only a limited number of monthly mean fields. The data is available to ECWMF Member States

for official duty use, and is also available for research and education, according to the terms of a data policy negotiated between the partners and ratified by ECMWF Council. This rich and common dataset is a valuable resource for many types of research, as well as supporting operational uses of real-time forecast data. ECMWF also creates, archives and disseminates multi-model data products, i.e. products derived by combining data from the different models. EUROSIP also provides support to the international community, for example by providing WMO access to the multi-model web products and supplying data to the EUROBRISA project in Brazil.

The most recent change to EUROSIP has been the entry of NCEP as associate partners in the EUROSIP project. Data from CFSv2 has been integrated into EUROSIP to give a four-member multi-model seasonal forecast system. This involved a number of significant technical changes to the system, allowing the CFSv2 data to be processed and archived as data with daily starts, and then combined into a "lagged average" forecast, as opposed to the existing EUROSIP models which are all archived as an ensemble starting from the 1st of each month (even if they originate from a lagged average forecast). There were also some scientific challenges, due to the well documented non-stationary SST biases in NCEP. This was dealt with by defining a stable sub-period of the available re-forecasts, and only using re-forecasts from this period (1999 onwards) to calibrate the real-time ENSO forecasts from CFSv2. Because both CFSv2 and the Met Office model have very restricted and only partially overlapping ENSO re-forecast periods, the multi-model system was further generalized to allow forecasts/re-forecasts to be produced when one of the input models is not available. This allows re-forecast statistics to be calculated over a fuller period, and prevents missing data from a single model crippling the calculation of statistics from the multi-model system.

Motivated by the introduction of variance scaling in ECMWF system 4, a slightly modified version of variance scaling was introduced into the EUROSIP system, for calculating El Nino SST plumes. Seasonal forecast systems are always corrected for model mean error or bias, estimated from the set of re-forecasts. However, models sometimes suffer from errors in variance as well as errors in the mean. This is especially true for SST variability in the equatorial oceans, where errors in the mean state of the coupled system easily result in SST anomalies being either systematically too small or systematically too big. It is straightforward to calculate the model SST variance over the re-forecast set, as a function of lead time and start date, and compare it with the corresponding observed SST variance. This gives a multiplicative scaling which can then be applied to the real-time forecast model SST anomalies. In practice, this approach has some problems, because of sampling. Although the mean bias is rather well sampled by O(20) years or more of re-forecasts, the variance is less certain. Cross-validated estimates are used when estimating skill for past cases, and these can be quite noisy depending which years are in or out of sample. In the past this has led to a conservative approach at ECMWF, only applying scaling in systems where the variance error was so large that it was obviously needed. In the revised EUROSIP system, we now apply variance scaling to all the models, but with some modifications to limit the possible impact of sampling effects. Firstly, the maximum scaling is limited to 1.4 - i.e. the model variance cannot be increased to more than 40% above its initial

level. Secondly, there is a gradual reduction of scaling greater than one applied in the case where the forecast value is much bigger than anything seen in the re-forecast set. The argument here is that we cannot be sure that variance errors are linear (in fact, we know that in many cases they are not), and that if our model is now in a physically different regime to the one used for estimating the variance scaling, we can no longer trust that scaling. If we are cannot trust the scaling, it would be dangerous to apply a large positive scaling to an already unprecedentedly large anomaly.

This modified approach to variance scaling seems to be successful and robust in handling output from a range of models. With full cross-validation, we find that the variance scaling improves the forecast statistics of every individual model; improves the consistency between the models; and improves the multi-model mean. Fig. 3 shows the improvement in Nino3.4 SST forecast skill from adding CFSv2 as an additional model (left), and from adding the model and revising the processing (right panel). The new EUROSIP system with NCEP is now much improved, but most of the improvement in fact comes from the revised processing.

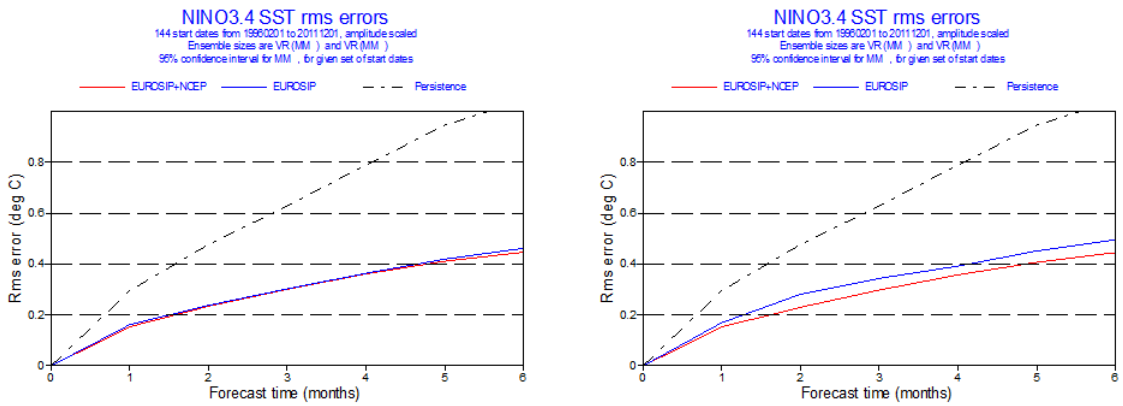

Figure 3: Nino 3.4 SST r.m.s. error reduction in EUROSIP, as measured using cross-validation over the longest available common period. Left panel: the improvement from including the CFSv2 model. Right panel: the improvement from including CFSv2 and including a robust variance scaling in the processing.

5 The EUROSIP calibrated p.d.f.

The original set of multi-model forecast products from EUROSIP are conceptually simple, in that they represent the simplest possible combination of the forecasts from each model. For example, the tercile probabilities for spatial maps are formed by averaging the corresponding probabilities from the individual models, and for the ENSO SST plumes, the ensemble members from the different models are all plotted together, as in Figure 4a below. There are two ways in which we would like to improve on this. Firstly, we believe that models can vary in skill, and that it is not always appropriate to weight the models equally. Secondly, since we know that even multi-model ensembles are not expected to be perfectly reliable, we want to translate model output into credible statements about the future - i.e. issue calibrated forecasts.

Both “improvements” are tricky to achieve. Since we have only a small set of past data (very small for the EUROSIP system), any attempt to find optimum weights for the models is fraught with the risk of over-fitting to the past. For atmospheric variables,

where predictability tends to be lower and therefore the sampling problems worse, we will not even attempt any weighting at the present time. But for ENSO SST, forecasts have a relatively high level of determinism at least in the shorter seasonal range, and since skill can vary appreciably between models (e.g. due to different mean state errors or different initialization procedures), it is worth considering whether we can move to a robust non-uniform weighting.

For the second improvement, to interpret the forecast output in terms of real-world probabilities, we need to construct a p.d.f. There are several ways this could be attempted, but we aim to keep things as simple as possible to minimise over-fitting problems. So, we assume the p.d.f. we want is a normal distribution, whose mean and variance must be determined. Since we don't know the variance, we will in fact end up with a "t" distribution.

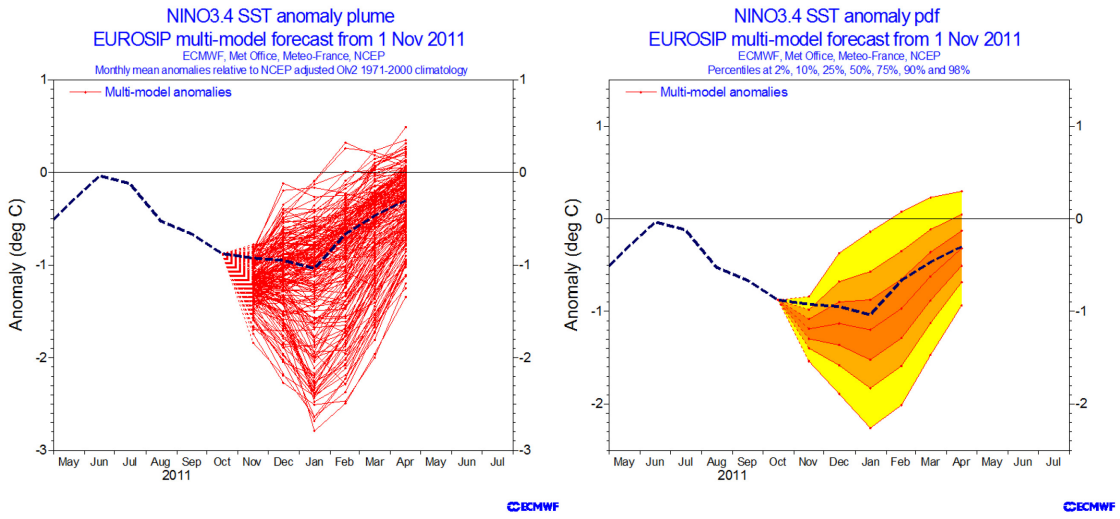

Figure 4: (a) A multi-model forecast comprising variance-scaled anomalies from each model plotted independently, and (b) the corresponding multi-model calibrated p.d.f., whose construction is explained in the text.

The method proceeds as follows. First, calculate a robust skill-weighted ensemble mean. We do not attempt a multivariate fit due to the small number of data points. We estimate the weights as proportional to 1/(error variance) for each model. This would be an optimal estimate in the case that the errors were independent. In reality, we expect the errors to be correlated, and the use of this weighting will not discriminate sufficiently between the different models. (To see this, consider that the part of the error variance that is common to models should ideally be ignored, but by use of this formula is added to the denominator of every term, leading to a more uniform weighting). However, this conservative approach to assigning weights is what we want. To make the method yet more conservative, the final weights are assigned as 50% uniform weighting, and 50% skill dependent as outlined.

Several comments are in order. All calculations are strictly cross-validated, which means it is easy to make the error statistics worse by having skill dependent weightings. (The case where model A is worst is the case where model A has its highest

weighting, because all the other years know about this worst case, but the worst case itself does not). Various other possibilities were tried. Rank weighting instead of error variance weighting did not help. A "Quality Control" term was tried, using likelihood estimates to downplay the impact of outliers (where if one model is very different to all the others in the forecast, it is considered suspect and given less weight), but again this did not help. An interpretation of this is that although outliers are usually wrong, there are some cases in which they are not, so ignoring them is dangerous. Finally, although we have attempted to produce a robust method of combining the models, most of the time the models actually agree reasonably well, so tweaking the method used to calculate weights actually has very little impact on calculated scores.

Having calculated the desired mean of our p.d.f., we must now specify its variance. As a first step, all of the multi-model ensembles (i.e. for the re-forecast dates and the forecast) are re-centred on their new mean values by adjusting the lower-weighted models. That is, the model with the highest weight is unchanged, and the others are moved so as to give an ensemble with the desired mean while having the least possible impact on the ensemble spread for each date. It is then possible to calculate the error variance of the mean forecast over the re-forecast period, and compare with the ensemble variance of the multi-model ensemble. Unbiased estimators are used in both cases. The multi-model ensemble spread can then be scaled so that the past forecast variance matches the past error variance. We have a choice at this point: do we scale the variance for the particular forecast (e.g. multiply by 0.9 or 1.1 or whatever), or do we use the climatological estimate of the forecast error variance from the re-forecasts. The choice depends on how much confidence we have in the multi-model's ability to predict interannual variability in forecast skill. We choose to take 50% of the variance from the scaled climatological value, and 50% from the scaled forecast value. This might be justified in various ways (not least by the results it gives), but note that it is inherently robust since if these estimates differ, the calculated forecast uncertainty will be dominated by the larger estimate. Thus, if either the re-forecasts or the real-time forecast suggest a large error variance, the real-time forecast will be given a large spread, and the risk of a disastrous over-confidence is reduced.

It is noteworthy that a simple analysis suggests that for the multi-model ensemble, some use of the predicted uncertainty improves the results, whereas for a single model with a comparable level of ensemble-mean skill, this is not the case. This suggests that even if the multi-model has comparable skill in the ensemble mean to a high performing single model, the multi-model has better information on how the uncertainty is varying.

The final stage in determining the p.d.f. is to specify the degrees of freedom for the "t" distribution. There are two main contributors to this, namely the limited number of years that the predicted and observed error variances are compared over, and the finite ensemble sizes used by the model. For the EUROSIP multi-model system, it is the limited number of years which is the main contributor, but the method is general enough to be applied to very small ensemble sizes, in which case the ensemble size becomes relevant. The use of the "t" distribution means that as either the number of past years or the ensemble size becomes small, the p.d.f. naturally broadens due to the increased uncertainty in the calibrated forecast.

The p.d.f. is plotted showing the 2nd, 10th, 25th, 50th, 75th, 90th and 98th percentiles, as in Fig. 4b above.

Several important points should be noted about interpretation. The p.d.f. has been calculated by calibrating the forecast system using past errors. The fact that the sampling period is limited is accounted for by the "t" distribution, so the risk of e.g. an El Nino event larger than previously observed is in principle already included. However, the risk of a real-time forecast having a new category of error is not covered. For example, a Tambora scale volcanic eruption (last seen in 1815) would have a major impact on SST, but the risk is not included by the analysis of recent forecast errors. For this and other reasons, we do not plot beyond the 2% and 98% limits of the distribution. Specifying the extreme tails of the distribution is a hard problem. A more likely reason for the forecast p.d.f. to become invalid is the risk of a change in the bias of the real-time forecasts relative to the re-forecast period. Of course, the re-forecast period itself is likely to contain some inhomogeneities, but real-time systems probably carry a higher risk of error due to changes in input data or analysis technique.

The p.d.f. refers to Bayesian probabilities describing our knowledge of what might happen in the future, accounting for errors previously seen in the models we use. A different multi-model system (using either a different set of models, or possibly a different combination methodology) would be expected to calculate a different p.d.f. – and both are correct. It is also clear that the better we are able to characterize the errors in our forecast system, the more precise our forecasts can become; and that our knowledge of the future is dictated more by the accuracy with which we specify the uncertainty than the precise value of the "best guess" that we derive from the ensemble mean.

If we purport to show a p.d.f., we should provide some validation. The main way in which the results of the method have been checked is by calculation of rank histograms. Cross-validated rank histograms show that the derived p.d.f.'s are remarkably accurate, including in the tails where one might worry that the risk of going wrong is higher. The chances of the observations lying in bins near the tails are what they should be, neither more nor less, and our use of the "t" distribution seems to validate well. Sensitivity tests show that verifying different periods (e.g. whole out of sample periods) rather than just using cross-validation can distort the p.d.f. in the case that the calibration and verification periods have a different mean bias. This may simply be a question of sampling (we don't have that many years to work with), but it is a reminder that if the mean bias in the forecast system changes, the reliability of our forecasts will be degraded.

6 Practical matters

Operational forecasting systems need to consider practical matters as well as the scientific basis of the forecasts and products. Experience has shown that quality control is an important part of a multi-model system, because a wide variety of problems and errors can and do occur. The ECMWF system is designed to be as automated as possible, because of our limited human resources. Although automatic systems can pick up some problems e.g. with incoming data, it must be admitted that

so far there is no complete substitute for manual input to the QC process. Particular care is needed when new systems are being introduced.

Operational systems need to run to a timetable. For EUROSIP, this means that the multi-model graphical and data products are released at 12Z on the 15th of each month, without fail. To enable this, we request that contributor data arrive a number of days before this, but there are occasions when data is late. A safety margin in the schedule allows some limited flexibility, and in the case that problems are found with the data, gives an opportunity for data to be re-sent. Nonetheless, the system allows a model to be excluded from the system "on the fly", should this be necessary. Our operational schedules also allow for weekends and the fact that there are times when computer systems are down.

7 The future

There are many ways in which the EUROSIP multi-model system can be expected to improve in the years ahead. Firstly, the constituent models will improve over time. Météo-France have a new system running, which will become operational imminently. The Met Office also has a new high-resolution forecast system, which they expect to introduce before the end of 2012. Beyond these immediate plans, all models will be slowly refined and improved over time. Better input data to the multi-model system will allow better multi-model forecasts.

More models are expected to join EUROSIP in the future. One important future contributor is DWD, who are working together with colleagues in Hamburg to develop an operational seasonal forecast system to contribute to EUROSIP. There are also other operational forecasting centres outside Europe who are interested in joining EUROSIP.

Finally, there is still important scope to improve the way in which multi-model products are calculated. The new Nino SST calibrated p.d.f. product gives an idea of the improvements that can be made - robust but differentiated combination of output from different models, and intelligent calibration against past performance, with full allowance for sampling errors due to the limited number of past cases. It should also be remembered that EUROSIP is but one part of the overall European and global infrastructure for seasonal prediction. Important work remains to be done in improving the flow of information from numerical models to end users, as discussed elsewhere in this Seminar.

Acknowledgements

The Met Office, Météo-France and NCEP are ECMWF's partners in EUROSIP, and have contributed to the data from which the EUROSIP multi-model products are derived.