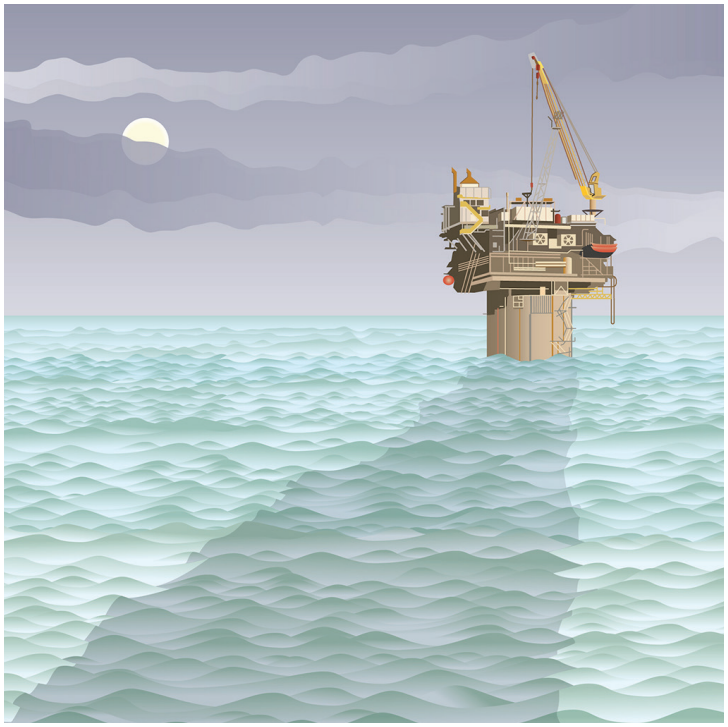
ECMWF Feature article

from Newsletter Number 118 – Winter 2008/09

METEOROLOGY

EUROSIP: multi-model seasonal forecasting



www.ecmwf.int/en/about/news-centre/media-resources

doi:10.21957/7wc0nybvir

This article appeared in the Meteorology section of ECMWF Newsletter No. 118 – Winter 2008/09, pp. 10-16.

EUROSIP: multi-model seasonal forecasting

Tim Stockdale, Francisco J. Doblas-Reyes, Laura Ferranti

ECMWF has run a seasonal forecast system for more than 10 years, and is presently on its third generation system. The forecast system is good at predicting El Niño related sea surface temperature (SST) anomalies in the Pacific, which are the major source of predictable seasonal variations in the weather around the globe. However, performance in predicting actual weather anomalies in many parts of the world is still substantially below the theoretical limits of what is possible.

Research has shown that combining forecasts from several different coupled ocean-atmosphere models is a robust and effective way to increase seasonal forecast skill. This is because combining models both averages out some of the individual model forecast errors, and also gives a better idea of the uncertainties in the forecast. ECMWF, the Met Office and Météo-France agreed some time ago to work together to develop an operational multi-model seasonal forecasting system, and EUROSIP was born. (Seasonal forecasting is often referred to in the research community as Seasonal to Interannual (or S/I) Prediction, which is the reason behind the "EUROSIP" name.) The cooperation of the scientific teams at Météo-France and the Met Office in developing the EUROSIP project is acknowledged.

Here we review the present status of EUROSIP and the performance of the multi-model forecast system.

How does a multi-model forecast help?

Despite successful El Niño predictions and constant efforts at model development, model error is still a major problem for seasonal forecasting. By 'model error' we mean the generic inaccuracies in the numerical model's representation of the real world. The problem is not specific to ECMWF – all existing models in the world have errors that limit the accuracy of seasonal prediction. Model error gives rise to model-induced forecast errors; these are errors in the individual forecasts that are due to the model. On a seasonal timescale, forecasts are inherently probabilistic, and so even a hypothetical 'perfect model' ensemble forecast of a particular variable would only give a probability density function (pdf). An ensemble of forecasts from an imperfect model will also generate a pdf, but it will differ from the 'true' pdf. We will refer to the difference between the pdfs as the model forecast error.

The most obvious model forecast error is bias – a model may be systematically too warm or too cold at some location, for example. Bias is estimated from a set of hindcasts (or reforecasts) made with each model, and this estimate can easily be removed from the real-time forecasts. However, the non-stationary component of forecast errors and non-linear effects are not accounted for by bias removal, and empirical corrections for these errors cannot easily be estimated from the limited number of past cases. Further, the forecast 'signal' that we are trying to predict is in most cases a relatively modest shift in the pdf from its climatology. Model forecast errors can easily overwhelm these signals, particularly away from areas of strong forcing such as the equatorial Pacific. Although model errors are being reduced, and will continue to be so in the future, the requirements for model accuracy are so exacting that model error is expected to be the dominant problem in seasonal prediction for decades to come.

So model forecast errors are endemic, hard to reduce, impossible to eliminate by a posteriori correction, and have a major impact on our forecasts. What can we do? A pragmatic approach starts by noting that although all models have errors, different models have different errors. Thus a multi-model combination can be useful – if we average the forecasts of several different models, some of the model forecast errors will be averaged out, while the forecast signal will remain undiminished. In practice, model forecast errors are likely to be partly correlated, and so averaging even a large number of models will not eliminate the error entirely. The number of independent models available is also limited. Nonetheless, averaging is able to reduce model forecast error to some extent.

A multi-model forecast system also helps by giving better information on the uncertainty of the forecast. A forecast pdf derived from a multi-model combination will typically be broader than one derived from a single model because the multi-model pdf naturally takes account of model uncertainty. The broader pdfs of multi-model forecasts increase their reliability, and allow them to gain higher verification scores when probabilistic measures are used. Forecast pdfs from a single model can be relaxed towards the climatological pdf to increase reliability, but this comes at the expense of 'damping' the forecast signal. Multi-model forecasts typically have increased reliability without loss of the mean forecast signal captured by the models.

 doi:10.21957/7wc0nybvir

A final benefit of a multi-model system is as a safeguard against the (hopefully small) risk of a real-time forecast system being corrupted in some way so as to produce misleading forecasts. For example, a real-time system might be inadvertently changed so that it systematically differs from the hindcasts; or it might fail to handle correctly a change in an external data stream; or data might be corrupted. Diagnosis of a problem by verification of the real-time forecasts is likely to be slow, and comparison with other forecasts might help to identify a problem much more quickly. Even for unrecognized errors, robustly constructed multi-model products will be much less impacted than the single affected model.

The benefits of multi-model combination

The EU-funded DEMETER project coordinated at ECMWF was a major step forward in establishing the practical benefits of multi-model seasonal prediction. Seven European coupled ocean-atmosphere models were used to make seasonal 'forecasts' covering recent decades, and the results of multi-model combinations were examined.

Key conclusions from DEMETER

Figure 1 shows a comparison of skill between multi-model combinations and single model ensemble forecasts from the DEMETER project. The ranked probability skill score (RPSS), a measure of probabilistic forecast quality, is shown for forecasts of June/July/ August seasonal mean 2-metre temperature at points in the northern hemisphere extratropics as a function of the ensemble size.

- ***Single model (blue lines).*** These results are drawn from a 54 member ensemble of forecasts from the ECMWF model, which for this particular forecast quantity and verification period is the best individual model. Each blue horizontal tick mark shows the result of one possible combination of members drawn from the total set of 54 members actually run. This means that the vertical spread in the blue lines represents sampling uncertainty in the generation of the ensemble.
- ***Multi-model combination (red lines).*** Results are shown for possible multi-model combinations, drawn from between 2 and 6 models (the number of models is shown in brackets by the side of the ensemble size). Since each model has nine ensemble members, the total ensemble size used is the same for the multi-model combination and the corresponding single model ensemble.

The vertical spread of the results from the multi-model combination (red tick marks) is typically larger than that of the single model (blue tick marks) since the skill of a multi-model combination depends on which models are combined.

The results in the left-most column of Figure 1 (coloured tick marks) show the skill of the individual 9-member model ensembles. Note that even a combination of three other models is likely to be better than a similar sized ensemble using a single model. By the time we get to four models, even the worst example of a four model combination beats the best result possible from the best single model. This general result is robust across many different variables, regions and seasons – for probabilistic forecasts, a multi-model combination is surprisingly effective when compared against a single model.

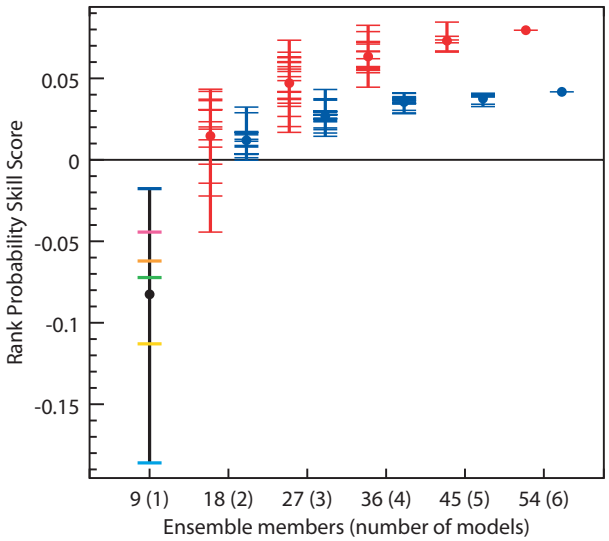


Figure 1 The ranked probability skill score (RPSS) for forecasts of June/July/August seasonal mean 2-metre temperature at points in the northern hemisphere extratropics as a function of the ensemble size, from the DEMETER project. The forecasts start from 1 May for the years 1987–1999. Red lines show the skill of multi-model combination, which is generally higher than the skill of similarly sized ensembles of the best single model shown by the blue lines. See text for details.

doi:10.21957/7wc0nybvir

Key conclusions from DEMETER are that:

- Multi-model combinations are more skilful than single models.
- The benefit is not just from having a larger total number of ensemble members.
- Adding a model with less-than-average skill to a multi-model combination is still usually of some benefit.
- A simple unweighted combination of models is usually the best approach, given the typically small sample sizes available for estimating model skill.

These very robust conclusions on the practical benefit of a multi-model combination were what drove ECMWF and its partners towards establishing an operational multi-model seasonal forecasting system.

Quality of the multi-model forecasts

The first consideration of a seasonal forecast system is the quality of the El Niño SST forecasts. Figure 2 shows the root mean square (rms) error of SST forecasts for the NINO3.4 index for the individual models of the present operational EUROSIP configuration (blue, green and orange) and the multi-model combination (red). The rms error of a simple anomaly persistence forecast is shown in black for reference. The multi-model combination is much better than the average of the models, and is fractionally better than the best single model.

Also shown in Figure 2 is the standard deviation of the ensemble forecast for a single model (dashed blue) and the multi-model combination (dashed red). The single model underestimates the uncertainty in its own forecasts, but the average spread of the multi-model forecasts almost matches the rms error of the forecasts. Despite this, the forecasts from the multi-model combination are still not properly calibrated – inspection of the individual forecasts shows that sometimes the multi-model combination clearly overestimates the uncertainty of a forecast, and sometimes it strongly underestimates the uncertainty. Preliminary results from a Bayesian calibration of the Niño plumes developed at ECMWF show a better scaling of the ensemble spread.

A further comparison between scores of operational ECMWF-only seasonal forecasts and those of the operational EUROSIP multi-model system is shown in Figure 3. This shows the ROC (Relative Operating Characteristic) skill scores of June/July/August seasonal mean 2-metre temperatures predicted from May for the years 1987–2005 for (a) the ECMWF model alone and (b) EUROSIP. The ROC skill score is effective at measuring the signal contained in a set of probabilistic forecasts and it does not punish forecasts for having poorly calibrated probabilities. Overall, the EUROSIP multi-model scores more highly, although the effect is relatively modest. The skill in summer temperature forecasts over southern Europe is apparent in both plots, and again the multi-model combination brings only modest gains. Note that the scores are fairly noisy and are only based on nineteen years of data – sampling uncertainties mean that detailed local comparisons are not appropriate. Plots of other variables and other seasons tell the same story – the EUROSIP forecasts are, overall, modestly more informative than the ECMWF-only forecasts.

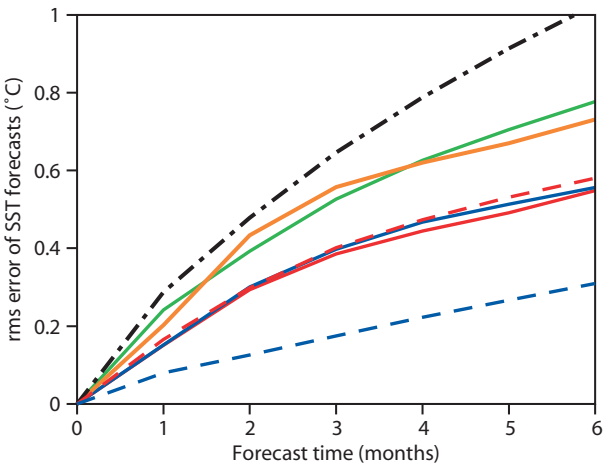


Figure 2 Root mean square errors of Nino 3.4 SST index forecasts from the EUROSIP multi-model combination (red line), anomaly persistence forecast (black line) and individual models (blue, green and orange lines). The multi-model combination is much better than the average of the models, and is fractionally better than the best single model. Also shown is the ensemble spread of the multi-model combination (dashed red) and the best single model (dashed blue).

doi:10.21957/7wc0nybvir

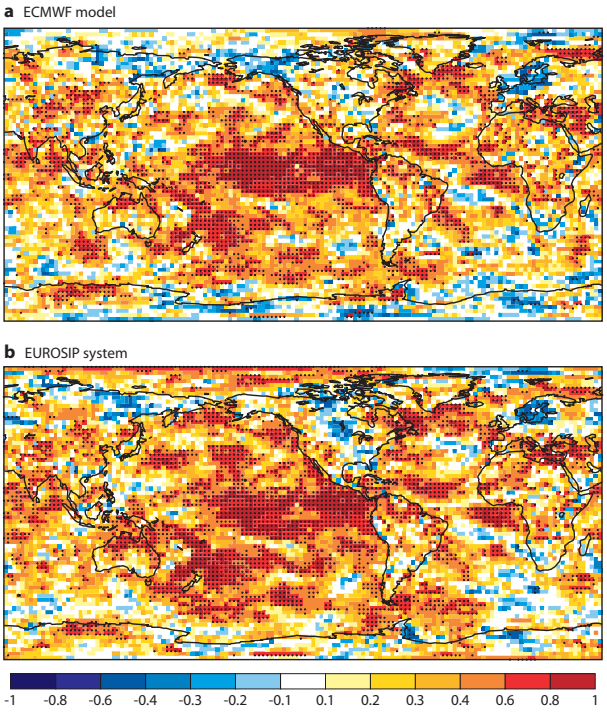


Figure 3 ROC skill scores for (a) the ECMWF model and (b) the EUROSIP system, for the event of the June/July/August seasonal mean 2-metre temperature being above the climatological median, as forecast from 1 May for the years 1987–2005. Black dots indicate values significantly different from zero with 95% probability. Scores are locally noisy, but the overall skill level of the EUROSIP system is higher.

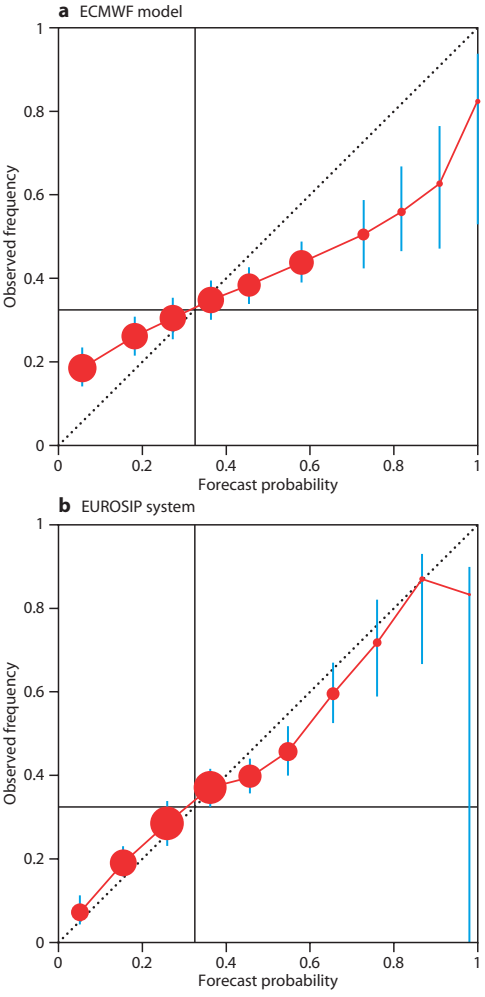


Figure 4 Reliability diagrams for (a) the ECMWF model and (b) the EUROSIP system for the event of the December/January/February seasonal mean 2-metre temperature being below the lower tercile of climatology, as forecast from 1 November, for the years 1987–2005. The red dots show the observed frequency of the event binned according to the probability its occurrence was predicted to have, with the blue error bars showing the effect of sampling uncertainty. The relative size of the red dots indicates the number of cases included in each bin. The black lines show the climatological frequency of the event. For a reliable system, the observed frequency should match the forecast probability in each bin. The EUROSIP system has a substantially higher reliability than the single model.

The reliability diagram in Figure 4 demonstrates the benefit of the multi-model system for the reliability of probability forecasts. Figure 4a shows the reliability of probability forecasts from the ECMWF model for seasonal mean 2-metre temperature being below the lower tercile for December/January/February for forecasts in the northern hemisphere extratropics. A reliable set of forecasts would have the observed frequency of occurrence matching the forecast probability – i.e. the points would be on the diagonal. Although the ECMWF forecasts have some ability to discriminate between different likelihoods of a cold winter, they are clearly a long way from being reliable. Figure 4b shows the result from the multi-model forecasts. The result is still not perfectly reliable, but is a big improvement on the single model result: note, for example, the change in the forecasts of a very low probability of a cold winter. The multi-model combination makes such a prediction less often (the size of the plotted circle represents the frequency with which the forecast probability is issued), but when such a forecast is made, it is much more reliable.

The substantial improvement in reliability relative to the ECMWF model is a general property of the EUROSIP forecasts, seen across other seasons and variables. Scores which are sensitive to the reliability of the probabilistic forecasts, such as RPSS, also benefit from the multi-model combination.

The present status of EUROSIP

The EUROSIP project presently involves ECMWF, the Met Office and Météo-France as partners – each partner contributes forecasts from a coupled atmosphere- ocean model to the multi-model system. Other organizations from ECMWF Member States or Co-operating States who would like to contribute can request to become EUROSIP partners. The German weather service (DWD) in collaboration with the Max-Planck-Institute for Meteorology intends to join the EUROSIP project in the future. Since spring 2005 graphical products from the multi-model system have been available to users in Member States. A formal data policy for EUROSIP was established by the ECMWF Council in December 2006, and in December 2007 the Council authorized the addition of a selection of EUROSIP multi-model data to the commercial catalogue.

The multi-model system works by combining the data from the operational versions of each contributing model. The main output of the multi-model system is a set of graphical forecast products that are discussed in the next section. Whenever one of the individual models is upgraded, the EUROSIP system will include the updated version. Typically, test data from a new model is made available for several months before the actual operational change, although this is not guaranteed. Each individual model is used to produce forecasts and also a corresponding set of hindcasts (or reforecasts). The hindcast data is used to estimate both model biases and also forecast skill. EUROSIP multi-model products always use the hindcast data corresponding to the real-time forecast data, so when a model version changes a new set of hindcast data is used. Information on the dates of changes in the various model components is available on the web.

In addition to graphical multi-model products on the web, certain EUROSIP products – based on the combined output of all of the models – are made available in digital form. These EUROSIP multi-model products are created together with equivalent hindcast multi-model products to allow skill estimation of the products.

The raw data for each individual model belongs to the contributing centre, and any commercial use of this data requires negotiation of terms with the owner. However, permission is granted to all Member States and Co-operating States to use the data for their official duty, and the data is also available for non-commercial research and education.

Full documentation of the EUROSIP system, including details of MARS access to the various datasets, is available on the web at www.ecmwf.int/products/forecasts/seasonal/documentation/eurosip/

 doi:10.21957/7wc0nybvir

Graphical forecast products and issues of interpretation

The EUROSIP graphical products are similar to those of the ECMWF System 3 forecasts, though with some differences. Products available include SST anomalies for key regions of the Equatorial Pacific (Niño plumes), probability maps for a range of atmospheric parameters, and predictions of tropical storm activity. The products are published on the web on the 15th of each month at www.ecmwf.int/products/forecasts/d/charts/seasonal/forecast/eurosip.

Niño Plumes

El Niño predictions are an important tool to anticipate the relative likelihood of regional climate anomalies. Niño plumes show the full set of SST anomaly plumes from all of the models, plotted together but without any adjustment or calibration.

Probability maps

Figure 5 shows an example of a multi-model forecast of 2-metre temperature for June/July/August 2008, and how it compares to the corresponding ECMWF-only forecast. The maps represent the probability of the most likely category of the seasonal mean 2-metre temperature being either above the 67% or below the 33% value of the model climate distribution. The forecasts are reasonably consistent, and the general tendency for the multi-model forecast to give slightly weaker probabilities (i.e. to be less confident) than the ECMWF forecast is visible. Sometimes the consistency between the ECMWF and multi-model forecast is lower than in this figure, reflecting the fact that the models disagree.

The consistency between the forecasts is not a reliable guide to either accurate or inaccurate forecasts, but it can give some information additional to that of the average past performance. In some cases inconsistencies between forecasts are related to the way individual models represent specific physical processes.

Tropical storms

EUROSIP predictions of tropical cyclones are produced by combining the calibrated forecasts of the individual models using equal weights. As discussed in *Vitart et al.* (2007), the skill of EUROSIP forecasts is generally higher than that of the individual models.

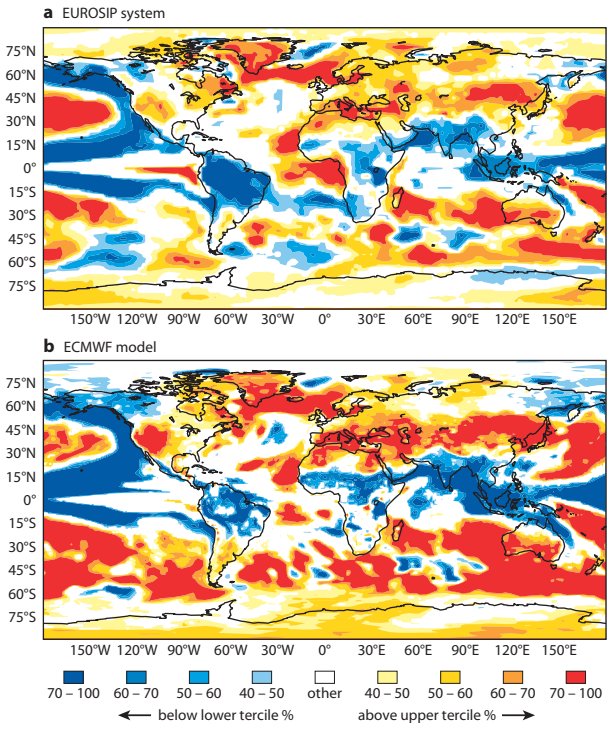


Figure 5 Forecasts for seasonal mean 2-metre temperature tercile categories for June/July/August 2008 from (a) the EUROSIP system and (b) the ECMWF model issued in May 2008. The forecasts are generally consistent, but EUROSIP tends to shift some of the higher probabilities (e.g. 70–100%) downwards towards lower values.

doi:10.21957/7wc0nybvir

Outlook for EUROSIP

The EUROSIP multi-model forecast system will continue to be maintained, and will benefit from each new forecast version that the contributors provide. Any increase in the number of models will also be beneficial.

There is also much scope to improve the accuracy, robustness and optimality of the combination methods used, and to consider the most effective ways of representing graphically the estimated signals and their uncertainties. Indeed, the proper calibration of probabilistic forecasts to account for model error is an issue for both single model and multi-model products. Collaboration with our Member States and others will be crucial in this area.

Finally, multi-model combination is not the ultimate tool for improving seasonal forecasts - there is no substitute for improving the individual forecasting systems themselves. Better models, run at the appropriate resolutions, will enable the impact of SST anomalies on the atmospheric circulation to be more accurately captured. Also more careful inclusion of other time-varying processes in the climate system (e.g. soil moisture, sea-ice, stratospheric dynamics, ozone, tropospheric and stratospheric aerosols) may lead to additional sources of non-negligible seasonal predictability. They may also give a better representation of the decade to decade changes in the Earth's climate that form an important part of the practical seasonal prediction problem. A multi-model combination will remain a valuable tool for many years to come, but it is only a complement to much other work that is needed.

Further Reading

Anderson, D., T. Stockdale, M. Balmaseda, L. Ferranti, F. Vitart, F. Molteni, F. Doblas-Reyes, K. Mogensen & **A. Vidard,** 2006: Seasonal Forecast System 3. *ECMWF Newsletter No. 110,* 19–25.

Molteni, F., L. Ferranti, M. Balmaseda, T. Stockdale & **F. Vitart,** 2007: New web products for the ECMWF Seasonal Forecast system 3. *ECMWF Newsletter No. 111,* 28–33.

Palmer, T.N., F.J. Doblas-Reyes & **R. Hagedorn,** 2003: DEMETER: Development of a European multi-model ensemble system for seasonal to interannual prediction. *ECMWF Newsletter No. 99,* 8–17.

Vitart, F., T. Stockdale & **L. Ferranti,** 2007: Seasonal forecasting of tropical storm frequency. *ECMWF Newsletter No. 112,* 16–22.

 doi:10.21957/7wc0nybvir