Medium-range Ensemble Forecasts at the Met Office

Christine Johnson, Richard Swinbank, Helen Titley and Simon Thompson

Met Office, Exeter, U.K.
christine.johnson@metoffice.gov.uk

ABSTRACT

The Met Office has recently implemented a regular 24-member 15-day global ensemble forecast. This is an extended-range version of the Met Office Global and Regional Ensemble Prediction System (MOGREPS), originally developed for short-range forecasts over the UK and Europe. It is now being run as a member state "time critical" suite on the ECMWF supercomputer. The medium-range ensemble forecasts are being stored in the TIGGE archive, where they are available to the THORPEX research community.

The Met Office ensemble forecasts, together with ECMWF and NCEP ensemble forecasts, are currently being used to develop techniques for multi-model ensemble forecasting. A key application is the use of both single- and multi-model ensemble forecasts to improve the prediction of high-impact weather. We show examples of the use of ensemble products that are tailored to the forecasting of severe weather, such as feature-based diagnostics. Finally, we outline our plans for future research based on the medium-range ensembles.

1. Introduction

The key aim of THORPEX (WMO, 2005) is accelerating improvements to forecasts of high-impact weather for the 1-day to 14-day range. As a major part of our contribution to THORPEX, the Met Office has recently implemented regular medium-range global ensemble forecasts. The Met Office medium-range ensembles contribute to the THORPEX Interactive Grand Global Ensemble (TIGGE), which is an international framework for the collaboration of research in ensemble forecasting, a particular aspect of which is the development of multi-model ensembles.

The Met Office is currently using the medium-range ensemble, together with forecasts from TIGGE partners (ECMWF and NCEP) to develop techniques for multi-model ensemble forecasting. Coupled with this work, we are developing a range of diagnostic products designed to alert forecasters and other users when high-impact weather events are forecast.

In section 2 of this paper, we give an overview of the MOGREPS ensemble forecast system, and how it has been extended to 15 days, running as a time critical suite at ECMWF. In section 3, we describe the development and verification of the multi-model ensemble. In section 4, we give some examples of diagnostics that we have developed in conjunction with operational forecasters, to highlight when high-impact weather is predicted. In section 5, we give a brief outline of our plans for future research and development based on the medium-range ensembles, with a view to the THORPEX aim of developing a Global Interactive Forecast System.

2. Medium-range Ensemble Forecast System

The medium-range ensemble forecast system is based on the 3-day Met Office Global and Regional Ensemble Prediction System (MOGREPS). MOGREPS comprises a regional model (covering the North Atlantic and Europe) with boundary conditions taken from global ensemble forecasts. The 24-member ensemble forecasts are run twice daily. Perturbations to the initial conditions are generated using an Ensemble Transform Kalman Filter (ETKF) method (Bowler, 2006).

For the medium-range ensemble forecasts, the global component of MOGREPS has been extended to 15 days and is run on the EMCWF supercomputer. The global Unified Model is run at a horizontal resolution of 0.83° latitude by 1.25° longitude (N144, nominally 90km), with 38 levels. For comparison, the current operational deterministic global model resolution is N320 (~40km) and 50 levels in the vertical. Initial conditions for the ensemble forecasts are generated by the MOGREPS system at the Met Office, and transferred to ECMWF (see Figure 2.1).

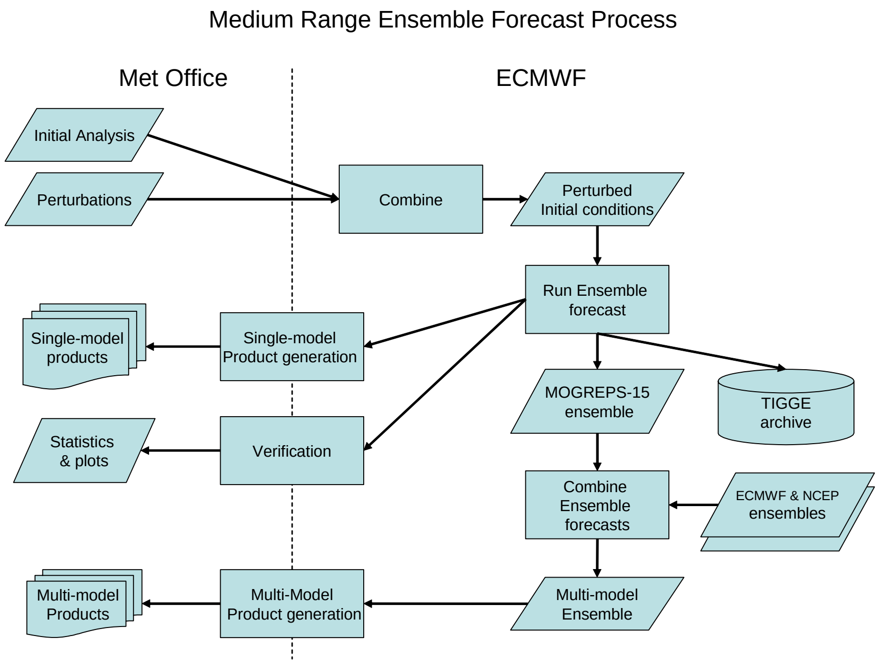

Figure 2.1: Schematic of the Met Office medium-range ensemble forecast system. At present experimental products and verification statistics are being produced from both the single-model forecasts and the multi-model forecasts.

The suite has been running since March 2006 and the forecasts have been archived in TIGGE since October 2006. The forecasts take 1¼ hours, when running each ensemble member on 4 nodes (64 processors), and running 8 members at once. The archiving of data also takes 1 ¼ hours and further processing of the diagnostic output fields takes a further hour so that for the 00 UTC run starting at 05 UTC, the products are available to Met Office forecasters by 8.30UTC, and similarly for the 12 UTC forecasts. As of the end of July 2007, the suite has been dedicated 'time-critical' under the 'Framework for Member State time-critical activities'. The suite is also split into two parts with different priorities, so that in the event of a delay, the forecasts for the next run can start even if archiving or other lower priority tasks are still in progress.

When originally implemented in late March 2006, the physical parameterizations were similar to those used by the global deterministic model in early 2005, with a limited implementation of stochastic physics. Following upgrades to the deterministic system in Spring 2006, a new package of physical parameterizations (Savage et al, 2007) was implemented, in conjunction with compatible stochastic physics schemes, early in 2007. The new stochastic physics includes a new version of the random parameters scheme and a scheme known as the Stochastic Kinetic Energy Backscatter (SKEB), resulting in significantly more spread in the extra-tropics. Further physics upgrades, including updating the SSTs daily and bringing the model physics into line with the global MOGREPS model, will be introduced soon.

The forecast model produces the TIGGE output fields directly in GRIB format, and these are converted to GRIB2 and transferred to the TIGGE archive via the ECMWF MARS archive system (see Fig. 1.1). Other output streams are used for verification and to generate products - particularly aimed at forecasting high-impact weather, as described in section 4.

3. Multi-model Ensembles

As shown in Fig. 2.1, we are currently combining the Met Office medium-range forecasts with forecasts from other TIGGE partners. Johnson (2006) has used an ensemble test-bed based on simple models to help guide the approach that is used to calibrate and combine the ensemble forecasts. Based on that work, we are currently producing an experimental bias-corrected multi-model ensemble that combines three ensembles: 24 member Met Office ensemble, 51 member ECMWF ensemble and 21-member NCEP ensemble. We are currently post-processing three variables (mslp, 2m temperature and 500hPa height), but this may be extended to more variables in the future.

3.1. Calibration and combination procedure

The multi-model ensemble is simply formed by the union of the single-model ensembles, so that the multi-model pdf is an average of the pdfs from the single models:

$$p(y) = \sum_{k=1}^{N} p(y \mid M_k) p(M_k) \qquad 3.1$$

where $p(y \mid M_k)$ is the probability based on model M_k and $p(M_k)$ is the probability of M_k being the best model.

The calibration and combination procedure is split into three steps: bias correction of the single-model means, model-dependent weighting and variance adjustment of the multi-model ensemble. This ensures that the bias correction can be applied to the single-models separately. To estimate the calibration and combination parameters, we use a running-mean based on past data. For example the running mean update equation for the estimated bias is given by

$$b_s^n = (1-\mu) b_s^{n-1} + \mu e_s^{n-s} \qquad 3.2$$

where b_s^n is the estimated bias for the ensemble mean starting at time t_n with a lead time of t_s, e_s^{n-s} is the ensemble mean forecast error for the forecast starting at time t_{n-s} with a lead time of t_s and $\mu = 0.01$ is the update parameter that controls the amount of temporal smoothing.

This procedure allows for the fact that there is a very limited set of past data and also allows for seasonal and weather-dependent variations. To allow for the fact that forecasts tend to verify better with their own analysis we use a multi-model analysis as the verification data. This multi-model analysis is a simple average of the analyses from the three component models.

Previous idealized studies with the Lorenz model (Johnson, 2006) had compared different methods for calculating weights and showed that different methods for calculating weights give surprisingly similar results. Therefore, the simplest method has been implemented here: the time-averaged mean-square-error is used to calculate the model-dependent weights

$$p(M_k) = \gamma / MSE_k \qquad 3.3$$

where γ is a normalization factor so that $\sum_k p(M_k) = 1$.

The mean-square-error of the bias-corrected ensemble mean is updated in a similar way to the estimated bias:

$$MSE_s^n = (1-\mu)MSE_s^{n-1} + \mu(\varepsilon_s^{n-s})^2 \qquad 3.4$$

where MSE_s^n is the time-average mean-square-error starting at time t_n with a lead time of t_s, ε_s^{n-s} is the forecast error for the bias corrected ensemble mean starting at time t_{n-s} with a lead time of t_s and $\mu = 0.01$ is the update parameter.

A similar running-mean process will be used to calibrate the multi-model ensemble variance.

3.2. Products

The multi-model ensemble software is run in real time using forecasts starting at 0Z and with forecast ranges from 12 to 360h. The forecasts are used to create ensemble product charts, primarily focussed on the European area. As the multi-model ensemble is formed from the three single-model ensembles we choose to show both the single model products and the multi-model products together. This aids the user to understand the forecast uncertainty due to inconsistency between models. The multi-model probability is simply an average of the three component model probabilities.

An example of the probabilities obtained from the multi-model ensemble is shown in Fig. 3.1 In this example, the weight given to each model is identical, as illustrated by the pie chart. However, we have also developed the capability to compute and use model-dependent weights. In this case there is a large difference between the probabilities obtained from the three component models, with the NCEP ensemble giving large probabilities of high pressure over the UK, and the ECMWF and Met Office ensembles giving smaller probabilities. It is not possible to determine which model gives the best probability forecast at this particular time. The multi-model ensemble gives an average of these probabilities, which should be a better estimate because it accounts for the errors in both the initial conditions and the forecast models.

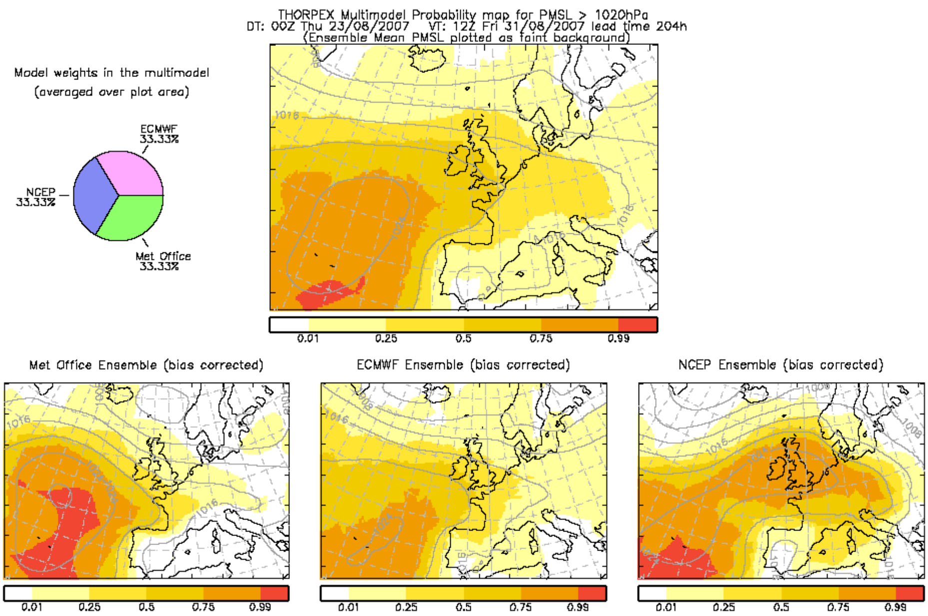

Figure 3.1 Probability map for the multi-model ensemble (top) and the three component model ensembles (bottom three).

Other products include the multi-model mean and spread (not shown). However instead of plotting the component model spreads, we plot the spread of the multi-model ensemble around the single model mean, as this gives a better estimate of the uncertainty associated with the single model means. These plots can also be used to easily identify when one model is behaving in a very different way to the other component models.

3.3. Verification

To verify whether the multi-model ensemble does give an improved estimate we verify the data using multi-model analyses. Fig 3.2a shows RMS error of the ensemble mean, as a function of lead time, for the raw Met Office ensemble and for the bias corrected ensemble. We find that the bias correction gives improvements at all lead times, but with the largest impact at short lead times. There is a relatively large reduction at short lead times – this is because a multi-model analysis is being used as the verification data. The RMS errors for the other variables (500mb height and mslp, not shown) do not show such a significant impact.

The RMS errors of the multi-model ensemble mean for 2m temperature, compared with those of the bias corrected single model ensembles (Fig 3.2b) show that the model combination does give a further improvement, as the multi-model ensemble mean has lower RMS errors than of any of the component model ensembles. One might expect that because we can identify the ECMWF model as being the best model in the RMS error sense that the combination of all three models would give worse performance. However, because the models have different kinds of errors, these errors cancel out in the multi-model combination, giving a reduction in the overall error. Again, the improvements are less significant for the other variables.

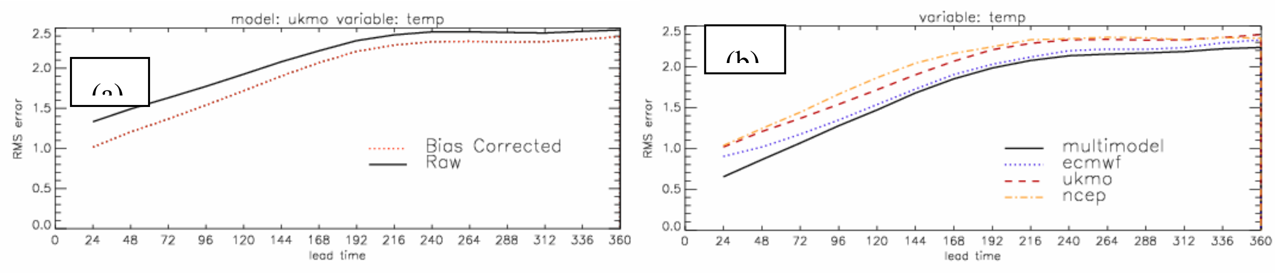

Figure 3.2: RMS errors of the ensemble means globally averaged over 40 days, for the (a) bias corrected and raw Met Office ensemble, and (b) bias corrected component models and unweighted multi-model.

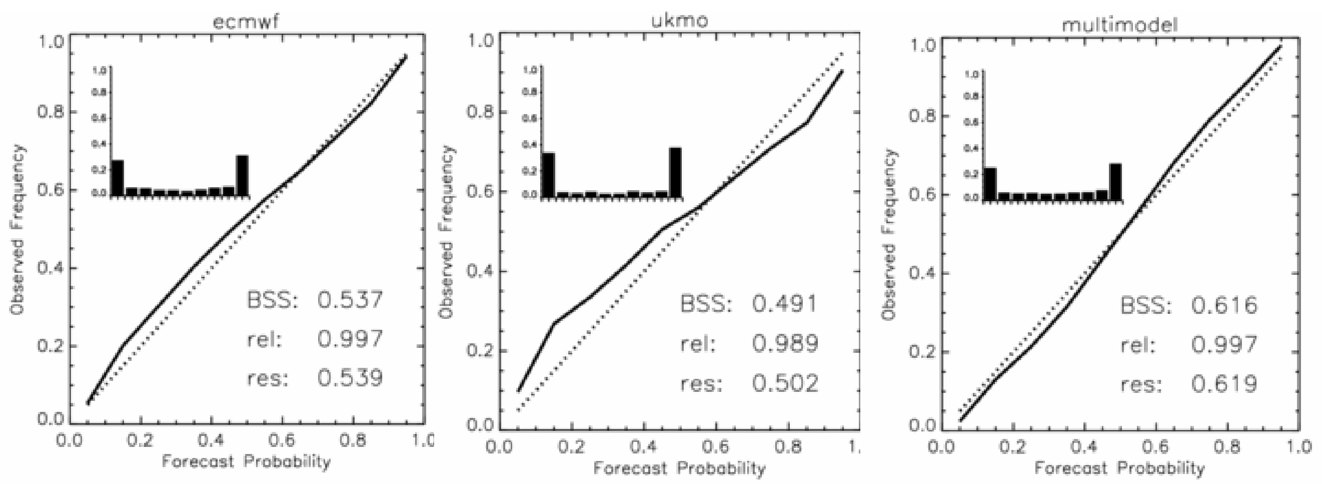

Figure 3.3 Reliability diagrams for the probability of 2m temperature greater than the climatological mean, at a lead time of 72 hours, globally averaged over 15 days. The bar chart shows the corresponding sharpness diagram.

To evaluate the probabilistic ensemble products, reliability tables have been derived using climatological thresholds. An example of the resulting reliability diagrams for two of the component models and for the multimodel ensemble (all bias corrected versions) is shown in Fig 3.3. The reliability diagram is a graph of the observed frequency against the forecast probability. That is, the probability that the event is observed given the forecast probability, p(o|q), as a function of the forecast probability, q. For a perfectly reliable ensemble, the observed frequency would match the forecast probability, as indicated by the dashed line. Also shown is the corresponding sharpness diagram, which shows the frequency of the forecast probabilities, p(q) against q. A system with perfect sharpness would give high frequencies only in the q=0 and q=1 categories. The plots show reliability diagrams for temperature greater than the climatological mean, derived from ERA-40 data (courtesy Martin Leutbecher, ECMWF). By using the climatological mean we can average globally without introducing artifical skill that might occur if using a fixed threshold.

All three ensembles depart from the perfect diagonal line. In particular, the forecast probabilities are too low (high) when the event is frequent (infrequent), meaning that the systems have poor resolution (the ability of the forecast system to discriminate between different events). For the multi-model (92 member) ensemble, the reverse is true meaning that the system has good resolution. This improvement could be due to the benefit from combining different models, or from the increase in ensemble members. This is summarised by the Brier skill scores (BSS) which show that the Met Office and NCEP models have the lowest BSS and the ECMWF the largest, but the multi-model BSS exceeds all of these with 0.62. In particular, we see that the reliability component is identical to that of the most skilful model, and the resolution component is larger than all three.

4. Forecasting High-impact Weather

Ensemble forecasts produce a vast amount of data which is hard for forecasters and end-users to interpret. In order to facilitate this process, a suite of intelligent diagnostics needs to be developed, to highlight when high-impact weather is forecast.

4.1. Parameter-based diagnostics

A range of products that were originally developed for displaying output from the short-range MOGREPS ensemble have been extended to medium-range forecasts, giving the forecasters the benefit of a seamless range of products across the short and medium range. These include *probability maps* (probabilities of forecasts exceeding given thresholds), *mean and spread* charts, *postage stamps*, and *spaghetti* charts (all available as animations). In addition, *heatwave maps* have been developed to display the probabilities of successive day-time and night-time temperatures exceeding the thresholds defined in the England and Wales Heat-Health Watch. To produce an *overview of expected warm/cold/wet/dry spells* in the 15-day forecast, probability charts are produced summarising the weather over 5-day periods. Finally, using an ERA-40 climatology dataset developed by Martin Leutbecher at ECMWF, products have been developed to highlight *forecast anomalies relative to climatology*. This is especially useful for forecasting high-impact weather events that fall in the tails of the climatological distribution.

4.2. Feature based-diagnostics

High-impact weather is often associated with synoptic features, and although the high-impact weather (severe gales, torrential rain etc) often occurs on scales below those that the ensemble members can resolve, they can usually represent the causal features. This necessitates the use of a feature-based approach, whereby feature development, density and attributes can be interpreted in the context of potential high impact events.

The Met Office currently objectively identifies two types of features: extra tropical cyclones and tropical cyclones.

(a) Extra-tropical cyclones

At the Met Office, the ensemble forecast output is automatically analysed to identify extra-tropical cyclones (at various stages in the life-cycle) and fronts, with feature-point attributes stored in a cyclone database (Watkin and Hewson, 2006). This information is processed with tailored tracking software, allowing the creation of products showing the potential for high-impact weather. The tracking system uses a combination of forward and backward tracking. It uses extrapolation and the 500hPa steering wind to estimate positions, and matches features based on separation distance, type and thickness. Forecasters can click on particular cyclones in an *interactive map* of the control analysis and see the range of tracks forecast by the ensemble, along with *feature-specific plumes* of intensity measures to show how that cyclone may develop.

These diagnostics currently run in real-time on the MOGREPS-15 forecast, and for case studies using the ECMWF EPS. They are illustrated with use of a case study: that of the New Year's Eve 2006 gales over N. Ireland, Scotland and northern England. This extra-tropical cyclone (low 969 hPa) was of a particularly high-impact for its intensity because it fell on New Year's Eve, and led to the cancellation of several high-profile events. The performance of both ensembles was good in week one. At short ranges the spread of tracks was relatively low, with the true solution still falling well within the spread (Fig 4.1). Combined together the ensembles provide strong evidence for the rapid intensification of this feature, and would have provided good supporting evidence to forecasters on the location and intensity of the gales during New Year's Eve.

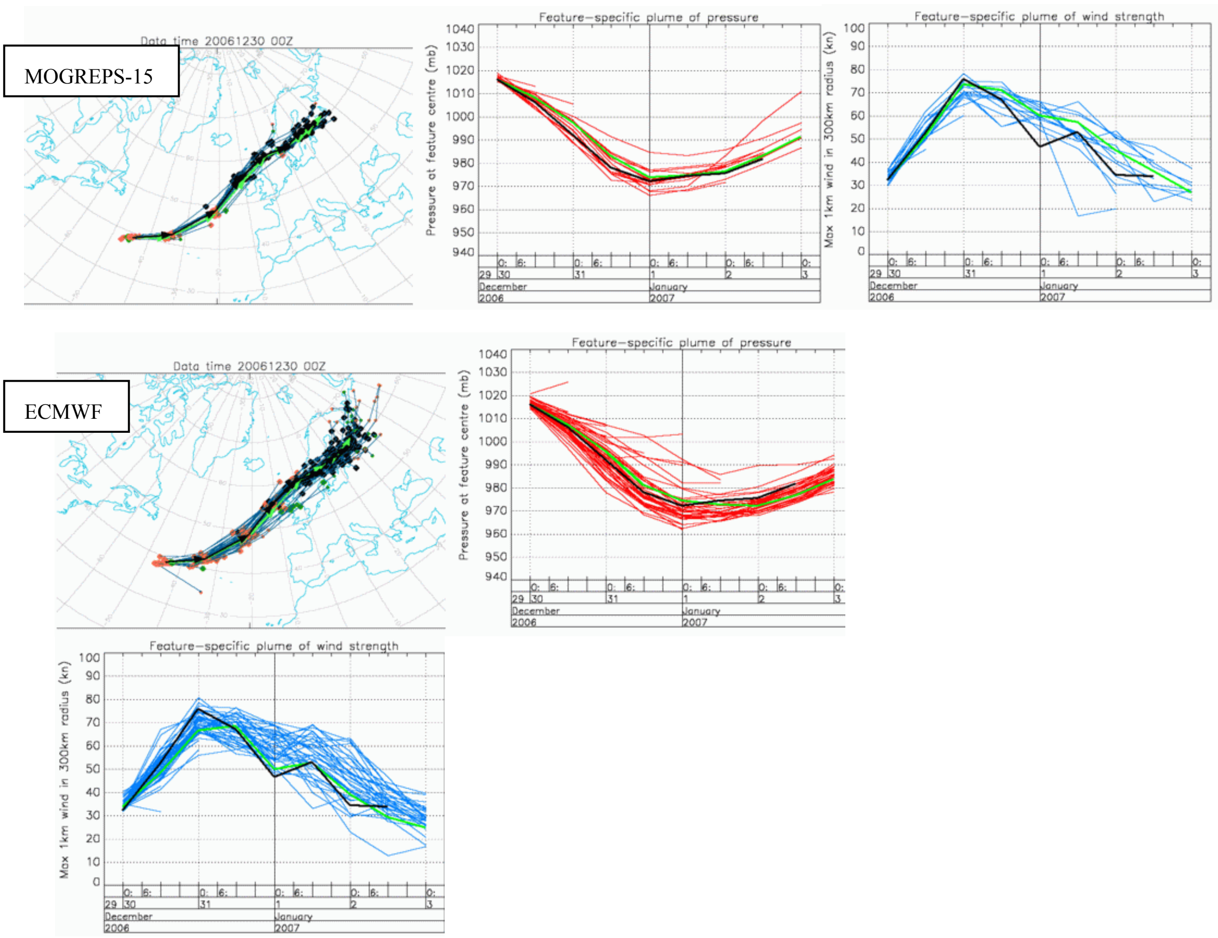

Figure 4.1 Left: Cyclone database feature-specific tracks for key frontal wave from 00Z 30/12/2006. Middle/right: Corresponding plumes of pressure/1km wind strength (control=green, analysis=black)

At longer lead times, the uncertainty over the timing and the track of the low increases, and the traditional probability charts of high winds in a particular location at a particular time start to show this apparent loss of signal in the ensemble. However, although the detail of the situation becomes more uncertain at longer lead times, ensemble forecasts can still provide useful guidance of the broader risk of a storm affecting some part of the UK within a set period. The *strike probability plots* produced by the cyclone database are one diagnostic that can be used to assess this risk, as they show the number of members predicting a feature with the potential to cause damaging wind gusts to track within 300km during a 24-hour period. The strike probability plots from the MOGREPS-15 ensemble (Fig 4.2) show that at a six day lead time (T+144) between 50-60% of members are still indicating a risk of a strong storm over the UK.

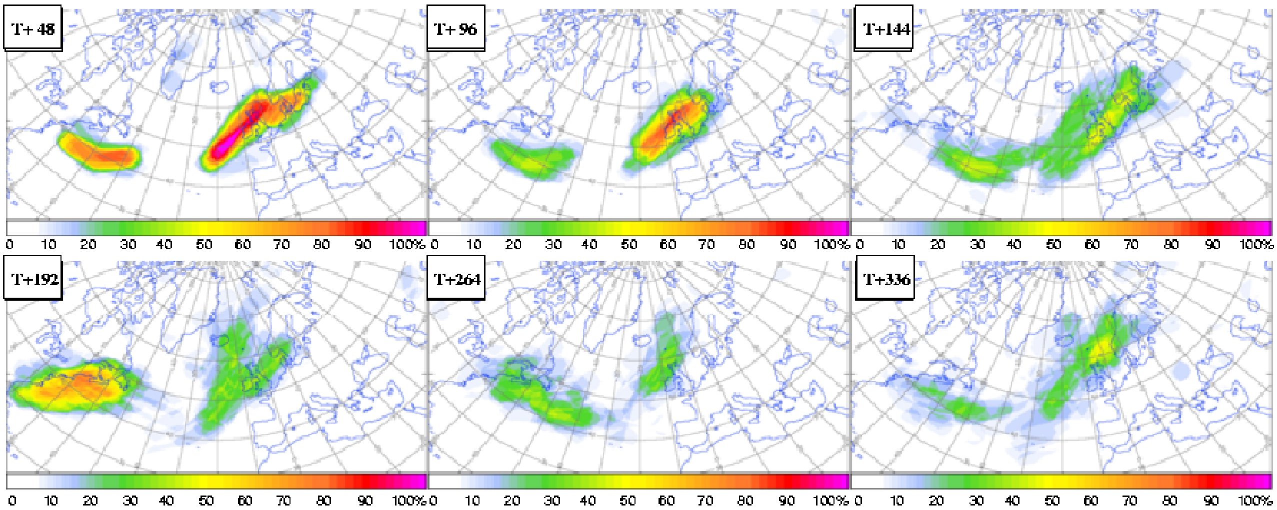

Figure 4.2: Strike probability plots showing probability of a cyclonic feature with 1km winds > 60 knots tracking within 300km in a 24-hour period centred on VT 12Z 31/12/2006 from MOGREPS-15 ensemble

(b) Tropical cyclones

Tropical cyclones are identified and tracked in the ensemble forecast data using 850hPa relative vorticity maxima. Plots for each basin show the tracks of both existing storms, and those predicted to develop in the first week of the forecast. For named storms, forecast tracks from the ensemble are displayed alongside strike probability plots, and the tracks from the deterministic model, the control and the ensemble mean.

An example tropical cyclone forecast from the MOGREPS-15 ensemble is given in Figure 4.3 for Tropical Cyclone George from March 2007, which made landfall near Port Headland with winds up to 195km/hr, resulting in 3 deaths. Fig 4.3a shows the tracks of all the ensemble members, coloured according to the lead time, and Fig 4.3b shows the corresponding strike probabilities with the associated analysed track overlain, showing that the abrupt left turn to hit Australia was captured by the ensemble at this lead time. Fig. 4.3c shows the key tracks, including the deterministic (green) which interestingly did not forecast the abrupt left turn, the control (blue) and mean (purple).

Verification results comparing the average track errors from the Met Office deterministic model and the mean of the MOGREPS-15 tracks are shown in Table 4.1. The errors are similar until T+72, but the ensemble mean has reduced the track error by 12% at T+96 and 23% at T+120 with respect to the deterministic model.

	T+0	T+24	T+48	T+72	T+96	T+120
Deterministic track	58	120	222	332	509	605
Mean of ensemble tracks	39	125	233	334	450	467

Table 4.1: Average tropical cyclone track errors in km (from 7 Feb to 11 Sep)

4.3. Regime-based diagnostics

In order to highlight forecast changes of weather type, the large-scale circulation patterns forecast by MOGREPS-15 over Europe are objectively characterized using a method based on Grosswetterlagen (GWL) regime typing (James, 2007). Web-based displays show which category is being forecast by each ensemble member, while regime-based clustering provides a summary of the forecasts. The GWL products are now available using the Met Office, ECMWF and NCEP ensembles, and are also produced from the monthly ensemble forecasts, improving consistency in forecast products across models and timescales.

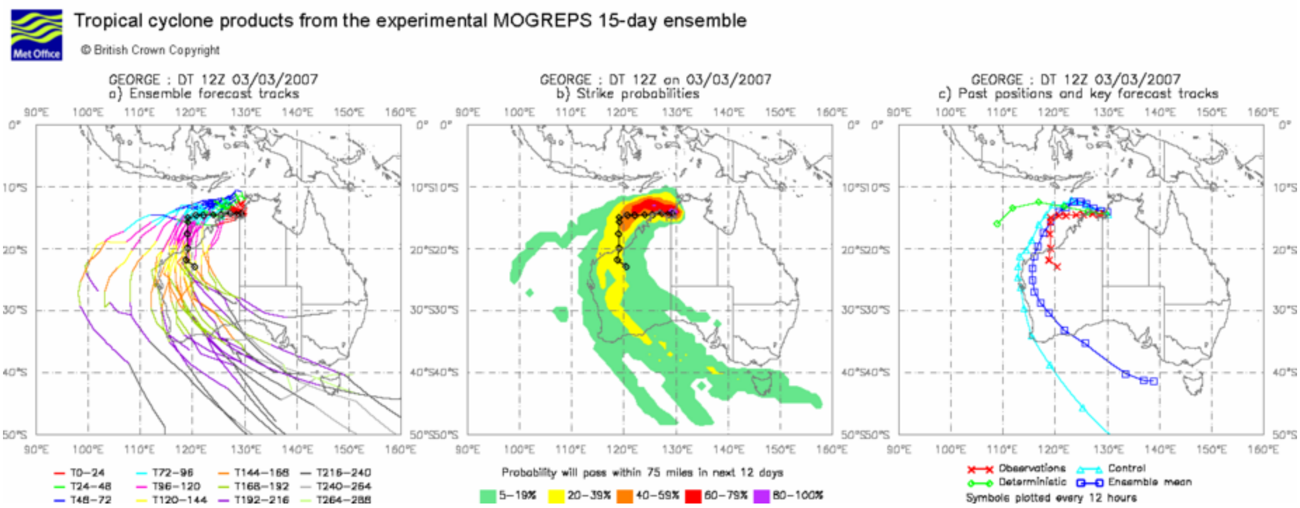

Figure 4.3: Example tropical cyclone ensemble product from MOGREPS-15, for Tropical Cyclone George

5. Summary and Future Plans

This paper gives an overview of work on medium-range ensemble forecasts that has been carried out at the Met Office over the past two and a half years as part of our contribution to THORPEX. We have developed a medium-range ensemble forecast system that contributes to the TIGGE project. It forms the basis of the development of techniques to produce multi-model ensembles combining Met Office forecasts with those from ECMWF and NCEP. Results from both the MOGREPS-15 ensemble and the multi-model ensemble are being used as the basis for new products to forecast high-impact weather events and provide additional non-operational guidance to Met Office forecasters.

In the next few years, we intend to develop the MOGREPS system to take advantage of increases in supercomputer capacity. As a first step, we expect to move from a 38-level configuration to a 70-level configuration (spanning the stratosphere) in about two years' time. Using the same set of levels as the deterministic forecast model will facilitate the use of ensemble techniques to estimate forecast error covariances for the variational data assimilation system, and the additional levels should also improve the medium-range forecasting skill. We also plan to improve the horizontal grid-length from the current 90 km to 60 km. We are also considering options to run a set of reforecasts which should allow improved calibration and downscaling of the ensemble forecasts, and/or to use a simple coupled ocean model which should also improve the skill of the forecast model. Introduction of these changes would entail the use of significantly more computer resources, so it is likely that we will need to change to running the medium-range ensembles once, rather than twice, a day.

The ultimate aim of the TIGGE project is the development of a Global Interactive Forecast System (GIFS), using the fruits of the THORPEX research programme to develop innovative approaches to weather forecasting. The precise form of the GIFS is as yet unclear, but we anticipate that it would entail an interactive approach to weather forecasting, with additional forecasts and diagnostics being produced on demand in response to expected high-impact weather events. Predictions of the likelihood of high-impact events will be based on ensemble techniques, such as those described in this paper. TIGGE entails the exchange of forecast data, while GIFS will also entail the exchange of products focused on high-impact weather.

While this paper summarises work on medium-range ensemble forecasts at the Met Office, it also gives a taste of the approaches that could be included in a future Global Interactive Forecasts System that it is hoped to develop over the next 5-10 years.

5.1. Acknowledgements

We would like to thank all the Met Office staff who contributed to this work, in particular the Ensemble Forecast Research team, Nick Savage, Tim Hewson, Paul James and Julian Heming. We also thank Martin Leutbecher, Paul Dando and all those at ECMWF who have helped us set up the MOGREPS-15 ensemble forecast system.

6. References

Bowler, N., A Arribas, K Mylne and K Robertson, 2007: The MOGREPS short-range ensemble prediction system Part I: System description. *Met R&D Technical Report 497*, Met Office.

James, P.M., 2007: An objective classification method for Hess and Brezowsky Grosswetterlagen over Europe, *Theor. App. Climatol*., **88**, (1-2): 17-42.

Johnson, C., 2006: On the benefits of multi-model ensembles: idealized studies using the Lorenz 1963 model. *Met Office Forecasting Research Technical Report*, FRTR 492.

Savage, N.H., S Milton, D Walters and J Heming, 2007: An assessment of the impact of new physical parameterisations on the performance of the Global Unified Model in THORPEX 15-day forecasts, *Met Office Met R&D Technical Report* No. 496.

Watkin, H. A. and T. D. Hewson, 2006: The development of feature-based diagnostics to assess TIGGE handling of high-impact extra-tropical cyclones. *Second THORPEX International Science Symposium, Landshut, Germany (4-8 December 2006).* WMO/TD No. 1355 WWRP/THORPEX No. 7, 258-9.

WMO, 2005: THORPEX international research Implementation Plan, WMO/TD-No.1258 WWRP/THORPEX No.4, WMO.