JP1.4 SPATIAL AND LEAD TIME ACCURACIES OF THUNDERSTORM FORECASTS FOR AIR TRAFFIC CONTROL

Helen A. Watkin*, Tania R. Scott, and Deborah J. Hoad
Met Office, Bracknell, UK

1. INTRODUCTION

For reasons of safety and passenger comfort aircraft often fly around thunderstorms, and thus air traffic management (ATM) need advance notice of thunderstorm activity to enable them to adjust air traffic flow before major disruption occurs. This project assesses the Met Office's ability to forecast air traffic disruption due to thunderstorm activity.

A previous Met Office project investigating short range thunderstorm forecasts from the GANDOLF (Generating Nowcasts for the Deployment of Operational Land-based Flood forecasts) system, showed that GANDOLF forecasts could provide useful information up to about 90 minutes ahead (Hoad, 2000). Feedback on the project from the UK ATM authority suggested they would like more information on the quality of forecasts at a number of different spatial and temporal scales in order to determine the value and utility of the forecasts in the context of their own operations. Hence, this follow-up project aims to quantify the trade-off between the spatial and lead time accuracies of Met Office disruption forecasts. The forecast lead times tested start with an analysis (T+0), then proceed hourly up to T+12. The spatial resolutions at which the disruption forecasts are prepared are 15km, 30km, 60km and 120km.

Air traffic disruption occurs when pilots perceive a threatening storm ahead and decide to deviate around it. Studies of pilots' storm penetration/deviation behaviour have shown storm intensity and coverage to be two key decision factors (Rhoda et al, 2000; Rhoda and Pawlak, 1999). Information on these factors is largely taken from aircraft radar reflectivities, which are displayed in the cockpit using three colours: green (corresponding to a rain rate of 1-4mm/hr), amber (4-12mm/hr) and red (>12mm/hr) (Seymour, 2002).

Using this display a pilot will look at the maximum intensity of an upcoming storm, both horizontally and vertically to determine if it is dangerous. UK pilots will not fly through amber or red and also try to avoid green where possible, but will fly through green if it is difficult to avoid the area (Rankin, 2001), confirming the importance of both intensity and coverage. In order to create accurate forecasts of the level of disruption this paper attempts to take account of pilot behaviour in its forecasting methods.

2. METHOD

2.1 Identifying thunderstorm events

Lightning data were used to identify thunderstorm events across the UK during summer 2001. Around 20 cases were chosen, but only 5 events were available for this preliminary analysis.

2.2 Replicating pilot radar displays

The first step taken was to replicate pilot radar displays using forecast and verification data. A 15km resolution was chosen to match the resolution of several of our forecast data fields and to correspond to the size of large thunderstorms which can be between 10 and 20km wide. The maximum convective rainrate at the ground was calculated from three types of forecasts, and the maximum observed convective rainrates obtained from radar data to be used for verification, using the following processing steps.

2.2.1 Forecast data

UK Mesoscale Model

The Mesoscale numerical weather prediction model is a limited-area model of the Met Office Unified Model. It covers the British Isles and surrounding areas with a spatial resolution of 12km, and is run every 6 hours. Mesoscale convective rainrate forecast data were converted to a 15km resolution.

Nimrod nowcasting system

The Nimrod nowcasting system integrates recent observations with numerical weather prediction model products to provide frequent forecasts over the UK up to 6 hours ahead (Golding, 1998). In this case a forecast of total rain rate at the ground is used because convective rainrate is not available. Dynamic rain is eliminated by masking with convective cloud cover from the Mesoscale model.

CDP

The Convection Diagnosis Project (CDP) provides forecast probability diagnoses of the intensity, distribution and duration of convective showers from Mesoscale model outputs (Hand, in submission). It runs every 6 hours providing hourly forecast products from T+6 hours. The forecast data used here are for peak rainfall rate for probabilities of shower occurrence of greater than 70%.

2.2.2 Applying thunderstorm reflectivity profile

The values of convective rainrates at the ground for each of the above three forecast types were converted to maximum convective rainrate in the vertical using the idealised reflectivity profile of a

* Corresponding author address: Helen A. Watkin, Met Office, London Road, Bracknell, Berkshire, RG12 2SZ, United Kingdom;
e-mail: helen.watkin@metoffice.com

mature thunderstorm cell. The profile was scaled logarithmically from the average figures found in UK thunderstorms of 6mm/hr at the ground corresponding to a 16mm/hr maximum in the cloud (Hand, 1996).

2.2.3 Verification data

Radar

The radar data used for the verification is taken from the nearest radar site to the event in question. Data from four beams are available, each at a different angle and hence reflecting a different height in the storm. The maximum reflectivity from these beams is taken and converted into mm/hr to give an estimate for the maximum rain rate in the cloud.

2.2.4 Classifying by radar colour

The maximum rain rate for the three forecasts and the radar observations were then classified into the colour they would appear on the aircraft radar. A total area of 240 by 240 km (16 by 16 grid cells) was analysed for each event, centred on the radar site closest to the storm activity. Figure 1 shows example replicated pilot radar displays over Eastern England.

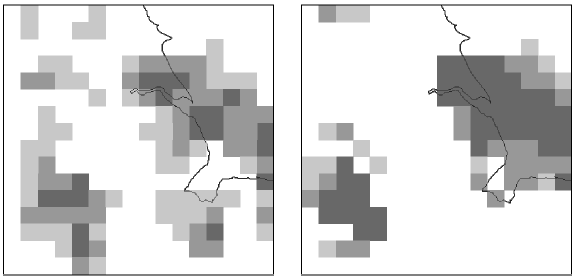

FIGURE 1: Replicated pilot display, 1900, 15/06/2001; using radar data (left), a Nimrod forecast (right)

2.3 Calculating air traffic disruption

Once estimates of the maximum rain rates were obtained, the data were analysed for air traffic disruption. This was done in the same way for Mesoscale, Nimrod, CDP or radar data. Two methods for calculating air traffic disruption were used, one only taking into account storm intensity, and the other including coverage.

2.3.1 Intensity method

Rainrate pixels were classed as deviations if their aircraft radar classification was amber or red. The level of air traffic disruption was categorised for the four spatial resolutions (covered by 1, 4, 16, and 64, 15km pixels) according to the percentages of deviation pixels in that region, as shown in Table 1.

TABLE 1: Classifying air traffic disruption (1st method)

% of deviation pixels (D)	Level of disruption
D=0	None
0<D<25	Minor
25≤D<37.5	Major
D≥37.5	Severe

2.3.2 Intensity and coverage method

The type of convective activity in the 75 by 75 km area surrounding each pixel was classified as shown in the headings of Table 2. The time taken to fly across an area of this size is approximately 5 minutes, assuming a cruise speed of about 250 m/s, allowing a reasonable timescale for decision making. Table 2 shows how each cell was classified as leading to “no deviation (N)”, “small deviation (S)”, or “large deviation (L)” from the expected flight path, depending on its combination of radar colour and coverage type.

TABLE 2: Estimating grid cell deviation characteristics

	Type of convective activity in nearby area		
Radar return	Not widespread <20% non-zero pixels	Widespread weak >20% non-zero, <20% amber-red	Widespread strong >20% non-zero, >20% amber-red
Black	N	N	N
Green	S	N	L
Amber -Red	S	S	L

Table 3 shows how the level of air traffic disruption in each airspace was classified into the same four categories as in the simpler method, but this time using percentages of the deviation types classified in Table 2.

TABLE 3: Classifying air traffic disruption (2nd method)

%Large (L)	Operator	%Total (T) NB. T=S+L	Disruption
L=0	AND	T=0	None
L≤12.5	AND	T<25	Minor
L<25	AND	T<37.5	Major
L≥25	OR	T≥37.5	Severe

Contingency tables were created from the results, for the different forecasts, methods, lead times and spatial resolutions.

3 ANALYSIS OF THE RESULTS

To analyse the results several statistical forecast quality measures were calculated from the contingency tables using the methods and formulae set out in Wilks (1995). Four measures of accuracy were calculated: Hit Rate (fraction of forecasts correct); Critical Success Index (CSI) (hit rate once ‘disruption type not forecast and not observed’ values are removed); Probability of Detection (POD) (fraction of times label was forecast when it had been observed); and False Alarm Rate (FAR). For the first three of these perfect forecasts give a value of one and completely incorrect forecasts give zero. These numbers are reversed for the FAR. The Bias ratio was also calculated to determine if a particular disruption category was overforecast (Bias>1) or underforecast (Bias<1). Finally skill was calculated using the Kuipers Skill Score (KSS), which compares the accuracy of the forecasts with that of unbiased random reference forecasts. The KSS is equal to one for a perfect forecast, zero for random or constant forecasts, and negative for forecasts that are inferior to random forecasts. The Hit Rate and the KSS give one value for each 4 by 4 contingency table whereas the other measures give a separate value for each of the disruption classes.

When space allows the key results quoted below have corresponding figures, but this is not always possible. The results will be published fully in a later report.

3.1 Comparison of the two disruption classification methods

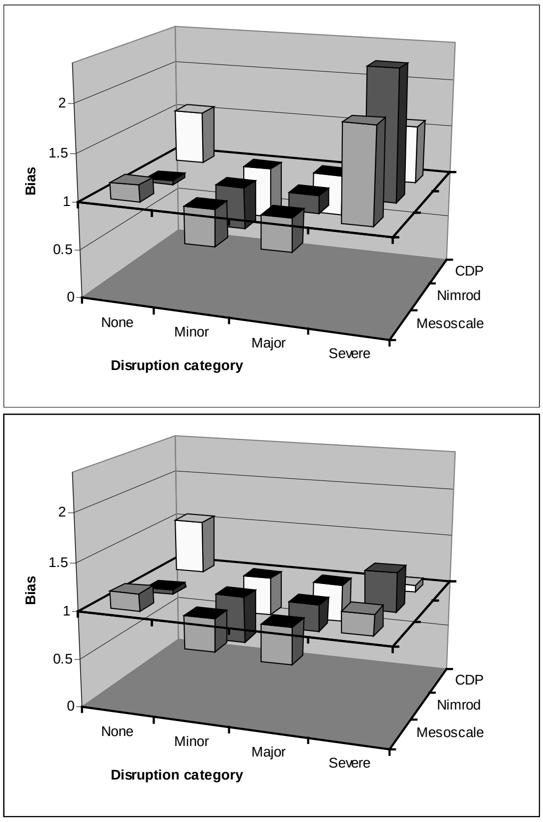

FIGURE 2: Bias values (averaged over lead time and resolution) for each forecast type and disruption category for the intensity-only method (top) and the intensity and coverage method (bottom)

The two disruption classification methods (intensity only vs intensity and coverage) perform similarly in terms of Hit Rate and KSS, but Figure 2 shows that the method which incorporates both intensity and coverage has a smaller degree of bias, with all categories lying within ±0.6 of the unbiased value of 1. The 'Severe' category is still overforecast, but to a more acceptable level. In addition the CSI scores for the 'Severe' category, particularly important in terms of ATM impact, are raised by an average of around 0.2. Finally, in the Nimrod forecasts the drop off of the POD scores for the 'Severe' class with increased lead time is slower in the more complex method, confirming that the forecasts produced using both intensity threshold and coverage levels are of the greater quality. The remaining results will therefore be for the more complex method only.

3.2 Comparison of the different forecast types

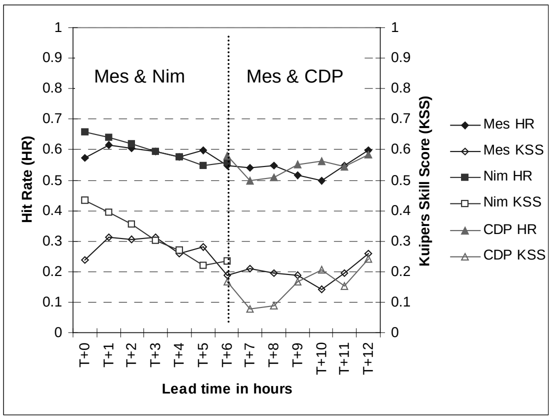

FIGURE 3: Hit Rate and Kuipers' Skill Score (KSS) for the three different types of forecast (averaged over spatial resolution)

The differences in forecast quality across forecast type vary depending on the measure considered. Figure 3 shows that the average Hit Rate for each lead time is similar for the three types of data, varying between 0.7 and 0.5. However the Kuipers Skill Scores do show significant differences with the Nimrod forecasts performing significantly better than the Mesoscale data up to T+4, and the Mesoscale and CDP data both only achieving values of 0 to 0.2 after T+6. The Nimrod forecasts also yield higher CSI and POD scores than the other forecasts. The False Alarm Rate is also lower for the Nimrod forecasts. After T+6 the degree of scatter in both the CDP and Mesoscale forecasts (due to the less frequent runs of these forecast types) makes it difficult to compare but the Mesoscale CSI results do appear consistently above those of the CDP disruption forecasts.

Thus it appears that between T+0 to T+6 the disruptions forecasts produced from the Nimrod data are of the highest quality hence we will now look at the trade-off between spatial and lead time accuracies for these.

3.3 Spatial and lead time accuracies

Figure 4 shows that for the 'Severe' category forecasts, the POD declines steadily with lead time before plateauing around T+5 (the CSI shows similar results but at a lower level), while the FAR rises. The graphs also indicate that accuracy decreases as spatial resolution increases. The KSS (the best overall measure of forecast skill), shown in Figure 3, remains around 0.2 for all resolutions, even at T+6, indicating the forecasts are still significantly better than random and may still be of some use in ATC management.

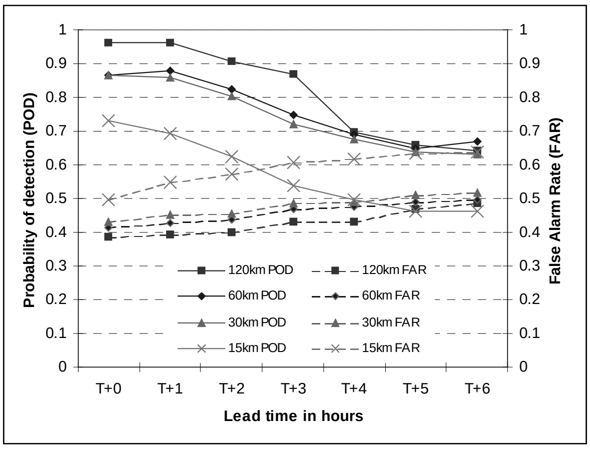

FIGURE 4: Comparing accuracy over lead time and resolution for ‘Severe’ air traffic disruption forecasts using Nimrod data: POD (top) and FAR (bottom).

4 DISCUSSION

4.1 Trends in forecast quality

The drop off in the quality of the forecasts as lead time increases reflects the difficulty of forecasting small scale features such as thunderstorms over a long lead time. The reduction in accuracy with increased spatial resolution is also to be expected, as over smaller areas there is more scope for small inaccuracies to have a large impact.

The forecasts are generally good at picking up the location of the storms, but often overestimate their intensity, which may cause several results to show a higher quality for “None” and “Severe” than for the intermediary categories. Inaccuracies in forecasting the location of the strongest parts of the storm, or in picking up certain areas of activity also restrict the quality. The forecasts using CDP data vary greatly in their quality between cases, with some of the thunderstorm events not being forecast at all, thus affecting the overall results for this forecast type.

4.2 Limitations of the project

There are several limitations of this project that may have influenced the accuracy of the results. Several of these relate to radar data processing. Firstly, in the area directly above the radar site, the beams would be below the convective cloud base and thus accurate figures for the maximum rainrate in the cell might not be achieved. Secondly, the radar beam attenuates as it passes through rainfall or cloud. A correction had been applied in the radar data used in this study, but this could be inaccurate in cases of severe attenuation. This means that reflectivity of storms far away from the radar may be incorrect in such conditions. Thirdly, the radar data used had not been corrected for the overestimate in radar reflectivity caused by melting. Lastly, an overestimate of reflectivity may also be caused by anomalous propagation. This occurs under certain weather conditions (such as temperature inversions), when the radar beam is bent downwards from its normal path, possibly intersecting the ground.

Although steps were taken to eliminate dynamic rain from the forecast data, these may not always have been effective. Thus the Nimrod forecasts, which used total rainrate masked with convective cloud cover, may compare better to the radar data which also used total rainrate, than the other two forecast types. Finally, the logarithmic profile of the idealised mature thunderstorm cell used to convert the rainrate at the ground to the maximum in the vertical makes the assumption that some mature cells are positioned in each grid cell. This may not always be the case and as cells at different life stages have different profiles this may limit the accuracy of the disruption forecasts.

5 CONCLUSIONS

This project has quantified the spatial and lead time accuracies of forecasts of air traffic disruption due to thunderstorm activity for three different types of Met Office forecast. Forecast quality has shown to be greatest for disruption forecasts prepared from Nimrod data and following a method which uses both storm intensity and coverage to predict deviations. Forecast accuracy and skill decline with increasing lead time and spatial resolution up to T+5, and remain fairly constant at lead times greater than T+6. ATM authorities will need to interpret these results in the context of their own operations. For example severe disruption may not be a problem if the density of traffic is low, or not going to impinge on other air space sectors. In addition, they need to establish the costs and benefits involved in taking action based on forecasts for a given lead time and spatial resolution.

6 REFERENCES

Golding, B.W., 1998: Nimrod: A system for generating automated very short range forecasts. *Meteorol. Appl*., **5**, 1-16

Hand, W.H., Submitted: The UK Met Office Convection Diagnosis Scheme. *Meteorol. Appl.*

Hand, W.H., 1996: An object-orientated technique for nowcasting heavy showers and thunderstorms. *Meteorol. Appl.,* **3**, 31-41

Hoad, D.J., 2000: Verification of short range thunderstorm forecasts using radar data to assess their benefit to the aviation community. *9th Conf. on Aviat., Range, and Aerosp. Meteorol.,* 188-191

Rankin, J., 2001: Personal communication

Rhoda, D.A. et al, 2000: Commercial aircraft encounters with thunderstorms in the Memphis terminal airspace. *9th Conf. on Aviation, Range, and Aerosp. Meteorol.,* 37-42

Rhoda, D.A. and Pawlak, M.L., 1999: The thunderstorm penetration/deviation decision in the terminal area. *8th Conf. on Aviation, Range, and Aerosp. Meteorol.,* 308-312

Seymour, J., 2002: Personal communication

Wilks, D.S., 1995: Statistical methods in the atmospheric sciences. Academic Press.