Numerical Weather Prediction

The Value of Field Modifications in Forecasting

Forecasting Research Technical Report No. 437

Tim Hewson

email: nwp_publications@metoffice.com

The Value of Field Modifications in Forecasting

Tim Hewson. Operations Centre / Verification / NWP, Met Office, Exeter. *First version 6/2/04. Revised 25/2/04.*

Abstract

Field modification techniques have now been used operationally in Met Office forecasting for about 5 years in the medium range, and about 15 months in the short range. This report has examined, using a wide range of objective and subjective measures, the magnitude of value added by the modifications, and proposes a way of combining these measures into a single ‘modifications index’ which represents, as a function of lead time, the ‘lead time gain’ of the modified forecast over the unmodified. This measure shows a positive modification impact across all lead times, albeit with something of a minimum around T+24/48. The early fall in the value-added profile probably reflects the reducing relevance, as lead time increases, of differences between our model and observations (including imagery); conversely the rise beyond T+24/48 reflects the increasing utility and, considering issue times, availability of other model runs. There seems to be justification for continued use of modifications at all lead times on three grounds. Firstly, the lead time gain is significantly greater, at all lead times, than the time taken to consider and effect any modifications. Secondly, by most measures the added value has, over time, increased. And thirdly, whilst it will require an investment of approximately 200 NWP man years (over the next 1 to 2 years) to make the raw model product match the accuracy of the current modified product, over this same period total expenditure on forecaster time for applying modifications will be just 2 man years. Though this final difference may seem striking, it should not be forgotten that our raw NWP output provides the essential foundation for successful forecasting, and is, by many measures, the best in the world (a position we must strive to retain). Field modification builds on this, clearly helping to maintain our world-leading position in *forecasting*, in a cost-effective manner: its impact could and should be quoted in advertising and marketing, in the same way that the high quality NWP output currently is.

In the short range the forecaster is best able to add value for low cloud and snow, and for showers in cold air convective outbreaks, though conversely modifications seem to make forecasts of relatively light rain slightly worse. In the medium range, use of the so called poor man’s (multi-model) ensemble improves the mslp field and substantially improves the positioning of bands of significant weather. Some recommendations for changes in forecasting practice have been included, such as more cautious use of modifications in warm air (summer) convective outbreaks.

1. Introduction

Textual and graphical guidance provided by chief and deputy chief forecasters in the Operations centre in Exeter lie at the heart of Met Office forecasting activities. Almost all other forecast guidance, for the 0 to 10 day period, stems from this. The standardised graphical guidance component, which is becoming increasingly important, comprises either raw model output, or modified model output. Various tools, developed by Eddy Carroll, and referred to as ‘On-Screen Field Modification’ (OSFM) (see Carroll (1997)) facilitate dynamically consistent modification of the model fields when it is deemed necessary. Whilst modifications are not made to model output without good reason (see section 6 for examples), for the purposes of forecaster feedback, and indeed justification, it is important to verify the value, or otherwise, of any changes made. Both objective and subjective measures are required, at all lead times. This is because all measures have intrinsic weaknesses, and slavishly relying on one would encourage production of misleading forecasts. Root mean square (RMS) errors in mslp, for example, tend to penalise the retention of deep lows, whilst for rainfall rates they have no relevance at all. This report brings together verification results from various sources relating to field modification, for lead times up to 10 days.

The products referred to above are essentially ‘deterministic’ – there are also important probabilistic components to guidance, although these will not be verified here. Suffice it to say that for the vast majority of meteorological probability distributions – ie those which are unimodal - the centre (or mode) can be very conveniently represented by the ‘deterministic’ forecast. This forecast then provides an essential focal point about which any lower probability alternatives can be visualised to lie.

Whilst field modification may improve the forecast, it is important to consider whether any improvements are sufficiently large to counteract the time spent effecting modifications. Similarly, what is the cost of making modifications, and how does this compare with the costs of achieving similar improvements through the

NWP route? Indeed should resources be re-allocated? These questions are all discussed, quantitatively, within this report.

We focus primarily on forecasts issued during the period July 2002 to January 2004, although for various technical/procedural reasons some statistics are for shorter or longer periods. Previous, related reports are Hewson (April 2002) which looked at medium range guidance, and Hewson et al (March 2003) which concentrated on a very limited period of short range guidance. This report is a comprehensive update of both of these. Table 1 shows the types of verification referred to in this and previous reports. As regards subjective verification, neither chief nor deputy take any part in this – ie no-one verifies their own forecasts.

Lead times, and run data times (label used in text)	**Parameters**	**Area**	**O**bject-ive/ **S**ubje-ctive	**Availability:** **2000**	**2001**	**2002**	**2003**
(A) T+24 (00Z,12Z runs)	Mslp	Mesoscale	O		—	—	—
(B) T+24 (00Z,06Z, 12Z, 18Z)	Mslp, fronts	Mesoscale	S		—	—	
(C)* T+6,12,18,24 (06Z,18Z)	Mslp (wind), total cloud, ppn rate, ppn type	British Isles	S			—	—
(D)* T+6,12,18,24,30 (00Z,06Z,12Z,18Z)	Low cloud, ppn rate, ppn type	British Isles	O			—	—
(E) T+36,48,60,72,96,120 (00Z (not 96,120), 12Z (all))	Mslp,700mb RH, 850mb theta-W, 500mb ht, 250mb ht	Mesoscale (ASXX)	O		—	—	—
(F) T+84,108,132 (00Z)	Mslp, 700mb RH, 850mb theta-W, 500mb ht, 250mb ht	Mesoscale	O			—	—
(G) T+48,72,96,120 (12Z)	Mslp, low depth / position, front positions, 1000-500mb thickness	Extended Mesoscale	S	—	—	—	—
(H) T+144 to T+240	Mslp	Extended Mesoscale	S	—	—	—	—

Table 1: Verification types. The three sets of rows (heavy outline) correspond to different lead time ranges – top set, A to D, is short range (section 3), middle set, E to G, medium range (section 4) and bottom set, H, longer range (section 5). Parameters in C and D are all depicted on standard graphical guidance – for examples see Figure 12 (and also the archive accessible internally within the Met Office via http://www-cf/~cfth/NMC_subj_verifn/Link_Page.html).

In section 2, two simple new measures of *relative forecast accuracy* - 'lead time gain' and 'percentage gain' – are introduced. These aim to unify and thereby facilitate direct intercomparison of any accuracy measures derived by objective or subjective means. Thereafter we deal with the detailed short range (~0 to 1 day) forecasts prepared by the chief in section 3, and less detailed medium range forecasts provided by the deputy (~2 to 5 day) in section 4. Brief reference is made to longer term trend forecasts (~6 to 10 day) in section 5. Section 6 provides examples of changes that are typically made, and reasons for them, whilst section 7 summarises results and section 8 makes recommendations.

2. A unifying measure of relative forecast accuracy

Consider the situation where we want to compare two or more different forecasts - such as modified and unmodified, or forecasts by two different models – but have a variety of ways of measuring the accuracy of those forecasts (such as RMS errors, hit rates, equitable threat scores, subjective skill scores etc). How can we reconcile the different measures and indeed potentially combine them into one? Provided the measures of accuracy satisfy two simple constraints, then one unifying measure, called 'lead time gain', which refers to the performance of one forecast *relative* to the other, can be very easily calculated. The constraints are (i) that the accuracy measures, when applied to a sufficiently large sample, decrease monotonically with lead time, and (ii) that these measures are available for at least two separate lead times for each forecast.
Figure 1 shows hypothetical accuracy traces for two forecast systems, as represented by the thick black and thick blue lines, which satisfy the above constraints. The lead time gain, Ψ, of the blue forecast system (F) over the black forecast system (G), at time T_1, is given by:

$$\psi_{T_1} \quad = \quad (a+b)/2$$

$$\therefore \qquad \psi_{T_1} \quad = \quad \frac{1}{2}(F_1 - G_1)\left(\left(\frac{T_1 - T_0}{G_0 - G_1}\right) + \left(\frac{T_2 - T_1}{F_1 - F_2}\right)\right)$$

This essentially represents the mean horizontal separation, at time T_1, of the blue and black curves, or the lead time difference beyond the black forecast at which the blue forecast would likely achieve a similar accuracy. When quoting a 'lead time gain' the lead time at which that applies (T_1 here) must always be quoted too. The same formula would also apply if the accuracy measure were an error measure, which increased with lead time. However, if the curves cross considerable care is required in computation. Where the forecast accuracy curves are less complete, a simplified version of the above equation, based on just *a* or *b*, can also be usefully applied. Lead time gain is closely linked to one of the Met Office's key NWP targets, namely the 'Core Capability Component' of the 'Efficiency Index', which measures how far ahead in lead time, on average, today's forecasts can go before accuracy deteriorates to match that of the average 1-day forecast in 1980. This is re-visited in section 7.

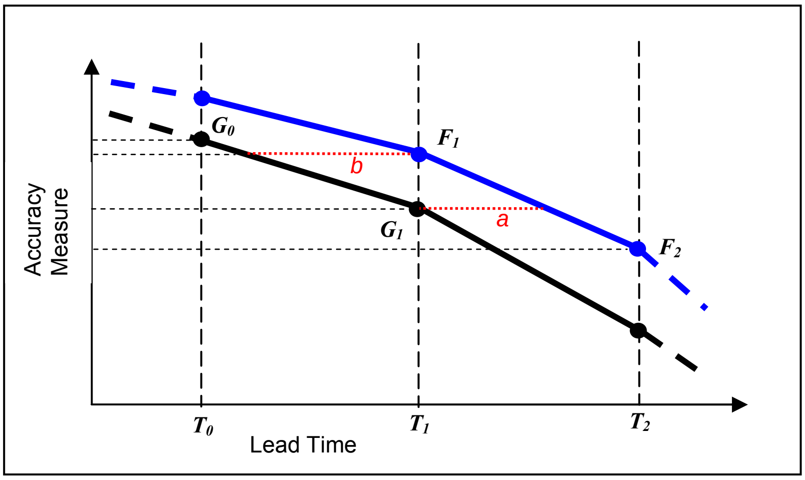


Figure 1: The derivation of 'lead time gain' from two accuracy curves (thick black and blue lines).

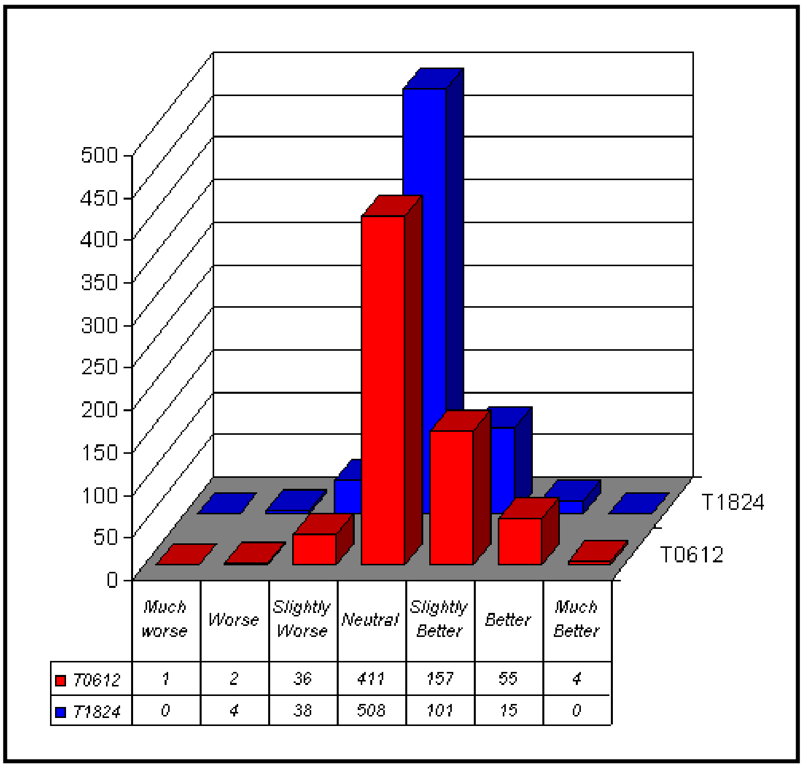
To enable further 'unification' one can also compute the 'percentage gain', as given by:

$$\psi_{T_1}(\%) \quad = \quad \left(\frac{\psi_{T_1}}{T_1}\right)\times 100\%$$

Across a range of lead times percentage gain might be expected to vary rather less than lead time gain. In turn, 'mean percentage gain' might then be used to encompass all lead times, and thereby express in simple terms the skill of one forecast system relative to another.

3. Short Range (~day 1)

The chief currently issues 4 sets of short range guidance per day, one to tie in with each model run. Table 1, rows A to D, show items being verified – i.e. (part of) the ASXX, and all the elements of the standard graphical guidance (described in detail in Hewson et al (2003)). Figure 2 shows the net results of the current subjective verification ((C) in Table 1) for two lead time bins. In both bins a positive skew indicates a net improvement in modified forecasts. It is easier to improve at short range, though substantial improvements are also seen on occasion at T+18 and 24. A large proportion of forecasts however are either not modified, or are modified in such a way that there is no net improvement or degradation - about 2/3 of cases at T+6 and 12, and about 3/4 at T+18 and 24. This may indicate judicious use of modifications. Some forecasts have been made worse, but overall it is 4 times more likely that the forecaster will improve upon the model. This compares with 3 times more likely in the smaller dataset represented in Hewson (2003), implying perhaps an improvement in forecaster contribution. This is despite a change in operational practice in 2003 which means that the forecasts are now issued an hour or so earlier than they used to be. This would be expected to slightly degrade the accuracy of the modified fields.



	Much worse	Worse	Slightly Worse	Neutral	Slightly Better	Better	Much Better
🟥 T0612	1	2	36	411	157	55	4
🟦 T1824	0	4	38	508	101	15	0

Figure 2: The quality of modified forecasts relative to unmodified, for two lead time bins (subjective).

To highlight what the forecaster is doing, Figure 3, from the subjective verification, divides up those modification types that have had a positive impact (left of bar) from those that have had a negative impact (right). The following conclusions can be drawn:

i) Forecasters are much more likely to improve the areal **cloud** coverage than make it worse, and they are more likely to do this by adding cloud

ii) It is difficult to improve the **mslp** field, although this parameter seems to be rarely changed.

iii) **Precipitation timing** errors are much easier to correct at very short range, though some skill is still apparent at T+18 and 24.

iv) Forecasters are good at improving the models **areal coverage of precipitation**; at longer lead times this generally involves ‘thinning out’ (often this is done in cold air convection, or in light rain and drizzle in warm sectors). Curiously, at short lead times the areal coverage seems to be just as likely to be underdone, possibly indicating model drift.

v) Attempts to change **precipitation intensity** are generally successful, with a clear bias towards successful *enhancement* of rates at short lead times. There may be merit in also enhancing rates at longer lead times, in certain situations (cold air convection, for example).

In general objective verification results agree with the subjective results above. Figure 4 shows the **mslp** root mean square (RMS) error difference between modified and unmodified forecasts, as represented on the T+24 FSXX (row A on table 1). Whilst there has been an improvement in modified fields, over the years, partly attributable to feedback from the subjective T+24 verification (row C in Table 1, now ceased), in most months modified forecasts are still very slightly worse than the unmodified global model (GM) forecasts. However, the mesoscale model (red line) continues to perform worse than the GM by this measure (and indeed worse than the modified forecast), in part due to a larger negative mslp bias (not shown) and perhaps also because of higher resolution and a tendency to produce more intense lows. *It should also be noted that whilst subjectively a deeper low in a particular forecast might score very well, rms errors will heavily penalise such a forecast unless the low’s position is very accurate*.[*] Given a clear remit to try to accurately forecast severe weather, in general forecasters will not fill an extreme low in a short range model forecast unless there is a very good reason. Because of this the relative performance of modified forecasts in the latter part

[*] (for an example see Met Office (2003), Figure 17, where the root mean square error in the very heavily modified forecast is, remarkably, the same as in the much poorer unmodified forecast (also available at http://www.metoffice.com/corporate/scitech0203/6_forecasting/man_machine_mix.html)).

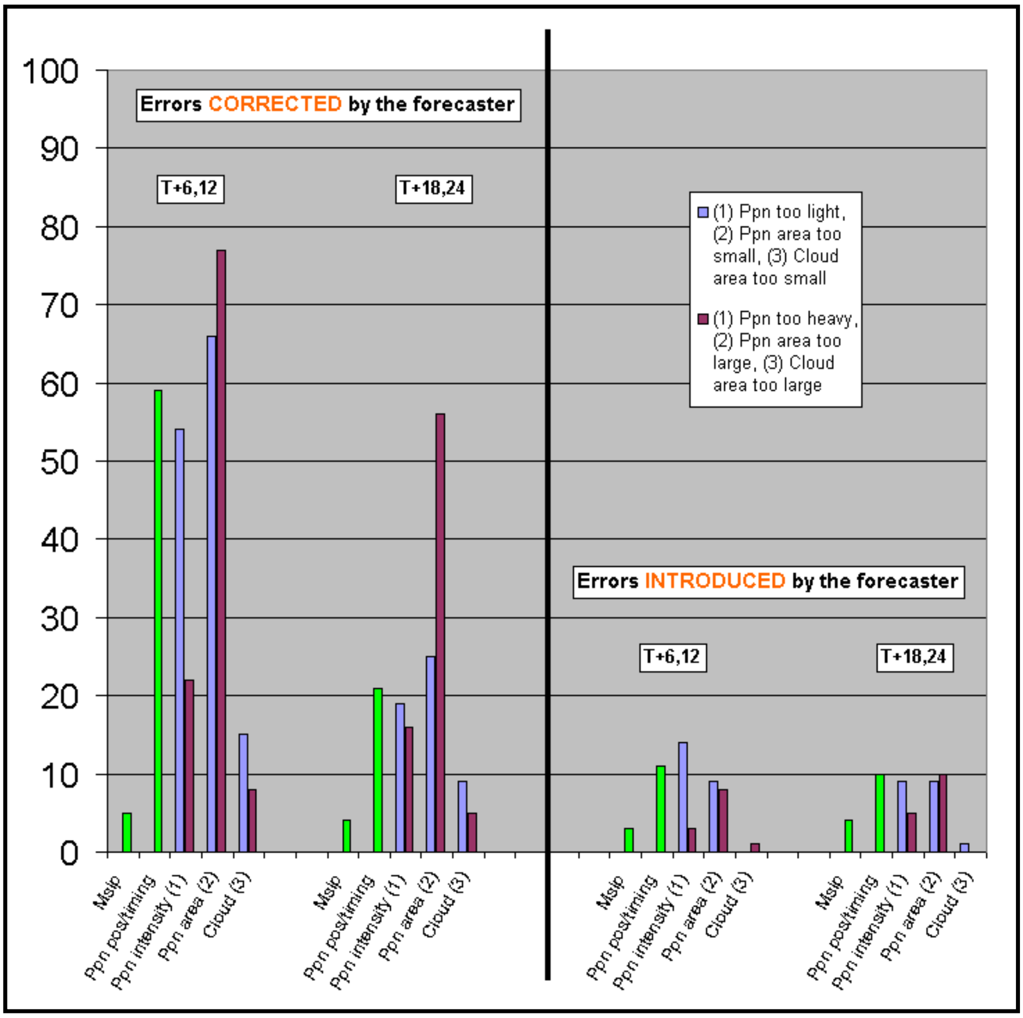


Figure 3:: Short range forecast errors (subjective)

of 2003 may well be as good as it will ever get (unless a new model comes along which is vastly superior to the GM – unlikely). Indeed it may be counter-productive to try to improve the mslp rms errors much further. Daily plots (not shown) and (ii) above indicate that it is rare for the mslp field to be changed at T+24 – in the last month 10% of such fields were modified, all applying a negative change (ie deepening lows), probably mainly in response to the Meso signal.

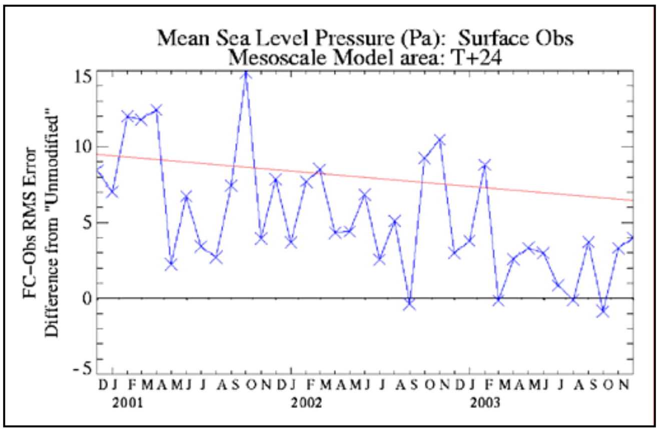


Figure 4: Rms mslp errors at T+24. Blue line compares modified with GM, red line (approx trend) compares Meso with GM. Values above zero indicate GM 'better'.

Errors in **low cloud** amount, objectively assessed against SYNOP observations, are depicted on Figure 5. These also concur with subjective results, but the improvement achieved by the forecaster is quite striking, corresponding to a lead time gain, Ψ, for T+6 to T+18 of about 6 hours. The gain may be partly due to the forecaster reducing the negative bias in model low cloud (top panel). To check consistency, the last 40 days of results were examined more closely – at T+12 about 10% of forecasts were significantly modified, and 90% of those changes improved the forecast.

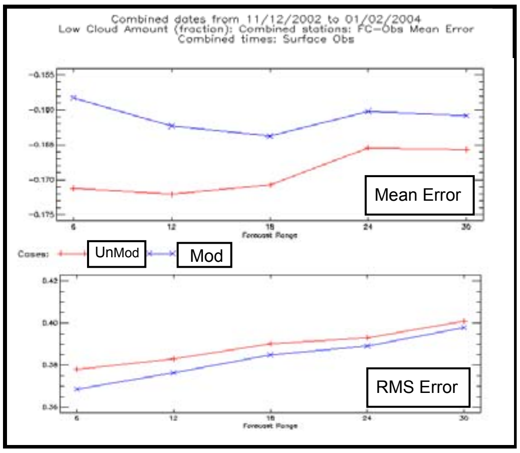


Figure 5: Errors in short range low cloud forecasts.

There is a multitude of ways of objectively verifying **precipitation rate** forecasts. The most useful of these arguably rely on 2x2 contingency tables for different rate thresholds. A standard Met Office measure derived from such tables is the equitable threat score (ETS) – see Forrester (2001). Figure 6 shows, for two rate thresholds - 0.5mm/hr (relatively light) and 2mm/hr (heavier) - modified and unmodified ETS values. This can be directly compared with Figure 4 in Hewson (2003), which shows validating data for a much shorter period. Figure 6 shows that modifications continue to make forecasts for the lighter rate threshold slightly worse, although the differential is noticeably less than it was. T+12 stands out as being the worst lead time – this may perhaps be due to extrapolation, by the forecaster, of current radar signatures beyond a valid time range, or over mountains – and probably also represents a bias in enhancing rates more often than reducing them. This assertion is supported by a brief manual analysis of 6 winter months of forecasts for Glasgow, Heathrow and Plymouth. When modified was worse than unmodified it was 6 times more likely to be due to the modified rate being too high than to it being too low. Similarly the mean rate error for mod, for all stations, ranges from +0.028mm/hr at T+6 to +0.021 at T+24; for unmodified the equivalent values are 0.019 and 0.018. A change in practice is probably warranted.

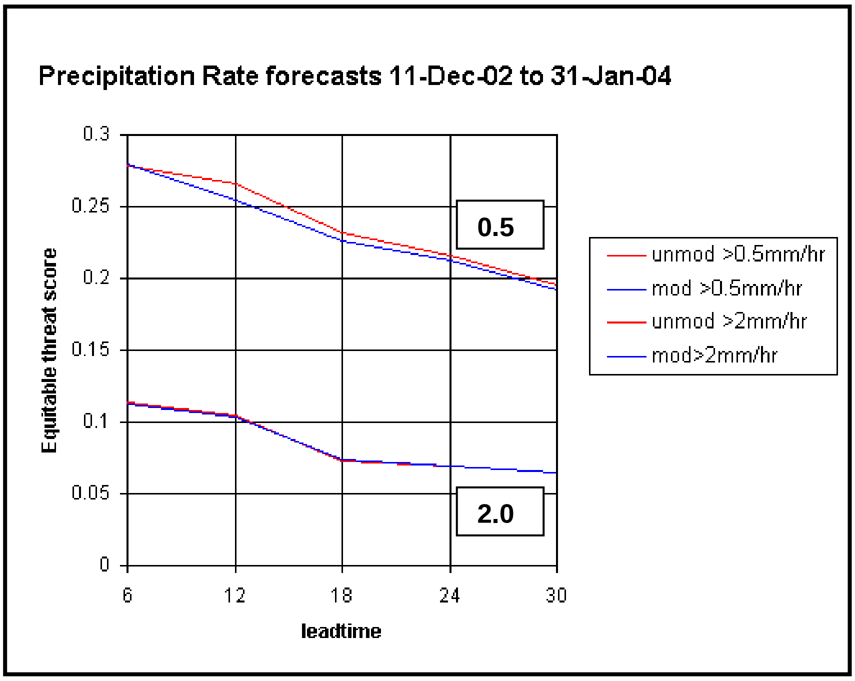
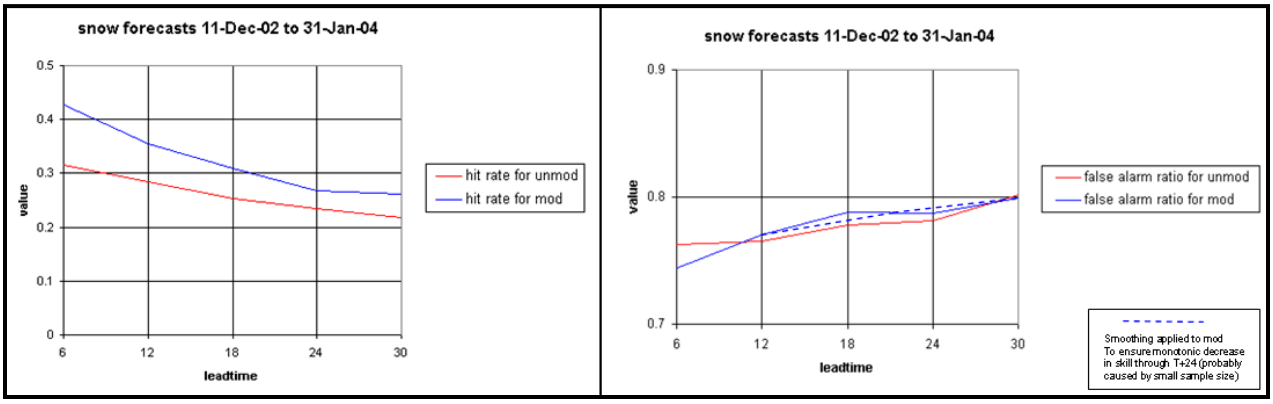


Figure 6: Equitable threat scores for two rainfall rate thresholds, for all UK SREW-reporting stations.

For the higher 2mm/hr rate thresholds there is minimal difference between mod and unmod. Previously, in Hewson (2003), mod had been better at 3 lead times, the same at 1 and worse at 1; however the ETS for the heavier rate had not shown a fully monotonic decrease with lead time, in either mod or unmod, implying that the sample size then was rather small – now we have more data that problem seems to have gone. As regards biases for *incidences* of rainfall these are low, being 2% overprediction (over all lead times) for mod, and 3% overprediction for unmod, when verified against SYNOP present weather.

Correctly capturing **snow** events is a key goal for the chief. Whilst it would be advantageous to verify snow intensity this is very problematic, being complicated by lack of snow depth measurements, wind blown snow, infrequent occurrence of heavier snow etc. For this reason we choose to verify snow using simple contingency tables for snow or no snow, irrespective of intensity, with SYNOP present weather again providing validation. Model output in standard Ops Centre format (Figure 12) represents precipitation as snow at any gridpoint where the proportion of convective or dynamic snow is non-zero (even if more rain is diagnosed). This is in both unmodified and modified fields. Thus prior to making any modifications the meso fields have already been pre-processed. In part because of this approach, biases for snow prediction are non-trivial, showing a 45% overprediction for mod and 15% over-prediction for unmod (although these biases have reduced markedly since Hewson et al (2002)). Figure 7 shows standard hit rates and false alarm ratios. There is a clear advantage of the modified over the unmodified, in terms of hit rate, although this is partly counteracted, at T+12, 18 and 24, by modified having a greater false alarm ratio, due primarily to the large bias noted above. The behaviour of the unmodified fields is peculiar, showing a 20% reduction in frequency of snow forecasts between T+6 and T+12, which then persists; unmod shows a more gradual, though notable, reduction with lead time. Both should probably be addressed. Such behaviour also adversely impacts on a number of verification measures that have been investigated – odds ratio, for example, increases dramatically for unmod between T+6 and T+12. Indeed overall verification measures for snow, both modified and unmodified, exhibit much more volatility, as a function of lead time, than many other measures. This is only likely to improve with more data. Despite the problems, results overall – notably for hit rate - are similar to those in Hewson et al (2003).



Smoothing applied to mod To ensure monotonic decrease in skill through T+24 (probably caused by small sample size)

Figure 7: Hit rates and false alarm ratios for point forecasts of snow, using all available SYNOP reports (including automatic observations).

The final part of the subjective verification scheme deals specifically with higher profile weather events - ie recording 'serious errors from the perspective of **hazardous weather**' – see Table 2. There are fewer occurrences in the modified (blue) column, indicating a positive contribution by the forecaster. Cases have been categorised synoptically – see the two letter labels. It seems that errors in heavy convective precipitation account for many of the reports. The forecaster is good at improving upon the model in cold air convection, but seems unable to add value in warm air convective situations. This is probably due to well-documented limitations of parametrised convection - such as insufficient rate variation, and insufficient inland penetration - being successfully addressed in cold air regimes, whilst the finely balanced situations that often characterise (summer) warm air convective outbreaks prove very hard to second guess. Frontal waves and intense lows, based on a small sample, also seem to cause equal problems for forecaster and model.

Data time	UnMod error?	Mod error?	Description of Error
06Z, 10/1/03	CC		Icy roads after showers not forecast to come inland
18Z, 16/1/03	CC	CC	Very heavy frontal, orographic, convective rain W. Scotland greatly underestimated
06Z, 17/1/03	CC		Very heavy frontal, orographic, convective rain W. Scotland greatly underestimated

06Z, 18/1/03	WV	WV	Heavy rain on cold front way out due to frontal wave
18Z, 18/1/03		LO	Frontal wave deepened far too much by T+12 implying gales that didn't occur
06Z, 21/1/03	CC		Heavy rain from small convective low S. England underestimated
18Z, 29/1/03	CC		Poor positioning of frontal rain/snow, and heavy convective snow *
18Z, 4/2/03		CC	Too much penetration of snow showers through Cheshire gap
06Z, 1/3/03	CC		Large area of heavy rain and thunderstorms completely missed at short range
18Z, 6/3/03	CC		Intensity of cold front rain and cold air convection following both lacking
06Z, 29/5/03	WC	WC	Localised severe thunderstorm activity in N of UK missed; warm air, slack flow
06Z, 30/5/03	WC	WC	Isolated thunderstorms and flash flooding in S Scotland missed
06Z, 1/6/03	WC	WC	Large continental thundery rain areas E England missed at T+12 and 24
18Z, 1/6/03	WC	WC	Insufficient intensity in thundery rainband; mod slightly worse
18Z, 26/6/03	TI	TI	E'ward movement of heavy rainbands at T+24 too fast, due to low too deep
06Z, 1/7/03	CC		Cold air convection lacked intensity and areal extent (including Wimbledon)
06Z, 17/7/03	CC		Convection around cold low lacked intensity
06Z, 24/7/03	WV	WV	Very active frontal wave in SW missed (parent low near NW Scotland too deep)
18Z, 9/8/03	WC	WC	Exceptional thunderstorms in NE in hot airmass missed
18Z, 17/11/03	LO		Intense low with storm force winds N Scotland completely missed
06Z, 12/12/03	WV	WV	Very active frontal wave S of UK completely missed at T+24
18Z, 21/12/03	CC		Cold air convection (mostly snow) lacked intensity and inland penetration in E
06Z, 9/1/04	CC		Heavy, organised cold air convection completely missing over S of UK at T+6
06Z, 11/1/04	LO	LO	Major cyclonic storm in S never materialised *

Table 2: All incidents of 'Serious errors', as reported in the subjective verification scheme. Coloured bars show whether these were flagged in unmodified or modified forecasts. Labels signify 'synoptic type': CC = cold air convection, WC = warm air convection, LO = deep low, WV = frontal wave, TI = timing. * denotes very high profile media events.

4. Medium Range (~days 2 to 5)

For obvious reasons medium range guidance is less detailed than its short range counterpart. Thus medium range verification should focus more on broadscale errors in the main synoptic features, such as cyclones and fronts. Instructions for the subjective verification scheme (which is very different to that used in the short range) encapsulate these features well. In the objective scheme we can use mslp as an indicator of the synoptic pattern, and 700mb relative humidity (RH) as a reasonable indicator of frontal activity and significant weather (higher values correlating with cloud and rain). Verification based on objective fronts has unfortunately floundered due to resourcing problems.

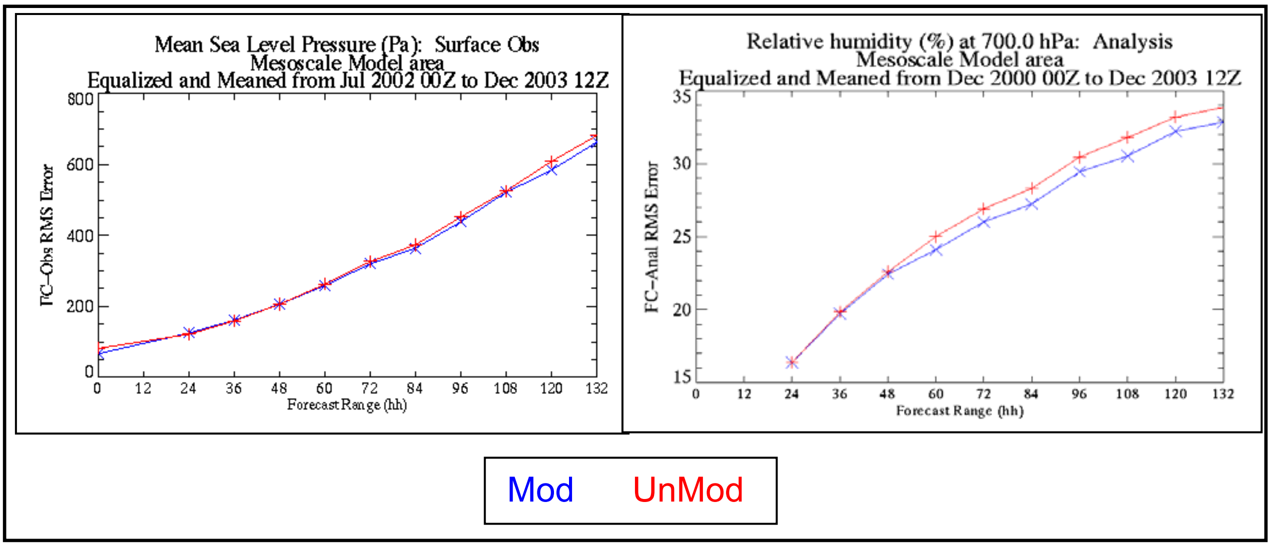


Figure 8: Modified and unmodified average rms errors in mslp and 700mb RH. Note the longer averaging period for RH – a shorter period plot was unavailable.

Figure 8 shows long period rms errors over the mesoscale model area in mslp and 700mb RH (rows E and F in Table 1). There is a positive and generally increasing lead time gain, Ψ, for the modified forecasts at all lead times from T+48, ranging from a few hours for mslp, to an impressive 14 hours by T+120 for RH. The time series in Figure 10 show, despite the odd bad month, a general positive trend in forecasters' ability to make beneficial modifications, with this trend being most striking in the RH field at T+48 (recall also the positive trend on Figure 4). Plots similar to Figures 9 and 10 exist also for other variables (not shown). Those for 850mb wet bulb potential temperature show similar signatures, including lead time gain, to those for mslp. Those for 500mb and 250mb height show generally negligible differences between mod and unmod. Given that the forecaster spends no time focussing on these levels this is not surprising – modifications targeting such levels probably could produce improvements. The neutral result is however important in that it indicates that the quasi-geostrophic and other assumptions implicit in the field modification process are not having a net detrimental effect on fields remote, in height, from those being targeted.

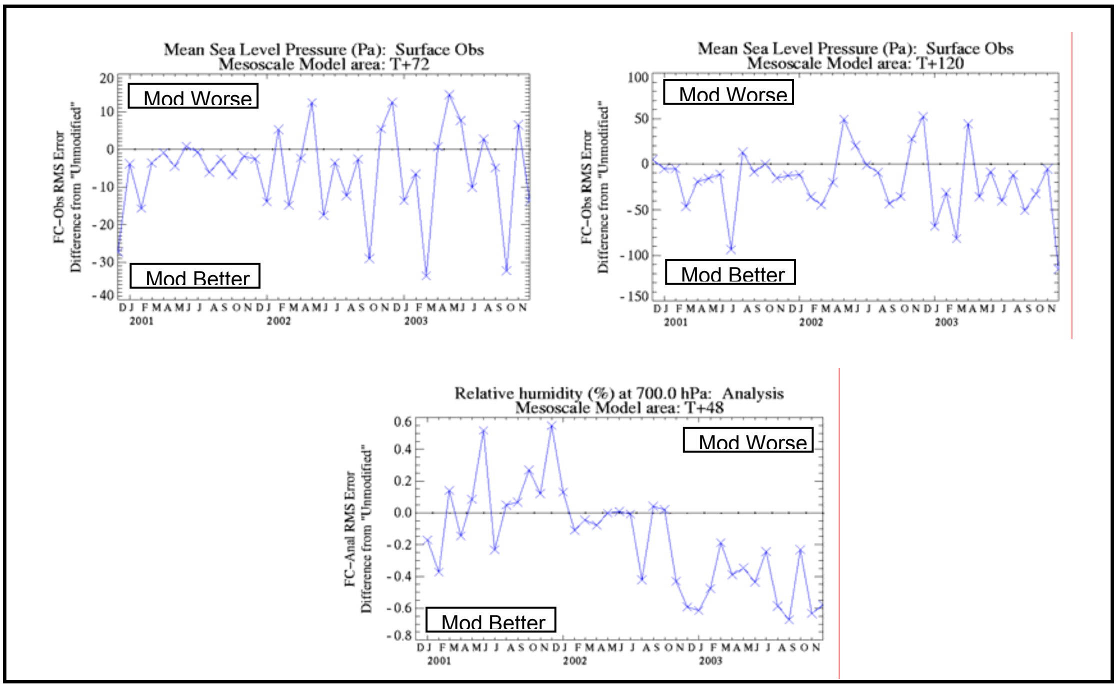


Figure 9: Difference between modified and unmodified rms errors, for mslp at T+72 (top left) and T+120 (top right), and RH at T+48 (below). Black line at zero is the dividing line between modifications improving / degrading the forecast.

Figures 4, 8 and 9 all derive from the mesoscale model area verification. Two other areas are also routinely verified objectively, one larger, the ASXX area, which covers much of the North Atlantic, Europe, and some way beyond, and one smaller, the British Isles area. As a general guide, the lead time gain between mod and unmod for the larger area reduces by about 33% compared to the mesoscale area, whilst for the smaller area it increases by about 33%. This is symptomatic of forecaster focus on the British Isles – impact is watered down when larger areas are considered.

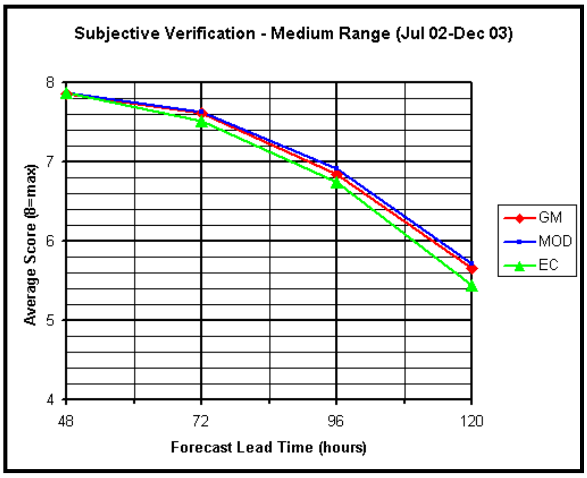


Figure 10: Average nominal skill score for 12Z GM, Modified GM (MOD) and ECMWF (EC) forecasts, from the medium range subjective verification scheme (each forecast frame is allocated a skill score of 0,2,4,6 or 8).

Figure 10 shows the latest subjective verification mean scores (row G on Table 1), on which the lead time gain for modified, whilst being positive at all times, is noticeably less than it was in Hewson (2002). Although in a longer subjective verification series the year 2001, discussed in Hewson (2002) does appear to have been especially good for modifications, the newer results are nonetheless at odds with the signal from the objective measures, where over the mesoscale area the difference between modified and unmodified is equal to or greater than it was previously. On Figure 10 the average scores have increased considerably since Hewson (2002) – eg 7.6 for GM at T+72, compared to 7.1 previously. This leaves less scope for recording improvements within the scoring scheme (which for each forecast can only register 0,2,4,6 or 8). In turn the step change in scores may be at least partly due to a step change in verifier, where the verifier now has no real involvement in the medium range. Indeed the verifying procedure remains very cumbersome; many paper copies have to be compared (by comparison the short range scheme is more user-friendly, allowing simple animation and overlaying via a web interface). Also of note on Figure 10 are the lower scores achieved by EC at all lead times, compared to GM. This is opposite to the results in Hewson (2002), and also differs from objective CBS scores which show EC to be more accurate than GM - *if* the later data cut-off is ignored. Recall also that EC forecasts always arrive too late (because of the data cut-off) to contribute to any modifications represented on Figure 11.

5. Longer Range (~days 6 to 10)

Although longer range 'trend' guidance is issued daily by the medium range forecaster, based on a blend of operational and ensemble model guidance from different centres, no verification of the forecaster input has so far been performed. Arguably, the main components in the trend forecast are the EC operational and ensemble runs. These are verified (row H in Table 1). The results in Figure 11 show that neither the main run nor the ensemble mean provided useful guidance beyond day 6 more than 20% of the time. Indeed in the view of the (verifying) forecaster the operational run, though not very useful, is generally no less useful than the ensemble mean. This differs from conventional wisdom. It would be illuminating to somehow add the forecaster's steer to the verification scheme.

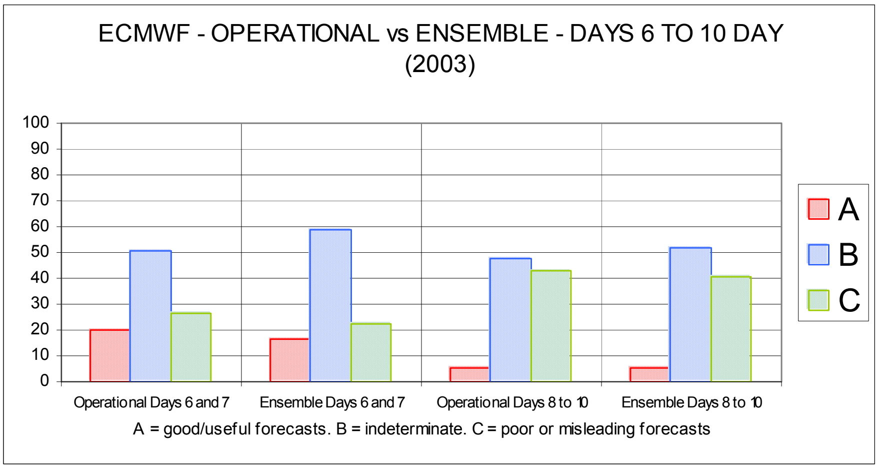


Figure 11: ECMWF longer range guidance subjective verification.

6. Modification Example

Figure 12 shows an unmodified and a modified forecast alongside one another, together with the validating data. This is in a format used in the short range subjective verification scheme. The Figure illustrates both the standard graphical output now being used, and some of the problems the forecaster encounters and tries to deal with. The weather situation depicted is more active than 'average'; similarly the number of modifications made is probably also greater than average. Labels have been added to illustrate some of the forecasting problems:

'A' signifies regions of cold air convection that have propagated a long way inland. These are largely missing from the unmodified forecast, but have been pasted in, reasonably successfully, on the modified forecast. This is a common model error.

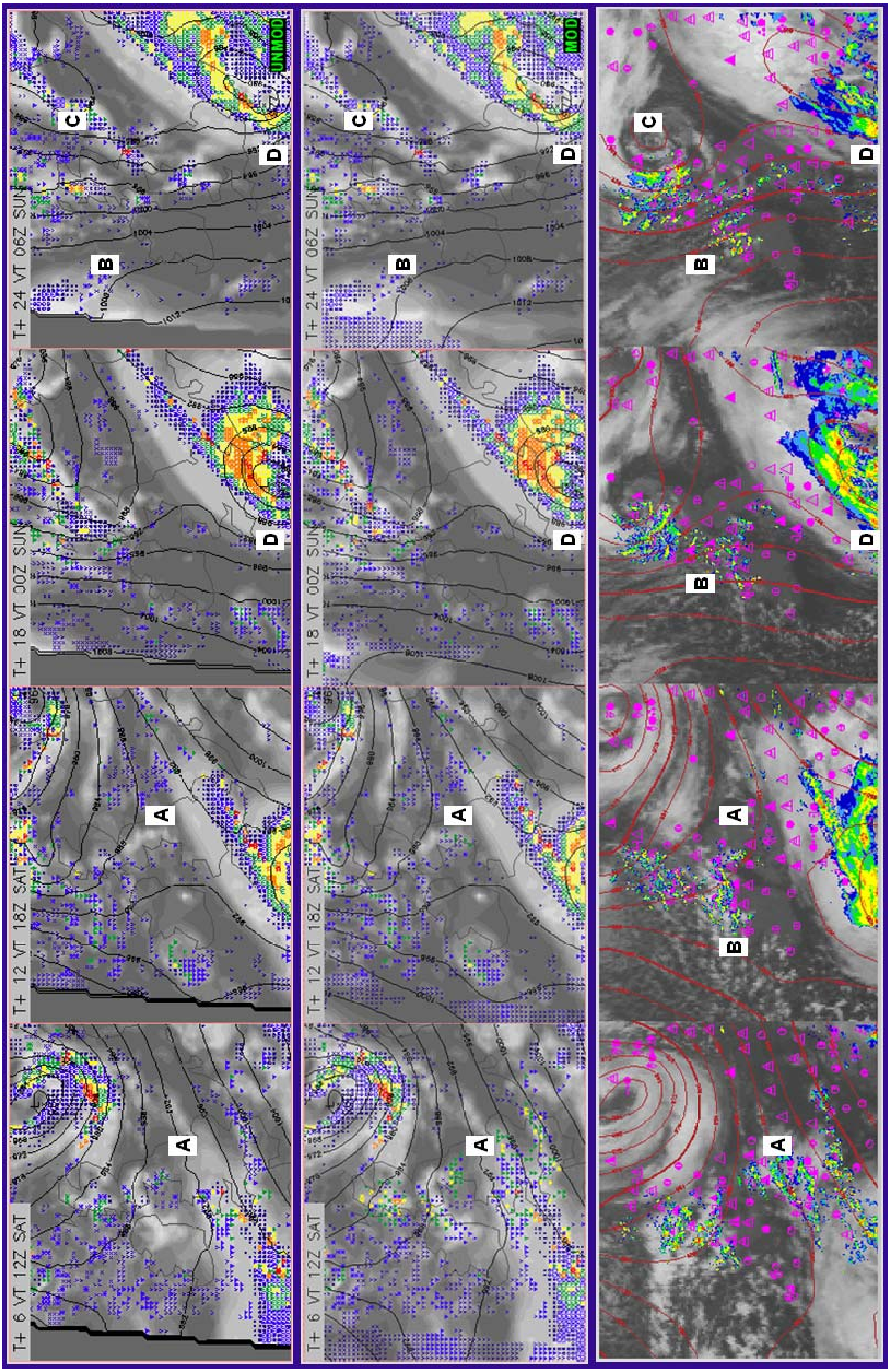


Figure 12: Operational short range forecast example from DT 06Z 27th December 2003 of unmodified (top panels), modified (middle panels) and verifying data (lower panels – covers identical area – some coasts are visible). On the top 8 panels black lines show mslp at 4mb intervals, grey shading indicates cloud (lighter = more), symbols indicate convective or dynamic rain or snow (V's and diamonds for convective rain, circles and stars for dynamic rain, spiky symbols for snow)), with symbol colour indicating precipitation rate in mm/hr (dark blue up to 0.5, blue 0.5-1, green 1-2, yellow 2-4, orange 4-8, red 8-16). On the lower four panels pressure is red (4mb interval, 980 and 1000mb thicker), Meteosat IR image is grey, present weather and total cloud observations are pink, and radar precipitation rate is coloured (same scale as model panels). Labels are referenced in the text.

‘B’ denotes, at longer lead times, a further area of cold air convection, containing heavy echoes, that has been missed on both unmodified and modified forecasts. The forecaster probably chose to not modify here due to lack of contrary data at the time of issue.

‘C’ shows a deep polar low type feature that the model has badly misplaced, and also underdeepened. The forecaster also failed to capture this feature. It is not uncommon for the model analysis to be in error by +4mb or more in such situations.

‘D’ denotes another deep low, with very strong gradients on its western flank in the model. Whilst there may have been doubts about the validity of this, it would take a brave forecaster to iron out such a low. High resolution short range ensemble runs might ultimately be the best way of dealing with uncertainties in the detailed evolution of lows such as this (and ‘C’).

7. Discussion and Conclusions

Appendix 1 puts forward a way of potentially combining the various scores described in sections 3 and 4 into single measures, or modification indices, which are conveniently also expressed as a lead time gain. Table 3 shows the component contributions and the modification indices computed using this methodology. *All* forecasts are represented, not just those that have been modified.

Lead	Short range: Ψ components (h)						Modification Index Ψ_T (h)	Ψ_T(%)
	Low cloud (rms)	Rain ETS ≥0.5 mm/hr	Rain ETS ≥2.0 mm/hr	Snow (HR & FAR)	Ψ_{subj}	Ψ_{haz}		
T+12	6.12	-3.89	-0.40	2.99	3.19	4.5	2.08	17.3
T+24	6.67	-1.35	0.00	2.99	1.33	4.5	2.36	9.8

Lead	Medium range: Ψ components (h)			Modification Index Ψ_T (h)	Ψ_T(%)
	Mslp (rms)	700mb RH (rms)	Subjective Skill Score		
T+48	0.00	1.56	0.74	0.77	1.6
T+72	1.24	5.66	1.20	2.70	3.8
T+96	2.53	7.95	1.63	4.04	4.2
T+120	3.88	14.32	1.19	6.46	5.4

Table 3: Modification indices and their components (in units of lead time gain (h)), and percentage gain.

Calculation was generally straightforward, though for snow the curves were less well behaved (see Figure 7), probably due to too small a sample, which in turn made the lead time gains very volatile. For this reason we had to compromise, using the smoothed dashed line on Figure 7 to ensure monotonic reduction in accuracy with lead time, then averaging out the net lead time gains at T+12, 18 and 24, then using these as both entries in Table 3. With a larger sample of wintry weather this should not be necessary.

Figure 13 shows lead time gain, and percentage gain, as represented by the modifications indices, in Table 3, as a function of lead time. Whilst positive at all times, there seems to be a minimum in forecaster contribution around T+24/48. This probably represents the crossover between using current trends to improve upon the short range forecast, and other models to improve upon the medium range forecast.

It could be argued that the improvements shown in the above table, and indeed on Figure 9 - mslp especially - are not, on first impression, all that great. So two key questions arise:

i) Is the lead time gain achieved by effecting the modifications greater than the time taken to do them?

ii) In the context of NWP progress how big are these improvements, and what is the relative cost of achieving them?

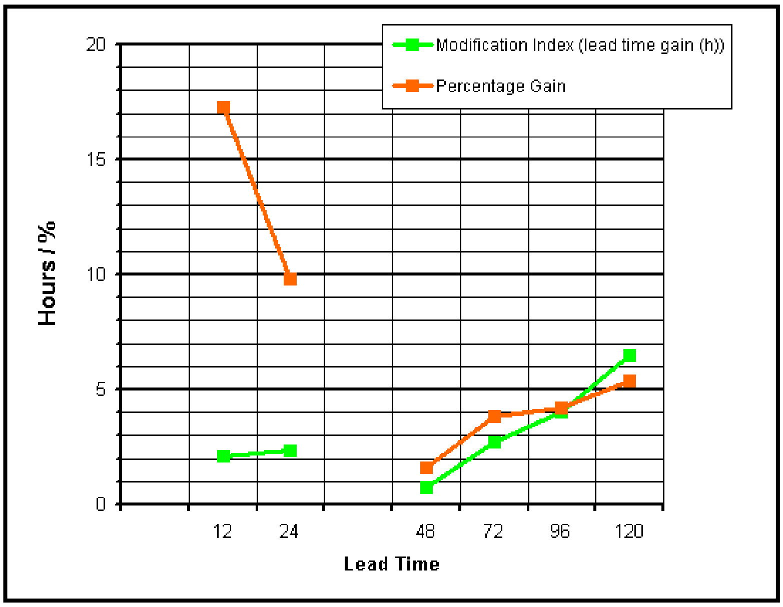


Figure 13: Modification index and percentage gain for short range forecasts (left group) and medium range forecasts (right group), for approx July 2002 to Jan 2004.

Regarding (i), time spent specifically on the modification process (that is assessing and incorporating information from other models, imagery and observations, and actually effecting any modifications) amounts to no more than about 5 hours per day – about 2 for the chief and 3 for the deputy. Though forecasters' approaches vary, typical values are shown in Table 4. Apparent discrepancies between the 'modification times' are largely due to schedules and deadlines. However in all cases the benefit, in terms of lead time gain, exceeds loss in terms of time expended (comparing with Table 3 and Figure 13).

In considering (i) above we have effectively compared with the hypothetical situation in which the forecaster issued all the normal guidance - graphics, frontal charts, text etc – starting at the same time, but did not effect any modifications. We have also not considered here the time spent *delivering* fields to the forecaster. Maximum delays currently tend to occur with 12Z mesoscale run data - up to about 20 minutes. Though this is inconvenient and needs addressing (and is intrinsically an IT issue) even when added to modification time, the 'total delay' is still much smaller than the lead time gain. Indeed there is no reason to believe that any customers receiving unmodified fields that weren't routed via the forecaster would not suffer similar delays.

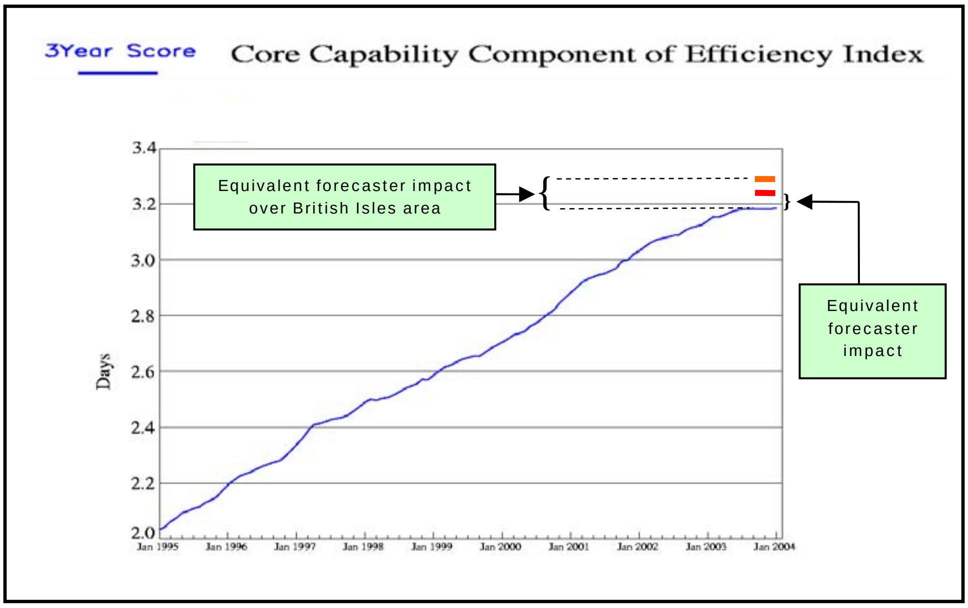
Forecast lead time	T+24, & short range graphical guidance (chief)	T+36,48 AM issue (deputy)	T+60,72 AM issue (deputy)	T+84 AM (deputy)	T+108,132 AM (deputy)	T+36,48 PM issue (deputy)	T+60,72 PM issue (deputy)	T+96,120 PM (deputy)
Modification time	0.4h for each of 4 issues	0.25h	0.4h	0.15h	0.25h	0.25h	0.6h	0.75h
Number of other recent quality model runs available	~1	~3	~4	~6	~6	~2	~4	~3

Table 4: Time spent on modifications, and approximate number of other recent international model runs available prior to issue time (the ECMWF ensemble suite has been nominally counted as 2 runs).

To address (ii) consider one aspect of standard NWP accuracy measures, the 'core capability component' (CCC). This effectively represents the raw NWP lead time gain (minus 1 day), over a baseline 1980 Met Office 1 day forecast (see Figure 14), calculated using mslp over an extended North Atlantic area. This area approximates to the 'ASXX area', one of the three standard modified fields areas verified objectively. The CCC currently stands at 3.2 days. Modified mslp lead time gain, over the ASXX area, at 3.2 days lead time, has been plotted directly alongside the CCC for comparison (red bar). In turn this suggests it will be another year or two, at the current rate of progress, before raw NWP improves to reach the accuracy level of the current modified mslp field. The orange bar illustrates the relative impact the forecaster is having over the (more focussed) British Isles area. As regards costs, for raw NWP output to reach the red bar will cost approximately 200 man years. In that time the expenditure on forecaster time, which for modifications runs at no more than 5 hours per day, will be less than 2 man years. This is two orders of magnitude less and seems to represent worthwhile investment in forecasting. Indeed perhaps we should be focussing more on modifications? Moreover, the improvements made by the forecaster are clearly helping to maintain our world-leading position; such figures could and should be used for advertising and marketing purposes.

In conclusion, in terms of value and cost to the organisation, and in terms of the relative advantage of delaying products slightly, there seems to be a strong case for continued use of field modification in both the short and medium range. Slight changes in practice and scheduling also have the potential to bring further improvements.

Figure 13: Met Office 'core capability component', and relative improvements achieved by the forecaster.



8. Recommendations

Table 5 gives a probable reason for the smaller modified-unmodified accuracy differential for T+84, 108 and 132 forecasts apparent on the left of Figure 9 - time pressures mean less time spent addressing these products. In turn this suggests that on accuracy grounds there would probably be justification for pushing back the issue time, at least for T+108 and 132 forecasts when there is some slack in the schedule (notwithstanding, of course, specific customer requirements for timeliness). This may be all the more fruitful given that the number of other quality models to consider peaks around the time of issue of T+84, 108 and 132 (5's on Table 5) – primarily because the full suite of EC products is then available. Hewson (2002) discusses how the extent and magnitude of any modification improvements depend strongly on the availability of other model runs – the so-called 'poor mans ensemble'. It should also be highlighted that ECMWF now do two operational runs per day (using some Met Office money!). Were the 00Z ECMWF operational runs also made available to us there would be scope for further improvements, especially in the afternoon and evening issues (noting that ECMWF is intrinsically an accurate model, rendering its contribution to any consensus forecast as positive on average - section 4 in Hewson (2002) gives a more quantitative discussion).

In the short term the only significant negative contributor to the modification index is lighter rainfall. Although at odds with the subjective verification this does seem to be correct and suggests the chief forecaster could perhaps take more care with modifications in this category. A possible reason for the degradation is perhaps a tendency to amplify rainfall rates more often than reduce them. This tendency may stem, understandably, from time pressures and the desire to not miss hazardous weather. Increased familiarity with modification software, and improvements to that software may reduce the time pressure element.

As regards convective rainfall, there is strong evidence of scope to improve upon rate distributions in cold air convective outbreaks - parametrisation forces a rate distribution that is too peaked in the light rate category, with both insufficient dry gridboxes and insufficient heavy rates. Forecasters should aim to address this, not just when the error can already be seen, say for T+6 to T+12, but probably also anticipating it, given the right airmass, at longer lead times. Modification tools for such changes could also be made less cumbersome. The tendency also to not move maritime convection inland is well known, and continues to need addressing.

Conversely little success seems to have been achieved in modifying warm air (summer) convection, suggesting this should not be adjusted without very good reason.

Cloud modifications tend to be fruitful and should be continued, remembering that the meso still underpredicts low cloud. Pressure errors, and indeed frontal waves, continue to be problematic, with no clear evidence, except perhaps at very short lead times, of a net ability of the forecaster to improve upon the model. However there are some notable exceptions, particularly with very small scale lows, which even the mesoscale analysis can fail to capture, thereby impacting detrimentally on the subjective verification which uses mesoscale analyses as truth.

In future the downstream impact of modifications on screen temperature and road ice forecasts (for example) is likely to become increasingly important. Whether derived through a site specific forecast model or other means these downstream effects will also need to be verified. The positive verification results for low cloud cover and cold air convection offer grounds for optimism that impacts in the sphere of 'Open Road' will be positive.

As regards subjective verification, the short range scheme has been failing in recent weeks, on occasion, due to lack of Horace workstations in the confined Exeter environment. Resource allocation will be required to ensure it continues. We cannot rely on objective measures alone.

The medium range subjective verification probably needs a rethink, to move to a web-based system. Also with improving models, or at least improving scores (!), the scoring system now has insufficient resolution. This could be quickly and easily addressed by using scores of 0,1,2,3,4,5,6,7 and 8 rather than just 0,2,4,6 and 8.

A method of verifying the forecaster's input to the trend guidance (days 6 to 10) should be sought.

Finally, it should be re-iterated that the time resource required to effect modifications in a considered manner should be carefully guarded into the future (under the re-engineering project, for example). This must also permit slight future changes in scheduling that could bring further improvements to the modified fields.

Appendix 1 - A possible 'Modifications Index'

By facilitating the direct intercomparison of many different types of accuracy measures, using the notions of lead time gain and percentage gain (section 2), we have taken a large step towards potentially formulating an easy-to-compute index to describe the value of field modification. There are however still problems in dealing with the short range subjective verification scheme, as it currently stands. This appendix suggests ways of dealing with these, and thereby arriving at a notional lead time gain for this subjective component. We then propose ways in which measures of lead time gain, from different subjective and objective sources, can be composited into a single score, giving due weight to those aspects of the forecast considered most important. We will call the final index a 'modifications index'.

Two parts of the subjective verification process, alluded to in section 3 above, and discussed more fully in Hewson (2003), seem to warrant inclusion in any short range 'modifications index' – namely (a) the net impact of modifications, as represented on Figure 2, and (b) the ability to adjust forecasts to correctly deal with severe weather, as represented on Table 2. A simple measure for (a), the 'notional subjective lead time gain' (Ψ_{subj}) can be calculated as follows:

$$\Psi_{subj} \ (h) \ = \ 12 \times (N_{better} - N_{worse}) / N_{total}$$

where N_{better} is the number of forecasts that were rated 'slightly better', 'better' or 'much better', N_{worse} is the number rated 'slightly worse', 'worse' or 'much worse' and N_{total} is the total number of cases. This would give a value of 12h if every forecast were improved, -12h if every forecast were made worse, and a value between -6h and +6h, whenever half or more of the modified forecasts fell into the neutral category. As such the index properties seem to be sensible, with likely values in about the right range compared to other objective values quoted in section 3. We can then deal with (b) by computing the 'notional hazardous weather lead time gain' (Ψ_{haz}) as follows:

$$\Psi_{haz} \ (h) \ = \ 12 \times (N_{unmoderror} - N_{moderror}) / N_{total}$$

where $N_{unmoderror}$ is the number of reports of serious errors in unmod (=red boxes on Table 1), $N_{moderror}$ reports of serious errors in mod (=blue boxes) and N_{total} the total number of cases with serious error reports (=number of rows on Table 1, which is *not*, generally, red plus blue). This would have a maximum value of 12 if no modified forecast had a serious error (unlikely!), and goes negative if unmod is overall better. Given the volatility of such a measure in quiescent weather, a further constraint to apply is that computation is not possible unless, say, N_{total} >= 10.

A1.1 Short range Index

Table A1.1 shows a proposed method of computing a short range modifications index. This broadly gives weight to available parameters in proportion to perceived importance to the forecasting process. Rain appears directly twice. Snow, which is arguably more important, appears only once - this is because over the year as a whole it is relatively rare. By giving one sixth weight to Ψ_{haz} the greatest single positive impact a forecaster can have on the index, in one forecast, would be to remove a serious error, from the perspective of hazardous weather, from the unmodified forecast. Equally they would be penalised, by the same amount, for adding one. Bearing in mind the high annual revenue derived from the national severe weather warning service this seems very appropriate.

Lead	Lead time gain components, Ψ, to be averaged					
	Objective				Subjective	
T+12	Low cloud (rms) (T12)	Rain ETS ≥0.5mm/hr (T12)	Rain ETS ≥2.0mm/hr (T12)	Snow (0.5 * (hit rate + false alarm ratio)) (T12)	Ψ_{subj} (T612)	Ψ_{haz}
T+24	Low cloud (rms) (T24)	Rain ETS ≥0.5mm/hr (T24)	Rain ETS ≥2.0mm/hr (T24)	Snow (0.5* (hit rate + false alarm ratio)) (T24)	Ψ_{subj} (T1824)	Ψ_{haz}

Table A1.1: Recommended components for a 'short range modifications index'

A1.2 Medium range Index

Table A2 shows a proposed method of computing a medium range modifications index. This focusses more on broadscale aspects, with 700mb RH delineating significant weather regions, mslp representing the broadscale pattern, and subjective scores representing both of these.

Lead	Lead time gain components, Ψ, to be averaged		
	Objective		Subjective
T+48 to 120	Mslp, mesoscale model area (rms)	700mb RH, mesoscale model area (rms)	Skill score (fig 11)

Table A1.2: Recommended components for a 'medium range modifications index'

Acknowledgements

- to Julie Oakley for maintaining the objective verification, and providing results, to Keith Sutlieff for providing medium range subjective verification data, and to the three reviewers.

References

Carroll, E.B. 1997. A technique for consistent alteration of NWP output fields. *Meteorological Applications*. Vol 4.

Forrester, D.A. 2001. NWP Verification System – Derivation of Verification Statistics. NWP Verification Document Paper 5. Met Office internal report.

Hewson, T.D. April 2002. Modified Forecast Accuracy during 2001. OSFM1 Project document. Met Office.

Hewson, T.D. 2002. Categorical Precipitation Type Verification. OSFM2 Project document. Met Office.

Hewson, T.D., Oakley, J.A. and Harris, G. March 2003. Verification report for NMC Field Modification project (OSFM2): Analysis Period: 11 Dec 2002 – 4 Mar 2003. Met Office.

Met Office, 2003. Annual Scientific and Technical Review 2002/3. HMSO.