Meteorology Research and Development

The impact of observation monitoring on Met Office NWP performance

Technical Report No. 533

Richard Dumelow
Colin Parrett

email: nwp_publications@metoffice.gov.uk

Document History for NWP Technical Report No. 533

Date	Version	Action/comments	Approval
24/3/09	0.1	Initial draft	Andrew Lorenc

ABSTRACT

An experiment is run to investigate the impact of the rejection/acceptance lists and bias corrections obtained from the observation monitoring process. The experiment uses a version of the Met Office operational, global NWP system and observational data over a one month period from 24th May 2007 to 24th June 2007. The NWP system is run as a continuous cycle of assimilations and forecasts with a 6-day forecast being run from 12UTC each day. Four runs are carried out. The CONTROL, in which all observation acceptance/rejection lists and bias corrections are applied to the observations; NOMON in which all observation acceptance/rejection lists and bias corrections are removed; NOMONPLUS, which is identical to NOMON except an acceptance list of wind profilers is included; and AUTO, which is identical to CONTROL except that the radiosonde acceptance/correction list is produced by an automatic rather than the manual method. Verification of the forecasts from each run is done by comparison with observations and analysis including a calculation of the impact on the global NWP index.

For the NOMON run, results indicate that short-range MSLP forecasts in the Southern Hemisphere are degraded by an amount that is statistically significant at the 90% level. However, the mean skill of MSLP forecasts in other regions and temperature and wind forecasts away from the surface do not reduce in a statistically significant way. Global error maps indicate that the T+24 forecasts from the NOMON run are less skilful in regions where there appear to be isolated low quality reports (such as Antarctica) or groups of reports (such as India). The skill of forecasts of the fields that make up the NWP index are reduced mainly in the tropics and Southern Hemisphere.

A comparison of results from the NOMONPLUS and NOMON runs reveals the impact of the wind profiler acceptance list. Overall, the use of this list has no statistically significant impact at the 90% level on mean forecast scores but a small deterioration in the NWP index and regionally in the T+24 hour forecasts is observed. This result suggests that the method of assimilation of wind profiler data should be reviewed.

There is no statistically significant difference at the 90% level in the forecasts of most fields from the AUTO and CONTROL runs. A decline in the NWP index against analysis in the AUTO run is considered to be approximately within the ‘noise’ level for operational changes. Hence the operational introduction of the automated system for producing upper air station lists is reasonable based upon results from this experiment.

It is concluded that the main beneficial impact of the acceptance/rejection lists and bias corrections produced by the monitoring process is small but significant for some short-range forecasts in regions of the tropics and Southern Hemisphere where there tends to be poor quality observations that are isolated or clustered in small groups which the objective quality control scheme has difficulty flagging. The potentially more beneficial aspect of monitoring, which is to provide data producers with information that can enable them to correct poor quality observations at source, is not measured by this experiment.

1. Introduction

The Met Office routinely carries out monitoring of radiosonde, aircraft, SYNOP, buoy and ship observations in order to identify stations that frequently report observations of poor quality. For each observation the difference between its value and that obtained from a short (6-hour) model forecast at the same point is calculated. By calculating such observation minus background (O-B) values from each model run over a period of time, a mean O-B value, or bias, can be obtained.

If the bias is large and consistent then the reports from a station may be permanently rejected. The rejection can be done selectively by level and at specific times of the day. If the bias is considered to be fairly large but not large enough for the station's reports to be permanently rejected, then the initial error assigned to observations from the station may be raised to reflect the increased likelihood that the reports may be erroneous.

In the case of surface pressure reports that have a consistent bias, a correction may be applied to the observations to remove the bias.

Bias statistics are calculated from O-B values taken over a one-month period. By looking at biases calculated over the latest three months, a list of observations that are to be rejected or corrected over the coming month can be compiled. These 'station lists', which get updated monthly, can then be used by the NWP system to influence quality control decisions about observations.

Note that the monitoring system is not designed to detect infrequent, random errors in reports as such values are unlikely to influence the monthly-mean O-B statistics. Random errors are dealt with by using the NWP system's automatic quality control scheme. The monitoring system is intended to help the automatic quality scheme flag observations with consistently large errors.

Observation monitoring requires the development and maintenance of sophisticated software combined with a degree of human intervention to create new station lists and is thus quite a time-consuming and expensive activity.

One of the most time consuming manual parts of the monitoring system is the setting up of the upper air station lists. Thus a scheme to automate this process has been devised as described in the Appendix.

The aim of this study is to determine how much beneficial impact the monitoring effort has on Met Office NWP forecasts and to examine the effectiveness of the automated scheme for producing radiosonde station lists compared with the manual method. Section 2 describes the design of the experiment, section 3 gives some results that are discussed in more detail in section 4. In section 5 the overall conclusions from the study are stated with recommendations for the further use of the observation monitoring process.

2. Experimental set up

In order to measure the impact of monitoring, runs of the operational NWP system that included station lists containing all monitoring information were compared with runs using station lists containing no monitoring information. The choice of how to define a station list containing no monitoring information was not straightforward as the operational station lists contain a list of wind profiler and weather radar winds that are accepted for use with all other reports being permanently rejected. This acceptance list approach for wind profiler and weather radar wind reports contrasts to the rejection list approach used for all other observing systems. To assess the impact of the acceptance/rejection lists plus bias corrections and the production of the radiosonde station list by an automatic method, the following four runs were carried out:

(i) using all monitoring information (CONTROL)
(ii) using no monitoring information and no accept list for wind profilers/weather radar winds (NOMON)
(iii) as (ii) except including the accept list (NOMONPLUS)
(iv) as (i) except with the upper air station list set up using the automatic method (AUTO).

The NWP system used in the experiments was the global version that was in operational use at the time (October 2007). This included the 4D-Var data assimilation scheme with the forecast model run at approximately 40km horizontal resolution and 50 levels in the vertical.

For convenience, the period chosen for the study was 24th May 2007 to 24th June 2007. Forecasts out to 6-days were run from 12UTC on each day of the experiment. Objective verification of forecasts was done by comparing forecast values with radiosonde and surface observations, as well as ‘own run’ analyses. The observing stations used for verification were taken from a WMO approved list and their reports were required to pass the objective quality control checks before use. The mean verification statistics were calculated over all the forecasts that were run during the study period.

3. Results

3.1 Impact on the NWP index

The impact of the runs on the global NWP index is shown in Figure 1. The figures in the left hand column show the impact against observations and the right hand column against analyses.

Removing all the monitoring information, including the removal of wind profiler accept list (NOMON), caused a marginal reduction of 0.058 in the index against observations and a larger reduction of 1.527 in the index against analysis with most of the negative impact occurring in the tropics and Southern Hemisphere (Figures 1(a), 1(b)).

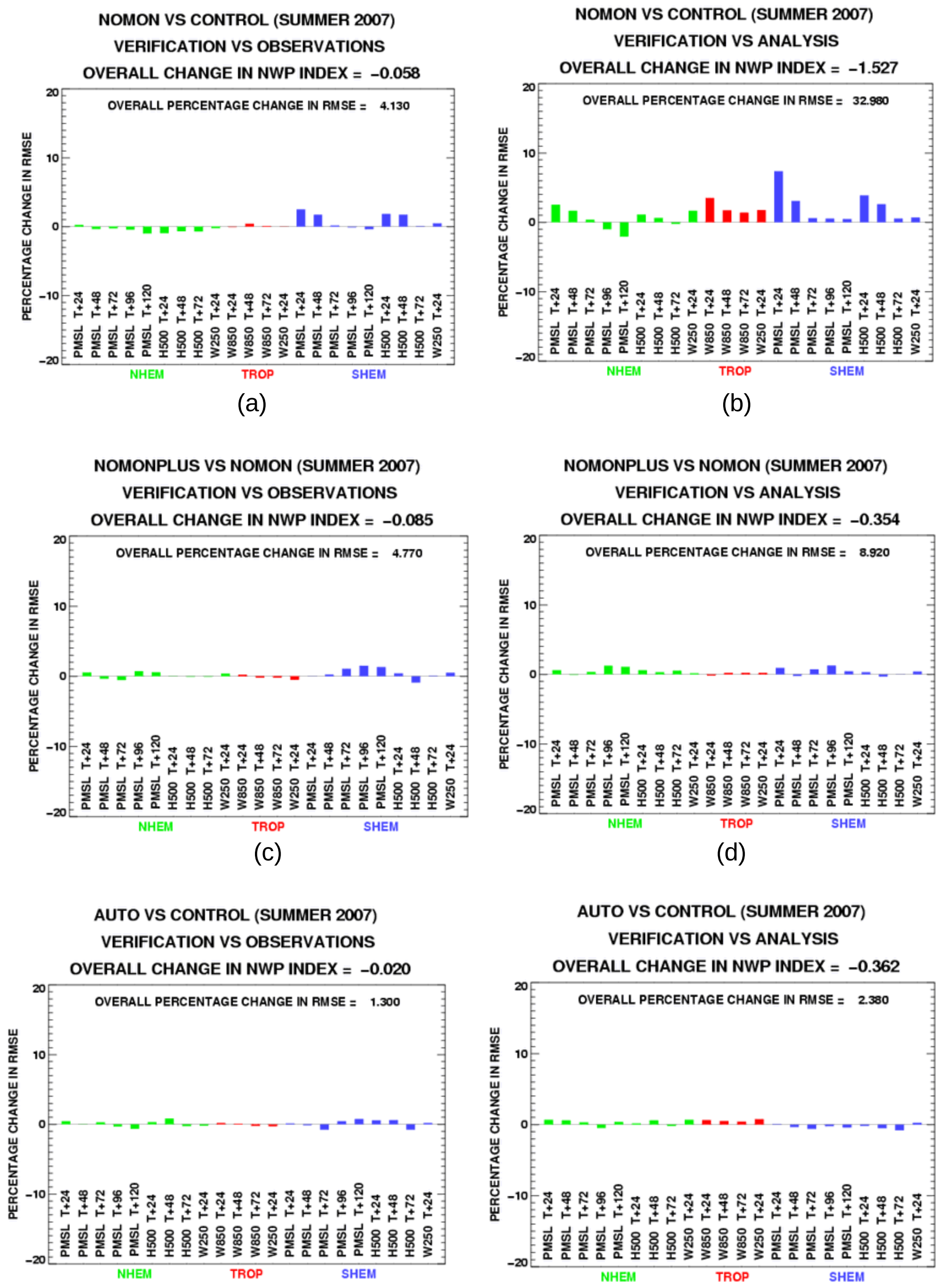
Figures 1(c) and 1(d) look at the effect of including the wind profiler acceptance list by comparing NOMONPLUS with NOMON. It can be seen that index against observations declines by 0.085 (Figure 1(c)) and the index against analysis by 0.354 (Figure 1(d)) and this negative effect from the wind profilers occurs throughout the globe. Comparing with Figures 1(a) and 1(b), it can be seen that the removal of wind profiler data has a positive impact verified against analyses and a slightly negative impact when verified against observations.

The effect of automating the production of the upper air station list is shown in Figures 1(e) and 1(f). There is a marginally negative impact of 0.020 against observations and a larger negative impact of 0.362 against analyses. These figures are approximately within the range that is taken to indicate that the impact of an operational change is ‘neutral’.

3.2 Geographical distribution of impact

The geographical distribution of the impact of monitoring is shown in Figure 2 which shows the difference in mean RMS error of the 24-hour forecasts of PMSL and 250 hPa temperature and wind speed. The figures in the left hand column compare the NOMON and CONTROL runs and those in the right hand column compare NOMONPLUS with NOMON. The 250 hPa level has been chosen to see the impact of the monitoring on Indian radiosondes which tend to have poor data quality and are often excluded. However, any differences in the region of India are affected by differences in the flow around the Himalayas, so at 250hPa these flow differences should be minimised.

In Figure 2(a), it can be seen that the PMSL errors in the NOMON run increase in the Southern Hemisphere mainly around Antarctica whereas the errors are similar elsewhere. The impact of adding in the wind profiler data is fairly neutral on the T+24 PMSL forecast (Figure 2(b)).



NOMON VS CONTROL (SUMMER 2007)
VERIFICATION VS OBSERVATIONS
OVERALL CHANGE IN NWP INDEX = −0.058

(a)

NOMON VS CONTROL (SUMMER 2007)
VERIFICATION VS ANALYSIS
OVERALL CHANGE IN NWP INDEX = −1.527

(b)

NOMONPLUS VS NOMON (SUMMER 2007)
VERIFICATION VS OBSERVATIONS
OVERALL CHANGE IN NWP INDEX = −0.085

(c)

NOMONPLUS VS NOMON (SUMMER 2007)
VERIFICATION VS ANALYSIS
OVERALL CHANGE IN NWP INDEX = −0.354

(d)

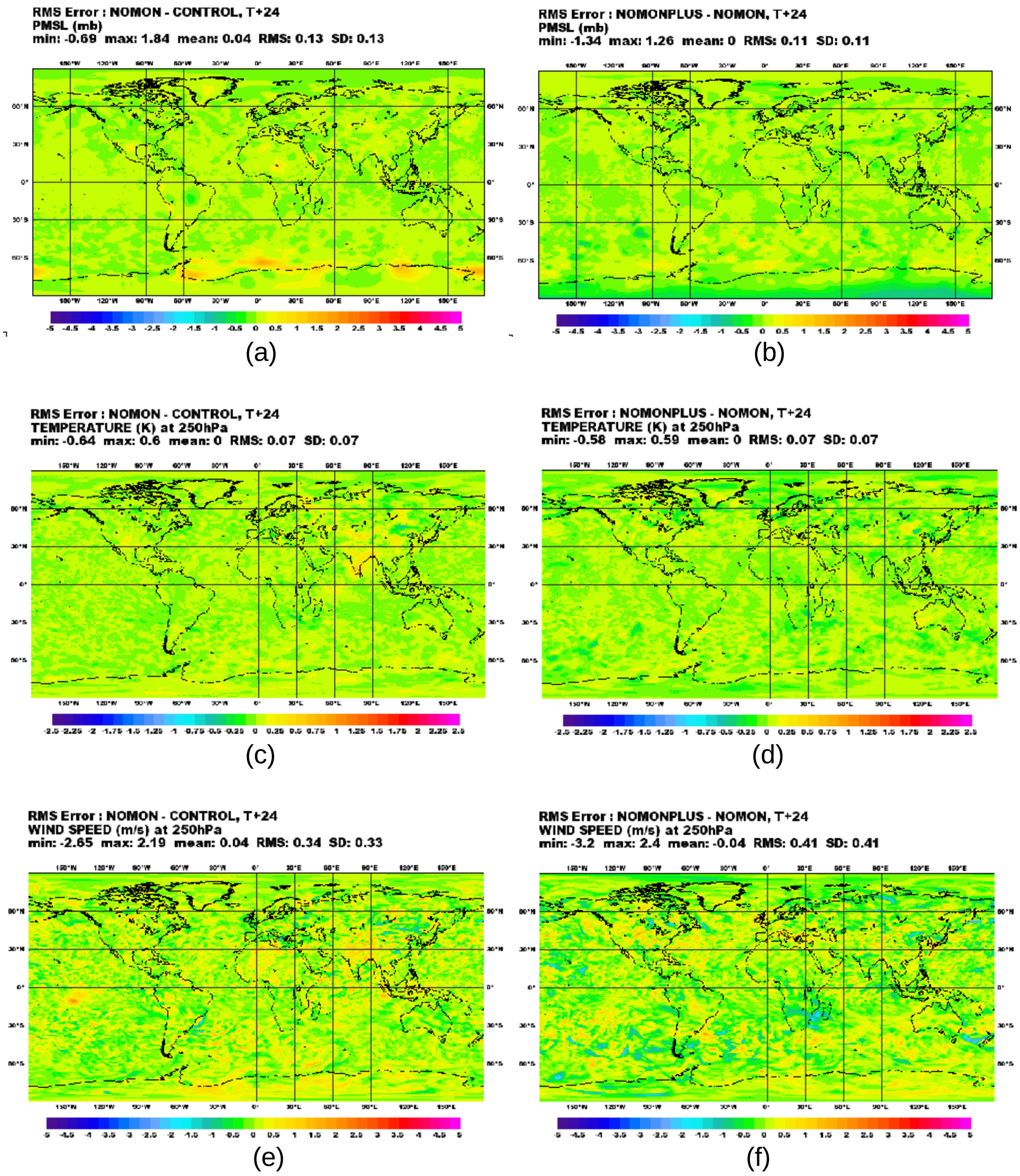
AUTO VS CONTROL (SUMMER 2007)
VERIFICATION VS OBSERVATIONS
OVERALL CHANGE IN NWP INDEX = −0.020

(e)

AUTO VS CONTROL (SUMMER 2007)
VERIFICATION VS ANALYSIS
OVERALL CHANGE IN NWP INDEX = −0.362

(f)

Figure 1. Impact of monitoring on the NWP index against observations and analysis.
(a) NOMON vs CONTROL (observations) (b) NOMON vs CONTROL (analysis)
(c) NOMONPLUS vs NOMON (observations) (d) NOMONPLUS vs NOMON (analysis) (e) AUTO vs CONTROL (observations) (f) AUTO vs CONTROL (analysis)


RMS Error : NOMON - CONTROL, T+24
PMSL (mb)
min: -0.69 max: 1.84 mean: 0.04 RMS: 0.13 SD: 0.13
RMS Error : NOMONPLUS - NOMON, T+24
PMSL (mb)
min: -1.34 max: 1.26 mean: 0 RMS: 0.11 SD: 0.11
-5 -4.5 -4 -3.5 -3 -2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
RMS Error : NOMON - CONTROL, T+24
TEMPERATURE (K) at 250hPa
min: -0.64 max: 0.6 mean: 0 RMS: 0.07 SD: 0.07
RMS Error : NOMONPLUS - NOMON, T+24
TEMPERATURE (K) at 250hPa
min: -0.58 max: 0.59 mean: 0 RMS: 0.07 SD: 0.07
-2.5 -2.25 -2 -1.75 -1.5 -1.25 -1 -0.75 -0.5 -0.25 0 0.25 0.5 0.75 1 1.25 1.5 1.75 2 2.25 2.5
RMS Error : NOMON - CONTROL, T+24
WIND SPEED (m/s) at 250hPa
min: -2.65 max: 2.19 mean: 0.04 RMS: 0.34 SD: 0.33
RMS Error : NOMONPLUS - NOMON, T+24
WIND SPEED (m/s) at 250hPa
min: -3.2 max: 2.4 mean: -0.04 RMS: 0.41 SD: 0.41
-5 -4.5 -4 -3.5 -3 -2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

(a) (b)

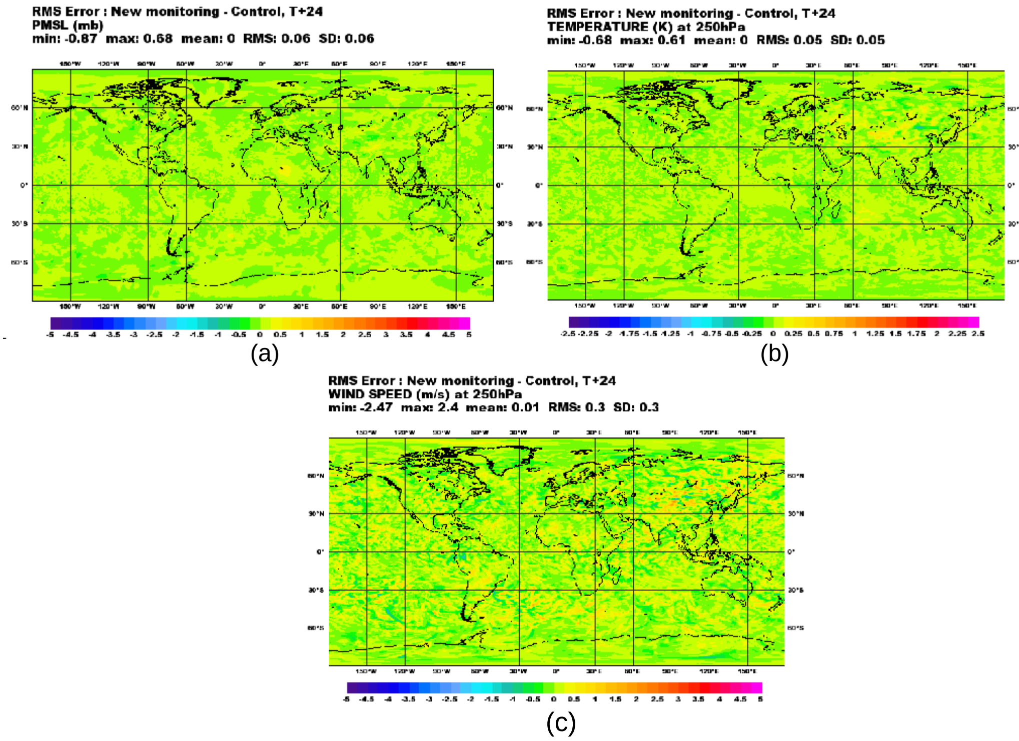
(c) (d)

(e) (f)

Figure 2. Impact of monitoring on 24-hr forecasts of PMSL, 250hPa temperature and 250hPa wind speed.

In the temperature field, an increase in RMS error over and around India appears to show a benefit from the use of the radiosonde rejection list whereas the impact is positive/neutral elsewhere (Figure 2(c)). The wind profiler data has little influence on the 250 hPa temperature field (Figure 2(d)).

In the wind speed field, the impact of monitoring is positive over India and at many other places such as the central Pacific where the benefit of monitoring island stations may be evident (Figure 2(e)). Some areas of negative impact can also be seen, for example, over Australia. Adding in the profilers generally increases the RMS errors in the 250 hPa wind speed (Figure 2(f)).


RMS Error : New monitoring - Control, T+24
PMSL (mb)
min: -0.87 max: 0.68 mean: 0 RMS: 0.06 SD: 0.06
RMS Error : New monitoring - Control, T+24
TEMPERATURE (K) at 250hPa
min: -0.68 max: 0.61 mean: 0 RMS: 0.05 SD: 0.05
RMS Error : New monitoring - Control, T+24
WIND SPEED (m/s) at 250hPa
min: -2.47 max: 2.4 mean: 0.01 RMS: 0.3 SD: 0.3

(a) (b)

(c)

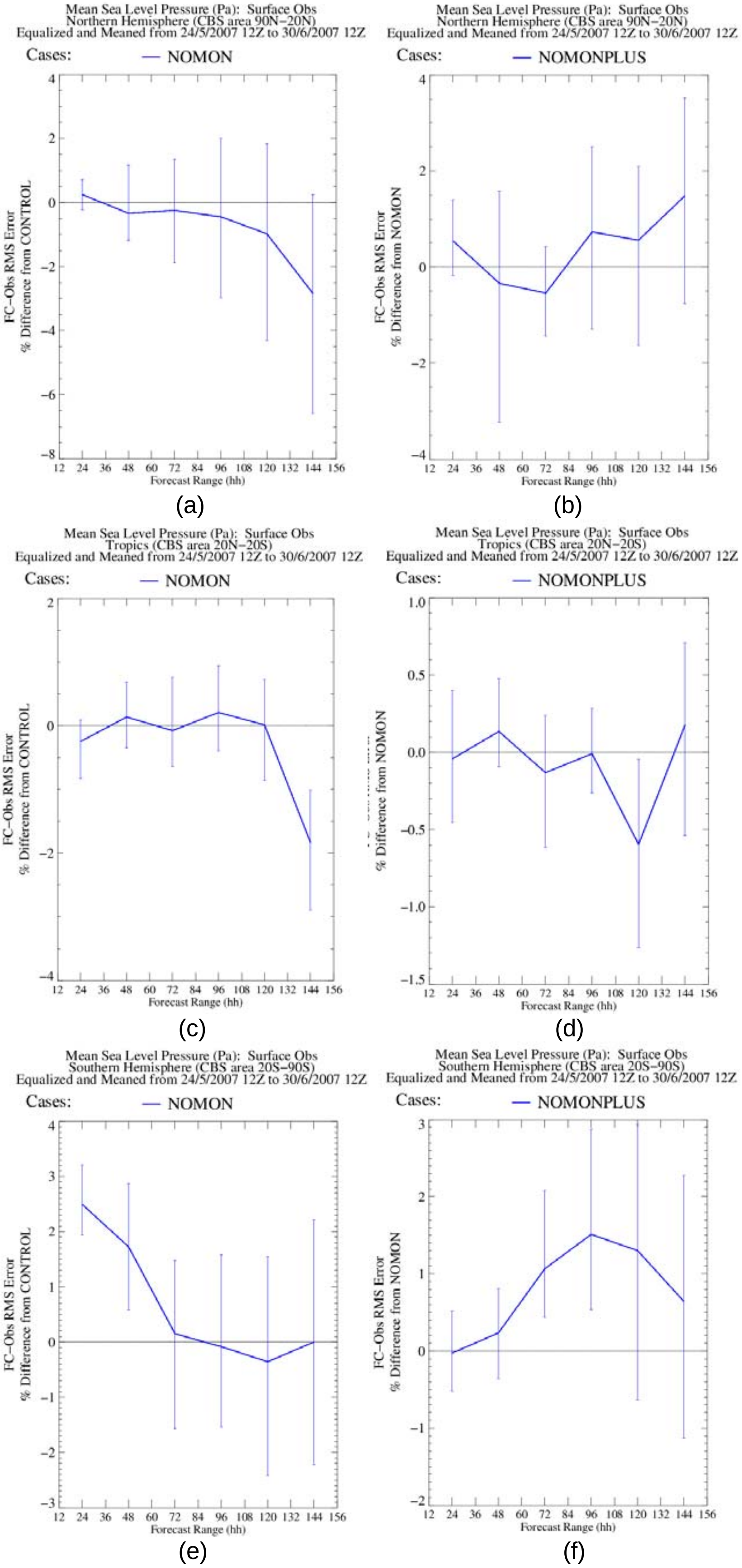
Figure 3. Automated monitoring system versus manual for 24-hr forecasts of (a) PMSL (b) 250 hPa temperature (c) 250 hPa wind speed.

In Figure 3 the impact of automating the production of the radiosonde station list is shown. This process has had an overall neutral impact on the T+24 forecast of MSLP (Figure 3(a)). A slightly negative effect can be seen on 250 hPa temperature, particularly in regions such as central Asia (Figure 3(b)), and a greater negative impact on the 250 hPa wind speed (Figure 3(c)). It should be noted that such negative impacts are not clearly seen in verification against observations (see section 3.3).

3.3 Mean impact against observations

In Figure 4 the mean impact of monitoring on PMSL forecasts for all ranges up to T+144 is shown where the verification is against surface observations. The error bars on the plots denote 90% statistical significance so when the bars fail to cross the zero line then the differences are statistically significant at the 90% level. The averages have been calculated over all forecasts throughout the trial and for the three geographical regions the Northern Hemisphere (90N – 20N), tropics (20N – 20S) and Southern Hemisphere (20S – 90S).

Figures 4(a), 4(c), 4(e) show the impact in the Northern Hemisphere, tropics and Southern Hemisphere respectively of the NOMON run. In the Southern



(a)

(b)

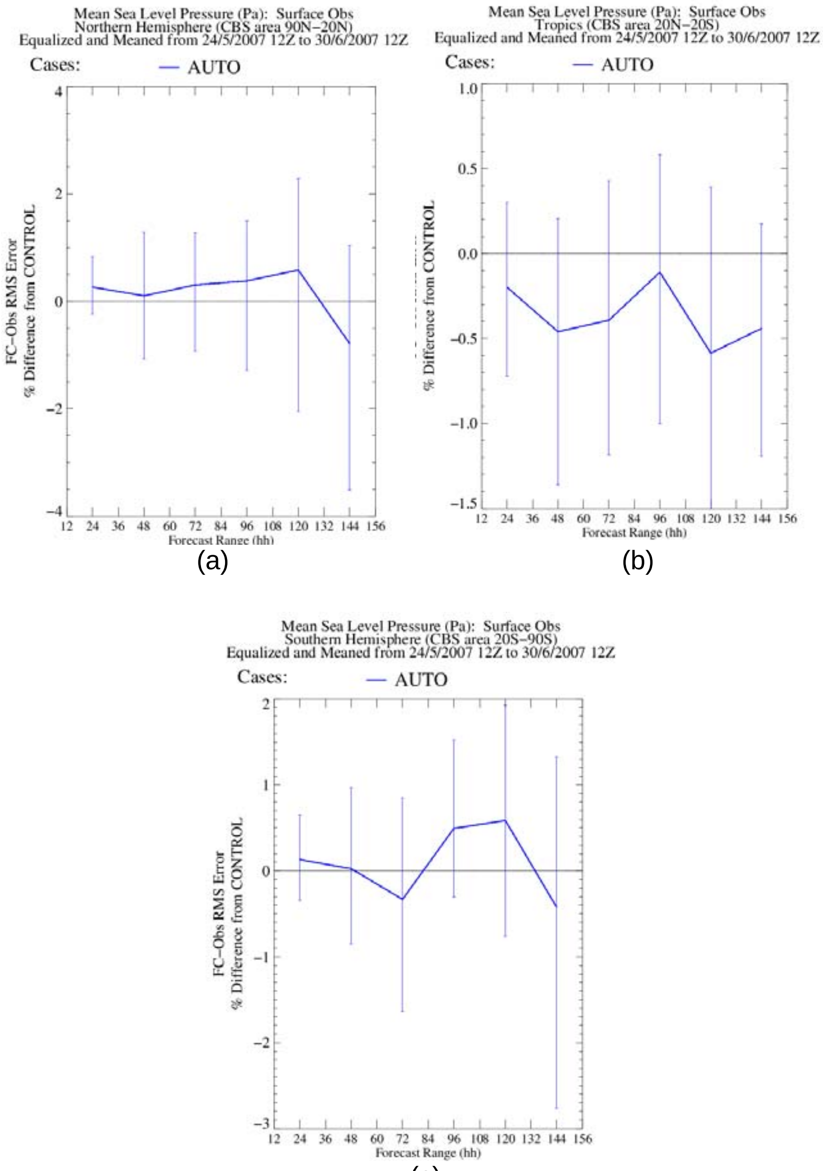
(c)

(d)

(e)

(f)

Figure 4. Impact of monitoring on MSLP. Difference in mean RMS errors for the Northern Hemisphere, tropics and Southern Hemisphere. Error bars denote 90% statistical significance.



(a) (b)

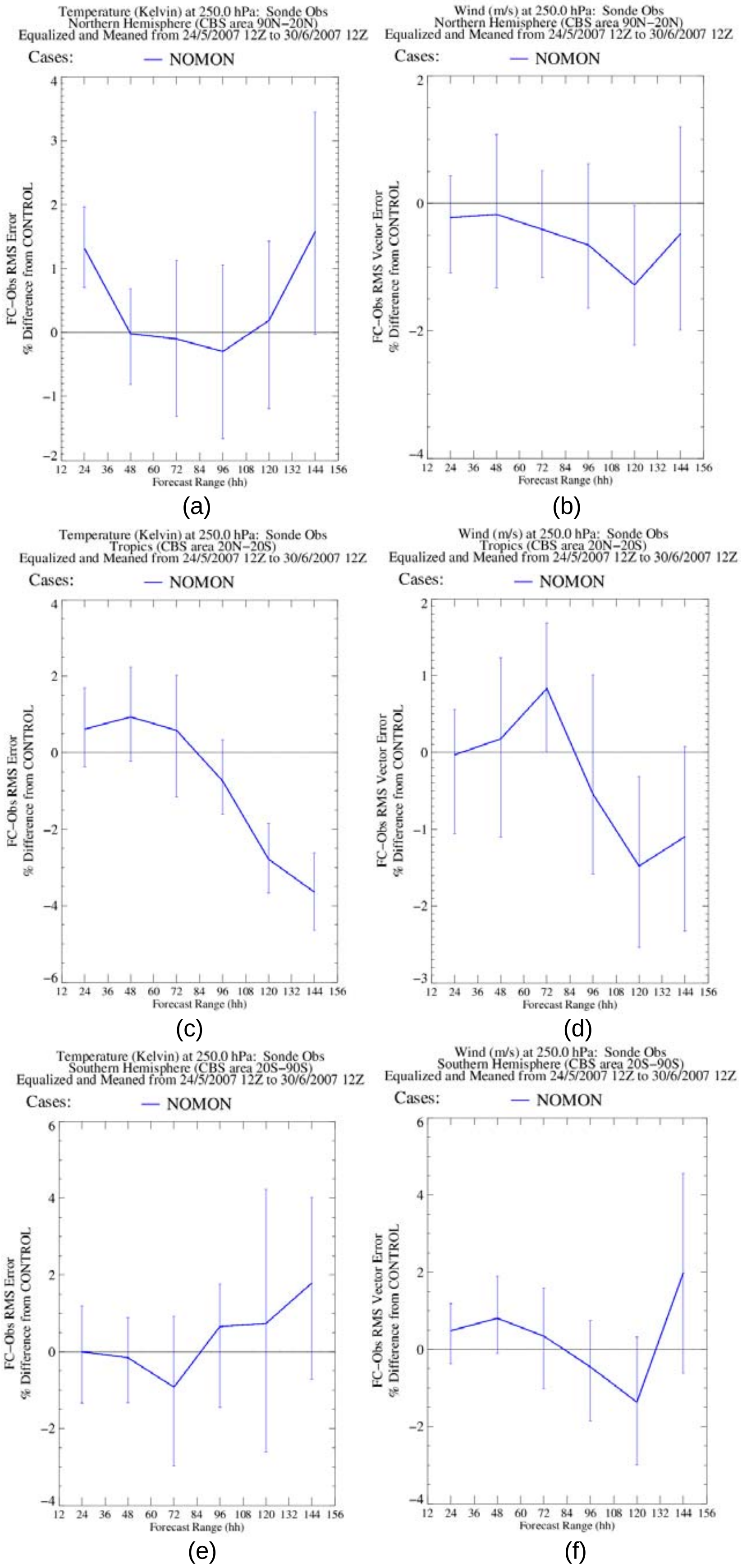
(c)

Figure 5. Impact of the automated monitoring system on MSLP. Difference in mean RMS errors for the Northern Hemisphere, tropics and Southern Hemisphere. Error bars denote 90% statistical significance.

Hemisphere, there is a statistically significant positive impact on PMSL at forecast ranges up to T+48. In the Northern Hemisphere there is no statistically significant positive impact at any forecast range whilst a small negative impact is seen at T+144 in the tropics.

The impact of including the wind profiler observations is shown in Figures 4(b), 4(d) and 4(f) which indicate that the data do not have a statistically significant impact, at the 90% level, on the PMSL forecast at almost all ranges in the Northern Hemisphere and tropics, but a slight negative impact at T+72 and T+96 in the Southern Hemisphere.

In Figure 5 it can be seen that the effect of automating the production of the upper air station list makes no statistically significant difference in all forecast ranges in all geographical areas.

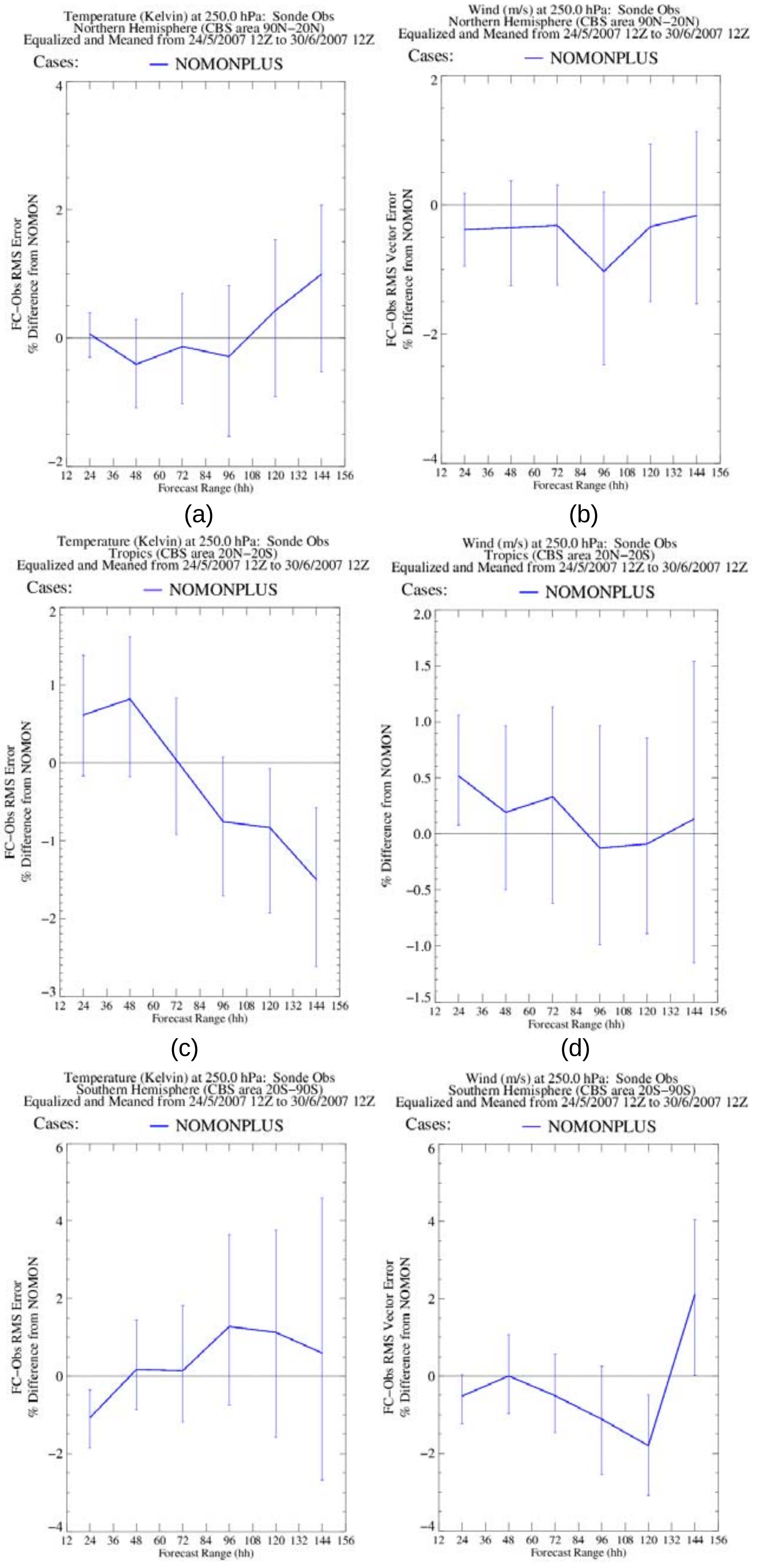


(a) (b)

(c) (d)

(e) (f)

Figure 6. Difference in mean RMS errors, between NOMON and CONTROL, for forecasts of temperature and wind for the Northern Hemisphere, tropics and Southern Hemisphere. Error bars denote 90% statistical significance.



(a) (b)

(c) (d)

(e) (f)

Figure 7. Difference in mean RMS errors, between NOMONPLUS and NOMON, for forecasts of temperature and wind for the Northern Hemisphere, tropics and Southern Hemisphere. Error bars denote 90% statistical significance.

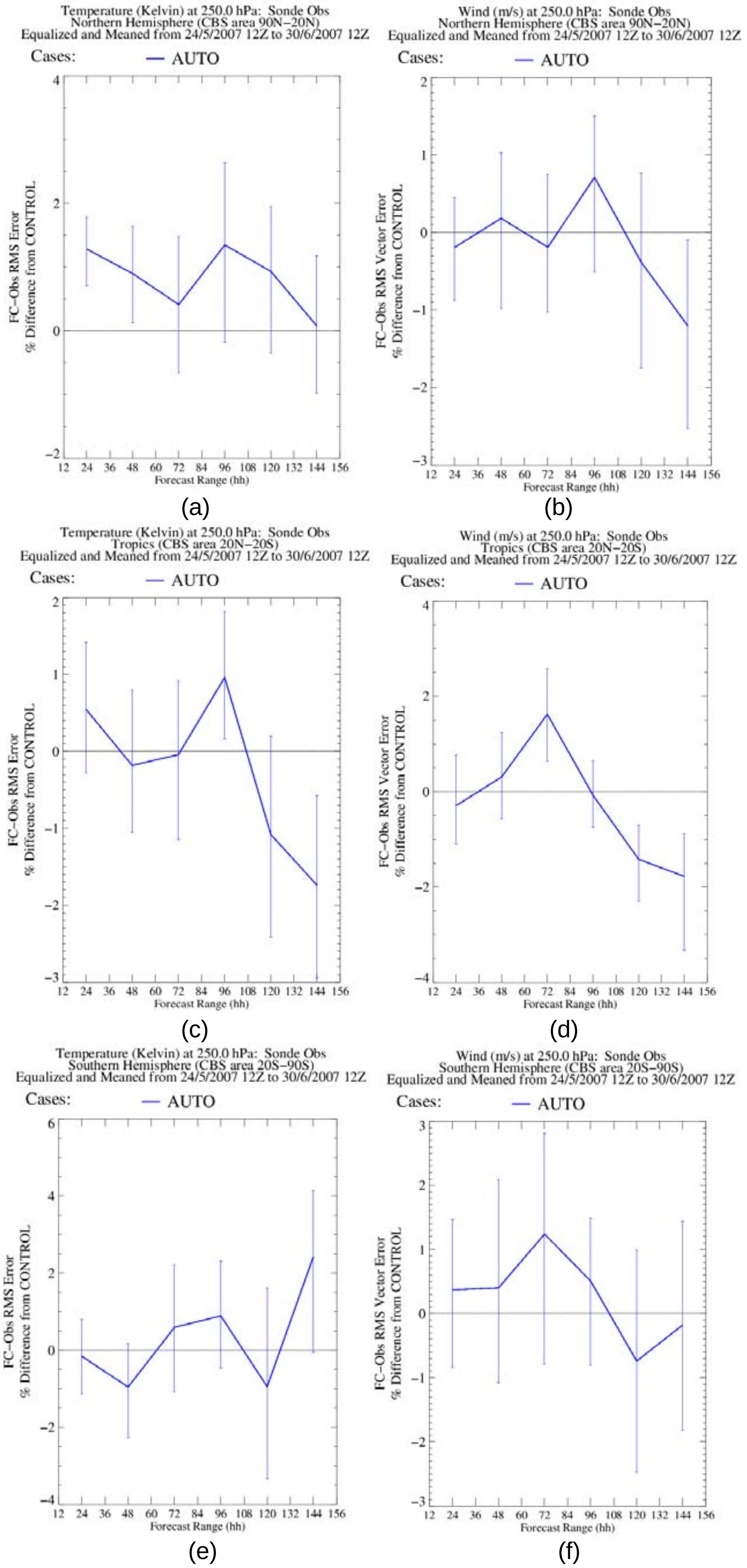


Figure 8. Difference in mean RMS errors, between AUTO and CONTROL, for forecasts of temperature and wind for the Northern Hemisphere, tropics and Southern Hemisphere. Error bars denote 90% statistical significance.

In line with the geographical distribution maps, the impact of the different runs on the forecasts of temperature and wind at 250 hPa are shown in Figures 6, 7 and 8. Here, as in Figures 4 and 5, differences are plotted with error bars denoting 90% statistical significance.

In Figure 6 the difference between the NOMON and CONTROL runs is shown. For both temperature and wind forecasts, there is no statistically significant difference between the forecasts at most ranges in most areas. The effect of the NOMON run over India, shown in Figure 2(c), extends from within the tropics to Northern Hemisphere. A statistically significant difference can be seen in the RMS errors averaged over the tropics for longer forecast ranges (Figures 6(c) and 6(d)) but not the Northern Hemisphere (Figures 6(a) and 6(b)).

In Figure 7 the difference in mean RMS values between the NOMONPLUS and NOMON runs is plotted thus showing the impact of the wind profiler acceptance list. In all regions at most forecast ranges the wind profiler data make no significant difference to the mean scores although a small positive impact can be seen T+120 and T+144 in the forecasts of tropical temperature (Figure 7(c)) and at T+120 in the Southern Hemisphere wind field (Figure 7(f)).

In Figure 8 the (AUTO-CONTROL) differences in the mean scores are plotted. Positive values show a degradation in the AUTO run and negative values an improvement. Once again, there is no statistically significant difference between most of the values from the two runs. One exception is for forecasts in the tropics where there is a slight improvement in forecasts from the AUTO run at T+120 for wind and T+144 for temperature and wind (Figures 8(c), 8(d)).

4. Discussion

Any conclusions from this experiment have to be tentative given that only a limited number (about 30) forecasts have been assessed. Different results may have been obtained if the experiment had been run for longer or if a different period had been chosen as the impact of the runs on a greater sample of weather regimes would have been assessed. Furthermore, all forecasts were run from 12UTC so to investigate the possibility of diurnal variation in the results, it would be preferable to run forecasts from 00UTC as well.

Most of the positive impact from monitoring can be seen in regions of the southern hemisphere and tropics where poor quality reports appear to occur in isolation (such as around Antarctica) or in close groups (such as in India). For such situations, the operational objective quality control scheme is less likely to be effective as there are insufficient good observational data to flag poor quality data via the ‘buddy’ check. Furthermore, consistently poor quality isolated or grouped reports are likely to be consistent with the background field which may have no alternative observational information.

The lack of statistically significant difference in the forecast scores from the NOMON and CONTROL runs may be due to the effect of averaging the scores around entire latitude bands that tends to mask the effect of regional variations within the bands. The NOMONPLUS versus NOMON comparison indicates that the wind profiler data do not have a statistically significant benefit on forecasts in general suggesting that their assimilation needs to be improved.

A greater impact from monitoring might be obtained if the methods by which the rejection/acceptance lists and bias corrections are calculated and applied were improved. For example, surface pressure corrections are calculated and applied on a monthly bias by which time they may be out of date or not needed at all if the data producer has already eliminated the bias. Calculating and applying the bias corrections more frequently, such as weekly, might get around this problem.

Automating the production of the upper air station list appears to have had an overall neutral impact, noting that the ensuing reduction in the NWP index against analysis was by an amount that would be considered to be approximately within the ‘noise’ range for an operational change. Given the time saving gained by the automation, introducing the method into operations is justified.

One of the most important parts of the observation monitoring process is to give feedback on data quality to observation producers enabling them to correct observation errors at source. This process has the potential to have a very positive benefit on NWP forecasts by reducing analysis errors, but its impact is not measured by this experiment. It could therefore be argued, that observation monitoring is an important process even if it were not used as a means of providing rejection/acceptance lists and bias corrections for operational use.

5. Conclusions and recommendations

The main conclusions from this study are:

- Monitoring has a small positive impact on forecast skill mainly in the Southern Hemisphere but a statistically significant impact is difficult to detect in scores averaged over an entire latitude band.
- The positive impact is likely to be caused by an improvement in the quality control of low quality reports that occur in isolation or small groups. It is likely that reports in other areas can usually be dealt with by the automatic quality control procedures such as the buddy check.
- Wind profilers do not have a clear statistically significant positive impact on mean forecast skill.
- An automatic method for generating the radiosonde station list has no significant detrimental effect on forecast scores and so is preferable to the manual method as it takes less time and is more objective and consistent.

It is recommended that:

- The operational use of rejection/acceptance lists and bias corrections is continued and the procedures for producing them continue to be made as efficient as possible.
- Improvements to the system, such as the automatic weekly updating of the rejections and bias corrections, should be considered.
- The assimilation of wind profiler data should be improved, possibly by the adjustment of the observation error profiles particularly for European profilers.

APPENDIX

Automatic method for producing the upper air station list

For each station, monthly O-B profiles are obtained and each level checked against set criteria; those report levels exceeding the criteria are labelled as suspect. For temperature and relative humidity both mean and RMS O-B are checked, whereas for wind there are checks on speed and direction bias and RMS vector wind O-B values.

The criteria for including a report level as suspect vary in the vertical, with generally larger limits in the boundary layer and near jet levels, where the background and representivity error tends to be larger. The limits for individual stations' monthly mean and RMS O-B values have been made proportional to the standard deviation of O-B for all stations combined (SD). These SD statistics were averaged over a whole year of data (Oct07-Sep08). The SD values were multiplied by a factor of 1.3 to obtain the limits for the mean and multiplied by a factor of 1.8 - 2.0 to obtain the limits for the RMS, with some additional tuning using comparisons with the manual system. The O-B limits are increased for stations with fewer than 15 reports (e.g. for stations with only 5 reports the limits are increased by 20%), since these statistics are less reliable; and also increased by 15% for African, south Asian and all southern hemisphere stations, to allow for larger background errors in these areas where the data density is lower. The numerical values used in the process were obtained by empirical tuning of the automatic system so that it produced similar results to those obtained by the manual method.

If all suspect levels for each station were rejected, then there would be many (possibly most) stations with some levels rejected, leaving many 'holes' in their assimilated data profiles, and the holes would probably change from month to month. To remove this undesirable feature the suspect levels are combined to form a smoother, filtered final rejection list. To this end the report levels are combined into 13 pressure layers (with boundaries at 1050, 950, 850, 700, 600, 500, 400, 300, 200, 100, 50, 25, 10, and 0.1 hPa), which are similar to the layers previously used in the manual rejection procedure.

To provide some smoothing in time, the procedure to detect suspect levels is run a second time, using somewhat lower limits (currently reduced by 20%), to find 'nearly-suspect' stations/levels. These ‘nearly-suspect’ levels are included in the suspect list if they (or adjacent levels) were suspect in the previous month. This process should help prevent stations/levels with O-B values close to the suspect limit going on and off the reject lists too frequently.

If the percentage of suspect report levels in a pressure layer is less than 25% the layer is not labelled as suspect, unless the layers above or below are labelled as suspect.

Having obtained sets of suspect pressure layers, the choice of which layers to reject is made as follows. If 5 or more of the 7 pressure layers between 850hPa and 100hPa are suspect, the whole report profile is rejected. Otherwise, where there are several adjacent suspect layers, but much of the profile is not suspect, only part of the station's profile is rejected. The partial profile rejects are split into three: rejection of all layers below a pressure layer boundary ('lower layer rejects'), rejection of all layers above a pressure layer boundary ('upper layer rejects') and rejection of all layers between two pressure boundaries ('mid layer rejects'). Any layer in between two suspect layers is also labelled as suspect and forms a continuous rejected layer.

Where there is a single suspect pressure layer, not at the top or bottom of the model, it is usually ignored, unless there are nearby suspect layers (possibly at different times) when it will form part of a multi-layer rejection. If the layer next to the top/bottom layer is suspect then both that and the top/bottom layer are rejected. If there are 3 or more single suspect layers, not top or bottom layers, the whole profile is rejected.

All stations that have at least one layer rejected also have their initial probability of gross error (PGE) increased, to reflect their increased risk of having errors in other layers.

As there are few sonde reports at 06Z and 18Z, and all 06Z/18Z statistics are included with the 00Z/12Z statistics, if stations are included on the reject list for 00Z/12Z they will also be rejected at 06Z/18Z.