Evaluation of short-term weather forecasts in South Africa

Estelle Banitz
South African Weather Service, Private Bag X097, Pretoria 0001, South Africa

Abstract

In this paper a brief overview will be given for the reasons for doing evaluations of short-term weather forecasts as well as the methodology thereof. Short-term weather forecasts are defined as a forecast valid for the current day as well as the next day. In other words up to 48 h ahead. Results are given for South African Weather Service temperature, rainfall and severe weather forecasts as issued by head office in Pretoria. Temperature forecasts generally tend to be accurate to within a limit of 2.3°C. A comparison is made between temperature forecasts for an inland station, a coastal station and a station influenced by the escarpment. Tendencies of rainfall forecasts show that rain is forecast more often than it occurs. Comparative rainfall forecasts for a summer and winter rainfall region are shown. Severe weather events are sometimes captured well, but severe thunderstorms are not predicted with great accuracy. Once again the tendency is to over-forecast. With one of the scientific aims of forecasting evaluations being to concentrate on areas of under-performance, these statistics show that a better observation network would improve conditions for evaluation of forecasts. Further research should be focused on alternative or better techniques to forecast precipitation (general and severe)with greater accuracy.

Introduction

Weather forecasts are important in our everyday lives for planning of various activities. It is important to know what the weather forecast is in order to plan our day. However, no weather forecast has much value if one cannot rely on the information. The next question then is: How accurately can we forecast the weather in South Africa? The South African Weather Service is in the process of commercialisation and the accuracy of the forecasts will become more important to a client who will have to pay for the service in the near future. Aside from financial motivation to do evaluations, objective evaluation of weather forecast quality is done for a variety of reasons. Brier and Allen (1951) categorised these as serving administrative, scientific and economic purposes:

- **Administrative**
  Comparing the reality with a forecast should be part of the procedure in every forecasting office around the country. If long-term trends of evaluation at different stations are kept, it should be easy to see if a station's performance is improving or deteriorating. If the forecasts from a specific station appear to be below the standards of accuracy previously attained, one needs to investigate the reasons for the dropping of standards. The mere existence of a checking scheme - however simple and imperfect - tends to keep the forecasters more alert and interested in maintaining and improving the accuracy of forecasts.
- **Scientific**
  Together with the increase in the understanding of the physical processes of the atmosphere, one would expect more accurate forecasts. Evaluation statistics can be used to monitor the trend in forecast accuracy. Another scientific purpose is to investigate the forecast errors to determine their nature and cause. It can serve to identify the synoptic conditions under which forecasts are most likely to be wrong or when numerical weather prediction models are not capturing certain weather phenomena adequately. This knowledge can then be used to discover the weaknesses of forecasting systems in order to decide where research emphasis is needed. Analysis of verification statistics can also help in the assessment of specific strengths and weaknesses of forecasters or forecasting systems (e.g. numerical weather prediction models). Forecasters should be given feedback on the performance of their forecasts in different situations that will hopefully lead to better forecasts in the future.
- **Economic**
  The uses and users of forecasts are so diverse that it becomes problematic to determine the economic value of a forecast. In this case, the reliability of weather forecasts can be measured by their approach to the truth and expressing the result in terms of degrees Celsius or percentage of hits.

Ultimately the justification for any forecasting enterprise is that it supports better decision-making (Wilks, 1995).

Pitfalls of verification

The purpose of verification should not be to create negative competition between forecasters of forecasting offices. It should be used as a positive measure to inspire forecasters to better accuracy. According to Brier and Allen (1951) one of the greatest dangers lies in attempts to compare the relative abilities of forecasters on the basis of forecasts which are not comparable because of differences in location, season and time of day. The degree of forecasting difficulty varies so much from one forecasting circumstance to the next that a very large sample of forecasts is needed to ensure that the average weather has been approximately the same in the two sets of forecasts being compared. Even if the forecasts being compared are for the same event, there may be other factors to be considered such as whether or not equal map facilities were available to each forecaster.

Brier and Allen (1951) also mentioned that it should be decided ahead of time what measures of accuracy are needed. If the tolerances are set too wide, the verification will fail to discriminate

☎(012) 309-3081;fax (012) 323-4518 e-mail: estelle@weathersa.co.za
Received 13 October 2000; accepted in revised form 28 May 2001.

between better and poorer forecasts. If the tolerances are too narrow, the forecasters will feel that they can never reach the desired accuracy.

Verification methods and scores

Forecast verification is perhaps easiest to understand with reference to categorical forecasts of discrete predictands (Brier and Allen, 1951).

Categorical means that the forecast consists of a flat statement that one and ONLY one of a set of possible events will occur. Categorical forecasts contain no expression of uncertainty in distinction to probabilistic forecasts.

A discrete predictand is an observable variable that takes on one and only one of a finite set of possible values.

When forecasts are made in categorical classes, a useful summary of forecast and observed weather can be presented in the form of a contingency table. Such a table provides the basis from which a number of useful pertinent scores or indices can easily be obtained. Conventionally, categorical verification data are displayed in an I x J contingency table of absolute frequencies, or counts, of the I X J possible combinations of forecast and event pairs. Perfectly accurate forecasts in the 2 X 2 categorical forecasting situation will clearly exhibit B = C = 0 with all Yes-forecasts for the event followed by the event and all No-forecasts for the event followed by a non-occurrence (Wilks, 1995).

Obs. vs forecast	Yes-Forecast	No-Forecast	Total
Observed	A	B	A+B
Not Observed	C	D	C+D
Total	A+C	B+D	A+B+C+D

Heidke skill score

The information contained in the contingency table is often combined into a single index called a skill score (Heidke, 1926). It is defined by:

$$SS = \frac{A - E}{N - E}$$

where E is the number of (yes and no) forecasts expected to be correct, based on some standard such as chance, persistence or climatology; A is the number of correct (yes and no) forecasts and N is the total number of forecasts. This score has a value of one when all forecasts are correct and a value of zero when the number correct is equal to the expected number correct (Brier and Allen, 1951).

Hit rate

Ratio test (hit rate, or HR) is defined (Noone and Stern, 1995) as the ratio of the total number of correctly forecast events and the non-events to the total number of forecasts:

$$R = \frac{A + D}{N}$$

A perfect forecast system would yield R=1 and for a system that was always wrong R=0. The HR satisfies the principle of equivalence of events, since it credits correct Yes and No forecasts equally. This is, however, not always a desirable attitude. The HR also penalises both kinds of errors equally. Sometimes HR is multiplied by 100 and referred to as the percentage correct, or the percentage of forecasts correct (PFC) (Wilks, 1995).

Bias

Bias (Fraedrich and Leslie, 1988) is a measure of the predictive scheme's climate vs. the observed climate:

$$Bias = \frac{A + C}{A + B}$$

Persistence, of course, shows no bias which also holds for forecasters' categorical predictions. Unbiased forecasts will exhibit Bias = 1, indicating that the event was forecast the same number of times that it was observed. Bias greater than one indicates that the event was forecast more often than observed, which is called over-forecasting. Conversely, bias less than one indicates that the event was forecast less often than observed, or was under-forecast (Wilks, 1995).

Probability of detection

Probability of detection (POD), which is referred to in the older literature as prefigurance, is another possible score to calculate. The POD (Wilks, 1995) is simply the fraction of those occasions when the forecast event occurred on which it was also forecast. That is, the POD is the likelihood that the event will be forecast, given that it occurred:

$$POD = \frac{A}{A + B}$$

The POD for a perfect forecast is one and the worst POD is zero.

False alarm rate

The false alarm rate (FAR) is that proportion of forecast events that fails to materialise (Wilks, 1995):

$$FAR = \frac{C}{A + C}$$

The FAR has a negative orientation, so that the smaller values of FAR are to be preferred. The best possible FAR is zero and the worst FAR is one.

Data used in this study

Temperature

Minimum and maximum temperatures are forecast twice daily, early in the morning (the AM forecast) and in the late afternoon (the PM forecast). The AM forecast is valid for the remainder of the day, while the PM forecast is valid for the next day. Temperatures are forecast for more than 40 stations around the country as shown in Fig. 1. The minimum and maximum temperatures are observed at 08:00 (SA time) every morning. This observation gives the minimum temperature which occurred earlier that morning (from the minimum thermometer) and the maximum temperature which was recorded the previous day (from the maximum thermometer). Temperature evaluation has been performed since 1992. Evaluation statistics (in the form of absolute errors) to February 2001 are included in this study.

Rainfall

In the late afternoon the Central Forecasting Office in Pretoria compiles a rainfall forecast in the form of a map valid for the following day. This is the same map that is shown on television in the evening broadcast. Observed rainfall is a 24-h precipitation total from 08:00 (SA time) until 08:00 the next morning. There is thus a discrepancy in the time for which rainfall is forecast (i.e. midnight to midnight) and the time when it is observed (i.e. 08:00

until 08:00). Rainfall is evaluated by means of rainfall reports from 1 702 rainfall reporting stations across the country divided into 19 geographical regions (Fig. 2) in South Africa. If rain was reported at any one of the rainfall stations in such an area, and it was forecast that there would be rain, then the forecast is considered to be correct. Rainfall is evaluated by means of a Yes/No contingency table. This kind of evaluation has been done since 1998. Unfortunately, the data for July 1999 is missing, but otherwise the evaluation figures are given up to February 2001.

Severe weather

Severe weather warnings can be divided into several different parameters: heavy rain, extreme cold, fire index, gale force winds, extreme heat, snow, high seas, severe thunderstorms, sand storms, discomfort index and heat waves. These parameters are forecast three times per day, twice for the current day and once for the following day, and consequently, evaluation thereof occurs thrice daily.

Figure 3 shows a pie chart of the percentage of total warning in each of these categories issued from January 1999 to the end of February 2001. A total of 1 701 warnings were issued in this period of which most were for fire danger (26.7%), gale-force winds(20.5%), heavy rain (16.7%) and discomfort index (15.1%). Severe weather is evaluated for the country as a whole, i.e. if reports of severe weather come in from anywhere in the country they are validated against the severe weather warnings which were issued for different regions. The severe weather network is extremely inadequate and most of the time one has to rely on reports of such events from the media or members of the public. Severe weather events have only been evaluated since 1999, which makes it a relatively small data set.

Results

Temperature

Temperature evaluation over the past nine years shows a definite seasonal trend (Figs. 4 and 5). Minimum temperatures are more difficult to forecast correctly in winter (greater absolute error), while maximum temperatures are more difficult to forecast during the summer months (greater absolute error).

Minimum temperatures forecast in the early morning (AM forecast) for the current day are closer to reality than minimum temperature forecasts in the late afternoon valid for the following day (Figs. 4a and 4b). The absolute error of the AM forecast is mostly less than 1.75°C while the PM forecast remains within a 2.3°C range. This can be expected due to the shorter lead time of AM forecasts.

Maximum temperature forecasts in the early morning show a departure of generally less than 2.25°C, while the absolute error (for maximum temperatures) increases to around 2.3°C with the afternoon forecast (Figs. 5a and 5b).

The trend in the absolute error for the past nine years, is upward for the early morning minimum temperature forecast (absolute errors are increasing), while for the other forecasts, the errors are relatively stable with a slight tendency to lower values in the past few years.

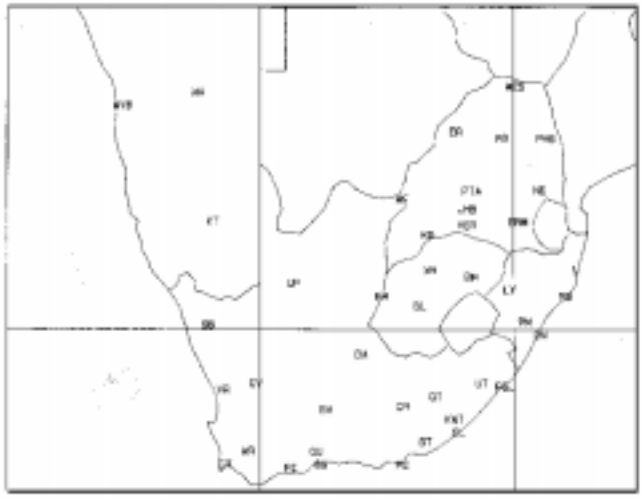

Figure 1
Map of South Africa showing the stations for which a temperature forecast is performed every day

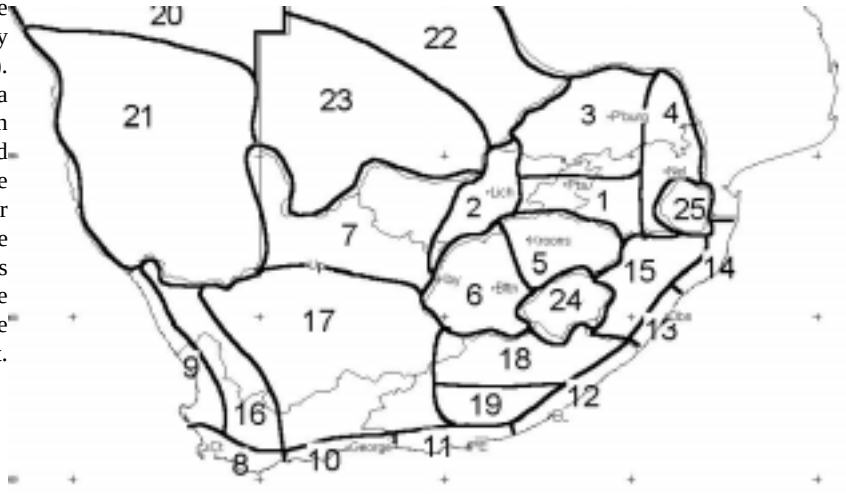

Figure 2
Map of South Africa showing the 19 rainfall regions where rainfall forecasts are available

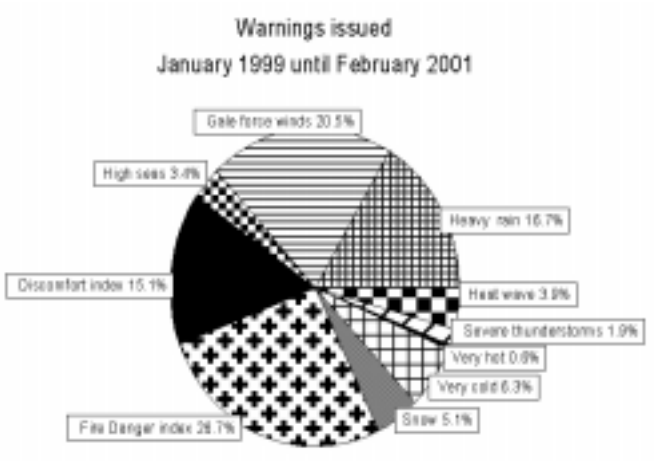

Figure 3
Pie chart of the percentage warnings issued for severe weather for the different parameters in the time period January 1999 to February 2001

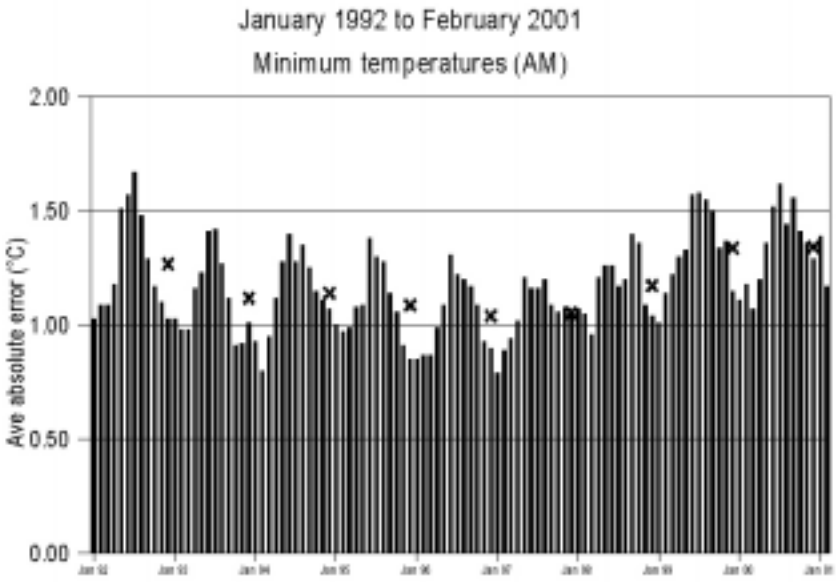

Figure 4a
Absolute error for minimum temperature from early morning (AM) forecast, valid for the current day for January 1992 to February 2001 as a monthly average of all the forecasting stations. The 'x' indicates the annual average for a year.

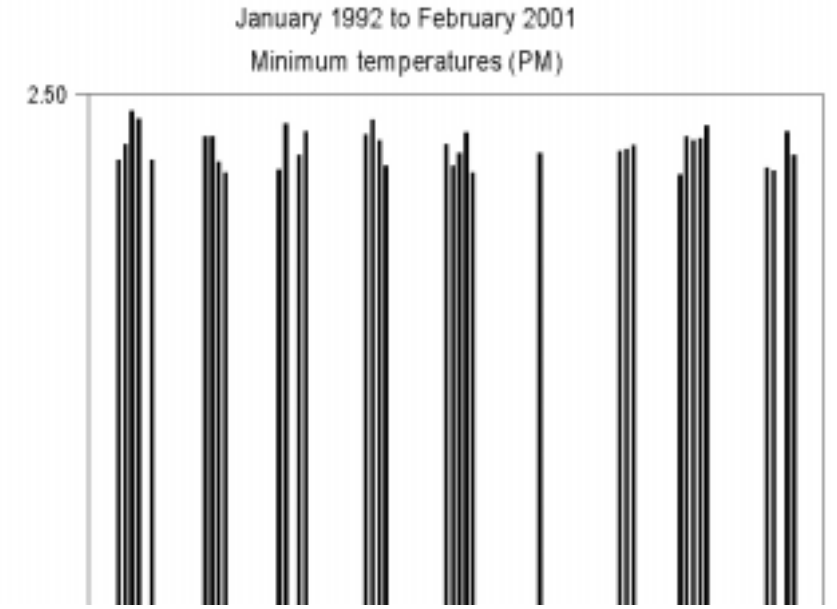

Figure 4b
Absolute error for minimum temperature from late afternoon (PM) forecast, valid for the following day for January 1992 to February 2001 as a monthly average of all the forecasting stations. The 'x' indicates the annual average for a year

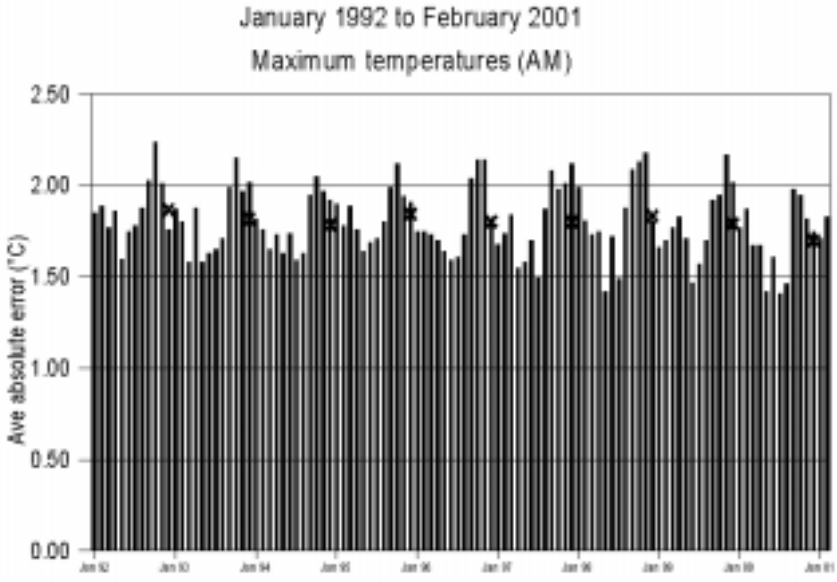

Figure 5a
Absolute error for maximum temperature from early morning (AM) forecast, valid for the current day for January 1992 to February 2001 as a monthly average of all the forecasting stations. The 'x' indicates the annual average for the year.

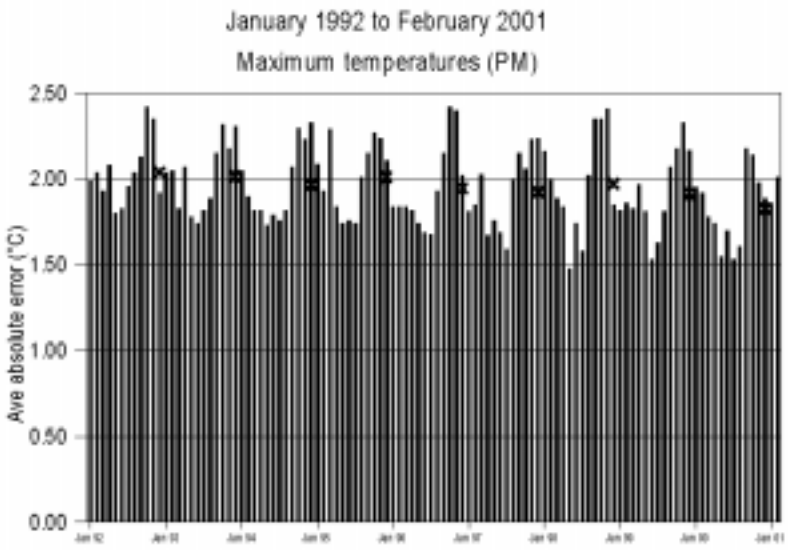

Figure 5b
Absolute error for maximum temperature from late afternoon (PM) forecast, valid for the following day for January 1992 to February 2001 as a monthly average of all the forecasting stations. The 'x' indicates the annual average for the year

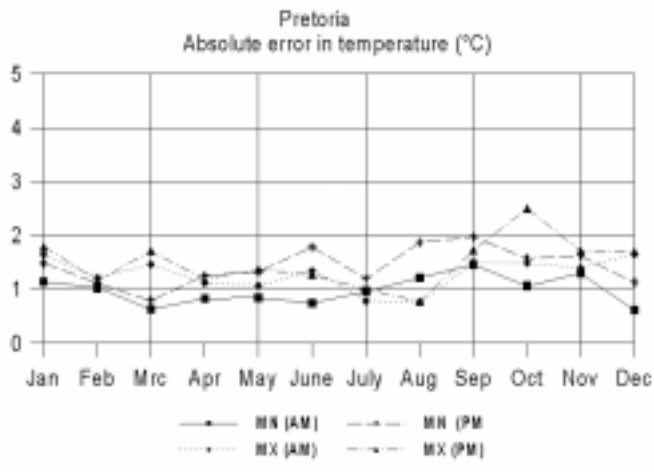

Figure 6a
Pretoria temperature forecasting statistics for January to December 2000

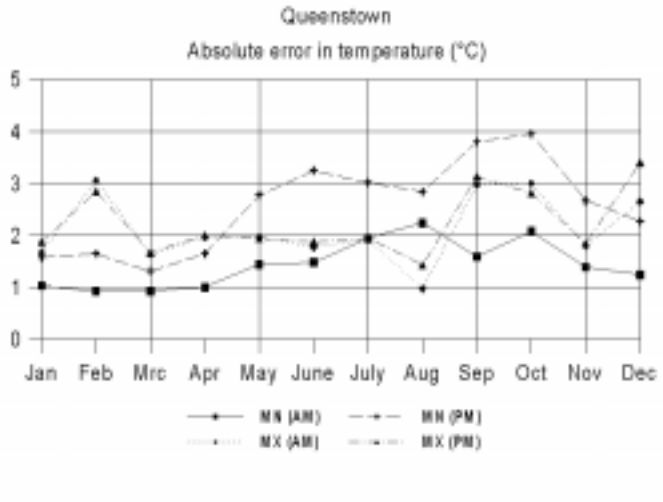

Figure 6c
Queenstown temperature forecasting statistics for January to December 2000

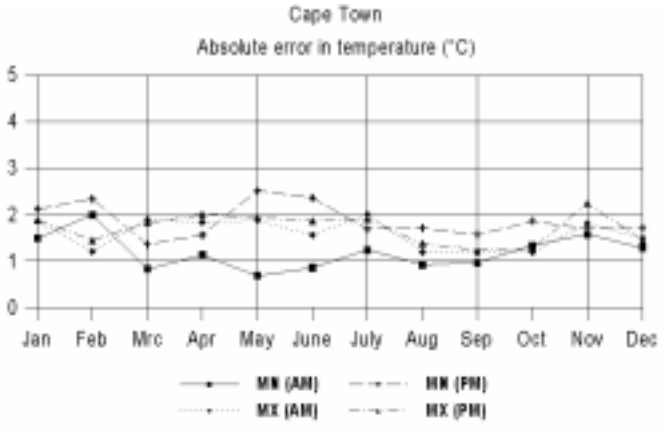

Figure 6b
Cape Town temperature forecasting statistics for January to December 2000

Looking at the temperature evaluation statistics for the year 2000 in Pretoria (Fig. 6a), we see the following:

- In general, the PM forecasts of maximum and minimum temperatures were the worst, while the AM forecasts were a little better.
- The PM forecasts of maximum temperature were the worst in the summer months (October to March), while the PM forecasts of the minimum temperature were the worst in winter (June to September).
- The AM forecasts of minimum temperature seemed to be the best overall, but one should bear in mind that this forecast is made at a time very close to the actual occurrence of the minimum temperature.
- The worst error made in Pretoria was in October 2000 (2.5°C) for the PM forecast of the next day's maximum.

Cape Town (Fig. 6b):

- The largest errors occurred in the PM forecasts, while the AM forecasts for minimum temperature were the best.

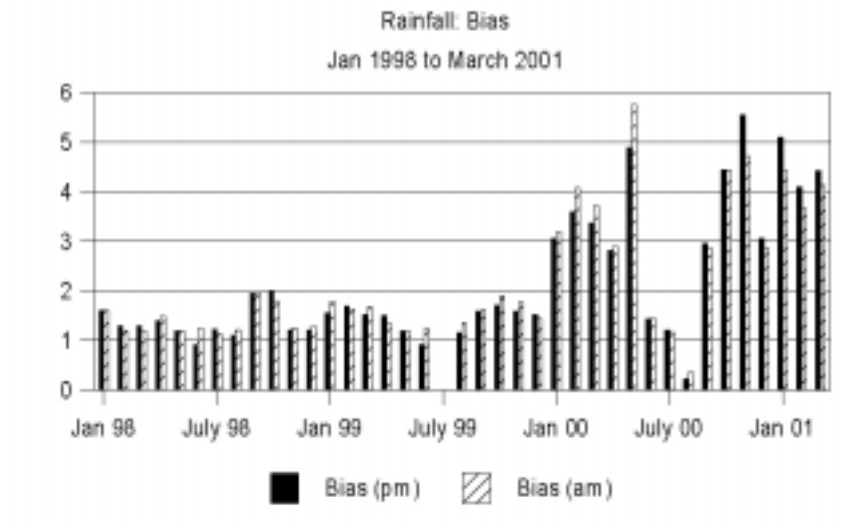

Figure 7a
Bias for rainfall (AM forecast light bar and PM forecast darker bar) for January 1998 to March 2001

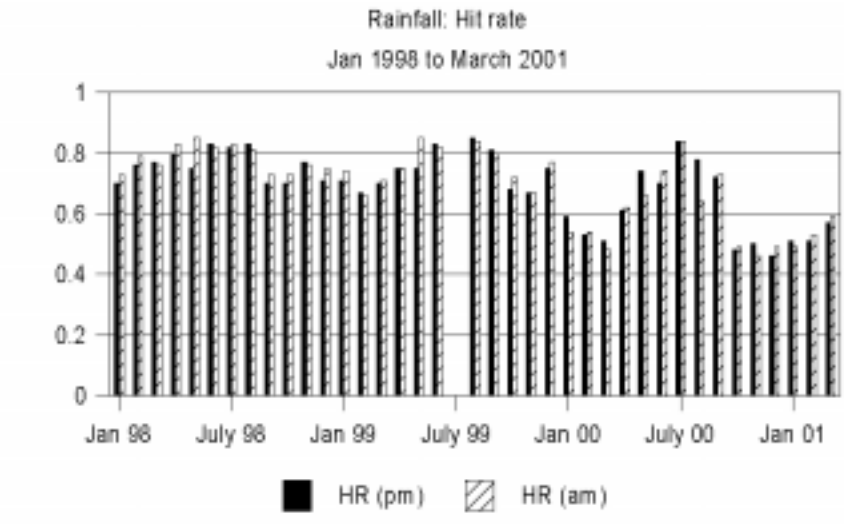

Figure 7b
Hit rate for rainfall (AM forecast light bar and PM forecast darker bar) for January 1998 until March 2001

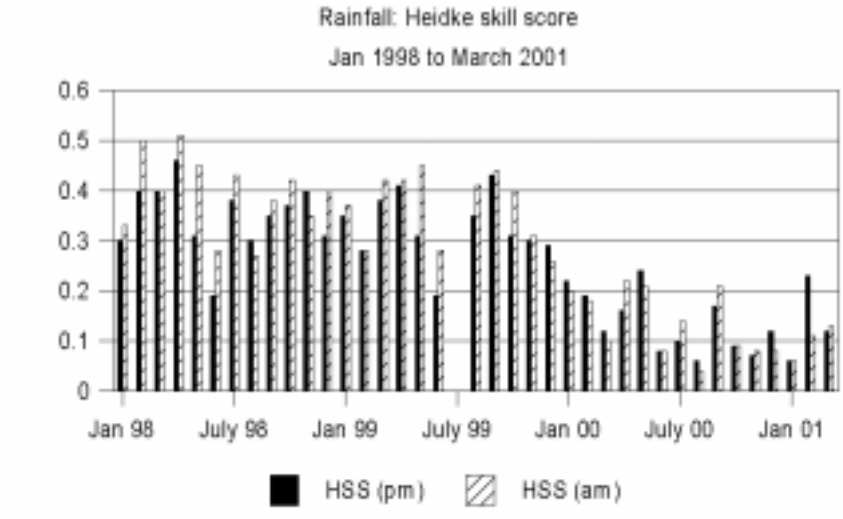

Figure 7c
Heidke skill score for rainfall (AM forecast light bar and PM forecast darker bar) for January 1998 until March 2001

- The errors in the PM forecasts of minimum temperatures were the largest in all the months, with the exception of March, April, July and November. The maximum error, which occurred in May 2000 was 2.5°C.

Queenstown (Fig. 6c):

- The largest error (4.0°C) occurred in October 2000 in the PM forecast for the following day's minimum.
- The AM forecasts of minimum temperature were generally the best, with the exception of August 2000.
- In general, the magnitudes of the errors were well above 2.5°C at this station, which is greater than at either Pretoria or Cape Town.

A few general conclusions:

- AM forecasts are more accurate than PM forecasts, especially for the minimum temperatures.
- Pretoria's errors were negligible compared to those at Cape Town, but Queenstown's errors were more significant and this is clearly a more difficult station to forecast.
- Another point of value is that at both Pretoria and Cape Town, the forecasters are situated at the location, whereas the temperature at Queenstown is forecast by the Port Elizabeth weather office.

Rainfall

The bias (Fig. 7a) remains fairly low, but less accurate forecasts were again noticeable in May 2000 when rain was definitely over-forecast. The hit rate (Fig. 7b) since 1998 remained fairly stable at more than 60%, but a period with less accuracy occurred at the beginning of 2000 when torrential rain occurred in the country. The Heidke skill score for rainfall (Fig. 7c) seems to diminish towards the end of the period.

Comparing a summer rainfall region like Gauteng and the eastern highveld (Area 1 in Fig. 2) with a winter rainfall region like the southwestern Cape (Area 8 in Fig. 2) for the year 2000, one notices the following:

- The bias of the AM forecast (Fig. 8a) was the greatest for both the regions in May 2000, when it reached a value of nine.
- The summer rainfall region's PM bias (Fig. 8b) was the greatest (6-7) during May 2000, with minimum bias values in the winter months. The winter rainfall region's bias was generally higher, with the highest value (8) occurring in November.
- The AM forecasts' hit rate (Fig. 8c) was the highest in January 2000 (100%) in the winter rainfall region, while the summer rainfall region's highest value occurred in July 2000

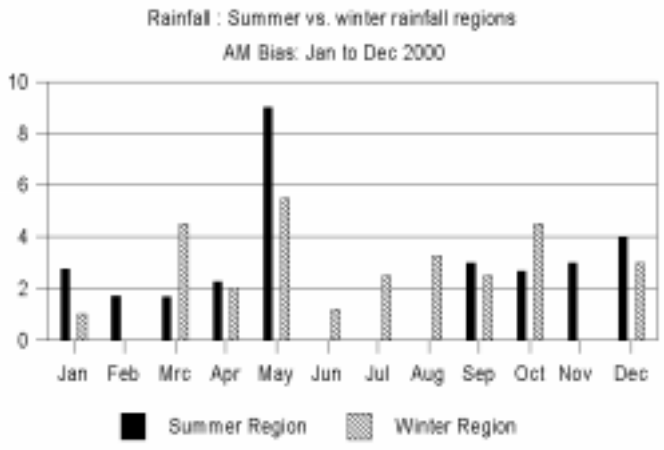

Figure 8a
Bias for rainfall comparing a summer rainfall region to a winter rainfall region for the January to December 2000 period for the AM forecast

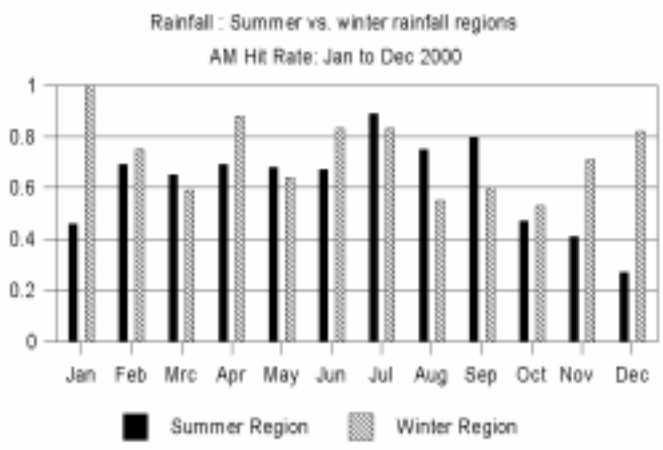

Figure 8c
Hit rate for rainfall comparing a summer rainfall region to a winter rainfall region for the January to December 2000 period for the AM forecast

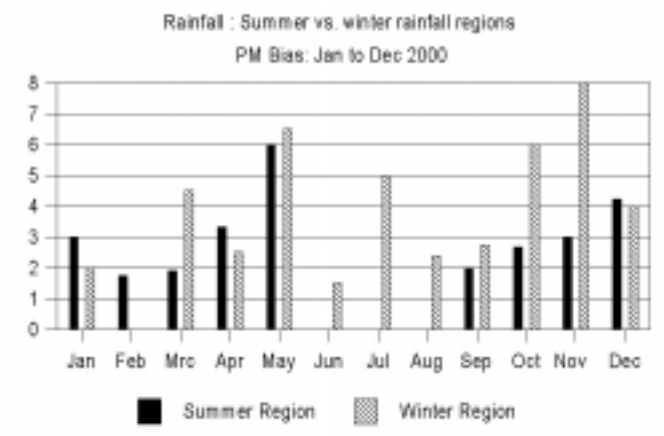

Figure 8b
Bias for rainfall comparing a summer rainfall region to a winter rainfall region for the January to December 2000 period for the PM forecast

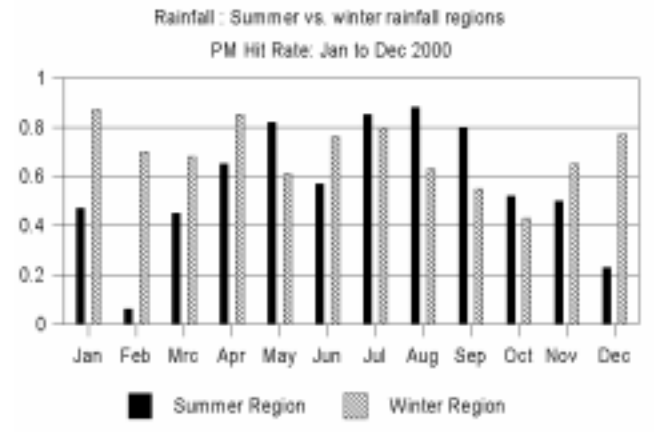

Figure 8d
Hit rate for rainfall comparing a summer rainfall region to a winter rainfall region for the January to December 2000 period for the PM forecast

(almost 90%). The summer rainfall region's worst hit rate was in December 2000, while the worst winter rainfall region's hit rate was in October.

- From the PM forecasts' hit rate (Fig. 8d) the best hit rate for the summer rainfall region was in August 2000, while the best hit rate for the winter rainfall region was in January 2000.

Severe weather

Fire danger index

The fire danger index is calculated by considering dry bulb temperature, relative humidity, wind speed and how recently rain occurred. This index must exceed 75 to be seen as a "hit" . Fire danger index is well forecast, with high PODs and low FARs (Fig. 9a). An unfortunate increase in the FAR is noted towards the end of 2000 and the beginning of 2001. A trend has been noted that the forecast onset of such fire danger events often occurs a day too late, while cessation is also frequently not mentioned.

Gale force winds

If the 10-min average (as reported synoptically) exceeds 35 knots, it is seen as a report of gale-force winds in this evaluation. POD (Fig. 9b) for gale-force winds generally remain above 30%, and are often more than 50%. The FAR (also on Fig. 9b) was below 50% for most of the period.

Heavy rain

For the purpose of evaluation heavy rain is defined as at least 50 mm of rain reported by at least two stations in the relevant geographic region. The probability of detection (POD) shown in Fig. 9c for heavy rain events was above 80% in the beginning of 2000 and even 100% in April. Persistence probably played a role in this increased forecasting skill, given a frequent recurrence of tropically-sourced heavy-rain-producing systems affecting the north-eastern part of the country as well as Mozambique. The POD was much lower in the beginning of 2001 than at the start of 2000. At the beginning of 2000 the false alarm rate (also shown in Fig. 9c) was generally below 50%, but in the subsequent few months too many warnings for heavy rain were issued resulting in a high FAR. In October 2000

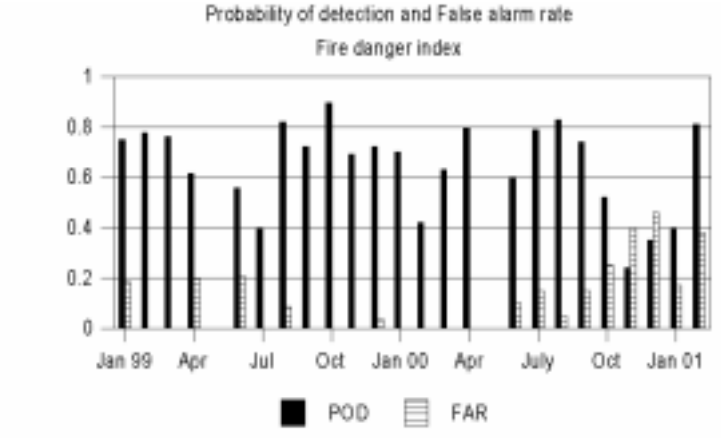

Figure 9a
Probability of detection and false alarm rate for fire danger index for January 1999 to February 2001

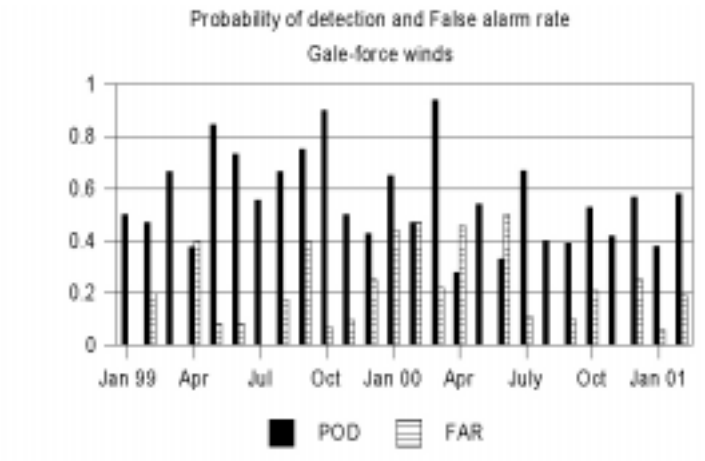

Figure 9b
Probability of detection and false alarm rate for gale-force winds for January 1999 to February 2001

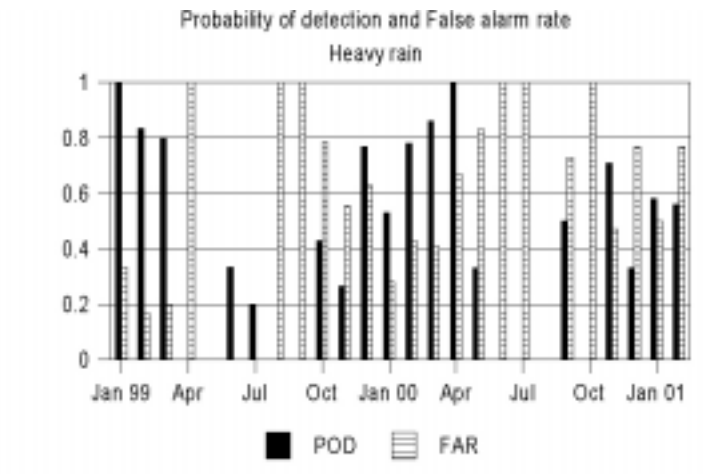

Figure 9c
Probability of detection and false alarms rate for heavy rain events for January 1999 to February 2001

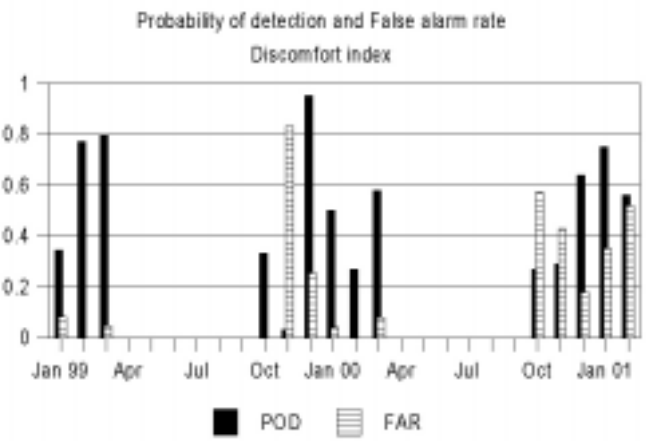

Figure 9d
Probability of detection and false alarm rate for discomfort index for January 1999 to February 2001

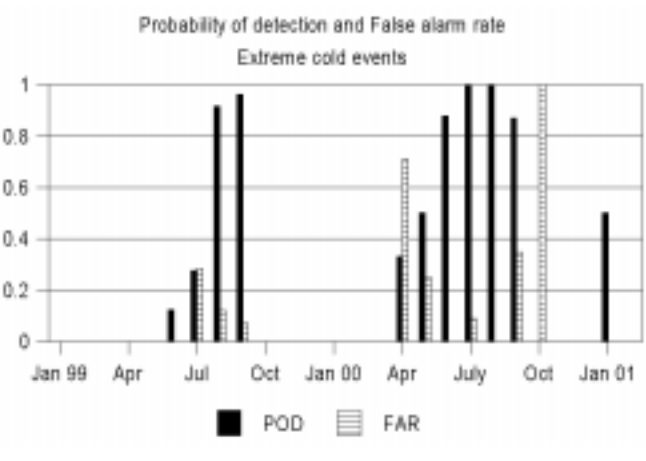

Figure 9e
Probability of detection and false alarm rate for extreme cold events for January 1999 to February 2001

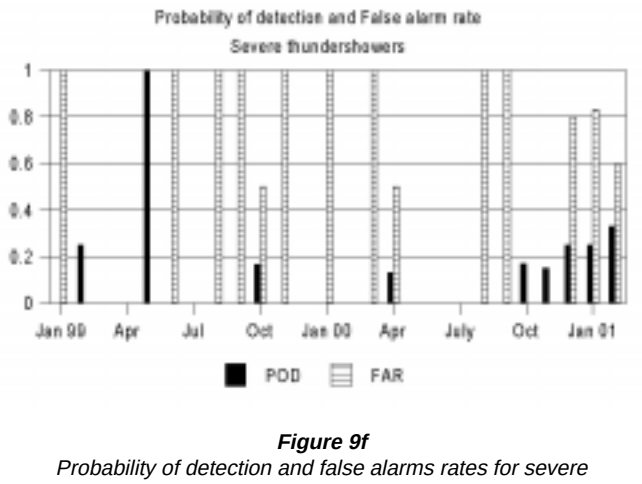

Figure 9f
Probability of detection and false alarms rates for severe thunderstorms for January 1999 to February 2001

the FAR was very high again, but it dropped towards the end of summer in 2001.

Discomfort index

The discomfort index is calculated by a relationship between temperature and relative humidity. A warning is issued if the index is equal to or exceeds 42. A lower temperature with high relative humidity can meet the threshold criterion, as well as a high temperature and a lower relative humidity. During the mid-summer months (December to February) the probability of detection is rather high (more than 70% shown in Fig. 9d). In this time the false alarm rate is fairly low, but there seems to be a tendency to higher false alarms rates earlier in summer (October and November).

Extreme cold

For this study an extreme cold event is seen as one where the daytime maximum temperature did not exceed 10°C in at least two places. Extreme cold events are usually predicted well, with a low FAR (Fig. 9e), with the exception of April and October 2000.

Severe thunderstorms

The National Severe Storms Laboratory (NSSL) in the United States of America defines a severe storm as one which has one or more of the following features:

- a tornado
- strong wind gusts, exceeding 50 knots
- hail larger than 19 mm in diameter (Doswell, 1982).

The reports of severe thunderstorm events rely heavily on the media and members of the public who call the Weather Service. If any report (synoptical, via the public or via the press) is available, then the event is seen as a severe weather event. Severe thunderstorms, including wind damage, hail or tornadoes are not anticipated well and usually have a high FAR (Fig. 9f). A positive observation is that the POD increased in the past summer.

The prediction of the severity of thunderstorms relies almost completely on radar information. This kind of prediction can only be made once the storm has already started to develop and radar signals (such as high reflectivity, echo shape, reflectivity gradient and storm movement) can be established from the radar image. Radar coverage in South Africa is not yet adequate and the improvement of this situation can lead to better forecasts in this category. This parameter is far more difficult to anticipate, both in terms of space and time, with any consistent degree of success. One must be cautious when using FAR and POD under such circumstances, since prediction of events with a low frequency of occurrence can be misleading.

Conclusions

Temperature forecasts lie within a 2.3°C range of accuracy with minimum temperature forecasts performing best early in the morning. Stations closer to the forecast office might have a bias towards better forecasts, since the forecasters are situated at the location and know the area well. Stations close to or on the escarpment (such as Queenstown) are more difficult to forecast and the errors are usually greater.

Rainfall tends to be over-forecast with the predictions only slightly better than chance. Although more than 1 700 rainfall reporting stations are spread across the country, smaller, more severe rainfall events can still be missed between stations. The forecasts tend to improve at the time of the year without significant rainfall. For example, the summer rainfall region's best results are in winter, and the winter rainfall station's results are best in summer. Satellite and radar are not yet used in the verification process and can be recommended for future use to make up for the gaps between rainfall stations.

Severe weather events are sometimes handled well, but severe thunderstorms are not predicted with great accuracy. The lack of an adequate observation network for severe weather events can contribute to problems in the verification system. It is, however, useful to take note of the statistical results bearing the constraints in mind.

One of the scientific purposes of verification is to identify the areas where more attention and/or research should be focused. From these presented statistics, such areas are:

- Maximum temperatures (more so than minimum temperatures) - reduce the absolute error to below 2°C.
- Rainfall (especially heavy rain and severe thunderstorms) - aim to get a higher POD and lower FAR. The forecasting and meaningful evaluation of these kinds of events relies on remote-sensing techniques such as satellite and radar, which are only available to a limited degree in this country. Additional technology and information would be needed to enhance skill in the case of severe thunderstorm events.
- Any improvement in the observation network (both for everyday weather and severe weather events) will not only benefit the forecasters, but will also make evaluation much easier: With more real time data at his/her disposal, a forecaster will have a better idea of the current situation and will be able to forecast with greater accuracy. The availability of more data to evaluate might also prove beneficial to the evaluation statistics, since fewer events will be 'missed', from an observation point of view.

With the South African Weather Service in the process of commercialisation, the accuracy of the product delivered will become increasingly important. Forecast verification statistics will have to be kept up to date, trends will need to be noted and it will also be necessary to have a benchmark of performance to present to prospective clients. Forecast evaluation is currently only done (officially) at head office in Pretoria, but could also be expanded to be done in a similar manner at the outstations. In order to increase the degree of accuracy, all operational forecasters should be equipped with technological expertise and training in the areas where improvement is desired.

Acknowledgments

I would like to acknowledge the following Weather Service staff members who perform the monthly evaluations and provide the monthly statistics which are used in this paper: Kevin Rae (severe weather warnings), Anastasia Demertzis (rainfall forecasts) and Louis van Hemert (temperature forecasts). Without their input, this paper would not have been possible. Hilarie Riphagen also provided valuable editorial advice.

References

BRIER GW and ALLEN RA (1951) Verification of Weather Forecasts. In: Malone TF (ed.) *Compendium of Meteorology. Am. Meteor. Soc*. 841-848.

DOSWELL CA (1982) *The Operational Meteorology of Convective Weather*. **1**: *Operational Mesoanalysis*. NOAA Technical Memorandum NWS NSSFC-5.

FRAEDRICH K and LESLIE LM (1988) Real-time short-term forecasting of precipitation at an Australian tropical station. *Weather Forecasting* **3** 104-114.

HEIDKE P (1926) Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst. *Geog. Ann. Stockh.* **8** 310-349.

NOONE D and STERN H (1995) Verification of rainfall forecasts from the Australian Bureau of Meteorology's Global Assimilation and Prognosis (GASP) system. *Aust. Meteor. Mag*. **44** 275-286.

WILKS DS (1995) *Statistical Methods in the Atmospheric Sciences*. Academic Press. 465 pp.