MOS PROBABILITY OF PRECIPITATION FORECASTING EXPERIMENTS IN THE MEDIUM RANGE

L.J. Wilson

European Centre for Medium Range Weather Forecasts
Reading, U.K.

ABSTRACT

A set of computer programs has been developed at ECMWF to facilitate statistical verification and interpretation studies on the output of the model. The software is designed for maximum flexibility in the choice of predictors and predictands to be used in individual studies, and includes such features as spatial and time averaging of predictands and predictors, horizontal derivation and time differencing of predictors. The statistical programs also include provision for several types of computed predictors such as non-linear transformations.

Using this software, statistical interpretation tests were carried out for two European locations with considerably different climatologies, a watershed area of southern Norway, and Barcelona, Spain. The available datasets consisted of one winter and one summer season of model output parameters and observations for use as dependent data, and shorter datasets for use as independent data. The predictand was probability of precipitation amount (POPA) over 24 hours, ending at 96 and 120 hours into the forecast. For comparison purposes, some tests were carried out on shorter range forecasts of 48 and 72 hours, and on forecasts of 3-day total precipitation accumulation for days 3 to 5 of the forecast. Experiments were also carried out using different averaging options of the software. The Multiple Discriminant Analysis statistical technique was used throughout, and all experiments were conducted in the Model Output Statistics formulation.

Results indicate, for the short samples available, that there is skill in the model forecasts for use in interpretation up to at least day 5 for the Norwegian area, but that skill is non-existent after day 2 for Barcelona. Changes in the averaging do not produce significantly different results. The statistical procedure generally was able to improve on the direct model output precipitation forecasts.

1. INTRODUCTION

Statistical interpretation of model output is usually carried out in order to develop a forecast product that is both able to improve on the direct model output and is nearer in format to the final forecast product issued by a forecaster to his users. Statistically based forecasts are therefore usually associated with an operational forecasting program, helping to provide a link between the model and the forecast, and permitting an objective combination of the large scale model predictions with local scale effects contained in the predictand datasets, but generally not adequately handled by the model.

Statistical interpretation studies have a second benefit that is of interest to model developers. They provide a means of verifying the model. If the predictand is one of several that are directly forecast by the model, the statistical selection program should pick up the direct model output parameter. Information on phase errors and other biases can be obtained by studying the predictor sets that are selected by the statistical program. If the direct model output is selected, information can be obtained by examining the other predictors selected to improve on the direct model output.

Statistical interpretation studies at ECMWF are important for the latter purpose, but the software developed to carry out these studies can, it is hoped, benefit meteorologists of the various Member States in carrying out statistical development intended for operational implementation. An important purpose of the studies described below was to test the new software; the results cannot be considered definitive because of the small datasets involved.

A third purpose of carrying out studies in statistical interpretation is to test the application of MOS, and perhaps perfect prog. techniques to the medium range forecast problems. This has not been tested yet; all MOS studies have so far been limited to shorter range forecasts. It is entirely possible that application methods that work in the short range, may not translate directly to the medium range. Many studies will be needed to determine the best means of using statistical interpretation procedures in medium range forecasting.

The studies described below were limited to precipitation amount as a predictand and use only the discriminant analysis technique. The development of equations followed the Model Output Statistics procedure, where relationships are derived between the observed predictand values and model output from an earlier forecast run. For example, equations for a 96 hour forecast are derived by relating the observations to output from a model run initialised 96 hours previously. In general, 96 hour forecast values of the predictors would be used from this run, but predictors

for other valid times are offered as well to catch model timing biases. For testing purposes, values of the predictors from the current model run are entered to produce a forecast valid 96 hours from now.

For each of the two stations, the data and station location are discussed in terms of the model structure, verification of the direct model output precipitation forecasts is described, the statistical experiments are discussed and results shown. For the Norwegian station, one interpretation experiment is described in some detail to illustrate the use of the statistical technique, and the results of other trials are summarised.

2. SIRA-NORWAY

2.1 The Problem

Sira is a watershed area in Southern Norway which contains a number of reservoirs and hydroelectric power developments. The power company that controls the plants is interested in precipitation forecasts for the watershed area. They presently receive forecasts out to 36 hours from the Norwegian Weather Service, but their resource management program calls for a planning cycle of one week, with daily updates. There is therefore a need for medium range precipitation forecasts as well.

The goal of the power company is to sell power when prices are high, while at the same time avoiding the risk of flooding. The major input to their planning is a hydrological model that uses as input not only precipitation forecasts, but also temperature and humidity information (for evaporation). The potential savings from accurate forecasts can be of the order of £5,000 per day.

Figure 1 shows the location of the Sira area on a detailed topographical map of Norway. The area is characterised by steeply sloping topography, rising from 200m in the south to about 1000m in the north. For comparison, the nearest 4 gridpoints from the European archive of model data are plotted. The Sira area is quite small compared to the model grid, and can nearly be considered as a point location.

Because of the steeply sloping terrain, it is expected that orographically induced precipitation is important, especially with southerly circulations. The new and old (before 1 April, 1981) model topography are shown in figure 2. The new topography ranges from 200 to 400m in the Sira area, much less steeply sloping than the actual topography. It would be expected from this that the model might tend to underforecast precipitation in the Sira area. The old model topography contains practically no slope at all in the Sira area.

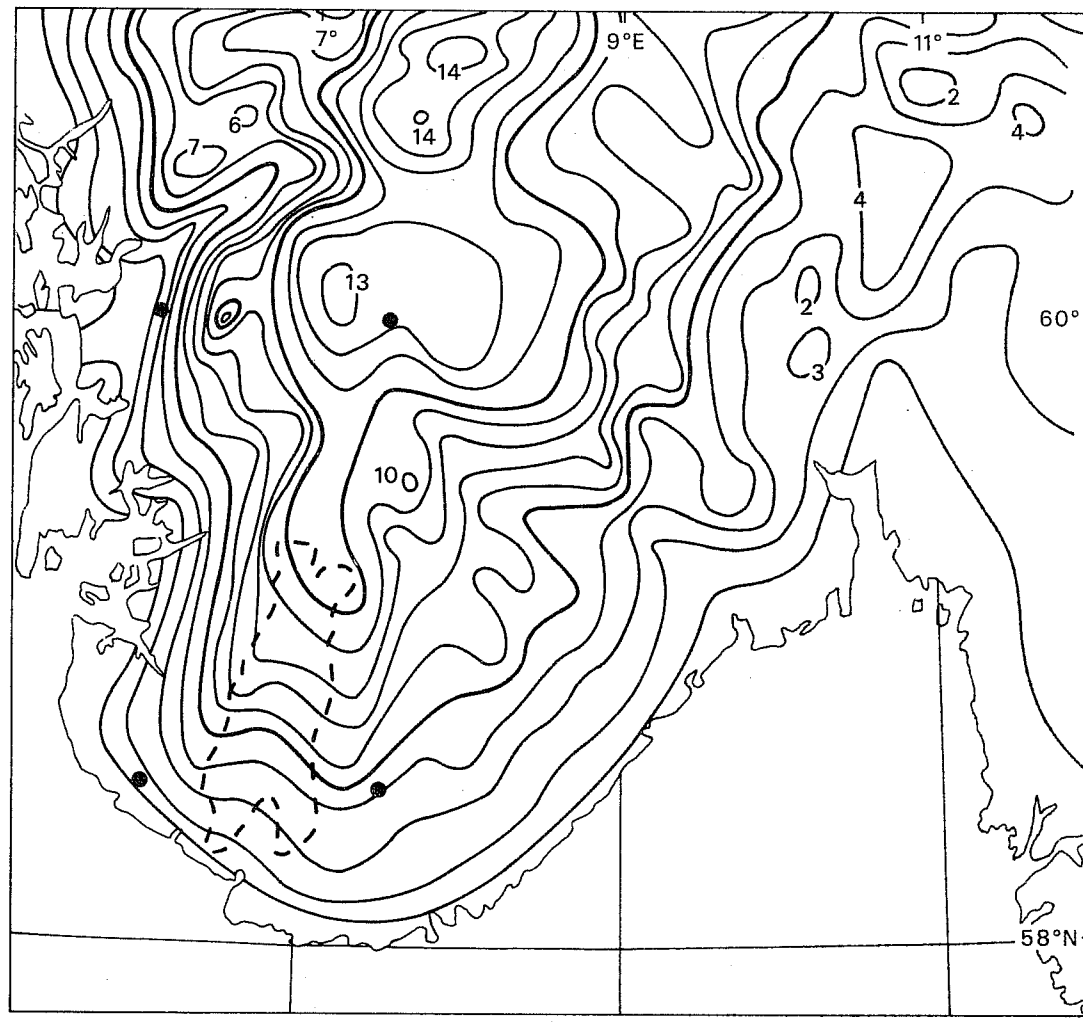

Fig. 1 Topography of southern Norway from data with a resolution of 10 km. Contour interval is 100 m with each 500 m contour represented by a heavy line. The SIRA area is outlined with a dashed line and the nearest four gridpoints of the European Archive are shown by dots.

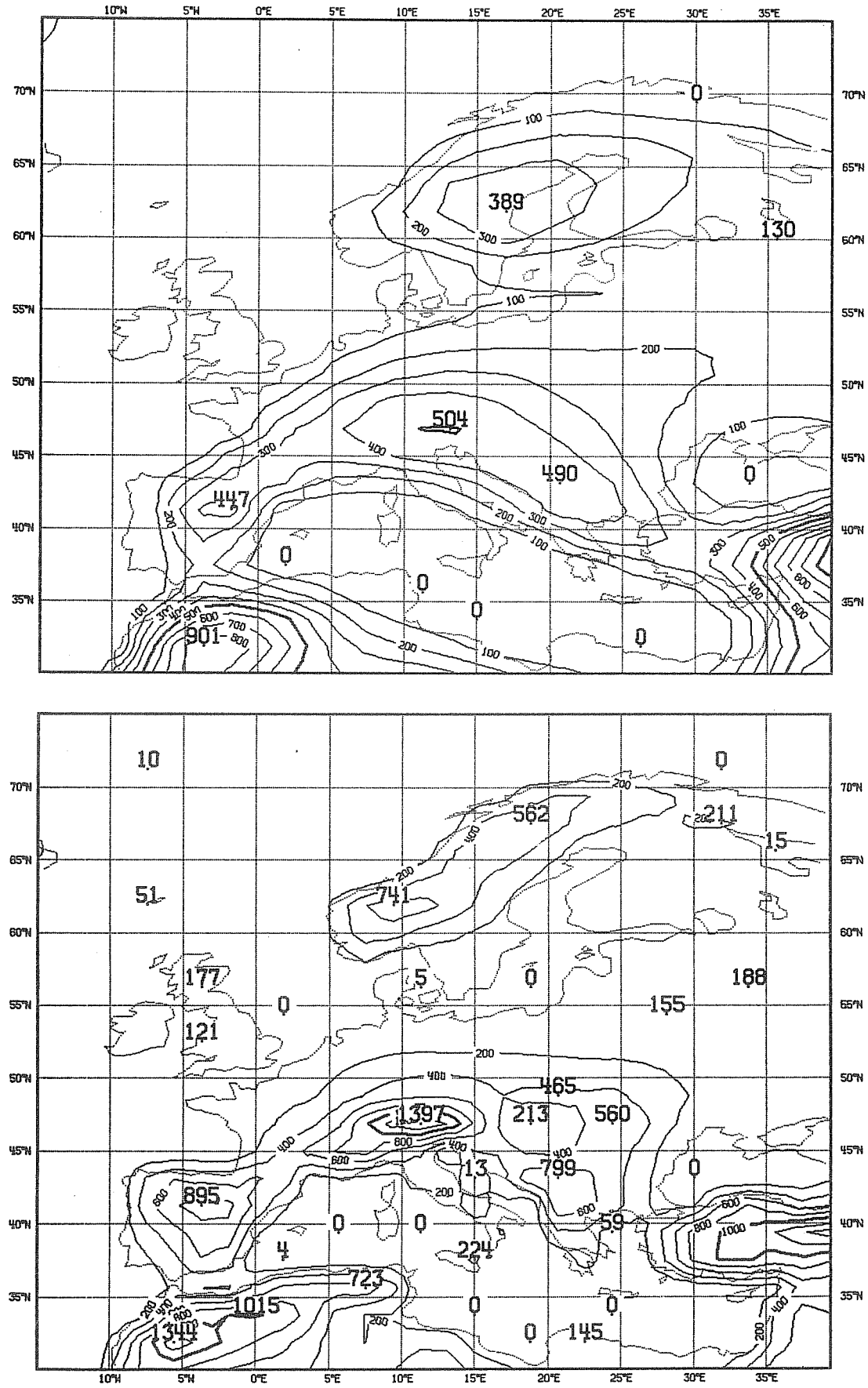

Fig. 2 ECMWF model orography over verification area, metres. Maxima and minima indicated. Top: Prior to April 1, 1981. Contour interval 100m. Bottom: From April 1, 1981. Contour interval 200m.

2.2 Datasets

We were fortunate to have access to good quality precipitation observations from the Sira area. The 4 observing sites that were used are marked on figure 3. The predictand set consists of observations of 24 hour precipitation accumulation averaged over the 4 sites, ending at 06 GMT. Although the sites are rather close together in terms of the model resolution, the use of an average of 4 point observations should stabilise the dataset and help ensure a representative observation for the whole Sira area.

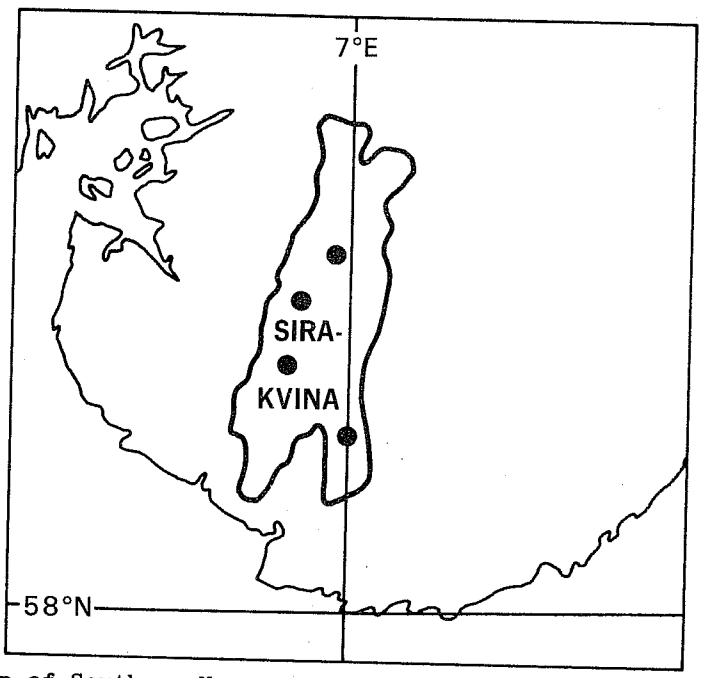

Fig. 3 Map of Southern Norway with SIRA watershed area outlined by a heavy black line. The four observing sites used to obtain the precipitation data are indicated by black dots.

The model output datasets that were available are shown schematically in figure 4, in comparison with model changes that might have a bearing on precipitation forecasting. The two dependent sets were for 6 month periods ending on 15 October, 1981 and 15 April 1982. Summer 1982, from 16 April to 30 June was used as the summer independent set and winter 1980/1981 was used as the winter independent set. The diagram shows that some differences can be expected in model output between dependent and independent sets. The independent winter set is before the topography change while the dependent set is after, and the dependent summer set is during a period when there were problems with the diffusion scheme causing overforecasting of convective precipitation near steep mountains. This problem has been corrected for the 1982 summer.

It is worthwhile to examine the datasets as time series to check that the independent observation samples and dependent samples are reasonably consistent. Figures 5 to 8 are time series plots of the observations (solid lines) and 96 hour forecast total precipitation (dashed lines) for the four datasets. It can be immediately seen from these plots that precipitation at Sira is frequent, and there is a relatively frequent occurrence of daily amounts in excess of 20mm. Furthermore, allowing for differences in the horizontal scale of the graphs, the dependent and independent samples have similar climatologies. Significant precipitation accumulations occur in summer as well as in winter. Although the predictand distribution is similar between seasons, the physical processes leading to precipitation are different, and it was not possible to combine the two seasons into one sample.

The predictor data consisted of model forecast data extracted at 16 points on 1.5^{o} latitude-longitude interval, centred on the Sira area. The gridpoint northeast of the area has the highest associated topography in the model and in reality. The predictor set generated from the data is described below.

2.3 Model Verification

Examination of figures 5 to 8 reveals some of the characteristics of the model precipitation forecasts for the Sira area. Firstly, there is a general tendency to underforecast the extreme amounts, in all seasons. This is most notable in winter 1980/81 (figure 6), presumably because the topography had not yet been changed. In all other seasons, the model occasionally forecasts large amounts, not necessarily correctly. The vertical diffusion problem is evident in the summer 1981 sample (figure 7), with an abundance of relatively small, but incorrect peaks.

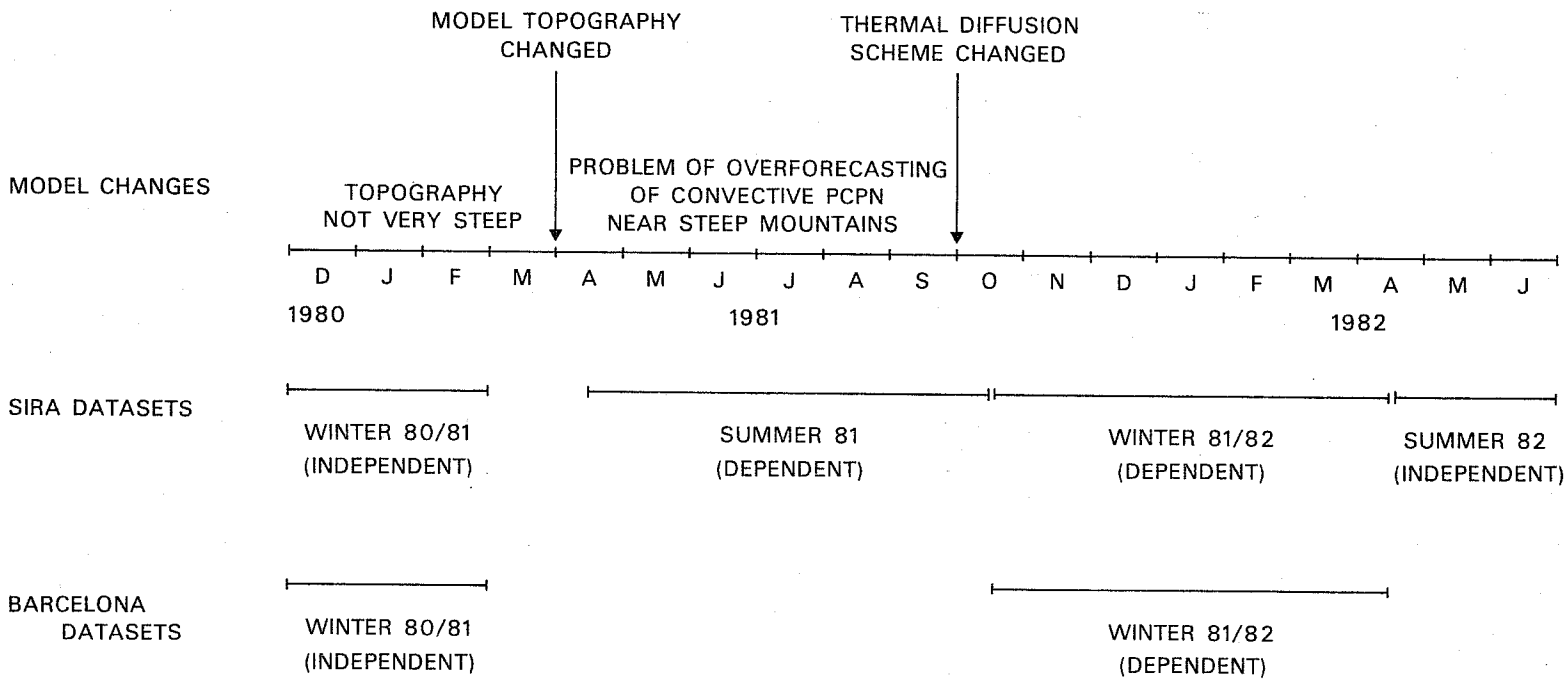

Fig. 4 Schematic representation of available datasets in comparison with model changes of significance to precipitation forecasting.

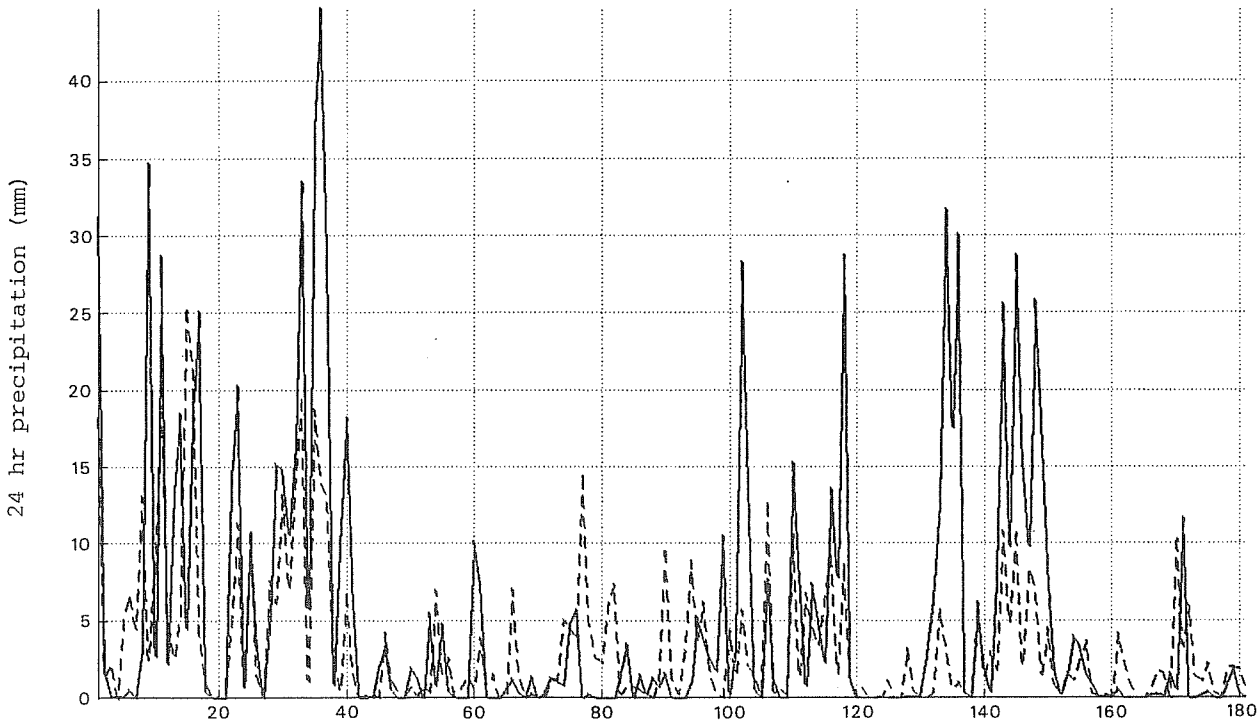

Fig. 5 Time series plot of one day total precipitation amounts for SIRA, Winter 81-82. Solid line is observed precipitation. Dashed line is 96 hour forecast total precipitation.

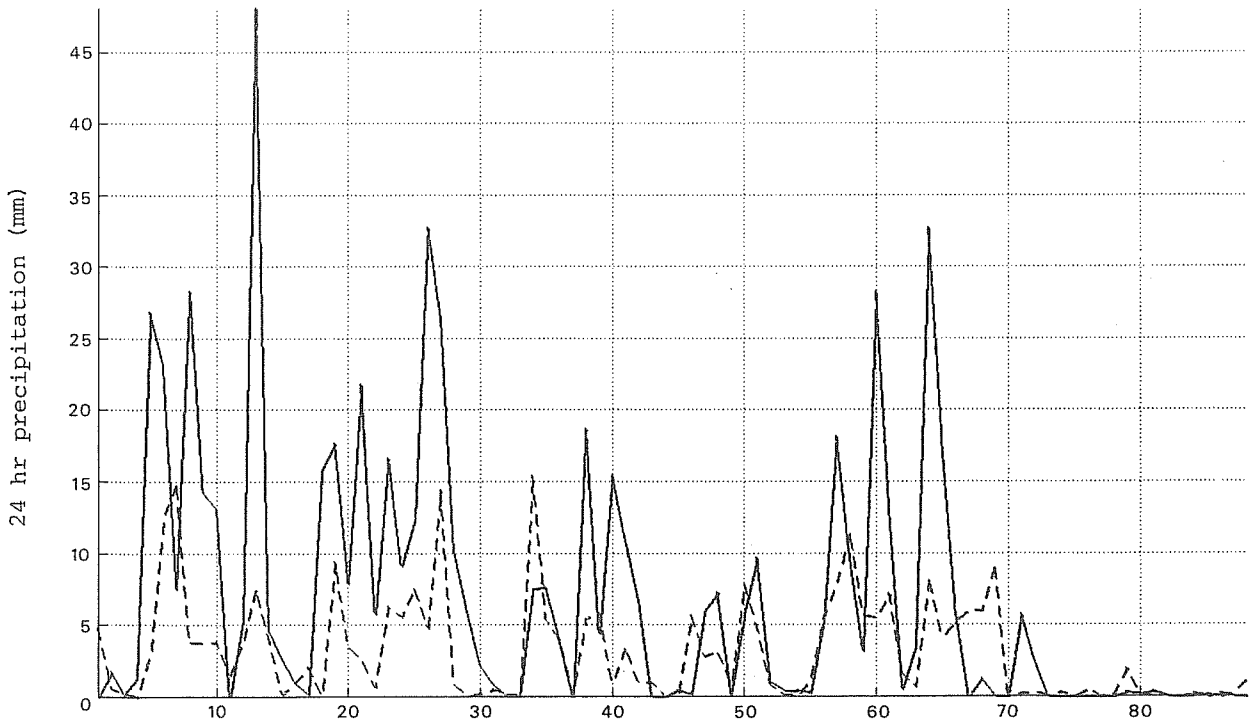

Fig. 6 Same as for Fig. 5 but for Winter 80-81.

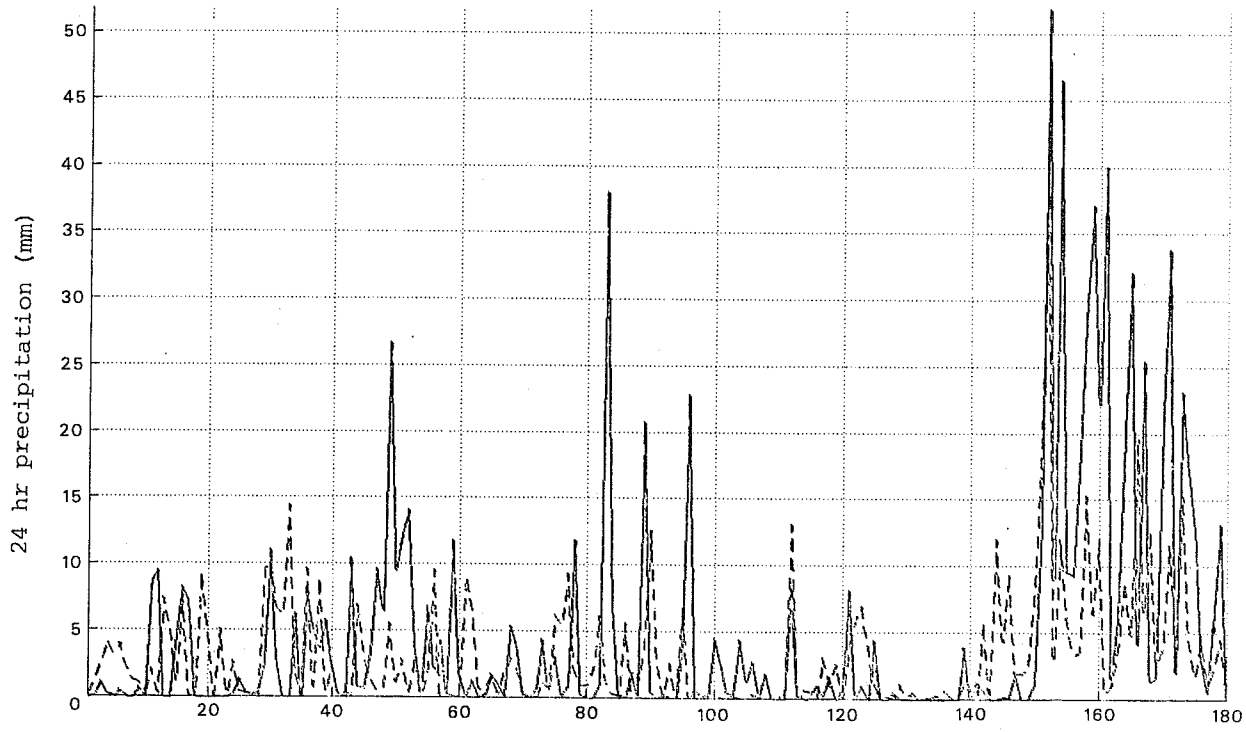

Fig. 7 Same as for Fig. 5 but for Summer 1981.

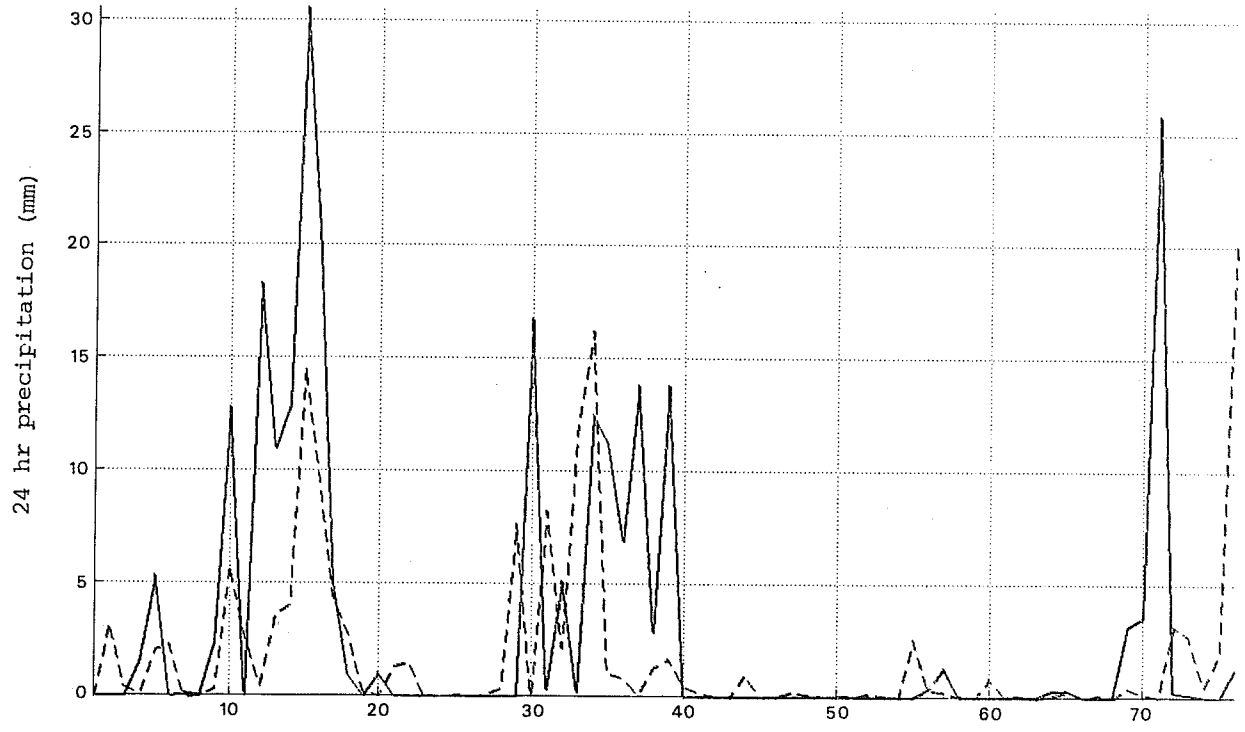

Fig. 8 Same as for Fig. 5 but for Summer 1982.

The general tendency to underforecast is also demonstrated by the scatter plot shown in figure 9. The abundance of points to the right of the "forecast observation" line demonstrates this. Also notable is the fact that there are some significant overforecasts as well. By underforecasting extremes and overforecasting at other times, the model tends to predict toward the mean in a statistical sense.

Summary statistics of the model verification for Sira are shown in figures 10 to 12. These generally confirm expectations based on knowledge of model changes that took place during the period. Specifically, winter 1981/1982 should be better than winter 1980/1981 because of the topography change and summer 1982 should be better than summer 1981 because the diffusion problem has been corrected. Root mean square error(RMSE) and mean absolute error (MAE) are shown for summer samples (figure 10) and winter samples (figure 11). In winter, improvements from 1980/1981 to 1981/1982 are smaller than between the two summers, but noticeable. The RMSE is somewhat higher (8 to 10mm) than the MAE, suggesting relatively frequent occurrence of large errors, as seen in figures 5 to 8. Both figures show the expected increase in errors with increasing forecast projection time. Despite improvements to the model, the errors are still quite large in magnitude, typically RMSE of 8mm and MAE of 4mm compared to average observed precipitation of about 5mm. There is still considerable room for improvement.

The bias characteristics (figure 12) show negative values (underforecasting) for both seasons. Improvement is most dramatic between winter 1980/1981 and 1981/1982, confirming the effect of the topography change to increase the forecast precipitation for the area. The summary verification results are for 4 point averages of the model output; point value verification in winter showed little difference in MAE and RMSE, and summer point value verification was not tried.

Experiments were also tried using 3 day total precipitation accumulation for days 3 to 5 of the forecast, using corresponding time averages of the predictands. Time series plots of the 3 day total precipitation are shown in figures 13 and 14 for the two winter samples. These figures show that very large 3 day accumulations do occur (>100mm) occasionally, and that the model's tendency to underforecast extends to 3 day totals, in fact, the magnitude of underforecast is increased. Also of interest is the tendency for the averaging process to smooth the observations, but not the forecasts. The 1981/1982 dataset especially shows a much more variable model forecast than the observed amounts would indicate. Perhaps this is an indication of a tendency of the model solution to change quite radically from one day's run to the next. The summer three day accumulation for 1982 (figure 15) still shows this effect, but to a lesser extent.

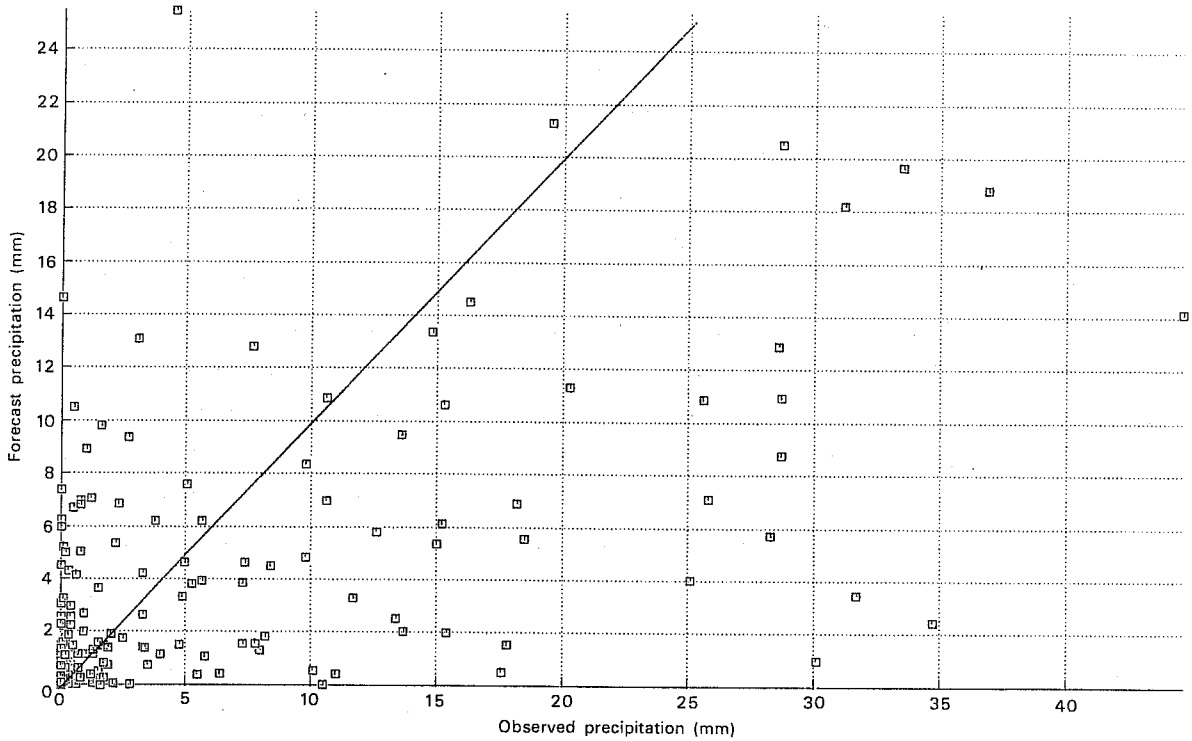

Fig. 9 Scatter plot of observed precipitation for SIRA (abscissa) and 96 hour forecast precipitation (ordinate). The diagonal line gives the locus of correct forecasts.

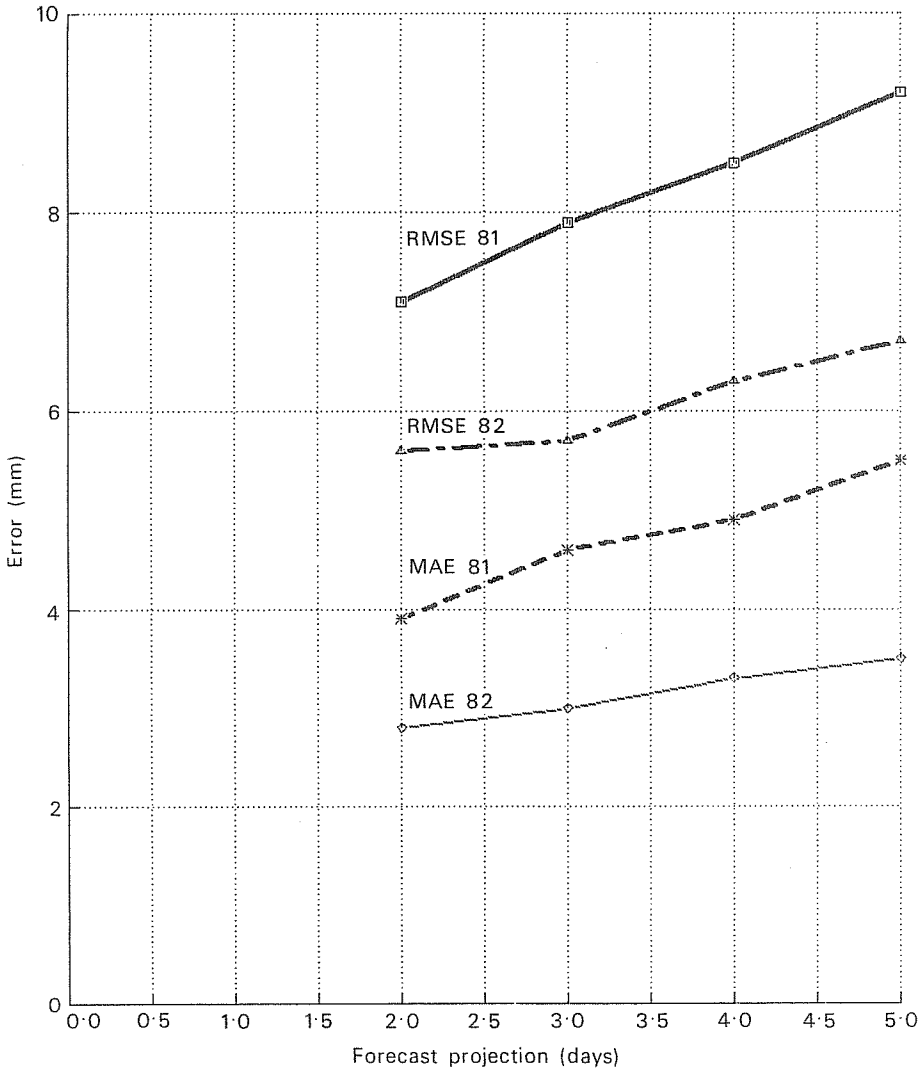

Fig. 10 Model verification statistics for Summer data for SIRA.

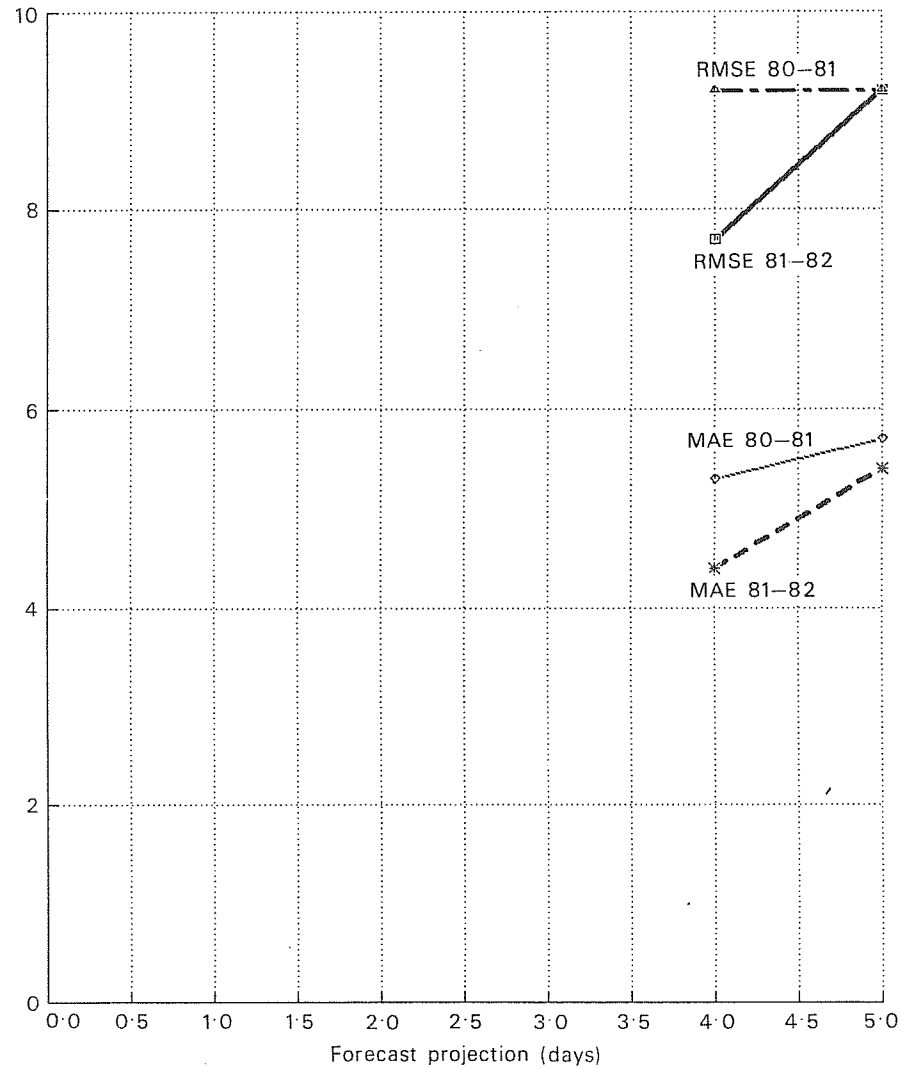

Fig. 11 Same as for Fig. 10 but for Winter data.

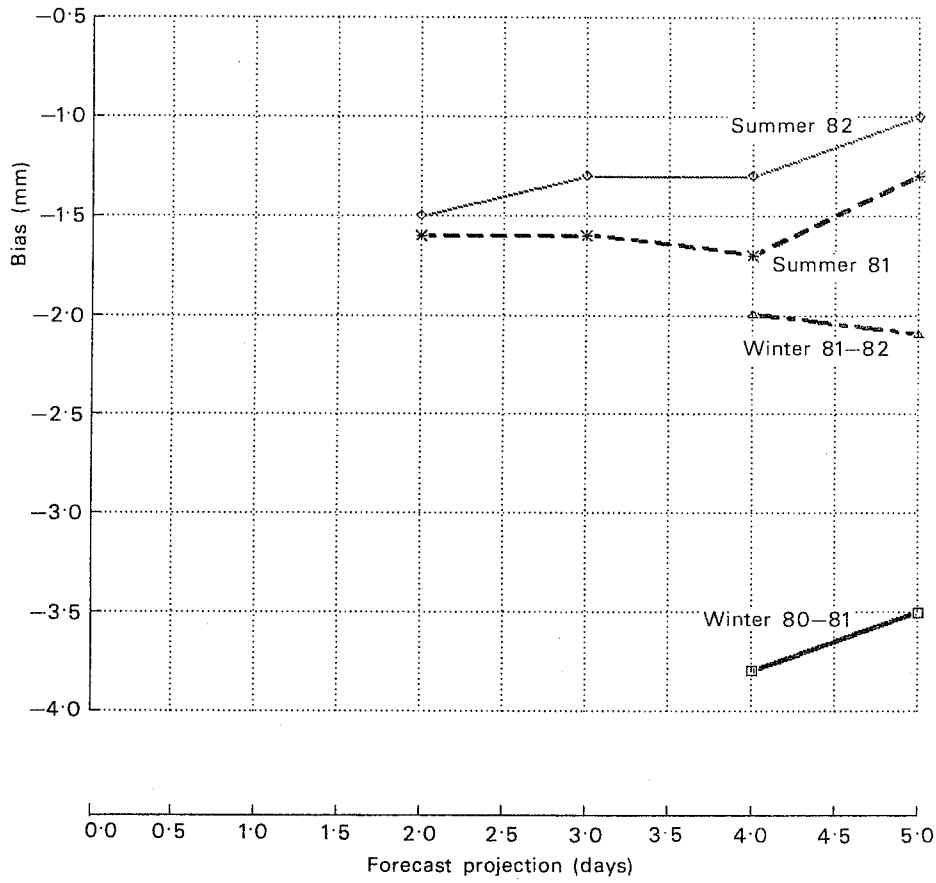

Fig. 12 Bias of model precipitation forecasts for SIRA.

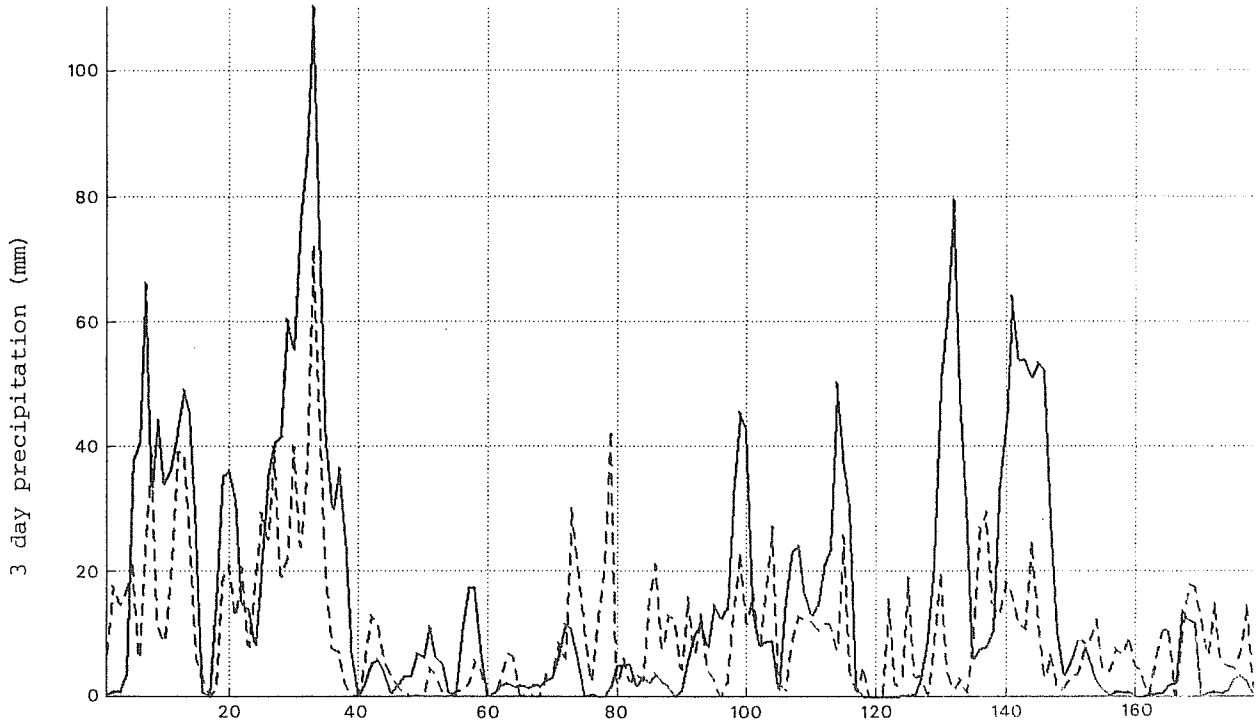

Fig. 13 Time series plot of three day accumulated precipitation for SIRA (solid line) and forecast of 48 to 120 hour precipitation (dashed line) for Winter 81-82.

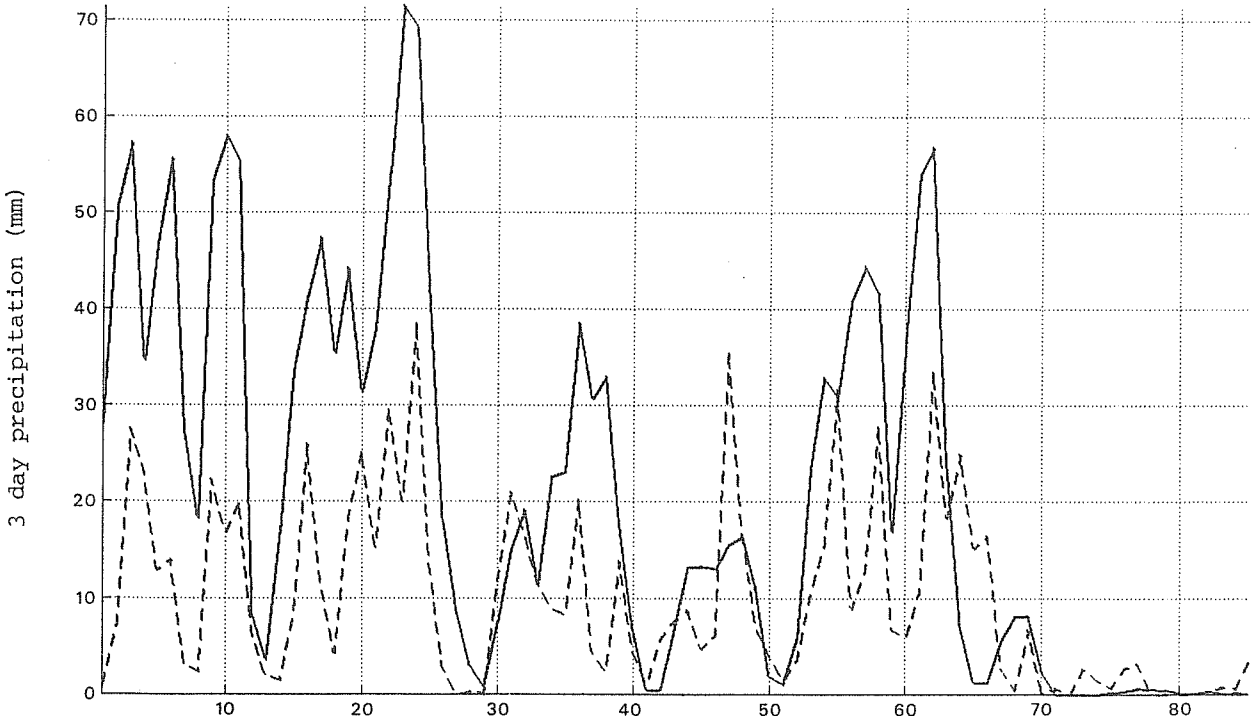

Fig. 14 Same as Fig. 13 but for Winter 80-81.

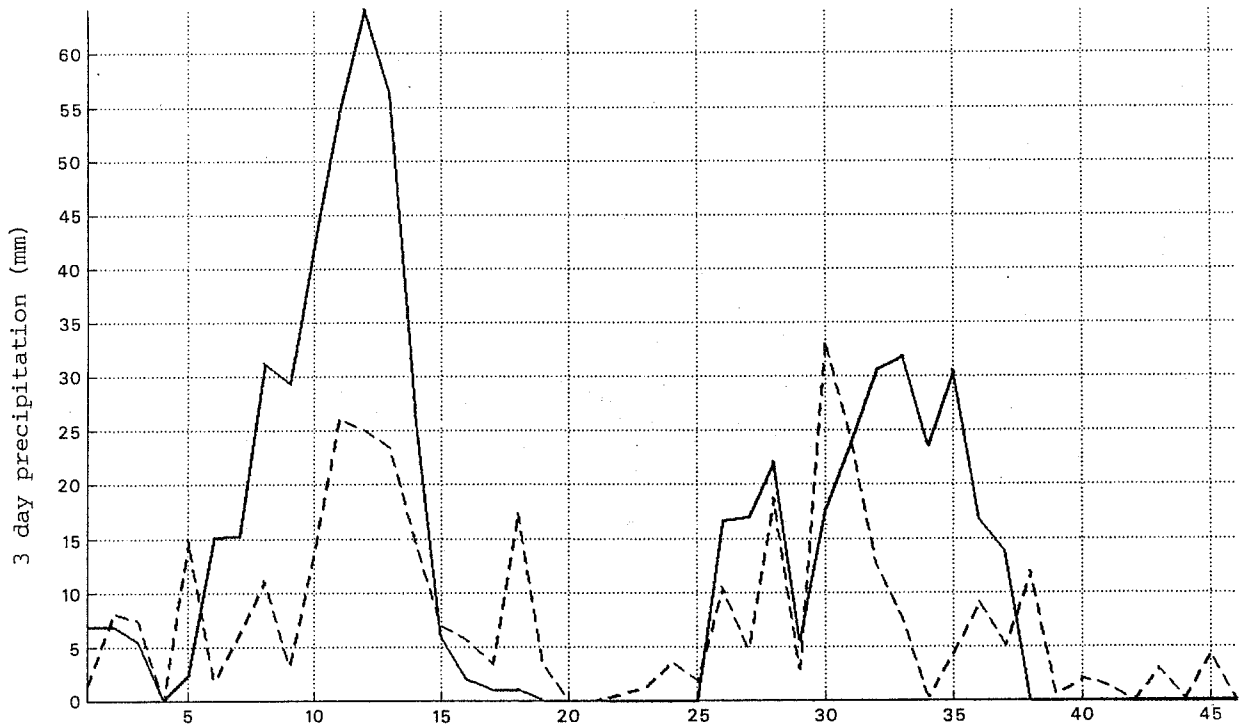

Fig. 15 Same as Fig. 13 but for Summer 82.

Summary statistics for three-day total precipitation are shown in table 1. Average precipitation is higher than for one day amounts, but errors are also larger. The three statistics RMSE, MAE and Bias tell essentially the same story as for the one-day totals - improvement between earlier and later datasets for both seasons, with greatest improvement in the bias of the winter forecasts.

SUMMER	1981	1982
RMSE (mm)	19.5	14.0
MAE (mm)	11.9	9.5
BIAS (mm)	-8.0	-5.6
WINTER	80/81	81/82
RMSE (mm)	19.2	18.4
MAE (mm)	14.0	12.0
BIAS (mm)	-12.0	-6.6

TABLE 1 SIRA MODEL VERIFICATION, 3 DAY TOTAL AMOUNTS

2.4 Statistical interpretation trials on 96 hr. precipitation forecasts

The experiment selected for a detailed discussion was the forecast of winter probability of precipitation amount (POPA) in 3 categories for forecast day 4. The predictand categories were <1.0mm, 1.0 to 20.0mm and >20.0mm in one day. This categorisation produced sample sizes of 90, 75 and 16 events respectively for a total dependent sample size of 181 events. The independent dataset from winter 80/81 consisted of 88 events.

Selection of predictors for the experiments was an iterative process. After an initial set was selected on the basis of expected physical relationships with precipitation forecasting, they were offered to the MDA program. After several runs with different timesteps and different category definitions, it was possible to identify the predictors favoured by the MDA screening program. Those predictors were retained, and a new set was defined to replace predictors that were ignored by the MDA with new predictors similar in type to those often selected. The final set offered to the MDA is shown in tables 2 and 3. All the MDA experiments for both Sira and Barcelona used this set of predictors, with minor changes mostly in the model forecast projection times offered. Iteration in predictor definition and screening is nearly always necessary because of the very large number of possible formulations available. Even a large computer cannot screen all possible predictors available after a careful pre-screening. An example of a predictor that was usually overlooked was vertical velocity, presumably because it was too noisy.

1. Geopotential Height 1000 mb
2. Geopotential Height 500 mb
3. Laplacian of Geopotential 1000 mb height (deleted from some runs)
4. Temperature 1000 mb
5. Temperature 500 mb
6. U velocity 850 mb
7. V velocity 850 mb
8. Vertical velocity 700 mb
9. Vertical velocity 500 mb
10. Cloud cover
11. V Velocity W-E gradient 1000 mb
12. V Velocity W-E gradient 500 mb
13. U Velocity N-S gradient 1000 mb
14. U Velocity N-S gradient 500 mb
15. Temperature 850 mb W-E gradient
16. Temperature 850 mb N-S gradient
17. Large scale rain
18. Convective rain
19. Persistence

TABLE 2 Basic set of predictors offered for screening for 96 hour one day POPA forecasts

1. Relative vorticity $\left(\frac{\partial v}{\partial x}-\frac{\partial u}{\partial y}\right)$ 1000 mb
2. Relative vorticity $\left(\frac{\partial v}{\partial x}-\frac{\partial u}{\partial y}\right)$ 500 mb
3. $\left[\text{Relative vorticity 1000 mb}\right] > 0$
4. $\left[\text{Relative vorticity 500 mb}\right] > 0$
5. Stability (T1000 - T500)
6. $\left[\text{U velocity 850 mb}\right]^2$
7. Wind speed 850 mb
8. Wind speed 850 mb * cloud cover
9. Wind speed 850 mb * stability
10. Total precipitation
11. V velocity 850 mb >0
12. $(\text{V velocity 850 mb} > 0)^2$
13. $\left[(\text{V velocity 850 mb}) + 50\right]^3$

TABLE 3 Predictors computed from basic set and offered to MDA screening program

Tables 2 and 3 demonstrate the flexibility of the interpretation software for predictor definition. Many non-linear transformations were used, and various types of differencing are available. The square and the cube of the V velocity component are examples of predictors added because of preference for the first power of V velocity in early runs. The basic set of predictors represent those created for the experiments and stored as a dataset. The computed predictors are those derived from the basic set on input to the MDA program. All predictors were averaged over 4 nearest gridpoints, and predictors for 84 hours, 96 hours, and 108 hours were offered. Those forecast times were chosen after a comparison of predictor-predictand correlation coefficients revealed highest values generally at 84 or 96 hours, but rarely before 72 hours. All predictors for the winter 96 hour experiment were averaged over 3 timesteps, the valid period of the predictand. Persistence was offered as a control, defined as the observation at model initialisation time (4 days earlier).

The predictors selected for the winter 96 hour equations are shown in table 4, along with the Mahalanobis value reached at each step, and the group means. A table such as this can be used to examine the separation of the group means for the predictors selected, and to check that the expected relationship is depicted in the data. For example, the first predictor, model total precipitation at 84 hours (centred on observation valid time) has group means increasing from 1.79mm for the category 1 cases to 10.61 for the category 3 cases. There is good separation between groups, and the means increase as the observed precipitation increases. Predictor 2, 1000 mb geopotential height tends toward negative values (low pressure area) for high precipitation events, again a reasonable distribution. Predictor 3 apparently was chosen to help separate category 3 from the other two, as was predictor 5. The V velocity predictors were probably chosen to represent the upslope contribution to precipitation in the Sira area. The fourth predictor likely was chosen to catch frontal precipitation. The final Mahalanobis value is reasonably high for a sample of this size, and probably would pass a significance test.

Predictor	Time	Mahalanobis Value	Group means 1(<1.0mm)	2(1.0-20.0mm)	3(>20.0mm)
1. Total precipitation	0	65.58	1.79	4.27	10.61
2. Geopotential 1000mb	0	82.00	93.53	54.11	-37.18
3. $[U \text{ velocity } 850mb]^2$	+24	96.60	56.30	68.85	180.3
4. T_{850} N-S gradient	+12	105.56	-1.33	-1.57	-0.86
5. $(V_{850} >0)^2$	0	111.33	42.19	49.02	93.08
6. $(V_{850} >0)^2$	+12	127.26	41.31	57.24	39.80

TABLE 4 Predictors selected for winter, 96hr forecast, one day total precipitation in 3 categories

Figure 16 is a scatter plot of the data as a function of the first two predictors selected. Category 1 events are open squares, category 2 events are black squares, and category 3 events are represented by crosses. A casual scan of the data suggests that assumption of a normal distribution within groups would not be too unreasonable, especially for category 3. The group means are nearly collinear, but the discriminant functions do not lie parallel to the line through the group means because they take into account the dispersion within groups. There are 2 discriminant functions because there are 3 groups.

For all experiments, verification of the probability forecasts was carried out using the Brier Score (Brier, 1950) and the Rank Probability Score (Epstein, 1969), and skill scores based on both. The skill score based on the Brier Score is in fact the reduction of variance for the probability forecasts because the Brier Score for climatology is the variance of the binary predictand, and it is identified as such in the results presented below. For independent sample verification, the dependent sample climatology is used in the skill score computations. The skill score based on the RPS is defined elsewhere in this volume.

For purposes of comparison with the direct model output, all probability forecasts are converted to categorical forecasts and verified using contingency tables and some of the associated scores.

Table 5 summarises the verification scores for the winter 96 hr experiment. The two skill scores show positive skill, and as expected, skill is lower on the independent data. The three scores based on the contingency tables show improvement over the direct model output, especially in the threat score. More information on the way in which the addition of other predictors has modified the direct model output (predictor 1) is available by examining the 4 contingency tables (table 6). The model, on the dependent data sample, significantly overpredicts category 2 at the expense of the other categories. The MDA procedure has largely corrected for this on dependent data by moving many events from category 2 forecasts to category 1 forecasts and to category 3 forecasts. On the independent data, the model still shows a slight tendency to overforecast the middle category, and the MDA overcorrects slightly for it. Of particular interest is the fact that, although the model predictors are different in 80/81 due to model topography differences, the MDA has still managed to use equations developed on 81/82 data to correct for category 3 underforecasting. The model never forecasts an extreme event, while the MDA procedure correctly forecasts 4 of the 9 events and gets 3 others close.

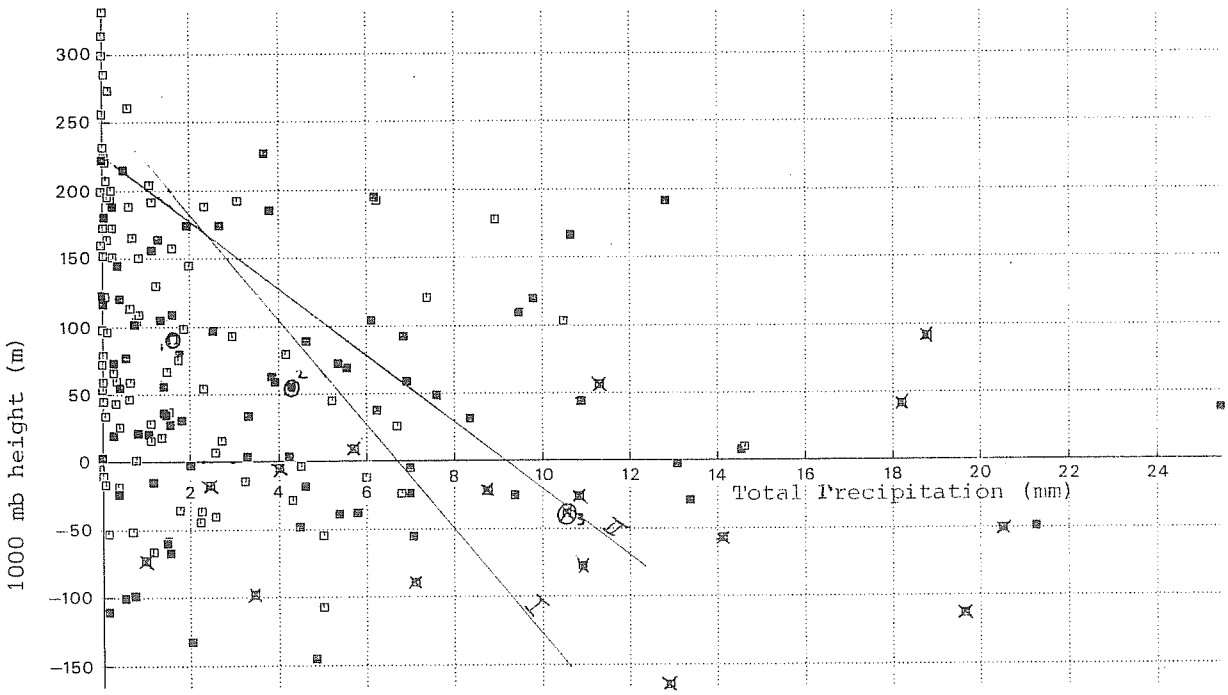

Fig. 16 Scatter plot of the winter 96 hour dependent data as a function of the first two predictors selected. Category 1 cases are open squares; Category 2 cases are black squares and Category 3 cases are crosses. Means are indicated with category symbol and circle. Discriminant functions are the two oblique lines.

Measure	Dependent		Independent	
	MDA	MODEL	MDA	MODEL
Reduction of variance (%)	19		18	
Skill score based on RPS (%)	26		24	
Contingency tables				
Percent correct (%)	65	58	64	62
Threat score (%)	49	31	42	33
Heidke skill against chance (%)	38	25	39	32

TABLE 5 Verification results for Sira, 96 hr winter forecasts

DEPENDENT SAMPLE

INDEPENDENT SAMPLE

MDA

OBSERVED

	1	2	3	
1	73	37	2	112
2	17	35	4	56
3	0	3	10	13
	90	75	16	181

Pct. correct = 65.1

Threat = .57,.36,.53

OBSERVED

	1	2	3	
1	29	11	2	42
2	9	23	3	35
3	0	7	4	11
	38	41	9	88

Pct. correct = 63.6

Threat = .57,.43,.25

MODEL

OBSERVED

	1	2	3	
1	50	19	1	70
2	40	54	14	108
3	0	2	1	3
	90	75	16	181

Pct. correct = 58.0

Threat = .45,.42,.06

OBSERVED

	1	2	3	
1	27	13	0	40
2	11	28	9	48
3	0	0	0	0
	38	41	9	88

Pct. correct = 62.5

Threat = .53,.46,0.0

TABLE 6 Contingency tables for Sira, winter forecasts, 96 hours

The MDA verification program prints out average forecast probabilities for each category for the dependent or independent sample stratified according to observed category. That is, averages are taken of the forecast probabilities for all category 1 events, then for all category 2 events and so on. If the forecast has high resolution, a high average forecast probability should appear for the verifying category and low average forecast probabilities for the others. An example of such a table is shown in table 7 for the winter 96 hour experiment. On dependent data, it can be seen that the MDA had some problems distinguishing the category 2 events from the category 1 events, while on independent data there was a problem separating category 3 from category 2.

In summary, the MDA procedure has used the direct model output, corrected for bias, and improved upon the model output for the winter, 96 hour forecast. Improvements are not large, but significant when it is considered that most improvement is in the forecasts of extreme precipitation.

		DEPENDENT SAMPLE				INDEPENDENT SAMPLE		
		AVG FCST PROB				AVG FCST PROB		
Cases		1	2	3		1	2	3
90	1	.607	.386	.007	1	.638	.357	.005
75	2	.465	.473	.062	2	.367	.468	.164
16	3	.183	.245	.572	3	.263	.376	.362

TABLE 7 A simple measure of resolution of the probability forecasts for Sira, winter forecasts, 96 hr.

2.5 Summaries of results of other interpretation trials on Sira data

A total of 16 experiments were run on the Sira data, for various combinations of summer and winter, one day total accumulations from 48 to 120 hours, averaged and point value predictors, three and four categories, and for 3 day total accumulation. Figure 17 shows a comparison of percent correct for all runs compared to the direct model output (independent data only). Points to the right of the 45° line indicate superiority for the MDA runs. Overall there is slight superiority for the MDA output; but the graph shows that with MOS, it is difficult to do very much better than the model output in the medium range, but it is also difficult to do very much worse. With stable and larger datasets, it ought to be possible to tip the balance more in favour of the MOS output by careful tuning of relationships and careful predictor selection. The one serious failure to beat the model is an example of the effect of selection of a completely unreasonable predictor due to a chance relationship in the dependent sample. The poor predictor was persistence (4-day persistence) which was selected first due to a chance 4-day cycle of precipitation

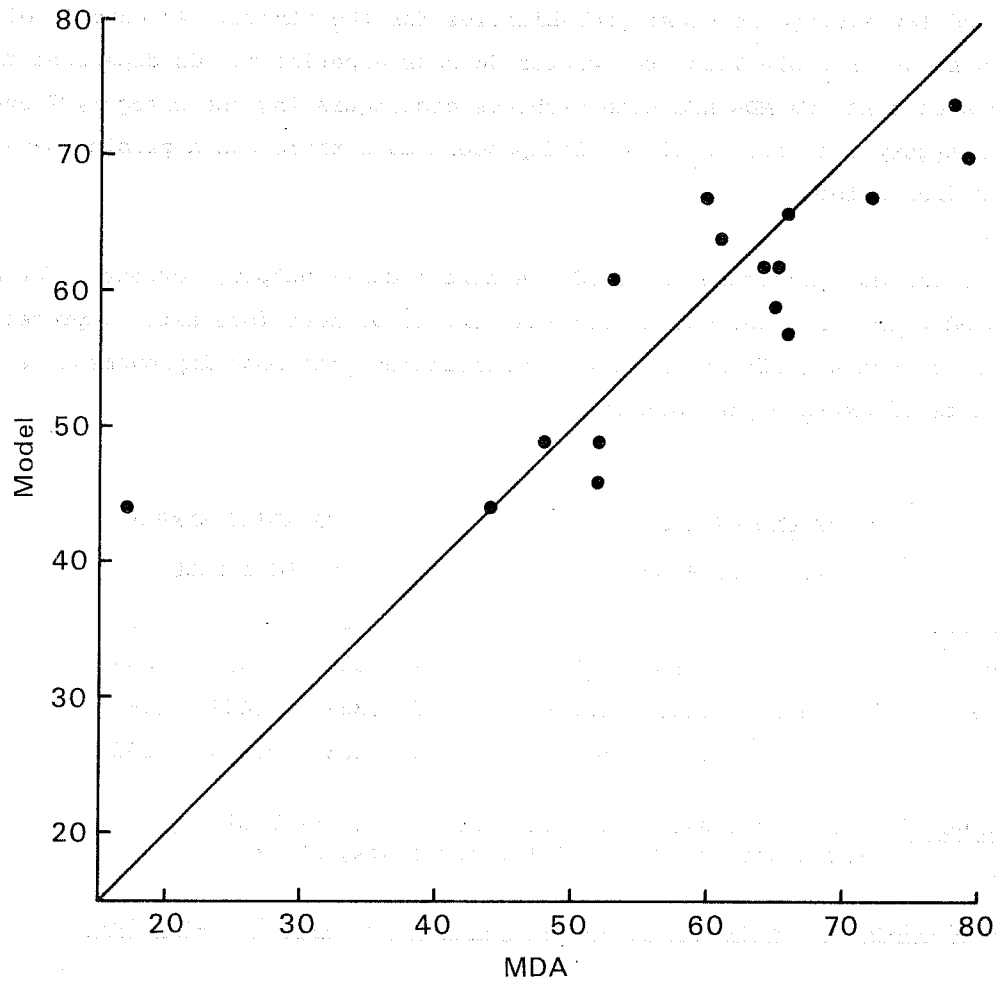

Fig. 17 Percentage correct for MDA on independent data and direct model output for all SIRA experiments. Points to the right of the line indicate superiority for MDA equations.

events in the dependent sample that was visible on a time-series plot of the dataset. After elimination of the offending predictor, the point which resulted from a rerun of the case moved to the right of the 45° line. This experience emphasises the hazards of using small datasets, and points out the need for careful pre-screening of predictors.

In general, there was consistency among the test runs in the predictor sets chosen. The model's precipitation forecast was nearly always selected as the first predictor, suggesting that there is useful information in the direct model output for forecasting one day and three day precipitation amounts out to day 5. The predictors selected tended to differ by season, but were not affected significantly when only the smoothing is changed, and also were usually not affected by changes in the forecast projection time. There was also evidence that model skill for precipitation forecasting is concentrated in the direct precipitation output. If the precipitation predictor was overlooked, it was usually difficult to find a satisfactory relationship with other predictors. In other words, it was not any easier to beat the direct model output without using it than it was by using it. The following sections highlight briefly the results of the test runs.

2.5.1 Point value vs smooth

For 96 hr winter forecast, comparison runs were done between smooth (Sm) predictor data (4 point average, 3 timestep average) and point value (Ptv) predictor data (no averaging in either time or space). The point used was the nearest gridpoint northwest of the Sira area. In both runs, the MDA screening was allowed to select the best set of predictors with a 5% cutoff criterion, and more predictors were required to reach that criterion for the Ptv run than for the Sm run.

Results of the comparison are given in table 8. The Sm run provides a better fit to the data in terms of all scores, and the gap widens when independent data is used, suggesting that the extra information contained in the specific point values is more noise than signal. Smoothing is therefore useful for the Sira area.

Measure	Dependent		Independent	
	MDA Smooth	MODEL Ptv	MDA Smooth	MODEL Ptv
Reduction of variance (%)	19	14	18	6
Skill score based on RPS (%)	26	21	24	13
Contingency tables				
Percent correct (%)	65	60	64	60
Threat score (%)	49	41	42	38
Heidke skill (%)	38	39	39	36

TABLE 8 Comparison of scores for forecasts using point value predictors with forecasts using 4 point averages of predictors. Sira, 96 hr, winter season

2.5.2 Trials on 3-day total precipitation as predictand

Three day total precipitation accumulation might be an important predictand for a user such as a power company interested in the threat of floods. It has been shown above that the Sira area can receive very large amounts of precipitation in three days, up to 110mm in one case. It was also shown that the model's under-forecasting problem extends to the three day amounts, and is, in fact, amplified in absolute terms. A statistical interpretation would be a useful asset if it could help identify some of the extreme events in a forecast mode.

Tests were carried out on both the winter and summer Sira data to forecast 3 day accumulation of precipitation for days 3 to 5 of the forecast (the 72 hour period from 48 to 120 hours). Because of the altered range of the predictand, 4 categories were used, with thresholds at 2mm, 15mm and 55mm. This left about 10 cases in the extreme category in the dependent sample. The predictors selected are shown in table 9, along with the category means for each. Reference times for the predictors are with respect to 84 hours, the midpoint of the valid period. The winter predictors selected follow the pattern set by the one-day trials, except that persistence was selected. For the summer, persistence was selected first, producing a poor equation. The case was rerun with persistence eliminated, resulting in the only run where the model total precipitation was not chosen first. Large scale rain was however chosen second. It is interesting to note that the large scale rain predictor does not discriminate well among the first 3 groups, but heavy rain requires a model forecast of relatively high large scale rain.

WINTER

	PREDICTORS SELECTED	MEANS 1	2	3	4
1.	Total precipitation (-12)	5.871	8.812	16.03	28.09
2.	Geopotential 1000 mb (0)	105.9	73.42	27.89	-60.59
3.	Temperature 1000 mb (0)	1.83	2.32	4.84	6.68
4.	Persistence	10.98	15.56	25.92	21.81
5.	$[\text{U velocity}]^2$ 850 (+36)	60.66	54.35	60.97	156.8

SUMMER

	PREDICTORS SELECTED	MEANS 1	2	3	4
1.	Geopotential 1000mb (-36)	127.1	94.31	22.56	-1.06
2.	Large scale rain (+12)	13.36	1.730	2.61	10.42
3.	U Velocity (500) (0) N-S gradient	0.765	-2.275	-2.64	-2.59

TABLE 9 Predictors selected by MDA screening program for 3 day total precipitation for Sira. There were 4 categories with thesholds 2mm,15mm,55mm Figures in brackets give predictor-predictand timing relationship in hours. (0) refers to T+84, the central point of the 3 day valid period.

Results of the trials on dependent and independent data are shown in table 10. In this case, the skill actually went up from dependent to independent data, probably due to differences in the sample climatologies. It is not likely to be due to improvements in the model, because the winter independent sample was taken before many improvements were made to the model. The comparison with the direct model output reveals noticeable superiority for the statistical technique in all cases, and in terms of all scores. This product, based on a larger sample, could therefore be quite useful in economic terms.

Measure	Dependent		Independent	
	MDA	MODEL	MDA	MODEL
SUMMER				
RV (%)	19		18	
Skill	30		36	
Contingency tables				
Percent correct	55	38	52	46
Threat	37	17	27	23
Heidke	24	6	32	25
WINTER				
RV (%)	21		19	
Skill	33		38	
Contingency tables				
Percent correct	57	42	52	49
Threat	41	22	33	27
Heidke	37	13	34	28

TABLE 10 Verification results for 3 day precipitation (48 to 120hr) forecasts for Sira

2.5.3 Summary of trials on one day accumulation, summer

Trials were carried out on one-day total precipitation from 48 hours to 120 hours as an attempt to compare model and MOS skill as a function of projection time. Early runs were made using three categories, but it was found that enough useful information was present to justify 4 categories, with thresholds at 1.0, 10.0, and 25.0mm.

The skill scores against climatology for the dependent and independent data are plotted in figure 18. The skill on both samples is positive through to 120 hours and again suggests there is useful predictive information in the model for the Sira area to at least day 5. As expected, independent sample skill is lower then dependent sample skill, and skill generally decreases with increasing projection time, albeit slowly.

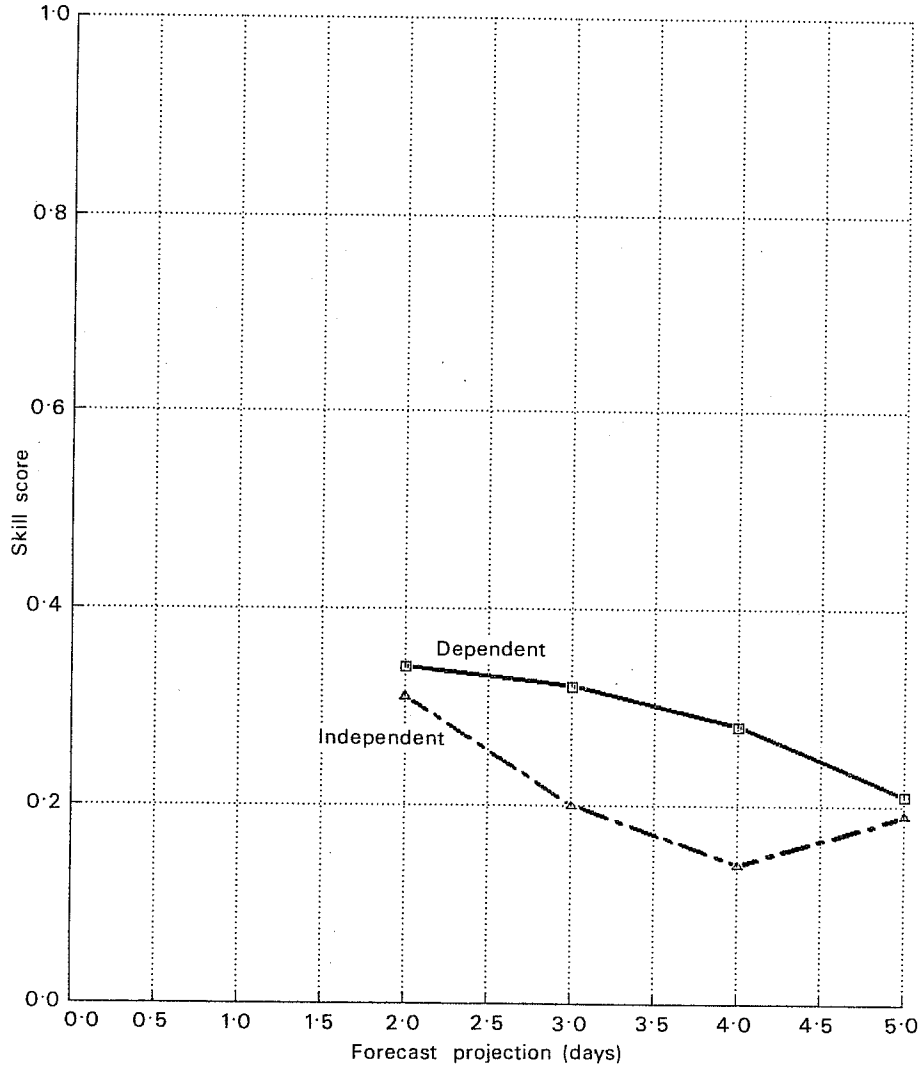

Fig. 18 Skill of summer one day precipitation forecasts for SIRA as a function of projection time.

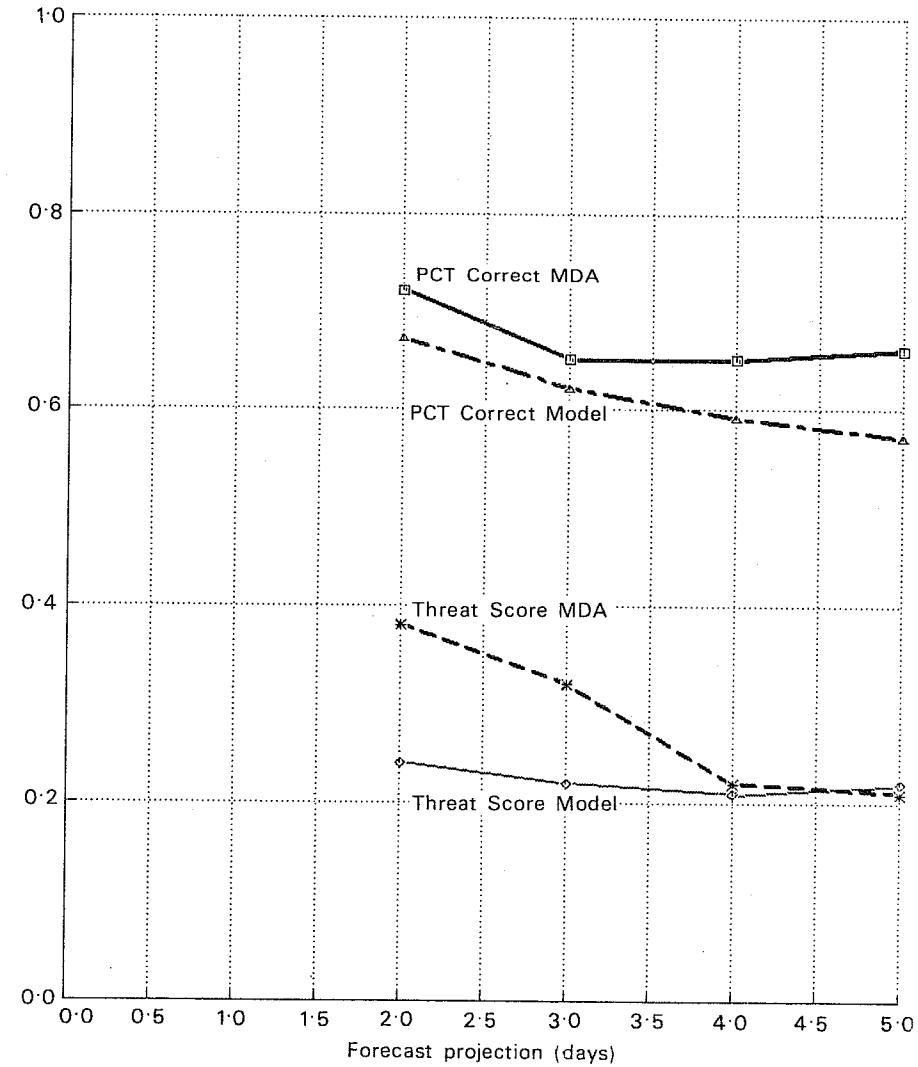

Fig. 19 Comparison of MDA and model output precipitation forecasts in terms of contingency table scores, as a function of projection time.

A comparison with model skill in terms of contingency table scores is shown in figure 19, for independent data. Again, the statistical technique is able to improve on the model output in terms of both percent correct and threat score. Again, the decrease of skill with increasing projection time is rather less than might have been expected.

3. STATISTICAL INTERPRETATION TRIALS FOR BARCELONA, SPAIN

3.1 The Problem

Barcelona is situated on the east (Mediterranean) coast of Spain. It has a low elevation, but the terrain slopes towards the northwest to the Pyrenees. The map in figure 20 shows that the city is nearly in the centre of the grid square formed by the nearest 4 gridpoints of the European archive. Only one of these points (the northwest point) is on land, in the foothills of the mountains. The model terrain indicates an elevation of about 400m. for this point, while Barcelona itself is given an elevation of 200m. As was the case for Sira, the model topography is not as steep as the actual topography.

The Barcelona dataset was of immediate interest. There are relatively few occurrences of precipitation at Barcelona, but when it occurs it is likely to be heavy. The area all along the east coast of Spain is subject to flash floods in winter, due to these relatively infrequent, but significant, rainfalls.

The synoptic situation favourable for heavy rain at Barcelona involves the existence of a sharp, deep upper trough just on or off the west coast of Spain which leads to strong surface development in the downstream south westerly or southerly winds. The low moves into the Mediterranean and occludes, producing a cutoff low which may remain in the area for some time. Essential ingredients for significant rainfall at Barcelona are easterly or southeasterly winds through a deep layer. The onshore upslope flow enhances the storm-produced rainfall, causing floods if sustained.

3.2 Datasets

For Barcelona, single station observed data was used, extracted from the regular 6-hourly synops. This data proved to be not particularly reliable, and almost half of the potential dependent and independent samples were lost due to missing observations. Total sizes were 96 events (dependent) and 55 events (independent). Approximately 90% of the events in both samples were no-rain cases ($\leq$1.0mm in 24 hours). The remaining 10% were approximately evenly split between the two rain categories, 1.0 to <20.0mm and $\geq$20.0mm. With small samples such as these, the statistical methods are not likely to be stable, especially for the rain categories,

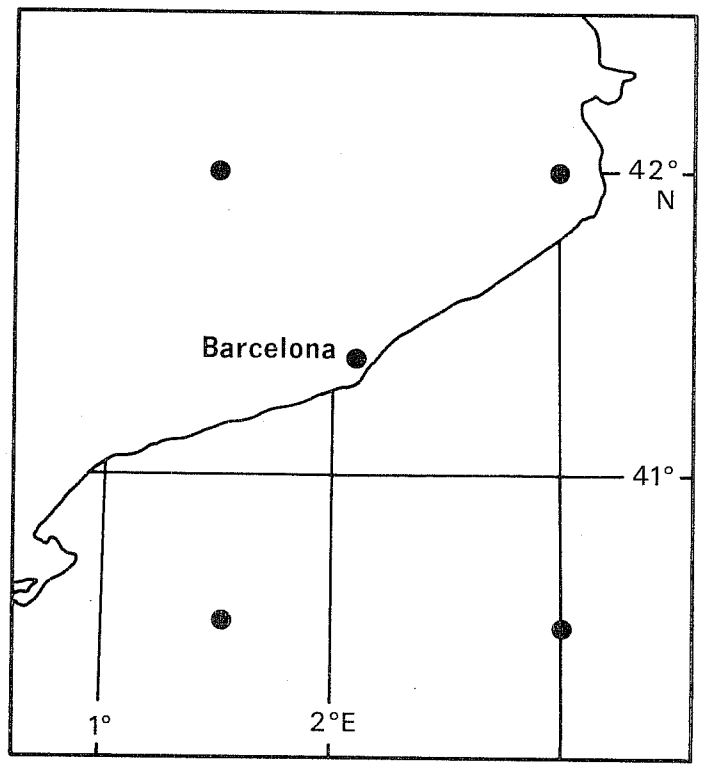

Fig. 20 Map of part of Eastern Spain showing location of Barcelona and four nearest grid points of the European archive.

and the results presented below should only be considered as an indication of the performance of both model and statistical method.

Figures 21 and 22 show the dependent and independent samples as solid lines. There are 5 events of 20mm or more accumulation in the dependent sample and 2 such events in the independent sample. The events are generally spread in time except that they are on adjacent days in the independent sample.

3.3 Model verification

Figures 21 and 22 also show, as dashed lines, the direct model output forecasts for 96 hours. As missing events have been left out of the plot, the data cannot be considered a time-series, and model forecasts that are horizontally close are not necessarily near misses in time. Forecast and actual can only be compared on a one-to-one basis at verifying time. It can be seen from the two plots that there is not much skill in the model for catching the precipitation events. Only one in each sample is caught, and even then is seriously underforecast. The value of this is obscured by the significant number of false alarms in the model forecasts in both samples. Other aspects of the model performance are shown below in comparison with the MDA test results.

3.4 Statistical interpretation trials

Several trials were run on winter data only, using different time projections and different types of smoothing. As with Sira data, it was found that smoothing did not make a great deal of difference in performance or predictors chosen, but the point value performance (northwest point) was slightly superior for Barcelona. This may be because the higher elevation of the northwest point allowed the upslope effects to be caught by the model and subsequently by the statistical technique. For comparison purposes, results of the test runs for 48 hour forecasts and 96 hour forecasts are shown below.

The predictors selected by the MDA program are listed in table 11. The model precipitation forecasts were not selected in any of the trials. The chosen predictors do represent the synoptic requirements for precipitation quite well however. For both 48 and 96 hours, most of the predictors are either vorticity or vorticity components, or wind components. The positive V component squared represents the requirements for a southerly component for extreme rainfall, and the distribution of the group means confirms the occurrence of strong positive V components for heavy rain cases. The U velocity at 850mb was selected for the 96 hour equations, with negative values (easterlies) relating to precipitation cases. It would seem, therefore, that the correct predictors are chosen according to the

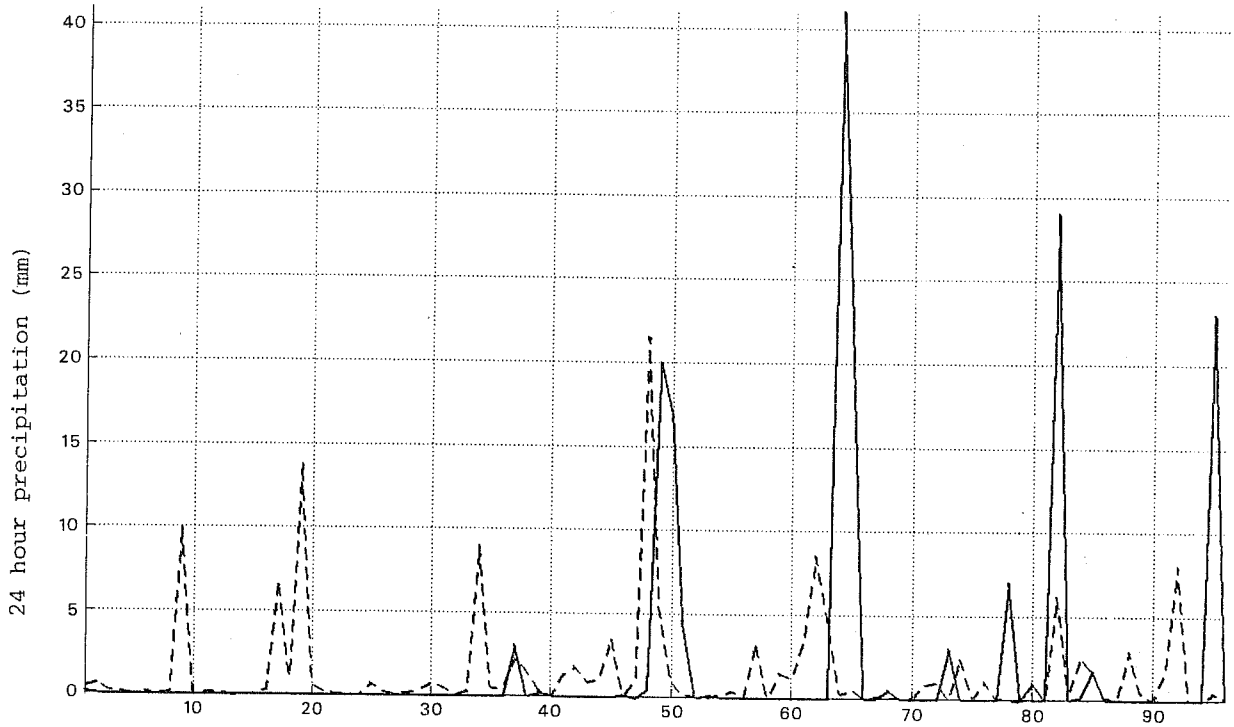

Fig. 21 Plot of dependent dataset for Barcelona. Solid line is observed one day precipitation amount. Dashed line is 72 to 96 hour direct model forecast of precipitation.

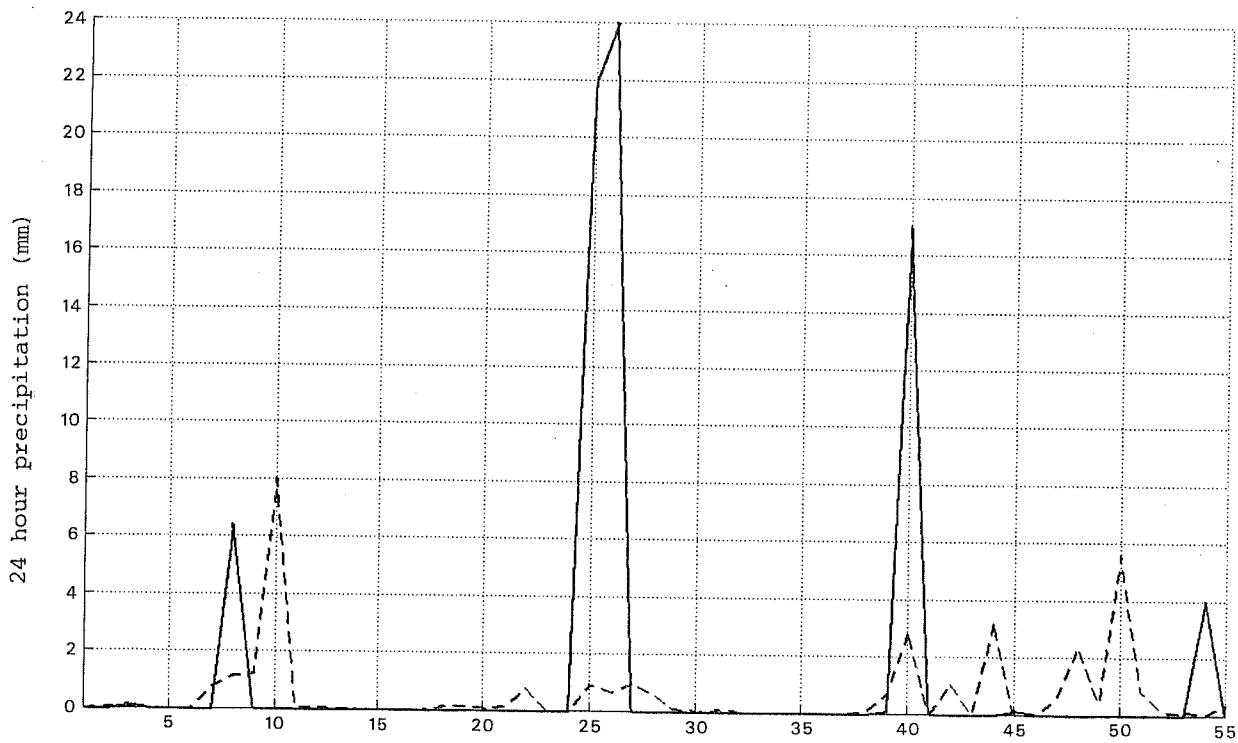

Fig. 22 Same as Fig. 21 but for Winter 80-81.

synoptic climatology described above.

48 Hour

	PREDICTORS SELECTED	MEANS		
		1	2	3
1.	V_{850}^2 >0 (0)	11.61	8.38	80.07
2.	N-S gradient U velocity	-0.007	-1.55	-11.10
3.	Vorticity (500) (+12)	-2.6	5.99	2.93
4.	U velocity (700) (+12)	4.14	-0.79	2.81
5.	Vorticity (1000) >0 (+24)	1.90	3.28	1.89
6.	Wind speed (700) (+12)	12.31	6.88	8.08
7.	Wind speed (700) (0)	12.42	8.79	8.17
8.	V_{850}^2 >0 (+12)	14.57	14.15	50.39
9.	Vorticity (1000) (+24)	-0.71	-0.31	-0.45

96 Hour

	PREDICTORS SELECTED	MEANS		
		1	2	3
1.	V_{850}^2 >0 (-12)	14.78	20.74	82.76
2.	V_{850}^2 >0 (0)	14.12	31.14	32.01
3.	Temperature (500) (+12)	-21.03	-25.19	-23.87
4.	U velocity (850) (-12)	5.02	-0.68	-7.8
5.	V_{850} >0 (-12)	2.018	2.75	6.79
6.	Convective rain (+12)	1.12	0.73	2.15
7.	Geopotential (1000) (0)	163.2	137.7	124.4

TABLE 11 Predictors selected by screening MDA program for 48 hour and 96 hour forecasts of winter precipitation for Barcelona. Predictand is one-day accumulation ending at 12 GMT. Figures in brackets give predictor-predictand timing relationship in hours. 0 means the time-averaged predictor is centred on the valid period of the predictand.

The results of runs on dependent and independent data (figure 23), however, show that these equations are not particularly satisfactory for predicting the precipitation events. Skill is positive at 48, 96 and 120 hours on dependent data, but is positive only at 48 hours on independent data. The skill score is sensitive in this case because precipitation is a rare event and the climatological forecast will be hard to beat. Nevertheless, the negative scores indicate difficulty in forecasting the probabilities of both the precipitation and no-precipitation events.

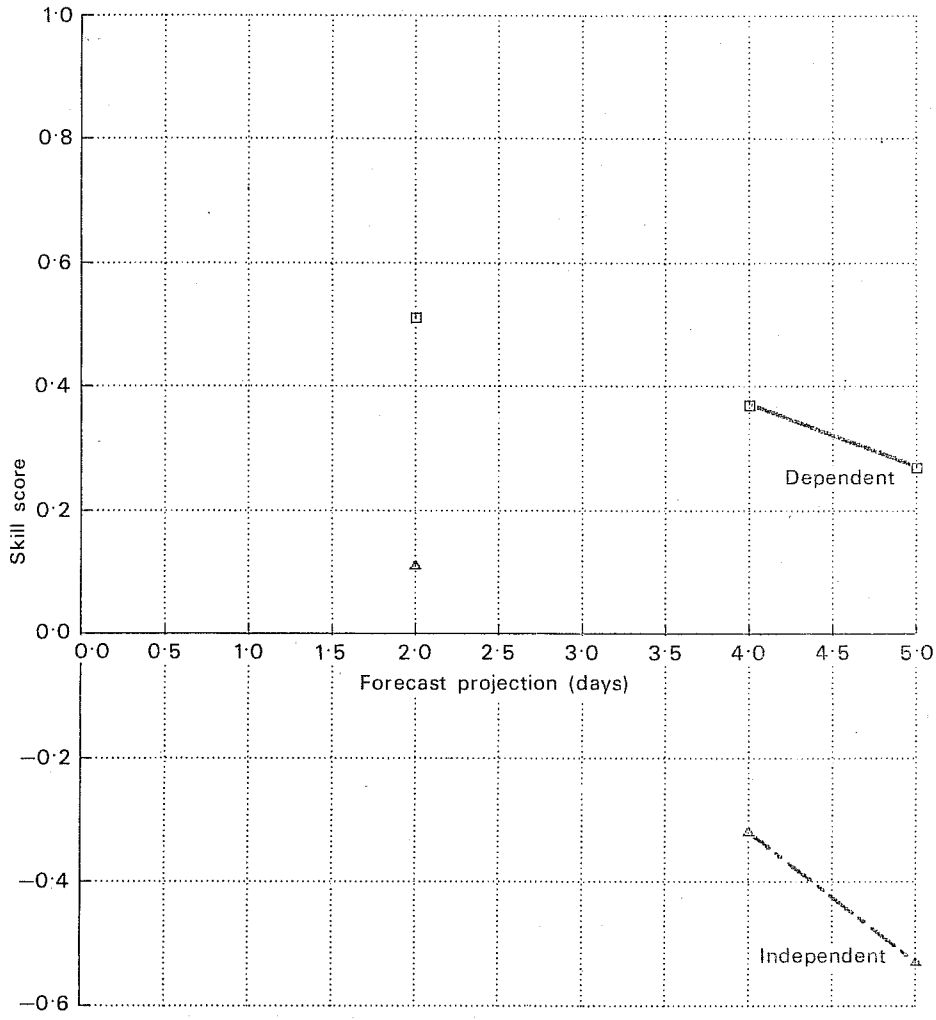

Fig. 23 Skill of probability forecasts for Barcelona one day precipitation, winter, as a function of forecast projection.

Further evidence of the behaviour of the equations can be obtained from tables 12 and 13, which give contingency tables and scores for the dependent and independent data, respectively. On dependent data, at both 48 and 96 hours, it can be seen that the model seriously overforecasts category 2, at the expense of category 1. This is the high false alarm rate referred to earlier. The MDA (top 2 tables of table 12) has clearly corrected for this by replacing most of the category 2 forecasts into category 1 without significant sacrifice to correct forecasts of category 2. Category 3 events have been fit at 48 hours, but only 2 out of 5 were successfully fit at 96 hours.

48 HOUR

MDA

OBSERVED

	1	2	3	
1	86	4	1	91
2	2	4	0	6
3	0	0	2	2
	88	8	3	99

Percent correct = 93

Threat = 0.75, 0.40, 0.50

HS = 0.60

96 HOUR

MDA

OBSERVED

	1	2	3	
1	82	5	2	89
2	1	1	1	3
3	2	0	2	4
	85	6	5	96

Percent correct = 89

Threat = .89, .12, .29

HS = 0.34

MODEL

OBSERVED

	1	2	3	
1	67	4	1	72
2	21	4	2	27
3	0	0	0	0
	88	8	3	99

Percent correct = 72

Threat = .72, .13, 0.0

HS = .15

MODEL

OBSERVED

	1	2	3	
1	61	4	3	68
2	23	2	2	27
3	1	0	0	1
	85	6	5	96

Percent correct = 66

Threat = .66, .06, .00

HS = 0.03

TABLE 12 Contingency tables for Barcelona forecasts, dependent data. MDA forecasts are top two tables and direct model output is given below HS is Heidke Skill score.

On independent data (table 13) there is still some tendency for the model to over-forecast category 2, but the effect is reduced. The MDA again corrects for this, placing most category 2 forecasts into category 1. By doing so, it achieves a high percentage correct at 48 hours, but misses all precipitation events. At 96 hours, skill is lost in category 1, resulting in a product that does not beat the model. The Heidke skill score, which takes into account forecasts correct by chance, is very low, indicating an unskilled forecast, despite the relatively high percent correct. It is evident from these results that the skill at 48 hours is positive only because of sharpening of the probability forecasts of no precipitation, not because of any ability to forecast the precipitation events.

MDA

48 HOUR

	OBSERVED			
	1	2	3	
1	50	3	1	54
2	0	0	1	1
3	0	0	0	0
	50	3	2	55

Percent correct = 91

Threat = .93, 0.0, 0.0

HS = 0.15

96 HOUR

	OBSERVED			
	1	2	3	
1	46	3	2	51
2	4	0	0	4
3	0	0	0	0
	50	3	2	55

Percent correct = 84

Threat = .84, 0.0, 0.0

HS = -0.07

MODEL

48 HOUR

	OBSERVED			
	1	2	3	
1	43	3	0	46
2	7	0	2	9
3	0	0	0	0
	50	3	2	55

Percent correct = 78

Threat = .81, 0.0, 0.0

HS = 0.05

96 HOUR

	OBSERVED			
	1	2	3	
1	44	1	2	47
2	6	2	0	8
3	0	0	0	0
	50	3	2	55

Percent correct = 84

Threat = .83, .22, 0.0

HS = 0.24

TABLE 13 Contingency tables for Barcelona forecast independent data. MDA above and Model output below

It appears from the evidence presented here that the model does not contain sufficient accurate information in any of its output to successfully forecast the relatively infrequent precipitation events at Barcelona. Although the selected predictors make sense in synoptic terms, there is perhaps too much noise in these predictors to provide a useful forecast. It may be that the model has difficulties in handling synoptic patterns that produce precipitation at Barcelona. It may also be that the data was overfit in this case. There is mounting evidence that a relatively small number of predictors provide best results for MOS in medium range forecasting. If too many are included, the additional noise, which is inevitable near the model's skill limit, may result in very unreliable equations.

4. CONCLUDING REMARKS

A data handling and statistical program set has been successfully applied to develop and test a discriminant analysis procedure on two European sites with different climatologies. The experience gained from the tests can be summarised in the following comments.

1. For some stations, there is useful information in the model's precipitation output to at least day 5 of the forecast.

2. The model output statistics procedure can improve on the direct model output by using it in combination with other predictors.

3. Although it is possible to improve on the direct model output without using it in the statistical equations, it is difficult to do so. If the direct model output is bypassed as a predictor, it seems to mean that there are model errors in synoptics etc., that affect other predictors as well.

4. It is useful to begin MOS studies on a particular station with a short range forecast projection when model errors are relatively small, to find the best predictor set to offer.

5. Much more testing and experimentation is required to determine how MOS techniques should be formulated for medium range forecasting. Issues that need to be addressed are:

   (a) stopping criteria and the number of predictors that should be included in the face of increasing model error variance.

   (b) Alternative formulations such as quasi-perfect prog where MOS equations are on short range forecasts and applied to medium range forecasts from the same model

5. REFERENCES

1. Brier, G.W., 1950: Verification of forecasts expressed in terms of probability. Monthly Weather Review, Vol. 78, 1-3.

2. Epstein, 1969: A scoring system for probability forecasts of ranked categories, Journal of Applied Meteorology, Vol. 8, 985-987.