

Evaluation of quantitative precipitation forecasts over southern South America

Andrea Celeste Saulo

Centro de Investigaciones del Mar y la Atmósfera (CIMA) (CONICET/UBA) and
Departamento de Ciencias de la Atmósfera y los Océanos, Facultad de Ciencias
Exactas y Naturales, Universidad de Buenos Aires

and

Lorena Ferreira

Departamento de Ciencias de la Atmósfera y los Océanos, Facultad de Ciencias
Exactas y Naturales, Universidad de Buenos Aires

(Manuscript received September 2002; revised February 2003)

This paper examines current status of Quantitative Precipitation Forecasts (QPFs) over southern South America and particularly over Argentina, of 36 and 60-hour forecasts corresponding to the LAHM/CIMA regional model and the NCEP global medium-range forecast model (MRF) during a three-month period. Two indexes of skill for QPFs, the Equitable Threat Score (ETS) and the bias score, have been calculated over two subregions.

Better ETS were obtained with MRF for predicting areas with precipitation intensities ranging from weak to moderate, while LAHM/CIMA performs better for higher amounts of precipitation. Both models show a tendency to a loss of accuracy as precipitation increases. Bias scores are relatively small for both models, except at larger rain limits, where LAHM/CIMA tends to overestimate the areas and MRF to underestimate. Limitations include diversity of precipitation regimes over the larger domain, and the scarcity of observations. While the former is common to similar studies, the latter becomes critical over southern South America and could affect the representativeness of interpolated fields. Generating a gridded dataset from an uneven (and coarse) observational network tends to spread precipitation leading to spurious rain in weak precipitation areas, and unreliable scores. The scores varied significantly when they were calculated against individual station observations, giving a more realistic depiction of model performance. These findings suggest that measures of skill should be applied and interpreted with caution, since different forecast verification strategies may lead to contradictory results.

Introduction

Forecast verification focusses on the use of the verification process to obtain insight into the basic strengths and weaknesses in forecasting performance.

Corresponding author address: Dr Celeste Saulo, CIMA 2do piso, Pabellón II, Ciudad Universitaria, 1428 Buenos Aires, Argentina
email: saulo@at.fcen.uba.ar

Tel: (5411) 4787 2693 Fax: (5411) 47883572

This process is an essential component of any effort to assess overall forecast quality. Since forecast quality is an important determinant of forecast value, detailed assessments of the various aspects of quality are a desirable adjunct to studies of the absolute and/or relative value of weather forecasts (Murphy 1997).

Despite the unquestionable importance of forecast verification, there are not many efforts directed to its documentation over regions such as South America, and this deficiency is even more noticeable when QPFs are considered.

On the other hand, the availability through the Internet of daily global model forecasts such as those issued by NCEP (National Centers for Environmental Prediction) or by several research and operational centres over the Americas has facilitated the usage of these products, which are, sometimes, the unique source of data with adequate areal and temporal coverage to undertake particular investigations. This broadened use of model forecasts poses many questions for forecast verification, such as how can these forecasting systems be compared in a rational manner, or how can users be advised to get the best of these predictions. Though these questions have been previously identified and, maybe, partially addressed over several regions worldwide (e.g. Saarikivi et al. 2000; Mass and Kuo 1998) it is clear that each situation may need different strategies and diverse answers or recommendations.

There are many examples of different model response arising from particular types of forecast verification. At the NCEP Environmental Modeling Center web page, the root mean square error (RMSE) of the 500 hPa height five-day forecasts from five different global models is displayed daily. One interesting result from these errors is that the various models do not perform equally well over both hemispheres, with weaker performance in the southern hemisphere. Additionally, from the models considered in that verification, the ECMWF generally exhibits the lower RMSE. To illustrate the extent to which forecast quality is dependent on area, period and variable (among many others), the study of QPFs by McBride and Ebert (2000) shows that the NCEP model (and not the ECMWF one) exhibits the best performance over Australia. Focussing in southern South America, Saulo et al. (2001) show that the NCEP model is better than the LAHM/CIMA regional model for predicting some variables but not for others.

As stated before, almost no documentation related to QPF evaluation is found for South America, except for the work of Chou and Justi da Silva (1999), who evaluated the Eta/CPTEC (Centro de Previsão de Tempo e Estudos Climáticos) model forecasts over the corresponding model domain for one year.

In order to examine the current status of QPFs over southern South America and particularly over Argentina, this paper considers a three-month period of 36 and 60-hour forecasts corresponding to the NCEP global model and the LAHM/CIMA regional model.

Besides the interest to determine the quality of their respective QPFs, other questions will be addressed, mainly those arising from the scarcity of observations that characterises the selected region. One of these questions is whether the measures of skill most widely used by forecast centres (e.g., the Equitable Threat Score (ETS) or the bias score) correctly quantify model performance over this particular region. It is expected that this research will progress the forecast verification problem over South America, which may be a first step towards the identification of limitations to the use of precipitation fields derived from coarse datasets such as the ones available here.

In the next section, we describe the selected methodology with a brief description of the models employed and the characteristics of the observed precipitation data. We then present the scores obtained for two selected subdomains. The sensitivity of the results to changes in the interpolation method applied to the observed precipitation data set is then analysed, together with a comparison against individual station data. The final section is devoted to the discussion of results.

Methodology for verification of QPFs and brief model descriptions

The improvement in numerical weather prediction in the last few decades has led to extensive use of many of its products. Consequently, adequate tools to estimate their quality had to be developed. One of the most requested of these products, given its impact upon economic activities, is the amount of precipitation. The evaluation of QPFs has experienced significant progress since the first related work by Anthes (1983). Some of the papers devoted to QPF verification over the United States are those of Anthes (1989), Schaefer (1990), Olson et al. (1995), Mesinger (1996) and Johnson and Olsen (1998), while fewer can be found over other regions (e.g., Chou and Justi da Silva (1999) over part of South America and McBride and Ebert (2000) over Australia).

While, as stated in the introduction, the results for different regions cannot be strictly compared, the selection of the same measures of skill as those adopted by previous studies, allows a better judgement of the results. For this reason, two of the most commonly used indexes of skill have been adopted in this work: the Equitable Threat Score (ETS) and the bias score. The ETS as defined by Schaefer (1990) is:

$$ETS = \frac{H - F \frac{O}{N}}{O - [H - F (1 - \frac{O}{N})]} \quad \dots 1$$

where H is the number of hits, F and O are the number of grid points where precipitation equal to or above a given threshold has been forecast and observed respectively, and N is the total number of points verified. Mesinger (1996) describes the ETS as the fractional number of correctly forecast points of a precipitation event above random, normalised by the total number of observed and forecast points, also above the number of hits in a random forecast. □

This score is similar to the Threat Score in Anthes (1983) except that the expected number of 'hits' in a random forecast ($E = F \cdot O / N$) is subtracted both from the numerator and the denominator.

This measure of skill is, in general, complemented with the bias:

$$\text{BIAS} = \frac{F}{O} \quad \dots 2$$

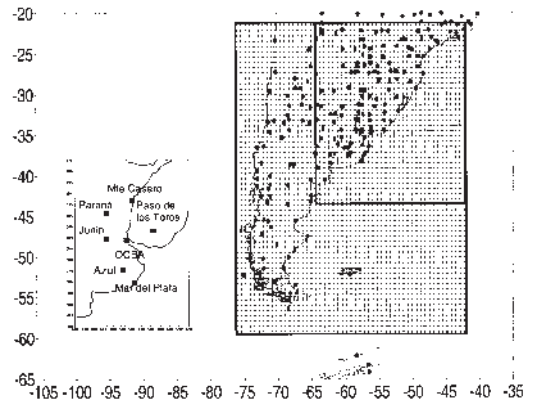
in order to evaluate if the model overestimates or underestimates the number of points with precipitation above a certain threshold.

From the expressions above, a perfect forecast is represented by an ETS and a bias = 1, while a random forecast will result in an ETS = 0. The bias score defined in this way has a singularity when observed precipitation equals zero. Another drawback of the selected scores is that neither of these indexes is designed to measure the model ability to correctly forecast the 'non-occurrence' of precipitation.

It is expected that the inclusion of E in the ETS could take into account the differences between wet and dry regimes occurring over model domain. However, it is not possible to completely remove the undesired effect of distinct precipitation regimes and of the model's own climatology, as discussed by Woodcock (1976). Ideally, these scores (and particularly the ETS) should be calculated over areas with similar precipitation regimes. In addition, their dependency on the spatial smoothness of the verification analysis, which in turn, is a function of the volume of data available per verification box, should be taken into account. These are clear restrictions to the validity of the aforementioned scores. Nevertheless, they have been widely used over entire model domains (i.e. the requisite of similar precipitation regimes is not fulfilled). This can be partially justified by the necessity of operational forecast centres to provide an overall measure of model skill with a single number.

In this work, similarly to the methodology applied by McBride and Ebert (2000) and Chou and Justo da Silva (1999), the scores were calculated over all the region of interest (see Fig. 1), that encompasses the continental region south of 20°S, and a subregion indicated in Fig. 1, analogous to the Río de La Plata Basin (see Garcia and Vargas (1996) for a precise def-

Fig. 1 Verification gridded domain (larger rectangle), with the La Plata basin area, denoted by the smaller rectangle. To the left, an enlargement of part of the La Plata basin, with positions of particular stations identified.



inition of this basin), where the theoretical conditions for ETS computation are more closely met. It is believed that this domain partition will help both to evaluate model performance in a broad way (despite network and regional limitations) and to know, with higher confidence on the methodology, QPF quality over an area with a critical economic and societal impact within the Americas.

The models to be verified are the Aviation Run (AVN) of the Medium Range Forecast (MRF) global model from NCEP and the LAHM/CIMA regional model. In both cases, 24-hour accumulated precipitation corresponding to the 12 to 36-hour and 36 to 60-hour forecasts (resulting from forecast cycles initialised at 0000 UTC) for a three-month period starting at 1 October 2000 and ending at 31 December 2000 are employed. The MRF global spectral model is widely used by the meteorological community, and is extensively documented. The reader is referred to Kalnay et al. (1998); and also to the NCEP's Environmental Modeling Center web pages (www.emc.ncep.noaa.gov/modelinfo/index.html) for a detailed description of model characteristics. The AVN products used in this work are available daily through the Internet with a 1°x1° horizontal resolution. The model's performance over southern South America has recently been evaluated in Saulo et al. (2001), where it was shown that 24 and 48-hour forecast quality is good.

The LAHM/CIMA regional model was originally used for research purposes and since 1998 it has been run daily to provide up to 72-hour forecasts over South America. This effort is a central part of a project aimed at the design of a forecast system adequate to fulfill local users' requirements. This forecast cycle

uses 0000 UTC NCEP operational analyses as initial conditions, while the AVN run is used to provide boundary conditions at 12-hour intervals. Model behaviour has been tested under a variety of case studies (Orlanski et al. (1991), Menendez (1994), Nicolini and Saulo (1995), Seluchi and Saulo (1998) among others) and also for the provision of daily forecasts (Saulo et al. 2001). The LAHM/CIMA regional model is hydrostatic and currently runs with a $0.65^\circ \times 0.65^\circ$ horizontal resolution, and 18 fixed-sigma levels in the vertical with a horizontal domain that encompasses continental South America and surrounding oceans south of 4°S .

The observational data set has been provided by the Argentine National Weather Service and consists of daily accumulated precipitation available from the regional surface synoptic network, which includes a total of 300 stations from Argentina, Uruguay, Paraguay, Chile and Brazil (denoted by dots in Fig. 1). The dataset is coarser than those used by similar studies such as Mesinger (1996), which has 10 000 rain gauge stations over the US, or McBride and Ebert (2000) which includes around 6000 stations over Australia. In both cases, the networks were denser by a factor greater than eight, even taking into consideration the smaller areal coverage of the present work.

According to several studies, precipitation forecasts usually improve with increasing resolution. In particular, previous studies carried out with the LAHM/CIMA regional model show that its performance over southern South America is better at 0.65° resolution than at lower resolutions (Seluchi and Saulo 1998), in terms of position, intensity, rain amounts and timing of relevant synoptic features associated with the analysed events. In order to calculate the scores, both model outputs and the observational data set had to be interpolated to a common grid. Taking into consideration that one critical issue for the design of a forecast system is the improvement of local precipitation forecasts, and given the interest in evaluating if the regional model higher resolution had any impact on precipitation forecasts, it was decided to use the regional model grid size as the common one, i.e., all the information was remapped to a $0.65^\circ \times 0.65^\circ$ horizontal resolution.

There are different interpolation techniques to obtain a regularly spaced dataset from an irregular distributed network. In this case, the Kriging method (e.g., Davis 1973) was used. Also, an alternative dataset based on a modified Cressman (1956) scheme (Glahn et al. 1985) applied to the same rain-gauge data has been used for comparison. This dataset is distributed through the Internet by NCEP and is increasingly being adopted as representative of observed daily precipitation.

Verification over the selected regions

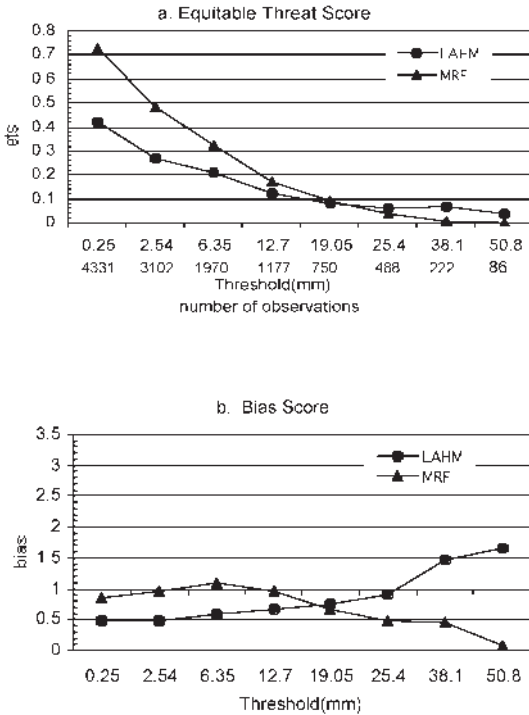
Southern South America

The selected period corresponds to the austral spring, a season characterised by important synoptic activity over southern South America. Frontal passages and cyclogenesis are the most significant phenomena affecting this region during this part of the year, both leading to large amounts of precipitation, mainly over the subregion denoted as the La Plata basin. North of 30°S , convective activity, both organised and isolated, is frequent during these months. Particularly for the year 2000 spring, the number of days with precipitation above 0.1 mm, was above the climatological mean, while the accumulated precipitation for the whole period was approximately 30 per cent larger than the mean value over subtropical areas. This behaviour is related to the enhanced frequency of convective systems observed during this period and introduces an additional complexity for QPFs.

The ETS and bias have been calculated for eight thresholds which have been standardised as in Mesinger (1996): 0.25 mm, 2.54 mm, 6.35 mm, 12.7 mm, 19.05 mm, 25.4 mm, 38.1 mm and 50.8 mm. Figure 2 shows the results for the larger domain averaged over the three month period, and includes the amount of data considered at each threshold in the abscissas. Only 36-hour forecasts are included in this figure since no significant differences are found with respect to the 60-hour forecasts, except for slightly smaller scores in the latter. Better ETS were obtained with MRF for predicting areas with precipitation intensities ranging from weak to moderate, while LAHM is better at higher amounts of precipitation, indicating some advantage of the higher resolution for the representation of heavily precipitating systems. Both models show a tendency to a loss of accuracy as precipitation increases, which is an expected result from the definition of the scores. The MRF seems to provide useless rain area forecasts (i.e. ETS = 0) for amounts above 38.1 mm. Bias scores are relatively small at lower thresholds particularly for the LAHM model. At larger rain limits, LAHM tends to overestimate the area and MRF to underestimate it. An interesting aspect is the very large divergence between models, mainly above 25 mm. This implies that there are major deficiencies in the operational models tested for the simulation of heavy rain events. Individual inspection for each month coincides with the average behaviour.

Perhaps one of the most remarkable differences between these scores and those calculated at other regions is the high score for predicting areas with weak rain (Mesinger (1996), evaluating a six-month period of global forecasts, obtained an ETS around

Fig. 2 The (a) ETS and (b) BIAS scores for LAHM/CIMA and MRF 12 to 36-hour forecast precipitation from October to December 2000 over the larger domain.



0.3 over the US, and McBride and Ebert (2000), an ETS below 0.4 over two Australian subregions). In contrast, the NCEP model exhibits an inferior performance over the southern hemisphere compared with that over the northern hemisphere, which seems to be inconsistent with the result obtained for these QPFs.

The high score at the reduced amounts of rain can be largely induced by the use of interpolation to create the observed precipitation grid. The interpolation smooths the original fields and creates spurious precipitation. The model would not be penalised in the case of small displacement of the predicted precipitation regions and this would result in higher scores (Chou and Justi da Silva 1999). This kind of problem could benefit the MRF more than the LAHM/CIMA, since the latter has a bias (Fig. 2(b)) towards under-prediction of rain at smaller thresholds. Further discussion of this effect is included in the following section, but, to examine whether this could be the case, a simple experiment in the interpolated observed precipitation data set was performed: at random days, the ratio between the number of grid-points with an

amount of precipitation above 'x' (over an specific area) and the number of grid-points being verified, was calculated. This ratio was compared with the one corresponding to the number of stations that observed an amount of rain above 'x' (over that same area) and the total number of stations. Clearly, this comparison is valid when the gauge spatial distribution is nearly homogeneous, and this could not be assured over the complete domain but only over smaller areas that were subsampled to perform this control. The result of this analysis was that these ratios were not similar for thresholds below 5 mm, i.e., a property of the observed dataset was not conserved in the interpolated dataset for weak rain values. It is worth mentioning that when the experiment was done over the whole domain which has a non-homogeneous gauge distribution, the ratios were even more dissimilar.

This preliminary inspection indicates that verification procedures applied at these ranges using interpolated data, should be interpreted cautiously.

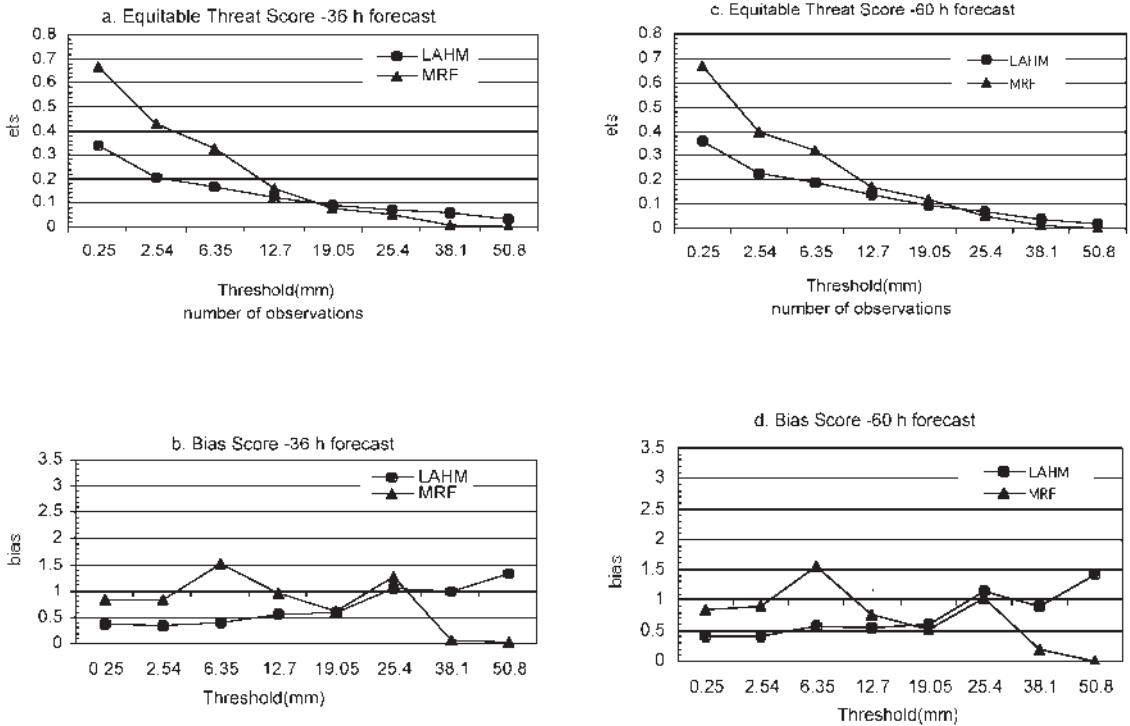
The La Plata basin

In the previous sections, some arguments have been provided to justify the selection of this subregion. Besides the interest in evaluating models over an area that is important both for its economic impact and for the increasing amount of research focussing on it, it is worth mentioning that the subregion has a denser network and more homogeneous precipitation regimes.

Figure 3 shows the average ETS and bias for the La Plata basin corresponding to 36 and 60-hour forecasts. There are not considerable differences between the behaviour of ETS at this subregion compared with that over the entire domain, except for somewhat lesser scores, which indicates the intrinsic difficulty that characterises QPFs over this particular area and/or corroborates that scores tend to be more realistic when there are more observational data. Models compare in the same manner, with MRF showing better skill at weak and moderate rain limits, and LAHM/CIMA slightly better performance for higher thresholds. There is almost no change in the ability of both models to predict precipitation with increasing forecast length. This may be another indication of the lack of sensitivity of these scores to capture observed differences in model performance, since in Saulo et al. (2001) it was shown that there was a quality loss for longer periods in all the analysed variables.

While evaluating bias scores (Fig. 3(b) and (d)), a similar response is detected as over the larger domain for LAHM/CIMA, though the tendency to overpredict at higher amounts of rain has been diminished. This is not the case for the MRF forecast which does not exhibit a similar skill and provides an almost useless precipitation forecast at higher thresholds.

Fig. 3 The ETS and BIAS scores for LAHM/CIMA and MRF from October to December 2000 at the La Plata basin: (a-b) 12 to 36-hour forecasts and (c-d) 36 to 60-hour forecasts.



Alternative verification procedures

Comparison with different gridded data

Through the previous sections, some limitations in the selected methodology have been mentioned, and have been mainly related to the characteristics of the observational network used for verification. Up to this point, ETS and bias scores, calculated over two domains, indicate that MRF performance is better than LAHM/CIMA except at higher precipitation amounts, the last being an encouraging result for future QPF improvement through the use of mesoscale models. Additionally, as evaluated by the ETS, the MRF over this region predicts areas with low rain amounts better than at any other region documented in preceding papers. Though this could be a fortuitous result just representative of the period chosen, other causes should be inspected since, as pointed out, the period was characterised by above normal convective activity and rain, which suggests that models would not be favoured during the particular three-month interval used.

One alternative to check the results obtained in the previous section, is to calculate the same scores but using as 'ground truth' an alternative precipitation field. The one employed here is that generated by the Climate Prediction Center (CPC) and freely available through the Internet. It should be noted that the network used by this analysis is similar (i.e., comprises the same surface synoptic stations) to the one considered in the interpolation obtained by the Kriging method used above. The corresponding results for the larger domain, again corresponding to 36-hour forecasts, are shown in Fig. 4, and do not display major differences to those obtained previously (see Fig. 2), at least in the ETS. Biases are larger and suggest a different response of LAHM/CIMA that now seems to overpredict areas even with weak and moderate rain intensity (opposite to the results shown in Fig. 2(b)). In turn, MRF overestimates areas with rain at these thresholds even more than Kriging. Comparing Figs 2(b) and 4(b), larger biases could be related to a tendency in the Cressman-generated field to produce less precipitation than the Kriging-generated one. Surprisingly, taking into con-

Fig. 4 As in Fig. 2 but using Cressman interpolated data as the observed field.

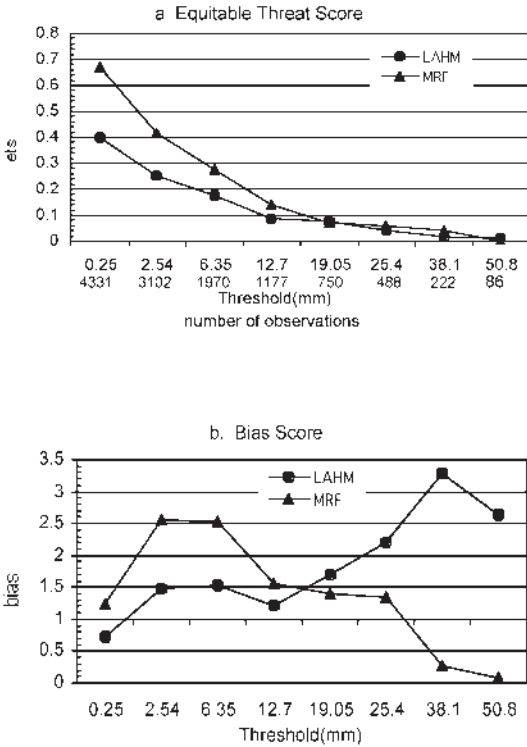
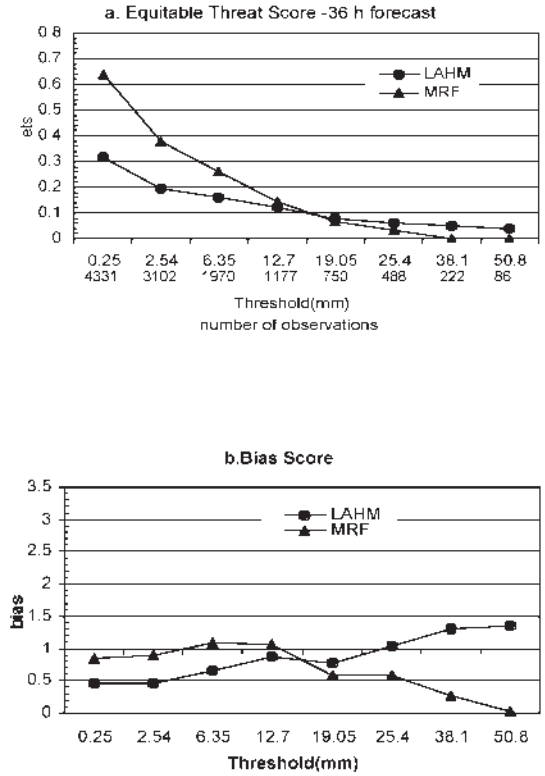


Fig. 5 As in Fig. 2 but using a 1.95° resolution verification grid.

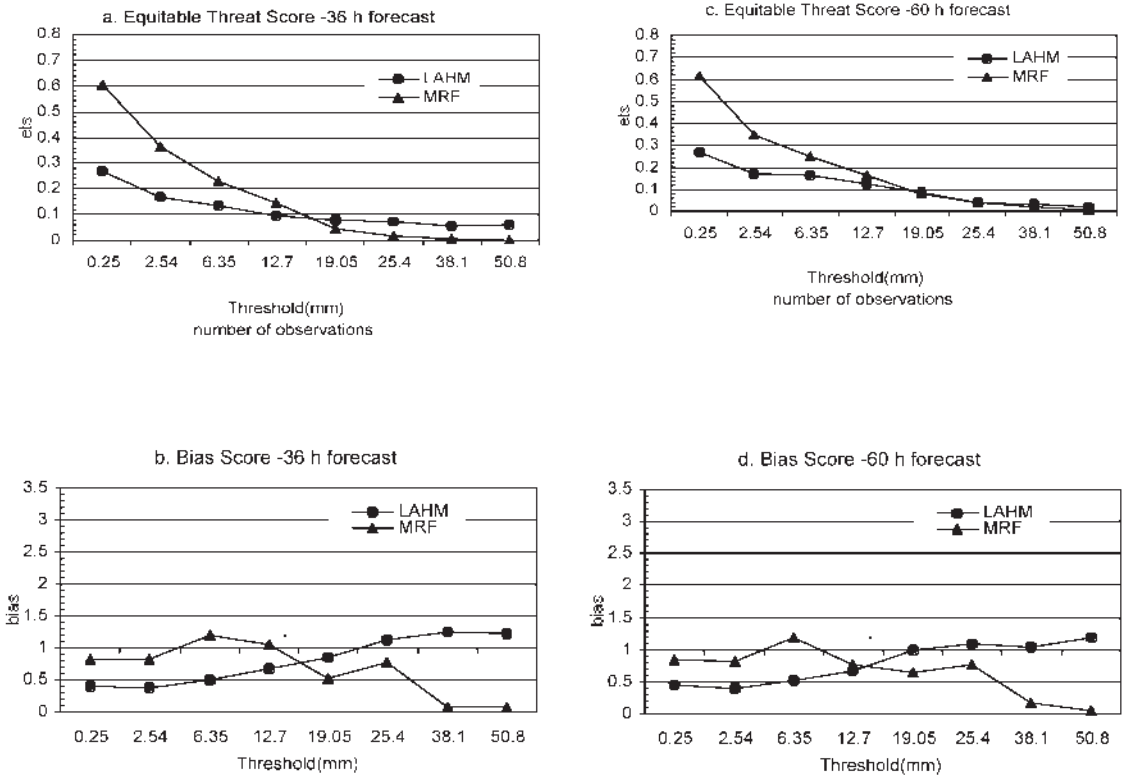


sideration that Cressman uses MRF analysis as the first guess field, it is not evident from Fig. 4 that this leads to a better scoring of MRF.

Given that the spatial distribution of the gauge data may not be adequate to support performing an objective analysis at 0.65 degree grid spacing, it was decided to calculate the same scores but remapping model output and the observed data (again interpolated using the Kriging method), to a 1.95° horizontal grid spacing. Comparing the scores in Fig. 5 with those in Fig. 2 a similar model response at the various thresholds can be recognised, except for smaller ETS particularly at lower rain amounts and lesser biases at higher ones. Over the La Plata basin (Fig. 6) the impact of decreasing interpolated data resolution upon the scores is similar, but exhibits a weaker loss in ETS values. The results are consistent in the sense that interpolation resolution mainly diminishes bias at higher thresholds (i.e., the heaviest rain peaks have been smoothed). On the other hand, provided that more rain gauge data are available over the La Plata basin, the lower resolution of the analysis affects the results to a lesser extent.

To explore the idea of how these forecasts work compared with gridded precipitation analyses and also to get some insight into the spatial structure of this field, three-month accumulated precipitation fields for the period are shown in Fig. 7. Comparing LAHM/CIMA forecasts (Fig. 7(c)) with any of the observations, it is seen that they overestimate the area with higher rain intensity over northeastern Argentina and southern Brazil, confirming the results with the bias score. Both models tend to overestimate precipitation over the northern portion of the La Plata basin (see the area encircled by the 400 mm contour). Over areas with poor data coverage, such as highlands in the border region between Bolivia and Argentina and near the Andes around 30°S, there are large amounts of precipitation forecast by both models that are not detected in the observations. Nevertheless, given the lack of observations in this region, it is difficult to decide whether the large amounts of forecast rain in this area are realistic or not. Finally, both models fail to represent the secondary maximum observed at Buenos Aires province (centred around 37°S, 60°W in Fig. 7(b)). In general, more similarities can be recog-

Fig. 6 As in Fig. 3 but using a 1.95° resolution verification grid.



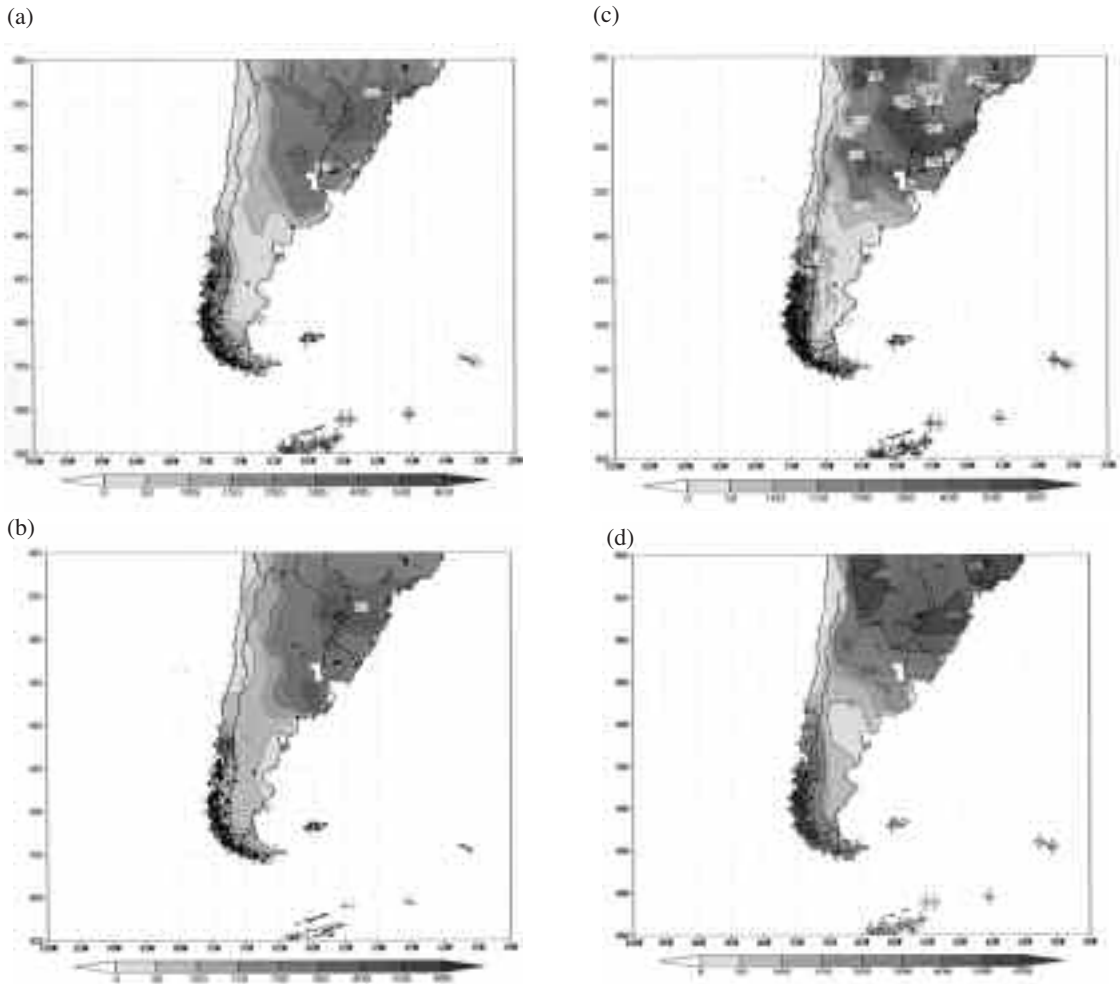
nised between model forecasts than between any model and the observed precipitation field, while it is evident that LAHM/CIMA fields provide more detail of the spatial structure. Both are expected results, the former probably being linked to the fact that LAHM/CIMA takes its boundary conditions from MRF forecasts and both models are initialised with the same analyses.

One clear disadvantage of the interpolated fields is apparent in Fig. 7(b): the relatively large amounts of precipitation to the east of the Andes south of 40°S. This precipitation pattern suggests that both interpolation methods spread observed rainfall from the heavy rainfall region west of the Andes into the region east of the Andes which is known to be dry. Therefore, both models are more realistic in this region than the interpolated observations. Nevertheless, it should be mentioned that the Kriging method does a better job in this area than the Cressman method.

Both analyses denote a similar spatial structure in the accumulated field, except for the larger amount of rain in the Cressman-analysed field that leads to differences of up to 100 mm over particular areas, with

larger discrepancies over Patagonia (lowlands in Argentina to the south of 40°), and north of 30°S, mainly over the area where the maximum amounts of accumulated precipitation were reported (not shown). It should be mentioned, however, that the three-monthly accumulated field calculated with Kriging (Fig. 7(a)) has been obtained by interpolating the three-month accumulated precipitation data at each station, and not by accumulating the day-by-day interpolated field. This was done in order to minimise the spurious rain generation. While this comparison may look unfair between methods (this could not be done with the Cressman-analysed field, since it is provided each day), it has been designed to highlight the effect that any interpolation technique will produce in the accumulated field, generating larger amounts of rain than observed when integrated over a larger period. The deficiencies denoted above are also clear when the accumulated field that corresponds simply to the sum of the daily Kriging-interpolated field is used (not shown). In this case, it is observed that the Kriging accumulated field is somewhat larger than the Cressman one, confirming the differences in the bias score discussed above.

Fig. 7 Accumulated precipitation from October to December 2000: (a) observed (interpolated using the Kriging method), (b) observed (interpolated using Cressman's method), (c) LAHM/CIMA 12 to 36-hour forecast and (d) MRF 12 to 36-hour forecast.



Comparison with station data

In order to complement the assessment of how these models represent QPFs, independently of the drawbacks evident in both interpolation techniques, daily accumulated precipitation at individual stations located inside the smaller domain has been compared with forecast values at the nearest grid-point and the results will be shown in term of ETS and bias, as in the previous section.

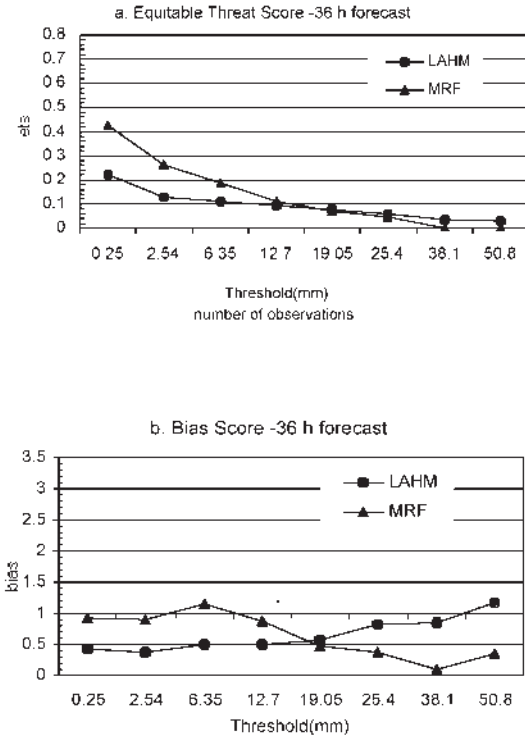
Figure 8 shows more realistic ETS values for both models in the La Plata basin, confirming the hypothesis that lower thresholds were affected spuriously by the data interpolation technique. The interpretation of model performance and how they compare is still valid, in the sense that MRF is better at lower rain amounts and LAHM/CIMA at higher ones. Bias score seems to be less affected by the use of individual sta-

tion data and LAHM/CIMA lower biases are confirmed through this alternate verification procedure. The weak sensitivity of these scores to increasing forecast length still holds (not shown).

Taking into consideration the strong sensitivity of forecast verification to the applied methodology, it is of interest to show a detailed analysis of precipitation forecast quality for individual stations over the La Plata basin. In what follows, we will concentrate on some of these stations that illustrate the mean behaviour over the central part of the La Plata basin, in which we are particularly interested (see Fig. 1 for their location).

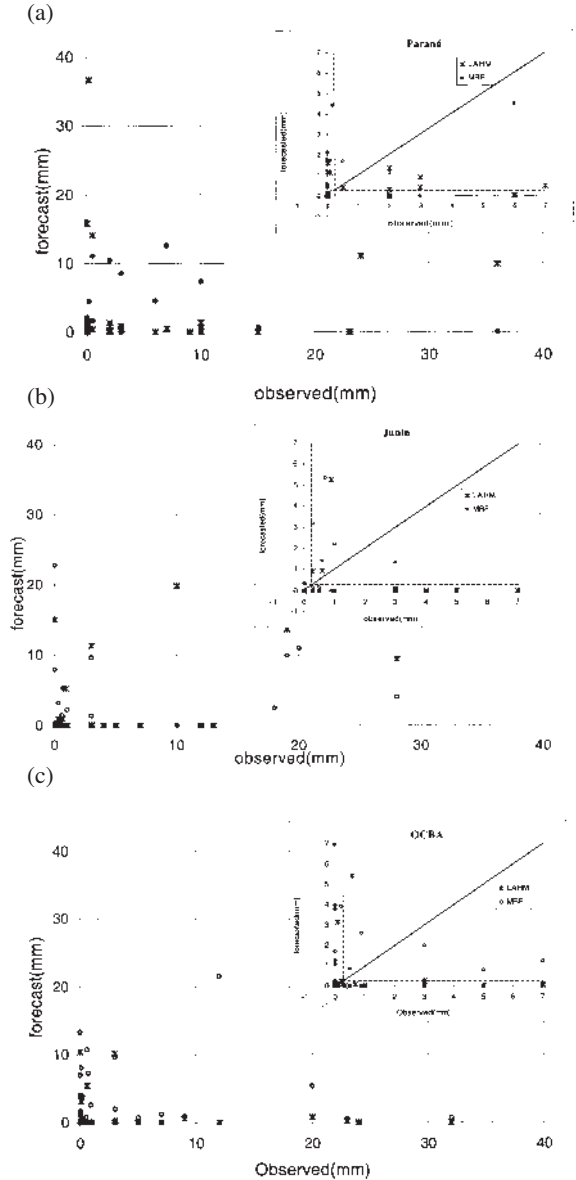
A simple diagram showing forecast rain in the ordinates and observed values in the abscissas at some of these stations is included in Fig. 9. This type of diagram was selected in order to detect to what

Fig. 8 The ETS (a) and BIAS (b) scores for LAHM/CIMA and MRF 12 to 36-hour forecast precipitation from October to December 2000 over the La Plata basin using daily accumulated precipitation at individual stations.



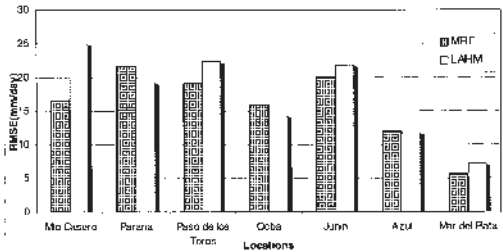
extent the large values of ETS found over this sub-region, indeed correspond with "very good" forecasts at random points. Following results described earlier, it could be expected that MRF points lie close to the diagonal, mainly for small amounts of rain. This behaviour is not clearly recognised in the figures, even at lower precipitation values (the 0.25 mm threshold has been highlighted, since ETS is calculated when observed precipitation is above that amount). In addition, a large dispersion is evident around the "perfect forecast" for both models, which does not allow a clear conclusion about their accuracy. This may be explained by the fact that ETS measures model ability to forecast precipitation above a given threshold over an area, but does not quantify by how much the actual forecast value differs from the observation. Nevertheless, these figures suggest that the skill measured by ETS may not be appropriate for some applications.

Fig. 9 Observed vs forecast precipitation for (a) Paraná, (b) Junin and (c) OCBA (see Fig. 1 for their location): circles – MRF, crosses – LAHM. The zooming corresponds to amounts between 0 and 7 mm (the 0.25 mm value is highlighted in both axes).



The root mean square error (RMSE) can be used as a suitable tool to quantify the aforementioned forecast dispersion and consequently to complement the evaluation of model performance that has been previously discussed in terms of ETS and bias. RMSE at selected stations over this three month period is presented in Fig. 10. Both models' behaviour looks sim-

Fig. 10 Precipitation RMSE for seven stations (see Fig. 1 for their location) from October to December 2000: dotted – LAHM/CIMA 12 to 36-hour forecast, hatched lines – MRF 12 to 36-hour forecast.



ilar, with greater differences (and larger errors) for stations located to the north, which could be due to the enhanced convective character of precipitation at this latitude. Again, it is hard to decide which model exhibits the best performance.

Summary and discussion of results

As stated in the introduction, forecast verification is a critical issue in any effort to assess forecast quality which, in its turn, is a basic stage to determine whether forecasts are of value or not. Among the variety of model products that are currently available, this study concentrates on the verification of QPF quality of two different models over southern South America for a three-month period. The MRF model is used worldwide, and its skill over diverse regions has been calculated and reported in many related papers. The other model included in this study is the LAHM/CIMA regional model, which is run daily to provide short-range forecasts over southern South America.

There are different alternatives to evaluate QPFs, but, in order to compare the results with those obtained in related studies, two of the most commonly adopted scores were selected: the ETS and the bias score. These indexes were calculated for a three-month period encompassing the austral spring over two subregions. The results indicate that MRF better predicts areas with weak and moderate precipitation, but LAHM/CIMA does better at higher thresholds. While both models biases are close to one, they increase at higher thresholds, displaying an opposite trend: MRF to underestimation and LAHM/CIMA to overestimation. The results were similar over both

subdomains, though the La Plata basin scores were slightly lower. The comparison of 36 and 60-hour forecasts did not show significant differences, a result that does not agree with previous evaluation of these same models (Saulo et al. 2001). It is believed that this is another indication of the low sensitivity of the adopted scores to capture observed differences in model performance. The fact that the regional model seems to provide better forecasts at higher thresholds is an encouraging result from the perspective of our ulterior objective: the design of a forecast system adequate to fulfill local users requirements. It is expected that the augmented resolution together with the use of more sophisticated mesoscale models, will lead to improved QPFs, mainly in highly precipitating events, and their implementation will be a key aspect of future development.

The unusually high ETS obtained in this study compared with those obtained over other regions (particularly at low rain thresholds), motivated different experiments to determine the dependence of this score on the interpolation technique, on the interpolated data resolution, and on the use of individual station data.

In order to evaluate whether the results obtained could be affected by the interpolation technique adopted to generate the ‘observed’ precipitation field, an alternative data set to that derived with the Kriging method, was employed. After performing the same calculations using a Cressman-generated precipitation field, similar conclusions could be derived, except for higher biases found for the Cressman-interpolated field.

When the interpolated data resolution was dropped to 1.95 degrees, the results showed a trend towards lower ETS as expected, but the values at small rain amounts remained unrealistically high. Two limitations for the correct use of the selected scores were highlighted: the diversity in the precipitation regimes over the larger domain, and the scarcity of observations. While the former is common to other similar studies, the latter becomes critical over southern South America and could also affect the representativeness of the interpolated fields.

To bypass the limitations arising from verification against gridded precipitation data, the scores were calculated using individual station data. This option produced more realistic scores and should be recommended when the region is characterised by coarse and inhomogeneous networks as in this case.

While the period length does not guarantee a statistical significance of the results, which should be validated over longer time periods, preliminary conclusions from the variety of experiments carried out to assess QPFs over this region are:

- despite the method employed to generate a gridded data set from an uneven (and coarse) observational network, the tendency to spread precipitation leads to a spurious rain generation that predominantly modifies areas characterised by weak precipitation. This alters the values of the scores calculated at each day, which at low thresholds may not be reliable. If interpolated fields are summed for various days this deficiency is accentuated, and the use of accumulated fields obtained in this way is not recommended.
- Given that measures of skill are intended to compare model behaviour it is highly desirable to adopt widely used indexes, although they should be applied and interpreted with caution. Moreover, different forecast verification strategies may lead to complementary and/or opposite results. This study shows that the ETS score gives an unrealistic measure of model performance, and is significantly affected by the characteristics of the data set employed as 'real data'.

From this evaluation, alternative methods used to examine QPF quality (such as RMSE or accumulated precipitation subjective comparison), confirmed some of the conclusions drawn from using ETS and bias, but not universally. The fact that neither the ETS nor the bias consider the success for predicting the non-occurrence of precipitation, may partially account for such differences, while the artificially generated rain by interpolation techniques may explain the unrealistic high MRF scores at lower thresholds. In brief, it may be concluded that the usefulness of each of these forecasts should be determined by the requirement of the user, since the alternative approaches adopted for their verification did not show which is 'the best forecast' conclusively. In any case, it is important to keep in mind that no single verification measure can assess all potentially relevant aspects of forecast quality.

Acknowledgments

This work has been supported by ANPCYT-PICT 04447, ANPCYT-PICT 06671, UBACYT X055 and IAI-CRN055. The anonymous reviewers are also acknowledged for their comments which helped to improve the work.

References

- Anthes, R.A. 1983. Regional models of the atmosphere in middle latitudes. *Mon. Weath. Rev.*, *111*, 1306-35.
- Anthes, R.A., Kuo, Y-H, Hsie, E-Y, Low-Nam, S. and Bettge, T.W. 1989. Estimation of skill and uncertainty in regional numerical models. *Q. Jl R. met. Soc.*, *115*, 763-806.
- Arakawa, A. and Schubert, W.H. 1974. Interaction of a cumulus cloud ensemble with large-scale environment, Part I. *J. Atmos. Sci.*, *31*, 674-701.
- Chou, S.C. and Justí da Silva, M.G.A. 1999. Objective evaluation of ETA model precipitation forecast over South America. *Climanálise*, *1*, Vol. *14*.
- Cressman, G.P. 1959. An operational objective analysis system. *Mon. Weath. Rev.*, *87*, 367-74.
- Davis, J.C. 1973. *Statistics and data analysis in geology*. J. Wiley Ed., New York, 550 pp.
- Fels, S.B. and Schwarzkopf, M.D. 1975. The simplified exchange approximation: A new method for radiative transfer calculations. *J. Atmos. Sci.*, *32*, 1475-88.
- García, N.O. and Vargas, W.M. 1996. The spatial variability of runoff and precipitation in the Rio de la Plata basin. *J. Sci. Hydrol.*, *41*, 279-99.
- Glahn, H.R., Chambers, T.L., Richardson, W.S. and Perrotti, H.P. 1985. Objective map analysis for the local AFOS MOS Program. *NOAA Technical Memorandum NWS TDL 75*, NOAA, US Department of Commerce, 34 pp.
- Johnson, L.E. and Olsen, B.G. 1998. Assessment of quantitative precipitation forecasts. *Weath. forecasting*, *13*, 75-83.
- Kalnay, E., Lord, S.J. and McPherson, R.D. 1998. Maturity of operational numerical weather prediction: Medium range. *Bull. Am. met. Soc.*, *79*, 2753-69.
- Mass, C. and Kuo, Y.-H. 1998. Regional real-time numerical weather prediction: Current status and future potential. *Bull. Am. met. Soc.*, *79*, 253-63.
- McBride J.L. and Ebert, E. 2000. Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia. *Weath. forecasting*, *15*, 103-21.
- Menendez, C.G. 1994. Análisis de un ciclón subtántico. *Meteorologica*, *19*, 33-42.
- Mesinger, F. 1996. Improvements in quantitative precipitation forecasts with the ETA regional model at the National Centers for Environmental Prediction: the 48 km upgrade. *Bull. Am. met. Soc.*, *77*, 2637-49.
- Murphy, A.H. 1997. *Economic Value of Weather and Climate Forecasts*. Edited by R. W. Katz and A. H. Murphy, Cambridge University Press, Cambridge, United Kingdom, 222 pp.
- Nicolini, M. and Saulo, A.C. 1995. Experiments using LAHM/CIMA model over Argentina in convective situations: Preliminary results of precipitation fields. *Programme Weather Prediction Research Rep. No. 7*, WMO/TD, No. 699, World Meteorological Organization, 5pp.
- Olson, D.A., Junker, N.W. and Korty, B. 1995. Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Weath. forecasting*, *10*, 498-511.
- Orlanski, I. and Katzfey, J.J. 1987. Sensitivity of model simulations for a coastal cyclone. *Mon. Weath. Rev.*, *115*, 2792-821.
- Orlanski, I., Katzfey, J.J., Menendez, C. and Marino, M. 1991. Simulation of an extratropical cyclone in the southern hemisphere: Model sensitivity. *J. Atmos. Sci.*, *48*, 2293-311.
- Saarikivi, P., Söderman, D. and Newman, H. 2000. Free Information Exchange and the future of European Meteorology: A private sector perspective. *Bull. Am. met. Soc.*, *81*, 831-6.
- Saulo, A.C. and Nicolini, M. 1995. Inclusión de la difusión vertical en un modelo regional de pronóstico: efecto sobre la precipitación. *Meteorologica*, *20*, 25-36.
- Saulo, A.C., Seluchi, M.E., Campetella, C. and Ferreira, L. 2001. Error Evaluation of NCEP and LAHM Regional Model Daily Forecasts over Southern South America. *Weath. forecasting*, *16*, 697-712.
- Schaefer, J.T. 1990. The critical success index as an indicator of warning skill. *Weath. forecasting*, *5*, 570-5.
- Seluchi, M. and Saulo, A.C. 1998. A sensitivity study to tune a limit-

ed area hydrostatic model over eastern South America. World Meteorological Organization - CAS/JSC Working group on numerical experimentation. Research activities in atmospheric and oceanic modelling. *Report No. 27 WMO/TD-No. 865*, 5.51-5.52.

Woodecock, F. 1976. The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Weath. Rev.*, 104, 1209-14.

