340

Verification of precipitation forecasts over the Alpine region using a high density observing network

Tiziana Cherubini [1], Anna Ghelli and François Lalaurette

Operations Department

[1:] Current affiliation: University of L'Aquila, Italy

Submitted toWeather and Forecasting

August 2001

340

Verification of precipitation forecasts over the Alpine region using a high density observing network

Tiziana Cherubini [1], Anna Ghelli and François Lalaurette

Operations Department

[1:] Current affiliation: University of L'Aquila, Italy

Submitted toWeather and Forecasting

August 2001

Verification of precipitation forecasts over the Alpine region using a high density observing network

Tiziana Cherubini [1], Anna Ghelli [2] and François Lalaurette [2]

[1] *University of L'Aquila*
[2] *European Centre for Medium Range Weather Forecasts*

Abstract

The demand for verification of forecasting systems to ascertain their strengths and weaknesses is increasing dramatically as models evolve more rapidly. Precipitation forecasts have always been of great interest to forecasters as they influence daily life. The recent flooding over Europe has also shown how important it is to know how models can reproduce these events. In this paper we address the issue of precipitation verification, starting from the assumption that model spatial scales have to be verified against data representing similar scales. Only in this way we are able to determine the skill of our forecasting system. A high-resolution observing network over the Alpine region has been used to reconstruct the observed precipitation field. It contains smoothed small-scale variability and represents with sufficient accuracy the average behaviour of the observed field in the model grid box. It is shown that verification against irregular and scattered observations are highly influenced by the variability of the precipitation in a grid-box. A precipitation analysis is, therefore, important if model skills have to be defined.

1. INTRODUCTION

Forecast verification is the process of determining the quality of forecasts through the assessement of the degree of similarity between forecasts conditions and observation conditions (Murphy and Winkler 1987; Murphy 1993).

Verification processes should give an insight into weaknesses and strengths of forecasting systems, thus allowing a more complete use of the information contained in the forecasts. Forecast verification encompasses many different methodologies to define the quality of the forecast, using either a deterministic or a probabilistic approach. For an extensive review see Katz and Murphy (1997).

In this paper we address the question of verification of deterministic forecasts against observations. The traditional way to verify precipitation forecasts using data available on the Global Telecommunication System (GTS) is liable to misinterpretation, as models predict precipitation on scales different from the observed ones.

Moreover, interpolation of model fields to station locations is necessary to compare forecasts to observations; this process does not create new information, it only increases the spatial precision of the field meanwhile introducing further uncertainties (Skelly and Handerson-Sellers 1996; hereafter referred to as SHS1996). Furthermore, the interpolation methods commonly used assume that the underlying field is continuous. This assumption is not generally true of precipitation fields.

Ghelli and Lalaurette (2000, hereafter referred to as GL) analysed model performance using a high-resolution observing network over France. They found that gridded observations better represent the grid-box behaviour described by the model. The question of whether general circulation model (GCM) results pertain to grid-point or

grid-box area has been largely discussed especially in the context of climate change. SHS1996 pointed out that when dealing with variables that are implicit areal (as they result from sub-grid parameterisations like convection, precipitation, radiation, etc.), GCM outputs should be treated as areal quantities. A 'grid-point approach' is probably most suitable when dealing with finite difference or spectral methods that produce point rather than areal values. Kim *et al.* (1984) and Karl *et al*. (1990) have treated simulated precipitation as areal quantity. Alternatively Wigley and Santer (1990) and Wilson and Lettermaier (1996) have accepted for the simulated quantities the 'grid-point approach'. In the present paper, simulated precipitation is treated as areal quantity.

The high-resolution precipitation data used in this study have been collected as part of the Mesoscale Alpine Programme (MAP) (Binder *et al.* 1996), during the Special Observation Period (SOP). The up-scaling procedure is as used in GL, and it consists of a simple averaging procedure of all the observations contained in a model grid box.

The up-scaled observations are used to verify ECMWF precipitation forecasts over the Alpine region. The verification period starts on September 8, 1999 and ends on November 16, 1999. Two areas have been taken into consideration:

1. a region extending from 42N to 50N and from 0E to 20E (hereafter referred to as MAP-LARGE);

2. an area centred on the Alps from 44N to 48N and from 7E to 14E (hereafter referred to as MAP-ALPS).

Verification against SYNOPs has also been carried out for both the areas during the same period. Three different measures of model skill are examined to assess the two verification approaches.

The paper is organised as follows: section 2 describes briefly the ECMWF model, the observation dataset specifications and the up-scaling techniques used; verification procedures are discussed in section 3. Results are presented in section 4. In section 5 conclusions are drawn.

2. MODEL OVERVIEW AND DATA SPECIFICATION

2.1 The ECMWF model

At the time of this study the ECMWF Global Circulation Model has spectral horizontal resolution TL319.

A detailed description of the model can be found in Simmons *et al.* (1989) and Ritchie *et al.* (1995). During the verification period the model vertical resolution changed (on October 12, 1999) from 50 (Untch *et al.* 1999) to 60 levels (Teixeira 1999; Jakob *et al.* 2000). The increased number of levels doubles the vertical resolution below 1500m and brings the lowest level down to 10m above the surface as opposed to 33 m above the surface in the previous 50-level version.

Stratiform and convective clouds are represented with a prognostic cloud scheme (Tiedtke 1993). Clouds are generated by large-scale ascent, diabatic cooling and boundary layer turbulence. They are dissipated through evaporation, turbulent mixing with unsaturated environmental air and precipitation processes. Convective precipitation, vertical momentum fluxes, temperature changes in the atmosphere due to release of latent heat or

cooling in connection with evaporation, are calculated in the convective scheme. The sub-grid vertical fluxes of mass heat water vapour and momentum are computed at each model level using a simple mass flux model interacting with its environment. The scheme is applied to penetrative convection, shallow convection and mid-level convection (Gregory *et al.* 2000; Tiedtke 1989).

In October 1999 a new parameterisation scheme has been introduced in the model (Jakob and Klein 2000). It differs from Tiedke's original in that the formulation of the precipitation/evaporation explicitly accounts for the vertical distribution of cloud layers. A more detailed description of the model's parameterisation can be found in Teixeira (2000).

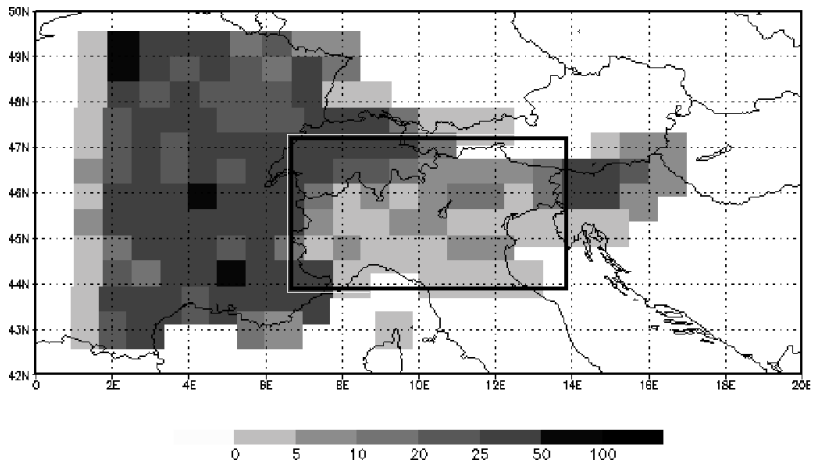

Figure 1: Station density for 0.5 x 0.5 grid boxes, averaged over the whole period under investigation. The rectangle define the area selected for the MAP-ALPS experiment.

2.2 Verification data

High-resolution precipitation observations (from the MAP dataset) are available at the MAP Data Centre (Hirter and Richner 1996) and consist of precipitation reports collected by National and Regional Meteorological Centres for about 4000 stations (synoptic, automatic and climatological).

The subset of the MAP dataset used in this study covers the period September 8, 1999 to November 17, 1999 (Bougeault *et al.* 1998; Frustaci *et al.* 2000). The precipitation observations are accumulated over 24h from 06UTC. A simple quality control (QC) has been carried out on the data to remove possible suspect reports.

The mean and median of local precipitation distributions, built using all the stations comprised within a 30-km radius around each grid point, are calculated. Whenever mean and median are more than a fixed threshold apart, the station reports correspondent to such local distribution are flagged (thresholds used: 25, 50 and 100 mm). Flagged reports are discharged or kept after an analysis of satellite imagery and a consistency check with neighboring stations.

Fig. 1 depicts the mean spatial distribution of the station density for the Alpine region considered in this study. The highest station density is on the French and Swiss side of the Alps, while the Italian side is less sampled. The

Slovenian network also provides a good coverage. Data for the Northeast side of the Alps were not available at the time of this study. The box in Fig.1 defines the area selected for the MAP-ALPS experiment.

Each station of the high-density observing network is assigned to a gridbox and an average of all the observations within each box is calculated. The up-scaled observations (hereafter referred to as USO) are assigned to the relative grid point and used to verify the model precipitation forecast. Other averaging methods, i.e. function of the station-gridpoint distance and function of precipitation intensity, have been tested and the results showed little or no dependence on the averaging method chosen.

The up-scaling is applied if there are more than three stations in a gridbox. This threshold has been chosen to compromise between the increased accuracy of the verification analysis as the station density increases (Mullen and Buizza 2000) and the loss of information on the Alpine area for higher thresholds, as the station density is not so high in the Alps (Fig. 1).

Fig. 2 shows the precipitation distribution for a chosen day as reported by the high-resolution observing system (panel a) and the distribution after up-scaling (panel b). The up-scaled field maintains the characteristics of the initial field but the very small-scale variability is lost in the averaging procedure.

The precipitation values obtained from the stations available on the GTS are accumulated over a 24-hour period from 6 UTC. These observations are hereafter referred to as GTS-SO. In order to compare the GTS-SO to the forecast value four grid points surrounding the station location are chosen and the values linearly interpolated to the station location itself. The GTS-SO distribution is presented in Fig.2 (c).

Table 1 shows, for each area and dataset, the frequency of occurrence of the observed precipitation values equal or greater than a fixed threshold. GTS-SO and USO frequency of occurrences, for each threshold and area, are different. The difference becomes very small for large threshold, as the sample size is small. Moreover Kuiper's statistical test, a variation of the Kolmogorov-Smirnov's test, (Kuiper 1962, Stephens 1970) applied to the GTS-SO and USO distributions shows that the two distribution are indeed statistically different. This has cautioned us to draw conclusion on the superiority of one verification approach versus the other. They represent two different methods and one of them (GTS-SO) suffers from ‘representativity’ problems, as discussed in the introduction.

		Precipitation thresholds (mm)					
		0.1	1.0	2.0	4.0	8.0	16.0
MAP-LARGE	USO	0.57	0.38	0.31	0.24	0.14	0.07
	GTS-SO	0.43	0.29	0.25	0.18	0.12	0.06
MAP-ALPS	USO	0.51	0.34	0.28	0.23	0.18	0.10
	GTS-SO	0.50	0.29	0.26	0.21	0.14	0.07

Table 1: Frequency of occurrence of observed precipitation events equal or greater than a fixed threshold, for the MAP-LARGE and MAP-ALPS area

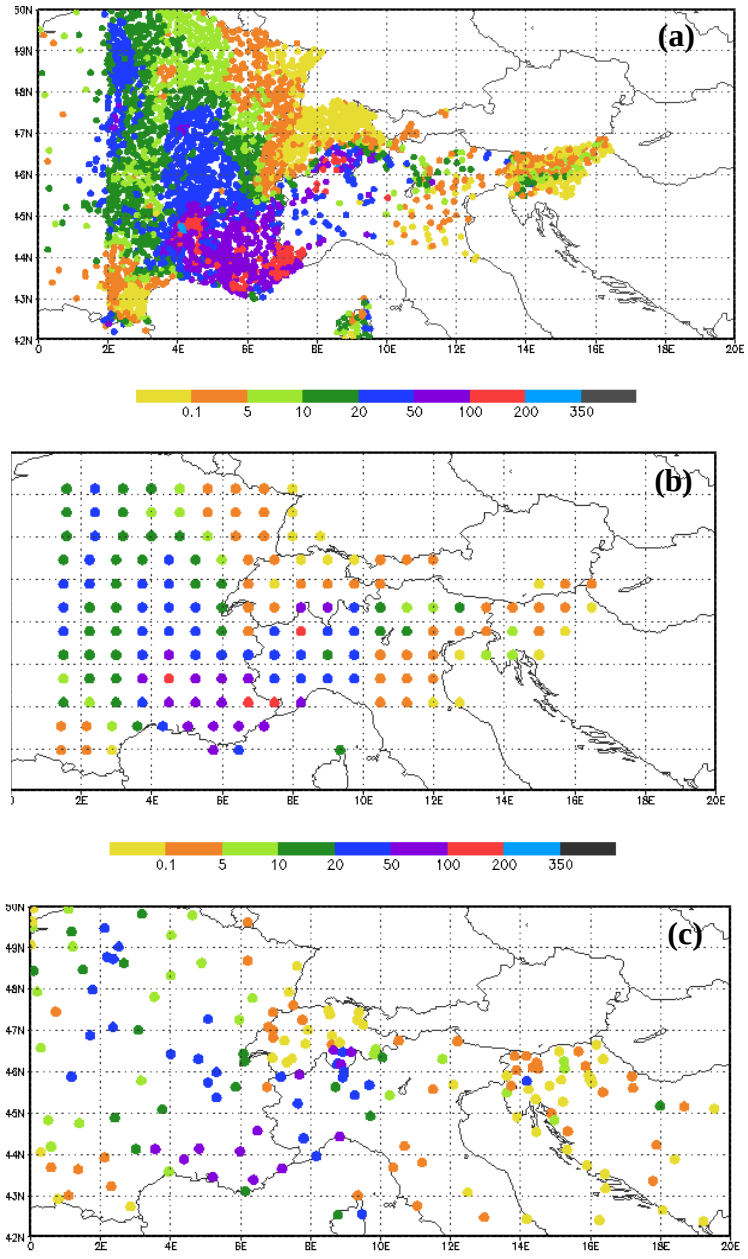

Figure 2: Distribution of the 24 h accumulated precipitation on September 20, 1999 (06 UTC) as observed from the high-resolution network stations (a), and after up-scaling technique has been applied (b). Panel (c) represents the same field as observed from the GTS-synop stations.

3. VERIFICATION SCORES

In this section the accuracy of a deterministic forecasting system in a dichotomous situation (yes/no) is under investigation. Murphy and Winkler (1987) and Murphy (1991) state that, while testing the performance of an individual forecast system, it is important to remember the complexity and dimensionality of the verification problem. Therefore, enough accuracy measures have to be used to fully estimate the value of a forecast. In this

study three measures are used to evaluate the skill of the ECMWF model forecasts. Following GL, contingency tables for different thresholds have been built for both the USO and SO.

Table 2 is an example: *a* is the number of correct forecasts of a precipitation category, *b* is the number of forecasts incorrectly predicting precipitation, *c* is the number of forecasts failing to predict an observed event, and *d* is the correct forecast of no precipitation.

The precipitation thresholds chosen for this study were 0.1, 1, 2, 4, 8, 16 mm / 24 hours. The forecasting system has been verified against the two data-sets using the Frequency Bias (FBI), the Equitable Threat Score (ETS) and the Hanssen-Kuipers Discriminant (also called True Skill Statistics, TSS). A detailed explanation of such scores can be found in Wilks (1995).

The measures used are written as:

$$FBI = \frac{(a+b)}{(a+c)}$$

$$ETS = \frac{a - R(a)}{(a+b+c-R(a))}$$

where:

$$R(a) = \frac{(a+b)\cdot(a+c)}{(a+b+c+d)}$$

and finally:

$$TSS = \frac{ad - bc}{(a+c)\cdot(b+d)}$$

The FBI measures the event frequency with no regard for the forecast accuracy. Its value is one for a perfect forecast, larger (smaller) than one if the system is over-forecasting (under-forecasting). The ETS (Schaefer 1990) is a modified version of the Threat Score rendered equitable by taking away the random forecast (R(a)). Therefore, a chance forecast will score zero as well as a constant forecast. A perfect forecast will have ETS equals to one. The TSS can also be written as the probability of detection (a/(a + c)) minus the probability of false detection (b/(b + d)). Like in the ETS, the random and constant forecasts receive a zero score, while a higher score is obtained if a rare event is forecast correctly.

	Observed YES	**Observed NO**
Forecast YES	a	b
Forecast NO	c	d

Table 2: Contingency table for observed and forecast precipitation categories. The symbols a, b, c, and d are referred to in the text.

4. RESULTS

4.1 The MAP-LARGE experiment

In this section the 12UTC forecasts (range t+42 and t+66) for the period September 8, 1999 to November 16, 1999 are verified against the GTS-SO (dashed line) and the USO (solid line). Fig. 3 represents the mean observation value relative to each forecast category. The forecasts are divided into categories (from 5 to 35mm/24h every 5mm/24h). For each forecast category the observations pertaining to the couple forecast-observation are averaged. This gives an indication on the under/over-estimation of the rainfall amounts. The diagonal represents the perfect forecast; values above (below) the diagonal indicate under (over)-forecasting. Both ranges, t+42 (panel a) and t+66 (panel b), show that the forecast overestimates precipitation for values larger than 10mm/24h. An average overestimation of 8mm/24h is assigned to a forecast (range t+66) of 30mm/24h for the GTS-SO (dashed line), while the overestimation is smaller for USO (solid line). Small amounts of rainfall are forecast accurately on average.

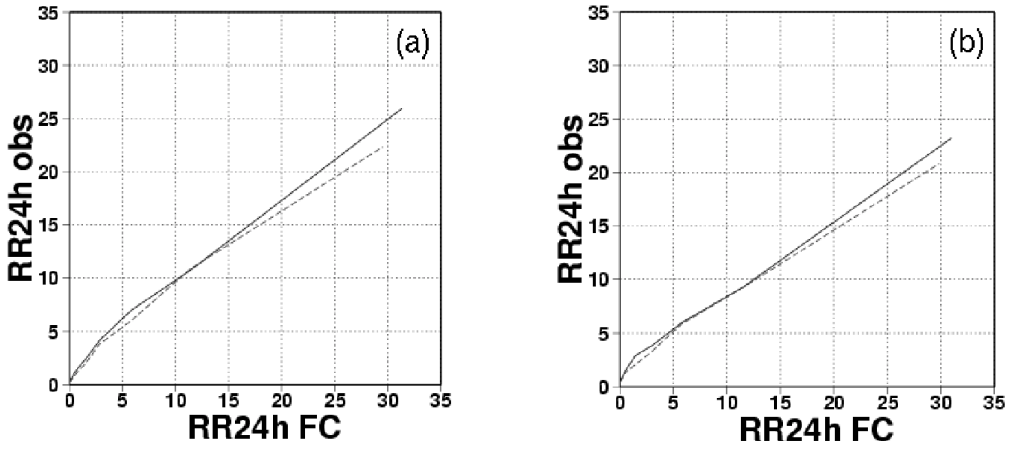

Figure 3: Mean observed 24h cumulated precipitation for each forecast category for the period 8 September to 16 November 1999 plotted against the forecast categories. The forecast range is t+42h (panel a) and t+66h (panel b). Solid line for USO and dashed for GTS-SO.

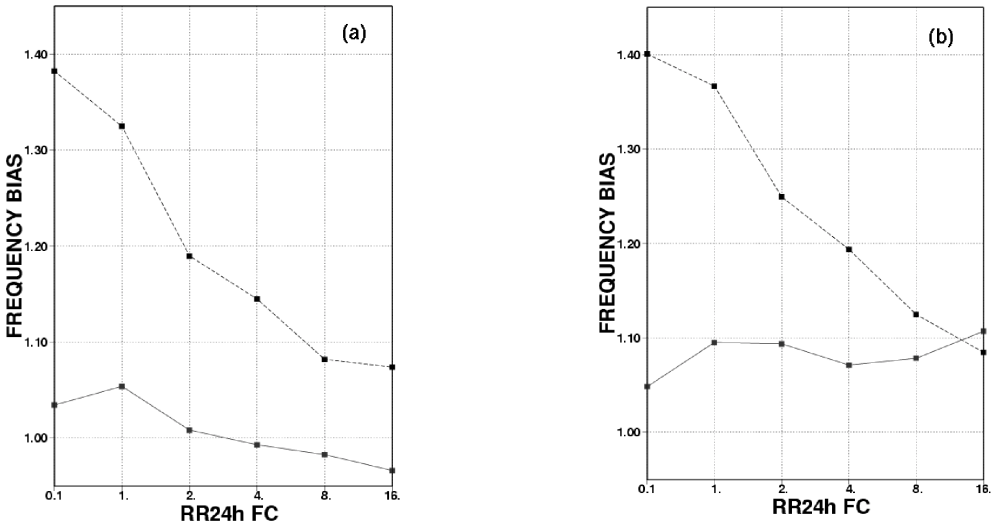

Figure 4: FBI for the period September 8, to November 16, 1999. Forecast range is t+42h (panel a) and t+66h (panel b). Solid line for USO and dashed line for GTS-SO.

Two verification approaches show substantial differences in terms of FBI. The FBI for the USO (solid line) and GTS-SO (dashed line) is depicted in Fig. 4. Ideally, for a perfect forecast the FBI should be equal to one (the event is forecast as often as is observed). For a FBI greater than 1 the event is forecast more often than is observed (over-forecast), and vice-versa for FBI less than 1 (under-forecast).

The FBI for the GTS-SO indicates overestimation for both forecast ranges and all the thresholds. A much better picture is obtained when verifying model forecasts against the USO. The model over-forecasts rainfall events for small thresholds at t+42 and for all the thresholds at t+66, though the overestimation is substantially reduced.

Apart from a few well-defined events, the SOP has been characterized by dry periods alternated with localized precipitation events, the latter being associated with weak systems or shallow troughs. In these situations the representativity issue becomes important, and the model skill will not be high if compared to irregularly distributed GTS-SO, unless its resolution is able to resolve very local effects. A typical example is when small portions of the grid box have rainfall events. The model may predict this fraction accurately, but if it is verified to a SYNOP located where there was no precipitation the model will have bad scores and in general it will show over-forecasting. Many stations inside the same gridbox will represent the behaviour of each portion of the gridbox, and their average is likely to compare better with the model forecast. Therefore, the GTS-SO might depict a distorted scenario leading to overestimation of the precipitation forecast in periods similar to the present SOP, while the USO describe the average behaviour in the grid-box and transfer this information onto the model gridpoint.

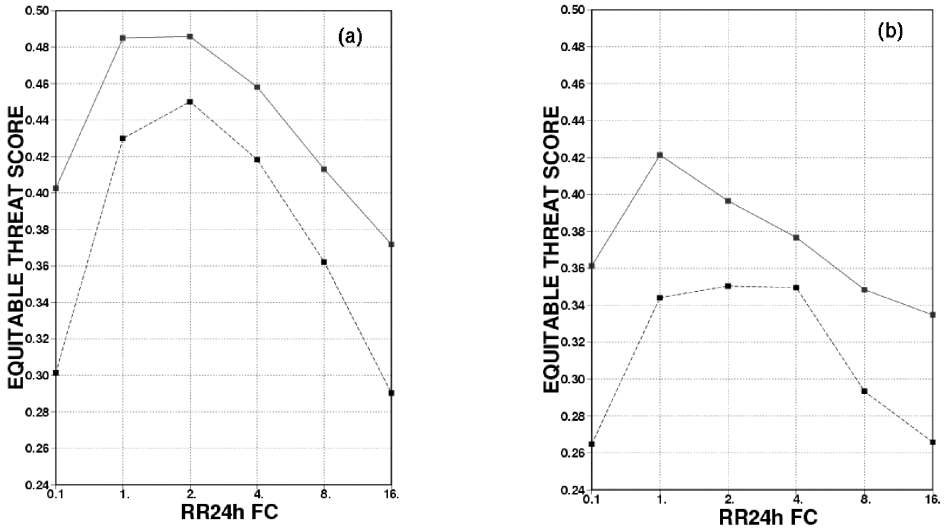

Figure 5: ETS for the period September 8, to November 16, 1999. Forecast range is t+42h (panel a) and t+66h (panel b). Solid line for USO and dashed line for GTS-SO.

Fig. 5 shows the ETS for the ranges t+42 (panel a) and t+66 (panel b). High values of the score indicate a more skilful forecasting system. Thresholds 1 and 2mm/24h give the highest forecast accuracy, while at the extreme thresholds the forecasting system is less skilful for both GTS-SO (dashed line) and USO (solid line) verifications.

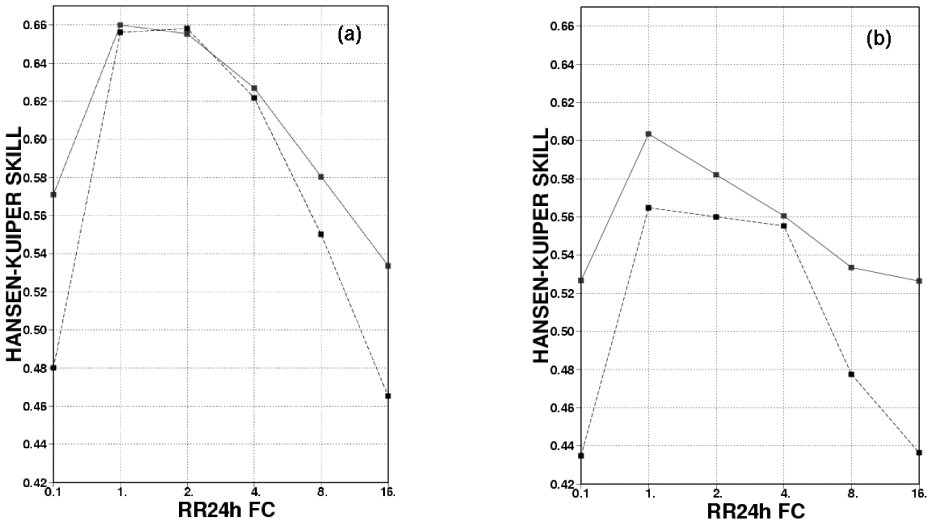

Figure 6: TSS for the period September 8, to November 16, 1999. Forecast range is t+42h (panel a) and t+66h (panel b). Solid line for USO and dashed line for GTS-SO.

As already observed for the FBI, the model shows better scores when verified against the USO. Moreover the model is less penalized at the extreme categories when it is verified against USO. The slight deterioration in the accuracy measure for the t+66 range reflects the 'natural' deterioration of the forecast with time.

The TSS is depicted in Fig. 6 (t+42 panel a and t+66 panel b) and offers a picture similar to the ETS. Extreme categories score better when USO (solid line) are used in the verification.

4.2 The MAP-ALPS experiment

The 'representativity' becomes of primary importance in regions with complex orography. Because of the coarse representation of mountain ranges in a model, gridpoints and station locations have significantly different orographic height, thus rendering verification results very difficult to interpret. The up-scaling technique partly overcomes this problem.

In this section the beneficial effect of verifying model forecast against the USO for the alpine area selected in Fig.1 is investigated. The area is centred over the Alps with 64 GTS-SO and 35 USO. FBI, ETS and TSS have been calculated for the range t+42 and t+66. The results for the two ranges are similar therefore, only scores for the range t+42 will be discussed.

Fig. 7 describes the observed distribution of precipitation for seven closed classes. About 20% of the cases in the SOP show precipitation above the 8mm/24h threshold, while the majority of the cases have very little to no rainfall.

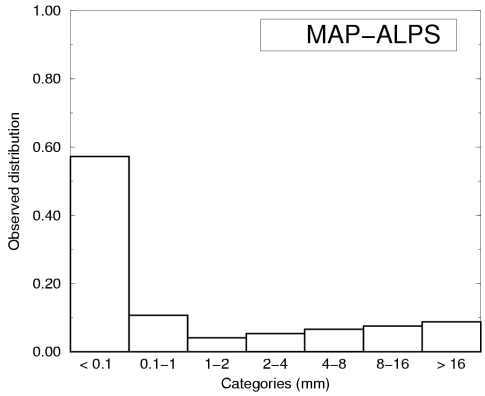

Figure 7:Observed distribution of precipitation for seven closed classes for the MAP-ALPS experiment.

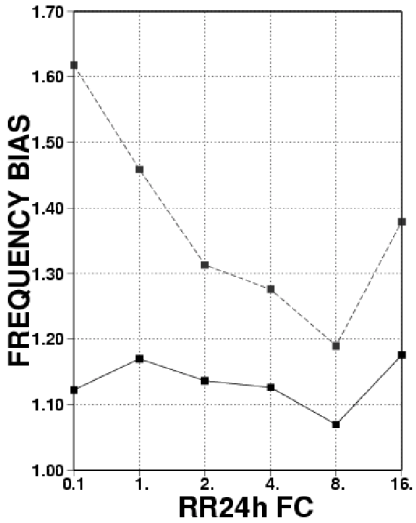

Figure 8: as in Fig. 4 (a) but for the MAP-ALPS experiment (forecast range t+42h).

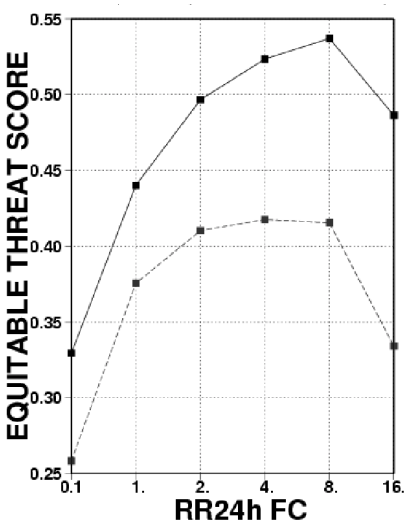

Figure 9: as in Fig.5 (a) but for the MAP-ALPS experiment (forecast range t+42).

Fig. 8 shows the FBI for the verification against the USO (solid line) and against the GTS-SO (dashed line). Both verification techniques indicate overestimation of precipitation in the Alpine area. In the GTS-SO case, small and large thresholds are largely overestimated. The improvement in these two extreme thresholds is substantial when the USO are used.

The ETS and TSS scores for the same area are depicted in figures 9 and 10. They provide a measure of accuracy of the forecast and confirm that for both verification against GTS-SO (solid line) and USO (dashed line) the model is less skilful for the lowest threshold value. The model skill for the SOP improves as the threshold increases, indicating that for the cases with rainfall above 8mm/24h (20% of cases) the forecast has been quite accurate. The accuracy is higher in the verification against USO, as the latter represent the gridbox behaviour rather than a local phenomenon.

The scatter plot of forecast versus observed values (not shown) indicates a strong correlation between forecast and observed precipitation values (0.74) in the case of the USO. Such correlation decreases if the model is verified against the GTS-SO (0.58).

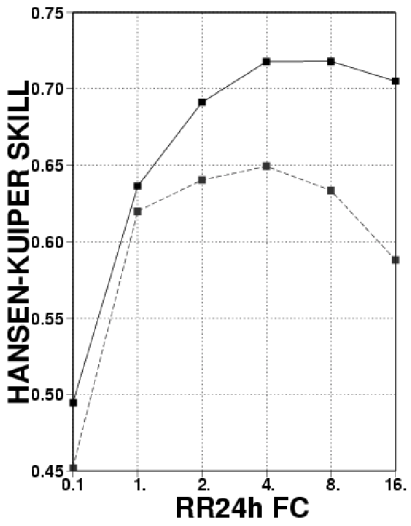

Figure 10: as in Fig.6 (a) but for the MAP-ALPS experiment (forecast range t+42).

4.3 Score time series: cases study

In this section the time-series of the FBI for the whole SOP is discussed as a possible alternative way to look at scores. The period considered is relatively short, therefore to support any conclusion we have analysed the synoptic situation in details.

The timeseries relative to the verification against USO, has been calculated for the MAP-LARGE experiment using a seven-day moving sample mean (FBI is calculated for a seven-day period) over the 70-days period.

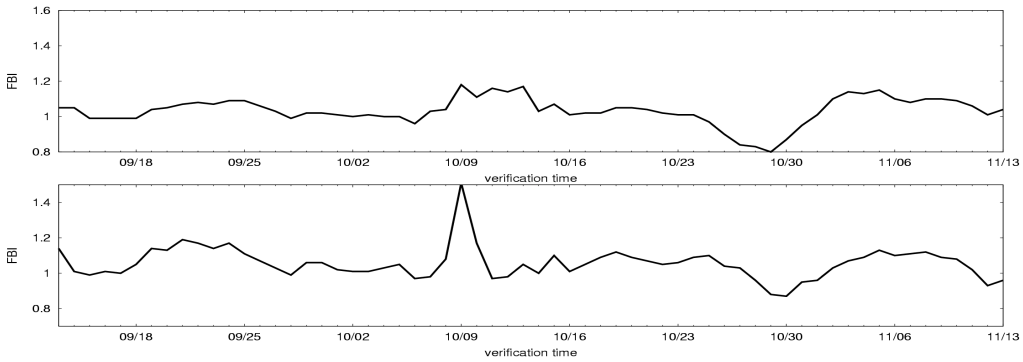

Figure 11: Time-series of FBI (weekly moving average) for the 0.1 mm (top line).

Fig. 11 shows the scores relative to the t+42 range forecast for 0.1 (top line) and 1 mm (bottom line) thresholds. Time-series for higher thresholds are affected by noise due to the small sample size.

The FBI indicates, for the period under investigation, over-forecasting of rainfall events in the Alpine region. This seems to disagree with the hypothesis of underestimation of the orographic effects on the precipitation field put forward by GL. We believe, though, that there is no contradiction in the two results. GL have examined a mixed terrain area using a high-resolution observing network that had a large amount of stations covering France and, therefore, the French side of the Alps. In this paper the attention is concentrated on the Alpine region with a large number of meteorological stations located on the Alpine area. Moreover, both studies concentrate on specific time period and, therefore, drawing general conclusions concerning the behaviour of the forecast would be inappropriate.

The FBI has higher values (both ETS and TSS indicate loss of accuracy for the same period, not shown) around the middle of October. The period was characterised by lack of precipitation events, and the forecast overestimated the precipitation on the only rainy day. The model change on October 12, 1999 cannot be responsible for the deterioration of the scores, as subsequent forecasts show relatively good skill. Synoptic analysis shows that, from the second week in October, a high pressure dominated the Mediterranean basin for several days. On October 13, a trough approached Europe affecting the Alps with high precipitation mainly on the windward side of the mountains. The trough moved eastward slowly because of the presence of the area of high pressure. The 12UTC forecast of October 11 1999 (range t+42) shows a trough and the associated rainband to the South of the observed position. The following 12UTC forecast corrects the position but overestimates the amount of precipitation.

Fig. 12 depicts the t+42 precipitation forecast of October 12 at 12UTC, verifying on October 14 1999 at 6UTC. The forecast field is shaded and numbers denote the USO. The maximum value of observed precipitation is 30mm/24h, while the forecast value goes up to 41mm/24h. Moreover, the precipitation field is overestimated near the mountains.

The second half of September was dominated by a heavy precipitation event associated with a perturbation characterised by a well-defined large-scale forcing. A trough extended South into North Africa with a strong warm and moist advection over the Alpine area and a southerly flow at low levels. Fig. 13 shows the infrared

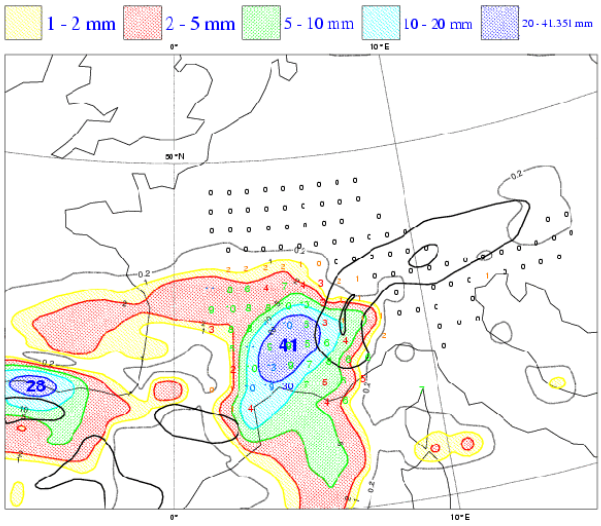

Figure 12: Forecast (shaded) and up-scaled observations (numbers) for the 0.1 mm (top line) and 1 mm (bottom line) threshold. The forecast range is t+42.

Figure 13: IR satellite image at 03:00 UTC of September 20, 1999.

satellite imagery relative to September 20 1999. The cloud band covers the Western Mediterranean, Central Europe and the British Isles. The precipitation patterns ranged from light stratiform rain to strong orographic precipitation (see reports at MDC for a detailed description). Fig. 14 shows the observed 24h cumulated precipitation (06:00 UTC) averaged over the period 15th to 30th September superimposed on the forecast field (shaded). The 24h accumulated precipitation forecast field has been obtained by averaging all t+42h forecasts that verify in the period mentioned. An indication of the model performance can be gained by the positive bias (forecast-minus-observed value) in the alpine region. The model overestimates precipitation amounts, as already observed in the FBI time series.

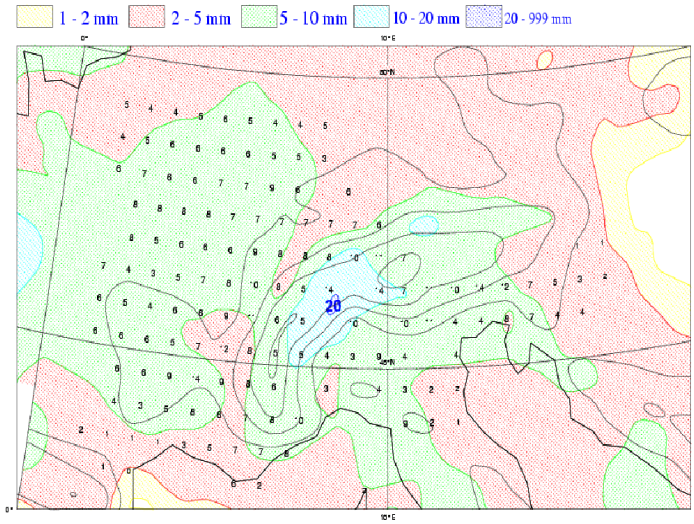

Figure 14: Observed 24h cumulated precipitation at 06UTC averaged over the period 15th to 30th September 1999 superimposed onto the 24h accumulated precipitation (range t+42h; shaded). The forecast field is an average of all the t+42 forecasts verifying between the 15th and 30th September. Shading in the legend.

The beginning of November was characterised by heavy precipitation events in the Alpine region and in the central Mediterranean basin. An upper-level trough with an associated cold front dominated the synoptic scenario. A strong convergence area (South-westerly and easterly flow) was observed over the Italian Po-Valley and maxima of precipitation were recorded along the Italian side of the Alps.

The FBI time series indicates that, in general, the forecast slightly overestimates precipitation amounts.

5. DISCUSSION AND CONCLUSIONS

In this paper two approaches to precipitation verification are discussed. Precipitation forecasts are usually validated using SYNOP data available in real time on the GTS. The spatial distribution is quite irregular, with areas of intense coverage alternated by areas of very little sampling. Verification against such data implies an interpolation of model fields to station location. Representativity becomes an important issue, and its relevance increases, if verification over complex terrain is carried out.

The second approach uses a high-resolution observing network to reconstruct an observed field on the model grid. The technique to obtain the gridded observed field is described in the paper. The method consists of assigning each high-resolution observation to a gridbox. All the observations within the same gridbox are averaged and the mean value is assigned to the gridpoint. Different averaging techniques have been tested, but sensitivity studies have shown little to no difference to the averaging techniques used.

The high-resolution data are a subset of the MAP dataset and cover a period of 70 days between September and November 1999. The ETS, TSS and FBI have been used to assess the skill of the forecasting system.

The model overestimation of precipitation amount is a common feature of the two verification approaches. Overestimation is smaller when the precipitation forecast is compared to the USO and, interestingly the FBI does not seem to change substantially between the lowest and the higher categories. On the contrary validation against the GTS-SO shows overestimation be larger for smaller amounts of precipitation. An accurate analysis of the period under examination shows high variability of the observed field with small and localised precipitation events. The up-scaling procedure takes away part of the observed small-scale variability, therefore comparing better to the model forecast.

Both the ETS and TSS have shown higher values when the model is verified against the USO, as opposed to verification against the GTS-SO, for the forecast ranges considered in the study. Another interesting aspect shown by both ETS and TSS is the better performance of the model when forecasting rainfall in the 1 and 2 mm/24h thresholds. Extreme thresholds are always scoring worse.

Similar results have been obtained for the smaller area centred over the Alps. The question of how representative stations are becomes crucial in this case because of the complexity of the terrain. We believe the up-scaling technique is actually able to produce a valuable alternative to verify the model. The averaging method used reduces the small-scale variability that the model is unable to forecast and, at the same time, the averaged value retains some information on the complexity of the terrain. The FBI has shown overestimation in all categories for both verification procedures, but smaller for the USO. One relevant aspect of the verification against the USO is that the ETS also has large values for large thresholds of precipitation, which leads us to conclude that the model has some skill in forecasting such events on the Alpine region.

The overestimation of the observed precipitation noted in this study, does not contradict the hypothesis in GL that orographic precipitation is underestimated. It must be taken into account that the chosen areas in the two studies are different, the periods analysed in GL are those defined as standard seasons, and the two datasets do not necessarily have similar climatological frequencies.

Time series of the FBI using a seven-day moving sample for small precipitation thresholds (0.1 and 1.mm/24h) have been shown for the verification against the USO. A spike around the middle of October 1999, indicates a general loss of skill of the model. The synoptic analysis has revealed that the lack of precipitation events renders the time-series noisy. Moreover an inconsistent forecast and an overestimation of the precipitation for the only rainy day of the period has led to the spike in the FBI trend.

We feel that the use of high-resolution observations addresses the problem of representative stations correctly. The up-scaling technique is designed to reconstruct an observed precipitation field smoothing small-scale variability that general circulation models are not able to simulate yet. The paper does not discuss in any great length possible errors associated with the up-scaling technique. More work would be needed in this direction, as well as a complete discussion on the statistical properties of the observation dataset. In particular, a more detailed investigation on climatological frequency would be welcome, but historical data are not available for the two areas considered in the present paper.

The improvement in the model behaviour when verified against the USO emphasizes the need for an analysis of precipitation. The differences between the two verification approaches shows a large contribution for the variability within the gridboxes, therefore great care must be taken when assessing a forecasting model to the

GTS-SO, as forecast errors could be due to 'representativity' problems. A better understanding of weaknesses and strengths of a forecasting system would be gained if model spatial scales were compared to similar observed scales.

Acknowledgments

We thank MAP Data Centre for providing datasets of precipitation used in the study. We thank the reviewers for their careful revision that has helped to improve the manuscript. Moreover useful discussions with one of the reviewer are acknowledged.

References

Binder, P., and Coauthors, 1996: MAP Mesoscale Alpine Programme design proposal. MAP Programme Office, 77 pp. [Available from MAP Programme Office c/0 Swiss Meteorological Institute, Krähbülstrasse, 58, CH-8044 Zürich, Switzerland.]

Bougeault, P., and Coauthors, 1998: MAP Mesoscale Alpine Programme science plan.MAP Programme Office, 64 pp. [Available from MAP Programme Office c/0 Swiss Meteorological Institute, Krähbülstrasse, 58, CH-8044 Zürich, Switzerland.]

Frustaci, G., C. Reina, C. Battaglia, R. Ranzi, M. Ferri, C. Cacciamani, L. Sandri, T. Cherubini, and G. Giuliani, 2000: ITAMAP: The data set of non-GTS surface hourly data gathered during the MAP field phase (Sep-Nov 1999): Version 1.2, September 2000. MAP Programme Office, 64 pp. [Available from MAP Programme Office c/0 Swiss Meteorological Institute, Krähbülstrasse, 58, CH-8044 Zürich, Switzerland.]

Ghelli, A., and F. Lalaurette, 2000: Verifying precipitation forecasts using upscaled observations. *ECMWF Newsletter*, **87**, 9-17.

Gregory, D., J.-J. Morcrette, C. Jakob, A. Beljaars, T. and Stockdale, 2000. Revision of convection, radiation and cloud schemes in the ECMWF Integrated Forecasting System., *Quart. J. Roy. Meteor. Soc*., **126**, 1685-1710.

Hirter, H., and H. Richner, 1996: MAP Mesoscale Alpine Programme Data Centre Description. MAP Programme Office, 77 pp. [Available from MAP Programme Office c/0 Swiss Meteorological Institute, Krähbülstrasse, 58, CH-8044 Zürich, Switzerland.]

Jacob, C. and S. A. Klein, 2000c: A parameterisation of the effects of cloud and precipitation overlap for the use in general-circulation models. *Quart. J. Roy. Meteor. Soc.*, **126**, 2525-2544.

Jakob, C., E. Andersson, A. Beljaars, R. Buizza, M. Fisher, E. Gèrard, A. Ghelli, P. Janssen, G. Kelly, A. P. McNally, M. Miller, A. Simmons, J. Teixeira, and P. Viterbo, 2000: The IFS cycle CY21r4 made operational in October 1999. *ECMWF Newsletter*, **87**, 2-9.

Karl, T. R., W.-C. Wang, M. E. Schlesinger, R. W. Knight and D. Portman, 1990: A method of relating General Circulation Model simulated climate to the observed local climate, Part I: seasonal statistics. *J. of Climate*, **3**, 1053-1079.

Katz, W. R., and A. H. Murphy, 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press. 222 pp.

Kim, J.-W., J.-T. Chang, N. L. Baker, D. S. Wilks and W. L. Gates, 1984: The Statistical Proble of Climate Inversion: Determination of the Relationship between Local and Large-Scale Climate. *Mon. Wea. Rev.*, **112**, 2069-2077.

Kuiper, N.-H., 1962: Test concerning random points on a circle. *Proceeding of the Koninklijke Nederlandse Akademie van wetenschappen*, **63**, 38-47.

Mullen, S. L., and R. Buizza, 2000: Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, in press.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature ofgoodness in weather forecasting. *Wea. Forecasting*, **8**, 281-293.

----- , 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590-1601.

----- , and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330-1338.

Ritchie, H., C. Temperton, A.J. Simmons, M. Hortal, T. Davies, D. Dent, and M. Hamrud, 1995: Implementation of the Semi-Lagrangian method in a high resolution version of the ECMWF forecast model. *Mon. Wea. Rev.*, **123**, 489-514.

Shaefer, J. T., 1990: The Critical Success Index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570-575.

Skelly, W. C., and A. Henderson-sellers, 1996: Grid box or grid point: what type of data do GCMs deliver to climate impacts researchers? *Int. J. of Climate*, **16**, 1079-1086.

Simmons, A.J., D.M. Burridge, M. Jarraud, C. Girard, and W. Wengen. 1989: The ECMWF medium-range prediction models: Development of the numerical formulations and the impact of increased resolution. *Meteor. Atmos. Phys.*, **40**, 28-60.

Stephens, M. A., 1970: Use of the Kolmogorov-Smirnov Cramer-Von Mises and related statistics without extensive tables. *J. Royal Statistical Soc.*, **B, 32**, 115-122.

Texeira, J., 2000: Boundary layer clouds in large scale atmospheric models: Cloud scheme and numerical aspects. PhD dissertation, University of Lisbon, 190 pp. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, England.]

Teixeira, J., 1999: The impact of increased boundary layer resolution on the ECMWF forecast system. *ECMWF Technical Memorandum*, **268**, 55 pp.

Tiedtke, M., 1989: A comprehensive mass flux scheme for cumulus parameterisation in large-scale models. *Mon. Wea. Rev.*, **117**, 1779-1800.

Tiedtke, M., 1993: Representation of clouds in large-scale models. *Mon. Wea. Rev.*, **121**, 3040-3061.

Untch, A., Simmons, A., Hortal, M., and Jakob, C., 1998: Increased stratospheric resolution in the ECMWF forecasting system. In *Proceedings of the workshop on chemical data assimilation, KNMI de Bilt, Holland.*

Wigley, T. M. L. and B. D. Santer, 1990. Statistical comparison of spatial fields in model validation, perturbation, and predictability experiments. *J. Geophys. Res.*, **90**, 8995-9005.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 269-272, 467 pp.

Wilson, L. L. and D. P., Lettermaier, 1992: A hierarchical stochastic model of large-scale atmospheric circulation patterns and multiple station daily precipitation. *J. Geophys. Res*., **97**, 2791-2809.