Verification of physical parameters: upscaling or downscaling?

F. Lalaurette

European Centre for Medium-range Weather Forecasts
Reading, England

1. Introduction

Most validations of physical parameterisations rely on dedicated field experiments addressing specific issues that are known to be weak points of the model formulations. Among such recent campaigns, PYREX, TOGA-COARE, MAP or the ARM programs have significantly contributed to our better knowledge of physical processes such as radiation, orographic drag or convection. Satellite instruments on the other hand have been or are about to be launched that describe several aspects of radiation, cloud or precipitation distributions with a coverage and a spatial resolution that could only be dreamed of a few years ago. It should not be remembered however that a very consistent global set of surface observations (SYNOP) is routinely exchanged over the GTS. It provides a unique opportunity to evaluate the direct model output against a reference that has been kept very consistent throughout several decades, making it possible to track model improvements or drifts. In this paper examples will be shown of some of the information that can be gained using this dataset. Some of the problems associated with the limited resolution of numerical models will be discussed, and both upscaling methods bringing the observation data to the model grid scale and downscaling methods coupling observed subgrid scale probability distributions to the Ensemble Prediction System (EPS) output will be presented.

2. Routine verification at ECMWF: the basics

In a way somewhat similar to numerical weather prediction models that have always been demanding computing resources at the forefront of technology, the exchange of meteorological reports from surface observations has relied on the best available telecommunication techniques since the second half of the nineteenth century. At this time, the first reports were exchanged using telegraphic lines but it is only by mid twentieth century that technology was there to allow the set-up of a real time, Global Telecommunication System (GTS). Nowadays more than 50,000 reports are exchanged everyday (Figure 1). The information that

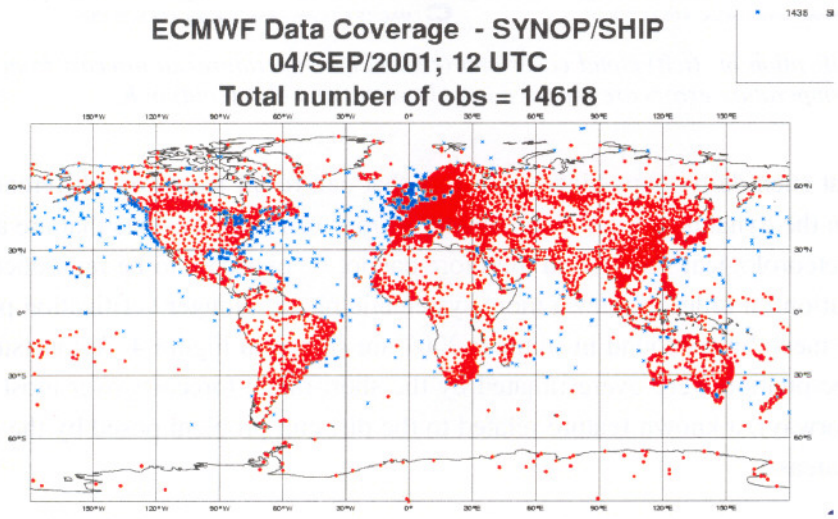

Figure 1: SYNOP message reception map(4 Sept 2001, 09-15UTC)

is extracted currently for routine evaluation of ECMWF model forecast covers precipitation (6, 12 or 24h accumulation), screen-level (2m) temperature and humidity, 10m wind speed and direction and Total Cloud Cover. Wind gusts, sunshine duration and partial cloud cover (high, medium and low level) will be considered for implementation in the near future.

Verification can start from very basic plots where both model fields and local observations are reported using consistent colour codes (Figure 2, left panel). Without the help from any objective analysis techniques, but relying on visual interpretation alone, one can get from there a quick overview of areas where the forecast was successful or failing. On that particular example, cloud cover is clearly underestimated by the model in some parts of the warm sector associated to the front passing over Germany and Poland. Not surprisingly, this day time underestimation of cloud cover is associated with a warm bias of the 2m-temperature forecast (Figure 2, right panel). Similar plots are generated on a daily basis at ECMWF for precipitation (Figure 3), 2m-specific humidity and 10m wind speed (not shown) for areas covering Europe, Africa, North America, South America, North Asia and South-East Asia.

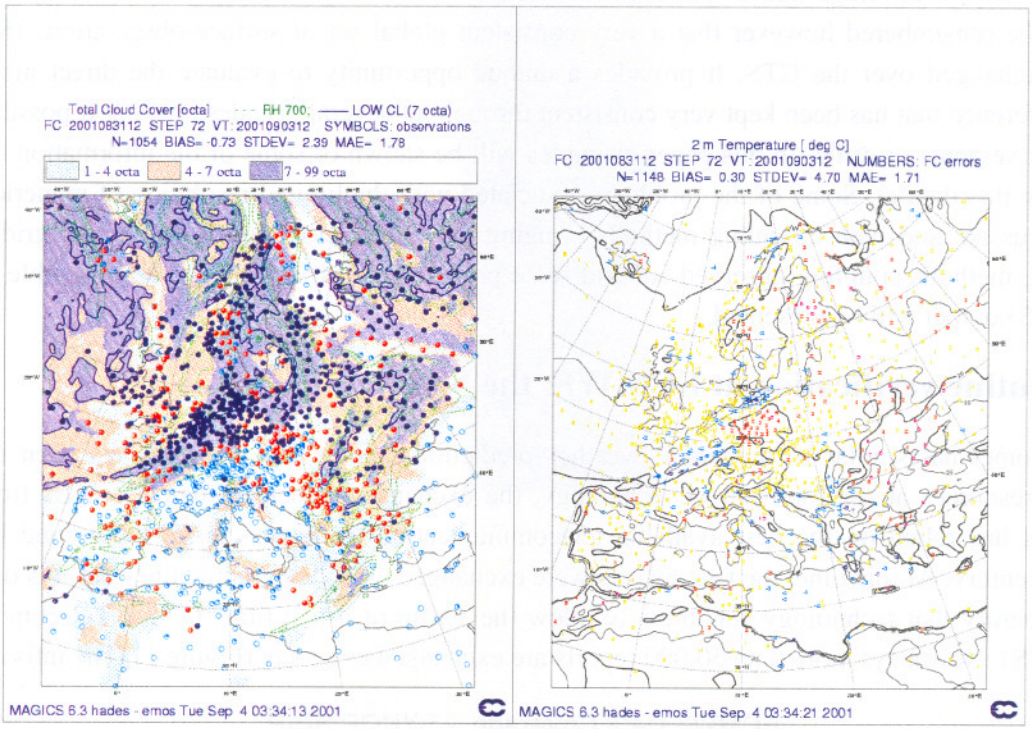

Figure 2: Verification of (left) cloud cover and (right) 2m-temperature72h forecast from 31 Aug. 2001 12 UTC. 2m-temperature errors are reported in red (warm) and blue (cold) in K.

It is clear hover that such daily maps highlight errors that can result from very different causes. These can range from errors in the dynamical forcing, parameterisations or isolated events that are affecting very small areas around the meteorological station and therefore cannot be expected to be reproduced by a model with around 40km resolution. In order to draw a more systematic picture, longer verification periods are required. As an example, the mean biases found in summer 2001 are shown in Figure 4. A consistent picture emerges there where the precipitation were overestimated by the short-range forecasts over most mountainous areas (Alps, Pyrenees, Norway), a known feature related to the difficult problem posed by the parameterisation of convection in such areas.

Another powerful and simple way to gather information about the impact of model change is through the compilation of time series of stable statistics. Time series of monthly biases of rainfall data averaged over

Europe are shown in Figure 5 since 1993. A very clear diurnal cycle can be found in summer, when the convective scheme clearly triggers too much convective activity in the morning (red curve) and not enough in the evening (blue curve). Changes introduced in the physics in October 1999 together with increased boundary layer resolution clearly improved on this feature, while reducing the wintertime positive bias. Clear signatures brought by model changes have also been found in March 1995 following the introduction of explicit microphysics (reduction of the model underestimation of cloud coverage) and in September-December 1996 following several changes in the boundary layer/ land surface parameterisation (reduction of wintertime cold bias).

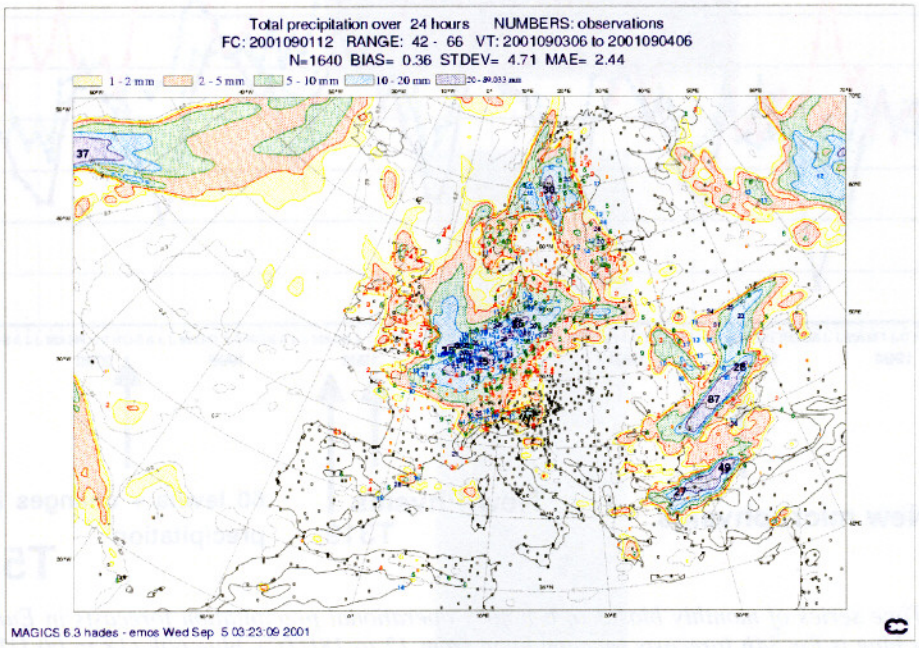

Figure 3: Precipitation daily map (forecast range 42-66h, colour code convention as referenced in caption)

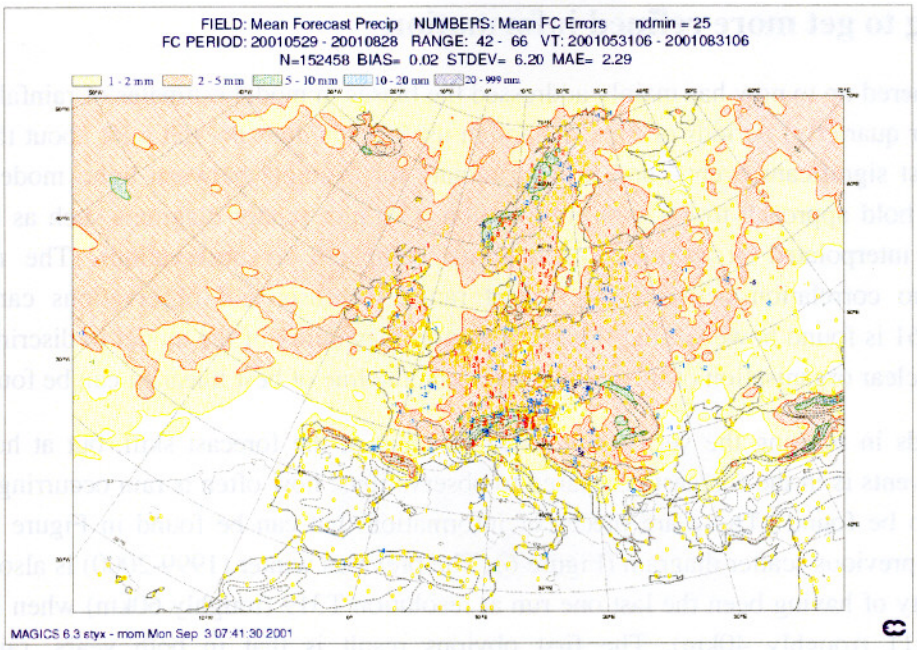

Figure 4: Quarterly bias map showing mean precipitation forecasts in summer 2001(shaded areas, see caption) and mean errors at SYNOP stations (blue is negative, red positive)

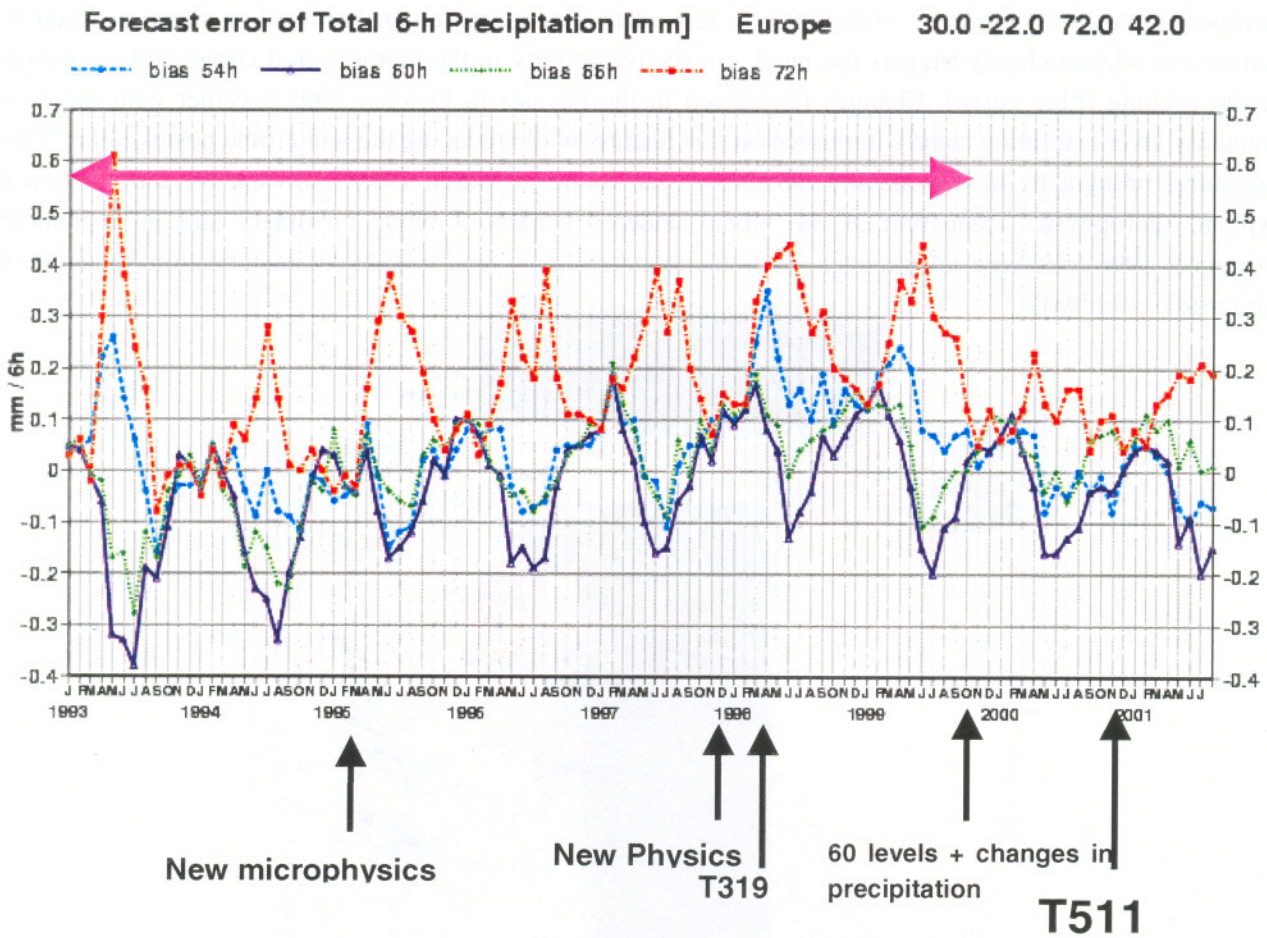

Figure 5: Time series of monthly biases of 6-hourly operational precipitation forecasts in Europe since 1993. Cyan blue is for 54h forecasts (accumulated from 12 to 18UTC), blue 60h (18 to 00 UTC), green 66h (00 to 06 UTC) and red 72h (06 to 12 UTC). Major model changes likely to affect the precipitation forecasts are reported as well.

3. Trying to get more refined information

Information gathered up to now has mainly addressed the biases in model estimates of rainfall averaged over long (monthly or quarterly) periods of time. Although useful, this does not tell a lot about the ability of the model to forecast significant events, and to discriminate for example between light, moderate and strong precipitation. A bold approach to get such information is to plot scatter diagrams such as Figure 6 where daily forecasts interpolated at weather stations are compared to observations. The result is much disappointing: no correlation between even short range forecasts and observations can be found. A correlation of 0.61 is found however, but this reflects more the ability of the model to discriminate dry from rainy events: no clear organisation of the scatter plot along the line of best linear fit can be found.

A step backwards in refining the verification is to look not at the forecast skill, but at how realistic the distribution of events is in the forecast compared to observations: how often is rain occurring, how often are large amounts to be found? These are pieces of information that can be found in Figure 7 for the same dataset as in the previous scatter diagram (Figure 6). The previous winter (1999-2000) is also reported there: it has the property of having been the last one run at resolution T319 (roughly 60km) when DJF 2000-2001 was already T511 (roughly 40km). The first obvious result is that in both years, rain is occurring significantly more often in the forecast than observed: the proportion of dry cases (less than 0.1mm/day) is 39% in the forecast, 58% in SYNOP reports. The frequency bias remains positive up to around 10mm/day, when it reverses sign: not surprisingly, intense precipitation is not occurring in the model as often as

observed. For events above 20mm/day however, the model however has still a relatively good representation of reality - the forecast proportion of such events is still 90% of the observed one in DJF 2000-01, a result that improves significantly on 1999-00 statistics (75%) and can be at least partly be seen as the result of increased model resolution. For higher thresholds though, this proportion drops quickly: the model proportion of rainfall events exceeding 40mm/day is only around 60% of what is observed (it was 50% in 1999-00).

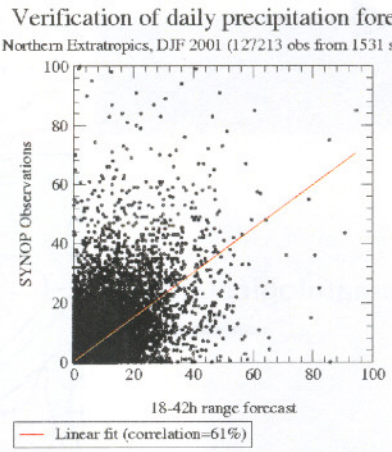

Figure 6: Scatter diagram of daily precipitation forecast (18-42h range) in 1531 weather stations from the Northern Extratropics in winter (DJF 2000-2001)

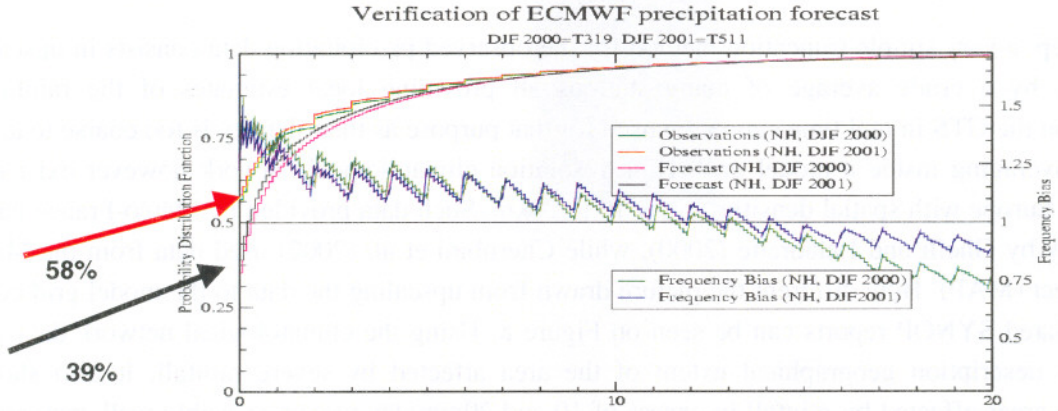

Figure 7: Distribution of daily rainfall events in winter (left scale); frequency biases (ratio of forecast over observed frequency)are also reported (right scale); see caption for more details; the "saw teeth" comes for the quantisation of the observations, usually reported every full mm only. Horizontal scale in mm. The proportion of dry events both in observations (red) and forecasts (black) in DJF 2000-2001 is highlighted on the left

4. How should model forecast be compared to observations?

The question whether or not the type of validation developed in the previous section makes sense needs to be addressed. For such a multi-scale parameter like rain, the model cannot be expected to provide an explicit representation of the full spectrum: at best, the model rainfall fluxes should be taken as an average flux inside a grid box - in practice, the representation of scales up to a few grid boxes is likely to be poor. A fair comparison therefore should not be between the model precipitation forecast interpolated at the weather station with single observations, as this would be like comparing a highly truncated representation of the truth with the truth itself. The only valid comparison therefore is between the forecast and the observed precipitation fields both truncated to the same spatial resolution.

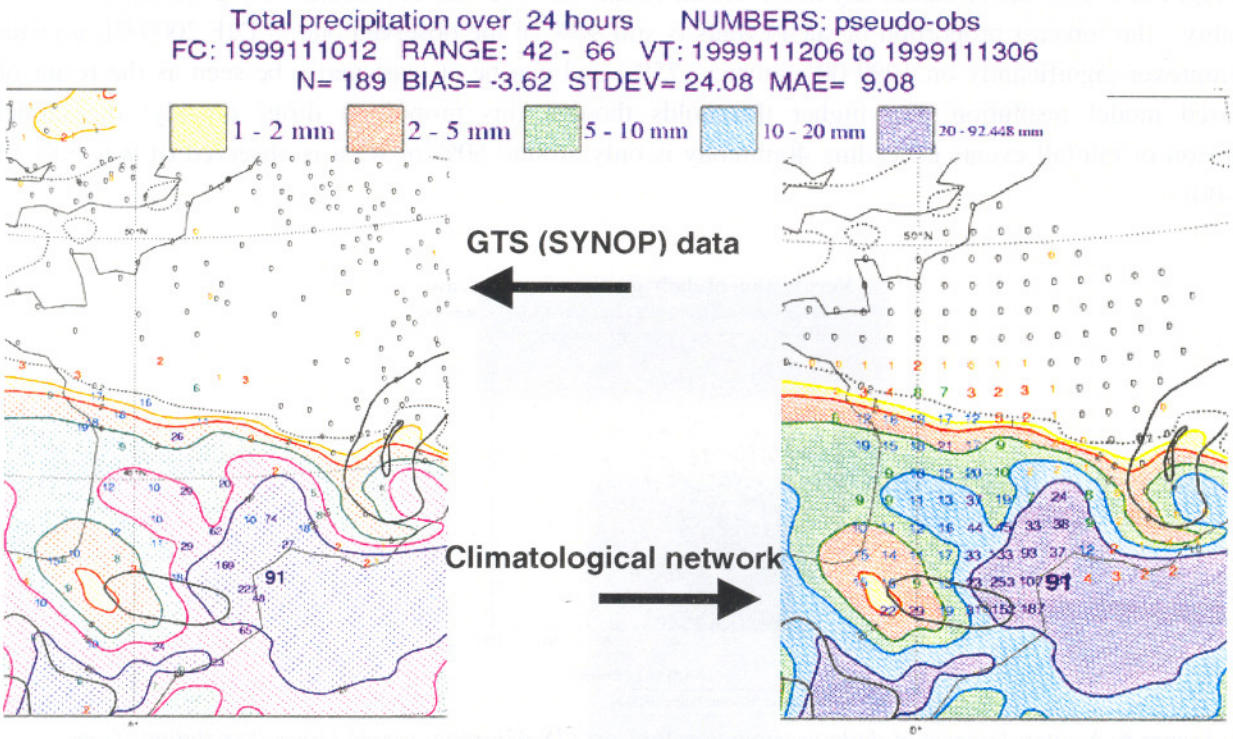

Figure 8: Maps valid for 12 Nov. 1999 (Flash floods in the Aude valley, France). Left panel is a daily verification map (same format as in Figure 3) while the right panel uses Meteo-France climatological network data upscaled on the model grid

As a first step, a very simple truncation method for the observed precipitation data consists in upscaling the observations by a crude average of many stations all providing local estimates of the rainfall. Data exchanged on the GTS in real time cannot be used for that purpose as their density is too coarse to allow any significant averaging inside a model grid. High resolution climatological network however exist in many countries of Europe with spatial densities every 10 to 20km. Such data provided by Météo-France have been used recently by Ghelli and Lalaurette (2000), while Cherubini et al. (2002) used data from the Mesoscale Alpine Project (MAP). How different the picture drawn from upscaling the data to the model grid compared to using isolated SYNOP reports can be seen on Figure 8. Using the climatological network data allows a much better description geographical extent of the area affected by severe rainfall. It also shows that although the areas affected by rainfall in excess of 10 and 20mm/day were reasonably well captured by the model, much more severe accumulations (in excess of 100mm/day) were affecting not just isolated stations, but areas covering several model grid grid meshes. It is therefore fair in that case to state that even at the model scale, the intensity of the event was underestimated.

To assess the impact of using upscaled observations rather than isolated weather station reports, frequency distributions have been collected over three months in winter 1999-2000 over France were high resolution data were made available. Results are shown in Figure 9. Clearly the distribution of upscaled precipitating events is very different from the distribution of SYNOP report amounts, with more rainy events. This can be interpreted by the fact that rain will occur in the "upscaled" observations as soon as at least one station from the high resolution network in a grid box reports rain: the likelywood of this to happen clearly is more than for rain to happen at a single station. As a result, the model frequency bias is reduced from a around 1.35 to 1.15 (15% too many rainy events when compared to upscaled observations, compared to 35% when compared to local observations). Indeed the lessons to be learnt using the upscaled dataset are almost reversed compared to the crude comparison to local observations: the main feature is that the model does not generate enough events with an intensity of 5mm/day and more, when the local reports would have

suggested the main feature to be an overestimation of the light intensity events, moderate events (up to 15mm/day) being seen as non biased.

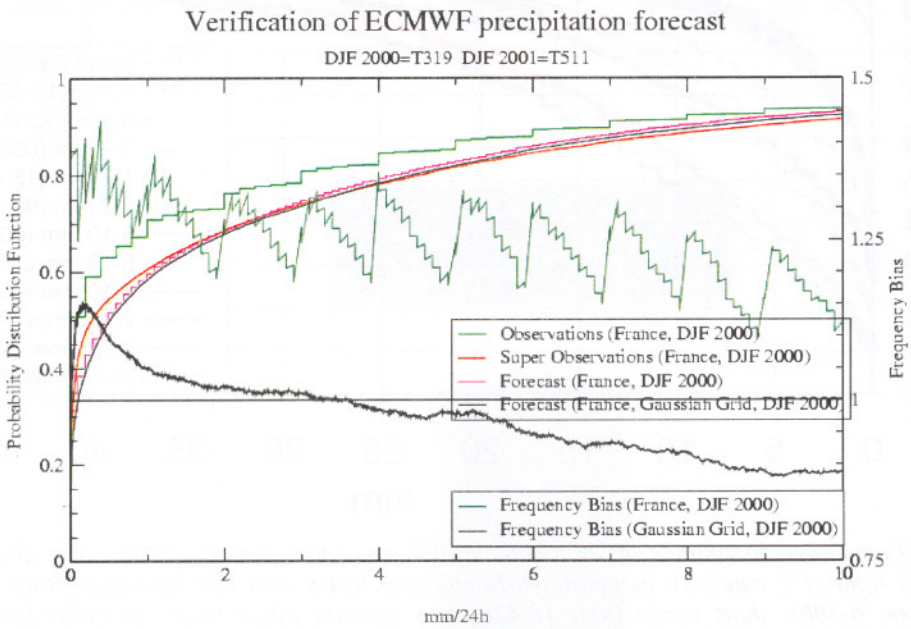

Figure 9: Same as Figure 7 but over France only; in addition to the comparison of interpolated forecasts to SYNOP reports (green and magenta curves), high resolution data upscaled to the model grid ("super observations") are compared to the model grid values (black curve). Corresponding frequency biases are also shown (scale is on the right).

5. Refining the verification: Probabilistic downscaling

Although it was made clear in the previous section that rainfall amounts predicted by the model could only be seen as average fluxes over one or several grid boxes, the ultimate goal remains to provide users of the forecast with quantities that can be directly compared to the local observations. There is no way however to provide this in a deterministic mode with a finite resolution model: the only way forward is to accept that the model forecast is only in control of a limited portion of the scales of atmospheric motion, and to describe smaller scales in a statistical or probabilistic mode.

The distribution of observations corresponding to different intensities of rainfall in the forecasts can be found in Figure 10. They show for example that when the model forecasts no rain in the medium range (90-114h accumulation), there are 17% of the cases when rain is observed at the station. This proportion goes down to 10% in the short and very short range. On the other side of the intensity range, 75% of the amounts of rain reported when the forecast is between 20 and 30mm in the very short range are below 20mm, while 12% are above 30mm. This of course may be the consequence of model errors or limited predictability of the synoptic forecast. It is more likely though to be the result of large subgrid scale variations: heavy rain hardly ever covers uniformly a grid box 60x60km^2: it makes more sense to picture the situation in such case as one when much more intense rainfall rates are observed in a small part of the grid box, while others get a less than average amounts. If that is correct, one can probably make as a working hypothesis the assumption that in the very short range, distributions from Figure 10 are representative of subgrid scale distribution rather than model errors or synoptic scale unpredictability.

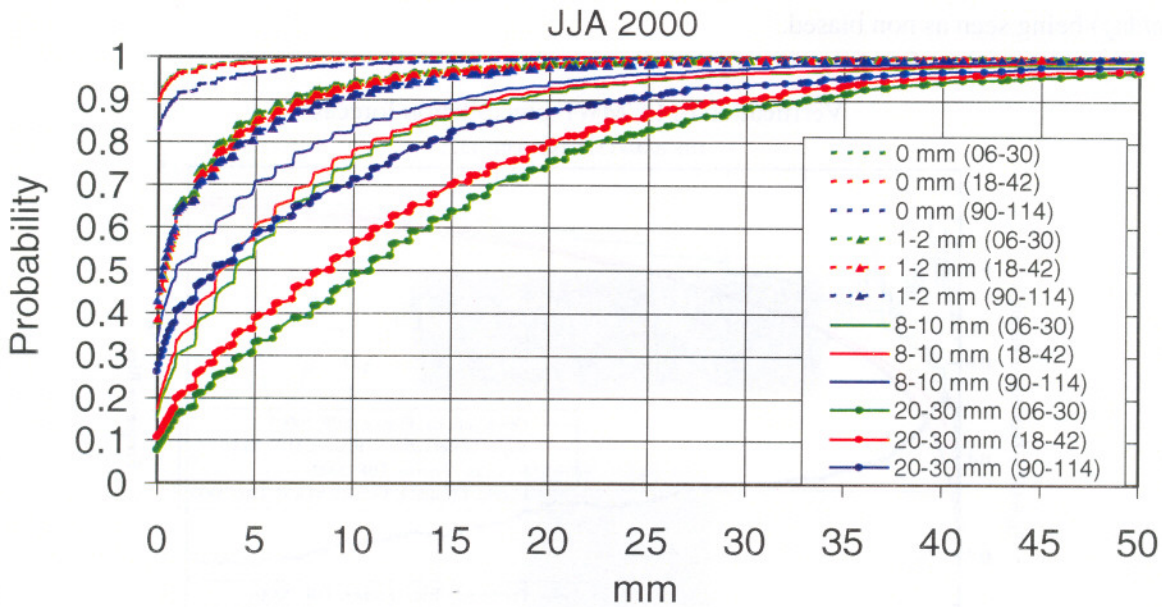

Figure 10: Cumulative distribution of observations subject to different forecasts amounts: dry (less than 0.1mm/day), light (1-2 mm/day), moderate (8-10mm) and heavy rain (20-30mm/day).Both very short range (green, 6-30h), short range (red, 18-42h) and medium range (blue, 90-114h) forecasts are shown.Period is summer2000, weather stations are covering the whole of the Northern Extratropics (about 1500 stations).

An independent check whether or not this is a valid assumption can be made using the Ensemble Prediction System (EPS) that runs 50 equally likely perturbed forecasts every day up to 240h. One of the main products processed from these ensemble are probabilities for predefined thresholds, e.g. for precipitation amounts exceeding 1, 5, 10 or 20mm/day. These probabilities are usually computed using the very simple "democratic voting" method:

$$P(X_{grid} > \alpha) = (1/\mathrm{N}) \sum_{i=1}^{\mathrm{N}} H(\alpha_i - \alpha) \qquad (1)$$

which states that the probability of having rain in a grid box is given by the proportion of ensemble members going for this scenario ($P(X_{grid} > \alpha)$ being the probability to exceed the threshold α, $(\alpha_i)_{i=1,N}$ being the values forecasted by the N ensemble members). Now if instead of forecasting the rain flux in a model grid, the forecast is for the local amount of rain, a formulation such as:

$$P(X_{loc} > \alpha) = (1/\mathrm{N}) \sum_{i=1}^{\mathrm{N}} p(X_{loc} > \alpha \mid X_f = \alpha_i) \qquad (2)$$

where the r.h.s conditional probabilities are coming from verification distribution such as sketched in Figure 10 is more likely to account for the subgrid sale uncertainties not explicitly sampled by the ensemble (Zsoter, 2001). An example of such a processing is given on Figure 11: it can be seen there that the downscaling in effect enhances the likelihood of having large amounts of rain to a bigger area than the democratic method, as even members that do not explicitly forecast 5mm of rain can contribute to the probability of the event to happen.

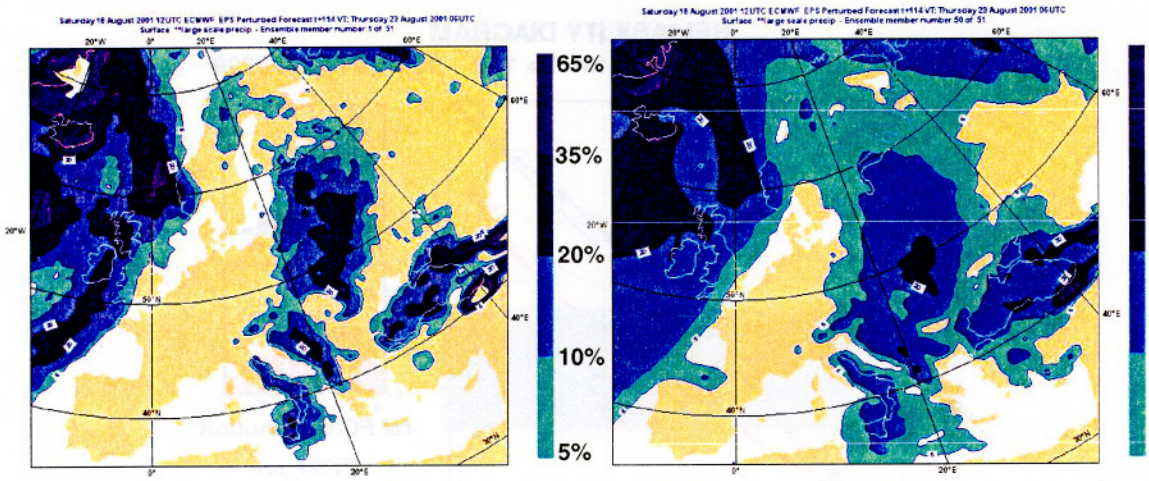

Figure 11: EPS Forecast Probability to have more than 5 mm of rain in the 90-114h forecast range; left shows probabilities processed using the democratic voting method (Eq.1), right using the downscaling method (Eq.2)

Verification of probabilistic forecasts have been introduced in routine operations at ECMWF (Strauss and Lanzinger, 1996), and the quality of the downscaling procedure can therefore be tested against observations. The most basic measure relates the forecast probabilities to the frequency of occurrence of the event. Such a reliability diagram is shown in Figure 12. While the classical "democratic vote" probabilities overestimate the probabilities of small amounts of rain, downscaling the forecast clearly gives a better estimate. This result is consistent with those found in the previous section (Figure 7): rainfall model estimates tend to overestimate the frequency of light rain events at the station. A widely used measure of skill for a probabilistic forecast is through the Brier Skill Score (e.g. Wilks, 1995) – the positive impact of the downscaling procedure is illustrated through the use of this score for several forecast ranges and rainfall intensities in Figure 13. The downscaling procedure proposed here is a simple way to provide an unbiased probabilistic forecast for local rainfall intensities. A deterministic framework does not allow such a correction, and therefore the observations have to be upscaled in order to provide a fair deterministic verification.

6. Summary

There is a large amount of information to be gained for the validation of physical processes from data routinely exchanged on the GTS. In particular biases can be computed that give useful indications as soon as they are averaged over monthly periods of time, and aggregated over subcontinental areas. Time series of such biases are particularly useful in showing the impact of model changes. The verification of physical parameters is not that straightforward as soon as one needs even moderately refined information such as conditional errors: then the representativeness of both model variables and observations need to be addressed. A very crude method to do the job for precipitation data has been presented here. It requires data that are not currently exchanged over the GTS, and that are usually only available to meteorological services with a two to three months delay. It has been shown that the model fits better these upscaled data from high resolution hydro-meteorological networks than local observations, as could be expected. More accurate methods to explicitly account for the representativeness of both observations and model data however need to be designed, maybe using techniques already used for data assimilation in numerical weather prediction. It should be stressed that if such an explicit account of representativeness differences is not identified, differences between model and observations can be wrongly interpreted - it has been shown for example that the crude comparison of model precipitation distribution with locally observed distributions was overestimating significantly the tendency of the model to forecast rain too often.

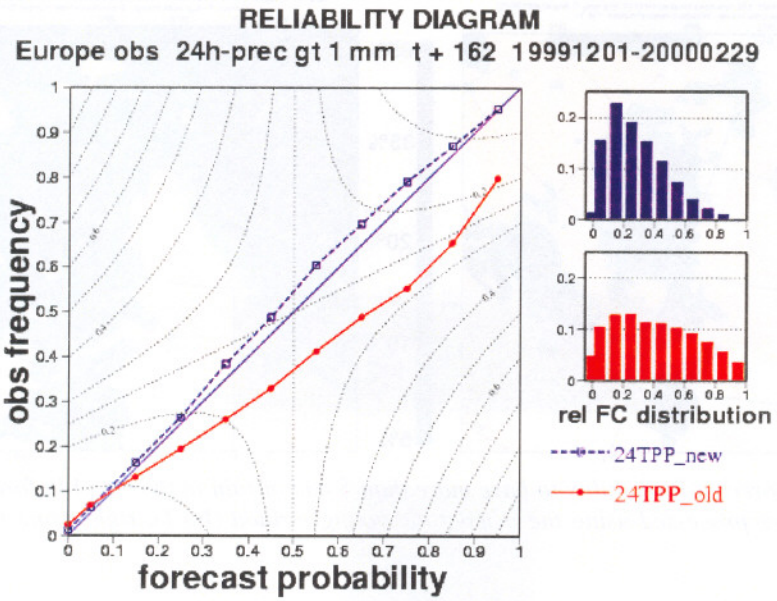

Figure 12: Reliability curves for probability forecasts of rain occurrence(>1mm/day) using the democrating vote method (red) and the downscaling method (dashed blue).

Although the verification of deterministic forecasts from a numerical model can only be achieved in a fair way once the observations have been put in a format that can be compared to what the model resolved scales are, users of the forecasts have a different view. Their interest is usually for local occurrences of weather events. A probabilistic framework has been presented here where the limited predictability of such events is explicitly recognised, but split in two different categories: on one hand, a dynamical method (the EPS) is used to account for the uncertainties of the dynamical environment, while on the other hand, subgrid scale variability of the local precipitation rates is statistically modelled thanks to the collection of local observations sampled over different categories of very short range forecasted precipitation rates (downscaling). Preliminary results show not only that this improves the reliability of the probabilistic forecasts, but also that their predictive skill is improved. Most of this contribution has focused on

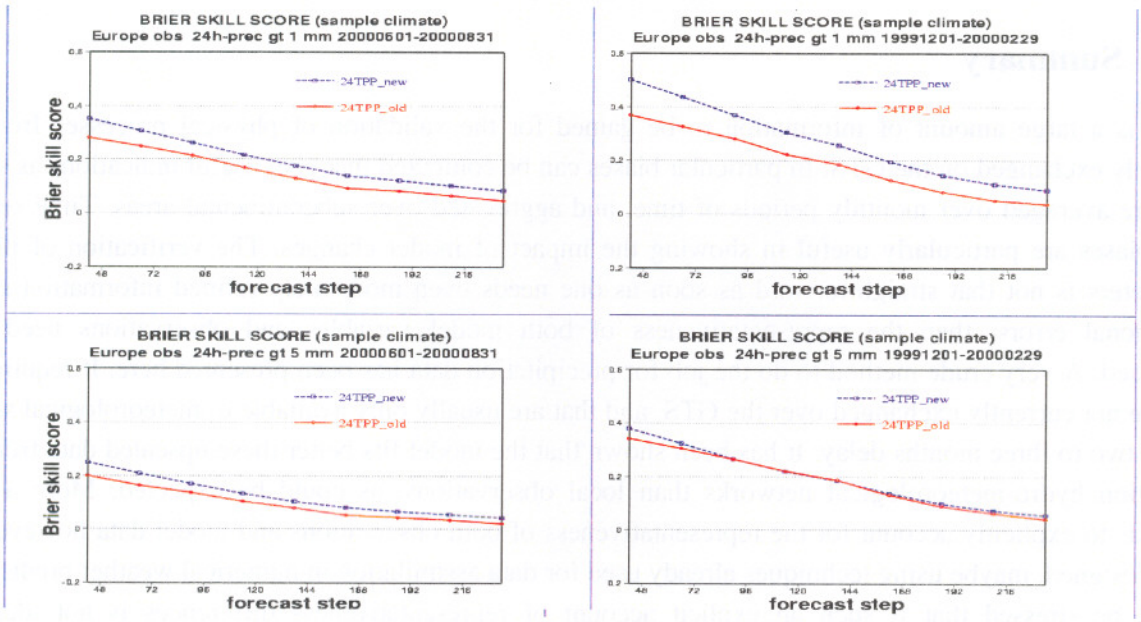

Figure 13: Brier Skill scores for probability forecasts of rain occurrence(>1mm/day on upper row, >5mm/day on lower row) using the democrating vote method (red) and the downscaling method (dashed blue).Left column: summer 2000; right column: winter 1999-2000.

precipitation, not only because of the multi scaling properties of such a variable, but also because high resolution network gathering daily rainfall data have been in place over a long historical period. This makes it relatively easy to test upscaling methods for rainfall. The multi-scaling problem is most certainly also one that should be tackled explicitly in the case of surface wind verification. Upscaling techniques could however be difficult in that case as the density of the networks gathering wind observations in routine is rather sparse. Downscaling should however be possible, maybe using some of the knowledge we have of statistical properties of the turbulence in the surface boundary layer.

Acknowledgements

Results shown in this paper have been collected thanks to the verification software and observation datasets built up and maintained along the years by Anna Ghelli, Andreas Lanzinger and Milan Dragosavac. Ervin Zsoter from the Hungarian Meteorological Institute has produced the results on the downscaling of EPS probabilities during a 3-months summer visit at ECMWF in summer 2001; data used in the upscaling studies have been kindly provided by Météo-France

References

Cherubini, T., A. Ghelli and F. Lalaurette, 2002: Verification of Precipitation Forecasts over the Alpine Region Using a High-Density Observing Network; *Weather and Forecasting*, **17**, 238-249

Ghelli, A. and F. Lalaurette, 2000: Verifying precipitation forecasts using upscaled observations; *ECMWF Newsletter*, **87**, 9-16

Strauss, B. and A. Lanzinger, 1996: Verification of ensemble prediction; *ECMWF Newsletter*, **72**, 9-15

Wilks, D. S., 1995: Statistical Methods in the Atmospheric Sciences; *International Geophysics Series*, **59**, Academic Press .

Zsoter, E., 2001: Downscaling EPS probabilities using SYNOP precipitation data; *ECMWF Memorandum, Operations Department, O/MOP/30.*