Adv. Geosci., 25, 29–36, 2010
www.adv-geosci.net/25/29/2010/

Advances in Geosciences

The forecaster's added value in QPF

M. Turco[1,*] **and M. Milelli**[1]

[1]ARPA Piemonte (Regional Environmental Protection Agency), Torino, Italy
[*]now at: GAMA (Meteorological Hazards Analysis Team), Department of Astronomy & Meteorology, Faculty of Physics, University of Barcelona, Barcelona, Spain

Received: 15 October 2009 – Revised: 22 February 2010 – Accepted: 13 February 2010 – Published: 9 March 2010

Abstract. To the authors' knowledge there are relatively few studies that try to answer this question: "Are humans able to add value to computer-generated forecasts and warnings?". Moreover, the answers are not always positive. In particular some postprocessing method is competitive or superior to human forecast. Within the alert system of ARPA Piemonte it is possible to study in an objective manner if the human forecaster is able to add value with respect to computer-generated forecasts. Every day the meteorology group of the Centro Funzionale of Regione Piemonte produces the HQPF (Human Quantitative Precipitation Forecast) in terms of an areal average and maximum value for each of the 13 warning areas, which have been created according to meteo-hydrological criteria. This allows the decision makers to produce an evaluation of the expected effects by comparing these HQPFs with predefined rainfall thresholds. Another important ingredient in this study is the very dense non-GTS (Global Telecommunication System) network of rain gauges available that makes possible a high resolution verification. In this work we compare the performances of the latest three years of QPF derived from the meteorological models COSMO-I7 (the Italian version of the COSMO Model, a mesoscale model developed in the framework of the COSMO Consortium) and IFS (the ECMWF global model) with the HQPF. In this analysis it is possible to introduce the hypothesis test developed by Hamill (1999), in which a confidence interval is calculated with the bootstrap method in order to establish the real difference between the skill scores of two competitive forecasts. It is important to underline that the conclusions refer to the analysis of the Piemonte operational alert system, so they cannot be directly taken as universally true. But we think that some of the main lessons that can be derived from this study could be useful for the meteorological community. In details, the main conclusions are the following:

– despite the overall improvement in global scale and the fact that the resolution of the limited area models has increased considerably over recent years, the QPF produced by the meteorological models involved in this study has not improved enough to allow its direct use: the subjective HQPF continues to offer the best performance for the period +24 h/+48 h (i.e. the warning period in the Piemonte system);

– in the forecast process, the step where humans have the largest added value with respect to mathematical models, is the communication. In fact the human characterization and communication of the forecast uncertainty to end users cannot be replaced by any computer code;

– eventually, although there is no novelty in this study, we would like to show that the correct application of appropriated statistical techniques permits a better definition and quantification of the errors and, mostly important, allows a correct (unbiased) communication between forecasters and decision makers.

Correspondence to: M. Turco
(mturco@am.ub.es)

1 Introduction

The goal of this study is to evaluate the forecaster added value in comparison with the direct model output (DMO). Within the alert system of ARPA Piemonte it is possible to study in an objective manner if the human forecaster is able to add value with respect to computer-generated forecasts. Before introducing the alert system of ARPA Piemonte, few preliminary remarks are needed:

– the conclusions refer to the analysis of the Piemonte operational alert system, so they cannot be directly taken as universally true;

– it is not obvious that the forecaster has an added value: some post processing method is competitive or superior to human forecast (see for instance Baars et al., 2005; Charba et al., 2003; Sanders et al., 1986; Roebber et al., 1996c).

The weather forecast component in the alert system of Arpa Piemonte is described here. The starting point for the forecasters is the output of the meteorological models COSMO-I7 and IFS. After a synoptic evaluation, the forecasters can see the QPF of the guidance model interpolated on each of the 13 warning areas (see Fig. 1), which have been created according to meteo-hydrological criteria. There are 11 regional warning areas plus Valle d'Aosta and Ticino which are included for hydrological reasons. The QPF of the models are expressed in terms of area average and maximum value, up to +72 h, every 6 h. Then the forecasters produce and deliver the weather forecast bulletin and their subjective QPF, hereafter named HQPF, also in terms of area average and maximum value over the warning areas, from +12 h to +72 h, every 6 h. It is important to underline that the warning period is limited to the first 36 h of forecast, that is from +12 h to +48 h. For this reason and, for sake of brevity, in this work we focus our attention on the +24 h/+48 h time interval. It has to be stressed here that in the Centro Funzionale of Regione Piemonte, the forecasters have more than five years of experience and are always in couple, excluding the given holidays when there is one person only. The Centro Funzionale of Regione Piemonte is based on the comparison between these HQPFs with predefined rainfall thresholds in order to allow the decision makers to produce an evaluation of the expected effects, according to the results of a flood forecasting model (Rabuffetti and Barbero, 2005), shallow landslides model and snow impact analysis (Campus et al., 2007). Since the HQPFs are in agreement with the precipitation forecast reported in the regional bulletin (www.arpa.piemonte.it), they can be considered as a proxy of weather forecast expressed in this bulletin, which can be useful for a large variety of end-users, interested not only in heavy precipitation (as Civil Protection Department might be) but also in light and moderate thresholds. An additional aim of this paper is to recognize the importance of the appropriated statistical techniques to quantify the errors and to communicate the forecast to the users. This paper is organized as follows: Sect. 2 describes the observational and forecast data used; Sect. 3 describes the method used to verify the forecaster's added value; Sect. 4 presents the main results and, eventually, conclusion and final remarks are given in Sect. 5.

2 Verification data

2.1 Observed data

The very dense non-GTS network of ~ 350 rain gauges (see Fig. 1), managed by ARPA Piemonte makes possible a high resolution verification. A two steps quality control is applied to the observed precipitation data: the first step automatically checks for internal consistency, while in the second step suspicious measurements are manually verified. Figure 1 shows the total daily coverage for each rain gauge, that is the percentage of days in which the gauge was active: it can be stated that almost all the stations were active for the whole considered period. We define the average (QPEA) and maximum (QPEM) quantitative precipitation estimate in this way:

$$\mathrm{QPEA}_{T,j} = \frac{\sum_i^{N_j} \mathrm{prec}_{T,ij}}{N_j} \quad (1)$$

$$\mathrm{QPEM}_{T,j} = \max\{\mathrm{prec}_{T,ij}\} \quad (2)$$

where:

– $\mathrm{QPEA}_{T,j}$ is the average quantitative precipitation estimate over the warning area j, at time T;

– $\mathrm{QPEM}_{T,j}$ is the maximum quantitative precipitation estimate over the warning area j, at time T;

– N_j is the number of rain gauges in the warning area j;

– $\mathrm{prec}_{T,ij}$ is the precipitation observed at time T (the interval in this study is 24 h) from the rain-gauge i of the warning area j.

2.2 Forecast data

The forecast data are the QPF derived from the two meteorological models COSMO-I7 (the Italian version of the COSMO Model, a mesoscale model developed in the framework of the COSMO Consortium, see www.cosmo-model.org for a more comprehensive description of the activities and of the organization) and IFS (the ECMWF global model). Since IFS increased the resolution from T_L511 to T_L799 in February 2006, the verification period starts from March 2006 and ends in February 2009, with four complete seasons for three years. For each model the area average and the maximum value are calculated in this way:

$$\mathrm{QPFA}_{T,j} = \frac{\sum_i^{N_j} \mathrm{prec}_{T,ij}}{N_j} \quad (3)$$

$$\mathrm{QPFM}_{T,j} = \max\{\mathrm{prec}_{T,ij}\} \quad (4)$$

where:

– $\mathrm{QPFA}_{T,j}$ is the average quantitative precipitation forecast over the warning area j, at time T;

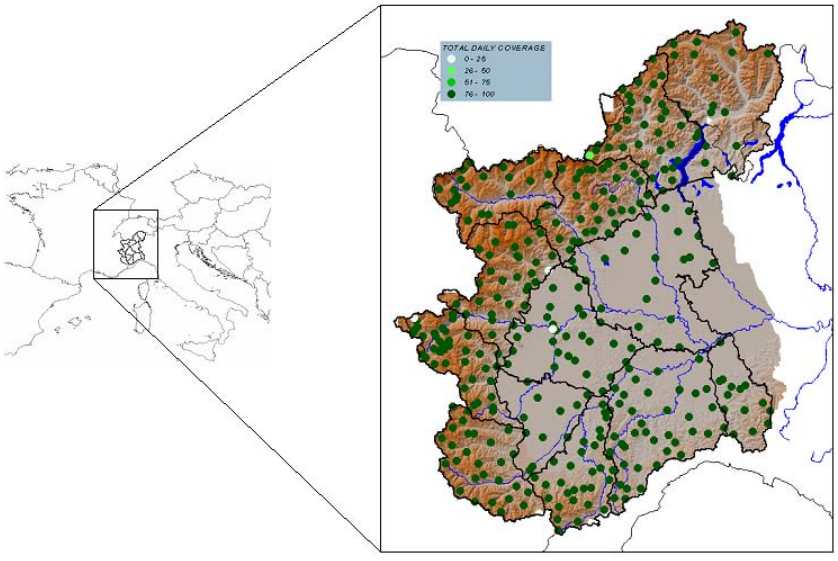

Fig. 1. Domain of the verification, with the 13 warning areas and the network distribution (dots). The colors of the dots indicate the total daily coverage of the rain gauges during the considered period, that is the percentage of days in which the gauges were active.

– $\mathrm{QPFM}_{T,j}$ is the maximum quantitative precipitation forecast over the warning area j, at time T;

– N_j is the number of model gridpoints in the warning area j;

– $\mathrm{prec}_{T,ij}$ is the precipitation forecast at time T for the gridpoint i of the warning area j.

The HQPF, as said in Sect. 1, is expressed in terms of area average and maximum value over the warning areas, from +12 h/+72 h, every 6 h. In this study we consider only a period of 24 h because the hypothesis test implemented in this study (Sect. 3.3) requires that the samples are independent (see Accadia et al., 2003a and Accadia et al., 2003b).

3 Verification method

3.1 Quality

In this context the most useful verification approach is the measure of the QPF and HQPF skills by first converting precipitation expressed as continuous amounts into "exceedance" categories (yes-no statements indicating whether precipitation equals or exceeds selected thresholds) and then computing the performances for each threshold. The thresholds are 1, 10, 20, 30, 40 mm/24 h. The latter is the greatest threshold which permits to have reliable verification statistics, at least in the considered period of time. The quality of the forecasted precipitation field is then evaluated according to statistical indices based on contingency tables (see Table 1). For more details see Wilks (2006) and Joliffe and Stephenson (2003). Table 1 shows a classical contingency table that is needed to define the indices cited or used in the following, and summarized in Table 2. We underline that the total number of elements in the contingency table is obtained by multiplying the number of available days and the total number (13) of warning areas (j in Eqs. 1 to 4). This has been done for having a more solid statistics, despite the fact that in this way we lose the information of the error at warning area level.

Table 1. 2×2 contingency table.

		Obs.	
		Yes	No
For.	Yes	a	b
	No	c	d

Table 2. Verification indices based on 2×2 contingency table. See Table 1 for the definition of a, b, c and d.

Index	Formulation
BIAS	$\frac{a+b}{a+c}$
POD=H	$\frac{a}{a+c}$
POFD=F	$\frac{b}{b+d}$
FAR	$\frac{b}{a+b}$
ODDS	$\frac{ad}{bc}$
HK=H-F	$\frac{ad-bc}{(a+c)(b+d)}$

In order to fully describe the table, one needs three indices suitably chosen (Stephenson, 2000). It is important to underline that an incorrect set of indices can lead to conflicting indications of the skill, as shown in Harvey et al. (1992). We have chosen to use the BIAS to compare the marginal probabilities of the forecasts and observations, so we have still to select the other two indices in a suitable way. Harvey et al. (1992) illustrate the utility of ROC diagram that derives from the theoretical framework of the signal detection theory. The ROC index has became a standard during recent years (Casati et al., 2008). The ROC diagram is the relation between POD (the probability of detection, i.e. the relative number of times an event has been forecasted when it actually occurred: $p(F|O)$) and POFD (the probability of false detection, i.e. the relative number of times the event has been forecasted as "event" when it did not occur). These two indices (see Table 2 for their mathematical formulation) measure the skill of the forecast from an "observations" point of view but it is also possible to choose stratification based on the forecasts, i.e. considering the relative number of times an event has been observed when it has been forecasted, $p(O|F)$ and also the relative number of times a "non-event" has been observed when it has been forecasted as event. The $1-p(O|F)$ is also named the "false alarm ratio" which should not be confused with the previously defined POFD (named false alarm rate) (e.g., Wilks (2006)). So it is necessary not to mix the stratification based on the observations and the one based on the forecasts, for example the representation based on BIAS, POD and FAR (i.e. the false alarm ratio, see Table 2) can get contradictory indications, especially when the base rate (i.e., climatological probability) is very low or very high (Göber et al., 2004). It is not obvious to distinguish $p(F|O)$ from $p(O|F)$ but their confusion can lead to decision making errors in every day life (Gigerenzer, 2002). For example, the percentage $p(F|O)$ of events ($\mathrm{QPEA}_{T,j} > 40$ mm/24 h) correctly forecast by HQPF (forecast time +24 h/+48 h) is around 70% (this is also called the POD score), but the percentage $p(O|F)$, i.e. the probability to observe an event when it has been forecasted is only around 50%. It is important to underline that $p(O|F)$ depends on the accuracy $p(F|O)$ and on the base rate $p(O)$ as shown from the Bayes theorem:

$$p(O|F) = \frac{p(O)}{p(F)} p(F|O) = \frac{p(O)}{p(F)} \mathrm{POD} \tag{5}$$

This equation shows that the $p(O|F)$ is a function of the accuracy (POD) and also of the frequency of the forecasted and observed events. The frequency of the events could be varying among years, seasons, zones, etc., so the forecaster has to know this "climatology" which is of crucial importance. Summarizing, as a first step in our analysis, we used the BIAS, POD and POFD representations to describe the contingency table, but then we had to consider that:

- the POFD is very low due to the fact that we consider events with small frequency (e.g. climatological probabilities around 0.9% for $\mathrm{QPEA}_{T,j} > 40$ mm/24 h);
- the triplet BIAS, HK, ODDS (Table 2) maintains the independence on the base rate and offers several advantages in respect to BIAS and ROC representation, see Stephenson (2000) for more details.

Therefore, in the following we consider BIAS, HK, ODDS to describe the contingency table.

3.2 Value index

Thornes and Stephenson (2001) show that in order to choose among different forecasts it is useful not only to consider the quality of the forecast (measured by the afore mentioned indices) but also to consider the value of the forecast, in a way that takes into account the different customers and the forecast error. Thus we adopt the Value index proposed by Richardson (2000):

$$V = \frac{\min(\alpha, \bar{o}) - \mathrm{F}(1-\bar{o})\alpha + H\bar{o}(1-\alpha) - \bar{o}}{\min(\alpha, \bar{o}) - \bar{o}\alpha} \tag{6}$$

where:

- α = C/L, the user's cost/loss ratio;
- H =POD;
- F = POFD;
- $\bar{o} = a + c$ = base rate (see Table 1).

It is easy to show that the maximum Value is in correspondence of C/L=$\bar{o}$, and the maximum Value is H-F, i.e. the HK index. This simple cost/loss model is widely used in verification of probability forecasts and it is quite simple to understand. The Value index is indubitable intuitive, interpretable and synthetic and it is useful not only for scientific purposes but also for communication to non-specialists (Mason, 2008).

3.3 Hypothesis test

In this study we applied the hypothesis test developed by Hamill (1999), in which a confidence interval is calculated with the bootstrap method in order to establish the real difference between the skill scores of two competitive forecasts. The hypothesis of this test is that the time series have negligible autocorrelations. In Accadia et al. (2003a) and Accadia et al. (2003b) there is the implementation of this test in northwest Italy and it is shown that model forecast errors are negligibly autocorrelated in time if observations are accumulated in 24 h; also Hamill (1999) considers this time period in his example of test application. A level of significance of 5% is used, so we add the 95% confidence intervals for

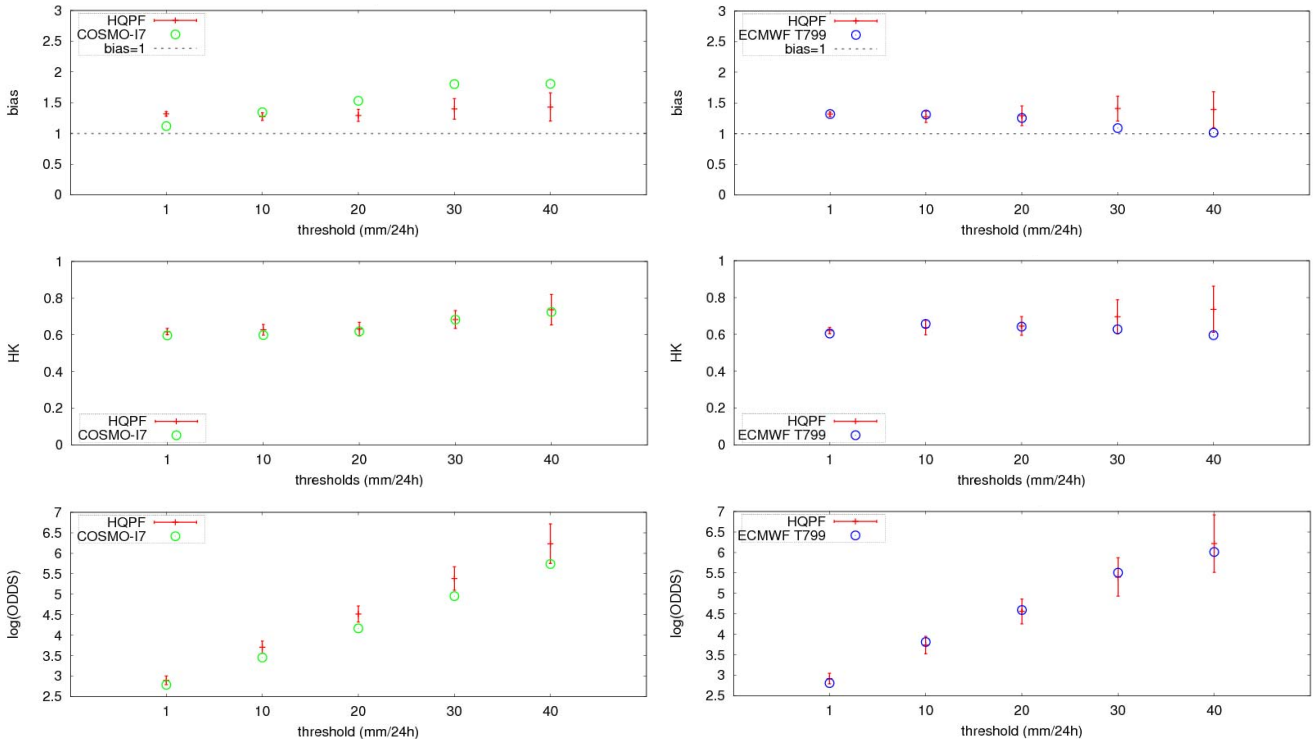

Fig. 2. Indices relative to QPFA (see text for explanation) in the interval +24 h/+48 h as a function of the threshold. In the left column there is the comparison between HQPF (in red) and COSMO-I7 (in green), while in the right column the comparison between HQPF (in red) and IFS (in blue). The BIAS score, HK score, and ODDS ratio score in logarithmic scale are indicated in the top line, middle line and bottom line panels, respectively. The plotted error bars indicate 2.5th and 97.5th percentiles of a resampled distribution for the difference itself.

the difference itself added to the metric of a selected forecast system: when the metric of the other forecast system is outside this interval, the differences may be considered statistically significant with a confidence of 95% (see Figs. 2 and 3). Note that this test is symmetric, that is we can also add the confidence intervals to the other forecast system.

4 Results

We have verified three different forecast systems (COSMO-I7, IFS, Human) and two different variables for each system (QPFA and QPFM). The samples have been stratified into quasi-homogeneous subsets by threshold, and by the forecast time (e.g. we verified QPFA and QPFM for 24 h, from +24 h/+48 h and +48 h/+72 h, for threshold from 1 mm/24 h to 40 mm/24 h). The comparison is always made between HQPF and model QPF and in the follow we show a synthesis of the results. For sake of brevity, in this work, we focus our attention to the +24 h/+48 h time interval, also because the warning time period is limited up to +48 h.

4.1 Quality of the forecast

Figure 2 shows the BIAS, HK and ODDS for the three forecast systems (for +24 h/+48 h and QPFA) compared in pairs in order to better visualize if there are any statistical significance differences. The COSMO-I7 model overestimates the QPF, more than the HQPF except for the lowest threshold, while the IFS model tends to overestimate in a similar way to HQPF until 20 mm/24 h, and for greater threshold it has a BIAS around 1. Regarding HK index, there are significant differences only between HQPF and IFS for the highest thresholds, with better score for the HPQF. For almost all thresholds the ODDS ratio is greater for HQPF than for COSMO-I7, while there is no significant difference between HQPF and IFS.

4.2 Value of the forecast

Figure 3 shows the comparison between the QPFA of COSMO-I7 vs HQPF and IFS vs HQPF considering only the 40 mm/24 h threshold for sake of brevity. In these plots it is possible to see not only the difference in Value but also the absolute values of the two forecast system competitors. Note that HQPF for +24 h/+48 h has a higher Value for C/L

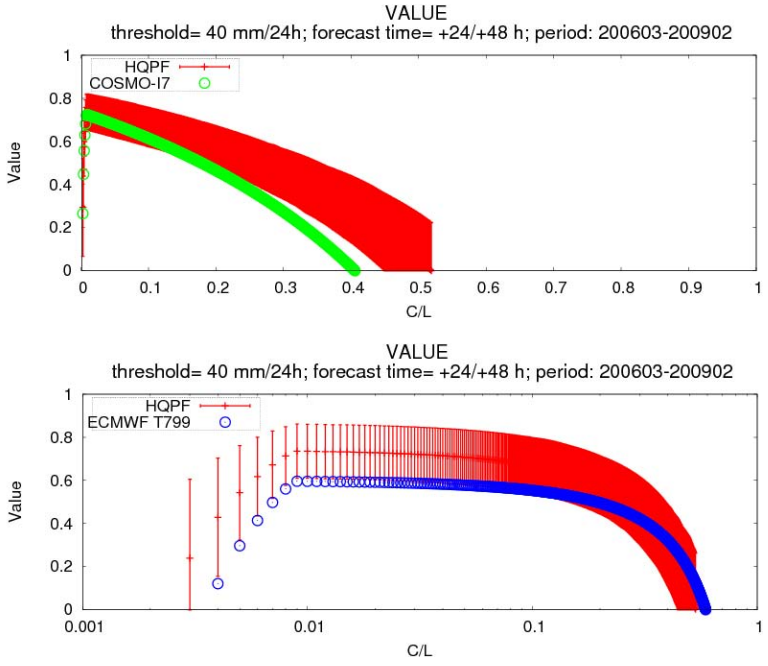

Fig. 3. Comparison between the Value of HQPF and, respectively, COSMO-I7 (top) and IFS (bottom), for the 40 mm/24 h threshold and different C/L ratios, considering QPFA and the forecast time +24 h/+48 h. Note that the x-axis of the bottom plot is in logarithmic scale. The plotted error bars indicate 2.5th and 97.5th percentiles of a resampled distribution for the difference itself.

ratios larger than ~ 0.2 with respect to COSMO-I7 (Fig. 3 top) while, in comparison with IFS, this is valid for small C/L ratios (Fig. 3 bottom).

In Fig. 4 we show a different way to plot the Value as a function of the C/L ratio and of the thresholds. In these graphs there are the differences among HQPF and model QPF only when this difference is statistically significant. In this way we can summarize the results obtained in this study. It is interesting to note that all the three forecast systems have Values and the HQPF is superior to IFS and COSMO-I7 for the lowest and the highest thresholds (except for some C/L with 30 mm/24 h). Probably the reason of this added value is that the HQPF is a good compromise between these two different models, which have two different error behaviors. The human added value appears in the decrease of the COSMO-I7 false alarms without loosing hits.

Regarding the maximum value (Fig. 5), it has to be pointed out that there is a Value and the best forecast is HQPF (for any threshold with C/L ratios less than 0.2–0.3). The QPFM is a very challenging target and the model errors are large. The forecaster is aware of this, he knows the territory and the climatology and reflects the uncertainty (for instance) in extending precipitation from a warning area to a neighboring one.

5 Conclusions

It is important to underline that the conclusions refer to the analysis of the Piemonte operational alert system, so they cannot be directly taken as universally true. But we believe that some of the main lessons that can be derived from this study could be useful for the meteorological community. In details, the main conclusions are the following:

– the three different forecast systems have Value in forecasting basin average values and maximum values;

– despite the overall improvement in global scale and the fact that the resolution of the limited area models has increased considerably over recent years, the QPF produced by the meteorological models involved in this study has not improved enough to allow its direct use at least for civil protection purposes: the subjective HQPF continues to offer the best performance for the period +24h/+48h (i.e. the warning period in Piemonte system) in forecast the average value. Moreover for all the considered periods but for C/L less than 0.2/0.3, also the maximum values confirm this hypothesis;

– in the forecast process, the step where humans have the largest added value with respect to mathematical models is the communication. In fact the human characterization and communication of the forecast uncertainty to end users cannot be replaced by any computer code. It is

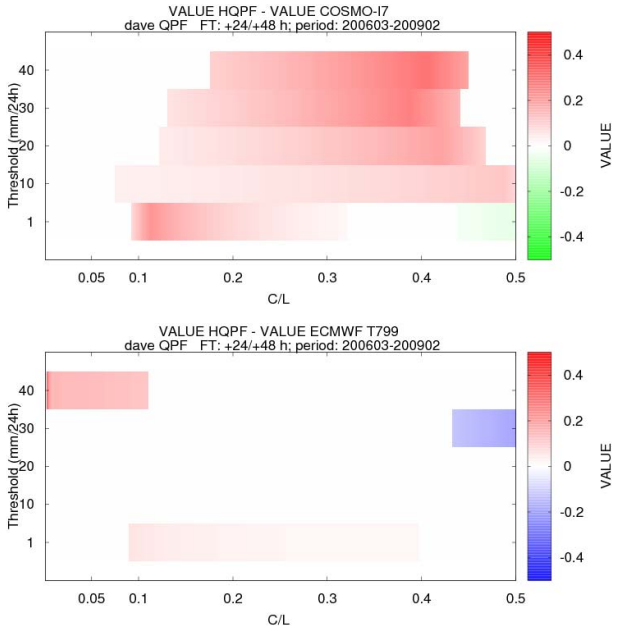

Fig. 4. Maps of the Value differences between HQPF and, respectively, COSMO-I7 (top) and IFS (bottom), for different thresholds and different C/L ratios, considering QPFA and the forecast time +24 h/+48 h. Note that only when the difference between HQPF Value and model QPF Value is statistically significant this difference is plotted. The bar indicates that when there are positive values (red color), HQPF has more Value than the model.

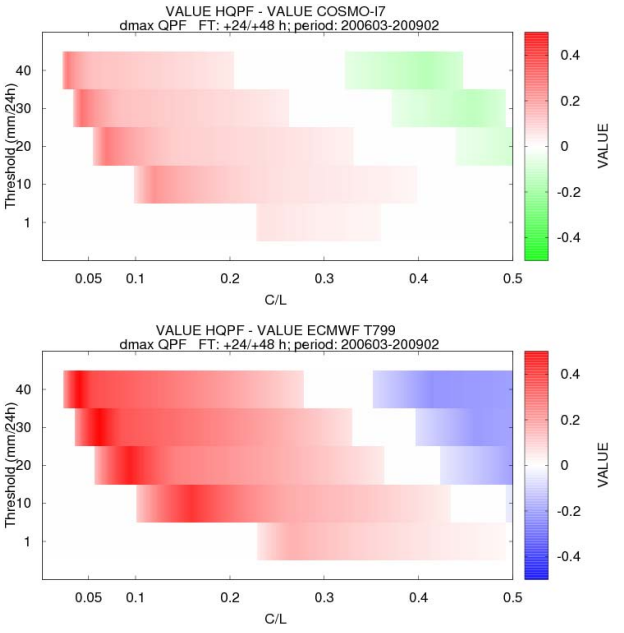

Fig. 5. Maps of the Value differences between HQPF and, respectively, COSMO-I7 (top) and IFS (bottom), for different thresholds and different C/L ratios, considering QPFM and the forecast time +24 h/+48 h. Note that only when the difference between HQPF Value and model QPF Value is statistically significant this difference is plotted. The bar indicates that when there are positive values (red colour), HQPF has more Value than the model.

important to know and to communicate the forecast error to the user for several reason (see WMO (2008)). In this context we want to underline that the uncertainty is intrinsic in the forecast process: the ratio between false alarms and misses could be defined in agreement between the forecaster and the decision makers, remembering that the misses can be reduced only increasing the false alarms and vice versa. So it is important that users know the QPF error since the choice of the thresholds is a function of the skill and of the C/L ratio: the forecast chain can be improved by the communication between forecaster and user;

– the QPF verification has feedback for modelers, forecaster, decision maker and general public. The correct application of appropriated statistical techniques permits a better definition and quantification of the errors and, mostly important, allows a correct (unbiased) communication between forecasters and users. For example:

  – without confidence bars it is difficult to interpret the result;

  – the confusion between $p(F|O)$ and $p(O|F)$ may bring to decision maker errors in every day life (Göber et al., 2004, Gigerenzer, 2002);

  – the incorrect use of the indices may bring to incorrect conclusions, as shown in Harvey et al. (1992);

  – the Value index is important, but it should be used together with the classical indices.

– following the studies of Roebber et al. (1996a) and Roebber et al. (1996b) and the forecaster added value in QPF we conclude that the experience is important in forecast skill, in particular in a small but complex domain like the one studied here.

In order to complete the current work, based on an accumulation period of 24h and on the forecast time +24 h/+48 h, a future analysis will consider other accumulation periods (for instance in 12 h) and forecast time. It would be interesting to classify the data seasonally and to try to analyze a case study in economical terms, evaluating the forecast Value in a real (and therefore complex) situation.

Acknowledgements. This work is supported by the Italian Civil Protection Department. We thank MeteoSwiss for the availability of the Ticino rain gauges. We are grateful to our colleagues at ARPA Piemonte for the useful discussions and, in particular, Renata Pelosini for her support.

Edited by: S. C. Michaelides
Reviewed by: two anonymous referees

References

Accadia, C., Casaioli, M., Mariani, S., Lavagnini, A., Speranza, A., de Venere, A., Inghilesi, R., Ferretti, R., Paolucci, T., Cesari, D., Patruno, P., Boni, G., Bovo, S., Cremonini, R.: Application of a statistical methodology for limited area model intercomparison using a bootstrap technique, Il Nuovo Cimento No. 26, 61–77, 2003a.

Accadia, C., Mariani, S., Casaioli, M., and Lavagnini, A.: Sensitivity of precipitation forecast skill scores to bilinear interpolation and simple neareast-neighbor average method on high-resolution verification grids, Weather Forecast., 18, 918–832, 2003b.

Baars, J. A. and Mass, C. F.: Performance of National Weather Service forecasts compared to operational, consensus, and weighted model output statistics, Weather Forecast, 20, 1034–1047, 2005.

Campus, S., Barbero, S., Bovo, S., Forlati, F.: Evaluation and Prevention of Natural Risks, Taylor & Francis, 2007.

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich M., Damrath, U., Ebert, E. E., Brown, B. G. and Mason, S.: Review Forecast verification: current status and future directions, Meteorol. Appl., 15, 3–18, 2008.

Charba, J. P., Reynolds, D. W., McDonald, B. E. and Carter, G. M.: Comparative Verification of Recent Quantitative Precipitation Forecasts in the National Weather Service: A Simple Approach for Scoring Forecast Accuracy, Weather Forecast., 18, 161–183, 2003.

Gigerenzer, G.: Reckoning with Risk. Learning to Live with Uncertainty, London Penguin, 2002.

Göber M., Wilson, C. A., Milton, S. F., and Stephenson, D. B.: Fairplay in the verification of operational quantitative precipitation forecasts, J. Hydrol., 288, 225–236, 2004.

Harvey, L. O. Jr., Hammond, K. R., Lusk, C. M., and Mross, E. F.: The application of signal detection theory to weather forecasting behavior, Mon. Weather Rev., 120, 863–883, 1992.

Hamill, T. M.: Hypothesis tests for evaluating numerical precipitation forecasts, Weather Forecast., 14, 155–167, 1999.

Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification: A Practitioner's Guide in Atmospheric Science, John Wiley and Sons, 2003.

Mason S.: Understanding forecast verification statistics, Meteorol. Appl., 15, 31–40, 2008.

Rabuffetti, D. and Barbero, S.: Operational hydro-meteorological warning and real-time flood forecasting: the Piemonte Region case study, Hydrol. Earth Syst. Sci., 9, 457–466, 2005, http://www.hydrol-earth-syst-sci.net/9/457/2005/.

Richardson, D. S.: Skill and relative economic value of the ECMWF ensemble prediction system, Q. J. Roy. Meteor. Soc., 126, 649–667, 2000.

Roebber, P. J. and Bosart, L. F.: The contributions of education and experience to forecast skill. Weather and Forecasting No. 11, 21-40, 1996a.

Roebber, P. J., Bosart, L. F. and Forbes, G. J.: Does distance from the forecast site affect skill? Weather Forecast., 11, 582–589, 1996b.

Roebber, P. J. and Bosart, L. F.: The Complex Relationship between Forecast Skill and Forecast Value: A Real-World Analysis, Weather Forecast, 11, 544–559, 1996c.

Sanders, F.: Trends in skill of Boston forecasts made at MIT, 1966–84. B. Am. Meteorol. Soc., 67, 170–176, 1986.

Stephenson, D. B.: Use of the odds ratio for diagnosing forecast skill. Weather Forecast., 15, 221–232, 2000.

Thornes, J. E. and Stephenson, D. B.: How to judge the quality and value of weather forecast products, Meteorol. Appl., 8, 307–314, 2001.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, 2006.

World Meteorological Organization (WMO): Guidelines on communicating forecast uncertainty. WMO/TD No. 1422, available at: www.wmo.int, 2008.