Nonlinear Processes in Geophysics (2001) 8: 401–417

Nonlinear Processes in Geophysics

Verification of intense precipitation forecasts from single models and ensemble prediction systems

F. Atger[1]

[1]Météo-France, Toulouse, France

Received: 22 September 2000 – Revised: 3 May 2001 – Accepted: 29 May 2001

Abstract. The performance of single models and ensemble prediction systems has been investigated with respect to quantitative precipitation forecasts. Evaluation is based on the potential economic value of +72 h/+96 h forecasts. The verification procedure consists of taking into account all precipitation amounts that are predicted in the vicinity of an observation in order to compute spatial, multi-event contingency tables. A probabilistic forecast from an ensemble can thus be compared to a probabilistic forecast from a single model run. The main results are the following: (1) The performance of the forecasts increases with the precipitation threshold. High levels of potential value reflect high hit rates that are obtained at the expense of a high frequency of false alarms. (2) The ECMWF ensemble performs better than a single forecast based on the same model, even when the resolution of the ensemble is lower. This is true for the NCEP ensemble as well, but only for morning precipitations. (3) The ECMWF ensemble performs better than the 5-member NCEP ensemble running at 12:00 UTC, even when the population of the former is reduced to 5 members. (4) The impact of reducing the population of the ECMWF ensemble is rather small. Differences between 51 members and 21 members are hardly significant. (5) A 2-member poorman ensemble consisting of the control forecasts of the ECMWF and the NCEP ensembles performs as well as the ECMWF ensemble for afternoon precipitations.

1 Introduction

An important aspect of the performance of weather prediction systems is their ability to accurately forecast intense precipitation events, i.e. those events whose intensity is sufficiently exceptional to cause public disruption. Floods, for instance, represent an important loss for human communities all around the world. The increase in model resolution is believed to be an important factor for improvement with respect to the forecast of intense precipitations (Buizza et al., 1999). The impact of the resolution is particularly in question when comparing a high resolution single model to an ensemble prediction system (EPS) that is generally based on a lower resolution model (Buizza et al., 1997). On the other hand, it has been mentioned that a large number of ensemble members is required for successful detection of rare events (Buizza and Palmer, 1998). A densely populated ensemble distribution seems indeed more adequate than a single model run to detect those events that are located in the tails of the climate distribution.

A number of studies have been devoted, at least partially, to the comparative performance of EPSs and single models with respect to quantitative precipitation forecasts (QPFs). Richardson (2000) compared the relative economic value of a single model to the ensemble forecasts from the European Centre for Medium-range Weather Forecasts (ECMWF) with respect to QPF. Zhu et al. (2001) did similar work in the U. S. for the National Centers for Environmental Prediction (NCEP) operational forecasting system. In these studies, deterministic forecasts based on a single model are compared to probabilistic forecasts based on an EPS. The information content of a probabilistic forecast is essentially higher than that of a deterministic forecast, since it allows the user to select the right probability threshold that corresponds to his concern (Murphy, 1985). The results of most comparative studies are thus not surprising: EPS probabilistic forecasts are more accurate, skillful and valuable than deterministic single model forecasts (Toth et al., 1998).

Operational forecasters, however, use information from a single model as probabilistic guidance. This is particularly obvious when dealing with extreme events, whose a priori probability is very low, such as intense precipitations. Physical processes that are involved in extreme precipitation events are not taken into account very well in atmospheric models, due to approximations introduced, for example, by the parameterization of the convection, the limited horizontal and vertical resolution, and the poor representation of topography. As a consequence, events that rarely occur are

Correspondence to: F. Atger (frederic.atger@meteo.fr)

even less often predicted by operational models. However, the lack of performance of numerical models in that respect has never prevented operational forecasters to successfully forecast rare, extreme events, on occasion. Forecasters are apparently able to extract from the model output information indicating that an extreme event, although not explicitly predicted by the model, might still occur with significant probability.

Forecasters' judgments are essentially probabilistic (Murphy, 1993), even when the technical information available is purely deterministic, as is the case when a forecaster interprets a single model output. In the case of QPF, the forecaster's judgment can take the form of probabilities of certain thresholds being reached. For example, given a model forecasting of 10 mm/12 h, a forecaster might consider a 5% probability of less than 1 mm/12 h and an 80% probability of more than 8 mm/12 h, with the numbers depending on expected model biases and uncertainties. This probabilistic judgment is not necessarily stated explicitly, but it represents the basis of any statement, including the very deterministic "12 mm/12 h" that may be required for operational purposes. Furthermore, an experienced forecaster would not elaborate a QPF at a given location from the precipitation amount predicted by the model at only that location. Forecasters are well aware of the limitations of numerical weather prediction, especially the consequences of insufficient resolution or poor representation of the topography, as well as the effect of errors in the initial conditions. They generally consider the whole model output in order to obtain an opinion about the expected value of a meteorological variable at a given point. In other words, a forecaster takes advantage of the spatial distribution of a forecast variable in order to predict its local probability density function (pdf). In practice, a high amount of precipitation predicted by the model at a short distance from a given point may indicate a considerable risk of a high precipitation amount, even if no precipitation is predicted by the model at that point. The forecaster's experience, as well as considerations about orography, the expected weather pattern, the model resolution and other characteristics, play an important role in the way this indication is inferred from the available information. The forecaster's judgment can still be facilitated by the use of model output statistics (MOS), especially when explicit probabilistic forecasts are required (Carter, 1989).

Probabilistic EPS forecasts perform undoubtedly better at all lead times than deterministic forecasts based on a single model (Zhu et al., 2001). On the other hand, it has been shown that it is possible to beat an ensemble at early lead times with a probabilistic forecast based only on a single model output and model statistics, when considering upper level variables, such as 850 hPa temperature (Talagrand, 1997) or 500 hPa geopotential height (Atger, 1999). In the present article, a forecast procedure is designed in order to mimic the way in which an operational forecaster infers a QPF from a single model output. Single models with different resolutions, operational ensembles and a "poorman ensemble" (consisting of single model runs) are compared in order to: (i) assess the performance of operational forecasting systems for the prediction of intense precipitations; (ii) evaluate the usefulness of an EPS when used in conjunction with one or several higher resolution models; (iii) investigate the relative impact of model resolution and ensemble population on the performance of an EPS.

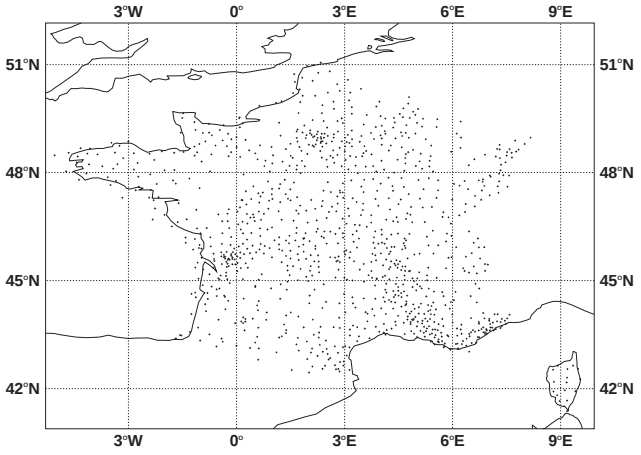

Fig. 1. The French network of rain gauges used in this study. The 1194 stations have reported at least one 12 h-precipitation amount during the winter of 1998–1999.

The article is organized as follows. The methodology is described in Sect. 2. The results are presented in Sect. 3, discussed in Sect. 4, and summarized in Sect. 5.

2 Methodology

2.1 Data

2.1.1 Observations

Observed precipitation data from the French rain gauges network have been collected from winter 1998–1999, i.e. 90 days from 1 December 1998 to 28 February 1999. Original data are 6 h accumulations at 1194 stations in France (Fig. 1). The final set consists of 12 h accumulations from 00:00 UTC to 12:00 UTC, and from 12:00 UTC to 00:00 UTC every day. Selected observations have successfully passed quality controls, so that gross departures from the climate are excluded. Due to missing or rejected data, the final set contains 194 191 values.

2.1.2 Forecasts

The verification procedure has been applied to single runs from the ECMWF model which was operational in winter 1998–1999 (Simmons et al., 1989; Courtier et al., 1991) in its high resolution version ECH (T_L319) and its lower resolution, ensemble prediction version ECL (T_L159). Both versions run at 12:00 UTC. The verification procedure has also been applied to single runs from the NCEP model, running at 12:00 UTC NC12 (T126 resolution up to +84 h,

T62 resolution afterwards) and at 00:00 UTC NC0 (T62 resolution). Concerning ensemble prediction, the verification procedure has been applied to the ECMWF EPS (Palmer et al., 1993; Molteni et al., 1996), which consists of 51 integrations of the T_L159 ECMWF model running at 12:00 UTC (ECEPS) and to the NCEP EPS (Tracton and Kalnay, 1993; Toth and Kalnay, 1997), which consists of 5 integrations of the T62 NCEP model running at 12:00 UTC (NCEPS12) and of 11 integrations of the T62 NCEP model running at 00:00 UTC (NCEPS0). Smaller ensembles have been constructed from the ECMWF EPS by retaining the control forecast and the first 10, 20, 32 perturbed members (ECEPS11, ECEPS21, ECEPS33). A 2-member "poorman ensemble" (ECNC) has been constructed from the single model runs described above. It consists of the ECMWF T159 model forecast ECL and the NCEP T126/T62 model forecast NC12.

Prior to verification, all forecasts have been interpolated onto the same 2.5°/2.5° grid that roughly corresponds to the horizontal resolution of the NCEP T62 model (the lowest resolution of the considered models). Forecasts have been retrieved over a large area surrounding France (56° N, 12° W, 36° N, 15° E) so that forecast data are available in a 500 km circular area around every available observation. In this section, precipitation accumulated from +72 h to +84 h and from +84 h to +96 h have been considered together. Morning precipitations (valid from 00:00 UTC to 12:00 UTC) and afternoon precipitations (valid from 12:00 UTC to 00:00 UTC) have been verified separately in Sect. 3.

2.1.3 Observed and forecast distribution

A cumulative distribution of 12 h precipitations has been computed from the whole set of selected observations. Approximately 100 cases (0.05%) have been identified with an accumulation exceeding 50 mm, 1000 observations (0.5%) with an accumulation exceeding 20 mm, and 10 000 observations (5%) with an accumulation exceeding 5 mm. Since the definition of an intense 12 h precipitation event is rather arbitrary, the 5 mm, 20 mm and 50 mm thresholds have been used as detection thresholds for verification in this study.

For precipitation thresholds from 1 mm to 50 mm (12 h accumulation), Fig. 2 shows the cumulative distribution of the observations and the corresponding forecasts of the different models used in the study, obtained from a bilinear interpolation at the observations. The impact of model horizontal resolution is clearly visible, with the forecast distribution that is closer to the observed distribution corresponding to ECH, the ECMWF T319 model. Note that different results might have been obtained with forecasts interpolated onto a more accurate grid, especially for ECH, whose horizontal resolution is much sharper than the 2.5°/2.5° grid used in the study.

Figure 2 also shows that morning precipitations (00:00–12:00 UTC) and afternoon precipitations (12:00–00:00 UTC) have a different distribution. Intense precipitations are more frequent during the afternoon, probably because convection

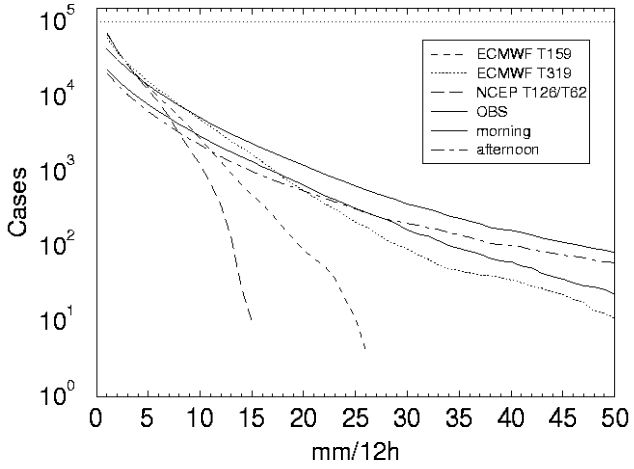

Fig. 2. Cumulative distribution of observed and forecast precipitation amounts for 50 thresholds from 1 mm/12 h to 50 mm/12 h. Observations: all (thick solid line), 00:00–12:00 UTC (thin solid line), 12:00–00:00 UTC (thin dash-dotted line). Forecasts (+84 h) interpolated at all observations: ECMWF T_L159 model (dashed line), ECMWF T_L319 model (dotted line), NCEP T126 model (long dashed line).

is more important after 12:00 UTC (this effect would have probably been emphasized if a summer season had been included in the considered period).

2.2 Contingency tables and relative operating curve

In a wide sense, forecast verification consists of a comparison of a distribution of forecasts $p(f)$ to a distribution of observations or analyses $p(x)$. The level of correspondence between $p(f)$ and $p(x)$ indicates how accurate the forecasting system is. There exist a number of methods to estimate this level of correspondence; the most widely used is the computation of the moments of the distribution of errors $p(f-x)$, which leads to the mean error, the mean square error, the standard deviation of the error, etc. The most informative approach, however, consists of a double factorization of the joint distribution of forecasts and observations (Murphy and Winkler, 1987).

$$p(f,x) = p(x|f)p(f) = p(f|x)p(x)\ , \qquad (1)$$

where the joint distribution $p(f,x)$ contains all of the information that is needed to evaluate the forecast's accuracy. By stratifying the data according to the forecasts, the joint distribution can be seen as the product of the distribution of forecasts $p(f)$ and the conditional distribution of observations, given the forecast $p(x|f)$. Similarly, by stratifying the data according to the observations, the joint distribution can be seen as the product of the distribution of observations $p(x)$ and the conditional distribution of forecasts, given the observation $p(f|x)$.

In the case of the deterministic forecast of a meteorological event, e.g. a precipitation amount exceeding 5 mm/12 h, the joint distribution is most generally represented as a 2×2

Table 1. Contingency table based on ECL (ECMWF T159 model, +72 h to +96 h) for the 5 mm/12 h observed threshold. H: number of Hits. FA: number of False Alarms. M: number of Misses. CR: number of Correct Rejections. $HR = H/(H+M) = 0.29$ (Hit Rate). $FAR = FA/(FA+CR) = 0.05$ (False Alarm Rate)

ECL 5 mm/12 h	Observed	Not observed
Forecast	$H = 4094$	$FA = 9426$
Not forecast	$M = 10061$	$CR = 170610$

contingency table. This table indicates, for a given observed (not observed) event, the number of times this event was predicted (non-predicted). Table 1 shows, for example, the contingency table for the 5 mm/12 h threshold and ECL. From this table, the stratification according to observations leads to two useful indicators: the hit rate (HR), which is the proportion of observed events that were successfully predicted and the false alarm rate (FAR), which is the proportion of non-observed events that were erroneously predicted.

In the case of a probabilistic forecast, the joint distribution can be represented similarly as a contingency table built from a number of probability categories. This table indicates, for a given observed (non-observed) event, the number of times every probability category is predicted (non-predicted). When verifying EPS forecasts, the categories are generally defined according to the number of ensemble members that forecast the event, from 1 to N (if N is the number of ensemble members). Table 2 shows, for example, an extract of the contingency table for the 5 mm/12 h threshold and ECEPS for a selection of probability categories based on the number of ensemble members. HR and FAR are computed separately for every category, so that the contingency table leads to an ensemble of pairs (FAR, HR). Every pair indicates the performance of a deterministic forecast that would be based on the fact that at least a certain number of ensemble members forecast the considered event.

It is convenient to plot these (FAR, HR) pairs as an ensemble of points on a diagram, forming the so-called Relative Operating Curve (ROC) (Mason, 1982). The relative position of the ROC obtained from a probabilistic forecasting system and the single point (FAR, HR) obtained from a deterministic forecasting system indicates their relative accuracy (Stanski et al., 1989). A single point above (below) the curve indicates that the deterministic system is more (less) accurate than the probabilistic system. Similarly, the relative position of the ROCs obtained from two probabilistic forecasting systems indicates their relative accuracy. Figure 3 shows, for example, the ROC for the 5 mm threshold and ECEPS. The (FAR, HR) point for ECL is plotted on the same figure. The position of the latter with respect to the former indicates a very similar overall performance of the two systems. Nevertheless, higher HRs (lower FARs) are attained by ECEPS for certain probability categories at the expense of higher FARs (lower HRs). For example,

Table 2. Contingency table based on ECEPS (ECMWF EPS, +72 h to +96h) for the 5 mm/12 h observed threshold. H_j (FA_j): number of Hits (False Alarms) for more than j members forecasting the event. $HR_j = H_j/\Sigma H_j$ (Hit Rate). $FAR_j = FA_j/\Sigma FA_j$ (False Alarm Rate). Example: $HR_2 = 0.78$; $FAR_2 = 0.28$. The number of forecast categories is 51 (ensemble members)

ECEPS 5 mm/12 h	Observed	Not observed
Forecast at least by 1 forecast	$H_1 = 12263$	$FA_1 = 67534$
Forecast at least by 2 forecasts	$H_2 = 11031$	$FA_2$50410
⋮	⋮	⋮
Forecast at least by j forecasts	H_j	FA_j
⋮	⋮	⋮
Forecast at least by 40 forecasts	$H_{40} = 105$	$FA_{40} = 123$
⋮	⋮	⋮
Forecast at least by 51 forecasts	$H_{51} = 0$	$FA_{51} = 0$

$HR = 0.78$ and $FAR = 0.28$ for the second probability category based on ECEPS ("at least 2 members are forecasting more than 5 mm/12 h"), while $HR = 0.29$ and $FAR = 0.05$ for the deterministic forecast based on ECL.

2.3 The cost-loss ratio

Figure 3 shows clearly that the advantage of a probabilistic forecast comes primarily from the fact that certain probability categories lead to higher HRs or lower FARs than those obtained with a single deterministic forecast. This is at the expense of an increase in the FAR or a decrease in the HR. For certain forecast users, a higher HR is valuable enough to tolerate a larger number of false alarms, typically when a user has the power to avoid a high loss L by protecting at low cost C. An example is given by the protection of Bordeaux vineyards from the frost in early spring: given the importance of the potential loss and the relatively low cost of protection, vineyards are protected as soon as the risk of frost exists, even when this risk is low. The so-called cost-loss ratio C/L is low. Another extreme example of low C/L is the protection of human life in the case of the risk of a dangerous meteorological event (e.g. storm, flood). The loss of a human life is incredibly high and the cost to protect it is generally low, so that C/L tends toward zero.

Other users do not tolerate false alarms. Due to a high C/L, they require a FAR as small as possible, even if this condition implies a decrease in the HR. High C/L are typical of long-term decision making situations, for example, the management of energy production: activation/deactivation of a nuclear reactor unit costs a lot, but the expected loss (or benefit, in this case) is limited.

Although all forecast users would benefit from points of the ROC that are ideally located close to the top left corner of Fig. 3, high and low C/L users do not benefit from the same part of the curve; low C/L users benefit from points

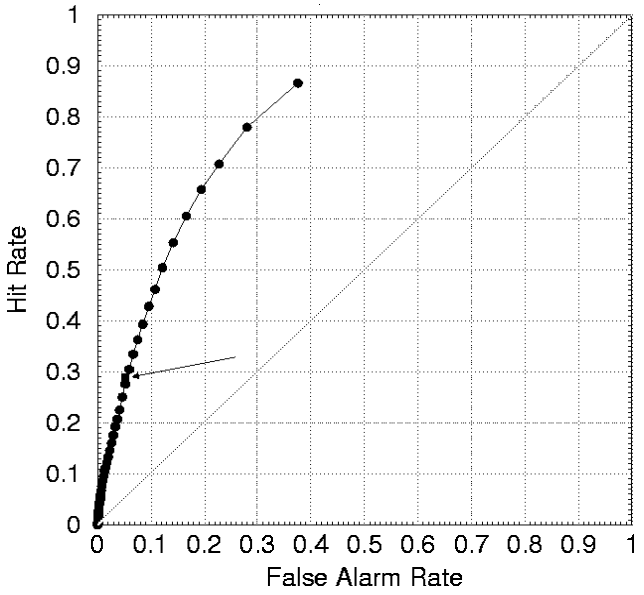

Fig. 3. Relative Operating Curve (ROC) for the 5 mm/12 h observed threshold, drawn from the contingency table shown in Table 2, based on ECEPS (ECMWF EPS, +72 h to +96 h). From the top of the curve, every point indicates the performance of a deterministic forecast based on the fact that at least 1, 2, 3, etc., ensemble members forecast at least 5 mm/12 h (51 points). The single square indicated by the arrow is the (FAR, HR) point drawn from the contingency table shown in Table 1, based on ECL (ECMWF T159 model, +72 h to +96 h). It indicates the performance of a deterministic forecast based on the fact that the model forecasts at least 5 mm/12 h.

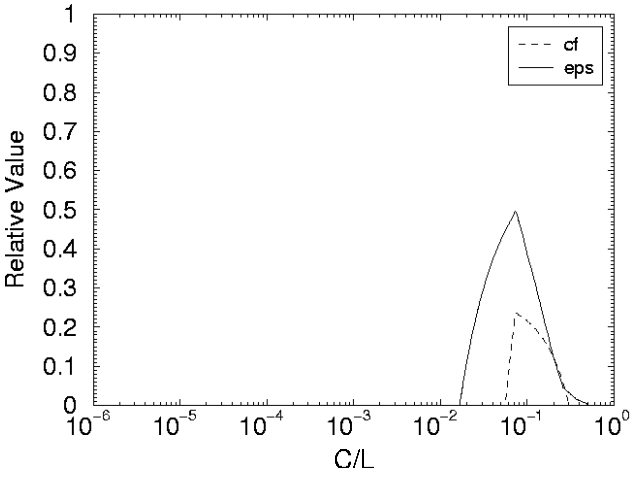

Fig. 4. Relative value for the 5 mm/12 h observed threshold, as a function of the user C/L, based on the contingency tables shown in Table 1 and Table 2. Dash line: ECL (ECMWF T159 model, +72 h to +96 h). Solid line: ECEPS (ECMWF EPS, +72 h to +96 h). By construction, the maximum value is obtained in both cases for $C/L = f$, the frequency of occurence of the event (see Eq. 3).

of the ROC that are located in the upper part of the curve (higher HR), while high C/L users benefit from points in the lower part of the curve (lower FAR).

2.4 Relative value

Ångstrøm (1919) was probably the first who introduced the concept of value in the field of weather forecasting (Liljas and Murphy 1994). After Murphy (1977), several authors explored multiple aspects of the usefulness/value problem in the 70's and 80's (Katz and Murphy, 1997). According to this approach, users of weather forecasts are "decision makers": they have to take different decisions according to the expected weather conditions. The usefulness of a weather forecast can thus be quantified by considering the occasions when the use of the forecast has been beneficial, detrimental or neutral to the user, with respect to the process of decision making.

Here we consider the particular situation when a user requires a forecast in order to avoid potential damages caused by adverse weather conditions, e.g. intense precipitations. A simple economic model can be applied when the user has just two alternatives: to protect or to do nothing. The cost of protection C is known, as well as the expected loss L occurring in case of damage. If no weather forecast is available, the decision to protect is likely to be based on climate knowledge. If f_c is the expected climatological frequency of the event, it is easy to show that the user should always protect if $C/L < f_c$, otherwise the user should never protect. Let f be the actual frequency of the event during the considered period. Under the assumption that $f_c = f$, the mean expense (per unit loss) $\mathrm{ME}_{\mathrm{climate}}$ is, therefore, the min of C/L and f.

On the other hand, a perfect knowledge of the future weather would allow the user to protect only when the event occurs, so that $\mathrm{ME}_{\mathrm{perfect}}$ would be the product of C/L and f. The relative economic value of a weather forecast (V) is defined as the amount of money that is saved by the user, normalized by the amount of money that he could save by using a perfect (hypothetical) forecast:

$$V = \frac{\mathrm{ME}_{forecast} - \mathrm{ME}_{climate}}{\mathrm{ME}_{perfect} - \mathrm{ME}_{climate}} \,. \quad (2)$$

Relative value is thus a skill-score based on the mean expected expense, according to the usual definition of the forecast skill (e.g. Stanski et al., 1989). The maximum value $V = 1$ is obtained by a perfect forecast, and $V = 0$ for the climate forecast. From the above discussion about the relative importance of higher HR and FAR for different categories of users, the relative value can be expressed as a function of the user's C/L on the one hand, and as a function of the forecast FAR and HR, on the other hand. Richardson (2000) has demonstrated the following relation:

$$V = \left(\min\left(\frac{C}{L}, f\right) - FAR\frac{C}{L}(1-f) + HR \cdot \left(1 - \frac{C}{L}\right) f - f \right) \left(\min\left(\frac{C}{L}, f\right) - f\frac{C}{L} \right)^{-1} . \quad (3)$$

It is important to note that this formulation is correct under the assumption that $f_c = f$, as mentioned above. In prac-

Table 3. Multi-event contingency table based on ECL (ECMWF T159 model, +72 h to +96 h) for the 5 mm/12 h observed threshold. H_k (FA_k): number of Hits (False Alarms) for the k mm/12 h forecast threshold. $HR_k = H_k/\Sigma H_k$ (Hit Rate). $FAR_k = FA_k/\Sigma FA_k$ (False Alarm Rate). The number of forecast categories is 20 (forecast thresholds)

ECL 5 mm/12 h	Observed	Not observed
Forecast > 1 mm/12 h	$H_1 = 11797$	$FA_1 = 60225$
Forecast > 2 mm/12 h	$H_2 = 9312$	$FA_2 = 34145$
⋮	⋮	⋮
Forecast > k mm/12 h	H_k	FA_k
⋮	⋮	⋮
Forecast > 5 mm/12 h	$H_5 = 4094$	$FA_5 = 9426$
⋮	⋮	⋮
Forecast > 20 mm/12 h	$H_{20} = 5$	$FA_{20} = 70$

Table 4. Multi-event contingency table based on ECEPS (ECMWF EPS, +72 h to +96 h) for the 5 mm/12 h observed threshold. $H_{j,k}$($FA_{j,k}$): number of Hits (False Alarms) for more than j members forecasting more than k mm/12 h. $HR_{j,k} = H_{j,k}/\Sigma H_{j,k}$ (Hit Rate). $FAR_{j,k} = FA_{j,k}/\Sigma FA_{j,k}$ (False Alarm Rate). The number of forecast categories is 20 (forecast thresholds) × 51 (ensemble members) = 1020

ECEPS 5 mm/12 h		Observed	Not observed
⋮	⋮	⋮	⋮
Forecast > k mm/12 h	At least by 1 forecast	$H_{k,1}$	$FA_{k,1}$
Forecast > $k$mm/12 h	⋮	⋮	⋮
Forecast > k mm/12 h	At least by 51 forecasts	$H_{k,51}$	$FA_{k,51}$
⋮	⋮	⋮	

tice, f is not known before the end of the verification period. The climate forecast is based only on the knowledge of f_c and is not as reliable as it might be if it was based on the knowledge of f, the actual frequency of occurrence of the event. $ME_{climate}$ is, therefore, underestimated in Eq. (2), which has a slight impact on the computed value. The above formulation has, however, been used in most studies, since the computation can be done from the sample only, with no need for independent climatological data. It has been used in the present study for the same reasons.

When considering a probabilistic forecast, there are as many (FAR, HR) pairs as probability categories. For a given C/L, it is, therefore, convenient to consider the maximum value that is attained for the probability category that is optimal for the user, i.e. that leads to the better compromise between a low FAR and a high HR (Richardson, 2000). Figure 4 shows, for example, the value as a function of C/L, for the 5 mm/12 h observed threshold, for the deterministic forecast based on ECL and the probabilistic forecast based on ECEPS (same forecasts as Fig. 3). The ECEPS curve is, in fact, the envelope of the 51 curves of value that are obtained for every forecast category, from "at least 1 member forecasting the event" to "all members forecasting the event". The better performance of the probabilistic forecast based on ECEPS is clearly visible, especially for lower C/L.

2.5 Multi-event contingency tables

In the previous subsection, it has been described how the performance of a deterministic forecast based on a single model can be investigated from a simple 2 × 2 contingency table, which gives a simplified representation of the joint distribution of forecasts and observations, limited to the forecasts and observations of one specified event (see Sect. 2.2). Multi-event contingency tables give a more complete representation of the joint distribution. A table indicates, for a given observed (non-observed) event, the number of times different events are predicted (non-predicted). This approach has been followed in seasonal prediction verification studies (e.g. Mason and Graham, 1999).

Table 3 shows, for example, the multi-event contingency table of ECL for the 5 mm/12 h observed threshold, based on 20 forecast thresholds from 1 mm/12 h to 20 mm/12 h. Higher forecast thresholds are not used since they occur very rarely, partly due to the coarse interpolation grid that has been used. Similar to a probabilistic forecast contingency table (e.g. Table 2), a multi-event contingency table leads to several (FAR, HR) pairs, each of which indicates the performance of a deterministic forecast that would be based on the fact that a specified forecast threshold is reached by the model. Therefore, the ensemble of (FAR, HR) pairs indicates the performance of a probabilistic forecast based on a single model run. Figure 5a shows the ROC corresponding to Table 3. The performance is very similar to that shown in Fig. 3, corresponding to the probabilistic forecast based on ECEPS.

Multi-event contingency tables can be used for the verification of probabilistic forecasts based on an EPS as well. A table indicates, for a given observed (non-observed) event, the number of times different events are predicted (non-predicted) by at least a certain number of ensemble members. Table 4 shows, for example, an extract of the multi-event contingency table of ECEPS for the 5 mm/12 h observed threshold, based on 20 forecast thresholds from 1 mm/12 h to 20 mm/12 h. Figure 5b shows the ROC corresponding to Table 4. The performance is improved, compared to Fig. 3 (ECEPS) and Fig. 5a (ECL multi-event), in the upper part of the curves where forecasts are primarily beneficial to lower C/L users.

Figure 6 shows the relative value as a function of C/L for the multi-event contingency tables based on ECL and

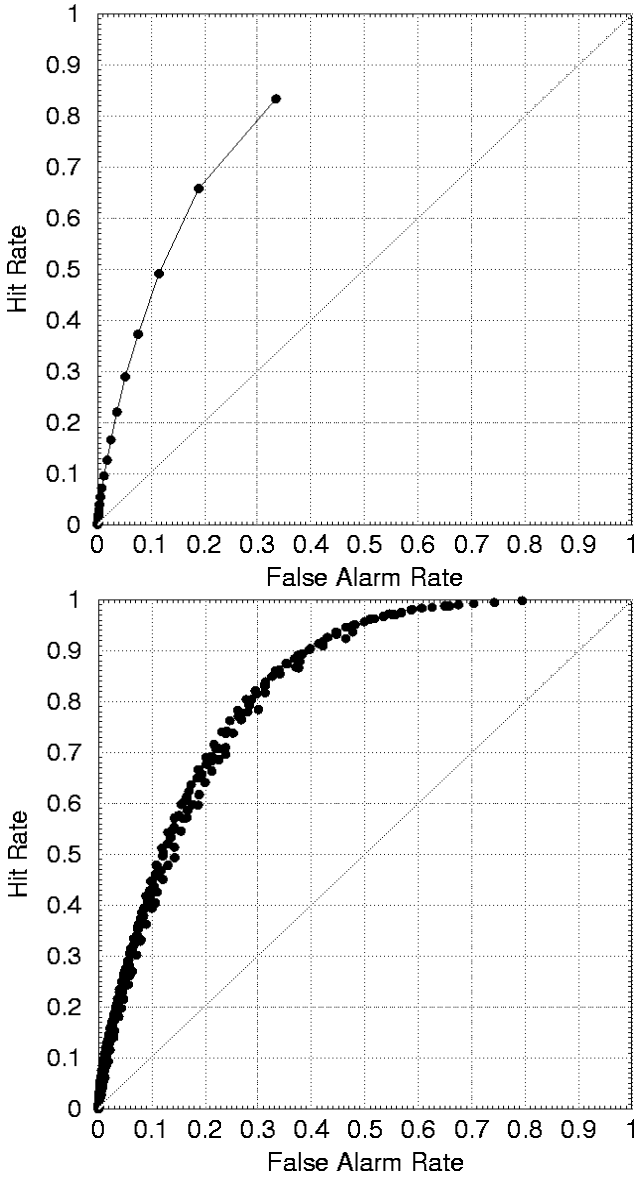

Fig. 5. Relative Operating Curve (ROC) for the 5 mm/12 h observed threshold, based on the multi-event contingency tables shown in Table 3 and Table 4. **(a)** ECL (ECMWF T159 model, +72 h to +96 h); from the top of the curve, every point indicates the performance of a deterministic forecast based on the fact that the model forecasts at least 1, 2, 3, etc., mm/12 h (20 points). **(b)** ECEPS (ECMWF EPS, +72 h to +96 h); every point indicates the performance of a deterministic forecast based on the fact that at least 1, 2, 3, etc., ensemble members forecast at least 1, 2, 3, etc., mm/12 h (51×20=1020 points).

ECEPS (same forecasts as in Fig. 5a and Fig. 5b). The ECL curve is, in fact, the envelope of the 20 curves of value that are obtained for every forecast threshold, from 1 mm/12 h to 20 mm/12 h. The ECEPS curve is the envelope of the 20×51=1020 curves of value that are obtained for every forecast category, from "at least 1 member forecasting more than 1 mm/12 h" to "all members forecasting more than 20 mm/12 h". The forecast based on ECEPS is only slightly

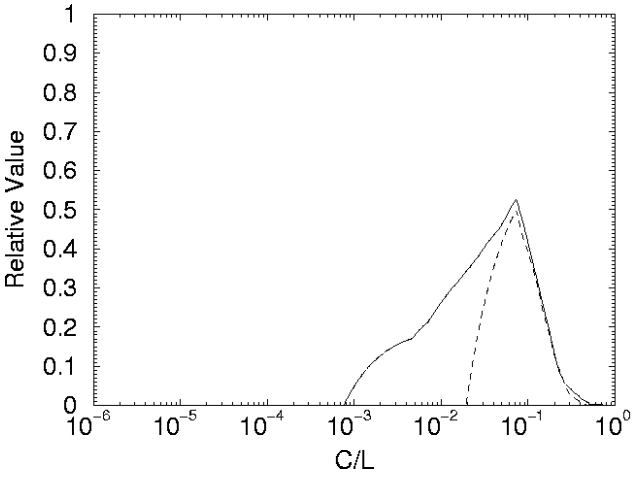

Fig. 6. Relative value for the 5 mm/12 h observed threshold, as a function of the user C/L, based on the multi-event contingency tables shown in Table 3 and Table 4. Dash line: ECL (ECMWF T159 model, +72 h to +96 h). Solid line: ECEPS (ECMWF EPS, +72 h to +96 h).

better overall than the forecast based on ECL, but is much better for lower C/L.

2.6 Spatial contingency tables

Verification of QPF as well as most quantitative weather forecasts would ideally require one to consider the correspondence between 3-dimensional distributions of forecasts and observations: two dimensions for the physical space, and one dimension for time. In practice, verification generally consists of an evaluation of the correspondence between (time distributions of) local forecasts and local observations, as described in the previous subsections. Space connections between local forecasts and local observations are rarely considered. A spatial approach of verification would consist of an evaluation of the correspondence between (time distributions of) spatial distributions of forecasts and spatial distributions of observations. One application of this approach is the evaluation of the correspondence between forecast and observed meteorological patterns, through the computation of Anomaly Correlation or the categorization of large-scale circulation patterns (Chessa and Lalaurette, 2000). Another application is the evaluation of the correspondence between a local observation and the local forecasts that are found in the vicinity of this observation.

Spatial multi-event contingency tables have been used in the present study. Each table indicates, for a given observed (non-observed) event, the number of times different events are predicted (non-predicted) at different distances from the observed event. Table 5 shows, for example, an extract of the spatial multi-event contingency table of ECL for the 5 mm/12 h observed threshold, based on 20 forecast thresholds from 1 mm/12 h to 20 mm/12 h, at 100 km, 200 km, 300 km, 400 km and 500 km from the observation.

Table 5. Spatial multi-event contingency table based on ECL (ECMWF T159 model, +72 h to +96 h) for the 5 mm/12 h observed threshold. $H_{k,l}(FA_{k,l})$: number of Hits (False Alarms) for more than k mm/12 h forecast at less than l×100 km from the observation. $HR_{k,l} = H_{k,l}/\Sigma H_{k,l}$ (Hit Rate). $FAR_{k,l} = FA_{k,l}/\Sigma FA_{k,l}$ (False Alarm Rate). The number of forecast categories is 20 (forecast thresholds) × 5 (distances to the observation) = 100

ECL 5 mm/12 h		Observed	Not observed
⋮	⋮	⋮	⋮
Forecast > k mm/12 h	At less than 100 km from observation	$H_{k,1}$	$FA_{k,1}$
Forecast > k mm/12 h	⋮	⋮	⋮
Forecast > k mm/12 h	At less than 500 km from observation	$H_{k,5}$	$FA_{k,5}$
⋮	⋮	⋮	⋮

Figure 7a shows the ROC corresponding to Table 5. Although most of the points of Fig. 7a are located below those of Fig. 5a, the envelope of the curves is almost identical, except in the upper part of the curve where a higher HR can be obtained at the expense of a higher FAR. This part of the curve is obtained with forecast categories that are very "sensitive" in detecting the occurrence of rain, with the most sensitive at 1 mm/12 h at 500 km from the observation.

Table 6 shows an extract of the spatial multi-event contingency table of ECEPS for the 5 mm/12 h observed threshold, based on 20 forecast thresholds from 1 mm/12 h to 20 mm/12 h, at 100 km, 200 km, 300 km, 400 km and 500 km from the observation. Figure 7b shows the ROC corresponding to this contingency table. Again, most of the points of Fig. 7b are located below those of Fig. 5b, but the higher density of points leads to an envelope that is slightly better. Figure 8 shows the relative value as a function of C/L for the spatial multi-event contingency tables based on ECL and ECEPS (same forecasts as in Fig. 7a and Fig. 7b). The curves are, in fact, the envelopes of the curves of value that are obtained for every forecast category (5 × 20=100 categories for ECL, 5 × 20 × 51=5100 categories for ECEPS). The forecast based on ECEPS is better than the forecast based on ECL for lower C/L, but the difference is reduced compared to Fig. 6.

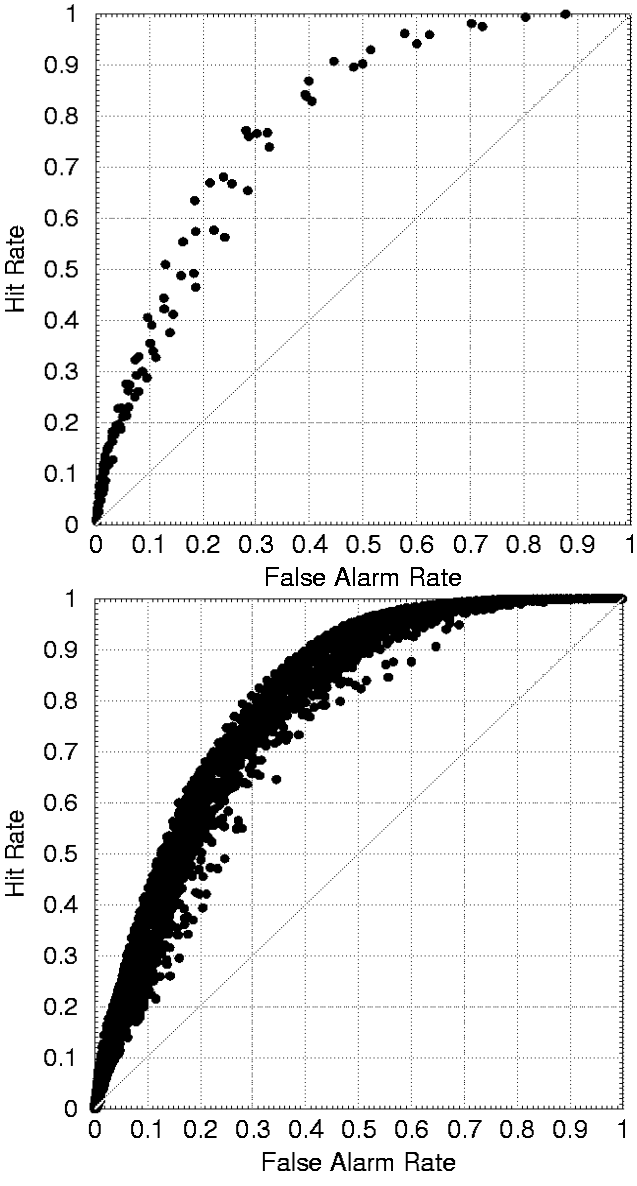

Fig. 7. Relative Operating Curve (ROC) for the 5 mm/12 h observed threshold, based on the spatial multi-event contingency tables shown in Table 5 and Table 6. **(a)** ECL (ECMWF T159 model, +72 h to +96 h); every point indicates the performance of a deterministic forecast based on the fact that the model forecasts at least 1, 2, 3, etc., mm/12 h at less than 100, 200, etc., km from the considered location (20 × 5=100 points). **(b)** ECEPS (ECMWF EPS, +72 h to +96 h); every point indicates the performance of a deterministic forecast based on the fact that at least 1, 2, 3, etc., ensemble members forecast at least 1, 2, 3, etc., mm/12 h at less than 100, 200, etc., km from the considered location (51 × 20 × 5=5100 points).

2.7 Significance tests

Figure 8 is an example of a comparison between two curves of value obtained from spatial multi-event contingency tables based on different forecasting systems. Some differences appear for a wide range of C/L ratios. Are these differences statistically significant? This question is particularly important when verifying the performance of forecasting systems with respect to extreme events, such as intense precipitation that rarely occurs in the data sample. Furthermore, the method of verification implies the use of a large number of forecast categories, which emphasizes the effect of insufficient sampling.

As pointed out by Hamill (1999), spatial correlation and the non-normality of errors make it difficult to use common

Table 6. Spatial multi-event contingency table based on ECEPS (ECMWF EPS, +72 h to +96 h) for the 5 mm/12 h observed threshold. $H_{j,k,l}(FA_{j,k,l})$: number of Hits (False Alarms) for more than j members forecasting more than k mm/12 h at less than $l \times 100$ km from the observation. $HR_{j,k,l} = H_{j,k,l}/\Sigma H_{j,k,l}$ (Hit Rate). $FAR_{j,k,l} = FA_{j,k,l}/\Sigma FA_{j,k,l}$ (False Alarm Rate). The number of forecast categories is 20 (forecast thresholds) × 51 (ensemble members) × 5 (distances to the observation) = 5100

ECEPS 5 mm/12 h			Observed	Not observed
⋮	⋮	⋮	⋮	⋮
Forecast > k mm/12 h	At least by j forecasts	At less than 100 km from the observation	$H_{j,k,1}$	$FA_{j,k,1}$
Forecast > k mm/12 h	At least by j forecasts	⋮	⋮	⋮
Forecast > k mm/12 h	At least by j forecasts	At less than 500 km from the observation	$H_{j,k,5}$	$FA_{j,k,5}$
⋮	⋮	⋮	⋮	⋮

hypothesis tests (e.g. t test) for assessing the significance of weather forecasting verification results. Computer-based methods of hypothesis testing have been used in this study to evaluate the significance of the results. A resampling method has been systematically applied in order to estimate the probability that differences in the relative value between two forecasting systems could have been obtained by chance.

The method consists of the construction of an empirical distribution of differences that are not statistically significant (Hamill, 1999). The probability that the actual difference belongs to this distribution, i.e. that the difference is not significant, is then evaluated. This null distribution is obtained by comparing the relative value for every C/L ratio of two sets of independent forecasts that should perform identically. These two sets are generated 1000 times by randomly choosing the forecasts from either one or the other forecasting systems. Since it is very likely that the errors are spatially correlated, all local forecasts valid for a given 12 h period are considered together as a unique case. Temporal correlation of errors is also probable. In order to limit the dependencies, forecasts valid for the 12:00–00:00 UTC period (afternoon precipitations) and for the 00:00–12:00 UTC period (morning precipitations) have been considered separately, so that no consecutive 12 h periods can be found in the sample.

The different steps of the procedure are as follows: (i) contingency tables are computed for every 12 h period for system A and system B; (ii) the sample of 12 h periods is randomly halved into 2 sub-samples; (iii) the relative value is computed separately from each sub-sample using the contingency tables; (iv) the difference between the relative value of the 2 sub-samples is computed; (v) the procedure is iterated 1000 times from (ii) to (iv); (vi) the probability that the difference between the actual value of system A and the actual value of system B is significant is estimated from the empirical distribution obtained at the end of step (v).

3 Results

Intense precipitation events occur more frequently during the afternoon. The results presented in this section are based solely on 12:00–00:00 UTC precipitations. Unless otherwise stated, the lead-time is +84 h (precipitations accumulated from +72 h to +84 h).

3.1 Ensemble vs. single run

An important requirement for an ensemble is that it performs better than a control single forecast based on the same model. As mentioned in Sect. 1, the superiority of probabilistic forecasts based on EPSs over deterministic control forecasts has been demonstrated. In this subsection, the value of an ensemble is compared to that of the control forecast on the basis of spatial multi-event contingency tables. This means that the performance of a probabilistic forecast based on the ensemble is compared to that of a probabilistic forecast based on the control single run. The latter is designed to represent the probabilistic judgment of an operational forecaster using a single model forecast. The comparison of these forecasts is meant to indicate the usefulness of an EPS in an operational environment, with respect to QPFs.

3.1.1 ECMWF ensemble vs. control forecast

Figure 9 shows the relative value of the ECMWF EPS (ECEPS) and the control forecast (ECL) for the total range of C/L ratios (0 to 1). For each C/L, the statistical significance of the difference between the curves of value has been evaluated through the resampling procedure described in Sect. 2. For the 5 mm/12 h observed threshold (Fig. 9a), ECEPS and ECL perform similarly for C/L above the optimal value (that corresponds to the sample frequency of the event, i.e. 0.07 approx.). For lower thresholds, as small as 10^{-4}, the superiority of ECEPS over ECL is confirmed by the curves of value, but the 90% significance level is reached only for a proportion of C/L ratios. By contrast, ECEPS is

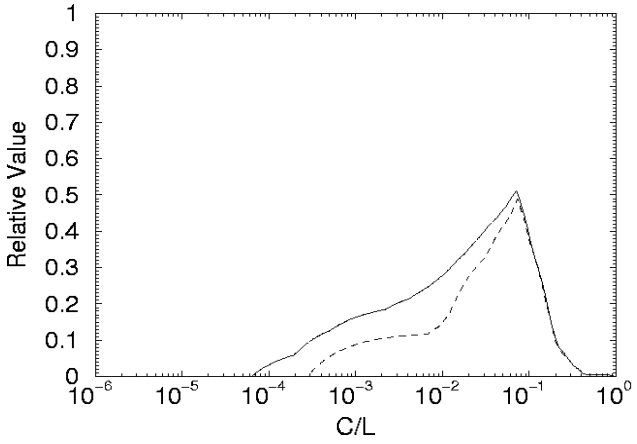

Fig. 8. Relative value for the 5 mm/12 h observed threshold, as a function of the user C/L, based on the spatial multi-event contingency tables shown in Table 5 and Table 6. Dash line: ECL (ECMWF T159 model, +72 h to +96 h). Solid line: ECEPS (ECMWF EPS, +72 h to +96 h).

significantly better than ECL with respect to the 20 mm/12 h threshold for a wide range of C/L ratios from approximately $2 \cdot 10^{-5}$ to 10^{-2} (Fig. 9b).

Although the 50 mm/12 h curves of value exhibit an advantage for ensemble forecasts for higher C/L ratios, significance tests show that ECEPS and ECL do not perform differently at the 90% level (Fig. 9c). Low significance of the results concerning the higher precipitation thresholds is probably due to, in this case as in many others presented below, the limited number of observed cases. For example, less than 100 cases of precipitations above 50 mm/12 h have been reported during the considered season. These 100 cases have occurred during 6 periods of 24 h, so that the number of independent observed cases is very small when only considering 12:00 UTC–00:00 UTC precipitations.

3.1.2 NCEP ensemble vs. control forecast

Figure 10 shows the relative value of the NCEP EPS running at 12:00 UTC (NCEPS12) and the corresponding control forecast (NC12). The differences are often not significant at the 90% level. When the differences are significant, NC12 is generally better than NCEPS12, except for smaller C/L ratios for 20 mm/12 h (Fig. 10b) and 50 mm/12 h (Fig. 10c). This result is rather disappointing, but indicates the importance of model resolution for QPF. The horizontal resolution of NC12 is T126, while the resolution of the 4 perturbed members of NCEPS12 is only T62.

The relative value of the NCEP EPS running at 00:00 UTC (NCEPS0) and the corresponding T62 control forecast (NC0) has been examined. The lead-time is +96 h, so that afternoon precipitations are considered (value is consequently lower than in the previous results where the lead-time is +84 h). Surprisingly, the result of the comparison (not shown) is similar to that obtained at 12:00 UTC, although NC0 and NCEPS0 have the same resolution (T62). This result seems

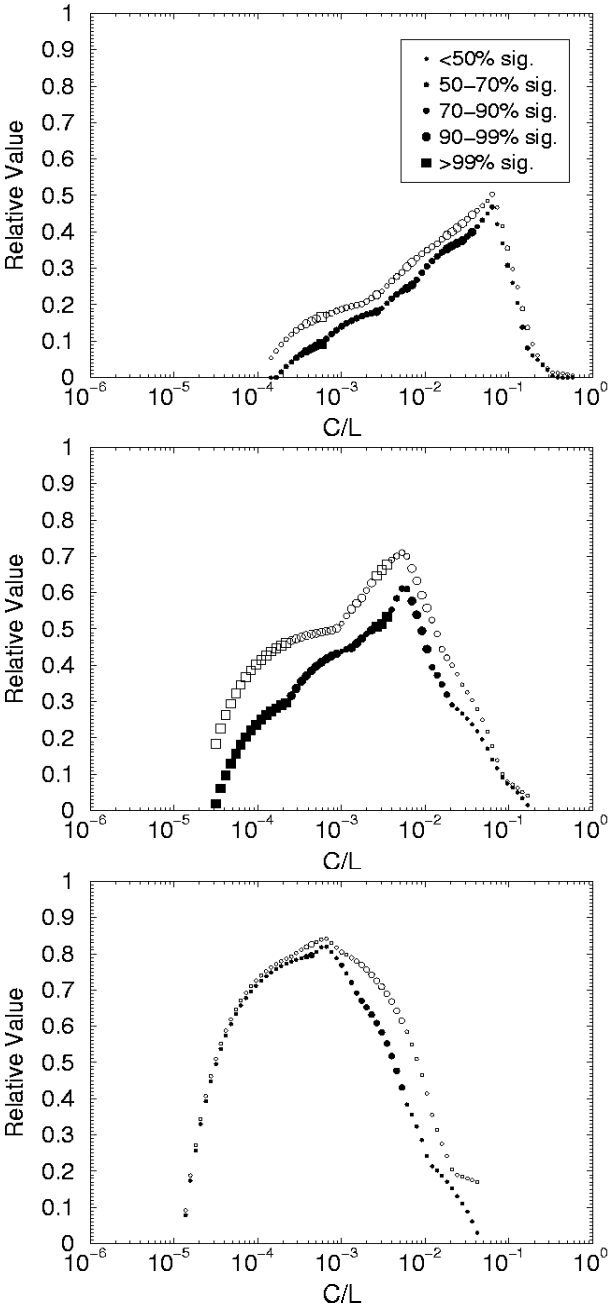

Fig. 9. Relative value as a function of the user C/L, based on spatial multi-event contingency tables. Afternoon precipitation only (+72 h to +84 h). Blank symbols: ECEPS (ECMWF EPS). Filled symbols: ECL (ECMWF T159 model). The size of the circles indicate the level of statistical significance of the difference in the value between the two forecasting systems: less than 50%, 50–70%, 70–90%, 90–99%. Squares indicate a level of significance above 99%. **(a)** 5 mm/12 h observed threshold. **(b)** 20 mm/12 h observed threshold. **(c)** 50 mm/12 h observed threshold. The legend indicated in Fig. 9a is valid for all figures from Fig. 9 to Fig. 16.

to contradict the conclusion of the previous paragraph about the importance of model resolution. It also contradicts the previous results based on 500 hPa geopotential height fore-

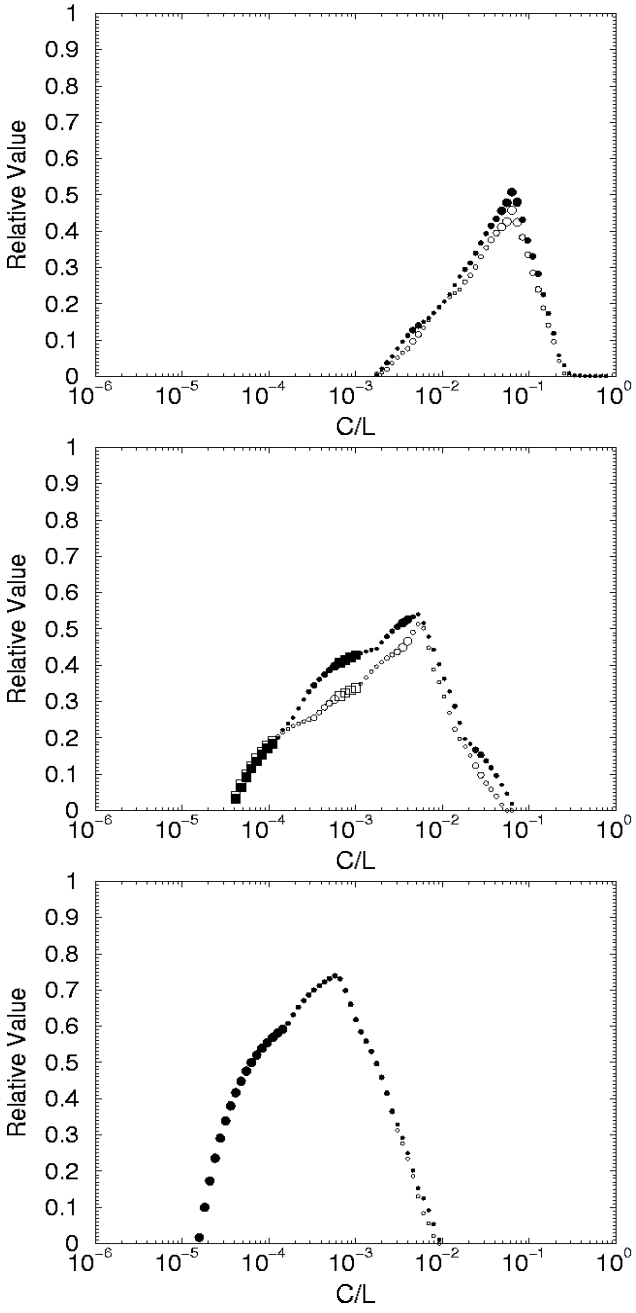

Fig. 10. Same as Fig. 9, but for NCEPS12 (NCEP EPS running at 12:00 UTC, blank symbols) and NC12 (NCEP T126 model, filled symbols). Differences are hardly visible, but there is a slight advantage for NCEPS12 in Fig. 10c.

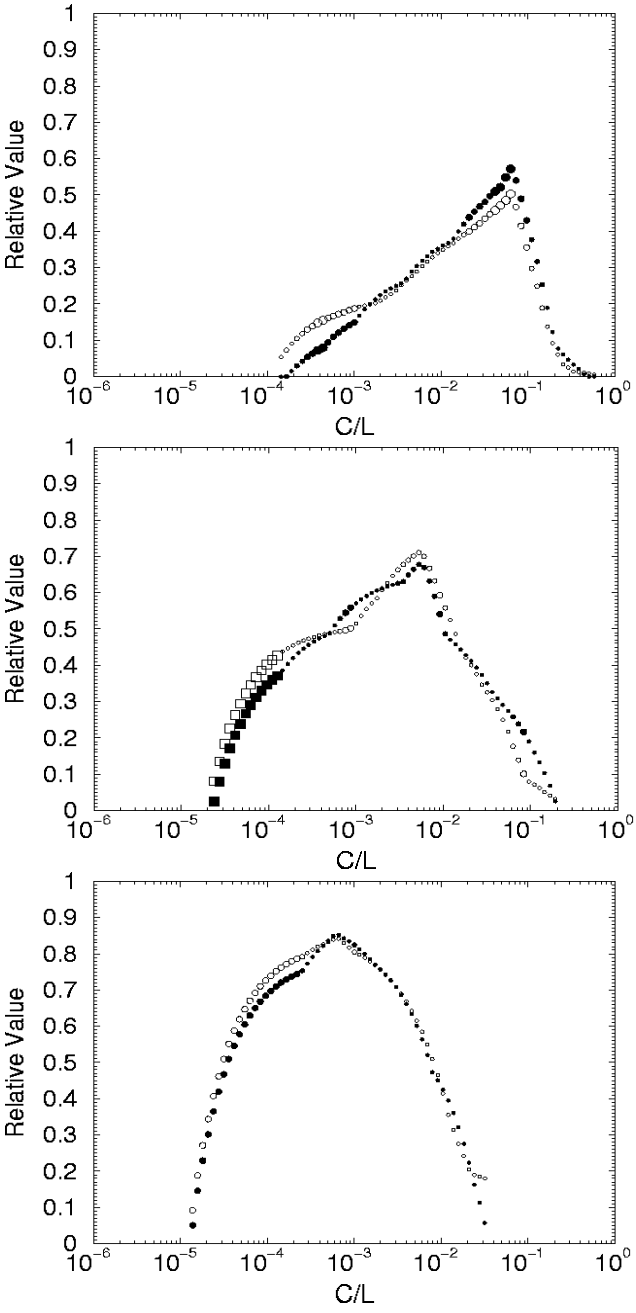

Fig. 11. Same as Fig. 9, but for ECEPS (ECMWF EPS, blank symbols) and ECH (ECMWF T319 model, filled symbols).

casts (Toth et al, 1998; Zhu et al, 2001). One should remain cautious in interpreting these contradicting results. The overperformance of the T62 control forecast might point to a special behaviour of the NCEP ensemble with respect to QPF. This possible weakness might be linked to essential differences in the ECMWF ensemble: (i) the method of generation of the perturbations; (ii) the lower resolution of the NCEP model; (iii) the limited population of the NCEP ensemble. The impact of (ii) and (iii) is examined in the next subsection.

3.1.3 ECMWF ensemble vs. higher resolution model single run

Figure 11 shows the relative value of the ECMWF EPS (ECEPS) and the higher resolution (T319) ECMWF model forecast (ECH). Most differences are not significant at the 90% level. For the 5 mm/12 h threshold (Fig. 11a), ECH is significantly better than ECEPS for the C/L corresponding to the maximum value, while the comparison is the opposite

for a lower C/L. For the 20 mm/12 h threshold (Fig. 11b), ECEPS is significantly better than ECH for C/L lower than 10^{-4} (at 99% level). No difference is significant for the 50 mm/12 h (Fig. 11c).

3.2 Ensemble vs. ensemble

In this subsection, ensembles running at 12:00 UTC have been compared in terms of the relative value computed from spatial multi-event contingency tables.

3.2.1 ECMWF ensemble vs. NCEP ensemble

Figure 12 shows the relative value of the ECMWF EPS (ECEPS) and the NCEP EPS running at 12:00 UTC (NCEPS12). ECEPS performs better than NCEPS12, with a high level of significance (often more than 99%) for the 5, 20 and 50 mm/12 h thresholds. This result is not surprising, given the difference in resolution of the models, on the one hand, and the difference in the number of members, on the other hand.

In order to evaluate the relative influence of these two factors, a smaller ensemble based on the ECMWF EPS has been constructed, consisting of the control forecast and the first 4 perturbed members. The value of this smaller ensemble (ECEPS5) has been compared to the value of the NCEP EPS (NCEPS12). The results (not shown) are very similar to those obtained with the fully populated ECMWF ensemble, indicating that the impact of the ensemble population might be much lower than the impact of the model resolution (or other characteristics of the ensembles, e.g. the method of generation of the perturbations).

3.2.2 ECMWF ensemble vs. smaller ensemble

In order to further investigate the impact of the ensemble population, smaller ensembles based on the ECMWF EPS control forecast and the 10/20/32 first perturbed members (ECEPS11, ECEPS21, ECEPS33) have been compared to the fully populated EPS (ECEPS). Note, however, that the first perturbed members' initial conditions are still obtained from all 25 singular vectors (SVs), since the perturbations are combinations of SVs (Molteni et al., 1996). This comparison thus favors smaller ensembles and only addresses the question of the number of integrations that are needed.

Value curves shown in Fig. 13 indicate that ECEPS21 performs as well as ECEPS, except for the 20 mm/12 h threshold and smaller C/L (order of 10^{-4}). Differences between ECEPS11 and ECEPS (not shown) are significant at the 90% level for a limited range of rather small C/L ratios for the 5 mm/12 h and 20 mm/12 h thresholds. No significant differences have been found for the 50 mm/12 h threshold and no significant differences have been found between ECEPS33 and ECEPS for any threshold (not shown).

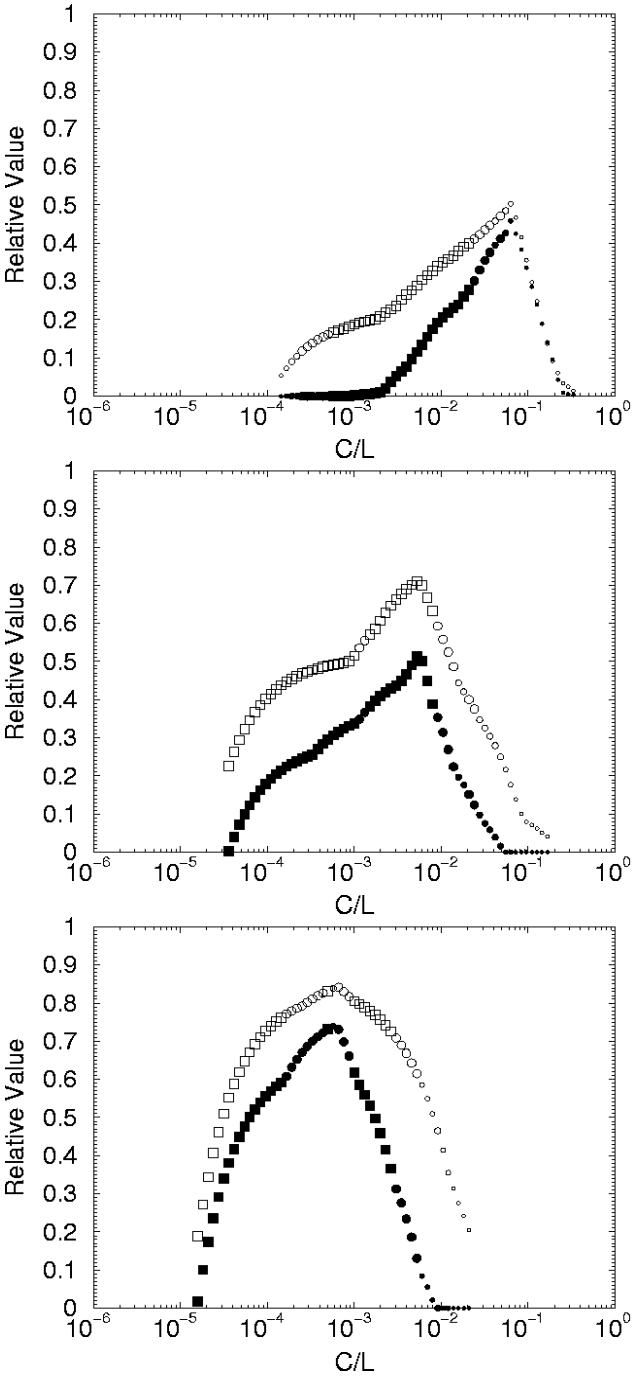

Fig. 12. Same as Fig. 9, but for ECEPS (ECMWF EPS, blank symbols) and NCEPS12 (NCEP EPS running at 12:00 UTC, filled symbols).

3.2.3 ECMWF ensemble vs. "poorman ensemble"

Since the impact of the ensemble population is limited, it might be interesting to consider the small "poorman ensemble" consisting of the ECMWF T159 control forecast and the NCEP T126 control forecast (ECNC). Since they simultaneously take into account the uncertainties of the initial conditions and model formulation, "poorman ensembles" have

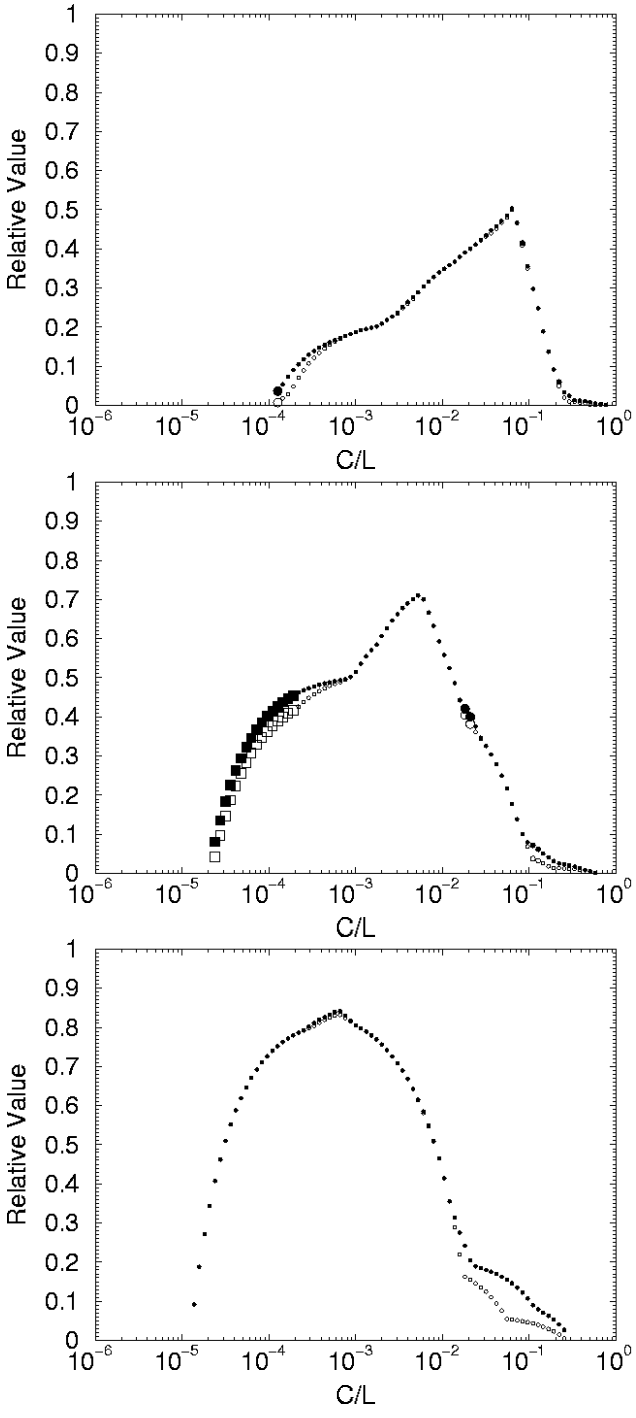

Fig. 13. Same as Fig. 9, but for ECEPS21 (ECMWF EPS control + 20 perturbed members, blank symbols) and ECEPS (ECMWF EPS, filled symbols).

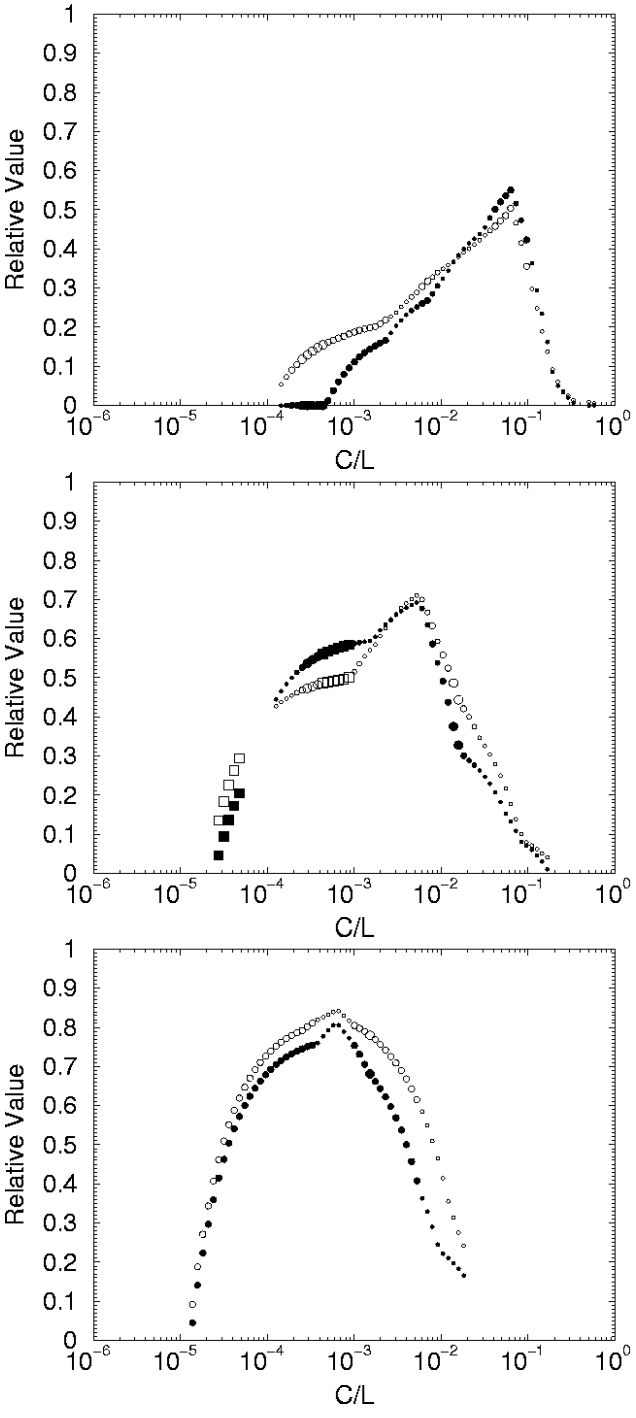

Fig. 14. Same as Fig. 9, but for ECEPS (ECMWF EPS, blank symbols) and ECNC ("poorman ensemble" consisting of the ECMWF T159 control forecast and the NCEP T126 control forecast, filled symbols).

proven to perform as operational EPSs for certain aspects of the prediction of upper level atmospheric parameters in the early medium-range (Ziehman, 2000; Atger, 1999).

Figure 14 shows the relative value of the ECMWF EPS (ECEPS) and the "poorman ensemble" (ECNC). Most differences are not significant at the 90% level. Significant differences indicate a superiority of ECEPS for smaller C/L ratios for 5 mm/12 h (Fig. 14a) and 20 mm/12 h (Fig. 14b). For the 20 mm/12 h threshold, ECNC is significantly better than ECEPS (at the 90% level) for a limited range of C/L ratios of the order of 10^{-3} (Fig. 14b). For the 50 mm/12 h threshold, ECEPS seems better than ECNC for any C/L, but no difference is statistically significant (Fig. 14c).

4 Discussion

4.1 Morning precipitations vs. afternoon precipitations

The results presented in the previous section only concern afternoon precipitations. Although the number of observations of intense precipitation is lower, the performance of ensembles and single model runs has been investigated with respect to the 00:00–12:00 UTC precipitations. Although most results are similar to those obtained for afternoon forecasts, some results differ with respect to comparisons between single models and EPSs.

Figure 15 shows the relative value of the ECMWF EPS (ECEPS) and the higher resolution (T319) ECMWF model forecast (ECH) for morning precipitations (+84 h to +96 h forecasts). In contrast with Fig. 11 (afternoon precipitations), ECEPS performs significantly better than ECH (at the 90% level) for a wide range of C/L ratios, especially for higher thresholds (20 mm/12 h and 50 mm/12 h). Similarly, the NCEP EPS running at 12:00 UTC (NCEPS12) performs as well as or even slightly better than NC12 (90% level significance for lower C/L ratios and for 50 mm/12 h), despite the higher resolution of the latter (T126) (not shown). As a consequence, ECEPS performs significantly better than the 2-member "poorman ensemble", consisting of the control forecasts of ECMWF and NCEP ensembles (ECNC), especially for lower C/L ratios and higher thresholds (not shown).

The difference in performance between morning and afternoon precipitations might come from the fact that operational ensembles are more likely to overperform a single run at longer lead-times (for a model running at 12:00 UTC: +96 h for morning precipitations, +84 h for afternoon precipitations). However, this hypothesis is not supported by the performance of the NCEP ensemble running at 00:00 UTC (NCEPS0), mentioned in Sect. 3.1.2, for +96 h forecasts of afternoon precipitations. For longer lead-times, the relative value of probabilistic forecasts of intense precipitation decreases and becomes close to zero for most C/L ratios, so that the performance of the ensembles over single runs cannot be demonstrated with confidence. This is the main reason for the choice of the 72–96 h range for the results presented in the previous section, although operational ensembles have been designed for use from day 3 to day 10 (ECMWF), and beyond (NCEP).

One intrinsic difference between the morning and afternoon precipitations is the frequency of convective activity. In France, during the winter season (as considered in this study), convective precipitations occur most frequently during the afternoon. Important precipitations occurring before 12:00 UTC are likely to originate from large-scale systems, while they are often a consequence of small-scale, convective activity when they occur after 12:00 UTC. The pdf of the morning precipitations is, therefore, primarily associated with large-scale dynamics uncertainty, while the intensity and location of the afternoon precipitations is more often largely unpredictable with operational global models. Operational ensembles have been designed for estimating variations in large-scale dynamics predictability. Probabilistic forecasts based on an EPS are thus likely to perform better for morning (large-scale) precipitations than afternoon (small-scale) precipitations. On the other hand, probabilistic forecasts based only on a single run take into account local uncertainties related to the location and intensity of the precipitation. Therefore, it is not surprising that they are more efficient with respect to the afternoon (small-scale) precipita-

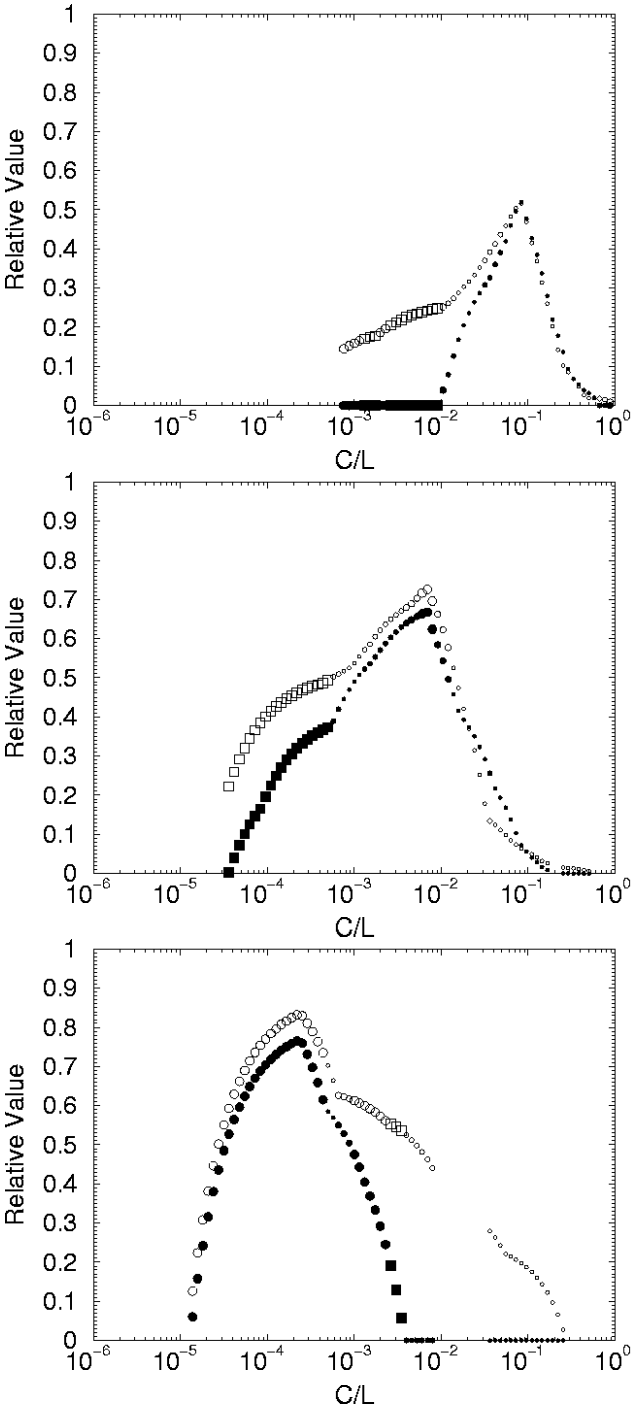

Fig. 15. Same as Fig. 11, but for morning precipitation (+84 h to +96 h).

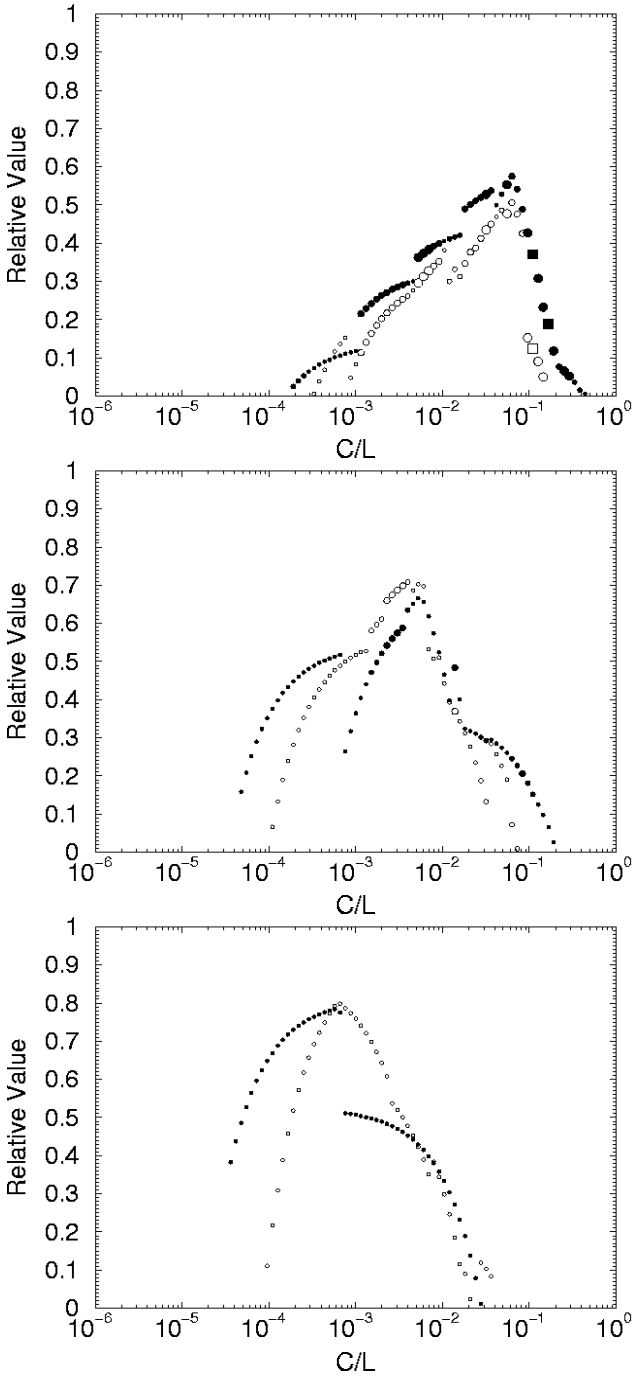

Fig. 16. Same as Fig. 11, but after halving the data in order to separate the sub-sample used for the derivation of the optimal forecast category for a given C/L, and the sub-sample used for verification.

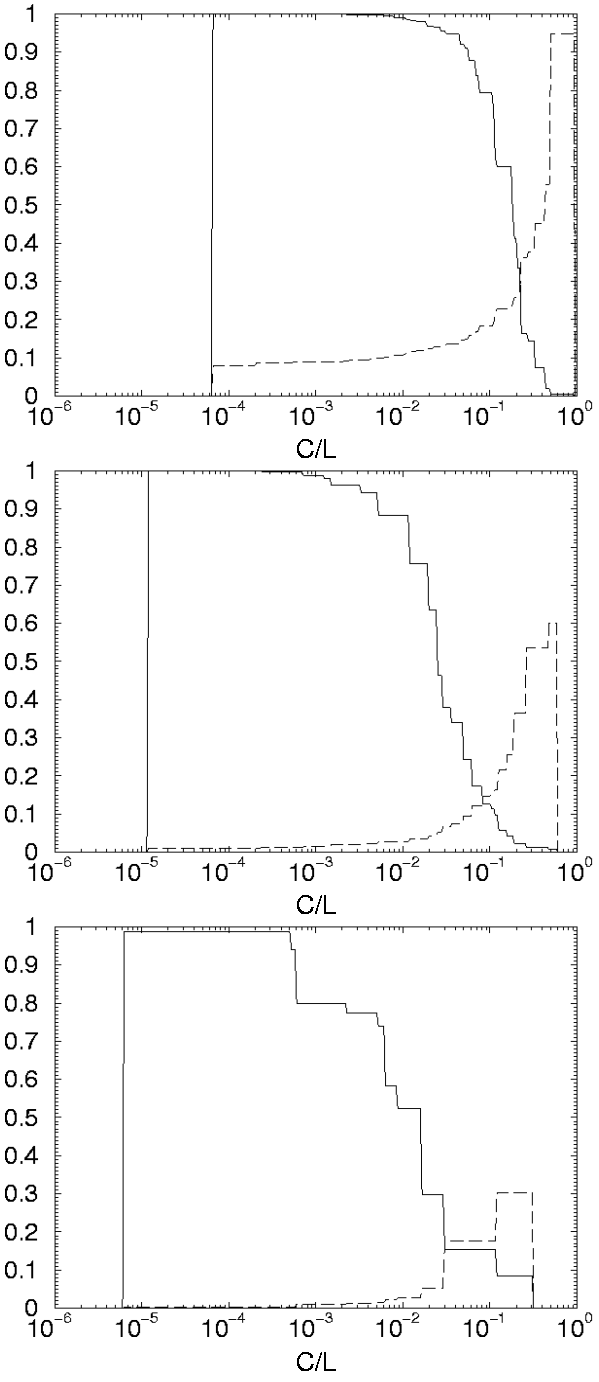

Fig. 17. Hit Rate (HR, solid line) and Correct Alarm Rate (CAR, proportion of forecasts of the event that are justified, dashed line) computed from spatial multi-event contingency tables based on the ECMWF EPS (ECEPS). For every C/L ratio, the HR and CAR are those obtained from the forecast category leading to the maximum value. Afternoon precipitation only ($+72$ h to $+84$ h). **(a)** 5 mm/12 h observed threshold. **(b)** 20 mm/12 h observed threshold. **(c)** 50 mm/12 h observed threshold.

tions, while uncertainties associated with large-scale systems are poorly estimated.

4.2 Computation of value from an independent sample

The results presented in the previous section have been obtained through an evaluation of probabilistic forecasts based on spatial multi-event contingency tables. As performed in most studies (e.g. Richardson 2000), the computation of the False Alarm Rate and Hit Rate, leading to the relative economic value of the forecast, has been performed under a

strong assumption: the forecast user is supposed to protect when the category forecast by the system leads to the maximum value that can be expected. For example, in the case of an EPS forecast evaluated from a standard contingency table, the user protects at least a certain number of ensemble members every time, with a forecast more than the considered threshold. The user can only know this certain number from the past. Proper evaluation should thus require an independent, representative sample, from which the optimal forecast category would be derived for every C/L. In practice, the available sample is generally small, so that it is used for both the evaluation and the determination of the optimal categories. One may qualify the result of this computation as a potential value (Richardson, 2000), i.e. the maximum value that is attainable in real conditions.

In order to evaluate the difference between the potential value and the actual value, the data have been randomly halved into 2 sub-samples. The first sub-sample is used for the derivation of the forecast category that leads to a maximum value for every C/L. The relative value is computed from the second sub-sample for the forecast category determined from the first sub-sample. Figure 16 shows the relative value of the ECMWF EPS (ECEPS) and the higher resolution (T319) ECMWF model forecast (ECH) when this procedure is followed. Most differences are not significant at the 90% level, except for the 5 mm/12 h threshold. The curves look rather noisy, with discontinuities reflecting the variability of the maximum value attained for a given C/L. This indicates that both sub-samples are too small (45 days each) to obtain conclusive results with respect to the actual value of the probabilistic forecasts of intense precipitations. When differences are significant, the actual value of the single forecast is higher. EPS forecasts probably suffer more than single forecasts, given the fact that the sample is small when compared to the number of forecast categories: $5 \times 20 \times 51$=5100 categories in the case of the EPS, but 5×20=100 categories for the single forecast. In other words, ensemble forecasts would have the potential to overperform single forecasts for the prediction of intense precipitations, but a larger sample would be needed in order to identify from past statistics the forecast category that leads, in effect, to the maximum value.

4.3 The meaning of very small C/L ratios

One of the aims of this study is to evaluate the usefulness of operational forecasting systems for the prediction of intense precipitations. The results presented in the previous sections show that the maximum potential value increases with the precipitation threshold. Impressively high levels of the potential value (80% of that attainable with a perfect forecast) are obtained for the 50 mm/12 h threshold. However, the range of users who benefit from intense precipitation forecasts is limited to very small C/L ratios. By construction, the maximum potential value is obtained for C/L ratios that are close to the frequency of occurrence of the event. The prediction of rare events thus is a benefit primarily to lower C/L users: those users facing a decision making situation that oblige them to protect (or take action, in a more general sense) as soon as the risk of potential damage exists, even if it is almost nil. This may be the case, for instance, of a mountaineer who requires a 99% probability of quiet weather, before deciding to go for a 3 day expedition in a remote area during winter.

When forecasting extreme events, the problem may just come from false alarms. The high maximum value obtained by operational forecasting systems for the prediction of rare precipitation events reflects the fact that high hit rates can be achieved provided that the frequency of false alarms is high. In practical terms, only well informed, professional users can tolerate a high frequency of false alarms. This is equivalent to saying that the C/L ratio of these users is small. By contrast, the majority of the users, especially among the public, hardly tolerate false alarms. The C/L ratio of these users is large, actually much larger than the climatological frequency of the considered event.

Figure 17 shows the hit rate and the Correct Alarm Rate (CAR, proportion of forecasts of the event that are justified) computed from spatial multi-event contingency tables based on the ECMWF EPS. For every C/L ratio, the HR and CAR correspond to the forecast category leading to the maximum value. Assuming that most users would require at least 30–50% of correct alarms, they could expect a 10–30% hit rate for 5 mm/12 h, but virtually no detection for 20 and 50 mm/12 h. This indicates that intense precipitation forecasts based on operational forecasting systems, although exhibiting high levels of maximum potential value, are only useful for a restricted category of users.

5 Summary

The performance of single models and ensemble prediction systems has been investigated with respect to quantitative precipitation forecasts, with a special emphasis on intense precipitation. Evaluation has been based on the relative economic value of the forecasts, computed from spatial multi-event contingency tables. A probabilistic forecast from an EPS can thus be compared to a probabilistic forecast based on a single model run. The latter is designed to represent the probabilistic judgment of an operational forecaster, from which any probabilistic or deterministic statement originates. The statistical significance of the comparisons between various forecasting systems has been estimated through a resampling procedure.

The relative value increases with the precipitation threshold. Impressively high levels of relative value (60–80% of that attainable with a perfect forecast) are reached for the 20 mm/12 h and 50 mm/12 h thresholds. These numbers reflect high hit rates that are obtained at the expense of a dramatic increase in the frequency of false alarms. The ECMWF EPS performs better overall than a single forecast based on the same model, even when the resolution of the ensemble is lower (T_L159 vs. T_L319). The difference is important for morning precipitation, especially for higher precipitation

thresholds and lower C/L ratios. On the other hand, the performance of the ECMWF EPS and single forecasts is rather similar for afternoon precipitation, probably due to more frequent convective events.

The NCEP EPS performs as well as a single forecast based on the same model for morning precipitation, even when the resolution of the ensemble is lower (T62 vs. T126). Higher resolution single forecasts perform better with respect to afternoon precipitation. The ECMWF EPS performs better than the NCEP EPS running at 12:00 UTC. This is still true when the population of the ECMWF ensemble is reduced to 5 members. More generally, the impact of reducing the number of members of the ECMWF EPS is rather small. No differences have been found between 51 and 33 members. The 11 member ensemble still performs as well as the fully populated ensemble for a limited range of C/L ratios. A "poorman ensemble", consisting of the control forecasts of the ECMWF and the NCEP EPSs, performs as well as the ECMWF EPS for afternoon precipitation. The ECMWF EPS is still significantly better with respect to morning precipitation.

Acknowledgement. C. Ziehmann, D. Richardson, B. Houdant and O. Talagrand, as well as an anonymous reviewer, provided helpful comments on an earlier version of this manuscript. Part of this work has been supported by ECMWF as part of an expert visit by the author in summer 2000.

References

Ångstrøm, A. K.: Probability and practical weather forecasting (in Swedish), Centraltryckeriet, Teknologföreningens Förlag, 11pp, 1919.

Atger, F.: The skill of ensemble prediction systems, Mon. Wea. Rev., 127, 9, 1941–1953, 1999.

Buizza, R. and Palmer, T. N.: Impact of ensemble size on ensemble prediction, Mon. Wea. Rev., 126, 2503–2518, 1998.

Buizza, R., Hollingsworth, A., Lalaurette, F., and Ghelli, A.: Probabilistic predictions of precipitations using the ECMWF Ensemble Prediction System, Wea. Forecasting, 14, 168–189, 1999.

Buizza, R., Petroliagis, T., Palmer, T. N., Barkmeijer, J., Hamrud, M., Hollingsworth, A., Simmons, A., and Wedi, N.: Impact of model resolution and ensemble size on the performance of an ensemble prediction system, Quart. J. Roy. Meteor. Soc., 124, 1935–1960, 1997.

Carter, G. M., Dallavalle, J. P., and Glahn, H. R.: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system, Wea. Forecasting, 4, 401–412, 1989.

Chessa, P. A. and Lalaurette, F.: Verification of the ECMWF Ensemble Prediction System forecasts: a study on synoptic patterns, to Wea. Forecasting, 16, 611–619, 2001.

Courtier, P., Freyder, C., Geleyn, J. F., Rabier, F., and Rochas, M.: The Arpege project at Meteo-France, in: Proceddings on Numerical Methods in Atmospheric Models, 2, (Eds) ECMWF, 192–231, 1991.

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

Katz, R. W. and Murphy, A. H.: Economic value of weather and climate forecasts, (Eds) Cambridge University Press, 1997.

Liljas, E. and Murphy, A. H.: Anders Ångstrøm and his early papers on probability forecasting and the use/value of weather forecasts, Bull. Amer. Meteor. Soc., 75, 1227–1236, 1994.

Mason, I.: A model for assessment of weather forecasts, Aust. Met. Mag., 30, 291–303, 1982.

Mason, S. J. and Graham, N. E.: Conditional probabilities, Relative Operating Characteristics, and Relative Operating Levels, Wea. Forecasting, 14, 713–725, 1999.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF ensemble prediction system: methodology and validation, Quart. J. Roy. Meteor. Soc., 122, 73–119, 1996.

Murphy, A. H.: The value of climatological, categorical and probabilistic forecasts in the Cost/Loss situation, Mon. Wea. Rev., 105, 803–816, 1977.

Murphy, A. H.: Probabilistic weather forecasting, in: Probability, Statistics, and decision making in the Atmospheric Sciences, (Eds) Murphy, A. H. and Katz, R. W., Westview Press, 337–377, 1985.

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

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

Palmer, T. N., Molteni, F., Mureau, R., Buizza, R., Chapelet, P., and Tribbia, J.: Ensemble prediction, in: Proceedings of Validation of Models over Europe (Eds) ECMWF, 1, 21–66, 1993.

Richardson, D. S.: Skill and economic value of the ECMWF Ensemble Prediction System, Quart. J. Roy. Meteor. Soc., 126, 649–668, 2000.

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

Stanski, H. R., Wilson, L. J., and Burrows, W. R.: Survey of common verification methods in meteorology, WMO/WWW Tech. Rep. 8, 1989.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, in: Proceedings of Predictability, (Eds) ECMWF, 1–25, 1997.

Toth, Z. and Kalnay, E.: Ensemble forecasting at NCEP and the breeding method, Mon. Wea. Rev., 125, 3297–3319, 1997.

Toth, Z., Zhu, Y., Marchok, T., Tracton, S., and Kalnay, E.: Verification of the NCEP global ensemble forecasts, Preprints, 12th Conf. on Numerical Weather Prediction, Phoenix, Arizona, Amer. Meteor. Soc., 286–289, 1998.

Tracton, M. S. and Kalnay, E.: Operational ensemble prediction at the National Meteorological Center: practical aspects, Wea. Forecasting, 8, 379–398, 1993.

Ziehmann, C.: Comparison of a single-model EPS with a multimodel ensemble consisting of a few operational models, Tellus, 52, 280–299, 2000.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., and Mylne, K.: On the economic value of ensemble based weather forecasts, Bull. Amer. Meteorol. Soc., submitted, 2001.