Accuracy and Potential Economic Value of Categorical and Probabilistic Forecasts of Discrete Events

ROBERTO BUIZZA

European Centre for Medium-Range Weather Forecasts, Reading, Berkshire, United Kingdom

(Manuscript received 11 April 2000, in final form 31 October 2000)

ABSTRACT

The accuracy and the potential economic value of categorical and probabilistic forecasts of discrete events are discussed. Accuracy is assessed applying known measures of forecast accuracy, and the potential economic value is measured by a weighted difference between the system probability of detection and the probability of false detection, with weights function of the cost–loss ratio and the observed ratio and the observed relative frequency of the event.

Results obtained using synthetic forecast and observed fields document the sensitivity of accuracy measures and of the potential forecast economic value to imposed random and systematic errors. It is shown that forecast skill cannot be defined per se but depends on the measure used to assess it: forecasts judged to be skillful according to one measure can show no skill according to another measures. More generally, it is concluded that the design of a forecasting system should follow the definition of its purposes, and should be such that the ensemble system maximizes its performance as assessed by the accuracy measures that best quantify the achievement of its purposes.

Results also indicate that independently from the model error (random or systematic) ensemble-based probabilistic forecasts exhibit higher potential economic values than categorical forecasts.

1. Introduction

Numerical weather predictions are often expressed in the form of categorical or probabilistic forecasts of discrete predictands (a discrete predictand is an observable variable that takes one and only one of a finite set of possible values). A typical example is the prediction of more than 10 mm of precipitation or of temperature below freezing. The prediction of discrete events can be based either on categorical forecasts (''the event will/will not occur'') or on probabilistic forecasts (''there is a 30% probability of occurrence''). Generally speaking, categorical forecasts are defined as forecasts consisting of a flat statement that one and only one of a possible set of events will occur (Wilks 1995). Probabilistic forecasts are forecasts given in terms of a probability that a considered event would happen.

Numerical weather forecasts are often used by decision makers to decide whether or not to take an action to protect against a possible loss. Typically, a decision maker would spend an amount C, if an event were predicted, to protect against a loss L (with $L > C$). The potential economic value of a forecasting system can be assessed by using skill measured defined by coupling contingency tables and cost–loss decision models (Katz et al. 1982; Murphy 1985; Wilks and Hamill 1995).

In this work, the *potential economic value* of a forecast is defined, as in Richardson (2000), by a function of the probability of detection and the probability of false detection of the system. Since this measure is defined for both categorical and probabilistic forecasts, it can be used to compare the potential economic value of a single forecast and of an ensemble forecasting system.

Forecasts and observed values are compared over set of $N_G = 1581$ contiguous points, which can be considered as representing Europe (latitude $30°\text{N} \leq \phi \leq 60°\text{N}$, longitude $20°\text{W} \leq \lambda \leq 30°\text{E}$) on a 1° regular grid. Synthetic observed and forecast patterns are defined as a combination of two-dimensional Gaussian functions and random fields. The sensitivity of different measures of forecast accuracy to imposed errors is investigated, and the potential economic benefit of an ensemble forecasting system instead of a single deterministic forecast is assessed.

The following aspects of the forecast accuracy and potential economic value of categorical and probabilistic forecasts are investigated in particular.

- The sensitivity of categorical deterministic and probabilistic forecasts to imposed model errors.
- The relative potential economic value of single deterministic and probabilistic forecasts.

Corresponding author address: Dr. Roberto Buizza, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.
E-mail: buizza@ecmwf.int

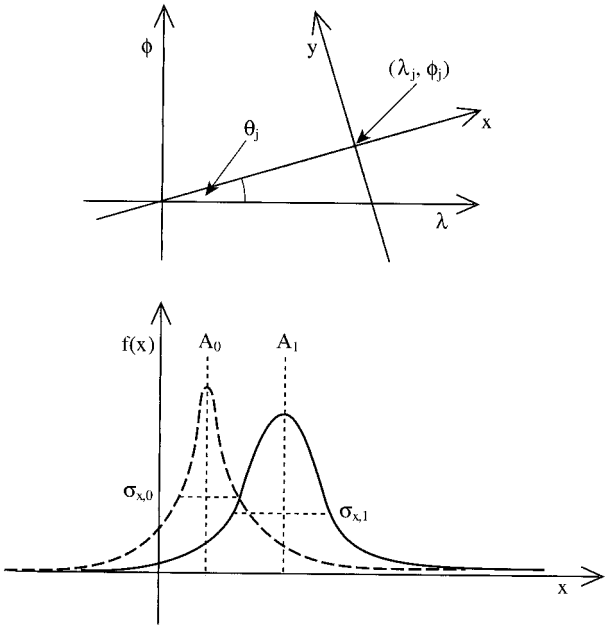

FIG. 1. Schematic of the definition of the forecast and observed fields (Gaussian functions).

- The sensitivity of the potential economic value of probabilistic predictions to ensemble size and model accuracy.

After this introduction, section 2 describes how the synthetic observed and forecast patterns are defined. The accuracy measures for categorical and probabilistic scores are introduced in section 3 and are applied to a single-case study in section 4. The average sensitivity (90-case average) of the accuracy measures and the potential economic value is investigated in section 5. The potential economic values of single deterministic and probabilistic forecasts are also compared in section 6. Conclusions are drawn in section 7.

2. Definition of synthetic forecast and observed fields

Denote by $g_j(\lambda, \phi)$ a Gaussian function of the longitude λ and the latitude ϕ:

$$g_j(\lambda, \phi) = A_j \exp\left\{-\left[\frac{\cos\theta_j \cdot (\lambda - \lambda_j) + \sin\theta_j \cdot (\phi - \phi_j)}{\sqrt{2} \cdot \sigma_{x,j}}\right]^2\right\} \times \exp\left\{-\left[\frac{-\sin\theta_j \cdot (\lambda - \lambda_j) + \cos\theta_j \cdot (\phi - \phi_j)}{\sqrt{2} \cdot \sigma_{y,j}}\right]^2\right\}, \quad (1)$$

defined by the maximum amplitude A_j, the rotation angle θ_j, the coordinate of the maximum value λ_j and ϕ_j, and the standard deviations $\sigma_{x,j}$ and $\sigma_{y,j}$ along the unrotated original axis. Figure 1a shows schematically how parameters λ_j, ϕ_j, and θ_j define the position and the orientation of the Gaussian function $g_j(\lambda, \phi)$, and Fig. 1b shows how parameters A_j, $\sigma_{x,j}$, and $\sigma_{y,j}$ define its shape.

TABLE 1. Contingency table for dichotomous event.

		Observed		Marginal distribution of the forecast
		Yes	No	
Forecast	Yes	a/n	b/n	$(a + b)/n$
	No	c/n	d/n	$(c + d)/n$
Marginal distribution of the obs		$(a + c)/n$	$(b + d)/n$	$n = a + b + c + d$

Equation (1) can be used to define an observed pattern $f_0(\lambda, \phi)$ and an ensemble of forecasts $f_j(\lambda, \phi)$, with each function defined by a different set of parameters A_j, θ_j, λ_j, ϕ_j, $\sigma_{x,j}$, and $\sigma_{y,j}$:

$$f_j(\lambda, \phi) = c_j(\lambda, \phi) \cdot g_j(\lambda, \phi), \quad (2)$$

where $c_j(\lambda, \phi)$ is either a constant function, specifically $c_j(\lambda, \phi) = 1$, or it is defined by a set of random numbers $c_j(\lambda, \phi)$ uniformly sampled in the interval [0, 2] [this choice guarantees a rescaling of the field $g_j(\lambda, \phi)$ by up to 100%]. Hereafter, $j = 0$ will identify the *verification pattern*, $j = 1, 51$ will identify the *ensemble of forecast*, and $j = 1$ will identify the *control forecast* (note that the control forecast is not different from any randomly chosen ensemble member).

The 51 ensemble forecasts can be used to construct categorical products like, for example, the forecast given by the ensemble mean,

$$\overline{f}(\lambda, \phi) = \frac{1}{51} \sum_{j=1}^{51} f_j(\lambda, \phi), \quad (3)$$

and probability forecasts, defined by the probability of occurrence of the event

$$p_f(\lambda, \phi) = \frac{n_f(\lambda, \phi)}{51}, \quad (4)$$

where $n_f(\lambda, \phi)$ is the number of forecasts predicting the event at the grid point with coordinates (λ, ϕ). The spread $s(\lambda, \phi)$ of the ensemble is defined as the ensemble second-order moment or standard deviation, that is, as the root-mean-square distance

$$s(\lambda, \phi) = \left\{\frac{1}{51} \sum_{j=1}^{51} [f_j(\lambda, \phi) - \overline{f}(\lambda, \phi)]^2\right\}^{1/2}. \quad (5)$$

Alternatively, the ensemble spread could be defined as the average distance of a randomly chosen pair of forecasts, or as the average distance of the control forecast ($j = 1$) from the other members.

3. Verification scores

A brief introduction of the scores used in this work to measure the skill of categorical and probabilistic fore-

casts of discrete dichotomous predictand (i.e., predictand allowed to be in only two possible states, yes or no) is reported hereafter. The reader is referred to Wilks (1995) for more details.

a. Scores for categorical forecasts

Categorical verification of dichotomous (i.e., binary) events can be based on the 2×2 contingency table that displays the number of all possible combinations of forecast and observed events (Table 1). The performance of each of the N_{ens} forecasts defined in Eq. (2) is measured by a contingency table constructed by adding an entry to the 2×2 contingency for each of the $g = 1, N_G$ grid points inside the area under investigation. A perfectly accurate forecast would clearly exhibit $b = c = 0$ in its corresponding contingency table.

1) SOME MEASURES OF FORECAST ACCURACY FOR BINARY EVENTS (HIT RATE, THREAT SCORE, PROBABILITY OF DETECTION, AND PROBABILITY OF FALSE DETECTION)

Accuracy measures summarize the correspondence between individual forecasts and occurred events. Denote by $p_{\text{cli}} = (a + c)/n$ the observed frequency of the event under consideration. Four of the most commonly used measures of *accuracy* (i.e., of the average correspondence between individual forecasts and the events they predict; see Wilks 1995) are the hit rate, the threat score, the probability of detection, and the probability of false detection (Doswell et al. 1990).

The hit rate

$$\text{HR} = \frac{a + d}{n} \tag{6}$$

measures the proportion of correct forecasts. Note that it gives the same credit to correct yes and no forecasts.

The threat score

$$\text{TS} = \frac{a}{a + b + c} \tag{7}$$

is the number of correct yes forecasts divided by number of cases when the event was forecast and/or observed. It can be viewed as the hit rate after removals of correct no forecasts.

The probability of detection

$$\text{POD} = \frac{a}{a + c} = \frac{a}{n}\frac{1}{p_{\text{cli}}} \tag{8}$$

is the probability that the occurred event would be forecast.

The probability of false detection

$$\text{PFD} = \frac{b}{b + d} = \frac{b}{n}\frac{1}{(1 - p_{\text{cli}})} \tag{9}$$

is the proportion of nonoccurrences incorrectly forecast (Doswell et al. 1990).

2) BIAS

The bias measures the correspondence between the average forecast and the average observed value of the predictand. The bias

$$B = \frac{a + b}{a + c} \tag{10}$$

is the ratio of the number of yes forecasts to the number of observed events. By definition, the bias does not depend on the individual correspondence between forecast and observed values, and so it is not a measure of accuracy.

3) THE KUIPERS SKILL SCORE

Forecast skill refers to the accuracy of a forecast with respect to a reference forecast. Standard choices for the reference are climatological average values, persistent forecasts, or random forecasts. For any chosen measure of accuracy A, the skill score SS_A of a forecast with accuracy A_f with respect to the reference forecast with accuracy A_{ref} is given by

$$\text{SS}_A = \frac{A_f - A_{\text{ref}}}{A_{\text{perf}} - A_{\text{ref}}} 100\%, \tag{11}$$

where A_{perf} is the accuracy of a perfect forecast.

One of the most commonly used skill scores used to summarize square contingency tables is the Kuipers skill score (Hanssen and Kuipers 1965; Murphy 1996), which is based on the HR as accuracy measure. The reference accuracy is the HR of an unbiased random forecast, that is, with $p_{\text{ref},D}(fc = \text{yes}) = p(\text{ob} = \text{yes})$ and, by definition, $a_{\text{ref},D} = (a + b)^2/n^2$ and $d_{\text{ref},D} = (b + d)^2/n^2$. Thus, the Kuipers skill score is

$$\text{KSS} = \frac{(ad - bc)}{(a + c)(b + d)}. \tag{12}$$

The Kuipers skill score can be written in terms of the probability of detection POD and the probability of false detection PFD:

$$\text{KSS} = \text{POD} - \text{PFD}. \tag{13}$$

Note that both random and constant forecasts receive the same zero score, and the contribution to KSS of correct no (yes) forecasts increases as the event is more (less) likely. It should be mentioned that the Kuipers skill score approaches the probability of detection POD when correct forecasts of no events dominate the contingency table, and therefore it is vulnerable to hedging in rare event forecasting (Doswell et al. 1990). Despite this, and the fact that Doswell et al. (1990) argue that the Heidke skill score (also called S statistics) should be preferred, the Kuipers skill score will be used in this

paper because of its relationship with the potential economic value (see section 3c).

b. Scores for probabilistic forecasts

The most common scalar measure of the accuracy of a probabilistic forecast of a dichotomous event is the Brier score,

$$\mathrm{BS} = \frac{1}{N_g}\sum_{g=1}^{N_G}(p_{f,g} - o_g)^2, \tag{14}$$

which is the mean squared error of the forecast probability $p_{f,g} = p_f(\lambda, \phi)$, where the index $g = 1, N_G$ denotes the forecast–event pairs of all considered grid points. The observed probability function is defined to be $o_g = 1$ if the event occurs and $o_g = 0$ if the event does not occur. The Brier score can be computed as the sum of three terms related to reliability, resolution, and uncertainty:

$$\mathrm{BS} = \mathrm{BS}_{\mathrm{rel}} - \mathrm{BS}_{\mathrm{res}} + \mathrm{BS}_{\mathrm{unc}}. \tag{15}$$

The Brier skill score BSS is defined as

$$\mathrm{BSS} = \frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{ref}}}{0 - \mathrm{BS}_{\mathrm{ref}}} = \frac{\mathrm{BS}_{\mathrm{res}} + \mathrm{BS}_{\mathrm{rel}}}{\mathrm{BS}_{\mathrm{unc}}}. \tag{16}$$

One of the properties of the BS is that it can be considered a strictly proper score, in the sense that the BS cannot be improved by forecasting something other than one's true beliefs about future weather events (i.e., hedging; see Wilks 1995).

Another measure of probabilistic forecast accuracy is the area under a relative operating characteristic (ROC) curve defined in signal detection theory (Mason 1982). Consider the forecast probability distribution $p_f(\lambda, \phi)$ defined in Eq. (4), stratified according to observation into 51 categories as in Table 2. For any given probability threshold j, the entries of this table can be summed to produce the four entries of a 2×2 contingency table:

$$a_j = \sum_{k=j+1}^{51} x_k \qquad b_j = \sum_{k=j+1}^{51} y_k \qquad c_j = \sum_{k=1}^{j} x_k$$

$$d_j = \sum_{k=1}^{j} y_k. \tag{17}$$

From each of the jth contingency tables, the probability of detection POD_j and the probability of false detection PFD_j can be computed. The 51 pairs (PFD_j, POD_j) can be plotted one against the other on a graph. The result is a smooth curve called the ROC curve.

As in Richardson (2000), the ROC area (ROCA) can be converted into a skill score

$$\mathrm{ROCAS} = \frac{\mathrm{ROCA} - \mathrm{ROCA}_{\mathrm{cli}}}{\mathrm{ROCA}_{\mathrm{per}} - \mathrm{ROCA}_{\mathrm{cli}}} = 2\mathrm{ROCA} - 1, \tag{18}$$

since $\mathrm{ROCA}_{\mathrm{per}} = 1$ for a perfect forecast, and $\mathrm{ROCA}_{\mathrm{cli}} = 0.5$ ($\mathrm{ROCA} = 0.5$ for a climatological forecast, since $\mathrm{PFD} = \mathrm{POD} = 0.5$).

TABLE 2. Table of occurrences/nonoccurrences for ROC area definition.

Category index	Probability range	Observed: Yes	Observed: No
1	$0 \le p_f < 1/N_{\mathrm{ens}}$	x_1	y_1
...	...	...	...
j	$(j-1)/N_{\mathrm{ens}} \le p_f < j/N_{\mathrm{ens}}$	x_j	y_j
...	...	...	...
N_{ens}	$(N_{\mathrm{ens}} - 1)/N_{\mathrm{ens}} \le p_f \le 1$	$x_{N\mathrm{ens}}$	$y_{N\mathrm{ens}}$

The ranked probability skill score (RPSS; Epstein 1969; Murphy 1971) is another measure of probabilistic forecast accuracy. Given a set of events related to the same variable and characterized by a different amount (e.g., consider precipitation events characterized by different rainfall amounts), the RPSS can be considered an extension of the Brier skill score to multicategory events (Wilks 1995). Let J_{ev} be the number of (ranked) forecast events, $p^j_{f,g}$ the forecast probability for the jth event, and o^j_g the observed probability function (with $o^k_g = 1$ if the kth event is observed, and $o^j_g = 0$ for $j \neq k$), where $g = 1, N_g$ denotes the gth grid point. The grid point ranked probability score RPS_g is computed from the squared error of the cumulative forecast and observed probabilities:

$$\mathrm{RPS}_g = \sum_{m=1}^{J}\left[\left(\sum_{j=1}^{m} p^j_{f,g}\right) - \left(\sum_{j=1}^{m} o^j_g\right)\right]^2. \tag{19}$$

The area-average ranked probability score RPS is defined as

$$\mathrm{RPS} = \frac{1}{N_g}\sum_{g=1}^{N_g}\mathrm{RPS}_g. \tag{20}$$

The RPSS is defined with respect to a forecast based on the sample

$$\mathrm{RPSS} = \frac{\mathrm{RPS}_{\mathrm{cli}} - \mathrm{RPS}}{\mathrm{RPS}_{\mathrm{cli}}}. \tag{21}$$

Another measure of ensemble performance is the percentage of observed values lying outside the ensemble forecast range, also called the percentage of outliers (POUTL; Strauss and Lanzinger 1995). As a reference value, $\mathrm{POUTL}_{\mathrm{ref}} = 2/52$ for an ensemble system with 51 members that randomly samples the forecast probability density function.

TABLE 3. Cost–loss decision model.

		Observed: Yes	Observed: No
Take action	Yes	C	C
	No	L	0

TABLE 4. Coefficients used to define the parameters of the observed and the forecast values [Eq. (1)].

Parameter	Observed value	Fc mean value	Fc range
Maximum A_j	100	100	40
Longitude λ_j	10	10	3
Latitude ϕ_j	45	45	3
Rotation θ_j	30	30	20
Standard deviation $\sigma_{x,j}$	1.5	1.5	0.4
Standard deviation $\sigma_{y,j}$	1	1	0.2

c. Potential forecast economic value

As in Richardson (2000), consider decision makers interested in protecting from the occurrence of the event under consideration. Suppose that if they take an action incurring a cost C they can avoid a loss L (with $L > C$). Table 3 summarizes this simple cost–loss model.

If the decision makers know only the observed frequency p_{cli} and assume that the sample observed frequency is equal to the long-term climatology, their optimal strategy would be to always protect if $C < p_{\text{cli}}L$, and their expected mean expense per unit loss would be

$$\text{ME}_{\text{cli}} = \min\left(\frac{C}{L}, p_{\text{cli}}\right). \tag{22}$$

If the decision makers have access to a perfect forecast, than their mean expense per unit loss would be

$$\text{ME}_{\text{perf}} = p_{\text{cli}}\frac{C}{L}, \tag{23}$$

since they would incur a cost C/L (per unit loss) only in the p_{cli} occasions when they protected themselves against the loss (always avoided).

Suppose now that the decision makers have access to a single deterministic forecast, which is taken at face value (i.e., without adjusting the forecast for estimated model errors), whose performance is summarized by Table 1, and suppose that the observed frequency was estimated from the sample, that is, $p_{\text{cli}} = (a + c)/n$. Then, from Tables 1 and 3 it follows that their mean expense (per unit loss) would be

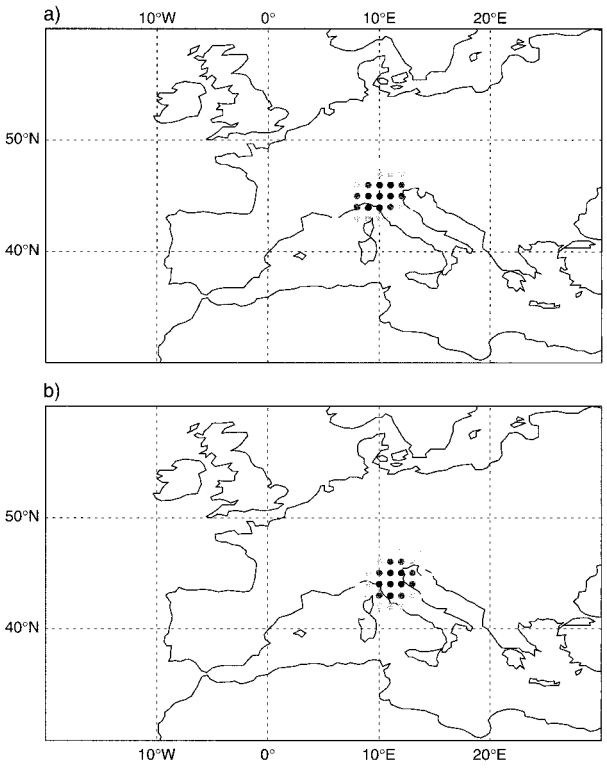

FIG. 2. (a) Observed and (b) control forecast fields for one case at each grid point.

$$\text{ME} = \frac{(a+b)}{n}\frac{C}{L} + \frac{c}{n}$$
$$= \text{PFD}(1 - p_{\text{cli}})\frac{C}{L} - \text{POD}p_{\text{cli}}\left(1 - \frac{C}{L}\right) + p_{\text{cli}}. \tag{24}$$

The potential economic value FV of the forecast is defined as the reduction of the mean expense with respect to the reduction of the expense that could be achieved by a perfect forecast:

$$\text{FV} = \frac{\text{ME} - \text{ME}_{\text{cli}}}{\text{ME}_{\text{perf}} - \text{ME}_{\text{cli}}}. \tag{25}$$

Applying Eqs. (22), (23), and (24) and the definition of the Kuipers skill score, the potential forecast economic value can written as

$$\text{FV} = \text{KSS} - \frac{(1 - \text{POD})\left[p_{\text{cli}} - \min\left(\frac{C}{L}, p_{\text{cli}}\right)\right] + \text{PFD}\left[\frac{C}{L} - \min\left(\frac{C}{L}, p_{\text{cli}}\right)\right]}{\min\left(\frac{C}{L}, p_{\text{cli}}\right) - p_{\text{cli}}\frac{C}{L}}. \tag{26}$$

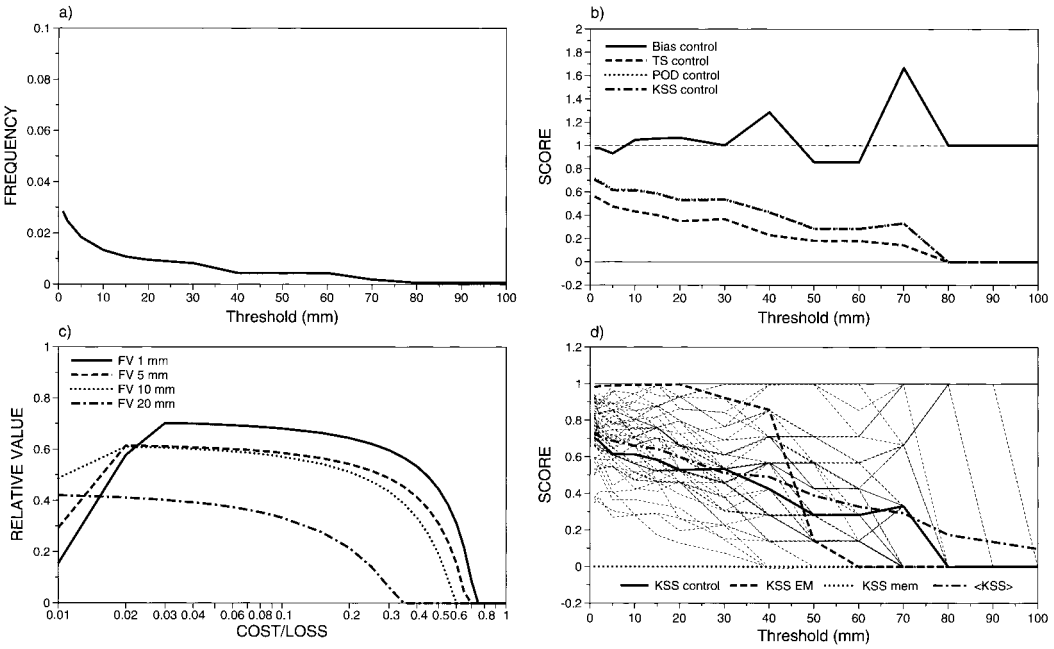

FIG. 3. (a) Observed frequency (i.e., sample climatology) p_{cli}. (b) Bias (solid line), TS (dashed line), POD (dotted line), and KSS (chain-dashed line) for the control forecast (note that the POD and the KSS curves are superimposed). (c) Forecast value for 1 (solid line), 5 (dashed line), 10 (dotted line), and 40 mm (chain-dashed line) for the control forecast. (d) KSS for the control forecast (solid line), the ensemble mean forecast (dashed line), the 50 ensemble members (dotted line), and 51-member average KSS (chain-dashed line).

From Eq. (26) it is easy to see that

- if $C/L > p_{\text{cli}}$ then

$$\text{FV} = \text{KSS} - \text{PFD}\frac{\left(\frac{C}{L} - p_{\text{cli}}\right)}{p_{\text{cli}}\left(1 - \frac{C}{L}\right)};$$

- if $C/L < p_{\text{cli}}$ then

$$\text{FV} = \text{KSS} - (1 - \text{POD})\frac{\left(p_{\text{cli}} - \frac{C}{L}\right)}{\frac{C}{L}(1 - p_{\text{cli}})};$$

- when $C/L = p_{\text{cli}}$ the forecast value is maximum, FV = KSS.

Equation (26) and subsequent equations highlight the fact that given the observed frequency of the event p_{cli} and the user cost–loss ratio C/L, the forecast value depends only on the probability of false detection and the probability of detection of the system [since by definition KSS = POD − PFD; see Eq. (13)]. Furthermore, these equations show that the potential economic value is a weighed difference between the probability of detection and of false detection of the system, with weights a function of the event observed frequency (climatology) and the user cost loss ratio. Equation (26) also indicates that the Kuipers skill score of the system can be considered as the maximum forecast value that can be obtained from the system.

Finally, suppose that the decision makers have access to the whole ensemble of forecast f_j, which is again taken at face value (i.e., without adjusting the forecast for estimated model errors). As for the computation of the ensemble ROC area, consider the 51 pairs (PFD_j, POD_j) computed from the 2×2 contingency tables associated with the 51 probability thresholds. Applying Eq. (26), the 51 forecast values FV_j associated with the jth probability threshold can be computed. Given the observed frequency of the event p_{cli}, for each cost–loss ratio the forecast value of the ensemble is defined as

$$\text{FV}_{\text{ens}}\left(\frac{C}{L}\right) = \max_{j=1,51} \text{FV}_j\left(\frac{C}{L}\right). \tag{27}$$

Equation (27) shows that each user can optimize the ensemble forecast value for each cost–loss ratio C/L by choosing the probability threshold that has maximum value at that specific ratio.

4. Scores of categorical and probabilistic forecasts: A single-case study

Denote by $f_0(\lambda, \phi)$ an observed pattern defined by ($A_0 = 100$, $\lambda_0 = 10$, $\phi_0 = 45$, $\phi_0 = 30$, $\sigma_{x,0} = 1.5$, $\sigma_{y,0} = 1$), and denote by $f_j(\lambda, \phi)$ an ensemble of fore-

casts defined by randomly sampling the parameters (A_j, λ_j, ϕ_j, θ_j, $\sigma_{x,j}$, $\sigma_{y,j}$) in the intervals defined in Table 4. These patterns can be considered to represent an observed and a forecast precipitation field. Since the parameters used to define the observed pattern are included in the range of the forecast parameters, the ensemble is reliable. Furthermore, by construction each ensemble member is, on average, equally skillful. It is worthwhile to remind the reader that the results discussed in this section refer to one case only and thus may present some peculiar features.

Figure 2 shows the observed pattern $f_0(\lambda, \phi)$ and the control forecast $f_1(\lambda, \phi)$ defined by ($A_1 = 102$, $\lambda_1 = 12$, $\phi_1 = 44$, $\theta_1 = 50$, $\sigma_{x,1} = 1.4$, $\sigma_{y,1} = 1.1$). Figure 3a shows the observed frequency p_{cli} of events characterised by a different value, that is, by a different amount of precipitation. Note that all frequencies are smaller than 0.03, indicating quite rare events for any threshold.

The control forecast is practically unbiased, with a higher threat score for low precipitation amounts and with a positive skill score KSS up to 70 mm (Fig. 3b). The fact that the KSS and the POD curves overlap indicates a very low PFD [see Eq. (13)] and is a consequence of the fact that correct forecasts of no event dominates the contingency table (Doswell et al. 1990). As mentioned in section 3, the forecast value depends on the cost–loss ratio C/L and on the event observed frequency (Fig. 3c), and its upper bound is given by the Kuipers skill score KSS. This is evident by the comparison of Figs. 3b and 3c. Figure 3d shows the Kuipers skill score KSS for the whole ensemble compounded of the control forecast and the 50 other forecasts. Figure 3d shows that the Kuipers skill score of the ensemble mean forecast (dashed line) is more skillful than the control forecast (solid line) for precipitation amounts up to 50 mm.

The 51 ensemble forecasts have been used to generate probability forecasts of different precipitation amounts. Figure 4a shows that the probability forecasts have a positive Brier skill score for amounts up to 70 mm. The BSS of the ensemble forecast decreases with the threshold amount in a way similar to the KSS and the TS of the control forecast. By contrast, the ROC area skill score is always equal to 1 and does not decrease with the precipitation threshold, due to very low probabilities of false detection. Figure 4a also shows that the RPSS is always positive (dotted line; constant since it is an integrated measures computed considering all the precipitation amounts). Figure 4b shows the potential forecast value FV_j for all the probability thresholds $j = 1$, N_{ens} for the prediction of more than 10 mm of precipitation. Note that different probability thresholds for each cost–loss ratio achieve the maximum potential forecast value.

By definition any score depends on the area onto which they are computed. This sensitivity to the area definition is shown by the comparison of the forecast

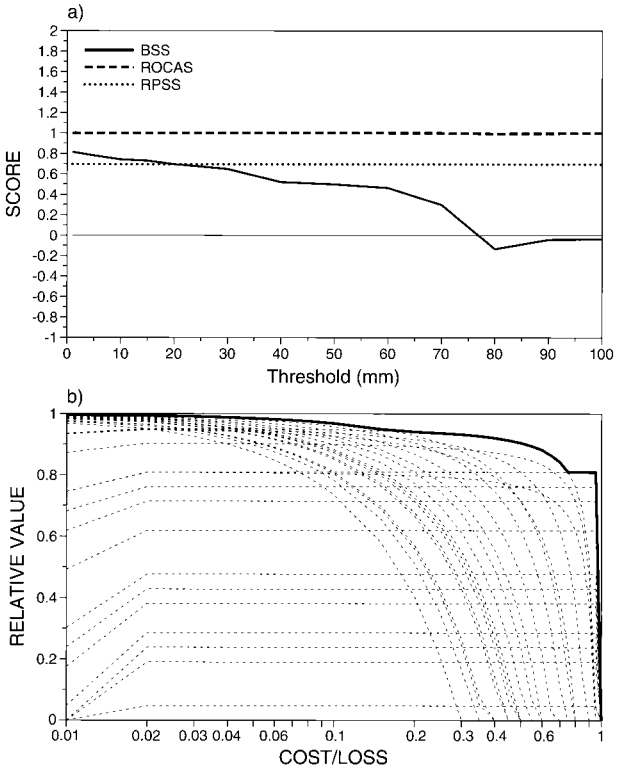

FIG. 4. (a) BSS (solid line), ROCAS (dashed line), and RPSS (dotted line) for the ensemble probabilistic prediction of different amounts of precipitation. (b) Forecast value FV_j of each of the $j = 1$, N_{ens} probability thresholds for the probabilistic prediction of more than 10 mm.

scores computed over a European subregion centered around the observed pattern (40° ≤ latitude ≤ 50°N, 0° ≤ longitude ≤ 20°E, 231 grid points) with the scores computed over the whole European area (1581 grid points). The reduction of the verification area induces an increase in the observed relative frequency p_{cli} by about a factor of 7 (see Fig. 5a for the small area and Fig. 3a for Europe). Since the contingency tables for the two areas differ mainly in the number of correctly forecast nonoccurrences, the only verification scores that are sensitivity to the area reduction are the scores that depend on the ''d'' entry of the 2×2 contingency table. Indeed, the bias, the threat score, and the probability of detection do not change (see Figs. 3b and 5b). By contrast, the probability of false detection changes, and this affects the Kuipers skill score (see Figs. 3b and 5b) and the forecast value (see Figs. 3c and 5c). Similarly, the Kuipers skill score of all the ensemble forecasts (not shown) and the forecast value of probabilistic forecast of precipitation amounts are affected (see Figs. 4b and 5d).

5. Scores sensitivity: A single-case study

The sensitivity of the forecast scores to a priori imposed errors due to amplitude under/overestimation, to

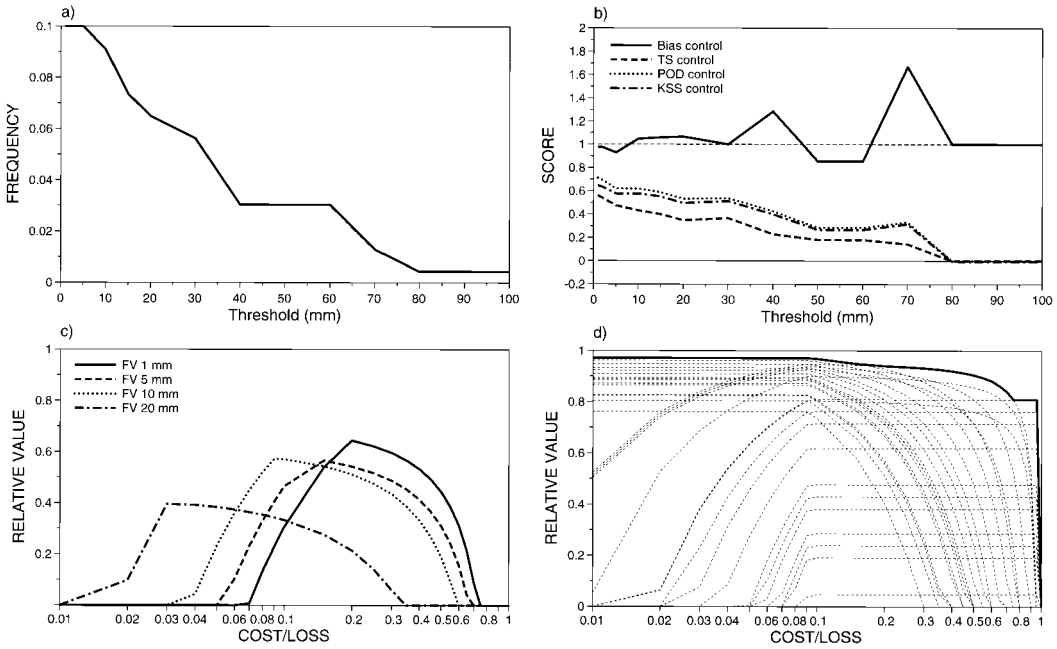

FIG. 5. Forecast scores over a small region. (a) Observed frequency (i.e., sample climatology) p_{cli}; (b) bias (solid line), TS (dashed line), POD (dotted line), and KSS (chain-dashed line) for the control forecast; (c) forecast value for 1 (solid line), 5 (dashed line), 10 (dotted line), and 40 mm (chain-dashed line) for the control forecast; (d) forecast value FV_j of each of the $j = 1, N_{ens}$ probability thresholds for the probabilistic prediction of more than 10 mm.

position errors and to ''shape errors'' (i.e., errors in the definition of the width of the Gaussian function) is investigated hereafter. All results discussed in this section refer to the prediction of 10 mm.

Figure 6a shows the sensitivity of the control scores to amplitude errors. Results have been obtained by setting all parameters apart for A_1 as in Table 4 (as in section 4) and with $0 \leq A_1 \leq 200$ (i.e., with $0 \leq A_1/A_0 \leq 200$ since $A_0 = 100$ for the observed field; see Table 4). It is interesting to note that for this precipitation amount (10 mm) the threat score is always positive for any $A_1 > 0$. The threat score relative, for example, to the prediction of 40 mm is zero for $A_1/A_0 < 0.6$ (not shown.)

Consider now an ensemble of 50 forecasts defined by $(A_j, \lambda_j, \phi_j, \theta_j, \sigma_{x,j}, \sigma_{y,j})$ generated as follows. For each A_1, the ensemble of 50 parameters A_j $(j = 2, 51)$ is sampled in the interval $\max(0, A_1 - 40) < A_j < (A_1 + 40)$, while all the other parameters are sampled according to Table 4 (e.g., by setting $7 \leq \lambda_j \leq 13$). Figure 6b shows the TS, the POD, and the KSS of the ensemble mean forecast. As for the control forecasts, the ensemble mean scores are sensitive to the amplitude errors. Compared to the control forecast (Fig. 6a), the ensemble mean has higher TS and KSS than the control for any $A_1/A_0 > 0.2$. Figure 6c shows the BSS, the ROCAS, and the RPSS for the probabilistic prediction. The BSS and the RPSS show a sensitivity to A_1/A_0 similar to the sensitive shown by the control and the ensemble mean forecasts, with smaller BSS and RPSS for $A_1/A_0 > 0.6$. By contrast, the ROCAS is 1 for any ratio $A_1/A_0 > 0$ and the POUTL is zero for any A_1/A_0 (not shown).

Figure 7a shows the sensitivity of the control scores to a position error in predicting the longitude of the precipitation maximum (parameter λ). All parameters but λ_1 were defined according to Table 4, while $0° \leq \lambda_1 \leq 20°$E [i.e., with errors $-10 \leq (\lambda_1 - \lambda_0) \leq 10$ since $\lambda_0 = 10$E for the observed field; see Table 4]. Results indicate that the (unbiased) control forecast is skillful only for a certain range of position errors, with this range depending on the precipitation amount and of the width (σ parameters) of the observed and forecast fields. As before, for each λ_1 an ensemble of 50 forecasts have been generated, each with a different parameter λ_j sampled in the interval $(\lambda_1 - 3) \leq \lambda_j \leq (\lambda_1 + 3)$. Figure 7b shows the sensitivity to the position error of the ensemble mean scores and Fig. 7c the sensitivity of the scores of the probabilistic predictions. Compared to the control forecast (Fig. 7a), the ensemble mean (Fig. 7b) has higher TS and KSS, but it has also a larger bias (see Hamill 1999 for a discussion on the impact of model bias on verification scores such as the equitable threat score). The scores of the probability forecasts show a strong sensitivity to the forecast position error (Fig. 7c). Large position errors induce low values of ROCAS, negative BSS and RPSS, and high POUTL. The fact that forecasts with a positive ROCAS have negative Brier skill score and negative RPSS confirm the fact that different measures of forecast quality give different results. Generally speaking, the comparison between

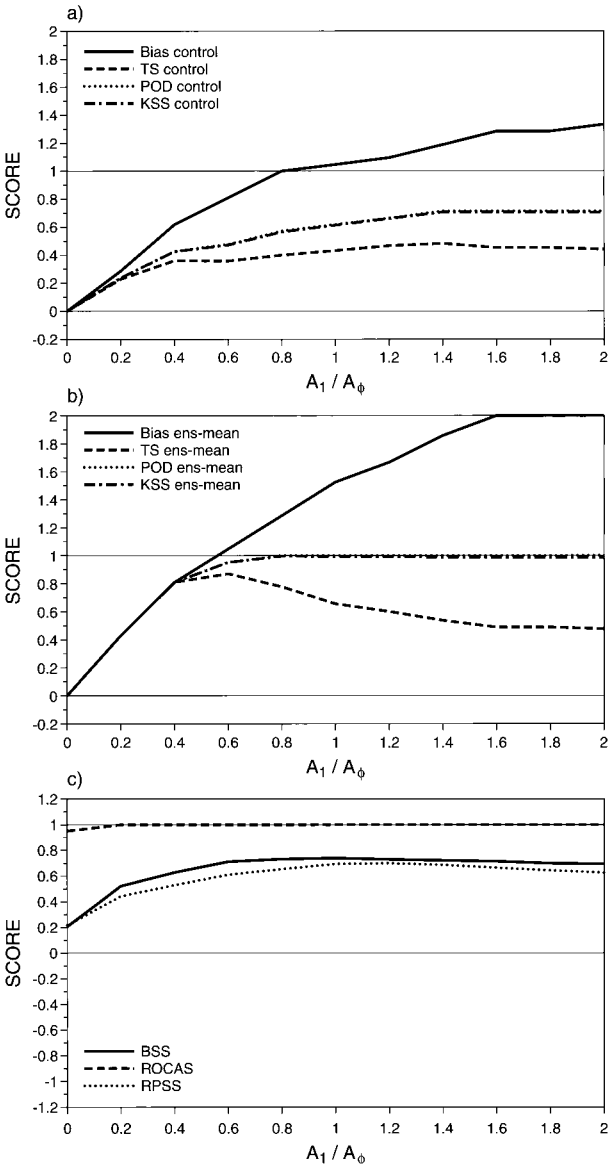

FIG. 6. Sensitivity of forecast scores of the control to amplitude errors. (a) Bias (solid line), TS (dashed line), POD (dotted line), and KSS (chain-dashed line) for the control forecast. (b) As in (a) but for the ensemble mean. (c) BSS (solid line), ROC area skill score (dashed line), and RPSS (dotted line) for the prediction of 10 mm. Ordinate: scores. Abscissa: imposed error in the control amplitude A_1/A_0.

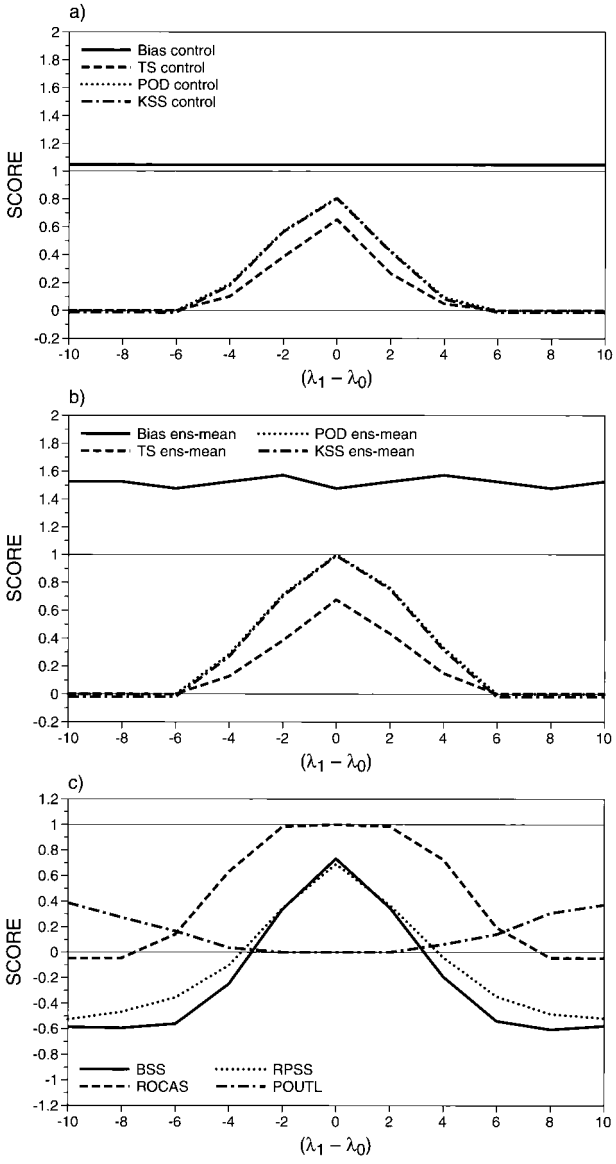

FIG. 7. Sensitivity of forecast scores of the control to longitude position errors. (a) Bias (solid line), TS (dashed line), POD (dotted line), and KSS (chain dashed line) for the control forecast. (b) As in (a) but for the ensemble mean. (c) BSS (solid line), ROC area skill score (dashed line), RPSS (dotted line), and POUTL (chain-dashed line) for the prediction of 10 mm. Ordinate: scores. Abscissa: imposed position error of the control forecast ($\lambda_1 - \lambda_0$).

Figs. 6 and 7 suggest that a position error has a more severe effect on the forecast scores than an amplitude error. Similar results would have been obtained by varying the latitude of the maximum value (not shown).

Figure 8 shows the sensitivity to errors in the prediction of the Gaussian function standard deviation along the (unrotated) x axis, that is, $\sigma_{1,x}$. All parameters but $\sigma_{1,x}$ were defined according to Table 4, while $0.75 \leq \sigma_{1,x} \leq 3.75°$ (i.e., with $0.5 \leq \sigma_{1,x}/\sigma_{0,x} \leq 2.5$ since $\sigma_{0,x} = 1.5°$ for the observed field; see Table 4). Results show that the ensemble mean scores are higher than the scores of the control forecast for any $\sigma_{1,x}$ (Figs. 8a,b) and that of the ensemble scores of the probabilistic precipitation prediction deteriorate for too small or too large $\sigma_{1,x}$ (Fig. 8c), that is, if each forecast field is too narrow or too wide.

In Fig. 8c, the fact that POUTL > 0 for large $\sigma_{1,x}$ is related to the way all forecast parameters are set (see Table 4). These results document the sensitivity of the different measures of skill to under/overestimation, to position errors, and to errors in the prediction of the

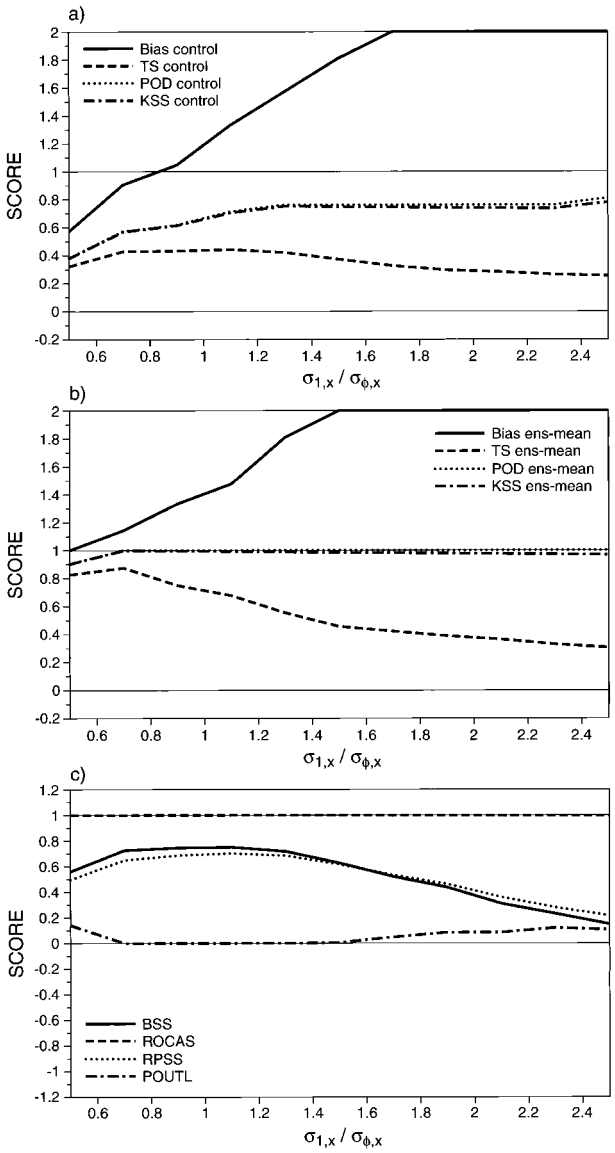

FIG. 8. Sensitivity of forecast scores of the control to errors in σ_x (area coverage). (a) Bias (solid line), TS (dashed line), POD (dotted line), and KSS (chain-dashed line) for the control forecast. (b) As in (a) but for the ensemble mean. (c) BSS (solid line), ROC area skill score (dashed line), RPSS (dotted line), and POUTL (chain-dashed line) for the prediction of 10 mm. Ordinate: scores. Abscissa: imposed error on the control forecast standard deviation $\sigma_{x,1}/\sigma_{x,0}$.

correct shape (σ parameter) of the observed pattern. They also show that forecasts judged to be skillful according to one measure of forecast skill could be judged not to have any skill according to others. Furthermore, they indicate that a reliable ensemble can be used to construct a single deterministic forecast (i.e., the ensemble mean) that is more skillful than each single ensemble member. This aspect is further analyzed in the following section, where categorical and probabilistic forecasts are compared considering a larger dataset.

6. Potential economic value of categorical and probabilistic forecasts: 90 cases average results

The impact of random and systematic model errors on the average performance (90 cases) is investigated hereafter.

a. Impact of random errors

A high quality forecasting system (no systematic model error and with equally skillful ensemble members) can be simulated by randomly sampling each of the parameters that define the observed and the forecast fields from the same interval.

Figure 9a shows the average scores of each single deterministic forecast given by one ensemble member (e.g., the control). The forecast has no bias, as expected by construction, and has a positive Kuipers skill score for all precipitation thresholds. Figure 9b shows the reliability, resolution, and uncertainty terms of the BS for the ensemble probabilistic predictions: it indicates that the ensemble is reliable (almost null Brier score reliability term). Figure 9c shows that the ensemble has a ROCAS above zero for all thresholds, a positive BSS for thresholds up to 70 mm, a positive RPSS, and an almost null percentage of outliers. Figures 9d–f show the potential forecast value of the control and the ensemble mean, and the potential forecast value of the ensemble probabilistic predictions for three different thresholds, 1, 10, and 40 mm. The comparison of the potential forecast value curves confirms the indications of section 5 that the ensemble mean is more valuable than the control forecast, and that the ensemble probabilistic prediction has a higher value than any deterministic forecast. It is worth pointing out that all potential economic value curves peak for very small cost–loss ratios since all events are very rare, even for 1 mm.

In other words, decision makers interested in predicting a binary event ''rainfall greater than an amount x'' would have a higher return if they make decisions (protect/nonprotect) according to the ensemble probabilistic forecast than to any single deterministic forecast. This result is summarized in the potential forecast value chess boards shown in Fig. 10. For any cost–loss ratio and any threshold amount, the potential forecast value is higher if actions are taken according to the ensemble probabilistic forecast rather than the control or the ensemble mean forecast (not shown).

Similar results have been obtained when a second source of random error [i.e., for forecasts defined applying Eq. (2) with c_j (λ, ϕ) set to be a random number uniformly sampled in the interval $0 \leq c_j$ (λ, ϕ) ≤ 1] has been introduced in the generation of the ensemble forecasts (not shown).

b. Impact of systematic over/underestimation

Consider now two ensemble systems characterized by random errors (as for the ensemble discussed in section

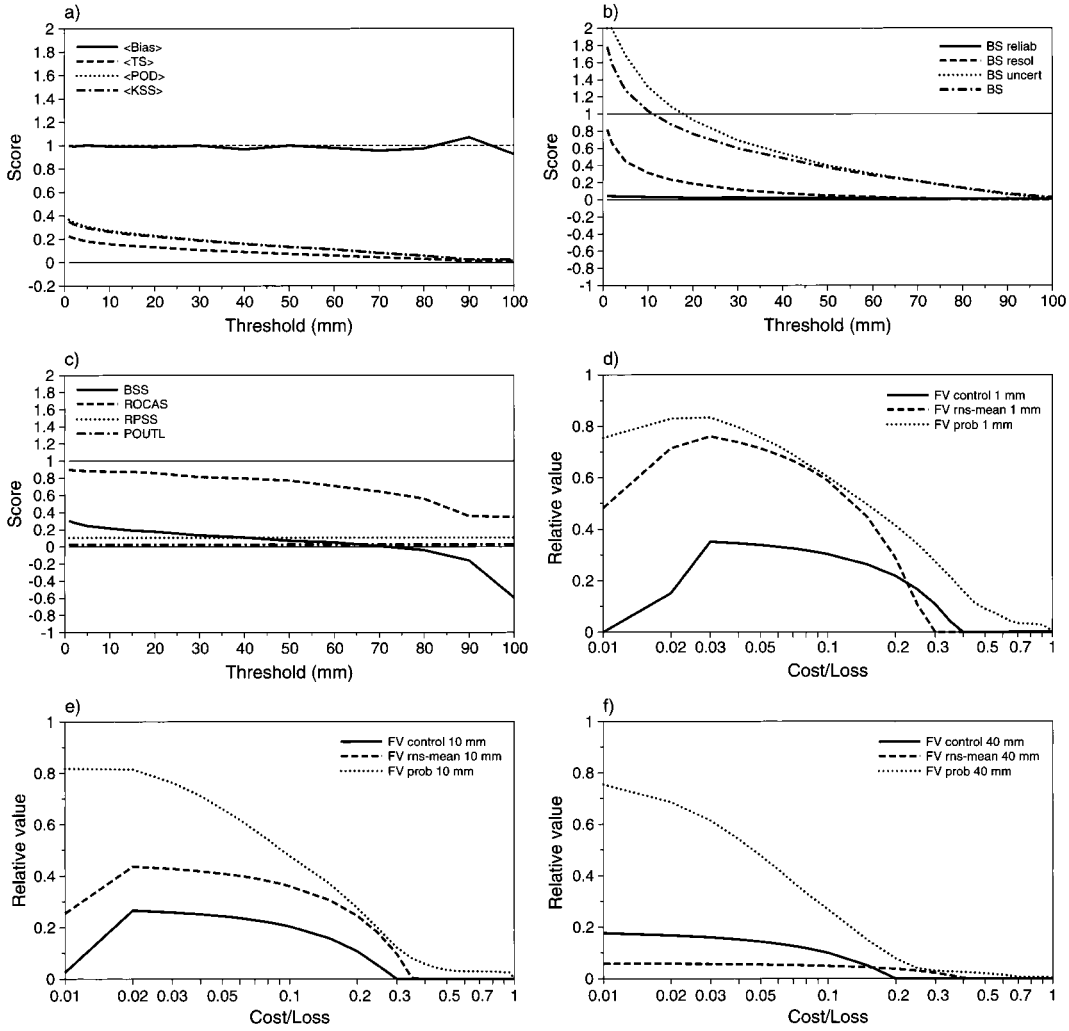

FIG. 9. Average performance (90 cases) of an ensemble system affected only by random errors. (a) Bias (solid line), TS (dashed line), POD (dotted line), and KSS (chain-dashed line) for the control forecast. (b) Brier score reliability (solid line, multiplied by 100), resolution (dashed line, multiplied by 100), and uncertainty (dotted line, multiplied by 100) terms, and full Brier score (chain dashed line, multiplied by 100) for probabilistic predictions. (c) BSS (solid line), ROC area skill score (dashed line), RPSS (bold dotted line), and POUTL (chain-dashed line) for probabilistic predictions. (d) Forecast value for the probabilistic prediction of 1 mm for the control (solid line), the ensemble mean (dashed line), and the ensemble probabilistic prediction (dotted line). (e) As in (d) but for 10 mm. (f) As in (d) but for 40 mm.

6a) and by a 40% under- or overestimation of the precipitation maxima. These results have been obtained by defining the forecasts as in section 6a but by multiplying the coefficient A_0 of the observation field by a factor of 0.6 or by a factor of 1.4. These examples can be thought to describe the performance of ensemble systems based on a model characterized by a poor simulation of moist processes that induces either a rainfall overestimation or underestimation. Figure 11 shows the performance of these two systems.

On average, under or overestimation errors induces a bias on each single deterministic forecast (see Figs. 11a,b and 9a), especially for thresholds larger than 60 mm, and it has a sizeable impact on the threat scores and the Kuipers skill scores for thresholds larger than 60 mm. The bias curve for the ensemble characterized by overestimation (Fig. 11b) drops to zero for precipitation values larger than 60 mm because, by construction, 60 mm is the maximum observed value. Considering the ensemble probabilistic predictions, both over- or underestimation have a sizeable impact on the ROCAS for thresholds larger than 60 mm, while overestimation has a larger impact than underestimation on the BSS for thresholds. The impact of over- or underestimation on the potential forecast value reflects the impact on the ROCAS, that is, small for small thresholds (say, up to 40 mm; see Figs. 11e,f for the 10-mm threshold) and large for larger values (not shown). It is worthwhile to point out that the potential forecast value curves for the system affected by underestimation are to the

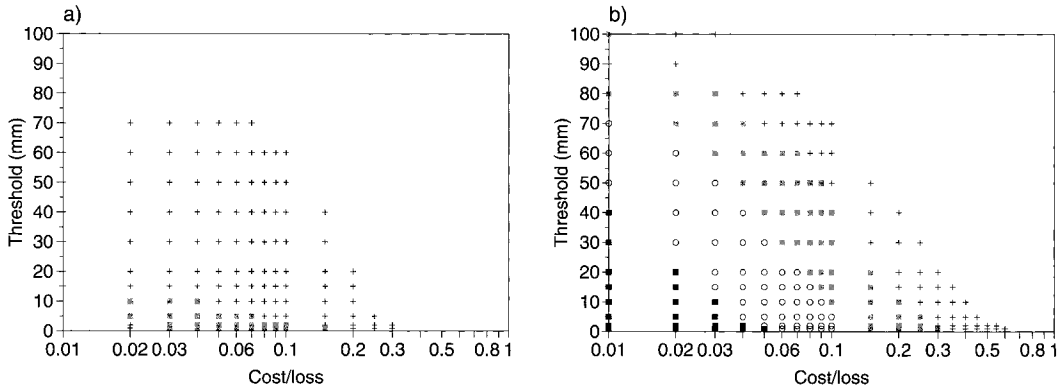

FIG. 10. Forecast value chess board (90 cases) of an ensemble system affected only by random errors for (a) the control forecast and (b) the ensemble probabilistic prediction. Symbols for forecast values are crosses for $0.05 < \text{FV} \leq 0.25$, squares for $0.25 < \text{FV} \leq 0.50$, full gray square for $0.50 < \text{FV} \leq 0.75$, and full circles for $0.75 < \text{FV}$.

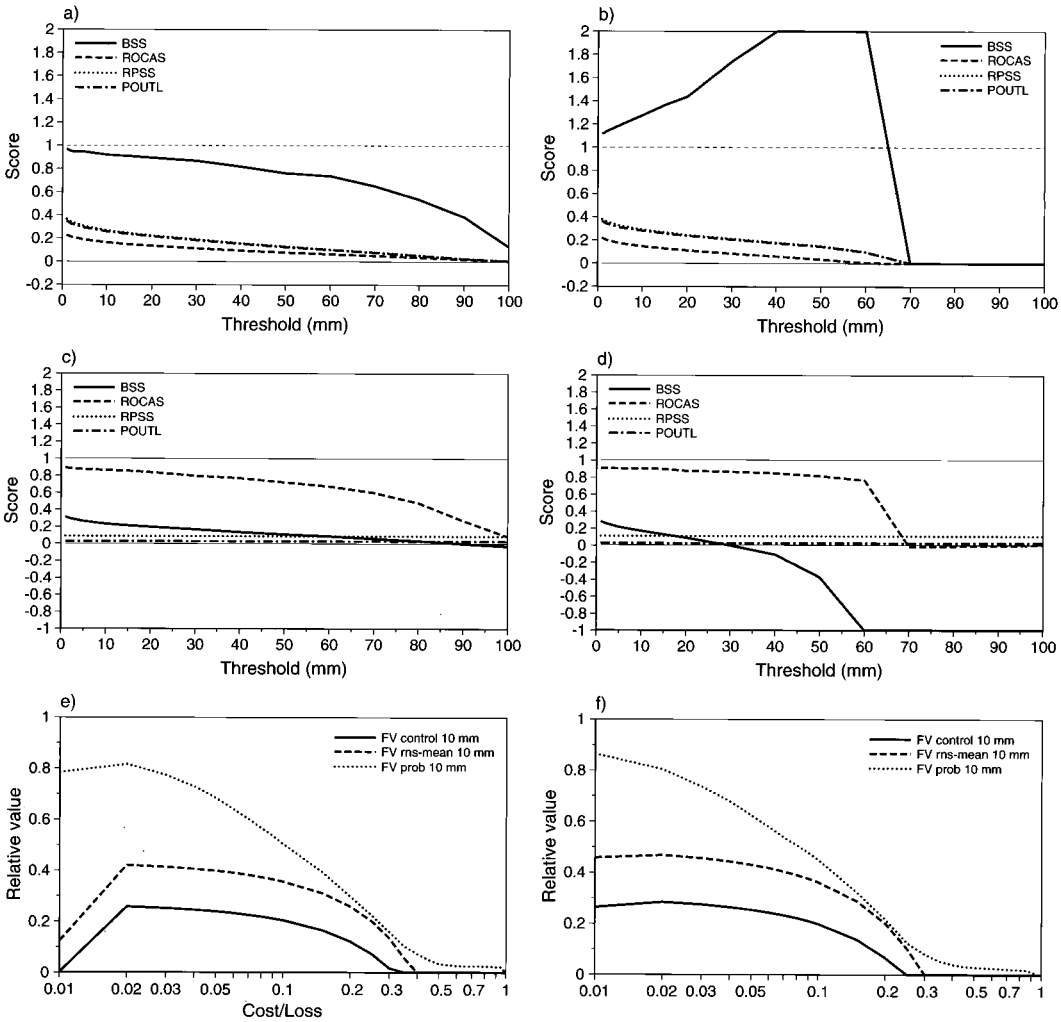

FIG. 11. Average performance (90 cases) of an ensemble characterized by a systematic 40% underestimation (left column) and by a 40% overestimation (right column). [(a), (b)] Seasonal average score of the control forecast: bias (solid line), TS (dashed line), POD (dotted line), and KSS (chain-dashed line). [(c), (d)] Seasonal average probability scores: BSS (solid line), ROC area skill score (dashed line), RPSS (bold dotted line), and POUTL (chain-dashed line). [(e), (f)] forecast value curves for the probabilistic prediction of 10 mm for the control (solid line), the ensemble mean (dashed line), and the ensemble probabilistic prediction (dotted line).

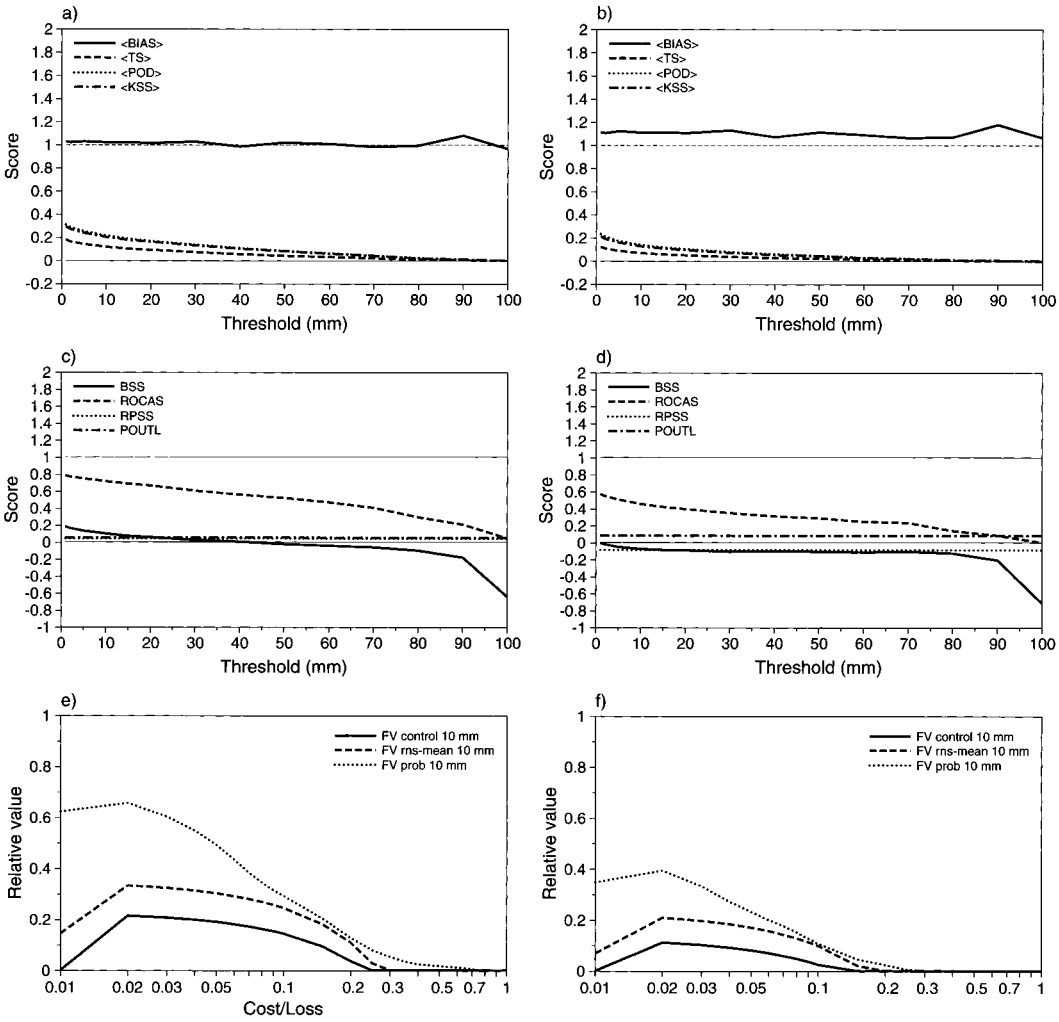

FIG. 12. Average performance (90 cases) of an ensemble characterized by a systematic 3° position error (left column) and by a 6° position error (right column). [(a), (b)] Seasonal average score of the control forecast: bias (solid line), TS (dashed line), POD (dotted line), and KSS (chain-dashed line). [(c), (d)] Seasonal average probability scores: BSS (solid line), ROC area skill score (dashed line), RPSS (bold dotted line), and POUTL (chain-dashed line). [(e), (f)] Forecast value curves for the probabilistic prediction of 10 mm of rain for the control (solid line), the ensemble mean (dashed line), and the ensemble probabilistic prediction (dotted line).

left of the forecast value curves of the system affected by overestimation (Figs. 11e,f). This indicates that, depending on whether a user has a high or low cost–loss ratio C/L, it could be more valuable to under- or overestimation.

c. Impact of systematic position errors

Consider now two ensemble systems characterized by random errors (as for the ensemble discussed in section 6a) and by a 3° or a 6° position error (which is two or four times the standard deviation of the observed pattern; see Table 4). These results have been obtained by defining the forecasts as in section 6a but by shifting the position of the maximum value of each day verification field by 3° or 6°. These examples can be thought to describe the performance of an ensemble of forecasts generated by a model with a tendency to predict too weak zonal flows. Figure 12 shows the performance of these systems.

The impact on the average scores of each single forecast of either a 3° or a 6° error is very small (see Figs. 12a,b and 9a) and almost undetectable on the threat score and the Kuipers skill score. The impact is larger on the probabilistic predictions (see Figs. 12c,d and 9c) especially for thresholds of 20 mm or more. Results show that for thresholds larger than 30 or 20 mm, respectively, a 3° or 6° position error makes the BSS negative. The impact on the potential forecast value reflects the impact on the ROCAS (see Figs. 12e,f and 9d).

It is interesting to compare these results of this section with the results obtained in the previous section. In par-

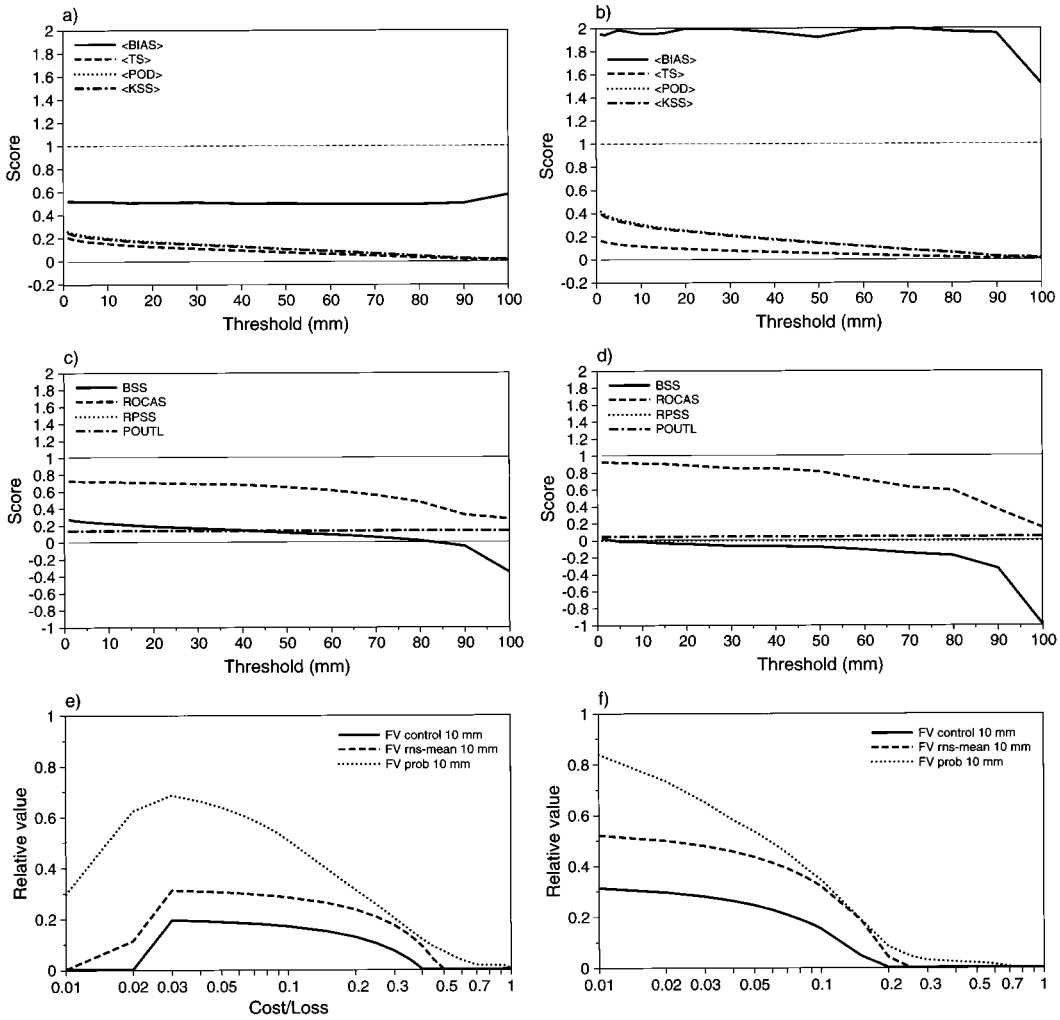

FIG. 13. Average performance (90 cases) of an ensemble characterized by a systematic two times too small prediction of $\sigma_{x,j}$ (left column) and by a two times too large prediction of $\sigma_{x,j}$ (right column). [(a), (b)] Seasonal average score of the control forecast: bias (solid line), TS (dashed line), POD (dotted line), and KSS (chain-dashed line). [(c), (d)] Seasonal average probability scores: BSS (solid line), ROC area skill score (dashed line), RPSS (bold dotted line), and POUTL (chain-dashed line). [(e), (f)] forecast value curves for the probabilistic prediction of 10 mm of rain for the control (solid line), the ensemble mean (dashed line), and the ensemble probabilistic prediction (dotted line).

ticular, the comparison of the potential forecast value for the 10-mm threshold (see Figs. 11e,f; 12e,f; and 9c) indicate that systematic position errors of 3° to 6° (i.e., of 2 to 4 standard deviation) reduce the potential forecast value of single deterministic and probabilistic forecasts more than systematic over- or underestimation errors of 40%.

d. Impact of systematic ''shape'' errors

Consider now two ensemble systems characterised by random errors (as for the ensemble discussed in section 6a) and by a systematic prediction of a too broad or too narrow precipitation field. More specifically, consider two ensemble systems defined by a two times too small or too large $\sigma_{x,j}$. These results have been obtained by defining the forecasts as in section 6a but by rescaling $\sigma_{x,0}$ by a factor of 2 or by a factor of 0.5. Figure 13 shows the performance of these systems.

The impact on the average scores of each single forecast of predicting two times too small or too large $\sigma_{x,j}$ is qualitatively similar to the impact of under or overestimation. The prediction of a two times too small $\sigma_{x,j}$ leads to an average bias of 0.5 (Fig. 13a), while the prediction of a two times too large $\sigma_{x,j}$ leads to an average bias of 2.0 (Fig. 13b). Qualitatively similar to the impact of under- or overestimation, the prediction of a two times too small $\sigma_{x,j}$ has a small impact on the probabilistic scores (Fig. 13c) while the prediction of a two times too large $\sigma_{x,j}$ leads to negative Brier skill scores for all thresholds (Fig. 13d). The impact on the ROCAS is smaller than the impact on the Brier skill score, and

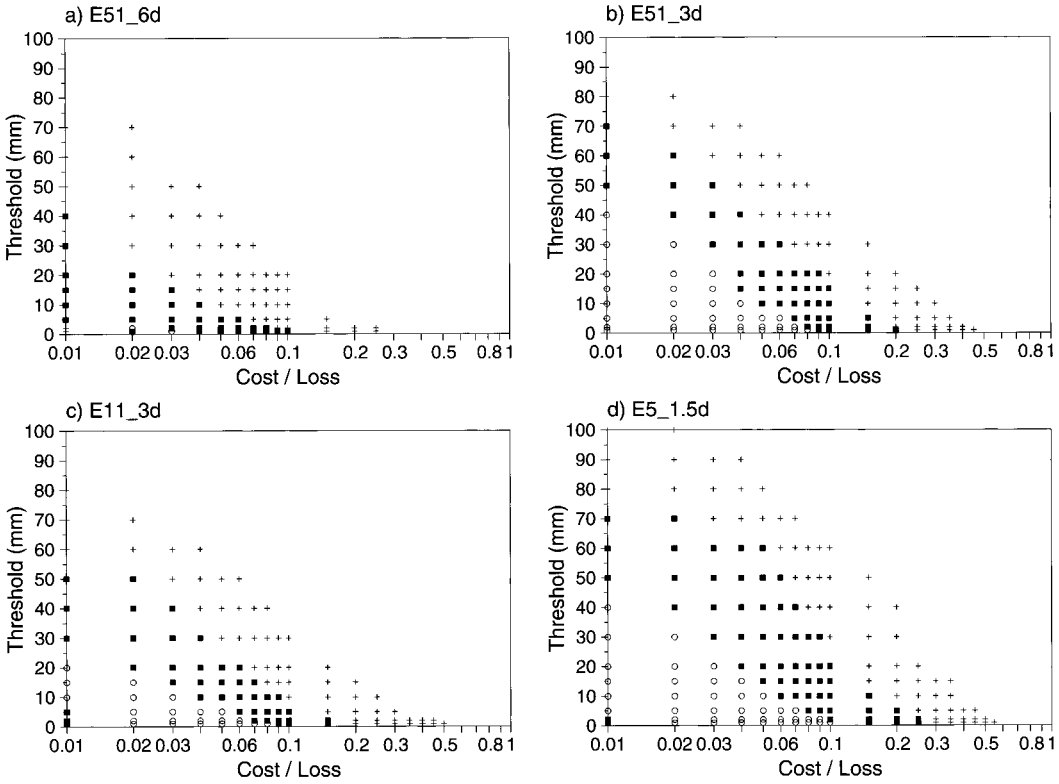

FIG. 14. Potential forecast value as a function of the precipitation amount and the cost–loss ratio for ensemble configuration (a) E51_6d, (b) E51_3d, (c) E11_3d, and (d) E5_1.5d. Symbols for forecast values are crosses for $0.05 < \text{FV} \leq 0.25$, squares for $0.25 < \text{FV} \leq 0.50$, full gray square for $0.50 < \text{FV} \leq 0.75$, and full circles for $0.75 < \text{FV}$.

it is similar for both too small or too large predicted $\sigma_{x,j}$ (Figs. 13c,d). It is worthwhile to point out that the potential forecast value curves for the system affected by a two times too small $\sigma_{x,j}$ are shifted to the right with respect to the forecast value curves of the system affected by two times too large $\sigma_{x,j}$ (Figs. 13e,f). Again, note that this shift is qualitatively similar to the shift of the forecast value curves of systems affected by over- or underestimation (Figs. 11e,f).

e. Impact of ensemble size and systematic model (position) errors

Consider now four ensemble configurations, each of them characterized by a different ensemble size and based on models with different systematic errors. Ensemble E51_6d has 51 members and uses a model affected by a 6° systematic position error; ensembles E51_3d and E11_3d have, respectively, 51 and 11 members with a 3° systematic position error; and finally E5_1.5 has only 5 members with a 1.5° systematic position error. The comparison of the performance of the four ensembles helps in addressing the question of whether an ensemble with a small ensemble size but based on an accurate forecast model has higher potential forecast value than an ensemble based on a larger set of integrations of a less accurate model. Figure 14 shows the potential forecast value for the four ensemble configurations.

The comparison of the potential forecast value of configurations E51_6d and E51_3d (Figs. 14a,b) confirms the results discussed above: that a reduction of model systematic error increases the forecast value. The comparsion of the potential forecast value of configurations E51_3d and E11_3d (Figs. 14b,c) shows that a 90% reduction of the ensemble size decreases the forecast value of an ensemble system based on an accurate model to almost the same level as configuration E51_6d, which is based on a poor model (Fig. 14a). On the other hand, Fig. 14d confirms that a further reduction of the model systematic error can increase the potential forecast value. In other words, these results indicate that the potential forecast value of an ensemble system is strongly dependent on both ensemble size and model accuracy (in this particular case the potential forecast value is more sensitive to model error than ensemble size).

f. Sensitivity of the ensemble performance to ensemble spread

The results presented so far did not analyze the sensitivity of the ensemble scores on the ensemble spread. This point has to be addressed since any ensemble system must have the right level of spread to be able to

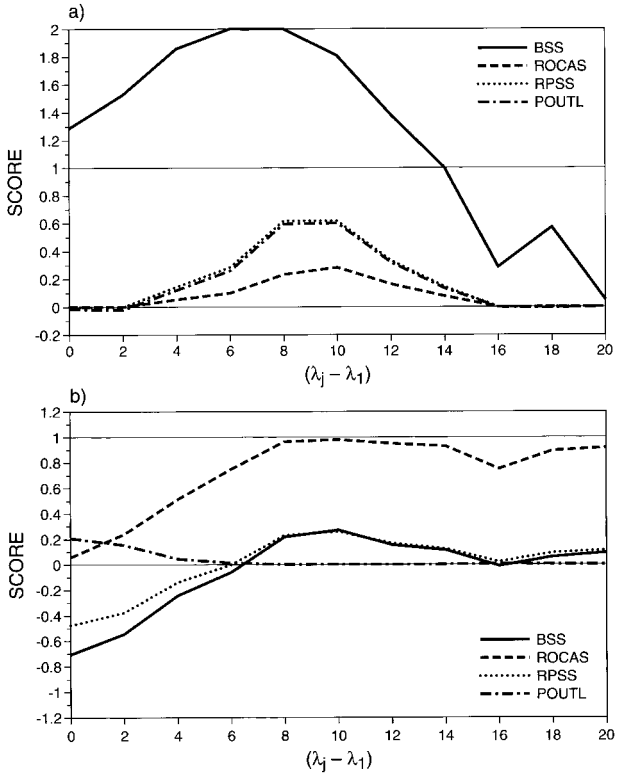

FIG. 15. (a) Sensitivity of the scores of the ensemble mean forecast of 10 mm to the ensemble spread: bias (solid line), TS (dashed line), POD (dotted line), and KSS (chain-dashed line). (b) Sensitivity of the ensemble probability scores for the prediction of 10 mm of precipitation: BSS (solid line), ROC area skill score (dashed line), RPSS (dotted line), and POUTL (chain-dashed line). Ordinate: forecast score. Abscissa: ensemble forecast range $(\lambda_j - \lambda_1)$.

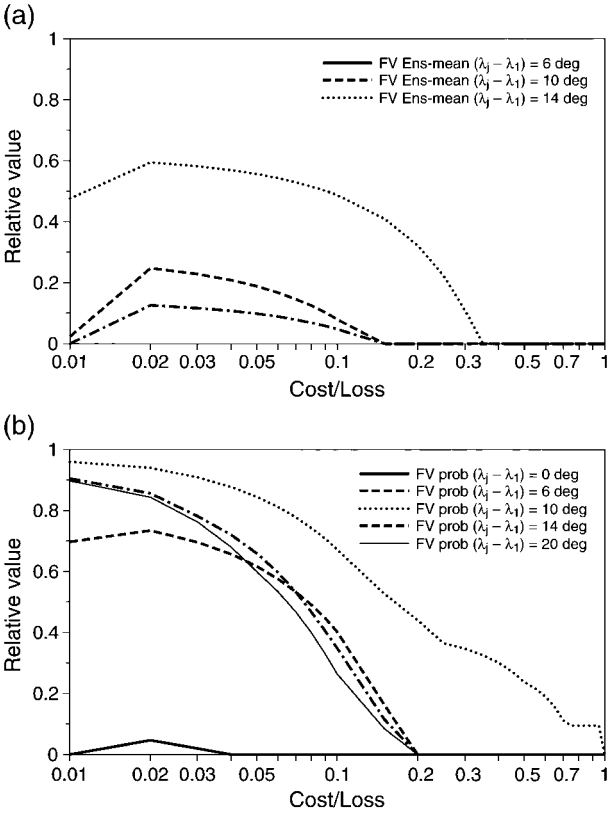

FIG. 16. Forecast value for (a) the ensemble mean and (b) the ensemble probabilistic prediction of 10 mm for different levels of spread simulated by varying the ensemble forecast range for the parameter λ_j (spread ranging from small to large): $(\lambda_j - \lambda_1) = 0$ (solid line), $(\lambda_j - \lambda_1) = 6$ (dashed line), $(\lambda_j - \lambda_1) = 10$ (dotted line), $(\lambda_j - \lambda_1) = 14$ (chain-dashed line), and $(\lambda_j - \lambda_1) = 20$ (thin solid line).

include the verification inside the range spanned by the ensemble forecast. On the other hand, it is worth investigating whether ensemble systems with a wrong level of spread may anyway provide probabilistic forecasts with a higher potential economic value than single deterministic forecasts.

Consider an ensemble of forecasts characterized by a mean position error of 6° in longitude ($\lambda_j - \lambda_0 = 6°$), with the observed and the predicted Gaussian distributions characterized by $1.1 \leq \sigma_{x,j} \leq 1.9$ and $0.8 \leq \sigma_{y,j} \leq 1.2$. Suppose that the ensemble has the right level of spread in the meridional direction, and consider the sensitivity of the ensemble performance to the spread in the zonal direction. This can be investigated by considering ensembles with different ranges for the parameters λ_j. Figures 15 and 16 show some results relative to the 10-mm threshold.

Due to the systematic position error the control forecast of 10 mm of rain has no skill (both in terms of TS or KSS, not shown). By contrast, the ensemble mean has a positive TS and a positive KSS if the ensemble spread is neither too small nor too large (Fig. 15a). Similarly, the ensemble probabilistic prediction of 10 mm of rain has a positive BSS and a high ROCAS only if the ensemble spread is neither too small nor too large (Fig. 15b). Similarly, the potential forecast value is positive for both the ensemble mean forecast (Fig. 16a) and the ensemble probabilistic prediction (Fig. 16b) only if the ensemble spread is neither too small nor too large. Note that results indicate that even if the ensemble spread is not properly tuned (but not outrageously wrong) ensemble probabilistic predictions can be skillful and have potential economic value.

7. Conclusions

Issues related to the verification of the accuracy of categorical and probabilistic forecasts of discrete dichotomous events (occurrence/nonoccurrence) have been discussed. Synthetic forecasts have been compared to synthetic verification fields. The accuracy of categorical and probabilistic forecasts has been assessed using a variety of accuracy and skill measures (hit rate, threat score, probability of detection and probability of false alarm, bias, Kuipers skill score, Brier score, and skill score, ROC area score and skill score, rank probability skill score, probability of outliers).

A simple decision model has been used to estimate the potential economic value of both categorical and probabilistic forecasts (as in Richardson 2000). It has been shown that the potential economic value can be written as a weighted difference between the system probability of detection and the probability of false detection. It has also been shown that the Kuipers skill score gives the maximum potential economic value.

Each forecast accuracy or skill measure summarizes in one number the differences between observed and forecast pattern. It has been shown how difficult it is to associate to one specific number the actual difference between the two patterns, and that the use of more than one accuracy or skill measure gives a more complete picture of the performance of a system. Different measures of forecast accuracy have been shown to have a certain degree of coherence in behavior, all showing a qualitatively similar response to increasing model errors. Nevertheless, it has been also shown that a quantitative disagreement of the response can occur, with forecasts judged to be skillful according to one measure judged to have no skill according to others. This supports Murphy's (1991) indication of the large dimensionality of the verification problem.

The sensitivity of accuracy and skill measures to imposed random and systematic errors has been investigated. It has been shown how accuracy and skill measures are sensitive to the area definition, thus indicating that care must be taken when comparing forecast scores for different regions characterized by different observed frequencies. The sensitivity of accuracy and skill measures to amplitude, position and ''shape'' errors have been studied. Considering the Brier skill score or the ROC area skill score, results indicated, for example, that position errors could have bigger effect than over/underestimation errors.

Ensembles with different size and systematic model errors have been compared to investigate the sensitivity of probabilistic forecasts to ensemble configuration. Results have shown that both model errors and ensemble size affect the accuracy and the potential economic value of an ensemble system, with small-size accurate ensembles performing on average similarly to large-size less accurate ensembles. Results have also confirmed that any ensemble system should have the right level of spread to be skillful and have high potential economic value.

The potential economic value of categorical forecasts generated by single deterministic forecasts given by one member of an ensemble of 51 forecasts has been compared with the potential economic value of the probabilistic forecasts given by the whole ensemble. Results indicate that, independently from the model random or systematic error, ensemble-based probabilistic forecasts exhibit higher potential economic values than categorical forecasts.

These results indicate that the design of a forecasting system should follow the definition of its purposes (i.e., the definition of the accuracy measures used to gauge its performance). The design should be such that the ensemble system maximizes its outcome as assessed by the accuracy measures that best quantify the achievement of its purposes.

Acknowledgments. I am very grateful to Robert Hine for all his editorial work, which improved substantially the quality of all the figures. David Richardson provided very useful comments to an earlier version of this manuscript. I am grateful to Steve Mullen and to two anonymous referees whose comments helped improve the first version of this manuscript.

REFERENCES

Doswell, C. A., III, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting,* **5,** 576–585.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

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

Hanssen, A. W., and W. J. A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Meded. Verh.,* **81,** 2–15.

Katz, R. W., A. H. Murphy, and R. L. Winkler, 1982: Assessing the value of frost forecasts to orchardists: A dynamic decision-making approach. *J. Appl. Meteor.,* **21,** 518–531.

Mason, I., 1982: A model for assessment of weather forecasts. *Austr. Meteor. Mag.,* **30,** 291–303.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.,* **10,** 155–156.

——, 1985: Decision making and the value of forecasts in a generalized model of the cost-loss ratio situation. *Mon. Wea. Rev.,* **113,** 362–369.

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

——, 1996: The Finley affair: A signal event in the history of forecast verification. *Wea. Forecasting,* **11,** 3–20.

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

Strauss, B., and A. Lanzinger, 1995: Validation of the ECMWF EPS. *Proc. ECMWF Seminar on Predictability,* Vol. 2, Shinfield Park, Reading, United Kingdom, ECMWF, 157–166.

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

——, and T. M. Hamill, 1995: Potential economic value of ensemble-based surface weather forecasts. *Mon. Wea. Rev.,* **123,** 3564–3575.