A Local Ensemble Prediction System for Fog and Low Clouds: Construction, Bayesian Model Averaging Calibration, and Validation

STEVIE ROQUELAURE AND THIERRY BERGOT

CNRM-GAME, Météo-France, CNRS, Toulouse, France

(Manuscript received 15 May 2007, in final form 28 April 2008)

ABSTRACT

At main international airports, air traffic safety and economic issues related to poor visibility conditions are crucial. Meteorologists face the challenge of supplying airport authorities with accurate forecasts of fog and cloud ceiling. These events are difficult to forecast because conditions evolve on short space and time scales during their life cycle. To obtain accurate forecasts of fog and low clouds, the Code de Brouillard à l'Echelle Locale (the local scale fog code)–Interactions between Soil, Biosphere, and Atmosphere (COBEL–ISBA) local numerical forecast system was implemented at Charles de Gaulle International Airport in Paris. However, even with dedicated observations and initialization, uncertainties remain in both initial conditions and mesoscale forcings. A local ensemble prediction system (LEPS) has been designed around the COBEL–ISBA numerical model and tested to assess the predictability of low visibility procedures events, defined as a visibility less than 600 m and/or a ceiling below 60 m. This work describes and evaluates a local ensemble strategy for the prediction of low visibility procedures. A Bayesian model averaging method has been applied to calibrate the ensemble. The study shows that the use of LEPS for specific local event prediction is well adapted and useful for low visibility prediction in the aeronautic context. Moreover, a wide range of users, especially those with low cost–loss ratios, can expect economic savings with the use of this probabilistic system.

1. Introduction

In the 1960s, studies by Lorenz (1963, 1969) revealed the chaotic nature of the atmosphere. In the context of numerical prediction, small errors in initial conditions grow inexorably and affect the predictability of the weather. Even with perfect numerical models, beyond a certain limit in time, any single deterministic forecast becomes useless. One way of circumventing this problem is to use ensemble forecasts. A benefit of ensemble forecasting is that it helps forecasters to predict the likelihood of unusual events such as fog. The use of ensemble is a practical way of estimating the uncertainty of a weather forecast. An ensemble forecast system is composed of multiple individual numerical forecasts (members) generated from a set of different initial conditions and/or different numerical configurations (Leith 1974). Thus, probabilistic forecasts can be obtained from the relative frequencies of events represented in the ensemble.

In the early 1990s, thanks to the increase in computer power resources, an interest in ensemble forecasting techniques developed and the production of probabilistic forecasts began to emerge. As a result, ensemble prediction systems (EPS) are now widely used operationally in meteorological centers around the world such as the National Centers for Environmental Prediction (NCEP) in the United States (Toth and Kalnay 1993), the European Centre for Medium-Range Weather Forecasts (ECMWF) in Europe (Buizza 1997; Buizza and Palmer 1998), and the Meteorological Service of Canada (MSC) in Canada (Houtekamer and Lefaivre 1997). Techniques have been developed in these centers, first for medium-range forecasts (3–7 days) and then later for short-range forecasts (12–72 h). The originality of the present study lies in the fact that a local ensemble forecast system is designed for the very short-term forecasting (0–12 h) of specific conditions such as fog and low ceiling conditions at a local airport area (some preliminary work of Stessel et al. 2000). The local ensemble system is based on a 1D

Corresponding author address: Stevie Roquelaure, CNRM-GAME, GMME/Météo France, 42 Avenue Coriolis, 31 057 Toulouse, France.
E-mail: stevie.roquelaure@cnrm.meteo.fr

DOI: 10.1175/2008JAMC1783.1

model forced with specific mesoscale forcings provided by a 3D numerical weather prediction (NWP) model, along with a 1D variational assimilation scheme forced by dedicated on-site observations.

Despite all the improvements in horizontal and vertical resolutions and the physics of 3D NWP models, the prediction of fog and low clouds still remains a challenge. Because fog events occur on relatively short space and time scales, forecasters are faced with the very difficult task of having to formulate space- or time-specific skillful forecasts. Murphy (1991) discussed the scientific and economic reasons for using probabilistic forecasts for predicting rare and severe weather. In the context of forecasting these events, forecasters have to formulate judgments regarding the likelihood of the occurrence of the events of interest. Following this procedure, these forecasts are typically biased because they mostly rely on the subjective judgment of forecasters, thereby possessing an inherently probabilistic nature. An explicit probabilistic forecast can provide a fair estimation of the risk that an event occurs, and users should be able to take advantage of this objective forecast depending on their needs.

At major international airports, aviation forecasters are concerned with rare events such as fog, and they also forecast the complete life cycles of low clouds. In practice at Charles de Gaulle International Airport in Paris, adverse ceiling and visibility conditions (visibility less than 600 m and/or ceiling below 60 m) lead to the application of low visibility procedures (LVP). The LVP application reduces the airport efficiency for take-off/landing by a factor of 2, causing aircraft delays. For these reasons, airport authorities require skillful local predictions of LVP to efficiently manage air traffic in a safe manner. As a part of the operational process, a local (1D) approach was implemented in 2005 at Charles de Gaulle airport to provide fog and low-cloud life cycle forecasts (Bergot et al. 2005). The 1D Code de Brouillard à l'Echelle Locale (COBEL) model (Bergot 1993) coupled with Interactions between Soil, Biosphere, and Atmosphere (ISBA; Boone et al. 2000; Boone 2000) is used together with a 1D variational assimilation approach based on local observations.

The goal of this study is to develop a local ensemble prediction system (LEPS) around COBEL–ISBA to estimate the likelihood of LVP occurrence at Charles de Gaulle airport. An evaluation of the predictability of local LVP conditions is obtained. Ensemble prediction is usually performed by running 3D NWP models. In this work, we propose a novel implementation of this technique to a local framework. In the next sections, we will describe LEPS from its construction to its validation. Section 2 will present the construction of the ensembles used in the study. Section 3 will describe the Bayesian model averaging (BMA) calibration applied to the ensembles to improve their reliability. Sections 4 and 5 are dedicated to the validation of LEPS. And finally, section 6 summarizes the results and concludes with LEPS' ability in and efficiency for LVP prediction.

2. The deterministic local forecast system, uncertainties, and sensitivity

a. The COBEL–ISBA numerical prediction system

In a local approach, a one-dimensional numerical model is used to describe the evolution of the boundary layer within a representative column of the atmosphere. These 1D numerical modeling strategies seem to be an interesting alternative for forecasting short space and time scale meteorological events. The mesoscale influences are considered by including mesoscale forcings in the column during the simulations. Thus, local approaches require two kinds of inputs: initial conditions (mainly atmospheric temperature and humidity profiles and soil temperature and water content profiles) and mesoscale forcings (mainly advection profiles and cloud cover). Currently, local approaches are used operationally to forecast the fog and low-cloud life cycles at San Francisco (Clark 2002) and Charles de Gaulle airports (Bergot et al. 2005). The same kind of strategy has been tested on the U.S. northeastern coast within the framework of the U.S. Federal Aviation Administration ceiling and visibility project (Herzegh et al. 2002).

The focus of this work is on the numerical prediction method used at Charles de Gaulle airport, which has some specific characteristics.

1) The 1D high-resolution COBEL (local scale fog code) atmospheric model (Bergot 1993; Bergot and Guédalia 1994) is coupled with the multilayer surface–vegetation–atmosphere transfer scheme ISBA (Boone et al. 2000; Boone 2000).
2) Specific observations from a 30-m-high meteorological tower (atmospheric temperature and humidity and short and longwave radiation fluxes) and soil measurements. These observations allow for a better description of the vertical structure of the boundary layer as well as the soil characteristics. They are used in a local assimilation scheme to construct initial conditions based on a 1D variational (1DVAR) assimilation scheme, together with a specific fog and low-cloud initialization.
3) The mesoscale influences are integrated at the local scale by taking the horizontal temperature and humidity advections, the geostrophic wind, and the

cloud cover from the Météo-France operational NWP model Aladin (information on Aladin is available online at http://www.cnrm.meteo.fr/aladin/).

COBEL–ISBA inputs are the atmospheric temperature and humidity profiles from the 1DVAR system, the geostrophic wind profiles, advection profiles, cloud cover and the soil temperature and water content profiles. The model computes the atmospheric temperature and humidity profiles, the wind profiles, the turbulent kinetic energy profile, and the atmospheric liquid water profile, from which visibility is diagnosed.

b. *Estimation of uncertainties and forecast sensitivity*

Despite the care in initializing COBEL–ISBA through the 1DVAR assimilation of on-site observations together with a fog and low-cloud specific initialization, uncertainties still remain with both the initial conditions and the mesoscale forcings. These uncertainties have been evaluated and quantified in a previous paper (Roquelaure and Bergot 2007). The results are summarized in the following subsections.

1) INPUT PARAMETERS

Mesoscale forcing uncertainty computation is based on a spatiotemporal strategy, using the hypothesis that uncertainty is correlated with the "intrinsic" variability of the 3D NWP model Aladin. The model variability is assessed in both space and time. The spatial variability is evaluated by comparing the forecast over an area of 3×3 grid points. This area is representative of the homogeneous surface conditions around the study area. The temporal variability is evaluated by comparing four Aladin runs (0000, 0600, 1200 and 1800 UTC) for the same verification time. At the end of this two-step procedure, the variability in both space and time is used to estimate the distribution of uncertainties on mesoscale forcings.

Initial condition uncertainties are estimated from errors on the observations for the soil and the lower part of the atmosphere, where site observations are available (below 30 m). At higher elevations, output from NWP model Aladin is used to provide both temperature and humidity profiles. As a consequence, uncertainties are assessed with the spatiotemporal methodology described earlier for mesoscale forcings.

2) FORECAST SENSITIVITY SUMMARY

At the airport location, the influence of these uncertainties on the COBEL–ISBA forecasts has been evaluated during the 2002/03 winter season (Roquelaure and Bergot 2007). The study has shown the time dependency of forecast dispersion (Fig. 1). The influence of uncertainties on initial conditions decreases during the first hours of the simulation (0–6 h), whereas the dispersion created by mesoscale forcings becomes more important in the second half of the simulation (6–12 h). The cloud radiative impact on dispersion is felt during the 12-h forecast period as well as during the low-cloud initialization. A heat and humidity budget analysis applied on the guess fields has permitted the quantification of the indirect impact of perturbations on the variational data assimilation scheme. Perturbations grow during the cycle and "feed" subsequent analyses through the assimilation process. Errors in the model initialization and forcing therefore propagate throughout the assimilation–forecast cycle.

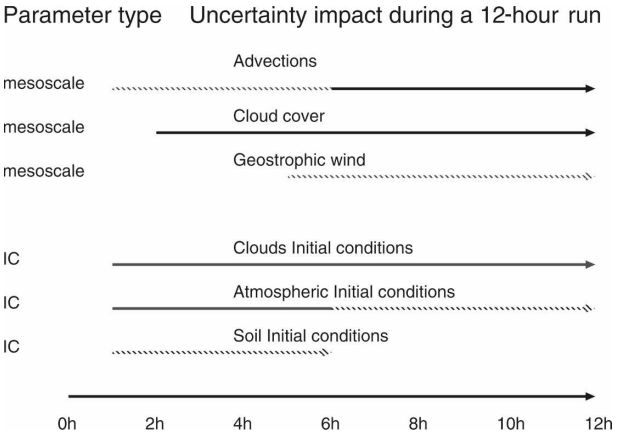

FIG. 1. Summary of the impact of uncertainty on fog and low-cloud forecasts (from Roquelaure and Bergot 2007). For each parameter type, the solid section of the arrow shows when the dispersion is higher and the dashed section shows when the dispersion is weaker during the 12-h run.

3. Construction of the ensembles

a. *Overview on ensembles*

Ensemble prediction techniques are designed to estimate the level of confidence in a particular forecast. Theoretically, the goal is to make an explicit computation—through the Liouville equations—of the probability density function (pdf) of a forecast from the pdf of the initial state (Ehrendorfer 1994). Ideally, multiple perturbed initial states, derived from a reference initial state, represent the pdf of the initial state pdf. However, even in a local approach, multiple model integrations of these perturbed states are costly and become rapidly prohibitive if a complete description of the forecast pdf is desired. The proposed forecast system has to be operationally used and, consequently, a pure Monte Carlo

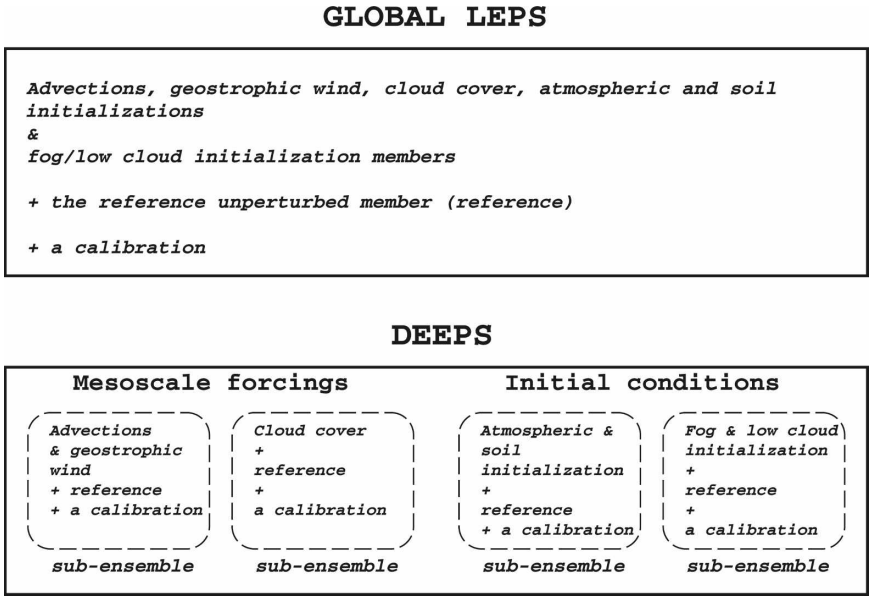

FIG. 2. Description of the ensemble construction: the global LEPS and DEEPS.

methodology cannot be applied. Therefore, the pdf has to be approximated using a finite sample of forecast scenarios. The sampling strategy is based on the perturbations of the initial conditions and mesoscale forcings. Ensembles are built using the "perfect model" assumption because the model physics are not perturbed. In a perfect ensemble system, forecast probabilities are reliable because they match the observed frequencies in a data sample. In practice, ensemble systems are not perfect mainly because of an imperfect model and sampling problems. Consequently, ensemble members are not equipropable, and the rough ensemble probabilities are not reliable. A calibration technique is used to circumvent this problem and increase the ensemble reliability. The calibration corrects the forecast probabilities using a statistical apprenticeship on a training dataset.

b. *Ensemble approaches: Global versus decomposed*

Two methodologies of ensemble construction have been tested in this study (Fig. 2). On the one hand, a global ensemble has been designed. All members are grouped into a single ensemble, and the calibration method is applied jointly on all members. This approach allows the prediction of the LVP pdf and, therefore, provides a probability of LVP occurrence. Each member has its own assimilation cycle.

On the other hand, a decomposed ensemble is also evaluated following the results of the sensitivity study. The members are split into four subensembles (or blocks) following the parameter, which is perturbed, and the calibration is applied within each subensemble (Tables 1 and 2). Four subensembles are designed for the mesoscale forcing, the initial conditions, the fog/stratus initializations, and the cloud cover. This is a reasonable approach because the dispersion of COBEL–ISBA LVP forecasts depends upon the time of the forecast as well as the parameter being considered. This result suggests that a decomposed approach can provide the LVP pdf but also a forecast diagnosis following each parameter and identify the physical drivers of a given meteorological situation. The decomposed ensemble could have the potential to indicate the sources of uncertainty in the forecast. Each member also has its own assimilation cycle.

4. Ensemble calibration

a. *The Bayesian model averaging calibration*

The calibration technique for the global LEPS and decomposed ensemble prediction system (DEEPS) subensembles follows the BMA method described in Raftery et al. (2005). The main idea underlying the BMA is that in any ensemble forecast there is a "best" member, but we do not know which one it is. The BMA is going to learn from a training dataset which members are the most efficient for the prediction of any quantity X (the occurrence of LVP in our case). Thanks to the apprenticeship, the BMA method will assign a weight to each member to improve the ensemble reliability. Consequently, each member is clearly identified and

TABLE 1. The 18 mesoscale forcing members used in both the 30-member global and decomposed ensemble approaches [STD means standard deviation following the variances of the uncertainty distributions evaluated in Roquelaure and Bergot (2007), ref means reference, and the word *block* is a synonym for *subensemble* in DEEPS].

Member	Ensembles: Global	Ensembles: Decomposed
1—reference member	Ref with Aladin cloud cover	Included in each block
2	Ref with clear sky	Cloud cover block
3	Ref with persistence	Cloud cover block
4	Ref + 1 STD on Aladin cloud cover	Cloud cover block
5	Ref − 1 STD on Aladin cloud cover	Cloud cover block
6	Ref + 1 STD on persistence	Cloud cover block
7	Ref − 1 STD on persistence	Cloud cover block
8	Ref + unperturbed advections	MF block
9	Ref + 0.5 STD on humidity advections	MF block
10	Ref − 0.5 STD on humidity advections	MF block
11	Ref + 1 STD on humidity advections	MF block
12	Ref − 1 STD on humidity advections	MF block
13	Ref + 0.5 STD on temperature advections	MF block
14	Ref − 0.5 STD on temperature advections	MF block
15	Ref + 1 STD on temperature advections	MF block
16	Ref − 1 STD on temperature advections	MF block
17	Ref + 1 STD on geostrophic wind	MF block
18	Ref − 1 STD on geostrophic wind	MF block
	Cloud cover subensemble: 7 members	
	MF subensemble: 12 members	

TABLE 2. As in Table 1 but for the 12 initial condition members used in both the 30-member global and decomposed ensemble approaches.

Member	Ensembles: Global	Ensembles: Decomposed
19	Ref + 1 STD on humidity atmospheric profile	IC block
20	Ref − 1 STD on humidity atmospheric profile	IC block
21	Ref + 1 STD on temperature atmospheric profile	IC block
22	Ref − 1 STD on temperature atmospheric profile	IC block
23	Ref + 1 STD on humidity soil profile	IC block
24	Ref − 1 STD on humidity soil profile	IC block
25	Ref + 1 STD on temperature soil profile	IC block
26	Ref − 1 STD on temperature soil profile	IC block
27	Cloud top + 1 grid point for fog and 2 for stratus	Fog/stratus IC block
28	Cloud top + 1 grid point for fog and 2 for stratus	Fog/stratus IC block
29	Ref + 1 STD on liquid water content	Fog/stratus IC block
30	Ref − 1 STD on liquid water content	Fog/stratus IC block
	IC subensemble: 9 members	
	Fog/stratus subensemble: 5 members	

has its own characteristics. If K members are available in the training dataset X^T, BMA takes into account that all members learn about each member's efficiency in forecasting the variable X. The law of total probability states that the forecast pdf $p(X)$ is given by

$$p(X) = \sum_{k=1}^{K} p(X|M_k)p(M_k|X^T), \quad (1)$$

where $p(X|M_k)$ is the forecast pdf based on member M_k, and $p(M_k|X^T)$ is the posterior probability of member M_k being correct on the training data. These posterior probabilities have to sum up to one, $\sum_{k=1}^{K} p(M_k|X^T) = 1$, and they can be interpreted as weights $[w_k = p(M_k|X^T)]$.

b. Estimation of BMA weight by maximum likelihood: The EM algorithm

The BMA weights w_k, $k = 1, \ldots, K$, and the variance σ^2 of the BMA pdf are estimated by maximum likelihood from the training data (Fisher 1922). The maximum likelihood estimator is the value of the parameter that maximizes the likelihood function, that is, the value of the parameter under which the observed data were most likely to have occurred.

The log-likelihood function, defined as l in Eq. (2), is generally maximized instead of the likelihood function itself for algebraic simplicity and to ensure numerical stability. Assuming independence of forecast errors in space s and time t, the general formulation for a variable X is

$$l(w_1, \ldots, w_k, \sigma^2) = \sum_{s,t} \log \sum_{k=1}^{K} w_k p_k(X_{st}|f_{st}). \quad (2)$$

For LVP forecast, the formulation is easier; the pdf $p_k(X_{st}|f_{st})$ is discrete and takes only two values: 1 for a

hit (LVP is observed and forecast, or LVP not observed and not forecast) and 0 for a miss (LVP is not observed and forecast, or LVP observed and not forecast). The log-likelihood function is maximized by the expectation–maximization (EM) algorithm (Dempster et al. 1977; MacLachlan and Krishnan 1997).

The EM algorithm is iterative and alternates between the E and M steps. It starts with an initial guess for the weights. In the E step, the unobserved quantities z_{kt} are estimated from the guess; in the M step, the weights are reestimated given the current values of the z_{kt}. For the BMA model given by Eq. (1), the E step is

$$z_{kt}^{(j+1)} = \frac{w_k^{(j)} p^{(j)}(X_t | f_{kt})}{\sum_{l=1}^{K} w_l^{(j)} p^{(j)}(X_t | f_{lt})}, \tag{3}$$

where the superscript j refers to the jth iteration of the algorithm, and $w_k^{(j)}$ refers to w_k at this iteration.

The M step is the estimation of w_k using the current estimates of $z_{kt}^{(j+1)}$ as weights, and n is the number of cases in training set,

$$w_{kt}^{(j+1)} = \frac{1}{n} \sum_{t} z_{kt}^{(j+1)}. \tag{4}$$

c. Ensembles and verification dataset

The test ensemble configuration consists of 30-member forecasts with perturbations on initial conditions and mesoscale forcings (Tables 1 and 2). COBEL–ISBA members have been obtained by running the model at Charles de Gaulle airport during three winter seasons, from 2002 to 2005. Runs of 12 h have been performed with a 3-h data assimilation frequency (about 1200 runs per winter). Observations have been collected during the same period, and the validation of model forecasts with observations was performed by considering 30-min time intervals. The first two winters are used for BMA weights computation, and the last winter (2004/05) is kept for validation to preserve the independence between the training and the verification datasets.

The dispersion of LVP forecasts has been shown to be dependent on forecast time (Roquelaure and Bergot 2007); consequently, BMA weights are also going to be computed as a function of the forecast time. However, LVP events are rare; the climatology frequency of LVP events is about 6% during winter 2004/05 (the climatology frequency of the training data is also 6% for winters 2002/03 and 2003/04). Because of this low climatological frequency, BMA weights have been computed by regrouping 30-min training data into 3-h time intervals to increase the number of observed LVP in each time interval (0–3, 3–6, 6–9, and 9–12 h).

d. Scores for validation: The Brier score

One of the most common measures of accuracy for verifying two-category probabilistic forecasts is the Brier score (BS; Brier 1950). The BS is used to evaluate an ensemble skill. It is defined as the mean square error of the probability forecast:

$$BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2, \tag{5}$$

where N is the number of forecasts, p_i is the forecast probability, and o_i is the verifying observation (1 if LVP occurs and 0 if it does not). The Brier score can be decomposed into three components: reliability, resolution, and uncertainty (Wilks 2006):

$$BS = BS_{\text{rel}} - BS_{\text{resol}} + BS_{\text{unc}}, \tag{6}$$

where

$$BS_{\text{rel}} = \frac{1}{N} \sum_{k=1}^{T} n_k (p_k - \overline{o}_k)^2, \tag{7}$$

$$BS_{\text{resol}} = \frac{1}{N} \sum_{k=1}^{T} n_k (\overline{o}_k - \overline{o})^2, \quad \text{and} \tag{8}$$

$$BS_{\text{unc}} = \overline{o}(1 - \overline{o}). \tag{9}$$

When a sample of N forecasts has been divided in T categories, each comprising n_k forecasts of probability p_k, o_k is the observed frequency of the forecast being found in that category, and $\overline{o}$ is the observed frequency in the whole sample.

The Brier skill score (BSS) can be defined as

$$\text{BSS} = \frac{\text{BS}_{\text{unc}} - \text{BS}}{\text{BS}_{\text{unc}}}. \tag{10}$$

The Brier skill score is 1 for a perfect forecast and 0 for a climatological forecast. Each component of the BS decomposition is described in Table 3.

5. Results: Validation of the ensembles

a. Skill of the 30-member ensembles

The Brier score and its decomposition into reliability, resolution, and uncertainty are analyzed to assess the ensemble skill (Fig. 3). The uncertainty part of the Brier score is a natural basis for comparisons with the ensemble probability scores (Fig. 3a). The uncalibrated ensemble provides a better forecast than the climatology up to 5 h. Beyond this forecast time, it produces higher errors than the climatology (Fig. 3a). For the

TABLE 3. The meaning of the reliability, the resolution, and the uncertainty components of the Brier score.

Component	Meaning
Reliability (BS_{rel})	The ability of the system to forecast accurate probabilities. Forecast probabilities have to match observed frequencies. For example, an observed LVP frequency of 40% in a sample is expected when a 40% probability is issued by the system. The reliability is negatively oriented like the Brier score (the lower the better).
Resolution (BS_{resol})	The ability of the system to differentiate between the different categories, whatever the probabilities. The resolution is positively oriented (the higher the better).
Uncertainty (BS_{unc})	The variance of observations. It indicates the intrinsic difficulty of forecasting the event and does not depend on the forecast system. Uncertainty is also the probability score of the sample climatology forecast.

decomposed approach (DEEPS), the Brier score is slightly worse than the uncalibrated ensemble. Nevertheless, the forecast performance is better than the climatology forecast up to 4 h. Conversely, the global approach (LEPS) provides a more suitable forecast than a forecast based only on climatological information during the 12-h forecast, and the mean improvement on the Brier score is 20%. For very short term forecasts (between 0 and 3 h), the mean improvement on the Brier score is 46% relative to the climatology (Table 6).

Most of the improvements on the Brier score as a result of the calibration come from the reliability part of the score (Fig. 3b). The BMA calibration is clearly efficient in the global approach but not in the decomposed one. During the 12-h forecast, LEPS' reliability is improved by 46% on average by the BMA calibration, whereas the calibration does not improve DEEPS' reliability, which is on average worsened by 14%.

The calibration is applied to improve the ensemble reliability and is not supposed to negatively influence the ensemble resolution. Figure 3c confirms that resolution is not affected by the calibration, and all the ensembles have about the same resolution.

b. Influence of the BMA calibration on the ensembles

Figure 3b has shown that calibration has successfully improved reliability for LEPS, but it has failed for DEEPS. BMA weights computed in both ensembles are examined in the next subsections to understand these different behaviors. The BMA method evaluates the weights in light of the predictive information from all members.

In two-category forecasts (yes–no, with two values, 1 or 0)—like here for LVP forecasts—the pdf of each member is discrete instead of being continuous like for more usual variables (e.g., Gaussian distribution for temperature). For a classical variable (e.g., temperature), the BMA pdf is a weighted sum of all pdf members; however, in binary cases such as LVP forecasts, the distinction between member solutions becomes impossible when, for example, all members predict a fog. The BMA method can only distinguish the relative frequency of each member to give the correct forecasts for the training data for each forecast period. Members that give similar information are ignored and are typically assigned weights equal to zero in a well-sampled ensemble with a wide range of parameters perturbed. But the strengths of the BMA method are the efficiency of the weight computation algorithm and its simplicity in the case of binary forecasts. The BMA is not a complex method in the case of LVP prediction. Because we are predicting a binary variable (LVP or no LVP; 1 or 0), the BMA calibration problem is simplified and reduced to the computation of appropriate weights for each ensemble member, so the BMA is a quite convenient and efficient in this context.

However, the BMA calibration can be affected by both "overfitting" the data and colinearity between members in the training data (Wilson et al. 2007). Overfitting occurs when the training data sample is too small; it damages the relationship between independent data despite improving the fit relationship with the training data. Colinearity of members occurs when ensemble members are not independent in the training data and leads to the failure of the ensemble sampling. Having a too-small dataset induces colinearity, which leads to de-weighting and to the exclusion of information from members (Hamill 2007).

1) BMA WEIGHTS FOR THE GLOBAL LEPS

Figure 4a shows the BMA weights more than 5% in the global approach. Only 8 of the 30 members have a contribution more than 5% on at least one of the four 3-h periods. The BMA selects among the 30 members the best members in the reliability meaning and split the weight between these members.

It is also shown that weights evolve with the forecast time. This result is in agreement with the conclusions of our prior study on LVP forecast sensitivity: initial condition members related to fog/stratus initialization prevail (the maximum weight contribution is approximately 42%) during the first 6 h and decrease rapidly afterward. The weight of the mesoscale forcing mem-

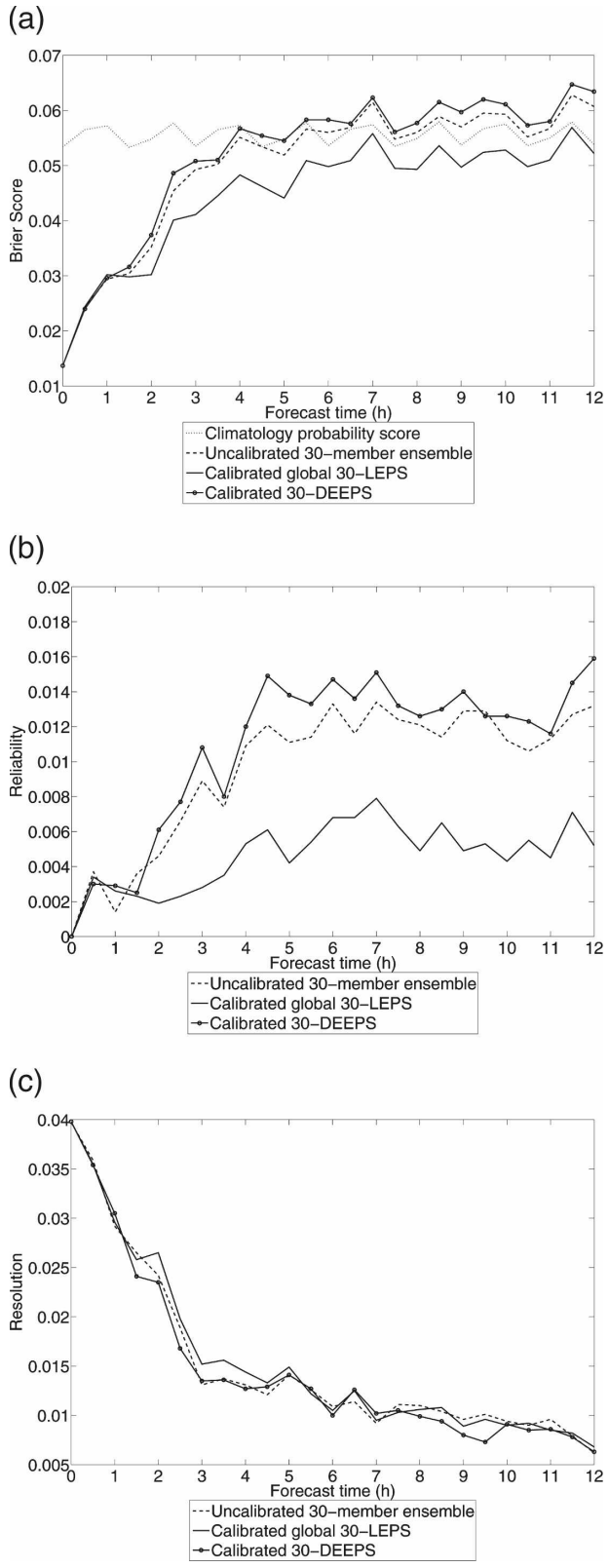

FIG. 3. (a) Brier score, (b) reliability, and (c) resolution of the 30-member ensembles for winter season 2004/05.

bers—especially humidity advections—increases during the run and becomes dominant between 6 and 12 h (the maximum weight contribution is approximately 52%). Cloud cover members act throughout the 12-h run, and initialization and mesoscale forcing members balance each other during run.

Most of the weights are nil, revealing the colinearity between some ensemble members. However, in the global approach the BMA calibration still manages to significantly improve the reliability (46% on average during the 12-h forecast period). Thanks to the variety of physical processes perturbed in the ensemble, some members can have totally different trajectories around the reference trajectory, and the BMA calibration is efficient. Consequently, the ensemble is relatively well sampled.

2) BMA WEIGHTS FOR DEEPS

Only the BMA weights supplied by the 12-member mesoscale forcing subensemble are shown (Fig. 4b), but conclusions are valid for the three other blocks. Almost all members in the block have nonnegative weights and contribute to the BMA weights, and the relative contributions of the members are lower than those in LEPS. Seven out of the 12 members have a weight more than 5%. The BMA assigns a weight to each member of the block, but they remain quasi-constant in time. This behavior is a result of the colinearity of members, in that each subensemble is composed of members with the same physical parameter perturbed; however, the members are still too dependent. Consequently, weights are lower (the weight contribution is less than 30%) than the main weights in the LEPS calibration, and they are also quasi constant during the run.

The training data sample is too small to recalibrate the four subensembles together to balance parameter block effects, like in the global LEPS when initial conditions and mesoscale forcings balance each other. Consequently, the calibration is applied within each block and afterward the blocks are assumed equiprobable. Each subensemble suffers from too much dependency of its inner members and, therefore, calibration in all the subensembles fails to improve the basic skill of the uncalibrated ensemble.

Individual scores of each subensemble are analyzed in Fig. 5 to help to understand each block's contribution. First, the mesoscale forcing block presents a Brier score comparable with the climatology uncertainty and the uncalibrated ensemble Brier score between 3 and 12 h. All other Brier scores are higher than the climatology uncertainty past the third forecast hour (Fig. 5a). Second, resolutions vary significantly after 2 h of simu-

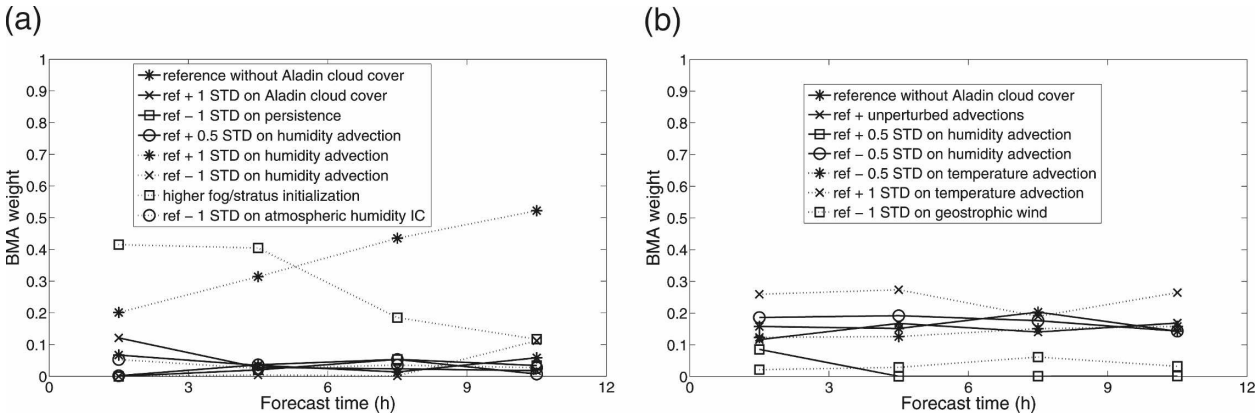

FIG. 4. BMA weights for both 30-member ensembles: (a) the global ensemble and (b) the mesoscale forcing block of the decomposed ensemble. Only the members with a weight contribution more than 0.05 (5%) are shown.

lation. Depending on the perturbed parameter (or perturbed physical process), a block may capture the event because of the decomposition strategy. Consequently, subensembles have less potential regarding the resolution than the global LEPS (Fig. 5c). Resolutions in subensembles are lower than in the LEPS, within a range of 30%.

c. *Ensemble sampling sensitivity*

To test this sampling representativity, a 54-member ensemble was built by adding perturbations up to two standard deviations (Tables 4 and 5) to the original 30-member ensemble. Tables 6–8 summarize the skill of the global LEPS for LVP prediction at Charles de Gaulle airport. Mean scores for the 12-h forecast period as well as for the first 3 h of the forecast are presented because high quality, very short term forecast are of particular interest to airport management.

Overall, the 30- and 54-member LEPS have the same skill. As a result, there is almost no improvement on the Brier score between both LEPS (Table 6)—either for very short periods or for during the 12-h forecast period. However, improvements appear with the decomposition of the Brier score.

First, for the very short term forecast (0–3 h), the calibration in the 54-member ensemble leads to a 41% improvement in reliability, whereas the calibration in the 30-member ensemble improves reliability by 25%. This result suggests a better sampling of the 54-member ensemble. The uncertainties seem to be more representative in the 54-member ensemble than in the 30-member ensemble, and the calibration leads to better reliability. Actually, the uncertainties of the atmospheric profiles were underestimated in the 30-member ensemble; it has been observed with the better results of the initial conditions subensembles for the 54-member DEEPS compared to the 30-member DEEPS (not shown).

Second, the calibration leads to a better resolution in the 54-member ensemble, especially for longer forecast times (9–12 h) when there is an improvement of 18%. Notice that the calibration has almost no impact (2%) on the resolution in the 30-member ensemble. This impact of the BMA calibration on the resolution of the 54-member ensemble could be explained by the fact that the dataset is too small for the validation of the 54-member ensemble. The same dataset is used to validate both the 30- and 54-member ensembles. If this dataset is sufficient to validate the 30-member ensemble, the dataset appears to be inappropriate for the validation of the 54-member ensemble. The larger the ensemble is, the larger the validation data needs to be. However, for both calibrated ensembles and longer-term forecasts (9–12 h), the 54-member LEPS improves by 29% relative to the 30-member LEPS. This result also confirms a better sampling representativeness in the 54-member ensemble. As a result, the 54-member ensemble is able to capture more events than the 30-member ensemble.

Because of its better sampling representativeness, the 54-member ensemble will be analyzed in the next sections to demonstrate its operational capabilities and its benefits relative to deterministic forecasts. Despite the negative consequences of using a small training data sample, the global LEPS ensemble methodology has been successful, and the BMA calibration has improved the basic skill of the uncalibrated ensembles. However, because we are dealing with the prediction of a rare event, the robustness of the results will have to be confirmed in the future using larger data samples—especially for longer forecast times (6–12 h) when few forecasts are issued with high probabilities.

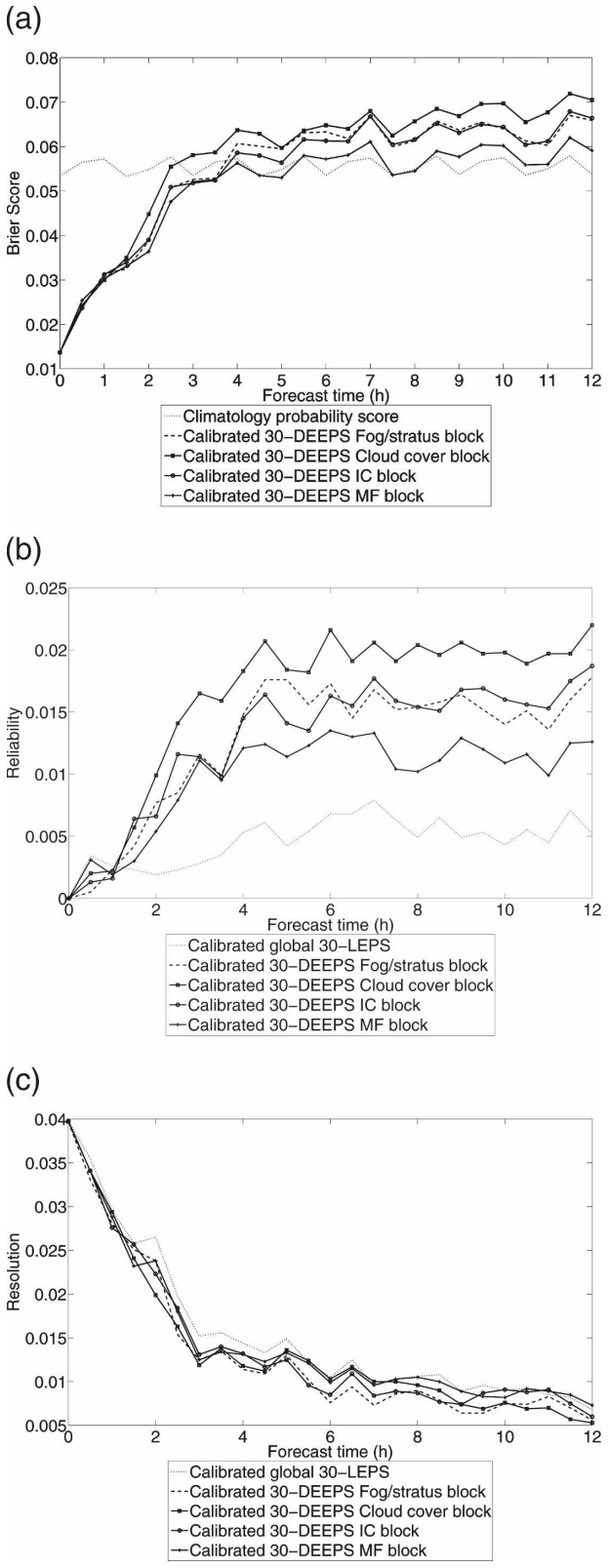

FIG. 5. (a) Brier score, (b) reliability, and (c) resolution of the four DEEPS subensembles for winter season 2004–2005 [fog/stratus, cloud cover, initial condition (IC), and mesoscale forcing (MF) blocks].

TABLE 4. As in Table 1 but for the 54-member ensembles LEPS and DEEPS.

	Ensembles	
Member	Global	Decomposed (blocks)
1—reference member	Ref with Aladin cloud cover	Included in each block
2–7	Same as members 2–7 in Table 1	Cloud cover block
8	Ref + 2 STD on Aladin cloud cover	Cloud cover block
9	Ref − 2 STD on Aladin cloud cover	Cloud cover block
10	Ref + 2 STD on persistence	Cloud cover block
11	Ref − 2 STD on persistence	Cloud cover block
12–22	Same as members 8–18 in Table 1	MF block
23	Ref + 2 STD on humidity advections	MF block
24	Ref − 2 STD on humidity advections	MF block
25	Ref + 2 STD on temperature advections	MF block
26	Ref − 2 STD on temperature advections	MF block
	Cloud cover subensemble: 11 members	
	MF subensemble:18 members	

6. LEPS operational characteristics and economic value

a. The 54-member LEPS versus the reference deterministic forecast

In this subsection, the 54-member LEPS is compared to the reference deterministic forecast to quantify the advantage of using an ensemble probabilistic forecast. Figure 6 compares the Brier score, the reliability and the resolution of the 54-member LEPS, and the reference deterministic forecast. The Brier score is not particularly adapted for the evaluation of deterministic forecasts; however, a deterministic forecast can be viewed as a probabilistic forecast with only two categories: 0 and 100%. The interpretation of the scores should take into account the different nature between deterministic and probabilistic forecasts. The Brier score results indicate that the 54-member LEPS is better than the reference by 8% on average during the 12-h forecast period (Fig. 6a). For very short-term forecasts (0–3 h), the improvement is 20% on average.

Figure 6b (reliability) and Fig. 6c (resolution) have to be analyzed together because of the different nature between deterministic and probabilistic forecasts. Be-

TABLE 5. As in Table 2 but for the 54-member ensembles LEPS and DEEPS.

Member	Ensembles	
	Global	Decomposed
27–36	Same as members 19–26 in Table 2	IC block
37	Ref − 2 STD on humidity atmospheric profile	IC block
38	Ref + STD on temperature atmospheric profile	IC block
39	Ref − 2 STD on temperature atmospheric profile	IC block
40	Ref + STD on humidity soil profile	IC block
41	Ref − 2 STD on humidity soil profile	IC block
42	Ref + STD on temperature soil profile	IC block
43	Ref − 2 STD on temperature soil profile	IC block
44–48	Same as members 27–30 in Table 2	Fog/stratus IC block
49	Cloud top + 1 vertical grid point	Fog/stratus IC block
50	Cloud top − 1 vertical grid point	Fog/stratus IC block
51	Cloud top + 2 vertical grid point	Fog/stratus IC block
52	Cloud top − 2 vertical grid point	Fog/stratus IC block
53	Cloud base − 1 vertical grid point	Fog/stratus IC block
54	Cloud base − 2 vertical grid point	Fog/stratus IC block
	IC subensemble: 17 members Fog/stratus subensemble: 11 members	

cause the ensemble is designed to forecast the LVP likelihood, the ensemble probabilistic system should have a better resolution than deterministic forecasts to be valuable. This prediction is shown in Fig. 6c, and LEPS resolution is superior to the reference by 93% on average for the 12-h forecast period. Consequently, LEPS is able to detect significantly more cases than the reference deterministic forecast, especially for longer-term forecasts (9–12 h) in which the mean improvement is 189%. However, because of their probabilistic nature, LEPS forecasts are less reliable than the reference forecast, with reliability worsening 210% on average during the 12-h forecast period. First, this result can be explained by the fact that LEPS forecasts more fog cases and therefore has a higher risk of producing false alarms. Second, a good reliability score is obtained when all the probability categories match the forecast probabilities. This requirement is easier to reach for the deterministic forecast, leading to a better reliability than LEPS. However, the improvement in resolution dominates the overall score, showing that 54-member LEPS forecasts have the potential to forecast more cases than deterministic forecasts.

TABLE 6. Summary of the results for the Brier score: percentage of improvement/damage [computed as $(BS_{ensemble} - BS_{baseline})/BS_{baseline}$] between ensembles or an ensemble and the uncertainty. Negative values correspond to an improvement and positive values correspond to damage to the quality of the studied ensemble.

Brier score	Percentage of improvement/damage (%)	
	Between 0 and 3 h	All forecast periods between 0 and 12 h
30-member uncalibrated ensemble vs uncertainty	−41	−10
54-member uncalibrated ensemble vs uncertainty	−42	−13
30-member LEPS vs uncertainty	−46	−20
54-member LEPS vs uncertainty	−46	−20
30-member LEPS vs 30-member uncalibrated ensemble	−6	−10
54-member LEPS vs 54-member uncalibrated ensemble	−4	−7
54-member uncalibrated vs 30-member uncalibrated	−2	3
54-member LEPS vs 30-member LEPS	0	0

b. *The LVP forecast efficiency: Pseudo relative operating characteristics curve*

The pseudo ROC curves provide an efficient way of representing the quality of dichotomous, categorical,

TABLE 7. Summary of the results for the reliability part of the Brier score: percentage of improvement/damage [computed as $(X_{ensemble} - X_{baseline})/X_{baseline}$, where X represents the reliability score from the Brier score's decomposition] between ensembles. Negative values correspond to an improvement and positive values correspond to damage to the quality of the studied ensemble.

Reliability	Percentage of improvement/damage (%)	
	Between 0 and 3 h	All forecast periods between 0 and 12 h
30-member LEPS vs 30-member uncalibrated ensemble	−25	−46
54-member LEPS vs 54-member uncalibrated ensemble	−41	−39
54-member uncalibrated vs 30-member uncalibrated	−6	−8
54-member LEPS vs 30-member LEPS	−8	−10

TABLE 8. As in Table 7 but for the resolution part of the Brier score.

Resolution	Percentage of improvement/damage (%)	
	All forecast periods between 0 and 12 h	From 9 to 12 h
30-member LEPS vs 30-member uncalibrated ensemble	2	−2
54-member LEPS vs 54-member uncalibrated ensemble	6	18
54-member uncalibrated vs 30-member uncalibrated	3	7
54-member LEPS vs 30-member LEPS	7	29

and probabilistic forecasts. The method is based on ratios that measure the proportions of LVP events and nonevents for which warnings were provided. It evaluates the skill of the forecast system by comparing the hit rate (HR) and the pseudo false-alarm rate (pseudo FAR) of LVP events for different thresholds.

For this study, four thresholds have been chosen: $P > 90\%$, $P > 50\%$, $P > 20\%$, and $P > 0\%$. The pseudo FAR is computed as the ratio of forecast and unobserved cases to LVP forecast cases. This calculation removes the influence of the "no–no good forecasts" (no LVP forecast and no LVP observed), which mostly dominates the data sample for rare events and hides the true skill of the LVP forecast system. The HR is computed as the ratio of forecast and observed cases to LVP observed cases. When defining HR and pseudo FAR (also called false-alarm ratio) in Eqs. (11) and (12), a is the number of observed and forecast events, b is the number of not observed and forecast events, and c is the number of observed and not forecast events:

$$\mathrm{HR} = \frac{a}{a+c} \quad \text{and} \tag{11}$$

$$\text{pseudo FAR} = \frac{b}{a+b}. \tag{12}$$

Figure 7 also shows the advantage of using probabilistic forecasts instead of the reference deterministic forecasts, which lie on or below the curves for both the 54-member uncalibrated ensemble and the 54-member LEPS, at any forecast time. Users can chose between two options according to their needs. Either users decide to take protective measures for low probability thresholds (like $P > 20\%$)—this is possible because the probabilistic forecast has higher detection capabilities

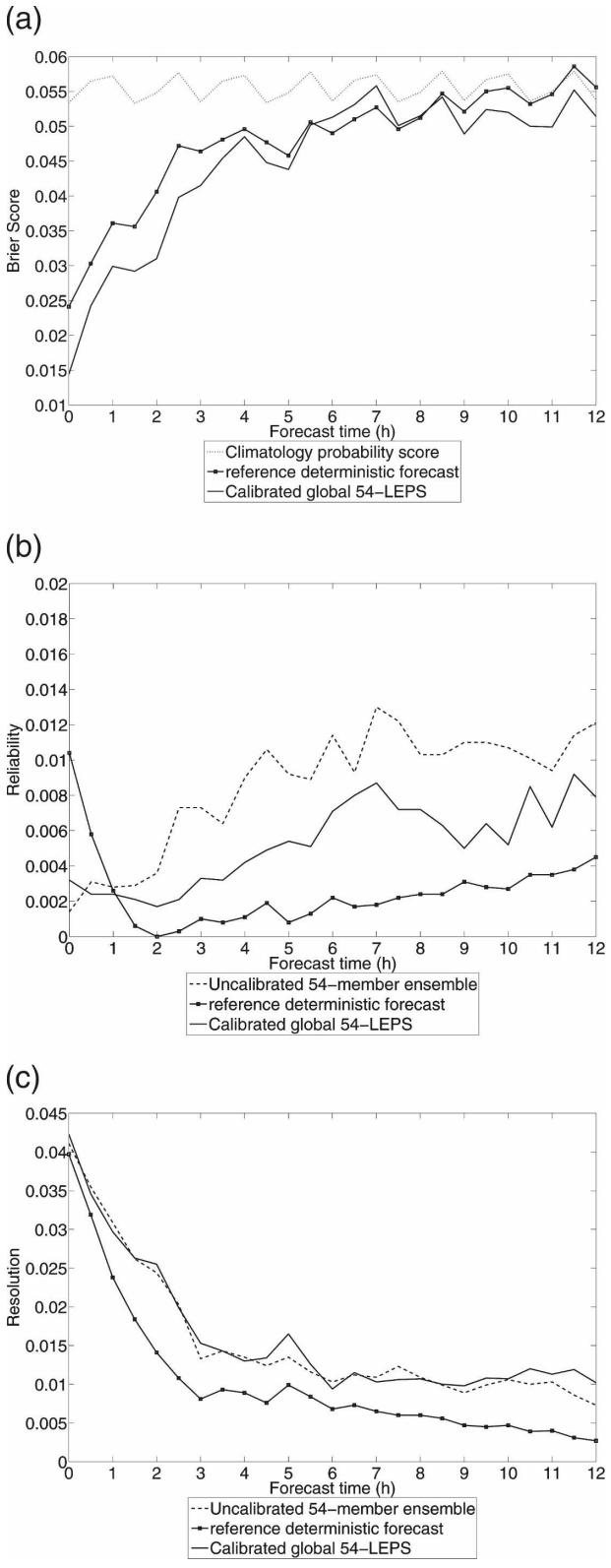

FIG. 6. (a) Brier score, (b) reliability, and (c) resolution comparisons between the 54-member global LEPS calibrated and uncalibrated ensembles and the reference deterministic forecast for winter season 2004/05.

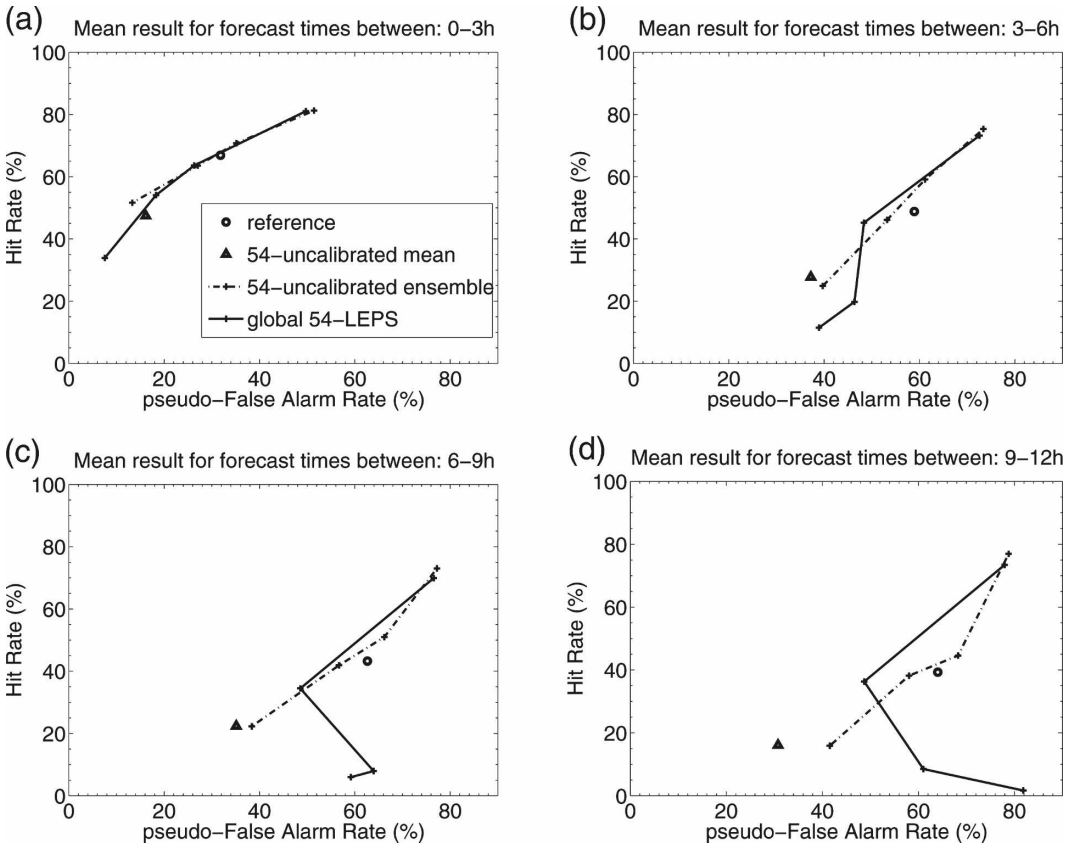

FIG. 7. Pseudo ROC curve during the 12-h forecast for the 54-member uncalibrated ensemble, the 54-member LEPS, the mean (HR and pseudo FAR for the mean cloud base height result) of the 54-member uncalibrated ensemble, and the reference deterministic forecast for (a) 0–3, (b) 3–6, (c) 6–9, and (d) 9–12 h for winter season 2004/05. Four probability thresholds are considered: $P > 90\%$, $P > 50\%$, $P > 20\%$, and $P > 0\%$. The mean ensemble solution is the ensemble mean value of ceiling; the visibility aspect of the LVP variable is represented by a value of ceiling of 0 m.

than the deterministic one but with more false alarms—or users can implement measures only when they have confidence in the forecast with high probability thresholds (like $P > 90\%$). This strategy is also possible because the probabilistic forecasts can provide high probabilities with significant hit rates but also provide fewer false alarms than the deterministic forecast between 0 and 6 h of the forecast. In a representative ensemble (with an adequate sampling), the mean solution of the ensemble eliminates the unpredictable components in the ensemble and preserves the predictable components. The mean ensemble solution is the ensemble mean value of ceiling; the visibility aspect of the LVP variable is represented by a value of ceiling of 0 m. Figure 7 shows that the ensemble mean always has a smaller HR but also a much smaller pseudo FAR than the reference forecast. Therefore, the ensemble mean is a more reliable deterministic forecast than the reference forecast. The pseudo ROC curves display the mean statistics of LEPS during the 2004/05 winter season for the eight daily runs (0000, 0300, 0600, 0900, 1200, 1500, 1800, and 2100 UTC). The statistics are presented in Fig. 7 according to four forecast time periods (0–3, 3–6, 6–9, and 9–12 h) on the pseudo ROC curves.

c. *LEPS economic value: A simple cost–loss decision model*

A simple cost–loss model can be applied to this context of local LVP probabilistic forecasts (Zhu et al. 2002; Richardson 2003). Consider a user or decision maker whose activities are sensitive to LVP forecasts. If the event occurs and the user has not taken any preventive action, then the user suffers a financial loss L, or the user could take action at a cost C, that could protect against this potential loss. It is shown in the appendix that the relative economic value is calculated as follows:

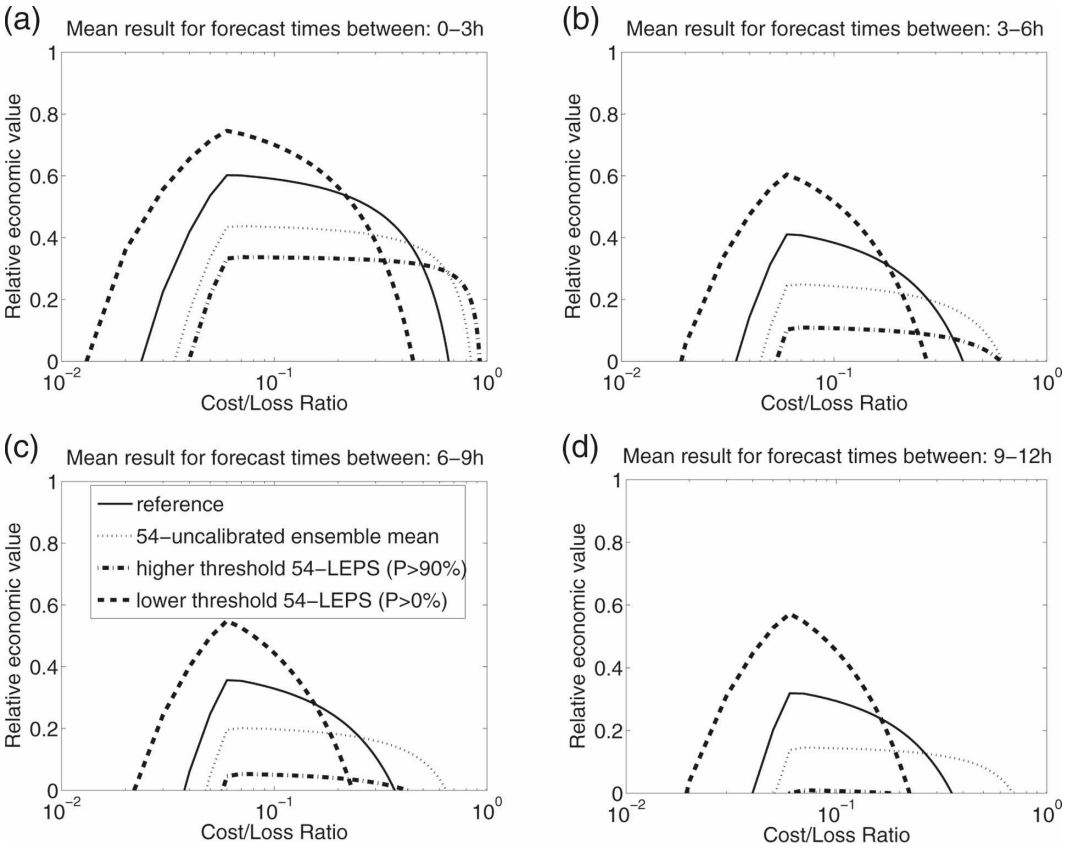

FIG. 8. Relative economic value during the 12-h forecast for the mean 54-member ensemble result, the $P > 90\%$ and $P > 0\%$ thresholds, and the reference member between (a) 0–3, (b) 3–6, (c) 6–9, and (d) 9–12 h for winter season 2004/05.

$$V(\alpha) = \frac{\min(\alpha, \bar{o}) - \alpha\bar{p}(\text{pseudo FAR}) + \bar{o}(1-\alpha)\text{HR} - \bar{o}}{\min(\alpha, \bar{o}) - \bar{o}\alpha}, \tag{13}$$

where α is the cost–loss ratio of any user, $\bar{o} = a + c$ is the fraction when LVP occurs (climatology), and $\bar{p} = a + b$ is the fraction of forecast LVP events.

Consequently, the economic value depends on the user cost–loss ratio. Figure 8 shows a wide range of users can benefit from the 54-member LEPS; however, the maximum saving is expected for users with low cost–loss ratios and with the lowest probability threshold.

The highest economic value is reached for users whose cost–loss ratio equals the climatology frequency of LVP events (about 0.06 or 6%)—between 0–3 h savings are up to 78% (threshold $P > 0\%$) and decrease to reach 58% during the 9–12 h forecast period. The mean ensemble forecast has a better skill and thus a better economic value than the highest probability threshold ($P > 90\%$), especially during the second part of simulation (between 6–12 h).

In the aeronautic context, losses are likely be much larger than the cost of taking protective precautions, so these users have an economic value (V) associated with low cost–loss ratios (α). These users take advantage of higher HR and can tolerate larger pseudo FAR. LEPS is worthwhile for users in the aeronautic sector who could rely on this probabilistic forecast system to manage the airport traffic and take appropriate actions according to the LVP likelihood (managing flight delays or cancellations; increasing number of staff, mostly for safety considerations; and increasing time and space interval security for takeoff and landing, among others).

7. Conclusions

The use of a local forecast system is well adapted for the prediction of events with small space and time scales such as low-cloud ceiling and visibility (Bergot 2007). Predictability of this type of sensitive events is not at all straightforward, and ensemble prediction was shown to be able to provide valuable estimates of the forecast skill as well as supplying confidence indices on

the forecasts. Probabilistic forecasts can help authorities in their daily decision-making responsibilities toward maintaining a high level of air traffic safety and cost efficiency. One way commonly used to produce probabilistic forecasts is through an ensemble approach. Two strategies have been tested at Charles de Gaulle International Airport for the prediction of LVP events: the global LEPS approach and the decomposed approach (DEEPS). For LEPS, all members are grouped into a single ensemble, and the calibration method is applied jointly on all members. For DEEPS, the members are split into "physically consistent" subensembles, and the calibration is applied in each subensemble. As with the LEPS, DEEPS provides the LVP pdf, but in addition the sources of uncertainty in the forecast can in principle be diagnosed by identifying the physical drivers of a particular event.

However, the decomposed EPS was not conclusive. The BMA calibration has not improved the reliability in the decomposed approach, with values remaining close to those of the uncalibrated ensemble. Increasing the amplitude of the perturbations and the ensemble size has led to improvements in reliability and resolution for DEEPS. But, these improvements were not sufficient to reach the 30-member LEPS skill.

The lack of training and verification data has been prejudicial in the decomposed approach. The decomposed approach, with the subensemble strategy, clearly requires more data to induce a forecast time dependency of the BMA weights within each subensemble and therefore permitting a useful recalibration of the four subensembles. The main weakness of the DEEPS is that the subensembles suffer from too much dependency of its inner members, and therefore the calibration in all the subensembles fails to improve the basic skill of the uncalibrated ensemble. Thus, the construction of the subensembles has to be reviewed and reconsidered to improve the skill of DEEPS.

On the other hand, the global LEPS is successful thanks to the BMA calibration. The BMA weights evolve with the forecast time, with initial condition members prevailing between 0 and 6 h and mesoscale forcing members prevailing between 6 and 12 h. A balance between the ensemble members is reached, which improves LEPS' reliability. LEPS' resolution is equivalent to the uncalibrated ensemble resolution because resolution is not affected by the calibration method in the 30-member ensemble. However, the calibration has influenced the resolution of the 54-member ensemble because of a too-small dataset for an appropriate validation of this ensemble.

Results show that the resolution is improved by the increase of the ensemble size. The ensemble sampling in the 54-member ensemble has a better representation of the uncertainties than the 30-member ensemble. Consequently, the 54-member LEPS has a higher resolution for longer-term forecasts than the 30-member LEPS. As the forecast time increases, the ensemble spreads out more and LVP events are forecast more frequently but with low probabilities. After 3 h of simulation, few LVP cases are forecast with probabilities more than 50%. Therefore, the calibration method becomes less efficient for high probability categories than small probability categories, which is a consequence of a lack of data. The calibration and the validation data samples are too small; small samples lead to the overfitting of the training data and colinearity between ensemble members, which affect the ensemble by deweighting and excluding information from members.

For very short term forecasts (0–3 h), the improvement in LEPS reliability is much better in the 54-member LEPS than in the 30-member LEPS. This result also confirms a better sampling of the uncertainties in the 54-member ensemble. The DEEPS subensembles analysis has been very useful in understanding the sampling representativeness point.

The advantage of using the LEPS probabilistic forecasts rather than the reference deterministic forecasts has been shown in the improvement in the resolution. LEPS has a much higher resolution than the reference deterministic forecast, thus it can detect more LVP cases.

Current operational LVP forecasts are accurate up to 3 h. Beyond this forecast horizon, the use of these operational forecasts leads to too many false alarms and they become useless (Bergot 2007). LEPS extends the limit of LVP predictability up to 12 h. The system is particularly reliable for very short term forecasts (0–3 h). Its ability remains significant up to 6 h for both reliability and resolution, and its potential for LVP detection is significant up to 12 h.

The quality of LEPS forecasts is very appreciable; the probabilistic information is reliable and complements the single deterministic forecast by adding a confidence index on the reference run. At all forecast times, complementary information from LEPS can be obtained depending on the user's interests. A simple cost–loss decision model has been applied to the 54-member LEPS, and it is shown that users with low cost–loss ratios can expect significant benefits from the system. For these users, it is important to avoid failings. They can tolerate more false alarms because the cost of taking protection measures is much smaller than the cost of losses caused by the occurrence of the event. Users can expect a maximum savings up to 78%, 60%,

58%, and 57% for the forecast time periods of 0–3, 3–6, 6–9, and 9–12 h, respectively.

This local probabilistic system has been designed with an operational purpose in mind. In addition to LEPS' forecast skill, some motivating strengths regarding more practical operational considerations have to be highlighted.

One-dimensional models are easy and relatively inexpensive to run. LEPS' results can rapidly be obtained in an operational environment on a personal computer (PC). The fact that numerous 1D forecasts can be obtained in a timely manner is essential because high-frequency runs are required for insightful very short-term probabilistic forecasts of LVP.

The methodology can be easily adapted to other airports. The strategy employed consists of three steps: the evaluation of the distributions of the uncertainty sources, the sampling of the ensemble, and the BMA calibration.

The BMA calibration procedure is efficient. New calibration weights have to be computed for each new location, and site observations are required for these computations. On-site, dedicated observations are also highly recommended for the 1D model initialization to obtain a finer vertical description of the boundary layer profiles and accurate forecasts (Bergot et al. 2005). The adaptability of the local ensemble prediction and its computational efficiency will always be an advantage compared to the important computational resources required to run 3D numerical weather prediction models for the local prediction purpose.

We have demonstrated that the LEPS has great forecast skill for short and very short-term forecasts of LVP conditions at Charles de Gaulle International Airport. Consequently, the system is well adapted and easy to run for this operational purpose.

Acknowledgments. The authors thank the Météo-France personnel involved in this research study for their help and fruitful comments. We also thank all of the people who have been involved in the COBEL and the ISBA model developments during the past decade. Without their hard work, this study would not have been possible.

APPENDIX

Simple Cost–Loss Decision Model

A simple cost–loss model can be applied to the present context of local LVP probabilistic forecasts (Zhu et al. 2002; Richardson 2003). Consider a user or decision maker whose activities are sensitive to LVP forecasts. If the event occurs and the user has not taken any preventive action, then the user suffers a financial loss L. Instead, the user could take action at a cost C that would protect against this potential loss. The costs and losses of all combinations of action and outcomes are described in Tables A1 and A2. The goal of the user is to minimize overall expense by deciding on which situations to protect against.

TABLE A1. Costs and losses associated with different actions and outcomes in the cost–loss model.

	Outcome	
Action taken	Yes	No
Yes	C	C
No	L	0

In a large number of cases, $\bar{o} = a + c$ is the fraction when LVP occurs in the sample (climatology), and $\bar{p} = a + b$ is the fraction of LVP events forecast. If the user always takes preventative action, then the cost will be C on every occasion and the average expense (per situation) will be

$$E_{\text{protect}} = C.$$

Alternatively, if the user never takes preventive actions, then the loss will only be incurred when LVP occurs and the average expense in this case will be

$$E_{\text{loss}} = \bar{o}L.$$

Assuming that the user knows only the climatological frequency of LVP $\bar{o}$, the optimal strategy is either always protect or never protect, depending on the strategy that gives the lower overall expenses. This is a baseline against which improvements gained by using the forecast information (reference run, ensemble mean, and probability thresholds) can be compared and evaluated. This is the climatological expense:

$$E_{\text{climatology}} = \min(C, \bar{o}L).$$

Another useful reference point is provided by the expense associated with perfect forecast information, which is obtained when the user only protects if LVP occurs,

$$E_{\text{perfect}} = \bar{o}C.$$

TABLE A2. Contingency for deterministic forecast of specified event of a set of cases showing fraction of occasions for each combination of forecast and outcome.

	Outcome	
Event forecast	Yes	No
Yes	a	b
No	c	d

The average expense of a deterministic forecast or a probabilistic forecast (using a probability threshold) is obtained by multiplying the corresponding cells in Tables A1 and A2,

$$E_{\text{forecast}} = aC + bC + cL.$$

The difference between $E_{\text{climatology}}$ and E_{forecast} measures the economic savings of the user when the forecast system is used relative to only having climatological information. The relative economic value (V) is defined by comparing this savings with the maximum possible savings that can be made with perfect deterministic forecasts:

$$V = \frac{E_{\text{climatology}} - E_{\text{forecast}}}{E_{\text{climatology}} - E_{\text{perfect}}}.$$

Replacing each expense by expression, the relative economic value is

$$V(\alpha) = \frac{\min(\alpha, \bar{o}) - \alpha\bar{p}(\text{pseudo FAR}) + \bar{o}(1 - \alpha)\text{HR} - \bar{o}}{\min(\alpha, \bar{o}) - \bar{o}\alpha},$$

where α is the cost–loss ratio of any user, HR is the hit rate and pseudo FAR is the pseudo false-alarm rate.

REFERENCES

Bergot, T., 1993: Modélisation du brouillard à l'aide d'un modèle 1D forcé par des champs mésoéchelle: Application à la prevision (Numerical simulation of fog with a 1D model forced by mesoscale parameters: Forecasting application). Ph.D. thesis, Université Paul Sabatier, No. 1546, 192 pp. [Available from Centre National de Recherches Meteorologiques, 42 Avenue Gapard Coriolis, F-31057 Toulouse Cedex, France.]

——, 2007: Quality assessment of the Cobel-ISBA numerical forecast system of fog and low clouds. *Pure Appl. Geophys.,* **164,** 1265–1282.

——, and D. Guédalia, 1994: Numerical forecasting of radiation fog. Part I: Numerical model and sensitivity tests. *Mon. Wea. Rev.,* **122,** 1218–1230.

——, D. Carrer, J. Noilhan, and P. Bougeault, 2005: Improved site-specific numerical prediction of fog and low clouds: A feasibility study. *Wea. Forecasting,* **20,** 627–646.

Boone, A., 2000: Modélisation des processus hydrologiques dans le schéma de surface ISBA: Inclusion d'un réservoir hydrologique, du gel et la modélisation de la neige (Modeling hydrological processes in the land-surface scheme ISBA: Inclusion of a hydrological reservoir, incorporation of soil ice and snow modeling). Ph.D. thesis, Université Paul Sabatier, 252 pp.

——, V. Masson, T. Meyers, and J. Noilhan, 2000: The influence of the inclusion of soil freezing on simulations by a soil–vegetation–atmosphere transfer scheme. *J. Appl. Meteor.,* **39,** 1544–1569.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78,** 1–3.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.,* **125,** 99–119.

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

Clark, D. A., 2002: The 2001 demonstration of automated cloud forecast guidance products for San Francisco International Airport. Preprints, *10th Conf. on Aviation, Range, and Aerospace Meteorology,* Portland, OR, Amer. Meteor. Soc., JP1.26. [Available online at http://ams.confex.com/ams/13ac10av/techprogram/paper_38862.htm.]

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.,* **39B,** 1–38.

Ehrendorfer, M., 1994: The Liouville equation and its potential usefulness for the prediction of forecast skill. Part I: Theory. *Mon. Wea. Rev.,* **122,** 703–713.

Fisher, R. A., 1922: On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London,* **222A,** 309–368.

Hamill, T. M., 2007: Comments on "Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging." *Mon. Wea. Rev.,* **135,** 4226–4230.

Herzegh, P., K. Petty, S. Benjamin, R. Rassmussen, T. Tsui, G. Wiener, and P. Zwack, 2002: Development of automated national ceiling and visibility products: Scientific and practical challenges, research strategies, and first steps. Preprints, *10th Conf. on Aviation, Range, and Aerospace Meteorology,* Portland, OR, Amer. Meteor. Soc., 3.2. [Available online at http://ams.confex.com/ams/pdfpapers/40395.pdf.]

Houtekamer, P. L., and L. Lefaivre, 1997: Using ensemble forecasts for model validation. *Mon. Wea. Rev.,* **125,** 2416–2426.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

——, 1969: The predictability of a flow which possesses many scales motion. *Tellus,* **21,** 289–307.

MacLachlan, G. J., and T. Krishnan, 1997: *The EM Algorithm and Extensions.* John Wiley and Sons, 274 pp.

Murphy, A. H., 1991: Probabilities, odds, and forecasts of rare events. *Wea. Forecasting,* **6,** 302–307.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174.

Richardson, D. S., 2003: Economic value and skill. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 240 pp.

Roquelaure, S., and T. Bergot, 2007: Seasonal sensitivity on COBEL-ISBA local forecast system for fog and low clouds. *Pure Appl. Geophys.,* **164,** 1283–1301.

Stessel, J. P., L. Frappez, and T. Bergot, 2000: Méthode interactive de prévision des brouillards denses: Définition et test de faisabilité (Interactive method of dense fog prediction: Definition and feasibility test). Institut Royal Météorologique de Belgique Publication Scientifique et Technique 12, 25 pp.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 91, Academic Press, 627 pp.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 1364–1385.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.,* **83,** 73–82.