On the Impact of Short-Range Meteorological Forecasts for Ensemble Streamflow Predictions

Guillaume Thirel, Fabienne Rousset-Regimbeau, Eric Martin, Florence Habets

On the Impact of Short-Range Meteorological Forecasts for Ensemble Streamflow Predictions

GUILLAUME THIREL
CNRM–GAME, Météo-France, CNRS, Toulouse, France

FABIENNE ROUSSET-REGIMBEAU
Direction de la Climatologie, Météo-France, Toulouse, France

ERIC MARTIN
CNRM–GAME, Météo-France, CNRS, Toulouse, France

FLORENCE HABETS
UMR–SISYPHE, ENSMP, CNRS, Fontainebleau, France

(Manuscript received 16 August 2007, in final form 18 March 2008)

ABSTRACT

Ensemble streamflow prediction systems are emerging in the international scientific community in order to better assess hydrologic threats. Two ensemble streamflow prediction systems (ESPSs) were set up at Météo-France using ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System for the first one, and from the Prévision d'Ensemble Action de Recherche Petite Echelle Grande Echelle (PEARP) ensemble prediction system of Météo-France for the second. This paper presents the evaluation of their capacities to better anticipate severe hydrological events and more generally to estimate the quality of both ESPSs on their globality. The two ensemble predictions were used as input for the same hydrometeorological model. The skills of both ensemble streamflow prediction systems were evaluated over all of France for the precipitation input and streamflow prediction during a 569-day period and for a 2-day short-range scale. The ensemble streamflow prediction system based on the PEARP data was the best for floods and small basins, and the ensemble streamflow prediction system based on the ECMWF data seemed the best adapted for low flows and large basins.

1. Introduction

The use of ensemble techniques for numerical weather prediction is now well developed. Ensemble prediction systems, based on a finite number of deterministic integrations, are used to predict an appropriate density function for the meteorological variables. The major meteorological centers integrated these techniques in the 1990s (Tracton and Kalnay 1993; Molteni et al. 1996). In hydrology, several studies have shown promise for using meteorological ensemble prediction to produce probabilistic streamflow forecasts. Roulin and Vannitsem (2005) tested the use of the Ensemble Prediction System (EPS) of the European Centre for Medium-Range Weather Forecasts (ECMWF) for two Belgian catchments. In Europe, the European Flood Alert System (EFAS) prototype (Ramos et al. 2007) uses the ECMWF EPS to predict floods at the European scale. The Hydrologic Ensemble Prediction Experiment (HEPEX) (information online at http://hydis8.eng.uci.edu/hepex/) brings together meteorologists and hydrologists to address the issue of hydrologic forecast uncertainty, including uncertainty in the meteorological forcing, the hydrological modeling, as well as the final user needs. In the United States, several studies have lead to the advancement of ensemble hydrologic forecasting (Wood et al. 2005; Schaake et al. 2007). In France, Rousset-Regimbeau et al. (2007; following Habets et al. 2004) used the ECMWF EPS to build an ensemble streamflow prediction system (ESPS) based on the three models Système d'analyse

Corresponding author address: Guillaume Thirel, CNRM–GAME, Météo-France, CNRS, GMME/MC2, 42 Ave. G. Coriolis, 31057 Toulouse, France.
E-mail: guillaume.thirel@meteo.fr

DOI: 10.1175/2008JHM959.1

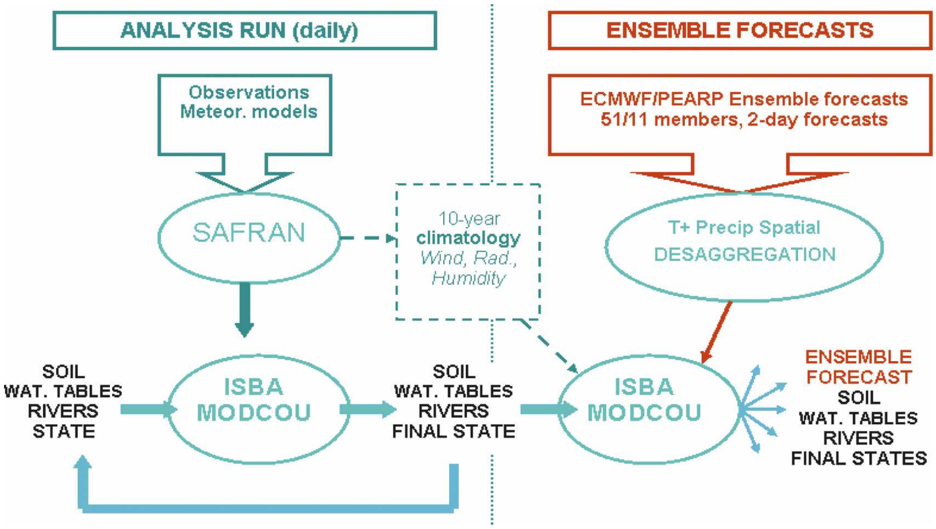

FIG. 1. Schematic of the Ensemble Streamflow Prediction System based on SIM: (left) the analysis run, which provides initial states of soil water tables and river flows to the ESPS; (right) the ESPS structure using disaggregated data from the ECMWF or PEARP EPS.

fournissant des renseignements atmosphériques à la neige (SAFRAN), the Interactions between Soil, Biosphere, and Atmosphere (ISBA) land surface model, and the distributed hydrological model Modélisation Couplée (MODCOU). This suite of models is known as the SAFRAN–ISBA–MODCOU model (SIM) (Habets et al. 2008). The system was tested over a 1-yr period, showing good performance for both high and low flows.

In the meantime, several EPSs dedicated to short-term forecasts have been constructed, such as the Prévision d'Ensemble Action de Recherche Petite Echelle Grande Echelle (PEARP) system (Nicolau 2002). This system was originally dedicated to predicting high impact storms in France. The aim of the present study is to evaluate the skill of the ESPS based on the PEARP input and to compare it to the results already obtained with the ECMWF EPS as input by using a large set of statistical scores and verifications. A conditional analysis has been performed, and an effort has been made, to analyze EPS and ESPS for dry and wet events and high and low flows. The scope of the study is then limited to short-range (48 h) streamflow forecasts for which predictions were available using both systems. Because ESPSs are a relatively new field of hydrometeorology, such a comparison is quite state of the art and can be informative in terms of our knowledge of ESPSs. After the description of the SIM model and the meteorological data used, the conditions of the test will be described. Then, the issue of the spatial disaggregation of the predicted rainfall is discussed. Finally, the probabilistic scores on the streamflows over all of France are presented, with a focus on the Seine River at Paris and the Ardèche River (a small basin in the south of France that is subject to extreme precipitation events).

2. Description of the SAFRAN–ISBA–MODCOU hydrometeorological model

The SAFRAN–ISBA–MODCOU (SIM) model (Fig. 1) is a suite of three independent models:

- SAFRAN (a French acronym for *Système d'analyse fournissant des renseignements à la neige*, which means analysis system that provides data to snow model) is a meteorological analysis system that uses meteorological model output combined with observations (from climatological and meteorological surface networks) to produce hourly values of meteorological variables near the surface. SAFRAN, originally developed by Durand et al. (1993) for the Alps, was extended to all of France and validated against observations from two recent years by Quintana-Seguí et al. (2008). SAFRAN analyses eight parameters: 10-m wind speed, 2-m relative humidity, 2-m air temperature, cloudiness, incoming solar and atmospheric/terrestrial radiation, snowfall, and rainfall. A detailed description and assessment of the SAFRAN analysis for France is presented in Quintana-Seguí et al. (2008). Only the main aspects are summarized here. The main purpose of SAFRAN is the use of climatologically homogeneous zones for its analysis (France is divided into 615 zones). Within each zone, the terrain elevation or altitudinal gradients are explicitly taken into account. In case of insufficient data in the zone, data from neighboring areas are used. More than 1000 meteorological stations are used for the 2-m temperature and humidity, and more than 3500 daily rain gauges for the precipitation analysis. An optimal interpolation method is used in the analysis every 24 h for precipitation and every 6 h for the other variables. All variables are disaggregated at an hourly time step. Finally, radiation terms are calculated using a radiation scheme.

- ISBA is a land surface model that simulates water and energy fluxes between the surface and the atmosphere (Noilhan and Planton 1989; Noilhan and Mahfouf 1996). It is used in numerical weather prediction, research, and climate models at Météo-France. To fulfill all its applications, the ISBA surface scheme is quite modular. In the SIM model, a three-layer force–restore model is used (Boone et al. 1999), together with an explicit snow model (Boone and Etchevers 2001). Moreover, a subgrid runoff scheme (Habets et al. 1999a) and subgrid drainage scheme (Habets et al. 1999b) are used. The latter parameterization is quite simple, and allows one to indirectly take into account the impact of unresolved aquifers on low riverflows based on a single parameter.

  The soil and vegetation parameters used by ISBA are derived from the Ecoclimap database (Masson et al. 2003). Only two parameters in ISBA are not directly associated with the soil and vegetation classification: the subgrid runoff parameter and the subgrid drainage parameter w_{drain} [see Habets et al. (2008) for more details].
- The MODCOU hydrogeological model computes the spatial and temporal evolution of the piezometric level of multilayer aquifers, using a diffusivity equation (Ledoux et al. 1989). Then, it solves for the interaction between aquifers and rivers and, finally, routes the surface water into rivers using a simple isochronism algorithm to compute riverflow. Riverflows are computed at a 3-h time step, and the evolution of the aquifer is computed daily.

SIM has been validated over the long term for several large French basins: the Adour (Habets et al. 1999), the Rhone (Etchevers et al. 2001), the Garonne (Morel 2003; Rousset et al. 2004), and then for France nationwide (Habets et al. 2007). It was shown that SIM was able to accurately reproduce water and energy budgets as well as observed streamflows, aquifer levels, and snowpack, in particular for basins with areas of over 1000 km^2. Since the end of 2003, SIM has been run on a daily basis at Météo-France.

3. The ensemble streamflow prediction system

a. Principles of the system

The ensemble streamflow prediction system (ESPS) is based on the SIM suite (Fig. 1) developed by Rousset-Regimbeau (2007) and Rousset-Regimbeau et al. (2007). The initial soil, river, and aquifer states are derived from the operational SIM suite. In this mode (called the analysis mode) SAFRAN is used to feed the coupled model ISBA–MODCOU. The only input data for the system are meteorological data. The soil moisture, soil temperature, river discharge, and aquifer levels are never reinitialized. This is the reason why Rousset-Regimbeau et al. (2007) decided to assess performance of the ESPS using the SIM analysis mode as the reference (and not the discharge observations). In prediction mode, ISBA–MODCOU is then forced by two types of data. Temperature and precipitation (including the snow/rain partition) are derived from the ensemble meteorological prediction system used. These data were previously disaggregated [see Rousset-Regimbeau et al. (2007) for the ECMWF EPS and section 5 for the Météo-France PEARP EPS for further details]. Other parameters (near-surface wind and humidity, radiation terms) are climatological values deduced from a long-term run of SIM. This choice was made because the operational Météo-France database received only the temperature and precipitation input for the ECMWF EPS. It is assumed that this probably leads to a reduced spread in the hydrological ensemble. This choice may be reconsidered later, but is beyond the scope of this paper. ISBA–MODCOU are identical in analysis and prediction modes (e.g., no modification has been introduced to account for the model uncertainty at this stage).

b. The ECMWF meteorological Ensemble Prediction System

The ECMWF EPS was implemented in 1992 and has been continuously improved and validated since then (Chessa and Lalaurette 2001; Buizza et al. 2007). The EPS consists of 51 10-day forecasts runs at resolution TL399L62, equivalent to a spatial grid of approximately 50 km and 62 vertical layers (it is now extended to 15 days at a lower resolution between days 10 and 15). The EPS is run twice a day, at 0000 and 1200 UTC. Initial uncertainties are simulated by perturbing the unperturbed analyses with a combination of T42L40 singular vectors (270-km grid mesh and 40 vertical layers), computed to optimize total energy growth over a 48-h time interval. Model uncertainties are simulated by adding stochastic perturbations to the tendencies due to pa-

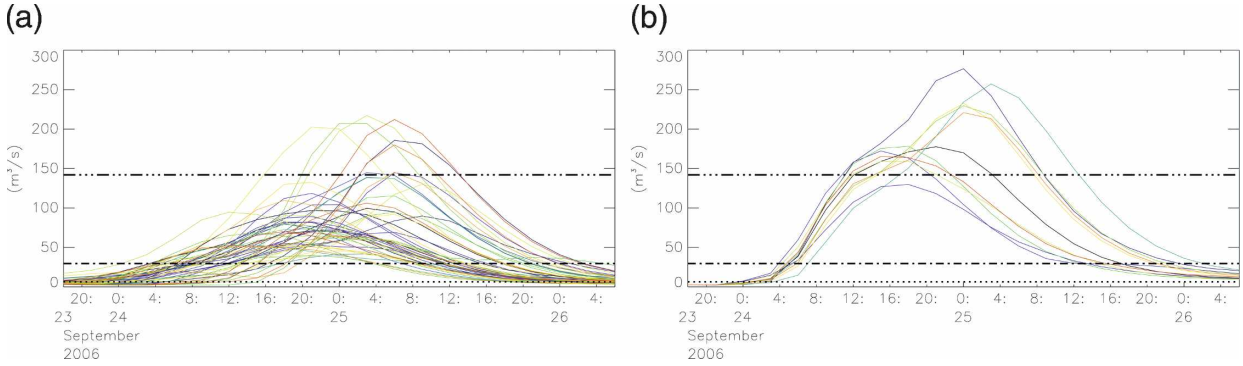

FIG. 2. Example of streamflows forecasted for the Ardèche River at Saint Martin d'Ardèche (2240 km^2) by (a) the ECMWF-based and (b) the PEARP-based ESPSs for the flood event on 24–25 Sep 2006: daily Q90 (dashed–dotted–dotted–dotted), Q50 (dashed–dotted), and Q10 (dotted) quantiles are indicated. No 3-h quantiles were available. Each solid line represents one member of the ESPS.

rameterized physical processes. In Rousset-Regimbeau et al. (2007) and in this study, we used the precipitation and temperature forecasts stored in the operational Météo-France database on a 1.5° × 1.5° grid.

c. *The Météo-France meteorological ensemble prediction system*

The Météo-France PEARP (Nicolau 2002) is based on the global spectral Action de Recherche Petite Echelle Grand Echelle (ARPEGE) model (Courtier et al. 1991). The model has a stretched grid, the resolution being higher over western Europe. Its resolution is TL358L46 with a stretch factor of 2.4 (equivalent to a mesh from 23 to 100 km, and 46 vertical layers). Initial perturbations are generated by a singular vector technique, as for the ECMWF EPS. As this system is focused on a short-term range, the area of interest for the singular vectors is limited to the Atlantic–western European area and the lead time is 12 h instead of 48 h. Forecast lead time is 60 h and the members are limited to 11 because of computer costs. The PEARP is run on a daily basis at 1800 UTC. Despite a lack of spread, the validation showed that the PEARP has a good skill for short-range prediction of severe events when compared to the ECMWF EPS (Nicolau 2002).

4. Characteristics of the test

We compared the results of the two ESPS over a common 48-h period from 0000 UTC of day 1 to 2400 UTC of day 2, corresponding to the forecast range 06–54 h for the PEARP and 00–48 h for the ECMWF EPS. This choice was also governed by the time step of the hydrological model (3 h for the river streamflows and 24 h for aquifers). The PEARP ensemble temperature and precipitation forecasts were available daily in the Météo-France operational database. The time step was 3 h for the first 54 h and the final time step 6 h. The ECMWF EPS output were stored using a 6-h time step over the whole period. A linear hourly interpolation of rainfall and temperature data was used.

We adopted the strategy used by Rousset-Regimbeau et al. (2007) to reduce computer cost of the ISBA–MODCOU model: the ISBA time step was extended to 20 min, instead of the usual 5 min. All validation presented here was based on results for a 569-day period from 11 March 2005 to 30 September 2006.

To compute streamflow statistical scores, a common SIM streamflow reference has been used. For the Brier skill scores (BSSs) and ranked probability skill scores (RPSSs), the reference was a streamflow model climatology from 1981 to 2004, as in Rousset-Regimbeau et al. (2007). For rainfall scores, the reference was the SAFRAN climatology from 1995 to 2004. Furthermore, in order to take into account the difference between ensemble sizes, which introduces a bias in skill scores, such as BSS and RPSS, an artificial bias (Weigel et al. 2006) was included afterward in the reference BS and RPS scores. This bias, described in appendix A, was included in the scores presented in this study. No significant changes were found in the comparison between EPS and ESPS, but this bias tended to increase PEARP scores more than ECMWF scores because the fewer members in an EPS biases scores more negatively (Weigel et al. 2006).

Figure 2 shows an example of results from the two ensemble streamflow prediction systems for a particular flood event of the Ardèche River (a small basin in southeastern France, with a surface of 2240 km^2). Each curve represents one of the members of the ESPS evolving each 3 h. The spread of the members can be seen for this event, and it is important to notice that the PEARP-based ESPS (Fig. 2b) forecast higher floods than the ECMWF-based ESPS (Fig. 2a). This figure shows that, despite the original low resolution of rainfall data, extreme events can be predicted in small basins. In Fig. 2, only daily quantiles are plotted because 3-h quantiles were not available. It can be assumed that the 3-h Q90 would be higher and the 3-h Q10 would be lower than daily quantiles: daily quantiles were plotted to provide an idea of how high the flows were for this event.

5. Spatial disaggregation and statistical analysis of rainfall forecasts

a. Reasons and methods for a rainfall disaggregation

Disaggregation methods are a compulsory way of providing model inputs at the same temporal and spatial scales as for the model itself. They are essential for taking into account physiographic characteristics that are not dealt with in low-resolution global models. There are two main classes of disaggregation models. First, statistical disaggregation methods can be used, such as the analog method (Boe et al. 2006) that relies on searching in climatological records for a day with the most similar atmospheric circulation pattern, in order to deduce the value of the variable to be forecast. This method is quite simple and successful, but does not rely on real physics mechanisms. Regression (Zorita et al. 1995) is a second method and links local variables with large-scale predictors. This method has difficulty in maintaining consistency between these variables. Second, dynamic disaggregation methods can be used, as described in Déqué (2007). The methods are numerous and rely mainly on unbiasing or resampling. They are mainly used in hydrology for climate change, as for example the delta method and the unbiasing method.

Because ECMWF and PEARP rainfall data are, respectively, available on 1.5° × 1.5° and a 0.25° × 0.25° grids and the ISBA grid is at a resolution of 8 km × 8 km, a spatial disaggregation of rainfall forecasts has been necessary for both ensemble forecast data. Indeed, some aspects like terrain elevation/altitude are represented in a very different way in both ensemble systems when compared with the ISBA grid. That is why two disaggregations of the EPS forecasts including homogeneity of areas and elevation/altitudinal effects have been set up (see sections 5b and 5c for more details). Moreover, all scores on rainfall data have been processed on the ISBA grid in order to have a consistent grid between both systems.

b. Disaggregation of the ECMWF EPS

The method proposed by Rousset-Regimbeau et al. (2007) was also used in this study for the ECMWF EPS. The disaggregation from a resolution of 1.5° × 1.5° to the 8 km × 8 km grid of SIM was motivated from the SAFRAN method. First, data from the ECMWF EPS are interpolated horizontally onto the SAFRAN zones (615 irregular zones) using distance-dependent weights ($1/r^2$ interpolation). Second, a fixed vertical gradient for precipitation is used to account for the elevation difference between the ECMWF EPS and the SIM orography. This vertical gradient was calibrated over approximately 1 yr (from 4 September 2004 to 31 July 2005) by Rousset-Regimbeau et al. (2007) using a trial and error method. The gradient for precipitation was 0.7 mm yr^{-1} m^{-1} beyond an altitude of 800 m and 2 mm yr^{-1} m^{-1} below 800 m. For temperature, the usual mean atmospheric lapse rate gradient [-0.65 K $(100\ \text{m})^{-1}$] was applied.

c. Disaggregation of the PEARP EPS

Even though the resolution of the PEARP EPS is better than in ECMWF EPS, in the operational database of Météo-France (0.25° × 0.25°), a disaggregation was applied. However, it was impossible to apply the same methodology as for the ECMWF EPS precipitation data. The differences between the 24-h precipitation forecast and the SAFRAN analysis were not explained by the elevation differences between the two model orographies. It has been decided to calculate the ratio

$$\frac{\text{SAFRAN rainfall}}{\text{PEARP rainfall}}$$

over 1 yr: from 11 March 2005 to 10 March 2006 (Fig. 3). This ratio was then applied to the PEARP EPS data in order to match the SAFRAN climatology. The results were validated using the 204 days following the calibration period (11 March 2006 to 30 September 2006). The PEARP rainfall (ensemble mean) overestimated the SAFRAN analysis by 12% for the first day of forecast and 9% for the second day. The overall shape of the precipitation field remained acceptable.

d. Statistical analysis of the rainfall forecasts

The maps of the disaggregated 24-h precipitation forecasts obtained for the two models are compared to the SAFRAN analysis reference (Fig. 4a) over the period from 11 March 2005 to 30 September 2006 in Figs. 4b and 4c. This period overlaps the period used by Rousset-Regimbeau et al. (2007) for the calibration of the elevation/altitudinal gradient (4 September 2004–31 July 2005) and the period used in the present study to calibrate the PEARP disaggregation (11 March 2005–10 March 2006). The spatial distribution over the whole

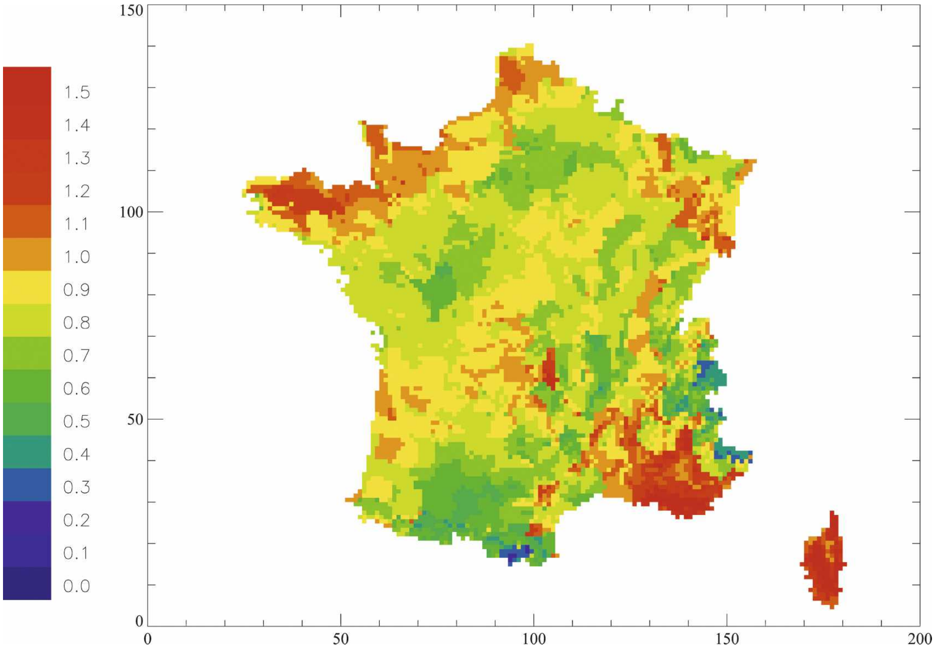

FIG. 3. Map of the SAFRAN/PEARP rainfall ratio for the period from 11 Mar 2005 to 10 Mar 2006 over all of France used for the disaggregation of PEARP rainfall.

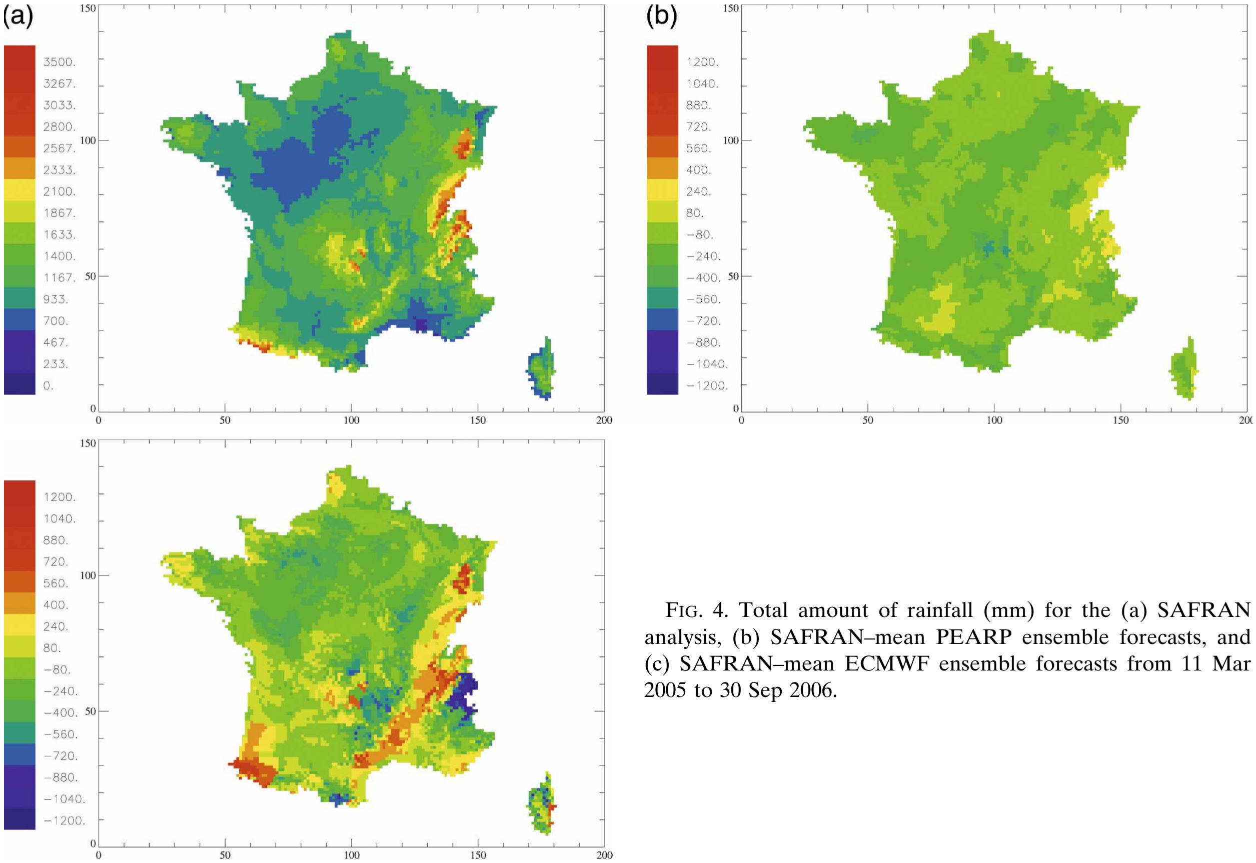

FIG. 4. Total amount of rainfall (mm) for the (a) SAFRAN analysis, (b) SAFRAN–mean PEARP ensemble forecasts, and (c) SAFRAN–mean ECMWF ensemble forecasts from 11 Mar 2005 to 30 Sep 2006.

TABLE 1. Rainfall mean Brier skill score for all of the period (569 days), the summer (366 days), and the winter (203 days). The evaluated certainty (%) of the significance of the differences between the ECMWF and the PEARP Brier Skill Scores corresponding for the *t* test (left) and the Wilcoxon test (right) appears in italics.

Threshold (mm day^{-1})	Period	ECMWF		PEARP		Statistical tests			
		Day 1	Day 2	Day 1	Day 2	Day 1		Day 2	
>5	All	0.56	0.56	0.63	0.62	*99.2*	*50.9*	*98.9*	*97.4*
	Summer	0.53	0.52	0.58	0.57	*71.0*	*85.7*	*99.9*	*98.0*
	Winter	0.60	0.62	0.69	0.67	*100*	*92.5*	*53.2*	*64.7*
>10	All	0.31	0.32	0.42	0.40	*99.1*	*75.7*	*97.2*	*99.1*
	Summer	0.28	0.27	0.38	0.36	*71.7*	*90.2*	*97.3*	*98.4*
	Winter	0.37	0.38	0.49	0.44	*99.9*	*74.2*	*73.9*	*82.0*
<1	All	0.70	0.70	0.77	0.78	*72.6*	*80.8*	*99.6*	*76.3*
	Summer	0.69	0.70	0.74	0.75	*100*	*93.5*	*100*	*99.4*
	Winter	0.72	0.71	0.84	0.82	*100*	*100*	*100*	*98.7*

period was quite accurate for the mean PEARP ensemble forecasts, with marked differences between plains and mountainous areas (Fig. 4b). The amount of rainfall averaged over all of France was 2.42 mm day^{-1} for the first day of PEARP forecasts and 2.33 mm day^{-1} for the second day, which was close to the mean amount of rainfall for the SAFRAN analysis, which was also 2.33 mm day^{-1} during this period. The mean ECMWF ensemble forecasts distribution for France was also accurate. However, some discrepancies appeared (Fig. 4c) in northwestern France (overestimation) and in southern France (both underestimation and overestimation) when compared to the SAFRAN analysis. The ECMWF rainfall ensemble forecasts averaged over all of France was close to the SAFRAN analysis (2.36 mm day^{-1}) for the first day of forecasts, but for the second day it was higher than for the SAFRAN analysis (2.45 mm day^{-1}).

Some statistical scores were calculated to describe the overall quality of the two ensembles and the significance of the differences encountered. The statistical scores are described in appendix B and the verification methods for the significance of the results are described in appendix C. All following scores were computed for each day, regardless whether precipitation was forecast, temperature also having an influence on hydrology. The rms error (RMSE) of the mean ensemble members used the reference rainfall data from SAFRAN. It varied from 3.76 mm day^{-1} for the first (forecast) day to 3.98 mm day^{-1} for the second day for the PEARP EPS, while it varied from 3.92 to 4.08 mm day^{-1} for the ECMWF EPS. According to the *resampling test,* the differences were significant for all days (see appendix C for more details on the resampling test). During our study period, the spread increased from 0.80 to 1.24 mm day^{-1} for the PEARP EPS, while it varied from 0.61 to 1.13 mm day^{-1} for the ECMWF, from day 1 to day 2. These scores confirmed that the PEARP forecasts were more appropriate for short-range forecasts, as could be expected. These scores were significantly different according to the resampling test. The spread improved for the second day. However, the spread appeared to be too low in all cases, according to the commonly admitted rule that the spread and the RMSE should be of the same order.

The Brier skill scores for three thresholds are given in Table 1. The BSS is the most common score used to quantify the quality of the prediction with respect to predefined thresholds for probabilistic forecasts (Brier 1950). The BSS scores vary from $-\infty$ to 1, positive values indicating an added value compared to the reference. The thresholds were chosen to cover a range of meteorological situations, from dry to wet: lower than 1 mm day^{-1}, higher than 5 mm day^{-1}, and higher than 10 mm day^{-1}. These three categories contain, respectively, 67%, 14.4%, and 6.5% of rainfall days in the mean for the SAFRAN analysis during these 569 days. The BSSs were computed for each grid point in France, then averaged. The reference was the SAFRAN climatology (1995–2004). When first looking at the global scores (on all 569 days), it can be seen that all scores were higher than 0, indicating an improvement when compared to the climatology. For all thresholds, the PEARP rainfall forecasts were better than the ECMWF forecasts, and the differences were significant according to both significance tests for all cases except for the 1 mm day^{-1} threshold on day 2, and in a few cases for one test. These results confirmed our confidence in the PEARP for short-term forecasts. When comparing the thresholds together, it appeared that both EPSs were better for the lower threshold than for the 5 mm day^{-1} threshold, and the higher threshold had the lower BSS. This can be explained by the fact that there were a nonnegligible number of days for the lower threshold when all

TABLE 2. Mean ranked probability skill score for rainfall during all of the period (569 days), the summer (366 days), and the winter (203 days). The evaluated certainty (%) of the significance of the differences between the ECMWF and the PEARP Ranked Probability Skill Scores corresponding for the *t* test (left) and the Wilcoxon test (right) appears in italics.

	ECMWF		PEARP		Statistical tests			
Period	Day 1	Day 2	Day 1	Day 2	Day 1		Day 2	
All	0.57	0.55	0.68	0.68	*98.6*	*89.2*	*99.9*	*86.5*
Summer	0.57	0.55	0.64	0.65	*98.2*	*80.5*	*100*	*98.3*
Winter	0.57	0.54	0.75	0.73	*100*	*99.9*	*84.1*	*87.4*

of the members predicted no rain, as for the SAFRAN analysis. But these cases cannot be rejected from the scores because of the temperature members, which has an influence on snow and evapotranspiration.

The Brier skill scores have been decomposed using its classical decomposition (see appendix B). The lowest reliability (i.e., the best one) was observed for the 10 mm day^{-1} threshold and the highest reliability for the 1 mm day^{-1} threshold (from 0.01 to 0.07), which means that there was probably an over- or underestimation of nonrain days. Reliability was twice lower for PEARP than for ECMWF EPS for each threshold. Concerning resolution, the lowest was for the 10 mm day^{-1} threshold, and the highest (that is to say, the best one) for the 1 mm day^{-1} threshold (from 0.02 to 0.11). This can be due to a bias toward heavy rain. The resolution was a little bit better for the ECMWF EPS than for the PEARP EPS. A slight decrease in resolution was seen from day 1 to day 2, as well as for reliability.

To assess the EPS rainfall forecasts over the whole range, ranked probability skill scores (RPSSs) were calculated (Table 2). The first row in Table 2 shows the global mean RPSS (i.e., calculated for the whole 569-day period). All of these scores show the skill of the EPSs, with scores over 0.5. When comparing both EPSs together, there was no doubt that the PEARP EPS performed better than the ECMWF EPS, and the differences were significant for day 1 and slightly less significant for day 2.

A seasonal study of these scores has been performed. Two seasons have been separated: summer (366 days including April–September periods) and winter (203 days including October–March periods). Both RMSEs and spreads showed higher values during summer than during winter, but did not change the hierarchy between EPSs (higher value for ECMWF for the RMSE and higher value for PEARP for the spread). The BSSs were better during winter than during summer (Table 1), and the hierarchy between the PEARP EPS and the ECMWF EPS did not change. But, it appeared that the differences for the two higher thresholds during the summer for day 1 were not significant, nor for winter on day 2. For the RPSS (Table 2), no seasonal dependence was found for the ECMWF EPS, whereas a better score was found during winter for the PEARP EPS. However, PEARP remained better than ECMWF regardless of season, with a high significance level except on day 2. Finally, rank (or Talagrand) histograms (not shown) were little modified when computed for the 569 days, or only winter or summer, showing a lack of spread, and not a conditional error, which could have been hidden over the whole period.

The influence of basin size on the results was also studied. Basin sizes were distributed into six classes: <600 km^2 (189 basins), 600–1000 km^2 (217), 1000–2000 km^2 (176), 2000–4000 km^2 (118), 4000–10 000 km^2 (93), and >10 000 km^2 (88). The BSS and RPSS scores were averaged over all grid points included in the basins and then averaged for all classes. The PEARP EPS had a higher BSS and RPSS than for the ECMWF EPS, and for both scores the larger the basin, the higher the score, except for the smaller class. The spatial distribution of the BSSs and RPSSs has been studied, showing lower scores for northwestern plains and for mountains (Pyrènèes, Alps, and Massif Central), mostly due to lower reference scores.

This statistical analysis highlighted higher scores for the disaggregated PEARP EPS than for the disaggregated ECMWF EPS. A weak seasonal dependence was found, especially for the PEARP, winter scores being better than summer scores. For both EPSs the higher the threshold, the lower the scores. This feature was consistent with Buizza et al. (1999). Finally, the larger the basin, the better the scores.

6. Statistical analysis of streamflow predictions

a. *Streamflow prediction over all of France*

A comprehensive statistical analysis of the ESPS was performed over the period from 10 March 2005 to 30 September 2006. Skill scores were calculated over the 881 gauge stations computed by SIM. Streamflow predictions were compared with the reference simulation of SIM.

We focused on high or moderate flows by defining

TABLE 3. Mean Brier skill scores for streamflows during all of the period (569 days), the summer (366 days), and the winter (203 days). The evaluated certainty in % of the significance of the differences between the ECMWF and the PEARP Brier Skill Scores corresponding for the resampling test appears in italics.

		ECMWF		PEARP		Statistical test	
Threshold	Period	Day 1	Day 2	Day 1	Day 2	Day 1	Day 2
>Q50	All	0.91	0.85	0.88	0.86	*100*	*100*
	Summer	0.91	0.84	0.87	0.83	*99.4*	*99.9*
	Winter	0.88	0.78	0.88	0.84	*100*	*100*
>Q90	All	0.90	0.85	0.91	0.88	*90.0*	*100*
	Summer	0.88	0.79	0.82	0.71	*100*	*73.8*
	Winter	0.89	0.83	0.90	0.87	*100*	*100*
<Q10	All	0.78	0.68	0.74	0.70	*93.4*	*100*
	Summer	0.86	0.78	0.77	0.73	*100*	*65.4*
	Winter	0.74	0.62	0.72	0.67	*100*	*100*

the thresholds using the 90th (Q90) and 50th (Q50) percentiles of daily streamflows from the study period. We also studied low flows by computing the scores for streamflows remaining below the 10th (Q10) percentile. As for the BS_{ref}, the 1981–2004 streamflow SIM climatology was used, as well as for determining the quantiles. Table 3 shows the mean BSS for both PEARP- and ECMWF-based ESPSs for the three thresholds. The BSSs were very high for both ECMWF- and PEARP-based ESPSs (a BSS over 0.7 can be considered as describing an EPS of quality). As expected, the predictions were less accurate for the second day. When comparing the two systems, the ECMWF-based ESPS was found to be the best for Q10 and Q50 for the first day. For the second day the PEARP-based ESPS scores were systematically better for all thresholds. The statistical test (resampling test) showed that the differences were significant for all thresholds (considering that an evaluated certainty higher than 90% describes a very high level of confidence of the differences between the scores). According to Table 4, RPSS showed a better score for the ECMWF ESPS for day 1 (but no significance) and a better score for the PEARP ESPS for day 2 (fully significant).

When examining the terms of the decomposition of the Brier score, no trend has been found concerning the reliability, this term being very good for each threshold (under 0.02). The resolution indicated best performance at the Q50 threshold (≥0.16), then at the Q10 threshold (around 0.10), and finally at the Q90 threshold (around 0.075). The systems may have problems separating probability classes for high streamflows (by putting all of them in the highest class or by missing the event), and for the Q10, it is not surprising that streamflows have difficulties with spread, since soil absorbs most of the moisture in such cases. No difference has been found for the resolution between the ESPSs, whereas the reliability was quite better for the PEARP version. A slight decrease in the resolution has been seen from day 1 to day 2, whereas reliability increased.

Reliability diagrams for streamflows are shown in Fig. 5. The reliability diagram for the PEARP ESPS (Fig. 5b) showed probabilities distributed into 12 classes, whereas probabilities were distributed into 52 classes for the ECMWF ESPS (Fig. 5a). It was of note that probabilities were overestimated for each case, implying many more floods predicted than observed, and many more low flows predicted than observed. This trend was more pronounced for the ECMWF than for the PEARP ESPS. Rank histograms, not included in this paper, had a U shape. As both reliability and rank diagrams had the same shape during the whole period, and during winter and summer, it can be deduced that it described a lack of spread for both ESPSs.

Next the spatial distribution of the BSS was studied. Figure 6 shows the 881 stations used in this study. They were distributed over all of France and included small and large basins. Figure 7 shows the stations where the two ESPSs differed significantly according to the resampling test (with a level of confidence of 90%) for

TABLE 4. Mean ranked probability skill scores for streamflows over the whole period (569 days), the summer (366 days), and the winter (203 days). The evaluated certainty (%) of the significance of the differences between the ECMWF and the PEARP RPSS corresponding for the resampling test appears in italics.

	ECMWF		PEARP		Statistical test	
Period	Day 1	Day 2	Day 1	Day 2	Day 1	Day 2
All	0.90	0.83	0.88	0.85	62.5	*100*
Summer	0.91	0.84	0.86	0.82	700	*52.0*
Winter	0.90	0.83	0.91	0.88	700	*100*

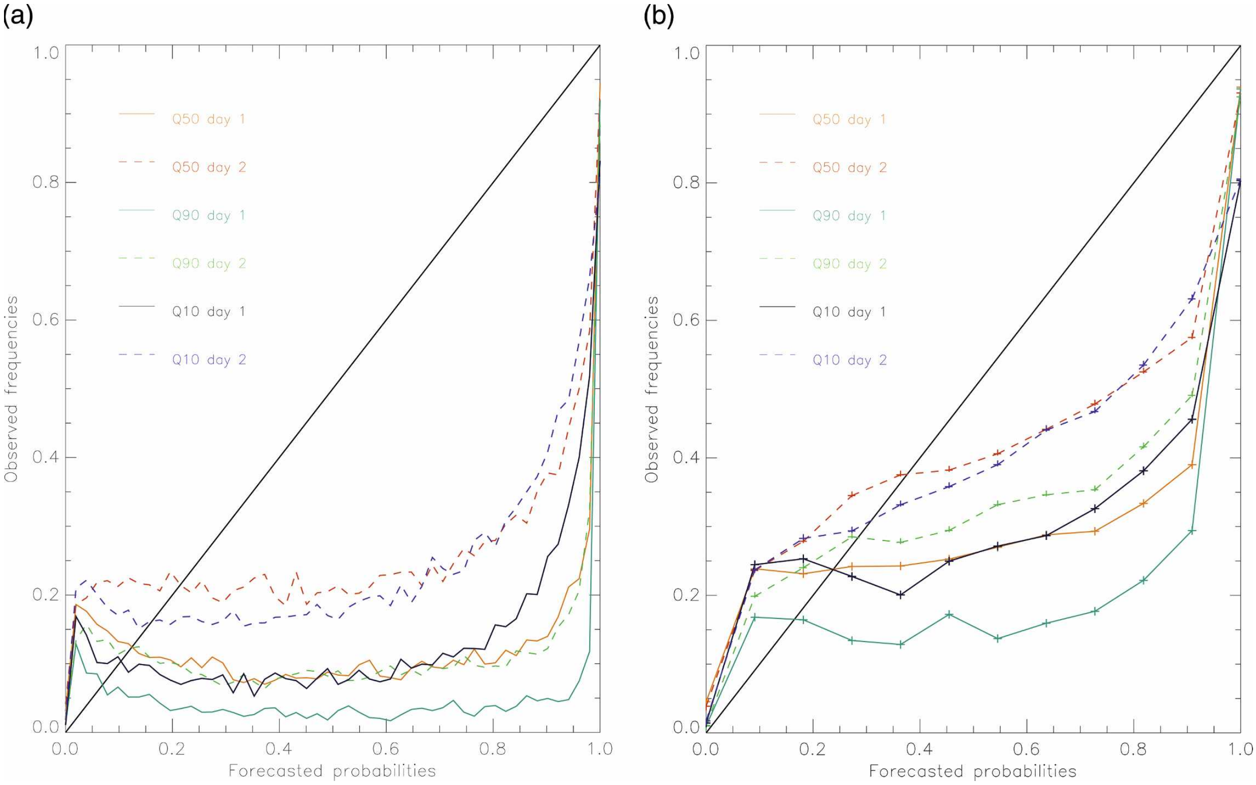

FIG. 5. Reliability diagrams for day 1 (solid lines) and day 2 (dashed lines) predictions for (a) the ECMWF and (b) the PEARP ESPSs of the Q10, Q50, and Q90 quantile thresholds.

the Q10 threshold. Stations where the PEARP-based ESPS was better are in red, while stations where the ECMWF-based ESPS was better are in blue. For the first day (Fig. 7a), the PEARP-based ESPS seemed better in most basins (184 stations against 98 for the ECMWF). The stations where the PEARP was better were irregularly distributed, with no region where one of the ESPSs was predominant. This was not the case for the second day (Fig. 7b); 329 stations for the PEARP versus 33 for the ECMWF). It must be noted that at a relatively high number of stations, the results of the two ESPSs were not significantly different and are not shown in Fig. 7a (599 out of 881 stations not significantly different) and Fig. 7b (519 out of 881).

For high flows (Figs. 8a and 8b) at the Q90 threshold, the PEARP-based ESPS confirmed good scores in almost all regions [338 stations versus 49 for the ECMWF ESPS for day 1 (panel a), and 486 versus 19 for day 2 (panel (b)], with the noticeable exception of the Seine basin (where the two ESPSs are not significantly different). This behavior can be attributed to the presence of a large aquifer in this region: In this case, the initial state has a major influence on streamflow predictions. Hence, the differences between the two ESPSs cannot easily be highlighted. It can be concluded that no ESPS was more adapted to simulate low flows, especially for day 1, while the PEARP-based ESPS is best for high flows. Despite a small number of members and a lack of ensemble spread, it was encouraging to note that the PEARP-based ESPS results were of the same order (or even better) as the ECMWF-based ESPS (it must be noted that this latter system was not calibrated for short-term predictions).

The false alarm (FA) ratios and hit rates (HRs) were calculated assuming that an event was predicted when 90% of the members predicted it (further details about the method of calculation are provided in appendix B). For the Q90, the FA was 4% (first day) and 8% (second day) for the ECMWF-based ESPS, whereas it remained under 4% for the PEARP-based ESPS. Both hit rates were above 75%, the PEARP-based ESPS always being best. For the Q10, the FA was 6% (first day) and 9% (second day) for the ECMWF-based ESPS, and remained around 9% for the PEARP-based ESPS. The HR decreased from 86% to 74% for the ECMWF-based ESPS from day 1 to day 2 and from 87% to 77% for the PEARP from day 1 to day 2. Both ESPSs can provide reliable information on high and low flows. The PEARP-based ESPS was best for high floods, whereas its advantage was less marked for the Q10. These good

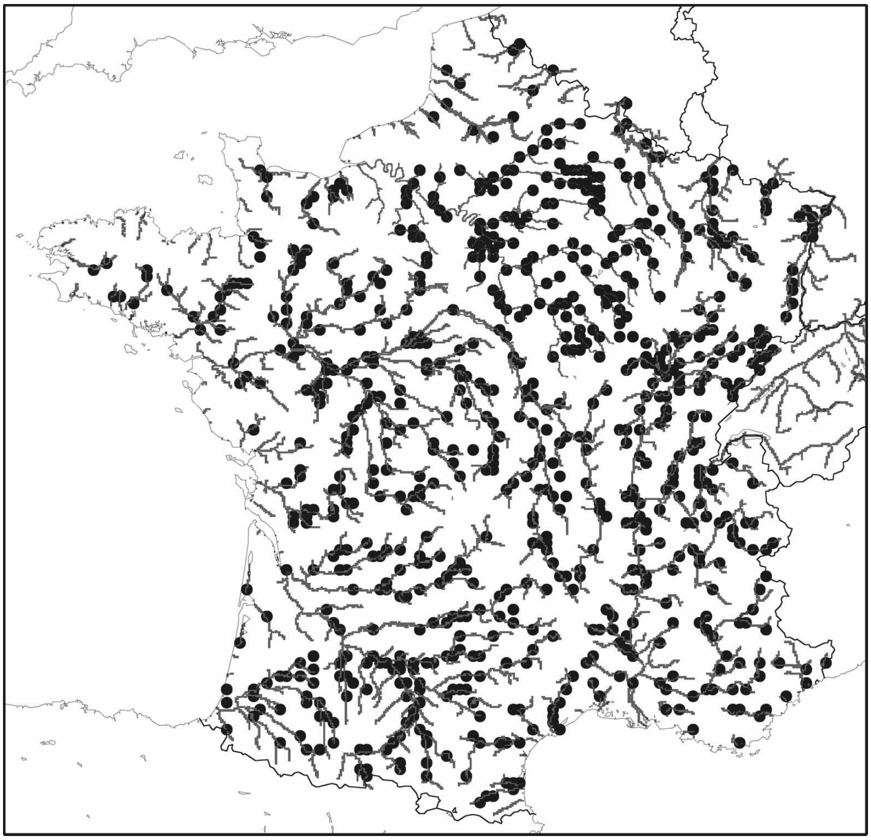

FIG. 6. Map of the 881 hydrological stations available nationwide in France for this study.

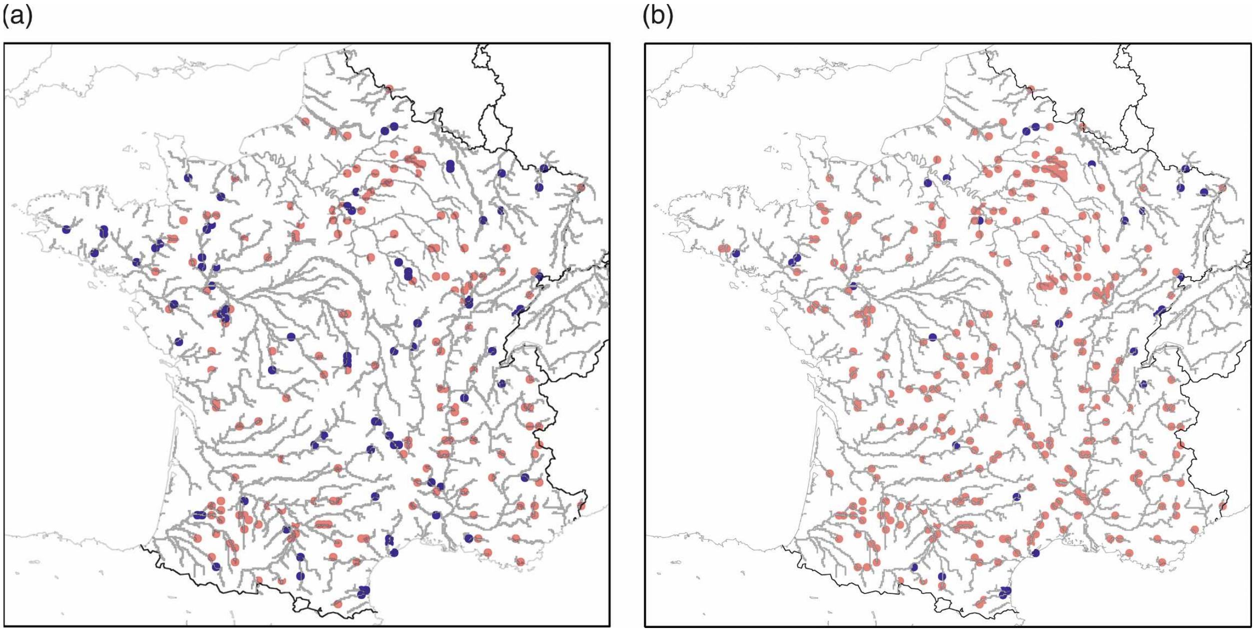

FIG. 7. Brier skill score over France for all stations having a statistically significant difference between the ECMWF- and PEARP-based ESPSs of the Q10 threshold, from 11 Mar 2005 to 30 Sep 2006, on (a) day 1 and (b) day 2: Stations are colored where the ECMWF-based ESPS (in blue) and where the PEARP-based ESPS (in red) is best with more than 90% significance according to the resampling test.

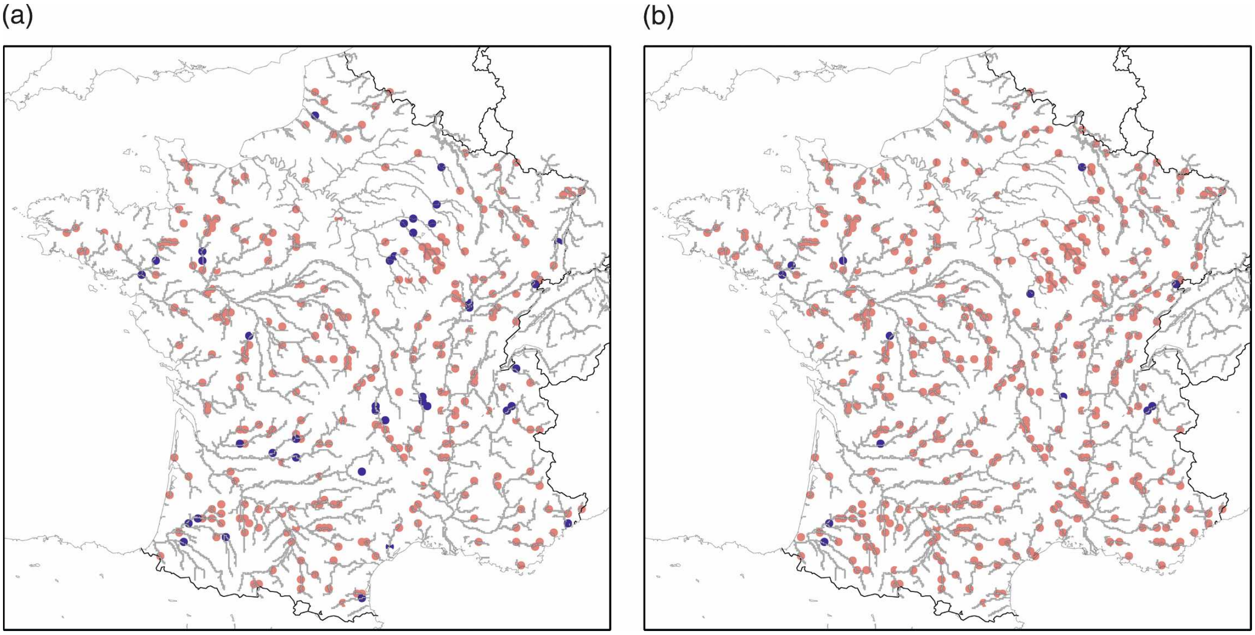

FIG. 8. As in Fig. 7 but of the Q90 threshold.

skills indicate that short-range ESPSs have the potential as a reliable decision tool for discharge forecasting.

Figures 9a and 9b show the evolution of the mean BSS for Q10 and Q90 at the basin sizes as described previously. As already seen, the mean BSSs were better for the first day than for the second for both thresholds. The BSSs were better for the floods (Q90, Fig. 9b) than for the low flows (Q10, Fig. 9a). But, in both cases, the mean BSSs were quite good (always greater than 0.6 and often above 0.7). The decrease in the BSS for the second day of forecast was much more important for the ECMWF-based ESPS than for the PEARP-based ESPS. The better score for ECMWF than for PEARP ESPS can be explained by a hypothetical activation of subgrid drainage with PEARP data because of greater spread error for rainfall. Finally, the BSS increased with basin size and faster for the ECMWF-based ESPS than for the PEARP-based ESPS. This increase could be explained by the fact that the larger a basin was, the lower the effect of rainfall before two days because of the structure of the hydrologic model, based on isochron zones. Statistical scores were calculated on the

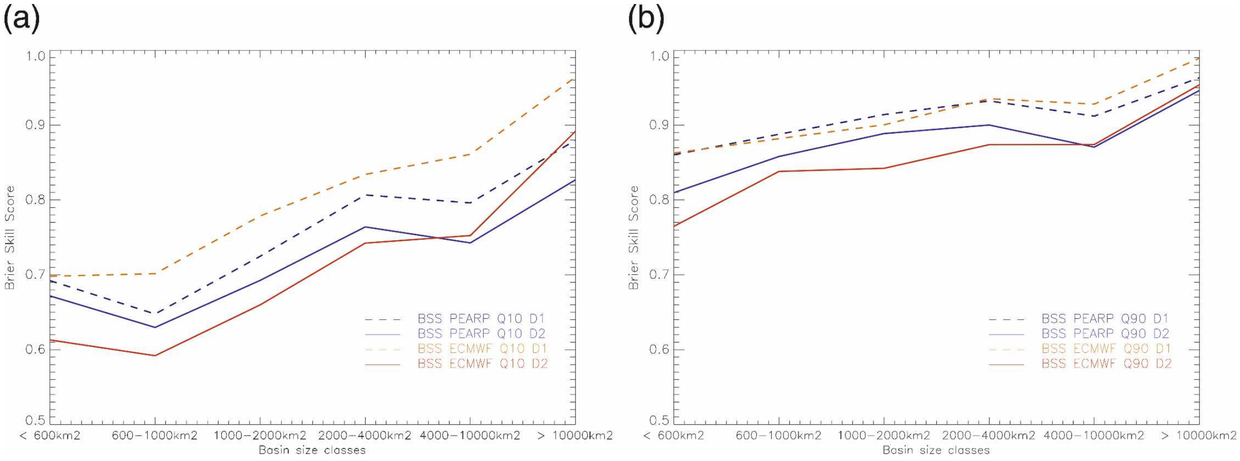

FIG. 9. Evolution of the BSS with basin size for (a) the Q10 and (b) the Q90 quantiles on day 1 (dashed lines) and day 2 (solid lines) for the ECMWF and for the PEARP streamflows. The BSS was averaged by basin size from 11 Mar 2005 to 30 Sep 2006.

scores presented in Fig. 9 and show that the differences were significant, except for the values where the curves crossed.

Here again, a seasonal study has been done of the BSSs and RPSSs for the streamflows. For summer, the ECMWF ESPS was always significantly best considering the BSS (Table 3), except on day 2 for Q90 and Q10. This is an interesting result since, over the whole period, the PEARP EPS is always best on day 2. During winter, the PEARP ESPS is always best except for Q50 and Q10 on day 1. Considering the RPSS (Table 4), the ECMWF ESPS is best for summer (but not significantly for day 2), and the PEARP ECMWF is best for winter. All of these results show that, despite a global score better for ECMWF ESPS on day 1 than for PEARP ESPS, the PEARP was more informative for winter.

b. *Results for two contrasted catchments: The Seine at Paris, and the Ardèche at St. Martin d'Ardèche*

The Seine at Paris and the Ardèche at Saint Martin d'Ardèche are two catchments with very different features. The Seine is a large basin (43 800 km^2 at Paris), with no significant orography and a large aquifer system (three layers of aquifers are simulated in the MODCOU model). On the other hand, the Ardèche is a small basin (2240 km^2) in the southeast of France in the hilly region of the Cévennes, which is susceptible to extreme precipitation events in autumn.

The detailed statistical scores for the two catchments are given in Table 5. The scores for the rainfall were computed over the whole basins (mean value over the area), while the scores for the streamflow were computed at the indicated gauge station. In most cases, the rainfall spread was higher for the PEARP EPS. Because of the different characteristics of the two basins, the streamflow spread was higher for the Ardèche (because of the small basin scale, the water routing time is less than 1 day; hence, streamflows were closely linked to rainfall). The streamflow RMSE on the Seine appeared to be remarkable for the ECMWF-based ESPS; the RMSE was about 0.84 m^3 s^{-1}, while it was 5.36 m^3 s^{-1} for the PEARP. For the second forecast day, the values for both ESPSs were closer (7.03 versus 9.00 m^3 s^{-1}). For the Seine basin, the rainfall RMSE stayed around 3 mm (24 h)$^{-1}$ for both EPSs. This apparent contradiction may be explained by the spatial distribution of rainfall over the catchment (the rainfall scores in Table 5 were based on an average rainfall over the basin), which can be located in another isochron zone. The spread of the rainfall RMSE on the 8 km × 8 km grid (representing the standard deviation of all grid-point RMSE values of the basin in comparison with the mean RMSE of the basin) was 22% higher for the PEARP EPS rainfall than for the ECMWF EPS rainfall. This indicated a more heterogeneous spatial distribution of the rainfall forecast errors of the PEARP EPS and highlighted the importance of the spatial distribution of rainfall for streamflow predictions.

TABLE 5. Streamflow statistical scores for two different-sized catchments: the Seine at Paris (43 800 km^2) and the Ardèche at Saint Martin d'Ardèche (2240 km^2).

		ECMWF		PEARP	
		Seine	Ardèche	Seine	Ardèche
Rainfall RMSE	Day 1	3.02	8.08	2.95	6.15
(mm day^{-1})	Day 2	3.06	7.71	3.06	7.53
Streamflow RMSE	Day 1	0.84	13.7	5.36	10.9
(m^3 s^{-1})	Day 2	7.03	19.4	9	18.2
Rainfall spread	Day 1	0.59	0.55	0.72	0.97
(mm day^{-1})	Day 2	1.12	1.10	1.11	1.63
Streamflow spread	Day 1	0.09	1.35	0.1	0.9
(m^3 s^{-1})	Day 2	0.58	2.47	1.1	2.6
Rainfall BSS	Day 1	0.31	0.43	0.45	0.39
>5 mm day^{-1}	Day 2	0.39	0.46	0.48	0.42
Rainfall BSS	Day 1	−0.03	0.11	0.08	0.04
>10 mm day^{-1}	Day 2	0.09	0.09	0.10	0.05
Rainfall BSS	Day 1	0.71	0.65	0.71	0.60
<1 mm day^{-1}	Day 2	0.72	0.69	0.71	0.63
Streamflow BSS	Day 1	1	0.91	0.97	0.95
>Q50	Day 2	0.99	0.87	0.96	0.95
Streamflow BSS	Day 1	1	0.98	1	0.96
>Q90	Day 2	1	0.97	1	0.96
Streamflow BSS	Day 1	1	0.87	0.99	0.91
<Q10	Day 2	0.98	0.85	0.98	0.89

The rainfall BSS indicated that the EPS for the Seine basin had similar behavior as for all of France, but for the Ardèche basin the BSS was much better for the ECMWF than for the PEARP EPS, despite the small size of this basin. The streamflow BSS showed very good skill (each of them over 0.85), with the ECMWF-based ESPS better than the PEARP-based ESPS in the Seine basin, and the opposite for the Ardèche basin.

7. Conclusions

The performance of two ensemble streamflow prediction systems was evaluated and compared for short-range scales. These systems were constructed based on the same hydrometeorological model (SIM). They differed only in the meteorological input data (from the PEARP and the ECMWF EPS). The two systems were run for the same 569-day period and were limited to a 2-day forecast period. The usual statistical scores were calculated in order to compare the two systems. The significance of the differences was assessed using a *t* test and a Wilcoxon test for rainfall, and a resampling test for streamflows.

The PEARP system showed higher spread of precipitation than for the ECMWF system, probably because it is mainly dedicated to short-term prediction as opposed to the ECMWF system, which was designed for medium-range forecasts. Despite the better results of the PEARP precipitation forecasts (e.g., on BSSs for all thresholds), the streamflows comparison revealed more complex results. The lower the streamflow threshold, the better the ECMWF-based ESPS performed when compared with the PEARP-based ESPS. With respect to basin size, it is clear that the bigger a basin, the better the EPS and ESPS performed. But, a major difference was that ECMWF rainfall always had lower scores than the PEARP EPS, but this trend was not true for streamflows, especially for large area basins. In our study, the PEARP was used at (almost) its native resolution, while the ECMWF EPS results were used at a coarse resolution. The ECMWF-based ESPS was somewhat spatially more homogeneous than the PEARP-based ESPS. This difference, together with the target time of the two EPSs, explained most of the differences encountered here. Finally, a seasonal study showed low dependency of the scores on season. The only exception was a conditional bias found on scores being lower over the northwestern plains and high mountains for rainfall.

It was very satisfactory to see that, despite low spread and few members, the PEARP-based ESPS was able to predict high flows for small basins. For large basins, especially the Seine, due to the hydrological response functions, the differences between the two ESPSs were not significant.

This study demonstrated the potential of using short-term meteorological ensemble prediction systems to produce short-term ensemble streamflow predictions. However, some limitations appeared during the test. First, the meteorological forecasts were not used optimally. Only temperature and precipitation variables were available in the Météo-France database. Their resolutions were coarser than the real outputs of the ECMWF EPS in order to reduce storage needs. For the other variables, it is assumed that the use of predicted instead of climatological values would lead to increased accuracy (including spread). Second, the rainfall disaggregation (which is critical) differed, and the gradient method used originally by Rousset-Regimbeau et al. (2007) could not be easily transferred to the PEARP EPS. Further research is needed in this field. Third, the ESPS spread was very low when compared to the RMSE. Part of this problem may come from the rainfall disaggregation and the use of climatological variables as input. The question of taking into account model errors by adding perturbations in the hydrological model may also need to be addressed. Finally, we only compared the predictions to a reference run of the hydrological model. To construct a fully operational application, the results should be compared to observed streamflows. For this purpose, an adapted streamflow assimilation system must be set up: work on this subject is planned for the near future.

Acknowledgments. We are grateful to Laurent Descamps (CNRM-GAME/GMAP/RECYF, Laboratoire de Météorologie Dynamique–ENS, Paris) for discussions on the PEARP EPS and statistical tests adapted to ensemble forecasts, and we thank the three anonymous reviewers for their careful revision, which helped improve this manuscript.

APPENDIX A

Analytic Formula to Account for the RPSS and BSS Biases due to Ensemble Sizes

Because of the ensemble sizes, biases are included in the calculation of skill scores, like BSS and RPSS, but can be treated by applying an analytic correction (Weigel et al. 2006). Weigel et al. showed that the fewer the members in an ensemble, the higher is this bias. The unbiased RPSS can be calculated with the formula

$$\mathrm{RPSS} = 1 - \frac{\mathrm{RPS}}{\mathrm{RPS}_{\mathrm{ref}} + D'}$$

with

$$D = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{k} p_i\left(1 - p_i - 2\sum_{j=i+1}^{k} p_j\right)$$

and n being the number of members, K the number of classes for the RPSS (here 10), and p_i the probability in the climatology of being in class i.

For the BSS, the formula can be deducted and becomes

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}} + D'}$$

with

$$D = \frac{1}{n}p(1 - p).$$

APPENDIX B

Statistical Tools

To assess the skills of ensemble predictions (rainfall and streamflows), the classical probabilistic scores used are the rms error (RMSE), the spread (σ), the Brier skill score (BSS), the ranked probability skill score (RPSS), and the false alarm (FA) ratio and hit rate (HR).

The RMSE is defined as

$$\mathrm{RSME} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(m_i - o_i)^2},$$

with N the number of days ($N = 569$), o_i the observation of the data for day i, and m_i the mean of the forecast members. This score quantifies the quality of the mean ensemble forecast.

The spread is computed as

$$\sigma = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_{k,i} - \overline{x}_i)^2},$$

with n the number of members, $\overline{x}_i$ the mean of the ensemble for day i, and $x_{k,i}$ the value of the member k for day i. The spread is the standard deviation of all the members.

The BSS is a score derived from the Brier score (BS) (Brier 1950), which is a widely used statistical score for ensemble predictions. The BS quantifies the ability of an ensemble forecast system to foresee a threshold exceedance:

$$\mathrm{BS} = \frac{1}{N}\sum_{k=1}^{N}(y_k - o_k)^2 \; 0 \le \mathrm{BS} \le 1$$

with y_k the probability of the forecasted event and $o_k = 1$, if the event is observed or $o_k = 0$, if not observed. For a perfect forecast, BS = 0 and BS is close to 1 for bad forecasts.

To make comparisons between the two EPSs, the BSS is used:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}, \quad -\infty \le \mathrm{BSS} \le 1,$$

with $\mathrm{BS}_{\mathrm{ref}}$ the BS of a reference experiment (often the climatology) and the BSS oriented positively, that is, a value close to 1 for better forecasts. A positive value describes an improvement of forecasts when compared with the climatology.

The BS can be decomposed as a sum of terms [see Murphy (1973) for a demonstration], called reliability, resolution, and uncertainty:

$$\mathrm{BS} = \mathrm{BS}_{\mathrm{rel}} + \mathrm{BS}_{\mathrm{res}} + \mathrm{BS}_{\mathrm{unc}}$$

with

$$\mathrm{BS}_{\mathrm{rel}} = \frac{1}{N}\sum_{t=1}^{n+1} N_i(y_i - \overline{o}_i)^2,$$

$$\mathrm{BS}_{\mathrm{res}} = \frac{1}{N}\sum_{t=1}^{n+1} N_i(\overline{o}_i - \overline{o})^2,$$

and

$$\mathrm{BS}_{\mathrm{unc}} = \overline{o}(1 - \overline{o}).$$

Here N_i is the number of forecasts in the category i and

$$\overline{o} = \frac{1}{N}\sum_{k=1}^{N} o_k.$$

TABLE B1. Contingency table of possible events.

Event	Observed	Not observed
Forecasted	a	b
Not forecasted	c	d

Reliability corresponds to the capacity of the system to predict right probabilities; a value of 0 means perfect reliability. Resolution describes the capacity of the system to separate the probability classes; it is oriented positively. The uncertainty is the variance of observations.

The ranked probability score (RPS) is a score derived from the Brier score. It assesses the ensemble predictions on the whole range of values of the parameter considered. The forecasts are divided into $J = 10$ classes in the study, which are determined by the values from 1 to 8 mm day^{-1} and 10 mm day^{-1} for rainfall and by the climatology for streamflows. Here y_j is the probability of the forecasted event for the class j. We define

$$Y_m = \sum_{j=1}^{m} y_j, \quad m = 1, \cdots, J,$$

$$O_m = \sum_{j=1}^{m} o_j, \quad m = 1, \cdots, J.$$

The RPS is then defined as

$$\mathrm{RPS} = \frac{1}{N}\sum_{k=1}^{N}\left[\sum_{m=1}^{J}(Y_m - O_m)^2\right]_k.$$

For ensemble prediction systems, the false alarm rate and hit rate can be defined. It is considered that the event is forecasted if p% of the members predict it; p can be adjusted to the user's needs and is taken equal to 90 in this study. Table B1 is used to define these scores: $a + b + c + d =$ the total number of cases.

So, we can define

$$\text{False Alarms} = \frac{b}{a + b}$$

$$\text{Hit Rate} = \frac{b}{a + c}.$$

APPENDIX C

Hypothesis Tests for Evaluating Ensemble Prediction Systems

Because of the difficulty to interprete the importance of probabilistic scores when comparing two ensemble prediction systems, three hypothesis tests are used in this study: the paired *t* test, the nonparametric Wilcoxon signed-rank test, and a resampling method (Hamill 1999). These tests are appropriate to test the skill of probabilistic forecasts. These tests are applied to the RMSE, spread, BSS for all the thresholds, and RPSS.

The method of computing these two statistic scores with a BSS is given here, while the methods for the other scores can be deduced easily. In this paper, the *t* test and the Wilcoxon test were computed for rainfall scores, and the resampling test was computed for streamflows scores in order to deal with the serial correlation of streamflow.

Hypothesis testing has to be performed on the BS (RPS) summed over all forecast locations rather than using the daily average BSS (RPSS), according to Hamill (1999).

a. The paired t test

For this test, a vector **q** containing the daily differences of the BS is needed. So, for each day, all BSs of the PEARP are summed on all simulated grid points, and all BSs of the ECMWF are then subtracted. A 569-day-long vector of the daily differences is constructed.

Then, the mean of the vector $\overline{\mathrm{BS}}$ and the standard deviation s_{BS} of this vector are computed. The paired *t* test assumes that $\overline{\mathrm{BS}}/(s_{\mathrm{SB}}/\sqrt{N})$ (with $N = 569$ days) is distributed as a *t* variable with $N - 1$ degrees of freedom (Hamill 1999). The location of the sample statistic is then compared against this distribution.

b. The nonparametric Wilcoxon signed-rank test

This test is described by Wilks (1995). The **q** vector of the *N* daily differences of the BS between the two models is used. Then, a vector **z** is filled with the absolute values of the elements of **q**. Then **t** is defined as the vector of the ranks of **z** from the lowest to the highest. For elements in **z** with equal value, **t** is assigned with the averaged value of the ranks, d_0 is the number of daily differences equal to zero in **q**, and d_i is the number of ties in nonsigned ranks other than zero, $i = 1, 2, \ldots, r$ with *r* sets of different ties in **z**. Then the ranks in **t** are given the signs of the original daily differences in **q** in order to build a new vector, **u**. The sum of all *positive* ranks in **u** is denoted U^+.

So, under the null hypothesis, the distribution of the positive ranks is Gaussian with mean:

$$\mu = \frac{N(N+1) - d_0(d_0+1)}{4}.$$

The standard deviation under the null hypothesis is

$$\sigma = \left[\frac{N(N+1)(2N+1) - d_0(d_0+1)(2d_0+1)}{4} - \frac{\sum_{i=1}^{r}(d_i^3 - d_i)}{48}\right]^{1/2}.$$

Finally, the comparison between $(U^+ - \mu)/\sigma$ and the Gaussian distribution is done.

c. A resampling test

Here, two vectors of *N* days containing the daily sums of RPS scores over all grid points are needed (Hamill 1999): $\mathrm{RPS}_{\mathrm{M1}}$ and $\mathrm{RPS}_{\mathrm{M2}}$. Then, a resampled null distribution *I* is created by generating a random *N*-long vector filled with 1 or 2 with equal probability. The value of $\mathrm{RPS}_{\mathrm{M1}}$ and $\mathrm{RPS}_{\mathrm{M2}}$ for day *j* ($j = 1, \ldots, N$) is exchanged when $I_j = 2$. This process is repeated a thousand times, and the values of

$$\sum_{j=1}^{N}(\mathrm{RPS}_{\mathrm{M1}} - \mathrm{RPS}_{\mathrm{M2}})$$

are stocked and define a null distribution. Finally, the initial value of

$$\sum_{j=1}^{N}(\mathrm{RPS}_{\mathrm{M1}} - \mathrm{RPS}_{\mathrm{M2}})$$

is compared to this distribution in order to assess the significance of the difference between both RPSSs.

REFERENCES

Boe, J., L. Terray, F. Habets, and E. Martin, 2006: A simple statistical–dynamical downscaling scheme based on weather types and conditional resampling. *J. Geophys. Res.,* **111,** D23106, doi:10.1029/2005JD006889.

Boone, A., and P. Etchevers, 2001: An intercomparison of three snow schemes of varying complexity coupled to the same land-surface and macroscale hydrologic models. Part I: Local-scale evaluation at an alpine site. *J. Hydrometeor.,* **2,** 374–394.

——, J.-C. Calvet, and J. Noilhan, 1999: Inclusion of a third soil layer in a land surface scheme using the force–restore method. *J. Appl. Meteor.,* **38,** 1611–1630.

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

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

——, J.-R. Bildot, N. Wedi, M. Fuentes, M. Hamrud, G. Holt, and F. Vitart, 2007: The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System). *Quart. J. Roy. Meteor. Soc.,* **133,** 681–695.

Chessa, P. A., and F. Lalaurette, 2001: Verification of ECMWF Ensemble Prediction System forecasts: A study of large-scale patterns. *Wea. Forecasting,* **16,** 611–619.

Courtier, P., C. Freydier, J.-F. Geleyn, F. Rabier, and M. Rochas, 1991: The ARPEGE Project at Météo-France. *Proc. ECMWF Workshop on Numerical Methods in Atmospheric Modelling,* Vol. 2, Reading, United Kingdom, ECMWF, 193–231.

Déque, M., 2007: Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. *Global Planet. Change,* **57,** 16–26, doi:10.1016/j.gloplacha.2006.11.030.

Durand, Y., E. Brun, L. Merindol, G. Guyomarch, B. Lesaffre, and E. Martin, 1993: A meteorological estimation of relevant parameters for snow schemes used with atmospheric models. *Ann. Glaciol.,* **18,** 65–71.

Etchevers, P., C. Golaz, and F. Habets, 2001: Simulation of the water budget and the river flows of the Rhone basin from 1981 to 1994. *J. Hydrol.,* **244,** 60–85.

Habets, F., and Coauthors, 1999a: The ISBA surface scheme in a macroscale hydrological model, applied to the HAPEX-MOBILHY area: Part 1. Model and database. *J. Hydrol.,* **217,** 75–96.

——, P. Etchevers, C. Golaz, E. Leblois, E. Ledoux, E. Martin, J. Noilhan, and C. Ottl, 1999b: Simulation of the water budget and the river flows of the Rhône basin. *J. Geophys. Res.,* **104,** 31 145–31 172.

——, P. LeMoigne, and J. Noilhan, 2004: On the utility of operational precipitation forecasts to serve as input for streamflow forecasting. *J. Hydrol.,* **293,** 270–288.

——, and Coauthors, 2008: The SAFRAN–ISBA–MODCOU hydrometeorological model applied over France. *J. Geophys. Res.,* **113,** D06113, doi:10.1029/2007JD008548.

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

Ledoux, E., G. Girard, G. de Marsilly, and J. Deschenes, 1989: Spatially distributed modeling: Conceptual approach, coupling surface water and groundwater. *Unsaturated Flow Hydrologic Modeling—Theory and Practice,* H. J. Morel-Seytoux, Ed., NATO ASI Series C, Vol. 275, Kluwer, 435–454.

Masson, V., J.-L. Champeaux, F. Chauvin, C. Meriguet, and R. Lacaze, 2003: A global database of land surface parameters at 1-km resolution in meteorological and climate models. *J. Climate,* **16,** 1261–1282.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Morel, S., 2003: Modélisation à l'échelle régionale des bilans énergétique et hydrique de surface et des débits: Application au bassin adour-garonne (Modeling at a regional scale of surface and streamflows energetic and hydric balance sheet). Ph.D. thesis, Université Paul Sabatier, Toulouse, France, 280 pp.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

Nicolau, J., 2002: Short-range ensemble forecasting at Météo-France—A preliminary study. *Proc. Tech. Conf. on Data Processing and Forecasting Systems,* Cairns, QLD, Australia, WMO/Commission on Basic Systems. [Available online at http://www.wmo.ch/pages/prog/www/DPS/TC-DPFS-2002/Papers-Posters/Topic1-Nicolau.pdf.]

Noilhan, J., and S. Planton, 1989: A simple parameterization of land surface processes for meteorological models. *Mon. Wea. Rev.,* **117,** 536–549.

——, and J.-F. Mahfouf, 1996: The ISBA land surface parameterization scheme. *Global Planet. Change,* **13,** 145–159.

Quintana-Seguí, P., and Coauthors, 2008: Analysis of near-surface atmospheric variables: Validation of the SAFRAN analysis over France. *J. Appl. Meteor. Climatol.,* **47,** 92–107.

Ramos, M.-H., J. Bartholomes, and J. T. del Pozo, 2007: Development of decision support products based on ensemble forecasts in the European flood alert system. *Atmos. Sci. Lett.,* **8,** 113–119.

Roulin, E., and S. Vannitsem, 2005: Skill of medium-range hydrological ensemble predictions. *J. Hydrometeor.,* **6,** 729–744.

Rousset, F., F. Habets, E. Gomez, P. Le Moigne, S. Morel, J. Noilhan, and E. Ledoux, 2004: Hydrometeorological modeling of the Seine basin using the SAFRAN–ISBA–MODCOU system. *J. Geophys. Res.,* **109,** D14105, doi:10.1029/2003JD004403.

Rousset-Regimbeau, F., 2007: Modélisation des bilans de surface et des débits sur la France, application à la prévision d'ensemble des débits (Surface balance and streamflow modeling on France, application to streamflow ensemble prediction). Ph.D. thesis, Université Paul Sabatier, Toulouse, France, 226 pp.

——, F. Habets, E. Martin, and J. Noilhan, 2007: Ensemble streamflow forecasts over France. *ECMWF Newsletter,* No. 111, ECMWF, Reading, United Kingdom, 21–27.

Schaake, J., and Coauthors, 2007: Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrol. Earth Syst. Sci. Discuss.,* **4,** 655–717.

Tracton, M. S., and E. Kalnay, 1993: Ensemble forecasting at NMC: Operational implementation. *Wea. Forecasting,* **8,** 379–398.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.,* **135,** 118–124.

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

Wood, A. W., A. Kumar, and D. P. Lettenmaier, 2005: A retrospective assessment of National Centers for Environmental Prediction climate model-based ensemble hydrologic forecasting in the western United States. *J. Geophys. Res.,* **110,** D04105, doi:10.1029/2004JD004508.

Zorita, E., J. P. Hughes, D. P. Lettemaier, and H. von Storch, 1995: Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. *J. Climate,* **8,** 1023–1042.