Postprocessing of Ensemble Precipitation Predictions with Extended Logistic Regression Based on Hindcasts

EMMANUEL ROULIN AND STÉPHANE VANNITSEM

Institut Royal Météorologique de Belgique, Brussels, Belgium

(Manuscript received 4 March 2011, in final form 26 August 2011)

ABSTRACT

Extended logistic regression is used to calibrate areal precipitation forecasts over two small catchments in Belgium computed with the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS) between 2006 and 2010. The parameters of the postprocessing are estimated from the hindcast database, characterized by a much lower number of members (5) than the EPS (51). Therefore, the parameters have to be corrected for predictor uncertainties. They have been fitted on the 51-member EPS ensembles, on 5-member subensembles drawn from the same EPS, and on the 5-member hindcasts. For small ensembles, a simple ''regression calibration'' method by which the uncertain predictors are corrected has been applied. The different parameter sets have been compared, and the corresponding extended logistic regressions have been applied to the 51-member EPS. The forecast probabilities have then been validated using rain gauge data and compared with the raw EPS. In addition, the calibrated distributions are also used to modify the ensembles of precipitation traces.

The postprocessing with the extended logistic regression is shown to improve the continuous ranked probability skill score relative to the raw ensemble, and the regression calibration to remove a large portion of the bias in parameter estimation with small ensembles. With a training phase limited to a 5-week moving window, the benefit lasts for the first 2 forecast days in winter and the first 5 or 6 days in summer. In general, substantial improvements of the mean error and of the continuous ranked probability score have been shown.

1. Introduction

The outputs of numerical weather prediction (NWP) models are increasingly used to estimate future precipitation amounts or probabilities. To deal with the large uncertainties associated with initial conditions and model errors, major operational model centers like the European Centre for Medium-Range Weather Forecasts (ECMWF) are running Ensemble Prediction Systems (EPS). Model errors maybe accounted for by appropriate specification and correction using statistics from previous forecasts and corresponding observations.

Applequist et al. (2002) have compared five different linear and nonlinear statistical models to produce 24-h probabilistic quantitative precipitation forecasts for accumulations exceeding thresholds of up to 2.5 mm. They have found that the logistic regression with cumulated precipitation forecast (0–24 h) and relative humidity at +12 h as predictors performs better. Hamill et al. (2004) also have shown a gain in skill and reliability by using the logistic regression using ensemble mean precipitation as predictor (6–10 day and 8–14 day, tercile probabilities). Hamill and Whitaker (2006) have shown that analog techniques are similar in skill to the logistic regression for the calibration of probabilistic forecasts of 24-h precipitation amount. For the logistic regression, they have used the square root of the ensemble mean precipitation and the column precipitable water ensemble mean as predictors. Wilks and Hamill (2007) have compared the logistic regression with the nonhomogeneous Gaussian regression and the Gaussian ensemble dressing. They have shown that logistic regression with the ensemble mean as the only predictor is best for medium-range precipitation forecasts.

Wilks (2009) has noted that fitting logistic regression for a set of thresholds has some drawbacks: a large number of parameters have to be estimated, probabilities at intermediate thresholds have to be interpolated, and separate equations for different thresholds may lead to probability forecasts inconsistent with each other.

Corresponding author address: Emmanuel Roulin, Institut Royal Météorologique de Belgique, 3 Avenue Circulaire, B-1180 Brussels, Belgium.
E-mail: emmanuel.roulin@meteo.be

DOI: 10.1175/MWR-D-11-00062.1

© 2012 American Meteorological Society

CORRIGENDUM

EMMANUEL ROULIN AND STÉPHANE VANNITSEM

Institut Royal Météorologique de Belgique, Brussels, Belgium

An editing error occurred in the caption of Fig. 4 of Roulin and Vannitsem (2012). The description of the curves were given in the reverse order. The figure and caption have been corrected in the version displayed below.

The staff of *Monthly Weather Review* regrets any inconvenience this error has caused.

REFERENCE

Roulin, E., and S. Vannitsem, 2012: Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Mon. Wea. Rev.,* **140,** 874–888.

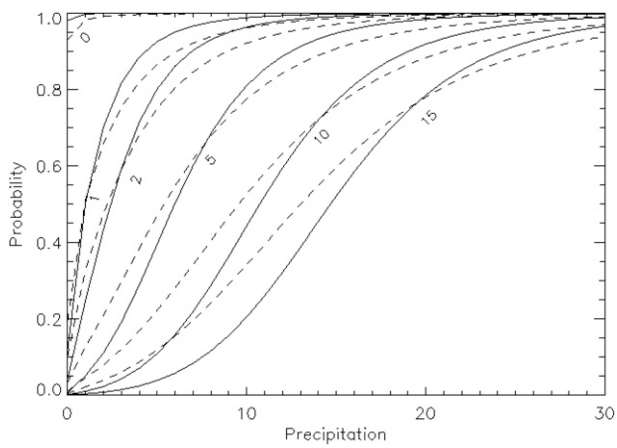

FIG. 4. Probability distribution [Pr($A \leq q$)] based on the extended logistic regressions (EPS-FULL) for the Ourthe during winter at forecast days 3 (solid line) and 6 (dashed line). (left to right) The curves correspond to ensemble averages of 0, 1, 2, 5, 10, and 15 mm day^{-1}.

DOI: 10.1175/MWR-D-12-00061.1

© 2012 American Meteorological Society

Therefore, he has introduced the extended logistic regression by which the logistic regression is estimated once by including the threshold itself as predictor. He has compared the Brier scores at different thresholds covering the distribution and has shown that these scores are similar for both methods if a long training period is used and that the extended version performs better when a short training period is used. Studying precipitation over the Netherlands and using an experimental reforcasting dataset produced by the ECMWF (see below), Schmeits and Kok (2010) have shown a similar skill improvement for the extended logistic regression and a modified version of the Bayesian model averaging for the first five forecast days.

To determine the error characteristics corresponding to a specific NWP model, it has been proposed to perform hindcasts of past meteorological situations using the same model. Such hindcast datasets have been built and tested for postprocessing for the National Centers for Environmental Prediction (NCEP) Global Forecast System (200-km grid spacing, 15 members, every day during a 25-yr period; Hamill et al. 2004; Hamill and Whitaker 2006; Wilks and Hamill 2007), and for the ECMWF EPS (80-km grid spacing, 15 members, weekly, during a 20-yr period; Hagedorn et al. 2008; Hamill et al. 2008; Schmeits and Kok 2010). Since March 2008, the ECMWF provides hindcasts for the operational EPS (five members, weekly, during an 18-yr period; Hagedorn 2008).

Even if the allocation of computer resources either in the number of past forecast situations or in the ensemble size has been analyzed by comparing the improvement in skill gained with the postprocessing (Hagedorn 2008), the impact of the number of members on the values of the parameters of the fitted statistical models has not been explored. Indeed the ensemble mean used as predictor is estimated with a large uncertainty because of the small number of members, inducing a bias in the estimate of the slope of a linear regression toward zero. This is commonly referred to as attenuation (Carroll et al. 2006). In the case of an additive error in a simple linear model, this bias can be accounted for with the so-called reliability ratio (the ratio of the variance of the predictors to the sum of this variance and the variance of the error). The bias can be corrected at the price of an increased variance error in the estimated slope. In multilinear or nonlinear problems, the effect of errors may be even more complex (Carroll et al. 2006). In logistic regression when the predictor is measured with additive error, attenuation does not always occur (Stefanski and Carroll 1985), but is typical.

In the present study, we apply a simple approach called regression calibration. In the case of replicate data, this method consists in replacing the predictor measured with error with a best linear approximation (Carroll and Stefanski 1990; Gleser 1990). Regression calibration has been applied to logistic regression by Rosner et al. (1989, 1990, 1992). Using simulation experiments, Thoresen and Laake (2000, 2003) have shown that the regression calibration compares favorably with other methods like the maximum likelihood.

Postprocessing of single ensemble precipitation predictions may also be provided. Indeed, bias correcting each member by adjusting for the cumulated distribution function has been implemented at NCEP in the mid-2000s (as mentioned by Hamill and Whitaker 2006). Other examples range from the correction using the rank histogram (Hamill and Colucci 1998), methods based on analogs (Hamill and Whitaker 2006), Bayesian techniques (e.g., Reggiani and Weerts 2008), and Bayesian model averaging (e.g., Sloughter et al. 2007; Schmeits and Kok 2010).

The first aim of this study is to investigate the extended logistic regression for the postprocessing of ECMWF-EPS precipitation forecasts using information contained in the five-member hindcasts. The second goal is to take advantage of the extended logistic regression as a simple method to postprocess the ensembles members so that ensemble precipitation traces for the medium range are reconstructed. The data and the methodology are first described in section 2. In section 3, the results are discussed and the conclusions are presented in section 4.

2. Data and method

a. Data

Postprocessing precipitation forecast data—like their verification—may be performed at different scales; for example, by interpolating NWP grid data to a finer grid (e.g., Hamill and Whitaker 2006) or to the rain gauge locations (e.g., Rodwell 2006) or by upscaling rain gauge data to $1° \times 1°$ grid boxes (e.g., Schmeits and Kok 2010). In the present study, daily totals of precipitation averaged over two small test catchments in Belgium are considered (Fig. 1). Observations are based on rain gauge data interpolated using the Thiessen method and forecasts of the NWP are weighted according to the grid cells' coverage on the catchment. The two catchments are the Kleine Gete (or Petite Gette, hereafter Gete) at Budingen, Belgium (area 276 km^2, elevation 25–170 m, and annual average precipitation 764 mm) and the Ourthe Orientale (hereafter Ourthe) at Mabompré, Belgium (area 317 km^2, elevation 280–650 m, and annual average precipitation 1029 mm).

For this study, we have analyzed EPS forecasts issued daily at 0000 UTC during a period when the ECMWF

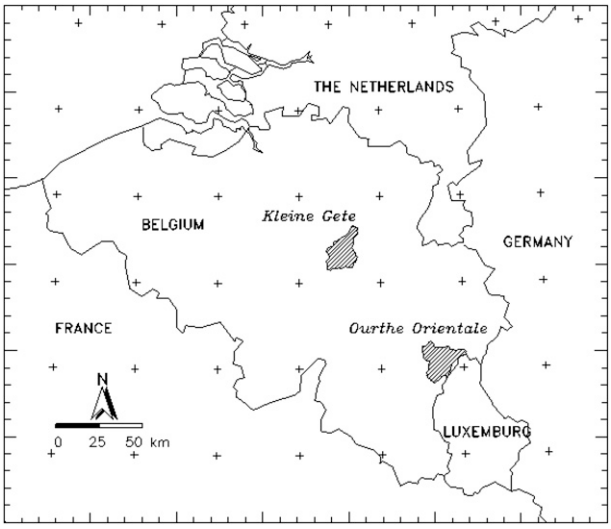

FIG. 1. Location map of the two test catchments in Belgium. The grid of the EPS is represented with the plus signs.

model had a spatial resolution of about 50 km and after a model change having a strong impact on precipitation forecast skill [i.e., the Integrated Forecast System cycle (cy31r1)], namely, from 12 September 2006 to 25 January 2010. Model biases might have changed even during this reduced period, but these do not seem to affect much precipitation forecasts according to ECMWF reports on preoperational tests. We have used 51-member ensembles: 1 control and 50 perturbed members.

Besides the operational EPS, a hindcast ensemble forecast has become available since 13 March 2008 (Hagedorn 2008). Every week, five forecasts (one control and four perturbed) starting at the same date for the past 18 yr are performed using the same model as for the operational EPS. In this study, we have used the hindcasts until 21 January 2010 (i.e., the last hindcast date before a major change in spatial resolution). (The different experiments based on these data are summarized at the end of this section together with Table 2, but let us now focus on the technique proposed.)

b. *Extended logistic regression and ensemble size correction*

The extended logistic regression relates the probability that the areal precipitation A is lower or equal to a threshold q to the predictor x and to the threshold itself:

$$F_A(q) = \Pr(A \le q) = H[f(x) + g(q)], \qquad (1)$$

where H is the logistic distribution function $H(t) = [1 + \exp(-t)]^{-1}$.

The ensemble mean of the fourth root transformation of areal precipitation forecasts a has been chosen as predictor:

TABLE 1. Areal precipitation thresholds (percentiles in mm day^{-1}) used for the extended logistic regression (Gete: 1978–2010 and Ourthe: 1978–2008).

Catchment	P50	P70	P80	P90	P95	P98	P99
Gete	0.23	1.7	3.5	6.6	9.8	14.6	18.7
Ourthe	0.43	2.7	4.9	8.6	12.9	18.2	22.8

$$x = \overline{(a^{1/4})}. \qquad (2)$$

The functions $f(x) = \beta_0 + \beta_1 x$ and $g(q) = \beta_2 q^{1/2}$ define a vector of three parameters, $\boldsymbol{\beta}$, estimated by maximizing the likelihood function:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N}\sum_{l=1}^{M}\{y_{il}(\beta_0 + \beta_1 x_i + \beta_2 q_l^{1/2}) - \ln[1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 q_l^{1/2})]\}, \qquad (3)$$

where N is the number of realizations in the calibration dataset, M is the number of thresholds, and the observations y_{il} are equal to 1 if the observed areal precipitation does not exceed the threshold q_l and are, otherwise, equal to zero. The precipitation thresholds have been selected such that they cover the upper quantiles of the distribution ($M = 7$; Table 1). Alternative power transformations and the use of ensemble spread as an additional predictor have been tested without improvement of the likelihood function. Schmeits and Kok (2010) have found that selecting the average of the transformed (square root) ensemble members as predictor performed slightly better than the transformed ensemble averages as in other studies (e.g., Wilks 2009), but this choice has been motivated here to enable the implementation of the ensemble size correction.

One important error in estimating the predictor x is due to the finite size of the ensemble:

$$\overline{w}_i = x_i + u_i, \qquad (4)$$

for the ith realization with $\overline{w}_i = (1/K)\sum_{j=1}^{K} a_{ij}^{1/4}$, K the size of the ensemble, and u_i the uncertainty on the ensemble mean evaluation, assumed to be a random process with $E(u) = 0$ and $\mathrm{Var}(u) = \sigma_u^2 = \langle\sigma_w^2\rangle/K$. The angle brackets denote the average over the N realizations. Here σ_w^2 is the variance of the power transformed ensemble members. This error may lead to inconsistencies in the estimation of $\boldsymbol{\beta}$ if $\overline{w}$ is used directly in Eq. (1) (Gleser 1990). The basis of regression calibration is the replacement of x by the regression of x on $(\overline{w}, q)$. A linear approximation to the calibration function is given in Carroll et al. (2006). Noting that we have only

one predictor x with error that is not correlated with q, this approximation can be written as

$$E(x_i \mid \overline{w}_i) \approx \mu_{\overline{w}} + \lambda(\overline{w}_i - \mu_{\overline{w}}), \tag{5}$$

where

$$\lambda = \frac{\sigma^2_{\overline{w}} - \sigma^2_u}{\sigma^2_{\overline{w}}}, \tag{6}$$

is the reliability ratio, $\mu_{\overline{w}} = \langle \overline{w} \rangle$, and $\sigma^2_{\overline{w}} = \langle (\overline{w} - \mu_{\overline{w}})^2 \rangle$. Then, $E(x_i \mid \overline{w}_i)$ is used in place of x_i in the extended logistic regression analysis.

c. Modification of the ensembles

In some applications like hydrological ensemble predictions, precipitation scenarios over the forecast range are needed rather than separate probability distributions at each forecast step. In this section, a methodology is proposed to modify the raw ensembles according to the postprocessing obtained with the extended logistic regression. Let us first consider the distribution function $F_A(q)$ whose verification should, on average, improve compared to the raw ensemble and how the ensemble members are modified. The form of the logistic equation and the choice of the functions f and g impose some of the distribution characteristics such as the mean and the variance. The variance of a logistic distribution is driven by the parameter β_2 but because of the change of variable in the function $g(q)$ [see Eq. (1)], the variance of the corrected ensembles is proportional to its mean as it can be seen with the following first-order approximations based on the properties of the logistic distributions and on Taylor expansion around $\mu_{q'}$ (e.g., Casella and Berger 2002) with $q = h(q') = q'^2$:

$$\mu_q \approx h(\mu_{q'}) = \left(\frac{\beta_0 + \beta_1 x}{\beta_2}\right)^2, \tag{7}$$

$$\sigma^2_q \approx [h'(\mu_{q'})]^2 \sigma^2_{q'} = \frac{4\pi^2(\beta_0 + \beta_1 x)^2}{3\beta_2^4}, \tag{8}$$

Therefore, $\sigma^2_q(x) \propto \mu_q(x)$. These equations are only valid for ensembles with no member forecasting zero precipitation. The probability of no precipitation is given by

$$p_0 = [1 + \exp(-\beta_0 - \beta_1 x)]^{-1}. \tag{9}$$

The raw ensembles are modified so that their probability distributions correspond to the fitted logistic function as follows. First, for a given forecast day, the ensemble members $a(j)$ are sorted with increasing predicted precipitation $a(k)$. Then, the probability is assigned:

$$p(k) = \Pr[A \le a(k)] = \frac{k + (1/2)}{K + 1}. \tag{10}$$

In this equation, the uncertainty due to the finite ensemble size is taken into account by adding an additional fictitious member distributed equally between the occurrence of the event and the nonoccurrence (Roulston and Smith 2002; Katz and Ehrendorfer 2006). Other probability estimates are discussed in Folland and Anderson (2002). Kernel density estimation (e.g., Peel and Wilson 2008) could also be appropriate in the context of small ensembles. In the present study, Eq. (10) is applied on 51-member ensembles and it has the interesting property of excluding probabilities of zero or one. Using the ensemble mean of transformed areal precipitation forecasts x, the precipitation value of the members ranked k is modified by inverting the logistic function as

$$a'(k) = \left\{ \frac{\ln\left[\dfrac{1 - p(k)}{p(k)}\right] - \beta_0 - \beta_1 x}{\beta_2} \right\}^2. \tag{11}$$

If the probability is lower or equal to the intercept of the logistic function [i.e., $p(k) \le p_0$], then the precipitation is set to zero (see Fig. 2, left). If zero precipitation in the raw ensemble is forecasted at a higher frequency than the one provided by the logistic distribution [i.e., $p(k_0) > p_0$], only a fraction $r_0 = p_0/p(k_0)$ of the corresponding k_0 members is randomly set to zero. The remaining members are randomly assigned to a value between p_0 and $p(k_0)$ and the precipitation is given by Eq. (11) (see also Fig. 2).

Finally, the modified values are reassigned to the corresponding members so that the ensemble traces $a(j, t)$ where t stands for lead time can be reconstructed (e.g., for later use in hydrological predictions).

d. Degree of mass balance

Several alternative correction methods are available. A comparative study of these methods is beyond the scope of the present work. However, the results obtained with a simple correction for a multiplicative bias or degree of mass balance (DMB; McCollor and Stull 2008) will highlight some aspects of the ensemble modification. This DMB factor is determined from the calibration dataset as the ratio of the average of the ensemble mean and the average of the corresponding observed areal precipitation for lead time t:

$$\mathrm{DMB}(t) = \langle \overline{a}(t) \rangle / \langle A(t) \rangle. \tag{12}$$

Then the correction is applied to the member j of the ensemble as

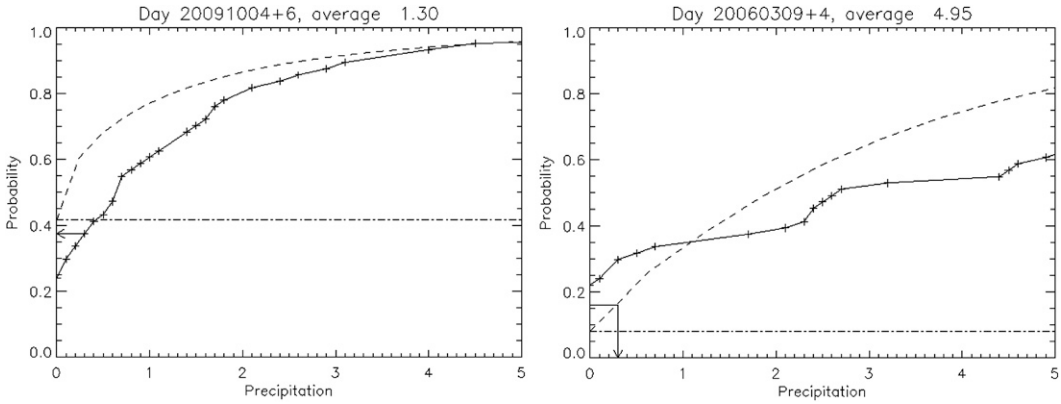

FIG. 2. Remapping of the members of the raw ensembles (continuous line with pluses) onto the fitted logistic distribution (dashed line, EPS-FULL). (left) The situation with $p(k) \leq p_0$ (Gete) and (right) the situation with $p(k_0) > p_0$ (Ourthe), where p_0 is the intercept of the logistic (horizontal mixed line) and k_0 is the number of members with zero forecast precipitation.

$$a'(j,t) = a(j,t)/\mathrm{DMB}(t). \tag{13}$$

e. Experiments

Eight experiments have been conducted and are summarized in Table 2. The first four are "in sample" (the same 51-member EPS are used for training and verification) and in the last four experiments, 5-member hindcasts are used for training, which are independent from the 51-member EPS used for verification. The first four are aimed at highlighting the effect of ensemble size on the logistic regression and test a correction method. The in-sample situation is used to show the maximum correction that the postprocessing can achieve. In the last four experiments, the extended logistic regression is fitted on the base of the hindcasts in order to postprocess the EPS, which is the main topic of the paper. In particular, the last two experiments reproduce operational settings, using information available at the time ensemble predictions are issued.

In the first four experiments (labeled "EPS" in Table 2), there is a pool of $N = 665$ ensemble forecasts or realizations for hydrological winter (October–March) and $N = 568$ for summer (April–September). In the first experiment, the fit is based on the raw 51-member ensembles (FULL). To test the role of ensemble size on the parameter estimation, bootstrap samples of 51 members (BOOT-51), 15 members (not shown), and 5 members (BOOT-5) have been drawn with replacement from the original 51-member ensembles. These analyses constitute experiments 2, 3, and 4. In the last one, the regression calibration (RC) method has been applied to small ensembles to reduce the bias in the parameters due to the uncertainty in the ensemble mean used as predictor.

The remaining experiments use hindcasts (HIN) for training. In experiments 5 and 6, hindcasts are used pooled (POOL) for each season. The 46 dates corresponding to hydrological winter and 52 to hydrological summer multiplied by 18 yr give $N = 828$ and $N = 936$ ensemble hindcasts, respectively. To mimic a more

TABLE 2. List of experiments. The names characterize which ensemble has been used to fit the parameters of the extended logistic regression. EPS refers to the daily (0000 UTC) operational 51-member ECMWF EPS ensembles; HIN to the weekly operational 5-member ECMWF hindcasts. FULL refers to the unchanged EPS; BOOT to bootstrap samples in which 51 or 5 members have been drawn from the full EPS. RC refers to the regression calibration method applied to the ensemble mean; POOL to the case where all hindcasts available are pooled; WIN5 and WIN7 to the cases where hindcasts corresponding to a moving window of 5 and 7 weeks, respectively, are used. Parameter sets are defined per forecast day, season, and basin for experiments 1–6 and per forecast day, week, and basin for experiments 7 and 8.

No.	Name	Training period	Verification period
1	EPS-FULL	12 Sep 2006–25 Jan 2010	12 Sep 2006–25 Jan 2010
2	EPS-BOOT-51	12 Sep 2006–25 Jan 2010	12 Sep 2006–25 Jan 2010
3	EPS-BOOT-5	12 Sep 2006–25 Jan 2010	12 Sep 2006–25 Jan 2010
4	EPS-BOOT-5-RC	12 Sep 2006–25 Jan 2010	12 Sep 2006–25 Jan 2010
5	HIN-POOL	13 Mar 1990–21 Jan 2009	12 Sep 2006–25 Jan 2010
6	HIN-POOL-RC	13 Mar 1990–21 Jan 2009	12 Sep 2006–25 Jan 2010
7	HIN-WIN5-RC	5 weeks × 18 yr	24 Mar 2008–10 Jan 2010
8	HIN-WIN7-RC	7 weeks × 18 yr	7 Apr 2008–10 Jan 2010

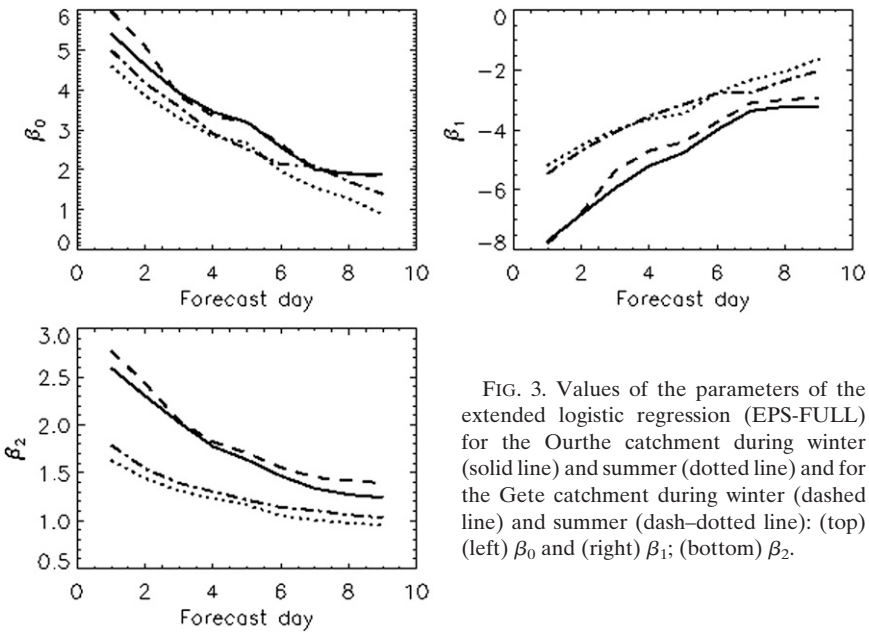

FIG. 3. Values of the parameters of the extended logistic regression (EPS-FULL) for the Ourthe catchment during winter (solid line) and summer (dotted line) and for the Gete catchment during winter (dashed line) and summer (dash–dotted line): (top) (left) β_0 and (right) β_1; (bottom) β_2.

realistic setting (Hagedorn 2008), the hindcasts are also pooled during a period of 5 weeks (WIN5, experiment 7), two before the target week and two after (i.e., for the 18 past years, a training dataset of 90 ensemble hindcasts). A 7-week moving window (WIN7, experiment 8) has been tested as well—four weeks before and two after the target week (i.e., 126 hindcasts in total). The use of moving windows allows minimizing the impact of model changes.

In Table 2, it can be seen that the verification of experiments 1–6 corresponds to the whole selected period. For experiments 7 and 8, since each parameter set is valid for 7 days, the verification period starts 3 days before the date of the third (WIN5) or the fifth (WIN7) hindcast and ends 3 days after the date of the antepenultimate hindcast. Therefore the verification period is slightly shorter than the period of availability of hindcasts. The results are averaged and presented for the winter and for the summer.

For the verification of the ensemble forecasts, the following quantities have been computed: the mean error (ME), the mean-squared error (MSE), the ensemble spread, and the continuous ranked probability score (CRPS). The continuous ranked probability skill score (CRPSS) has also been computed using the raw EPS as reference. For the CRPS of raw or modified ensembles, we followed the method described in Hersbach (2000). For the distributions resulting from the extended logistic regression, we integrate numerically:

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} [F_A(x) - F_{A_O}(x)]^2 \, dx, \tag{14}$$

where $F_A(x)$ is given by Eq. (1) and, with observed areal precipitation A_O,

$$F_{A_O}(x) = \begin{cases} 0, & x < A_O \\ 1, & x \geq A_O \end{cases}. \tag{15}$$

Using Eq. (14), it is easily seen that the CRPS is equivalent to the integral of the Brier score over the set of possible thresholds. The Brier skill score (BSS) relative to the sample uncertainty and its decomposition into resolution and reliability terms (e.g., Roulin and Vannitsem 2005) has been analyzed for the range of thresholds and some qualitative conclusions will also be reported on in the next section.

3. Results and verification

a. Logistic regression parameters

The parameters of the extended logistic regression fitted on the full EPS are presented in Fig. 3 for the two catchments, both for winter and summer. As the forecast range increases, β_0 decreases, β_1 increases, and β_2 decreases implying that the distributions are more dispersed and their means decrease as reflected in Fig. 4.

It is worth noting here that if the raw ensembles are first corrected for the multiplicative biases (or DMB) before fitting the logistic regression, the parameter β_2 is shifted upward (downward) if the correction factor is smaller (greater) than 1, and the distributions are shifted to the left (to the right).

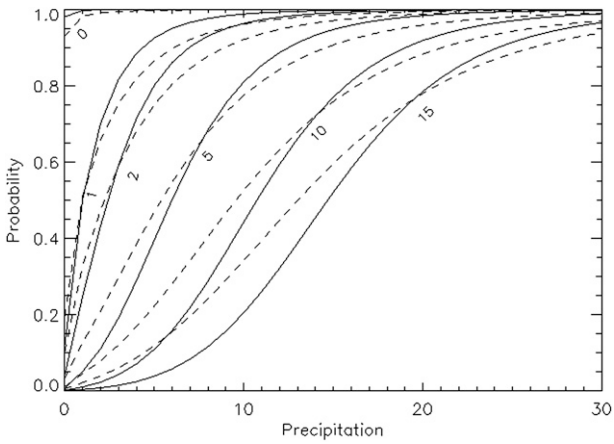

FIG. 4. Probability distribution [Pr($A \leq q$)] based on the extended logistic regressions (EPS-FULL) for the Ourthe during winter at forecast days 3 (solid line) and 6 (dashed line). (left to right) The curves correspond to ensemble averages of 15, 10, 5, 2, 1, and 0 mm day^{-1}.

The extended logistic regression has then been fitted to subensembles of 51, 15 (not shown), and 5 members drawn from the 51-member EPS. In Fig. 5, the parameter values obtained on the subensembles of five members are averages over 100 bootstraps and are compared with the values obtained with the full ensembles. The slope β_1 is reduced for a small ensemble size and the difference increases with lead time as the uncertainty on the subensemble mean increases. When the regression calibration method is applied on the subensembles [Eq. (5)], the attenuation of the slope β_1 is partly corrected (Fig. 5). A similar result is found for the intercept β_0. Note that the correction of the intercept is greater than the correction of the slope. Obviously, the quantiles q_l are not correlated with the x_i, and the regression calibration does not change the values of β_2. Similar results are found for the two catchments and both seasons, showing that the effect of small ensemble is appropriately corrected with the regression correction method. Note that the slope β_1 could also have been directly corrected with the reliability ratio λ [Eq. (6)], as it is common in epidemiology studies aimed at quantifying the effect of predictors (Rosner et al. 1992). Here, we are interested in all parameters. The parameter values obtained with the 51-member bootstrap ensembles are very close to the values obtained on the full ensemble showing a very small residual bias. Finally, parameter values obtained with 15-member bootstrap ensembles (not shown) are intermediate and somewhat closer to the values obtained on the full ensemble so that the issues of bias in parameter estimation would be less stringent with the 15-member hindcasts.

For the five-member hindcasts, the parameter values obtained with and without the regression calibration are compared to those fitted on the full EPS ensembles (see Fig. 6 for the Ourthe during winter). The effect of using raw hindcasts is to decrease β_0 and increase β_1 as for the sensitivity experiment presented in Fig. 5. However, the parameter β_2 is decreased and the parameter β_1 is increased even at the beginning of the forecast. A shift of β_2 can also be detected in the sensitivity experiment (Fig. 5) but to a lesser extent. This statistical difference

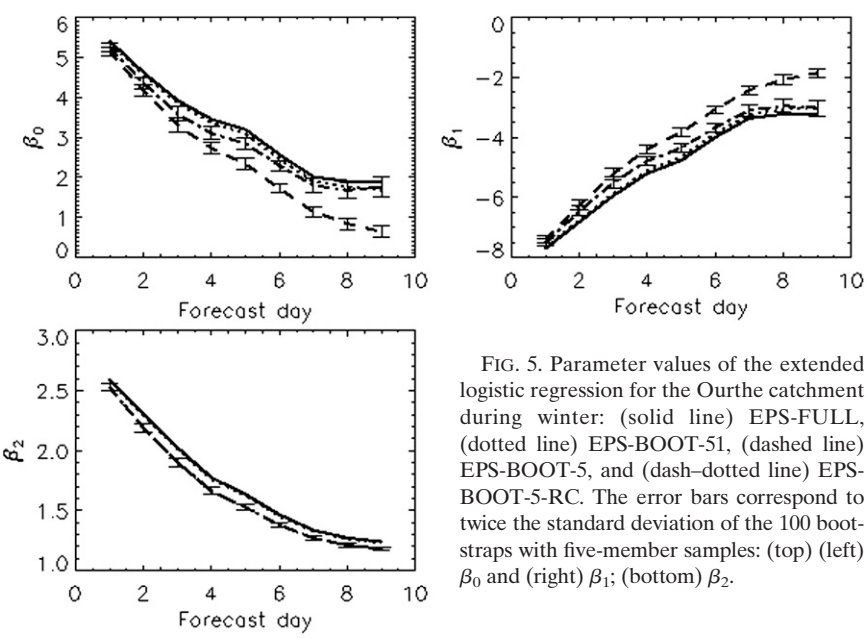

FIG. 5. Parameter values of the extended logistic regression for the Ourthe catchment during winter: (solid line) EPS-FULL, (dotted line) EPS-BOOT-51, (dashed line) EPS-BOOT-5, and (dash–dotted line) EPS-BOOT-5-RC. The error bars correspond to twice the standard deviation of the 100 bootstraps with five-member samples: (top) (left) β_0 and (right) β_1; (bottom) β_2.

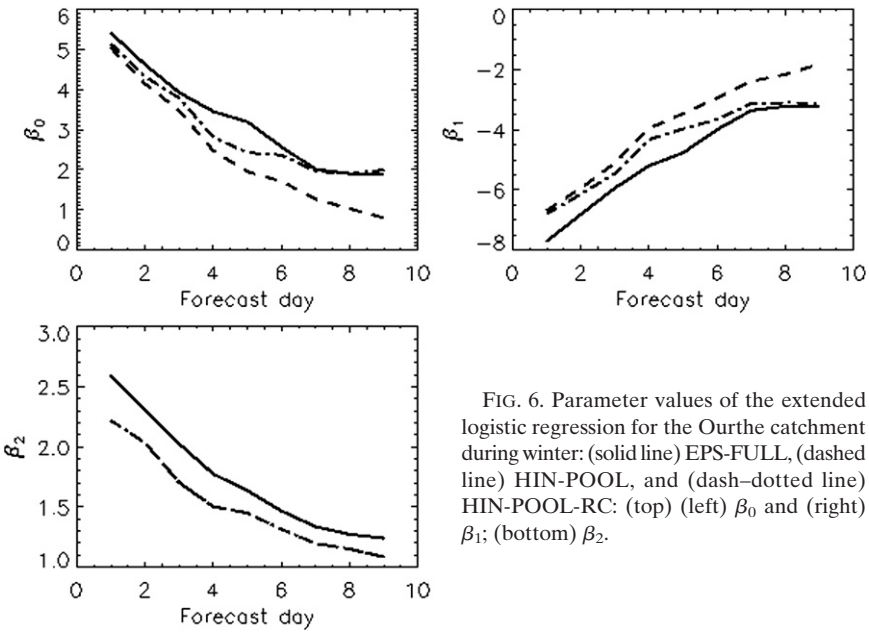

FIG. 6. Parameter values of the extended logistic regression for the Ourthe catchment during winter: (solid line) EPS-FULL, (dashed line) HIN-POOL, and (dash–dotted line) HIN-POOL-RC: (top) (left) β_0 and (right) β_1; (bottom) β_2.

between the EPS and the hindcasts is also found for the other catchment with winter data. For summer, the difference is less important. The regression calibration method applied to the hindcasts generally improves the parameter values toward their optimal values obtained on the full EPS. However, the shift of β_2 for winter remains unchanged since there is no correlation between the quantiles and the ensemble mean and this effect is still to be clarified.

b. *Verification of calibrated probabilities*

The overall benefit of the extended logistic regression is assessed with the CRPSS with respect to the raw ensemble CRPS. First the maximum attainable skill score is evaluated using the parameter fitted on the EPS ensembles used for the validation itself. The skill score is positive over the entire forecast range (Fig. 7) showing the usefulness of calibrating the forecast probabilities with the extended logistic regression. If only five-member ensembles are available, the skill score is deteriorated as compared to a large ensemble (dashed line of Fig. 7). For the test case of the Ourthe catchment during winter, the skill score is even negative beyond the sixth forecast day. Applying the regression-calibration method completely recovers the skill score values as obtained using the full ensembles, confirming the usefulness of the approach.

The postprocessing of EPS based on the regression of the hindcast dataset is displayed in Fig. 8. It improves the CRPS compared to its value on the raw ensemble except for the Ourthe catchment during winter from the seventh forecast day onward. If the regression calibration method is applied to take into account the small ensemble size of the hindcasts, the skill scores are improved most of the time.

When the extended logistic regression is trained on the hindcasts corresponding to a moving window of five weeks, the skill score is degraded especially for winter as illustrated in Fig. 9. The postprocessing has skill during the first 2 forecast days for winter and during the first 5 or 6 days for summer. Increasing the window from 5 to 7 days marginally improves the skill score. This clearly illustrates the necessity to have as much realizations as possible.

c. *Modified ensembles*

The ensemble members are modified based on the extended logistic regression as described in section 2c. The examples presented in Figs. 10 and 11 have been obtained using the parameters of the extended logistic regression fitted on the EPS ensembles. The logistic distribution corresponding to the ensemble mean is compared with that of the raw ensemble (Fig. 10). A new value is assigned to the member(s) according to its (their) probability as illustrated by the arrows in Fig. 10. Then the precipitation traces are reconstructed and the postprocessed ensembles may be compared with the raw ensembles in Fig. 11. For this particular forecast, the postprocessing has improved the CRPS averaged over the entire range of 9 days. However, it is obvious that for the first 6 days, the raw ensembles are performing very well and the postprocessing slightly degrades the CRPS.

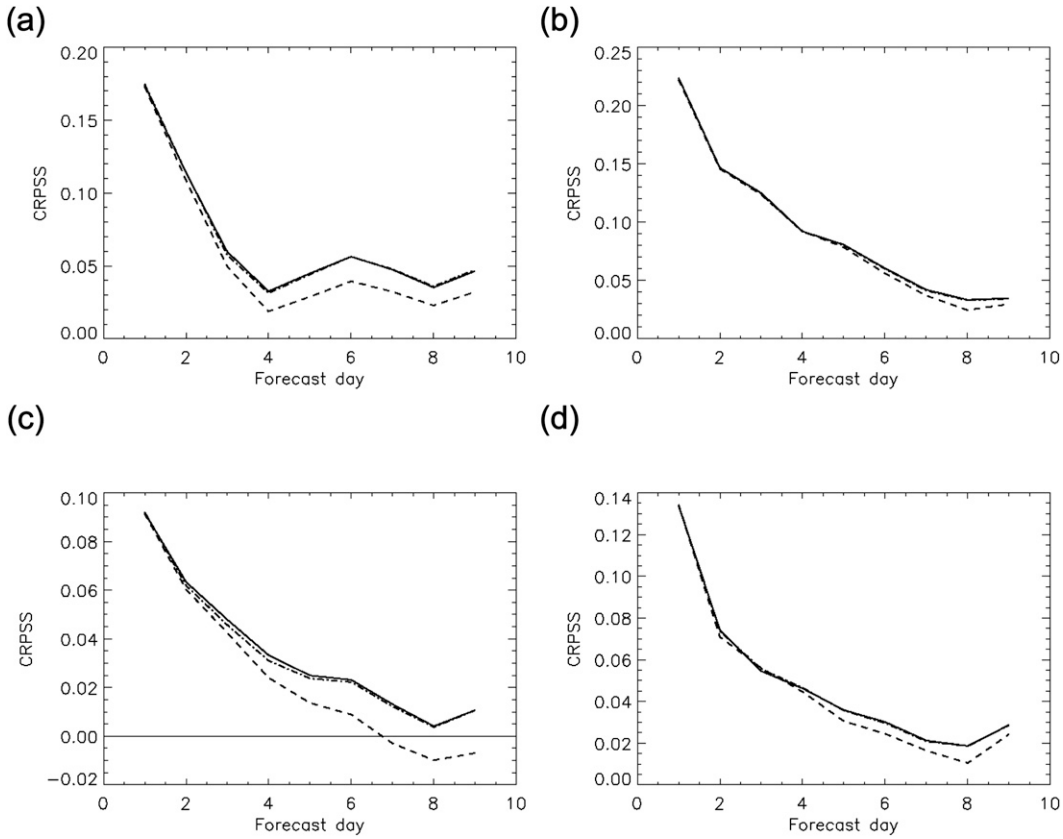

FIG. 7. CRPSS of the probability distributions estimated with the extended logistic regression relative to the raw ensembles: (solid line) EPS-FULL, (dashed line) EPS-BOOT-5, and (dash–dotted line) EPS-BOOT-5-RC. Test cases: (a),(b) Gete catchment and (c),(d) Ourthe catchment for (a),(c) winter and (b),(d) summer.

This can be seen in more details for the lead times of 6 and 7 days in Fig. 10. Rewriting Eq. (14) as

$$\mathrm{CRPS} = \int_{-\infty}^{A} F_A^2(x)\,dx + \int_{A}^{\infty} [F_A(x) - 1]^2\,dx, \quad (16)$$

one can note that the CRPS is made of two parts: the first, on the left of the observed value (vertical mixed line), is the integral of the square of the distance between zero and the curve defined by the distribution, and the second, on the right, is the integral of the square of the distance between the distribution and the unit probability. At day 6 (Fig. 10, left), the integral limited by the raw ensemble is less than the integral limited by the logistic; therefore, the raw ensemble is doing better regarding the CRPS. At day 7 (Fig. 10, right) it is the reverse and, as the observed value lies outside the raw ensemble, the longer tail of the logistic curve contributes to lowering the CRPS. For the case of the Ourthe during winter, a number of situations like the one depicted in Fig. 10 (right) implies a substantial improvement of the average CRPS.

The overall performances of the postprocessed ensembles compared to those of the raw ensembles are presented for the two catchments and both seasons in Figs. 12–15. The ME of the raw ensembles reveals a positive bias except for the Ourthe during winter with precipitation forecasts that are underestimated on average. Similar contrast between the ME values during winter for the same catchments taken downstream was discussed in Roulin and Vannitsem (2005) and Van den Bergh and Roulin (2009). Using the parameters obtained with the EPS themselves provides an optimal picture of the impact of postprocessing. Indeed, the bias is corrected in the postprocessed ensembles. However, for the Ourthe catchment, the negative bias tends to be overcorrected. We suspect that the postprocessed ensemble average is altered by the step of averaging transformed member's precipitation when preparing the predictors of the extended logistic regression (instead of transforming the ensemble average). It can be reminded that this choice was motivated to ease the regression calibration in the parameter estimation. Preliminary results on 51-member ensembles and using transformed ensemble

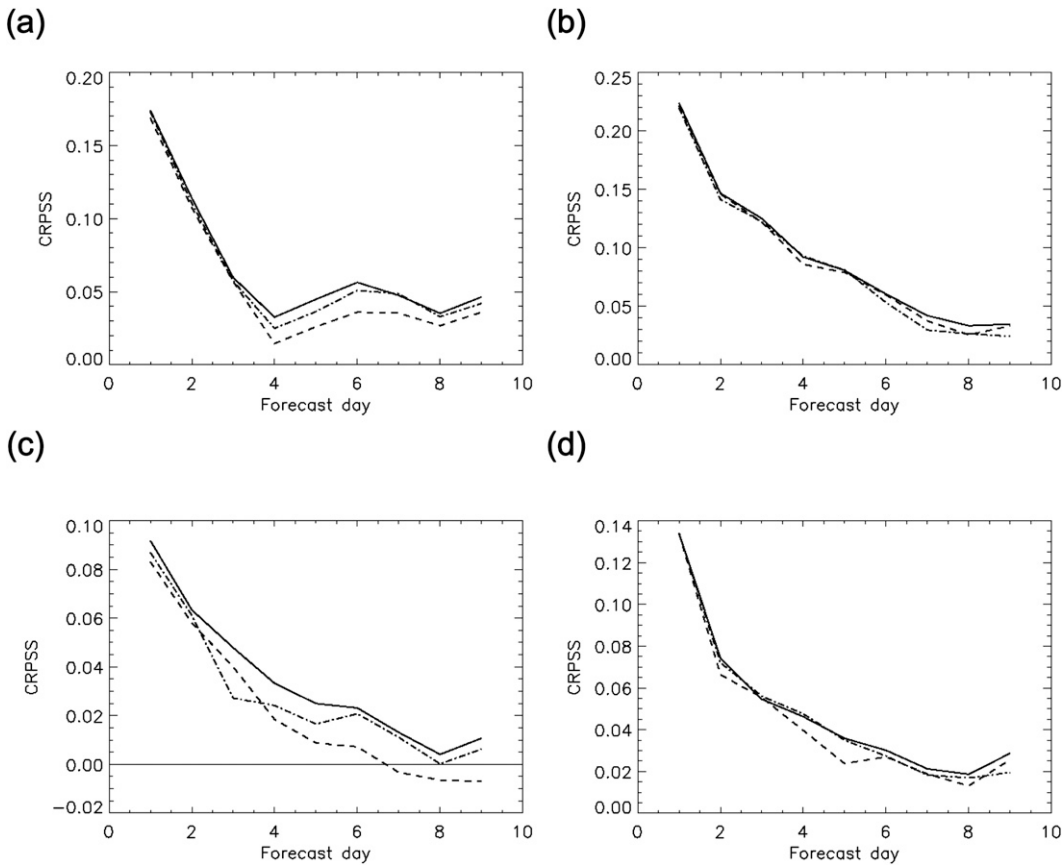

FIG. 8. As in Fig. 7, but for (dashed line) HIN-POOL and (dash–dotted line) HIN-POOL-RC.

average resulted in a better correction of the bias (ME) in the postprocessed ensembles. Note that using the multiplicative bias (DMB) correction, the correction would be simply proportional to the raw ensemble mean.

Using the parameters obtained with the hindcasts results in a correction of the bias during the first two forecast days for the Ourthe catchment during winter and the bias is degraded for longer lead times. For summer the bias is corrected up to 6 days. For the Gete catchment, the bias is intermediate between the values of the raw ensembles and those achieved with the parameters obtained with the EPS.

The MSE is increasing with lead time, the values being larger for summer than for winter and larger for the Ourthe than for the Gete. The MSE is only marginally reduced with the postprocessing during summer, either based on EPS or on hindcasts. The spread (variance) of the ensembles is also increasing with lead time. One important criteria of the quality of the ensemble forecasts is the necessity to have a spread equal to the MSE (see Leutbecher and Palmer 2008). Here, the raw ensembles are in general underdispersed. The postprocessing with the parameters based on EPS allows for an increase of the spread to a value closer to the MSE. The only exception is the Gete during summer for lead times of 4–7 days (Fig. 15). The postprocessing with the parameters based on hindcasts produces a too large spread on average compared to the MSE for winter (Figs. 12 and 14) whereas for summer the spread is more similar to the one resulting from the parameters based on EPS (Figs. 13 and 15). Note that the spread of the corrected ensembles does not use any information about the uncertainty of the raw ensemble. Using a multiplicative bias correction (results not shown), the spread would have been decreased in the situations where the ensemble means are overestimated and slightly increased otherwise.

The results on the CRPS are similar for the two parameter sets (51 or 5 members) based on EPS or hindcasts. The best improvement of the CRPS with the postprocessing is achieved for the Gete during summer (Fig. 15) especially during the first half of the forecast range. Some insight can be gained by comparing the BSS of the raw ensembles and of the postprocessed ensembles for different thresholds. For instance, for the Gete during summer at forecast day $D + 3$, results (not shown) reveal that the BSS is improved for the entire

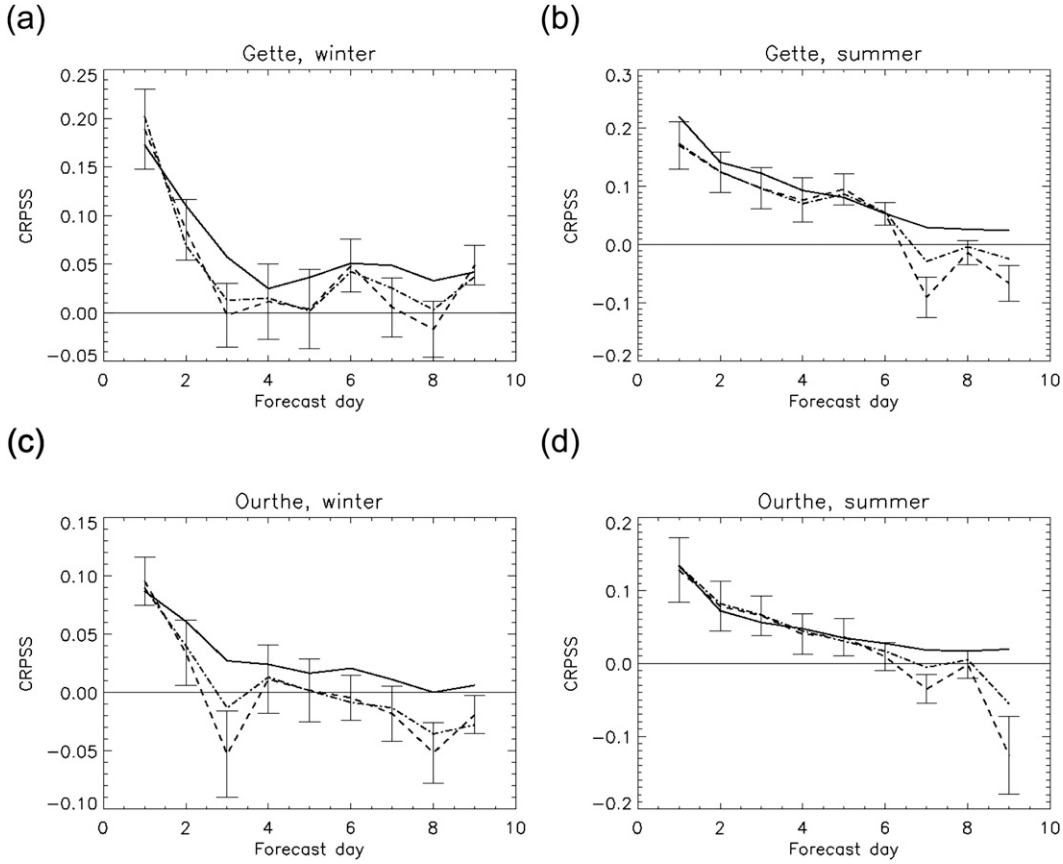

FIG. 9. As in Fig. 7, but for (solid line) HIN-POOL-RC, (dashed line) HIN-WIN5-RC, and (dash–dotted line) HIN-WIN7-RC. The error bars correspond to twice the standard deviation of 1000 bootstraps in the verification time series (HIN-WIN5-RC).

range of thresholds. This improvement can be explained by a better reliability for all thresholds and by a better resolution for low thresholds (<P60). The improvement is minor for the Ourthe during winter (Fig. 12). For this case, at forecast day $D + 3$, the decomposition of BSS (not shown) reveals almost no change in resolution neither in reliability—the raw ensemble are already reliable—except for a very low threshold (<P40). For this case, as

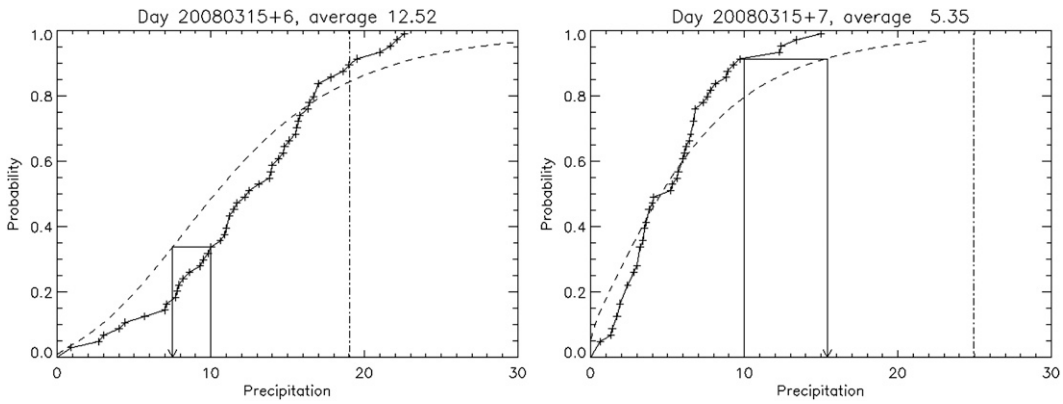

FIG. 10. Probability that the areal precipitation exceeds a threshold. Case study: Ourthe, forecast (left) days $D + 6$ and (right) $D + 7$ from 15 Mar 2008: (solid line) raw EPS, (dashed line) distribution (EPS-FULL) calculated with the extended logistic regression based on the mean of the power transformed raw ensemble, and (dash–dotted line) drawn at the observed value. The arrow shows the mapping from raw EPS members to fitted extended logistic.

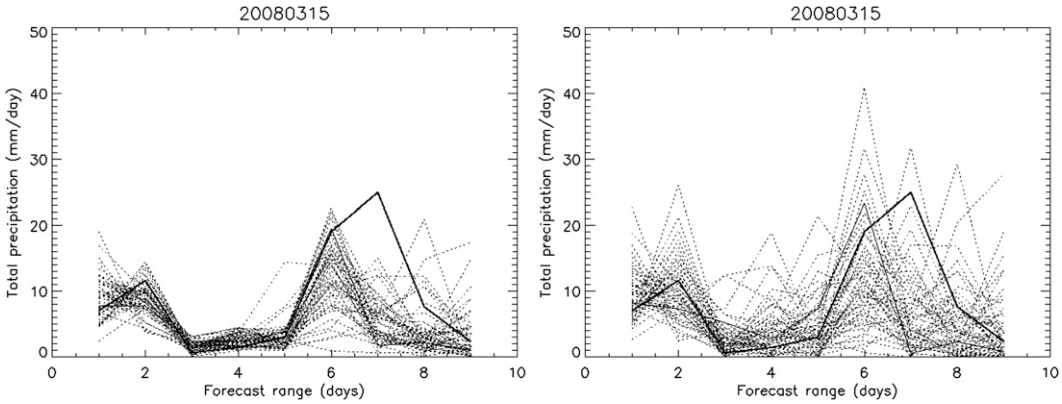

FIG. 11. Areal precipitation over the Ourthe catchment: ensemble traces (dotted lines) and observed (solid bold line); (left) raw ensembles and (right) postprocessed ensembles (EPS-FULL).

mentioned above, a finer analysis of the CRPS of each realization both for the raw and the postprocessed ensemble shows that a group of situations included in the verification dataset contributes to lowering the average CRPS and is characterized by all members of the raw ensemble underestimating the observed areal precipitation whereas the postprocessed ensemble has an increased spread and includes the observed value as in Fig. 10 (right). On the contrary, for all other cases (Ourthe during summer and Gete during winter and summer), groups of situations contributing to lowering the average CRPS may be isolated in the dataset, but they correspond to raw ensembles overestimating low observed values and to postprocessed ensembles with less outliers due to an increased spread and with a decreased ensemble mean.

4. Summary and discussion

The use of hindcasts to develop postprocessing techniques for ensemble precipitation predictions has been

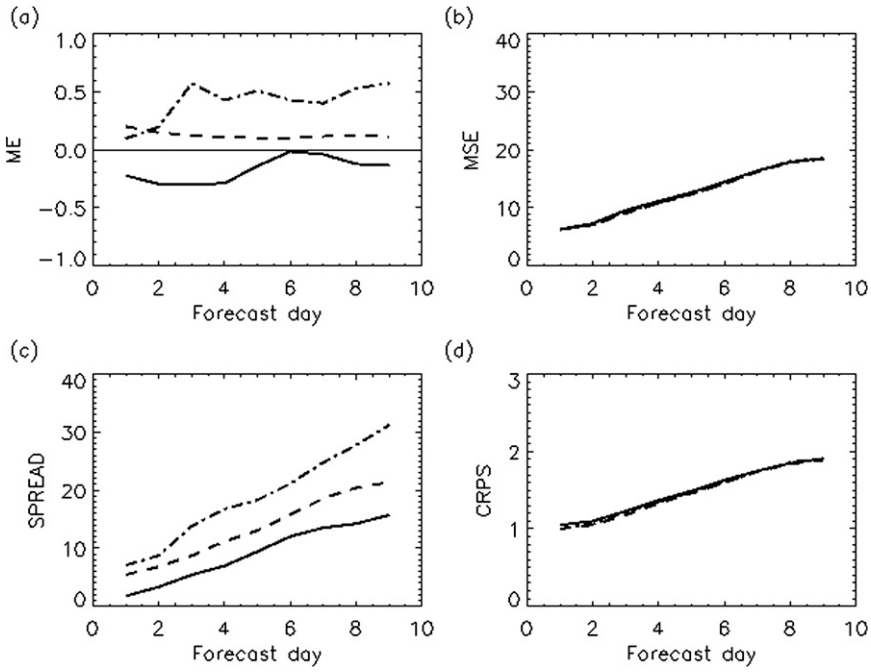

FIG. 12. Verification of the ensemble prediction for the precipitation: Ourthe catchment during winter. The (a) ME and (d) CRPS in mm day^{-1}; the (b) MSE and (c) ensemble spread in mm^2 day^{-2}: (solid line) raw ensembles, (dashed line) postprocessed using the extended logistic regression parameters fitted on the EPS (EPS-FULL), and (dash–dotted line) on hindcasts (HIN-POOL-RC).

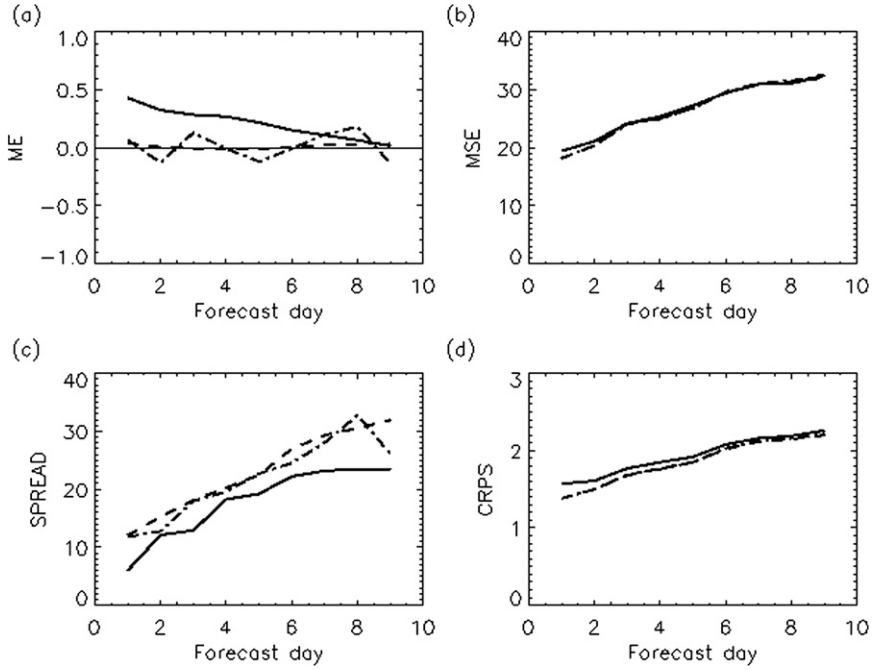

FIG. 13. As in Fig. 12, but for the Ourthe catchment during summer.

investigated for two Belgian catchments. The method tested is the extended logistic regression that keeps the coherence between the probabilities estimated for different thresholds. As the size of the hindcasts is low compared to the full operational ensemble, the regression parameters should be corrected for biases by using, for instance, the so-called regression calibration method. This method is efficient as revealed by the comparison of the parameters obtained on the full EPS and of subsamples.

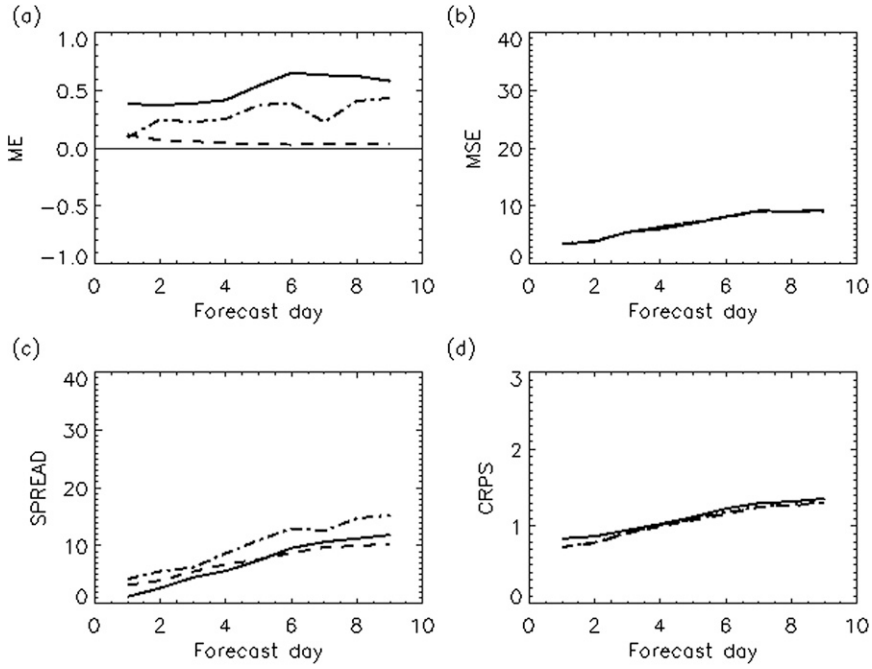

FIG. 14. As in Fig. 12, but for the Gete catchment during winter.

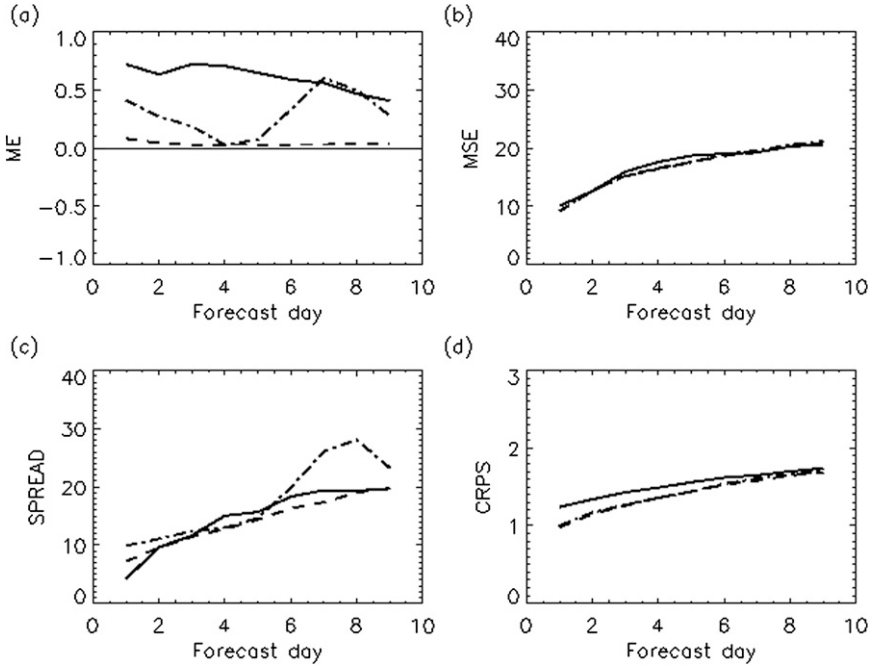

FIG. 15. As in Fig. 12, but for the Gete catchment during summer.

As the extended logistic regression aims at correcting the forecast probabilities over the full distribution, a natural extension of its application consists in the correction of the precipitation forecast ensemble members. Benefits have been found like the improvement of bias (ME), a better spread, a decreased number of outliers, and also decreased CRPS. The simulated logistic distributions and, therefore, their mean and variance depend solely on the raw ensemble average (more precisely, the mean of power transformed members); the ensemble spread does not improve much the likelihood function in the parameter estimation and is estimated with a large uncertainty on the small hindcast ensembles; however, further tests will be carried out to check whether this information can be transferred to the postprocessed ensembles. Finally, the method should be compared with other methods like the Bayesian model averaging.

Statistical differences between hindcasts and EPS have been shown. In an operational setting where the calibration database corresponds to hindcasts taken in a moving window, the postprocessing with the extended logistic regression has skill during the first two forecast days for winter and during the first five to six forecast days during summer.

Acknowledgments. We thank Tom Hamill and an anonymous reviewer for the useful remarks and suggestions. This work was partly supported by the Belgian Federal Science Policy Office under Grant MO-34-020.

REFERENCES

Applequist, S., G. E. Gahrs, R. L. Pfeffer, and X.-F. Niu, 2002: Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Wea. Forecasting,* **17,** 783–799.

Carroll, R. J., and L. A. Stefanski, 1990: Approximate quasilikelihood estimation in models with surrogate predictors. *J. Amer. Stat. Assoc.,* **85,** 652–663.

——, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, 2006: *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall/CRC Press, 455 pp.

Casella, G., and R. L. Berger, 2002: *Statistical Inference*. 2nd ed. Duxbury Press, 660 pp.

Folland, C., and C. Anderson, 2002: Estimating changing extremes using empirical ranking methods. *J. Climate,* **15,** 2954–2960.

Gleser, L. J., 1990: Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. *Contemp. Math.,* **112,** 99–114.

Hagedorn, R., 2008: Using the ECMWF reforecast dataset to calibrate EPS forecasts. *ECMWF Newsletter,* No. 117, ECMWF, Reading, United Kingdom, 9–13.

——, T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.,* **136,** 2608–2619.

Hamill, T. M., and S. J. Colucci, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.,* **134,** 3209–3229.

——, ——, and X. Wei, 2004: Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.,* **132,** 1434–1447.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.,* **136,** 2620–2632.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting,* **15,** 559–570.

Katz, R. W., and M. Ehrendorfer, 2006: Bayesian approach to decision making using ensemble weather forecasts. *Wea. Forecasting,* **21,** 220–231.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.,* **227,** 3515–3539.

McCollor, D., and R. Stull, 2008: Hydrometeorological accuracy enhancement via postprocessing of numerical wheather forecasts in complex terrain. *Wea. Forecasting,* **23,** 131–144.

Peel, S., and L. J. Wilson, 2008: Modeling the distribution of precipitation forecasts from the Canadian ensemble prediction system using kernel density estimation. *Wea. Forecasting,* **23,** 575–595.

Reggiani, P., and A. H. Weerts, 2008: Probabilistic quantitative precipitation forecast for flood prediction: An application. *J. Hydrometeor.,* **9,** 76–95.

Rodwell, M., 2006: Comparing and combining deterministic and ensemble forecasts: How to predict rainfall occurrence better. *ECMWF Newsletter,* No. 106, ECMWF, Reading, United Kingdom, 17–23.

Rosner, B., W. C. Willett, and D. Spiegelman, 1989: Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat. Med.,* **8,** 1051–1069.

——, D. Spiegelman, and W. C. Willett, 1990: Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariate measured with error. *Amer. J. Epidemiol.,* **132,** 734–745.

——, ——, and ——, 1992: Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *Amer. J. Epidemiol.,* **136,** 1400–1413.

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

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.,* **130,** 1653–1660.

Schmeits, M. J., and K. J. Kok, 2010: A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Mon. Wea. Rev.,* **138,** 4199–4211.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 3209–3220.

Stefanski, L. A., and R. J. Carroll, 1985: Covariate measurement error in logistic regression. *Ann. Stat.,* **13,** 1335–1351.

Thoresen, M., and P. Laake, 2000: A simulation study of measurement error correction methods in logistic regression. *Biometrics,* **56,** 868–872.

——, and ——, 2003: The use of replicates in logistic measurement error modeling. *Scand. J. Stat.,* **30,** 625–636.

Van den Bergh, J., and E. Roulin, 2009: Hydrological ensemble prediction and verification for the Meuse and Scheldt basins. *Atmos. Sci. Lett.,* **11,** 64–71.

Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.,* **16,** 361–368.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.,* **135,** 2379–2390.

CORRIGENDUM

EMMANUEL ROULIN AND STÉPHANE VANNITSEM

Institut Royal Météorologique de Belgique, Brussels, Belgium

An editing error occurred in the caption of Fig. 4 of Roulin and Vannitsem (2012). The description of the curves were given in the reverse order. The figure and caption have been corrected in the version displayed below.

The staff of *Monthly Weather Review* regrets any inconvenience this error has caused.

REFERENCE

Roulin, E., and S. Vannitsem, 2012: Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Mon. Wea. Rev.,* **140,** 874–888.

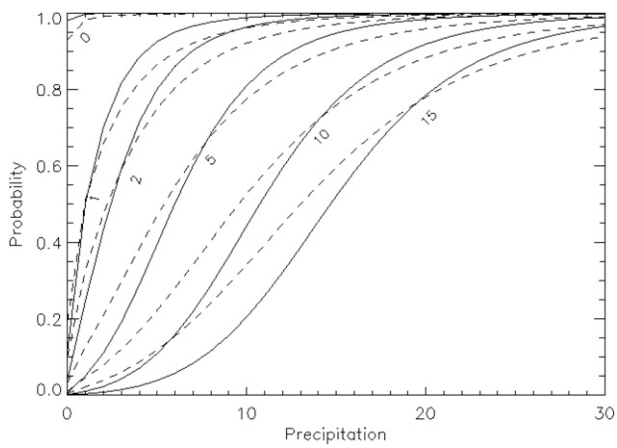

FIG. 4. Probability distribution [Pr($A \leq q$)] based on the extended logistic regressions (EPS-FULL) for the Ourthe during winter at forecast days 3 (solid line) and 6 (dashed line). (left to right) The curves correspond to ensemble averages of 0, 1, 2, 5, 10, and 15 mm day^{-1}.

DOI: 10.1175/MWR-D-12-00061.1

© 2012 American Meteorological Society