

A Comparison between Raw Ensemble Output, (Modified) Bayesian Model Averaging, and Extended Logistic Regression Using ECMWF Ensemble Precipitation Reforecasts

MAURICE J. SCHMEITS AND KEES J. KOK

Royal Netherlands Meteorological Institute, De Bilt, Netherlands

(Manuscript received 17 November 2009, in final form 19 May 2010)

ABSTRACT

Using a 20-yr ECMWF ensemble reforecast dataset of total precipitation and a 20-yr dataset of a dense precipitation observation network in the Netherlands, a comparison is made between the raw ensemble output, Bayesian model averaging (BMA), and extended logistic regression (LR). A previous study indicated that BMA and conventional LR are successful in calibrating multimodel ensemble forecasts of precipitation for a single forecast projection. However, a more elaborate comparison between these methods has not yet been made. This study compares the raw ensemble output, BMA, and extended LR for single-model ensemble reforecasts of precipitation; namely, from the ECMWF ensemble prediction system (EPS). The raw EPS output turns out to be generally well calibrated up to 6 forecast days, if compared to the area-mean 24-h precipitation sum. Surprisingly, BMA is less skillful than the raw EPS output from forecast day 3 onward. This is due to the bias correction in BMA, which applies model output statistics to individual ensemble members. As a result, the spread of the bias-corrected ensemble members is decreased, especially for the longer forecast projections. Here, an additive bias correction is applied instead and the equation for the probability of precipitation in BMA is also changed. These modifications to BMA are referred to as “modified BMA” and lead to a significant improvement in the skill of BMA for the longer projections. If the area-maximum 24-h precipitation sum is used as a predictand, both modified BMA and extended LR improve the raw EPS output significantly for the first 5 forecast days. However, the difference in skill between modified BMA and extended LR does not seem to be statistically significant. Yet, extended LR might be preferred, because incorporating predictors that are different from the predictand is straightforward, in contrast to BMA.

1. Introduction

Ensemble prediction system (EPS) forecasts are routinely produced by several institutes around the world. Despite their relatively high skill, medium-range EPS forecasts still suffer from underdispersion, especially in the early forecast range, so calibration is needed. In the last couple of years a number of studies have addressed the issue of how ensemble forecasts can best be calibrated (e.g., Wilks 2006a; Wilks and Hamill 2007). In Wilks (2006a) a comparison was made between a large number of statistical methods, such as logistic regression (LR; Brelsford and Jones 1967; Wilks 2006b), Bayesian model averaging (BMA; Raftery et al. 2005), nonhomogeneous Gaussian regression (NGR; Gneiting et al.

2005), and Gaussian ensemble dressing (GED; Roulston and Smith 2003; Wang and Bishop 2005).

Wilks (2006a) used ensemble forecasts based on the Lorenz (1996) model; the variable from that model has a Gaussian distribution. In his study, the following three methods turned out to be the best: LR, NGR, and GED. Subsequently, Wilks and Hamill (2007) investigated the skill of these three methods using a 25-yr ensemble reforecast dataset for temperature and precipitation on the basis of a 1998 version of the National Centers for Environmental Prediction (NCEP) Global forecast system (GFS) ensemble system. LR and NGR generally were the best methods. However, they concluded that differences in the lengths of the training periods usually produced larger skill differences than the three different statistical methods.

BMA did not turn out to be among the three best methods in the study of Wilks (2006a). He concluded that BMA generated less good forecasts, mainly because ensemble underdispersion was overcorrected for forecasts

Corresponding author address: Dr. M. J. Schmeits, Royal Netherlands Meteorological Institute, P.O. Box 201, 3730 AE De Bilt, Netherlands.
E-mail: schmeits@knmi.nl

of relatively rare events (with the threshold being the lowest decile). The best methods for these relatively rare events were NGR and LR. Bishop and Shanley (2008) recently proposed a paradigm shift to fix the problematic treatment of the extremes in BMA. In another study (Sloughter et al. 2007), in which multimodel ensemble forecasts of precipitation were calibrated, LR and the BMA version for precipitation show comparable Brier skill scores (e.g., Wilks 2006b) for a large range of precipitation thresholds, but BMA scores better for the extremes. Besides, BMA has the advantage, in contrast to (conventional) LR, that the total probability density function (pdf) can be estimated. However, Wilks (2009) extended LR to provide full-probability-distribution forecasts, which undoes this advantage of BMA.

BMA for Gaussian-distributed predictands (Raftery et al. 2005) was used in the comparison study of Wilks (2006a); BMA for gamma-distributed predictands like precipitation (Sloughter et al. 2007) was not investigated by Wilks and Hamill (2007). Sloughter et al. (2007) made a comparison between the performance of BMA and conventional LR for one forecast projection, but a more elaborate comparison has not yet been made. Therefore, our study compares the performance of the Sloughter et al. (2007) version of BMA and a modified version thereof (called “modified BMA”; section 3b) with the performance of extended LR (Wilks 2009) using a 20-yr (1982–2001) European Centre for Medium-Range Weather Forecasts (ECMWF) EPS reforecast dataset. This dataset was also used in Hamill et al. (2008), but they only used conventional LR to calibrate the precipitation forecasts.

The following questions will be addressed in this study. (i) How do the extended LR, BMA, and modified BMA methods perform in calibrating ECMWF EPS precipitation forecasts? Note that for this single-model ensemble the BMA weights and parameters of the component pdfs are constrained to be equal (Fraley et al. 2010). (ii) How is the behavior of these methods for the extremes? (iii) Does the expansion of the training period to include days from the same season in previous years improve the performance of (modified) BMA, as suggested by Sloughter et al. (2007)?

In section 2 the ECMWF EPS reforecast dataset and the predictands are described, and in section 3 the extended LR, BMA, and modified BMA methods are described. In section 4 an example of the performance of the BMA and extended LR forecast systems during a day with heavy rain is presented, and in sections 5 and 6 some of the verification results for area-mean and area-maximum precipitation forecasts are described, respectively. Finally, in section 7 the results are summarized and discussed.

2. Reforecast data and predictand definitions

The datasets used in this study are precipitation observations from the observation network of volunteers in the Netherlands (section 2b; Heijboer and Nellestijn 2002) and precipitation data from a reforecasting experiment with the ECMWF EPS system (section 2a; Hamill et al. 2008).

a. ECMWF EPS reforecast data

Hamill et al. (2008) also used this ECMWF reforecast dataset, which consists of a 15-member ensemble reforecast computed once weekly from 0000 UTC initial conditions, provided by the 40-yr ECMWF Re-Analysis (ERA-40; Uppala et al. 2005), for the initial dates of 1 September–24 November. The maximum lead time was 10 days. These forecasts were computed for the period 1982–2001. Model cycle 29r2 was used, with T255 horizontal resolution and 40 vertical sigma-coordinate levels. In our study the total precipitation forecast data are used with a $1^\circ \times 1^\circ$ resolution and accumulated from 0600 to 0600 UTC. The forecast horizons considered range from +30 h (i.e., 6–30-h projections) to +150 h (i.e., 126–150-h projections); and not longer because of the limited predictability of precipitation.

b. Observations and predictand definitions

For the observations both the area-mean and the area-maximum 24-h accumulated precipitation amount at 0800 UTC for $1^\circ \times 1^\circ$ grid boxes are calculated. These calculations are based on precipitation measurements, which are performed daily at 0800 UTC by a dense network of volunteers across the Netherlands (Fig. 1). There are 12 grid boxes containing observation sites. Note the different number of stations per grid box, as well as the inevitable 2-h time difference between observations and model output. This appears to have only a minor effect on the results and conclusions. The observation network of volunteers is used instead of radar data, because data from 1982 to 2001 are needed and the two current radar systems in the Netherlands have only been operated since 1997. The BMA predictand is defined as the pdf of the area-mean or area-maximum 24-h accumulated precipitation amount at 0800 UTC for a $1^\circ \times 1^\circ$ grid box. For the extended LR method, the same predictand is used as for BMA.

3. Statistical methods

a. Extended logistic regression

In the extended logistic regression method (Wilks 2009) the forecast probability p is given by

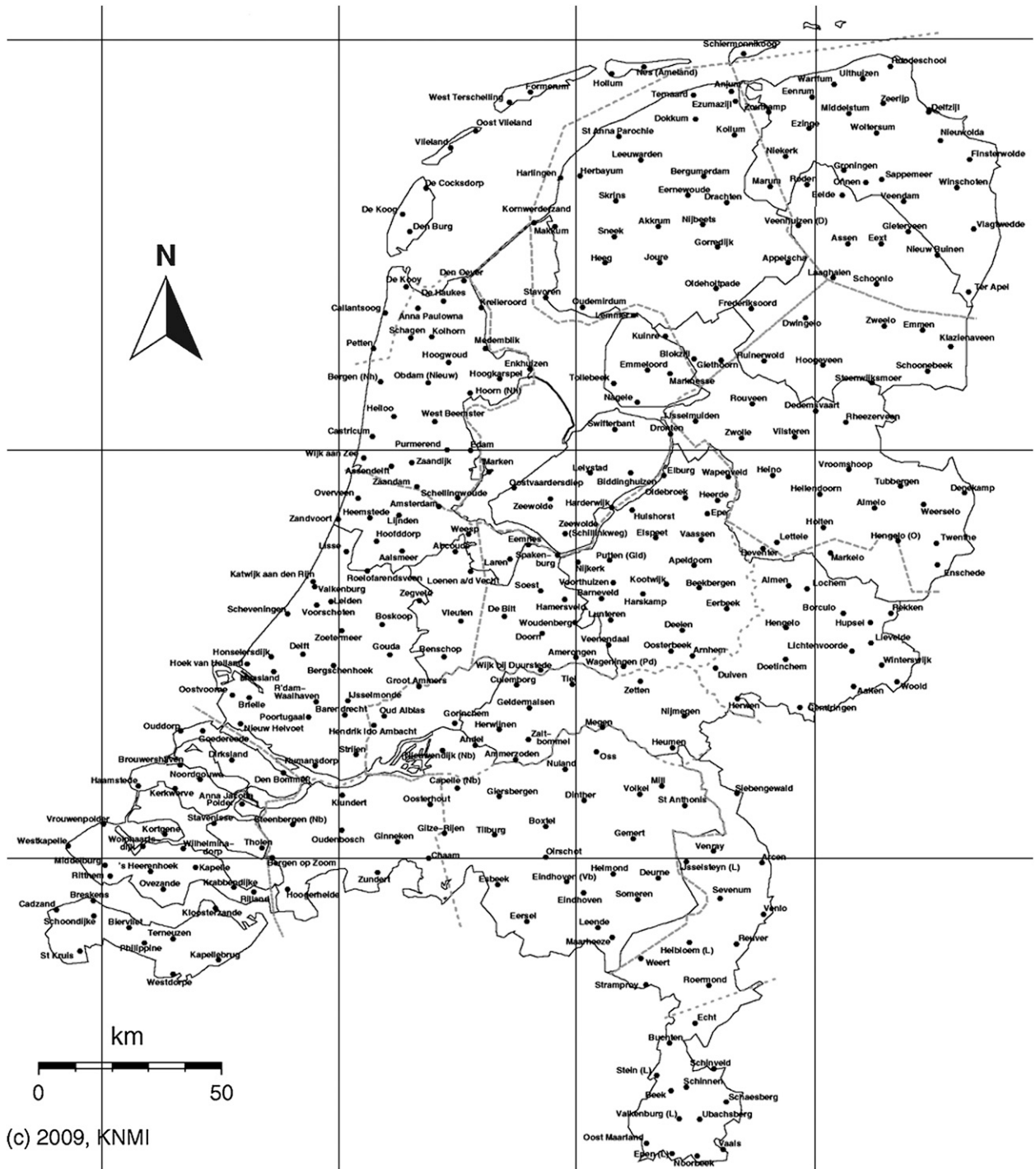


FIG. 1. Locations of precipitation stations (dots) in the Netherlands. The stations that were operational in the whole period from 1982 to 2001 are used in this study.

$$\ln \left[\frac{p(q)}{1 - p(q)} \right] = d_0 + g(q) + d_1 x_1 + d_2 x_2 + \dots + d_n x_n, \quad (1)$$

with $g(q)$ being a nondecreasing function of the threshold q . In this way the full probability distribution can be

estimated, in contrast to conventional logistic regression. The thresholds that are used are 0, 1, 5, 10, and 15 mm for area-mean precipitation and 0, 1, 5, 10, 15, 20, and 25 mm for area-maximum precipitation. The predictors x_i ($i = 1, 2, \dots, n$) and those in $g(q)$ are selected

via a so-called forward stepwise selection method (Wilks 2006b). As in Wilks (2009) the precipitation threshold and its square root are used as potential predictors in $g(q)$. At each step, a predictor is chosen that produces the best regression in conjunction with the predictors chosen on previous steps; hereby a significance probability threshold of 0.05 is specified. Each chosen predictor is kept in the equation unless the specified significance probability of 0.10 is exceeded at a following step. The regression coefficients d_i and those in $g(q)$ are determined using the maximum likelihood method (Wilks 2006b), an iterative method that maximizes the product of all computed probabilities of the (non)occurrence of the events in the training dataset. The training period is a sliding window of the previous w cases of forecasts and verifying observations. To increase the samples, data from the 12 grid boxes are pooled, so that the total length of the training set is $12 \times w$. We use cross validation (e.g., Wilks 2006b) for all verification results in this study (sections 5 and 6).

Apart from the precipitation threshold and its square root (as in Wilks 2009), the square root and cube root (Sloughter et al. 2007) of the ensemble mean total precipitation (TP) are used as potential predictors, as well as the ensemble mean of the square root of TP. Besides, the TP ensemble standard deviation, its square and cube root (Hamill et al. 2008) are used as potential predictors, as well as the percentages of the ensemble members exceeding TP amounts of 1, 5, and 10 mm, respectively. Nonprecipitation variables, like convective indices, could also have been used, but these were not included here to make the comparison with the raw ensemble output and BMA as clean as possible. The two predictors that are selected are the square root of the precipitation threshold and the ensemble mean of the square root of TP. These predictors are used as the only predictors in Eq. (1) for all forecast projections (+30, +54, +78, +102, +126, and +150 h). The first predictor is the same as in Wilks (2009) and the second one performed slightly better than the square root of the ensemble mean of TP, which was selected as the second predictor in his study.

b. (Modified) Bayesian model averaging

For an extensive treatment of BMA for precipitation the reader is referred to Sloughter et al. (2007). Here only a relatively short description of the method is given, largely based on that paper. However, a few important modifications to the conventional BMA method are proposed in this study.

The BMA model for the forecast pdf of the cube root of precipitation accumulation y for a K -member ensemble is (with $K = 15$ in this case):

$$p(y|f_1, \dots, f_K) = \sum_{k=1}^K W_k \{ P(y=0|f_k) I[y=0] + P(y>0|f_k) g_k(y|f_k) I[y>0] \}, \quad (2)$$

where the first part between the accolades computes the probability of zero precipitation as a function of f_k [i.e., the forecast from ensemble member k ($k = 1, \dots, K$)], and the second part computes the pdf of the precipitation amount given that it is nonzero. The weight W_k is the posterior probability of ensemble member k being best, and the general indicator function $I[\dots]$ is unity if the condition in brackets holds and zero otherwise. Besides, $P(y=0|f_k)$ is specified by

$$\text{logit}P(y=0|f_k) \equiv \log \frac{P(y=0|f_k)}{P(y>0|f_k)} = a_{0,k} + a_{1,k} f_k^{1/3} + a_{2,k} \delta_k, \quad (3a)$$

where δ_k is equal to 1 if $f_k = 0$ and equal to 0 otherwise. In this study also an alternative expression for $P(y=0|f_k)$ is applied, namely:

$$\text{logit}P(y=0|f_k) = \text{logit}P(y=0|f_1, \dots, f_K) = a'_0 + \frac{a'_1}{K} \sum_{k=1}^K f_k^{1/3}. \quad (3b)$$

Note that this expression for the probability of precipitation (POP) is similar to the one used often in ensemble model output statistics (e.g., Wilks and Hamill 2007; Hamill et al. 2008).

In Eq. (2) the conditional pdf $g_k(y|f_k)$ of the cube root precipitation amount y given that it is positive is a gamma distribution with pdf:

$$g_k(y|f_k) = \frac{1}{\beta_k^{\alpha_k} \Gamma(\alpha_k)} y^{\alpha_k-1} \exp(-y/\beta_k), \quad (4)$$

with Γ being the gamma function (e.g., Wilks 2006b). The parameters $\alpha_k = \mu_k^2/\sigma_k^2$ and $\beta_k = \sigma_k^2/\mu_k$ of the gamma distribution depend on f_k through relationships for the mean μ_k and the variance σ_k^2 of that distribution. The mean is given by

$$\mu_k = b_{0,k} + b_{1,k} f_k^{1/3}. \quad (5a)$$

For the mean an alternative relationship is also applied in this study; namely, an additive bias correction, which is similar to the ones proposed by Wilson et al. (2007a) and Hamill (2007) for Gaussian predictands:

$$\mu_k = b'_{0,k} + f_k^{1/3}, \quad (5b)$$

with $b'_{0,k} \geq -f_k^{1/3}$. The variance of the gamma distribution is given by

$$\sigma_k^2 = c_0 + c_1 f_k. \quad (6)$$

The BMA method with the modified POP [Eq. (3b)] and the additive bias correction [Eq. (5b)] is referred to as modified BMA in this paper. The reason for these modifications to BMA is as follows: in calibrating an ensemble, model output statistics (MOS; Glahn and Lowry 1972; Wilks 2006b) should not be applied to individual ensemble members, as this decreases the spread of the ensemble, especially for the longer forecast projections for which the individual ensemble members are regressed toward the climatological mean (e.g., Wilks 2006a). This is further discussed in section 5. For the variance of the gamma distribution [Eq. (6)] an alternative expression could have been used as well (e.g., only a constant), but it was not modified here, because in general the value of c_1 in Eq. (6) appears to be close to 0 anyway.

As is also done in the extended logistic regression model (section 3a), the parameters are estimated using data from a training period (i.e., a sliding window of the previous w cases of forecasts and verifying observations). The data from the 12 grid boxes are also pooled, so that the total length of the training set is $12 \times w$. The parameters $a_{0,k}$, $a_{1,k}$, and $a_{2,k}$ or a'_0 and a'_1 are estimated by logistic regression with yes–no precipitation as the binary predictand, and the parameters $b_{0,k}$ and $b_{1,k}$ are determined by linear regression with the nonzero precipitation observations as cases and the cube root of the precipitation amount as the predictand. Alternatively, $b'_{0,k}$ is determined by subtracting the mean value of the cube root of the forecast precipitation from the mean value of the cube root of the observed precipitation amount, also with the nonzero precipitation observations as cases.

The parameters W_k , c_0 , and c_1 are estimated by the maximum likelihood technique (Wilks 2006b) from the training data. The log-likelihood function for the BMA model [Eq. (2)] (Sloughter et al. 2007) is maximized numerically using the so-called expectation-maximization (EM) algorithm. Because in this study the 15 ensemble members come from a single model, they are exchangeable; therefore, the BMA weights W_k are constrained to be equal and have a value of $1/15$ each (Raftery et al. 2005; Wilks 2006a; Hamill 2007; Fraley et al. 2010). Hence, only c_0 and c_1 have to be estimated by the maximum likelihood technique. Besides, $a_{i,k} = a_i$ ($i = 0, 1, 2$), $b_{j,k} = b_j$ ($j = 0, 1$), and $b'_{0,k} = b'_0 \geq -\min(f_1^{1/3}, \dots, f_K^{1/3})$, which means that these parameters are also the same for each ensemble member in this case. In this study we use the same values of c_0 and c_1 for modified BMA as for

conventional BMA. Sensitivity tests indicate that this choice probably has a minor impact on the verification results, as will be discussed in section 7.

c. BMA versus extended LR

An advantage of BMA compared to extended LR is that one can start to produce the forecast pdfs readily, once the software is installed and a relatively short archive of forecasts and observations is available, without having to derive regression equations first, using a relatively long archive. A disadvantage of BMA is that it is not straightforward to use predictors that are different from the predictand because of the BMA model formulation [Eq. (2)], whereas it is straightforward to use those predictors in extended LR [Eq. (1)].

4. Example of a heavy rain case

In this section the BMA and extended LR forecasts for a heavy rain case (i.e., for the 24-h accumulated area-mean precipitation at 0800 UTC 25 September 1988) are presented. All forecasts are based on the 54–78-h accumulated EPS TP reforecasts from the 0000 UTC 22 September 1988 run.

Following Sloughter et al. (2007), the example in Fig. 2a illustrates how the BMA method works. Figure 2a shows the raw ensemble forecasts, the probability of zero precipitation, and the BMA predictive pdf and its components; namely, the weighted contributions from the bias-corrected ensemble members, for the grid box (51.5°–52.5°N, 3.5°–4.5°E) for a training window length of 247 cases (19 autumns). The probability of exceeding a given amount is given in Fig. 2a by the proportion of the area under the BMA pdf (top curve) to the right of it.

The 24-h accumulated precipitation probability forecasts for all 12 grid boxes, valid at 0800 UTC 25 September 1988, computed by the BMA system with a training window of 247 cases, are shown in Fig. 2b. The probabilities of exceeding a precipitation threshold of, for example, 10 mm are between 7% and 18% (dependent on the grid box). The corresponding probabilities from the extended LR system (with the same training window of 247 cases) are between 2% and 42% (not shown). On the other hand, the raw EPS output shows 0–6 members exceeding 10 mm of TP in this case, and the climatological probabilities of ≥ 10 mm are between 5% and 9%. Therefore, it can be concluded that in this case most forecast probabilities of ≥ 10 mm are higher than normal, and already give an indication of high precipitation amounts 2–3 days before. Yet, the highest BMA probabilities of exceeding the higher precipitation thresholds have been forecast in the northern grid boxes (Fig. 2b), whereas the highest amounts have been observed in the

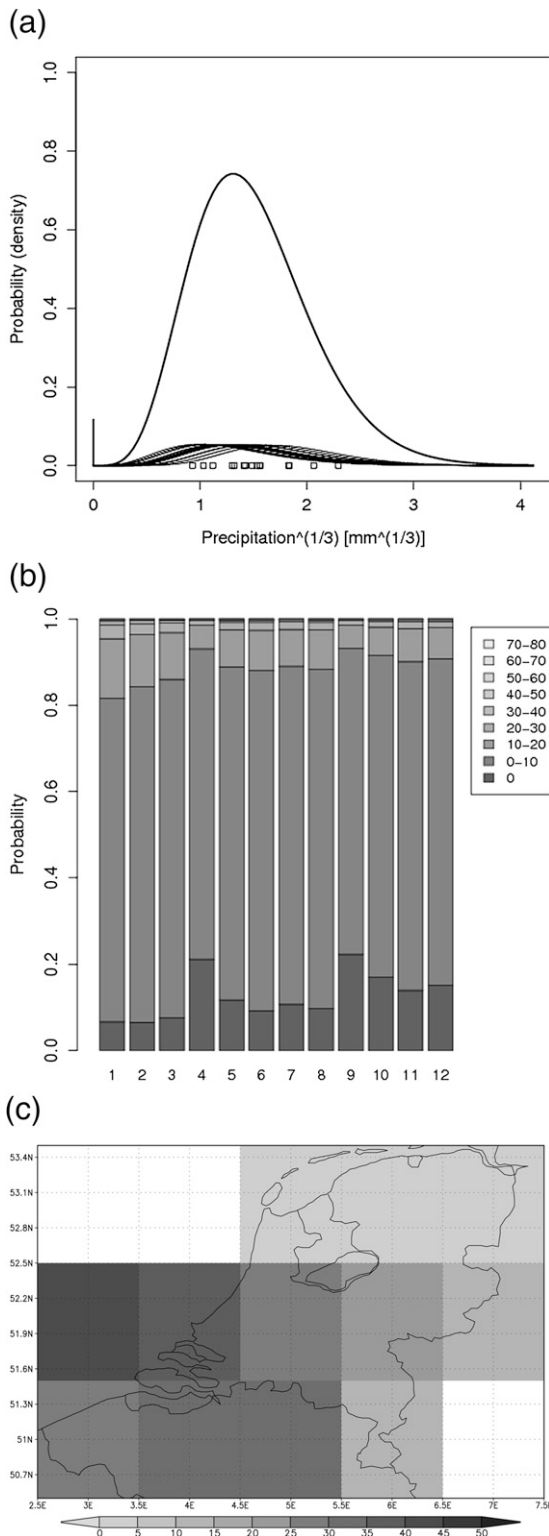


FIG. 2. (a) BMA-fitted pdf of 24-h accumulated mean precipitation in the grid box (51.5° – 52.5° N, 3.5° – 4.5° E) at 0800 UTC 25 Sep 1988 for a training window length w of 247 cases (after Sloughter et al. 2007). The thick vertical line at 0 represents the BMA estimate of the probability of no precipitation, and the top

southwestern grid boxes (Fig. 2c). In the 9 most southern grid boxes the amount of 10 mm has been exceeded and the maximum observed amount at a station is 83 mm. Of course, (probability) forecasts cannot be verified using only one case, so objective verification results are now presented.

5. Verification results for area-mean precipitation forecasts

In this section verification results are shown for the area-mean precipitation forecasts. All verification results are based on spatially pooled cross-validated data from the autumns of the 1982–2001 period ($N = 20 \times 13 \times 12 = 3120$).

a. Verification of the BMA system with the raw EPS as the reference

In this subsection the BMA system is verified and the optimum BMA training window length is determined. Following Raftery et al. (2005) and Sloughter et al. (2007), the continuous ranked probability score (CRPS) is used to determine the optimum training window length, but here the CRPS is transformed into a skill score (e.g., Wilks 2006b), the continuous ranked probability skill score (CRPSS), with the raw ensemble as the reference. First, in Fig. 3a the CRPSS of the BMA system is shown as a function of forecast projection for a training window length $w = 247$, and Fig. 3b shows the CRPSS as a function of the training window length w for the 126-h forecast projection. The behavior of the CRPSS as a function of w is similar for the other projections. Remember that the temporal resolution of the EPS reforecast dataset is 1 week (section 2a), so that the training window length is 1, 5, 10, 15, or 19 autumns.

The CRPSS is only generally positive for the 30-h forecast projection, indicating that the BMA system has more skill than the raw ensemble. From the 54-h projection on, the BMA system is as skillful as or less skillful than the raw ensemble, the latter being an undesirable

←

solid curve is the BMA pdf of the precipitation amount given that it is nonzero. The bottom curves are the components of the BMA pdf, namely the weighted contributions from the bias-corrected ensemble members, and the squares represent the ensemble member precipitation forecasts. (b) BMA stacked bar graph of 24-h accumulated precipitation probabilities for the 12 grid boxes at 0800 UTC 25 Sep 1988 for $w = 247$. The numbers 1–3 indicate the northern grid boxes from west (W) to east (E), 4–8 the central boxes from W to E, and 9–12 the southern boxes from W to E [see (c)]. (c) Observed 24-h accumulated mean precipitation (mm) in the 12 grid boxes at 0800 UTC 25 Sep 1988.

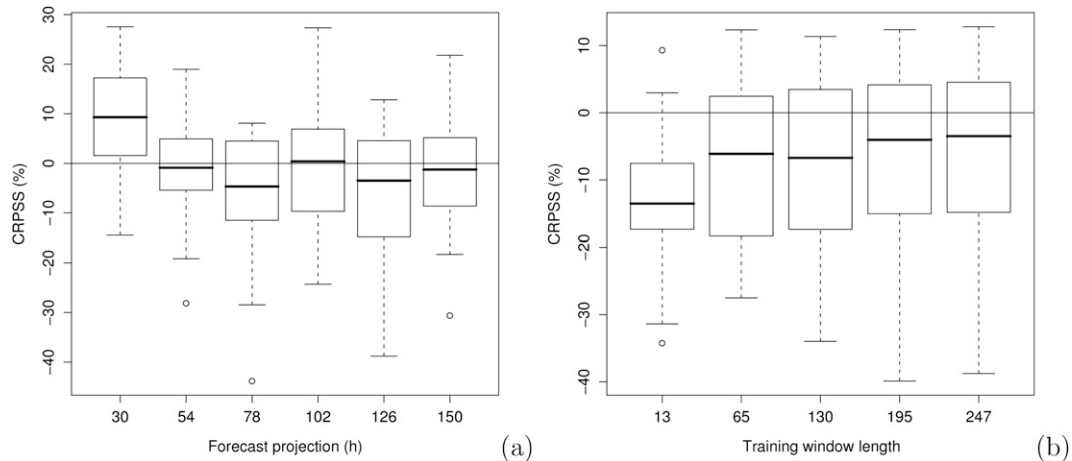


FIG. 3. CRPSS (%) of the BMA system for 24-h accumulated area-mean precipitation with respect to the raw EPS: (a) as a function of the forecast projection for a BMA training window length $w = 247$ and (b) as a function of the BMA training window length for the 126-h forecast projection. Each box-and-whisker plot shows the following statistics for the annual CRPSSs from 1982 to 2001 (autumns only): the lower limit of the box gives the first quartile ($q_{0.25}$) and the upper limit gives the third quartile ($q_{0.75}$) of the annual CRPSSs. The line inside the box is the median ($q_{0.50}$). The whiskers indicate the spread of the CRPSSs, apart from outliers. Outliers are represented by an open circle (o), when they are more than 1.58 box lengths away from $q_{0.25}$ or $q_{0.75}$.

property of a statistical postprocessing method based on that raw ensemble. We will investigate this issue further in section 5b. In general, the CRPSS increases as a function of w (e.g., Fig. 3b). The same is true for modified BMA and for extended LR (not shown), in accordance with Fig. 14 of Hamill et al. (2004). Therefore, the maximum window length $w = 247$ is used for all subsequent verification results.

b. Verification of raw EPS versus (modified) BMA and extended LR

In this subsection Brier skill scores (BSS), the Brier score (BS) decomposition terms reliability and resolution (e.g., Wilks 2006b), and reliability diagrams are used to study the differences between the raw EPS, BMA, modified BMA, and extended LR. Each panel of Fig. 4 shows the BSS as a function of the precipitation threshold for these methods, and the uncertainty in the BSS is indicated by 90% block bootstrap confidence intervals (e.g., Efron and Tibshirani 1993; Wilks 2006b), with the 12 grid boxes being blocked together. Figure 4a shows the BSS for the 30-h forecast projection, Fig. 4b for the 78-h projection, and Fig. 4c for the 126-h projection.

The BSS generally decreases as a function of the precipitation threshold, as expected. However, the lower BSS of especially the raw EPS TP for the 0-mm threshold, compared to the 1-mm threshold, is due to the fact that NWP models often produce spurious light precipitation. In accordance with Fig. 3a, it can be concluded from Fig. 4 that the BMA system improves on the skill

of the raw EPS for the 30-h projection, but it is much less skillful than the raw EPS for the 78- and 126-h projections.

When looking at the values of the coefficient b_1 in the equation for the mean of the gamma distribution [Eq. (5a)] it is clear what causes this bad performance of BMA for the longer forecast projections. While b_1 is about 0.6 for the 30-h forecast projection, it decreases to only 0.3 for the 126-h forecast projection. In other words, the spread of the bias-corrected ensemble members is decreased to a large extent for the longer forecast projections, leading to a postprocessed ensemble that is underdispersed. The reason is that MOS is used for the bias correction of the individual ensemble members, leading to a regression of the individual ensemble members toward the climatological mean.

If the additive bias correction [Eq. (5b)] is applied instead, together with the modified POP formulation [Eq. (3b)], the performance of BMA is substantially improved for the 78- and 126-h projections (Figs. 4b,c). On the other hand, the modified BMA system is about as skillful as the conventional BMA system for the 30-h projection (Fig. 4a). Generally, the difference in skill between the raw EPS, extended LR, and modified BMA does not seem to be statistically significant. An exception is the 0-mm threshold, for which the skill of extended LR and modified BMA is significantly better than the raw EPS, due to the spurious light precipitation in the EPS. The difference in skill between extended LR and conventional LR (with the ensemble mean of the

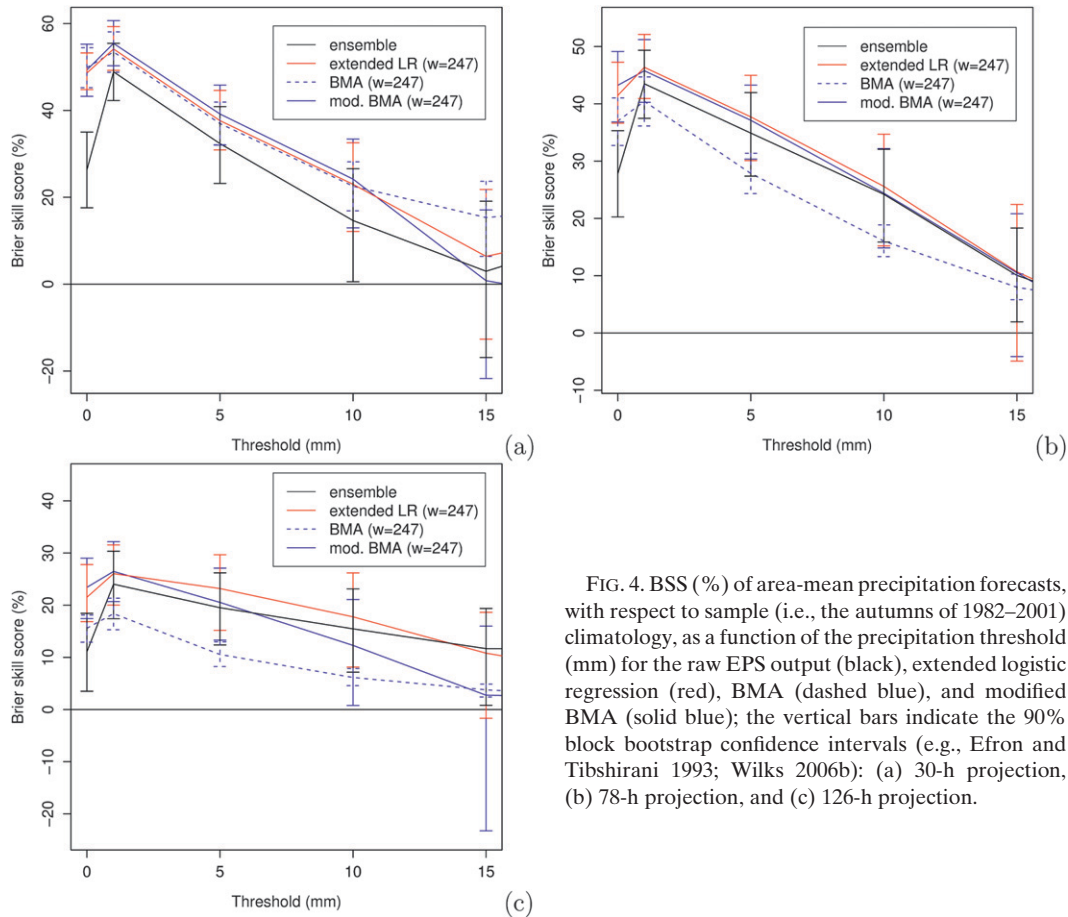


FIG. 4. BSS (%) of area-mean precipitation forecasts, with respect to sample (i.e., the autumns of 1982–2001) climatology, as a function of the precipitation threshold (mm) for the raw EPS output (black), extended logistic regression (red), BMA (dashed blue), and modified BMA (solid blue); the vertical bars indicate the 90% block bootstrap confidence intervals (e.g., Efron and Tibshirani 1993; Wilks 2006b): (a) 30-h projection, (b) 78-h projection, and (c) 126-h projection.

square root of TP as the only predictor for each threshold) is also not significant (not shown), in accordance with the results of Wilks (2009).

To explore the differences between the methods in more detail, reliability diagrams for the 30- and 126-h projections are shown in Figs. 5 and 6, respectively. In these figures the diagrams are shown for the raw EPS (Figs. 5a and 6a), the extended LR system (Figs. 5b and 6b), the conventional BMA system (Figs. 5c and 6c), and the modified BMA system (Figs. 5d and 6d), all for the 5-mm threshold. For the 30-h projection the raw EPS is rather well calibrated, but both its reliability and resolution can be improved by extended LR, BMA, or modified BMA, with the differences between the three postprocessing methods being small.

Even for the 126-h projection the raw EPS is well calibrated (Fig. 6a). While the extended LR system (Fig. 6b) shows both better reliability and better resolution than the raw EPS, the conventional BMA system (Fig. 6c) shows both worse reliability and worse resolution than the raw EPS, resulting in the worst BSS of all methods. The concentration of the forecast probabilities around

the climatological probability can be clearly seen. This lack of sharpness may be a result of the individual ensemble members being regressed toward the climatological mean as well as being dressed with gamma-distribution kernels that have rather constant and similar variance for each ensemble member. The latter is due to the value of the coefficient c_1 in Eq. (6) being close to 0 (i.e., the variance of the kernel hardly depends on f_k). On the other hand, the reliability diagram of the modified BMA system (Fig. 6d) shows much sharper probabilistic forecasts. Because the kernel variance in the modified BMA is the same as the kernel variance in the conventional BMA, this cannot explain the difference in sharpness. It is therefore the MOS-based bias correction that is largely responsible for the lack of sharpness in the probabilistic forecasts of the conventional BMA at long leads. The modified BMA system shows an enormous skill improvement compared to the conventional BMA system, and has better resolution but worse reliability than the raw EPS, which compensate each other to a large extent in terms of the Brier score.

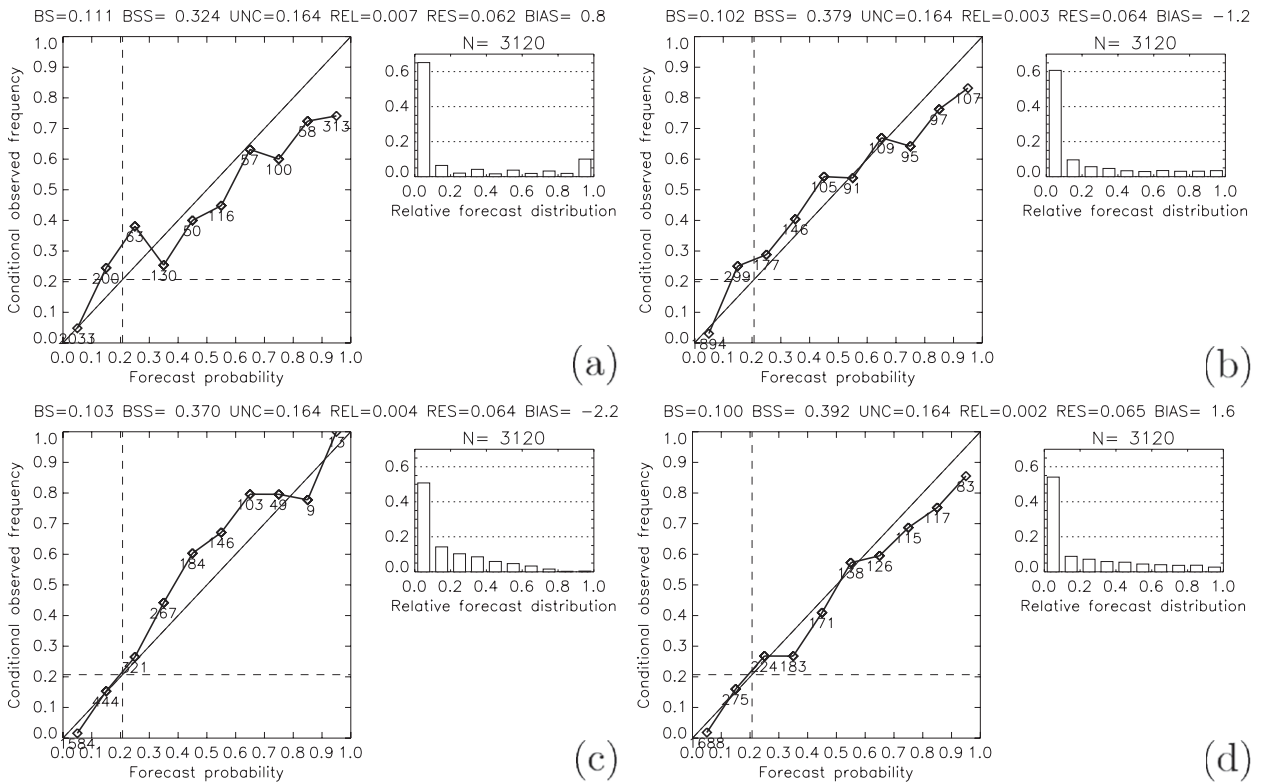


FIG. 5. Reliability diagrams of +30-h area-mean precipitation forecasts for the 5-mm threshold for (a) the raw EPS output, (b) extended logistic regression ($w = 247$), (c) BMA ($w = 247$), and (d) modified BMA ($w = 247$). The verification period is from 1982 to 2001 (autumns only). In these diagrams the observed frequencies are shown, conditional on each of the 10 forecast probabilities (indicated by diamonds). For perfectly reliable forecasts these paired quantities are equal, yielding all points in the diagram falling on the diagonal line. The dashed line indicates the sample climatology. The histogram on the right in each panel portrays the relative forecast distribution. Here BS is short for Brier score, UNC for uncertainty, REL for reliability, and RES for resolution (e.g., Wilks 2006b), and N is the total number of cases.

6. Verification results for area-maximum precipitation forecasts

Section 5 showed that the raw EPS is rather well calibrated using area-mean precipitation as the predictand. A less well-calibrated EPS can be simulated using area-maximum precipitation as the predictand. Moreover, probability forecasts for area-maximum precipitation are relevant in their own right. Therefore, verification results are shown for the area-maximum precipitation forecasts in this section. Again these results are based on spatially pooled cross-validated data from the autumns of the 1982–2001 period ($N = 3120$).

a. Verification of the BMA system with the raw EPS as the reference

Figure 7 shows again the CRPSS of the BMA system as a function of forecast projection, but now for the area-maximum precipitation forecasts. The CRPSS is now positive for the 30-, 54-, and 78-h forecast projections, indicating that the BMA system has more skill than the

raw ensemble for these projections, and it decreases with increasing projection. From the 102-h projection the BMA system is (about) as skillful as the raw ensemble, which can again be improved using modified BMA (section 6b).

b. Verification of raw EPS versus (modified) BMA and extended LR

Figure 8 shows the BSS as a function of the precipitation threshold for the raw EPS, the extended LR system, the BMA system, and the modified BMA system. Figure 8a shows the BSS for the 30-h forecast projection, Fig. 8b for the 78-h projection, and Fig. 8c for the 126-h projection.

In accordance with Fig. 7, it can be concluded from Fig. 8 that the BMA system improves on the skill of the raw EPS for the 30- and 78-h projections but only slightly for the 126-h projection. Again, if the additive bias correction [Eq. (5b)] is applied instead, together with the modified POP formulation [Eq. (3b)], the performance of BMA is substantially improved for the 78- and

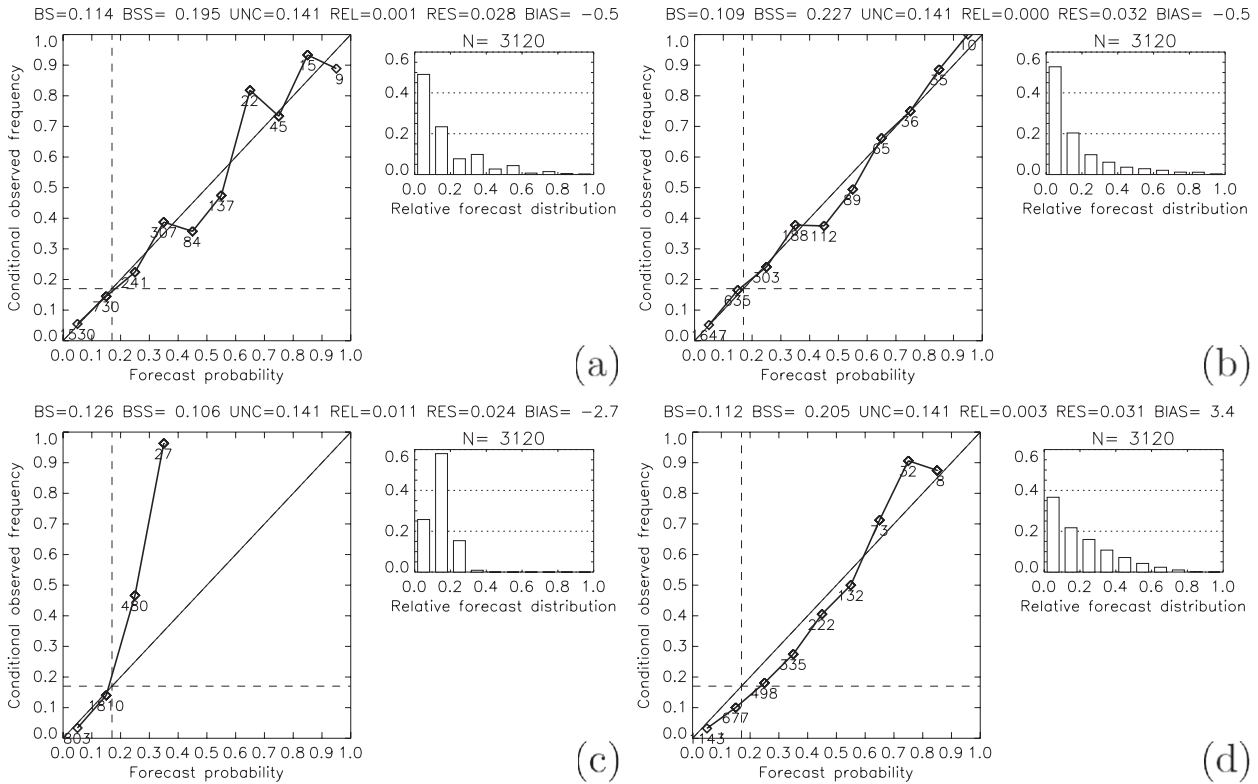


FIG. 6. As in Fig. 5, but for +126-h area-mean precipitation forecasts.

126-h projections (Figs. 8b,c). On the other hand, the modified BMA system is about as skillful as the conventional BMA system for the 30-h projection (Fig. 8a). For all projections the raw EPS can be improved significantly by statistical postprocessing. However, this is to be expected as the EPS is not intended to forecast area-maximum precipitation. Both modified BMA and extended LR are much more skillful than the raw EPS, but the difference in skill between the two postprocessing methods does not seem to be statistically significant. For the extremes these methods show a substantial skill improvement as well (compared with the raw EPS).

To explore the differences between the methods in more detail, reliability diagrams for the 126-h projection are shown in Fig. 9. It shows these diagrams for the raw EPS (Fig. 9a), the extended LR system (Fig. 9b), the conventional BMA system (Fig. 9c), and the modified BMA system (Fig. 9d), all for the 10-mm threshold. The raw EPS is not well calibrated of course, as it underforecasts the area-maximum precipitation amounts to a large extent (Fig. 9a). The extended LR system (Fig. 9b) shows both better reliability and better resolution than the raw EPS. The conventional BMA system (Fig. 9c) shows the same reliability and somewhat better resolution than the raw EPS. On the other hand,

the modified BMA system (Fig. 9d) again shows an enormous skill improvement compared to the conventional BMA system, and has both better reliability and better resolution than the raw EPS. However, it shows

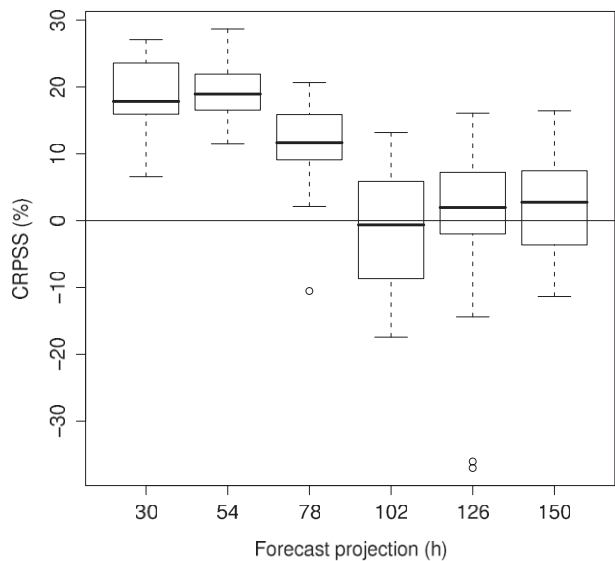


FIG. 7. As in Fig. 3a, but for 24-h accumulated area-maximum precipitation.

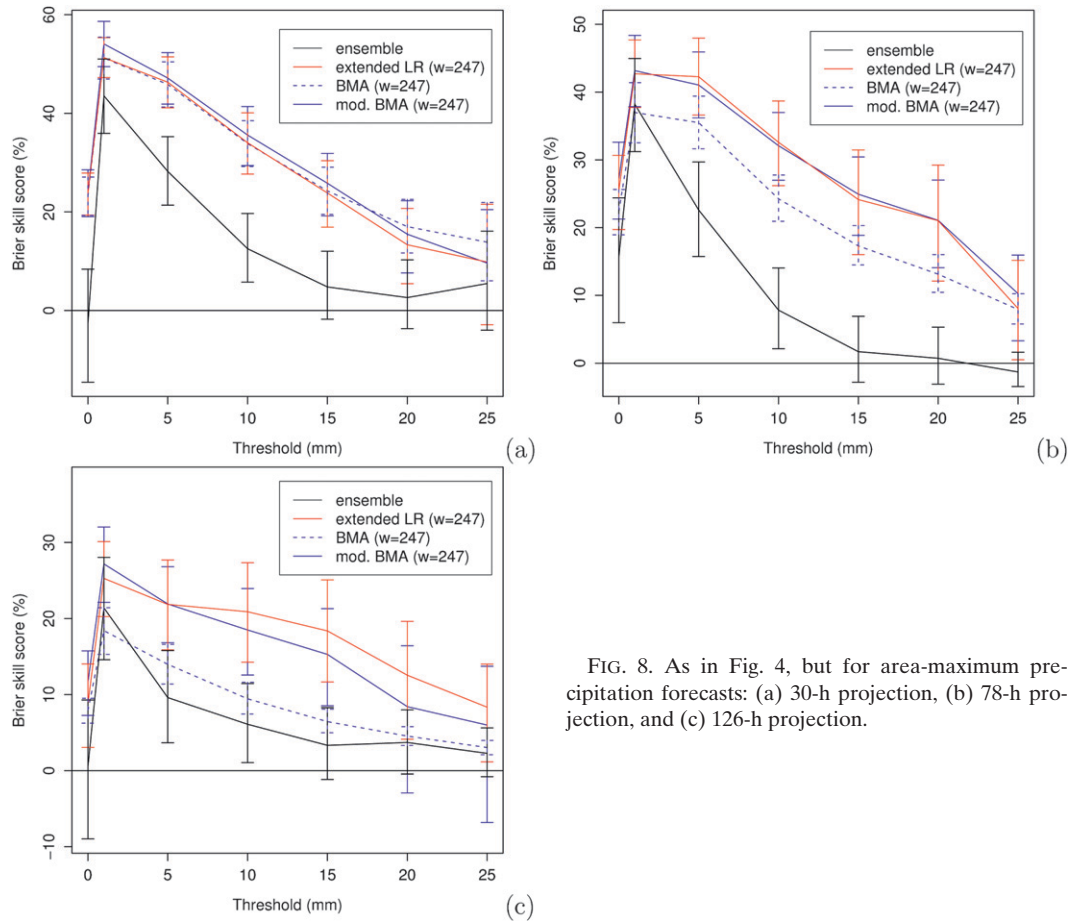


FIG. 8. As in Fig. 4, but for area-maximum precipitation forecasts: (a) 30-h projection, (b) 78-h projection, and (c) 126-h projection.

somewhat worse reliability and resolution than the extended LR system.

7. Summary and discussion

Using a 20-yr ECMWF EPS reforecast dataset of total precipitation (Hamill et al. 2008), and a 20-yr dataset of a dense precipitation observation network in the Netherlands (Fig. 1), a comparison has been made between the raw EPS output and EPS output postprocessed with either BMA (Sloughter et al. 2007; Fraley et al. 2010) or extended LR (Wilks 2009). The raw EPS output turns out to be generally well calibrated up to 6 forecast days, if compared to the area-mean 24-h precipitation sum. Surprisingly, BMA is less skillful than the raw EPS output from forecast day 3 onward (Fig. 3a). This is due to the bias correction in BMA, which applies MOS to individual ensemble members. As a result, the spread of the bias-corrected ensemble members is decreased, especially for the longer forecast projections. Here an additive bias correction has been applied instead, which is similar to the ones proposed by Wilson et al. (2007a)

and Hamill (2007) for Gaussian predictands. Besides, the equation for the probability of precipitation in BMA has also been changed. These modifications to BMA, referred to as “modified BMA,” lead to a significant improvement in the skill of BMA for the longer projections (Fig. 4). If the area-maximum 24-h precipitation sum is used as a predictand, both modified BMA and extended LR improve the raw EPS output significantly for the first 5 forecast days (Fig. 8).

Considering the questions posed in the introduction, the following additional conclusions can be drawn from this study. (i) Modified BMA performs better than conventional BMA (Sloughter et al. 2007) for the longer projections, when applied to a single-model ensemble (in this case the ECMWF EPS), in which case the weights and the parameters of the component pdfs are constrained to be equal (Fraley et al. 2010). Extended LR is also more skillful than conventional BMA for the longer projections, whereas the difference in skill between modified BMA and extended LR does not seem to be statistically significant (Figs. 4 and 8). Extended LR might be preferred, however, because incorporating predictors

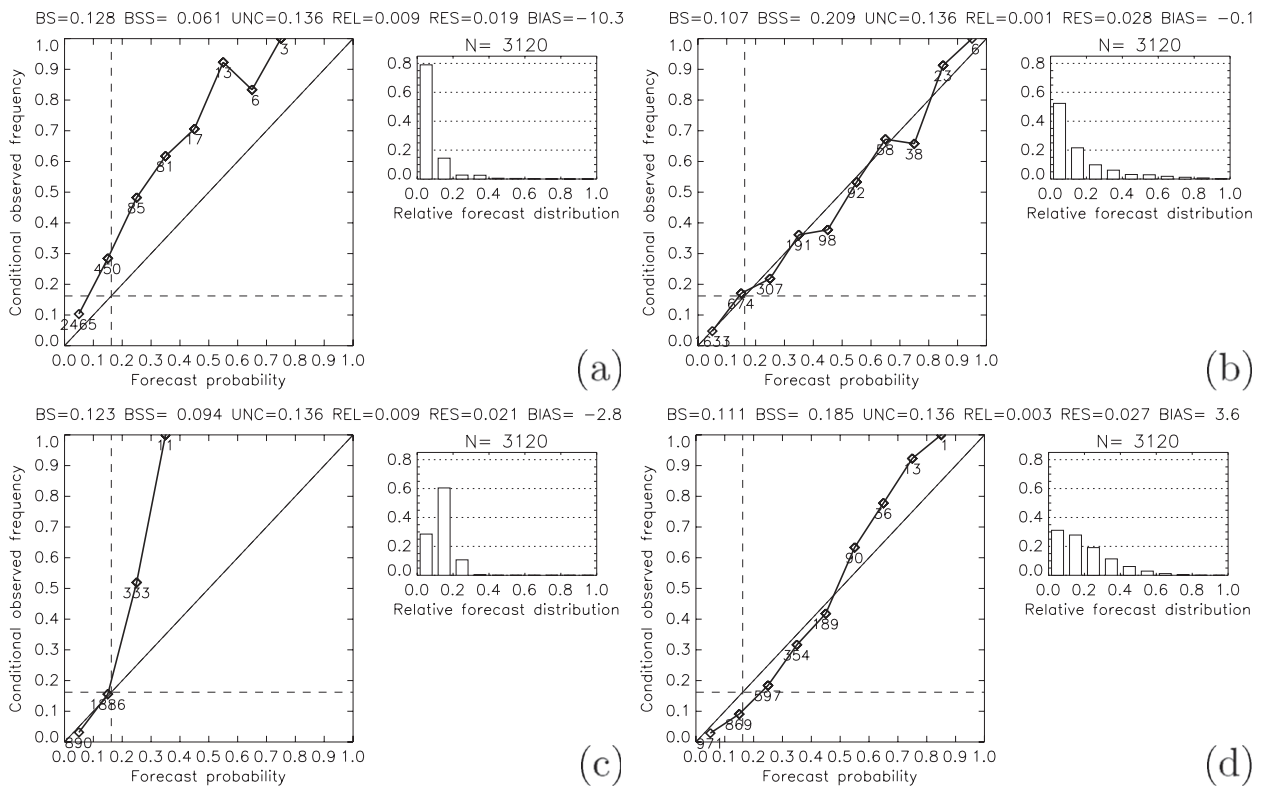


FIG. 9. As in Fig. 5, but for the +126-h area-maximum precipitation forecasts for the 10-mm threshold for (a) the raw EPS output, (b) extended logistic regression ($w = 247$), (c) BMA ($w = 247$), and (d) modified BMA ($w = 247$).

that are different from the predictand is straightforward, in contrast to BMA. (ii) For the extremes of the area-maximum precipitation, both modified BMA and extended LR show a substantial increase in skill compared to the raw EPS (Fig. 8), but they might be improved. As noted before, Bishop and Shanley (2008) adjusted the BMA method, so that it is more appropriate for the extremes, but to date that was only done for Gaussian predictands. The behavior of extended LR might be improved by including nonprecipitation variables (e.g., convective indices) as potential predictors and/or by including more extreme thresholds in the training set. Another approach might be to use conventional LR for the extremes, with conditional instead of absolute probabilities for the higher thresholds, which is an approach similar to Schmeits et al. (2005, 2008). (iii) The expansion of the training period to include weekly data from the same season beyond 5 previous years (i.e., 65 cases) generally only slightly improves the performance of BMA (e.g., Fig. 3b) and modified BMA (not shown); probably nearly the same number of training seasons are necessary if daily instead of weekly data would have been available (cf. Figs. 14 and 15 of Hamill et al. 2004).

Wilson et al. (2007a) and Hamill (2007) also proposed similar additive bias corrections, but those corrections

did not generally improve the skill of BMA for Gaussian predictands (Wilson et al. 2007a,b). This should be investigated further, as one would expect the BMA results to improve for the longer forecast projections if a non-MOS based bias correction is applied.

The additive bias correction [Eq. (5b)] improves the performance of BMA significantly, but it works only for underforecast biases or overforecast biases that do not violate either the constraint $b'_{0,k} \geq -f_k^{1/3}$ for nonexchangeable members or $b'_{0,k} = b'_0 \geq -\min(f_1^{1/3}, \dots, f_K^{1/3})$ for exchangeable members. To account also for larger overforecast biases than these constraints allow, another approach might be followed, for instance a correction based upon mapping from the forecast quantile to the observed quantile (Hamill and Whitaker 2006). However, Hamill and Whitaker (2006) noted that there are problems with that quantile remapping, too.

In our study we have used the same variance coefficients c_0 and c_1 [Eq. (6)] for modified BMA as for conventional BMA. Sensitivity tests, where the kernel variance has been increased and decreased by 20% compared to the original value, have shown that the Brier skill scores hardly change. Therefore, the impact of fitting the variance coefficients for modified BMA, using the EM algorithm, would probably be small.

Although we have found that the raw EPS is already well calibrated, Hamill et al. (2008) found that it could be significantly improved using conventional LR. This difference might be related to the differences in temporal and spatial discretization, which was finer in their study, as well as to the different orography in the countries for which the calibration was done: the Netherlands (flat with only a few hilly areas) versus the United States (with many mountainous areas). As is well known, orography plays an important role in the formation of precipitation, and (global) NWP models suffer both from an inaccurate representation of orography due to their relatively low resolution, as well as from deficiencies in the precipitation formation process.

It would be interesting to repeat this study for the latest reforecast dataset that is now available, in which the varEPS system has been used (Hagedorn 2008), and to extend it to other predictands as well. It would also be interesting to investigate how modified BMA and extended LR perform on a multimodel ensemble. Results from such studies could then be used to answer the question how ensemble forecasts can best be calibrated.

Acknowledgments. The Department of Statistics of the University of Washington is gratefully acknowledged for making the ensemble BMA software available online, Renate Hagedorn (ECMWF) for providing the ECMWF EPS reforecast dataset, Janet Wijngaard (KNMI) for the delivery of the precipitation observations, and Caren Marzban (University of Washington) for a useful discussion on verification. Tom Hamill (NOAA/ESRL), two anonymous reviewers, and Seijo Kruizinga are thanked for their useful comments on earlier versions of the manuscript.

REFERENCES

- Bishop, C. H., and K. T. Shanley, 2008: Bayesian model averaging's problematic treatment of extreme weather and a paradigm shift that fixes it. *Mon. Wea. Rev.*, **136**, 4641–4652.
- Brelsford, W. M., and R. H. Jones, 1967: Estimating probabilities. *Mon. Wea. Rev.*, **95**, 570–576.
- Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman and Hall, 436 pp.
- Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.*, **138**, 190–202.
- Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211.
- Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble Model Output Statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118.
- Hagedorn, R., 2008: Using the ECMWF reforecast dataset to calibrate EPS forecasts. *ECMWF Newsletter*, No. 117, ECMWF, Reading, United Kingdom, 8–13.
- Hamill, T. M., 2007: Comments on “Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging.” *Mon. Wea. Rev.*, **135**, 4226–4230.
- , 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.
- Heijboer, D., and J. Nellestijn, 2002: *Klimaatatlas van Nederland*. KNMI, 182 pp.
- Lorenz, E. N., 1996: Predictability: A problem partly solved. *Proc. ECMWF Seminar on Predictability*, Vol. I, Reading, United Kingdom, ECMWF, 1–18. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]
- Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174.
- Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16–30.
- Schmeits, M. J., C. J. Kok, and D. H. P. Vogelegang, 2005: Probabilistic forecasting of (severe) thunderstorms in the Netherlands using model output statistics. *Wea. Forecasting*, **20**, 134–148.
- , —, —, and R. M. van Westrhenen, 2008: Probabilistic forecasts of (severe) thunderstorms for the purpose of issuing a weather alarm in the Netherlands. *Wea. Forecasting*, **23**, 1253–1267.
- 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.
- Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.*, **131**, 2961–3012.
- Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteor. Soc.*, **131**, 965–986.
- Wilks, D. S., 2006a: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteor. Appl.*, **13**, 243–256.
- , 2006b: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.
- , 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.
- Wilson, L. J., S. Beaugard, A. E. Raftery, and R. Verret, 2007a: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 1364–1385.
- , —, —, and —, 2007b: Reply. *Mon. Wea. Rev.*, **135**, 4231–4236.