Asia-Pacific Journal of Atmospheric Sciences (2020) 56:503–514
https://doi.org/10.1007/s13143-019-00143-8

Online ISSN 1976-7951
Print ISSN 1976-7633

ORIGINAL ARTICLE

Bias Correction of Short-Range Ensemble Forecasts of Daily Maximum Temperature Using Decaying Average

Miloslav Belorid[1] · Kyu Rang Kim[1] · Changbum Cho[1]

Received: 25 March 2019 / Revised: 1 July 2019 / Accepted: 23 July 2019 / Published online: 4 December 2019

Abstract

In this study, we assessed the performance of the decaying average bias correction method in removing the systematic error in daily maximum temperature (dTmax) ensemble forecasts. We applied the technique to a short-range high-resolution limited-area ensemble prediction system of the Korea Meteorological Administration, which shows under-predictive and under-dispersive characteristics for dTmax. The bias correction was applied to the grid of the model using spatial interpolation of the decaying average bias from surrounding reference points. The method was verified by evaluating the accuracy of the ensemble mean, spread-skill relationship, and the performance of the probabilistic forecasts. The results showed that the decaying average technique minimized the systematic error in the ensemble mean and improved the performance of the probabilistic forecasts. The overall mean absolute error of the ensemble mean was lowered from 2.2 to 1.2 °C and the root-mean-square error from 2.5 to 1.6 °C. The continuous ranked probability score decreased from 1.9 to 1.0 °C. The reliability of three dichotomous events also improved and the Brier skill scores increased. However, the bias correction only slightly affected the ensemble spread, and the system remained under-dispersive.

Keywords Bias correction · Decaying average · Ensemble forecast · Daily maximum temperature

1 Introduction

The daily maximum temperature (dTmax) is not only an ordinary product of weather forecasts, but is also a widely used conditional parameter for identifying heat waves in many warning systems worldwide (Smoyer-Tomic et al. 2003; Tan et al. 2007; Lowe et al. 2011; Parker et al. 2014; Lyon and Barnston 2017). Although each warning system is different and configured according to local climatological or epidemiological characteristics, the methodology is largely the same. Typically, a heat wave warning is issued if the predicted dTmax exceeds a specific threshold for a certain period. For such systems, an accurate prediction of dTmax is an essential requirement.

Apart from the conventional single-model deterministic approach, several studies used ensemble prediction systems (EPSs) for applications, such as severe weather warning systems (Legg and Mylne 2004; Chessa et al. 2004; Neal et al. 2014; Matsueda and Nakazawa 2015). The main advantage of EPS is its ability to assess the predictability of a predicted event. However, ensemble forecasts, similar to single-model deterministic forecasts, have systematic errors. These errors should be removed in a post-processing step before any further use of the numerical weather prediction (NWP) output. One of the most common post-processing techniques is the model out statistics (MOS). The main disadvantage of this technique is that it requires long training period (commonly more than 2 years). As suggested by several previous studies, model performance changes with time and, therefore, information on recent model performance can be used to improve forecast accuracy (Woodcock and Engel 2005; Baars and Mass 2005). Several studies have already proposed various adaptive bias correction (BC) techniques for EPS that use short training periods. For example, Yussouf and Stensrud (2006) estimated the mean bias of a 12-day window to remove the systematic error in near-surface variables from short-range ensemble forecasts. In another study, they concluded that the BC forecasts of 2-m temperature and dewpoint were more accurate than the MOS of

Communicated by: Seok-Woo Son

✉ Miloslav Belorid
mbelorid@korea.kr

[1] Applied Meteorology Research Division, National Institute of Meteorological Sciences (NIMS), 33, Seohobuk-ro, Seogwipo-si, Jeju-do 63568, Republic of Korea

the Global Forecast System (GFS) (Yussouf and Stensrud 2007). Zhu et al. (2014) used a running mean bias correction approach with a 16-day sample window to correct near-surface variables. Boi (2004) used a simpler iterative method without a fixed sample window to adjust the 2-m temperature of each ensemble member of the ECMWF EPS. In the iterative approach, previous bias information is continuously accumulated and updated from the initial forecast. Cui et al. (2012) proposed another efficient adaptive BC method known as the decaying average BC, which is based on the Kalman filter (1960). They used this technique to correct the global ensemble forecasts of air temperature and geopotential heights. The results showed very good performance for short-range predictions (up to five days). Cui et al. (2012) also suggested that this method can be used either with a fixed-length training period or using the iterative approach. The decaying average was also used by Glahn (2014) to correct the MOS of GFS and Wang et al. (2018) applied this method to adjust the mesoscale EPS of the China Meteorological Administration. The decaying average BC method has mainly been studied for applications to hourly data but studies have not fully evaluated its applicability and optimization for daily forecasts.

In this study, we assessed the performance and applicability of the decaying average BC method for gridded dTmax forecasts from a high-resolution EPS. The BC method was first optimized and then evaluated in terms of the ensemble mean, ensemble spread, and probabilistic forecasts. The initial objective of this study was to integrate the dTmax ensemble forecasts with a new impact-based heat wave warning system for the Republic of Korea. Nevertheless, the results can be useful for any applications wherein dTmax need to be predicted accurately.

The rest of this study is organized as follows. In Section 2, we briefly describe the EPS and data used in this study. The BC and evaluation methods are explained in Section 3. The results of the BC optimization and verification are provided in Section 4. The verification results are divided into three subsections focusing on the ensemble mean, ensemble spread, and probabilistic forecasts. The discussion is given in Section 5, and in Section 6, we present a summary and the conclusions drawn from this study.

2 Ensemble Prediction System and Data

The ensemble dTmax forecasts were produced from the hourly output of a Limited-area ENsemble prediction System (LENS). The LENS was developed by the Korea Meteorological Administration (KMA) (Kim et al. 2015). The ensemble comprises 13 ensemble members produced by Unified Model of the Met Office (Walters et al. 2017). The LENS in the operational mode runs twice a day at 00:00 UTC and 12:00 UTC and provides forecasts 72 h ahead. The horizontal resolution of the grid of the LENS is 3 km, and the domain covers the entire Korean Peninsula, part of Japan, and part of China. The LENS is initialized and forced by downscaling the ensemble members from the Global Ensemble Prediction System (EPSG), and the boundary conditions from the EPSG are applied to the LENS every 12 h (Kim et al. 2015).

The observation-based dTmax was derived from the hourly near-surface air temperature of KMA's Automated Synoptic Observing System (ASOS) and Automatic Weather Station (AWS) Network. We only used stations with a 100% data coverage for the study period. Considering the good spatial resolution of the LENS, each station was paired with model grid points using a nearest-point approach. To account for the discrepancy between the grid point height and actual elevation of the observation site, we applied a standard atmospheric lapse rate (–6.5 °C km^{-1}) before the BC procedure. However, to eliminated large errors due to elevation discrepancy, stations with differences between the actual elevation and the corresponding grid point exceeding 500 m were excluded. The final dataset consisted of 82 ASOS stations and 302 AWSs. Figure 1a shows the locations of the ASOS stations and AWSs used in this study while Fig. 1b shows their elevations.

The 72-h LENS forecasts were divided into three 24-h windows to extract the dTmax value. The 00:00 UTC forecast cycle begins at midday according to local standard time (LST). To obtain the full 24-h window, the data gaps were filled by data from the previous 12:00 UTC forecasts.

3 Methodology

The decaying average BC can be applied to either hourly or daily forecast data. In the first approach, the dTmax is generated from the already bias-corrected hourly temperature forecasts. The disadvantage of this method is the increased computational cost and memory required. For example, the LENS would need to apply the BC 936 times per forecast (i.e., 72 h x 13 members). On the other hand, in the second approach, the dTmax is generated prior to the BC, which reduces the number of BC runs to 39 times per forecast (i.e., 3 d x 13 members). This not only improves the cost efficiency but also significantly reduces the amount of data stored for the training. For these reasons, we used the latter method. The BC was individually applied to each forecast cycle, lead day, and ensemble member.

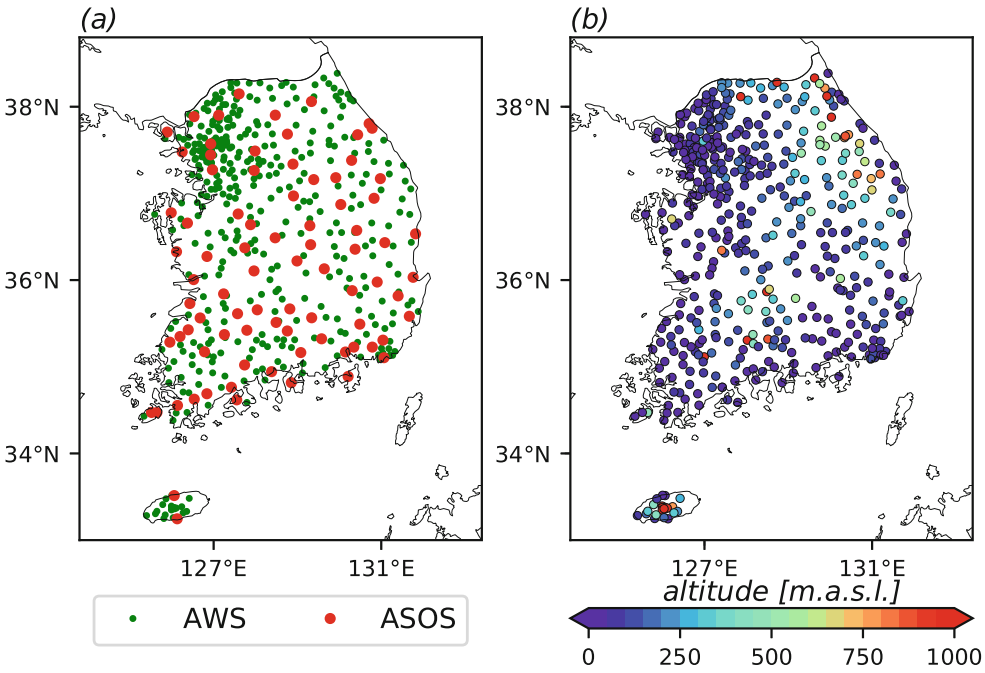

Fig. 1 Maps of the AWS and ASOS locations (**a**) and their actual elevations (**b**)

3.1 Decaying Average Algorithm

The first step in the decaying average algorithm is to estimate the forecast error as the difference between the most recent uncorrected (RAW) forecast at reference time t and corresponding observation y_t:

$$e_t = f_t - y_t. \tag{1}$$

The decaying average bias to correct the next forecast is then computed as

$$B_{(t'+1)} = (1-w)B_{t'} + we_t, \tag{2}$$

where t' indicates the time of most recent forecast, $B_{t'}$ is the accumulated bias from previous forecasts in the sample window until the last available forecast $f_{t'}$, and w is the weighting factor.

Finally, $B_{(t'+1)}$ is used to correct the subsequent RAW forecast $f_{(t'+1)}$:

$$F_{(t'+1)} = f_{(t'+1)} - B_{(t'+1)}. \tag{3}$$

To better understand the BC procedure, two examples are demonstrated in Fig. 2. In the first example (Fig. 2a), e_t is computed between first lead day (LD1) of the most recent forecast $f_{t'}$, which corresponds to reference time t of the last available observation y_t. The LD1 of the subsequent forecast $f_{(t'+1)}$ is then corrected by $B_{(t'+1)}$. In the second example (Fig. 2b), second lead day (LD2) of $f_{(t'+1)}$ is corrected in an identical manner, except the most recent forecast, which corresponds to y_t, is $f_{(t'-1)}$ and, for third lead day (LD3) it is $f_{(t'-2)}$ (not shown).

The main issue with respect to the decaying average BC is the choice of w. As Du and Zhou (2011) pointed out, the w value can be tuned depending on the nature and purpose of a forecasting system. Note that, if $w = 0$, $B_{(t'+1)} = B_{t'}$. Initially, if $B_0 = 0$, all future $B_{t'}$ values become zero; thus, no BC is applied. On the other hand, if $w = 1$, $(1-w)B_{t'} = 0$ and the unweighted e_t is applied to correct the forecast ($B_{(t'+1)} = e_t$). Therefore, to apply the decaying average BC, $w \in \mathbb{R}|(0 < w < 1)$. In this study, e_t was weighted with a w value of 0.14 (14%).

Another issue is the choice of the sample window size. To consider the relevant previous forecast performance information, we used a moving sample window technique. The size of the sample window was 35 days. More details on the selection of w and size of the sample window are provided in Section 4.1.

3.2 Spatial Interpolation

Typically, y_t in Eq. 1 is derived from an objective analysis of meteorological fields, wherein irregularly scattered observation data are interpolated onto a regular grid. In this study, instead of observed dTmax, $B_{(t'+1)}$ was interpolated onto the model grid. $B_{(t'+1)}$ was first computed on N number of grid cells with the corresponding AWS and ASOS. For the remaining grid cells, $B'_{(t'+1)}$ was estimated

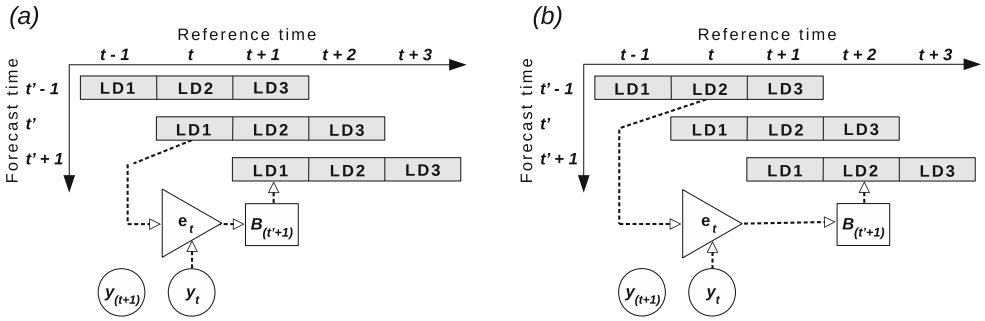

Fig. 2 An example flowchart of how the BC is applied to the LD1 (**a**) and LD2 (**b**) of forecast issued at $t' + 1$

by interpolating the values of known $B_{(t'+1)}$ using inverse distance weighting (IDW) following Shepard's algorithm (Shepard 1968):

$$B'_{(t'+1)} = \frac{\sum_{i=1}^{N} W_i B_{i(t'+1)}}{\sum_{i=1}^{N} W_i}. \tag{4}$$

Here, W_i is the weighting factor, which is inversely proportional to the distance $r(C, C_i)$ between the target grid cell C and the grid cell C_i with the $B_{(t'+1)}$ sample. In the Shepard's algorithm, the weighting factor is defined as

$$W_i = \frac{1}{r(C, C_i)^{\alpha}}. \tag{5}$$

The influence of interpolation can be controlled by power α. In this study, we used $\alpha = 2$.

3.3 Evaluation

To understand how the BC method performs at grid points without observations, we used the leave-one-out cross-validation. In other words, each grid cell with an available reference was validated in an independent test, where $B_{(t'+1)}$ at the station was omitted. Instead, the $B'_{(t'+1)}$ was used to adjust the dTmax forecasts. The BC forecasts at validation grid points were then evaluated against its observation in terms of the ensemble mean, ensemble spread, and probabilistic forecasts.

The accuracy of the ensemble mean was evaluated using standard statistical evaluations methods. As a major metric, we used the MAE. Moreover, the root-mean-square error (RMSE) and mean bias error (MBE) were computed. The RMSE was also used to assess the spread-skill relationship through the spread-error ratio (SER). The SER is simply a ratio of the spread to the RMSE. In an ideal case, the SER equals 1. If SER < 1, the ensemble is under-dispersive; if SER > 1, it is over-dispersive.

To evaluate the probabilistic forecasts, we used the continuous ranked probability score (CRPS), which is an integral of the Brier scores (BS) for all possible thresholds. The CRPS of a single forecast is defined as

$$\mathrm{crps}(F, y) = \int_{-\infty}^{\infty} [F(x) - H(x - y)]^2 dx, \tag{6}$$

where F is the predictive cumulative density function and $H(t - y)$ is the Heaviside function, which takes the value of 1 if $(x - y) \geq 0$; otherwise, it is 0. The CRPS is then averaged over all the forecasts:

$$\mathrm{CRPS} = \frac{1}{n} \sum_{i=1}^{n} crps(F_i, y_i). \tag{7}$$

The probabilistic forecasts of three dTmax thresholds were evaluated using the Brier skill score (BSS):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS_{clim}}}. \tag{8}$$

The BS is the Brier score of the forecast defined as

$$\mathrm{BS} = \frac{1}{n} \sum_{i=1}^{n} (p_i - o_i)^2, \tag{9}$$

where p_i is the predicted probability, and o_i is the binary observation, which takes a value of 1 if the event occurs; otherwise, it is 0. $\mathrm{BS_{clim}}$ is the BS of the sample climatology:

$$\mathrm{BS_{clim}} = \bar{o_i}(1 - \bar{o_i}). \tag{10}$$

The three dTmax thresholds of the probabilistic forecasts are ≥31 °C, ≥33 °C, and ≥35 °C. The ≥33 °C and ≥35 °C thresholds have been used by the KMA to issue heat wave advisory and warning, respectively. The dTmax ≥31 °C condition was recently implemented in the impact-based forecasting system of the KMA.

Apart from the statistical methods explained above, a graphical tool rank histogram was used to identify any major issues in the RAW and BC ensemble forecasts. In the rank histogram, each bin represents the fraction of the observations that fall between the two ranked ensemble members. A rank histogram has ideally a flat shape. An asymmetry in the rank histogram shape indicates bias in the ensemble mean or variance.

The statistical results for the 00:00 UTC and 12:00 UTC cycles were nearly identical. Therefore, for simplicity and

to assess the overall performance of the BC forecasts, the independently corrected data for two forecast cycles were combined and evaluated together as a single dataset.

4 Results

4.1 Optimization

An optimization of the decaying average BC is an essential step before applying the method in an operational forecasting system. In particular, the choice of the decaying weighting factor and size of the moving window largely impacts the performance of the BC. In this study, we performed numerous experiments with different settings for these two parameters. The optimal configuration was then identified based on experiments with the lowest MAE. The dataset used for these tests covered the period from 1 May 2017 to 30 September 2017. Figure 3 shows the distribution of MAE dependent on the choice of weighting factor and sample window size. The MAE was highest when the weight or sample size was too small. For each sample window size, we obtained the w value characterized by a minimal MAE, which yields a curve of optimal w relative to the sample window size (solid line in Fig. 3). The most optimal combination of the two parameters for a particular dataset lies on the curve and can be estimated again by minimizing the MAE of the curve. This approach was applied separately to each forecast cycle and forecast lead day, and the results show some differences (Fig. 3). The optimal w varies between 0.11 and 0.15. The optimal sample window size shows even greater variability, ranging from 22 to 46 days. However, in all six cases at sample size above 20 days, the MAE is nearly constant. For simplicity, a single configuration suitable for all lead days and forecast cycles, can be selected from set of configurations corresponding to a range of acceptable forecast error. For example, if we assume that the increase in MAE by 1% is negligible, we can specify an area of tolerance where MAE is 1% higher than the lowest MAE (see Fig. 3 within the areas indicated by the dashed lines). As shown in Fig. 3, the regions are relatively large, providing sufficient freedom of choice. In practice, any combination within these areas can yield similar results. A decision can now be made

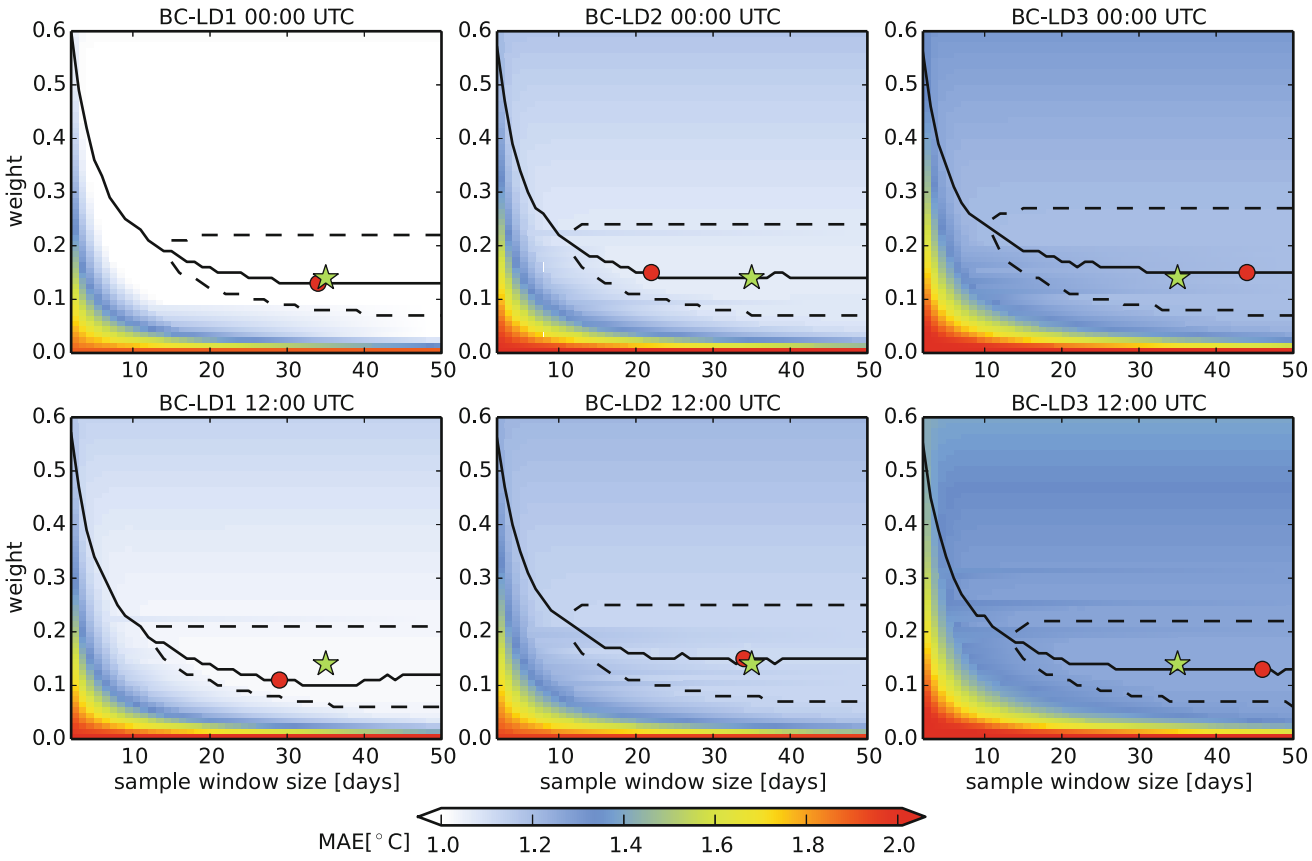

Fig. 3 Distribution of MAE dependent on the choice of weighting factor and sample window size for the LD1, LD2 and LD3 of 00:00 and 12:00 UTC forecast cycle based on the 2017 dataset. The solid line indicates the lowest MAE relative to the weighting factor and sample window size the red point indicates the lowest MAE. The area where the increase in MAE is below 1% is demarcated by dashed line. The final choice of the weighting factor and sample window size is indicated by the green star

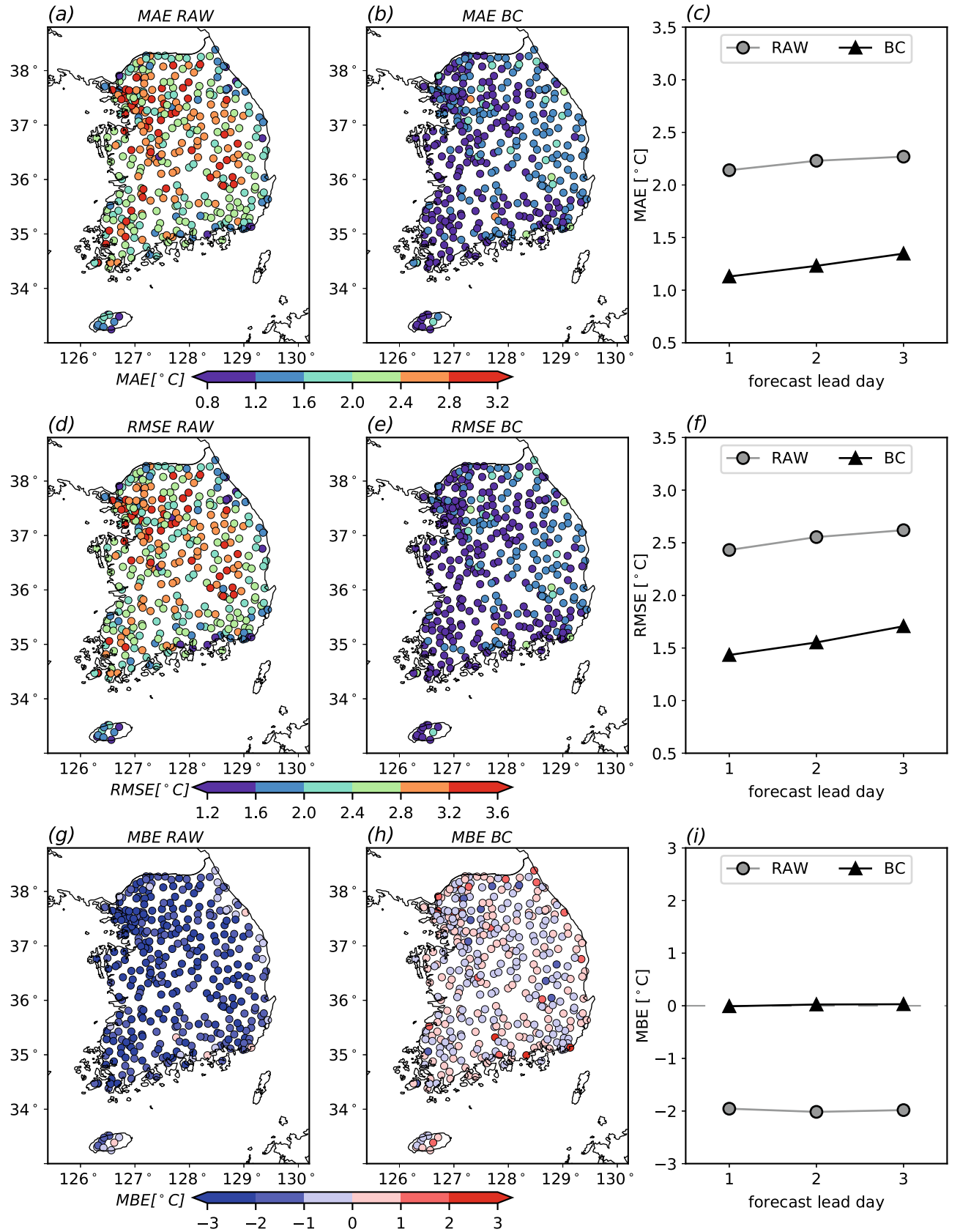

Fig. 4 The spatial distributions and temporal changes by lead day for the RAW and BC forecasts in terms of the MAE (**a**,**b**,**c**), RMSE (**d**,**e**,**f**) and MBE (**g**,**h**,**i**) for the 2018 dataset

that considers other relevant factors, such as computational cost or memory requirements. In this study, we used an average of the six results by lead day and forecast cycle. This method yielded a reasonable compromise, i.e., a combination located within the area of tolerance for all six datasets as follows: $w = 0.14$ and 35-day sample window size (see Fig. 3 indicated by a star).

4.2 Verification

The ensemble mean, ensemble spread, and probabilistic forecasts corrected using the optimized BC setup were verified against observation and compared with the RAW ensemble. To verify the BC, we used data covering period from 1 May 2018 to 30 September 2018. This was also an ideal period to evaluate the dTmax forecasts due to the occurrence of severe heat-wave that plagued northeast Asia. In South Korea, the dTmax reached a historical maximum of 41 °C on 1 August 2018.

4.2.1 Ensemble Mean

In the spatial distribution of the RAW ensemble, the MAE varied between 1.0 and 4.4 °C (Fig. 4a). For 66.2% of the stations, the MAE was above 2 °C. Moreover, at 26 stations, the MAE exceeded 3.0 °C and, even, 4.0 °C at three stations. In the remaining 33.8% of the stations, the MAE fell below 2.0 °C, but only at one station the MAE was lower the 1.0 °C.

The high MAEs are mainly located in north-western regions. We found a relationship between the MAE and dTmax. From the boxplot shown in Fig. 5a, it is clear that

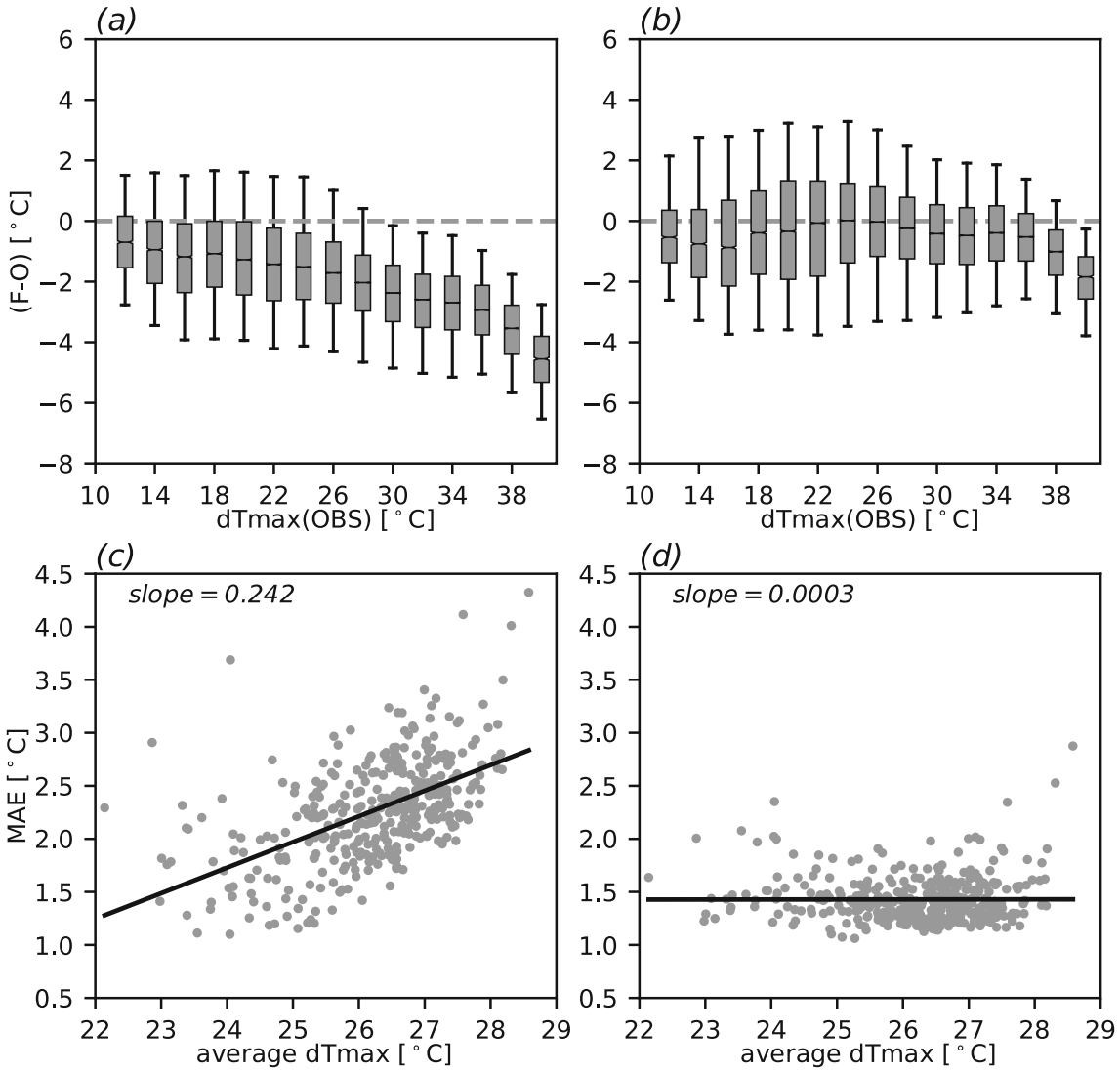

Fig. 5 Forecast errors for the RAW (**a**) and BC (**b**) forecasts at the range of observed dTmax (categorized with 2.0 °C interval), and the relationship between the average dTmax at stations and the MAE of the RAW (**c**) and BC (**d**) ensembles

negative forecast errors increase with respect to the dTmax. Consequently, the results for warmer regions were more erroneous, which is also shown in Fig. 5c, wherein the MAE increases with an increase in the average dTmax of a specific station.

After applying the BC, the MAE decreased at all the stations and was concentrated in a narrower range of 0.9-2.5 °C (Fig. 4b). In 97.9% of all the stations, the MAE fell below 2.0 °C and, at 31 stations, the MAE was even lower than 1.0 °C. In only 2.1% of all stations, the MAE fell between 2.0 °C and 3.0 °C. A slightly higher MAE was found in mountainous regions along the eastern part of Korea (Fig. 4b). This is an expected behavior since terrain complexity and related atmospheric processes in mountainous regions are generally not well resolved in NWPs.

The BC reduced the negative forecast error over the entire range of dTmax (Fig.5b). The relationship between the MAE and average dTmax of a station shows almost no trend (Fig. 5d).

The change in the MAE with respect to the lead days is as expected (Fig. 4c). The MAEs of the RAW and BC forecast increased with the forecast lead time. Overall, the BC helped decrease the total MAE (averaged over all stations and lead days) from 2.2 to 1.2 °C.

For the RMSE, the spatial (Fig. 4d and e) and temporal distributions (Fig. 4f) were similar to those of the MAE. The total RMSE decreased from 2.5 in the RAW to 1.6 °C in the BC forecasts.

According to the MBE, the RAW ensemble underestimates the dTmax at nearly all stations (Fig. 4g). The values varied between -4.3 and 0.8 °C. A negative MBE was obtained for 379 stations, whereas the MBE was positive only at five stations. However, after the BC, the MBE fell between –1.8 and 2.4 °C and the number of stations with a positive MBE increased to 180 compared with 204 stations that had a negative MBE (Fig. 4h). The MBE was nearly constant with respect to lead day for both RAW and BC forecast (Fig. 4i). Overall, the BC improved the total MBE from –2.0 to 0.02 °C.

4.2.2 Ensemble Spread

The ensemble spread is the mean standard deviation of the ensemble forecasts. It is an indicator of the theoretical ensemble forecast precision. A small spread indicates a high accuracy, while a large spread indicates a low accuracy. To identify a potential bias in the ensemble spread, we performed a rank histogram test. For the RAW forecasts, the rank histogram was right-side elevated (Fig. 6a), indicating under-predictive forecasts. After the BC, the rank histogram changed to a U-shape, which typically indicates an insufficient spread (Fig. 6a). If we compare the rank histograms of each lead day, we see that the ensemble forecasts at all three lead days were under-dispersive; however, their histograms got flatter with each lead day (Fig. 6b).

The spatial distribution showed the highest spread for coastal areas and near the north-west Korean border (Fig. 7a and b). The lowest spread was obtained in the inland regions, but also in the south-west coastal areas. The spatial distribution of the spread was affected minimally by the BC (Fig. 7b). In the RAW ensemble, the SER was highest for the east coast and lowest for the south-east inland and north-west regions.

The spread of both RAW and BC experiments almost linearly increased with respect to the lead days (Fig. 7c). After applying the BC, the spread slightly increased over the RAW ensemble on the LD2 and LD3; however, the difference was minimal.

For the RAW forecasts, the SER at most of the inland areas fell below 0.5. After applying the BC, the SER improved at almost all the stations (Fig. 7e), and the distribution was consistent with the distribution of the RMSE.

Figure 7f shows a comparison of the SERs at three lead days. Both the RAW and BC ensembles showed SER < 1, indicating an insufficient spread. The lower RMSE of BC ensemble positively contributed to SER; thus, the SER of BC ensemble tended to 1.0. The increase in the spread by lead time resulted in an improvement in the SER, and the

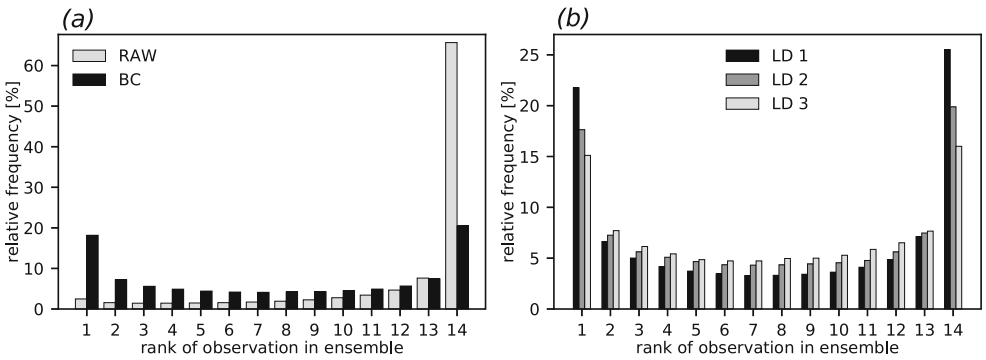

Fig. 6 Rank histograms of the RAW and BC ensembles, regardless of lead day (**a**), and histograms of the BC ensemble according to the lead day (**b**)

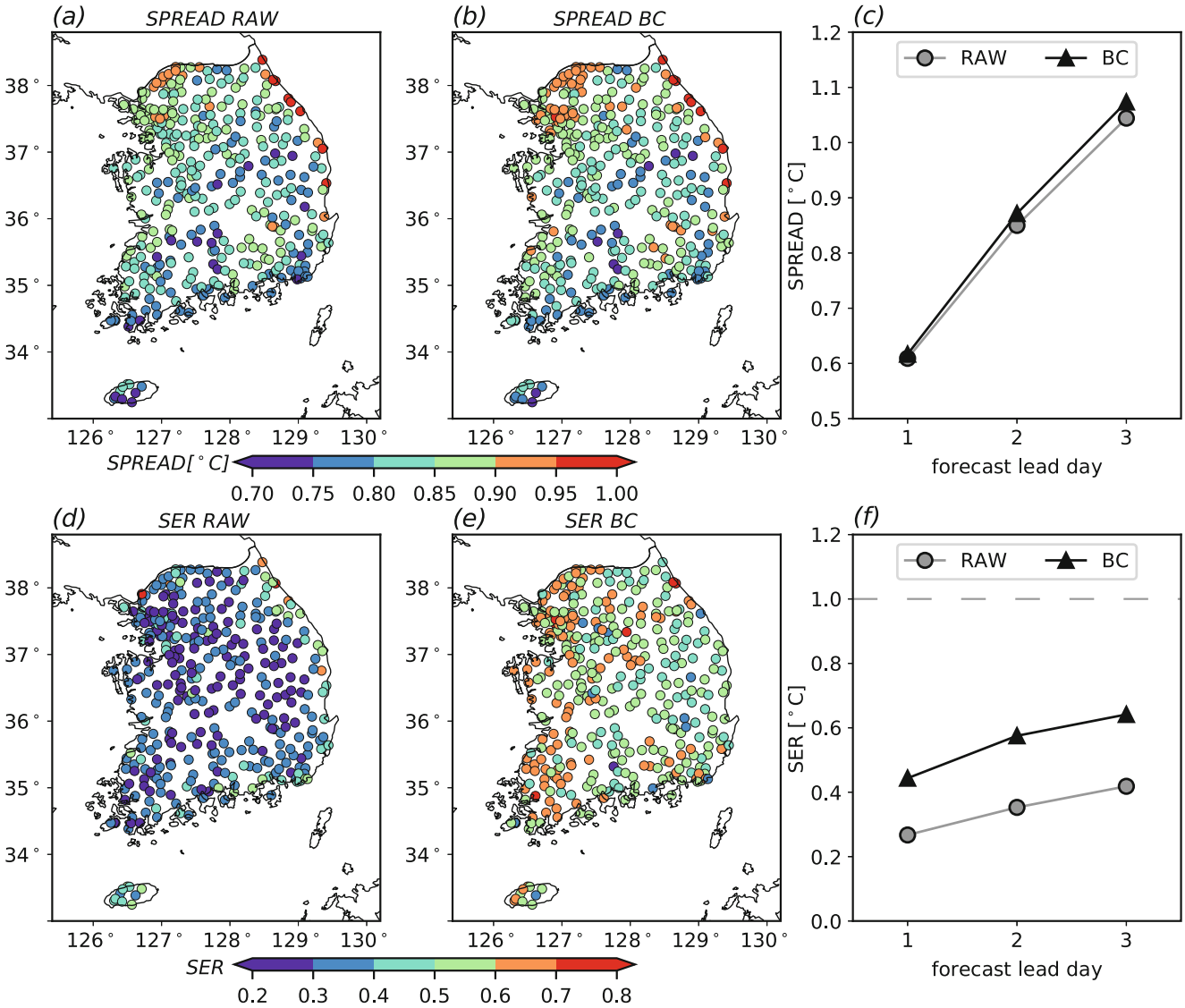

Fig. 7 The spatial distributions and temporal changes by lead day for the RAW and BC ensembles in terms of the SPREAD (**a,b,c**) and SER (**d,e,f**) for the 2018 dataset

best SER was reached on the LD3. This was in contrast to the skills of the ensemble mean, which showed the best performance on the LD1 and worst performance on the LD3 (see Fig. 4c and f).

4.2.3 Probabilistic Forecasts

The spatial variability of the CRPS (Fig. 8a and b) appears to be very similar to those of the MAE or RMSE (see Fig. 4c and f). In the RAW ensemble, stations with the highest CRPS were concentrated in the inland regions, whereas the ones with the lowest CRPS were found near the east coast. After the BC, the CRPS improved nearly at all stations. In the BC ensemble, a slightly higher CRPS was located along the mountainous region on the east of the country. The temporal variation in the spatially averaged CRPS showed a difference between the RAW and BC ensembles. In the BC ensemble, the CRPS increased with the lead day, whereas in the RAW ensemble, the CRPS decreased (Fig. 8c). Using the BC, the total CRPS decreased from 1.9 to 1.0 °C.

The probabilistic forecasts of ≥31 °C, ≥33 °C and ≥35 °C thresholds were obtained using normal distribution and evaluated using reliability diagrams and BSS (Fig. 9). The BSS may behave erratically at stations with rare events. Therefore, we did not analyze the spatial distribution of the BSS. For the RAW ensemble, the reliability curves of the three threshold probability forecasts lie far above the

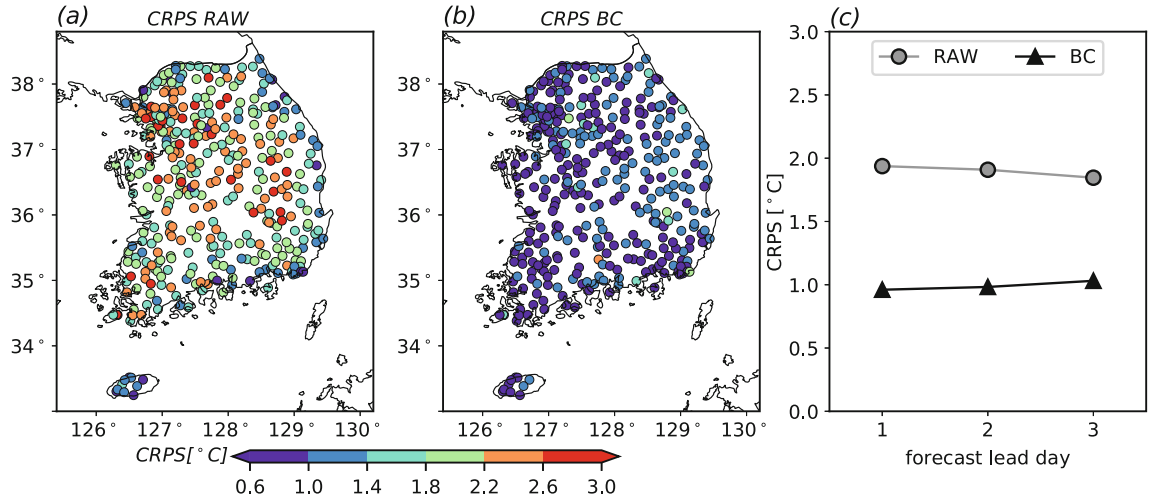

Fig. 8 The spatial distributions and temporal changes by lead day for the RAW and BC ensembles in terms of the CRPS (**a**,**b**,**c**) for the 2018 dataset

diagonal, indicating under-predictive forecasts in all the probability categories. On the other hand, the curves of the BC reliability diagrams shifted toward the diagonal. For the ≥31 °C threshold, the probability forecasts were almost perfectly reliable.

Based on the lead day, the forecasts on the LD1 appears less reliable. We believe that, the main reason for this is the under-dispersive characteristic of the ensemble, which is more serious on the LD1. In contrast, based on the BSS, the best performance was given by the ≥31 °C forecast on the LD1 (BSS = 0.738). Similarly, the ≥33 °C forecasts showed higher BSS on the LD1 compared to the LD2 and LD3. The worst BC ensemble performance was given by the ≥35 °C probabilistic forecast on the LD3 with BSS = 0.471. However, it still outperformed the RAW ensemble with very low BSS = 0.031.

5 Discussion

The main issue associated with the decaying average BC is the choice of the weighting factor. The value of the weighting factor used in this study was selected based on the performance of the dTmax forecasts by minimizing the MAE. The final choice, i.e., $w = 0.14$, differs significantly from values used in previous studies. For example, Wang et al. (2018) used a w value of 0.02 based on the 50-day sample window size (1/50). However, according to our results, a value of 0.02 would significantly under-perform with respect to the $w = 0.14$ setup if applied to LENS-based dTmax forecasts. In particular, if the 50-day sample window size is used, the overall MAE of a $w = 0.02$ setup would increase by 10% as compared with $w = 0.14$ setup. In most previous studies, the decaying average BC was applied to hourly forecasts but the final choice of weighting factor was done regardless of the hour of the day. However, errors in weather forecasts often show diurnal variability (Yussouf and Stensrud 2007; Glahn et al. 2009; Veenhuis 2013; Glahn 2014, etc.). Consequently, if the BC is applied to hourly datasets, and a single value of w is used for all forecast hours, then the choice of w must take into account the error at each hour of the day, which forces a compromise. For example, a weighting factor that is too large may perform well for middays but can possibly under-perform for evenings. This inhibits the full potential of the decaying average method. On the other hand, the dTmax forecasts typically refer to approximately the same hour of the day. For the LENS, the daily bias is relatively high and nearly always negative. The only compromises required in this study were between the performance of the two forecast cycles, which, in fact, show very small differences in the overall MAE (i.e., less than 0.2 °C) and 3 lead days, as was shown in Section 4.1. This allows to focus on only the bias during the part of the day when temperature reaches daily maximum. As a result, more weight can be given to the error of the last forecast.

Another interesting issue is the choice of the sample window. Identical to the weighting factor, we considered the performance of the dTmax forecast. The results showed that the sample window size should be sufficiently large. However, forecast performance is less sensitive to the choice of sample window size than the choice of the weighting factor, which allows more freedom with respect to the final decision. For this study, 35-day day sample window was sufficiently large while not overly expensive in terms of the computational cost or memory requirements.

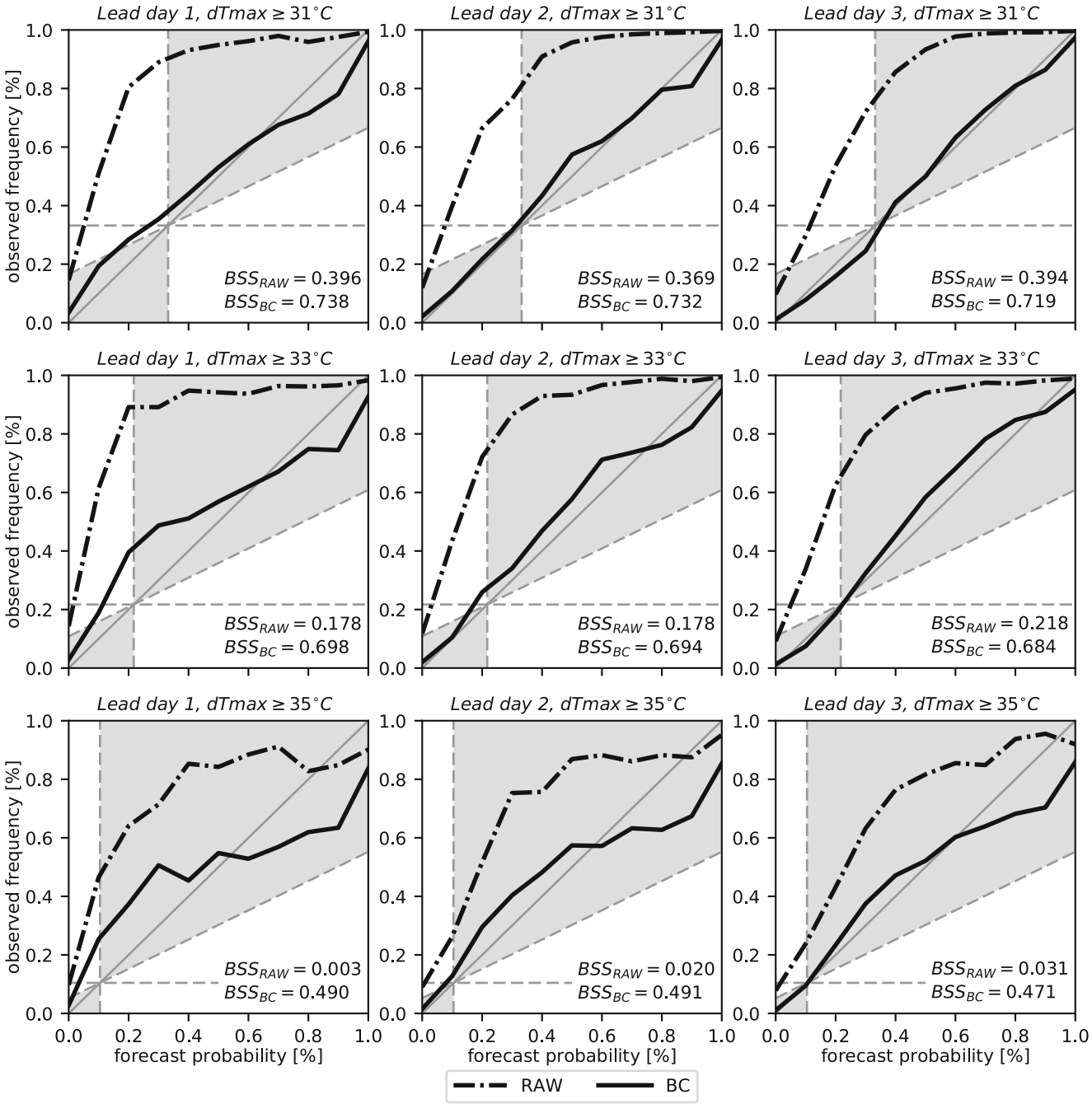

Fig. 9 Reliability diagrams of the RAW and BC ensemble probabilistic forecasts according to the lead days and dTmax thresholds. The solid gray line indicates a perfectly reliable forecast, vertical dashed line is the resolution line, and the dashed line between them is the no-skill line. The shaded area indicates a positive contribution from the forecast to the skill with respect to sample climatology

6 Summary and Conclusion

In previous studies, the decaying average method was used mainly to adjust the hourly data of various meteorological variables. In this work, we demonstrated that the decaying average BC can also be successfully applied to correct the systematic errors of dTmax. We tested the method on high-resolution short-range ensemble forecasts with a strong negative bias. The insufficiency in the spatial coverage of the observation points was solved by interpolating the decaying average bias using the IDW algorithm. Using this approach, the BC can be applied at any grid point of the domain.

By applying the decaying average BC to correct the dTmax of LENS, we were able to reduce the overall MAE and RMSE of the ensemble mean by approximately 1.0

°C. Moreover, the averaged MBE of the BC ensemble was adjusted near to 0.

We demonstrated that the decaying average BC is beneficial not only for a deterministic type of forecast but also for probabilistic forecasts. The BC helped improve the CRPS at all the stations. The BSS of all the three threshold exceedance events increased after BC. We found that the reliabilities of the forecast before and after the BC are worse on LD1 compared to the LD2 and LD3. This seems to be a result of the under-dispersive forecast, which is more evident on the LD1. As the BC did not improve the spread, the ensemble forecasts remained under-dispersive. The spread-skill relationship was improved, but only due to the reduction in the RMSE of the ensemble mean.

Overall, owing to its simplicity and low memory requirements, the decaying average BC method can be easily applied to EPS to correct systematic errors in the dTmax. Before applying the method on the EPS, we recommend optimizing the decaying weighting factor, because the optimal value for the target EPS may differ from the values found in literature. Moreover, the inability of this method to correct the spread of the under-dispersive EPS should be taken into account.

Acknowledgements This work was funded by the Korea Meteorological Administration Research and Development Program "Advanced Research on Biometeorology and Industrial Meteorology" under Grant (1365003004). The authors wish to also thank Sun-il Kwon for providing guidance on LENS dataset, Dr. Ju-Young Shin for valuable comments on statistical evaluation and two anonymous reviewers for very constructive suggestions.

References

Baars, J.A., Mass, C.F.: Performance of national weather service forecasts compared to operational, consensus, and weighted model output statistics. Weather Forecast. **20**, 1034–1047 (2005)

Boi, P.: Probabilistic temperature forecast by using ground station measurements and ECMWF ensemble prediction system. Meteorol. Appl. **11**, 301–309 (2004)

Chessa, P.A., Ficca, G., Marrocu, M., Buizza, R.: Application of a limited-area short-range ensemble forecast system to a case of heavy rainfall in the Mediterranean region. Weather Forecast. **19**, 566–581 (2004)

Cui, B., Toth, Z., Zhu, Y., Hou, D.: Bias correction for global ensemble forecast. Weather Forecast. **27**, 396–410 (2012)

Du, J., Zhou, B.: A dynamical performance-ranking method for predicting individual ensemble member performance and its application to ensemble averaging. Mon. Weather Rev. **139**, 3284–3303 (2011)

Glahn, B., Peroutka, M., Wiedenfeld, J., Wagner, J., Zylstra, G., Schuknecht, B.: MOS uncertainty estimates in ensemble framework. Mon. Weather Rev. **137**, 246–268 (2009)

Glahn, B.: Determining an optimal decay factor for bias-correcting MOS temperature and dewpoint forecasts. Weather Forecast. **29**, 1076–1090 (2014)

Kalman, R.E.: A new approach to linear filtering and prediction problems. Trans. ASME J. Basic Eng. **82**, 35–45 (1960)

Kim, S., Kim, H.M., Kay, J.K., Lee, S.-W.: Development and evaluation of the high resolution limited area ensemble prediction system in the Korea meteorological administration. Atmosphere. **25**, 67–83 (2015)

Legg, T.P., Mylne, K.R.: Early warnings of severe weather from ensemble forecast information. Weather Forecast. **19**, 891–906 (2004)

Lowe, D., Ebi, K.L., Forsberg, B.: Heatwave early warning systems and adaptation advice to reduce human health consequences of heatwaves. Int. J. Environ. Res. Public Health. **8**, 4623–4648 (2011)

Lyon, B., Barnston, A.G.: Diverse characteristics of US summer heat waves. J. Clim. **30**, 7827–7845 (2017)

Matsueda, M., Nakazawa, T.: Early warning products for severe weather events derived from operational medium-range ensemble forecasts. Meteorol. Appl. **22**, 213–222 (2015)

Neal, R.A., Boyle, P., Grahame, N., Mylne, K., Sharpe, M.: Ensemble based first guess support towards a risk-based severe weather warning service. Meteorol. Appl. **21**, 563–577 (2014)

Parker, T.J., Berry, G.J., Reeder, M.J.: The structure and evolution of heat waves in southeastern Australia. J. Clim. **27**, 5768–5785 (2014)

Shepard, D.: A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 1968 23rd ACM National Conference, pp. 517–524. ACM (1968)

Smoyer-Tomic, K.E., Kuhn, R., Hudson, A.: Heat wave hazards: An overview of heat wave impacts in Canada. Nat. Hazards. **28**, 465–486 (2003)

Tan, J., Zheng, Y., Song, G., Kalkstein, L.S., Kalkstein, A.J., Tang, X.: Heat wave impacts on mortality in Shanghai, 1998 and 2003. Int. J. Biometeorol. **51**, 193–200 (2007)

Veenhuis, B.A.: Spread calibration of ensemble MOS forecasts. Mon. Weather Rev. **141**, 2467–2482 (2013)

Walters, D., Brooks, M., Boutle, I., Melvin, T., Stratton, R., Vosper, S., Wells, H., Williams, K., Wood, N., Allen, T., et al: The Met office unified model global atmosphere 6.0/6.1 and JULES global land 6.0/6.1 configurations. Geosci. Model Dev. **10**, 1487–1520 (2017)

Wang, J., Chen, J., Du, J., Zhang, Y., Xia, Y., Deng, G.: Sensitivity of ensemble forecast verification to model bias. Mon. Weather Rev. **146**, 781–796 (2018)

Woodcock, F., Engel, C.: Operational consensus forecasts. Weather Forecast. **20**, 101–111 (2005)

Yussouf, N., Stensrud, D.J.: Prediction of near-surface variables at independent locations from a bias-corrected ensemble forecasting system. Mon. Weather Rev. **134**, 3415–3424 (2006)

Yussouf, N., Stensrud, D.J.: Bias-corrected short-range ensemble forecasts of near-surface variables during the 2005/06 cool season. Weather Forecast. **22**, 1274–1286 (2007)

Zhu, J., Kong, F., Lei, H.: Bias-corrected short-range ensemble forecasts for near-surface variables during the summer season of 2010 in North China. Atmos. Oceanic Sci. Lett. **7**, 334–339 (2014)