Removal of Systematic Model Bias on a Model Grid

CLIFFORD F. MASS, JEFFREY BAARS, GARRETT WEDAM, ERIC GRIMIT, AND RICHARD STEED

Department of Atmospheric Sciences, University of Washington, Seattle, Washington

(Manuscript received 20 December 2006, in final form 19 October 2007)

ABSTRACT

Virtually all numerical forecast models possess systematic biases. Although attempts to reduce such biases at individual stations using simple statistical corrections have met with some success, there is an acute need for bias reduction on the entire model *grid.* Such a method should be viable in complex terrain, for locations where gridded high-resolution analyses are not available, and where long climatological records or long-term model forecast grid archives do not exist. This paper describes a systematic bias removal scheme for forecast grids at the surface that is applicable to a wide range of regions and parameters.

Using observational data and model forecasts over the Pacific Northwest, a method was developed to reduce the biases in gridded 2-m temperature, 2-m dewpoint temperature, and 12-h precipitation forecasts. The method first estimates bias at observing locations using errors from forecasts that are similar to the current forecast. These observed biases are then used to estimate bias on the model grid by pairing model grid points with stations that have similar elevation and/or land-use characteristics.

Results show that this approach reduces bias substantially, particularly for periods when biases are large. Adaptations to weather regime changes are made within a short period, and the method essentially "shuts off" when model biases are small. With modest modifications, this approach can be extended to additional variables.

1. Introduction

Virtually all weather prediction models possess substantial systematic bias, errors that are relatively stable over days, weeks, or longer. Such biases occur at all elevations but are generally largest at the surface where deficiencies in model physics and surface specifications are often substantial. For example, systematic bias in 2-m temperature (T2) is familiar to most forecasters, with a lack of diurnal range often apparent in many forecasting systems [see Fig. 1 for an example for the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5)].

In the United States, the removal of systematic bias is only attempted operationally at observation sites as a by-product of applying model output statistics (MOS) as a forecast postprocessing step (Glahn and Lowry 1972). In fact, it has been suggested by some (e.g., Neilley and Hanson 2004) that bias removal is the most important contribution of MOS and might be completed in a more economical way. As noted in Baars and Mass (2005), although MOS reduces the average forecast bias over extended periods, for shorter intervals of days to weeks, MOS forecasts can possess large biases. A common example occurs when models fail to maintain a shallow layer of cold air near the surface for several days to a few weeks; MOS is usually incapable of compensating for such transient model failures and produces surface temperature forecasts that are too warm. MOS also requires an extended developmental period (usually at least 2 yr) during which the model is relatively stable, a problem for a model undergoing continuous improvement. One approach to reducing a consistent, but short-term, bias using MOS is updatable MOS (UMOS) as developed at the Canadian Meteorological Center (Wilson and Vallée 2002). The method proposed in this paper is related to updatable MOS but extends it in significant ways.

It has become increasingly apparent that bias removal is necessary on the entire model grid, not only at observation locations. For example, the National Weather Service has recently switched to the Interac-

Corresponding author address: Prof. Clifford F. Mass, Dept. of Atmospheric Sciences, University of Washington, Box 351640, Seattle, WA 98195.
E-mail: cliff@atmos.washington.edu

DOI: 10.1175/2007WAF2006117.1

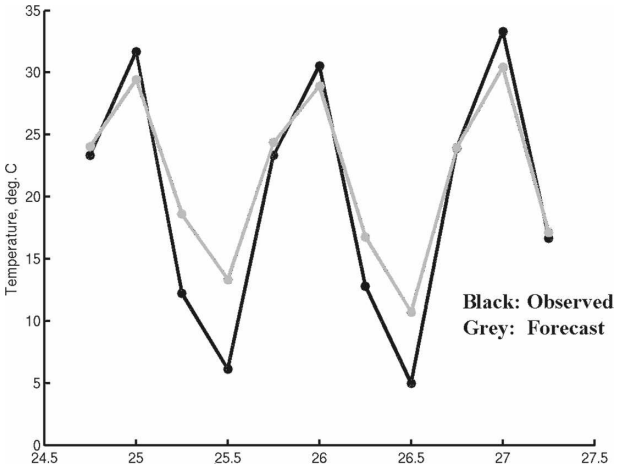

FIG. 1. Observed (black) and MM5 forecast (grey) T2 at Burns, OR. The MM5 simulation was initialized at 1200 UTC 24 Aug 2005.

tive Forecast Preparation System (IFPS), a graphical forecast preparation and dissemination system in which forecasters input and manipulate model forecast grids before they are distributed in various forms (Ruth 2002; Glahn and Ruth 2003). Systematic model biases need to be removed from these grids, and it is a poor use of limited human resources to have forecasters manually eliminating model biases if an objective system could do so. Additionally, it would be surprising if subjective bias removal could be as skillful as automated approaches, considering the large amount of information necessary to complete this step properly, and the fact that biases can vary in space and time. Removal of systematic bias away from observation sites is also needed for a wide range of applications from wind energy prediction and transportation to air quality modeling and military requirements, to name only a few. Finally, bias removal on forecast grids is an important postprocessing step for ensemble prediction, since systematic bias is knowable and therefore not a true source of forecast uncertainty. Thus, systematic model bias for each ensemble member should be removed as an initial postprocessing step or the ensemble variance will be inflated. Eckel and Mass (2005) demonstrated that a simple grid-based, 2-week, running-mean bias correction (BC) increased the reliability of the forecast probabilities from an ensemble system by adjusting the ensemble mean toward reality and increasing sharpness/resolution through the removal of unrepresentative ensemble variance.

The need for model bias removal has been discussed in a number of papers, with most limited to bias reduction at observation locations. Stensrud and Skindlov (1996) found that model (MM5) 2-m temperature errors at observation locations over the southwest United States during summer could be considerably reduced using a simple BC scheme that removes the average bias over the study period. Stensrud and Yussouf (2003) applied a 7-day running-mean bias correction to each forecast of a 23-member ensemble system for 2-m temperature and dewpoint; the resulting bias-corrected ensemble-mean forecasts at observation locations over New England during summer 2002 were comparable to Nested Grid Model (NGM) MOS results for temperature and superior for dewpoint. A Kalman filter approach was used to create diurnally varying forecast bias corrections to 2-m temperatures at 240 sites in Norway (Homleid 1995). This approach removed much of the forecast bias when averaged over a month, although the standard deviations of the differences between the forecasts and observations remained nearly unchanged.

Systematic bias removal on grids, as discussed in this paper, has received less emphasis. As noted above, Eckel and Mass (2005) applied bias removal to gridded MM5 forecasts used in an ensemble forecasting system before calculating the ensemble means and probabilistic guidance. The corrections were based on average model biases over the previous 2-week period using analysis grids [i.e., the 20-km version of the Rapid Update Cycle (RUC20) or the mean of operational National Centers for Environmental Prediction (NCEP) analyses] as truth. Yussouf and Stensrud (2006) interpolated the preceding 12-day biases at observing sites to a model grid using a Cressman (1959) scheme and used these biases to modify an ensemble of forecasts; the result substantially enhanced the prediction of surface variables over Oklahoma. The National Weather Service has recently developed a gridded MOS system that, like conventional MOS, reduces systematic bias (Dallavalle and Glahn 2005). This system starts with MOS values at observation sites and then interpolates them to a grid using a modified Cressman scheme that considers both station and gridpoint elevations. In addition, surface type is considered, with the interpolation using only land (water) MOS locations for land (water) grid points.

An optimal bias removal scheme for forecast grids should have a number of characteristics. It must be robust and applicable to any type of terrain. It must work for a variety of resolutions and particularly for grid spacings at which mesoscale models will be run in the near future (1–10 km). It should be capable of dealing with regions of sparse data, yet also be able to take advantage of higher data densities when they are available. It must be viable where gridded high-resolution

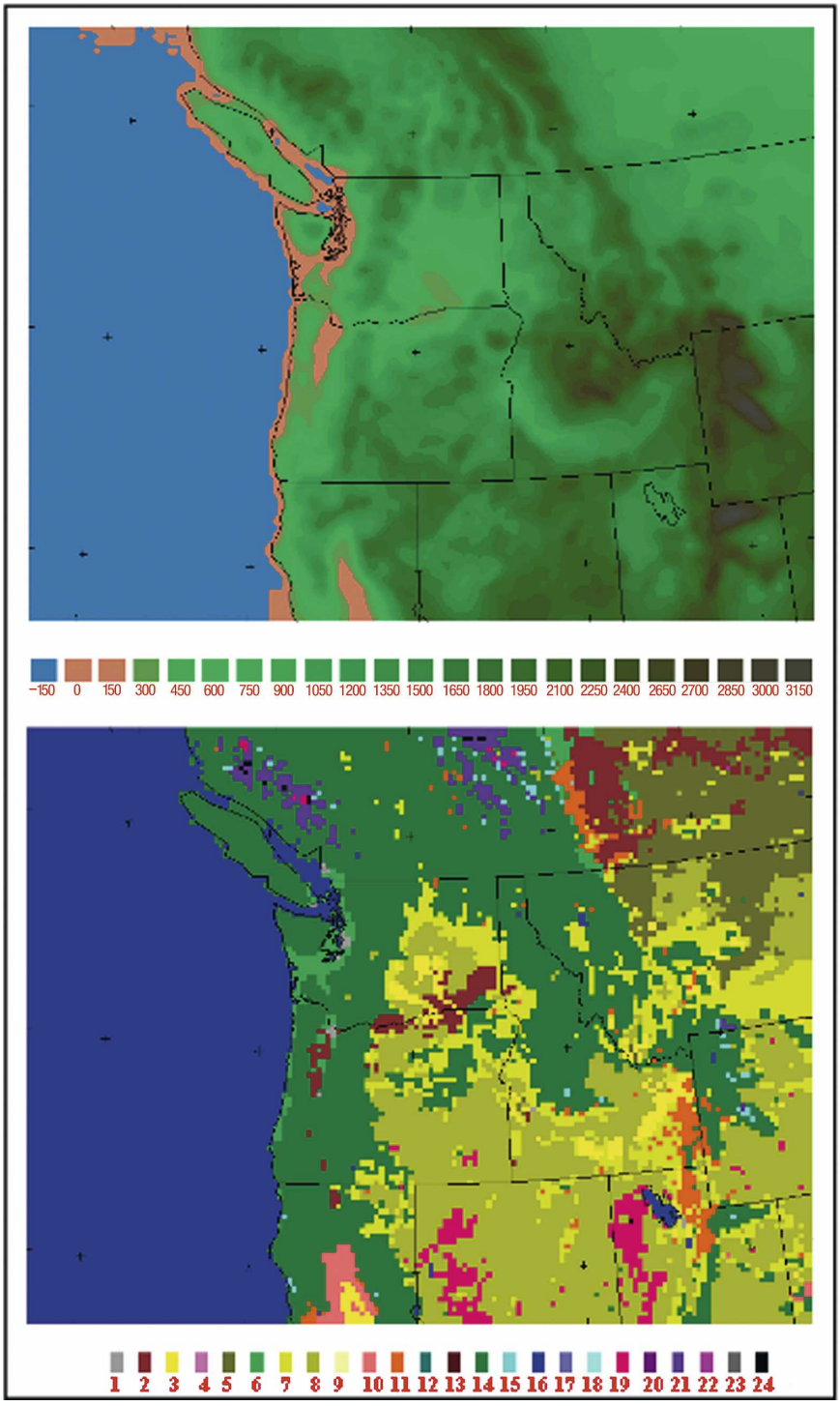

FIG. 2. The 12-km MM5 domain (top) topography and (bottom) land use. See Table 1 for a description of land-use categories.

analyses are not available or where long climatological records or long-term model forecast archives do not exist. Finally, it should be able to deal gracefully with regime changes, when model biases might change abruptly. This paper describes an attempt to create such a systematic bias removal scheme for forecast grids at the surface, and one that is applicable to a wide range of regions and parameters.

TABLE 1. MM5 land-use categories.

Vegetation integer identification	Vegetation description	Albedo (%)		Moisture available (%)		Emissivity (% at 9 mm)		Roughness length (cm)		Thermal inertia (cal cm^{-2} K^{-1} $s^{-1/2}$)	
		Summer	Winter	Summer	Winter	Summer	Winter	Summer	Winter	Summer	Winter
1	Urban	18	18	10	10	88	88	50	50	0.03	0.03
2	Dryland/cropland/pasture	17	23	30	60	92	92	15	5	0.04	0.04
3	Irrigated cropland/pasture	18	23	50	50	92	92	15	5	0.04	0.04
4	Mixed dry–irrigated cropland/pasture	18	23	25	50	92	92	15	5	0.04	0.04
5	Cropland–grassland mosaic	18	23	25	40	92	92	14	5	0.04	0.04
6	Cropland–wood mosaic	18	20	35	60	93	93	20	20	0.04	0.04
7	Grassland	19	23	15	30	92	92	0.12	0.1	0.03	0.04
8	Shrubland	22	25	10	20	88	88	10	10	0.03	0.04
9	Mixed shrubs–grassland	20	24	15	25	90	90	11	10	0.03	0.04
10	Savanna	20	20	15	15	92	92	15	15	0.03	0.03
11	Deciduous broadleaf	16	17	30	60	93	93	50	50	0.04	0.05
12	Deciduous needleleaf	14	15	30	60	94	93	50	50	0.04	0.05
13	Evergreen broadleaf	12	12	50	50	95	95	50	50	0.05	0.05
14	Evergreen needleeaf	12	12	30	60	95	95	50	50	0.04	0.05
15	Mixed forest	13	14	30	60	94	94	50	50	0.04	0.06
16	Water bodies	8	8	100	100	98	98	0.01	0.01	0.06	0.06
17	Herbaceous wetland	14	14	60	75	95	95	20	20	0.06	0.06
18	Wooded wetland	14	14	35	70	95	95	40	40	0.05	0.05
19	Barren/sparse vegetation	25	25	2	50	85	85	10	10	0.02	0.02
20	Herbaceous tundra	15	60	50	90	92	92	10	10	0.05	0.05
21	Wooded tundra	15	50	50	90	93	93	30	30	0.05	0.05
22	Mixed tundra	15	55	50	90	92	92	15	15	0.05	0.05
23	Bare ground tundra	25	70	2	95	85	95	0.1	5	0.02	0.05
24	Snow or ice	55	70	95	95	95	95	5	5	0.5	0.5

2. Data

The bias correction algorithm developed in this research was tested on forecasts made by the MM5, which is run in real time at the University of Washington (UW) (Mass et al. 2003). This modeling system uses 36- and 12-km grid spacing through 72 h, and a nested domain with 4-km grid spacing that is run out to 48 h. Using this system, the 2-m temperature (T2), 2-m dewpoint temperature (TD2), and 12-h precipitation (PCP12) forecasts on a grid were corrected for forecast hours 12, 24, 36, 48, 60, and 72 for model runs initialized at 0000 UTC during the 1-yr period from 1 July 2004 to 30 June 2005. For this work, only grids from the 12-km domain (Fig. 2) were bias corrected. Corresponding surface observations for the period were gathered from the UW NorthwestNet mesoscale network, a collection of observing networks from throughout the U.S. Pacific Northwest. Over 60 networks and approximately 1500 stations are available in NorthwestNet (Mass et al. 2003) for the region encompassed by the 12-km domain. As described in the appendix, the observations were randomly divided for use in either verification or bias estimation.

Extensive quality control (QC) was performed on all observations. Quality control is very important if a heterogeneous data network of varying quality is used, since large observation errors could produce erroneous biases that can be spread to nearby grid points. The QC system applied at the University of Washington includes range checks, step checks (looking for unrealistic spikes and rapid changes), persistence checks (to remove "flat lined" observations), and a spatial check that ensures that observed values are not radically dif-

TABLE 2. Combined land-use categories and their components.

Combined land-use category	Component MM5 land-use categories
1, urban	1
2, cropland	2, 3, 4, 5, 6
3, grassland	7, 8, 9, 10
4, forest	11, 12, 13, 14, 15
5, water	16
6, wetland	17, 18
7, barren tundra	19, 23
8, wooded tundra	20, 21, 22
9, snow–ice	24

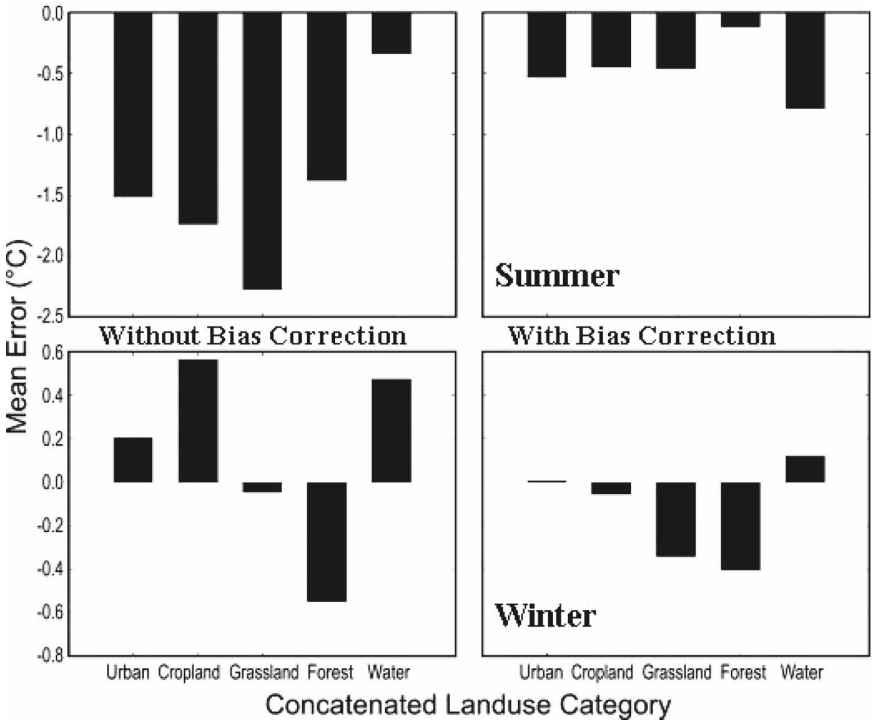

FIG. 3. Biases of T2 over the Pacific Northwest for (top) July–August 2004 and (bottom) December 2004–January 2005 (left) without and (right) with bias correction. The other concatenated categories (wetland, barren tundra, wooded tundra, and snow–ice) were not shown due to a lack of observations.

ferent from those of nearby stations of similar elevation. More information on this quality control scheme can be found online (http://www.atmos.washington.edu/~qcreport/).

3. An observation-based approach to bias removal on a grid

As noted above, an observation-based scheme is used because high-resolution analyses are only available for a small portion of the globe and even when available they often possess significant deficiencies. The gridded BC scheme for temperature and dewpoint described below is based on a few basic ideas.

1) It begins with estimating the forecast error at observing locations. The forecast error at each observing site is calculated by subtracting the observed value from the forecast, which is determined by bilinearly interpolating the four surrounding gridpoint forecast values to the observation location. The standard atmospheric lapse rate ($-6.5°C\ km^{-1}$) is used to account for the discrepancy between the model terrain height and the real station elevation when dealing with T2.
2) The BC scheme only uses observations of similar elevation to that of the model grid point in question and considers nearby observations before scanning at greater distances. As described below, although proximity is used in station selection, distance-related weighting is not applied, reducing the impact of a nearby station that might have an unrepresentative bias.
3) The BC scheme makes use of land-use type, applying only forecast errors from observing sites with similar land-use characteristics to the grid point in question to estimate the bias at the grid point. This approach is based upon the empirical observation that land use has a large influence on the nature of many surface biases; for example, water-covered regions have different biases than land surfaces, and desert regions possess different biases than irrigated farmland or forest. To illustrate this relationship, the 24 land-use categories used in MM5 (Table 1) were combined into nine that possessed similar characteristics (see Table 2). The biases in T2 for these combined land-use categories over the entire Pacific Northwest were calculated for 2 months of summer and winter. The summer results, shown in Fig. 3, indicate substantial differences in warm season

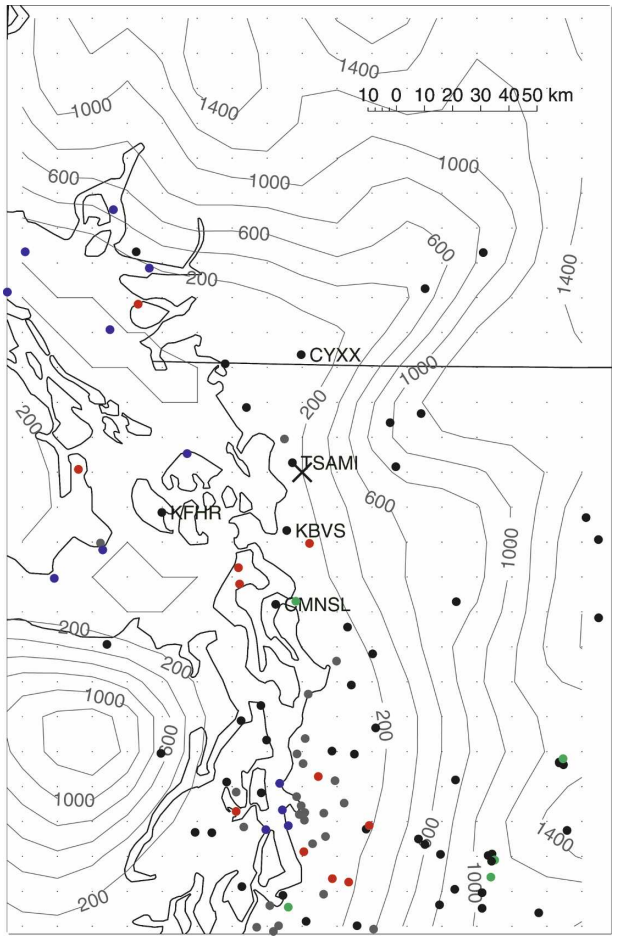

FIG. 4. Stations chosen to bias correct grid point (89, 66; "×" symbol) for T2 for forecast hour 48 of the forecast initialized at 1200 UTC 2 Mar 2005 are indicated by four- or five-letter identifiers. Observation locations are colored according to their combined land-use category, relative terrain height (m) is shown with the gray contour lines, and all model grid points within the region are shown as small black dots.

temperature bias among the various land-use categories, ranging from a small negative bias over water to a large negative bias over grassland. In contrast, during the winter season the signs of the biases vary from moderate positive biases over water, cropland, and urban areas to a moderate negative bias over forest and little bias over grassland. A Student's *t* test analysis revealed that the differences in bias among the categories were statistically significant. Based on these results, the nine combined land-use categories were applied in the bias correction method presented in this paper. The land-use type of each observation location is determined using a 1.33-km resolution land-use grid, while the land-use type on the 12-km model grid is defined by the predominant category within each 12-km grid box.

4) This scheme is designed to mitigate the effects of regime change, which is a major problem for most bias correction methods because they typically use a preceding period of a few days to a several weeks to estimate the bias. Using such preforecast averaging periods can result in very poor bias estimation when a large regime change occurs. Specifically, the nature of the bias can be altered during a regime change and can result in the application of the wrong corrections to the forecasts, thereby degrading the original predictions. The approach applied in this work minimizes the effects of such regime changes in two ways. First, only errors from forecasts of similar parameter value (and hopefully similar regime) are used in estimating the bias at a grid point. Thus, if the forecast of T2 at a grid point is 70°F, only errors from forecasts with T2s that are similar (say, between 65° and 75°F) are used in calculating bias estimates at that point. Additionally,

TABLE 3. The optimized settings for the bias correction method for T2 and TD2.

Setting	T2 value	TD2 value
No. of search dates for finding "similar forecasts"	59	73
No. of similar forecasts	11	10
No. of "similar stations" per grid point	8	9
Tolerance used to define a similar forecast (°C)	6.5	6.5
Max station error (QC parameter)	6.0	12.5
Max station-to-gridpoint distance (km)	864	1008
Max station-to-gridpoint elevation difference (m)	250	480

TABLE 4. Optimized settings for the bias correction of 12-h precipitation (PCP12).

Setting	PCP12 value
No. of search dates for finding similar forecasts	74
No. of similar forecasts	6
No. of similar stations per grid point	4
Similar forecast bin 1 (in.)	0.0000
Similar forecast bin 2 (in.)	0.0001–0.1101
Similar forecast bin 3 (in.)	0.1102–0.4482
Similar forecast bin 4 (in.)	0.4483+
Max station error (QC parameter, in.)	4.5508
Max station-to-gridpoint distance (km)	718
Max station-to-grid-point elevation difference (m)	1080

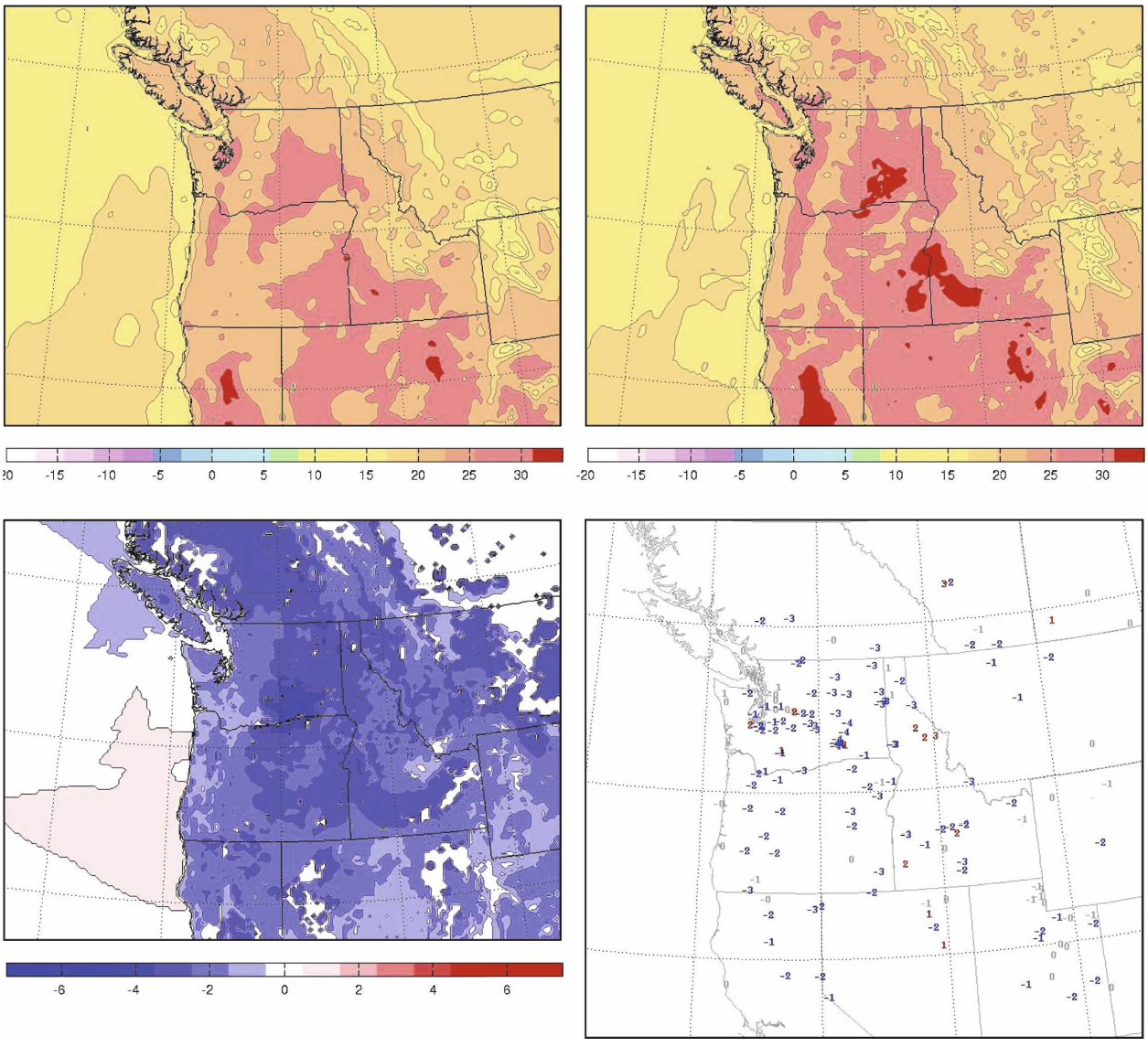

FIG. 5. The effects of bias correction on the 48-h forecast of T2 verifying at 0000 UTC 11 Aug 2004. Values of T2 (°C) (top left) with and (top right) without bias correction. (bottom left) The bias correction produced by the algorithm (°C) and (bottom right) the magnitude of the bias correction at observing sites, with blue (red) numbers indicating improvement (degradation) of the forecasts. Stations degraded or improved by less than 2°C are shown in gray.

only the most recent errors are used for estimating bias at a station as long as a sufficient number are available.

5) The biases are calculated for each forecast hour, since biases vary diurnally and the character of the bias can change with the forecast projection even for the same time of day.
6) The scheme estimates the bias at a grid point by using a simple average of the interpolated forecast errors from a minimum number of different sites that meet the criteria noted above and are within a set radius for each grid point. Figure 4 shows an example of the stations in the vicinity of a grid point (89, 66), including the five nearest stations that were considered similar to the grid point in question by the land-use, elevation, and forecast-value criteria. Simple averaging of several stations without distance weighting is used to avoid spreading the representational error of a single station to the surrounding grid points. By averaging different observing locations in the calculation of bias at each grid point, the influence of problematic or unrepresentative observing sites is minimized. Such an approach determines the underlying sys-

tematic bias common to stations of similar land-use, elevation, and forecast value. Furthermore, as an additional quality control step, stations with extremely large (defined later) forecast errors are not used.

Several parameter values that describe the various distances and levels of "similarity" need to be set to implement the BC method. These values may depend on the particular observation network density and configuration as well as the model forecast domain and resolution. In the initial development of the method, an empirical approach was taken in which parameter values were adjusted subjectively within physically reasonable bounds. However, in an attempt to improve upon the empirically determined settings, an objective optimization was undertaken. The optimization process used the Evol software module (Billam 2006), which employs a random-search strategy for minimizing large variable functions. The Evol routine minimizes a function by a single metric, which was chosen to be the domain-averaged mean absolute error (MAE). Minimization of this metric was found to be more effective than minimization of the domain-averaged mean error (ME), since optimizing domain-averaged ME sometimes resulted in a degradation of the MAE due to the existence of regional pockets of bias of opposite sign. Experimental use of domain-averaged MAE as the optimization metric was found to minimize both MAE and ME. A more detailed discussion of the optimization process is given in the appendix. The optimized settings for the seven parameters of the BC method for T2 and TD2 are shown in Table 3. Parallel testing of the empirical and objectively optimized settings produced similar results, with the optimized settings showing a small, but consistent, improvement. Thus, in this paper, only results based on the objectively optimized settings are presented.

Using the settings shown in Table 3, the algorithm works in the following way for temperature (T2) and dewpoint temperature (TD2). For each hour, it first interpolates the model forecast for these parameters to all available observation sites. At each site it then searches back in time to find at least 11 (10 for TD2) forecasts at that site that are similar to the current forecast. For temperature and dewpoint, "similar" means within 6.5°C. Looking for such similar forecasts, the algorithm will search back a maximum of approximately 2 months (59 and 73 days, respectively, for T2 and TD2). Such a time limitation ensures that the algorithm will not make use of data from a substantially different season. Errors from these similar forecasts are gathered and averaged to estimate the bias at each station. Next, the model uses the biases at the observation locations to estimate values on the grid. For each model grid point, the algorithm searches for at least eight stations (all of which have similar forecasts as noted above) with similar elevation and land use. If eight "similar stations" are found, it averages the interpolated forecast errors from those stations in order to estimate an appropriate bias correction to be applied at that grid point. To reject stations with potential problems, those with very large errors (6° and 12.5°C, respectively, for T2 and TD2) are not used. The algorithm is able to look fairly far afield, searching to a maximum distance of 864 (1008) km for T2 (TD2). Elevations for similar stations have to be within 250 m for T2 and 480 m for TD2.

TABLE 5. Verification statistics for the uncorrected and bias-corrected forecasts for T2 (°C) during July 2004–June 2005 at forecast hours 12, 24, 36, and 48 over the 12-km domain.

Forecast hour	Settings	ME (°C)	MAE (°C)	Fraction of stations improved (≥0.5°C)	Fraction of stations degraded (≥0.5°C)
F12	No BC	0.20	2.26	N/A	N/A
	BC	0.12	2.19	0.27	0.21
F24	No BC	−0.34	2.13	N/A	N/A
	BC	−0.10	1.98	0.30	0.19
F36	No BC	−0.30	2.36	N/A	N/A
	BC	−0.06	2.30	0.26	0.21
F48	No BC	−0.65	2.38	N/A	N/A
	BC	−0.16	2.19	0.32	0.20

For precipitation, some alterations in the algorithm for T2 and TD2 are made. One adjustment is to eliminate the use of land use when matching observing locations (and their corresponding errors) to grid points, since land use is not expected to strongly influence the forecast bias of precipitation. Thus, for a given grid point, the search algorithm only looks for observing stations within a given radius and elevation band. Another adjustment to the algorithm is to consider precipitation forecasts to be "similar" when they fall into one of four bins: no precipitation, greater than 0.00 in. but less than some value $p1$, greater than $p1$ but less than or equal to some value $p2$, and greater than $p2$. The value of $p1$ and $p2$ were initially determined through empirical experimentation, with values of 0.01 and 0.10 in. showing reasonable results. Similar to T2 and TD2, an optimization slightly alters the algorithm settings. Finally, the estimate of bias at a grid point is calculated using the *median* of the errors from its similar stations rather than the mean of those errors as was done for T2 and TD2. [This was suggested by T. Gneiting and C. Marzban (2007, personal communi-

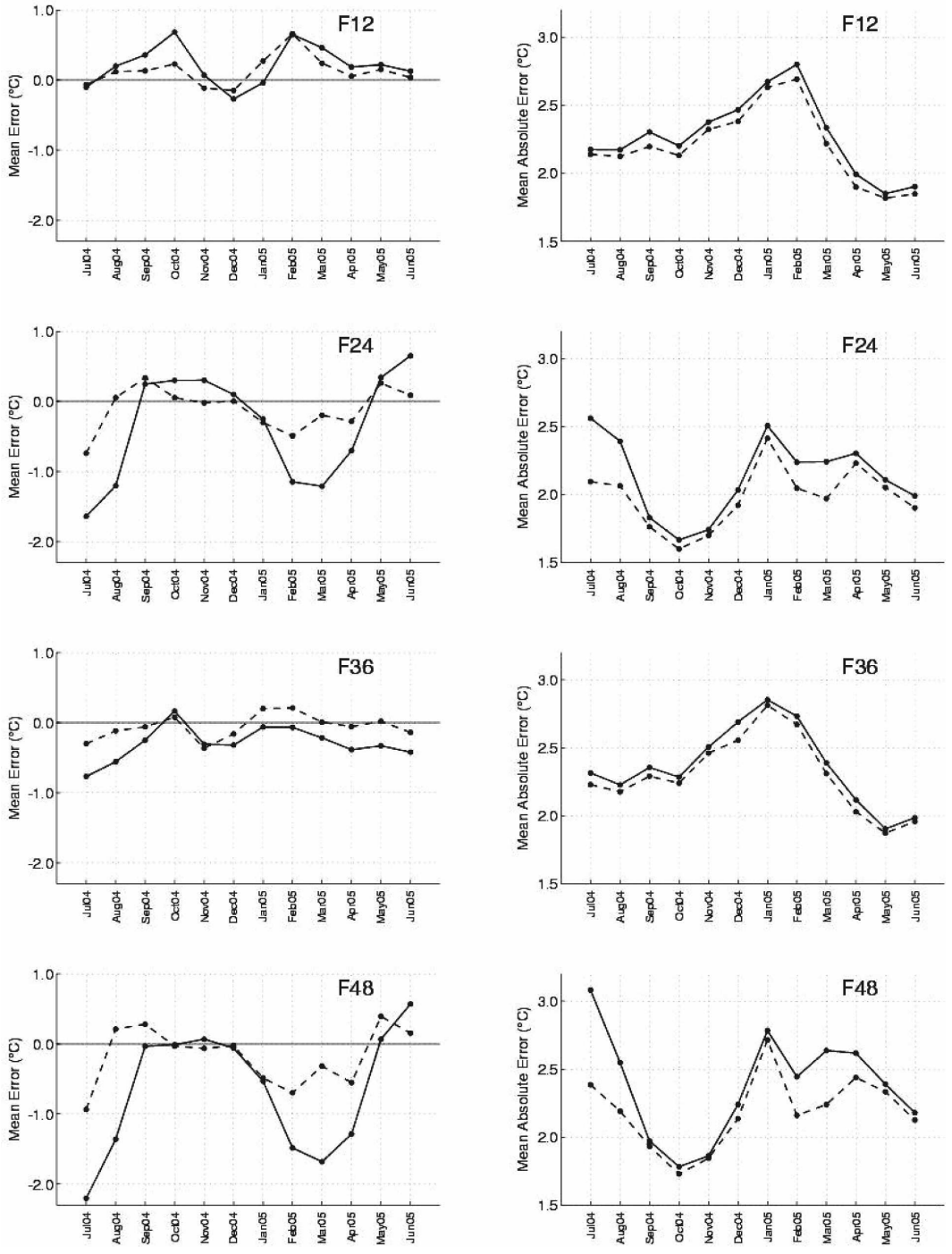

FIG. 6. Domain-averaged (left) mean error and (right) mean absolute error by month for T2 corrected (dashed) and uncorrected (solid) forecasts for 12, 24, 36, and 48 h.

cation) and showed better results than using means.] The optimized settings used for precipitation are shown in Table 4.

An example of the application of the bias correction scheme is found in Fig. 5, which shows the effects of such a correction on the 48-h forecast of T2 that verified at 0000 UTC on 11 August 2004. A major issue at that time was a significant warm bias for maximum temperature, a problem that was particularly large east of the Cascade Mountains. The bias correction scheme

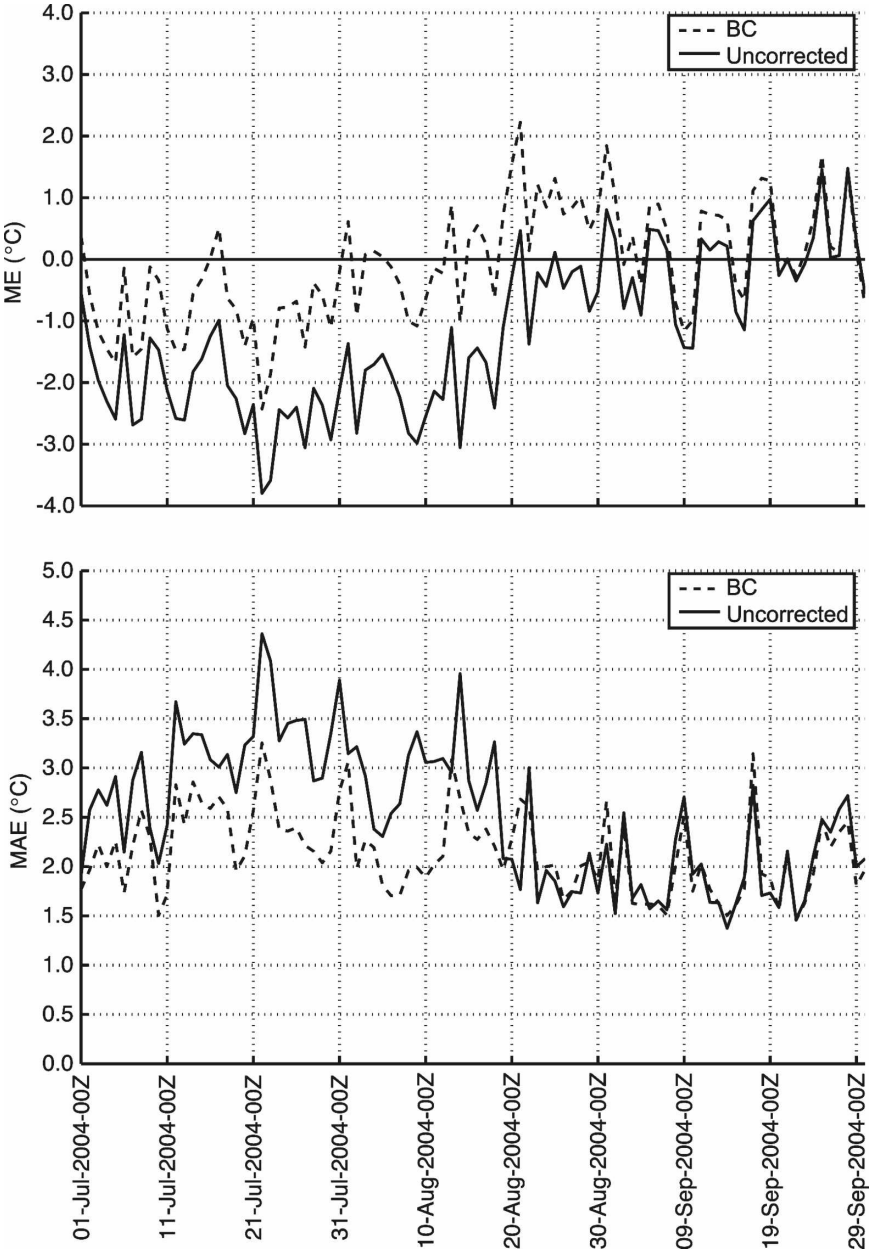

FIG. 7. Daily (top) mean error and (bottom) mean absolute error of T2 for 48-h forecast for the corrected and uncorrected forecasts during July–September 2004.

cooled the surface temperatures over nearly the entire domain, with the largest alterations (~4°C) in the Columbia Basin of eastern Washington (Figs. 5a–c). The bottom-right panel in Fig. 5 shows the difference between the bias-corrected forecast and the uncorrected forecast in degrees Celsius at the verification observation sites, with the color blue indicating improvements and red indicating degradations. On that day, the impact of the scheme was highly positive with forecast errors at nearly all sites being reduced by the bias correction.

4. Results

a. 2-m temperature (T2)

Domain-averaged verification statistics for the corrected and uncorrected 12–48-h forecasts of T2 over the period July 2004–June 2005 are shown in Table 5. A total of 57 514 model–observation data pairs were used in calculating each of these statistics, with the verification data being independent from those used to perform the bias correction (see the appendix for an ex-

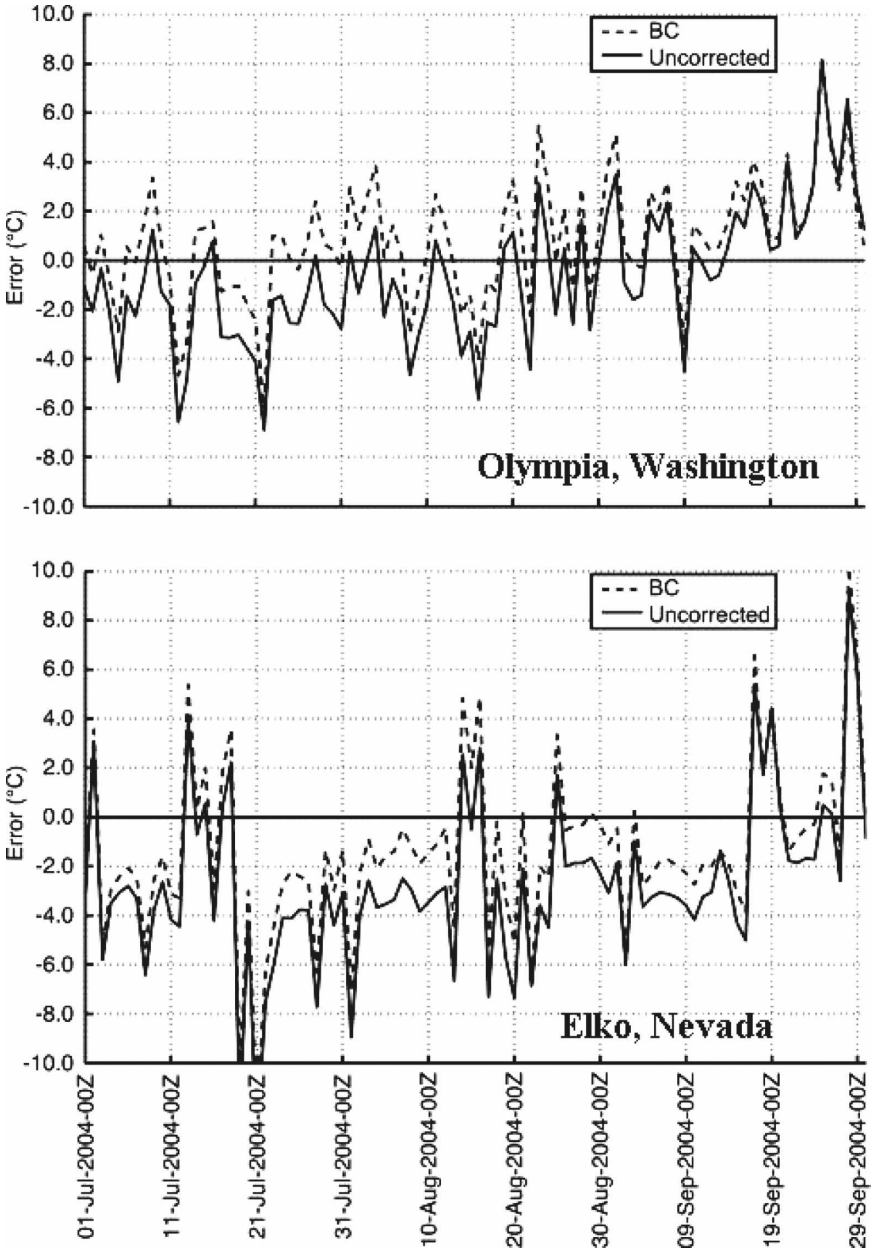

FIG. 8. The T2 mean error for corrected and uncorrected 48-h forecasts for (top) Olympia and (bottom) Elko for 1 Jul–30 Sep 2004.

planation). For all hours, the bias correction improved both the mean error (ME) and mean absolute error (MAE). The largest improvements were for the 24–48-h forecasts, which verified at 0000 UTC [1600 Pacific standard time (PST)], near the time of maximum temperature. For example, at 48 h the mean error for the uncorrected forecasts was −0.65°C, with the bias correction reducing the bias by about 0.5°C and the MAE by 0.19°C. It is worth noting that the algorithm degraded forecasts at some stations, but in all cases more stations were improved than degraded. A comparison of the bias histogram before (Fig. 3, left side) and after correction (right side) reveals a large drop in bias in four of the five combined land-use categories, with the largest improvements during the summer months.

Figure 6 shows the domain-averaged ME and MAE for T2 by month for the uncorrected and corrected forecasts for hours 12, 24, 36, and 48 for July 2004–June 2005. There is substantial diurnal variability in the amount of bias, with the largest errors occurring at the

time of maximum temperature (hours 24 and 48). Not surprisingly, the bias correction scheme makes the largest change (improvement) during the hours of largest bias. Even at times of minimum temperature and smaller bias, the bias correction scheme substantially reduces both the bias and mean absolute error. In addition to diurnal differences in bias, there are periods with much larger bias, such as July–August 2004 and February–April 2005. During such periods the bias correction makes large improvements to the forecasts during the daytime, reducing the mean error by roughly 1°C and the mean absolute error by 0.5°C.

Daily domain-averaged results of the bias correction method for T2 for a shorter period (July–September 2004) and one forecast hour (hour 48) are shown in Fig. 7. The bias correction algorithm decreased the ME by about 2°C from late July 2004 through the first half of August 2004, when a large cold bias was present. MAE was also substantially reduced during this period. The bias in the uncorrected forecast decreased to near zero around 20 August 2004, and for a few days a small warm bias correction was made. Subsequently, the uncorrected forecast bias remained near zero, and the bias correction essentially turned off.

The T2 mean errors for the uncorrected and bias-corrected forecasts for Olympia, Washington, and Elko, Nevada, for 1 July–30 September 2004, at forecast hour 48, are shown in Fig. 8. Compared to daily time series of domain-averaged ME (Fig. 7), the individual-site ME results show increased temporal variance in the bias. On some occasions, the bias correction algorithm degrades the forecast at a single station following major changes in bias. At Olympia, a warm (positive) correction was made during most of the period from 1 July through 30 September 2004. This led to an improved forecast on some days and a degraded forecast on others. Over the period shown, the uncorrected forecast ME was −0.62°C, while the bias-corrected forecast ME was 0.86°C. In short, for a forecast with only minimal bias and large variability, the scheme produced a slight degradation. At Elko, the forecast bias was much larger and consistent, and thus the bias correction greatly improved the forecast, with few days of degradation. Over this period, the uncorrected forecast ME at Elko was −2.68°C, while with bias correction it dropped to −1.36°C.

b. 2-m dewpoint temperature (TD2)

Verification statistics for the corrected and uncorrected forecasts for TD2 for the 12–48-h forecasts for July 2004–June 2005 are shown in Table 6. A total of 32 665 model–observation data pairs were used in calculating these statistics, with verification data being independent from those used to perform the bias correction (see the appendix). As for T2, the bias correction for TD2 improves the mean and mean absolute errors at all hours. However, for TD2, the uncorrected errors are much larger and the corrections are substantially greater and thus highly beneficial, with improvements in mean error exceeding 1°C at all hours. For TD2 the ratio of improved to degraded forecasts is much larger than for T2, with roughly half of all stations being improved by more than 0.5°C.

TABLE 6. Verification statistics for the uncorrected and bias-corrected forecasts for TD2 (°C) during July 2004–June 2005 for the 12-, 24-, 36-, and 48-h forecasts.

Forecast hour	Settings	ME (°C)	MAE (°C)	Fraction of stations improved (≥0.5°C)	Fraction of stations degraded (≥0.5°C)
F12	No BC	2.26	2.80	N/A	N/A
	BC	1.25	2.33	0.46	0.20
F24	No BC	2.55	3.23	N/A	N/A
	BC	1.36	2.65	0.48	0.21
F36	No BC	1.77	2.64	N/A	N/A
	BC	1.11	2.40	0.37	0.22
F48	No BC	2.35	3.25	N/A	N/A
	BC	1.32	2.79	0.46	0.23

Figure 9 shows the domain-average ME and MAE for TD2 by month for the uncorrected and bias-corrected 12–48-h forecasts for July 2004–June 2005. This figure shows that the improvement of the bias-corrected forecast is substantially greater for TD2 than T2 (Fig. 6). Uncorrected biases are large and positive, and generally decrease over the period at all forecast projections. The bias correction scheme provides substantial improvement (1.5°–3.0°C) over most of the period, with the only exception being at the end when the uncorrected bias had declined to under 2°C. The largest TD2 biases were during July 2004 and early spring 2005, with biases being greatest during the cool portion of the day (1200 UTC, 0400 PST). Results of the bias correction method for TD2 for July–September 2004 are shown in Fig. 10 for the 48-h forecasts and over the entire domain. The bias corrections are roughly 2.0°–2.5°C from mid-July through mid-August 2004, with nearly all days showing substantial improvement. Similar improvements are noted in the MAE. As found for temperature, the greatest improvements are made when the bias is largest (in this case, the first half of the period).

The spatial variations in the impacts of the bias correction scheme for TD2 for a sample forecast are shown in Fig. 11, which presents corrected forecast error mi-

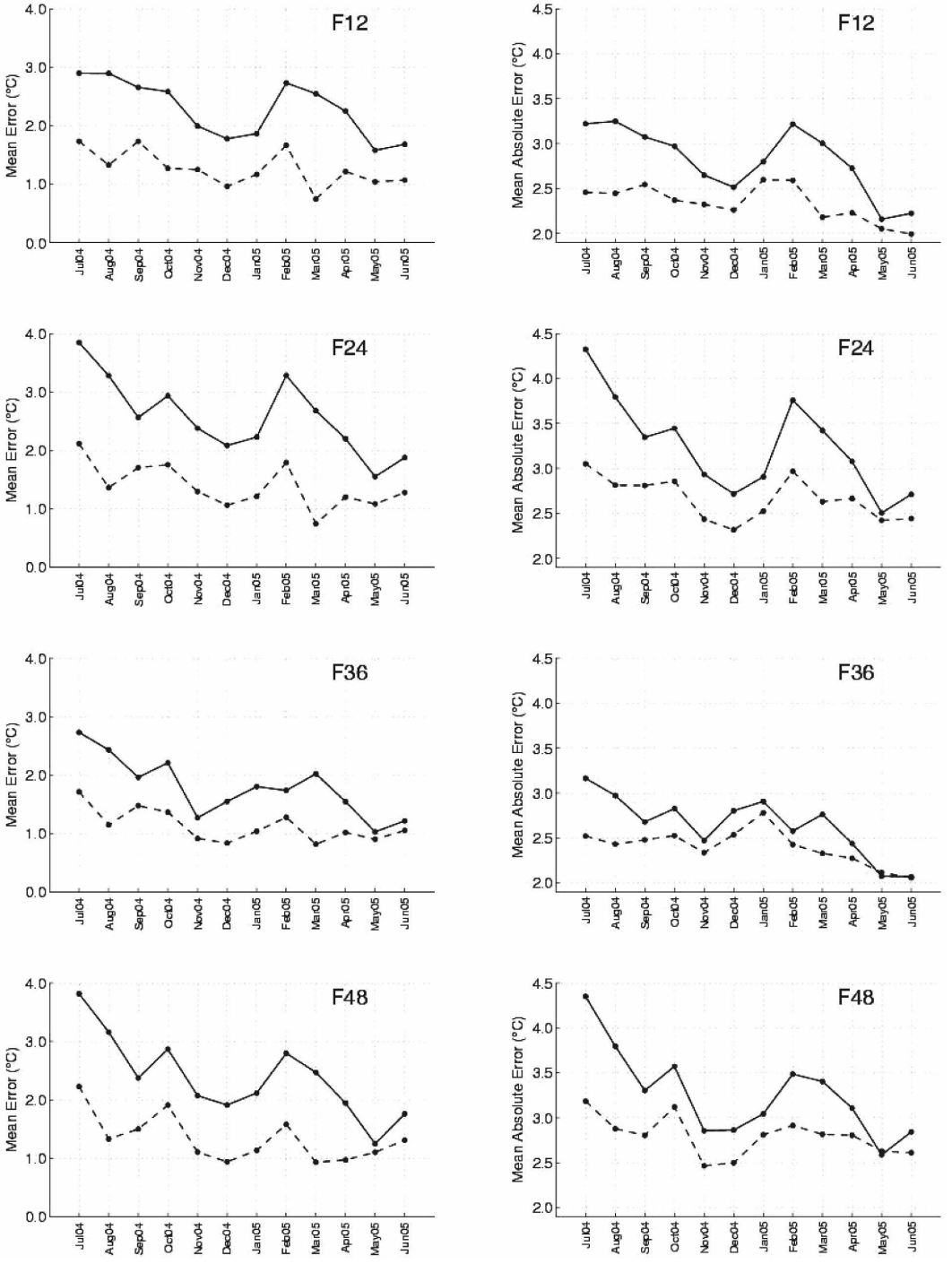

FIG. 9. Same as in Fig. 6 but for TD2. Shown are the results for 12, 24, 36, and 48 h.

nus the uncorrected forecast error at observing locations for the 48-h forecast verifying at 0000 UTC on 9 August 2004. Forecasts at observation locations that were "helped" by bias correction are shown by blue negative numbers, and those that were "hurt" are indicated by positive red numbers. For this forecast, the impact of the bias correction was overwhelmingly positive, with typical improvements of ~2°C.

An illustration of the influence of the bias-corrected scheme on TD2 at two locations over the summer of 2004 is provided in Fig. 12 (again for the 48-h forecast). One location had relatively little bias (Olympia), while

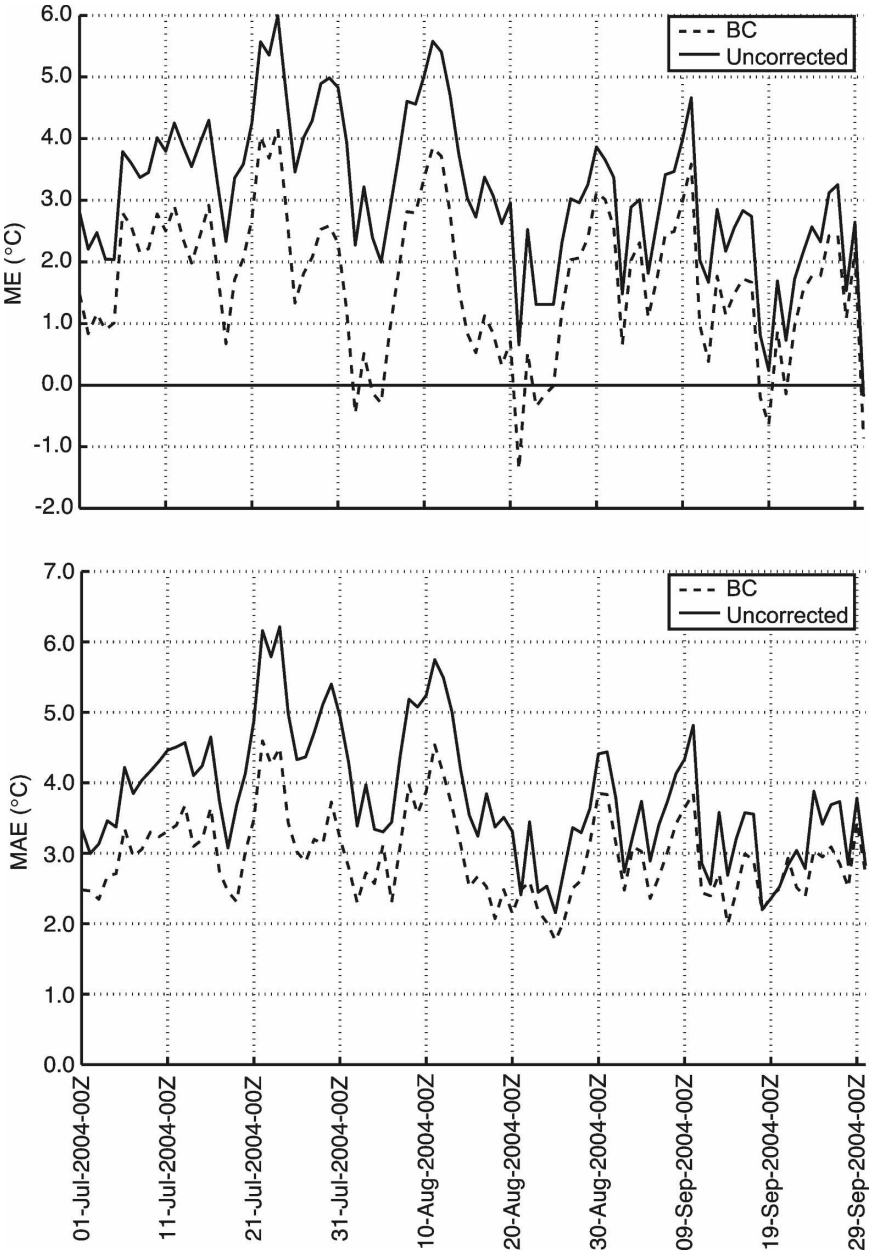

FIG. 10. Same as in Fig. 7 but for TD2.

the other (Elko) possessed an extraordinarily large bias. At Olympia, a modest 2°–3°C bias in the 48-h forecast during the first month was reduced by the scheme, while little was done during the second half of the summer when the dewpoint bias was small. The MEs for Olympia over that period were 1.26° and 0.09°C for the uncorrected and corrected 48-h forecasts, respectively. In contrast, at Elko the bias was far larger, averaging 5°–10°C, with transient peaks exceeding 15°C. At this location the bias correction scheme made large improvements of ~4°C, with the average ME being 7.86° and 3.39°C for the uncorrected and corrected forecasts, respectively.

c. 12-h precipitation amounts (PCP12)

As noted earlier, the algorithm for bias-correcting precipitation differs from that used for temperature and dewpoint, with land use not being considered. Also different was the binning of precipitation to deal with the need for sufficient information for bias correction for a parameter that occurs less frequently than continuous variables such as temperature or dewpoint. Table 7 pre-

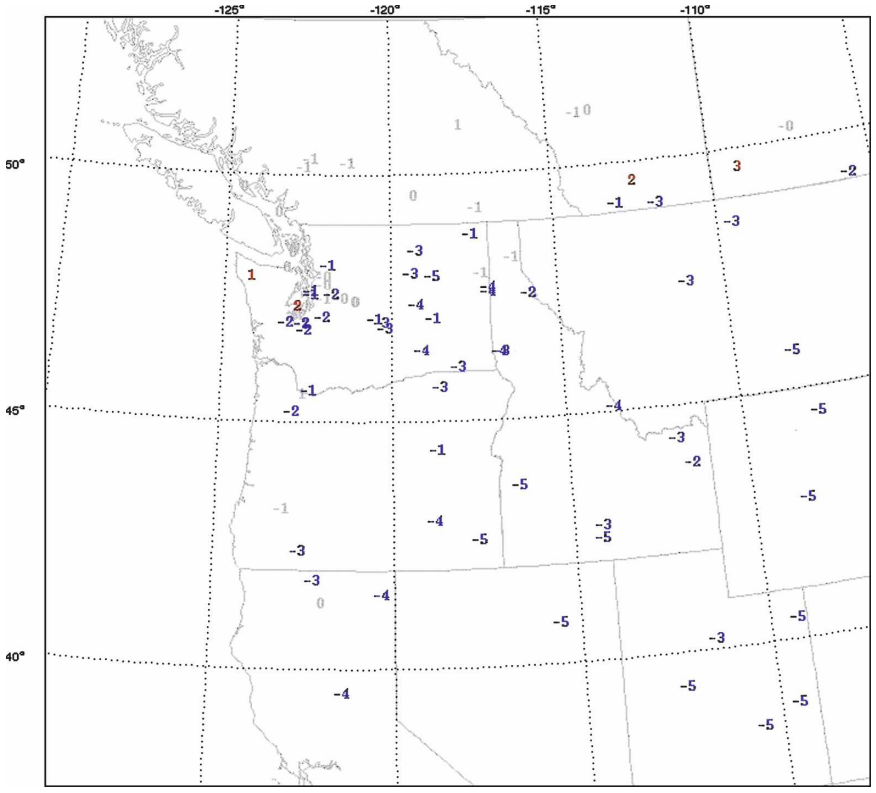

FIG. 11. Same as in Fig. 5, bottom-right panel, but for TD2.

sents a summary of the results for PCP12 for July 2004–June 2005 for hours 12, 24, 36, and 48 over the entire 12-km domain. Mean error is reduced for three of the four forecast hours, while mean absolute error is improved at all times, with the greatest improvements for the longest ranges. The ratio of improved to degraded forecasts (by at least 0.05 in.) due to the bias correction scheme increases from roughly 2.3 at hour 12 to approximately 3 at the longer projections.

Figure 13 shows monthly average mean and mean absolute errors for the 12-, 24-, 36-, and 48-h PCP12 forecasts for July 2004–June 2005 over the entire 12-km domain. Mean errors were generally improved from June 2004 through February 2005, with dramatic enhancement occurring for hours 36 and 48, when biases were highest. In contrast, average precipitation biases were occasionally worsened during the later periods when bias was low. Mean absolute errors are reduced for all months and for all forecast projections, with particular improvements at 36 and 48 h.

An example of the impact of the bias correction scheme for a shorter period of high precipitation (November–December 2004) is shown in a daily plot, which presents average values over the entire domain (Fig. 14). The mean error plot (Fig. 14, top panel) indicates persistent overprediction by the model. The bias correction scheme greatly reduces the excessive precipitation but occasionally goes too far, particularly during transient declines in precipitation. Mean absolute error (Fig. 14, bottom panel) shows general reductions in error for the bias-corrected values over nearly the entire period, even during the days of excessive compensation for the model overprediction.

Figure 15 shows the daily 12-h precipitation error for hour 48 at a single station, Seattle–Tacoma International Airport, for January–March 2005. The bias correction scheme generally improves the forecast, primarily by reducing overprediction, but occasionally degrades the precipitation prediction during spikes of model underprediction.

5. Discussion and summary

This paper has reviewed a new approach for reducing systematic bias in the forecasts of 2-m temperature, dewpoint temperature, and 12-h precipitation; one that is applicable to other parameters as well. This scheme is designed to be robust and flexible, adaptable to varying observation densities, and able to handle regime changes without large negative effects. An underlying

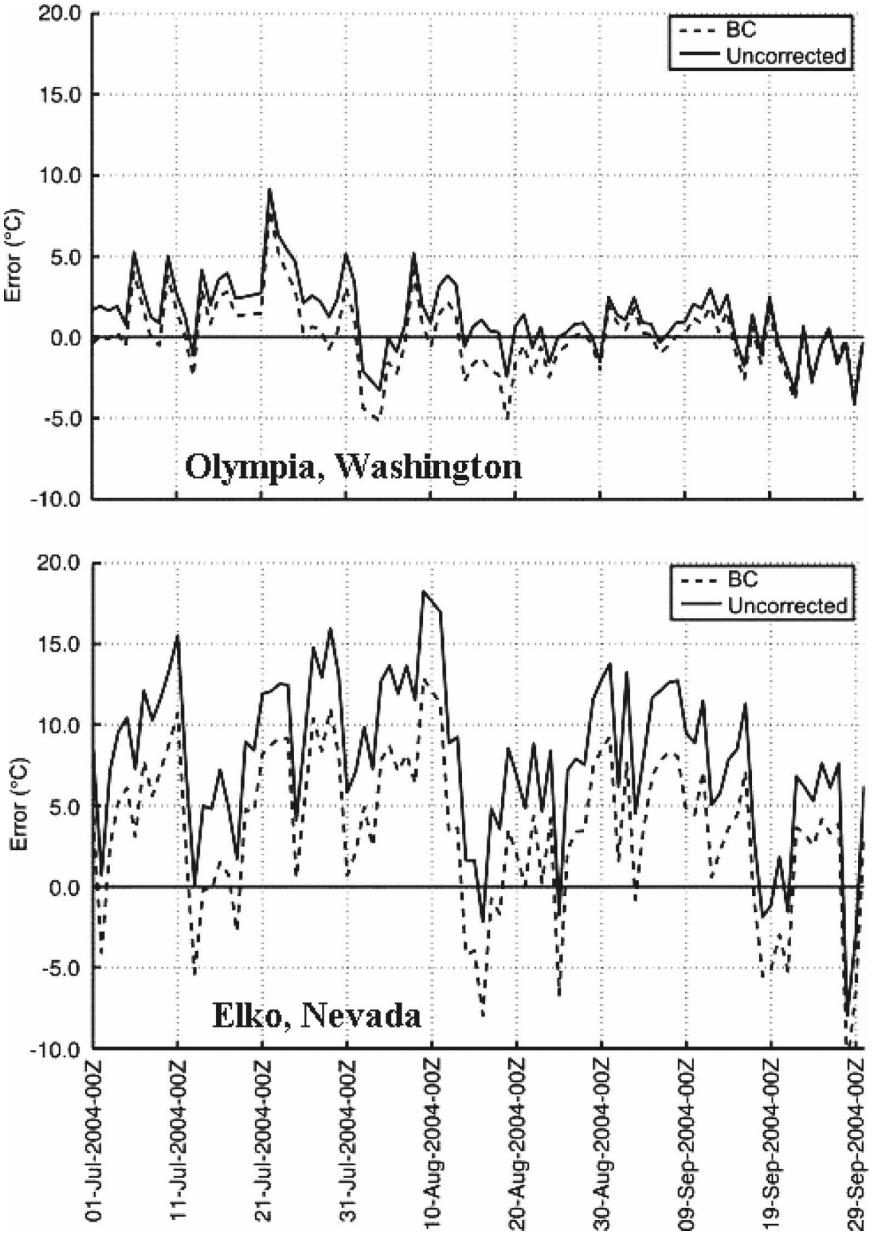

FIG. 12. Same as in Fig. 8 but for TD2. Mean errors are shown for (top) Olympia and (bottom) Elko.

rationale of the algorithm is that biases in surface variables such as temperature and dewpoint are highly related to surface land use and elevation. Furthermore, since changes in weather regimes can alter the systematic biases, bias correction should only make use of information for recent periods of similar parameter values. To lessen the impact of unrepresentative stations, an average of the forecast errors at several stations is used, rather than employing proximity-based weighting and allowing an individual station to heavily influence neighboring grid points. The algorithm uses stations close in time and space if their observations and site characteristics meet the above requirements.

Based on over a year's worth of real-time testing, it appears that this algorithm is highly promising, substantially reducing bias during periods of large and sustained bias, while doing little when bias is small. Changes in weather regime have a negative impact for only a day or two and far more stations are improved than degraded by this approach. Both the bias-corrected and the original MM5 forecasts have been distributed operationally to the Seattle office of the Na-

TABLE 7. Verification statistics for the uncorrected and bias-corrected forecasts for PCP12 (in) during July 2004–June 2005 12-, 24-, 36-, and 48-h forecast hours over the 12-km domain.

Forecast hour	Settings	ME (in.)	MAE (in.)	Fraction of stations improved (≥0.05 in.)	Fraction of stations degraded (≥0.05 in.)
F12	No BC	0.003	0.040	N/A	N/A
	BC	−0.011	0.035	0.07	0.03
F24	No BC	0.013	0.049	N/A	N/A
	BC	−0.011	0.040	0.11	0.04
F36	No BC	0.018	0.053	N/A	N/A
	BC	−0.010	0.040	0.132	0.042
F48	No BC	0.018	0.060	N/A	N/A
	BC	−0.013	0.046	0.134	0.047

tional Weather Service for over a year as an aid for initializing its gridded weather preparation system (the Interactive Forecast Preparation System, IFPS).

The approach described above is essentially univariate and single level. Thus, inconsistencies can result with other parameters (such as temperature being adjusted below the dewpoint) or with values at other levels (since the adjustment at only one level can produce an unphysical lapse rate). It would not be difficult to expand the algorithm to deal with such potential inconsistencies, but it is not clear that this is a serious problem since many applications are only dependent on surface conditions. The bias correction values produced by this algorithm have value beyond their use in weather prediction and other applications. Specifically, they show the nature and characteristics of model bias, and direct researchers and developers to deficiencies in model and surface physics.

It is clear that real-time bias correction on a grid deserves more attention and represents "low-hanging fruit" in the quest to improve surface weather prediction. Grid-based bias correction is a necessity for producing reliable and sharp probabilistic information from ensemble members with significant systematic biases, and bias removal will be an important component of future ensemble-based data assimilation schemes, such as ensemble Kalman filters (EnKFs). The relatively simple grid-based bias correction scheme presented above is only a beginning and its improvement and extension to other variables will be completed over the course of the next year.

Acknowledgments. This research has been support by grants from the NOAA CSTAR program and the Department of Defense (MURI and JEFS programs). Substantial assistance has been provided by David Ovens and Mark Albright, and valuable advice was provided by Professors Tilmann Gneiting and Adrian Raftery of the UW Department of Statistics.

APPENDIX

Optimization of the Bias Correction Settings

In an attempt to improve upon the results of the experimentally determined settings for the BC method, an objective optimization routine (Evol) was employed, using MAE as the metric to minimize. To achieve independent evaluation of each iteration during optimization, as well as evaluation of the final optimized settings, the observations were randomly divided into three groups: one for bias estimation during optimization (50% of all observations), one for metric calculation (verification) during the optimization (25% of all observations), and a final set for independently verifying the final, optimized settings (25% of all observations). A map showing the three groups of observations can be seen in Fig. A1.

The optimization process proceeded as follows:

1) Using the experimentally determined (subjective) settings as a first guess, the bias correction is performed on the model grid using a 50% subset of the observations for each day over a period in question (e.g., 1 month) for a given forecast hour.
2) The resulting bias-corrected grids are then verified using the second, 25% set of observations, producing a metric (domain-averaged MAE), which is returned to the Evol routine along with the settings that produced it.
3) Using its random search strategy, the Evol routine determines a new group of settings to test and the process is repeated until the domain-averaged MAE is minimized.
4) Convergence was assumed when the variance of the domain-averaged MAE over the previous 30 iterations was less than 0.5% of the variance of the domain-averaged MAE over all prior iterations. The settings at convergence were the final, "optimized" settings for that period and forecast hour.
5) Using the final optimized settings, the grids were bias corrected for each day in the given period and the results were verified with the third, 25% set of independent observations. The final verification allowed for a fair comparison of the performance of the optimized settings with other baseline settings. Figure A2 shows the MAE metric for each iteration during the optimization of July 2004 T2 at forecast hour 24.

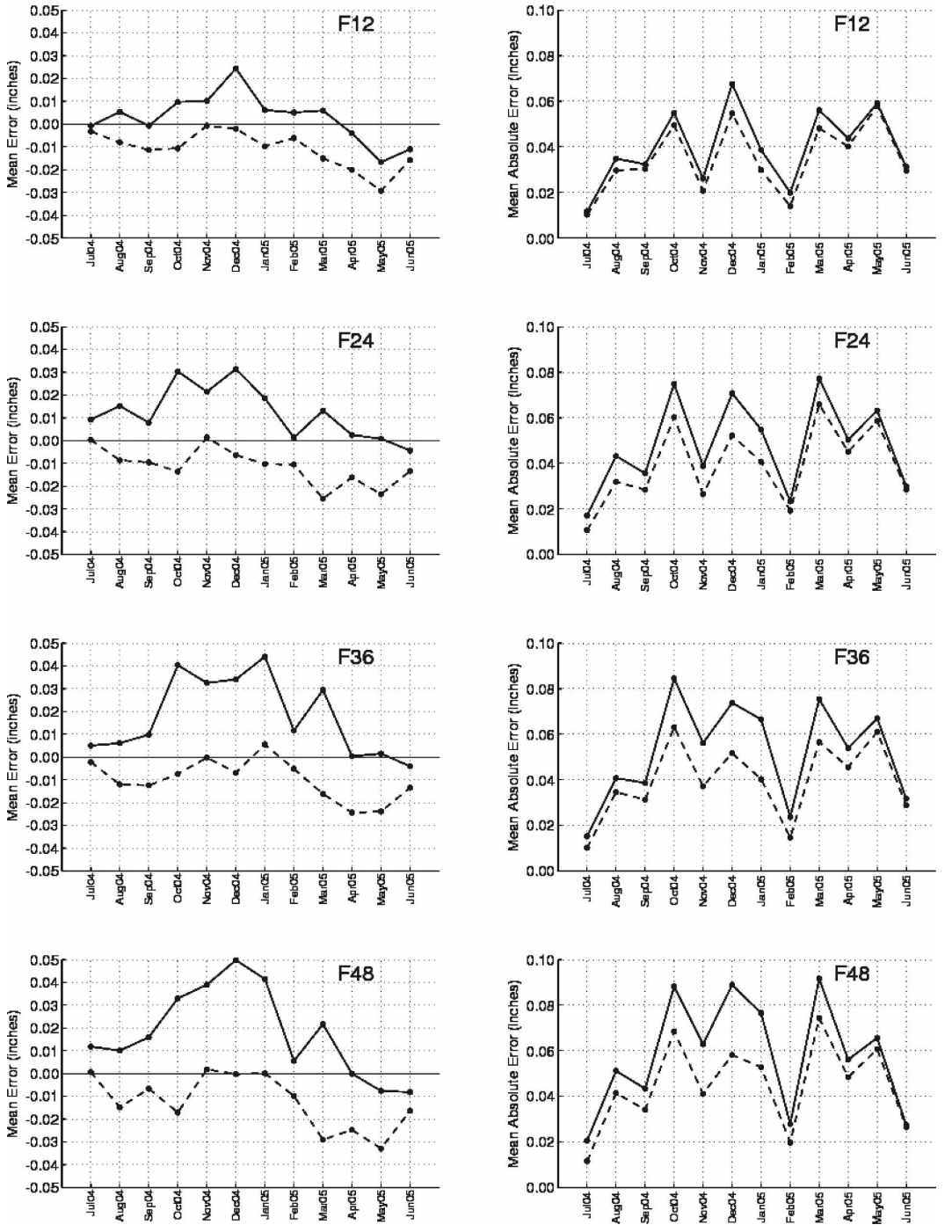

FIG. 13. (left) Mean error and (right) mean absolute error by month for PCP12 corrected (dashed) and uncorrected (solid) 12-, 24-, 36-, and 48-h forecasts.

Initially for T2 and TD2, optimizations were performed separately for each month of the July 2004–June 2005 period for forecast hours 12, 24, 36, 48, 60, and 72, totaling 72 monthly optimizations for each variable. For PCP12, optimizations were only performed for each month of the July 2004–June 2005 period for forecast hours 36 and 48. Verification of these optimized settings was then compared to that of the experimentally determined settings. Monthly optimized results were superior to the experimentally determined settings in terms of MAE for 62 of the 72 months of the July 2004–June 2005 period for T2 and TD2.

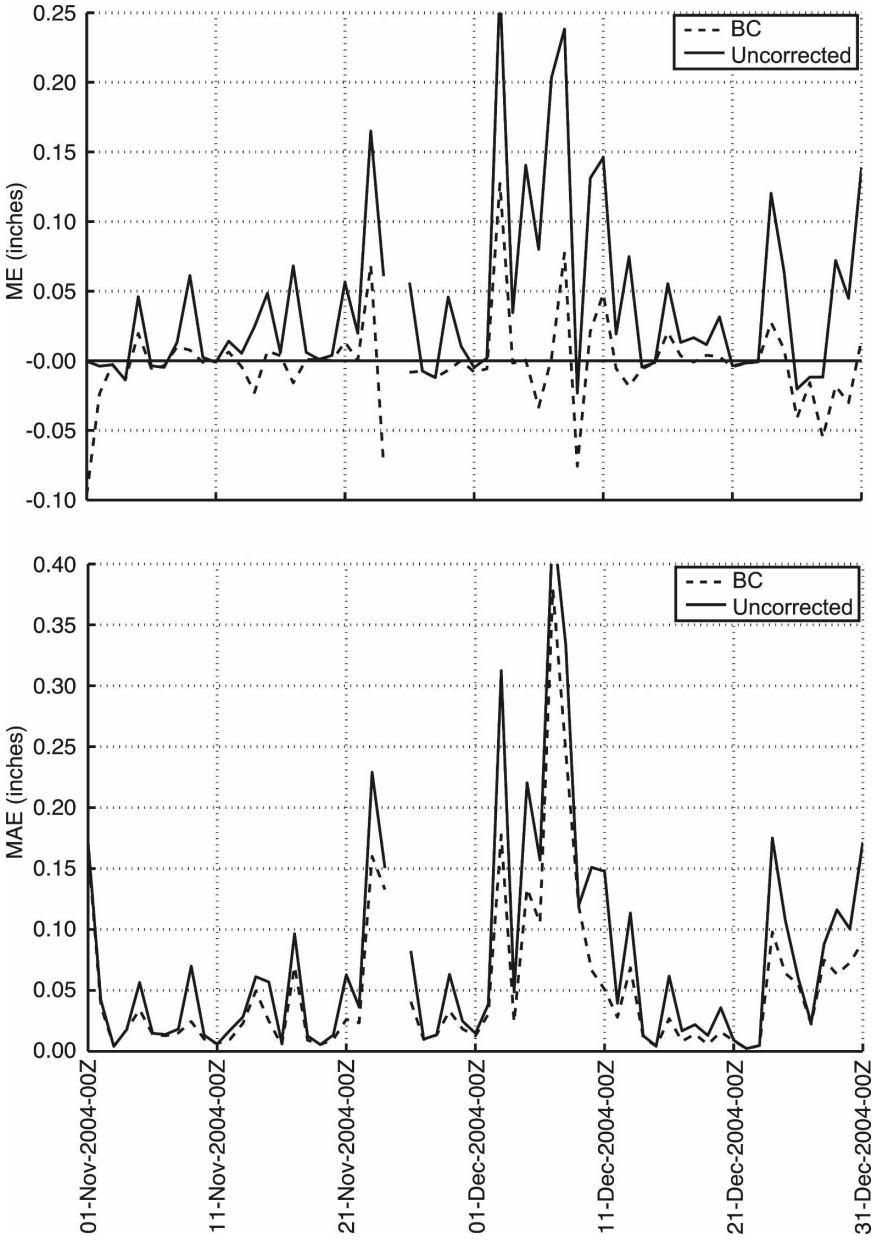

FIG. 14. (top) Mean error and (bottom) mean absolute error of PCP12 for the 48-h forecasts corrected and uncorrected during November–December 2004.

Optimizations were also run for the entire 12-month (July 2004–June 2005) period for T2 and TD2 for forecast hours 12, 24, 36, 48, 60, and 72. As with the monthly optimizations, the verification of the optimized settings was then compared to the verification of the experimentally determined settings. The optimized annual results were superior for all forecast hours tested. Annual optimizations were not run for PCP12

The optimized settings varied more over the 72 monthly optimizations than over the 6 annual optimizations. Figure A3 shows the optimized setting for the maximum distance between observation and grid point for each monthly optimization and each annual optimization verifying for forecast hours 24, 48, and 72 for T2. The monthly optimized maximum distance ranged from 288 to 1008 km, while the annual optimized maximum distance ranged from 660 to 816 km. The variability in the optimized value for this setting was similar to the variability seen for the other settings used in the BC method. In general, the optimized settings increased in value over the experimentally determined ones.

Given the variation of settings between the monthly

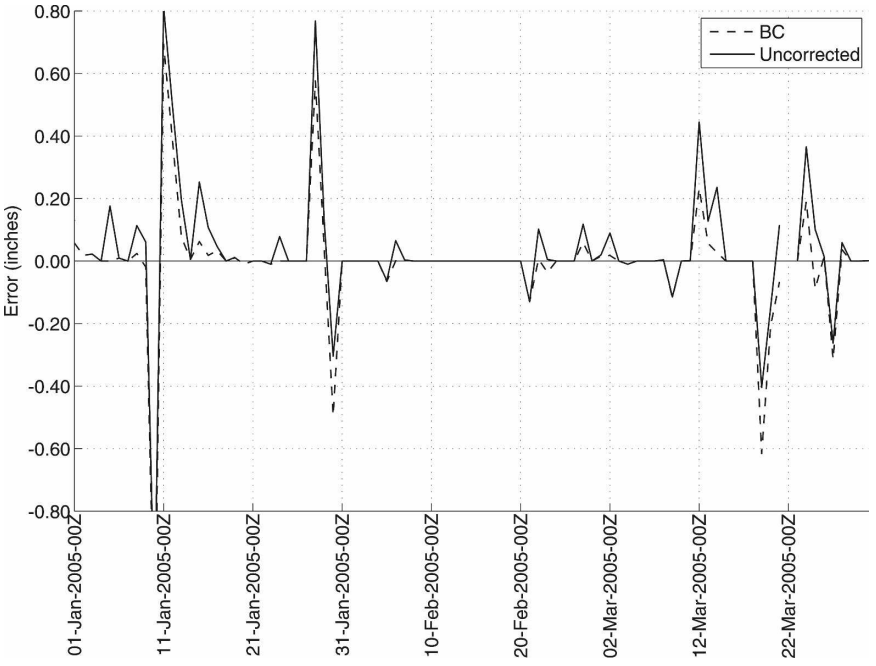

FIG. 15. Error in PCP12 for 48-h forecasts at Seattle–Tacoma International Airport.

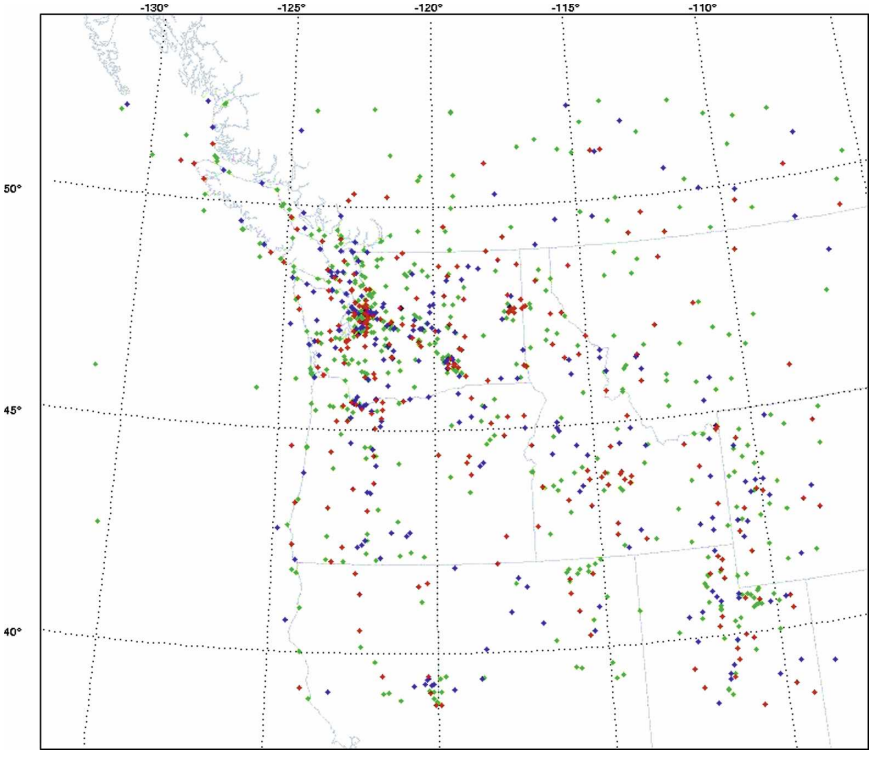

FIG. A1. Observation groups used for bias estimation, optimization, and final verification. Green stations (50% of total) were used for bias estimation during optimization, blue stations (25% of total) were used for verification during optimization, and red stations (25% of total) were used for independent verification of the final, optimized settings.

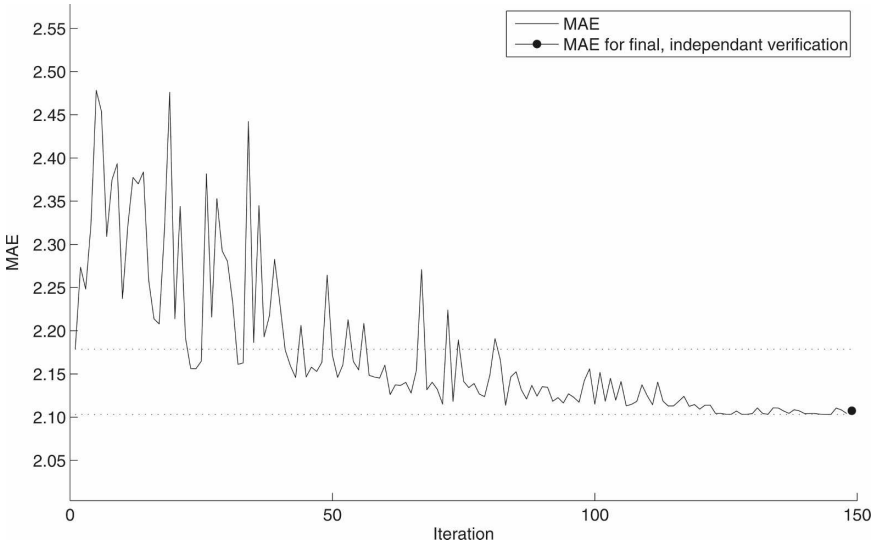

FIG. A2. MAE metric for each iteration of the Evol optimization for July 2004, T2, forecast hour 24.

and annual optimizations, tests were performed to evaluate the bias correction using averages or medians of the settings. For example, the average of the settings from optimizations for all forecast hours verifying at 0000 UTC were used to bias correct each forecast hour (even those verifying at 1200 UTC), with results compared to a given forecast hour's individually optimized settings. The average optimized settings showed competitive results with the forecast hour-specific optimized settings. Various average settings were tested, including the average of the monthly optimized settings for all forecast hours, the average of the annual settings for all forecast hours, and the median of the annual settings for all forecast hours. For T2 and TD2, an average of the annual optimized settings for all forecast hours performed as well as (and in some cases better than) the individual forecast hour-specific optimized settings. The competitive performance of the average annually optimized settings held even when the verification statistics were compared on a monthly basis. For

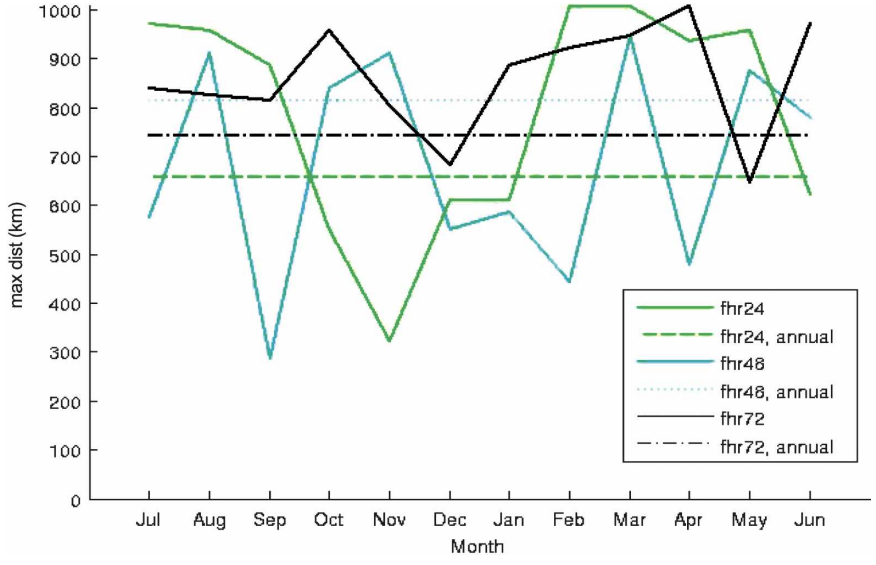

FIG. A3. Settings for the maximum distance between grid points and observations for various monthly and annual optimizations for forecast hours valid at 0000 UTC.

PCP12, an average of the monthly optimized settings for forecast hours 36 and 48 performed the best.

The performance of the BC method was not particularly sensitive to small or even moderate-sized changes to individual settings. Hence, the optimization surface appeared to be relatively "flat." A considerable benefit of this finding is that one group of settings appears to suffice, eliminating the need to vary the settings by season or time of day. Table 3 shows the final settings for T2 and TD2 and Table 4 shows the final settings for PCP12.

For T2 and TD2, the final, optimized settings are larger than the experimentally determined settings in most cases. An effect of these increases in setting values is to increase the data used for bias estimation at stations and grid points. The increases also slow the bias correction algorithm's response to changes in the uncorrected forecast bias, as the number of similar forecasts used increased from 5 to 11 for T2 and from 5 to 10 for TD2.

For PCP12, the initial settings used for optimization were essentially those of the final optimized settings for T2, with the exception being the empirically set precipitation bins. Settings for PCP12 did not change much from these initial settings for all but the precipitation bins, which all increased.

REFERENCES

Baars, J. A., and C. F. Mass, 2005: Performance of National Weather Service forecasts compared to operational, consensus, and weighted model output statistics. *Wea. Forecasting,* **20,** 1034–1047.

Billam, P. J., cited 2006: Math::Evol README and manual. [Available online at http://www.pjb.com.au/comp/evol.html.]

Cressman, G. P., 1959: An operational objective analysis system. *Mon. Wea. Rev.,* **87,** 367–374.

Dallavalle, J. P., and H. R. Glahn, 2005: Toward a gridded MOS system. Preprints, *21st Conf on Weather Analysis and Forecasting and 17th Conf. on Numerical Weather Prediction,* Washington, DC, Amer. Meteor. Soc., 13B.2. [Available online at http://ams.confex.com/ams/pdfpapers/94998.pdf.]

Eckel, F. A., and C. F. Mass, 2005: Effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting,* **20,** 328–350.

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.

——, and D. P. Ruth, 2003: The new digital forecast database of the National Weather Service. *Bull. Amer. Meteor. Soc.,* **84,** 195–201.

Homleid, M., 1995: Diurnal corrections of short-term surface temperature forecasts using the Kalman filter. *Wea. Forecasting,* **10,** 689–707.

Mass, C., and Coauthors, 2003: Regional environmental prediction over the Pacific Northwest. *Bull. Amer. Meteor. Soc.,* **84,** 1353–1366.

Neilley, P., and K. A. Hanson, 2004: Are model output statistics still needed? Preprints, *20th Conf. on Weather Analysis and Forecasting and 16th Conf. on Numerical Weather Prediction,* Seattle, WA, Amer. Meteor. Soc., 6.4. [Available online at http://ams.confex.com/ams/pdfpapers/73333.pdf.]

Ruth, D., 2002: Interactive forecast preparation—The future has come. Preprints, *Interactive Symp. on AWIPS,* Orlando, FL, Amer. Meteor. Soc., 3.1. [Available online at http://ams.confex.com/ams/pdfpapers/28371.pdf.]

Stensrud, D. J., and J. Skindlov, 1996: Gridpoint predictions of high temperature from a mesoscale model. *Wea. Forecasting,* **11,** 103–110.

——, and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.,* **131,** 2510–2524.

Wilson, L. J., and M. Vallée, 2002: The Canadian Updateable Model Output Statistics (UMOS) system: Design and development tests. *Wea. Forecasting,* **17,** 206–222.

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