Performance of National Weather Service Forecasts Compared to Operational, Consensus, and Weighted Model Output Statistics

Jeffrey A. Baars[1]

Clifford F. Mass

Department of Atmospheric Sciences

University of Washington

Seattle, Washington

Submitted September 13, 2004

Revised February 2, 2005

[1] Corresponding Author
Department of Atmospheric Sciences, Box 351640
University of Washington, Seattle, Washington 98195
jbaars@atmos.washington.edu

ABSTRACT

Model Output Statistics (MOS) guidance has been the central model post-processing approach used by the National Weather Service since the 1970s. A recent advancement in the use of MOS is the application of "consensus" MOS (CMOS), an average of MOS from two or more models. CMOS has shown additional skill over individual MOS forecasts and has performed well compared to humans in forecasting contests. This study compares MOS, CMOS, and WMOS (weighting component MOS predictions by their past performance) forecasts of temperature and precipitation to those of the National Weather Service (NWS) subjective forecasts. Data from 29 locations throughout the United States from 1 August 2003 through 1 August 2004 are used. MOS forecasts from the GFS (GMOS), Eta (EMOS) and NGM (NMOS) models are included, with CMOS being a simple average of these three forecasts. WMOS is calculated using weights determined from a minimum variance method, with varying training periods for each station and variable.

Performance is analyzed at various forecast periods, by region of the U.S., and by time/season, as well as for periods of large daily temperature changes or large departures from climatology. The results show that CMOS is competitive or superior to human forecasts at nearly all locations and that WMOS is superior to CMOS. Human forecasts are most skillful compared to MOS during the first forecast day and for periods when temperatures differ greatly from climatology. The implications of these results regarding the future role of human forecasters are examined in the conclusions.

1. INTRODUCTION

Since its advent in the 1970's (Glahn and Lowry 1972), the Model Output Statistics (MOS) approach, based on multiple linear regression, has demonstrated an ability to improve upon the skill of raw forecast model output. This increased accuracy is mainly the result of MOS correcting for model bias and taking into account some of the effects of terrain and surface conditions that are not resolved by the model. Furthermore, MOS has the added benefit of producing probabilistic forecasts based on deterministic model output.

Over time, MOS guidance has shown steady improvement as the skill of the underlying models have improved. Dallavalle and Dagostaro (2004) found that in recent years the forecasting skill of MOS has approached that of National Weather Service (NWS) forecasters, particularly for longer projections. In order to allow human forecasters to work more efficiently, spending time where they can make the greatest contributions, it is critical to understand how the skill of human forecasters compares to objective approaches such as MOS for a wide range of situations, locations, and parameters. Such is the goal of this paper.

It has long been recognized that a consensus of forecasts, be they human or machine produced, often performs better than the component predictions. Initially noted in academic forecasting contests with human forecasters (Sanders 1973; Bosart 1975; Gyakum 1986), these results were extended to objective predictions by Vislocky and Fritsch (1995) who demonstrated the increased skill from a consensus of MOS products. Subsequently, Vislocky and Fritsch (1997) showed that a more advanced consensus MOS-- combining MOS, model output, and surface weather observations-- performed well in a national forecasting competition.

Given that the simple average of different MOS predictions (CMOS) shows good forecast skill, and that prediction quality varies among the individual MOS forecasts, it seems reasonable that a system of weighting the individual MOS forecasts (termed WMOS) could show improvement over CMOS. However, Vislocky and Fritsch (1995), utilizing a simple weighting scheme for NGM and LFM MOS using a year of developmental data, found “no meaningful improvement” over simple averaging. This paper will examine this issue further.

In addition to examining simple measures-based verification statistics such as mean absolute error (MAE), mean squared error (MSE), and bias, it is informative to investigate the circumstances under which a given forecast performs well or performs poorly (Brooks and Doswell 1996; Murphy and Winkler 1987). Statistics such as MAE do not give a complete picture of a forecast’s skill and can be quite misleading in assessing its overall quality. For instance, one forecast may perform well on most occasions, giving a small MAE, but shows poor performance during periods of large departure from climatology. These periods may be of most interest to some users, such as agricultural interests or energy companies, if extreme weather conditions affect them severely. This paper will evaluate NWS and MOS performance for both mean and extreme conditions.

Section 2 describes the data and quality control used in this study. Section 3 details the methods used in generating statistics and how WMOS is calculated.

Results are shown in section 4, and Section 5 summarizes and interprets the results.

2. DATA

Daily MOS and NWS forecasts of maximum temperature (MAX-T), minimum temperature (MIN-T) and probability of precipitation (POP) were gathered from 1 August 2003 through 1 August 2004 for 29 stations spread across the U.S. (Fig. 1). The stations were primarily chosen to be at or near major weather forecast offices (WFOs) and to represent a wide range of geographical areas. Forecasts were taken from the subjective NWS local forecasts, as well as GFS (GMOS), Eta (EMOS), and NGM (NMOS) model output statistics.

MOS forecasts were taken from the 0000 UTC model cycle, and NWS subjective predictions were gathered from the early morning (~1000 UTC or about 0400 PST) forecast. This was done so that NWS forecasters would have access to the 0000 UTC model output and corresponding MOS data. Such an approach gives some advantage to the NWS forecasters, who not only have access to the MOS forecasts, but also have the advantage of considering 6-9 h further development of the weather. For each cycle, forecasts and verification data were gathered for 48 hours, providing for the evaluation of two MAX-T forecasts, two MIN-T forecasts, and four 12-hr POP forecasts.

During the study period, the NWS Meteorological Development Laboratory (MDL) implemented changes to two of the MOS predictions used in the study. On 15 December 2003, Aviation MOS was phased out and became the Global Forecast System (GFS) MOS, or GMOS. For this study, the output from the

AMOS/GMOS was treated as one MOS as the model and equations are essentially the same. This MOS is labeled as “GMOS” throughout this paper. The second change occurred on 17 February 2004 when EMOS equations were changed, being derived from a higher-resolution archive of the Eta model (Hirschberg 2004). As with the AMOS/GMOS change, EMOS was treated as one continuous model throughout the study.

The definitions of observed maximum and minimum temperatures used in this study follow the NWS MOS definitions (Jensenius et al. 1993), with maximum temperatures occurring between 7 AM through 7 PM local time and minimum temperatures occur between 7 PM through 8 AM local time. For probability of precipitation (POP), two forecasts per day were considered: 0000 – 1200 UTC and 1200 – 0000 UTC. Thus, since forecast periods out to 48 hours were considered in this study, precipitation data for four periods were examined (day 1, 1200-0000 UTC; day 2, 0000-1200 UTC, 1200–0000 UTC, and day 3 0000-1200 UTC). Definitions of MAX-T, MIN-T, and POP for the NWS subjective forecasts follow similar definitions (Chris Hill, personal communication, 2003).

While quality control measures are implemented at the agencies from which the data were gathered, simple range checking was also performed on the data used in the analysis. Temperatures below –85°F and above 140°F were removed, POP data were required to be in the range of 0 to 100%, and quantitative precipitation amounts used for verification had to be in the range of 0.0-in to 25.0-in for a 12-hr period. When the differences between forecast and observed temperatures exceeded 40°F, the corresponding data was removed from consideration, since differences greater than 40°F were considered an indicator of error in either the observations or the forecasts. A few such large forecast-observations differences were found and were apparently due to erroneous acquisition of forecast data.

The resulting data set was analyzed to determine the percentage of days when all forecasts and required verification observations were available; it was found that each station had complete data for about 85-90% of the days. There were many days (greater than 50%) when at least one observation and/or forecast was missing from at least one station, making it impossible to remove a day entirely from the analysis when all data were not present. Therefore, only individual station data) were removed from the analysis when missing data occurred. However, for each station and variable, data were required to be complete each day, i.e., *all* forecasts were required to be available for analysis for a given day.

3. METHODS

3.1 MOS Forecasts

CMOS was calculated by simply averaging GMOS, EMOS, and NMOS for MAX-T, MIN-T, and POP. A second CMOS (CMOS-GE) that only used GMOS and EMOS was also calculated. CMOS-GE was calculated in an attempt to improve the original CMOS by eliminating the weakest member (NMOS), which is based on a frozen model of limited horizontal resolution. All seven forecasts (NWS, CMOS, CMOS-GE, WMOS, GMOS, EMOS and NMOS) had to be available for a given day, station and variable to be included in the analysis.

WMOS was calculated using minimum variance estimated weights (Daley 1991). Using this method, weights for each MOS forecast can be calculated using the equation

$$w_n = \frac{\sigma_n^{-2}}{\sum_n \sigma_n^{-2}} \tag{1}$$

where *n* indicates the MOS (GMOS, EMOS and NMOS), w_n is the weight for MOS *n*, and σ^{-2} is the mean square error over a set training period for MOS *n*.

An optimum training period was determined for each station and variable by calculating the minimum squared error produced by a weighted MOS using equation (1) from 1 July 2003 through 1 July 2004. Training periods of 10, 15, 20, 25, and 30 days were tested. With eight variables (two MAX-T's, two MIN-T's and four POP forecasts) and 29 stations, there were 232 training periods stored. Using these pre-determined periods, WMOS was calculated for each station and variable for each day in the study, with weights calculated using equation 1. If any of the three individual MOS forecasts were missing for a given day and variable, WMOS was not calculated.

A table of the average weights across all stations and time periods is given in Table 1. NMOS (GMOS) had the smallest (largest) weights of the three MOS forecasts for all variables. A plot showing a typical time series of weights for MAX-T period 1 for one station (KBNA, Nashville, TN) is given in Figure 2. The training period for this station and variable is 30 days. GMOS has large weights for October and November 2003, and a more even distribution of weights is seen for the remainder of the year. Each model has periods with higher weights than the others, and weights for the three MOS forecasts generally fall between 0.2 and 0.5.

3.2 Verification

Bias, or mean error, is defined as

$$\frac{1}{n}\sum_{i=1}^{n}(f-o) \tag{2}$$

and mean absolute error (MAE) is defined as

$$\frac{1}{n}\sum_{i=1}^{n}\left|f-o\right| \tag{3}$$

where *f* is the forecast, *o* is the observation, and *n* is the total number of forecast/observation pairs. Precipitation observations were converted to binary rain/no-rain data (with trace amounts treated as no-rain cases). The Brier Score is defined as

$$\frac{1}{n}\sum_{i=1}^{n}(f-o)^2 \tag{4}$$

where *f* is the forecast probability of rain (0 to 1: 0 to 100%) and *o* is the observation converted to binary rain/no-rain data. Brier Scores range from 0.0 (perfect forecast) to 1.0 (worst possible forecast). The resolution of the MOS POP is 1% while the resolution of the NWS POP is generally 10%, although for low POP events the NWS occasionally forecasts 5%.

To better understand the circumstances under which each forecast performed well or poorly, a type of “distributions-based” verification was performed (Brooks and Doswell 1996; Murphy and Winkler 1986). Periods of large (+/-10°F) one-day changes in observed MAX-T or MIN-T change were examined, since such periods were expected to be challenging for the forecaster and the models. This study also examined periods when observed temperatures departed significantly from climatology, since it is hypothesized that human forecasters might have an advantage over statistical approaches during such times. Days showing a large departure from climatology were determined from monthly average maximum and minimum temperature data from the National Climatic Data Center for the 1971-2000 data period, which were then interpolated linearly to each date. A large departure from climatology was defined to be +/- 20°F.

As another measure of forecast quality, the number of days each forecast was the most or least accurate was also determined. A given forecast may have low MAE for a forecast variable but is rarely the most accurate forecast. Or a given forecast may be most accurate on more days than other forecasts, but least accurate on average due to infrequent large errors. This type of information remains hidden when considering only standard verification measures such as the MAE.

4. RESULTS

4.1 Temperature

Summary MAE scores were calculated using all stations, both for MAX-T and MIN-T, over all forecast periods (Table 2). It can be seen that WMOS has the lowest total MAE, followed by CMOS-GE, CMOS, NWS, GMOS, EMOS and NMOS. MAEs are notably lower than was found by Vislocky and Fritsch (1995)-- who reported MAEs of about 3.5°F for NWS, CMOS, limited-area fine mesh (LFM)-based MOS, and NMOS-- presumably due to more than ten years of model improvement. The NWS's National Verification Program, using data from 2003, reports similar MAEs for GMOS, the Medium Range Forecast (MRF) MOS and NMOS to those shown here (Taylor and Stram 2003).

Figure 3 shows the distribution of absolute errors over all stations and the entire period of the study. Bins are 1°F in size, centered on each whole degree. The most frequent error is 1°F, with NWS, CMOS, CMOS-GE, and WMOS having similar distributions, and NMOS and EMOS having more larger errors.

Figure 4 shows MAE for MAX-T and MIN-T for each of the forecast periods. Period 1, "pd1" is the day 1 MAX-T, period 2, "pd2," is the day 2 MIN-T, period 3, "pd3," is the day 2 MAX-T, and period 4, "pd4," is the day 3 MIN-T. The sample sizes for MAX-T pd1, MIN-T pd2, MAX-T pd3 and MIN-T pd4 are 8893, 8856, 8980, and 8947 respectively. WMOS is the most skillful forecast, with the lowest MAEs for all periods. NWS has a lower MAE than both CMOS and CMOS-GE for the period 1 MAX-T, but CMOS and CMOS-GE have similar MAEs to the NWS for period 3 MAX-T and lower MAEs for both MIN-T's. The individual MOS forecasts have higher MAEs than WMOS, NWS, CMOS and CMOS-GE for all periods and GMOS has the lowest MAEs of the individual MOS predictions. It appears that human intervention is most positive during the first period and even then a weighted MOS is superior.

To determine forecast skill during periods of large temperature change, MAEs were calculated on days having a 10°F change in MAX-T or MIN-T from the previous day. Results of these calculations are shown in Fig. 5. The sample sizes for MAX-T pd1, MIN-T pd2, MAX-T pd3 and MIN-T pd4 are 1465, 1452, 1406, and 1390 respectively. There is approximately a 1.0 to 1.5°F increase in MAEs for the seven forecast types compared to the statistics for all times (Fig. 4), and WMOS again has the lowest MAE for all periods except for period 4 MIN-T, when GMOS shows the lowest MAE. Apparently the poor performance of EMOS and NMOS affected the combined MOS forecasts enough to cause GMOS to show the lowest MAE score. NWS has lower MAEs than CMOS for all periods but similar or higher MAEs than CMOS-GE.

As noted in the methods section, MAEs were also calculated for days on which observed maximum or minimum temperatures departed by more than 20°F from the daily climatological values. Results from these calculations are given in Fig. 6. The sample sizes for MAX-T pd1, MIN-T pd2, MAX-T pd3 and MIN-T pd4 are 223, 213, 173, and 170 respectively. In general, errors are several degrees F larger than the unfiltered data set for all forecast types (Fig. 4). The NWS shows considerably higher skill (lower MAE) relative to the other forecasts for the period1 MAX-T, with MAEs of about 0.5°F lower than the next best forecast, WMOS. For MIN-T period 2 and MAX-T period 3, CMOS-GE performs the best; apparently the poor performance of NMOS increases MAEs for both CMOS and WMOS. As was seen for the one-day temperature change statistics, GMOS performs best out of all forecasts for MIN-T period 4. In short, when there are large deviations from climatology, human intervention can be quite positive for the first period maximum temperature, but subjective predictions drop back into the pack for longer period forecasts.

Figure 7 shows the number of days that each forecast was most accurate (i.e., had the smallest absolute forecast error). In Figure 7a, when two or more forecasts possessed the same error and were most accurate, each was awarded one day. Consensus and weighted MOS forecasts show the fewest number of days having the most accurate forecast for all periods and variables. Similar results for consensus MOS have been reported by Wilks (1998). For period 1 MAX-T, NWS was most frequently the most accurate forecast, with about 50% more days than CMOS, 40% more days than CMOS-GE and WMOS, and 15% more days than the individual MOS predictions. For the remaining periods, the individual MOS forecasts are more comparable to the NWS, and GMOS actually shows more days with the most accurate forecast for MIN-T period 4.

Since errors of 1-2°F are close to the magnitude of typical instrument error and small enough to pass without notice for most users, the calculation of the number of days each forecast was most accurate was redone, considering errors of less than or equal to 2°F a tie. The results, shown in Figure 7b, are considerably altered from Figure 7a, with the consensus MOS predictions having more most accurate days than the NWS and the individual MOS products at all projections.

Figure 8a shows the number of days each forecast had the *least* accurate prediction, considering ties when two or more forecast temperatures have the same worst forecast. For this situation, consensus or weighted MOS forecasts are far superior to the NWS and individual MOS predictions, with less than half the number of days being the worst forecast. Clearly, the averaging used to compute these consensus/weighted forecasts tends to eliminate extremes, so that they are seldom the least accurate forecast, relative to NWS and the individual MOS's (GMOS, EMOS, NMOS). The consensus and weighted MOS forecasts can only have a least accurate day in a "tie" situation where all MOS forecasts agree. NWS has considerably fewer least-accurate forecasts than the individual MOS's for most periods. Thus, it appears that human forecasters are often able to improve upon individual MOS forecasts and thus avoid being the worst prediction. Figure 8b, shows the worst forecast results using the relaxed definition of ties (all forecasts within 2°F of the worst forecast are counted as having a worst prediction). A substantial leveling of the performance of the various predictions results from this loosened definition. For the first two periods, the NWS has slightly fewer poor forecasts than WMOS, with WMOS modestly superior over the final two periods. In short, consensus or weighted MOS appears to gain some improvement upon the NWS on average by greatly reducing the number of times they are the worst forecast and only moderately reducing their frequency of being the best forecast.

Figure 9 shows a time series of MAE for MAX-T period 1 averaged over all stations for NWS, CMOS, and WMOS over the entire study. The average temperature over all stations is also shown with a dotted line. The correlation among the three forecast MAEs is quite evident. WMOS has the lowest MAEs for the most periods, followed by the NWS and CMOS. An increase in MAEs for all three forecasts occurs in the 2003 –2004 cold season. It appears than NWS forecasters reduced some of the peak errors of the MOS guidance, particularly during the cold period in early January 2004.

The nature of the human intervention is illustrated in Figure 10, which shows a time series of bias for the first period maximum temperature over all stations for NWS, CMOS, and WMOS. As in Fig. 9, a correlation among the three forecast biases is evident, as is a pronounced warm bias by CMOS and WMOS throughout much of the 2003-2004 cold season. In general, the NWS forecasts have the least bias, particularly for the coldest periods when the MOS biases are largest. This presumably shows that NWS forecasters understand MOS bias and can compensate for them to a substantial degree. The fact that the MOS predictions have such a large and consistent bias indicates the need to improve the MOS approach to correct for persistent short-term bias.

Figure 11 compares the performance of the various NWS forecast offices and two types of consensus MOS for MAX-T over the entire study period. The stations are sorted geographically, starting in the West and moving through the Inter-mountain West and Southwest, the Southern Plains, the Southeast, the Midwest, and the Northeast (see map, Fig. 1). MAEs are typically around 2-2.5°F and vary more spatially in the western U.S. than over the northeast U.S. Not surprisingly, tropical MIA (Miami, FL) shows the lowest MAE, with desert-southwest LAS (Las Vegas, NV) and PHX (Phoenix, AZ) also having low MAX-T MAEs. Higher MAEs are seen at high-elevation stations in the west, such as at Missoula-MSO and Denver-DEN, where cold-air outbreaks and upslope flows can create difficult forecasting situations. As noted later, such stations are also ones for which MOS has a distinct warm bias. CMOS appears to do worse, compared to NWS and WMOS, at western stations such as MSO and DEN, perhaps due to the coarse elevation in the NGM model.

Figure 12 shows biases for MAX-T, period 1, for each of the 29 individual stations in the study for the entire study period. A prominent feature is the positive (warm) biases through much of the western U.S., particularly at higher altitude stations in the Intermountain West and Southwest, with lesser positive biases extending through the Southern Plains and into the South. Small negative (cool) biases are observed in much of the Midwest. NWS forecasters are most effective in improving on MOS for the western stations, where they considerably reduce the warm MOS bias at the higher stations.

4.2 Precipitation

Brier Scores for the study period for the seven forecasts for all stations and forecast periods are given in Table 3. The scores do not vary greatly. WMOS and CMOS-GE show the lowest (best) scores, followed by CMOS, AMOS, NWS and EMOS, and NMOS.

Figure 13 shows the number of occurrences of forecasts of various precipitation probabilities when precipitation did (left side) and did not (right side) occur for NWS, CMOS, WMOS and GMOS during forecast period 1. The figure also presents normalized squared errors for the four forecasts (dashed lines), calculated by dividing a forecast’s total squared error (forecast minus observation) in each POP bin by the total squared error over all bins for that forecast type. Bins are 10% in size, centered every 10%. The occurrence distributions show that the four forecasts are very similar to each other and that the number of POP forecasts is more uniform as probability varies when precipitation is observed. Specifically, there are a nearly equal number of occurrences of POP forecasts of 40% through 100% during precipitation cases. For non-precipitation cases, the distributions are skewed towards 0% POP, with a nearly exponential drop towards higher forecast POPs. In other words, the predictions are sharpest when rain is not observed, with a strong tendency to forecast a POP of 0 to 20%, while when rain occurs the forecasts are more wide ranging.

The normalized squared error distributions (dashed lines in Fig. 13) show the situations contributing to the total squared error of each forecast. Much of the error for the forecasts comes from forecasting POP of 10% to 50% during times of precipitation, and POPs of 30 to 70% when precipitation does not verify. The former cases represent times when the forecaster (or MOS) feels there is an elevated chance of precipitation but much uncertainty exists.

Reliability diagrams for NWS, CMOS, WMOS and GMOS for the four forecast periods are shown in Fig. 14. During period 1 (Fig 14a), the predictions are quite similar, with all exhibiting a slight underforecasting bias for lower forecast probabilities. At the highest forecast probabilities, there is an overforecasting bias for all predictions, with the NWS forecasts somewhat worse than the others. At longer forecast periods (Fig. 14b, c and d), an “s-shaped” pattern develops, with an overforecasting bias at the lowest forecast probabilities and an underforecasting bias being seen at higher forecast probabilities (50-

90%). GMOS forecasts appear to be the most reliable, most closely matching the 1 to 1 line.

Figure 15 shows Brier Scores for each of the four 12-hr precipitation forecast periods. WMOS has the highest skill (lowest Brier scores) for all periods. There is a substantial increase in Brier Scores with increasing forecast projection for all forecasts. CMOS-GE has the second best Brier Scores, followed by CMOS and NWS and the individual MOS forecasts. Although the NWS outperforms all individual MOS predictions during the first period, that advantage is lost to GMOS for all subsequent forecasts. NMOS has particularly poor Brier Scores relative to the other forecasts.

Figure 16 shows a time series of Brier Scores for NWS, CMOS and WMOS averaged for all stations for period 1. A very high correlation among the three forecasts is seen with time, far more so than for temperature. A general increase in Brier Scores (decline in skill) is seen during the warm season when convection is more prominent.

Figure 17 shows Brier Scores by station. Low precipitation locations, such as at LAS, PHX and ABQ, show the lowest (most skillful) Brier Scores. MIA shows the highest Brier Scores due to the considerable convective precipitation at that location. In general, NWS forecasters have poorer scores than CMOS or WMOS, with only a handful of stations showing better human performance compared to the objective guidance.

5. CONCLUSIONS

This study compares the skill of NWS forecasts throughout the United States with the predictions of individual, composite, and weighted Model Output Statistics. Consensus Model Output Statistics (CMOS) was calculated by simple averaging of three individual MOS forecasts (GFS MOS-GMOS, Eta MOS-EMOS, and NGM MOS-NMOS), while a weighted MOS (WMOS) combines these MOS forecasts based on their previous performance. In general, CMOS shows equal or superior forecast performance in terms of overall MAEs and Brier Scores to that of the NWS and of individual MOS's. WMOS shows superior forecast performance to that of CMOS. Relative to individual MOS forecasts, NWS forecasts perform better for temperature than for precipitation, and even GMOS outperforms NWS for precipitation for all but the 12-hr forecast. The removal of the weakest model (NGM MOS) from the consensus forecasts (CMOS-GE) produces an increase in skill for some forecast variables.

Time series of NWS and WMOS/CMOS MAEs and biases show very similar temporal evolutions, with NWS forecaster adjustments to MOS indicating considerable awareness of seasonal temperature biases in the MOS. Regional variations in MAE and bias are apparent in the data, with larger errors at high altitude stations of the western U.S.

NWS forecasters performed particularly well for short-term temperature forecasts when there are large (+/- 20°F) departures from climatology. During periods of a large (+/- 10°F) one-day temperature change, CMOS and WMOS are competitive or have lower MAEs than the NWS.

Calculating the total number of days that each forecast predicted the most or least accurate temperatures revealed that for a loose definition of tie forecasts

(forecasts within 2°F of each other considered to be equally useful) the consensus and weighted MOS forecasts are most often the most accurate, while the NWS predictions are slightly less frequently the least accurate.

Reliability analysis reveals that all forecasts are generally reliable the first day, with a tendency for overpredicting precipitation probability when the forecasts are 80-100%. By the second day, an “s-shaped” reliability diagram is evident, with overprediction for low probabilities and underprediction for higher probabilities.

An interesting, and perhaps surprising, result of this analysis is the existence of systematic and sustained bias in some of the MOS forecasts. Often evident for high-elevation stations and during periods of sustained cold temperatures, such bias is an important source of NWS forecaster improvement over MOS. It might be expected that an improved or more sophisticated MOS, perhaps using previous bias over some training period as a predictor, might alleviate this systematic bias and greatly reduce the value of human intervention.

An essential finding of this paper is that it is getting increasingly difficult for human forecasters to improve upon MOS, a simple statistical post-processing of ever-improving model output. Humans cannot consistently beat MOS precipitation forecasts for virtually all of the locations and forecast projections examined in this study, and are only superior to MOS for short-term temperature forecasts during large excursions from climatology. These results are consistent with the recent results of Dallavalle and Dagostaro (2004), which showed that during the past two years, human and MOS skill in predicting short-term (24 and 48 h) probability of precipitation and minimum temperatures have become virtually equivalent, with only maximum temperature providing an arena in which human forecasts are marginally better (0.3 to 0.5°F).

These results have significant implications for the future of forecasters in the NWS and the transition to gridded forecast preparation/dissemination using the new Integrated Forecast Preparation System (IFPS) system. Currently, forecasters spend much of their time preparing forecast grids out to seven days using IFPS. Using this system, NWS forecasters can start with gridded model output, previous IFPS gridded predictions, or with MOS station forecasts spread throughout their domain, and then merge and modify these data as part of the forecast process. The need for constant updating of forecast grids often leaves little time for short-term prediction and nowcasting, a critical deficiency in NWS operations. As noted above, this study indicates that for all but the first 12-h it is very difficult for forecasters to consistently beat MOS, with MOS superiority being enhanced using a consensus or weighted MOS product. These findings imply it would be far better for forecasters to put less emphasis on creating forecast grids beyond 12 hours, leaving the such predictions in most cases to bias-corrected model output, which retrieves much of the skill increase of MOS but with less effort (Neilley and Hanson 2004), or MOS output distributed throughout their domain. The NWS is currently developing a grid-based MOS that will improve upon NCEP model output over large domains.

If NWS forecasters cannot beat MOS at observation locations where they can gain a deep familiarity with MOS verifications over a wide range of conditions, they are unlikely to improve upon a model/MOS gridded forecasting system. Furthermore, if the relatively primitive post-processing of the current

MOS, based on simple linear regression, is competitive or superior to subjective predictions for 12h and beyond, one can imagine the potential of more modern post-processing approaches such as neural networks. An implication of the transition to human/MOS equivalence in prediction skill for precipitation and temperature at 12h and beyond is that humans should spend most of their time on the short-term (0-12 h) forecasting problem, where the combination of superior graphical interpretation and physical understanding, coupled with the ability to communicate with the user communities, will allow profound improvements in the accuracy and usability of forecast information. Thus, this paper should not be seen as an excuse to reduce the number and responsibilities of forecasters, but rather as an indication that they should shift their efforts to important short-term forecasting problems and user interactions that are inadequately served today.

ACKNOWLEDGEMENTS

This research has been supported by grants from the NOAA Collaborative Science, Technology and Applied Research (CSTAR) program and the ONR Multidisciplinary University Research Initiative (MURI). Constructive comments have been provided by Chris Hill and Brad Colman of the Seattle office of the NWS. Mr. Paul Dallavalle of the NWS's Meteorological Development Laboratory provided updated information on current MOS performance, and Mark Albright conducted an initial investigation into this study and helped set up the basic framework for data gathering.

6. REFERENCES

Bosart, L. F., 2003: Whither the weather analysis and forecasting process? *Wea. Forecasting*, **18**, 520-529.

Brooks, H. E., and C. A. Doswell, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecasting*, **11**, 288-303.

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press, 457 pp.

Dallavalle, J. P., and V. J. Dagostaro, 2004: Objective interpretation of numerical weather prediction model output – A Perspective based on verification of temperature and precipitation guidance. Preprints, *Symposium on the 50th Anniversary of Operational Numerical Weather Prediction*, College Park, MD, *Amer. Meteor. Soc.*, CD-ROM, 6.1.

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.

Gyakum, J. R. 1986: Experiments in temperature and precipitation forecasting for Illinois. *Wea. Forecasting,* **1**, 77-88.

Hirschberg, P., cited 2004: Amended: Change to issuance time of Eta MOS Guidance Effective February 17 2004 at 1200 UTC. [Available online at http://nws.noaa.gov/om/notifications/tin03-48eta_mos_aaa.txt]

Jensenius, J. S., Jr., J. P. Dallavalle, and S. A. Gilbert, 1993: The MRF-based statistical guidance message. NWS Technical Procedures Bulletin No. 411, NOAA, U.S. Dept. of Commerce, 11 pp.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330-1338.

Neilley, P., and K. A. Hanson, 2004: Are model output statistics still needed? Preprints, *20th Conference on Weather Analysis and Forecasting/16th Conference on Numerical Weather Prediction,* Seattle, WA, Amer. Meteor. Soc., CD-ROM, 6.4.

Sanders, F. 1973: Skill In forecasting daily temperature and precipitation: some experimental results. *Bull. Amer. Meteor. Soc.*, **54**, 1171-1178.

Taylor, A., and M. Stram, cited 2003: MOS verification. [Available online at http://www.nws.noaa.gov/mdl/verif/].

Vislocky, R., and J. M. Fritsch, 1995: Improved model output statistics forecasts through model consensus. *Bull. Amer. Meteor. Soc.*, **76**, 1157-1164.

___________, and _____________. 1997: Performance of an advanced MOS system in the 1996-97 National Collegiate Weather Forecasting Contest. *Bull. Amer. Meteor. Soc*., **78**, 2851-2857.

Wilks, G. E., cited 1998: Intercomparisons among the NGM-MOS, AVN-MOS, and consensus temperature forecasts for West Texas. [Available online at http://www.srh.noaa.gov/topics/attach/html/ssd98-39.htm].

Table 1. Weights used in WMOS, averaged over all 29 stations and the entire study period.

Variable	GMOS	EMOS	NMOS
MAX-T pd1	0.368	0.322	0.310
MAX-T pd2	0.374	0.324	0.303
MIN-T pd1	0.373	0.334	0.294
MIN-T pd2	0.394	0.329	0.278
Precipitation, pd 1	0.332	0.372	0.279
Precipitation, pd 2	0.346	0.342	0.280
Precipitation, pd 3	0.358	0.355	0.274
Precipitation, pd 4	0.355	0.333	0.291

Table 2: MAE for each forecast type for 1 August 2003 to 1 August 1 2004. These averages include data for all stations, all forecast periods, and both maximum (MAX-T) and minimum (MIN-T) temperatures.

Forecast	MAE (°F)
WMOS	2.41
CMOS-GE	2.49
CMOS	2.50
NWS	2.54
GMOS	2.64
EMOS	2.84
NMOS	2.97

Table 3. Brier Scores for 1 August 2003 – 1 August 2004 for all stations and forecast periods.

Forecast	Brier Score
WMOS	0.091
CMOS-GE	0.091
CMOS	0.092
GMOS	0.094
NWS	0.096
EMOS	0.096
NMOS	0.105

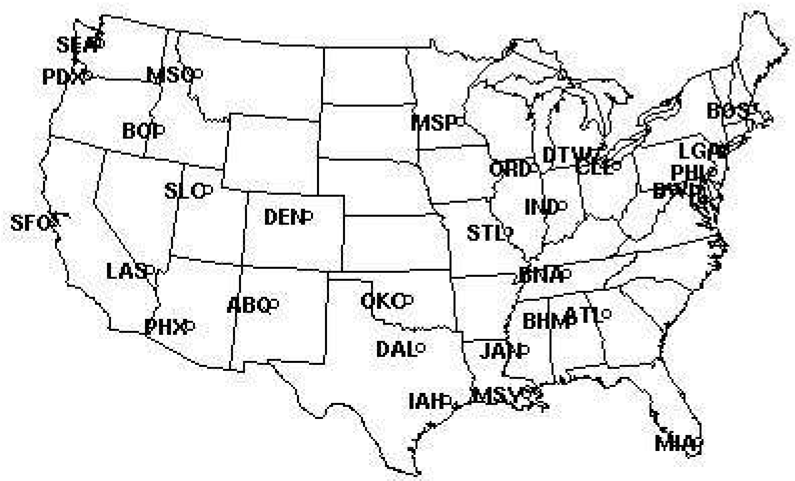

Fig. 1. NWS locations used in the study.

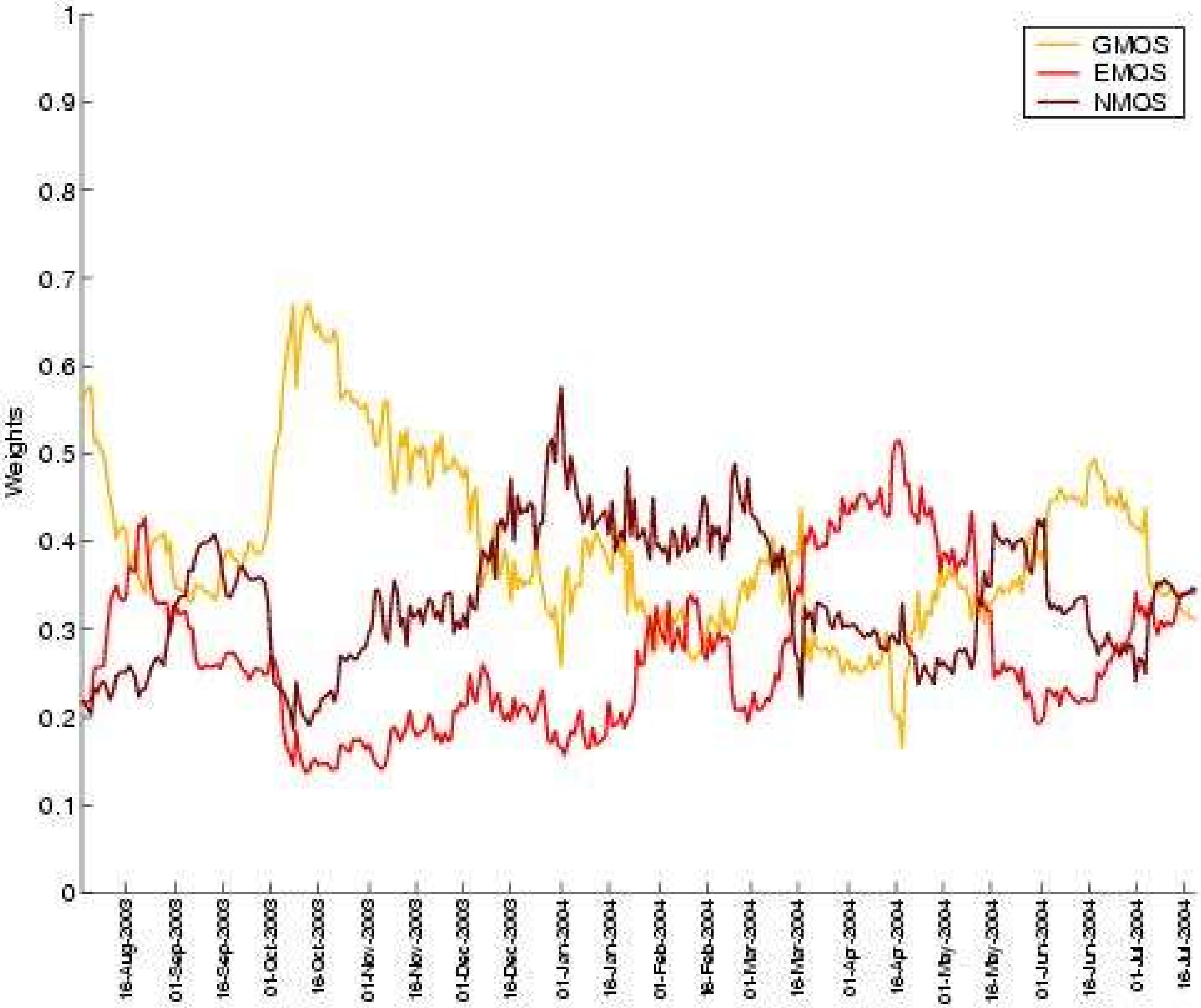

Fig. 2. Time series of weights used for each of the three MOS forecasts in WMOS for MAX-T period 1, Nashville, TN (KBNA) over the study period.

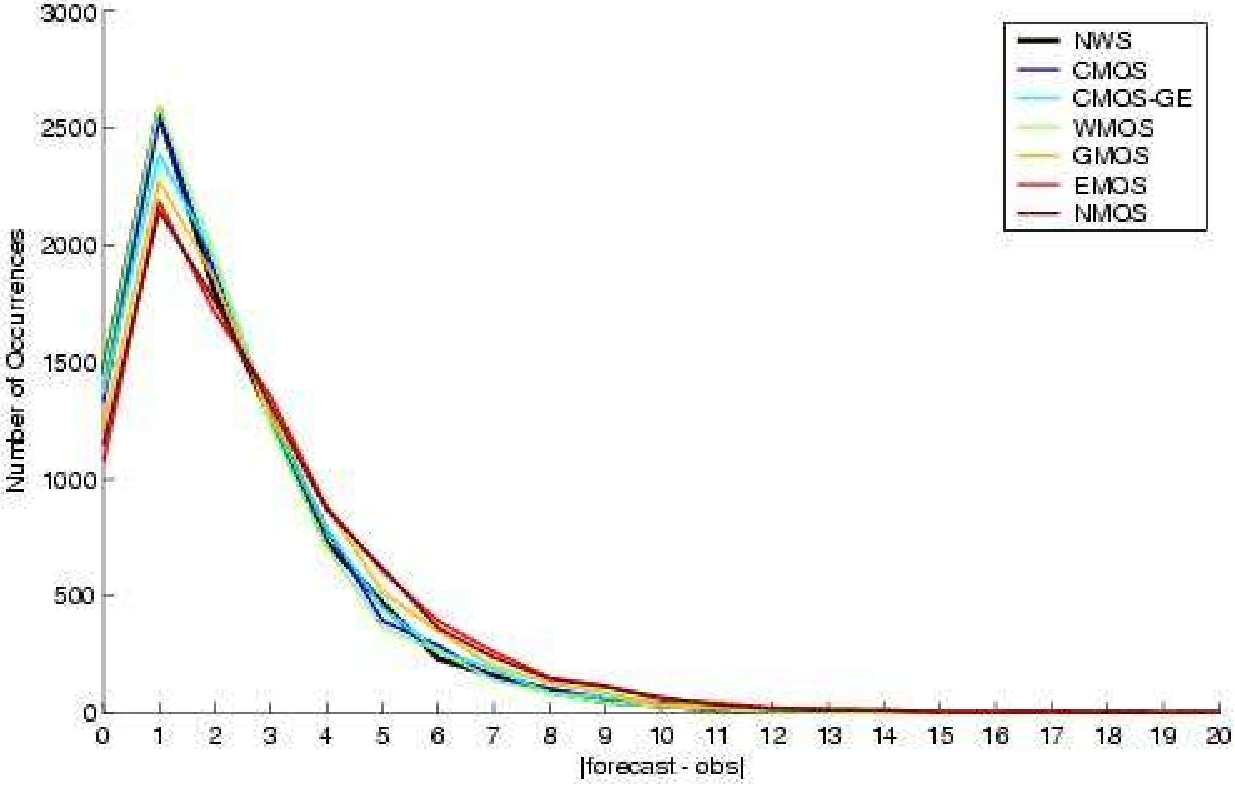

Fig. 3. The number of occurrences of absolute errors for first period maximum temperatures for 1 August 2003 through August 1 2004. Bins are 1°F in size, centered on each whole degree.

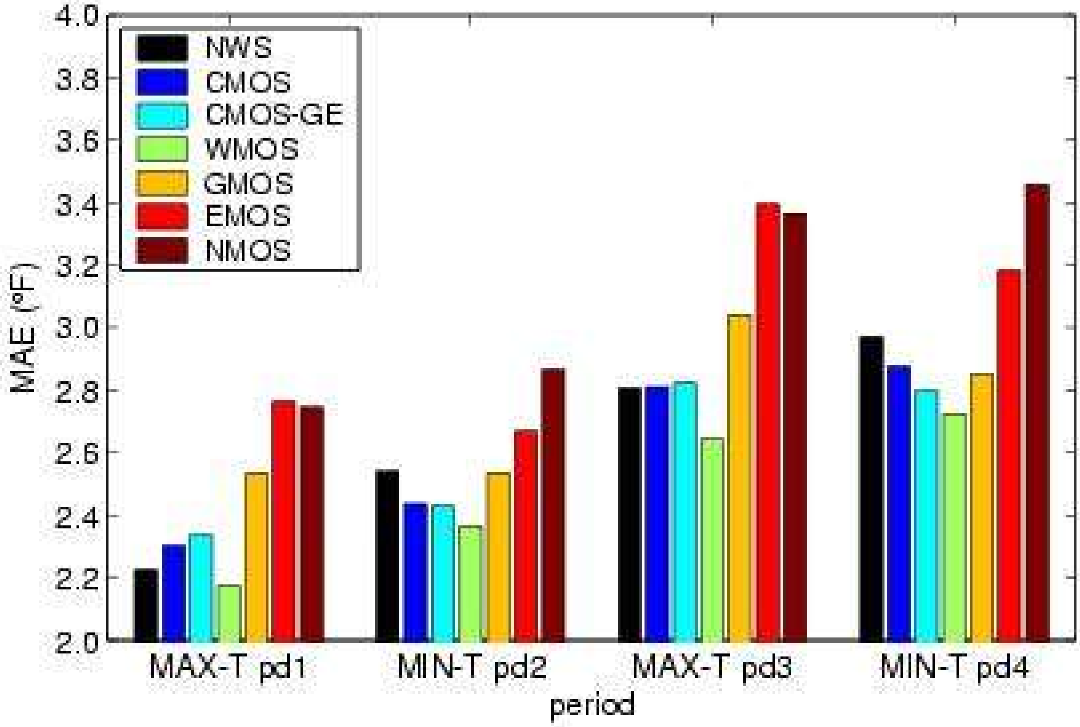

Fig. 4. MAE (°F) for the seven forecasts for all stations, all time periods, 1 August 2003 – 1 August 2004.

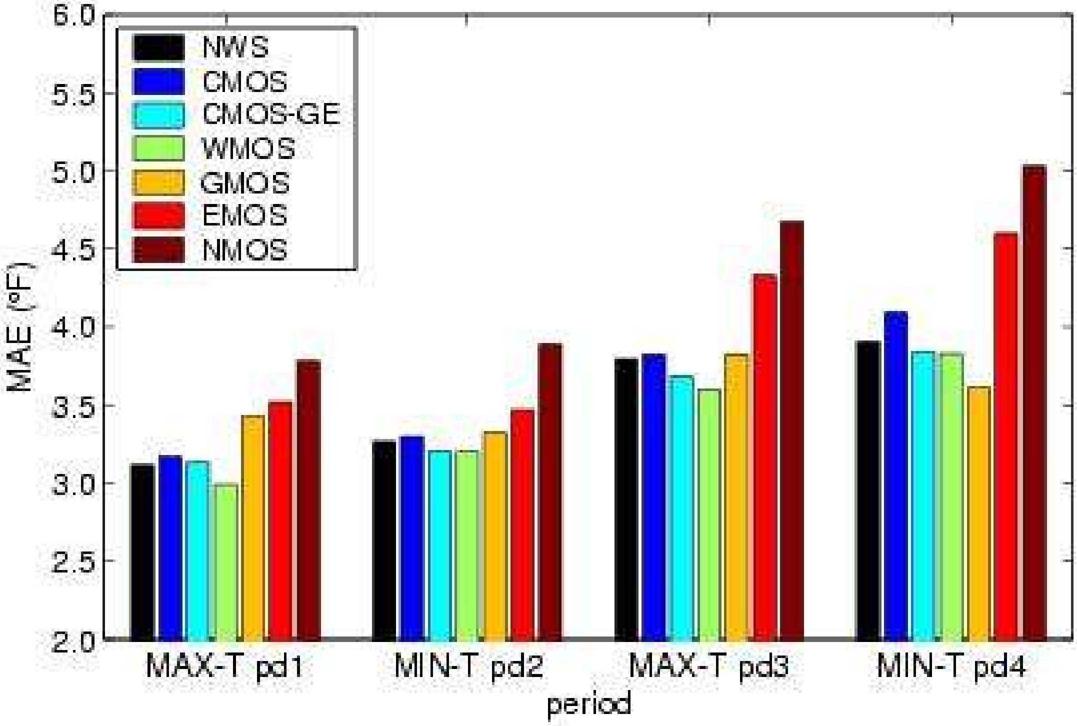

Fig. 5. MAE for each forecast during periods of large temperature change (10°F over 24-hr), 1 August 2003 – 1 August 2004. Includes data for all stations.

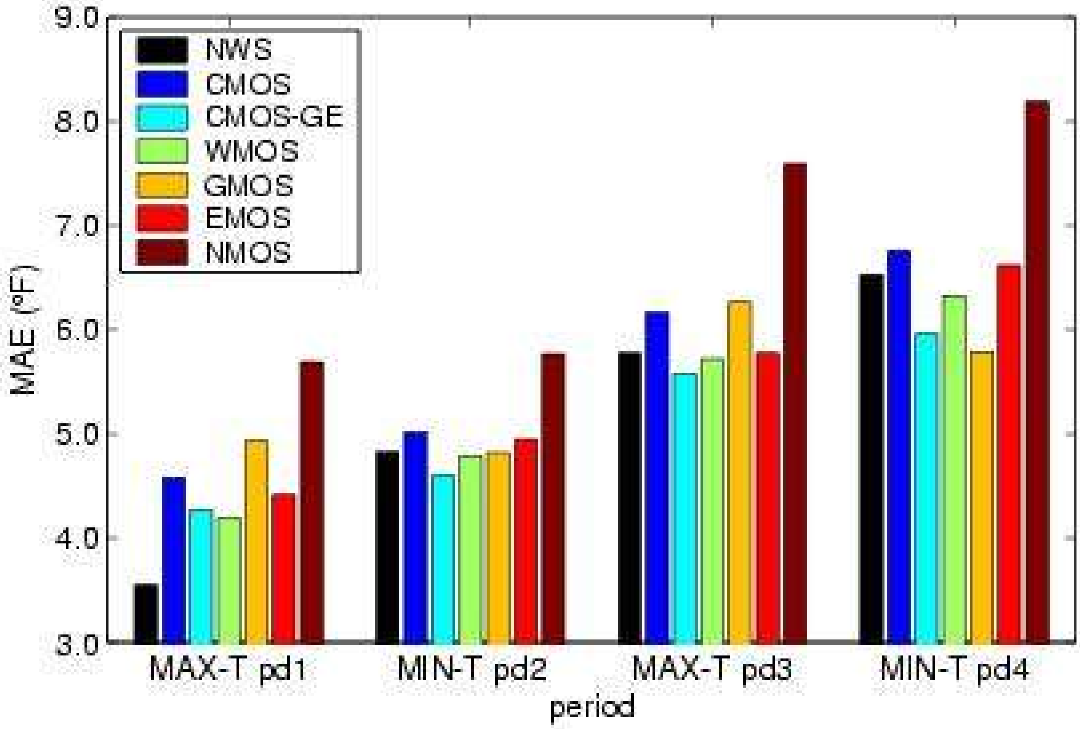

Fig. 6. MAE for each forecast during periods of large departure (20°F) from daily climatological values, 1 August 2003 – 1 August 2004, all stations.

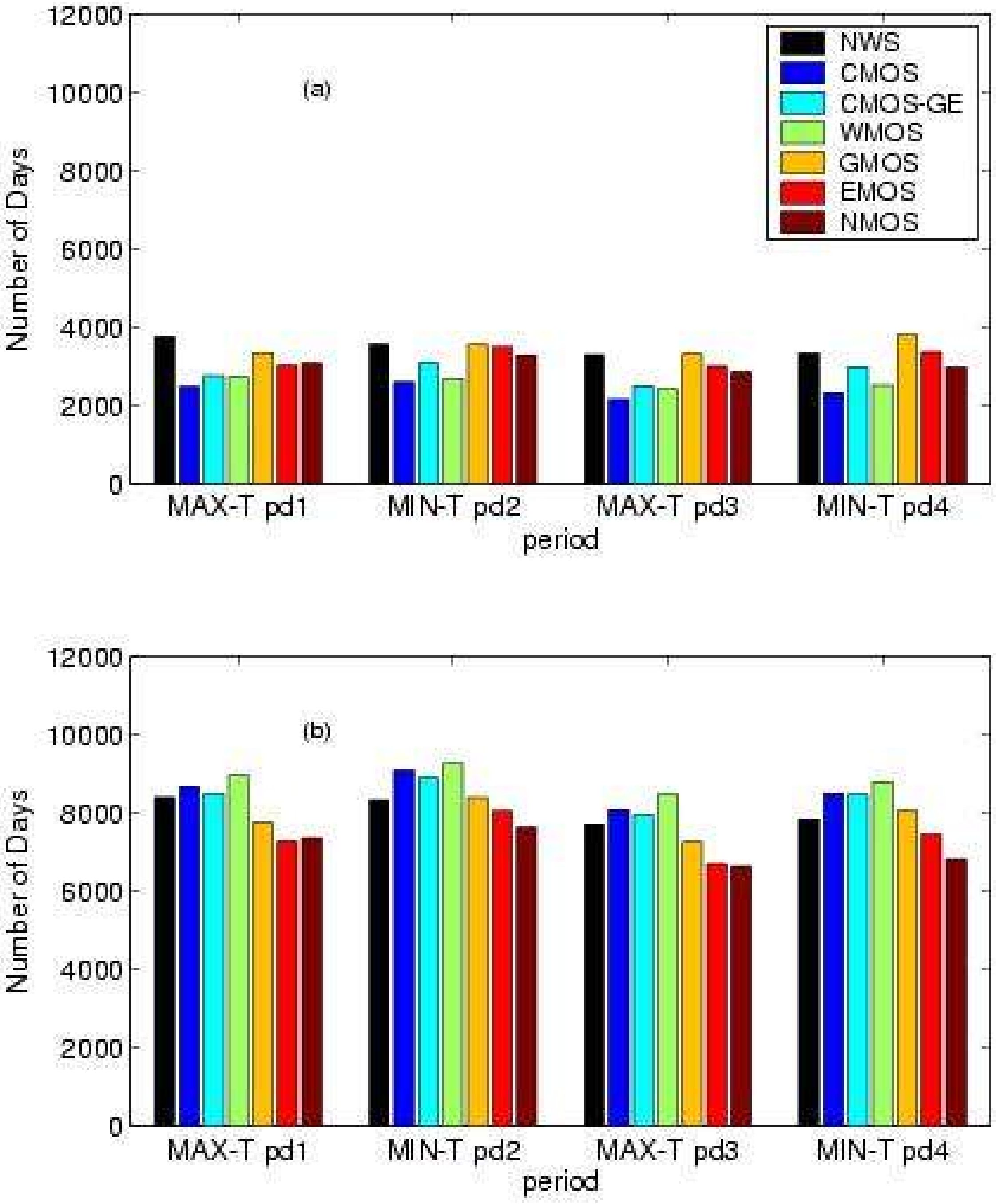

Fig. 7. Number of days each forecast is the most accurate, all stations, 1 August 2003 – 1 August 2004. In (a), tie situations are counted only when the most accurate temperatures are exactly equivalent. In (b), tie situations are cases when the most accurate temperatures are within 2°F of each other.

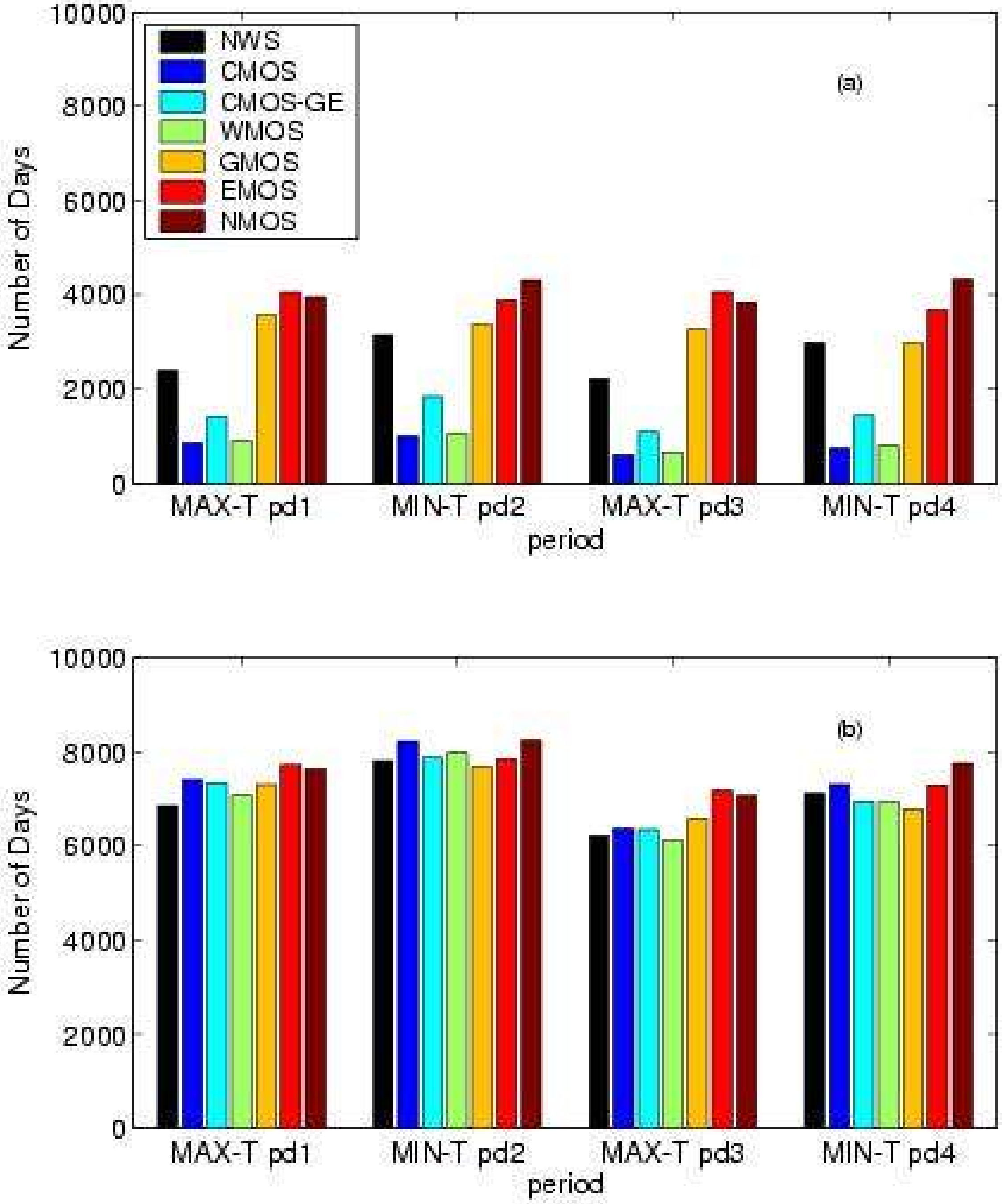

Fig. 8. Number of days each forecast is the least accurate, all stations, 1 August 2003 – 1 August 2004. In (a), tie situations are counted only when the least accurate temperatures are exactly equivalent. In (b), tie situations are cases when the least accurate temperatures are within 2°F of each other.

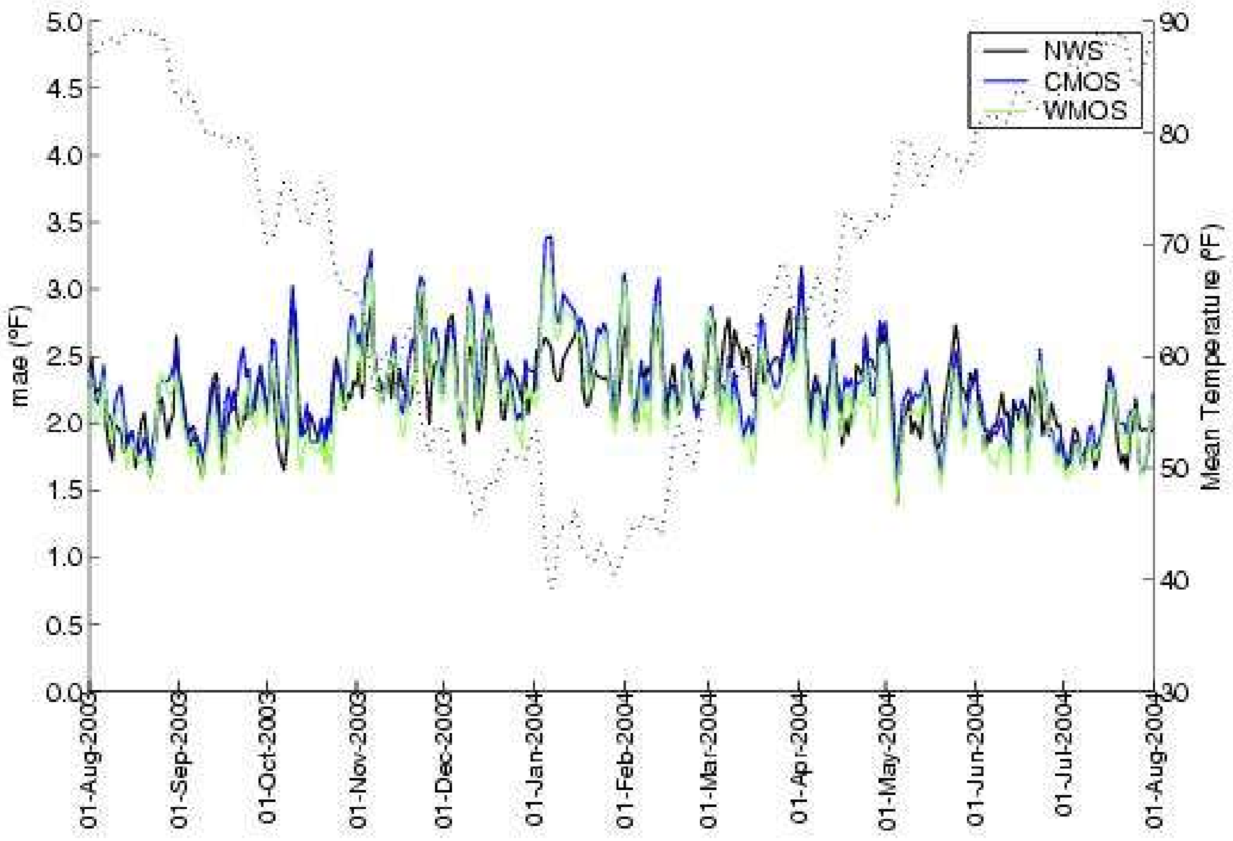

Fig. 9. Time series of MAE of MAX-T for period one for all stations, NWS, CMOS and WMOS forecasts, 1 August 2003 – 1 August 2004. The mean temperature over all stations is shown with a dotted line. 3-day smoothing is performed on the data.

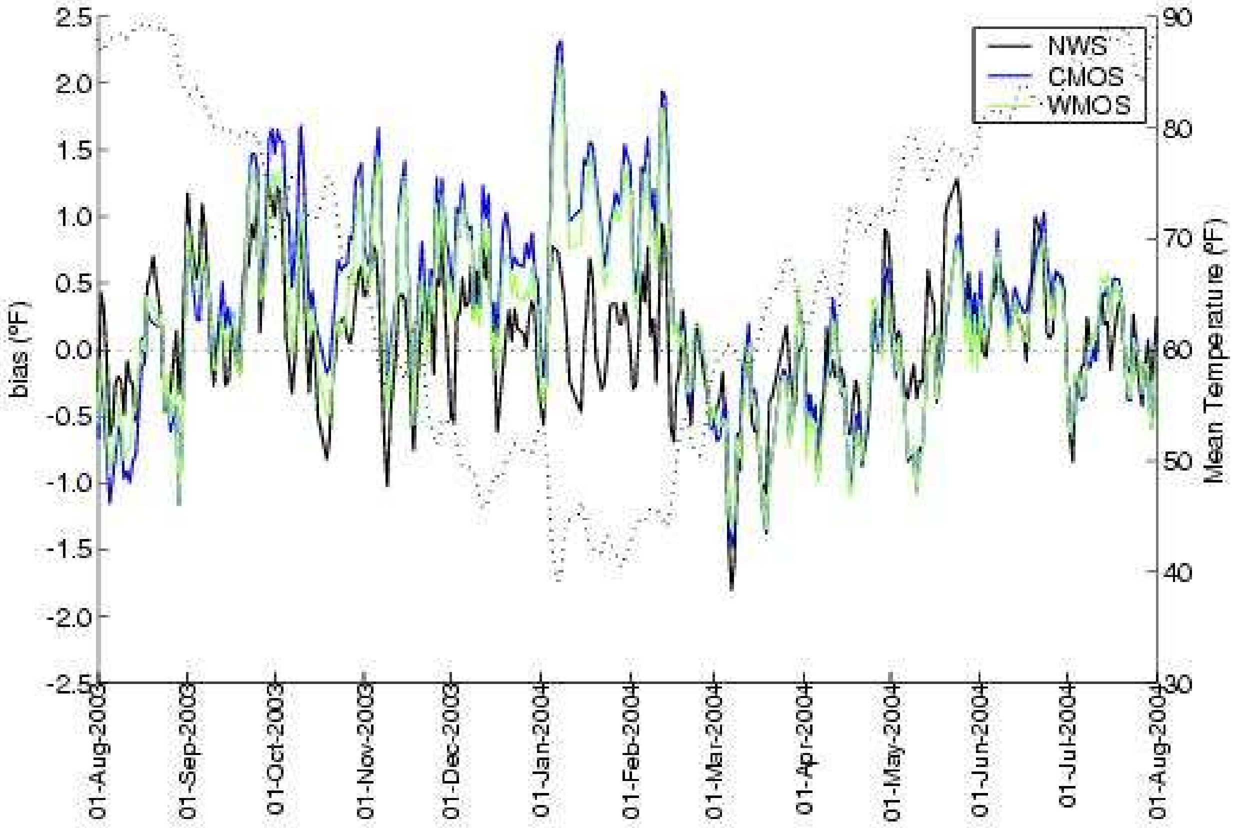

Fig. 10. Time series of bias in MAX-T for period one for all stations, NWS, CMOS and WMOS forecasts, 1 August 2003 – 1 August 2004. Mean temperature over all stations is shown with a dotted line. 3-day smoothing is performed on the data.

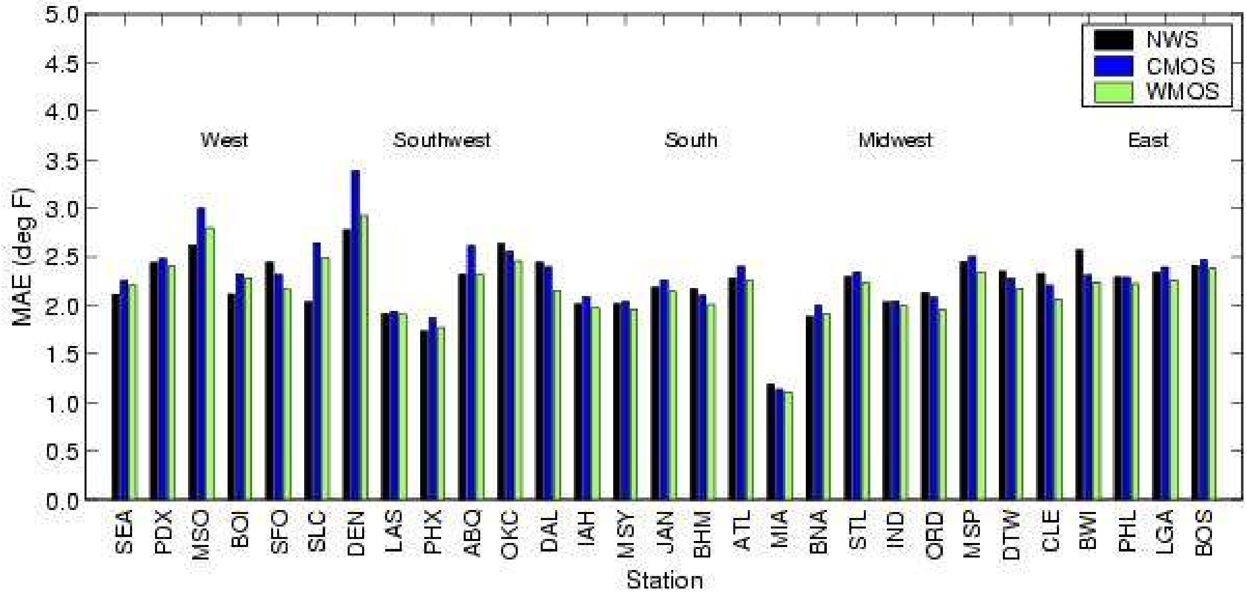

Fig. 11. MAE, MAX-T period 1 for all stations, NWS, CMOS and WMOS forecasts, 1 August 2003 – 1 August 2004, sorted by geographic region.

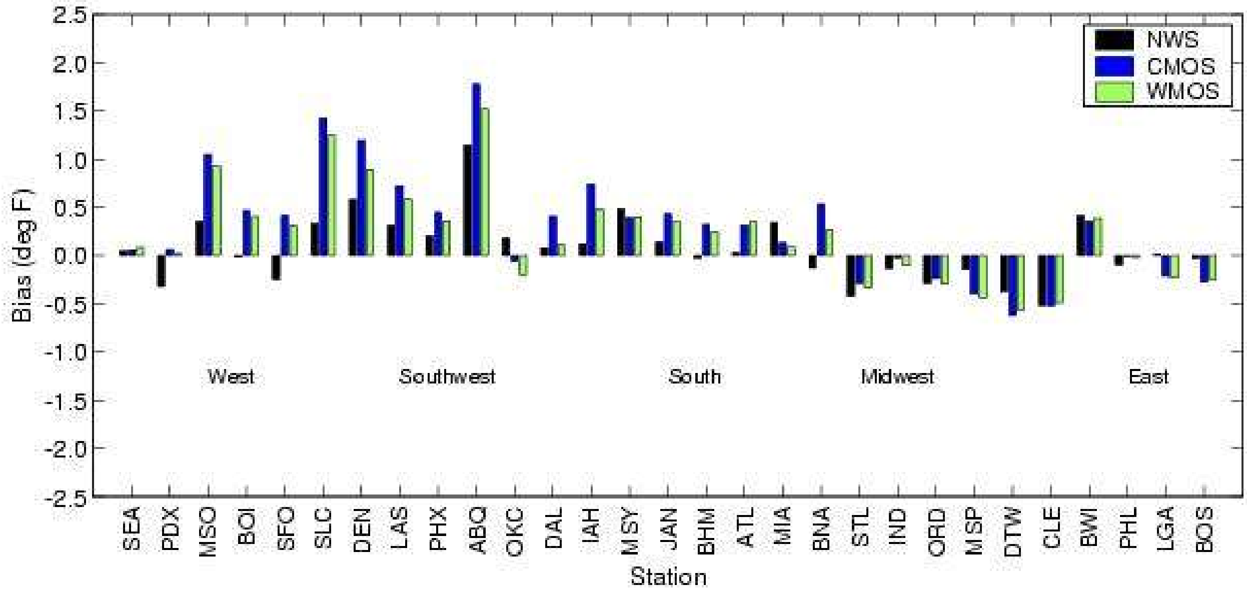

Fig. 12. Bias, MAX-T period 1 for all stations, NWS, CMOS and WMOS forecasts, 1 August 2003 – 1 August 2004, sorted by geographic region.

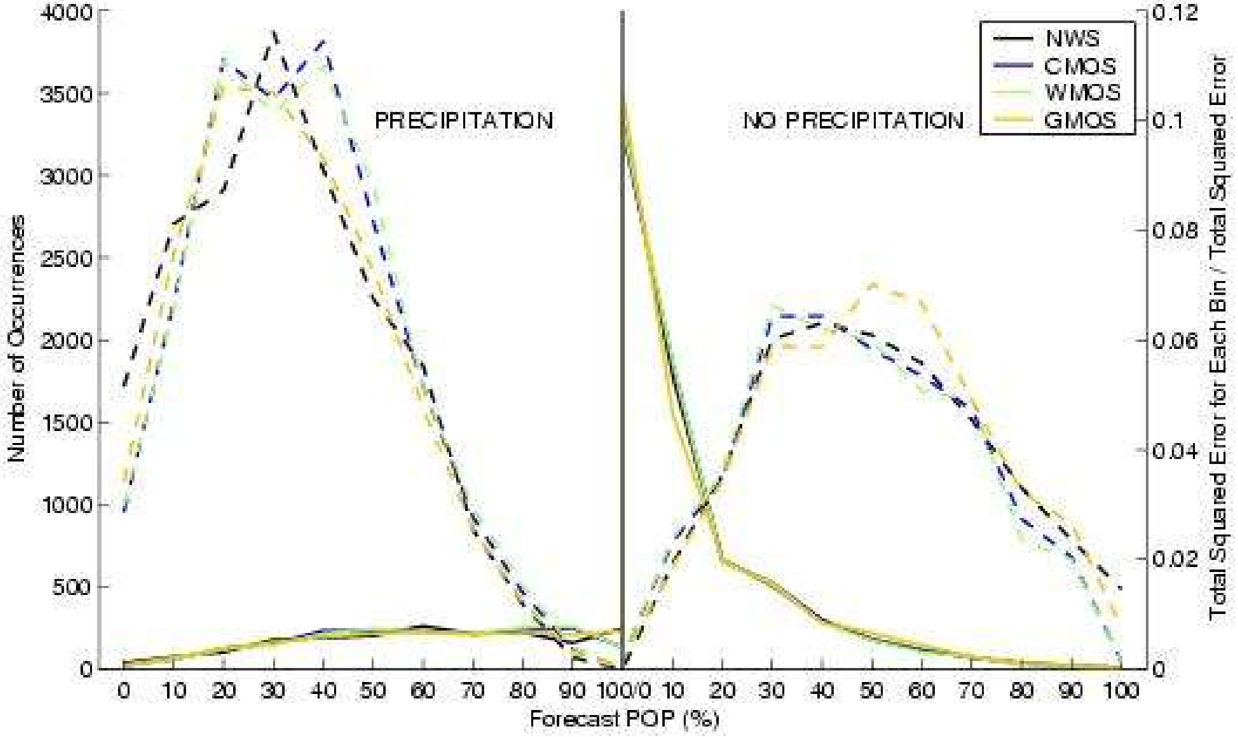

Fig. 13. Number of occurrences of various forecast probabilities for precipitation and non-precipitation events (solid lines) as well as normalized, squaredprecipitation error (dashed lines) for all stations, NWS, CMOS, WMOS and GMOS forecasts, all periods, 1 August 2003 – 1 August 2004. Bins are 10% in size, with data plotted in the center of each bin. Cases with observed "precipitation" and "no precipitation" are found on the left and right sides of the plot, respectively.

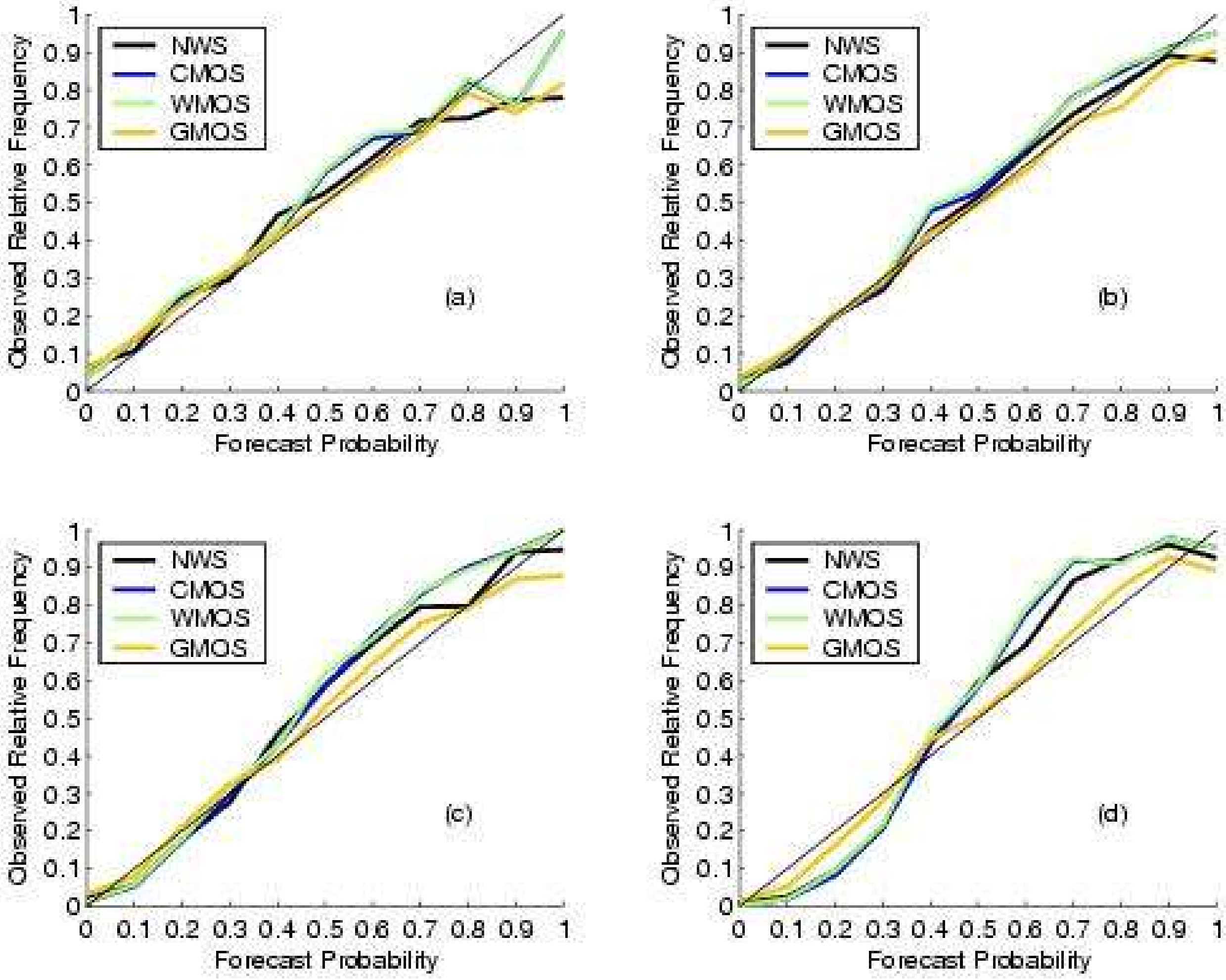

Fig. 14. Reliability diagrams for period 1 (a), period 2 (b), period 3 (c) and period 4 (d) for NWS, CMOS, WMOS and GMOS forecasts.

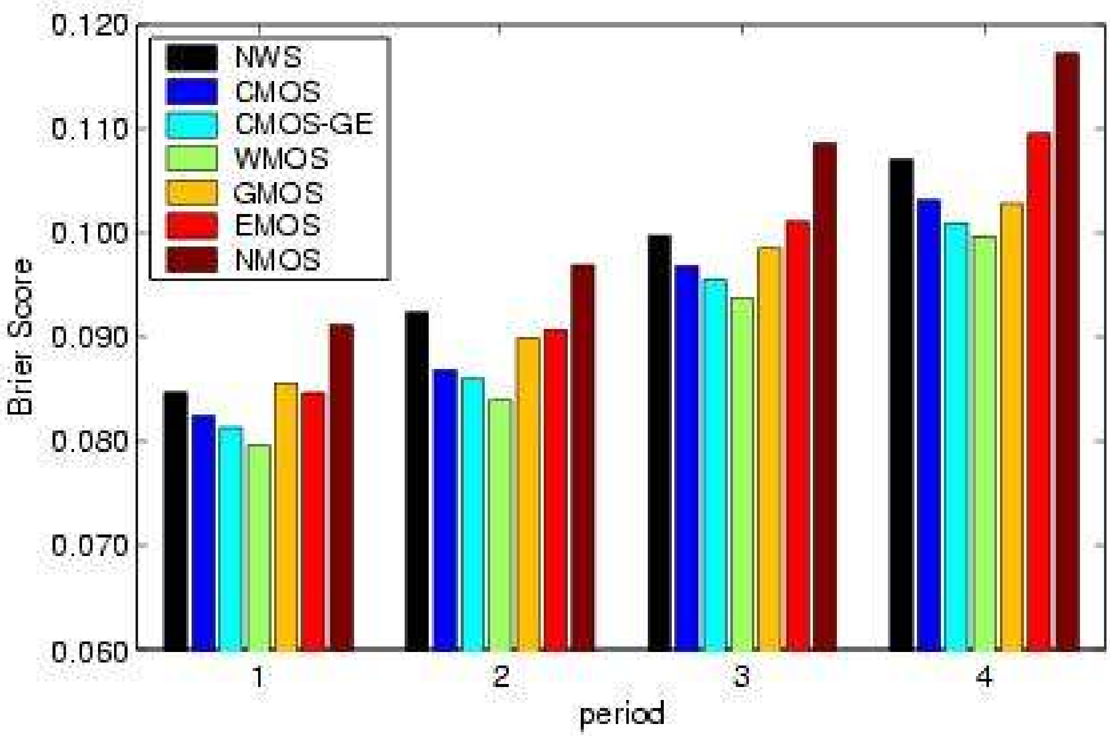

Fig. 15. Brier Scores for the seven forecast methods for all stations, 1 August 2003 – 1 August 2004.

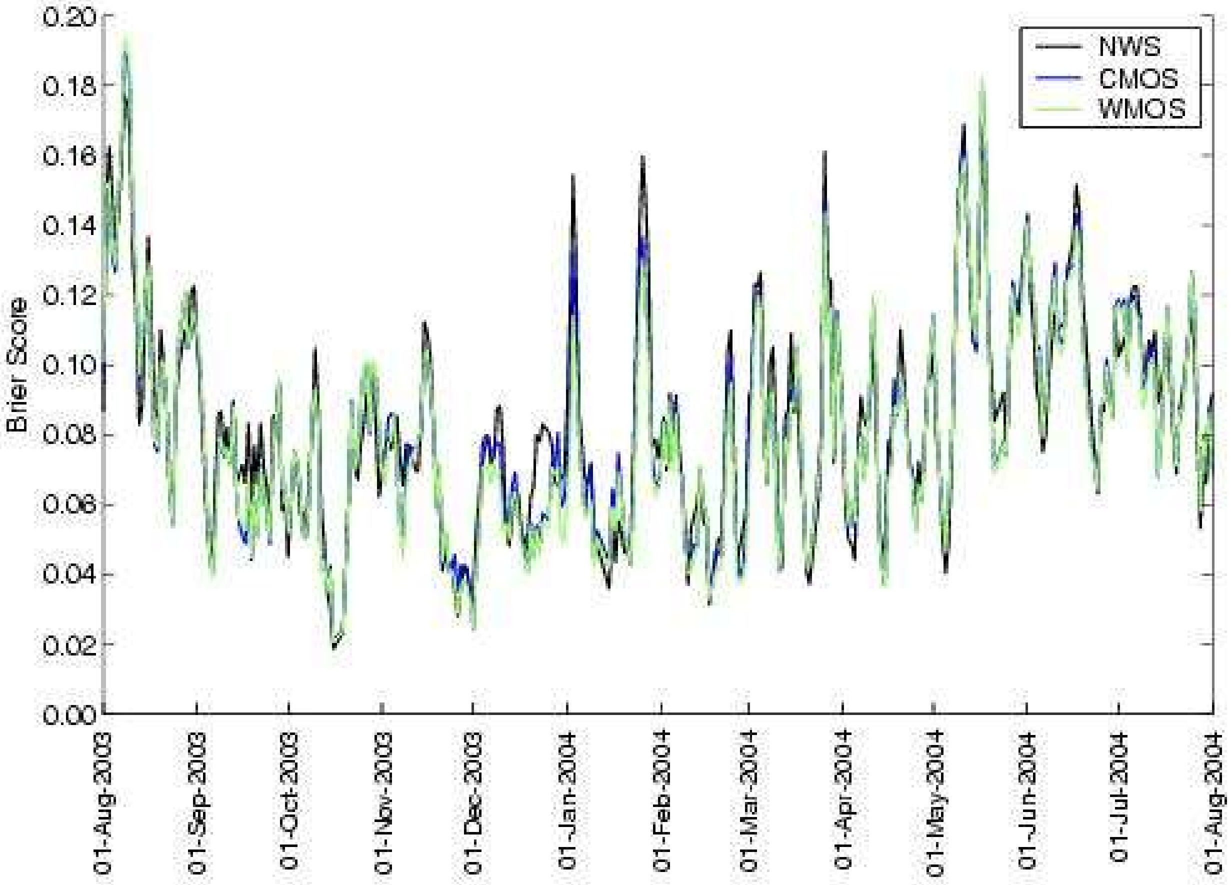

Fig. 16. Brier Score for all stations, NWS, CMOS and WMOS forecasts, 1 August 2003 – 1 August 2004. 3-day smoothing is performed on the data.

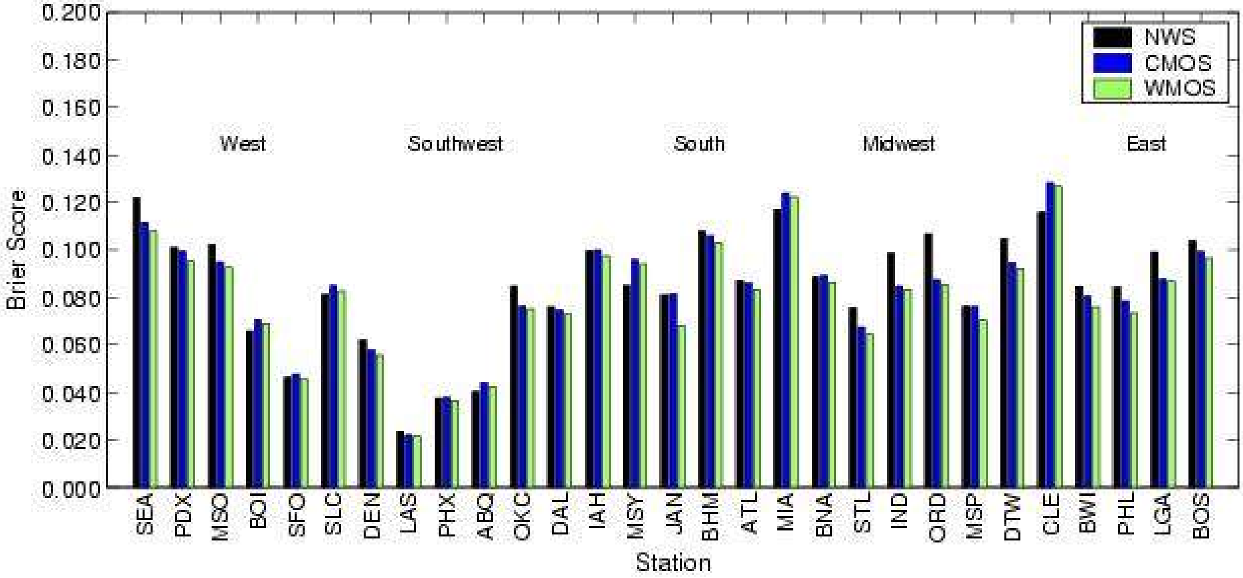

Fig. 17. Brier Score for all stations, NWS, CMOS and WMOS forecasts, 1 August 2003 – 1 August 2004, sorted by geographic region.