Sean A. Hribal

National Weather Center Research Experiences for Undergraduates
California University of Pennsylvania

RR1 Box 624 Scottdale, PA 15683
(724) 887- 6616
Hri9027@cup.edu

Stephen M. Leyton
Center for Analysis and Prediction of Storms
University of Oklahoma

More Accurate MOS Temperature Forecasts Using Bias Correction and Consensus

SEAN A. HRIBAL

National Weather Center Research Experiences for Undergraduates, Norman, Oklahoma
California University of Pennsylvania, California, Pennsylvania

5 August 2005

ABSTRACT

Bias correction and consensus have been applied to MOS temperature forecasts in effort to increase accuracy. Temperature forecasts obtained for KOKC and KPIT from 1 May 2002 to 30 January 2005 were classified according to maximum and minimum, projection, initialization, and season to examine model behavior between these divisions. All forecasts were verified using RMSE.

Compared to the uncorrected individual model forecasts, a seasonal bias correction showed a slight increase in RMSE values. A lagged bias correction decreased RMSE by approximately 0.1°F to 0.5°F compared to the RMSE for the uncorrected forecasts. An equally weighted consensus decreased RMSE by about 0.5°F to 1.0°F and 0.1°F to 0.5°F for maximum and minimum temperatures, respectively, over the uncorrected individual model forecasts. This method improved upon the lagged bias correction of individual model forecasts by several tenths of a degree. A linear regression consensus performed slightly worse than the equally weighted consensus. An unequally weighted consensus method based on lagged variance was the most accurate of all forecast enhancement methods, decreasing RMSE values by approximately 0.5°F to 1.5°F compared to uncorrected individual model forecasts and by several tenths of a degree over the lagged bias corrected individual model forecasts. Thus, based on the methods examined in this study, it is shown that a model consensus using a lagged correction based on past performance will provide the most significant MOS temperature forecast improvement.

1. Introduction

Consensus forecasts of Model Output Statistics (MOS) products have consistently demonstrated increased skill over their individual forecast counterparts (e.g., Vislocky and Fritsch 1995; Brooks and Doswell 1996). In these instances, the addition of different forecasts provided more information to the consensus, thereby decreasing the influence of individual discrepancies (Vislocky and Fritsch 1995). The incorporation of bias correction has also been shown to improve consensus performance (Woodcock and Engel 2005). Therefore, this study aims to determine the best methods for correcting bias and weighting consensus temperature forecasts.

Two main methods are used to create consensus forecasts. An equally weighted consensus averages individual forecasts together, whereas an unequally weighted consensus assigns varying weights to individual forecasts based on measures of error. Vislocky and Fritsch (1995) found that an unequally weighted consensus shows an insignificant improvement over an averaged consensus. However, Baars and Mass (2004), using a more refined methodology for constructing an unequally weighted consensus, examined model performance by geographic distribution, time/season, and climatological departure. The latter approach resulted in a more substantial forecast improvement. Woodcock and Engel (2005) showed that consensus forecasts can gain additional accuracy from applying a running bias correction to individual forecasts before assimilation (2005).

The hypothesis of the present paper follows the results of more recent studies (Etherton 2004; Baars and Mass 2004; Woodcock and Engel 2005) that indicate a small but valuable increase in performance for unequally weighted consensus temperature forecasts. The adoption of a running bias correction is also expected to benefit individual models. The models examined in this study include Nested Grid Model (NGM), Global Forecast System (GFS), and North American Mesoscale (NAM) (formerly known as the Eta Model). More information on observational and model data is covered in section 2. Methods of forecast verification and consensus construction are explained in section 3. Results are provided in section 4 and further analyzed in section 5.

2. Data

Two datasets were used for this study. The first was an archive of MOS forecasts from the NGM, GFS, and NAM models. The length of each archive varied depending on the model, but approximately spanned from 2001 to 2005. The second was an archive of hourly surface weather observations that spanned from 1982 to 2005. For the purpose of this project, MOS data

was used that was present for all models (1 May 2002 – 30 January 2005) and the corresponding surface observations to those dates. Temperature forecasts were then extracted from the model datasets. Maximum and minimum temperature forecasts were acquired for Day 1 and Day 2 projections from 0000 and 1200 UTC model initializations.

To verify model forecasts, 6-hourly maximum and minimum temperatures were used as reported at the site. Daily maximum and minimum temperatures were determined by checking for the highest value at 1800, 0000, and 0600 UTC and the lowest value at 1200 and 1800 UTC, respectively. Although daily maximum and minimum temperatures can occur at other times, these are the times for which the MOS equations were constructed. If missing observations or model forecasts occurred, all data from the associated maximum or minimum time period were disregarded. In effort to decrease the influence of outliers in the dataset, standard deviation of model errors was used to establish a 95% confidence interval for model errors. The distribution of error was checked to ensure that the errors followed a normal distribution, making this process reasonable (an example is shown in Fig. 1). If any forecast was deemed an outlier, all forecasts and observations for that time were considered outliers and removed from later use.

This study focused on model performance at Oklahoma City, OK (KOKC) and Pittsburgh, PA (KPIT). The former was chosen simply due to proximity to the author. The latter was selected as a basis of comparison to a site considered meteorologically different than the initial location.

3. Methodology

Forecast verification determines the accuracy of forecasts. This process employs measures that can be applied to compare forecasts and discern forecast characteristics. The present study used mean error (ME, also known as “bias”), mean square error (MSE), and root mean square

error (RMSE) for this purpose. No one of these measures can sufficiently represent all aspects of forecast accuracy. Therefore, it was important to use different measures with different strengths.

RMSE was the measure of choice for forecast verification purposes. This measure provided the best compromise between MAE and MSE. Because RMSE squares values during the summation, consistency is rewarded more than MAE (Brooks and Doswell 1996). Since the square root is applied to the average summation, values are less sensitive to forecasts errors and therefore less exaggerated than MSE.

This study is concerned with improving the MOS forecasts of individual models and then examining ways to create consensus forecasts. For individual forecasts, the RMSE was ultimately compared between the forecasts with no bias correction, the forecasts with a seasonal bias correction, and the forecasts with a running lagged bias correction. For consensus forecasts, an equal weighting was applied to each of the three forecast models on a daily basis, as well as an unequal weighting of the three forecast models, by using linear regression and a lagged weighting scheme described by Etherton (2004), which is discussed in more detail below.

For each of the procedures mentioned above, each combination of season (warm – April to September; cool – October to March), model initialization time (0000 UTC, 1200 UTC) and temperature forecast (Day 1 maximum and minimum; Day 2 maximum and minimum) was treated separately. Cross-validation was utilized to test procedures on multiple independent data sets. In this process, the data was separated into 3 distinct sets. Then, two of these sets were combined for the purpose of training data while the remaining set served as the validation data. This created three independent data sets; the results of which were then averaged.

a. Individual forecasts with no bias correction

In this process, the RMSE was calculated for each forecast model over the entire period for each combination of season, initialization time and temperature forecast. Determining the RMSE for these uncorrected individual forecasts was important for providing a basis of comparison for the bias-corrected individual forecasts and the consensus forecasts.

b. Individual forecasts with seasonal bias correction

To establish a seasonal bias correction for individual model forecasts, the ME (i.e., bias) was calculated over the entire season for each combination of model, initialization time and temperature forecast in the two-year training data set. Next, using the validation data set, each calculated seasonal ME was then subtracted from the respective forecast. The RMSE was then calculated for these seasonally bias-corrected forecasts.

c. Individual forecasts with lagged bias correction

A running lagged bias correction was also applied to individual model forecasts. For a given forecast date, this method calculates the ME for previous forecasts over a predetermined length of time (i.e., lag period). The bias over this lagged period is then subtracted from the given forecast. This lagged bias correction was performed for each forecast in the training data set. In order to determine the optimal lagged period for each combination of model, season, initialization time and temperature forecast, the performance of lag periods was spanning from 1 to 30 days was examined. As before, the RMSE was calculated for the bias-corrected forecasts.

d. Consensus forecasts with equal weighting

To establish a baseline of comparison for more sophisticated consensus forecast techniques, consensus forecasts in which each individual model was equally weighted were first examined. Therefore, the NGM, GFS, and NAM forecasts were simply averaged together. In a process

similar to that of the individual forecasts with no correction, the RMSE was simply calculated for forecasts over the entire period.

e. Consensus forecasts using linear regression

It was assumed that a consensus forecast using equal weighting of each forecast model would not create the optimal consensus forecast. Therefore, consideration was made for the means by which each individual model could be utilized, but allow for unequal weighting of these models toward the resulting consensus forecast. One method considered for this purpose was multiple linear regression. In this procedure, a predictand, Y, is fit by a linear combination of a set of predictors, X_i, was then applied in the following manner:

$$Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

in which "~" represents a numerically modeled fit using a set of coefficients, $ß_i$ (for i = 1, …, p). In this study, our predictand is the observed temperature while the predictors are the temperature forecasts by each model. The value of each coefficient (in addition to an equation intercept, $ß_0$) is computed so as to minimize the squared sum of the residuals. Residuals are the errors, or differences, that naturally arise between the modeled values of the predictand and its true, or observed, values.

The regression equations were derived from the training set and then tested upon the validation set. The RMSE was calculated for the forecasts over the entire period.

f. Consensus forecasts using a lagged weighting scheme

Minimum variance error was also used to weight individual model forecasts, using a technique described by Etherton (2004). This method employs the following series of equations to assign optimal weights, designated by "w". The variables "a," "b," and "c," represent each model and "σ^2" represents MSE.

$$w_a = (\sigma^2_b/\sigma^2_c)/(\sigma^2_a\sigma^2_c+\sigma^2_a\sigma^2_c+\sigma^2_b\sigma^2_c)$$
$$w_b = (\sigma^2_a/\sigma^2_c)/(\sigma^2_a\sigma^2_c+\sigma_a^2\sigma^2_c+\sigma^2_b\sigma^2_c)$$
$$w_c = (\sigma^2_a/\sigma^2_b)/(\sigma^2_a\sigma^2_c+\sigma^2_a\sigma^2_c+\sigma^2_b\sigma^2_c)$$

These equations produce weights ranging from 0 to 1, with the value of the weight increasing inversely to MSE. Therefore, the more accurate the model is, the higher the weight it receives. This gives greater influence to models that perform best and still allows the addition of beneficial information from weaker models. To create the consensus forecast, we then combine the weights and forecasts in the following manner:

$$\text{forecast}_{\text{consensus}} = w_a*\text{forecast}_a + w_b*\text{forecast}_b + w_c*\text{forecast}_c$$

The MSE used in these equations is obtained from a lag period. In a method similar to that applied to the individual forecasts with a running lagged bias correction, an examination was made for the performance of lag periods spanning from 1 to 30 days in order to determine the optimal lagged period for each combination of season, initialization time and temperature forecast. The RMSE was then calculated for these consensus forecasts.

4. Results

The key to making fair comparisons between the different individual model methods and consensus methods is using the same measure of forecast performance (e.g., RMSE) and classifications for verification. As discussed previously, we treated each combination of season (warm – April to September; cool – October to March), model initialization time (0000 UTC, 1200 UTC) and temperature forecast (Day 1 maximum and minimum; Day 2 maximum and minimum) separately. Thus, one can ascertain individual strengths and weaknesses of each model in the various procedures. For example, one model may have the smallest RMSE for warm season, 0000 UTC forecasts of day 1 maximum temperature but may have the greatest RMSE for warm season, 0000 UTC forecasts of day 2 maximum temperature.

a. Individual forecasts with no bias correction

1) KOKC

The dominant model for KOKC was GFS, outperforming the others forecasts for all classifications except for NGM Day 1 maximum temperature. GFS RMSE ranged from 2.91°F to 5.49°F. Minimum temperatures were typically more accurate than maximum temperatures for this model. NGM RMSE ranged from 3.2°F to 7.1°F and NAM ranged from 3.29°F to 6.39°F. NGM typically performed better for maximum temperatures and NAM for minimum temperatures (Table 1).

2) KPIT

NAM gained more accuracy for KPIT warm season temperatures at 0000 and 1200 UTC. However, GFS maintained dominance for the other classifications. All models showed narrower RMSE ranges for KPIT of 3.04°F to 5.18°F, 2.78°F to 4.57°F, and 2.75°F to 4.96°F for NGM, GFS, and NAM, respectively. NAM tended to provide more accurate forecasts than NGM, especially for the warm season (Table 1).

b. Individual forecasts with seasonal bias correction

1) KOKC

The GFS model showed better performance for most KOKC forecasts, excluding NGM maximum temperature forecasts for the warm season 0000 UTC initialization. RMSE values ranged from 3.3°F to 7.66°F, 2.96°F to 6.03°F, and 3.58°F to 6.6°F for NGM, GFS, and NAM, respectively. No distinct favor appears to exist between NGM and NAM for classifications (Table 2).

2) KPIT

For this station, the GFS was more accurate for every classification. RMSE ranges for all models were again smaller, with minimum and maximum values of approximately 3°F and 5°F, respectively. As indicated by the unbiased results for KPIT, the NAM model outperformed the GFS by a few tenths of a degree in most situations (Table 2).

c. Individual forecasts with lagged bias correction

1) KOKC

The GFS model showed lower RMSE values for a majority of KOKC forecasts but NAM produced lower RMSE values for certain warm season classifications. The RMSE range for GFS was smaller than other models at 2.88°F to 5.41°F. A pattern is also evident such that 0000 UTC forecasts were more accurate in the warm season and 12000 UTC were more accurate in the cool season (Table 3).

The previous forecasts assessment was made using the best forecasts for each classification. The lag period from which the best forecasts were derived, varied between most classifications. However as lag periods approached 30 days, their accuracy typically increased as indicated by the parabolic-shaped frequency curve on Fig. 2.

2) KPIT

The GFS provided the best maximum temperature forecast in all situations for KPIT. However, NAM dominated for the Day 2 minimum temperature. The ranges for all models were smaller and more consistent than those of KOKC. Minimum RMSE values for all models occurred for the Day 1 maximum temperature. Minimum RMSE values all occurred for the Day 2 maximum temperature. 0000 UTC forecasts were most accurate in the warm season and 1200 UTC forecasts were most accurate in the cool season (Table 3).

The plot of optimal lag periods for KPIT resembled the plot for KOKC. However, KPIT frequency was typically greater for lagged periods of less than 30 days. This suggests that the optimal lag period for KOKC was around 30 days and 25-30 days for KPIT (Fig. 2).

d. Consensus forecasts with equal weighting

1) KOKC

For an equally weighted consensus forecast for KOKC, the most accurate forecast occurred for Day 1 maximum temperature and the least accurate forecast occurred for Day 2 maximum temperatures with respective RMSE values of 2.64°F and 5.39°F. 0000 UTC forecasts provided the most accurate maximum temperature forecast, while 1200 UTC forecasts provided the most accurate minimum temperature forecast (Table 4).

2) KPIT

RMSE values for KPIT tended to be a few tenths of a degree lower than those for KOKC. Similarly, forecast accuracy for KPIT behaved in the same manner between classifications as for KOKC (Table 4).

e. Consensus forecasts using linear regression

1) KOKC

The lowest RMSE for KOKC was 3.02°F for Day 1 minimum temperature. The highest RMSE was 5.84°F for Day 2 maximum temperature. 0000 UTC forecasts were always more accurate for maximum temperatures and 1200 UTC forecasts were the most accurate for minimum temperatures. Warm season forecasts were more accurate than cool season forecasts for all classifications by about a 1°F to 2°F (Table 5).

2) KPIT

KPIT forecasts were overall more accurate by KOKC forecasts by a few tenths of a degree. Forecast characteristics were similar to KOKC for model initialization and season (Table 5).

f. Consensus forecasts using a lagged weighting scheme

1) KOKC

RMSE values for KOKC ranged from 2.49°F to 5.09°F. The former value occurred for Day 1 maximum temperature and the latter for Day 2 maximum temperature. 0000 UTC forecasts provided the lowest RMSE values for maximum temperatures. 1200 UTC forecasts were most accurate for minimum temperatures (Table 6).

The best forecasts for this station occurred most often for the 30 day period. Days 4 to 13 also provided the best forecasts for certain classifications. However, the frequency was low enough for these periods that they were regarded as insignificant (Fig. 3).

2) KPIT

KPIT showed less forecast accuracy consistency between classifications compared to KOKC. The RMSE range was significantly lower with values ranging from 2.43°F to 3.41°F. Overall, KPIT forecasts were more accurate, especially for 1200 UTC cool season forecasts (Table 6).

The optimal lag periods for KPIT appear to precede KOKC by a few days. This places the best KPIT forecasts in the range of 28 to 30 and KOKC at 30 days. This slight change for optimal lag periods suggests that different lag periods are needed to obtain the best forecasts for different locations.

5. Discussion

a. Individual model forecasts

Since an independent assessment of MOS forecast enhancement methods has been established, the next step is to compare these methods to determine which ones provide the

greatest improvement. After obtaining and verifying individual MOS forecasts, a seasonal bias correction was applied. Unexpectedly, RMSE values of seasonally biased forecasts were typically higher than unbiased forecasts by a few tenths of a degree. This decrease in accuracy for the seasonal bias correction did not differ greatly between classifications. This increase in RMSE suggests that while the models do have forecast errors, the bias is not consistently warm or cold long periods of time. It is suspected that this seasonal bias correction may be over sampling the data, resulting in an inability to ascertain any further predictive information.

A lagged bias method was also applied to individual model forecasts. In this, the bias was calculated over a lagged period of time and then applied to the current forecast. The best lagged bias forecasts did not did differ significantly between classifications, which is consistent with the results of the uncorrected and seasonally biased corrected forecasts. The increase in accuracy for lagged bias forecasts over unbiased forecasts was similar to that of the uncorrected forecasts over seasonally biased corrected forecasts; typically, a few tenths of a degree. Therefore, we deduce that by sampling the model performance over shorter periods of time prior to the forecast, we can obtain valuable predictive information. The fact that the optimal lagged period was approximately 30 days for both locations furthers support our reasoning for the shortcomings of the seasonal bias correction.

Overall, the results of these tests on the individual model forecasts are encouraging. We have shown that the MOS temperature forecasts produced for all models do result in errors but that there are ways in which to reduce that error over long periods of time. In fact, by examining the bias over short periods of time prior to the actual forecast, we can reduce the RMSE over an extended period of time.

b. Consensus forecasts

In addition to improving the individual model forecasts, we were also interested in developing consensus forecasts and comparing them to the individual model forecasts. In doing so, we examined both equal weighting of the individual forecasts as well as two methods of unequal weighting.

As expected, the equally weighted consensus of the NGM, GFS, and NAM models was typically more accurate than any individual model. The extent of improvement over the uncorrected individual models was more pronounced for maximum temperatures for both stations. Overall, the equal weighted model consensus decreased RMSE by about 0.5°F to 1.0°F for maximum temperatures and about 0.1°F to 0.5°F for minimum temperatures when compared to the best individual model results. This simple consensus also showed a general improvement over individual models with lagged bias correction by a few tenths of a degree. It is important to note that while the weighting of the individual model forecasts may be sub-optimal; this simple consensus outperformed the individual model methods tested.

One method of creating unequal weighting of the models is to use multiple linear regression to minimize the variance of the forecasts. Surprisingly, the linear regression with unequal weighting performed worse than the equal weighting. Moreover, the performance of the linear regression models was no better than that of the best uncorrected individual models. The reason for this has not been determined yet, though it is suspected to be related to the over sampling that weakened the seasonal bias correction of the individual model forecasts.

Another method of determining unequal weights of the individual models is to simultaneously calculate the weights based on lagged variance of errors, as described by Etherton (2004). This consensus method consistently showed improvement over the equally weighted consensus of a few tenths of a degree for all classifications. When compared to the best individual uncorrected

forecasts, this method showed a significant improvement; about 0.5°F to 1.5°F. It is interesting to note that in using this method, the RMSE values decreased for maximum temperature forecasts slightly more than for minimum temperature forecasts.

With all of these results in mind, we have answered our two original questions. Individual MOS temperature forecasts can be improved simply by considering past performance of approximately 30 days. Moreover, consensus forecasts that utilize past performance to calculate individual model weighting perform better in the long run than individual model forecasts. However, it is important to note that not all consensus forecasts will necessarily improve upon the individual model forecasts.

c. Future considerations

While we have shown that processes exist to improve the performance of MOS temperature forecasts, there is still plenty of room for further enhancement of the ideas discussed above. Perhaps the linear regression consensus would perform much better if more predictors were used than simply the individual temperature forecasts. Now knowing the value of the lagged model performance, past bias could be used to improve this model. Another possibility is to combine the methods described above. For example, the forecasts using the Etherton (2004) could be refined further through a second bias correction procedure. Lag periods beyond 30 days could also examined to determine if these longer periods, shorter than a season, can produce more accurate forecasts. The methods used in this study could also be used to assess extreme climatological departures or various meteorological situations. For verifying forecast methods, a distributions-oriented approach may also be useful for a more in-depth look at forecast characteristics.

Aside from improving the forecasts themselves, another important step would be to examine the performance of these methods at additional locations. Due to the time constraints of this project, only two stations (KOKC and KPIT) were examined. They were chosen due to the author's familiarity as well as them being meteorologically diverse. However, it would be beneficial to test other meteorologically different sites as well as locations that are meteorologically similar.

Acknowledgments: The Research Experience for Undergraduates (REU) at the University of Oklahoma is supported by National Science Foundation Grant 0097651. Thanks to Steve Leyton for his guidance and to all those involved with the National Weather Center REU program for their support. NCAR and MDL are also thanked for the datasets utilized in this study.

6. References

Baars, J. A., and C. F. Mass, 2004: Performance of national weather service forecasts compared to operational, consensus, and weighted model output statistics. Preprints, 21st Conference on Weather and Forecasting / 17th Conference on Numerical Weather Prediction. American Meteorological Society, 4 pp.

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.

Etherton, B. J., 2004: Model consensus and ensemble weighting for spot forecasts. Preprints, 17th Conference on Probability and Statistics in the Atmospheric Sciences. American Meteorological Society, 4 pp.

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

Woodcock F., and E.Chermelle, 2005: Operational consensus forecasts. Wea. Forecasting, **20**, 101-111.

Station Season	Run (UTC)	Day 1 Max			Day 1 Min			Day 2 Max			Day 2 Min		
		NGM	GFS	NAM	NGM	GFS	NAM	NGM	GFS	NAM	NGM	GFS	NAM
KOKC													
Warm	0	**3.20**	3.25	3.74	3.60	**2.91**	3.29	4.15	**3.98**	4.78	3.83	**3.13**	3.41
Cool	0	3.88	**3.72**	4.36	4.44	**3.88**	4.13	5.87	**4.96**	5.87	5.80	**4.85**	5.13
Warm	12	3.81	**3.67**	4.75	3.35	**2.83**	3.23	4.43	**4.22**	4.63	3.72	**3.14**	3.40
Cool	12	5.10	**4.57**	5.25	4.43	**3.65**	4.09	7.10	**5.49**	6.39	5.42	**4.32**	4.43
KPIT													
Warm	0	3.15	2.78	**2.75**	3.04	2.84	**2.82**	3.83	3.42	**3.37**	4.03	**3.40**	3.61
Cool	0	3.53	**3.24**	3.70	4.28	**3.68**	3.74	4.68	**4.01**	4.78	5.47	**4.29**	4.88
Warm	12	3.57	**3.10**	3.24	2.79	**2.63**	2.72	4.30	**3.63**	3.83	3.27	3.16	**3.10**
Cool	12	4.13	**3.54**	4.61	4.02	**3.51**	3.65	5.18	**4.57**	4.96	4.60	**3.94**	3.97

TABLE 1. Average RMSE (°F) results for individual model forecasts with no correction. Bold values indicate the most accurate models (i.e., the smallest RMSE)

Station Season	Run (UTC)	Day 1 Max			Day 1 Min			Day 2 Max			Day 2 Min		
		NGM	GFS	NAM	NGM	GFS	NAM	NGM	GFS	NAM	NGM	GFS	NAM
KOKC													
Warm	0	**3.30**	3.37	3.85	3.59	**2.96**	3.58	**4.27**	4.31	4.77	3.78	**3.21**	3.69
Cool	0	4.29	**3.99**	4.49	4.71	**4.16**	4.38	6.46	**5.52**	6.18	6.09	**5.12**	5.46
Warm	12	4.06	**3.79**	4.67	3.39	**2.86**	3.44	4.60	**4.47**	4.75	3.68	**3.19**	3.68
Cool	12	5.67	**4.95**	5.45	4.54	**3.84**	4.29	7.46	**6.03**	6.60	5.63	**4.67**	4.76
KPIT													
Warm	0	3.23	**2.96**	3.02	3.19	**2.90**	2.97	4.01	**3.66**	3.90	4.12	**3.48**	3.83
Cool	0	3.68	**3.37**	3.56	4.33	**3.70**	3.80	4.76	**4.17**	4.76	5.65	**4.28**	5.05
Warm	12	3.81	**3.33**	3.50	2.91	**2.69**	2.80	4.44	**3.91**	4.15	3.49	**3.20**	3.24
Cool	12	4.17	**3.83**	4.41	4.05	**3.53**	3.71	5.30	**4.72**	4.81	4.69	**3.98**	4.01

TABLE 2. Average RMSE (°F) results for individual models forecasts with seasonal bias correction. Bold values indicate the most accurate models (i.e., the smallest RMSE)

.

Station Season	Run (UTC)	Day 1 Max			Day 1 Min			Day 2 Max			Day 2 Min		
		NGM	GFS	NAM	NGM	GFS	NAM	NGM	GFS	NAM	NGM	GFS	NAM
KOKC													
Warm	0	3.15	2.90	**2.80**	3.38	**2.88**	2.94	4.02	3.60	**3.41**	3.75	**3.14**	3.33
Cool	0	3.95	**3.56**	3.85	4.74	4.10	**4.09**	6.01	**4.56**	4.84	5.90	**5.06**	5.10
Warm	12	3.76	3.14	**2.26**	3.13	**2.79**	2.85	4.28	**3.73**	3.95	3.55	**3.12**	3.20
Cool	12	5.10	**4.32**	4.50	4.57	**3.83**	4.03	7.37	**5.41**	5.79	5.63	**4.51**	4.55
KPIT													
Warm	0	2.84	**2.52**	2.59	2.89	2.83	**2.72**	3.28	**2.93**	3.22	3.72	**3.36**	3.52
Cool	0	3.47	**3.20**	3,43	4.30	3.76	**3.70**	4.75	**3.95**	4.48	5.43	**4.21**	4.94
Warm	12	3.21	**2.71**	2.89	2.66	2.60	**2.59**	3.70	**3.16**	3.48	3.14	3.13	**3.01**
Cool	12	4.10	**3.54**	4.16	3.97	**3.56**	3.64	5.19	**4.39**	4.55	4.66	**3.95**	4.03

TABLE 3. Average RMSE (°F) results for individual model forecasts with lagged bias correction. Bold values indicate the most accurate models (i.e., the smallest RMSE)

Station Season	Run (UTC)	Day 1 Max	Day 1 Min	Day 2 Max	Day 2 Min
KOKC					
Warm	0	2.64	2.91	3.34	3.10
Cool	0	3.25	3.76	4.37	4.75
Warm	12	3.20	2.74	3.67	2.99
Cool	12	4.12	3.66	5.36	4.19
KPIT					
Warm	0	2.53	2.57	3.00	3.36
Cool	0	2.81	3.58	3.72	4.42
Warm	12	2.78	2.48	3.44	2.87
Cool	12	3.29	3.45	4.10	3.74

TABLE 4. Average RMSE (°F) results for consensus forecasts using equally-weighted models. Bold values indicate the most accurate models (i.e., the smallest RMSE)

Station Season	Run (UTC)	Day 1 Max	Day 1 Min	Day 2 Max	Day 2 Min
KOKC					
Warm	0	3.02	3.00	3.82	3.20
Cool	0	3.82	4.03	5.46	4.86
Warm	12	3.55	2.85	4.10	3.16
Cool	12	4.89	3.79	5.84	4.47
KPIT					
Warm	0	2.81	2.77	3.55	3.45
Cool	0	2.97	3.55	3.89	4.33
Warm	12	3.16	2.57	3.87	3.03
Cool	12	3.51	3.48	4.12	3.81

TABLE 5. Average RMSE (°F) results for consensus forecasts using linear regression. Bold values indicate the most accurate models (i.e., the smallest RMSE)

Station Season	Run (UTC)	Day 1 Max	Day 1 Min	Day 2 Max	Day 2 Min
KOKC					
Warm	0	2.49	2.78	3.18	2.96
Cool	0	3.15	3.71	4.25	4.72
Warm	12	2.87	2.63	3.46	2.87
Cool	12	3.98	3.61	5.09	4.15
KPIT					
Warm	0	2.43	2.50	2.75	3.15
Cool	0	2.74	3.55	3.62	4.30
Warm	12	2.65	2.40	3.12	2.79
Cool	12	3.24	3.41	2.89	3.74

TABLE 6. Average RMSE (°F) results for consensus forecasts using the Etherton (2004) method. Bold values indicate the most accurate models (i.e., the smallest RMSE)

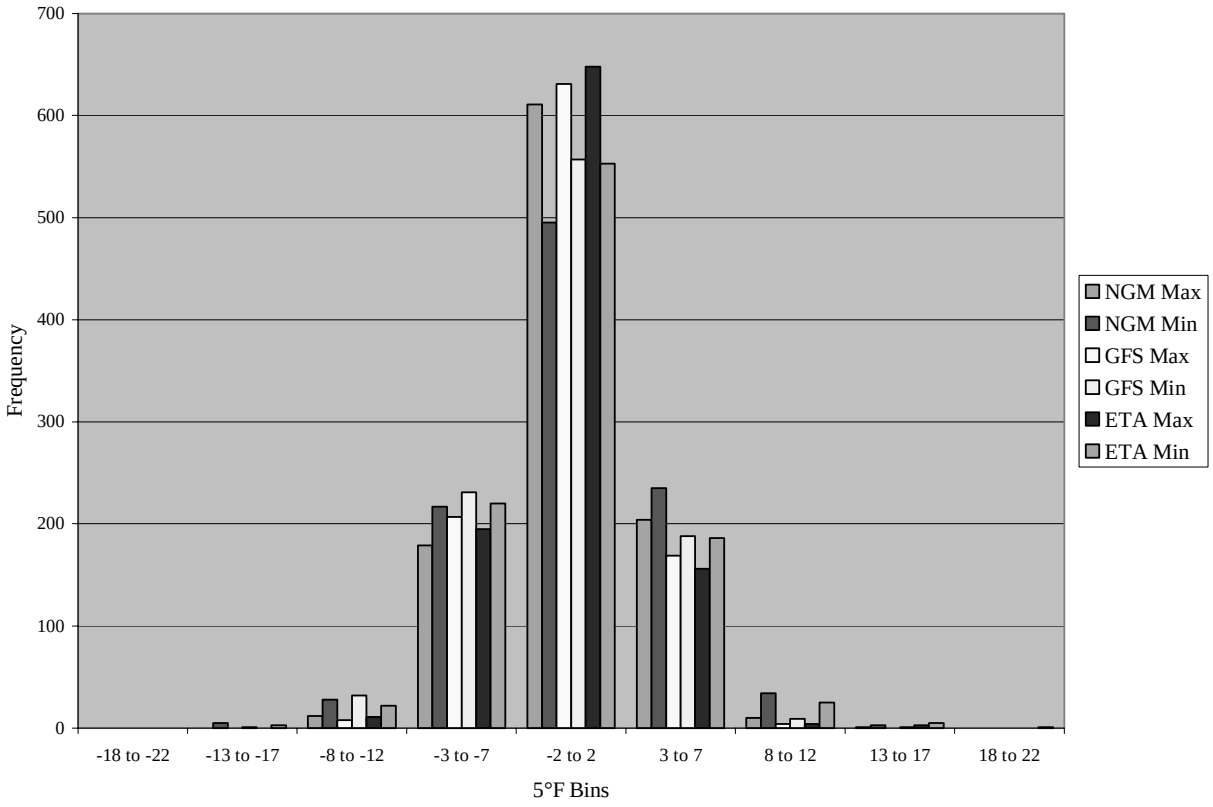

FIG. 1. Mean error distributions using 5°F bins for warm season KPIT model forecasts.

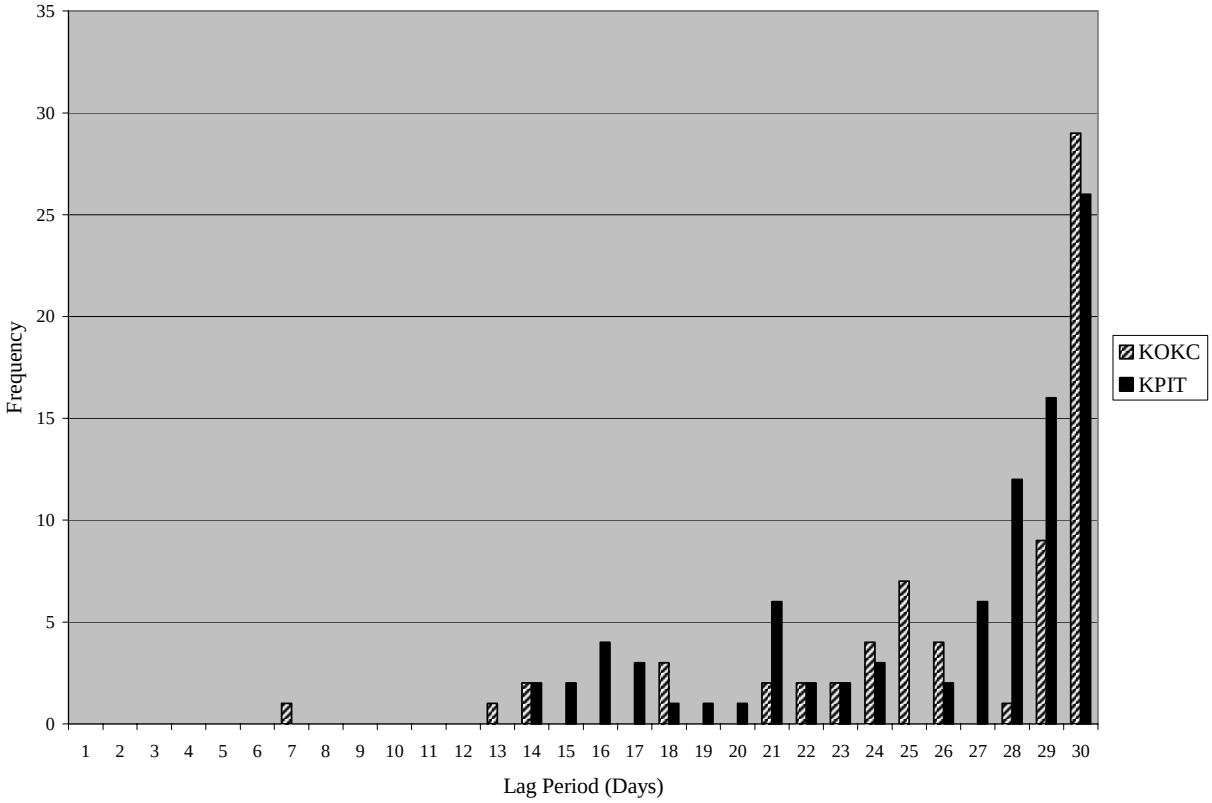

FIG. 2. Frequency of KOKC and KPIT optimal lagged periods for individual model forecasts.

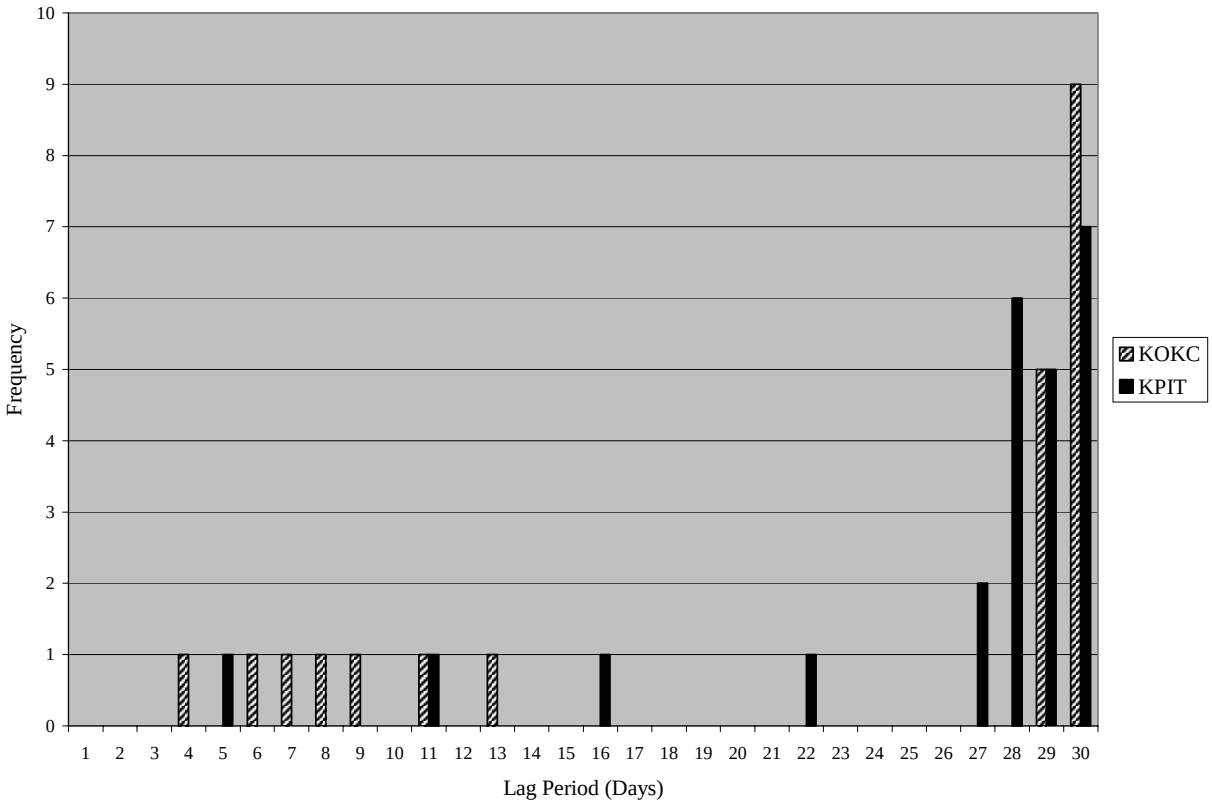

FIG. 3. Frequency of KOKC and KPIT optimal lagged period consensus forecasts, using the Etherton (2004) method.