USING ENSEMBLE PREDICTION SYSTEMS TO ESTIMATE MEDIUM RANGE TEMPERATURE FORECAST ERROR

Peter M. Finocchio[1], David Hogan[2]
[1]National Weather Center Research Experience for Undergraduates
University of Oklahoma, Norman, Oklahoma
And
University of Miami, Coral Gables, FL
[2]Atmospheric and Environmental Research, Inc. (AER), Lexington, MA

ABSTRACT

A salient benefit of an ensemble prediction system (EPS) is its ability to provide a means of estimating forecast error. This study tests the error prediction skill of three EPS features: spread (standard deviation) among ensemble members, consistency between MEX/MOS output and the ensemble mean, and consistency between consecutive 24-hour runs of an EPS. For 27 stations throughout the northeastern United States, 15-day high and low temperature forecasts from calibrated ECMWF ensemble output issued Feb 1 through May 31 2007 are examined. For each of the 15 forecast days, the significance (from r^2 statistic) and slope of the error predictor-forecast error relationship is used to determine how valuable each feature is in estimating forecast error.

For high temperature forecasts in the northeast, ensemble spread and run-to-run consistency are most effective at predicting forecast error through 9-day lead times. Although error prediction skill in both drops off in the longer ranges, ensemble spread is more useful beyond 9-day lead times. Spread and run-to-run consistency are more effective for low temperature forecast error prediction, with ensemble spread still performing best beyond 9-day lead times. Model-to-model consistency is only a moderate to weak error predictor for high temperature forecasts in the short range. For the sake of comparison of error predictability between two regions with disparate climates, data from 21 stations in the southwestern United States are also examined. In this region, all three error predictors are consistently effective in anticipating forecast uncertainty for both high and low temperatures.

1. INTRODUCTION

Numerical weather prediction in the medium and extended ranges requires consideration of the chaotic tendencies of the atmosphere. Lorenz (1963) showed the importance of chaos in non-periodic flow environments. As a result of chaotic systems in which a mere flap of a seagull's wing could alter weather events downstream, he surmised the limit of atmospheric predictability to be in the vicinity of 14 days (Lorenz 1963, Yoden 2007).

Ensemble forecasting in light of Lorenz's chaos theory becomes an increasingly important alternative to the common deterministic approach. An ensemble consists of a control forecast and an array of forecast possibilities generated from repeatedly perturbing the initial conditions and/or model physics of the control. The slight perturbation of initial conditions and/or model physics allows the model to better account for two main sources of rapidly growing forecast errors: 1) the natural error in atmospheric conditions read into the model and 2) the model's inherent error due to limits in resolution and complexity (Buizza 2000). Ensembles are probabilistic in that they provide a range of forecast possibilities that enables the user to obtain a probability of forecast verification. The probabilistic model is preferred in longer range forecasting because it provides an indication of forecast certainty unlike deterministic models (Buizza et al. 2004; Richardson 1999; Toth and Kalnay 1993). Leith (1974) has also shown that the "improvement in skill is appreciable for Monte Carlo (ensemble) forecasts as compared to conventional single forecasts" further emphasizing the benefits of ensembles. Most importantly, probabilistic forecasts have introduced the idea of predicting forecast error. By analyzing certain features in the ensemble, an estimate of forecast certainty for each lead time can be deduced. This study will focus on the effectiveness of three such features or forecast "error predictors."

The most well known and researched of the error predictors is ensemble spread. In this study, spread refers to the standard deviation of the ensemble members about the ensemble mean. Studies by Whitaker (1998) and Barker (1991) have shown correlations between ensemble spread and forecast error/skill through long range forecasts. These findings indicate that ensemble spread is a viable forecast error predictor. Forecast error sensitivity to two other possible error predictors, model-to-model and run-to-run consistency, will be examined. Although overlooked and often overshadowed by the

importance of ensemble spread in predicting forecast error, these two predictors could prove to be even more beneficial in the improved evaluation of ensemble forecasts. Model-to-model consistency refers to the level of agreement between the ensemble mean and other deterministic products. The purpose of testing inter-model consistency as an error predictor is to determine the possible value of a poor man's or grand ensemble. A group of deterministic products compose the members of a poor man's ensemble. Palmer et al. (2000), in their study of seasonal weather pattern predictability, found that a 9-member multi-model ensemble had higher forecast skill scores than the 9-member single-model ensemble. Discovering a relationship between inter-model variability and forecast error would improve the analysis and utilization of poor man's ensembles. Run-to-run consistency refers to the level of agreement between two consecutive ensemble runs for a common forecast time.

If forecast error is sensitive to any of the three error predictors and if the relationship is statistically significant, the predictor can be used to estimate temperature forecast error. Examining the nature of forecast error dependence upon the three predictors through a 15-day forecast establishes an ensemble evaluation "suite" that could improve the anticipation of error in medium range temperature forecasting.

Section 2 contains information on the ECMWF EPS which provides the raw dataset for this study, and a brief discussion of AER's *eCast*™ product. Section 3 outlines the methodology and analysis. Section 4 presents the results for the three error predictors being tested for high and low temperatures in the northeast and additionally in the southwest. Section 5 provides a discussion of the results and conclusions. Section 6 proposes a regional dependency of error prediction skill and possible new directions for research.

2. DATA

Raw data for this study is based upon the European Centre for Medium Range Weather Forecasting EPS. ECMWF EPS provides forecasts with lead times extending out to 15 days. The ensemble is populated from 50 perturbations of the control using singular vectors, which represent axes of maximum error growth in the atmosphere. ECMWF calculates 50 singular vectors in order to create the 50 perturbations along these axes (Buizza et al 2004). The larger ensemble is beneficial to this study as it has been proven to outperform smaller ensembles in medium to longer ranges forecasts (Atger 1999; Buizza et al. 2004). As of 2006, the EPS analyzes 62 vertical levels through the 15 days with horizontal grid resolution of 50 km (TL399 spectral resolution) for lead times of 0-10 days, reduced to 80km grid resolution (TL255 spectral resolution) for days 9-15 (ECMWF HTML User Guide).

Atmospheric and Environmental Research Inc. in Lexington, MA creates the *eCast*™ product as a user-friendly method of relaying ensemble information. *eCast*™ uses raw ECMWF EPS data and applies ensemble bias correction and calibration. Calibration of the ensemble is a spread amplification technique that is used to better simulate the actual spread in the forecast. It corrects the under-dispersive tendencies of an EPS. Data used in this study originates from a research and development version of *eCast*™ that was operating in February 2007.

eCast™ calibrated ECMWF 12 UTC high and low temperature forecasts are taken from Feb 1 through May 31, 2007 creating a dataset of 120 forecast validation days. Each validation day has 15 corresponding forecast days. Data from 27 stations across the northeastern United States is analyzed. The domain extends from Virginia to Maine, from 33.7˚N to 44.3˚N, 69.8˚W to 80˚W. Southwest data from 21 stations is later analyzed from California to Colorado, encompassing the area from 32˚N to 41˚N, 104.7˚W to 122.4˚W.

For inter-model comparisons, MOS/MEX 12 UTC high and low temperature forecast data is taken from Feb 1 through May 31, 2007 as well. Each of the 120 validation days has only seven corresponding forecast days. Data is drawn from the same 27 northeastern and 21 southwestern stations. High and low temperature observations are from the same time period (1 Feb – 31 May 2007) for the 27 northeastern stations and the 21 southwestern stations.

3. METHODOLOGY AND ANALYSIS

3.1 Quality Control

Before calculations are performed, all data must be quality controlled. Initially an ensemble mean is taken over the 51 ensemble members. Due to ensemble forecasts being collected from early tests of the research and development version of *eCast*™, there are missing data points flagged

with the value -999. These data points have higher frequencies in February and March, resulting primarily from data delivery problems by the ECMWF data provider. Fortunately such missing values occurred for all ensemble members in a given forecast so the ensemble mean could be safely calculated, producing a -999 for each validation day with missing forecasts. The observations similarly contain flagged missing points. Forecast error is calculated using the mean absolute error ($|$ *Ensemble Mean temperature – Observed temperature* $|$). Because error is a function of ensemble mean *and* observations, a quality mask that marks the positions of valid points is applied enabling an error calculation to occur only where there are valid observations and ensemble means. This shrinks the data set slightly below 120 validation days for each station. For each of the error predictors, this mask has to be tailored slightly. These alterations will be discussed in the respective sections below.

3.2 Ensemble Spread

This portion of the study serves to re-examine the relationship between ensemble spread and forecast error. The re-examination attempts to reveal more detailed characteristics of the spread-error relationship through the medium range and into the early extended range forecasts.

After calculating standard deviation of the ensemble members, the quality mask is applied to the set of ensemble spread to ensure that the same validation days are compared. The ensemble spread axis is divided into 20 one-degree bins. Each bin of ensemble spread contains an amount of corresponding forecast error values. Taking the mean of these forecast error values produces one forecast error point per bin. Bins with fewer than five forecast error values are skipped in the mean calculation because means taken over fewer than five points become excessively noisy.

A linear regression is performed for each bin's forecast error mean. The regression is weighted by the number of forecast error points in order to obtain a slope that represents the majority of the data. The correlation coefficient between the forecast error averages and the binned ensemble spread is also calculated and squared in order to obtain the coefficient of determination (r^2). This process is performed for each of the 15 forecast days for high and low temperature forecasts.

3.3 Model-to-model Consistency

This study further investigates the value of using poor man's or grand ensembles by testing the ability of model-to-model consistency to predict forecast error. The *eCast*™ ensemble mean is compared to deterministic MOS/MEX output. MEX provides forecasts out to seven days, so only ensemble forecast days 1-7 are considered in this portion of the study. The MEX data contains flagged missing values, so the original quality mask is expanded to exclude such points from calculations. In order to obtain a measure of agreement between MEX and ensemble means that can be easily plotted against forecast error and evaluated, a simple absolute model difference is taken ($|$ *Ensemble mean temperature – MEX temperature* $|$).

Forecast errors are plotted versus the forecast difference between the two models. Model differences are similarly divided into 20 one-degree bins. Forecast error means in each bin are calculated and plotted. A linear fit is performed, weighted to error means over bins containing larger amounts of forecast error points. Extracting the slope of the regression line and calculating the coefficient of determination (r^2) to obtain a measure of relationship significance, the process is repeated for each of the seven forecast days for high and low temperatures.

3.4 Run-to-run Consistency

Forecasters often take note of run-to-run consistency in observing model data, but studies on the effectiveness of run-to-run consistency in predicting forecast error are sparse. In order to calculate the difference between two consecutive forecast runs for a single forecast day, the 2-day forecast on a validation day is subtracted from the 1-day forecast on the following validation day ($|$ (*2 day temperature forecast on valid day 1*) – (*1 day temperature forecast on valid day 2*) $|$). This absolute run difference shows how forecasts for a specified lead time vary from one validation date to the next. However, this formula introduces some restrictions in that 1) only forecast days 2-15 for each valid day can be compared because forecast day 1 cannot be subtracted from forecast day 0 if such a lead time does not exist, 2) the quality mask must be expanded to prevent missing forecasts in both validation day's from being run into the equation and 3) forecast error must be calculated for the 1-day forecast issued on the later validation day.

Twenty one-degree bins of the difference between the consecutive runs are created and the mean of the forecast errors within each bin is again calculated and plotted. The linear regression line is weighted to bin error means taken over larger amounts of data. Extracting the slope of the regression line and calculating r^2, the process is repeated for 2-15 day lead times for high and low temperature.

4. RESULTS

Three pentads (5-day averages) of slope and r^2 divide the 15-day forecast period of high and low temperature into early, medium and long range lead times. For northeast high temperature forecasts, the ensemble spread- forecast error relationship showed generally high significance (r^2>.6) through the middle pentad (5 -10-day lead). Large slopes through the middle pentad represent high forecast error sensitivity to ensemble spread as well. There is a rather abrupt fall in relationship significance and error sensitivity to spread in the final pentad (11 - 15-day lead), where both tend to level off closer to zero (Fig. 1a). Low temperature forecasts show a different trend, with relationship significance remaining high (r^2>.6) and error sensitivity to ensemble spread increasing throughout the forecast days (Fig. 1b). A sensitivity max occurs in the middle pentad and a significance max occurs in the final pentad.

Due to having only seven forecast lead times, model consistency results are presented for each forecast day. For northeast high temperature forecasts, significance remains high through forecast day 5, decreasing only slightly thereafter while error sensitivity to model disagreement steadily drops throughout the period (Fig. 2a). Low temperatures on the other hand show drastic fluctuations in relationship significance, with relatively low forecast error sensitivity throughout the period (Fig 2b).

Similar to ensemble spread results, three pentads are taken again for run-to-run consistency results. The first pentad differs slightly however in that it contains averaged values for forecast days 2-5 due to the inability to compare a previous run for forecast day 1 as mentioned in section 3.3. For northeast high temperatures, significance remains extremely high through the first pentad, dropping off rapidly toward the end of the forecast period (Fig 3a). Slope is also quite large for the 1st pentad, and drops off even more rapidly through the middle range. Day-to-day analysis (Fig 3b) actually shows relationship significance over 60% and large slope through most of the middle pentad as well, with a precipitous drop at the 9-day forecast. Although significance and sensitivity drop in the latter days of the 2nd pentad for low temperatures as well, the decline is more gradual throughout the period. As with high temperature forecasts, there is a maximum significance and error sensitivity in the 1st pentad.

Southwestern results are presented in the same form for the sake of regional comparison. For both high and low temperature, forecast error is very sensitive to spread for all three pentads, showing high and unvarying significance throughout (Fig. 4a-b). Error sensitivity to model agreement (Fig. 5a-b) in the southwest is slightly greater than in the northeast, showing more significance for low temperatures than for high temperatures where sensitivity decreases in later forecasts. Error sensitivity to run-to-run consistency in the southwest is slightly greater for high temperatures while the significance of the relationship is greater for low temperatures (Fig. 6a-b). In general, error sensitivity (slope) and the significance of the sensitivity (r^2) are larger and more consistent in the southwest than in the northeast.

5. DISCUSSION AND CONCLUSIONS

The two values calculated in this study (slope and determination coefficient, r^2) give an overall indication of how well forecast error can be predicted by ensemble spread, inter-model agreement, and run-to-run consistency. The slope shows the sensitivity of forecast error to the predictor being tested. In the estimation of forecast error, the slope is essentially a “sensitivity multiplier.” Specifically, forecast uncertainty can be estimated using the linear model $ax + b$, where a is the slope of the linear regression, x is the value of the error predictor on a given forecast day, and b is the offset of the regression line. It is important to note that this study utilizes the absolute values of forecast error, thus providing information regarding forecast uncertainty rather than error. Corresponding determination coefficients indicate the significance of error estimates made with the linear model. Higher coefficients validate the use of error estimates made using the calculated slope and offset. This couplet of information is especially important in evaluating and ranking the value of the three error predictors

5.1 Ensemble Spread

For the northeast, this study generally verifies previous findings that indicate the ability of ensemble spread to predict forecast error, while providing greater detail on the nature of this spread-error relationship. For high temperature forecast error, ensemble spread is an excellent predictor through the middle pentad (specifically through 8 day lead times). Thereafter, ensemble spread becomes almost useless in predicting high temperature forecast error. Low temperature forecast error is better predicted by ensemble spread in middle and long range pentads, with lower ability to predict error in the early (1-5 day) lead times. In the day-to-day analysis, the increase in forecast error sensitivity to ensemble spread through the earlier forecast period is due to the fact that early lead times tend to have higher ensemble spread than forecast error. As lead times progress, forecast error grows more rapidly than ensemble spread causing increasing slopes in the spread-error analysis. This increasing trend leads to a maximum of error sensitivity to ensemble spread in lead times of 7-9 days for both high and low temperature. The leveling trend of error predictability for low temperature beyond forecast day 9 is not unusual as the lead time nears Lorenz's aforementioned limit of predictability. However, the more sudden drop off in error predictability for high temperature at the 9-day lead time (3rd pentad) certainly requires further investigation.

5.2 Model-to-model consistency

Results indicate that model-to-model consistency performs rather poorly at indicating error in high and especially low temperature forecasts. For high temperatures, significance and forecast error sensitivity to inter-model consistency decline steadily with increasing lead times suggesting this error predictor becomes less valuable in longer range forecasts. Low temperature forecast error shows consistently low dependence upon inter-model agreement, with a randomly fluctuating relationship significance throughout the period. The low skill in error prediction for low temperatures could be due to model deficiencies in capturing nighttime boundary-layer radiative flux (Betts et al. 1997; Morcrette 2001). The differences in each model's nighttime boundary layer schemes could also cause this loss of error predictability. Furthermore, the MOS product, which is largely governed by climatology at lead times as early as seven days introduces a new set of errors into the model comparison. Using a variety of other models that do not rely so heavily on climatology would be beneficial in making general conclusions regarding the value of model-to-model consistency in predicting error.

5.3 Run-to-run consistency

Run-to-run consistency proves to be the best error predictor for both high and low temperatures in the short range and early middle range forecasts, with moderate extended range skill in error prediction for low temperatures. The highest significance coupled with extremely high forecast error sensitivity verifies the ability of run-to-run consistency to predict forecast error in the first pentad. Low temperature results show similar trends, but with improved error prediction skill in the longer ranges. The significant loss of error predictability for low and especially high temperature in the long range forecasts again requires further investigation.

At this point, a summary of the most skillful error predictors for each pentad of lead times would fail to show an adequate error predictor for high temperature in the longer range forecasts. Because this study aims to improve error predictability in all parts of the 15-day forecast period, results from the southwest are used to compare results regionally and provide a broader understanding of error predictability.

6. REGIONAL DEPENDENCE OF ERROR PREDICTION AND NEW DIRECTIONS

Southwest results show higher significance and forecast error sensitivity to each error predictor. Comparing the 3rd pentad skill in error prediction, forecast error sensitivity to ensemble spread is almost equivalent to that of run-to-run consistency. However, slightly higher relationship significance is found for ensemble spread. Using a combination of northeast and southwest results, a summary of error prediction skill guidance through the 15-day forecast period is obtained (table 1).

It is worth discussing possible reasons for increased error predictability in the southwest. More stations in the northeast experience a drastic seasonal shift in the months observed in this study while some low-elevation stations in the southwest experience more gradual seasonal changes. Northeast stations are located in an area of greater springtime baroclinic instability, especially compared to the southernmost stations in the

southwest. Northeastern weather is more continental showing greater diurnal temperature fluxes, while some southwestern stations are highly influenced by a marine air mass. As a result, the northeast is more likely to have greater and more unpredictable forecast errors, causing the error predictors to lose value rapidly in the longer range forecasts.

Comparing two climatologically disparate regions was beneficial in broadening the applicability of this study. Such error prediction guidance may be general, but can be of great value to forecasters. Improving the anticipation of model error provides weather-sensitive corporations and futures traders with a better idea of forecast verification probability. Nonetheless, the study can be extended. Datasets encompassing multiple seasons could be analyzed in order to determine a possible annual cycle of error predictability. For inter-model consistency, comparing two 15-day forecast models instead of the 7-day MOS would be beneficial in assessing long range error prediction skill. In terms of data quality control, applying a single comprehensive mask to all three variables would ensure consistency in validation days being used in calculations. Finally, in order to better account for the possible regional dependence, normalizing the results with climatological forecast variability for each region may provide more universal conclusions.

Acknowledgments: This material is based upon work supported by the National Science Foundation under Grant No. ATM-0648566. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. At AER, the author would like to thank mentor David Hogan, Mark Leidner, Yuguang He for technical support and programs, and Christian Alcala, Edward Kennelly, and Scott Zaccheo for their assistance. At OU the author would like to especially thank Daphne LaDue for making the National Weather Center REU possible.

References

Atger, Frederic, 1999: The Skill of Ensemble Prediction Systems. *Mon. Wea. Rev.*, **127**, 1941-1953.

Barker, Timothy, 1991: The Relationship between Spread and Forecast Error in Extended range Forecasts. *Journal of Climate*, **4**, 733-742.

Betts, Alan K., Fei Chen, Kenneth E. Mitchell, and Zavisa I. Janjic, 1997: Assessment of the Land Surface and Boundary Layer Models in Two Operational Versions of the NCEP Eta Model Using FIFE Data. *Mon. Wea. Rev.*, **125,** 2896-2916.

Buizza, Roberto, 2000: Chaos and Weather Prediction. http://www.ecmwf.int.newsevents/training/course_notes/PREDICTABILITY/CHAOS/index.html.

Buizza, Roberto, P.L. Houtekamer, Zoltan Toth, Gerald Pellerin, Mozheng Wei, and Yuejian Zhu, 2005: A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems. *Mon. Wea. Rev.*, **133**, 1076-1097.

ECMWF HTML User Guide http://www.ecmwf.int/products/forecasts/guide1

Leith, C.E., 1974: Theoretical Skill of Monte Carlo Forecasts. *Mon. Wea. Rev.*, **102**, 409-418.

Lorenz, Edward N., 1963: Deterministic Nonperiodic Flow. *J. Atmos. Science*, **20**, 130-141.

Morrcrette, J.-J., 2001: Assessment of the ECMWF Model Cloudiness and Surface Radiation Fields at the ARM SGP Site. 11th ARM Science Team Meeting Proceedings, Atlanta, GA, March 19-23, 2001.

Richardson, D. S., 2000: Skill and Relative Economic Value of the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteo. Soc.*, **126**, 649-667.

Smagorinsky, Joseph, 1969: Problems and Promises of Deterministic Extended Range Forecasting. *Bull. Amer. Meteor. Soc.*, **50**, 286-311.

Toth, Zoltan and E. Kalnay, 1993: Ensemble Forecasting at NMC: The Generation of Perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317-2330.

Whitaker, Jeffrey S. and Andrew F. Loughe, 1998: The Relationship between Ensemble Spread and Ensemble Mean Skill. *Mon. Wea. Rev.*, **126**, 3292-3301.

Yoden, Shigeo, 2007: Atmospheric Predictability. *J. Met. Soc. Japan*, **85B,** (*in press*)

Figures 1-3: Northeast Results

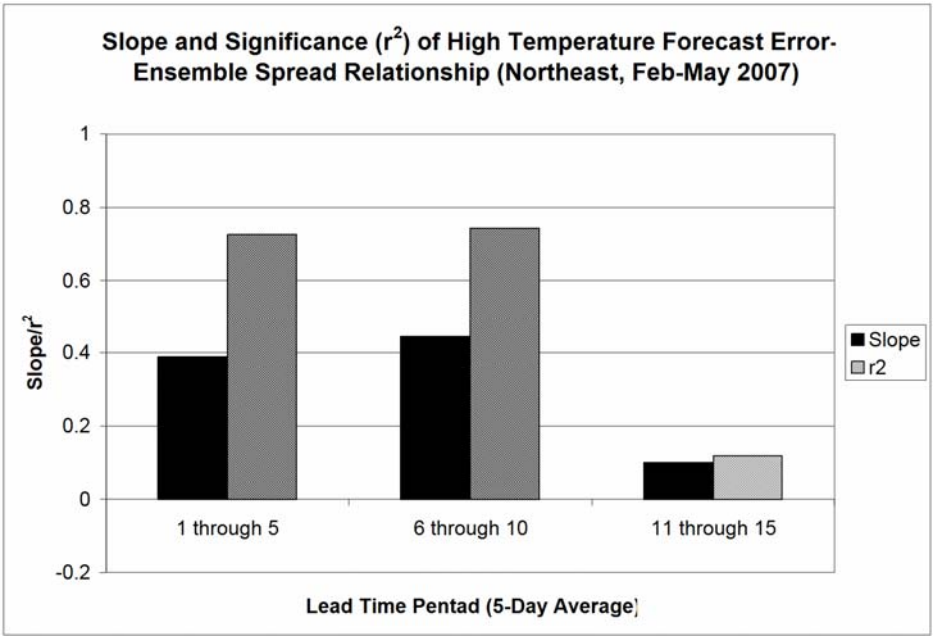

Fig. 1a. Northeast high temperature forecast error sensitivity to ensemble spread as indicated by the slope presented with the significance of that slope in predicting forecast error. Pentads with significance exceeding .6 have bold crosshatching. Ensemble spread is valuable for error prediction through middle range forecasts, and loses value in the final pentad.

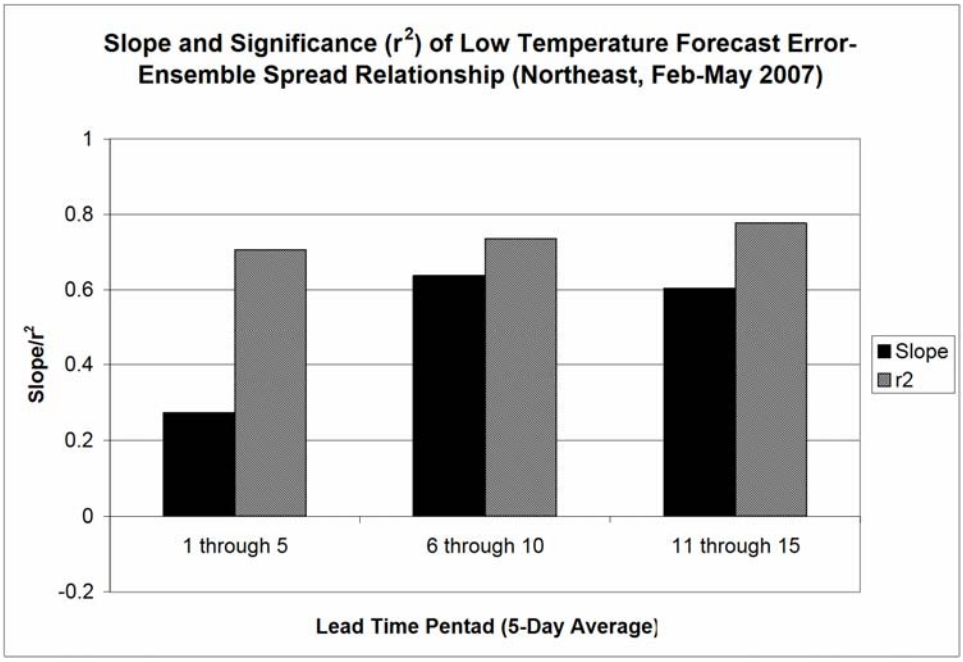

Fig. 1b. Northeast low temperature forecaster error sensitivity to ensemble spread (slope) with significance of the slope as a predictor of forecast error. Low temperature error sensitivity to ensemble spread is best for middle and long range forecasts as seen in large slopes in 2nd and 3rd pentads. Significance of ensemble spread as an error predictor is high (>.6) throughout all 15 days.

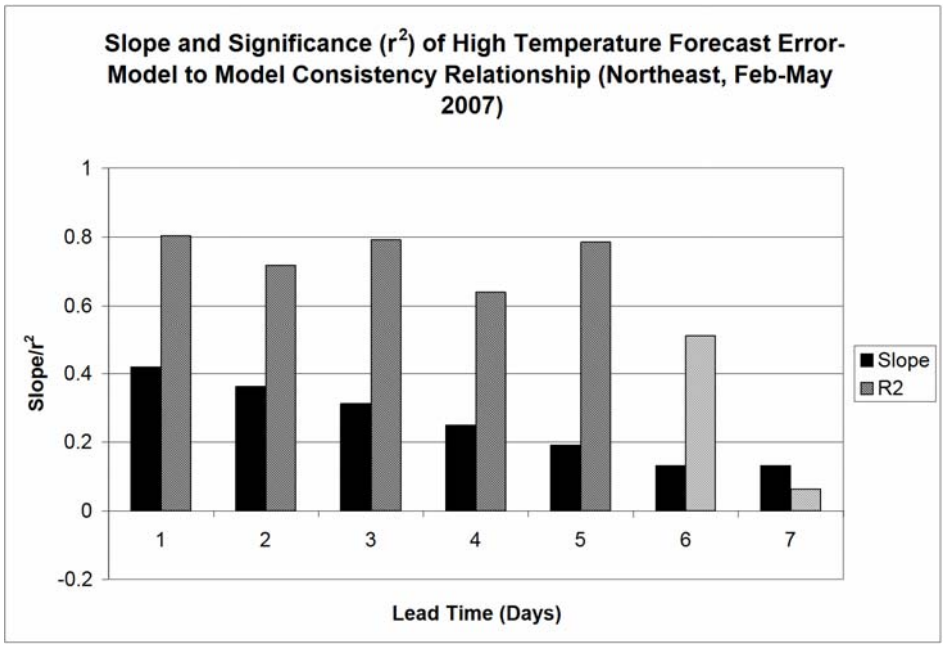

Fig. 2a. Northeast results for the ability of MOS/MEX and *eCast*™ ensemble mean temperature forecast agreement to predict high temperature forecast errors. Days with significance exceeding .6 have bold crosshatching. Inter-model agreement is only moderately useful as a predictor of forecast error for short range (1-5 day lead times). This is evident through moderate forecast error sensitivity to model agreement (slope) and high relationship significance through a 5-day forecast.

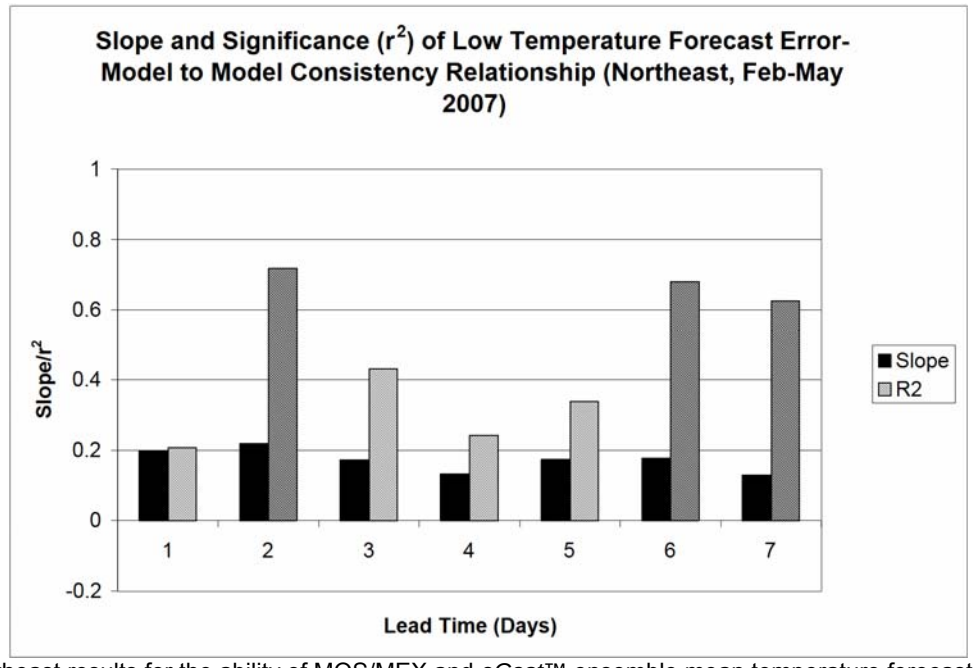

Fig. 2b. Northeast results for the ability of MOS/MEX and *eCast*™ ensemble mean temperature forecast agreement to predict low temperature forecast errors. Days with significance exceeding .6 have bold crosshatching. There is little sensitivity of low temperature forecast error to inter-model agreement for all 7 lead times as seen in consistently small slopes. There is also generally low yet variable significance to the low error sensitivity results.

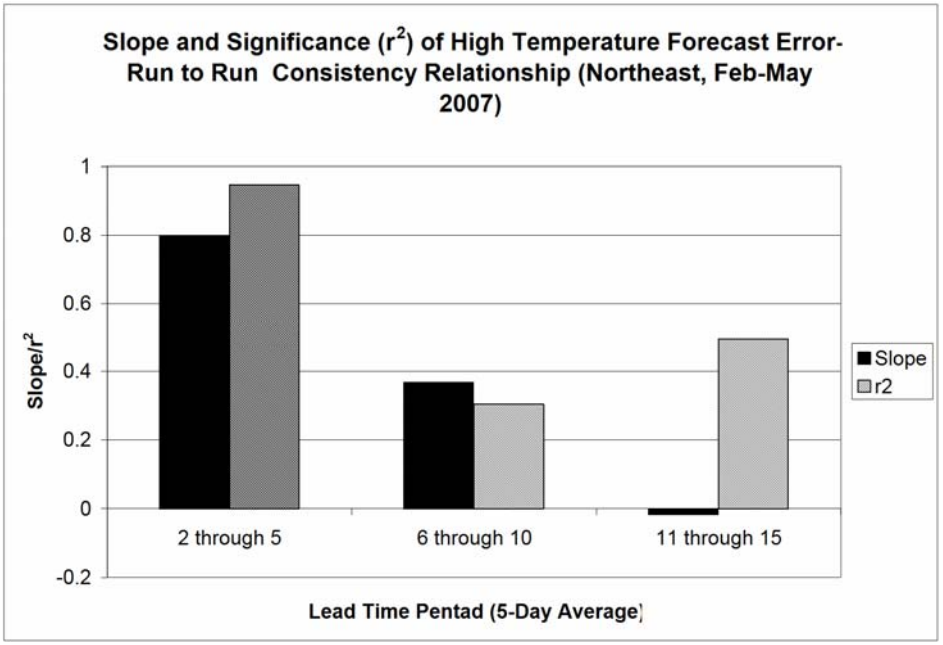

Fig. 3a. Northeast high temperature forecast error sensitivity to consistency between consecutive 24-hour runs of the ECMWF *eCast*™ ensemble. Pentads with significance exceeding .6 have bold crosshatching. Skill in predicting error is high for run-to-run consistency in the short range, with high sensitivity (large slope) and high relationship significance. Run-to-run consistency becomes less valuable in predicting forecast error in the middle range and especially the long range.

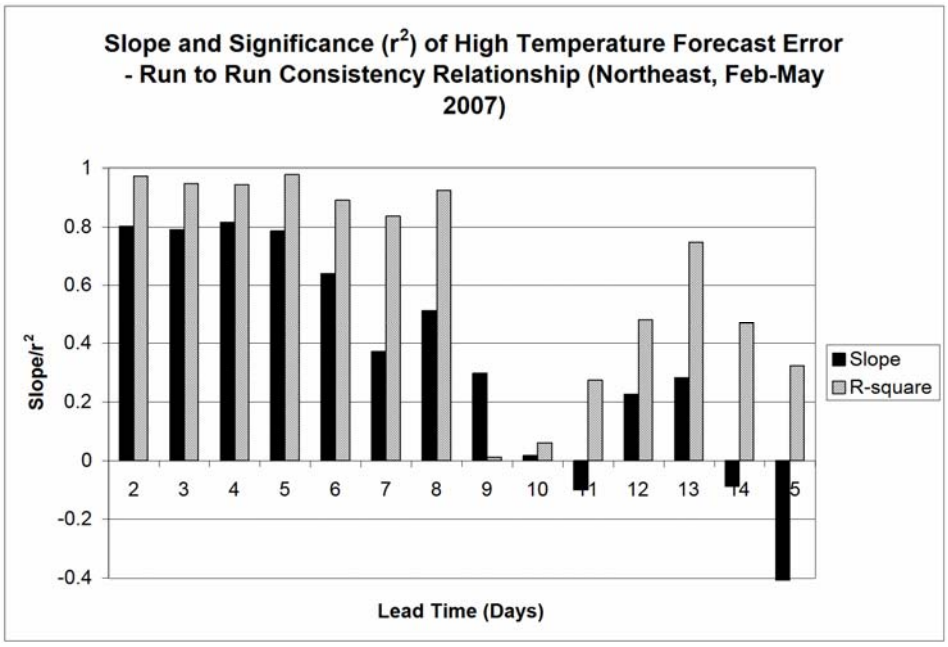

Fig. 3b. Multi-day averages (fig. 3a) are taken over this day-to-day analysis. Examining individual days indicates precisely where in the medium range a loss of skill in predicting error occurs for run-to-run consistency. Changes in ECMWF resolution or incorporation of climatology in *eCast*™ forecasts around 9-day lead times could be responsible for this sudden loss of error predictability.

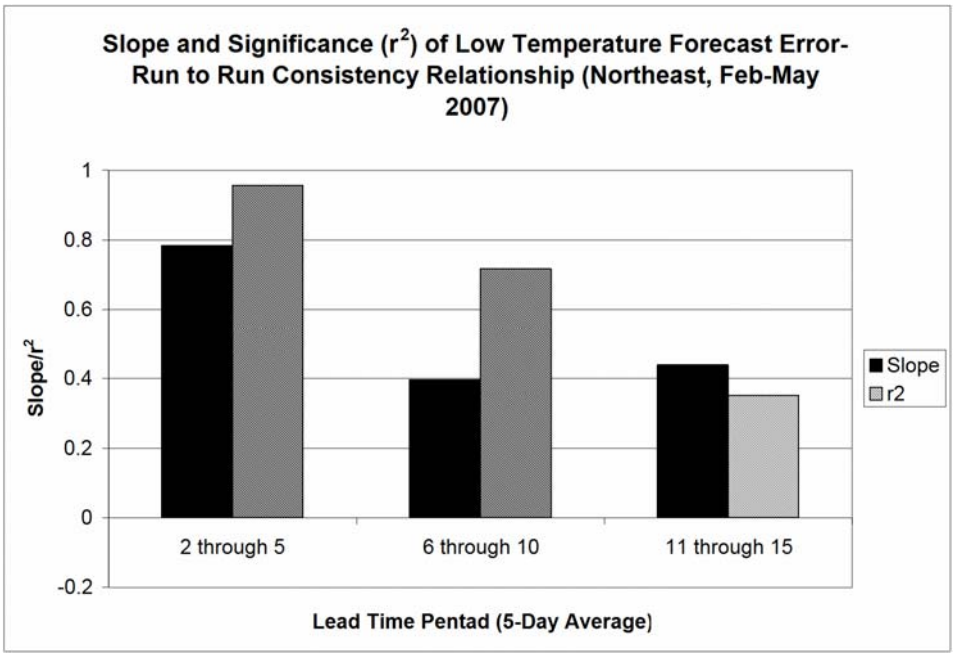

Fig. 3c. Northeast low temperature forecast error sensitivity to consistency between consecutive 24-hour runs of the ECMWF *eCast*™ ensemble. Pentads with significance exceeding .6 have bold crosshatching. Compared to high temperature forecast errors (fig. 3a), the predictability of error using run-to-run consistency remains high in the short range, and becomes moderate in the medium and long range forecasts. The drop off in error predictability of run-to-run consistency in middle range lead times is more gradual for low temperatures.

Figures 4-6: Southwest Results

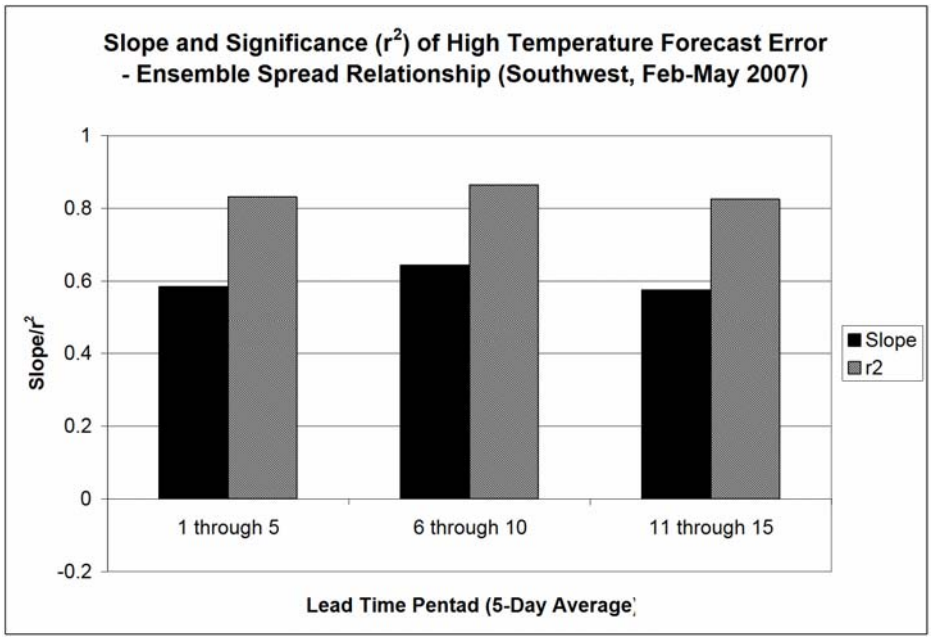

Fig. 4a. As in Fig. 1a but for the southwestern United States. Compared to northeast results (Fig. 1a), ensemble spread is more consistently valuable as an error predictor throughout the 15 forecast days, lacking the sudden loss of error prediction skill at the 9-day lead time.

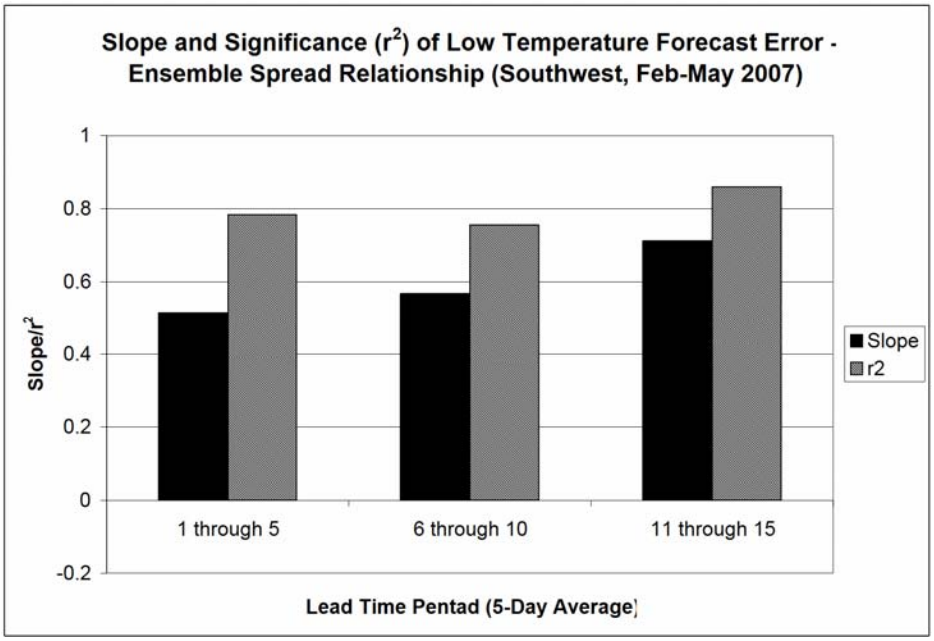

Fig. 4b. As in Fig. 1b but for the southwestern United States. Forecast error sensitivity to ensemble spread, and the significance of the spread-error relationship both reach a maximum in the long range, showing the increasing value of ensemble spread as an error predictor further out in a forecast. Compared to northeast results (Fig. 1b), ensemble spread is more valuable as an error predictor through the 15 forecast days.

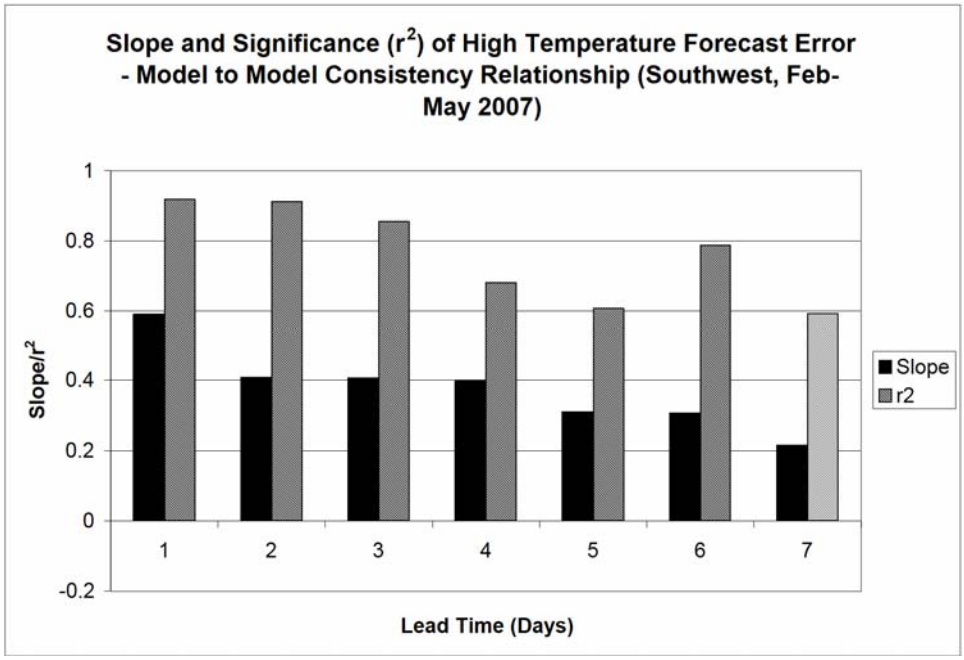

Fig. 5a. As in Fig. 2a but for the southwestern United States. Days with significance exceeding .6 have bold crosshatching. Similar to the northeast, inter-model consistency is a moderate error predictor in the short range, losing some skill in later lead times. Compared to northeast results (Fig. 2a), there is a noted improvement in error prediction skill as seen in larger slopes and higher significance indices.

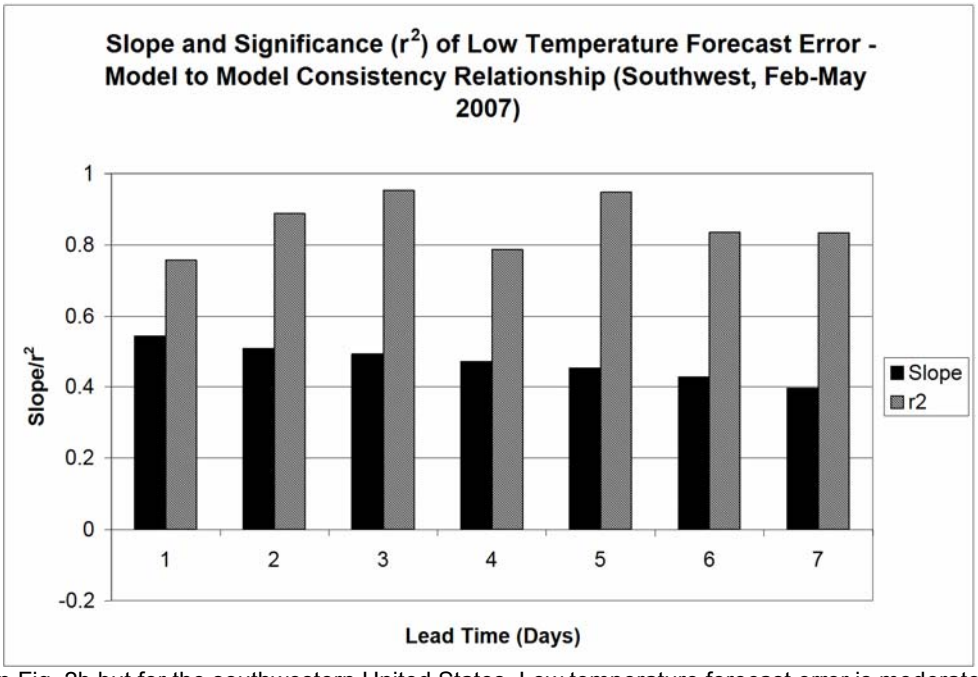

Fig. 5b. As in Fig. 2b but for the southwestern United States. Low temperature forecast error is moderately predicted by inter-model agreement, contrasting rather drastically with the low error sensitivities and variable relationship significance found in the northeast results (Fig. 2b)

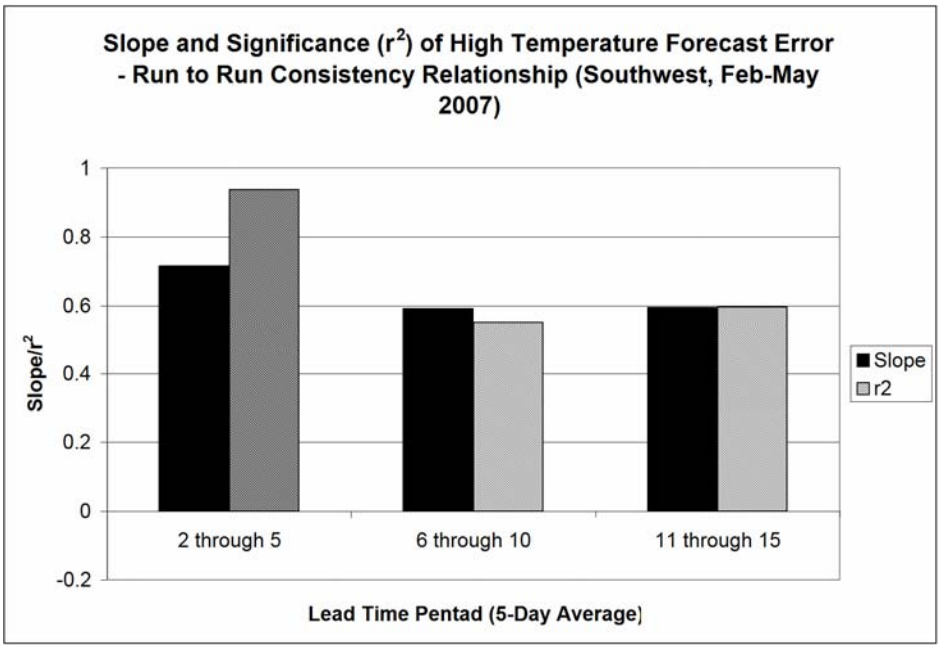

Fig. 6a. As in Fig. 3a but for southwestern United States. Pentads with significance exceeding .6 have bold crosshatching. Run-to-run consistency again predicts error best in the short range, losing skill in error prediction in the middle pentad. The loss of error prediction skill is similar, but not as dramatic as that found in the northeast for run-to-run consistency (Fig. 3a).

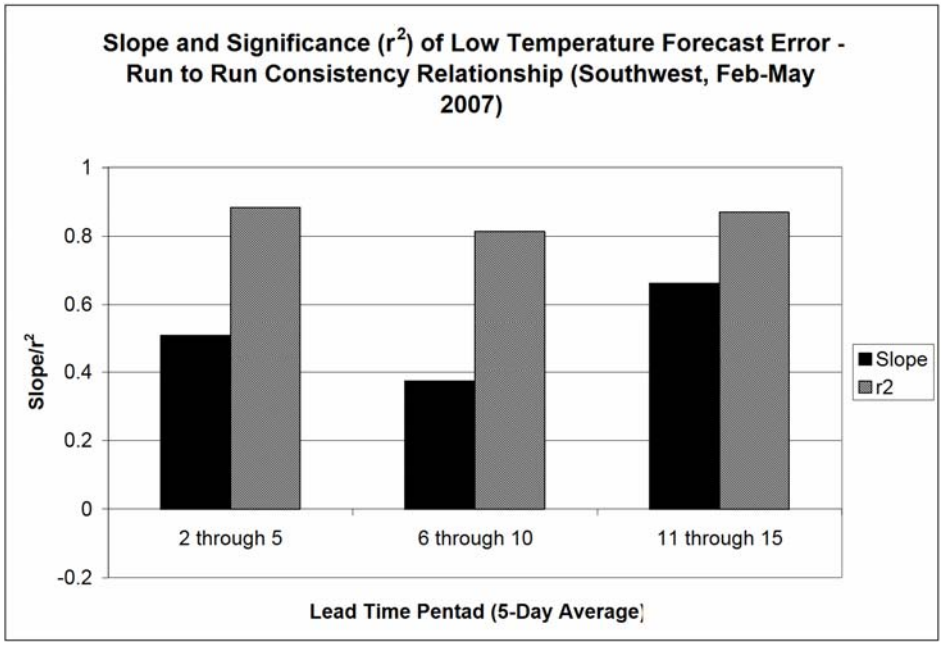

Fig. 6b. As in Fig. 3c but for southwestern United States. Although significance remains high throughout (>.6), forecast error sensitivity reaches a maximum in the final pentad. Greater forecast error sensitivity to run-to-run consistency in the longer range is inconsistent with northeast results that indicate a drop in sensitivity in longer range forecasts (Fig. 3c).

	Pentad 1 (1 – 5-day lead time)	Pentad 2 (6 – 10-day lead time)	Pentad 3 (11 – 15-day lead time)
High Temperature	Run-to-Run Consistency	Ensemble Spread	Ensemble Spread
Low Temperature	Run-to-Run Consistency	Ensemble Spread	Ensemble Spread

Table 1: The most effective predictors of error in the each forecast range based on combinations of highest relationship significance (r^2) and highest forecast error sensitivity to each error predictor (slope). These final conclusions are drawn from high and low temperatures results in the northeast and southwest combined. Ensemble spread clearly dominates error predictability in longer lead times, while run-to-run consistency is more valuable in short range error prediction.