Guidance Verification for the 2005-2006 NCWFC

Jon Moskaitis
April 21, 2006

1. Introduction

Here, I compare the performance of the GFS ensemble median high and low temperature forecasts with other available deterministic guidance for the 2005-2006 National Collegiate Weather Forecasting Contest (NCWFC). The verification metric I use to evaluate these deterministic forecasts is absolute error, as the contest scoring system is based on absolute error. In this year's edition, I also include results from a probabilistic verification method, in order to evaluate the quality of the continuous forecast probability distributions derived from the GFS ensemble MOS.

2. Deterministic verification

i. Types of guidance

Six different types of model-based guidance are evaluated in this note. Five of the guidance products are conventional model output statistics (MOS): the GFS ensemble control, the GFS operational[1], the AVN, the ETA, and the NGM. The other model-based guidance product, the GFS ensemble median, is derived from a continuous probability distribution fit to the discrete set of forecast values from the GFS ensemble MOS. If the aforementioned probability distribution was the true forecast probability distribution, the GFS ensemble median would be the deterministic forecast that minimizes expected absolute error. Significant changes were made to the distribution-fitting method between the 2004-2005 and 2005-2006 NCWFC seasons, in order to bring the GFS ensemble MOS-based forecast probability distributions closer to truth. The original ad hoc specification of the distribution parameters was replaced with a more defensible scheme of parameter estimation, based on past GFS ensemble MOS forecasts and the

[1] The ensemble control and operational versions of the GFS are integrated from the same analysis, but use different resolutions. The ensemble control uses a lower resolution than the operational, consistent with the rest of the ensemble members.

corresponding observations[2]. This change has only a slight effect on the distribution medians, but drastically alters the distribution shapes to attain more appropriate variances.

The model integrations from which the guidance products are derived are initialized at different times relative to the NCWFC forecast interval (06z – 06z). The GFS is initialized 30 hours prior to the beginning of the NCWFC forecast interval (00z), while the AVN, ETA, and NGM are initialized 18 hours prior (12z).

ii. Forecast set

High temperature guidance for all 13 NCWFC cities is considered here. I limited the verification to days when the high temperature had a daytime maximum, as this is the quantity the guidance is designed to forecast. This ended up excluding 10 of the 104 possible forecast days. Low temperature forecasts for all 13 cities were also verified, excluding the 31 days with afternoon or evening minima. For some cities, as many as half of the forecast days had to be excluded from the low temperature verification. See the bottom of Table 1 for the number of high and low temperature forecasts verified for each city.

iii. Results

Cumulative absolute error and mean absolute error (MAE) results for the year are shown in Table 1. For high temperature, the ETA and NGM attained the lowest values, consistent with the results from the 2004-2005 contest. The next best guidance product was the GFS ensemble median, which beat out the other 3 GFS products, including the 12z AVN. This was not the case last year, when the GFS ensemble median scored higher than both of its operational counterparts (12z AVN, 00z GFS operational). Perhaps the relative improvement of the GFS ensemble median this year can be partially attributed to

[2] Specifically, the new scheme calculates the parameters that maximize the probability that the observations are drawn from the forecast probability distributions (constructed using the parameters), for the training data. The training data was 9 months of GFS ensemble MOS forecasts and observations for Boston. Ideally, distribution parameters should be separately estimated for each NCWFC city.

the new formulation of the forecast probability distributions from the GFS ensemble MOS values.

Figure 1 shows a city-by-city analysis of high temperature forecast MAE for the 6 guidance products. It is clear that the city-specific relative performances vary significantly, and are not necessarily in line with the overall relative performance of the guidance products. There is no 'model of choice' for every combination of geographical location and flow regime. In fact, only the 12z AVN failed to be the best model for at least of the 1 of the 13 stations! The biggest differences between the performance of the various guidance products occurred at Fairbanks (FAI). It is interesting to note that if this city is removed from the overall statistics, the GFS ensemble median would edge the ETA and NGM for the lowest MAE.

Low temperature results for the year can also be found on Table 1. The ETA was by far the best, as opposed to last year when all the guidance products performed rather similarly. Last year, the GFS ensemble median was the worst low temperature guidance product, but this year it beat both the GFS control and operational. Figure 2 shows the city-to-city variability in the mean absolute error of the low temperature guidance. Again, there is no clear 'model of choice' that is the best for every city.

For information on the accumulated absolute errors for high and low temperature forecasts at each city, please see Table 2.

3. Probabilistic verification

Probabilistic verification of the GFS ensemble MOS-based forecast probability distributions is a new undertaking for this year's NCWFC guidance verification. To relate it as closely as possible to the previously discussed deterministic forecast verification, I have chosen to use a somewhat atypical probabilistic verification method. This method compares the actual absolute errors of the distribution medians (used in the last section) to *expected* absolute errors for those medians. For a given forecast

probability distribution, $p(x)$, the expected absolute error (EAE) of the median, x_{median}, is given by

$$\mathrm{EAE} = \int_{-\infty}^{\infty} p(x)\left|x - x_{median}\right| dx \,. \tag{1}$$

A similar expression can be formed for $q(x)$, the true forecast probability distribution. If $p(x)$ and $q(x)$ are the same distribution, as we would like, then their EAEs must also be the same. However, EAE equality is a necessary, but not sufficient condition. It is possible that $p(x)$ and $q(x)$ are different, even if they have the same EAE. We will conveniently ignore this possibility from now on, though, and assume EAE equality means $p(x) = q(x)$.

A 'forecast' EAE – 'true' EAE comparison would be very straightforward, but unfortunately, we do not have the true forecast probability distribution at our disposal. Instead, we have only one draw from it, the observation. So, the forecast probability distribution has an expected absolute error (EAE), but all we have to compare it to is an absolute error (AE), calculated from the observation. This is a fundamental difficulty in probabilistic verification.

We can deal with this difficulty by considering a large *set* of probabilistic forecasts and the corresponding observations, instead of just one forecast and one observation. For each forecast in the set, one can calculate the EAE according to Eq. 1, and calculate the AE using the value of the corresponding observation. A scatter plot can then be made for the EAE-AE pairs. Such a scatter plot is shown by the blue circles in Figure 3, for the set of 94 high temperature GFS ensemble MOS-based forecast probability distributions. Now, let us take the 94 pairs and sort them into 5 mutually exclusive bins, according to their EAE values. Thus, each bin has pairs all with roughly similar EAE values. For each bin, one can then calculate the mean EAE (MEAE) and the corresponding mean AE (MAE) for all the pairs in the bin. In effect, we are using the binning to try to do a crude 'forecast' EAE – 'true' EAE comparison: the MEAE is the 'forecast' EAE (which is roughly constant amongst all the pairs in a bin) and the MAE is

the ‘true’ EAE, calculated using many observations. This method circumvents the fact that there is only one observation per forecast probability distribution.

The red squares in Figure 3 show the 5 MEAE-MAE pairs, one for each bin. One can see that MAE increases monotonically with MEAE, meaning that, *on average*, the bigger the expected absolute error of the median of the forecast probability distribution, the bigger the realized absolute error of the median. This is exactly what one would expect from a set of probabilistic forecasts that can skillfully discern high-uncertainty situations from low-uncertainty situations.

Still, there is much room to improve the high temperature GFS ensemble MOS-based forecast probability distributions. Ideally, the MEAE-MAE pairs in Figure 3 should all fall on dashed red line, where MEAE=MAE. The MEAE-MAE pairs for the first 3 bins (ordered from lowest to highest MEAE) are all below the dashed red line, meaning that the forecast probability distributions are generally too wide (EAEs too large) for the low-uncertainty forecasts. The MEAE-MAE pair for the 5th bin is above the red-dashed line, meaning that the in situations of high uncertainty, the forecast probability distributions are generally too narrow. Overall, it appears that the high temperature forecast probability distributions are too ‘rigid’ in their variance: they do not get narrow enough in low-uncertainty situations and they do not get wide enough in high uncertainty situations. It is noteworthy that one could not reach this conclusion by comparing the overall MAE and MEAE, which are almost exactly the same (represented by the green square in Figure 3). One must use the binning technique to infer such conditional[3] bias in the distribution EAEs.

The exercise described above can also be carried out for the low temperature forecast probability distributions. The results are shown in Figure 4, which is in the same format as Figure 3. Here, there is not a discernable relationship between bin MEAE and bin MAE, much less a monotonic relationship. The low temperature forecast probability distributions appear to have no skill whatsoever in discerning high uncertainty from low

[3] Conditional on the forecast, specifically

uncertainty situations. The overall MEAE is not even correct, as it underestimates the overall MAE by 0.5.

4. Discussion and Conclusion

This year's verification of the deterministic high temperature guidance largely supports the results of last year's study, with the ETA and NGM best overall. The GFS ensemble median improved its relative performance this year, placing third best overall and equal to the ETA and NGM if Fairbanks is excluded from the MAE calculation. It is possible that the revised formulation of the GFS ensemble MOS-based probability distributions is responsible for this improvement, but it could also be the case that this year's stations and flow regimes were simply better suited to the relatively low-resolution GFS ensemble.

Probabilistic verification of the GFS ensemble MOS-based high temperature forecast probability distributions reveals that they show skill in discerning between high and low uncertainty forecasts, but are a bit too rigid in holding the EAE close to its mean value. This is partially a consequence of the methodology used to transform a discrete set of GFS ensemble members into a continuous distribution. The static parameters used in this process are designed to yield 'good' probability distributions *on average*, for a large set of training data. The conditionally biased EAEs lead me to think that the parameters must be conditional on the forecast, instead of static. Perhaps the development of 'adaptive parameters' is a worth exploring before next year's contest.

Low temperature verification did not yield quite as much insight as its high temperature counterpart. The ETA was the best deterministic guidance product this year, but was in the middle of the pack last year. This makes it difficult to draw any conclusions about its potential superiority, although its relatively high boundary layer resolution should sensibly give it an advantage in predicting nighttime temperature. The probabilistic verification of the GFS ensemble MOS-based low temperature forecast probability distributions showed that there was essentially no skill in discerning high

uncertainty from low uncertainty events. It is not clear that improved parameter estimation (of distribution characteristics) could ameliorate this problem.

Table 1: Cumulative Error for 2005-2006 NCWFC

		Accumulated Absolute Error		
Models	*Time*	*High Temp*	*Low Temp*	*Total*
GFS Ens Median	00z	304	253	557
GFS Control	00z	313	258	571
GFS Operational	00z	328	269	597
AVN	12z	323	247	570
ETA	12z	280	216	496
NGM	12z	278	241	519

		Mean Absolute Error		
Models	*Time*	*High Temp*	*Low Temp*	*Total*
GFS Ens Median	00z	3.23	3.47	6.70
GFS Control	00z	3.33	3.53	6.86
GFS Operational	00z	3.49	3.68	7.17
AVN	12z	3.44	3.38	6.82
ETA	12z	2.98	2.96	5.94
NGM	12z	2.96	3.30	6.26

	Number of forecasts	
City	*High Temp*	*Low Temp*
CHS	7	5
APN	6	5
VCT	8	6
ISN	8	6
EKN	6	4
PDX	8	6
SBN	6	4
BOS	8	6
FAI	5	6
SLC	8	7
PNS	8	6
ICT	8	6
TUS	8	6
Total	**94**	**73**

Table 2: City-by-city accumulated absolute error for each model

GFS Ens Median

City	Accumulated Absolute Error		
	High Temp	Low Temp	Total
CHS	13	11	24
APN	23	24	47
VCT	7	24	31
ISN	18	39	57
EKN	22	27	49
PDX	12	20	32
SBN	24	8	32
BOS	20	6	26
FAI	55	23	78
SLC	35	30	65
PNS	25	15	40
ICT	37	15	52
TUS	13	11	24
Total	**304**	**253**	**557**

AVN

City	Accumulated Absolute Error		
	High Temp	Low Temp	Total
CHS	16	9	25
APN	23	15	38
VCT	15	16	31
ISN	25	27	52
EKN	15	22	37
PDX	10	16	26
SBN	27	6	33
BOS	22	8	30
FAI	41	35	76
SLC	52	35	87
PNS	28	17	45
ICT	31	25	56
TUS	18	16	34
Total	**323**	**247**	**570**

GFS Ens Control

City	Accumulated Absolute Error		
	High Temp	Low Temp	Total
CHS	12	10	22
APN	25	26	51
VCT	8	24	32
ISN	17	38	55
EKN	26	29	55
PDX	13	20	33
SBN	22	8	30
BOS	22	8	30
FAI	60	25	85
SLC	36	30	66
PNS	26	13	39
ICT	39	16	55
TUS	7	11	18
Total	**313**	**258**	**571**

ETA

City	Accumulated Absolute Error		
	High Temp	Low Temp	Total
CHS	15	10	25
APN	21	19	40
VCT	18	15	33
ISN	27	21	48
EKN	16	14	30
PDX	17	16	33
SBN	22	9	31
BOS	22	7	29
FAI	30	25	55
SLC	25	34	59
PNS	25	19	44
ICT	27	13	40
TUS	15	14	29
Total	**280**	**216**	**496**

Table 2 (continued)

GFS Operational

City	Accumulated Absolute Error		
	High Temp	Low Temp	Total
CHS	26	8	34
APN	22	19	41
VCT	16	24	40
ISN	23	34	57
EKN	16	24	40
PDX	7	20	27
SBN	31	6	37
BOS	20	9	29
FAI	40	29	69
SLC	45	38	83
PNS	28	21	49
ICT	39	19	58
TUS	15	18	33
Total	**328**	**269**	**597**

NGM

City	Accumulated Absolute Error		
	High Temp	Low Temp	Total
CHS	7	11	18
APN	14	14	28
VCT	33	24	57
ISN	32	29	61
EKN	12	25	37
PDX	11	13	24
SBN	20	13	33
BOS	22	5	27
FAI	28	35	63
SLC	26	34	60
PNS	19	10	29
ICT	27	10	37
TUS	27	18	45
Total	**278**	**241**	**519**

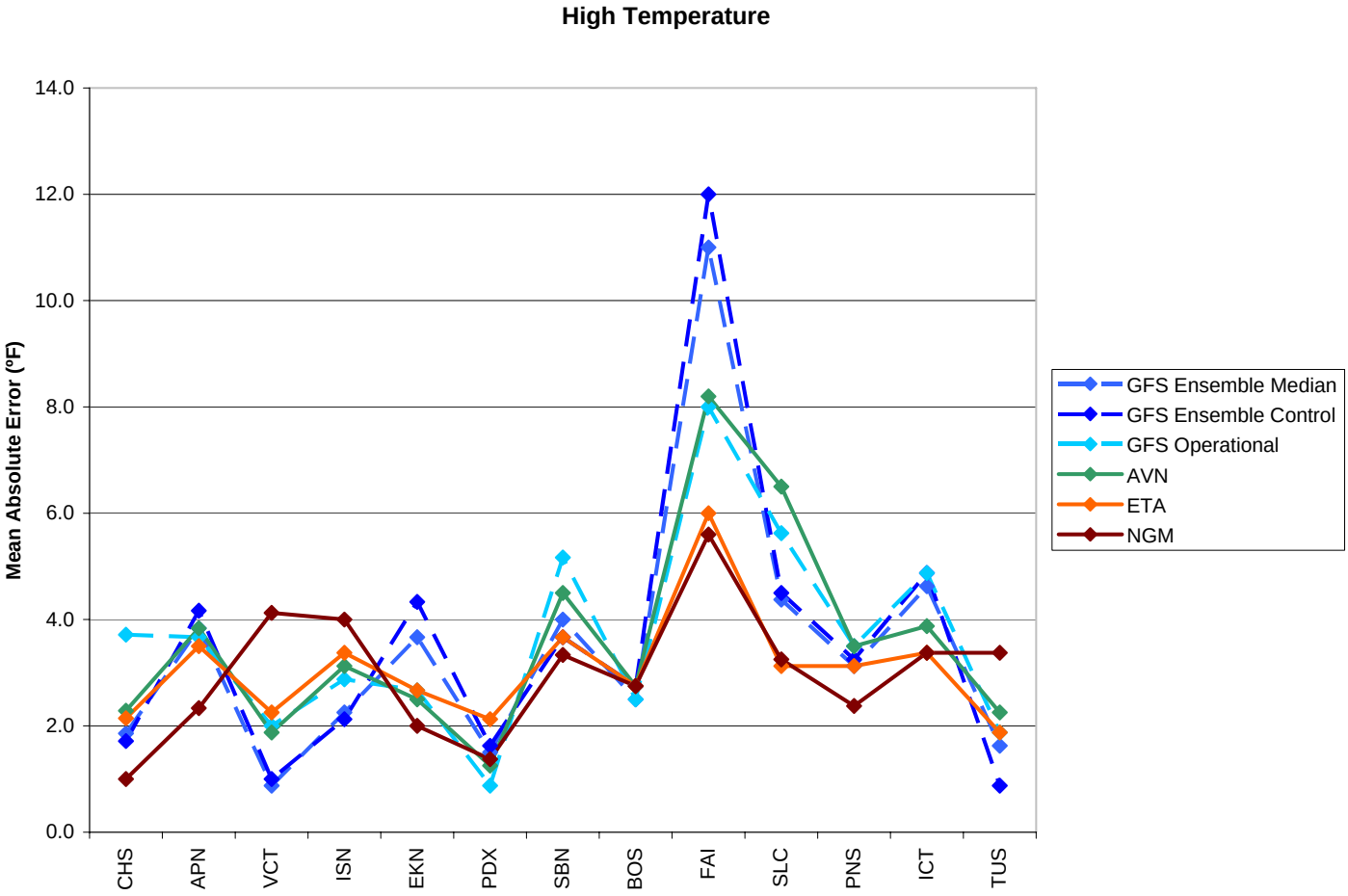

Figure 1: Mean absolute error for high temperature guidance at each of the 13 NCWFC cities in the 2004-2005 contest. The number of forecasts that went into each of these means is documented at the bottom of Table 1.

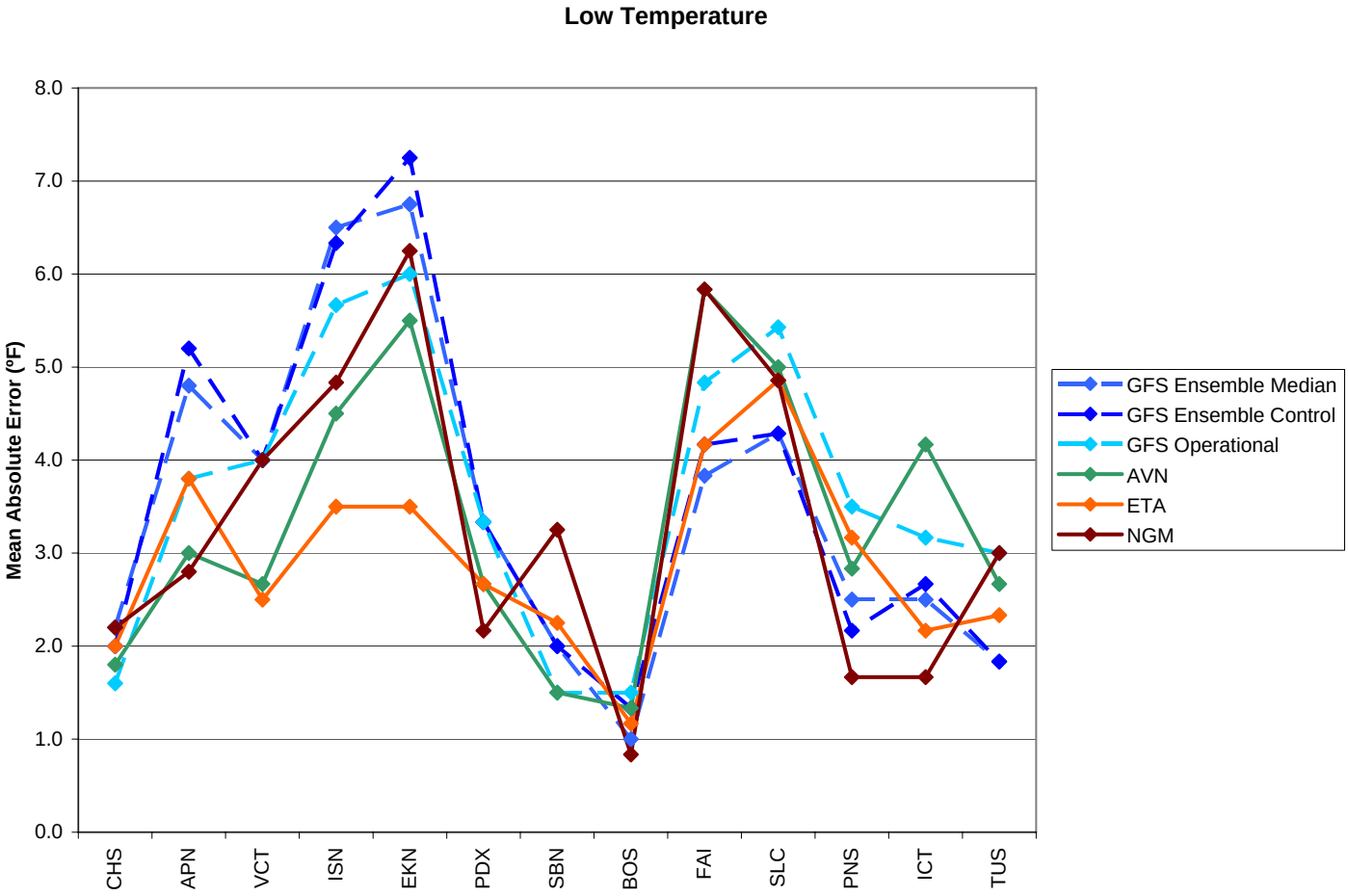

Figure 2: As in Figure 1, but for low temperature.

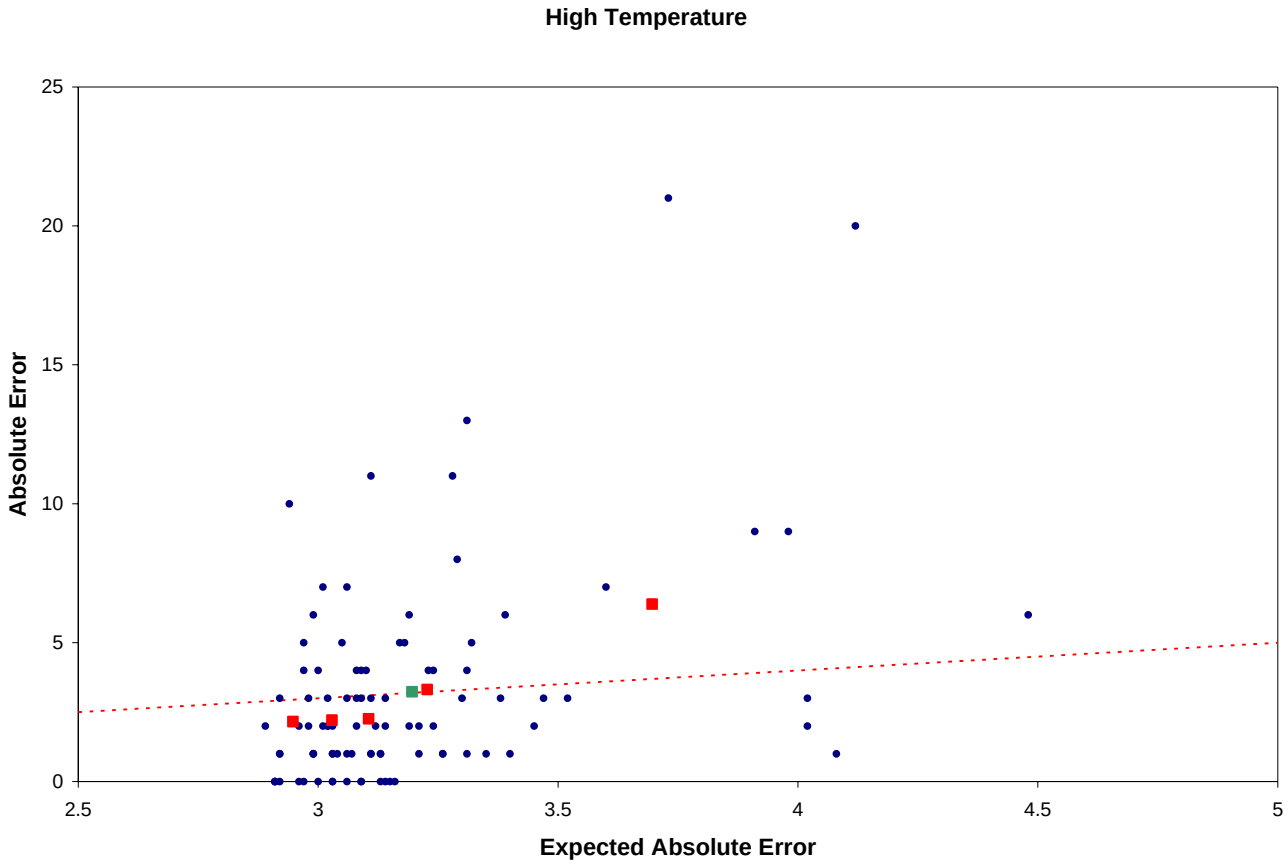

Figure 3: Comparison of the expected absolute errors (EAEs) and actual absolute errors (AE) for the GFS ensemble median high temperature forecasts. The blue circles denote the EAE-AE pairs for the 94 individual forecasts. One can organize these forecasts into 5 bins based on their EAE. The red squares are then the MEAE-MAE (M for mean) pairs for each of the five bins. If the bin MEAE matched the bin MAE, the red squares would fall near the dashed red line. The green square shows the MEAE-MAE pair for overall means.

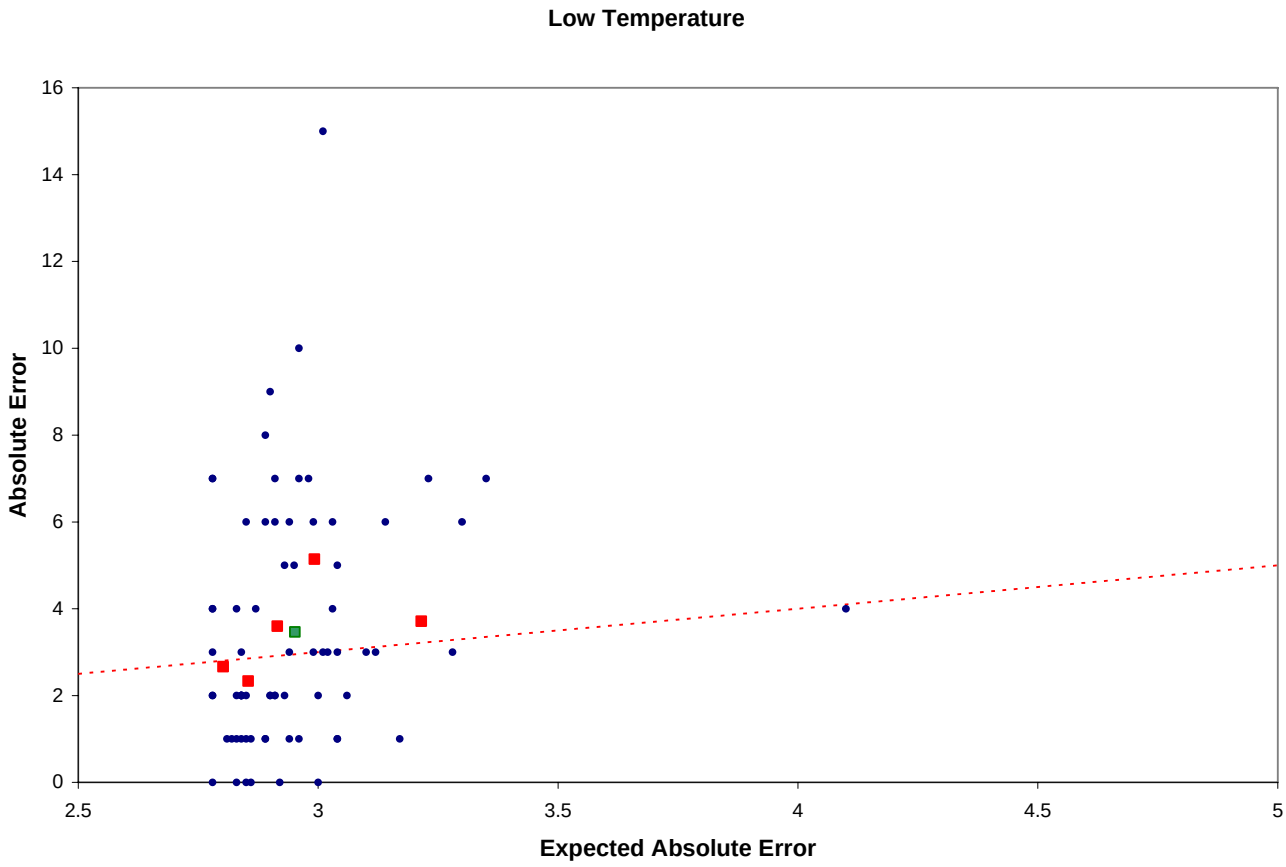

Figure 4: Like Figure 3, but for the low temperature forecasts. There are 73 individual EAE-AE realizations, represented by the blue circles.