Uncertainty in the Global Forecast System

Dr. David Werth

Dr. Alfred Garrett

Savannah River National Laboratory

Submitted to Monthly Weather Review

ABSTRACT

We validated one year of Global Forecast System (GFS) predictions of surface meteorological variables (wind speed, air temperature, dewpoint temperature, air pressure) over the entire planet for forecasts extending from zero hours into the future (an analysis) to 36 hours. Approximately 12,000 surface stations world-wide were included in this analysis. Root-Mean-Square- Errors (RMSE) increased as the forecast period increased from zero to 36 hours, but the initial RMSE were almost as large as the 36 hour forecast RMSE for all variables. Typical RMSE were 3ºC for air temperature, 2-3mb for sea-level pressure, 3.5°C for dewpoint temperature and 2.5 m/s for wind speed. Approximately 20-40% of the GFS errors can be attributed to a lack of resolution of local features.

We attribute the large initial RMSE for the zero hour forecasts to the inability of the GFS to resolve local terrain features that often dominate local weather conditions, e.g., mountain- valley circulations and sea and land breezes. Since the horizontal resolution of the GFS (about 1º of latitude and longitude) prevents it from simulating these locally-driven circulations, its performance will not improve until model resolution increases by a factor of 10 or more (from about 100 km to less than 10 km). Since this will not happen in the near future, an alternative for the near term to improve surface weather analyses and predictions for specific points in space and time would be implementation of a high-resolution, limited-area mesoscale atmospheric prediction model in regions of interest.

1. Introduction

The Global Forecast System (GFS) is a global climate model (GCM) developed at the National Center for Environmental Prediction (NCEP) (NCEP, 2003). The model is run continuously at NCEP and the products are placed online, providing users needing worldwide weather forecasts with a valuable resource. The GFS (formerly the AVN model) data has been used extensively (Saha et al, 2006; Hoffman and Leidner, 2005), and has been used to provide both forecast and analysis information for a variety of applications. These include the use of the data for the interpretation of surface information seen by satellites.

Given the widespread use of this data, it raises the question as to how good it is, and how its quality varies. The quality of a forecast is dependent on several factors- topography and proximity to the ocean, time of day and season. As these are known, (or vary predictably), knowledge of their effect on the forecast quality can allow us to determine the uncertainty in the GFS data. For example, Goff (2004) validated the GFS precipitation forecasts for the northeast US for the cold and warm seasons from 2 years and found that the quality depended on the season and the area being validated. The model data is at 1^{o} resolution, fairly fine for a GCM, but still unable to resolve many surface heterogeneities that can allow the meteorology to vary greatly within a small distance. The necessity of resolving such surface features was also highlighted in an article by Medvigy et al. (2008), in which they ran the OLAM global climate model and determined that a high-resolution mesh over the Andes is needed to properly reproduce the observed effect of ENSO on Amazon precipitation.

The resolution of small-scale variability is particularly important since small-scale atmospheric eddies can interact to produce large-scale features. Therefore, a model with a coarse resolution will not only miss small-scale events, but the large-scale events it can resolve will be in error as well. For example, Chen et al. (1990) ran the GLA forecast model at a fine (2^{o}x2.5^{o}) and coarser (4^{o}x5^{o}) resolution. Differences in large-scale stationary eddies were seen in the finer model, and these were linked to differences in the resolved tropical heating.

In order to determine the forecast quality, we apply a systematic validation algorithm to the GFS data to see how its quality varies throughout the year and over all locations. We expect that some variables are more difficult to predict than others, and that, as the values of the variables themselves undergo an annual cycle, so too may the errors in their prediction (though not necessarily the same cycle). Questions could also be asked as to how the resolution affects the consequent forecast accuracy, and how well the model does with respect to an unskilled forecast. To learn more about this, we obtained GFS data and station data against which to compare it. By quantifying the forecast quality for different times of the year, different times of the day, and different locations, we can determine where the forecast is weakest and what could be done to improve it.

2. The Global Forecast System and Integrated Surface Hourly (ISH) Station Data

The GFS (NCEP, 2003) comprises a T254 global spectral model that also uses a Gaussian grid of 768x385 (~0.5 degrees). It has 64 vertical layers, with smaller vertical spacing in the boundary layer (to resolve turbulent transfer) and at higher levels (to

resolve the dissipation of gravity waves). The model solves the primitive equations for vorticity, divergence, logarithm of surface pressure, specific humidity, virtual temperature, and cloud condensate as the dependent variables. The parameterization includes the determination of the momentum flux due to gravity waves at the surface, as well as at higher levels. The model uses a radiative transfer model with a correlated-k distribution (which solves for the cumulative effect of a large spectrum of absorption bands, saving time), and convection occurs when the cloud work function (the integrated buoyant instability) exceeds a certain threshold. The model also uses a K-theory PBL scheme in which eddy diffusivity (K) is a cubic function of boundary layer height.

The model data are obtained through NOAA's National Operational Model Archive and Distribution System (NOMADS) web site. We downloaded data from March 1, 2005 to Feb. 28, 2006, getting the 0, 18, and 36-hour forecasts, each made at 0Z, 6Z, 12Z and 18Z, and the data are at 1ºx1º resolution (coarser than the .5º resolution at which it is produced). The variables to be validated include 2 meter temperature, 2 meter dewpoint temperature, 10 meter winds, and sea-level pressure (SLP).

The data used to validate the GFS model are taken from global station readings that have been combined into a single data source. The Integrated Surface Hourly (ISH) dataset[1] was compiled by the National Climatic Data Center (NCDC), and includes data from approximately 12,464 stations. Not every station reports the weather every hour, however, so for any particular time, only about 6,000 stations are used. The stations report data for temperature, dewpoint, wind speed, and sea-level pressure. We have station data for 2005 and 2006, so these will be used to validate the GFS data.

[1] http://gcmd.nasa.gov/records/GCMD_gov.noaa.ncdc.C00532.html

3. Validation Procedure

The procedure for validating the GFS is as follows. First, the forecasts are grouped into the 0-hour, 18-hour and 36-hour forecasts (henceforth the forecast periods). Then, for each forecast, an algorithm is run which loops through each model forecast time (that is, the time the forecast was made) – 0z, 6z, 12z, 18z, for each day of the year. When a particular GFS time is selected, (e.g., the 18 hour forecast made at 6z on March 1, 2005), a file containing the ISH station data for the corresponding validation time is opened (in this case, the station data file for March 2, 2005 at 0z).

The program then loops over all stations in the ISH file, reading in the desired variables (rejecting stations with missing data), and noting the latitude and longitude of each. At each station, the four nearest GFS gridpoints that surround the station are identified, and the GFS data are converted so that they have common units with the station data. The GFS data are then interpolated to the station location using a bilinear interpolation (Raytheon, 2004), and the difference between the station value and the interpolated GFS value is squared and added to a running sum. The program then moves on to the next station, repeating the procedure. After all stations have been interrogated, the root mean square error (RMSE) value is calculated from the total sum of squares and the number of stations that were included. As each of the four forecast times for each day are assessed, we will have a time series that shows how the forecast quality varies throughout the year.

4. Results

a. Global Error Patterns

When calculated over the whole world, the annual and diurnal cycles should theoretically be less defined as each time will include stations under both day and night conditions, as well as experiencing different seasons. The world's stations are not distributed uniformly, however, so we may expect some variability in forecast quality during the year.

On a global scale, the RMSE temperature values for the 0Z forecasts show a strong downward trend from March to September (Fig.1), falling from 3.5K to 2.9K. The values remain relatively flat in September until they rise sharply in mid-November, reaching about 3.5K before March of the following year. A distinct layering of values for the different forecast periods can be detected, with the 0-hour forecast having the lowest errors and the 36-hour forecast the highest. The differences indicate the magnitude of the forecast 'correction' as we get closer to the validation time. The 'spikes' seen on Sep. 1 and Jan. 10 (and elsewhere) are due to there being only a few (~400) stations available at those times).

In the global mean (Fig. 2a), the stations are on average hotter from 12Z (~7pm in Europe) to 18Z (~12pm in North America). The model biases are generally too cool (Fig. 2b) except for the 6Z forecast from June to March. We also see from this that the annual cycle of the errors differs for forecasts initialized at different times of the day, with the values in April through October (the warmer time of the year in the global mean, which is dominated by stations in the Northern Hemisphere) highest at 18z (Fig. 2b). The model

biases for 18Z are too cool (Fig. 2b), so the model seems to be having difficulty reaching the warmer daytime maximum temperatures.

When the errors are averaged for the 3-month periods, we can compare the values for the different seasons as well as the different forecast periods. The temperature errors (Table 1a) show a clear (but small) increase as the forecast period increases, but also show a clear annual cycle, with larger values in DJF and smaller values in SON. The differences between the 36- and 0-hour forecasts are also larger for DJF (.29K) relative to SON (.18K).

The RMSE values of wind speed also show a strong annual cycle, and are generally lower in the warmer part of the year. Similar to temperature, the errors in 10m wind speed start at relatively high values (2.7m/s) in March and fall to about 2.3m/s in early June (Fig. 3). They are fairly steady until September and then rise until December, and remain high (2.6m/s) until March. This cycle is driven by the annual cycle in global mean wind speed (Fig. 4a) – speeds fall from March to June, then rise until the following March. We again see that the 0-hour analysis is superior to the 18 and 36-hour forecasts, but also that the differences are small (~0.2m/s). The model bias has an annual cycle that leads the annual cycle in wind speed (Fig. 4b), and is lowest at 12-18Z and greatest at 0-6Z.

When averaged over the entire period, the forecast errors (Table 2) reveal that the model does best in JJA and worst in DJF, in accordance with Fig. 3. We again see how the increasing forecast period degrades the forecast quality, with the largest increase in RMSE for MAM.

The SLP (Fig. 5) has a tendency towards larger errors during both the warm and cold seasons. They start from about 2-3mb in March, fall until May, rise in July, fall again until September, and rise as March approaches. Values are mostly below 3mb (Table 3). The annual cycle of SLP falls to a low in the NH summer and has its maximum in winter (Fig. 6a), and the bias has an opposite cycle (Fig. 6b) – the SLP values are too high during the minimum and too low during the maximum. This implies that the annual cycle of interhemispheric movement of air is too weak in the GFS. Table 3 shows that the maximum error is during DJF for the 0-hour forecast, while the greatest degradation (increase in RMSE) is during MAM (~0.78mb).

The dewpoint error cycle (Fig. 7) resembles that for temperature (Fig. 1), but starts with high values (4.2K), which fall in March to 3.8K. They rise to 4.2K in April and fall again to 3.4K by June 1. They hold steady at that value until November, when they rise to about 4.3K in January before falling and rising again in February. The annual cycle (Fig. 8a) has a maximum in JJA, and the model does worst during the drier part of the cycle (DJF). Similarly to the SLP, the model bias (Fig. 8b) is out of phase with (and, in this case, lags) the annual cycle – the model is too wet during the dry season and vice versa. This indicates that either the model precipitation/evaporation cycle is not responding to the annual forcing, the interhemispheric transport of moisture is inadequate or numerical diffusion is creating unphysical changes in water vapor, and this is resulting in an inadequate spread in dewpoint values. Also note that, as with temperature, the errors in the warmer months become greater as the forecast time shifts from 0-6z to 12-18z. Table 4 shows how the errors increase with forecast period, and tend to be largest in DJF (during which we see the greatest degradation).

Fig. 9 shows the distribution of all stations, with the stations with the largest RMSE values (averaged throughout the year) for the 0Z forecast highlighted (the maps for 12Z look similar). We see that the temperature errors (Fig. 9a) are greatest in mountain regions- the Rockies, the Alps, the Andes, and the Himalayas are all clearly visible in the plot. Mountain regions appear to create difficulties for the model when predicting SLP (Fig. 9c) and dewpoint (Fig. 9d) as well. For the speed errors (Fig. 9b), however, coastal regions seem to dominate, suggesting that resolution of the sea-breeze or the poor resolution of the surface roughness discontinuity could be the problem.

b. Comparison to Persistence

Quantitative forecast validations are usually expressed as the relative improvement over an unskilled reference forecast – usually a 'forecast' that involves the assumption that the past climate represents the future climate. One such forecast is persistence – the assumption that the tomorrow's weather will simply be a repeat of today's. Mittenmaier (2008) showed that persistence represents a sterner reference for testing forecast accuracy than a simple random forecast.

Using the station data, we can create a persistence forecast – the meteorological values for any time will be the same as the most recent validated time. For example, the 18Z temperature tomorrow will be the same as today's 18Z temperature, and the 18Z temperature two days from now will also be assigned today's 18Z temperature.

Can this forecast method do as well as the GFS, indicating that the latter has no skill? We can compare the 18-hr forecast to the 24-hour persistence, and the 36-hr forecast to the 48-hour persistence by calculating the RMSE values for the persistence

forecast and comparing them to the GFS forecast RMSE values. A measure of forecast skill that makes use of both RMSE values and a reference forecast is the skill score SS (Murphy and Epstein, 1989):

$$SS = 1 - \frac{MSE_{fcst}}{MSE_{ref}}$$

where MSE_{fcst} is the mean square error (the RMSE squared) of the actual forecast, and MSE_{ref} is that of the reference. An SS value of 1 indicates perfect model skill, while values of 0 (or below) indicate no forecast skill. Murphy and Epstein (1989) used climatology as the reference, but Mittenmaier (2008) suggested substituting persistence. Doing this, we can calculate the SS score for the 18hr and 36hr model forecasts and determine how skillful they are.

For temperature (Table 1), the model SS is highest in DJF and lowest in JJA, and we see that the 36 hour forecast is more skillful than the 18 hour forecast – though the RMSE values are larger, persistence is a worse forecast for this longer forecast period, making the SS value larger. The skill of the speed forecasts (Table 2) is about the same all year round, and changes little for longer forecast periods. The SLP SS values (Table 3) are generally higher than for the other variables, and have an annual cycle similar to that for temperature. They also increase for the longer period forecast. For dewpoint temperature (Table 4), the SS values are slightly lower than those for temperature, suggestive of the greater difficulty in forecasting this variable.

c. Effect of Resolution

The data on the model grid will tend to have a characteristic length-scale, which can vary among the model variables but never fall below the model grid spacing. The fact that station data varies on scales shorter than this is responsible for at least part of the model error. This can be seen in the fact that forecasts of fields dominated by small-scale spatial variability (e.g., temperature) are generally less accurate than those dominated by large-scale variability (e.g., SLP).

If the station data were fit to a grid identical to that of the forecast field, the resulting length scale would then be closer to that of the model, and the elimination of small-scale variability from the observed field should therefore yield lower errors when the model and observations are compared. The resulting change in forecast quality (with respect to that done by interpolating the model to the station locations) could then give us an idea as to how much the forecast errors are due to resolution issues, and how much are due to other problems.

We will repeat the 0-hr validation, but now fitting the station data to the same grid used by the forecast fields. To do the fitting, we elect to use a one-pass Cressman analysis scheme (Cressman, 1959). In this scheme, an initial guess field is assumed, and the errors between this first guess and the station values (which requires interpolation of the first guess to the stations) are calculated. The correction to each grid point is then calculated as a weighted average of the station errors within a given distance (D) from that gridpoint. The weighting is calculated as:

$$w_j = \frac{D^2 - d_j^2}{D^2 + d_j^2} \qquad (1)$$

where d_j is the distance from the gridpoint to the station. Each gridpoint value is corrected, and the resulting grid field is then used as the guess field for the next pass. With each new pass, the reference distance is reduced.

We will apply the scheme without the successive corrections – we assume a first guess field of zero, then apply the Cressman scheme once to fit the station data to the forecast grid. Once the analysis is done, we compare the forecast and the station analysis, looping over all gridpoints and calculating the RMSE values as before. By doing this repeatedly with different values of the reference distance, we can see which D value gives us the most improvement in the forecast. To maintain a fair comparison between error scores calculated at the largest value of D (300km) and scores calculated with the smallest value (50km), the number of validated gridpoints should be the same (in our analysis, only gridpoints with at least 2 stations within the distance D are counted in the error calculation). To control for the fact that, at lower values of D, fewer gridpoints will 'qualify' as having the minimum number of stations within the required distance, only gridpoints that qualify at D=50km will be included in the analysis for larger values of D, so the same gridpoints will be validated for all values of D.

As D is increased, the analysis at the gridpoints will be influenced by stations further and further away. If D is too large, stations that are uncorrelated with nearby stations will be included in and degrade the analysis. If D is too small, the analysis is susceptible to the influence of only a few stations and will fail to smooth out the small-scale variability that the GFS cannot capture. Both of these will result in a poorer RMSE score. If D is close to the scale at which the model data varies, however, the match between model and observed length scales will result in a better RMSE score.

For example, Fig. 10a shows that we get the lowest temperature errors if we create an observational analysis by averaging the station data within a 100-150km radius of each grid point, and this holds true for all seasons. This implies a characteristic length scale over which the GFS temperature forecast varies, since the model comparison is best by averaging the observed data to that same scale. Because the 1º GFS model is of roughly 100km resolution, this is about the minimum we could expect, so the model temperature field is at its maximum spatial variability. These values are about 15-20% smaller than the values in Table 1, which can represent the fraction of the error due to inadequate resolution. The greatest improvement (both in a relative and an absolute sense) is in SON (when the RMSE values are smallest), with smaller improvements in DJF (when the RMSE values are largest), suggesting that resolution is more of an issue in SON.

For the wind speed (Fig. 10b), the lowest errors come with D=200km, implying that the GFS speed field varies at larger scales than temperature, even though the model can resolve finer scales. This is especially true given that we get little increase in error when we average at 300km, suggesting that the model wind speeds vary little over the latter scale. These values represent an approximately 20-30% improvement in score, pointing even more to a model resolution problem. Notice once again that large improvements come when the errors are generally lowest (JJA), and the weakest improvements happen during the period when errors are highest (DJF), implying the effect of model resolution can vary during the year.

The SLP (Fig. 10c), which is coupled to wind speed on large scales, also sees the most improvement for D=250-300km for all seasons, and the values improve the forecast

by up to 45% for SON. Given the fact that this variable varies on large spatial scales and is therefore better resolved by the GCM, this implies that the model dynamical errors are smaller as well, making resolution errors a larger fraction of the total. Note however, that the annual cycle is different than the other variables - the greatest improvement happens during the period with the worst RMSE scores (DJF, see Table 1), and the lowest improvement occurs when RMSE scores are smallest (MAM). This implies that the resolution is the largest problem in DJF for SLP.

For Td (Fig. 10d), we see the lowest errors at 150-200km, and also shows little degradation at larger scales (again implying lower-than-expected spatial variability in this simulated variable). We see the largest improvements (~27%) for the NH summer months JJA and SON (when the forecast is best) at D=150km (similar to temperature).

Of course, the model forecast has not gotten any better, only the RMSE values. This highlights the importance of selecting a proper forecast metric - the Cressman scheme is useful for validating a model when it is the intent to forecast features at coarser resolution. Also, the use of this scheme demonstrates how much of the model error is due to resolution (in this case, between 15-45%), and how much is due to errors in the actual forecasting or in the data assimilation (both of which are used to create the 0 hour analysis). The use of a Cressman-type analysis is likely the best indicator of the true "model skill", that is, what the model *can* predict.

6. Conclusions

This process has shown the magnitude of errors expected when the GFS forecast system is used to analyze or predict the surface meteorology. We have seen that the

global and local errors vary throughout the year. It is not clear why the forecasts vary the way they do, particularly with regards to the annual variability. For example, why do the global temperature errors reach a minimum in one season (SON) and a maximum the following season (DJF)? And why does the minimum occur in SON and not MAM, which should have similar weather?

Ultimately, the value of any forecast is in its contribution to the decision-making process. At a 1º resolution, the GFS forecasts are adequate for forecasting large-scale weather systems, but will necessarily have problems when used to predict smaller-scale features, such as the meteorological values at individual stations where local terrain variability, e.g., a river valley, significantly affect surface wind speed and direction, and temperature to a lesser degree. Since the GFS cannot resolve these local terrain features, it cannot simulate their effects on surface weather. The results of the error analyses presented in this report make it clear that local variability is extremely important at the majority of the stations analyzed around the world, because in almost all cases the RMS error for the 36 hour forecast was only a little larger than the 00 hour forecast.

The method often used to evaluate forecast model skill validates model predictions at individual nodes by combining the measured surface variables, e.g., temperature, via a distance weighted average for all stations within the area represented by the individual model node. This procedure averages out the local terrain effects at individual stations and thus produces a validation data set that is appropriate for assessment of model predictions, which are necessarily space and time averaged. As computer power increases, model resolution will continue to improve, so the accuracy of these types of large-scale model predictions should continue to get more accurate.

7. References

Chen, T.C., J.M. Chen, and J. Pfaendtner, 1990: The Effect of Horizontal Resolution on Systematic Errors of the GLA Forecast Model. *Mon. Wea. Rev.*, **118**, 1371–1378.

Cressman, G., 1959: An operational objective analysis system. Mon. Wea. Rev., 87, 364-374

Goff, J., 2004: Reliability Trends of the Global Forecast System Model Output Statistical Guidance in the Northeastern US: A Statistical Analysis with Operational Forecasting Applications, Eastern Region technical Attachment No. 2004-05, October, 2004

Hoffman, R., and S, Leidner, 2005: An Introduction to the Near–Real–Time QuikSCAT Data, Weather and Forecasting, 4, 476-493

Medvigy, D., R. Walko, and R. Avissar, 1008: Modeling interannual variability of the Amazon hydroclimate, *Geo. Rev. Lett*., **35**, doi:1029/2008GL034941

Mittermaier, M.P., 2008: The Potential Impact of Using Persistence as a Reference Forecast on Perceived Forecast Skill. *Wea. Forecasting*, **23**, 1022–1031.

NCEP, 2003: NCEP Office Note 442, The GFS Atmospheric Model, November, 2003, National Center for Environmental Prediction, Global Climate and Weather Modeling Branch, EMC, Camp Springs, Maryland

Raytheon, 2004: A006 Technical Report – Software Analysis for Point Analysis Reengineering, Advanced Technology Support Program II, Raytheon Technical Services Company, 1610 Hughes Way, Long Beach, CA 90810

Saha, S., S. Nadiga, C. Thiaw, and J. Wang, W. Wang, Q. Zhang, H. M. Van den Dool, H.-L. Pan, S. Moorthi, D. Behringer, D. Stokes, M. Peña, S. Lord, G. White, W. Ebisuzaki, P. Peng, and P. Xie, 2006: The NCEP Climate Forecast System, J. Clim, 15, 3483-3517

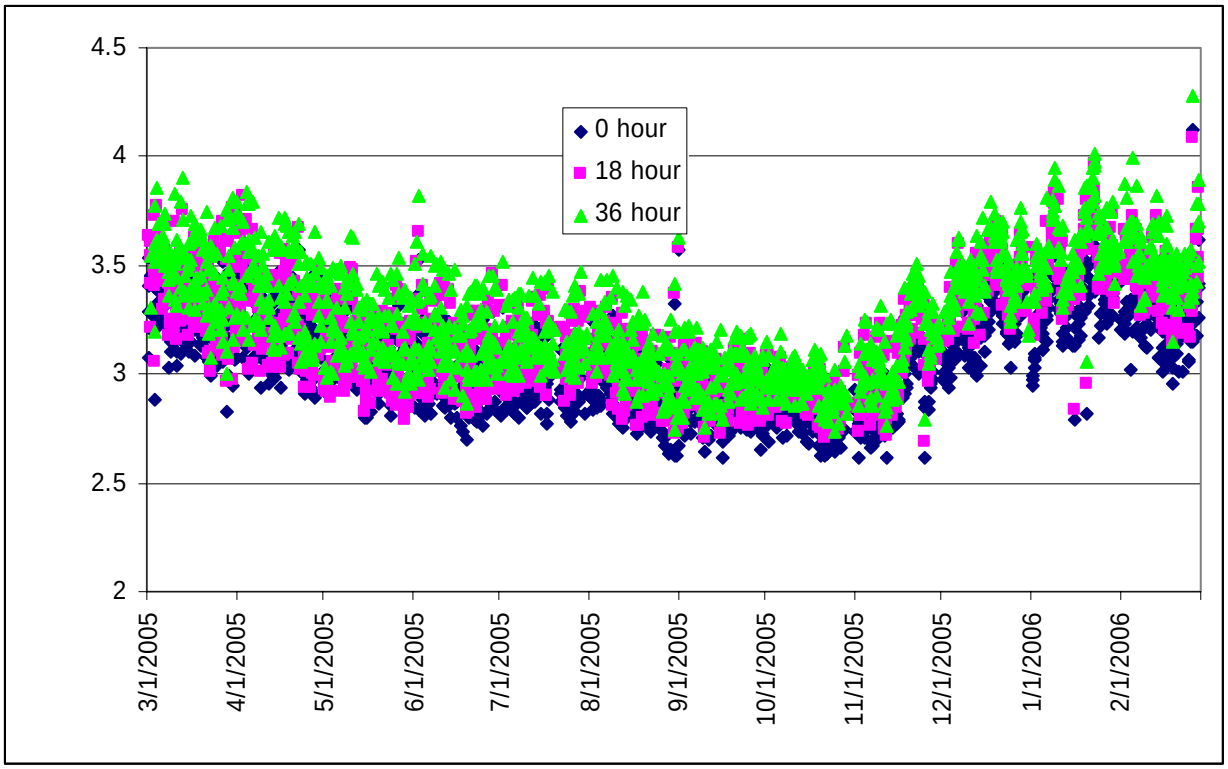
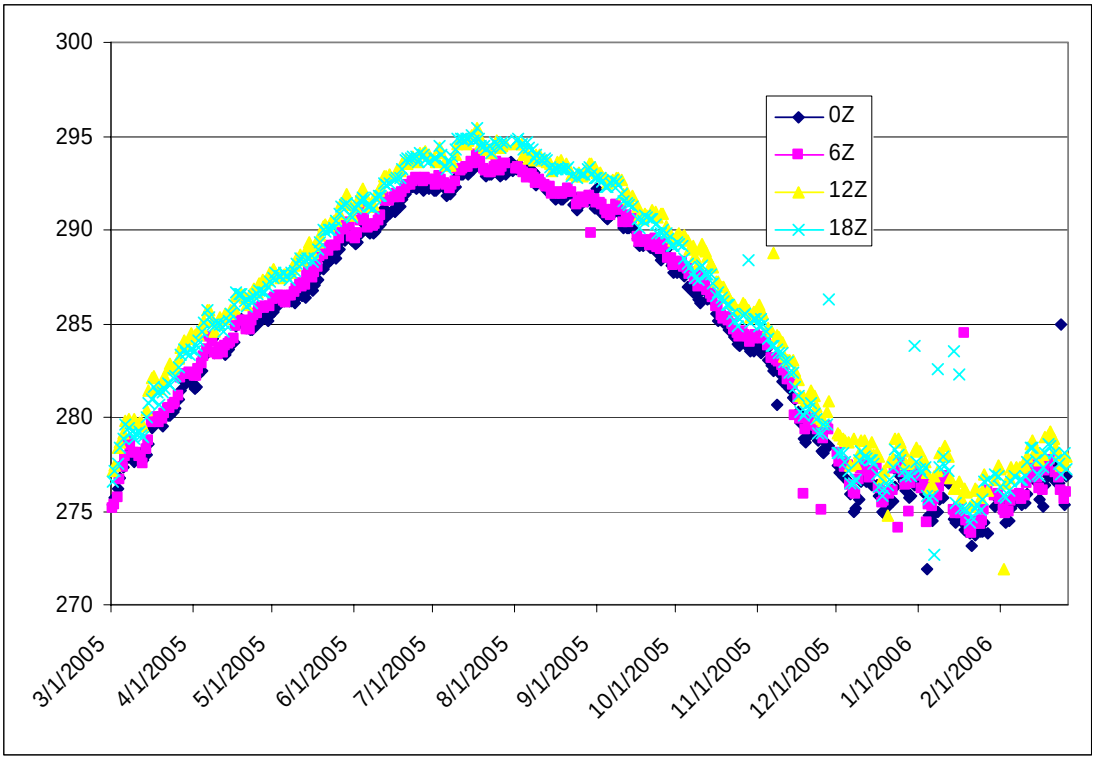


Figure 1 Global RMSE Temperature error (K), for the 2005/2006. Forecast periods are 0 hours (blue), 18 hours (pink), and 36 hours (green).

a)



b)

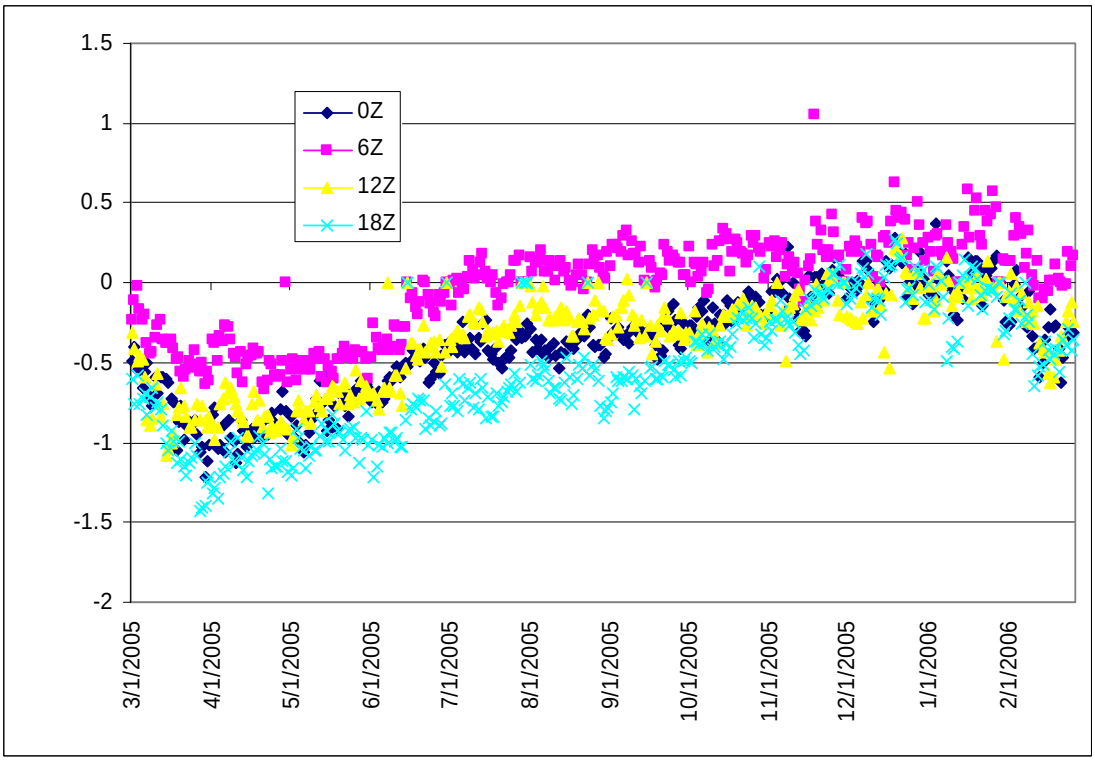


Fig. 2 a) Temperature station mean b) 0hr forecast bias (K).

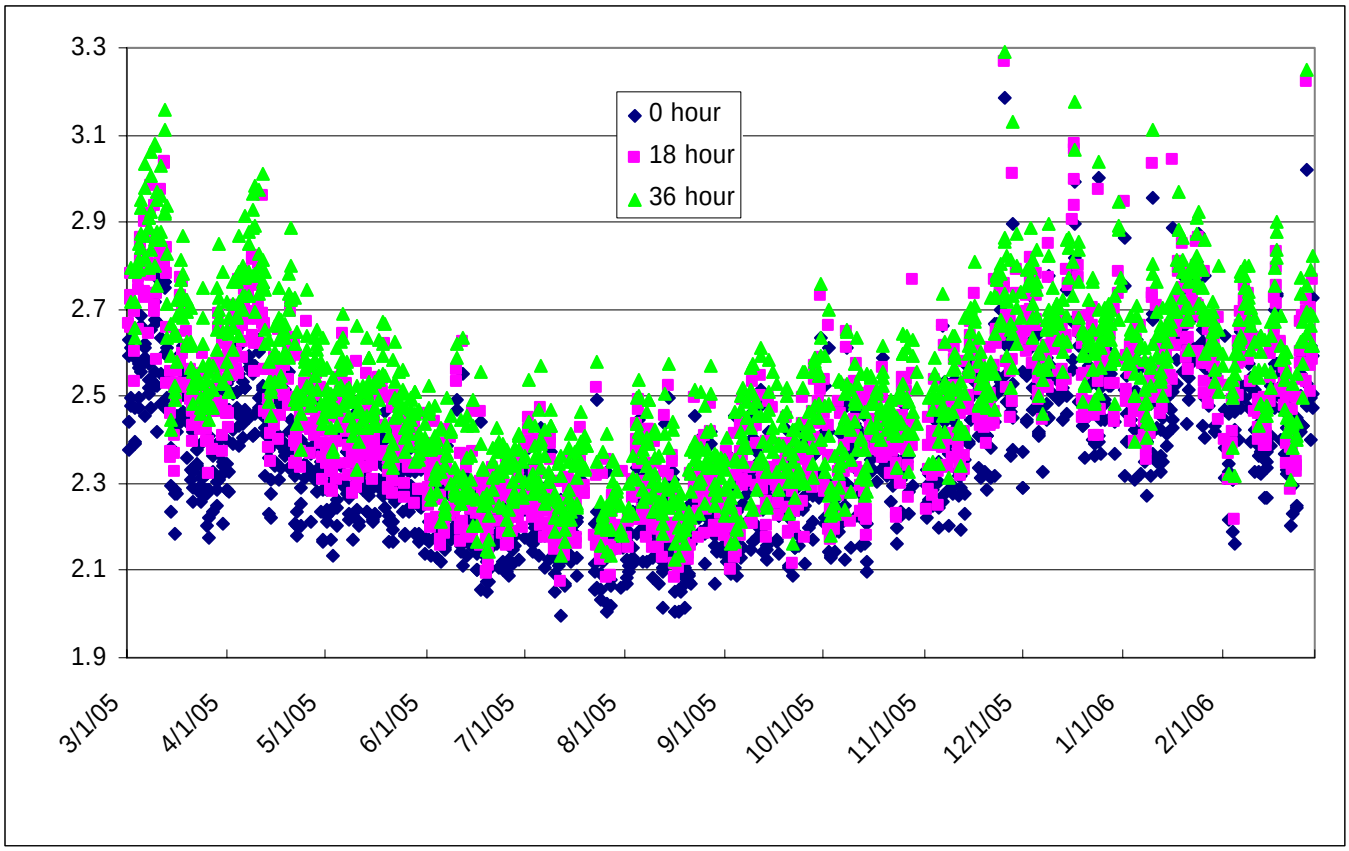
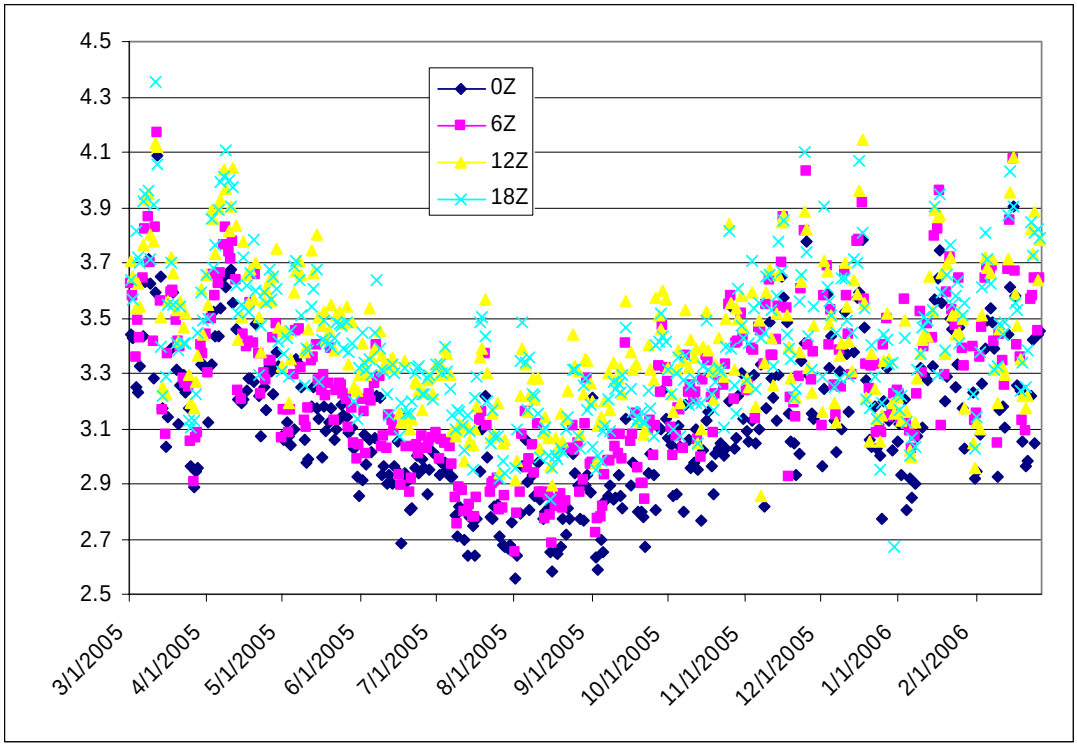


Figure 3 Global RMSE Speed Error (m/s), for 2005/2006. Forecast periods are 0 hours (blue), 18 hours (pink), and 36 hours (green).

a)



b)

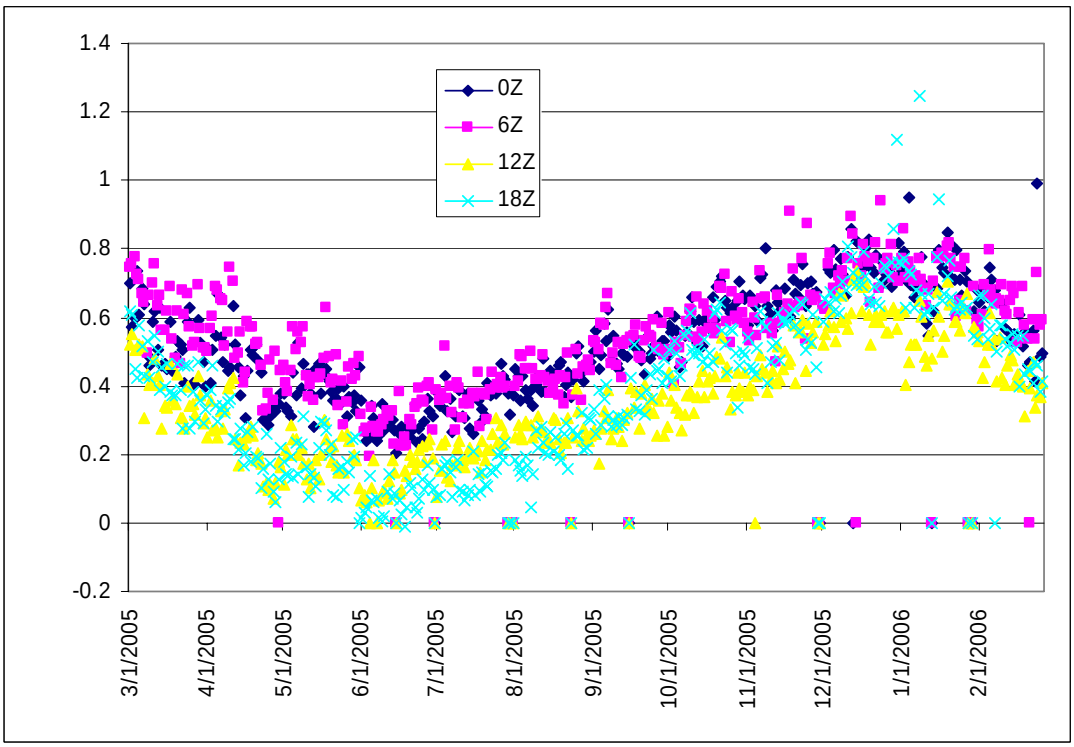


Fig. 4 a) Station mean speed b) 0hr wind speed forecast bias (m/s).

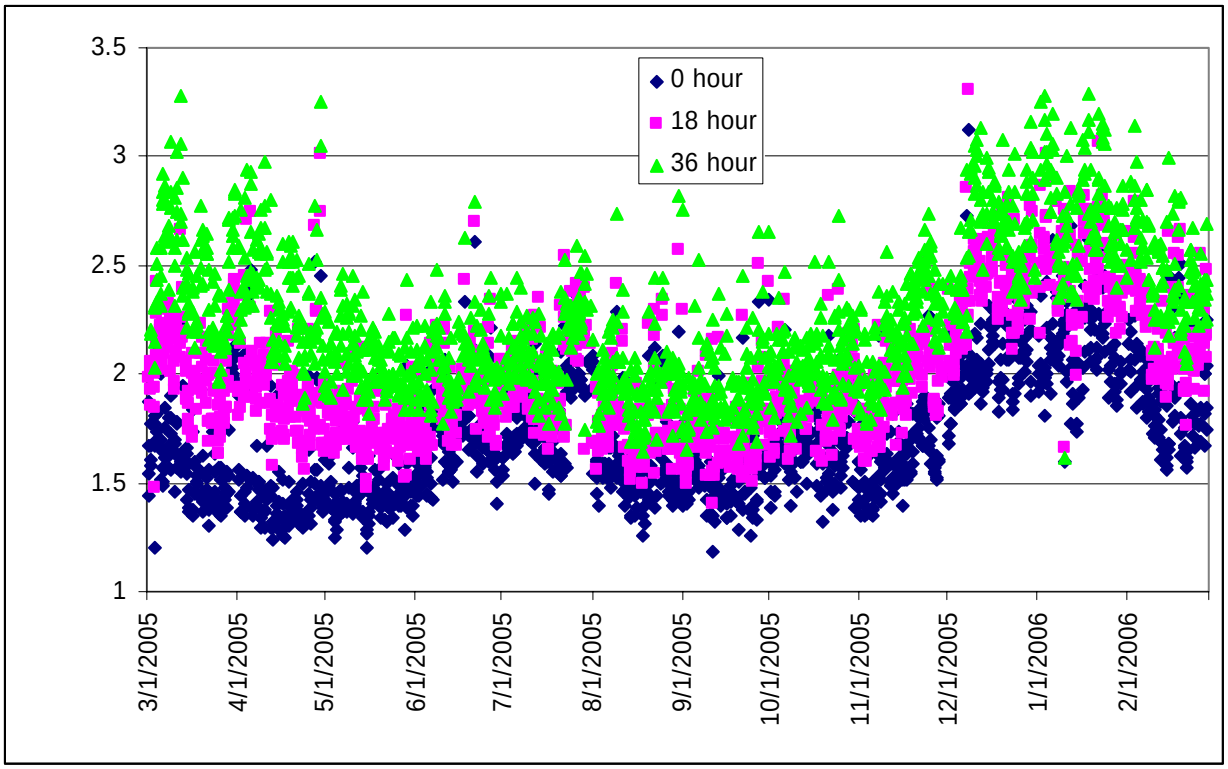
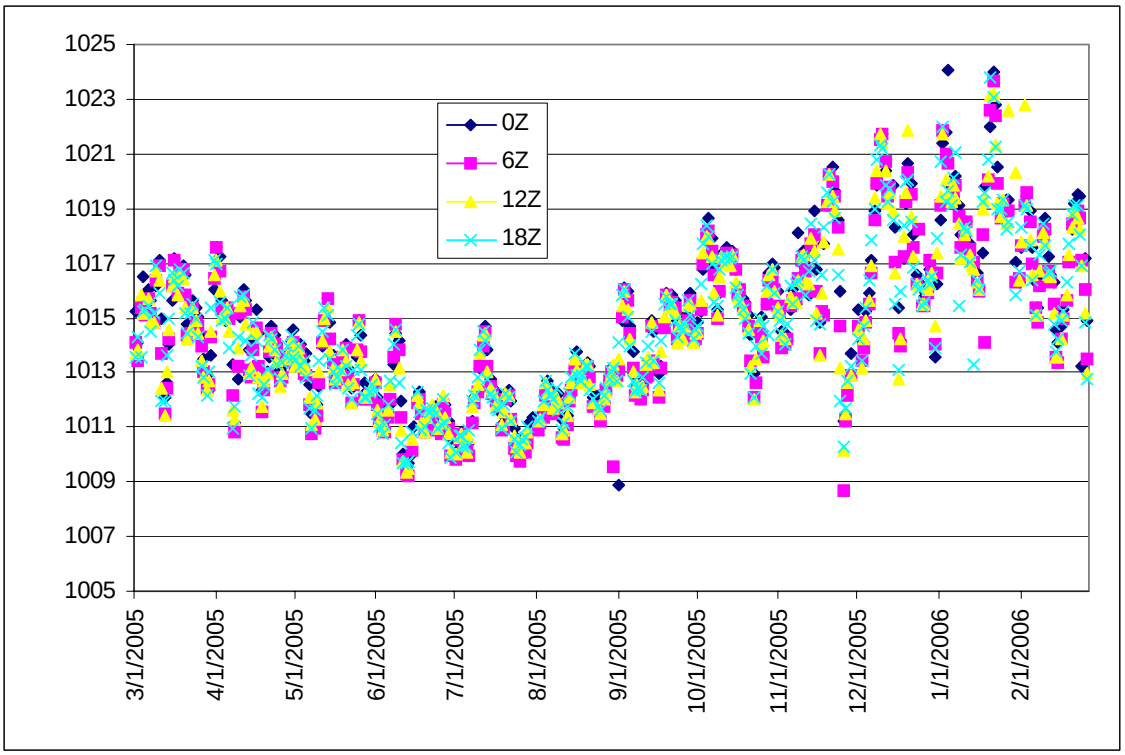


Figure 5 Global RMSE SLP error (mb), for the a) 0Z, b) 6Z, c) 12Z, and d) 18Z forecast. Forecast periods are 0 hours (blue), 18 hours (pink), and 36 hours (green).

a)



b)

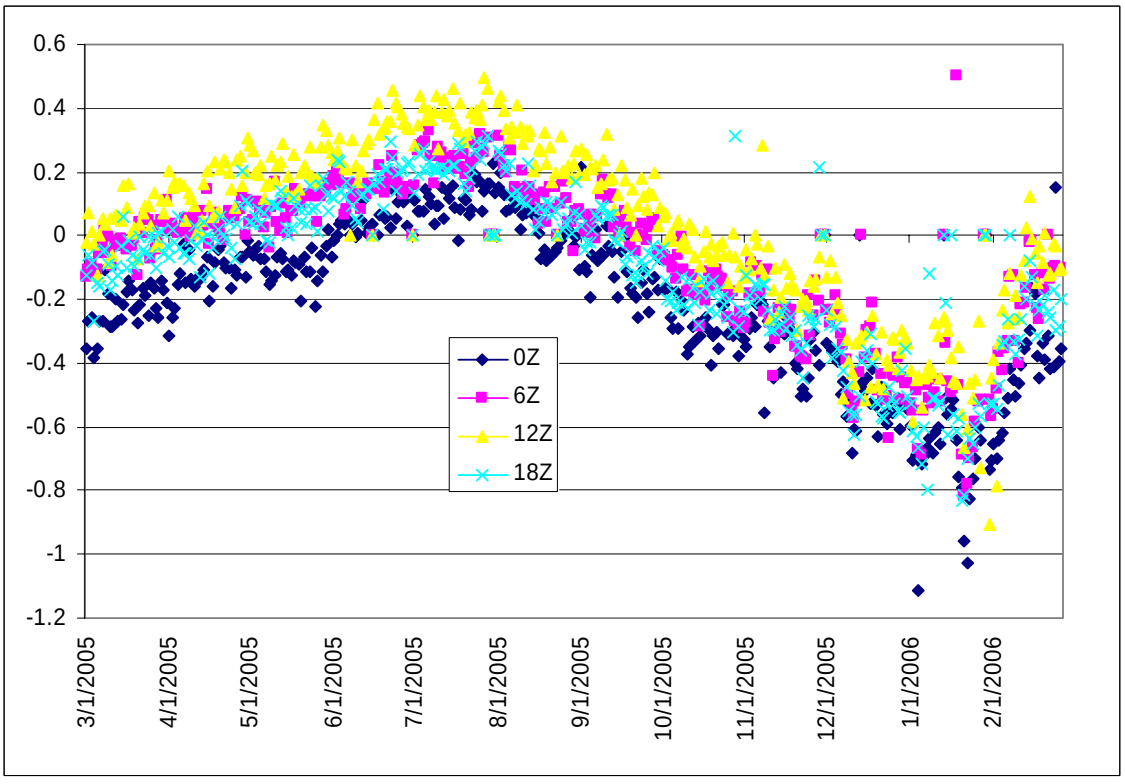


Fig. 6 a) Station mean SLP b) 0hr SLP forecast bias (mb).

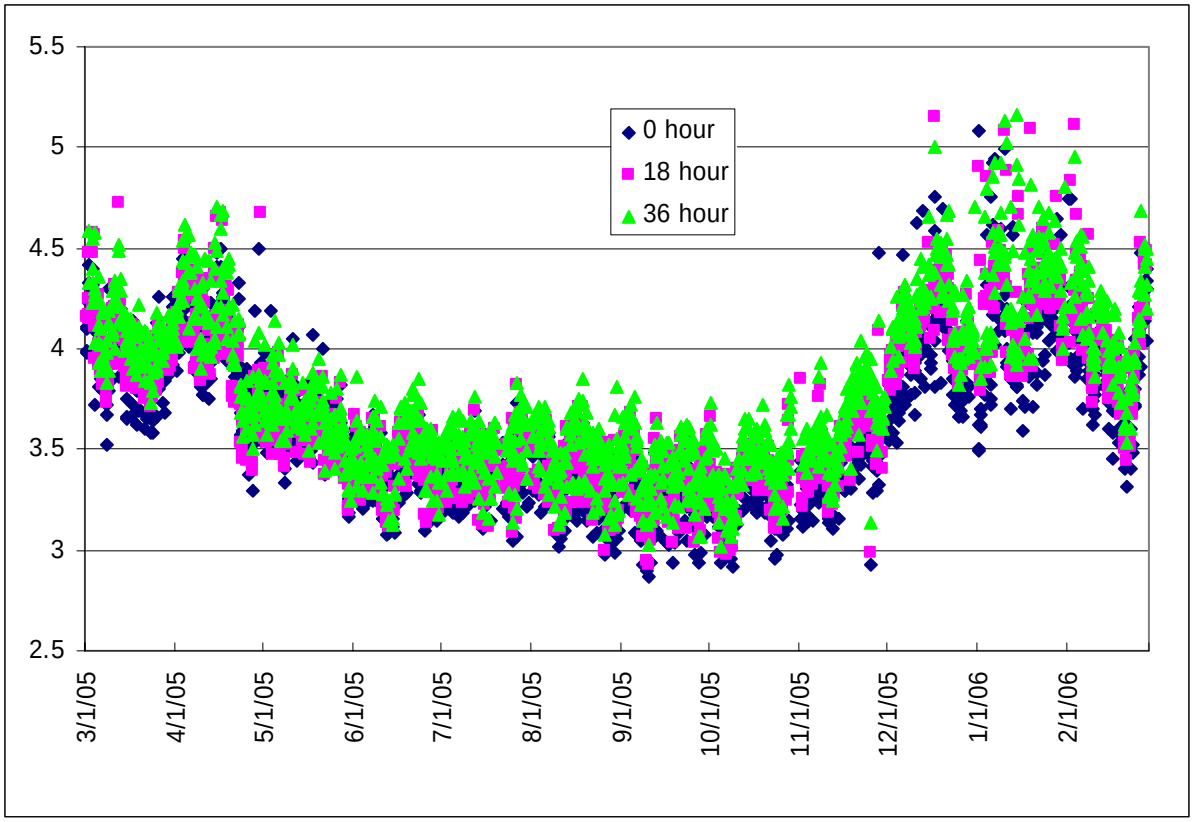


Figure 7 Global RMSE dewpoint errors (K), for the a) 0Z, b) 6Z, c) 12Z, and d) 18Z forecast. Forecast periods are 0 hours (blue), 18 hours (pink), and 36 hours (green).

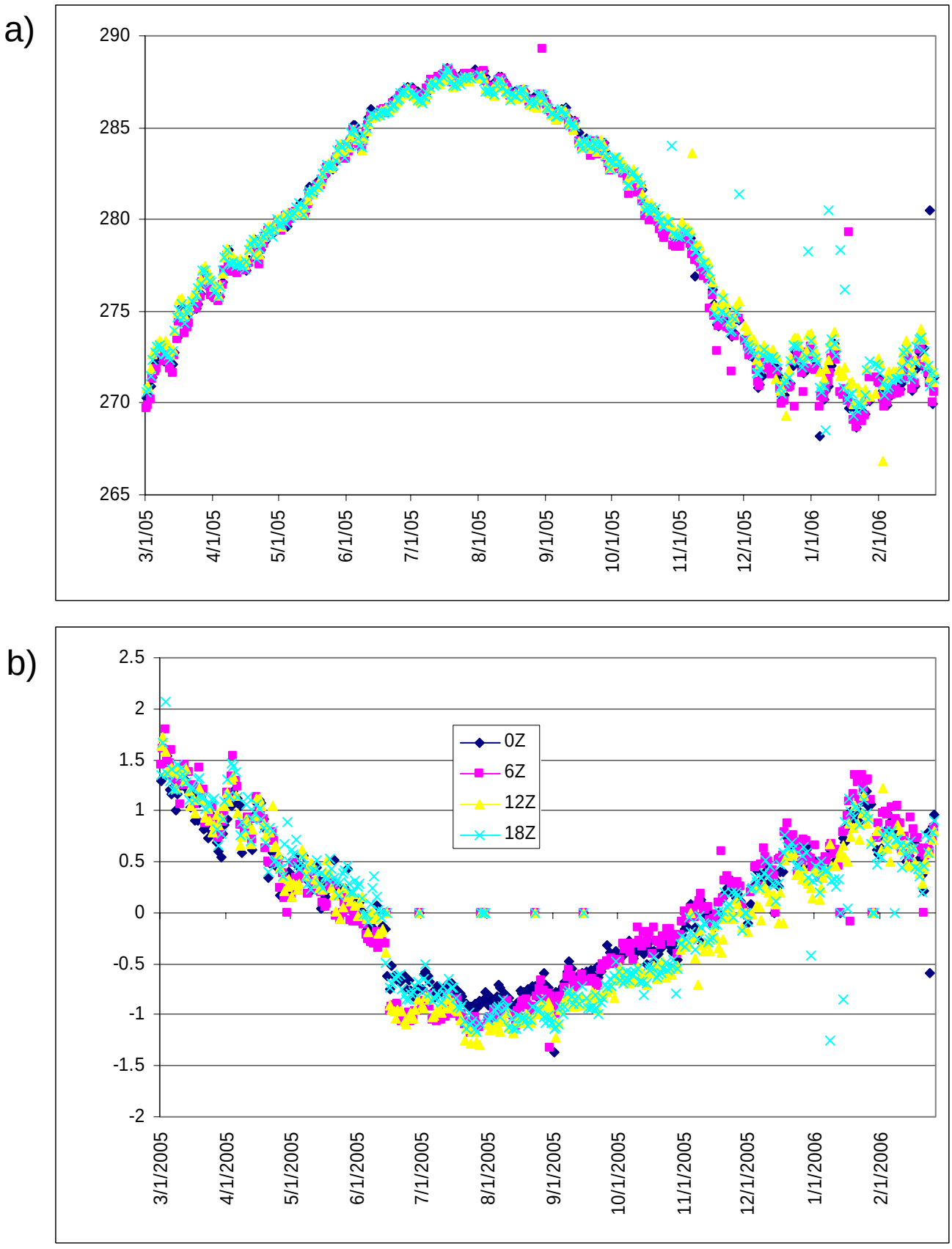
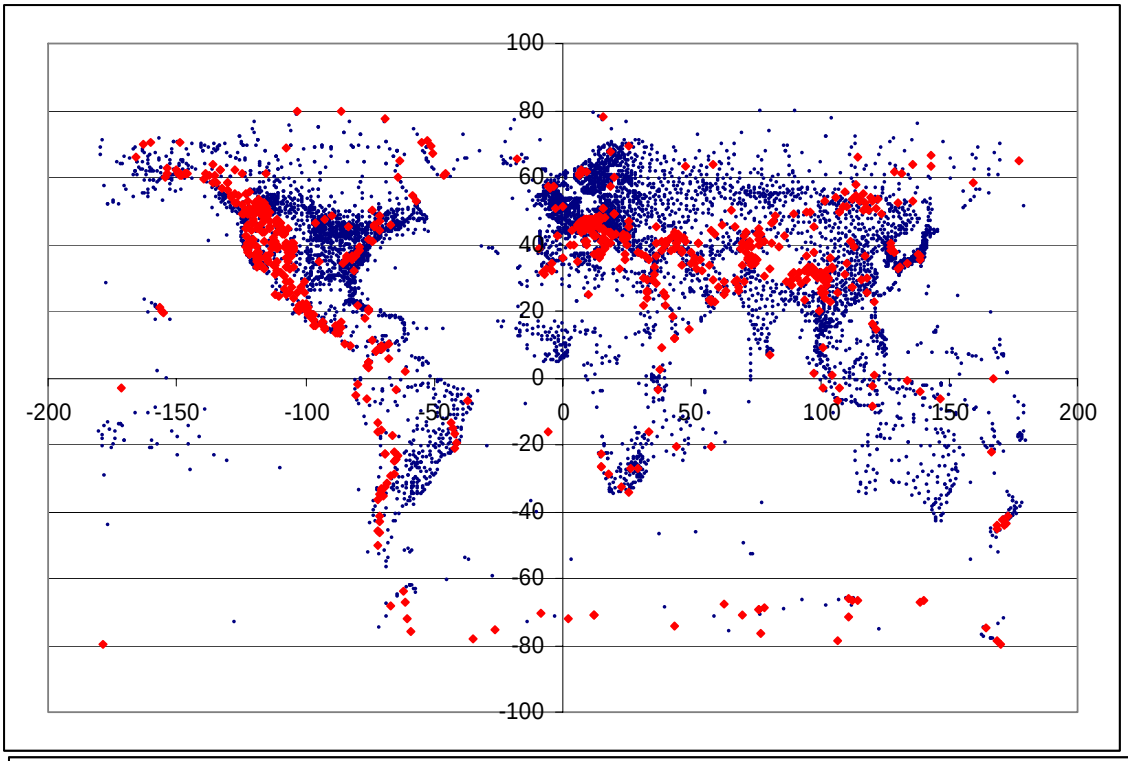


Fig. 8 a) Station mean Td, b) 0hr Td forecast bias (K).

a)



b)

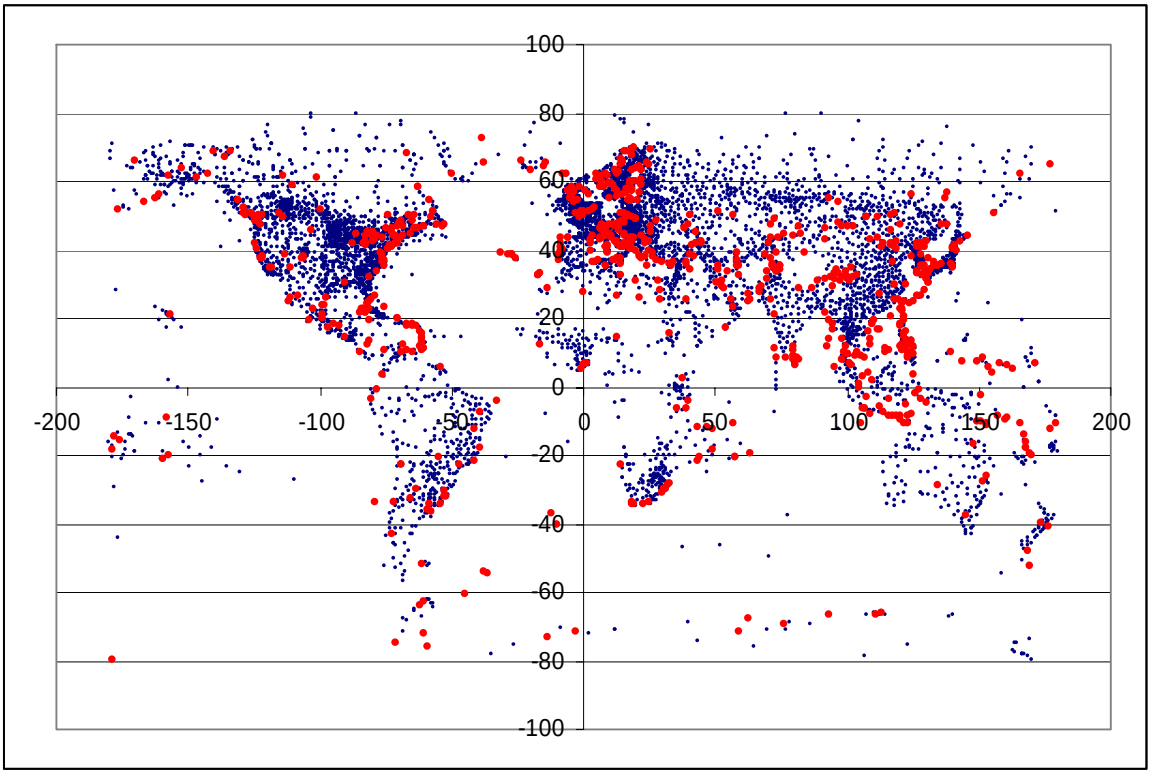
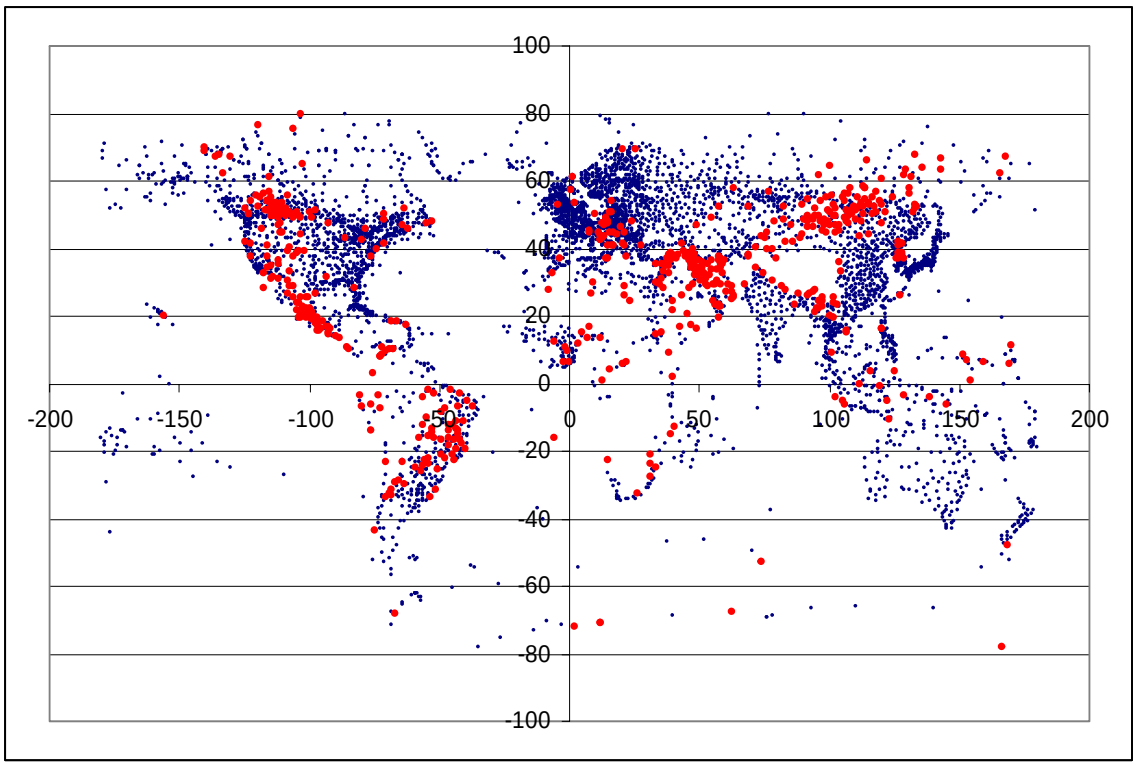
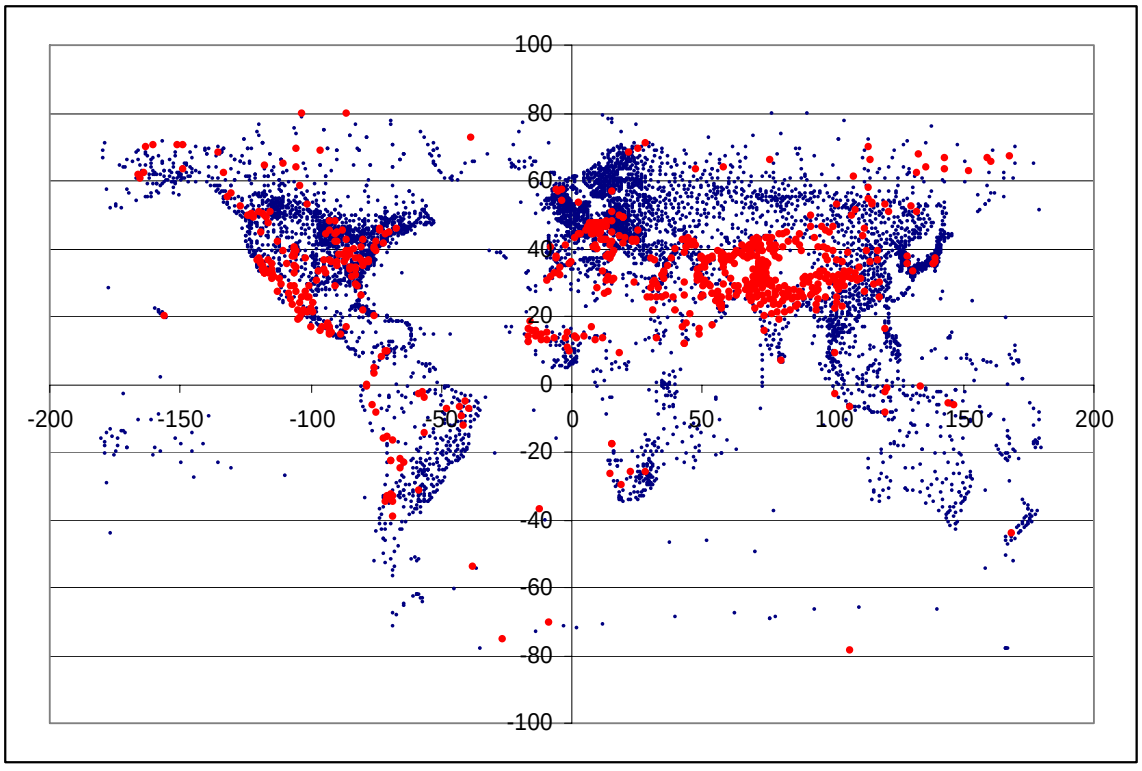


Fig. 9 Map of worst 10% of stations as determined by the mean RMSE 0hr 0Z values for a) temperature, b) speed, c) SLP, and d) dewpoint.

c)


100
80
60
40
20
0
-20
-40
-60
-80
-100
-200
-150
-100
-50
0
50
100
150
200

d)


100
80
60
40
20
0
-20
-40
-60
-80
-100
-200
-150
-100
-50
0
50
100
150
200

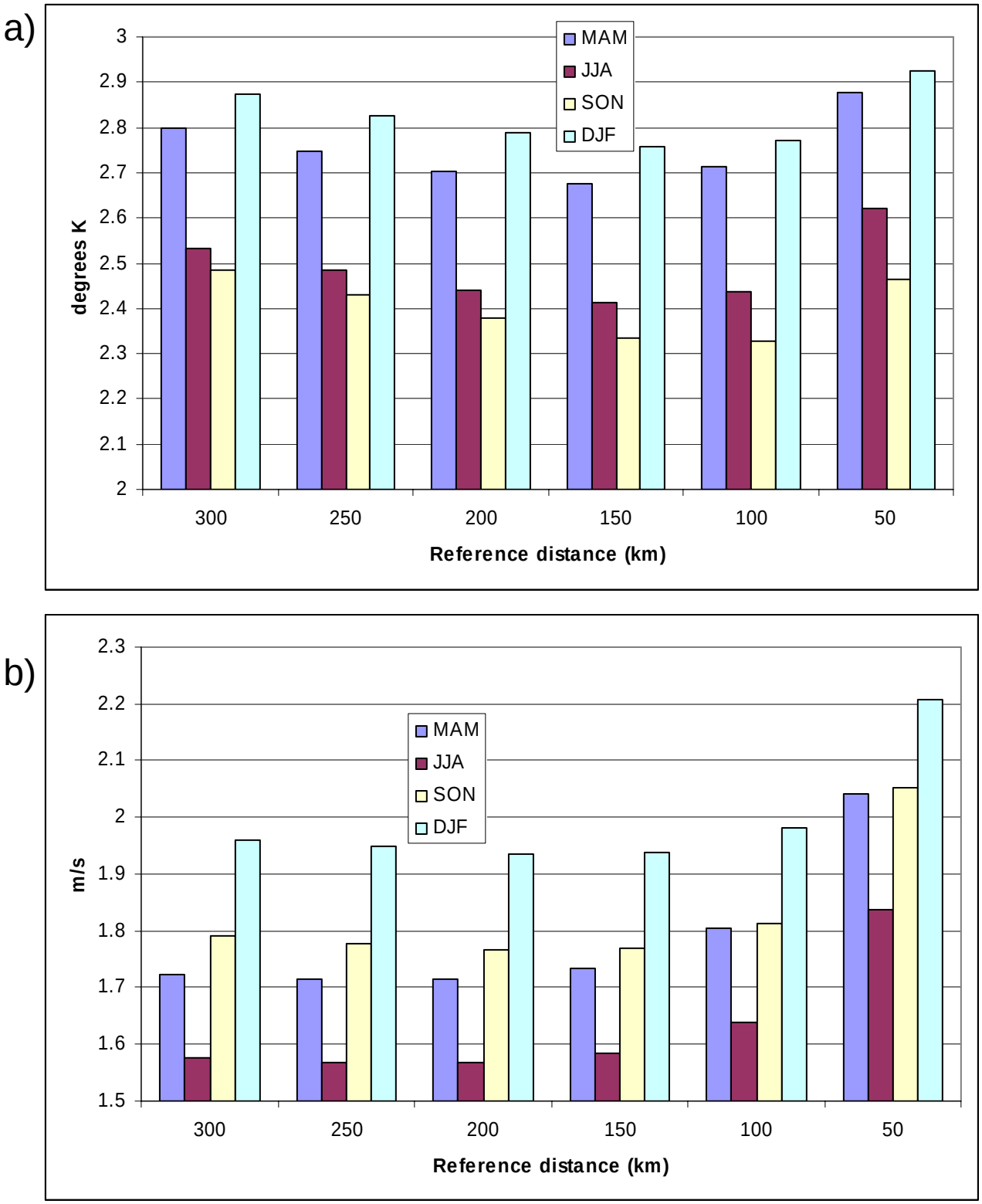
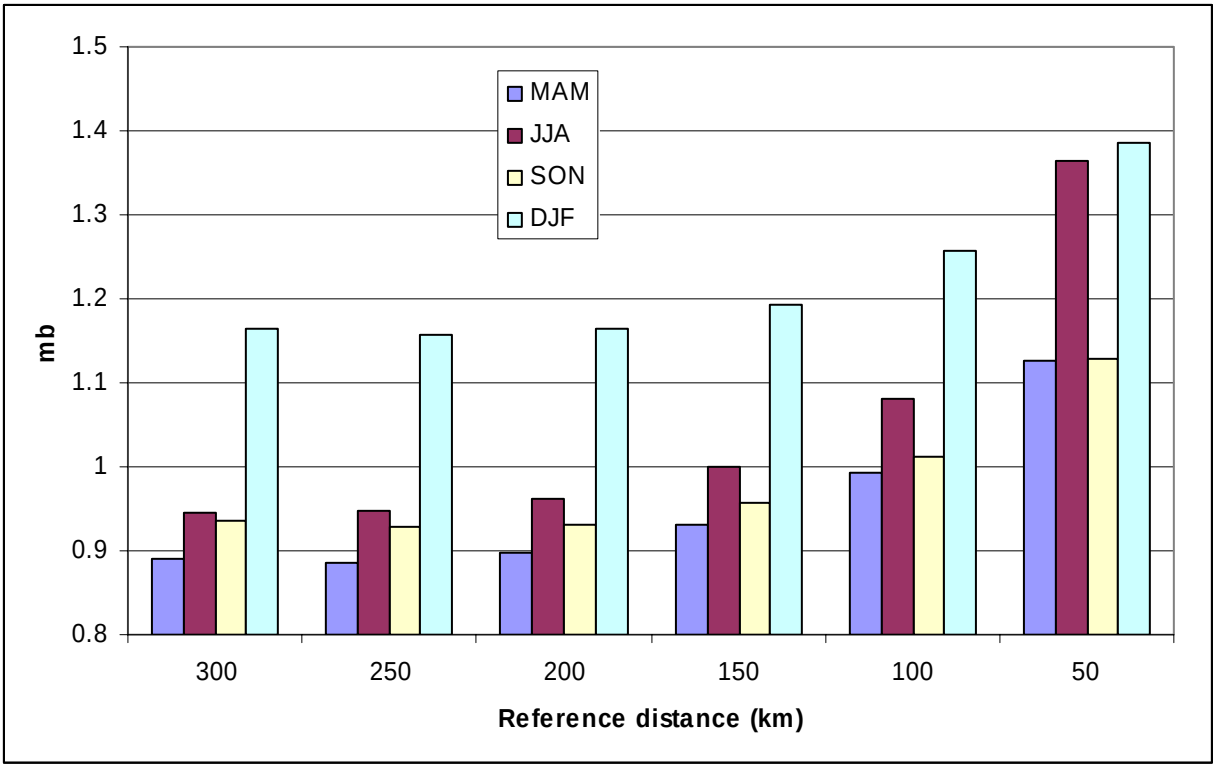


Fig. 10 RMSE values of a) temperature b) wind speed c) slp and d) dewpoint for the 0-hr forecast calculated by averaging station values to the GFS gridpoints.

c)



d)

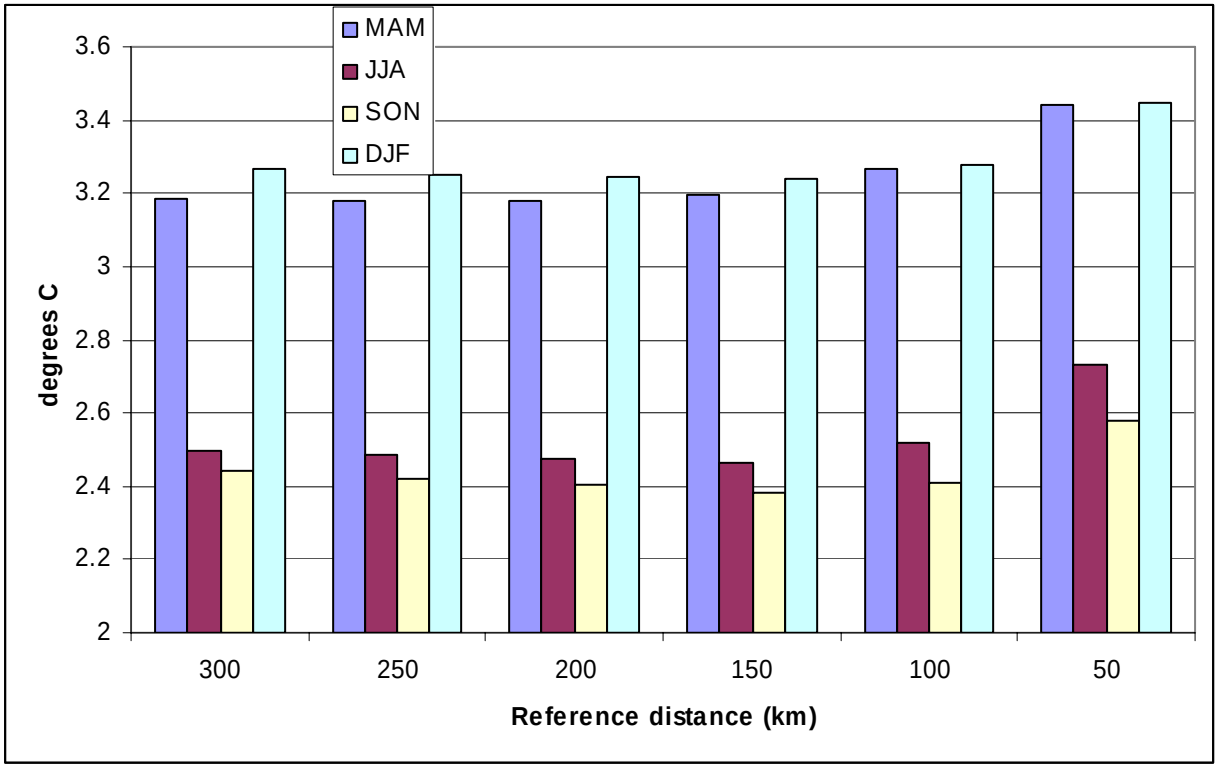


Fig. 10 continued

	0 hour	18 hour	36 hour
MAM	3.18	3.26 **.74**	3.37 **.82**
JJA	2.95	3.07 **.66**	3.15 **.76**
SON	2.86	2.97 **.76**	3.04 **.84**
DJF	3.26	3.45 **.77**	3.5 **.86**

Table 1 Temperature Error (K), and SS scores (boldface).

	0 hour	18 hour	36 hour
MAM	2.38	2.52 **.85**	2.62 **.86**
JJA	2.20	2.27 **.85**	2.32 **.86**
SON	2.35	2.42 **.86**	2.48 **.87**
DJF	2.51	2.6 **.86**	2.66 **.87**

Table 2 Wind Speed Error (m/s), and SS scores (boldface).

	0 hour	18 hour	36 hour
MAM	1.52	1.91 **.89**	2.30 **.92**
JJA	1.70	1.90 **.79**	2.06 **.87**
SON	1.61	1.86 **.91**	2.09 **.93**
DJF	2.05	2.38 **.89**	2.65 **.93**

Table 3 SLP Error (mb), and SS scores (boldface).

	0 hour	18 hour	36 hour
MAM	3.85	3.86 **.72**	3.95 **.81**
JJA	3.35	3.41 **.61**	3.49 **.72**
SON	3.31	3.39 **.76**	3.49 **.83**
DJF	3.99	4.13 **.75**	4.24 **.84**

Table 4 Dewpoint Error (K), and SS scores (boldface).