HIWPP Statistical Post Processing Report
October 2015

Isidora Jankov
Scott Gregory

Introductory Methodology: At the outset of the statistical post processing effort in HIWPP, the focus was on combining a few high-resolution deterministic weather models to obtain a skillful forecast result for 2m temperature, 10m winds, 500mb geopotential height, and 6 hour accumulated precipitation. In addition to creating a skillful forecast for these variables the forecast probability distribution was desired, such as one may get from an ensemble of forecast models.

We first considered the method for combining the models to arrive at a single forecast result. The choice was made to weight the models at every grid point, based on recent skill, to get an average. To do this, we analyzed the recent performance of each model relative to the GFS analysis for T2m, U/V10m, and Z500 and relative to CMORPH analysis for precipitation. Determining the amount of recent historical training data that is necessary to evaluate each model's relative performance was the first step in the process. Figure 1 shows the improvement in MAE (bias removed from each model) for various lengths of training data for 2m Temperature (similar curves for other variables). The optimal training period chosen for further work is 60 half days/ 30 days since it is the approximate amount where the MAE improvement is best.

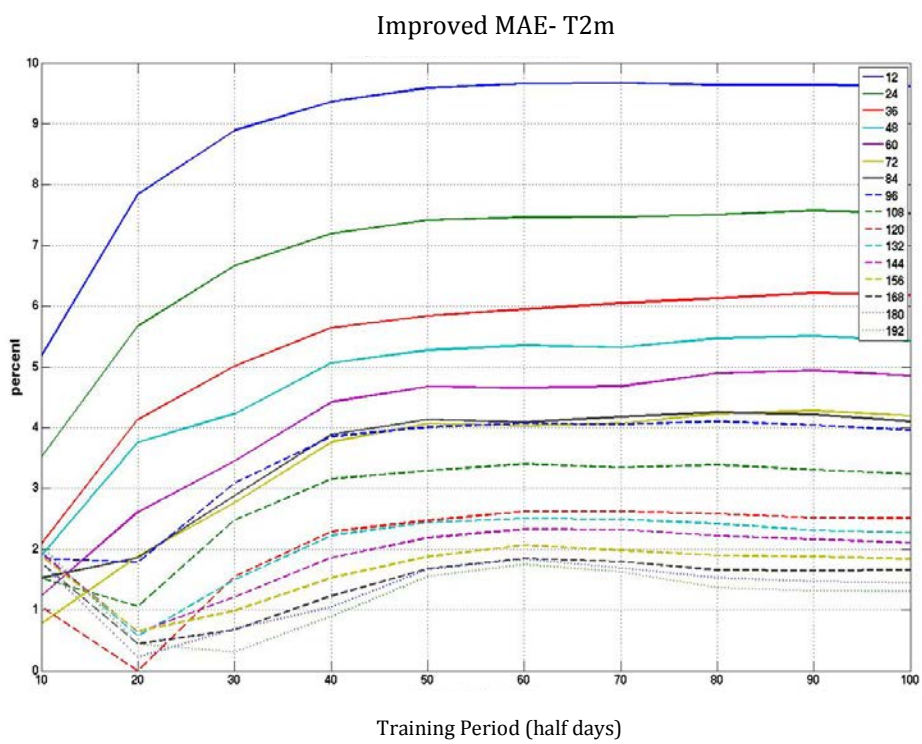

Figure 1: Reduction in MAE for various lengths of training data. 2m Temperature height using weighted model average VERSUS arithmetic average where weight is determined to be inversely proportional to MAE during training period.

In order to determine the probability distribution in the absence of an ensemble, the training data is used again. Once the weight of the models at each grid point is

determined, the weighted average forecast is produced in addition to the weighted hindcast of the training period. Under the assumption that the probability distribution of the forecast should be equal to the distribution of the training period, the distribution of the hindcast errors is determined. In the preliminary analyses, the hindcast error distributions were investigated for the distribution shape. As shown in Figure 2, the distribution of the actual hindcast errors in wind (similar for Temperature and Z500) is approximately a normal distribution; therefore the choice was made to represent the probability distributions as normal, characterized by the standard deviation of error in the training data.

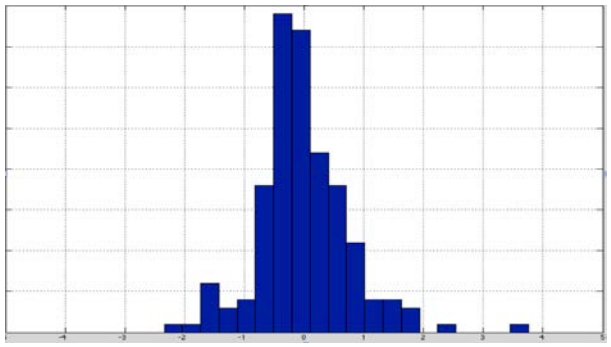

Figure 2: Histogram of sample hindcast of wind error used to determine shape of error distribution

Probability distribution for precipitation is approached in a different manner since the training data are generally sparse. The quantitative precipitation forecast (QPF) is calculated similarly to other variables (models weighted based on recent skill) and this QPF is used in producing a probability. Guidance for the probability was taken from the 'Probabilistic Quantitative Precipitation Forecasts Product/ Service Description Document (PDD)' (http://products.weather.gov/PDD/PDD-POE.pdf) and 'An Experiment in Probabilistic Quantitative Precipitation Forecasting' by Steve Amburn and James Frederick of the National Weather Service Forecast Office in Tulsa, Oklahoma (http://www.nws.noaa.gov/ost/nfuse/PQPF_AMSProbConf2006.pdf). We have chosen to represent the probability distribution with an exponential probability of exceedance (POE) of a selected rainfall amount (x). In Amburn and Frederick, the POE(x) for any point is given by:

$$POE(x) = \exp(-x/\mu)$$

where μ is the QPF, graphically shown in Figure 3.

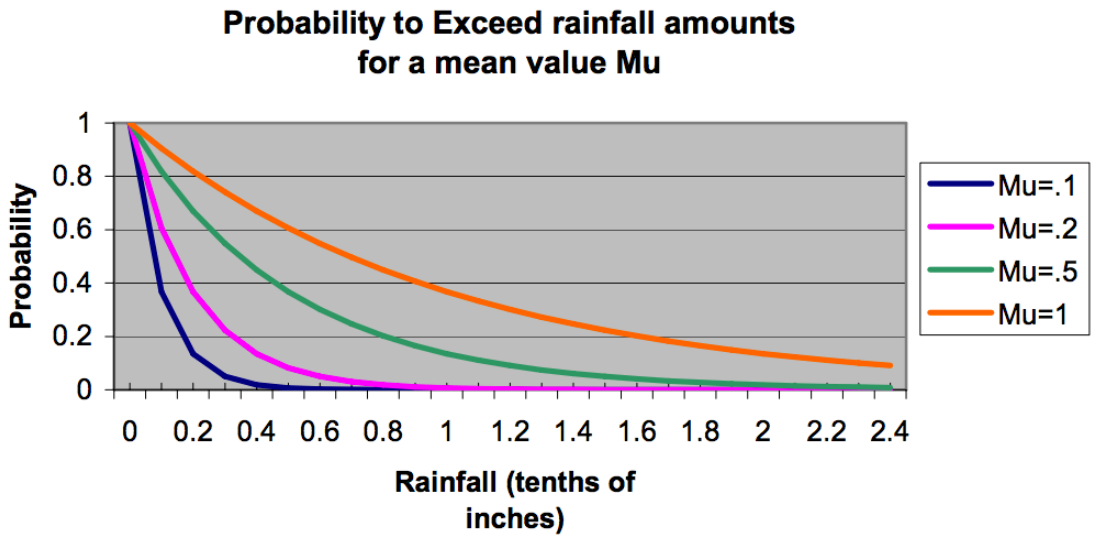

Figure 3: Probability of exceedance of Rainfall Amount

Production Code: In the final quarter of the calendar year of 2014, the HIWPP statistical post-processing task delivered the Production Version 1.0 of Statistical Post Processing (V1.0_SPP). Based on the results of analysis performed on surface temperature and wind (T2m and U/V 10m), geopotential height at 500mb (Z500), and 6 hour accumulated precipitation with two models, V1.0_SPP includes model weighting at each grid point that is inversely proportional to the Mean Absolute Error at the point in 30 days of training data. The other major result desired from the statistical post-processing task is a probability distribution of the forecast. To this end, also included in the output of the production code is the standard deviation of the weighted forecast error in the training period at each grid point for T2m, U/V10m, and Z500, variables currently assumed to be normally distributed. For precipitation, the probability of exceeding 1mm, 5mm and 10mm of precipitation are provided, based on an exponential distribution using the reciprocal of the weighted mean quantitative precipitation forecast (QPF) in the exponent. The performance of the weighted mean versus the arithmetic mean of the models for the period from December 24, 2014 thru January 7, 2015 is shown below in Figure 4. These are consistent with the results from sample data analyzed prior to the development of the production code and are therefore considered representative of the general performance of the weighting algorithm.

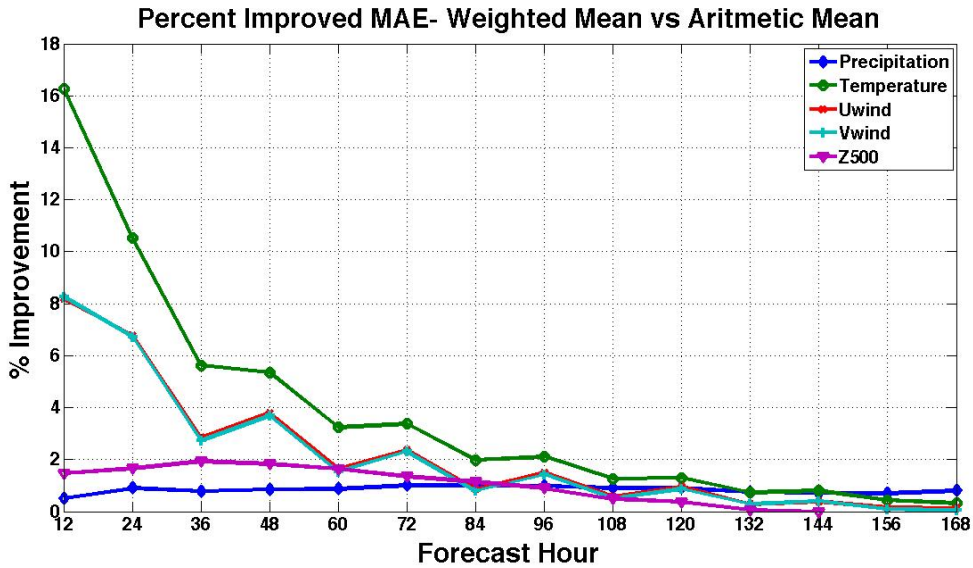

Figure 4: Weighted Mean Forecast Performance

The production code was monitored for stability and additional code was added to perform conversions of output from netCDF to GRIB1 for verification, and to GRIB 2 for users who prefer it.

In the latter portion of Q4 calendar 2014, implementation and testing of the year 2 concept of improving the forecast with field alignment began. The testing has included and will continue to include application of the field alignment algorithm upon the weighted and raw forecasts with various lengths of training cycles to determine the best application of the method. Testing will be complete by end of Q1 calendar 2015.

New Direction: After developing the methodology and production code for post processing, the HIWPP post processing task has had a change in trajectory from including only a very few high-resolution deterministic models to using actual ensembles of GEFS (20 members) and FIM ensemble (10 members). The purpose of the change in direction is to use the statistical post-processing code to facilitate answering the question: 'Can combining FIM with GEFS members add value to the overall ensemble performance?' To extend the HIWPP statistical post processing to ensembles of 20 or more members, additional code development had to take place. Specifically, the purpose of additional development was on reducing the memory burden of carrying the statistical basis for weighting all ensemble members.

Improved code was used to process data from a six-month long (June-July-August, November-December, 2014 and January, 2015) retro run, including the experimental version of GEFS and the matching configuration of FIM ensemble. The experimental version of GEFS represents the next implementation of the GEFS and it includes use of high-resolution GFS and improved initial conditions, higher in both spatial and temporal resolution and forecasts out to sixteen days. Ten FIM ensemble members were run using ten different initial conditions provided by GEFS and with the matching physics, spatial and temporal resolution. The idea was to compare performance of an ensemble that

consists of ten FIM members and ten GEFS members to performance of twenty GEFS members. The initial result was simply making an ensemble mean for the 20-member GEFS ensemble and making the ensemble mean from 10 FIM members plus 10 GEFS members. An example of preliminary results is presented in the figures below. Figures 5 and 6 illustrate a 'die-off' diagram showing change in the ensemble mean 500mb anomaly correlation for two ensembles with the lead-time and for global domain and for Northern and Southern Hemispheres. It can be seen that combining two different dynamic cores (GEFS and FIM) into the ensemble resulted in an improvement of 500mb ensemble mean anomaly correlation for the overall global domain and longer lead-times. Also, Figures 4 and 5 show that most of the improvement is coming from the Southern Hemisphere performance.

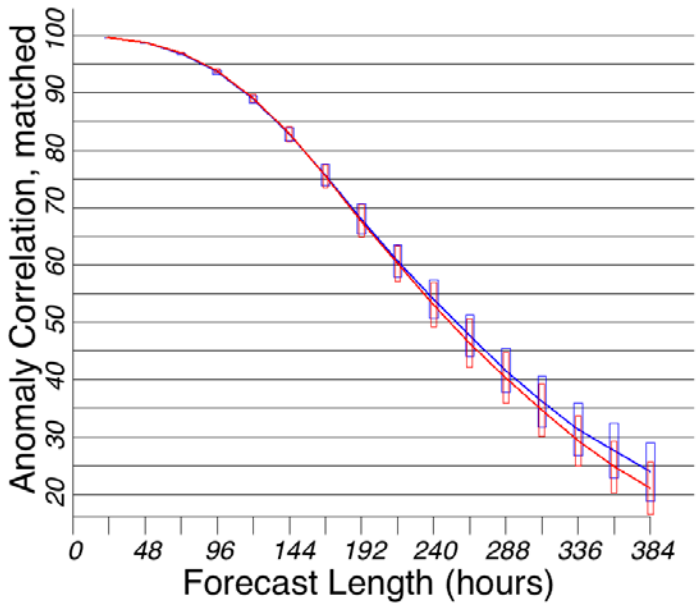

Figure 5: Mean Anomaly correlation versus forecast lead time for Global domain. Ensemble Mean- GEFS alone (red) and Combined FIM/GEFS (blue)

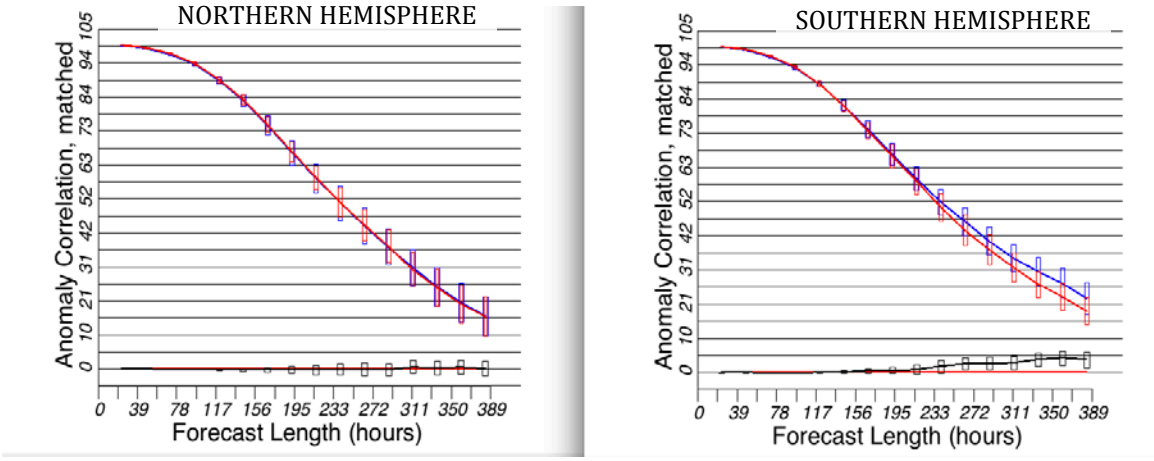

Figure 6: Ensemble Mean Anomaly Correlation with lead time, Northern and Southern Hemispheres, GEFS ensemble (red) and the mixed GEFS/FIM ensemble (blue).

Continued Ensemble Evaluation: The post processing effort proceeded to expand the statistics methodology of combining 10 FIM ensemble members with 10 GEFS members and compare the result to 20 GEFS members alone. Again, the data for the experiments included 2014 summer season (Jun, July, August) and winter season (November, December, January).

We evaluated ensemble statistics such as spread/error ratio and Continuous Rank Probability Skill Score (CRPSS). Figure 6 shows CRPSS for 2m temperature and 850 mb temperatures for summer months (June-July-August) for Southern and Northern Hemisphere. Combining FIM ensemble members with GEFS members resulted in statistically significant increase in skill for 2m temperature and 850 mb temperature for Northern Hemisphere. Combining FIM members with GEFS members also produced slight improvement in the Southern Hemisphere (Fig. 7).

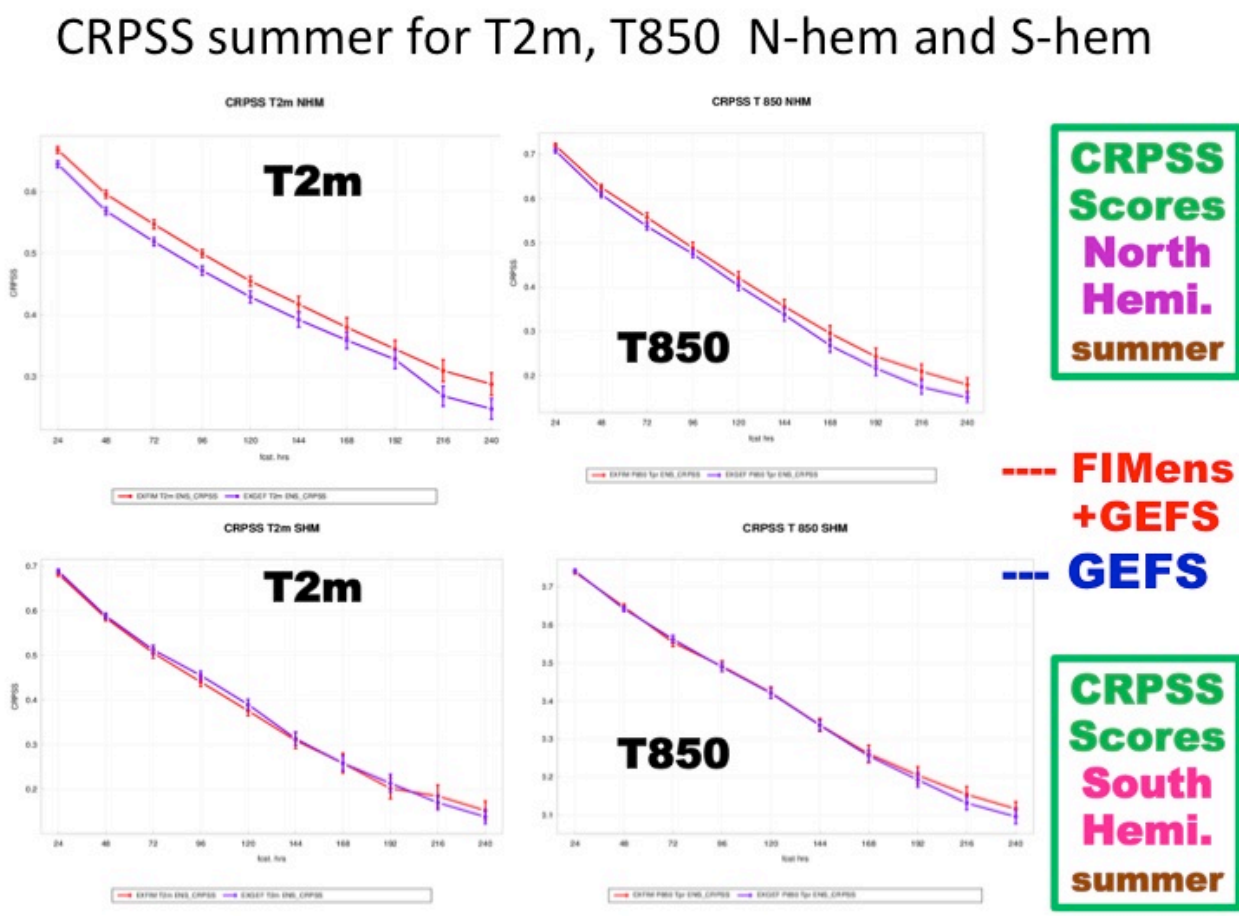

Figure 7: CRPSS for 2m temperature and 850mb temperature for warm season and both Northern and Southern Hemisphere.

Figure 8 shows 2m temperature spread and Root Mean Squared Error for the two ensembles and for both Southern and Northern Hemisphere. We can see that in the Northern Hemisphere GEFS ensemble is characterized with statistically significantly smaller spread for all lead times and significantly higher error at shorter lead times as

compared to the ensemble that combines GEFS and FIM ensemble members. In Southern Hemisphere the ensemble with combined FIM and GEFS members resulted in slightly improved spread for linger lead times.

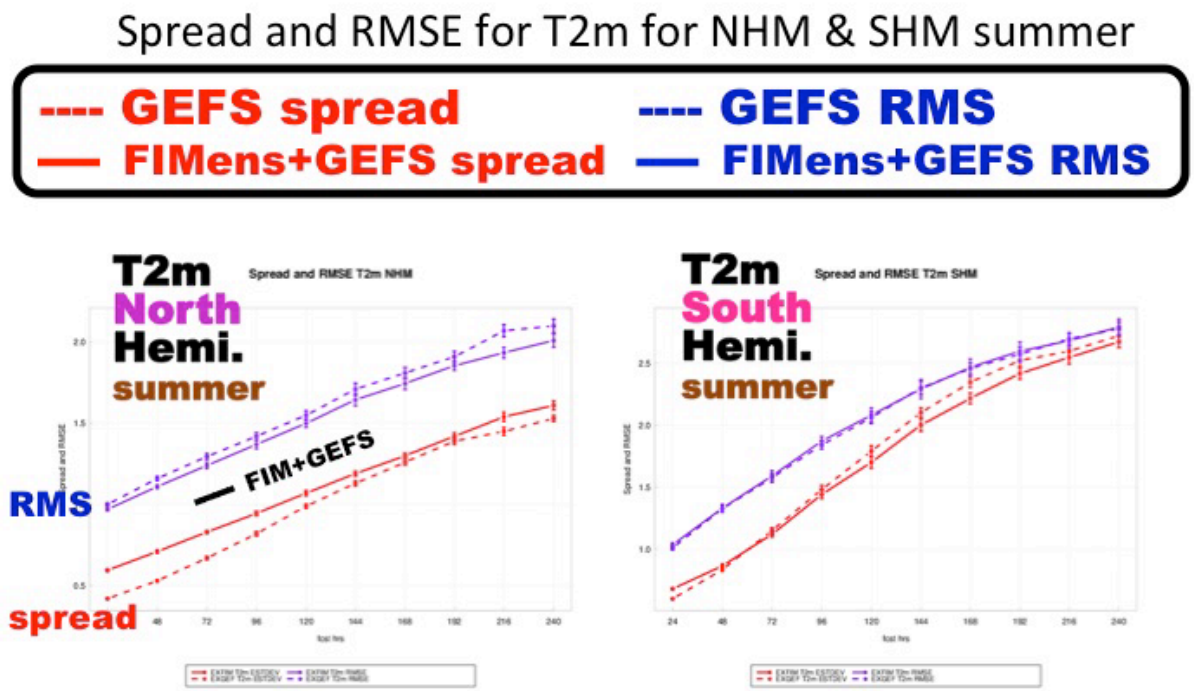

Figure 8. Spread and Root Mean Squared Error comparison for 2m temperature for the two ensembles and for both Northern and Southern Hemisphere.

With the positive result exhibited from combining the FIM ensemble with 10 of the GEFS members to make a new ensemble mean, the methodology developed initially for the fewer members, weighting based on skill in 30 days training, was applied to attempt to obtain an even better result. Several varieties of weighting were attempted; all of the results were very similar. The varieties include the original weighting developed- each grid point weighted inversely proportional to its MAE in the training period; weighting proportional to each point's correlation in time in the training set; and finally weighting each member by its mean anomaly correlations in the training period. In Figure 9, an example of these post-processing results is included (Note: since a portion of the time period is used for training, the results for ensemble means shown in Figure 8 differ from Figure 5).

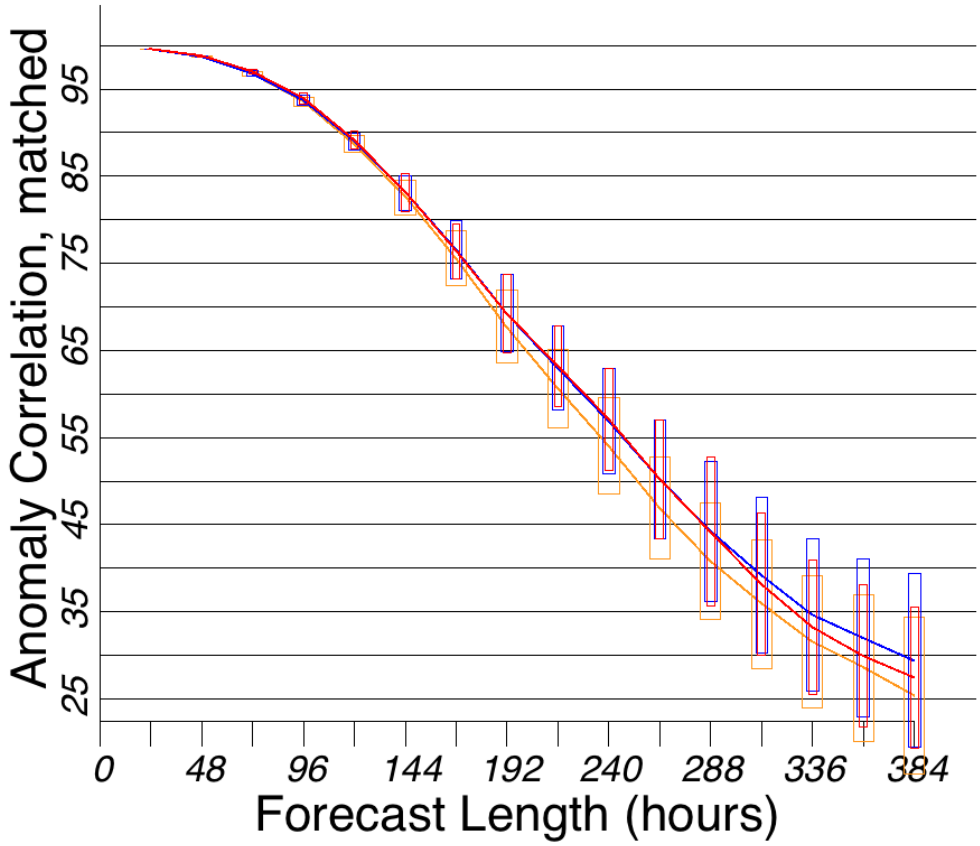

Figure 9: As in Figure 1, added the post processed example (orange).

The results of the experiments exhibit the benefit of combining the FIM ensemble with the GEFS ensemble and show that the ensemble variety is the source of the benefit.

The post processing continued to focus on mixed model ensemble including GEFS and FIM dynamic cores. Previous experiments included combining the two dynamic cores with perturbed initial conditions but exactly the same physics. Testing use of Grell-Freitas (GF) convective parameterization within FIM showed promising results. A decision was made to include variations in physics for FIM members. In addition to initial conditions perturbations, five out of 10 FIM members included the GF convective scheme. In agreement with our EMC colleagues, a decision was made to perform a full year of 2014 retro run with the new FIM ensemble configuration. The year-long retro run was completed and some of the data analyses were performed. Figure 10 compares 500mb anomaly correlation results between GEFS, the old GEFS+FIM ensembles, and the new GEFS+FIM ensemble on the global domain and for summer of 2014. We can see that adding variation in convective treatment resulted in a positive impact on the global domain.

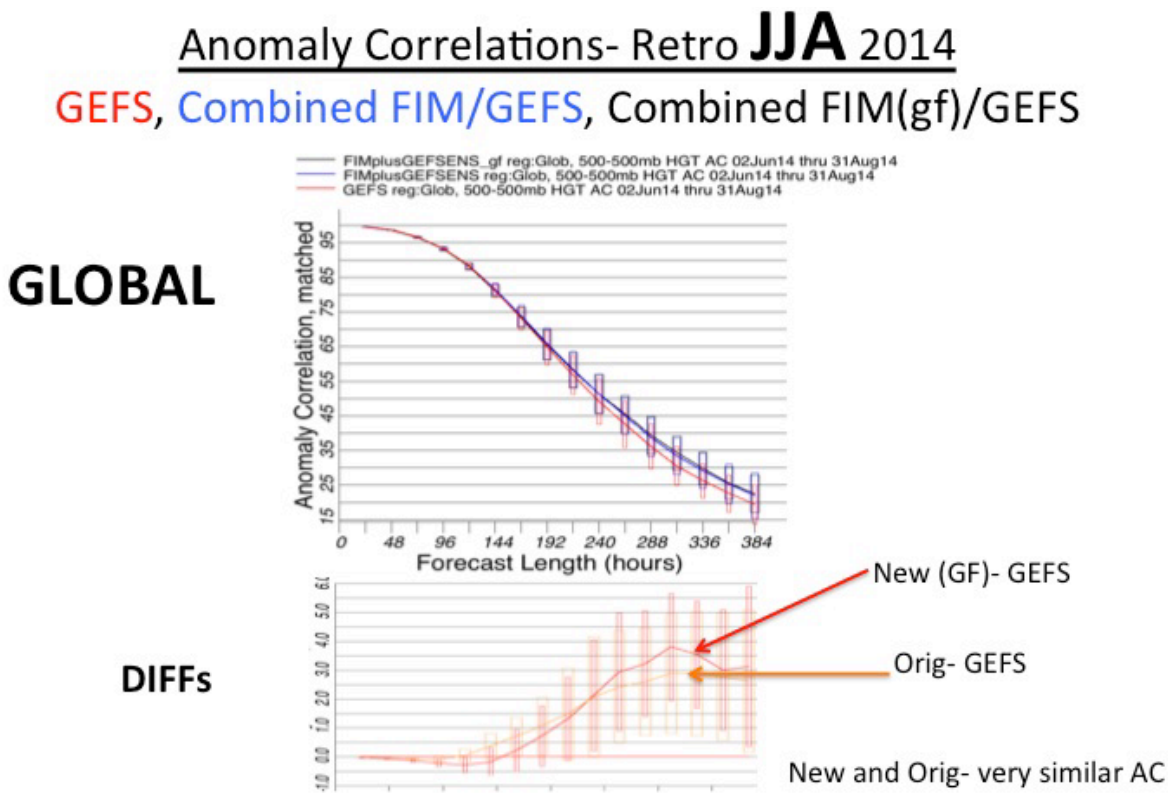

Figure 10. 500mb Anomaly Correlation for GEFS and old and new version of FIM ensembles over the global domain and for the summer of 2014.

By performing the same analysis for the Northern and Southern Hemispheres we can see that the majority of positive impact is associated with an improvement in the Southern Hemisphere Figure 11). Similar results were obtained for the month of November (Figure 12). The analysis for the rest of the retro run is still ongoing. Additional ensemble statistics such as skill, spread and error for various variables were also evaluated. An example of spread, ensemble mean RMSE, spread/error ratio and CRPSS statistics for 2m temperature, Northern Hemisphere, and for the summer period of 2014 is shown in Figure 13. We can see that combining FIM members with GEFS resulted in an improvement of all of the statistics.

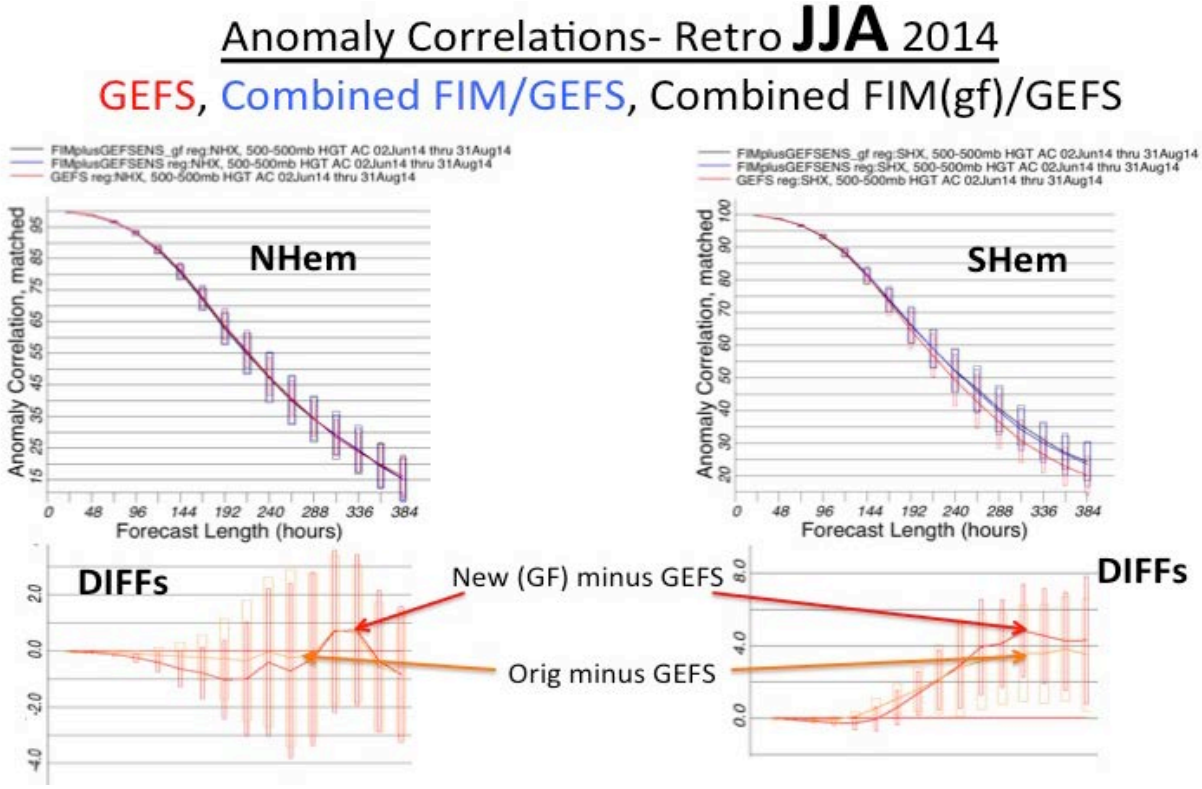

Figure 11. As in Figure 1, except for Northern and Southern Hemisphere domains.

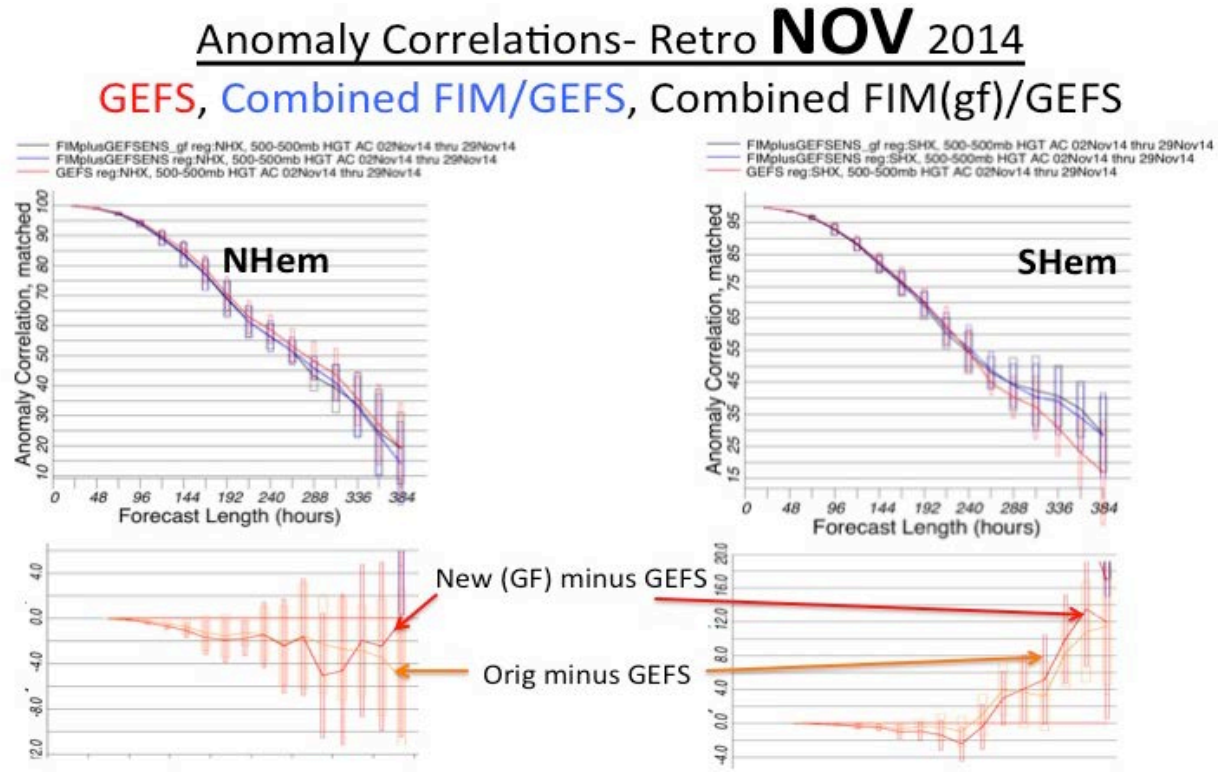

Figure 12. As in Figure 2, except for the month of November 2014.

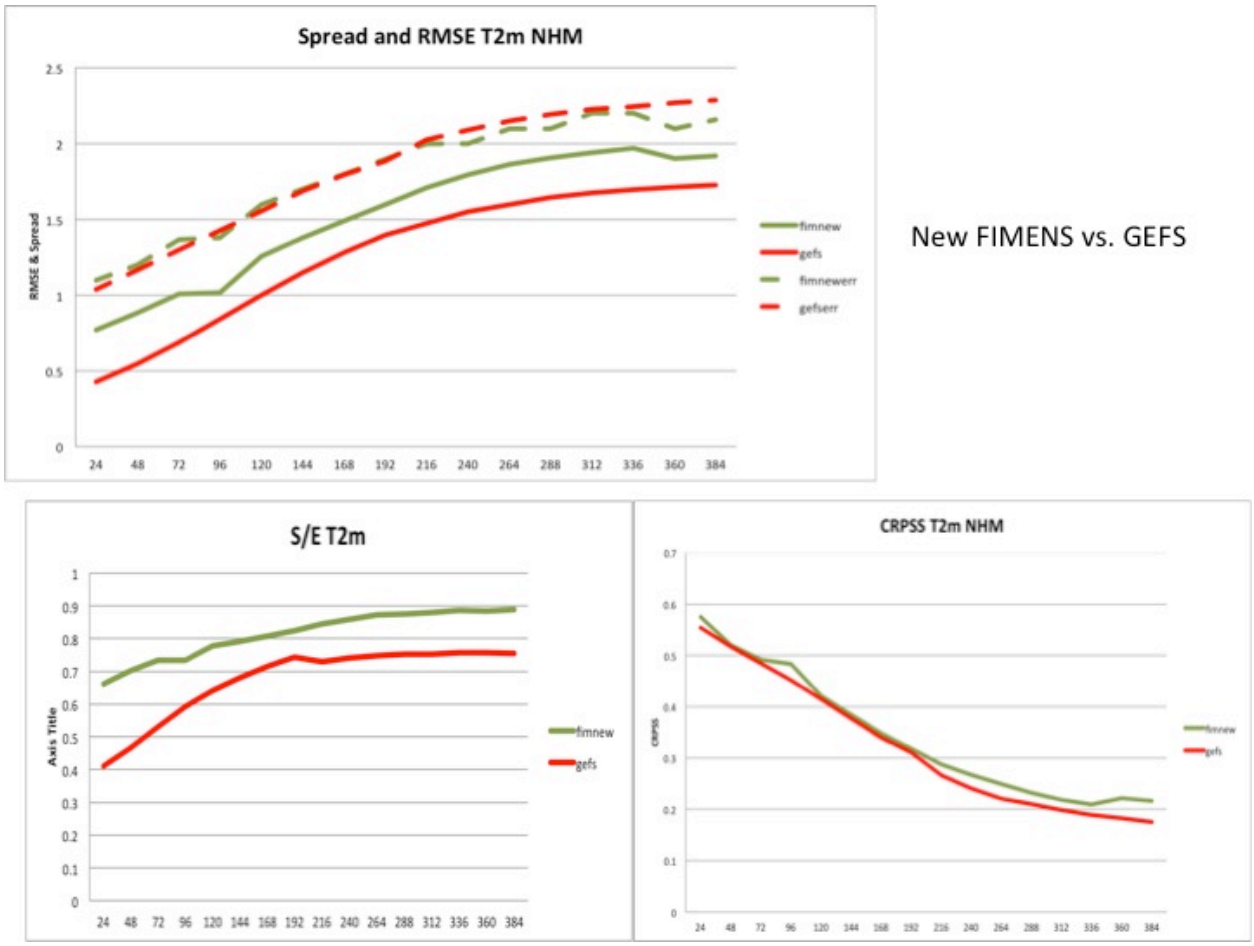

Figure 13. Spread and RMSE, spread/error ratio and CRPSS for Summer 2014, Northern Hemisphere and for the GEFS vs. FIM ensembles.

Most recently, the FIM ensemble has been running in real time production and the ensemble mean upper air temperature of the combined ensemble and the GEFS alone have been made for creating new verification statistics. Figure 14 shows the upper air temperature global Root Mean Square Error for the combined versus the GEFS. This plot, based ion a very limited (~2-3 weeks) recent history of forecasts, indicates that the results are similar in this statistic, while the combined ensemble has lower error than the GEFS alone at several pressure levels.

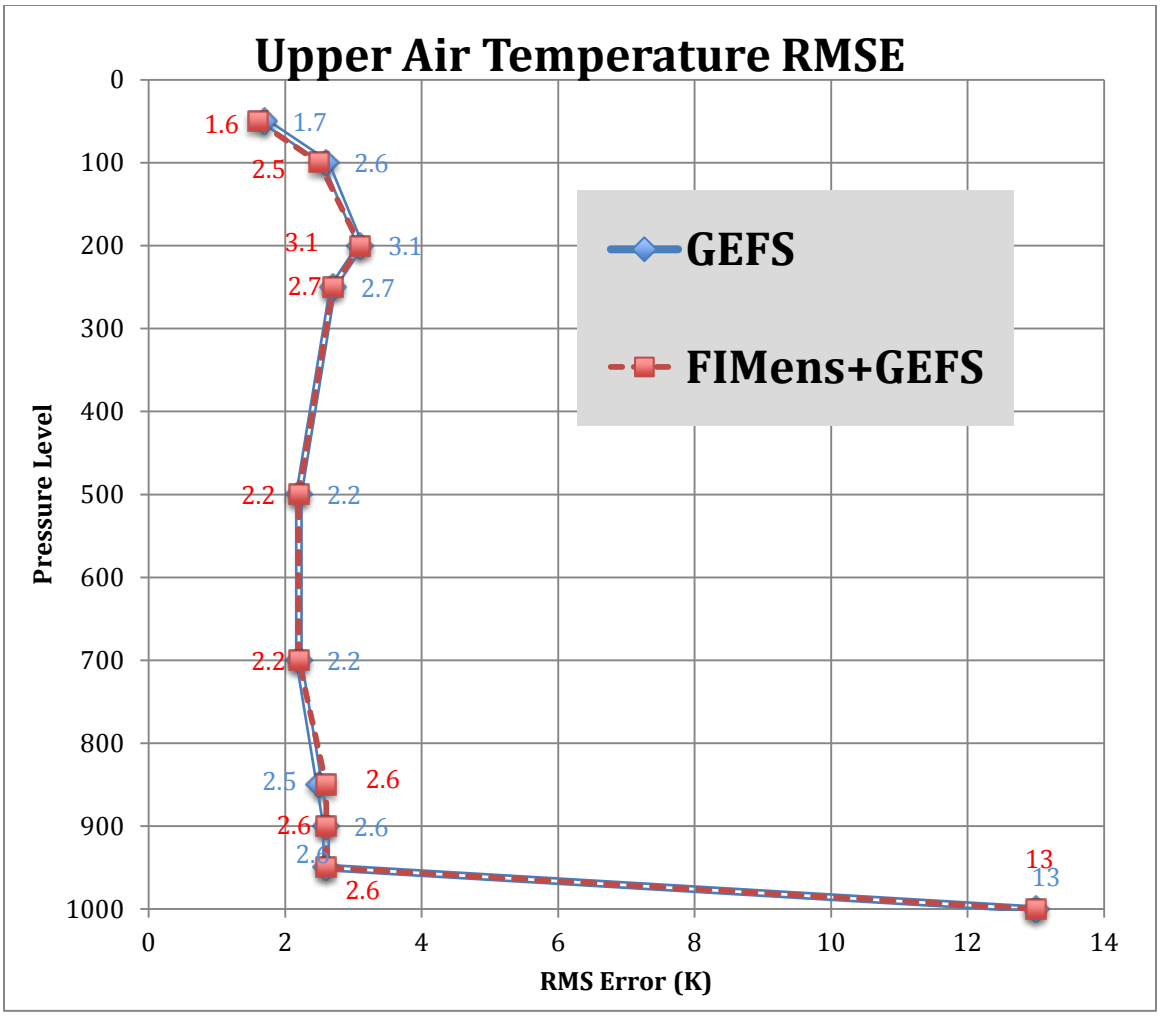

Figure 14: Real Tim Upper Air Temperature Error, GEFS and combined GEFS FIM ensemble.