On Improving the Output of a Statistical Model

Using GFS single point outputs for a linear regression model and improve forecasting

Mark Delgado
4/19/2016

i. Introduction

Forecast Modelling

- Using computers to create a simulation of the world.
- Modelling has improved dramatically over the years, with resolutions of the latest generations reaching 12 km gridded resolution.

Forecast Modelling

- Using computers to create a simulation of the world.
- Modelling has improved dramatically over the years, with resolutions of the latest generations reaching 12 km gridded resolution.

But a lot can happen in that 12 km!

- Models still can’t resolve spatially resolve small scale variations that can create significant deviations from the gridded output.

Forecast Modelling - Trying to Resolve the Spatial Resolution problem

- The models such as the gfs are also able to create model outputs for a single point.
- These points outputs improve the overall forecast of that point.
- But, there are ways to improve that forecast further.

That’s where the idea of a linear regression model to improve forecasting comes into play.

- Also known as a Model Output Statistics (MOS)

MOS and Linear Regression! How Do they Work??

- MOS takes various model outputs (Predictands) to find the data that most correlates with the desired output variable being modelled.
    - Predictands can be any number and types of variables. Temperature at the surface, or at a certain layer of the atmosphere, Relative Humidity, cloud cover, rainfall percentages are just a few examples
- The predictands with the highest correlations are used to generate the linear regression equation. Coefficient of Determination used as a basis for the weighting.

$$Y = c_0 + c_1 * X_1 + c_2 * X_2 + \ldots + c_N * X_N$$

So What Is the Goal?

Improve overall forecasting at a single point.

Show Linear Regression modeling can improve the performance of large scale models.

Keep the approach simple and easily understood.

Time permitting add more Predictands to the new LR Models and see if statistically significant improvement in forecasting is shown with each new Predictand.

ii. Methodology

Gathering and Processing the Data

- Found GFS point data on a specific spot (Brasstown Bald, GA).
- The same location as a Remote Automatic Weather Station (RAWS).
- Used the 18 hour forecast from the 00Z GFS model runs for every day from January 1st, 2011 through December 31st, 2014.
- Generated a linear regression and correlated the data of various atmospheric levels to determine which levels had the highest correlations.
- Further broke data down according to similar slopes and intercepts.
- Created a simple linear regression model based on the highest correlated levels.

Predicting the Temperature

- After running the new model, compared the results against the original GFS output.
- Used the Student's T-Test to determine if the mean errors were different.
- Found the confidence interval of the correlation coefficient.
- Determined the mean and standard deviation of the new model's slope and intercept with the bootstrap resampling method.
- Used the jackknife method to determine the amount of bias in the new model
- New Model Equation for Temperature:

$$PTF = PTGFS + COV * (PTGFS - (SLP * PTGFS + INT))$$

The long term goal is to add more variables to the Model.

Predicting Relative Humidity and Wind and Improving the overall linear Regression

The methods outlined previously are used to create the same models for Relative humidity, as well as the u and v components of the wind.

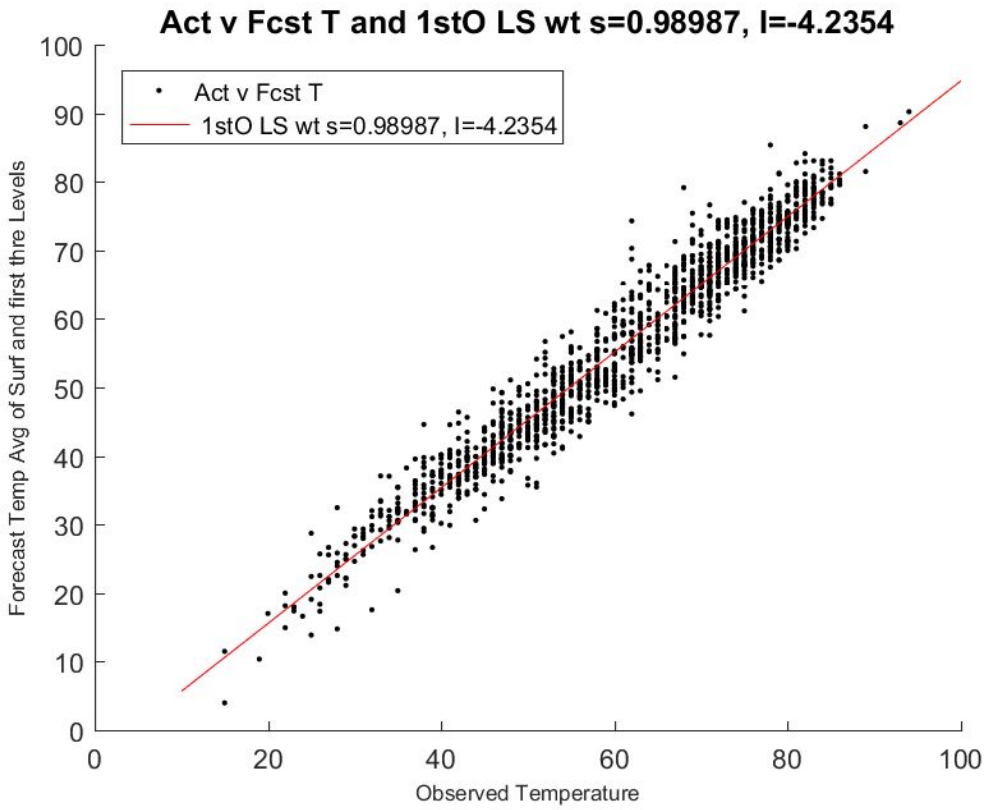
Each model output variable will have it's own unique set of Predictands with weights based on the correlation and (cross correlation) with the Predicand and the actual data.

Time permitting, the goal is to incorporate RH and wind vectors Predictands to all three models and see if adding more predictors will significantly improve the new model performance.

iii. Results

Creating the New Temperature Model

The Correlation and Coefficient of Determination for the 18 hour GFS model



Height	Correlation	Coefficient of D
Surface (Model Level 1)	0.973863	0.948408
Model Level 2	0.974080	0.948831
Model Level 3	0.973992	0.948660
Model Level 4	0.973659	0.948012
Model Level 5	0.972871	0.946478
Model Level 6	0.971185	0.943201
Model Level 7	0.968600	0.938185
Model Level 8	0.964489	0.930238
Model Level 9	0.958093	0.917943
Model Level 10	0.949469	0.901491

Table 1: Correlation and Coefficient of Determination of the 18 hour forecast versus recorded temperature. The highest correlations are at the surface and first 3 atmospheric levels.

Breaking Things Down Further

The Temperature seems to have three distinct groupings for slope and intercept.

The JJAS has the weakest correlation.

Time Period	Slope	Intercept	Correlation	Determination
February	0.79626	3.3014	0.920511	0.847341
March	0.78544	5.8362	0.938304	0.880415
April	0.79753	6.6887	0.905012	0.819209
May	0.85909	4.9582	0.843740	0.711897
June	0.80059	11.2081	0.843740	0.711897
July	0.73192	17.2998	0.808132	0.653077
August	0.72569	17.3567	0.792807	0.628543
September	0.77781	11.8401	0.861370	0.741959
October	0.8845	2.2489	0.937465	0.878841
November	0.84006	2.1614	0.947237	0.897258
December	0.85782	2.0324	0.902464	0.81442
January	0.95856	-4.0379	0.949041	0.900680
Full Year	0.98987	-4.2354	0.974031	0.948737

Table XX – Slope, Intercept, Correlation Coefficient, and Variance for each month as well as the cumulative S, I, CC, and V for the full data set. Note, January has been placed at the bottom of the table for ease of demonstrating grouping.

Time Period	Slope	Intercept	Correlation	Determination
ONDJ	0.94971	-2.8077	0.9624	0.9261
FMAM	0.91004	-0.29894	0.9593	0.9202
JJAS	0.80714	10.7775	0.8557	0.7323

Table XY – Slope, Intercept, Correlation Coefficient, and Variance for the four month groupings of October, November, December, and January; February, March, April, and May; and June, July, August, and September.

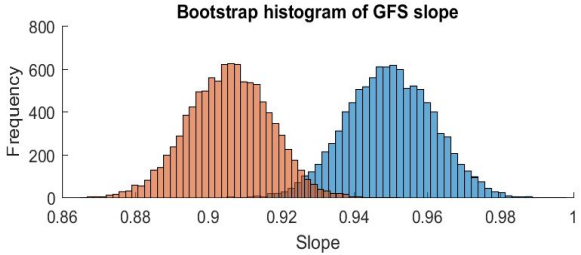
Time Period	Slope	Intercept	Correlation	Determination	RMSE
ONDJ	0.94971	-2.8077	0.9624	0.9261	3.19
Adjusted	0.90548	-5.2772	0.9624	0.9261	3.04

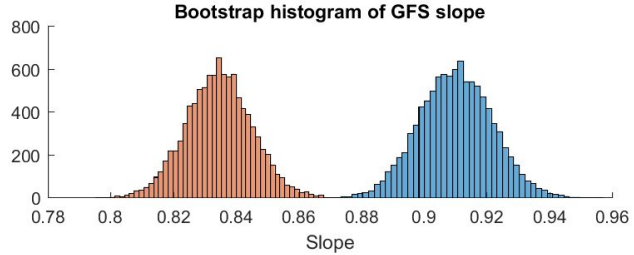
Time Period	Slope	Intercept	Correlation	Determination	RMSE
FMAM	0.91004	-0.29894	0.9593	0.9202	3.59
Adjusted	0.83471	-0.54928	0.9593	0.9292	3.29

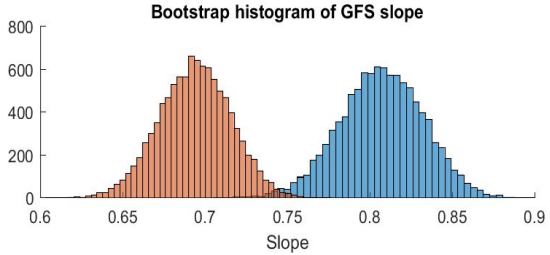
Time Period	Slope	Intercept	Correlation	Determination	RMSE
JJAS	0.80714	17.778	0.8557	0.7323	2.96
Adjusted	0.69377	17.099	0.8556	0.7321	2.54

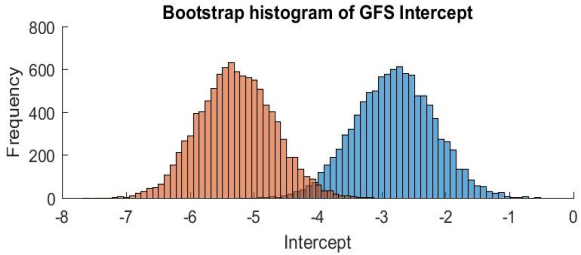
Table XZ. Comparison of the original fitted LR of the three groups with the new model forecast temperature. RMSE: Root Mean Squared Error.

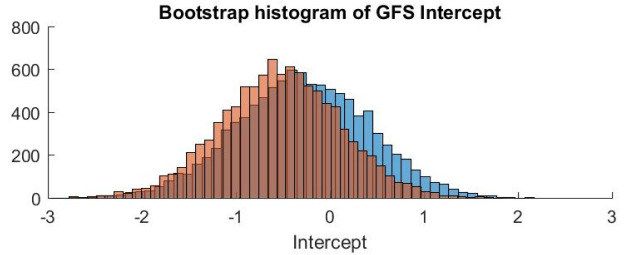
How does the New Temperature Model Compare to the straight GFS?

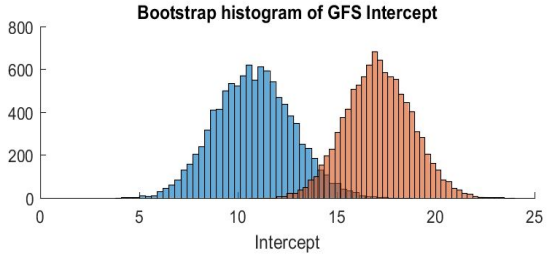












More Results

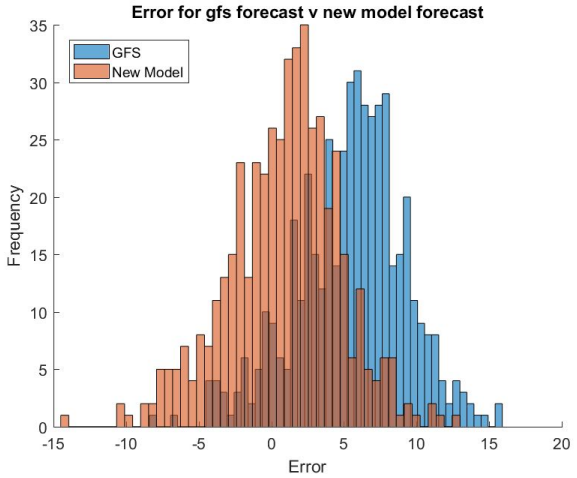
Student's T-Test was run on the raw error between the GFS versus the new LR Temperature Model

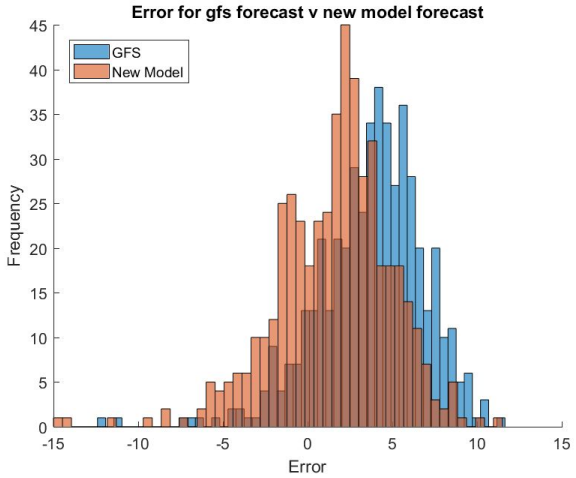
Months	Reject at 99%?	P value
FMAM	Reject	3.2924e-65
JJAS	Reject	1.5832e-24
ONDJ	Reject	3.5163e-89

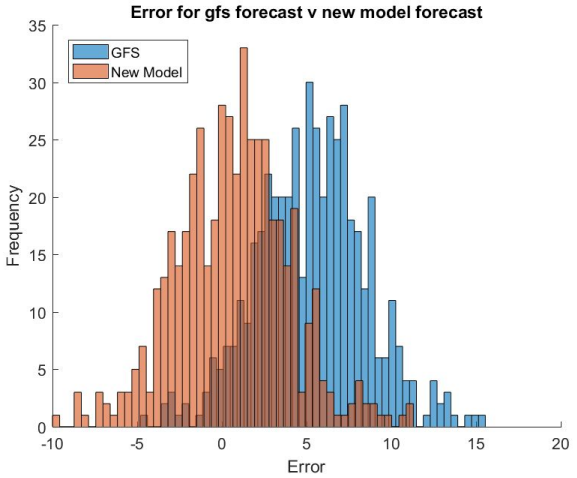
Student's T-Test at 99% confidence that the slope is different. Running a jackknife resample revealed a mean bias that was extremely small for all three data samples

Months	Reject at 99%?	P value	Jackknife Mean Bias
FMAM	Reject	8.6686e-09	Rho = -3.9380e-05
JJAS	Reject	1.2625e-10	Rho = -2.2698e-04
ONDJ	Reject	2.4254e-10	Rho = -7.5149e-05

Comparing and Contrasting the GFS versus the New LR Model







Comparison of Raw errror in 18 hour forecast for base GFS model versus the new LR model. The difference between the two models seems obvious on visual examination and the Student's T-Test confirmed the mean distributions were not the same at a 99% Confidence.

iv. Discussion and Conclusion

So What Does it All Mean?

- The new model showed improvement in the forecasting of actual temperatures
- The new model showed a difference in the RMSE between the base GFS model and the new LR model, reducing the RMSE for Temperature by 8.4% for FMAM, 13.9% for JJAS, and 4.1% for ONDJ.
- The new LR model also showed statistically significant improvements in forecasting at 99% confidence.
- It was surprising how much improvement a simple LR Model improved performance.

What’s Next?

- Add more Predictands
- Use SVD

v. References

“The Use of Model Output Statistics (MOS) in Objective Weather Forecasting”, Glahn, Harry and Lowry, Dale, Journal of the American Meteorological Society, Dec 1972, pp 1203-1211.

“Everything You Wanted to Know About MOS, But Were Afraid to Ask”, Maloney, Joe, Statistical Modeling Branch, 2005.