Application of Generalized Boosted Regression in the XM Tool Project

Ken Lau
Co-op Student
Environment Canada
Air Quality Science Unit

August 25, 2011

Outline

- List of Predictors
- Leaps BIC Predictor Selection
- Generalized Boosted Models (GBM)
  - Introduction
  - Parameters
  - Optimization
  - Results
- Conclusion and Future Work

List of Predictors

Predictors

84 AQ Model Predictors

3 **Persistence Predictors**

27 **Antecedent Predictors**

Persistence Predictors

OBS at Hr 00

Antecedent Predictors

Lag 24/48/72 hrs

Max

Min

OBS

Leaps BIC Predictor Selection

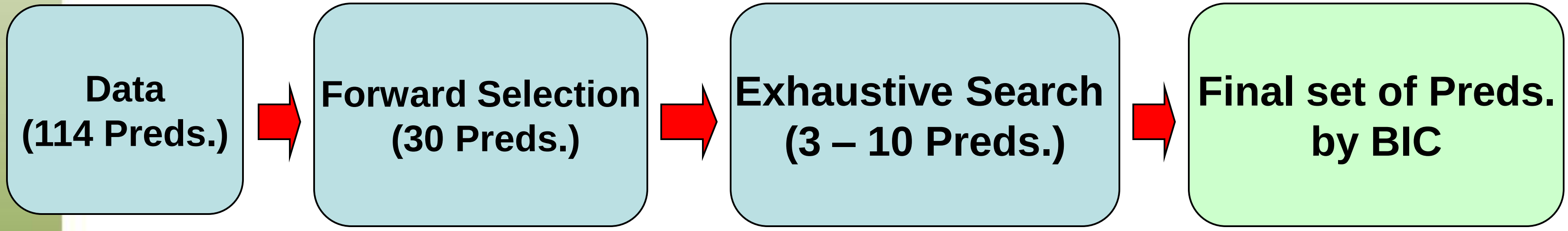
- Leaps Package in R Provides Automated Predictor Selection Routines

- Combination of Forward, Exhaustive, and minimum BIC Selection
  - Fast (Forward Selection)
  - Accurate (Exhaustive)
  - Avoids Overfitting (minimum BIC)

- $BIC=n*\log(RSS/n) + \boxed{(\log(n))}*(p+1)$

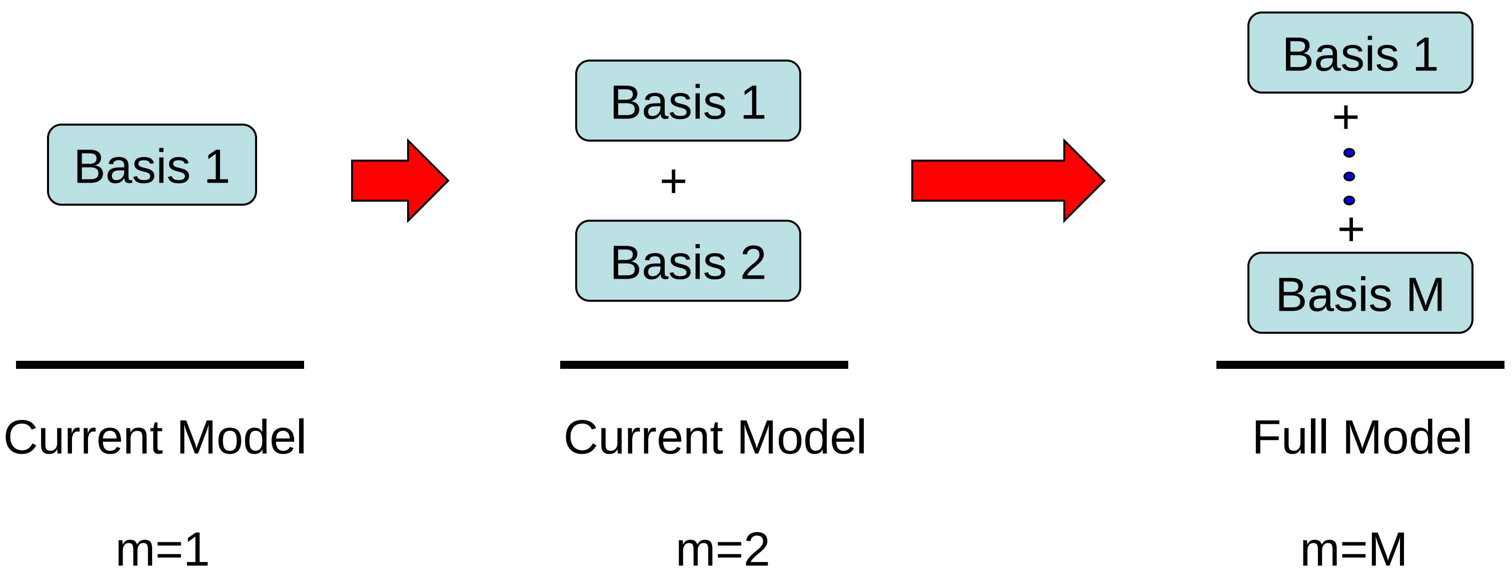
n = # Obs | p = # Predictors | RSS = Residual Sum of Squares

Flow Diagram for Leaps BIC Selection



Generalized Boosted Regression Model (GBM)

- Regression method that iteratively combines weak prediction models into one strong model
- Fit regression using current model, search basis function based on minimizing loss function, then update current model by adding basis function



Basis and Loss Function

Full Model	$f(x) = \sum_{m=1}^{M} \beta_m b(x)$
Loss Function	$\min \sum_{i=1}^{N} L(y_i, \beta b(x))$ $L(y, f(x)) = (y - f(x))^2$

b(x) represents basis function

β_m represents corresponding coefficient

Advantages of GBM

- Easy to compare between the 2 techniques (GBM and MLR)
  - Uses same predictors
  - Uses the same loss function

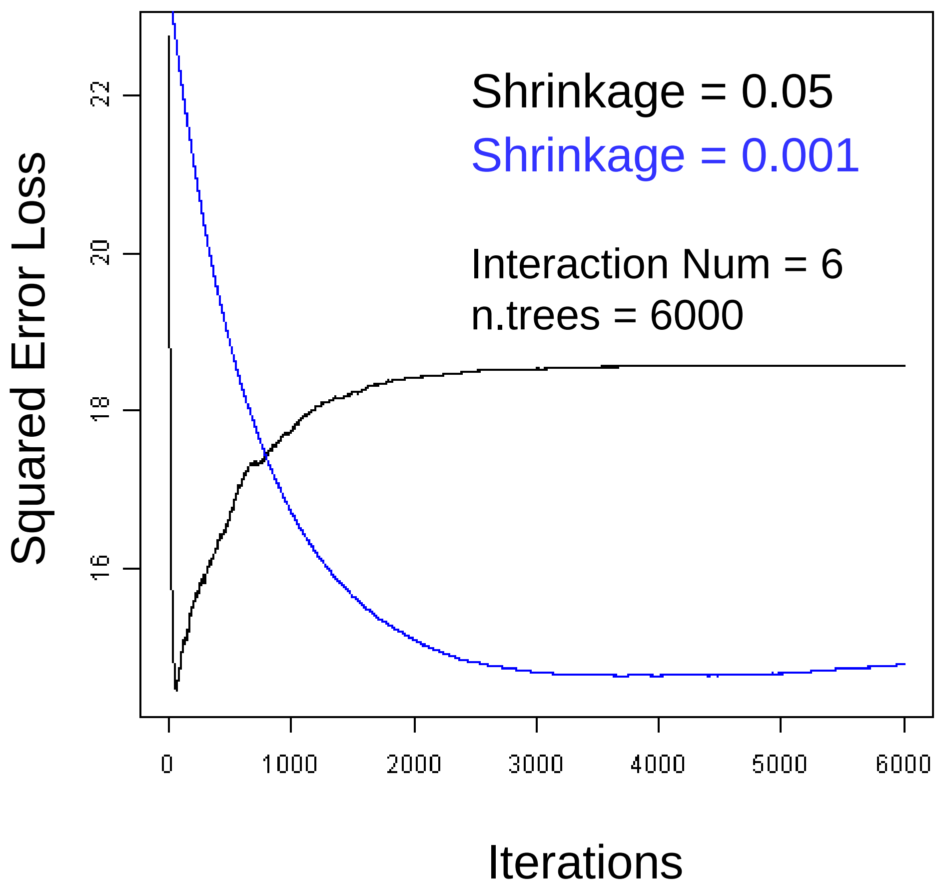
- Applicable on many techniques
  - Regression, classification, exponential, decision tree, support vector machine

GBM Parameters

	Distribution	Shrinkage	Interaction .depth	N.trees
Description	Specifying Loss Function	Learning Rate / Weight for Basis Function	Num. of Variable Interactions	Iterations
Recommended Specification	Gaussian	0.01 – 0.001	4 - 8	1000 - 10000

Rule of Thumb: # Trees x Shrinkage = [1:100]

Optimizing Shrinkage Parameter

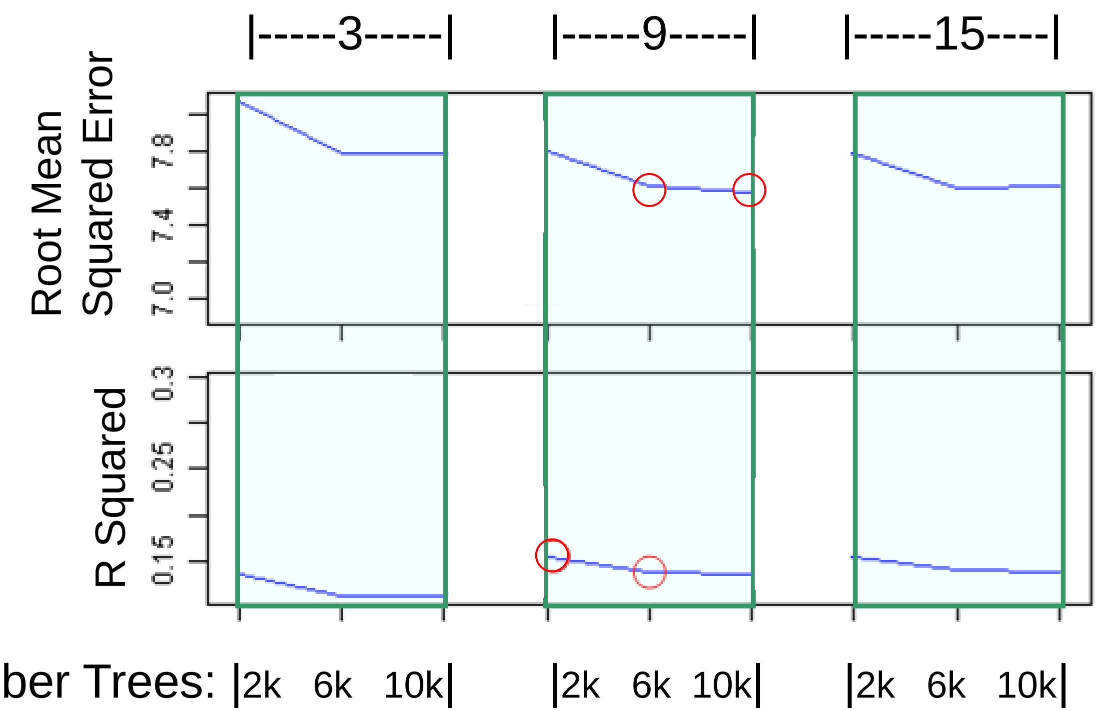


Advantages of Low Shrinkage	Disadvantages of Low Shrinkage
More consistent error loss improves Predictive Performance for new cases Avoids Overfitting	Computation Cost

Optimizing Interaction Depth and Number of Iterations

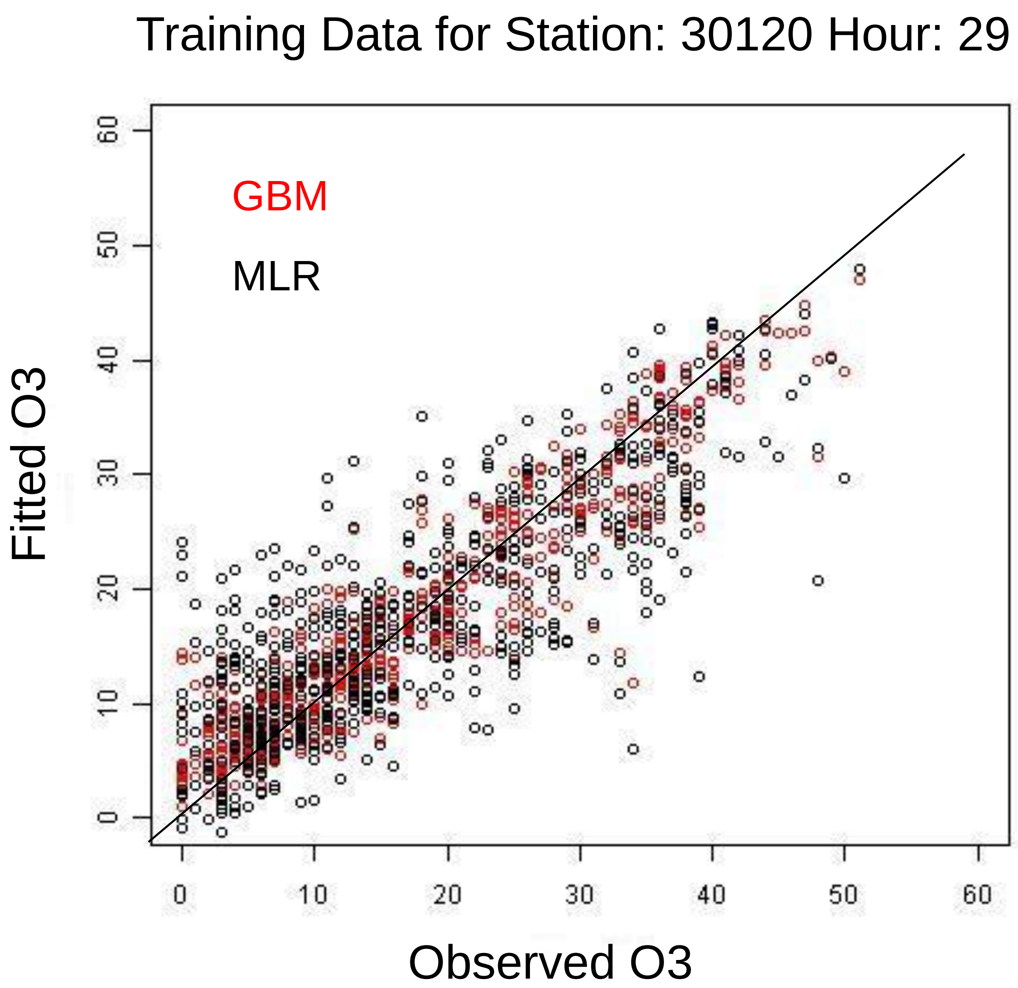
Station: 30120 Hour: 29 Pollutant: O3 Shrinkage: 0.001

Interactions



Above 75th percentile

Improvements in Fitted Model

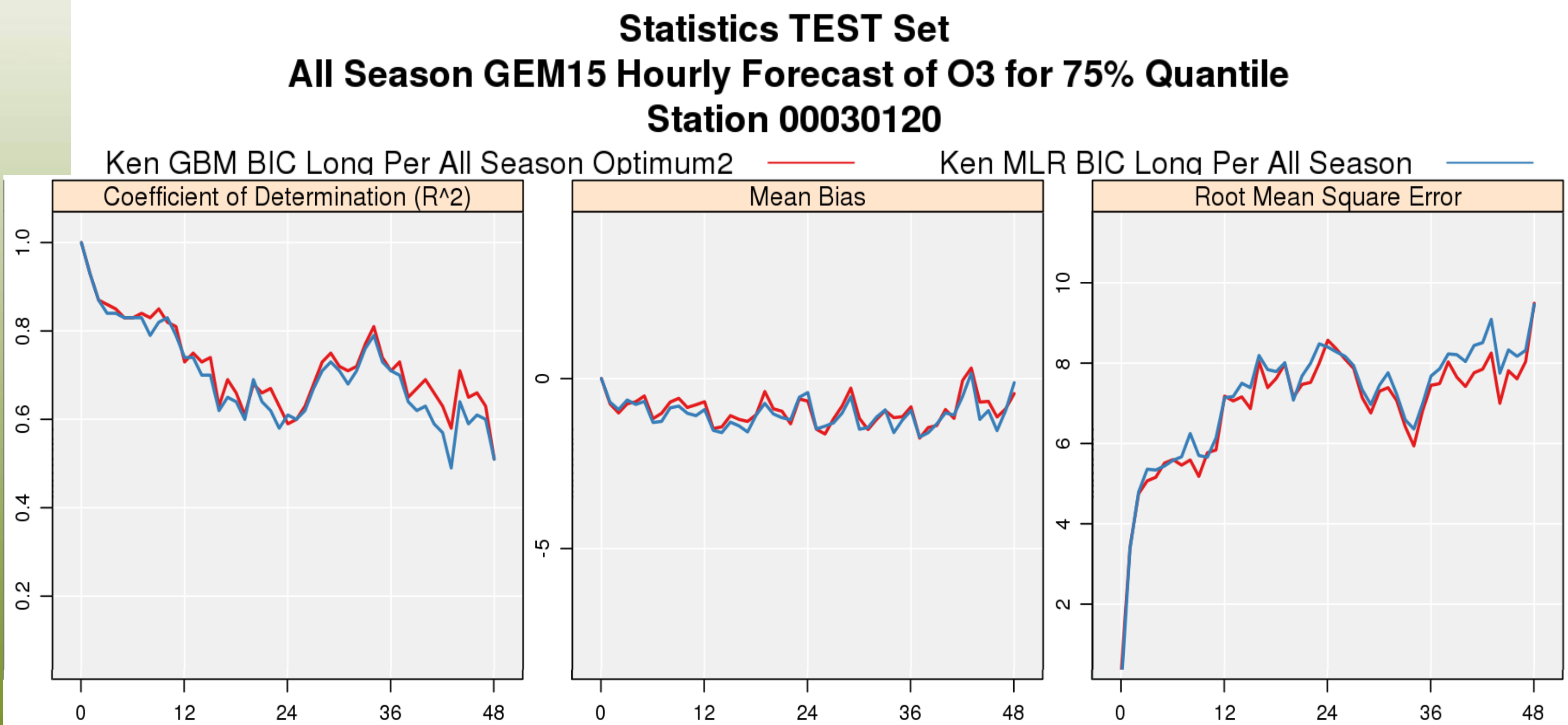


Shrinkage: 0.001

Iterations: 6000

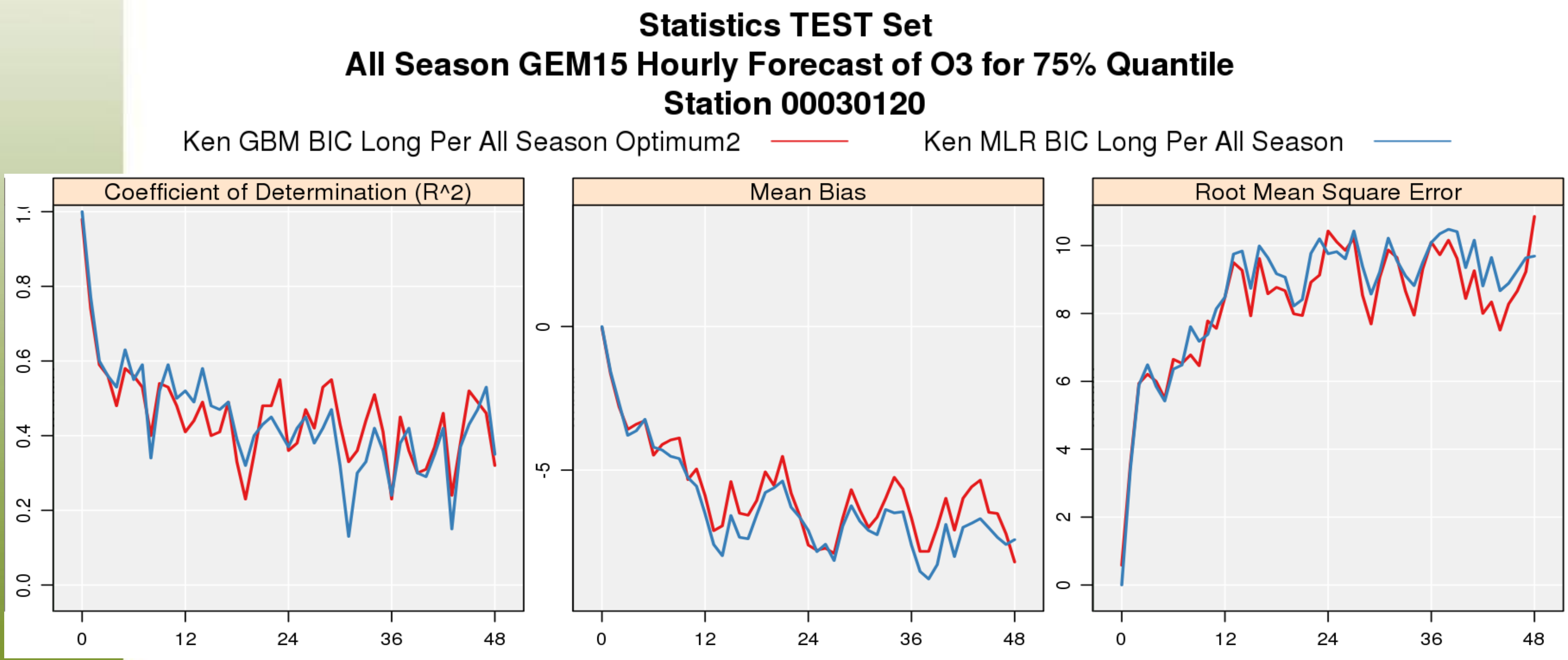
Interactions: 9

Verification Results for all cases



Shrinkage: 0.001 Iterations: 6000 Interactions: 9

Verification Results for Above 75th Percentile cases



Shrinkage: 0.001   Iterations: 6000   Interactions: 9

Conclusion and Future Work

Conclusions:

- GBM is an iterative regression method that searches for basis functions to minimize a selected loss function
- With proper optimization of parameters, GBM is capable of producing marginal improvements over MLR

Future Work:

- gbm.more function to find the optimal number of iterations
- Boosted Regression with other loss functions