Application Of Data Mining And Statistical Learning Approaches For Insights Into Dissolved Oxygen

Delaware Estuary Science & Environmental Summit
January 25-28, 2015
Cape May, NJ

Delaware River Basin Commission
DELAWARE • NEW JERSEY
PENNSYLVANIA • NEW YORK
UNITED STATES OF AMERICA
WWW.DRBC.NET

John Yagecic, P.E.

Thomas Fikslin, Ph.D.

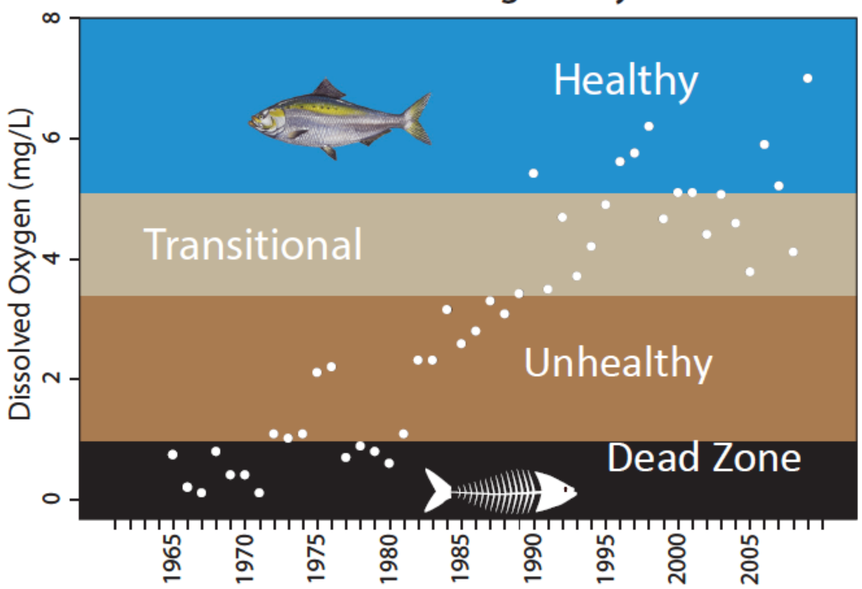
This Presentation

- Motivation for this effort
- The Data Sets
- Identification of Temporally Optimal Explanatory Variables
- The 3 resulting models
- Climate change assessment
- Interpretation and Conclusions

Delaware River Basin Commission
DELAWARE • NEW JERSEY
PENNSYLVANIA • NEW YORK
UNITED STATES OF AMERICA
WWW.DRBC.NET

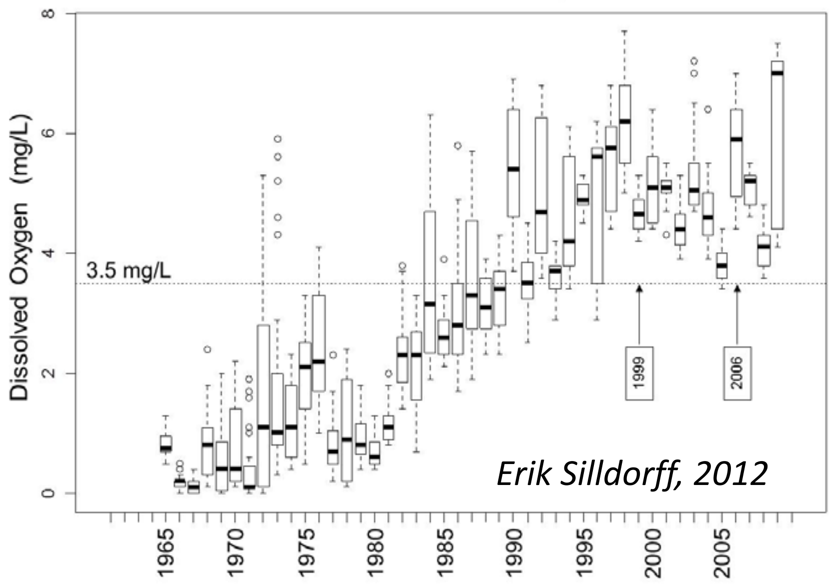
Status of Oxygen in the Delaware Estuary

Ben Franklin Bridge - July Data



RM 100

Ben Franklin D.O. - July Data



Erik Silldorff, 2012

- Designated use ≠ existing use, EPA Nov. 2009 at WQAC;
- Atlantic Sturgeon listed as endangered species, Feb. 2012;
- Delaware Riverkeeper, others, petition DRBC to upgrade uses, revise criteria, March 2013;
- STAC issues DO brief, current criteria too low for sturgeon, Feb. 2014;
- Delaware River Basin Fish and Wildlife Management Cooperative letter to DRBC to increase DO criteria, April 2014.

Action Moving Forward

- Nutrient Criteria Development Plan
  - Eutrophication model;
    - Deterministic model accounting for enough of the physical, chemical, biological processes to be able to link management scenarios to system response;
    - Long term;
    - High effort;

  - Data Mining & Statistical Learning Exercise (this project);
    - Can **not** be used to link management & system response (more later);
    - May (or may not) inform the Eutrophication model about important drivers, conditions;
    - Some additional insight (?);
    - Relatively quick;
    - Relatively low effort;

    1. Multiple term linear regression model;
    2. Regression tree model;
    3. Random Forest model.

Delaware River Basin Commission
DELAWARE • NEW JERSEY
PENNSYLVANIA • NEW YORK
UNITED STATES OF AMERICA
WWW.DRBC.NET

Data Mining & Statistical Learning

- Statistical learning – set of tools for modeling and understanding complex data sets *(James et al., 2013)*;
- Data mining - computational process of discovering patterns in large data sets involving interdisciplinary methods and tools

  (adapted from http://en.wikipedia.org/wiki/Data_mining)
- Extracting knowledge from data collected for *other purposes*;

- Improbable union of statistical methods and computer science / programming;
- Beware – new and exciting tools for getting a wrong answer.

Daily Data 2000 through 2010

NOAA TIDES & CURRENTS

- Median tidal elevation at Philadelphia;
- Minimum tidal elevation at Philadelphia;
- Maximum tidal elevation at Philadelphia;
- Tidal Range at Philadelphia;

- Air Temperature Mean at PHL;
- Air Temperature Max PHL;
- Air Temperature Min PHL;
- Dewpoint PHL;
- Precipitation Total PHL;
- Wind gust PHL;
- Relative Humidity PHL;
- Mean Wind PHL;
- Barometric Pressure PHL;
- Wind Direction PHL;
- Maximum Wind PHL;

The Pennsylvania State CLIMATOLOGIST

Home Data Features

- Daily radiation total at PHL;

National Solar Radiation Data Base

1991- 2010 Update

No loadings!

USGS science for a changing world

National Water Information System: Web Interface

- Discharge at Trenton;
- Specific conductance at Trenton;
- DO Sat. at Trenton;
- pH Median at Trenton;
- pH Range at Trenton;
- Water Temperature Mean at Trenton;
- Specific Conductance Max at Ben Franklin;
- Specific Conductance Min at Ben Franklin;
- Specific Conductance Mean at Ben Franklin;
- Specific Conductance Range at Ben Franklin;
- pH Max at Ben Franklin;
- pH Min at Ben Franklin;
- pH Median at Ben Franklin;
- pH Range at Ben Franklin;
- Water Temp Min at Ben Franklin;
- Water Temp Max at Ben Franklin;
- Water Temp Mean at Ben Franklin;
- Water Temperature Range at Ben Franklin;

- DO Sat. at Ben Franklin;

DRBC

Delaware River Basin Commission

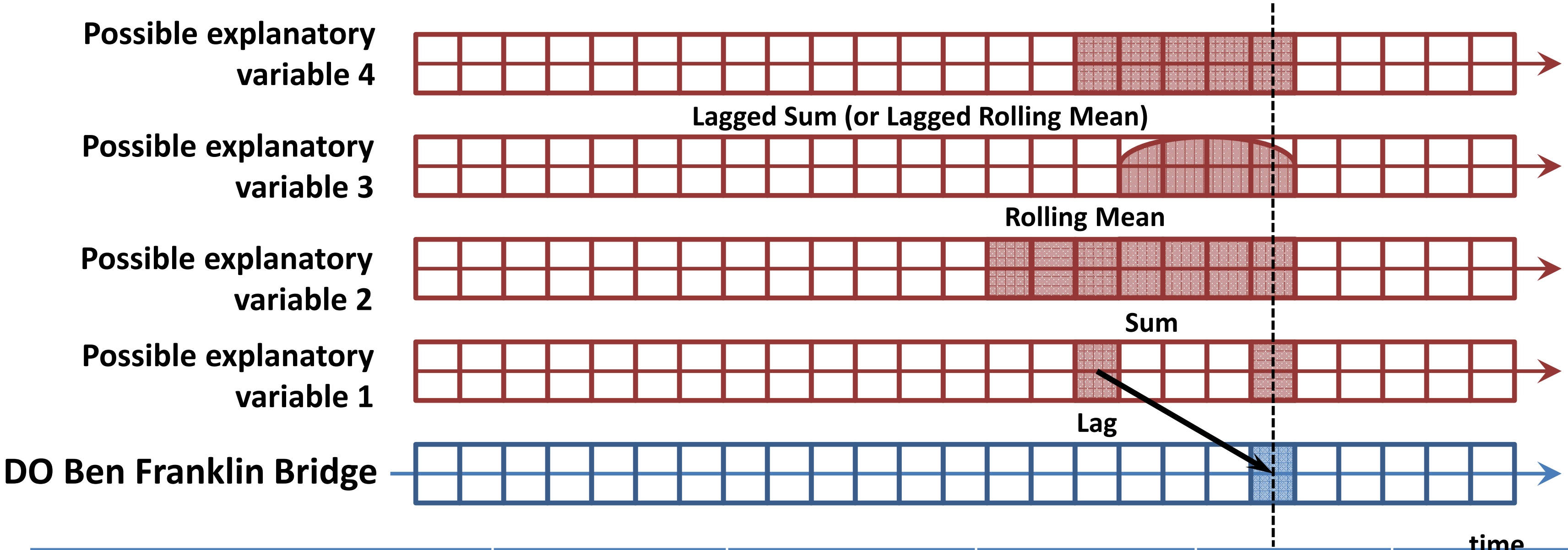
DELAWARE • NEW JERSEY
PENNSYLVANIA • NEW YORK
UNITED STATES OF AMERICA

WWW.DRBC.NET

Temporal Complexity Problem

Delaware River Basin Commission
DELAWARE • NEW JERSEY
PENNSYLVANIA • NEW YORK
UNITED STATES OF AMERICA
WWW.DRBC.NET

Video Link


Possible explanatory variable 4
Lagged Sum (or Lagged Rolling Mean)
Possible explanatory variable 3
Rolling Mean
Possible explanatory variable 2
Sum
Possible explanatory variable 1
Lag
DO Ben Franklin Bridge
time

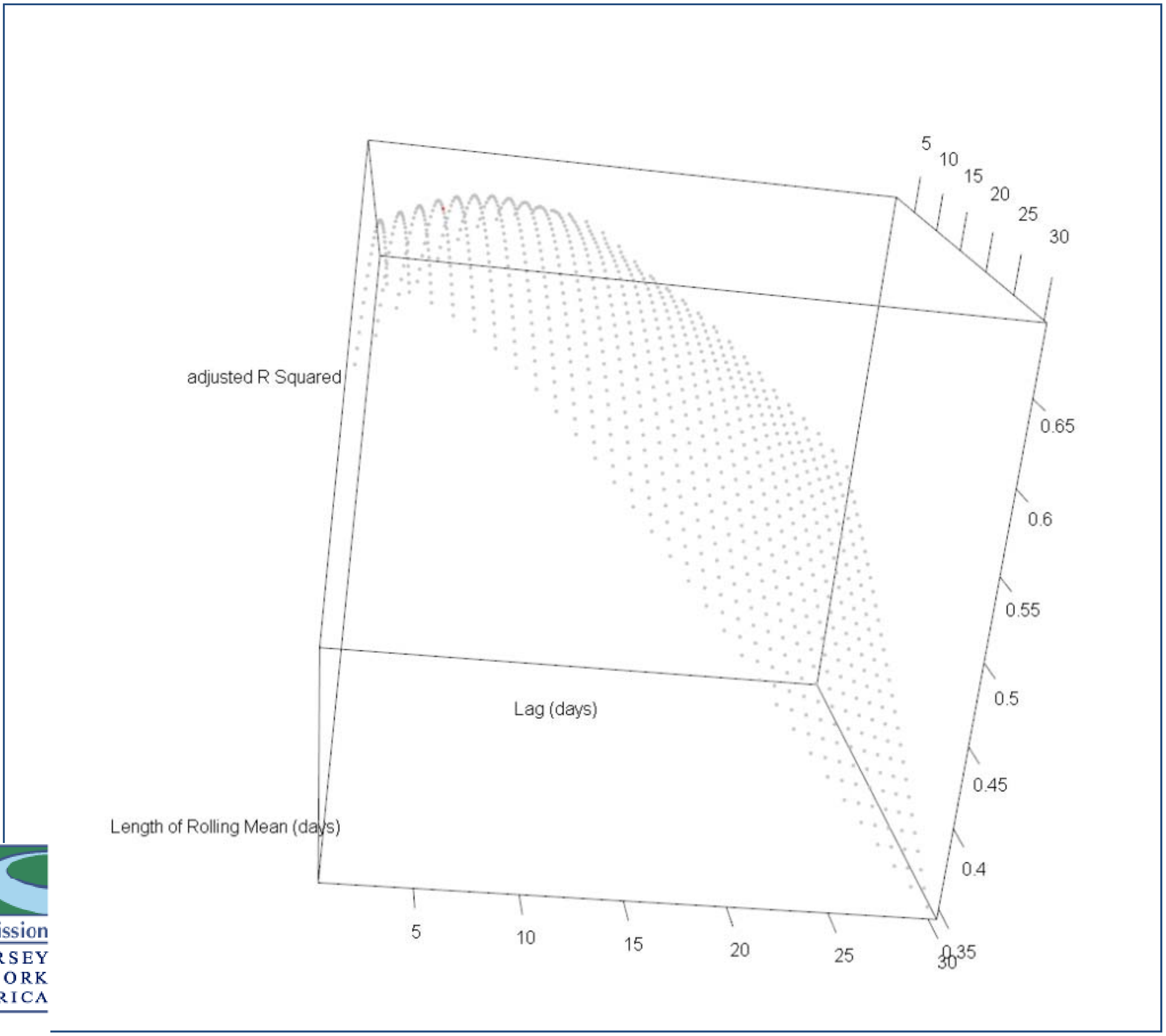
	Explanatory Variables	Averaging Periods	Summing Periods	Lagging Periods	Subtotal
Lag	35			30	1,050
Sum	17		30		510
Rolling Mean	35	30			1,050
Sum + Lag	17		30	30	15,300
Rolling Mean + Lag	35	30		30	31,500
		Possible Additional Explanatory Variables			49,410

Select Raw and Temporally Optimized Variables

Raw Variable	R^2	Temporally Optimized Version	R^2
Radiation	0.009	Radiation_Mean_27_Lag_5	0.220
AirTempMax	0.298	AirTempMax_Sum_22_Lag_5	0.622
DewPoint	0.325	DewPoint_Sum_22	0.566
dischargeTrenton	0.126	dischargeTrenton_Sum_30	0.353
MaxWind	0.027	MaxWind_Mean_26	0.292
spcTrenton	0.324	spcTrenton_Mean_16	0.413
MeanWind	0.026	MeanWind_Mean_27	0.343
tempTrenton	0.525	tempTrenton_Mean_16_Lag_4	0.686

Delaware River Basin Commission
DELAWARE • NEW JERSEY
PENNSYLVANIA • NEW YORK
UNITED STATES OF AMERICA
WWW.DRBC.NET

Computed R^2 value has Structure in response to variation and Rolling Mean and Lag

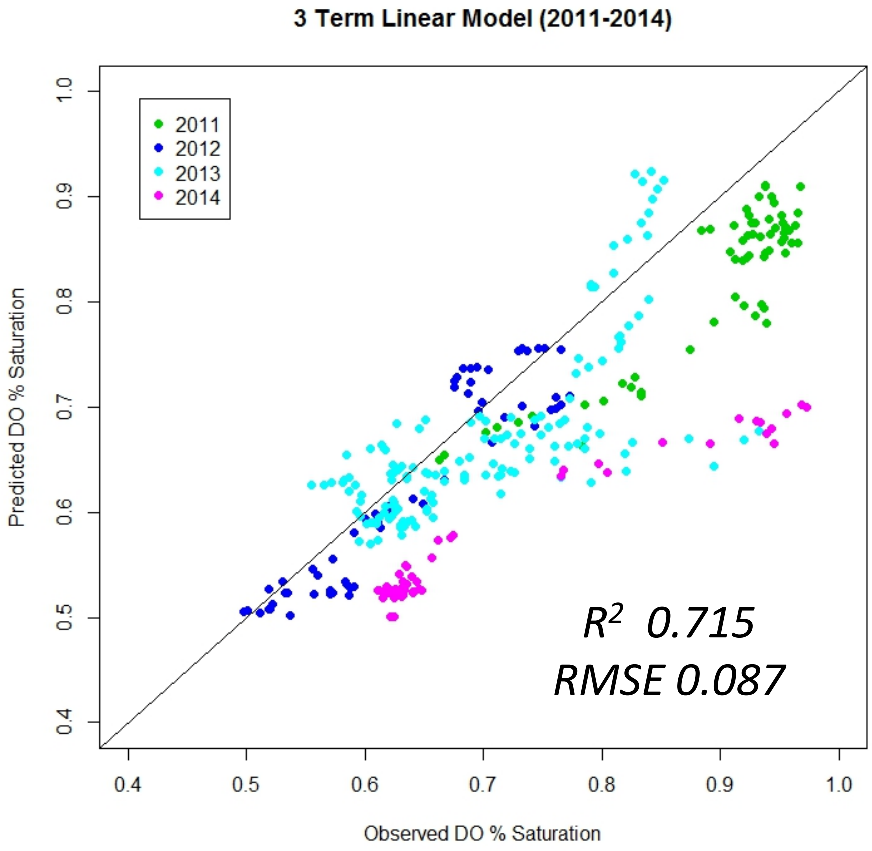


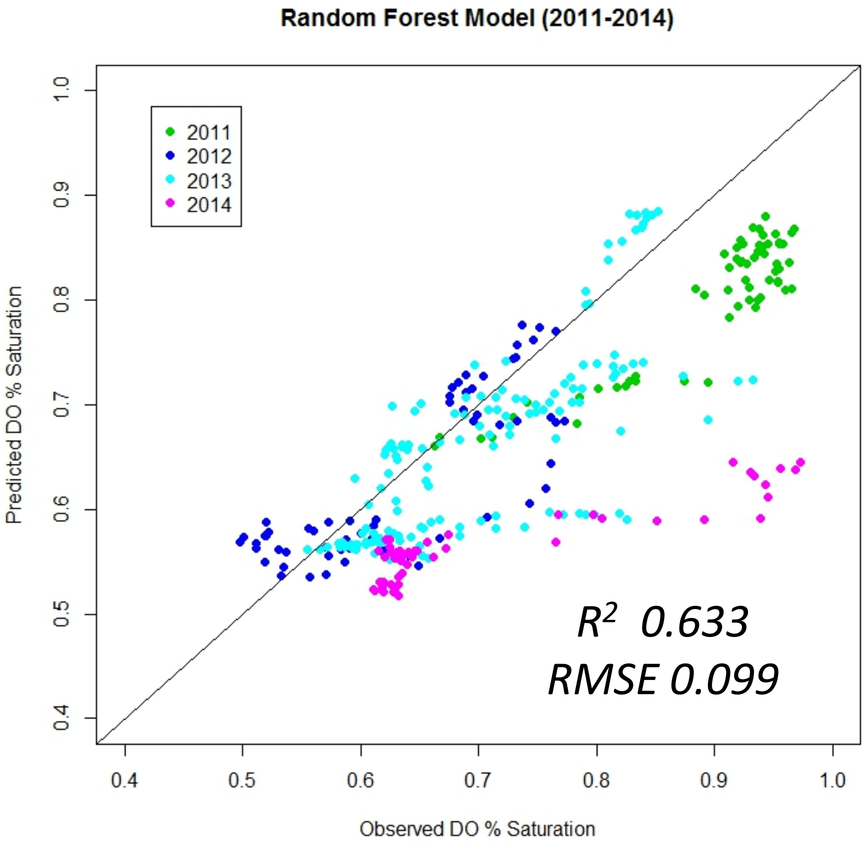
Video Link

The Resulting Models

- Linear Regression Model (3-Term);
  - Any term could be squared or logged (but not both);
  - 8.5 Million possibilities;
  - Cycled through all 8.5 Million to identify the best possible within the training set;
- Regression Tree (interim step toward….);
- Random Forest Model;

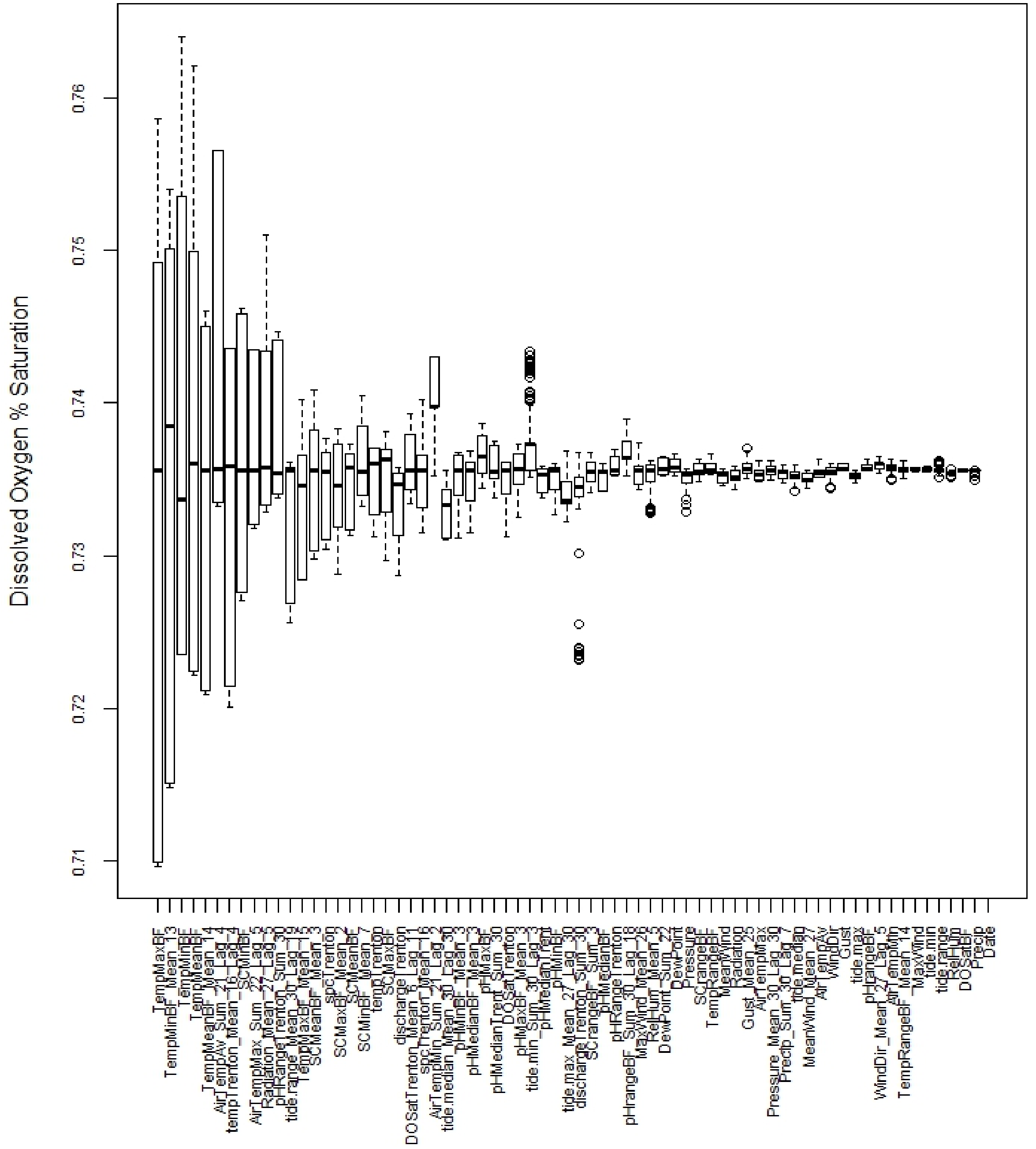
Model Performance
2011-2014 data (i.e. out of sample)





DOSatBF = 4.862E^{-1} - 1.089E^{-2} x TempMeanBF + 1.205E^{-2} x pHMedianBF^2 - 4.468E^{-6} x spcTrenton_Mean_16^2

Senstivity Analysis Results of Random Forest Model at Ben Franklin

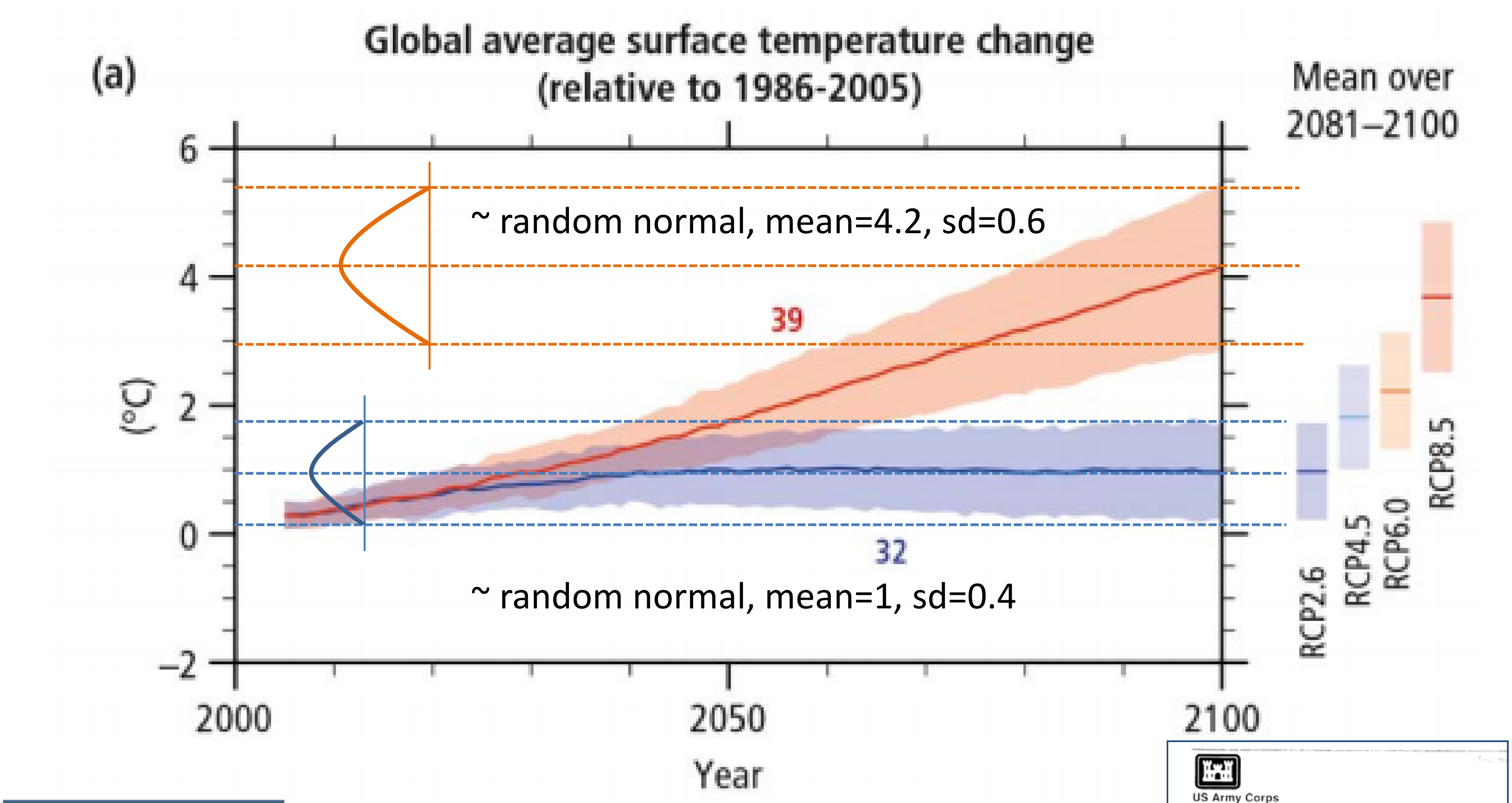


Sensitivity Analysis

- Re-run model sampling one input at a time, all other inputs held at median;
- Shows relative impact individual inputs

Delaware River Basin Commission
DELAWARE • NEW JERSEY
PENNSYLVANIA • NEW YORK
UNITED STATES OF AMERICA
WWW.DRBC.NET

Probing the models for Climate Change

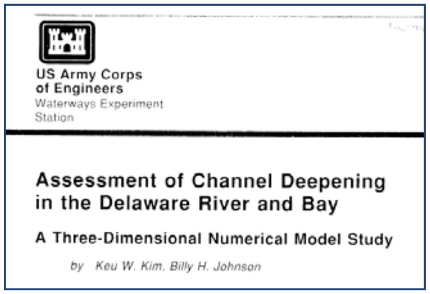


~ random normal, mean=4.2, sd=0.6

~ random normal, mean=1, sd=0.4

Delaware River Basin Commission
DELAWARE • NEW JERSEY
PENNSYLVANIA • NEW YORK
UNITED STATES OF AMERICA
WWW.DRBC.NET

Source: Climate Change 2014 Synthesis Report. Intergovernmental Panel on Climate Change. www.ipcc.ch Retrieved Dec. 22, 2014.



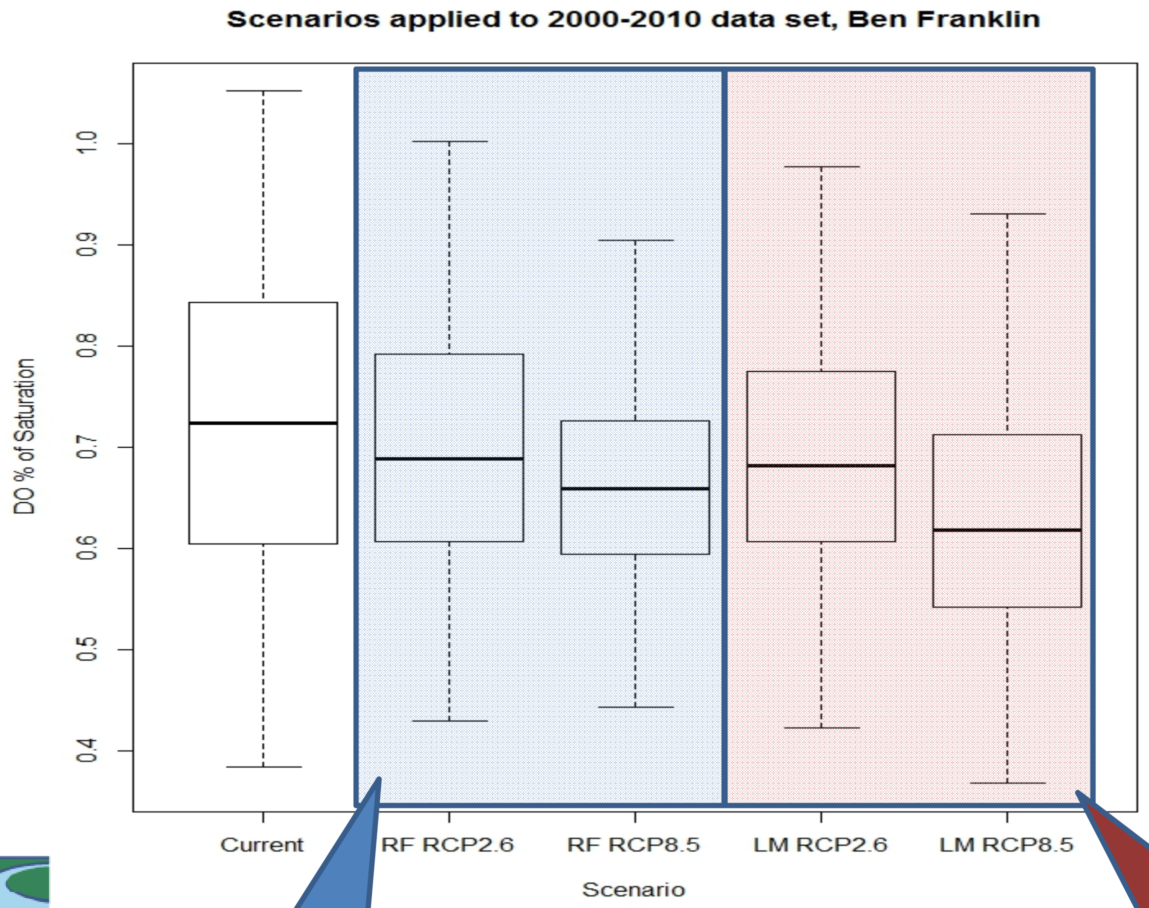
US Army Corps
of Engineers
Waterways Experiment
Station

Assessment of Channel Deepening
in the Delaware River and Bay

A Three-Dimensional Numerical Model Study

by Keu W. Kim, Billy H. Johnson

Climate Change Probe Results



Random Forest

3-Term Linear Regression

Delaware River Basin Commission
DELAWARE • NEW JERSEY
PENNSYLVANIA • NEW YORK
UNITED STATES OF AMERICA
WWW.DRBC.NET

Interpretation

- Both forms (3-Term LM and Random Forest) responsive to various expressions of temperature;
- Results suggest an un-accounted for variable (especially important in summer 2014);
- Re-emphasizes the need for deterministic Eutrophication model;
- Both forms unsuccessful at Chester;
- Probably not sufficient for forecasting;
- Climate change likely to exert downward pressure on dissolved oxygen at Ben Franklin;

Application Of Data Mining And Statistical Learning Approaches For Insights Into Dissolved Oxygen

Delaware Estuary Science & Environmental Summit
January 25-28, 2015
Cape May, NJ

Questions?

Delaware River Basin Commission
DELAWARE • NEW JERSEY
PENNSYLVANIA • NEW YORK
UNITED STATES OF AMERICA
WWW.DRBC.NET

John Yagecic, P.E.

Thomas Fikslin, Ph.D.