An Evaluation of Statistical Learning Methods for Gridded Solar Irradiance Forecasting

David John Gagne[1,2]
Sue Ellen Haupt[1]
Seth Linden[1]
Gerry Weiner[1]
Amy McGovern[3]
John Williams[4]

1. National Center for Atmospheric Research
2. OU CAPS/School of Meteorology
3. OU School of Computer Science
4. The Weather Company/WSI

NCAR

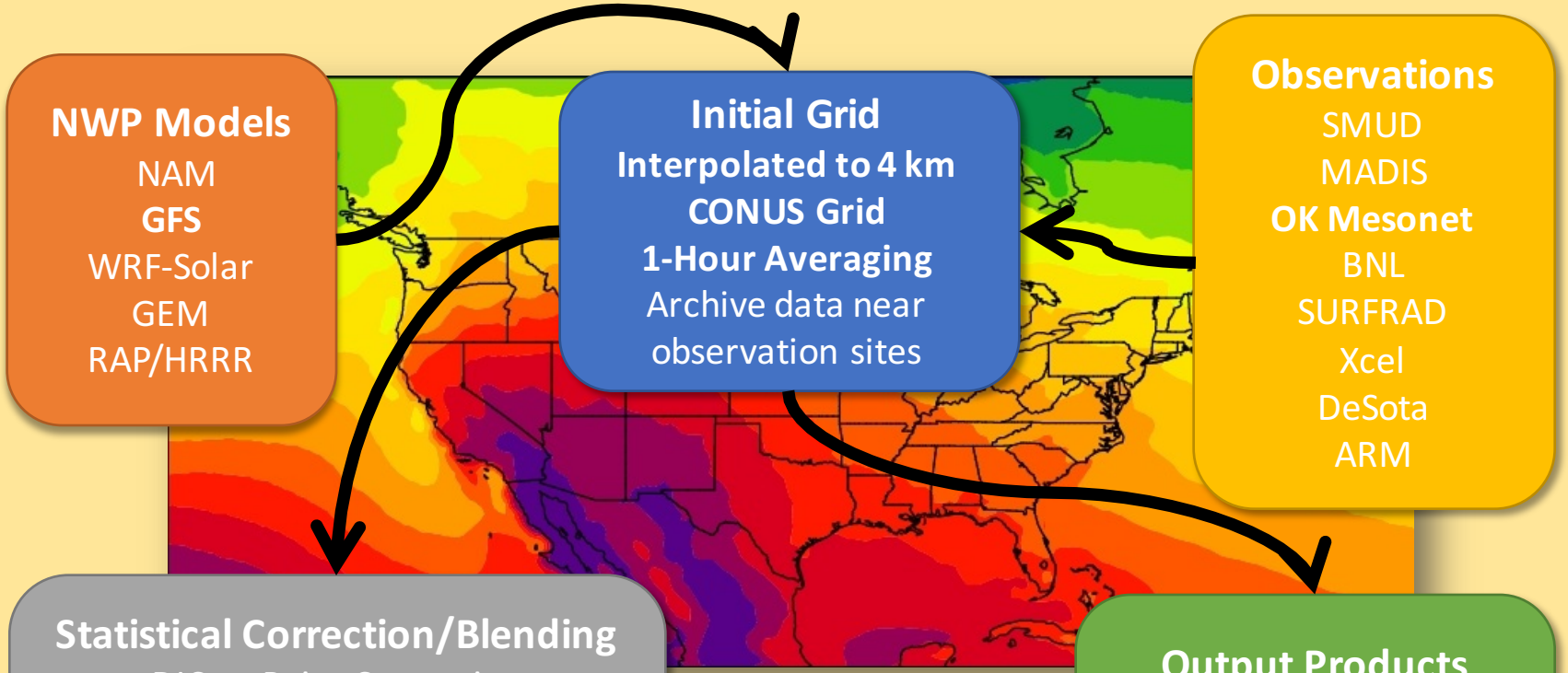
Motivation

- Solar power generation is growing rapidly and comes from a mix of utility scale and residential scale sources
- Long series of solar irradiance observations not available at many locations
- Gridded solar irradiance forecasts can provide necessary spatial coverage but need to be calibrated
- What is the best calibration strategy?
  - Choice of statistical learning model
  - Types of inputs
  - One model applied everywhere or blend of separate models?

Proposed Solution: GRidded Atmospheric Forecast System

(GRAFS)

Gridded Atmospheric Forecasts: GRAFS-Solar



NWP Models
NAM
GFS
WRF-Solar
GEM
RAP/HRRR

Initial Grid
Interpolated to 4 km
CONUS Grid
1-Hour Averaging
Archive data near observation sites

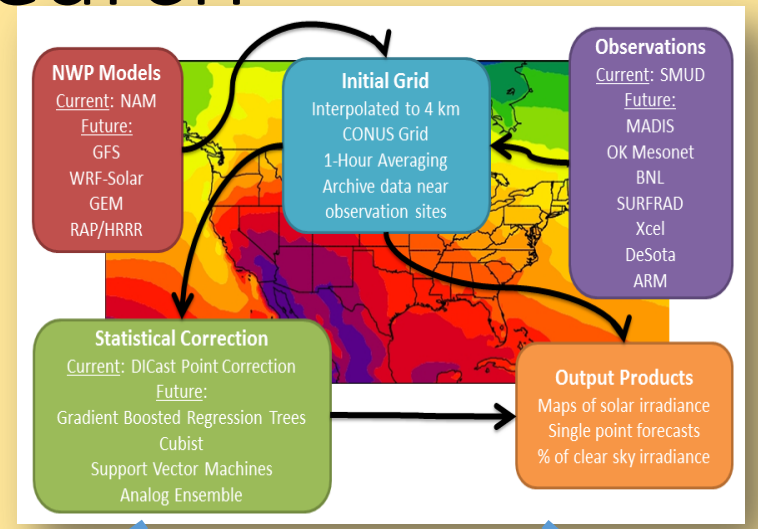
Observations
SMUD
MADIS
OK Mesonet
BNL
SURFRAD
Xcel
DeSota
ARM

Statistical Correction/Blending
DICast Point Correction
Gradient Boosted Regression Trees
Cubist
Random Forests
Analog Ensemble

Output Products
Maps of solar irradiance
Single point forecasts
% of clear sky irradiance
Other met. variables

NCAR

Modes of Research



Common Base Framework

Research Mode:

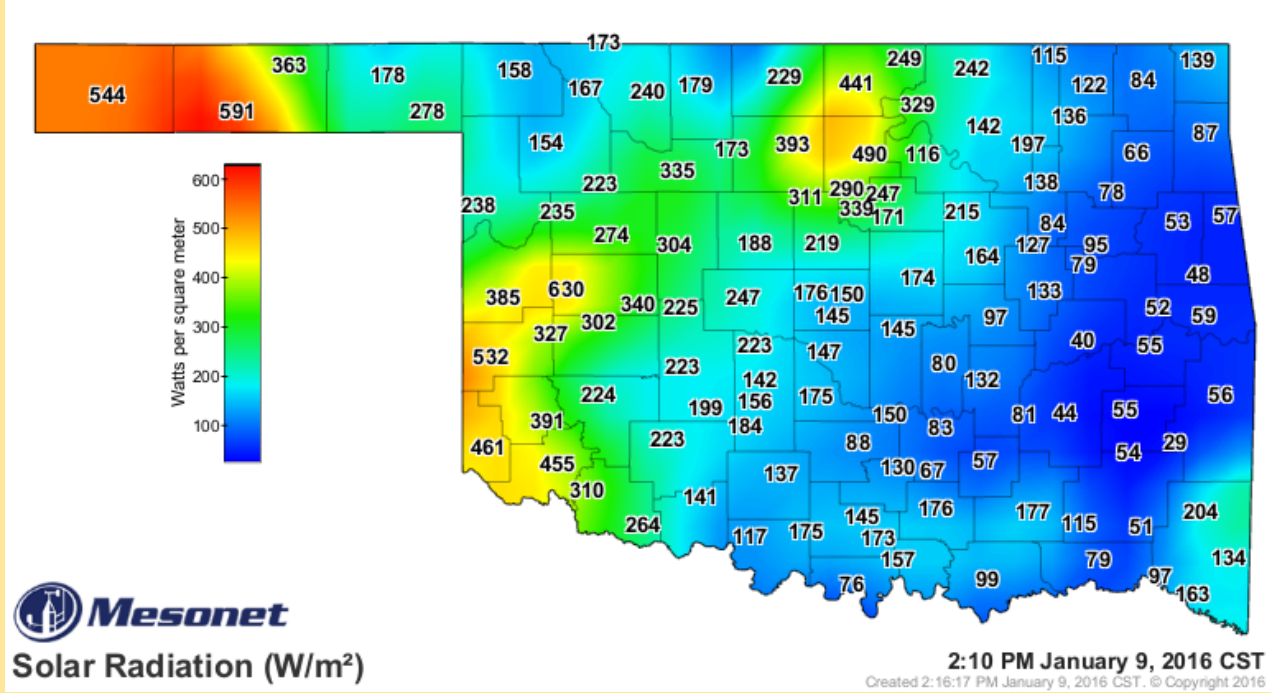
- Open Source
- Community Development
- Sharing paradigm

Application Mode:

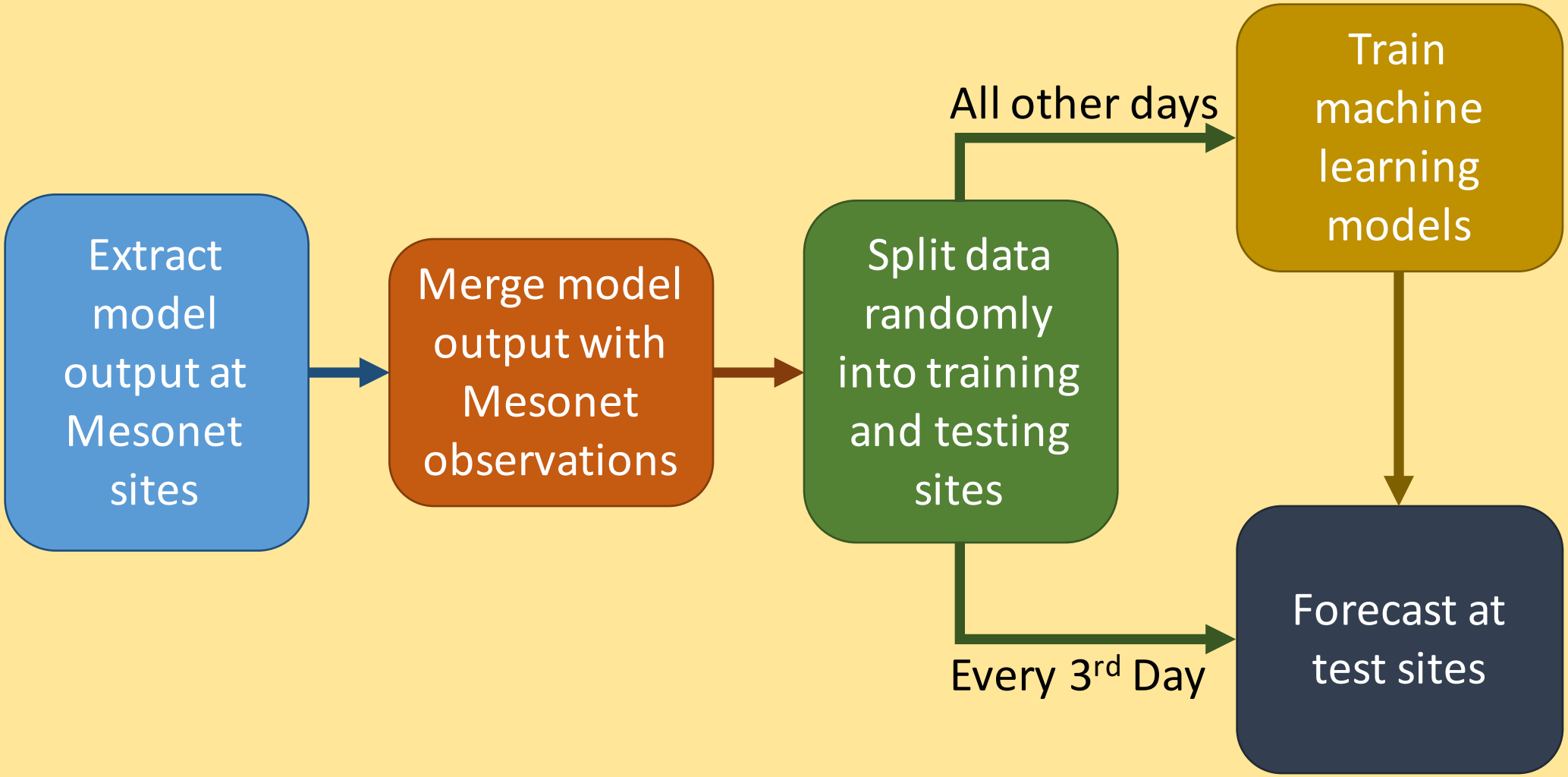
- Includes licensed IP
- Team collaborations
- Include proprietary data

Data

- Forecasts: NOAA NCEP Global Forecast System (GFS)
  - 4 June – 31 August 2015
  - 10-26 hour forecasts
  - Grids interpolated to 4 km spacing
  - Using downward shortwave irradiance, temperature, and cloud cover variables
- Observations: Oklahoma Mesonet (mesonet.org)
  - Li-Cor Pyranometer
  - ~32 km average station spacing



Experiment Procedure



Machine Learning Configurations

Multi Site Models

- Aggregate data from all training sites
- Train single ML model
- Apply at test sites using local information

Single Site Models

- Keep training site data separate
- Train ML model at each training site
- Forecast at training sites and nearest neighbor interpolation to test sites

ML Models (scikit-learn)

Random Forest

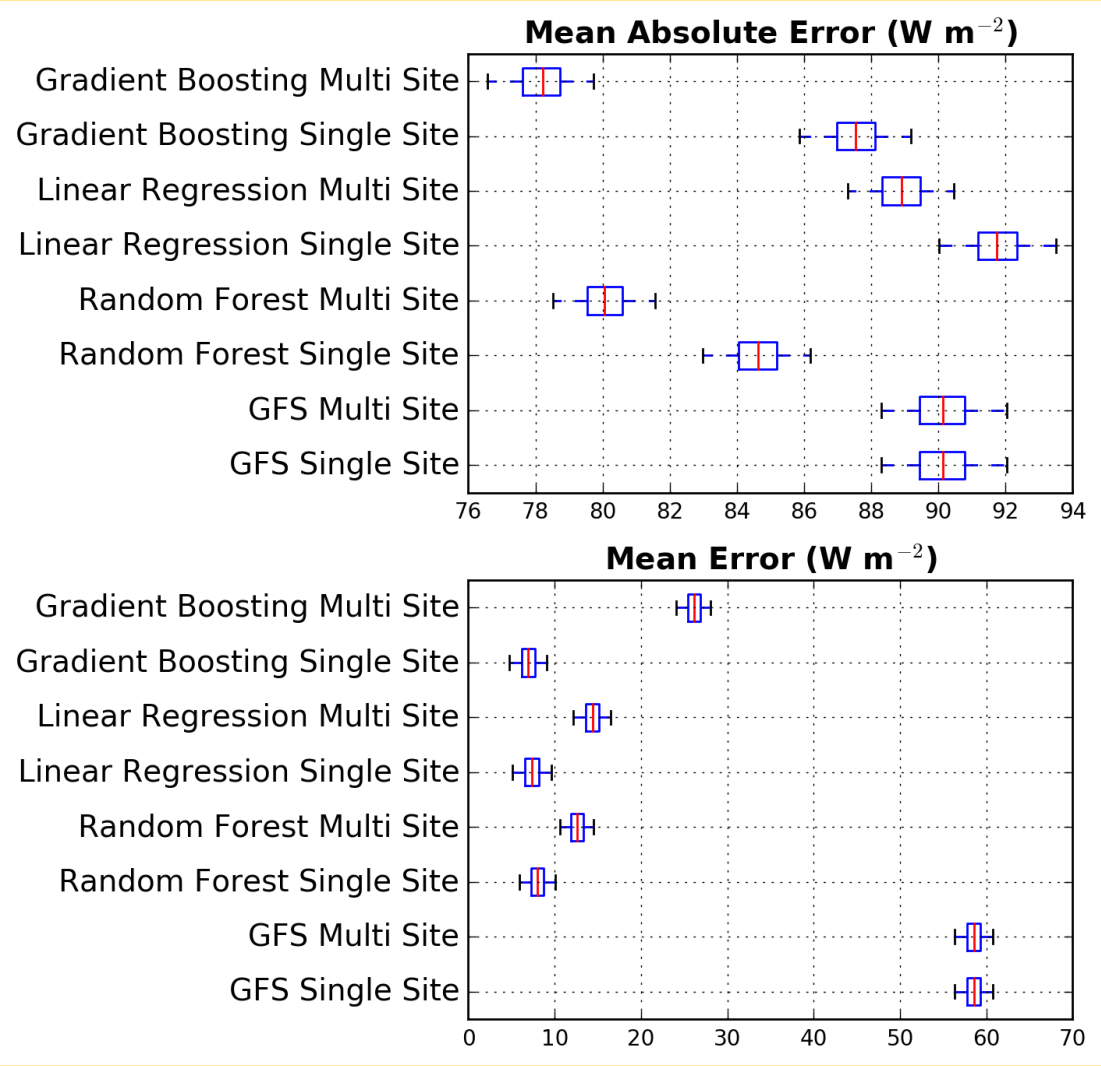
Gradient Boosting Regression

Lasso Linear Regression

ML Models

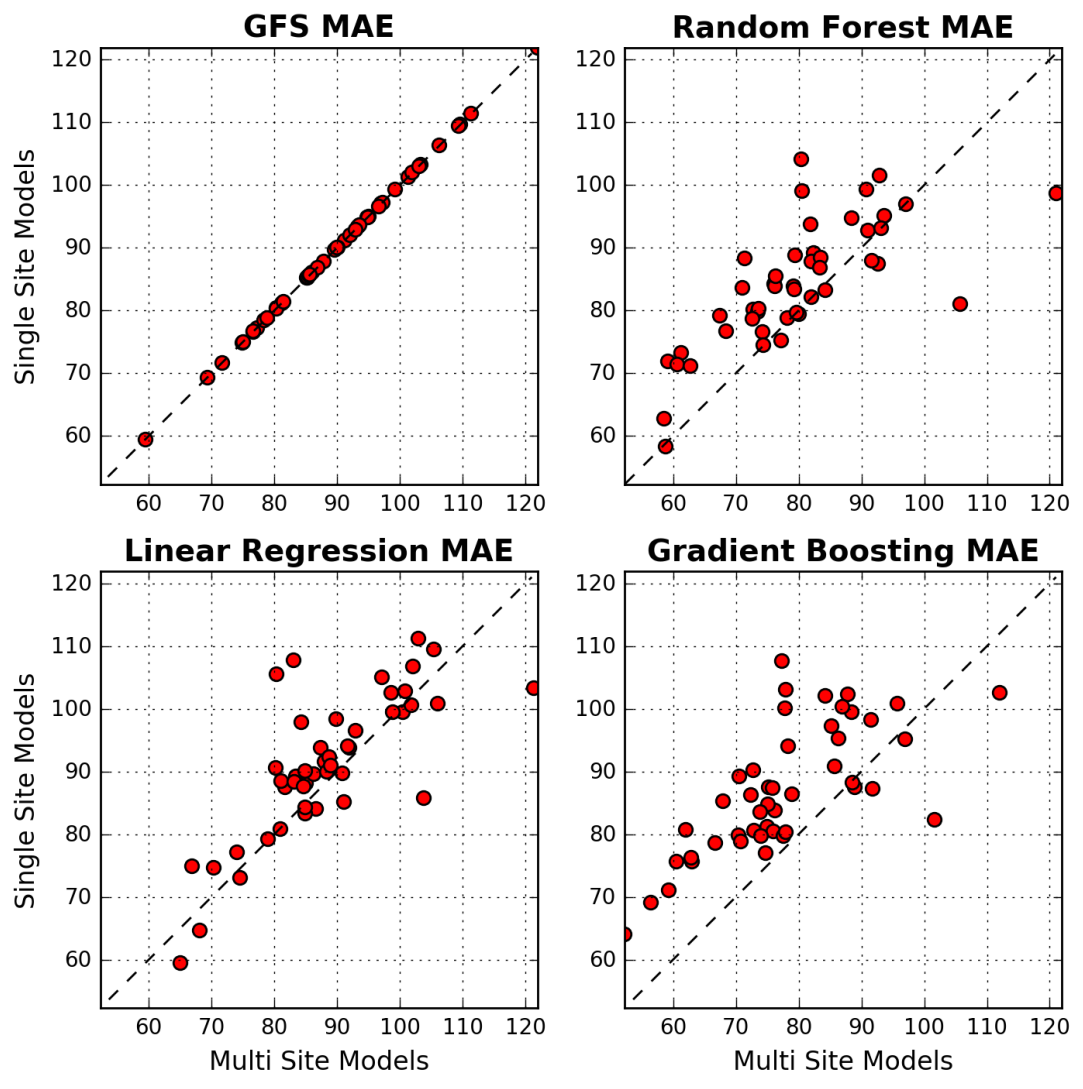
- Software Package: scikit-learn
- Random Forest
  - 500 trees, max depth=8, max features=square root
- Gradient Boosting
  - 500 trees, loss=least absolute deviation, max depth=8, max features=square root
- Linear Regression
  - Lasso regression on top 10 features

Overall Errors by Model



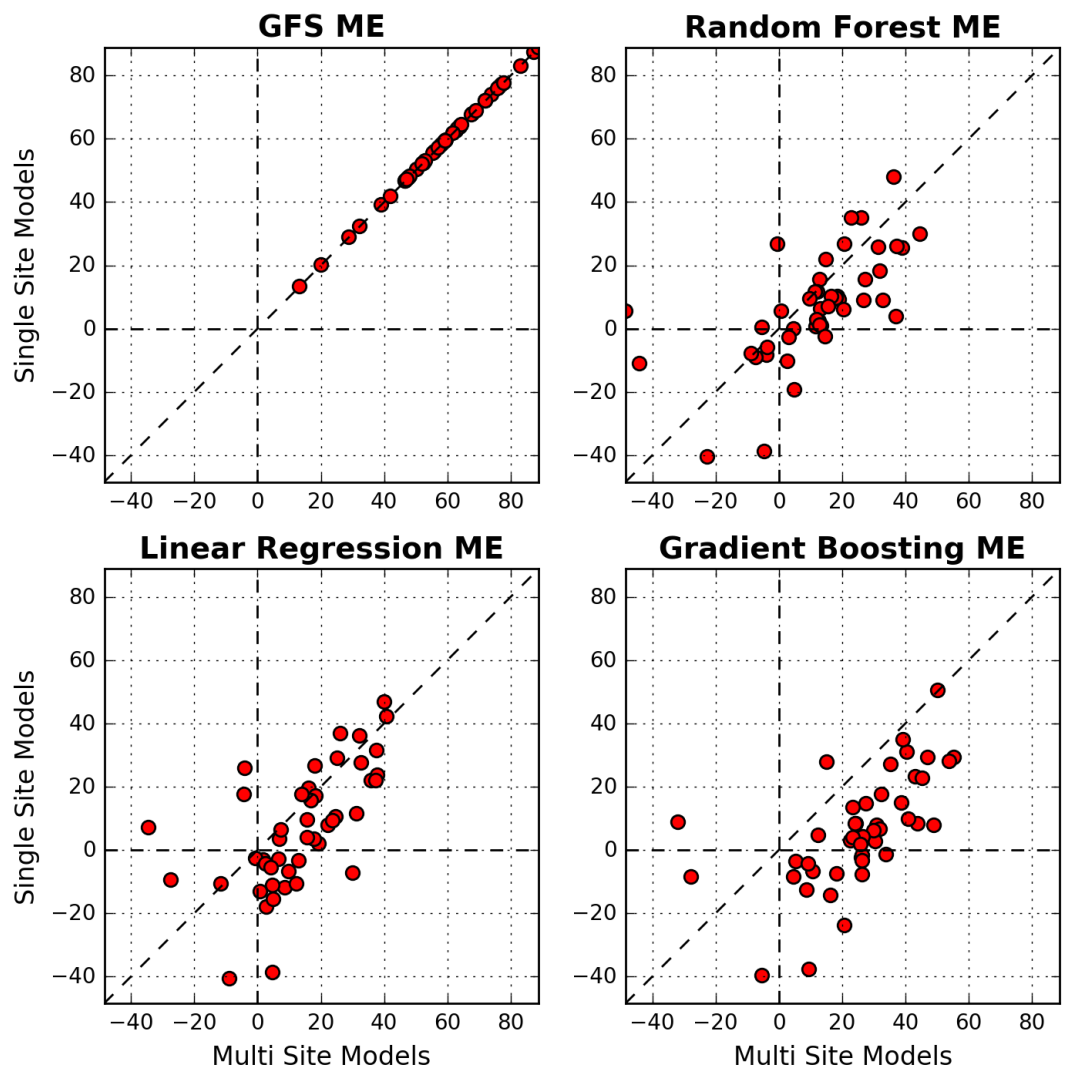
- Whiskers indicate 95% bootstrap confidence interval
- Random Forest and Multi Site Gradient Boosting significantly outperform raw GFS
- Single site linear regression actually worse
- All statistical learning models reduce bias
- Every model still has slight positive bias

Mean Absolute Error by Station



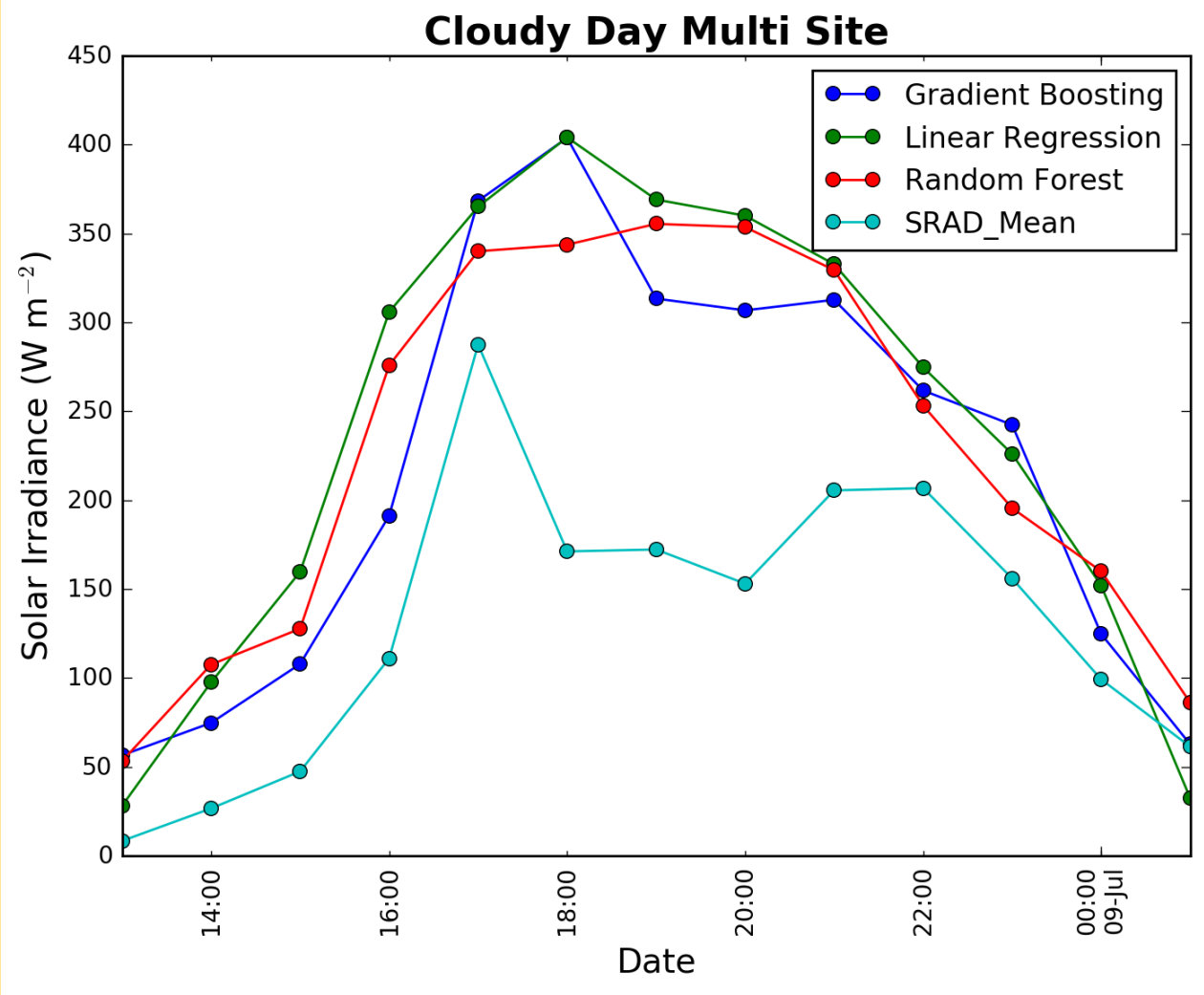
- Plot of mean absolute error by station for the single and multi site models
- Multi Site model provides better forecasts at most sites for the ML models

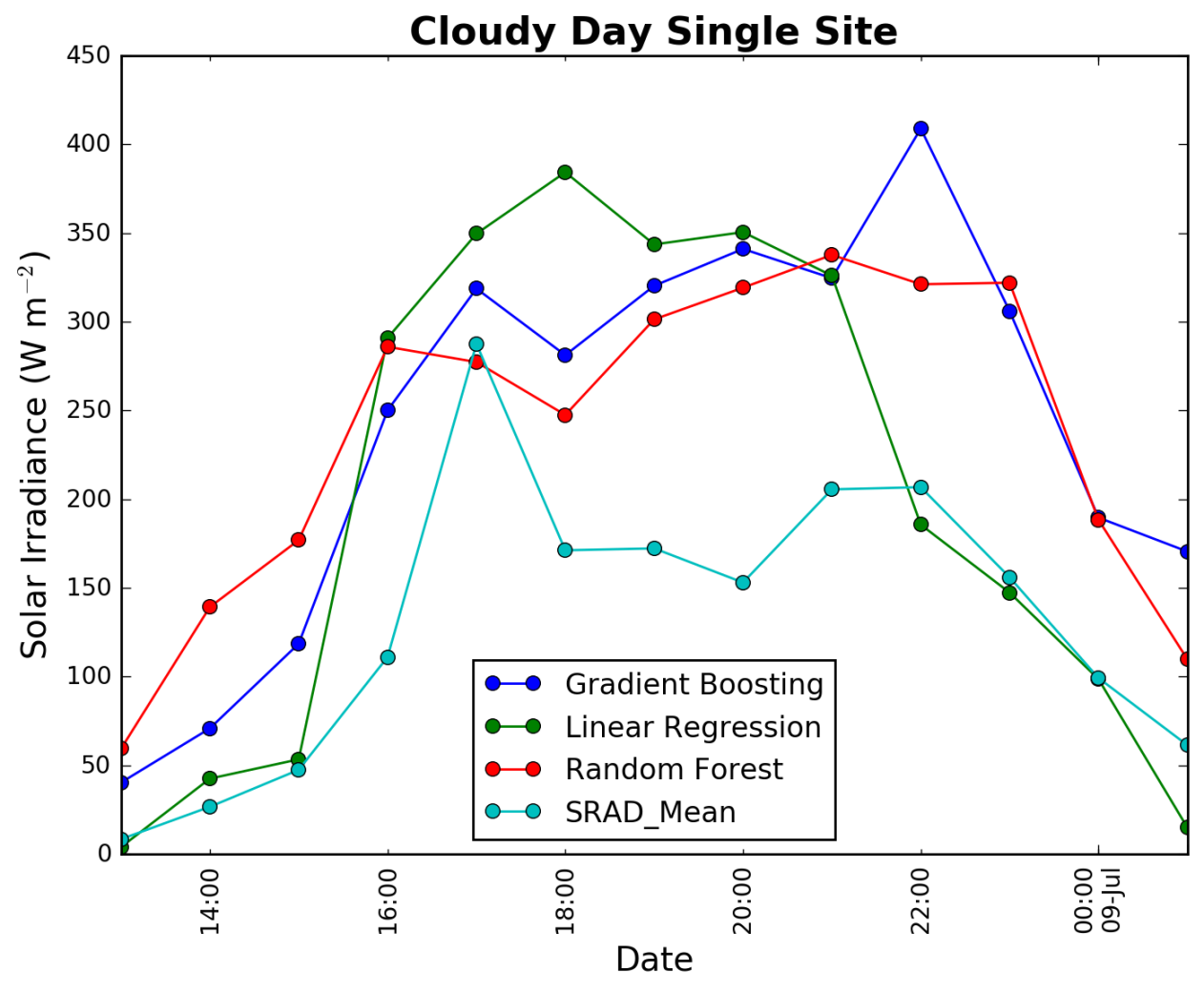
Mean Absolute Error by Station



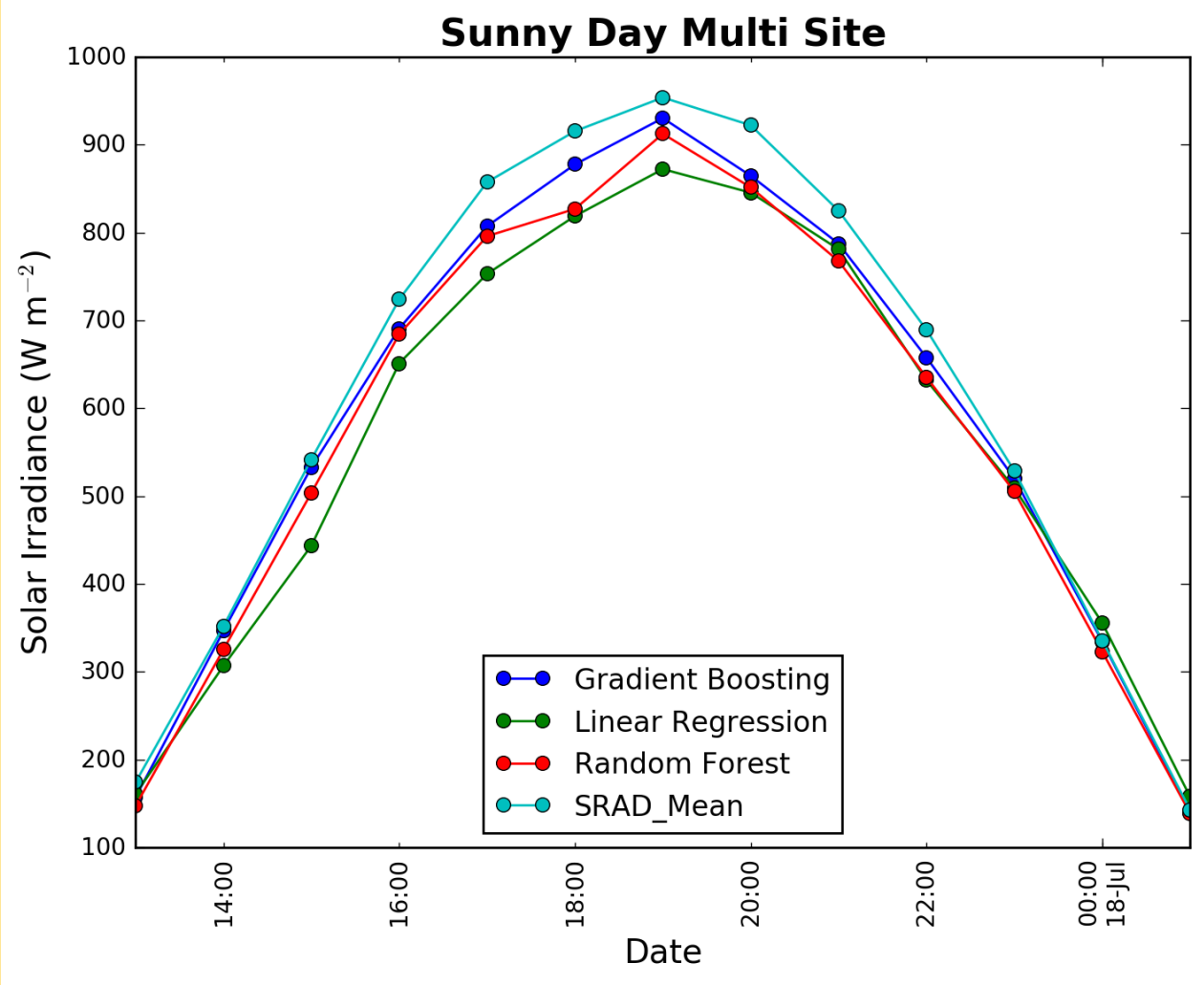
- Plot of mean error by station for the single and multi site models
- Multi Site models have higher biases than single site models for most sites
- Linear regression and gradient boosting have a number of sites with biases of opposite signs

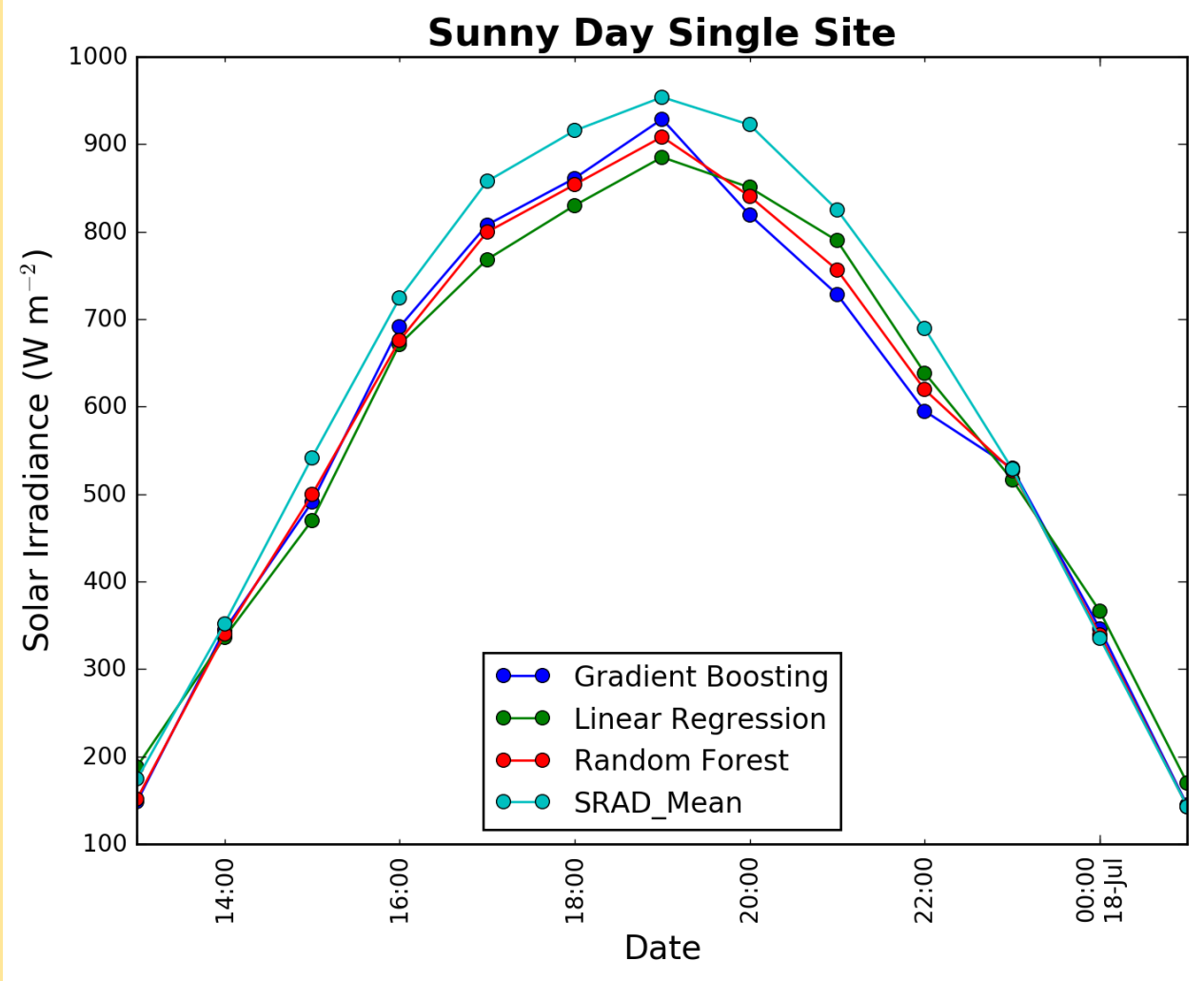
Solar Irradiance Forecast Example: Cloudy





Solar Irradiance Forecast Example: Cloudy





Variable Importance

Random Forest

Variable	Importance
Neighbor Mean Solar Radiation	0.165
Correlation Coef. Solar Radiation	0.155
Solar Radiation	0.141
Neighbor Max Solar Radiation	0.139
Neighbor Median Solar Radiation	0.126
Neighbor Min Solar Radiation	0.109
Forecast Hour	0.032
Valid Hour (CST)	0.025
Neighbor Min Temperature	0.017
Correlation Coef. Temperature	0.015

Gradient Boosting

Variable	Importance
Neighbor Max Solar Radiation	0.068
Neighbor Min Solar Radiation	0.066
Solar Radiation	0.064
Neighbor Gradient Solar Radiation	0.064
Neighbor Median Solar Radiation	0.063
Neighbor Mean Solar Radiation	0.061
Temperature Gradient	0.060
Correlation Coef. Solar Radiation	0.058
Correlation Coef. Temperature	0.054
Neighbor Mean Temperature	0.052

Summary

- Multi Site machine learning models provide lower error than single site models
- Multi Site models have a larger bias than Single Site models
- Gradient Boosting regression produces the lowest error
- All statistical models underestimate cloud cover, but Gradient Boosting captures the trends well
- Spatial neighborhood solar irradiance information most important

Contact Info
Email: djgagne@ou.edu
Twitter: @DJGagneDos