The Performance Impacts of Machine Learning Design Choices for Gridded Solar Irradiance Forecasting

Features work from “Evaluating Statistical Learning Configurations for Gridded Solar Irradiance Forecasting”, *Solar Energy*, Under Review.

David John Gagne II, *NCAR*

Sue Ellen Haupt, *NCAR*

Amy McGovern, *University of Oklahoma*

John Williams, *The Weather Company, an IBM Business*

Seth Linden, *NCAR*

Doug Nychka, NCAR

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

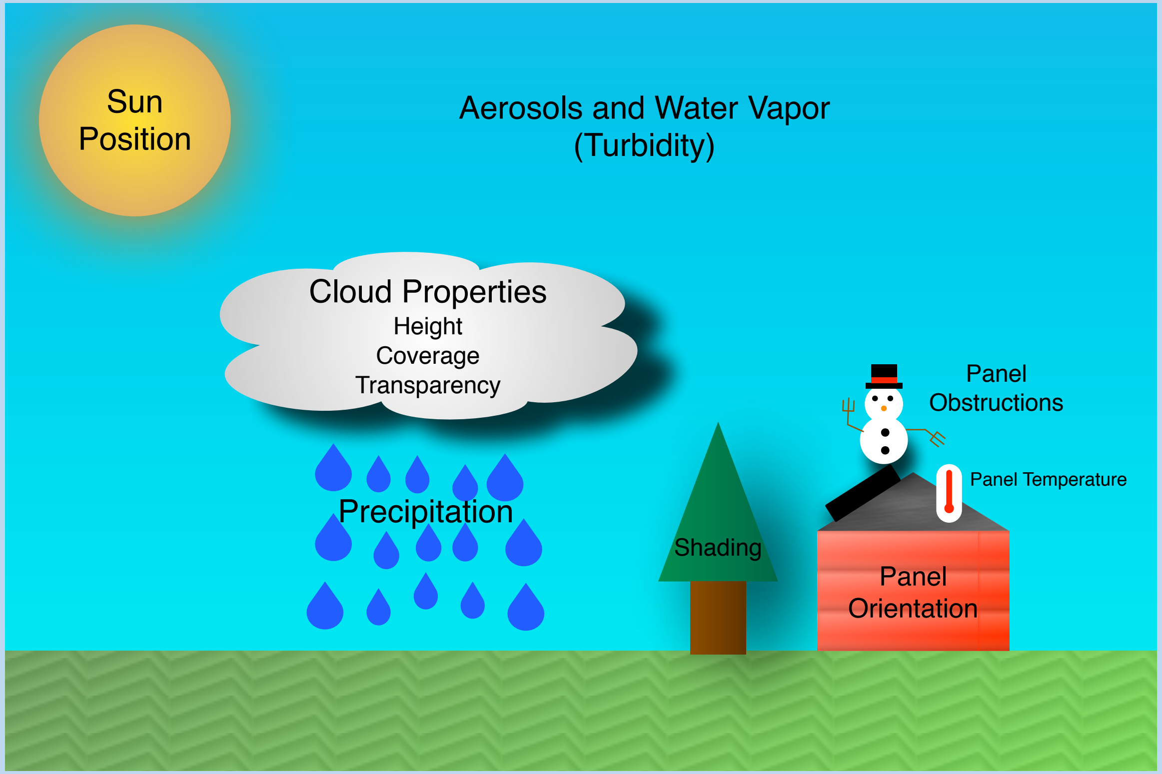
Motivation: Solar Irradiance

- Solar electricity generation continues to grow rapidly and decrease in cost
- Accurate solar irradiance predictions needed by electric utilities to balance supply with expected demand
- Solar power is being generated more at sites that do not have observations or historical records of irradiance
- Contributions
  - Developed a Gridded Atmospheric Forecasting System (GRAFS) for solar irradiance
  - Evaluated different machine learning model configurations for predictive accuracy at unobserved sites for day ahead solar irradiance forecasts

Solar irradiance predictions are needed for sites without historical data (Source: http://www.adventurecats.org/cat-tales/maine-coon-deaf-sailors-ears-sea/)

Solar Forecasting Ingredients

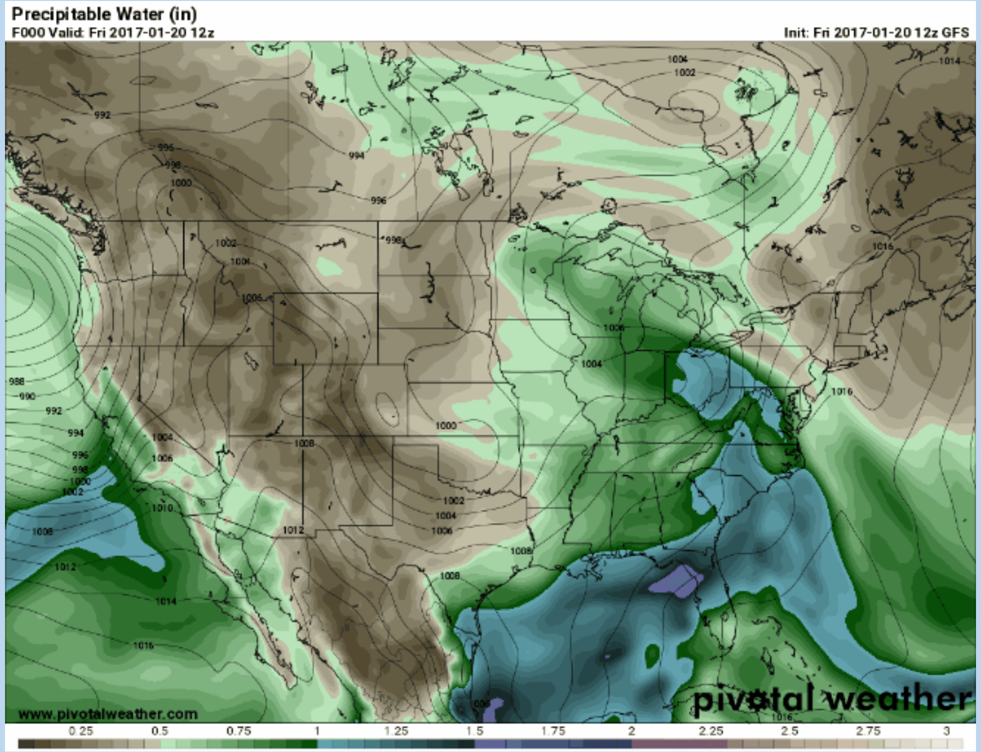
- Position of sun in sky
- Scattering by atmosphere & aerosols
- Cloud cover effects
- Precipitation
- Non-meteorological obstructions

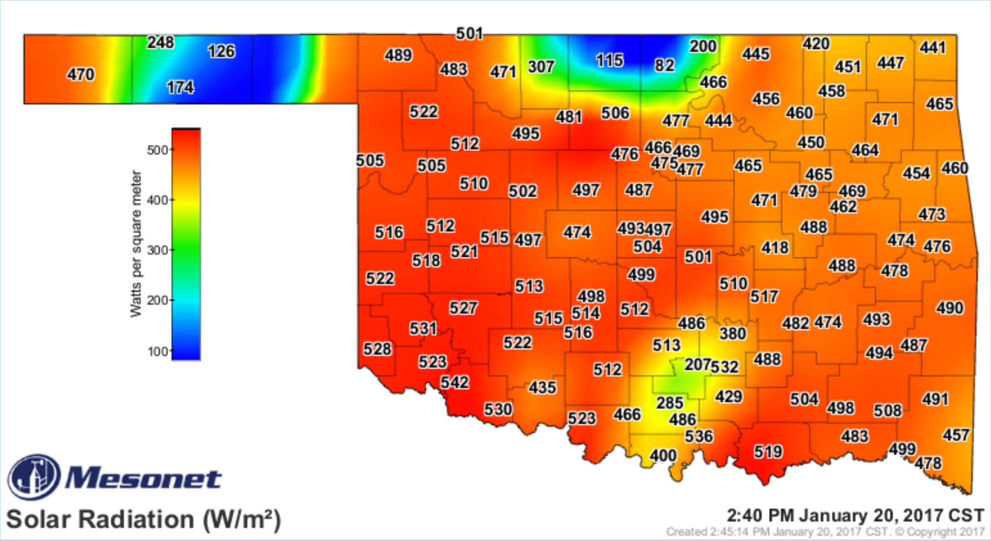


Solar factors diagram from Gagne (2014)

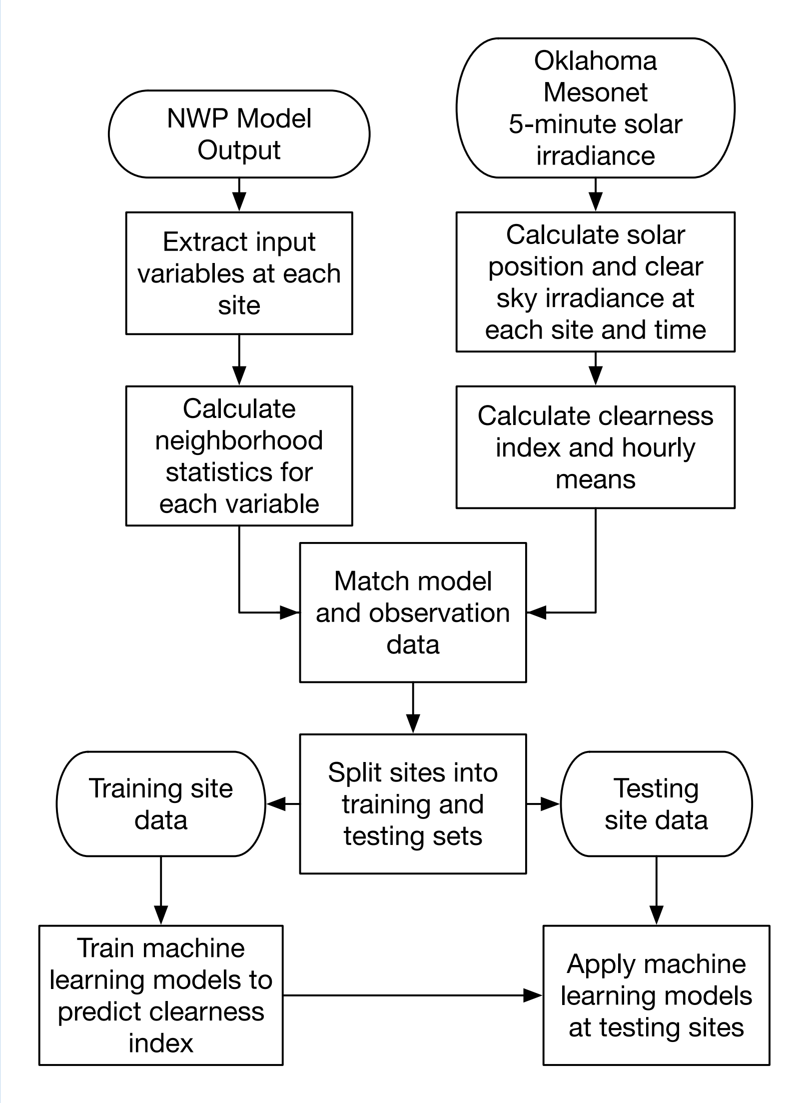
Solar Data

- NOAA Global Forecast System (GFS)
  - Interpolated to 4 km grid
  - 3 hourly output interpolated in time to hourly output
  - Variables: Solar irradiance, temperature, cloud cover, sun angles, spatial statistics
  - Evaluation Period: June-August 2015
- Oklahoma Mesonet (McPherson et al. 2007)
  - Sites record solar irradiance every 5 minutes with a Li-Cor pyranometer
  - Hourly-averaged irradiance and clearness index computed from raw observations
  - Clearness index: ratio of observed irradiance to top-of-atmosphere irradiance



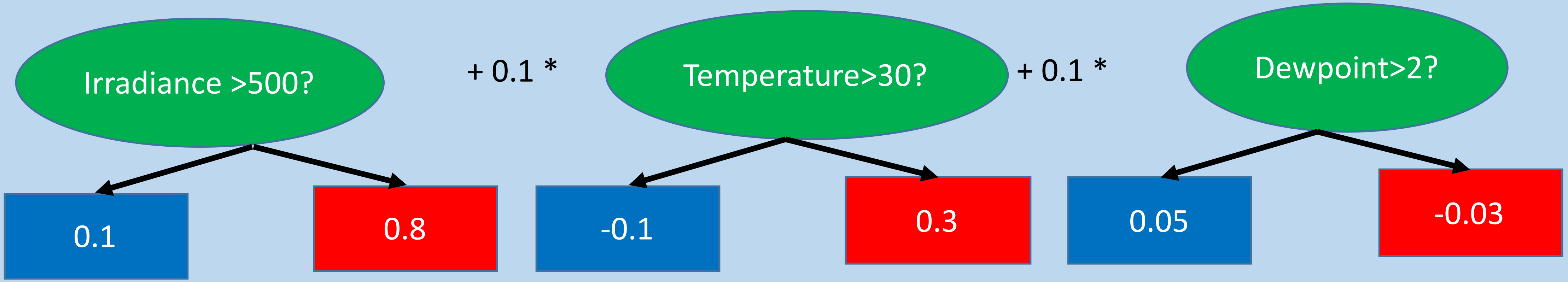


Machine Learning Configurations: Solar



- Mesonet stations randomly split into "training" and "testing" sites
- Evaluation period split into training and testing days: every 3rd day used for testing
- Models: Random Forest, Gradient Boosting, Lasso Linear Regression
- Multi Site Training
  - One machine learning model fitted with all training sites' data
  - Applied at testing sites using input data collocated with site
- Single Site Training
  - Separate machine learning models fitted at each training site
  - Predictions made at training sites and interpolated to testing sites with Cressman interpolation (Cressman 1959)
  - Similar to approach used by Gridded MOS (Glahn et al. 2009)

Gradient Boosting Regression

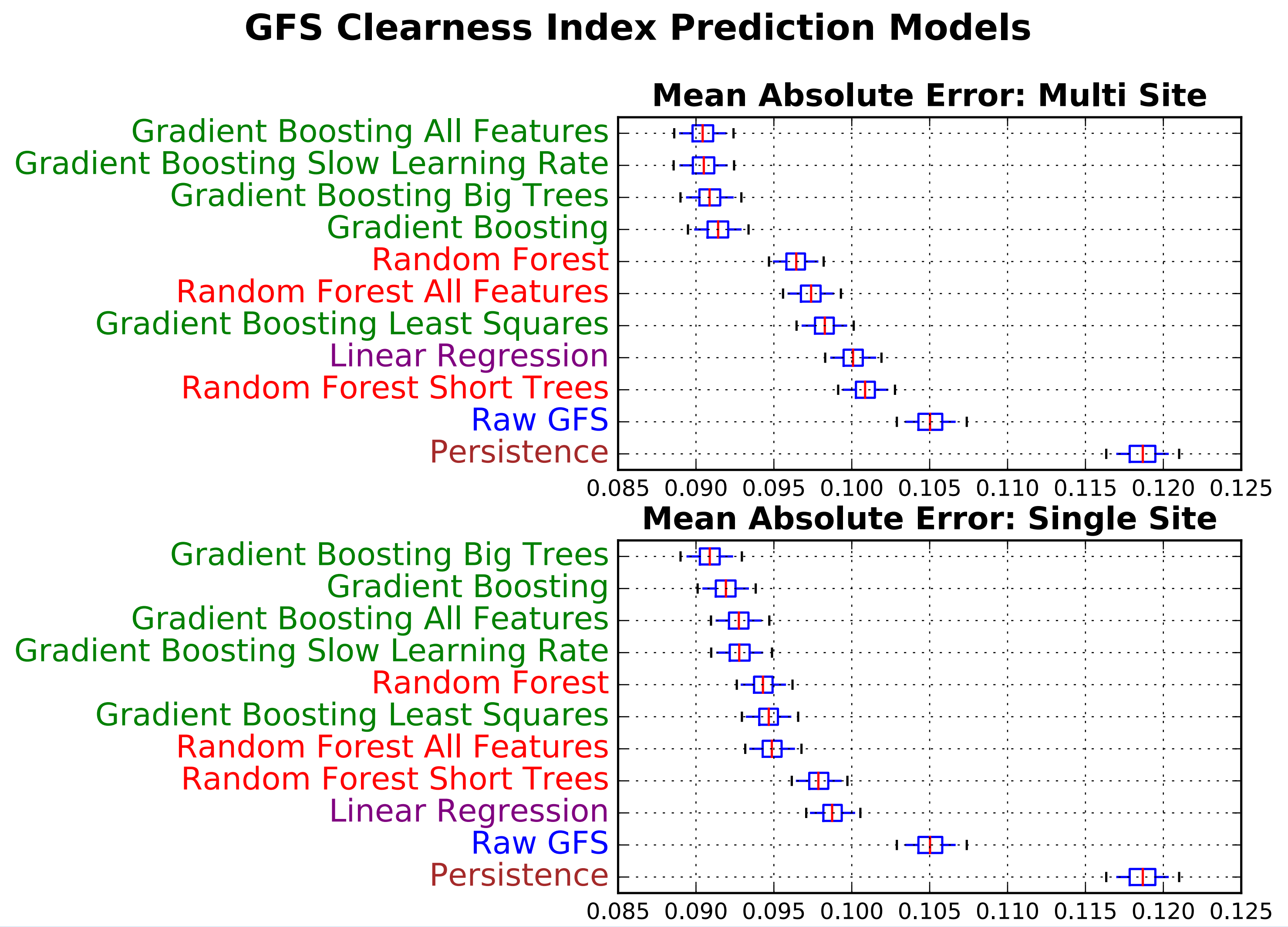


- Stagewise, additive decision tree ensemble
- Initial tree predicts exact value, subsequent trees predict residuals of total predictions from all previous trees
- Used by top 4 finishers of AMS Solar Energy Prediction Contest

Detailed Configuration

- Random Forest
  - Default: 500 trees, min samples split 10, features=sqrt
  - Short Trees: max depth 3
  - All Features: features = all
- Gradient Boosting
  - Default: loss="lad", 500 trees, max depth 5, features = sqrt, learning rate=0.1
  - Least Squares: loss = "ls"
  - Big Trees: min sample split = 10
  - All Features: features = "all"
  - Slow Learning Rate: learn rate = 0.01
- Lasso Linear Regression
  - Top 16 variables by F-Score, Alpha=0.5

Solar: GFS Clearness Index Error



Gradient Boosting: Optimizes with MAE, Tree Depth of 5, Samples subset of features

Gradient Boosting Least Squares: Uses MSE instead of MAE

Gradient Boosting All Features: Evaluates all input features

Gradient Boosting Slow Learning Rate: Uses a learning rate of 0.01 instead of 0.1

Gradient Boosting Big Trees: Allows trees to grow to minimize training samples in each branch

Random Forest: fully grown trees, evaluates subset of features

Random Forest All Features: evaluates all features

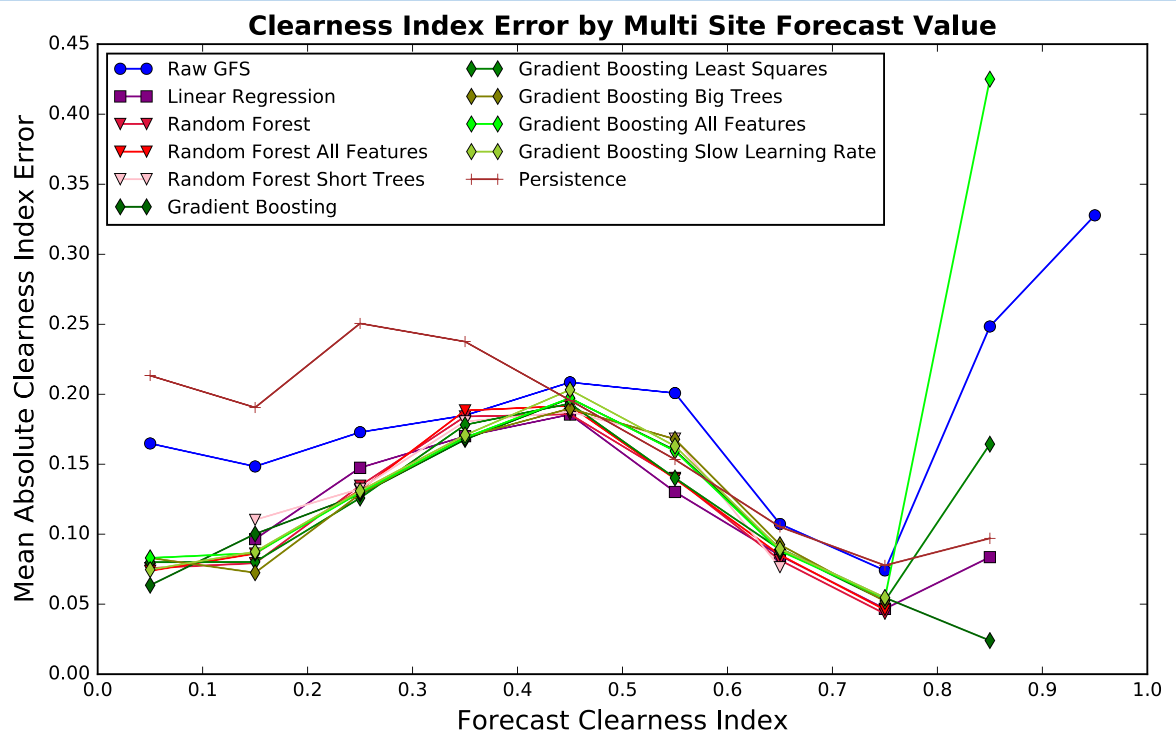
Random Forest Short Trees: tree depth of 3

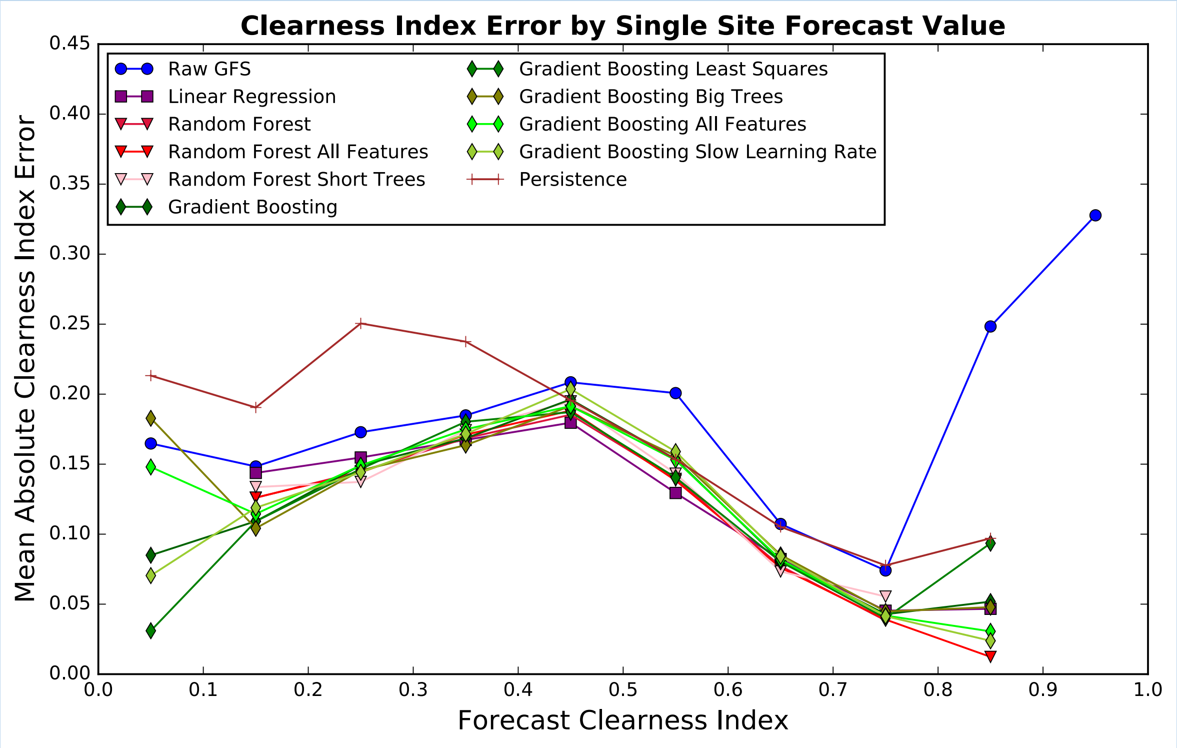
Linear Regression: Lasso with top 16 variables

Raw GFS: Downward shortwave irradiance

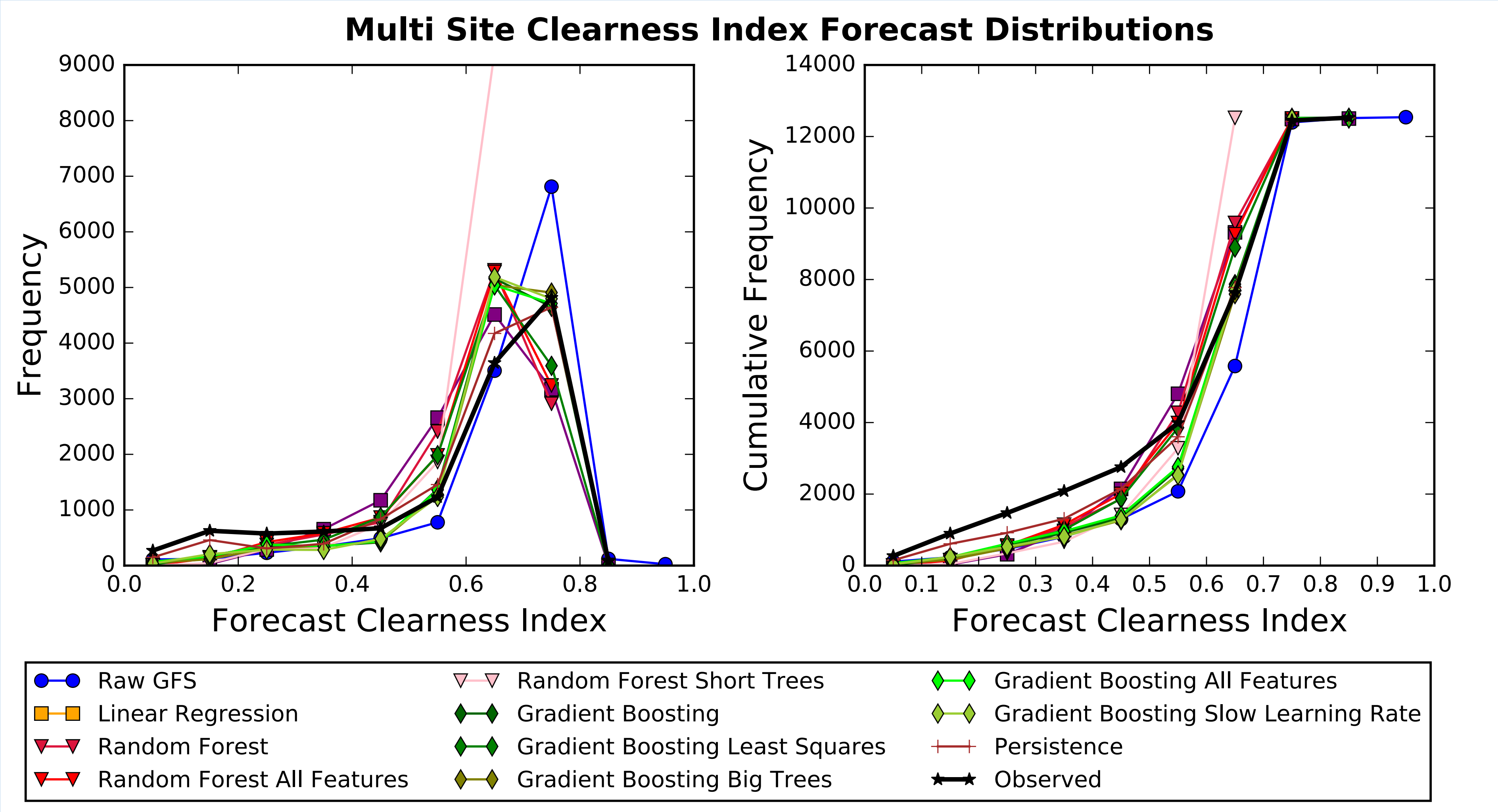
Persistence: Interpolated irradiance at test sites based on observations from 24 hours before

GFS Solar Distributions

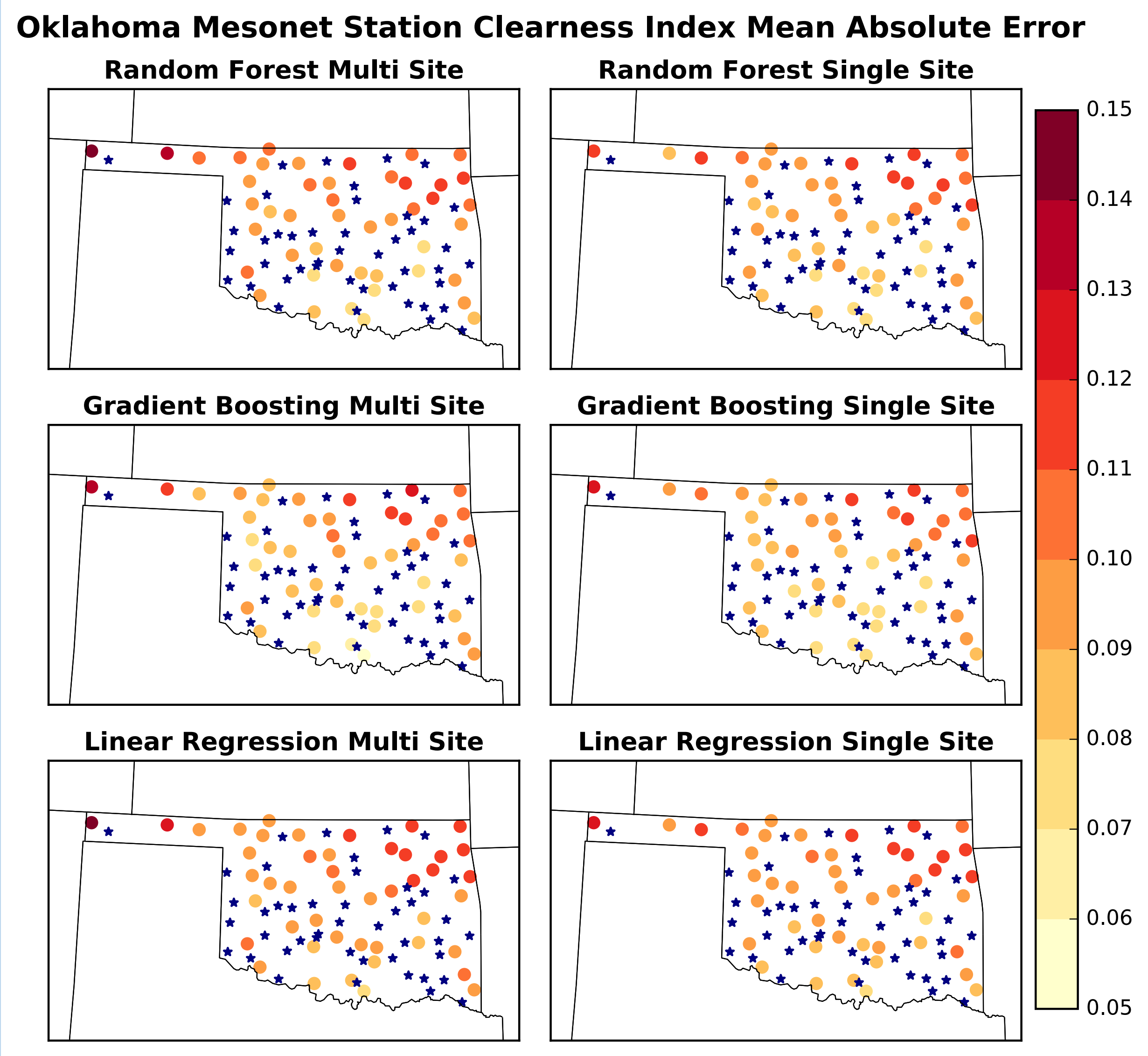


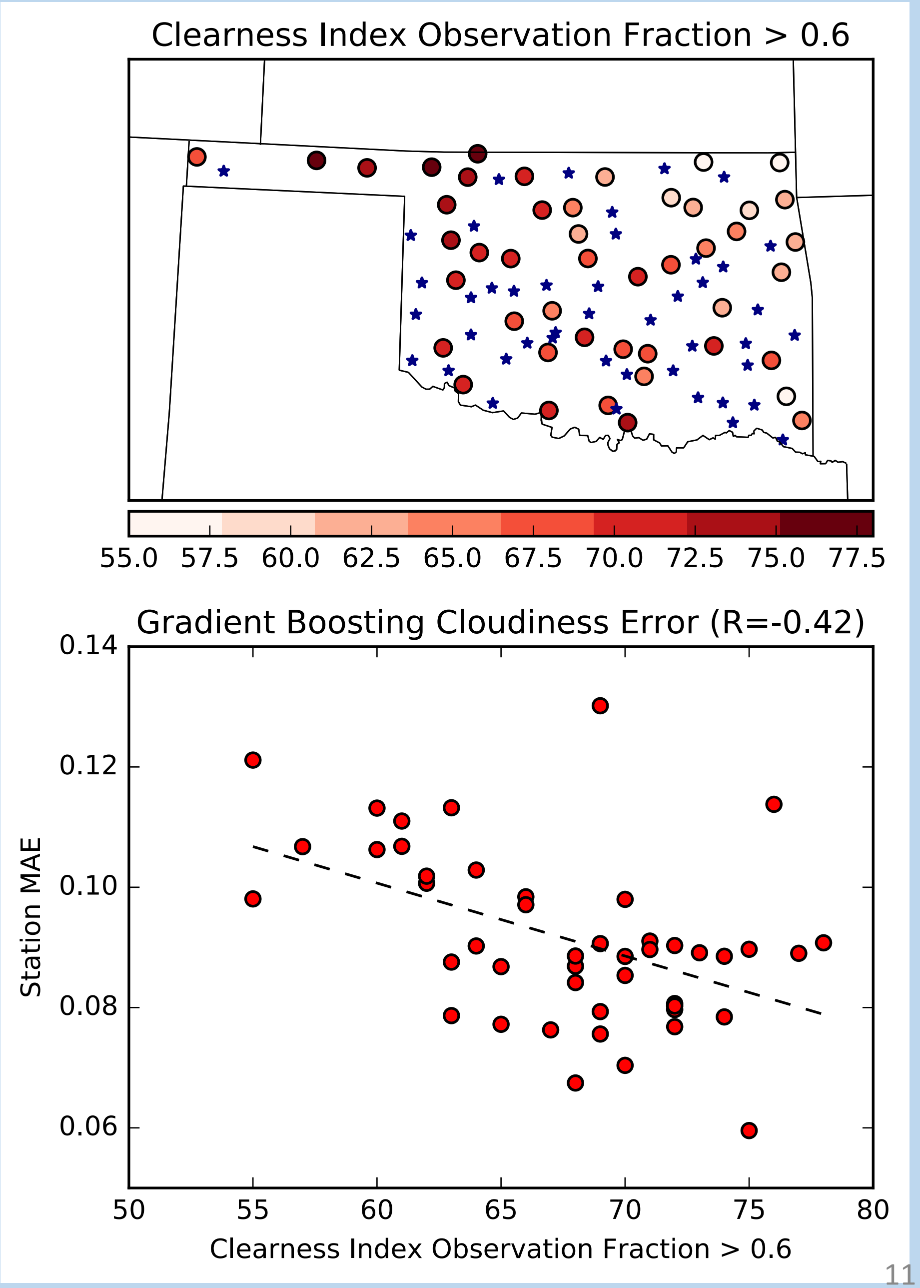


GFS Forecast Distributions



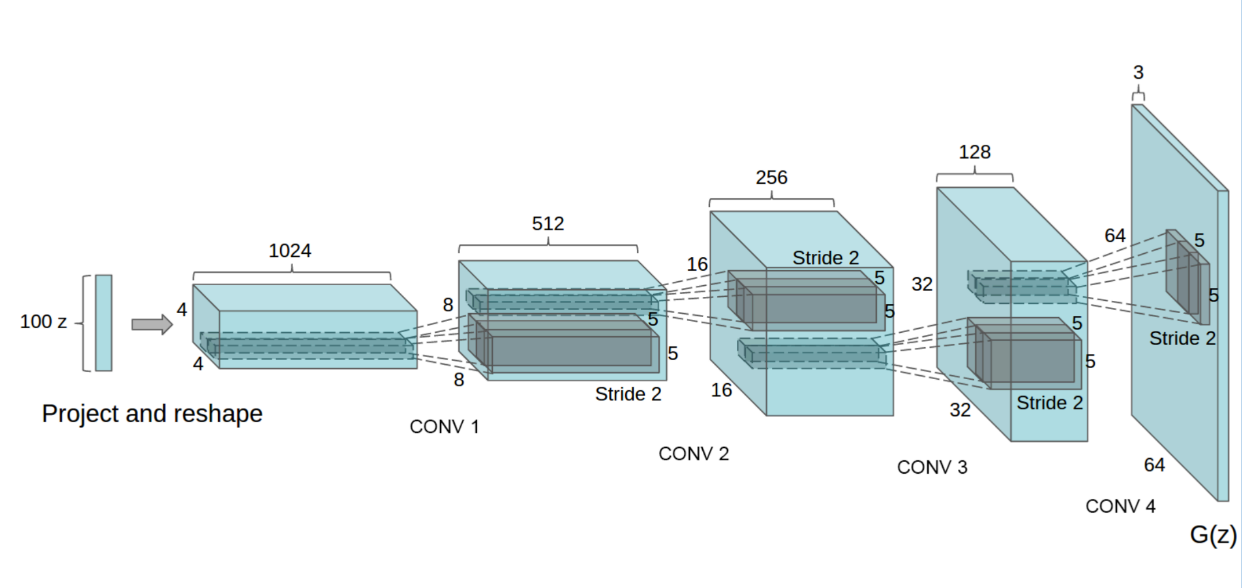
GFS Solar Station Errors





Next Steps: Deep Learning

- Investigating the use of deep learning models for weather feature and regime identification
- Goal: Train models to recognize multiscale features in NWP output
- Potential application for improved solar irradiance mean and variability forecasts based on weather regime
- Many other weather and climate applications



Deep Convolutional Generative Adversarial Network architecture from Radford et al. (2016)

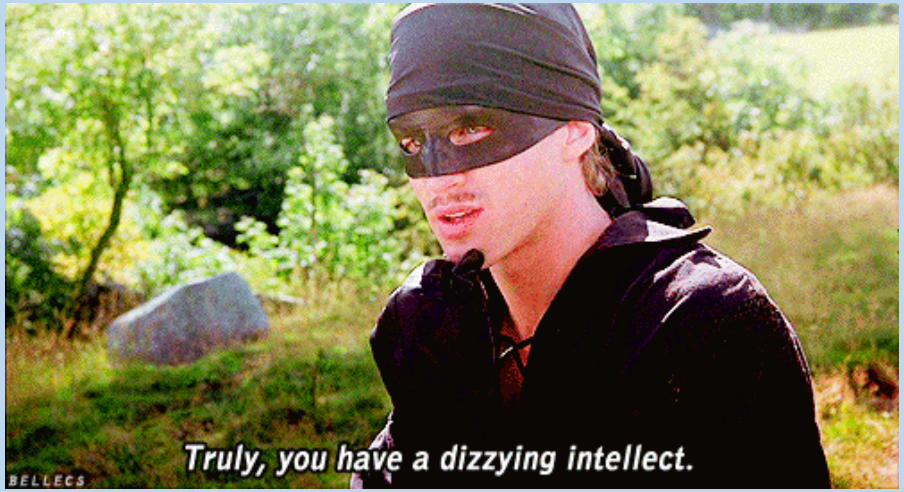
Generative Adversarial Networks

Unsupervised method of learning complex feature representations from data
Requires 2 deep neural networks

Discriminator: determines which samples are from the training set and which are not

Both networks have a "battle of wits" either to the death or until the discriminator is fooled often enough

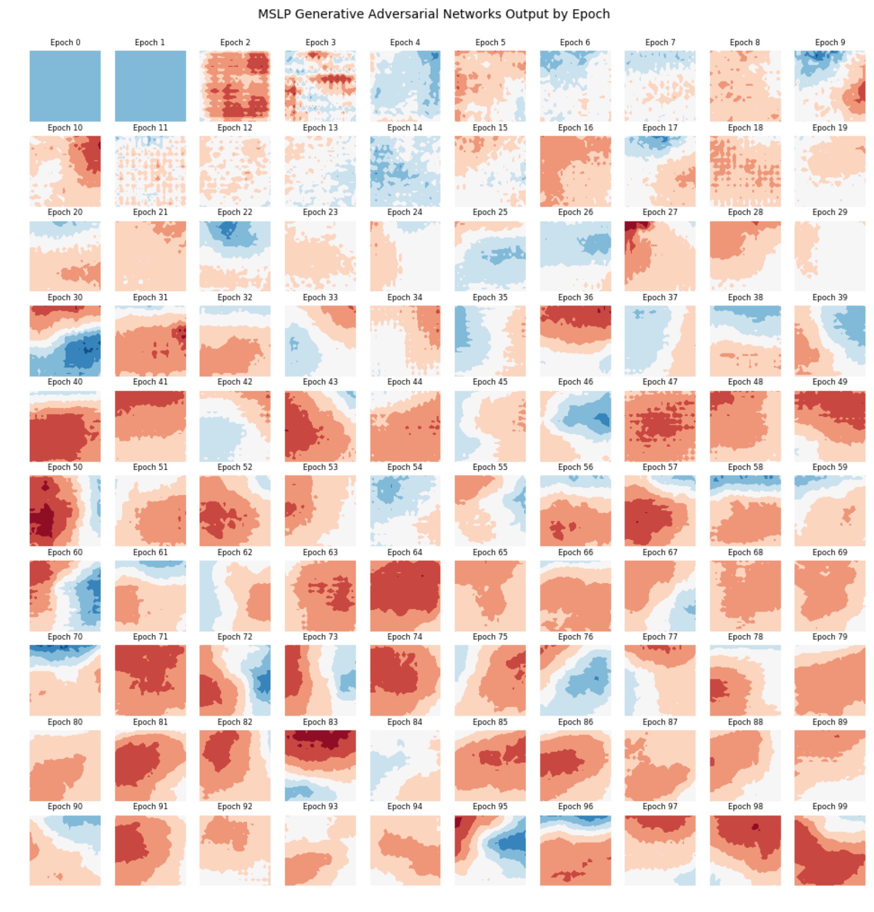
Generator: Creates synthetic examples similar to training data to fool discriminator

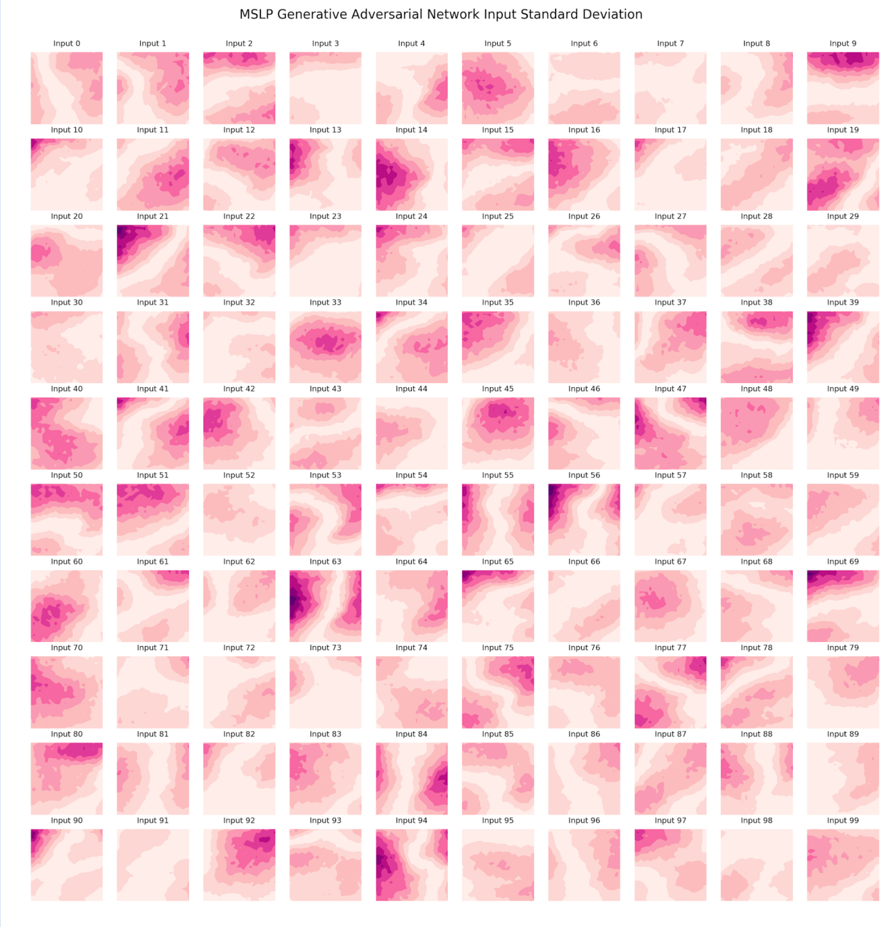


Advantages

- Unsupervised pre-training: learn features without needing a large labeled dataset
- Dimensionality reduction: reduce image to smaller vector
- Learns sharper, more detailed features than autoencoder models
- Do not need to specify a complex loss function

Preliminary Results: Mean Sea Level Pressure





- Trained on 4096 GEFS pressure forecasts
- Produces ”realistic” pressure fields after 100 epochs of training

- Generator uses 100-value vector as input
- Each input adjust different parts of field

Summary

- Developed gridded statistical forecasting system for solar irradiance
- Evaluated different machine learning models and configurations on their ability to predict irradiance at multiple sites
- Gradient Boosting consistently showed lowest errors
- All machine learning models underestimated cloud cover frequency
- ML models had lower errors at sites with fewer clouds
- Generative Adversarial Networks show potential for extracting information from weather data

Acknowledgements

- Rich Loft
- Tom Hamill
- The Oklahoma Mesonet

Contact Me

- Email: dgagne@ucar.edu
- Twitter: @DJGagneDos
- Github: github.com/djgagne