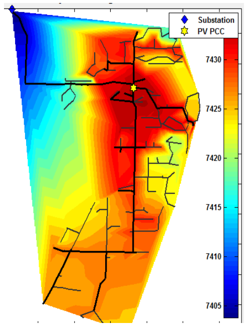
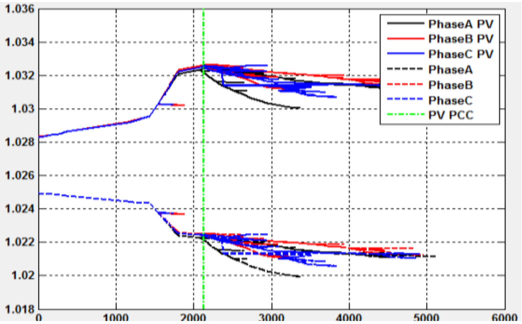
SAND2018-1607C





Decision Tree Ensemble Machine Learning for Rapid QSTS Simulations

Logan Blakely, Matthew J. Reno, Robert J. Broderick

Sandia National Laboratories

Sandia National Laboratories

This research was supported by the DOE SunShot Initiative, under agreement 30691. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Introduction

- Increasing penetration of distributed energy resources (DER) and the deployment of smart grid devices require new methods of distribution system analysis

- Quasi-static time-series (QSTS) analysis is necessary to capture the high-speed interactions of solar variability and other time-dependent aspects of the system

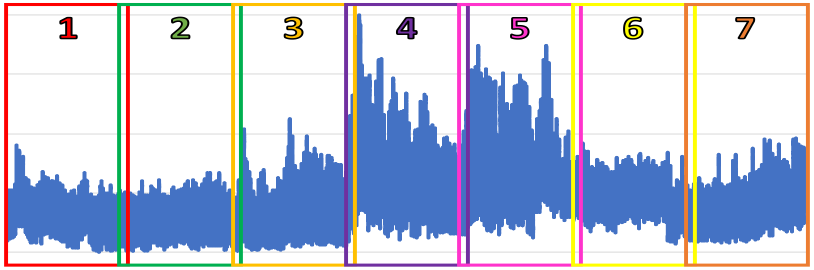
- Current QSTS algorithms are computationally intensive (10-120 hours per simulation). This dataset has 1-second resolution load and photovoltaic output (PV) profiles based on the 13-node IEEE test feeder.

- The goal of this project is to increase the speed of QSTS simulations

Methodology

Overview

- Goal - Reduce the computational time of QSTS simulations:
  - Separate the year-long data set into time periods. We are using periods of two-hours.
  - Identify the key time periods throughout the year to run with the QSTS simulation.
  - Use those chosen periods as input to a decision tree ensemble algorithm and predict the remainder of the time periods to accurately reproduce the full baseline QSTS analysis.



Methodology

Overview

Separate year-long data into periods

↓

Calculate features for all periods

↓

Intelligently sample a percentage of the year & run QSTS

↓

Train decision tree ensemble

↓

Use the ensemble to predict the remainder of the year

- Features – Statistics calculated from the 1-second resolution Load and PV data, used as input to the decision tree ensembles
  - 17 features - Load Mean, PV Mean, Load Standard Deviation, . . .
- Evaluation Metric - Voltage Regulator Tap Changes
  - With the increasing prevalence of highly variable DER, an increasing number of tap changes can increase Operations & Maintenance costs.
- Evaluation Accuracy – Compare predicted number of yearly tap changes to the actual number from a full-length QSTS simulation
  - A 10% error threshold in the year-long prediction has been deemed acceptable by our industry partners

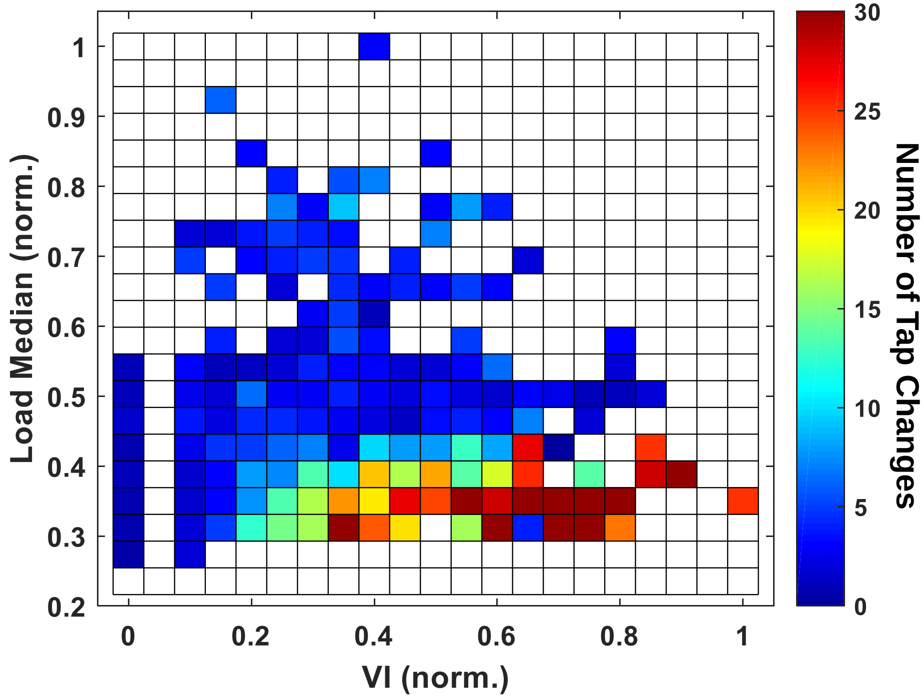
Methodology

Intelligent Sampling

- Randomly sampling time periods from the year to perform QSTS simulations can occasionally have a significant bias because only specific types of days (e.g. only clear sky days, or only winter days) are sampled. Random sampling requires >85% of the year.

- The objective of the intelligent sample selection is to select which periods are the most effective to simulate with QSTS to estimate the yearly impacts

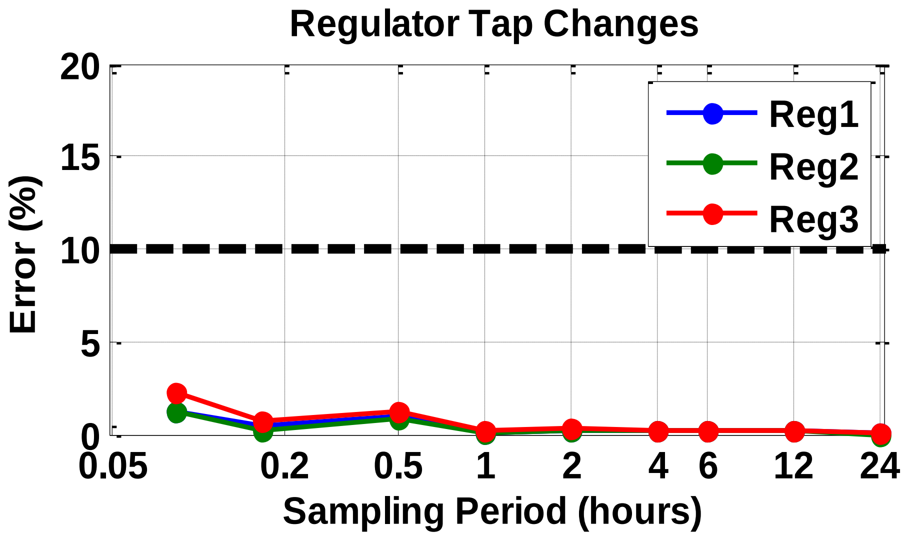
- Stratified sampling ensures that QSTS is run for at least one sample for each type of time period



Methodology

Error Introduced by Running the QSTS Simulations for Individual Periods

- In QSTS each power flow is dependent on the simulation before it. In sampling parts of the year to run individually, there is no information about what happened previously. This introduces some error compared to running the full QSTS simulation sequentially.



Methodology

Decision Tree Ensemble Machine Learning

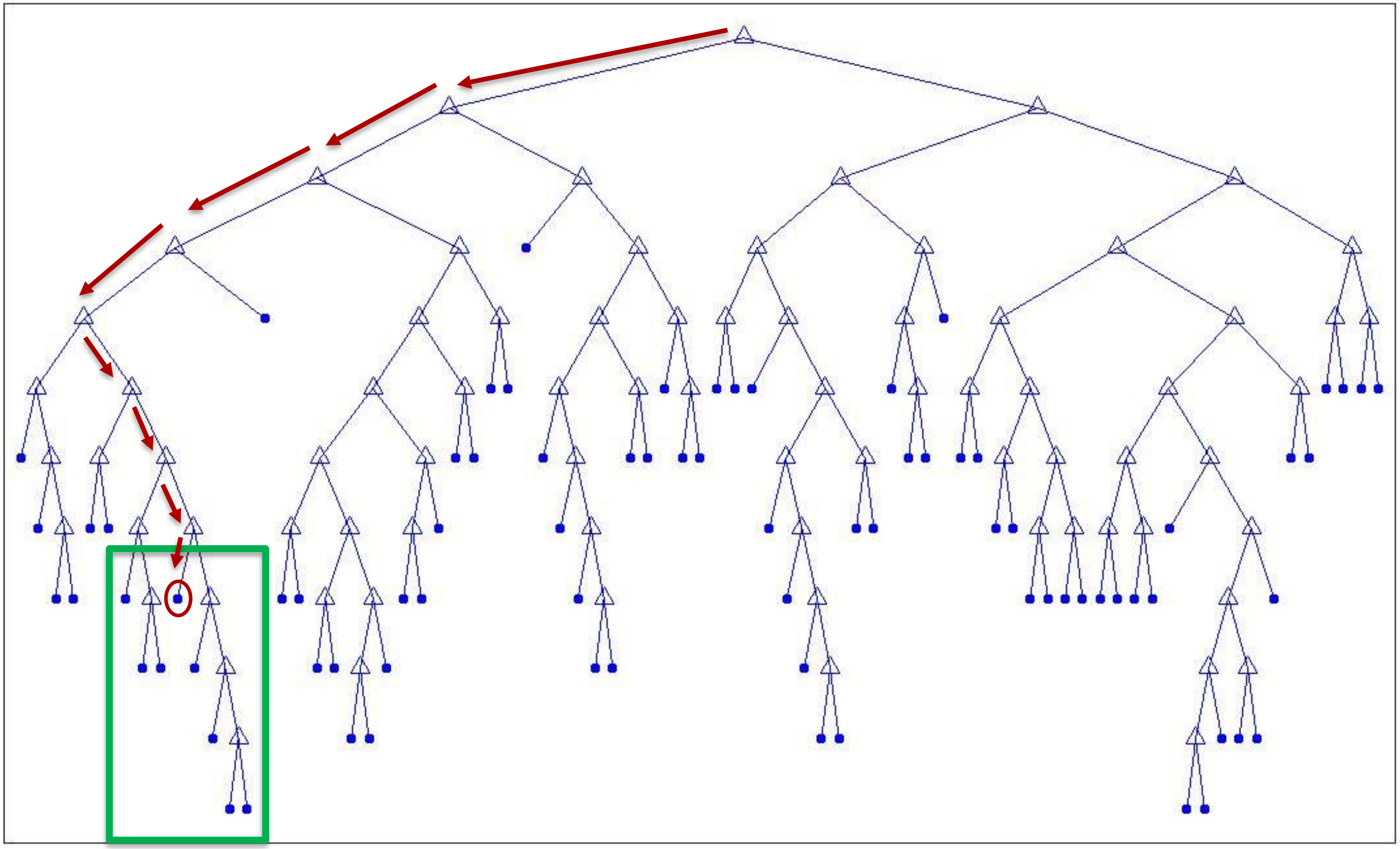
- We investigated two methods – random forests and boosted decision tree ensembles

- Decision tree ensembles have been shown to be competitive among other machine learning algorithms while remaining relatively fast

  - Two main types of decision tree ensemble methods – bagging ensembles (random forest) and boosting ensembles
  - Both use an ensemble of Classification And Regression Trees (CART)
  - Ensembles provide superior results compared to individual decision trees

Random Forest

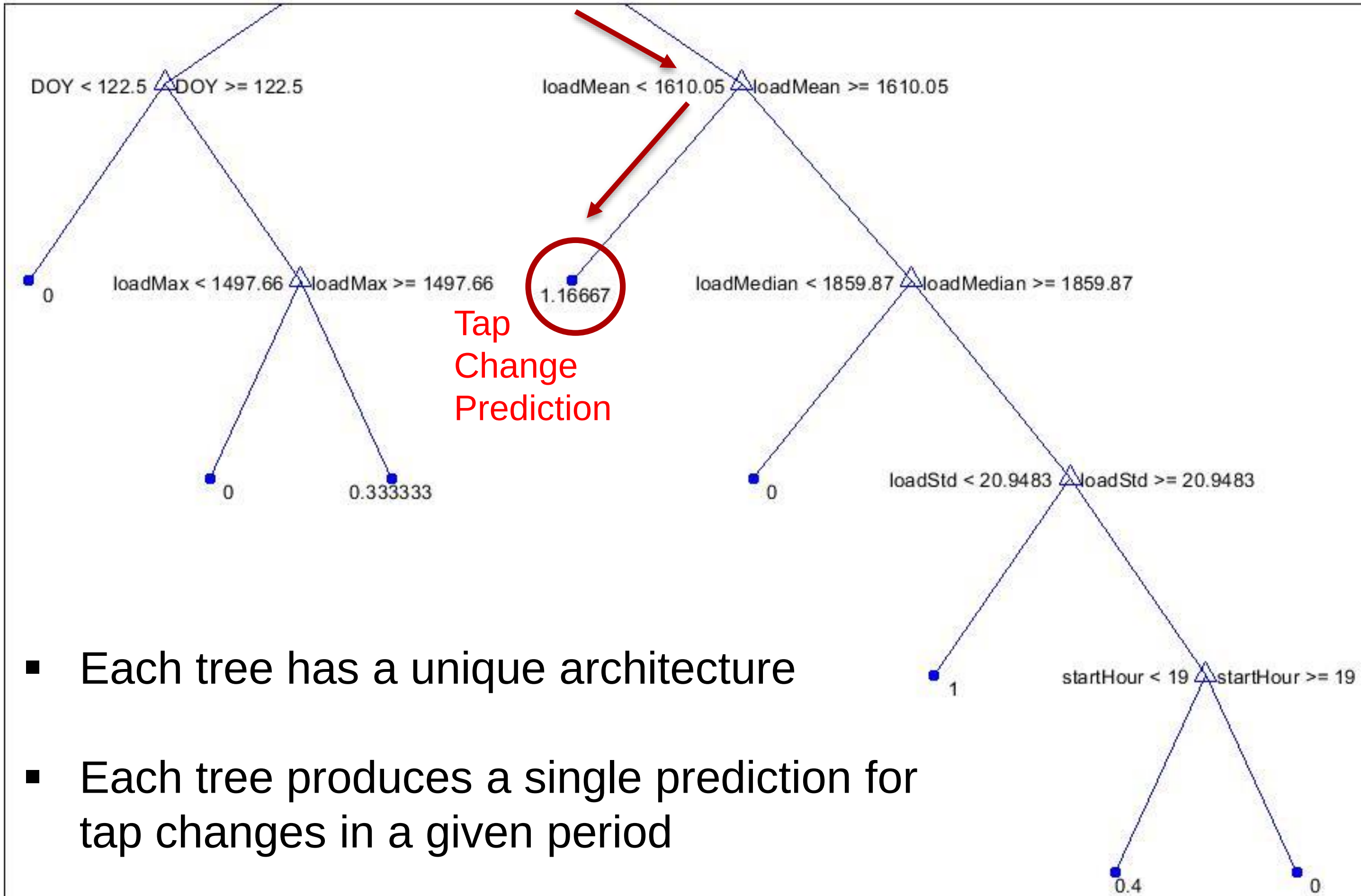
Individual Decision Tree



The “Forest” is a user-selected number of **independent** decision trees

Random Forest

Individual Decision Tree



- Each tree has a unique architecture
- Each tree produces a single prediction for tap changes in a given period

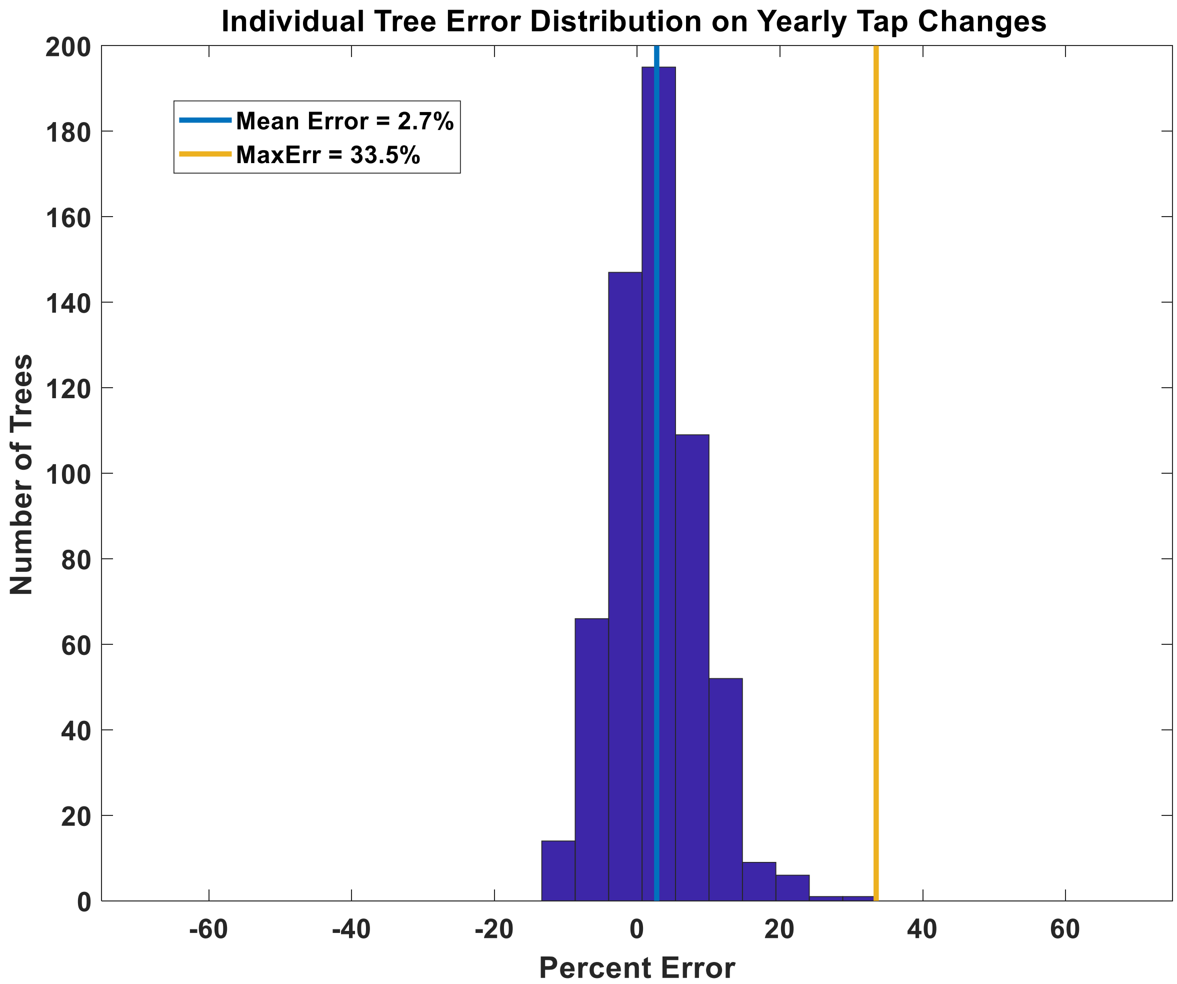
Random Forest

Individual Decision Tree

- Each tree attempts to make the most accurate **general** prediction possible

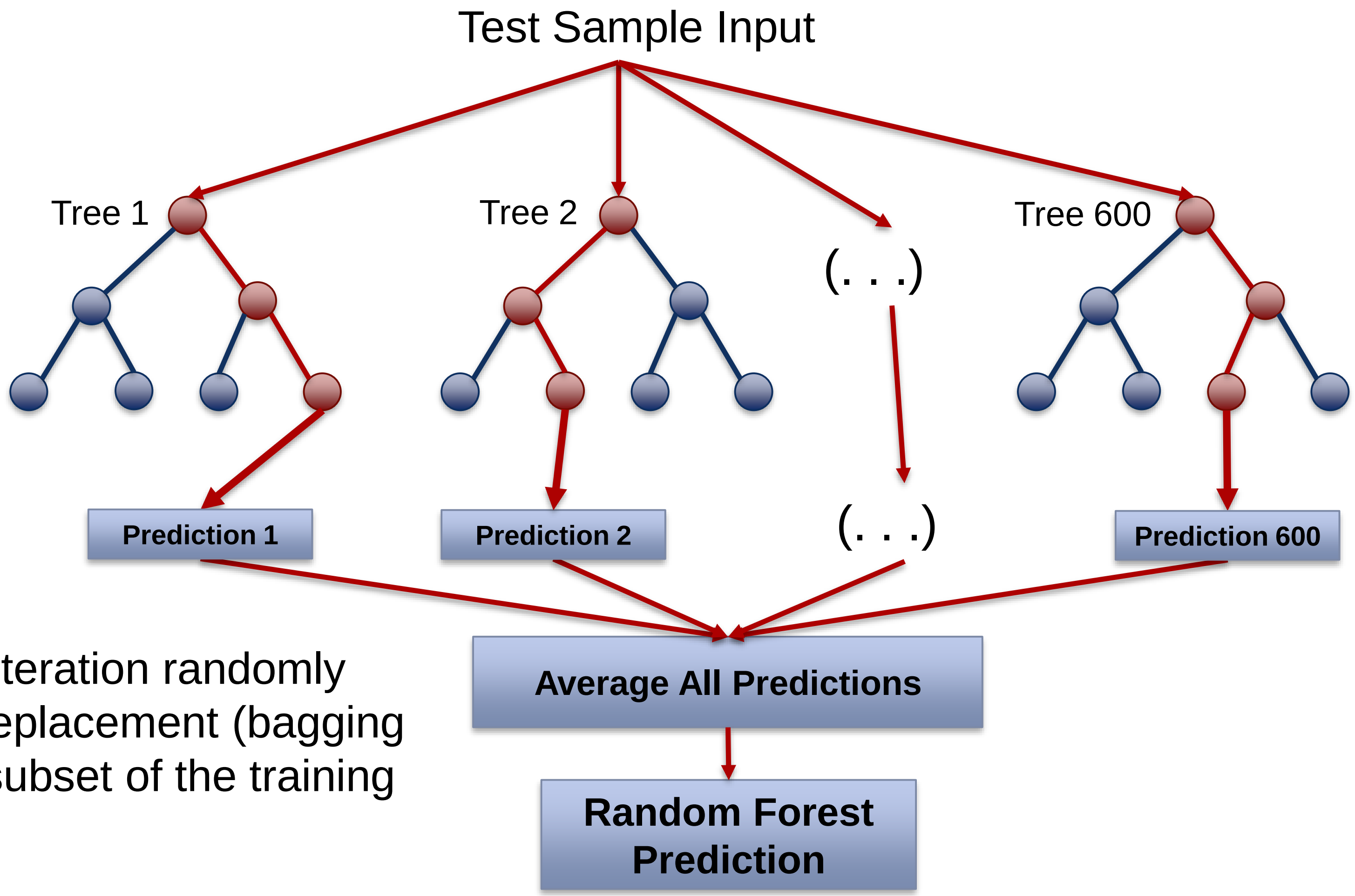
- The yearlong prediction is the sum of the training samples (QSTS simulations) and the sum of the predictions for each remaining period.

- The Mean Error of 2.7% corresponds to the error of the Random Forest prediction



Random Forest

Ensemble

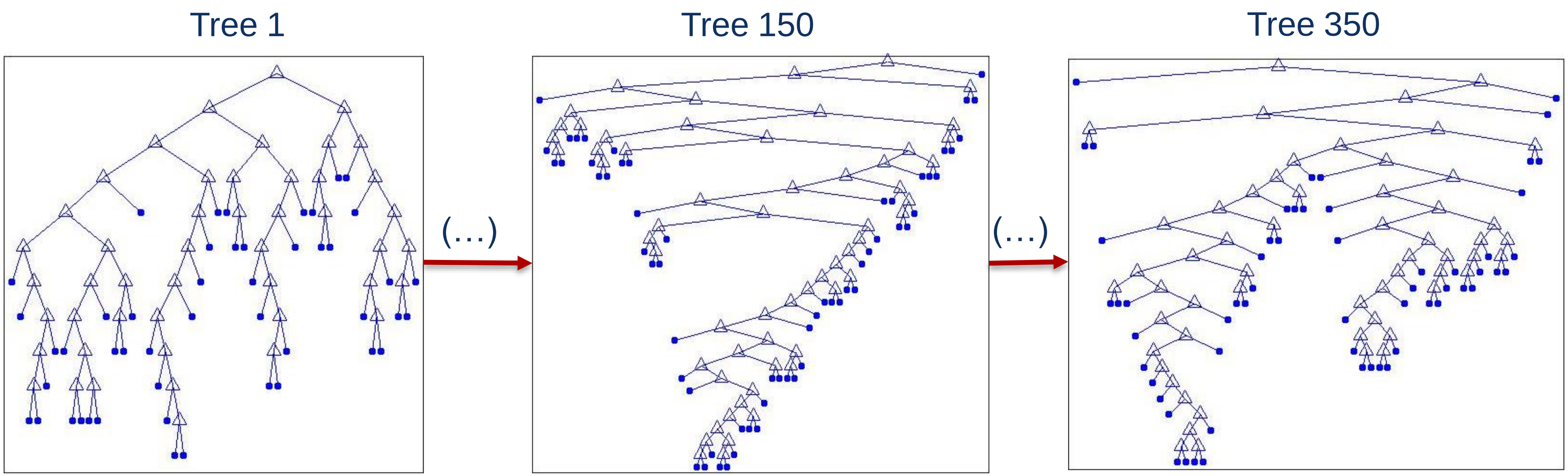


- Each training iteration randomly selects, with replacement (bagging approach), a subset of the training data to use
- Individual tree predictions are averaged for a final, ensemble prediction from the 600 trees

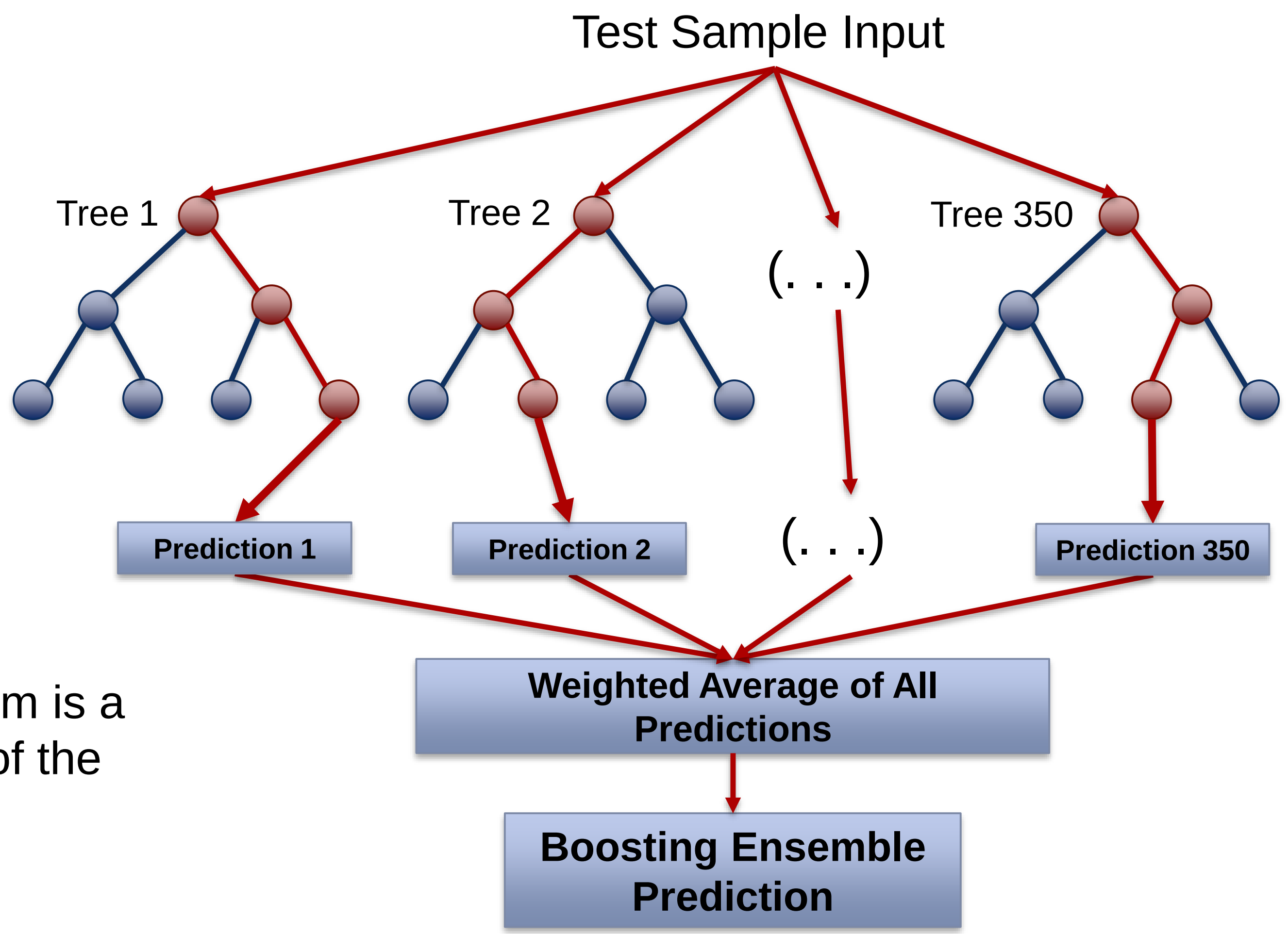
Boosting

Individual Trees

- Create trees **sequentially** and weight each tree according to its training error.
- Unlike Random Forest, the individual trees in a boosting ensemble perform poorly in isolation
- 'Incorrect' samples are given higher importance in further iterations
- Each subsequent tree is attempting to better predict a subset of the training data that was poorly predicted by previous trees



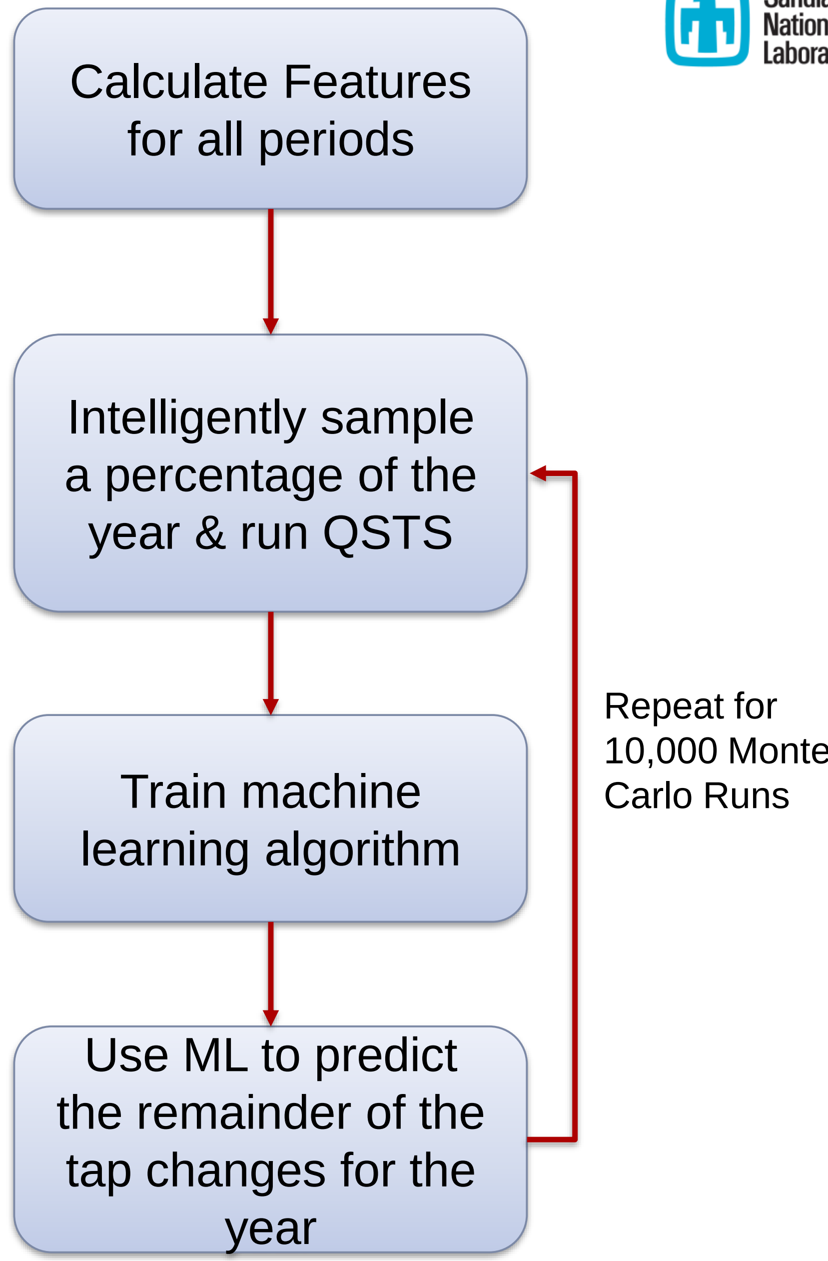
Boosting
Ensemble



- The boosting algorithm is a regression variation of the Adaboost algorithm.
  - The final result is the **<u>weighted</u>** average of all 350 models.

Results
Simulation Validation

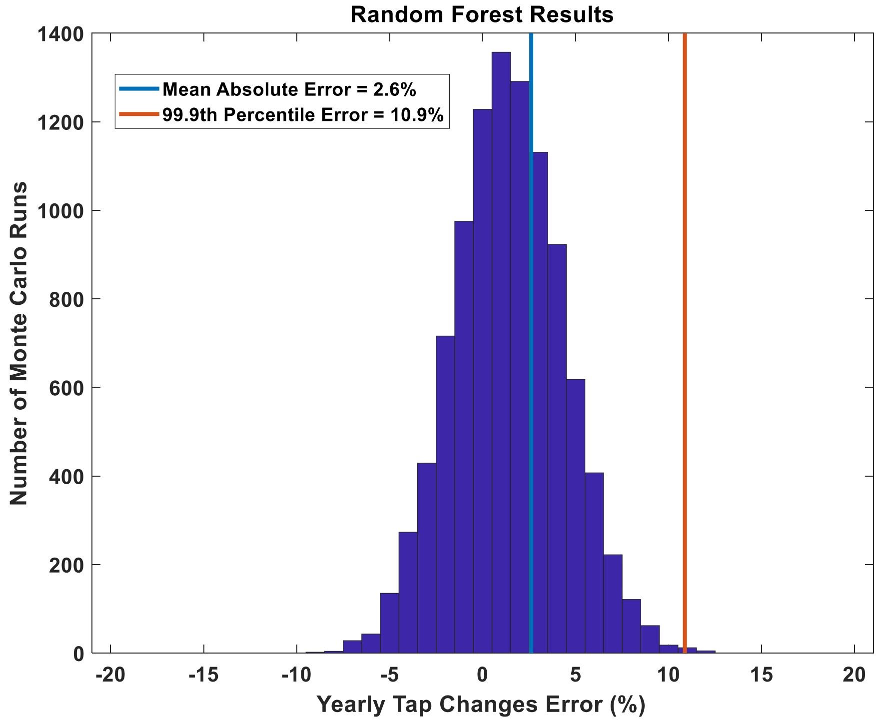
- Intelligent Sampling introduces randomness in the choice of training data
- Machine learning introduces randomness in the choice of training subsets and features to split on
- 10,000 Monte Carlo runs were done to validate the results for each method and the 99.9th percentile errors are shown

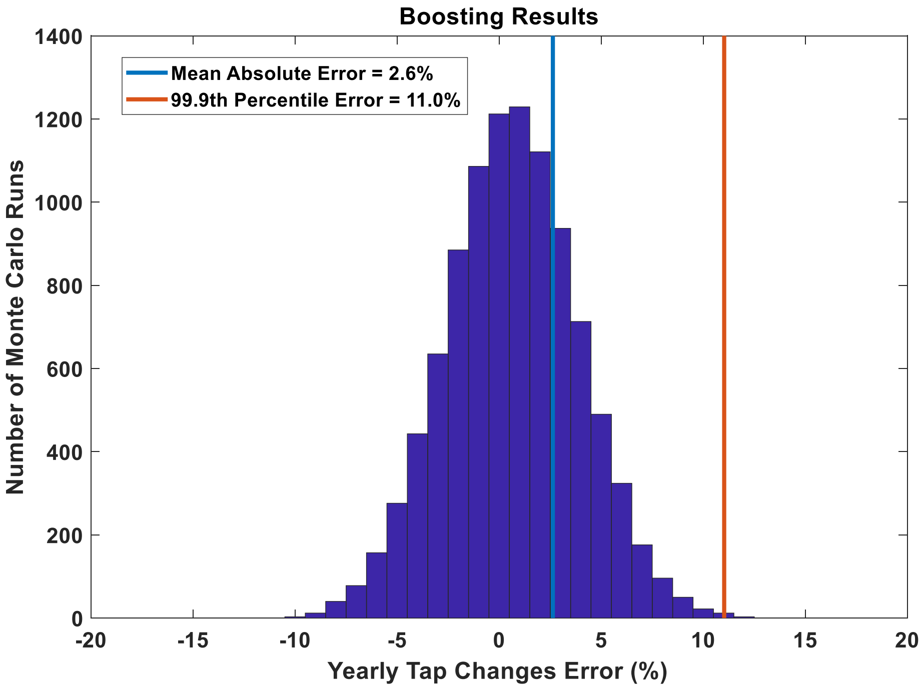


Results

Random Forest & Boosting Ensemble

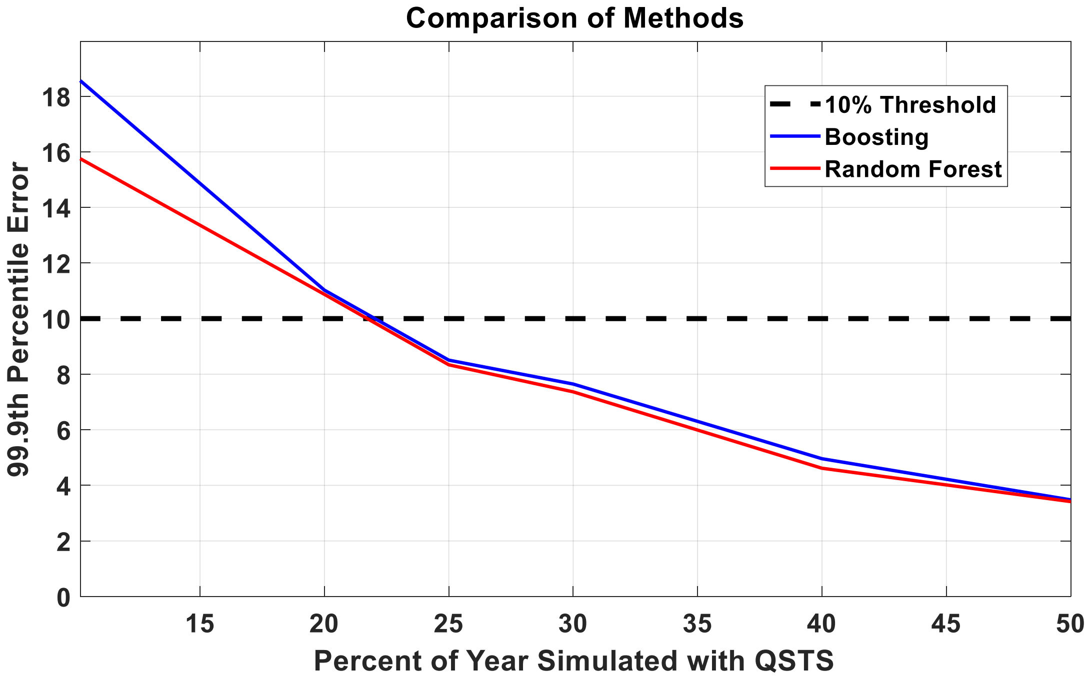
- Results shown use 20% of the year as training data (simulated with QSTS)





Results

Speed vs Accuracy



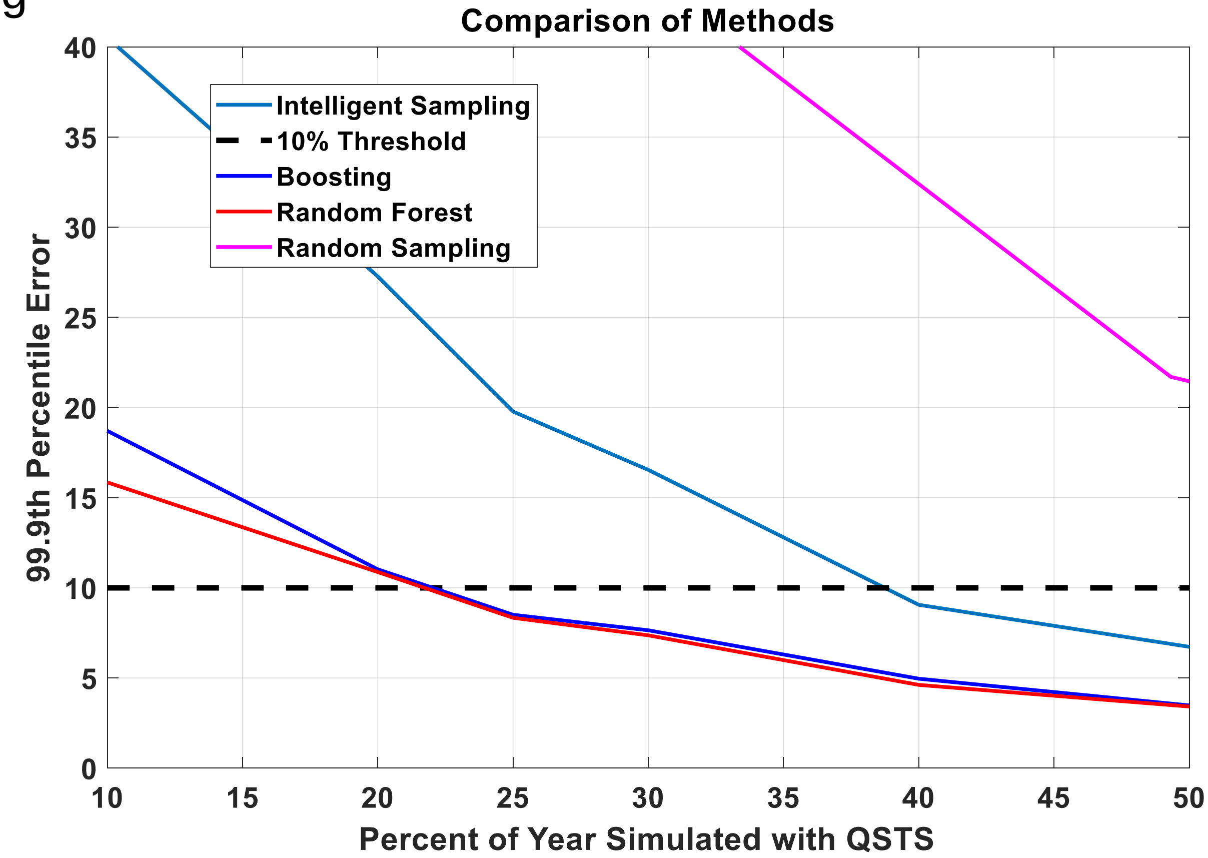
- ~25% of the year must be run using QSTS to guarantee that the 99.9th percentile error for the prediction of yearly tap changes will be under 10%

- By simulating more of the year with QSTS, the training data size and accuracy is increased, but additional computational time is required for running the longer QSTS simulation.

Results

Compared to Sampling Methods

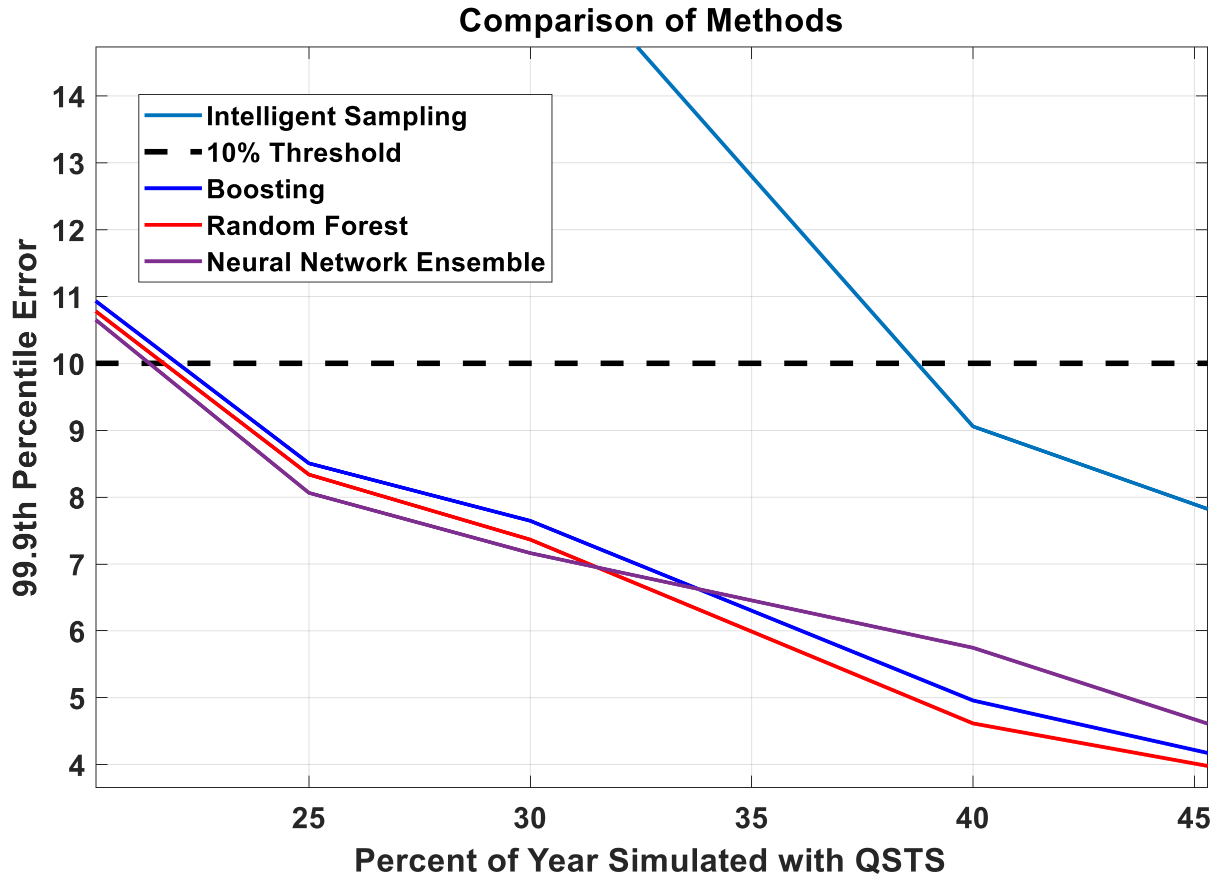
- Random forest and boosting ensembles can reduce computation time by ~4x compared to random sampling and ~1.5x compared to intelligent sampling alone while maintaining acceptable levels of error



Results

Compared to other Machine Learning Methods

- Similarity in the results for Random Forest and the Boosting Ensemble, as well as a Neural Network Ensemble, seem to suggest a limiting factor other than the machine learning algorithms.

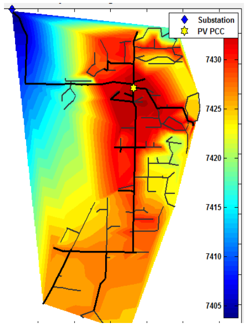


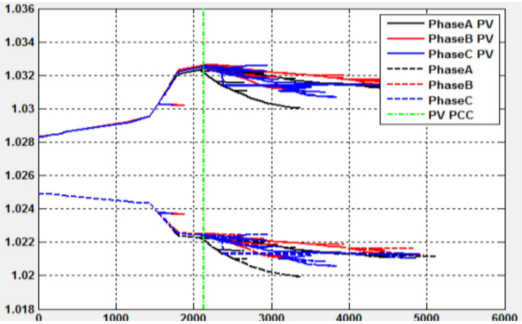
Conclusion

- QSTS simulations are necessary to accurately model the rapidly changing conditions introduced by solar power and other DER.

- Both Random Forest and a Boosted Decision Tree Ensemble reduce the computation time needed by ~4x relative to a full-year, sequential QSTS simulation

    - A representative portion of the year, ~25%, can be sampled and used as training data for a decision tree ensemble which can then predict the remainder of the year while staying under the 10% error threshold for yearly tap changes error.





Questions?

Logan Blakely

lblakel@sandia.gov

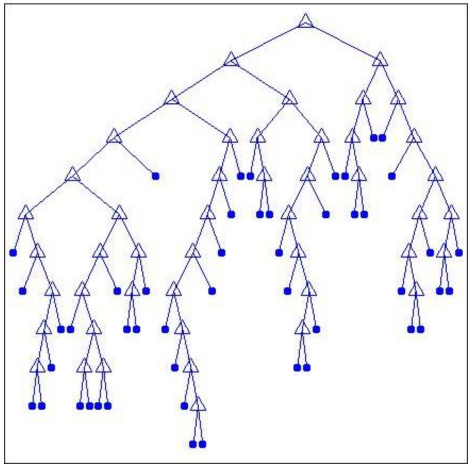
Sandia National Laboratories

This research was supported by the DOE SunShot Initiative, under agreement 30691. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

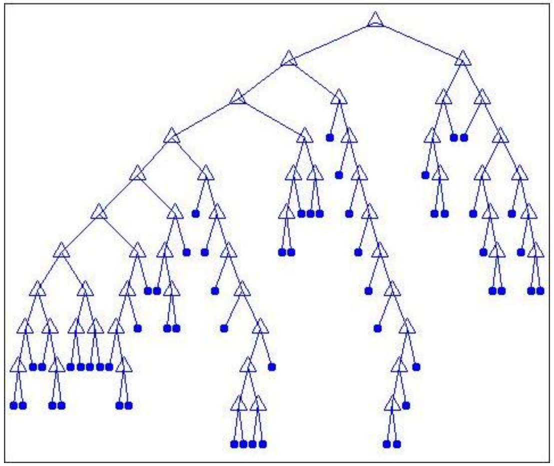
Supplemental Slides

Boosting

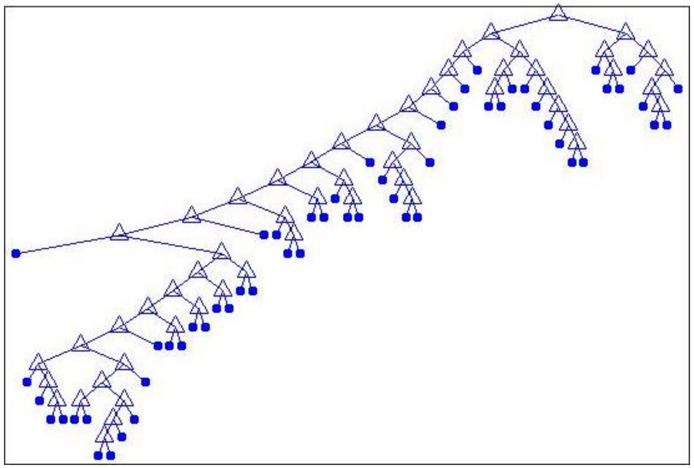
Individual Trees



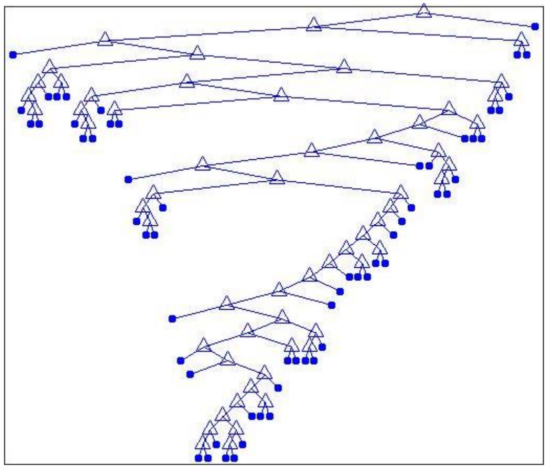
Tree 1



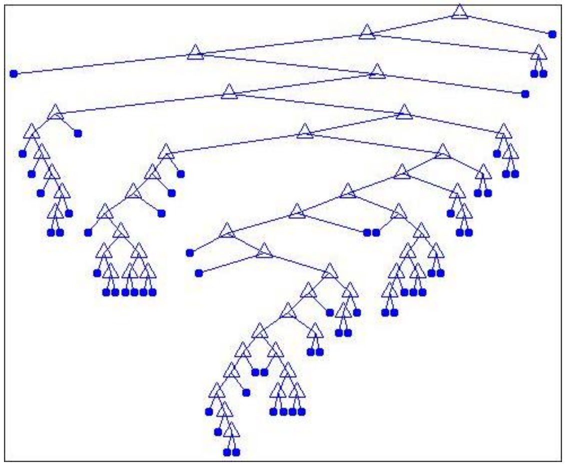
Tree 25



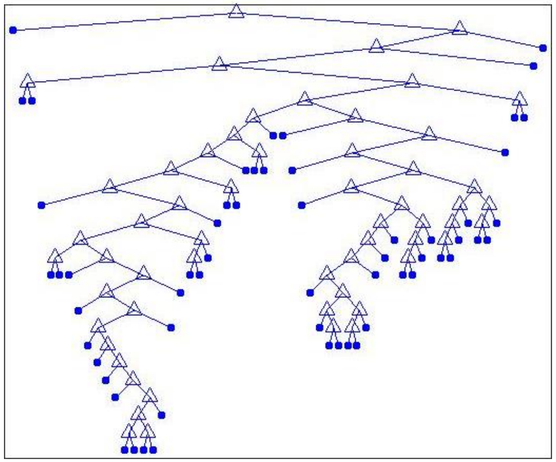
Tree 50



Tree 150



Tree 250



Tree 350

Boosting & Random Forest

Parameters

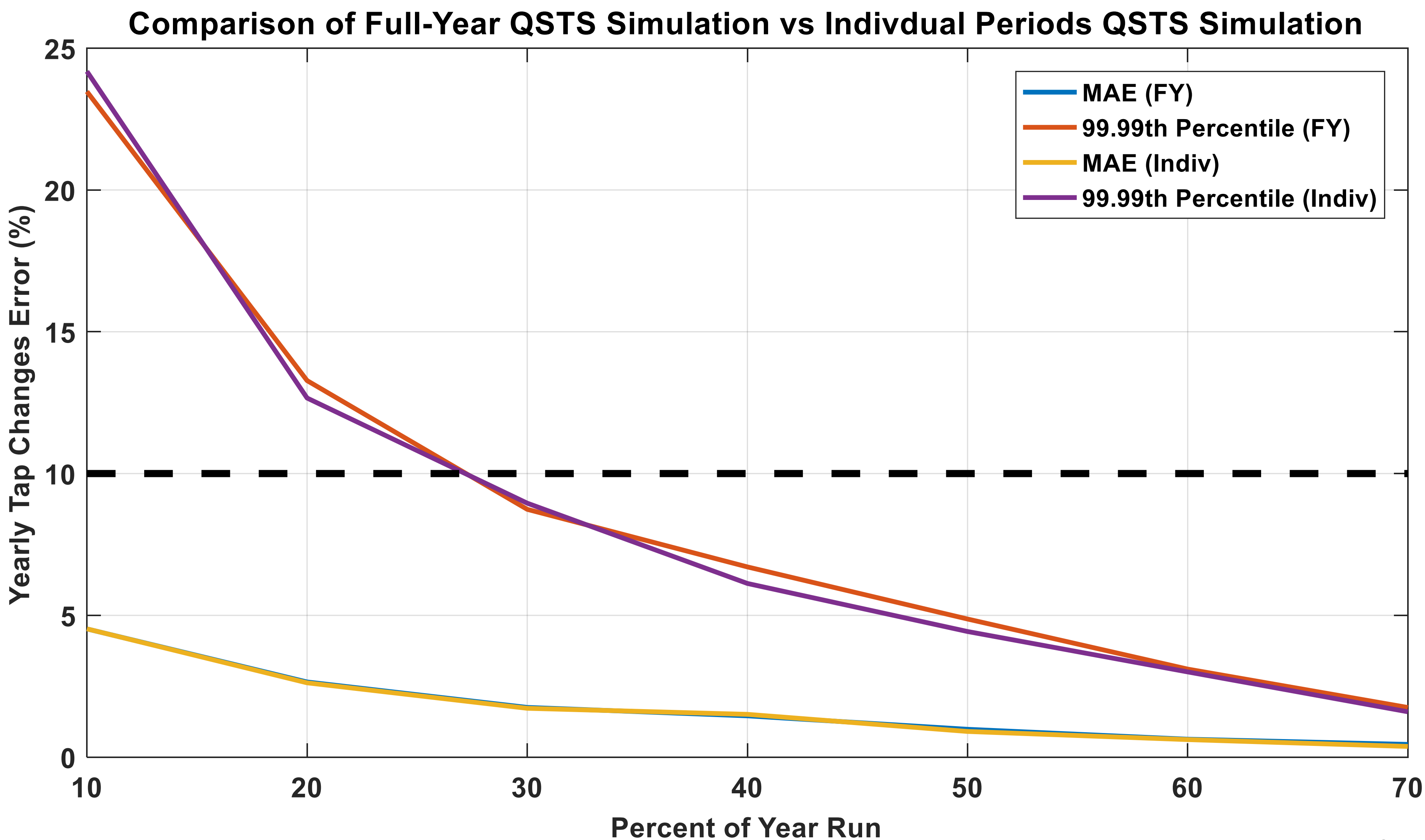
Random Forest

- Features to sample – 5 (1/3 of the total number)
- Minimum samples per parent node – 10
- Split Criterion – Mean Squared Error
- Sampling percentage – 100%

Boosting

- Features to sample - 11
- Minimum Leaf Size – 6 samples
- Split Criterion – Mean Squared Error
- Learning Rate – 0.08 (shrinkage)

- These parameters were obtained using a mix of optimizing algorithms, hand-tuning and parameter sweeps.
- Full Feature List: load max, pv max, load min, pv min, load mean, pv mean, load range, pv range, load length, pv length, load standard deviation, pv standard deviation, starting hour, day of the year, standard deviation of the two weeks prior



Comparison of Full-Year QSTS Simulation vs Indivdual Periods QSTS Simulation
Yearly Tap Changes Error (%)
Percent of Year Run
MAE (FY)
99.99th Percentile (FY)
MAE (Indiv)
99.99th Percentile (Indiv)
0
5
10
15
20
25
10
20
30
40
50
60
70