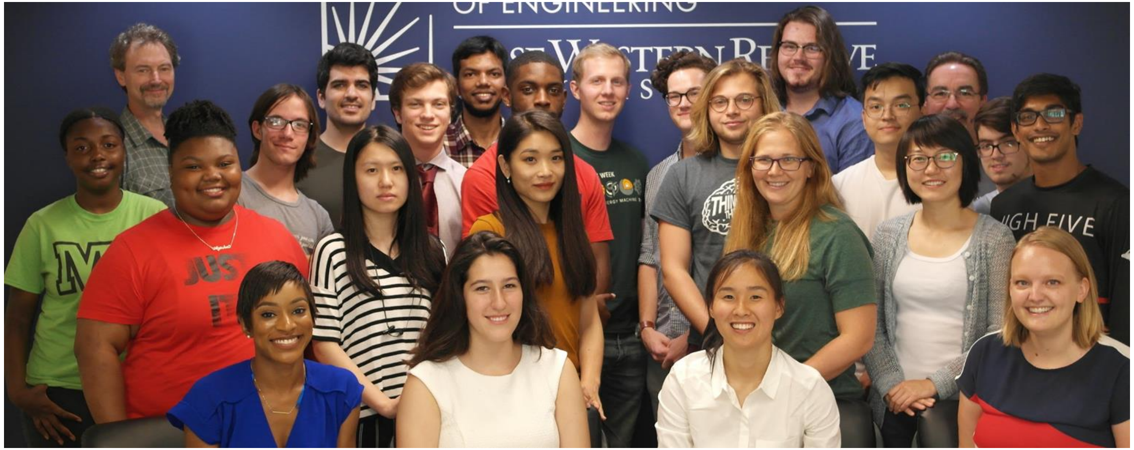
A Big Data Approach to Performance Loss Rate Determination of Commercial Photovoltaics

Laura S. Bruckman

Associate Research Professor
Case Western Reserve University
lsh41@case.edu

SDLE Research Center: Acknowledgements



CWRU Faculty

- Roger French, Laura Bruckman, Jeffrey Yarus, Jennifer Braid, Mehmet Koyutürk, Yinghui Wu, Alp S

Post-doctoral Research Associates

- Two openings: PV Degradation, Statistics & Data Science

Graduate Students

- Alan Curran, JiQi Liu, Arafath Nihar, Will Oltjen,Ben Pierce,Deepa Bhuvanagiri
- Raymond Wieser, Kunal Rath, Sameera Nalin Venkat, Tian Wang, Alex West, Steven Timothy

Undergraduates

- Tyler Burleyson, Carolina Whitaker, Minh Luu, Asher Baer, Daniel Arnholt, Hein Aung
- Medha Nayak, Cora Lutes, Alejandra Ramos,

High School:

SDLE Staff: Jonathan Steirer, Rich Tomazin

About the Data

1100 power plants distributed across North America with different features

1. Age
2. Suppliers
   a. Inverters
   b. Modules
3. Koppen-Geiger Climate Zones
   a. Major Type
   b. Temperature subtype
4. Module Placement
   a. Roof vs Ground

Each system measures the following components

- Time Stamp, power, temperature, windspeed, and irradiance
- Measurements made at 1 minute intervals

CRADLE v2.2 Architecture: Petabyte and Petaflop Computing

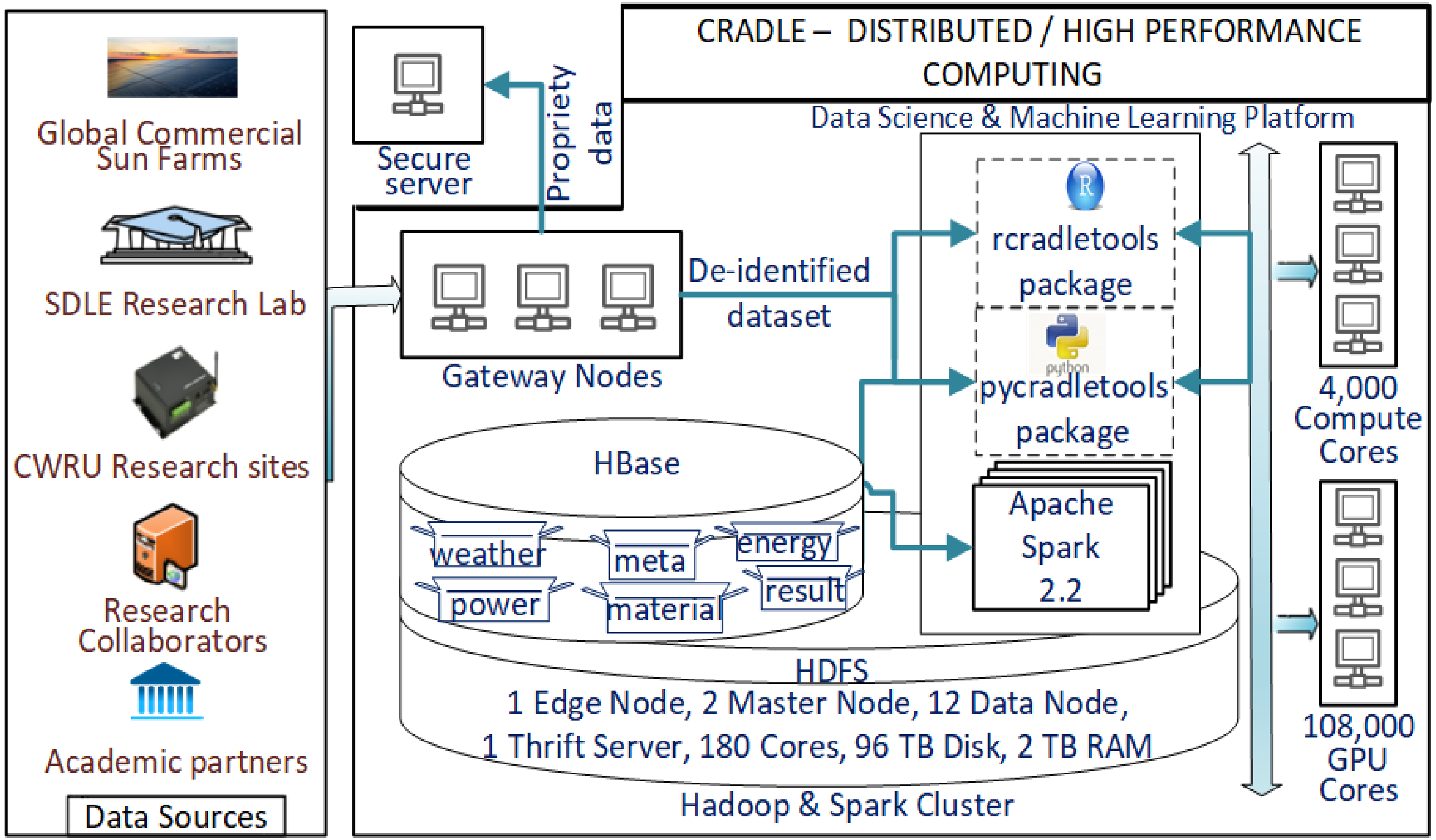
Hadoop Distributed Computing

With Hbase & Spark

Using R & Python For Analytics

In-place Analytics

Write-back All Results Into Hbase



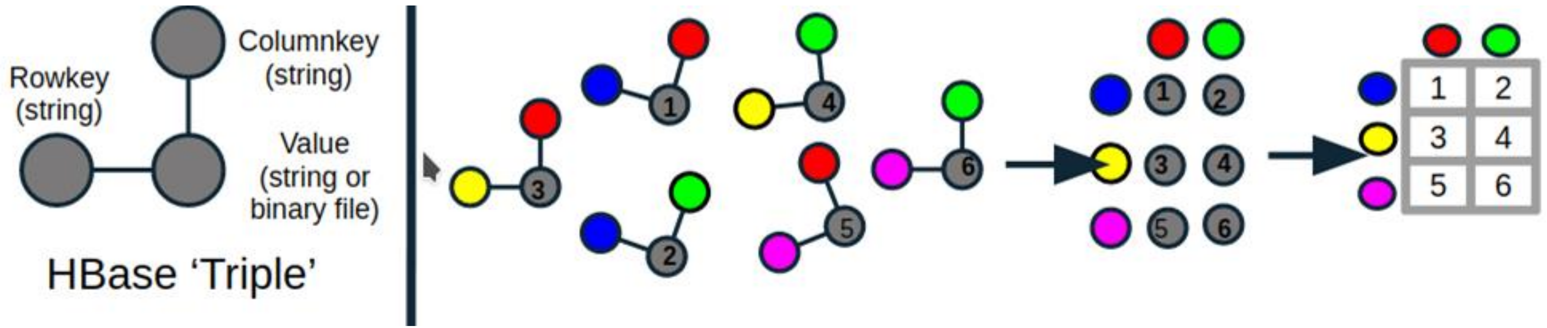
Data Handling

Hadoop/Hbase

Combine Lab data (Spectra, Images etc.)
With Time-series Data (PV Power Plant Data)

Petabyte Data Warehouse In A Petaflop HPC Environment

- Query Data
- Based on rowkey or columnkey
- All data related to PET
- Or All Images



FAIRification of Datasets and Models, Enables AI learning

Making Datasets & Models FAIR

- By “FAIRification”

Enables Models to find Data

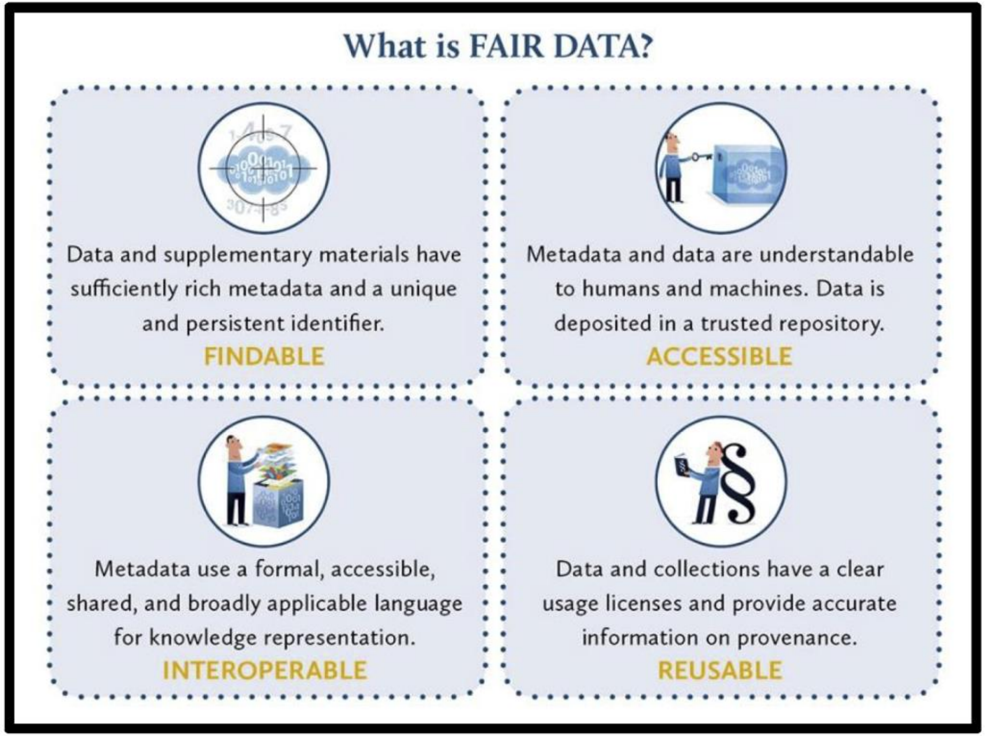
- And Data to find Models

So that they can advance

- Without human intervention

This is an aspect of the Semantic Web

- And Resource Description Framework
- Hbase triples are an example of RDF



We just received a DOE SETO AI award

- For st-GNN, that involves FAIRification

Enabling this in Hadoop/Hbase Environment

- **Can enable automation of analysis**

M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” Sci. Data, vol. 3, no. 1, pp. 1–9, Mar. 2016, doi: 10.1038/sdata.2016.18.
https://en.wikipedia.org/wiki/Resource_Description_Framework

Age

PV Systems

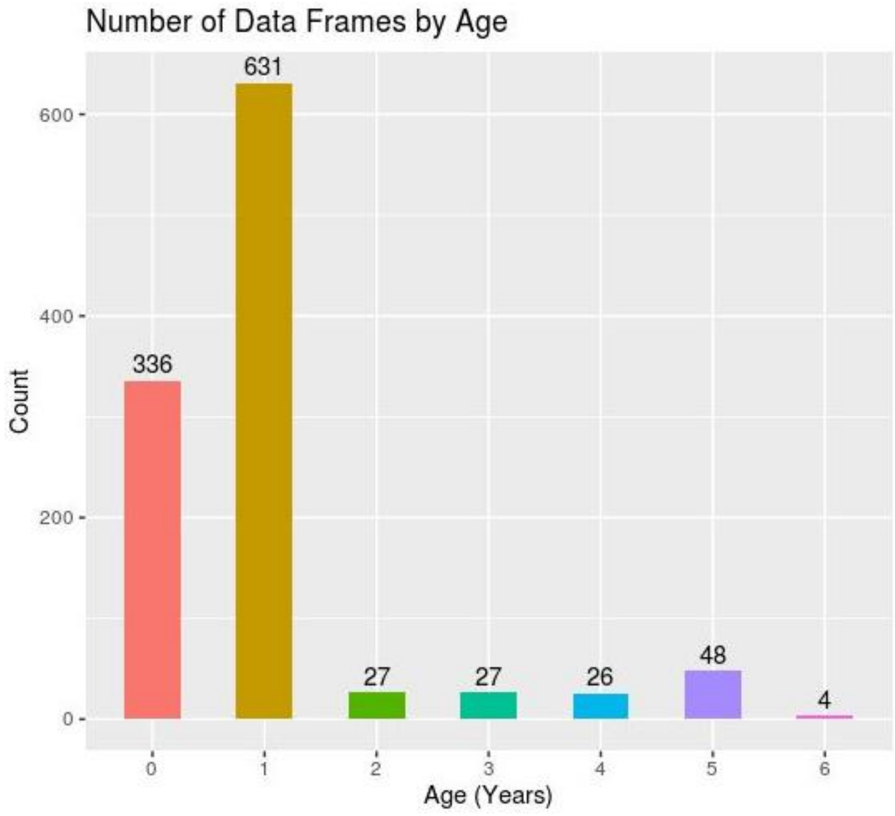
- Various patterns
- Due age

Profile of amazingly fast growth

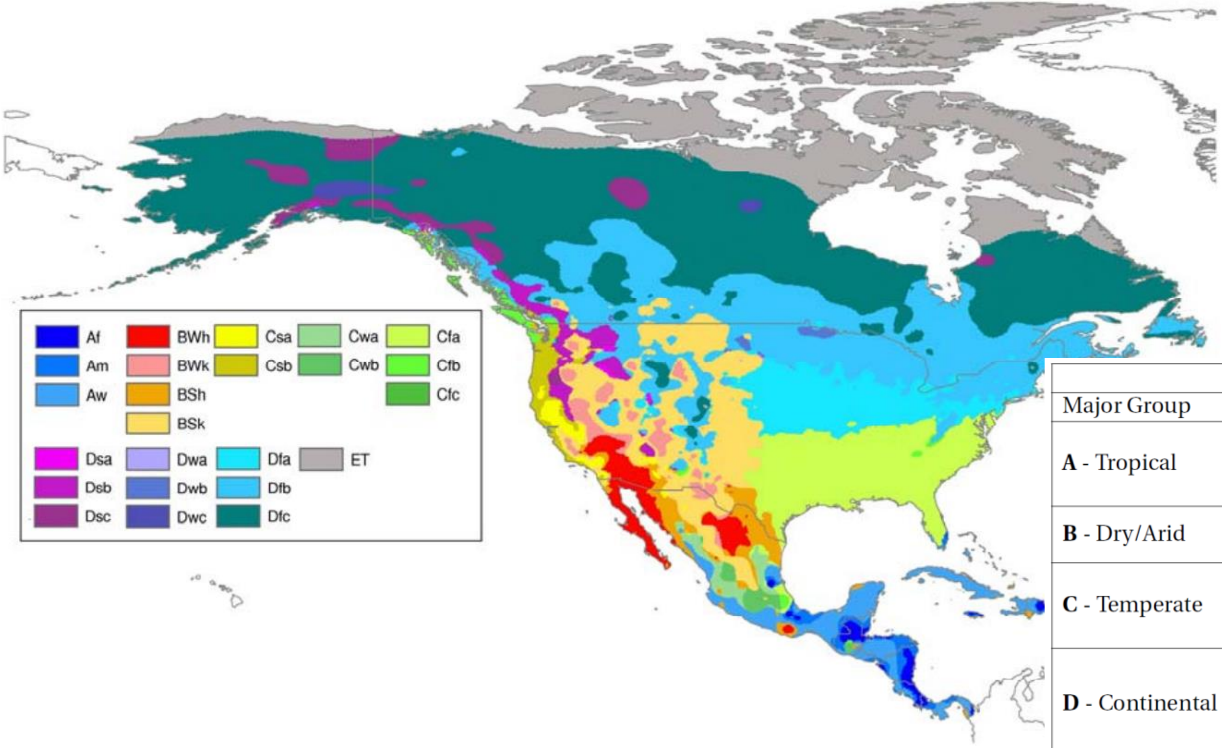
- In the US

12 inverters suppliers

24 module suppliers



Köppen-Geiger Climate Zones

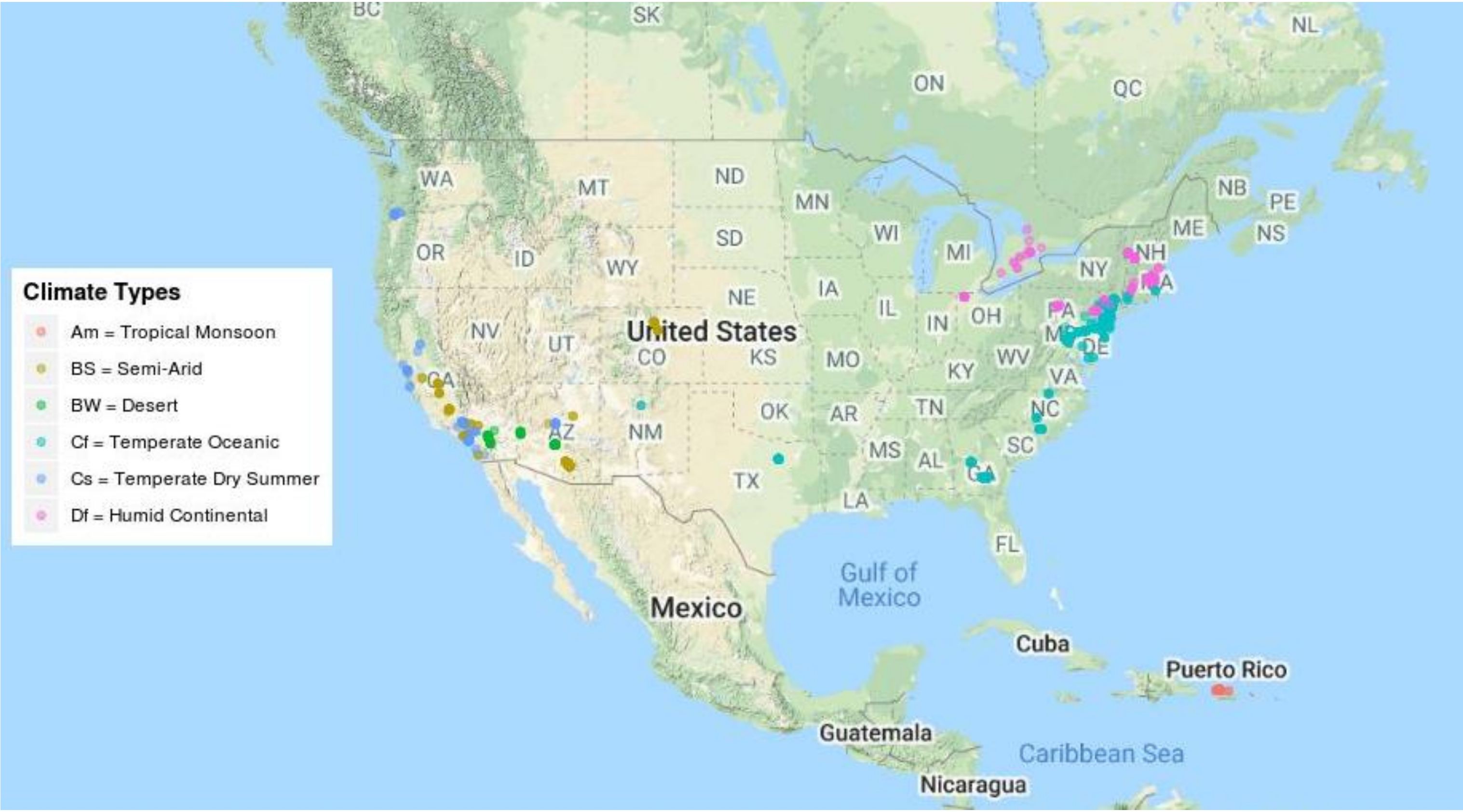


Koppen-Geiger Climate Zone Types			
Major Group	Subcategory	Temperature	Example
A - Tropical	**f** - Tropical Rainforest **m** - Tropical Monsoon **s** - Savannah		Af, Am, As
B - Dry/Arid	**W** - Desert **S** - Semi-Arid/Steppe	h - hot k - cold	BWh, BWk, BSh
C - Temperate	**f** - dry summer **w** - dry winter **s** - no dry season	a - hot summer b - warm summer c - cold summer	Cfa, Cwc, Csb
D - Continental	**f** - dry summer **w** - dry winter **s** - no dry season	a - hot summer b - warm summer c - cold summer d - very cold winter	Dfa, Dwc, Dsd
E - Polar	**T** - Tundra **F** - frost		ET, EF

Table 1.1. Summary of the combinations of the major group, subcategory, and temperature designations of Koppen-Geiger Climate Zones

Peel, M. C., Finlayson, B. L., and McMahon, T. A.: Updated world map of the Köppen-Geiger climate classification, Hydrol. Earth Syst. Sci., 11, 1633–1644, https://doi.org/10.5194/hess-11-1633-2007, 2007.

Climate Zones of Solar Plants in Our Data



Climate Types of Solar Farms in Our Data

Group A: Tropical

- m = Tropical Monsoon Climate

Group B: Dry

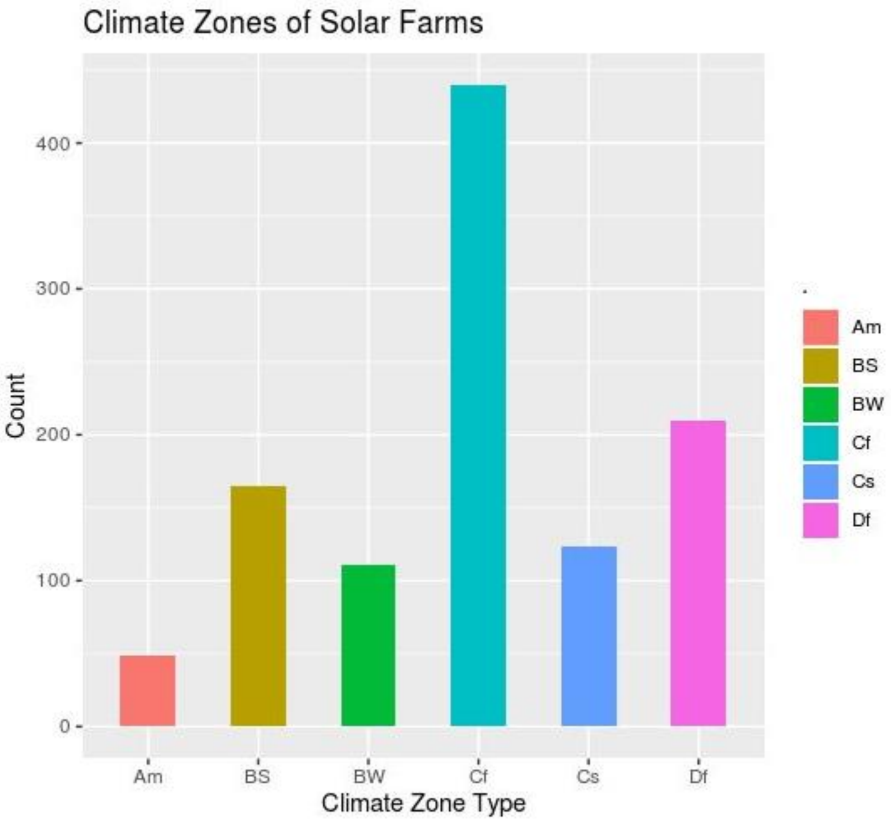
- S = semi-arid
- W = desert

Group C: Temperate Climates

- f = no dry season
- s = dry summer

Group D: Continental

- f = no dry season



Methods

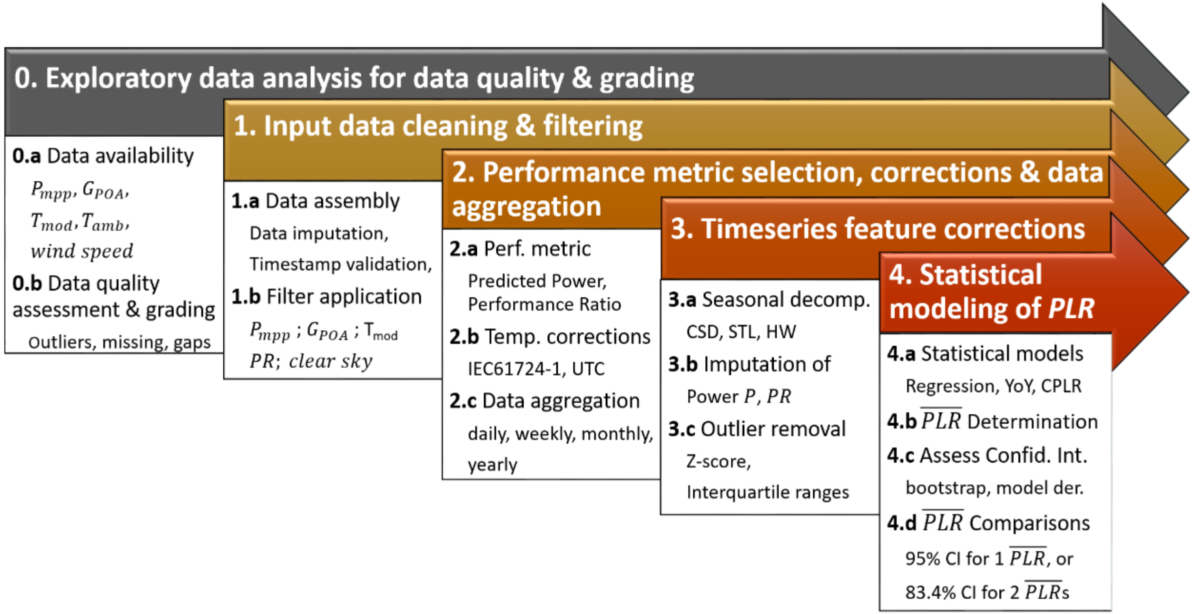
Performance Loss Rate Determination

Accurate determination of PV System's Performance Loss Rate (PLR)

- **Critical for assessing PV system operation, maintenance and production**

Four main steps in *PLR* determination

- **0. Data Quality assessment**
- **1. Cleaning & Filtering**
- **2. Metric Selection**
- **3. Feature Corrections**
- **4. Statistical Modeling**



Exploratory Data Analysis & Dataset Grading

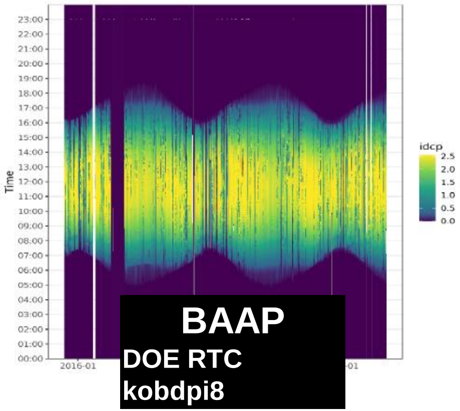
Performance of *PLR* algorithms, strong function of dataset "missingness"

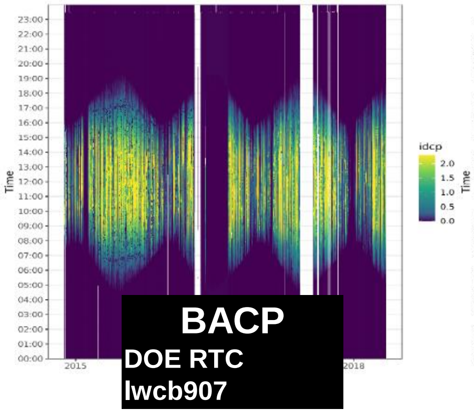
Missingness includes Outliers, Missing Datapoints, and Data Gaps

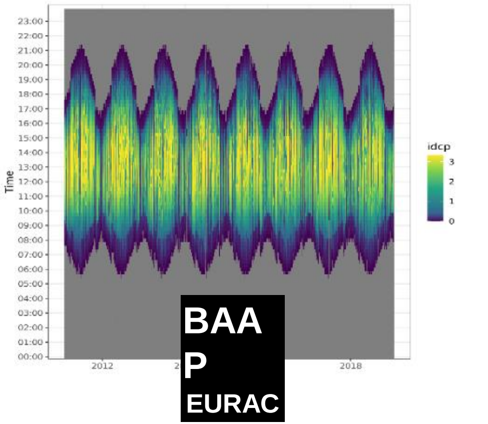
Dataset Grading

A A A P

Outliers Missing Gaps P/F







Outliers = Anomalies and Rapid Changes (can be Clouds)
Missing = 5 or less missing data points
Gaps = Missing data longer than 5 data points

Data Processing Pipeline

Retrieve Data from HBase → **Data Cleaning** → **Predictive Power Modeling** → **Cross-Sectional Analysis**

Retrieve Data from HBase
- Data Querying

Data Cleaning
- Year 0 Removal
- Nighttime Filtering
- Outlying Points (Sensor Failure)

Predictive Power Modeling
- XbX UTC Model
- Linear Fitting
- STL Decomposition
- YoY Performance Loss Rate (PLR)

Cross-Sectional Analysis
- ANOVA
- Random Forests
- AIC

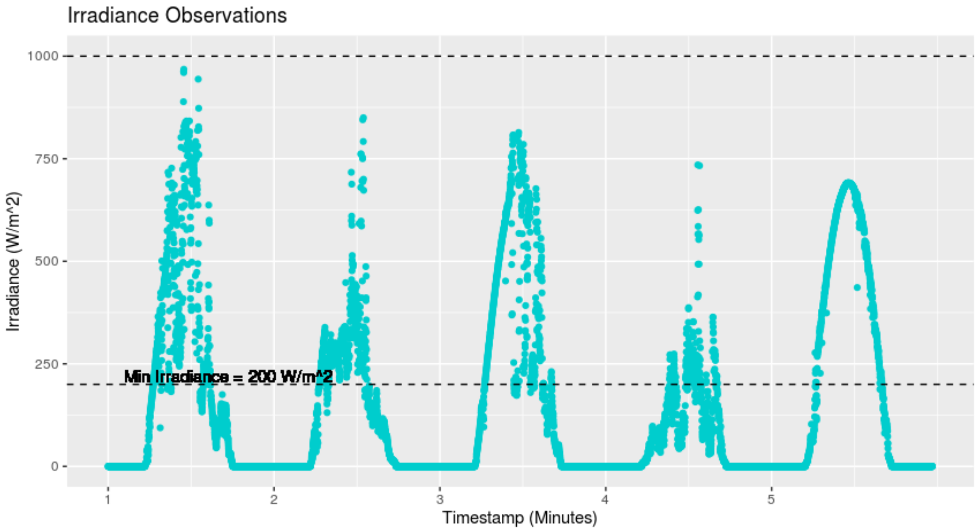
Data Filtering

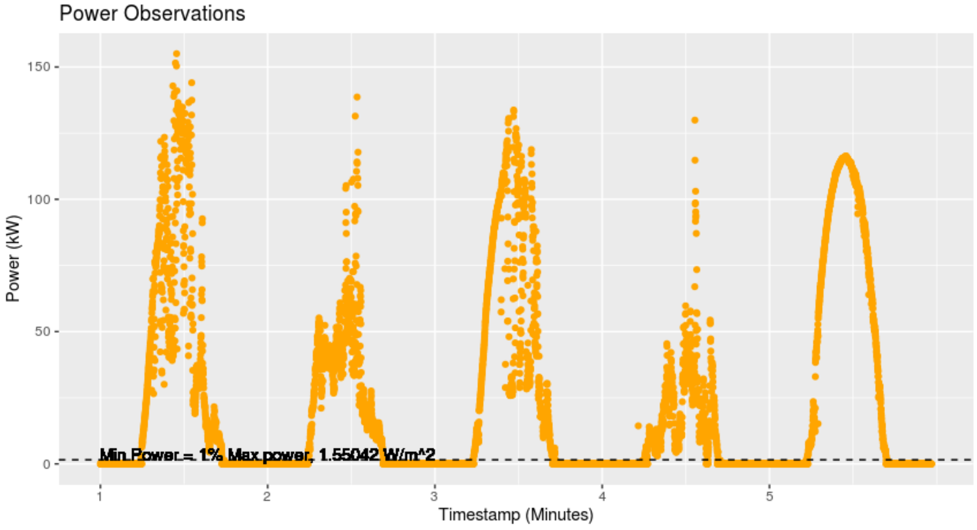
Irradiance Filter

- Minimum filtering at 200 W/m^2 to prevent nighttime effects
- Maximum filtering at 1000 W/m^2 to prevent effects of sensor failure

Power Filter

- Minimum filtering at 1% max power to prevent nighttime effects





“X by X” + UTC Model

“X by X” + Universal Temperature Correction (XbX + UTC) Model

- “X by X” indicates X as a time period
  - DbD for Day-by-Day
  - WbW for Week-by-Week
  - MbM for Month-by-Month
- DbD chosen as the time period
- Converts a measured temperature to a representative temperature
  - Corrects seasonal temperature variation
- Filters irradiance values by 900 ± 10 W/m^2 for consistent irradiance measurements for G_{rep} values

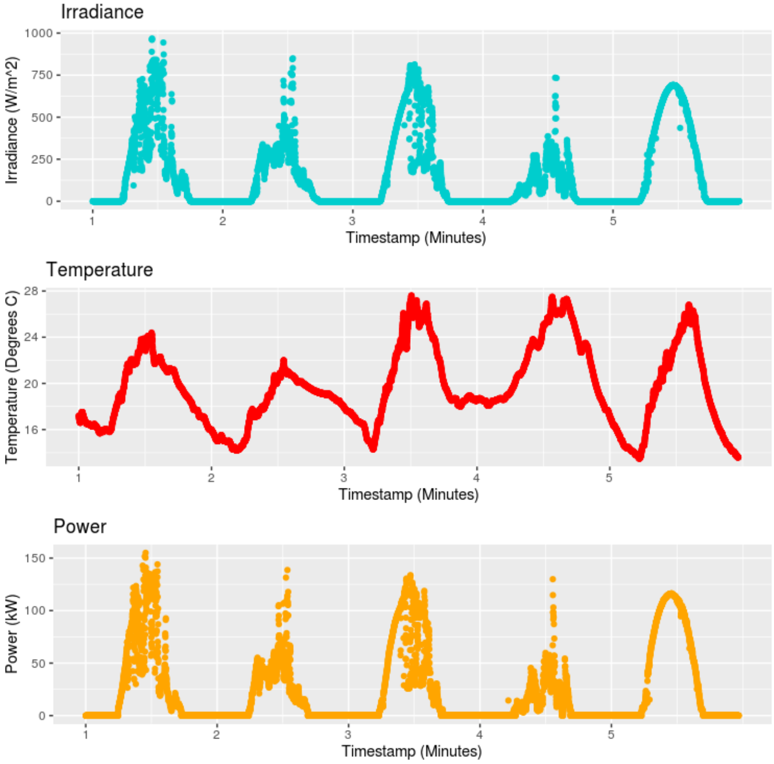
$$P_{cor} = \frac{P_{obs}}{1 + \gamma_T (T_{obs} - T_{rep}) (\frac{G_{obs}}{G_{rep}})}$$

$$P_{cor} = \beta_0 + \beta_1 G + \epsilon$$

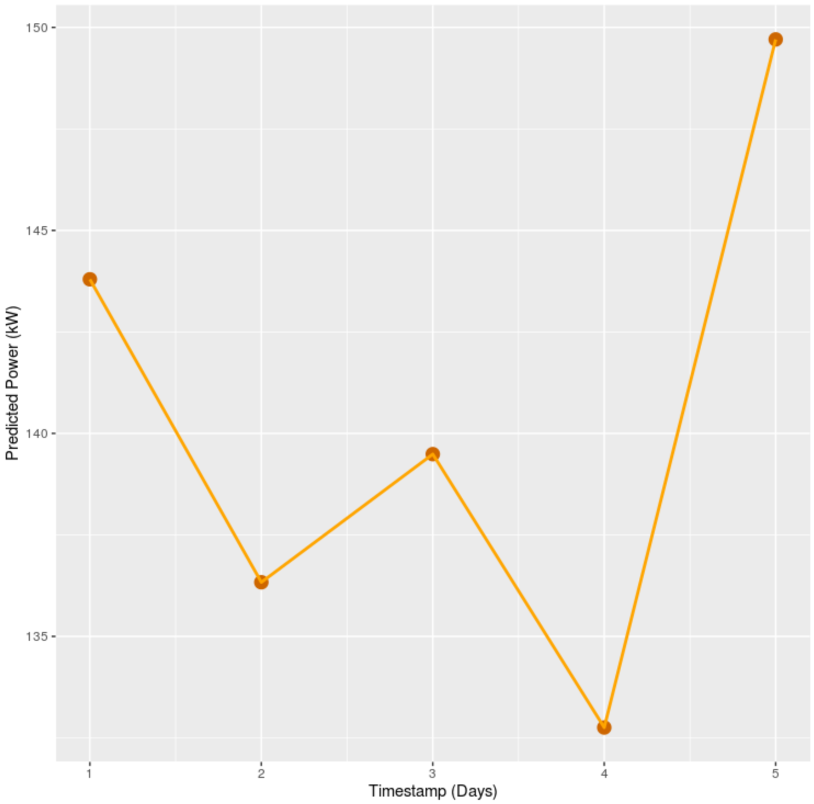
Curran, Alan & Jones, Christian & Lindig, Sascha & Stein, Joshua & Moser, David & French, Roger. (2019). Performance Loss Rate Consistency and Uncertainty Across Multiple Methods and Filtering Criteria. 1328-1334. 10.1109/PVSC40753.2019.8980928.

XbX + UTC Model

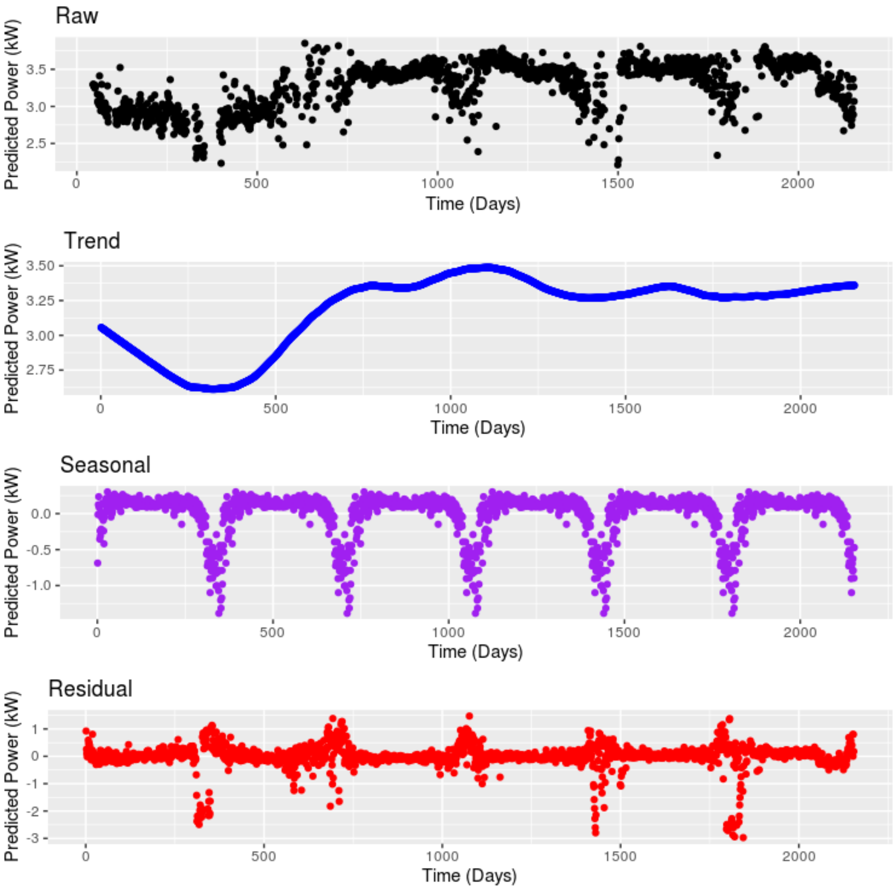
Observed Data



Predicted Power



STL Decomposition



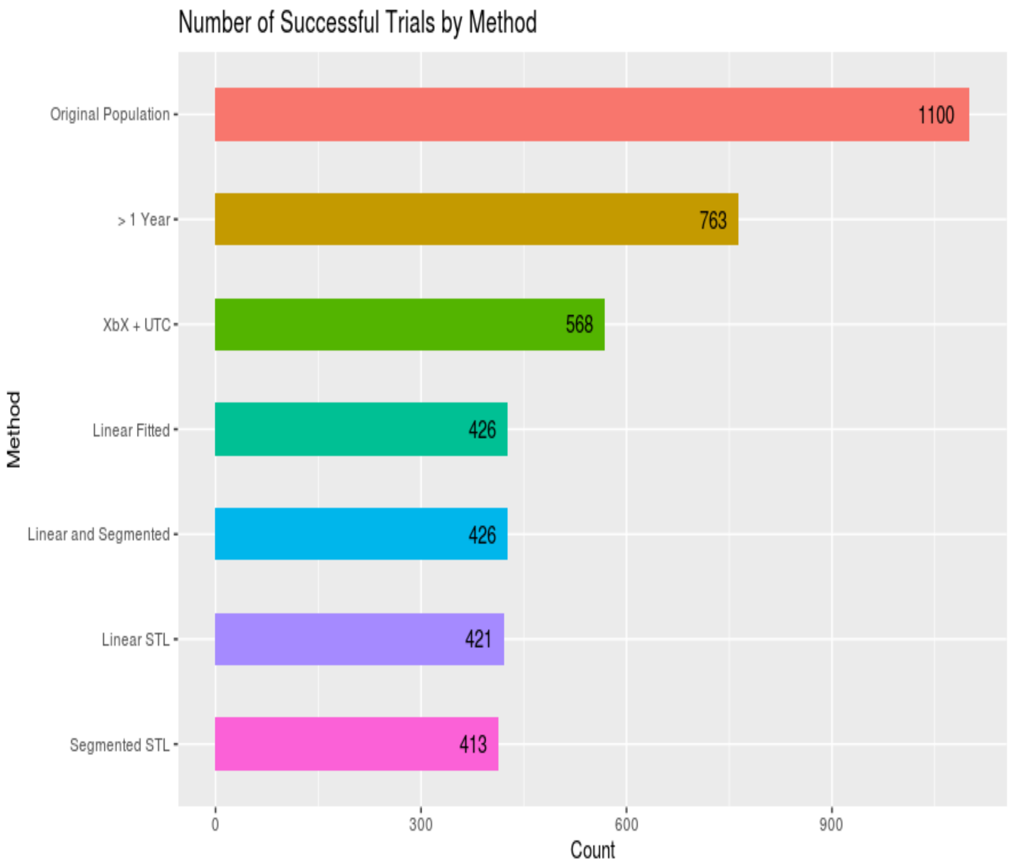
“Seasonal and Trend decomposition using Loess”

- Breaks down time series data into 3 components:
  - loess trend
  - Seasonal
  - Residual
- Raw data is the addition of all 3 components
- Decomposition may fail if
  - Same seasons are repeatedly missing values
  - Variation in data lacks predictable seasonal trends

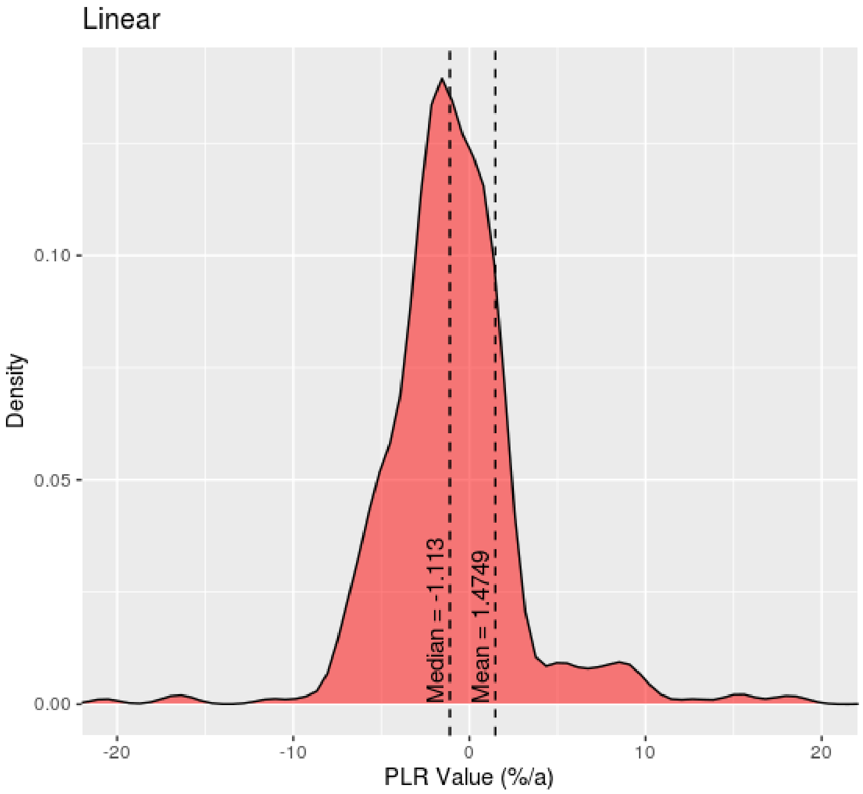
Data Loss

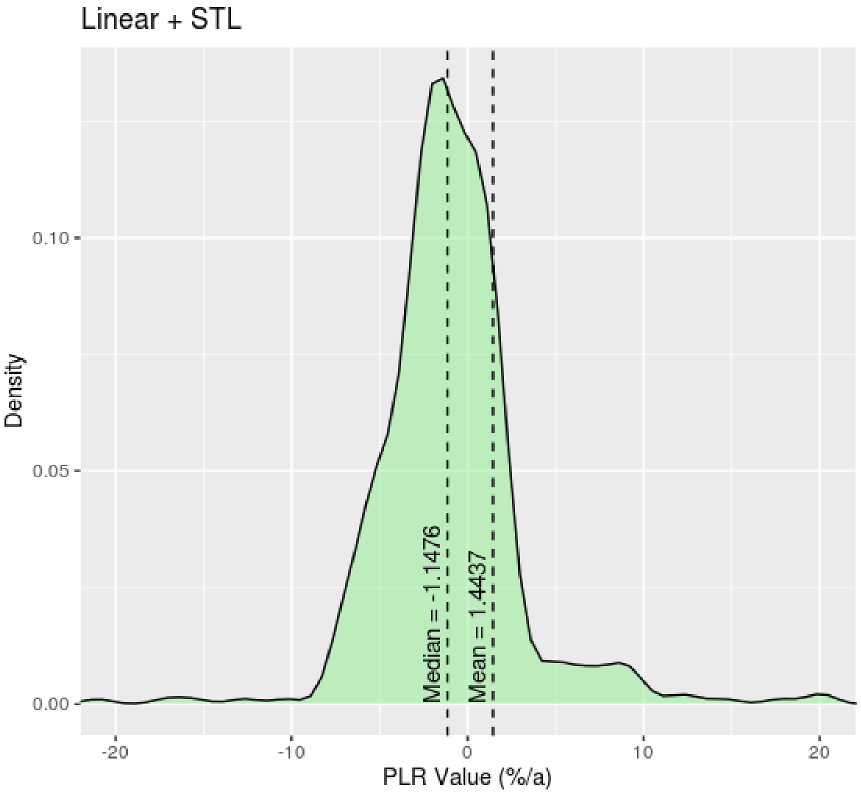
There are two ways we are tracking the loss of data

1. Removal of entire dataframes in each step of data processing pipeline
2. Percentage of missing data points in the XbX + UTC model



PLR Distributions - Linear





Conclusions

1. **Data Processing Pipeline**
    a. Preserved 426 dataframes for use out of the 763 in the sample
2. **PLR Determination Accuracy**
    a. Using a combined Segmented PLR determination with STL Decomposition yields a far more accurate model
        i. Linear PLR at median adjusted R^2 of 0.03, Segmented + STL median adjusted R^2 of 0.28
    b. Low overall adjusted R^2 values indicate that we are unable to capture the variance in our predicted power values with PLR determination methods
3. **PLR Value**
    a. Using our Segmented + STL model, our median PLR values are -0.18% per year and -1.6% per year for segment 1 and 2, respectively
    b. Usage of STL decomposition required for segmented methods for consistent predictions
4. **Cross Sectional Methods**
    a. Random forests and AIC methods both consistently chose module supplier as the primary factor

ST2.5 Performance Loss Rate Determination

Benchmarking of PLR Calculations Methods Using 23 Open Datasets: https://osf.io/vtr2s/

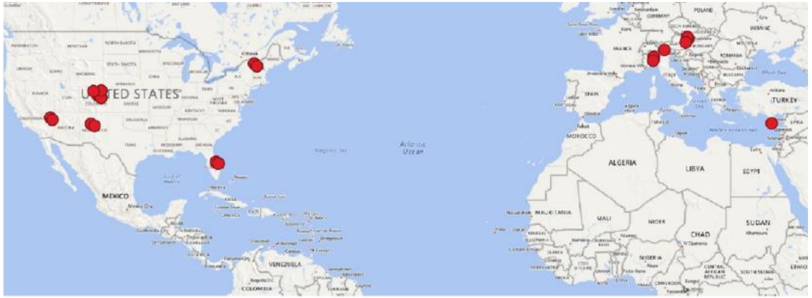
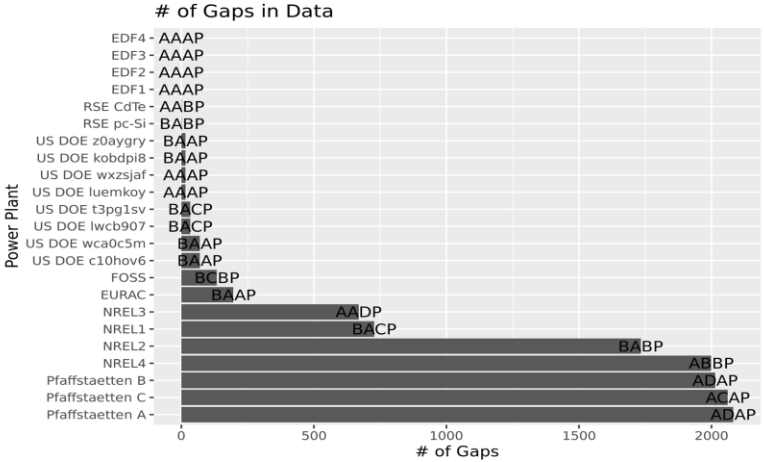
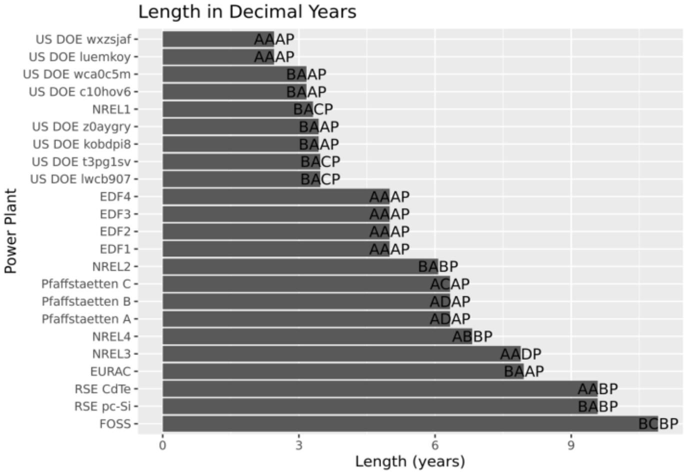
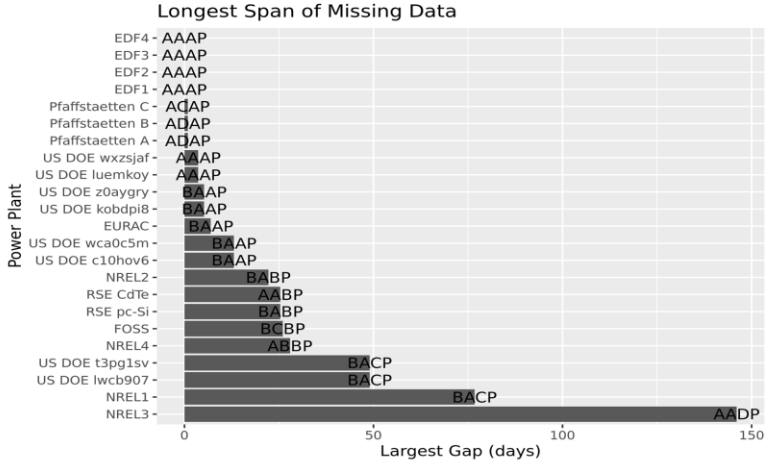


FIGURE 1 The locations of the PV Systems used for the benchmarking exercise.







Spatiotemporal Graph Neural Network for Performance Prediction of Photovoltaic Power Systems

Ahmad Karimi, Yinghui Wu & Mehmet Koyutürk,

Department of Computer and Data Sciences,
Case Western Reserve University

Laura S. Bruckman, Roger French
Department of Material Science and Engineering,
Case Western Reserve University

Spatiotemporal Graph Neural Network (st-GNN)

Interest

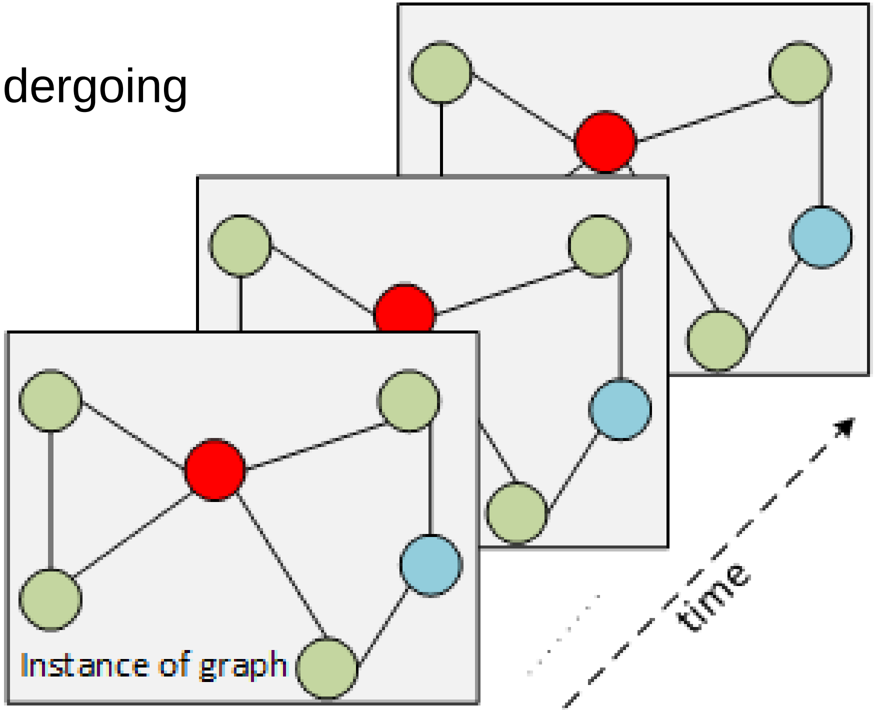
- Information from neighboring nodes undergoing similar exposure

Sequence of

- Graph convolution layer
- Temporal convolutional layer
  - 1-D convolution

Coherence

- Spatial dependencies
- Temporal dependencies

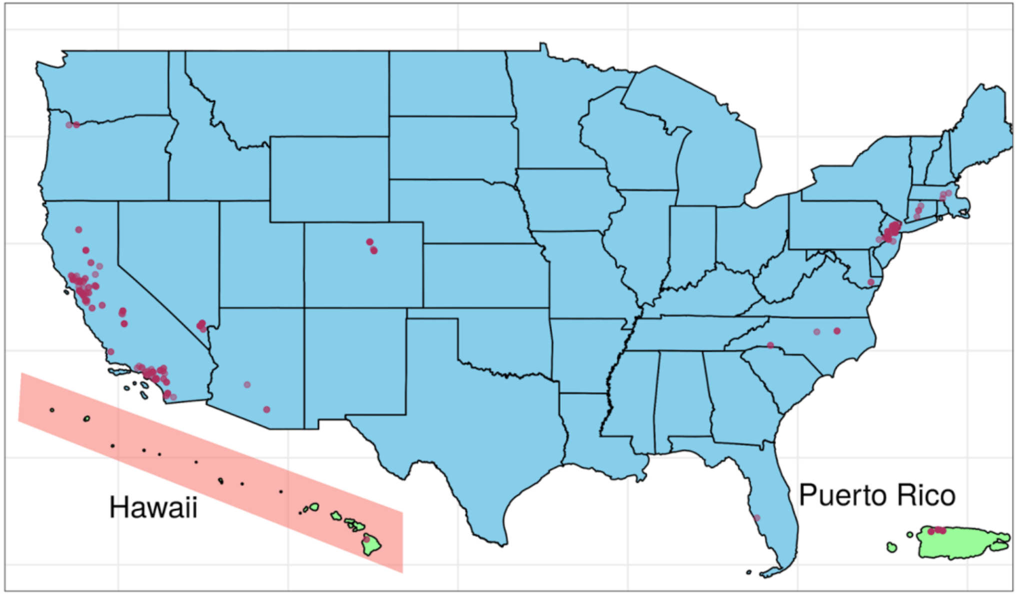


Spatio-temporal graph

Dataset

Dataset

- SS1 + SS2 dataset: 316 power plants
- 2 years of data (730 days)
- 5 minutes interval
- 288 points makes up a day
- 210,240 points for a system
- Data partition
  - 690 days training, 20 days validation, 20 days testing
- Input Features for modeling
  - Power timesries(P_{mp})



Location of PV systems on the map

Power forecasting models (2hrs in future)

- Power (P_{mp})

PV Network Representation

Calculate distance between two nodes

$$d_{lon} = lon_2 - lon_1, d_{lat} = lat_2 - lat_1$$

$$a = (sin(d_{lat}/2))^2 + cos(lat_1) * cos(lat_2) * (sin(d_{lon}/2))^2$$

$$d = 2 * R * arcsin(\sqrt{a})$$

where, R is radius of the earth

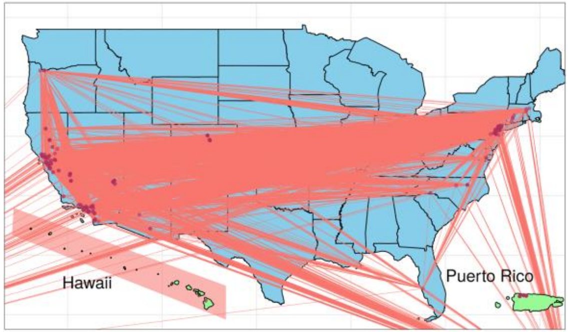
- Equation to convert element of distance matrix to weight matrix
- $\varepsilon_c = 0.5$

$$w_{ij} = \begin{cases} \exp(-\frac{d_{ij}^2}{\sigma^2}), \ i \neq j \text{ and } \exp(-\frac{d_{ij}^2}{\sigma^2}) \geq \epsilon \\ 0 \qquad\qquad\qquad , \text{ otherwise.} \end{cases}$$

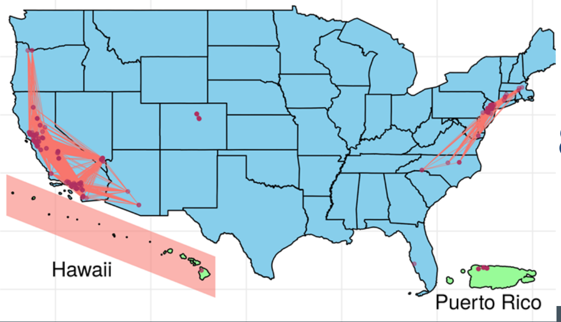
d_{ij} = distance between node i and node j

σ is normalizing constant

ϵ is constant which control graph sparsity



$\varepsilon_c = 0$



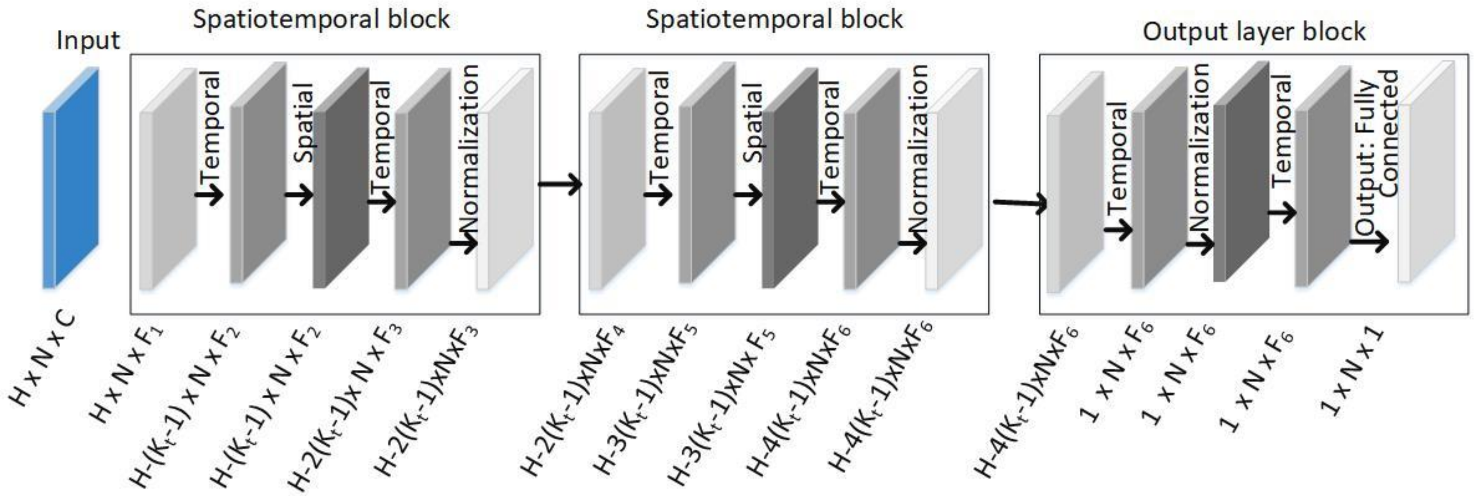
$\varepsilon_c = 0.5$

H.B. Goodwin, The haversine in nautical astronomy, Naval Institute Proceedings, vol. 36, no. 3 (1910)

Spatiotemporal Graph Neural Network (st-GNN) Representation

Two Spatio-temporal Block
- Two temporal convolution layer
- One spatial convolution layer

Output Layer Block
- Two temporal
- Fully connected layer



H: Number of previous time points, N: Number of PV systems, K_t: Kernel size, F_1-F_6: Filters

H = 24 number of time lag points
N = 316 PV Systems

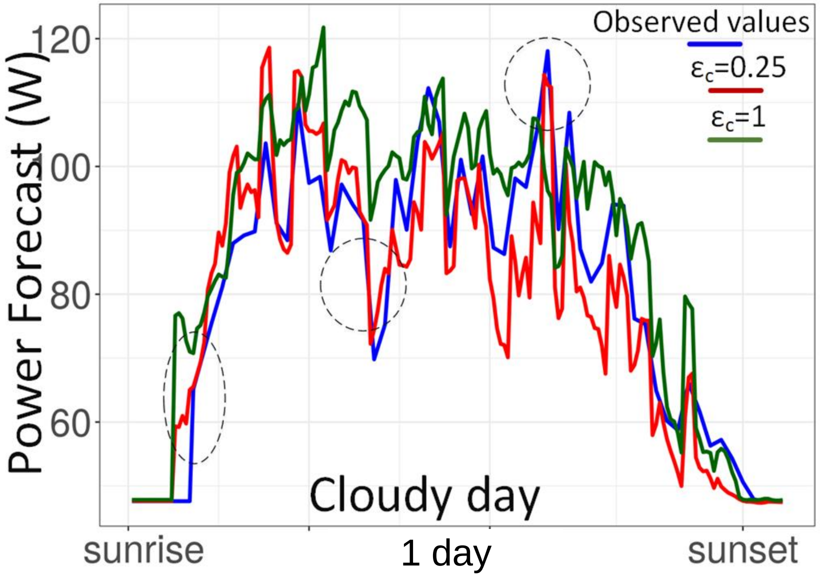
Trainable parameters:
1 Channel Network: 775,468

Result

Single System PV Power Forecast

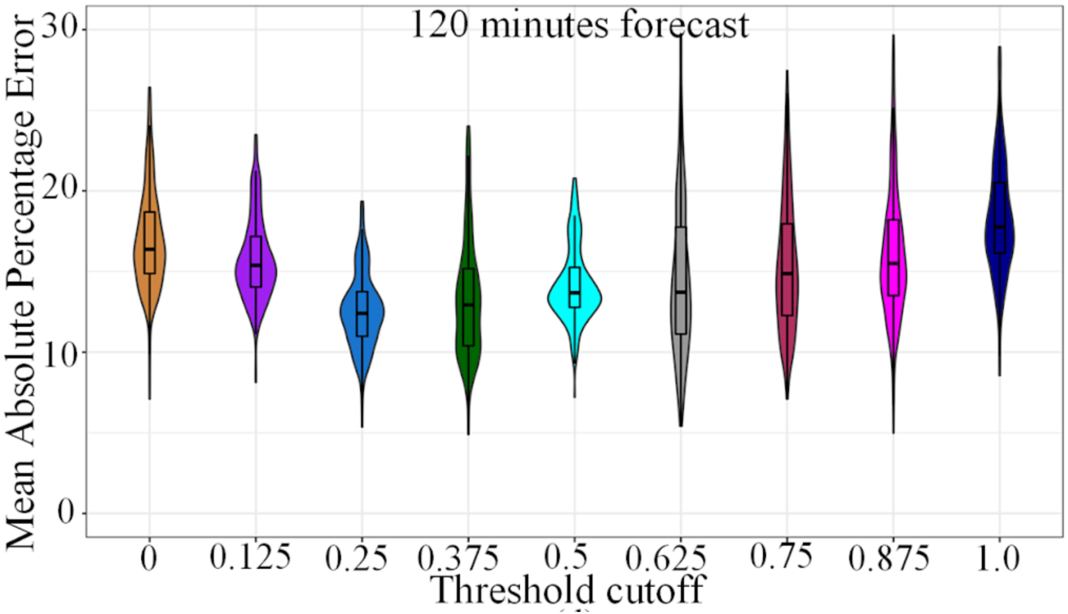
PV power forecast for one day

- Fluctuation in the curve due to cloud cover
- Forecast for spatiotemporal convolution (ε_c= 0.25)
- Forecast for temporal (1-D) convolution (ε_c= 1.0)
- Spatiotemporal curve follows observed values trend closely



GCN Model Accuracy

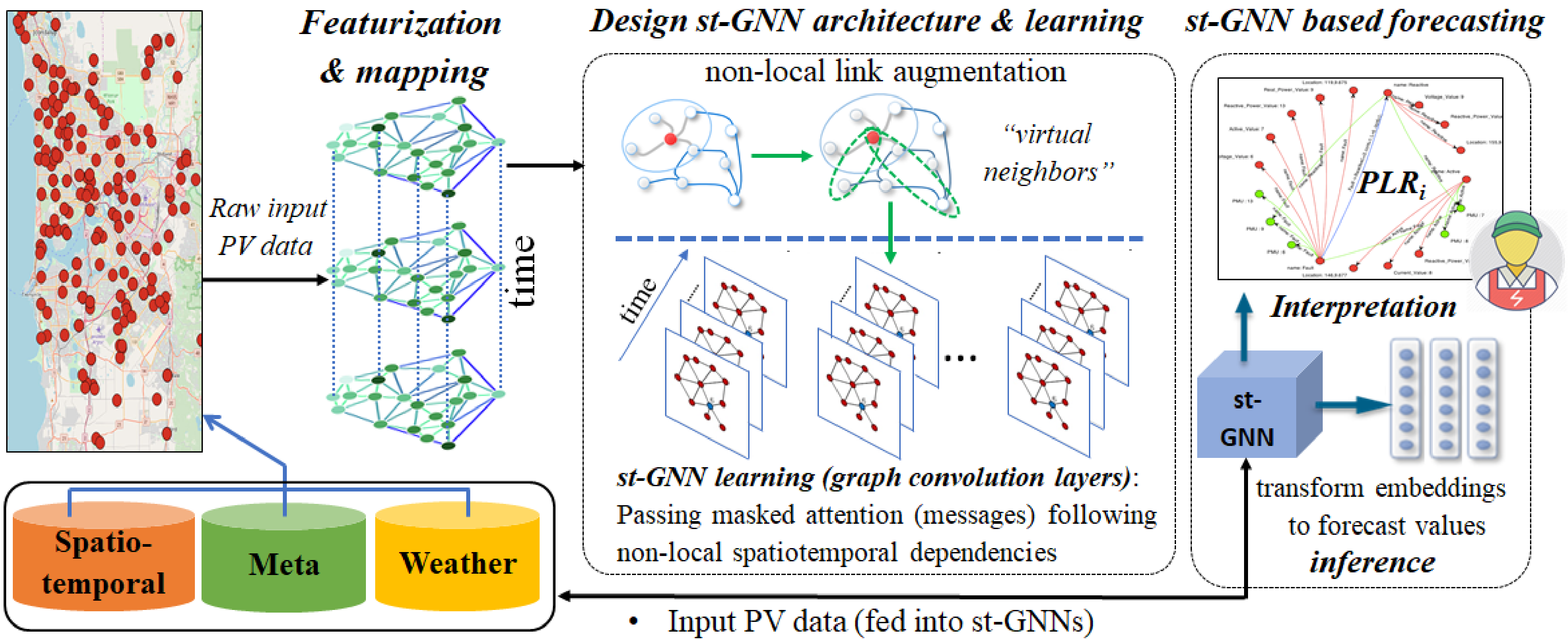
Spatiotemporal GCN & temporal convolution



Forecast (minute)	MAPE for 316 systems			
	s-t convolution		temporal convolution	
	ϵ_c=0.375		ϵ_c=1.0	
	mean	sd	mean	sd
120	11.01	5.04	18.98	5.15
105	9.31	4.36	15.63	4.57
90	8.39	3.87	13.62	4.07
75	7.67	3.36	11.78	3.54
60	7.24	2.87	10.12	2.96
45	6.28	2.61	8.42	2.45
30	4.68	2.48	6.65	2.13
15	2.75	2.37	3.92	2.01

Table 1: Mean and standard deviation of MAPE values for temporal convolution (standalone) vs spatiotemporal convolution for PV systems with optimum ϵ_c for st-GNN network.

New DOE-SETO AI-for-PV project

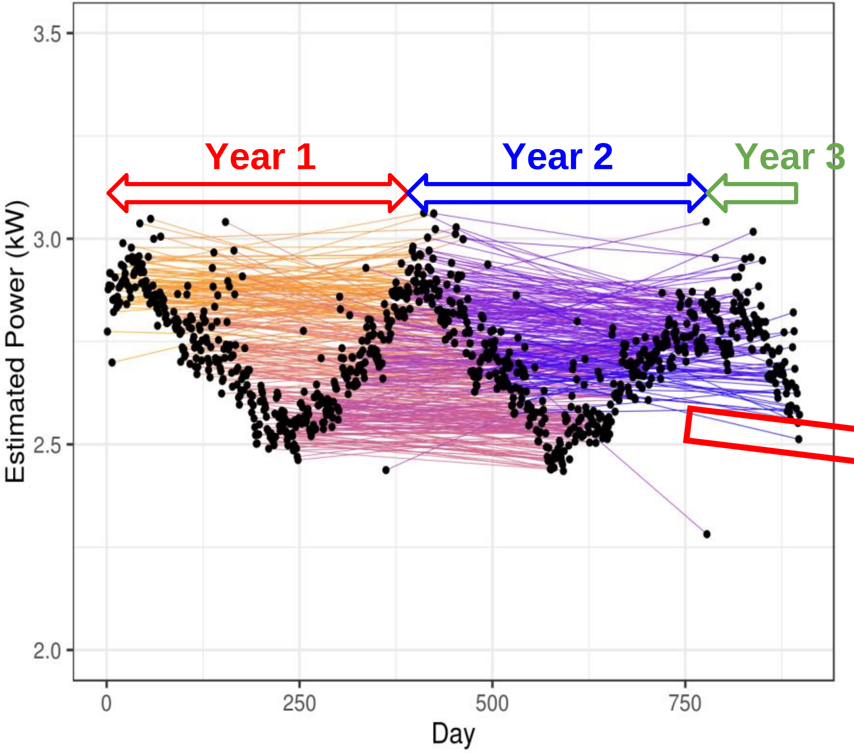


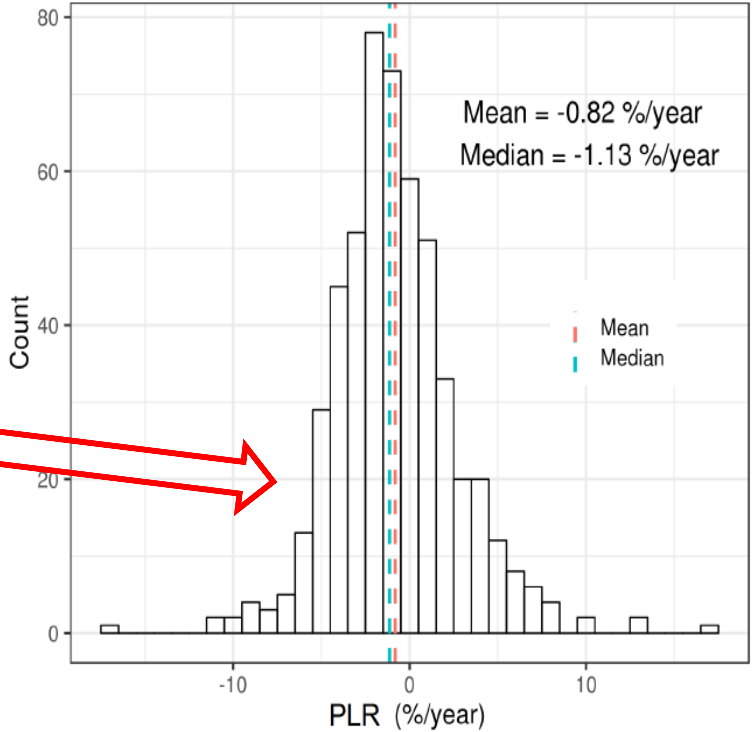
FAIRification of PV st-Data (CRADLE)

- Input PV data (fed into st-GNNs)
- Forecast results/models (back to data store)

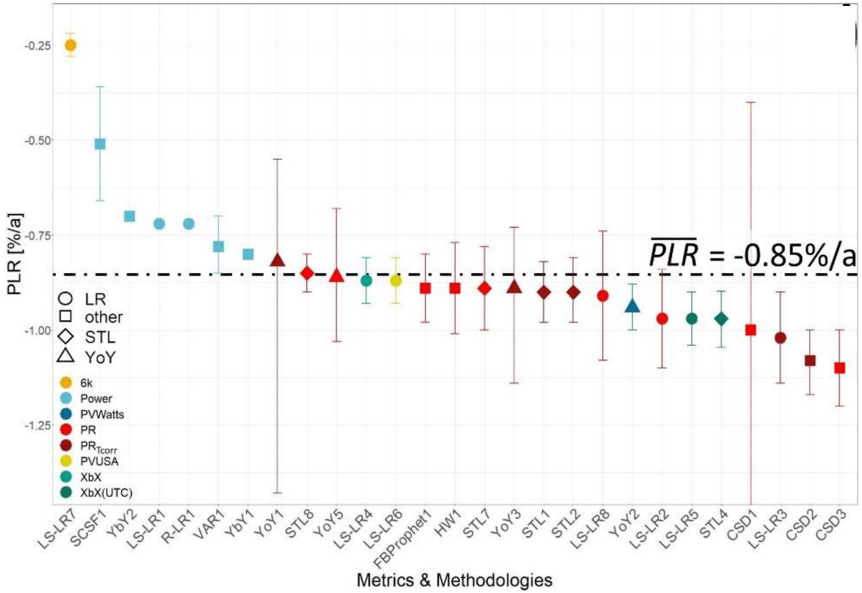
Questions

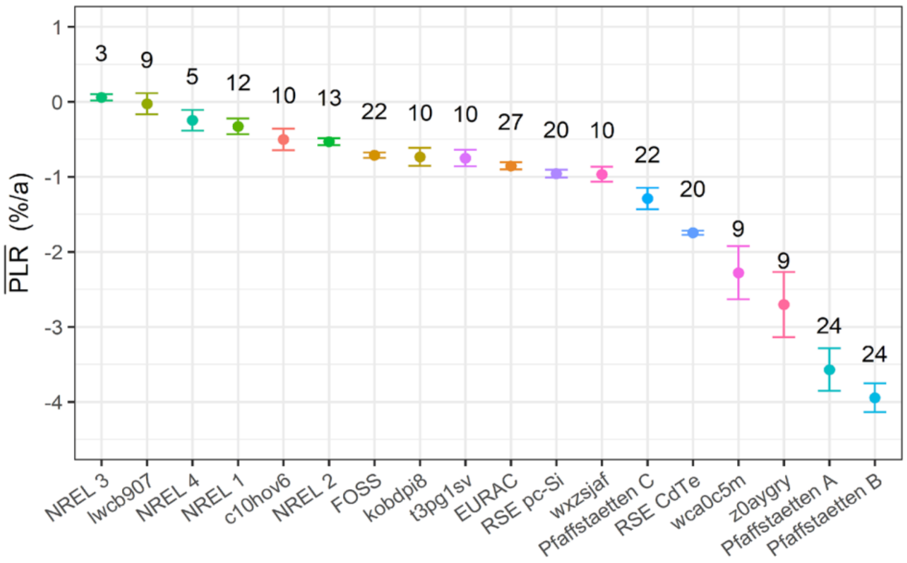
Year-on-Year Performance Loss Rate (PLR)





PLR of 1 System by 27 Methods. And of 18 Systems





PLR of the EURAC System

- By 27 Metric/Statistical Model Approaches
- Ensemble model yields mean *PLR* ($\overline{PLR}$)
- $\overline{PLR}$ = -0.85%/annum

$\overline{PLR_i}$ determined for 18 PV Systems

- Using ensemble model (voting) approach
- With 83.4% Confidence Intervals
- Significant Differences among these PV systems

Performance Loss Rate Determination

- Task 13 members and other PV researchers completed a benchmarking study of approaches for calculation of the Performance Loss Rates (*PLR*) of a large number of commercial and research PV power plants in diverse climatic zones, utilizing the PV systems' power and weather data.
- The combination of 1) data cleaning and filtering, 2) metrics (performance ratio (*PR*) or predicted power (*P*) based), temperature corrections, and data aggregation, 3) time series feature corrections, and 4) statistical modeling methods are benchmarked in terms of a) their deviation from the $\overline{PLR}$ value, and b) their uncertainty, standard error and confidence intervals.
- These results will inform standards development for *PLR* determination, which was previously attempted with an initial proposal for a new IEC 61724-4 standard. However, the results reported here suggest that proposing a specific standardized method is still premature.

PVPS

Report IEA-PVPS T13-22:2021, April 2021