Rate Of Change Determination of Real-world Commercial PV Power Plants Using Data-driven Modeling

Alan J. Curran, Ahmad Maroof Karimi, Kevin J. Nash, JiQi Liu, Yang Hu, Roger H. French

SDLE Research Center,
Case Western Reserve University,
White 536, 10900 Euclid Ave.,
Cleveland OH, 44106, USA

SDLE Research Center: Acknowledgements

CWRU Faculty
- Roger French, Laura Bruckman, Alexis Abramson, Jennifer Carter, Mehmet Koyüturk

Post-doctoral Research Associates
- Jennifer Braid, Nick Wheeler, Rojiar Haddadian, Wei-Heng Huang

Graduate Students
- Devin Gordon, Yu Wang, Donghui Li, Alan Curran, Addison Klinke
- Justin Fada, Arash Khalilnejad, Ahmad Karimi, Xuan Ma,
- Rhener Zhang, Menghong Wang, JiQi Liu,

Undergraduates
- Andrew Loach, Lucas Fridman, Yiyang Sheng, Jonathan Ligh, Silas Ifeanyi
- Noah Tietsort, Kevin Nash, Rachel Swanson, Abhi Devahti, Jonah Larson

High School: Sheina Cundiff, Dominique Gardner, Precious Flanders

SDLE Staff: Chris Littman, Rich Tomazin

Projects

Data Analytics of Complex Systems

Materials are parts of a Complex Systems:

- Coatings on Complex Substrates
- Used in Complex Environmental Exposures and Climate Zones

Materials Degradation Predictive & Mechanistic Models

- Predictive Modeling of Materials Degradation
- Mechanistic Network Models To Guide New Materials Development
- Cross-correlation of Real-world and Accelerated Studies for Service Life

Image Processing

- Develop Pipeline Methodology, Apply to Historical Datasets
- Cluster Output and Compare Cell-level Heterogeneity with I-V

Machine Learning

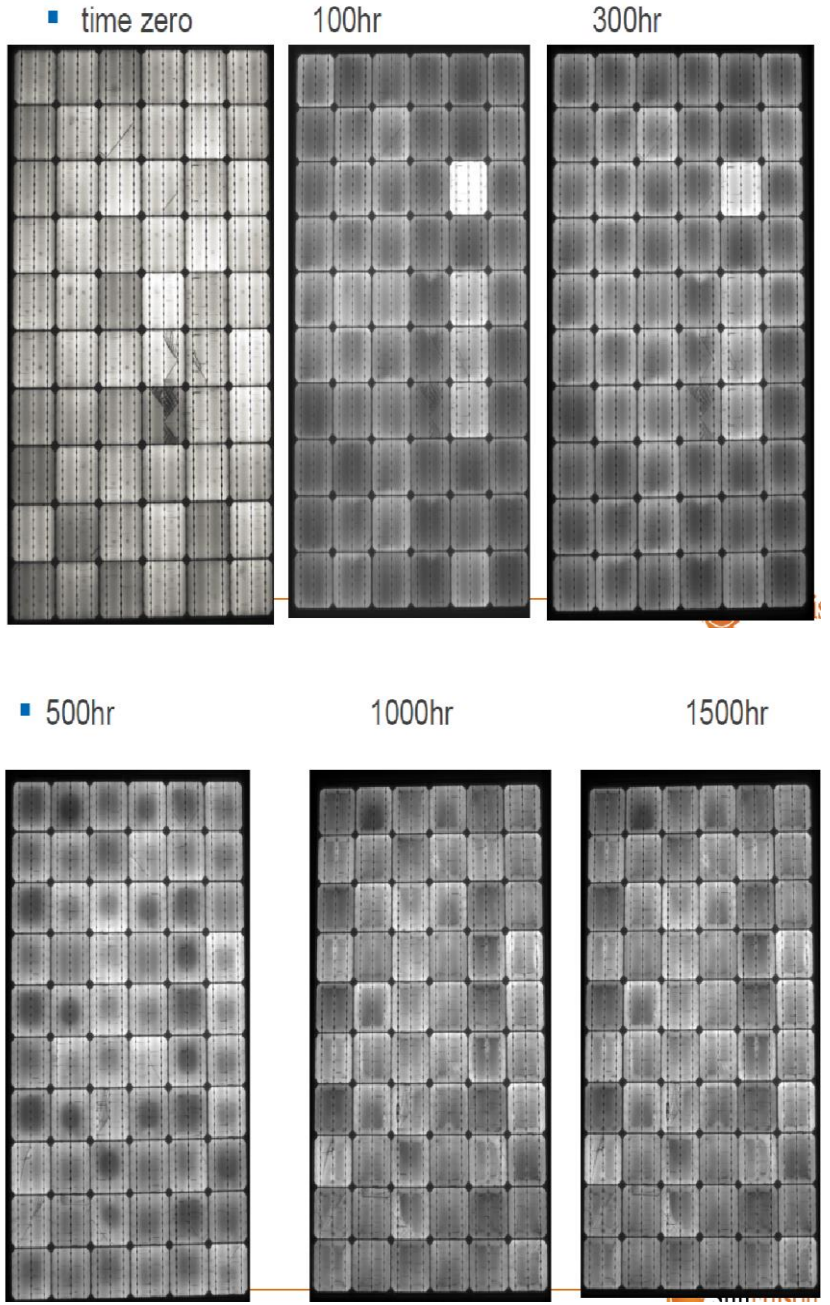
- Classifying Stages of Degradation: Identifying Feature Change Over Time
- Cluster Cell Behavior to Model of Ensemble Performance
- Determine Features Variation with Indoor Testing, Compare / Contrast

Time Series Analysis

- High Performance Computation / Data Pipelining for Rd Analysis
- Subset Datasets by Climate, Module Brand, Inverter Brand

Sample Sets of Systems such as PV power plants

- Sample Set Segmentation, Identify Performance Changes and Variations



Data Science:
Informatics and Analytics

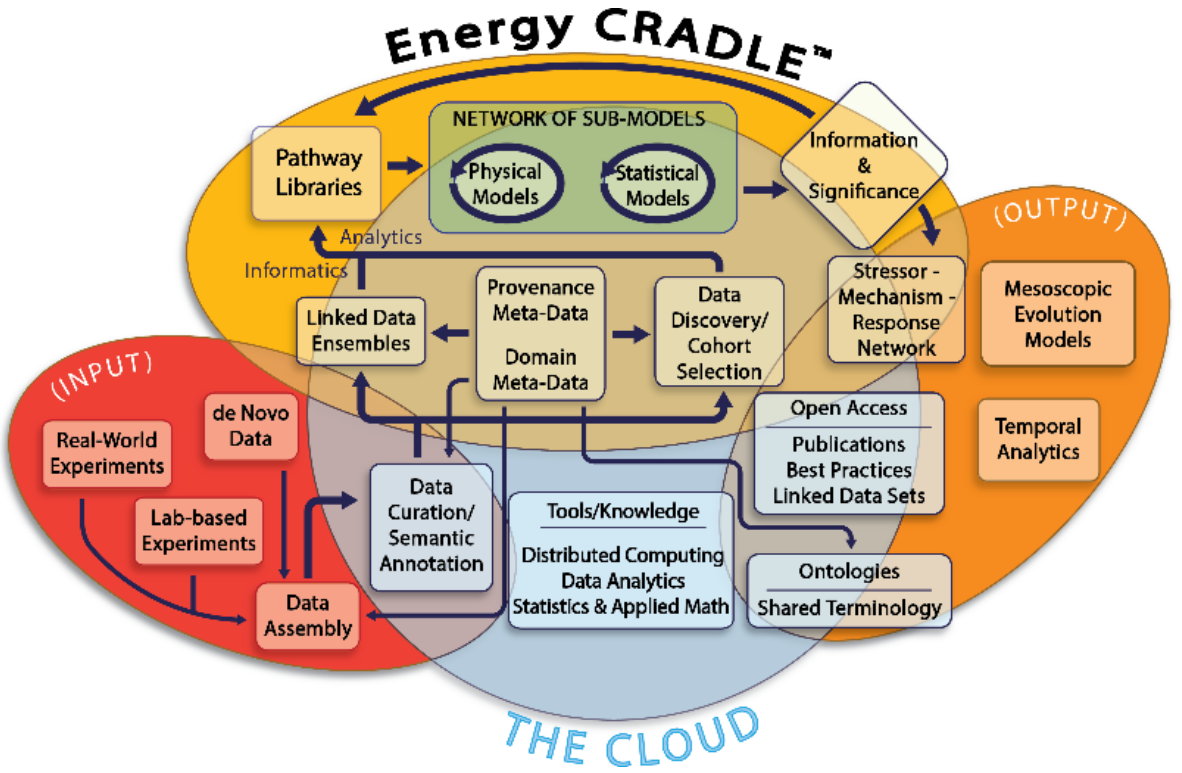
SDLE Research Center
Case Western Reserve University
Cleveland OH 44106

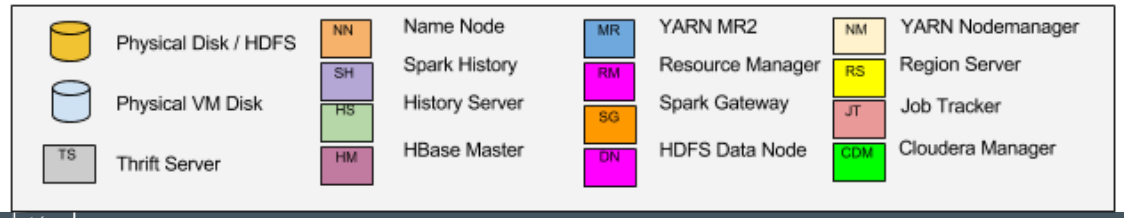
CRADLE v2.1 Architecture: Petabyte and Petaflop Computing

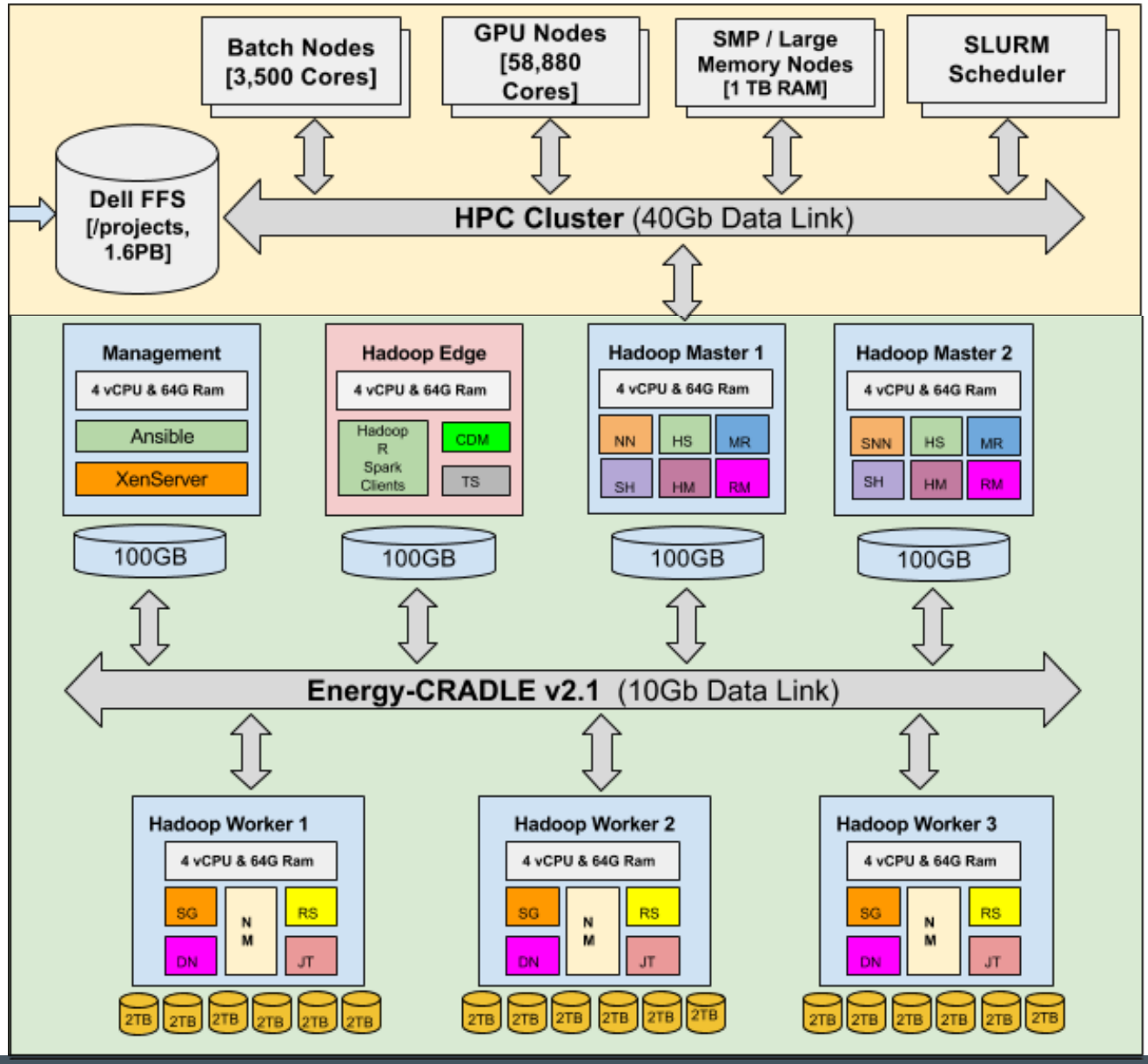
National Strategic Computing Initiative 2015

Hadoop/Hbase/Spark

Based on Cloudera CDH5 distribution





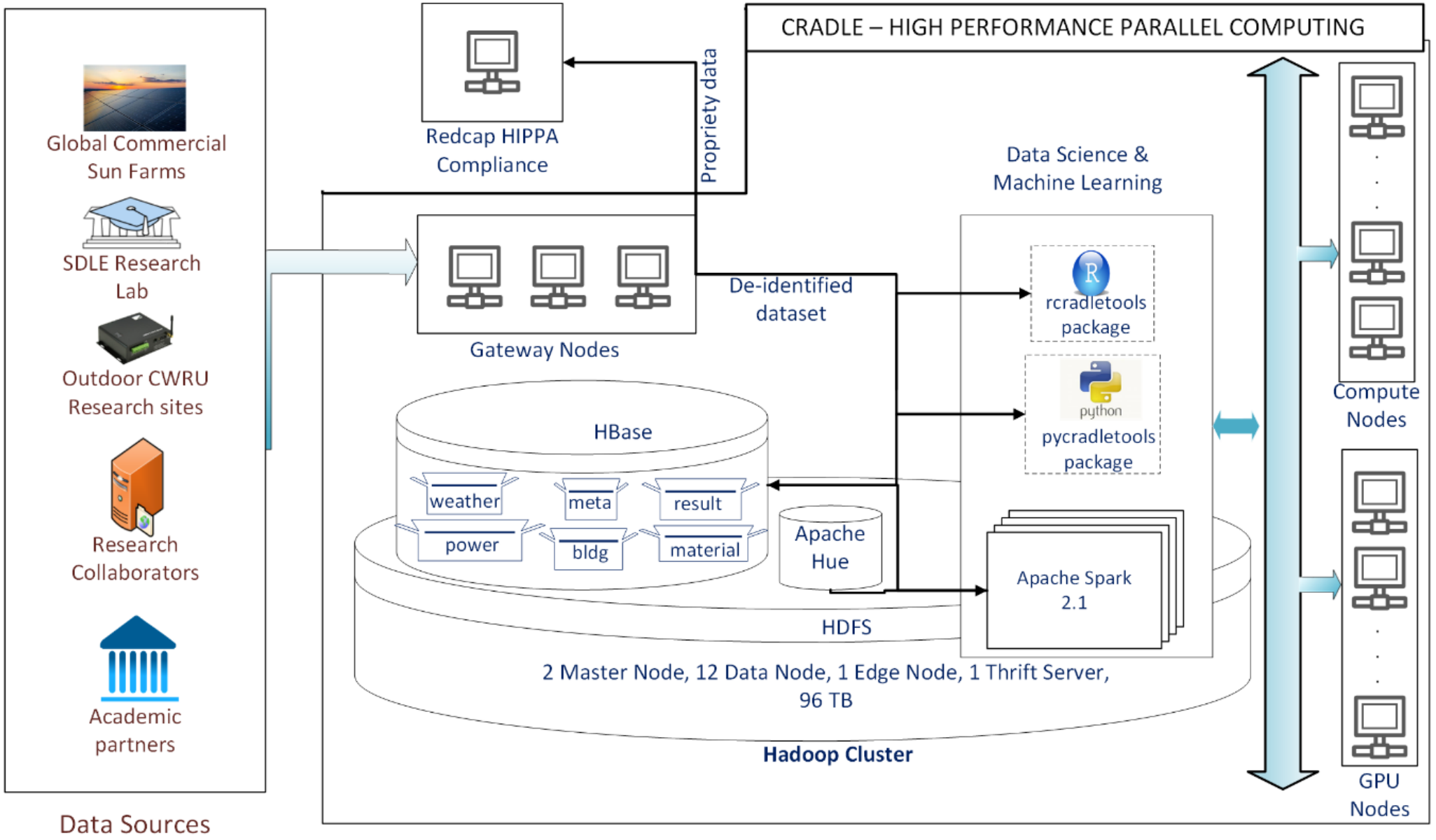


CRADLE v2.2 Architecture: Petabyte and Petaflop Computing

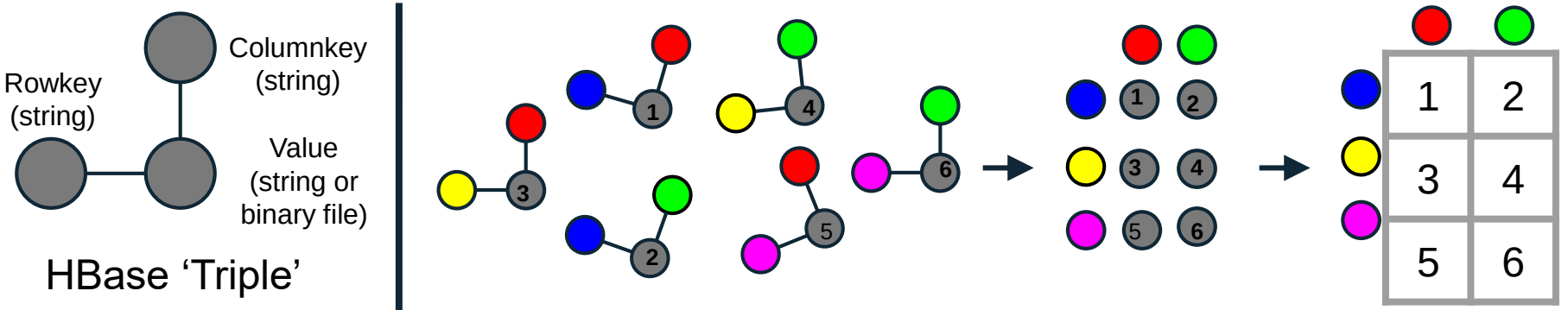
Using R & Python

In-place Analytics

Write-back
all Results
into Hbase



NoSQL DB Abstraction of Hadoop/Hbase



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

High Performance PV Data Analytics: Petabyte Data Warehouse In A Petaflop HPC Environment

- In-place Analytics: Distributed R-analytics in Hadoop/HDFS
- In-memory Data Extraction: To Separate HPC Compute Nodes

IEEE JPV

A non-relational data warehouse for the analysis of field and laboratory data from multiple heterogeneous photovoltaic test sites

Yang Hu, *Member, IEEE,* Venkat Yashwanth Gunapati, Pei Zhao, Devin Gordon, Nicholas R. Wheeler, Mohammad A. Hossain, *Member, IEEE,* Timothy J. Peshek, *Member, IEEE,* Laura S. Bruckman, Guo-Qiang Zhang, *Member, IEEE,* and Roger H. French, *Member, IEEE*

Real-world Data Source: CWRU SDLE Global SunFarm Network

SDLE PV Data Covers ~3.4 GW

Encompasses 1.92% of Global PV Power Production

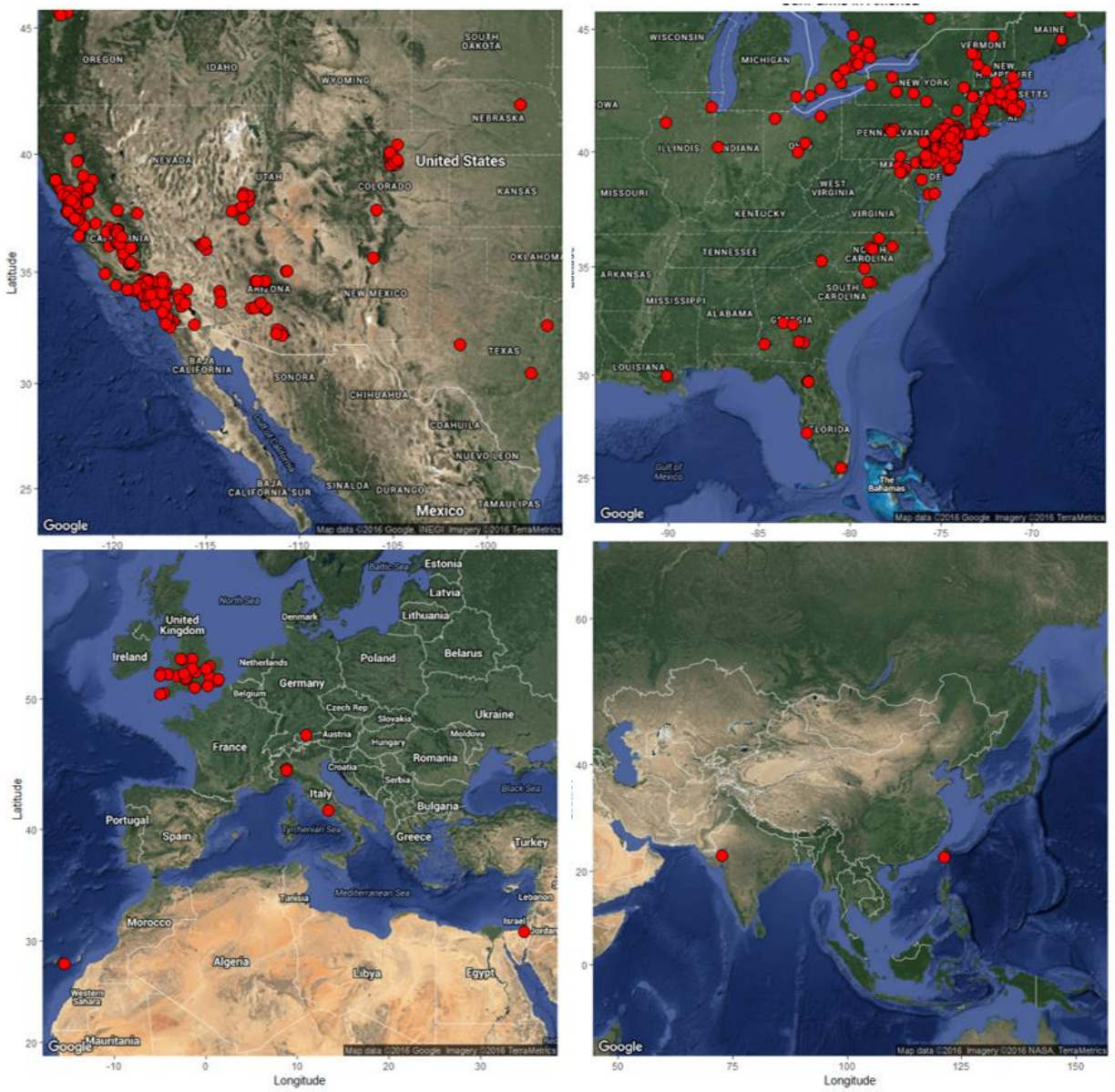
- 787 PV Project Sites
- 5638 PV Systems (Inv. & Modules)
- 60 PV Module Brands/Models
- 38 PV Inverter Brands/Models
- Across 13 Köppen-Geiger Climatic Zones
- Single Modules to 265 MW plants
- Going Back Up To 15 years

Epidemiological PV Populations

- Of Time-series data streams
- Real-world power production
- Real World Exposure Conditions
- Operating Over Real Time-scales

11 Different Companies Have Signed On

- To our Data Use Agreement



ETL and Data Ingestion to Hbase

ETL: Extract, Transform, Load

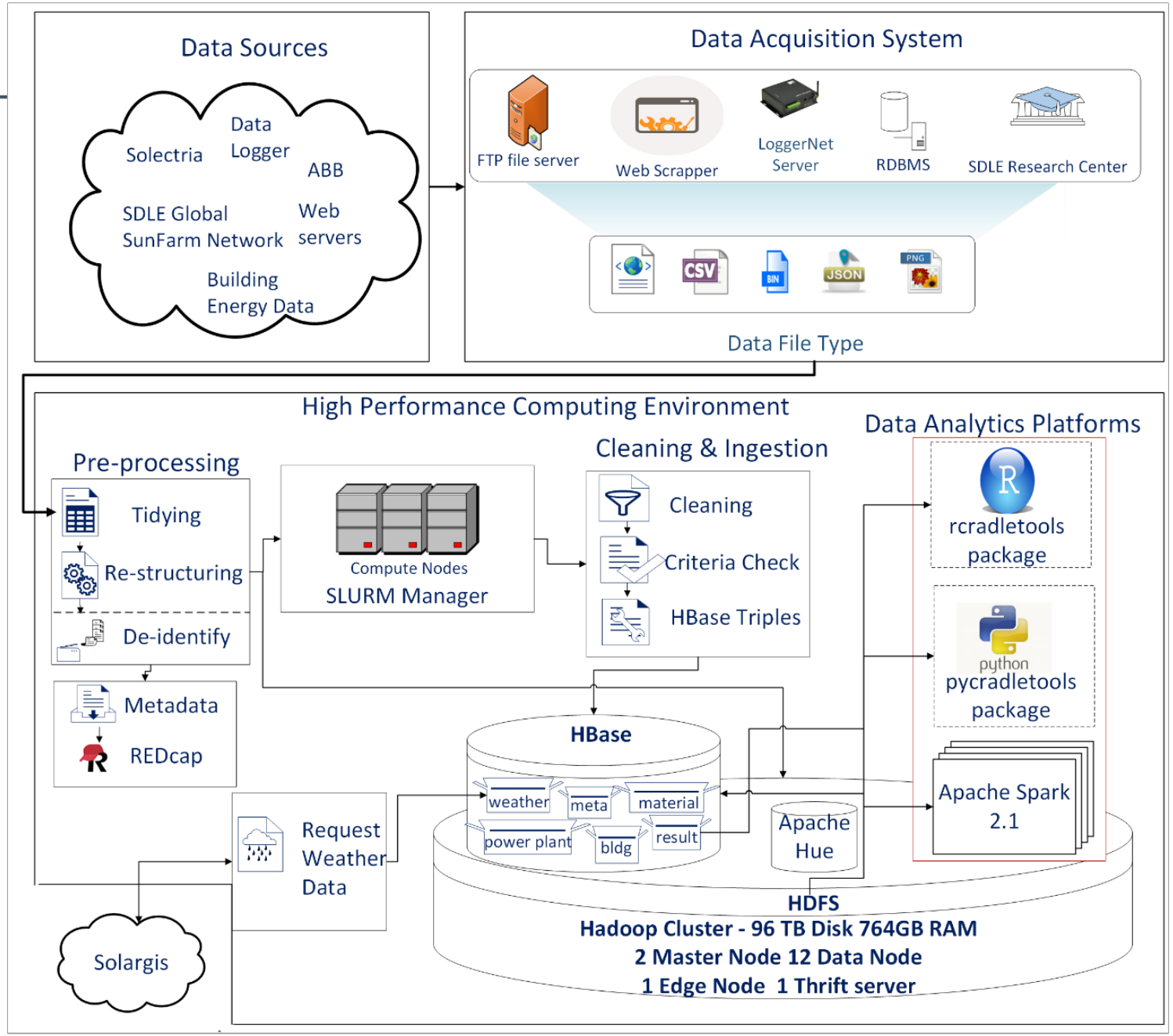
- Standard process for data acquisition
- Typically into an RDBMS system
  Relational Database Management System

Data Ingestion

- Used for NoSQL Databases like Hbase
- Preparing the data for inclusion in Triples
  Rowkey
  Columnkey
  Value

Hbase Tables

- Metadata: information about the data
- Weather: Weather & Irradiance Time-series data
- Power Plant: PV Power Plant Time-series data
- Buildings: Building Electrical Time-series data
- Materials: Spectral, Image data of Materials



Time-series Analysis of Real World PV Systems To Determine Rate Of Change (ROC)

Alan Curran, JiQi Liu, Yang Hu

SDLE Research Center
Case Western Reserve University

Other ROC Methods: Responsivity Method

The Responsivity is a standard method
For determining a system degradation rate

- Has started to fall out of favor recently

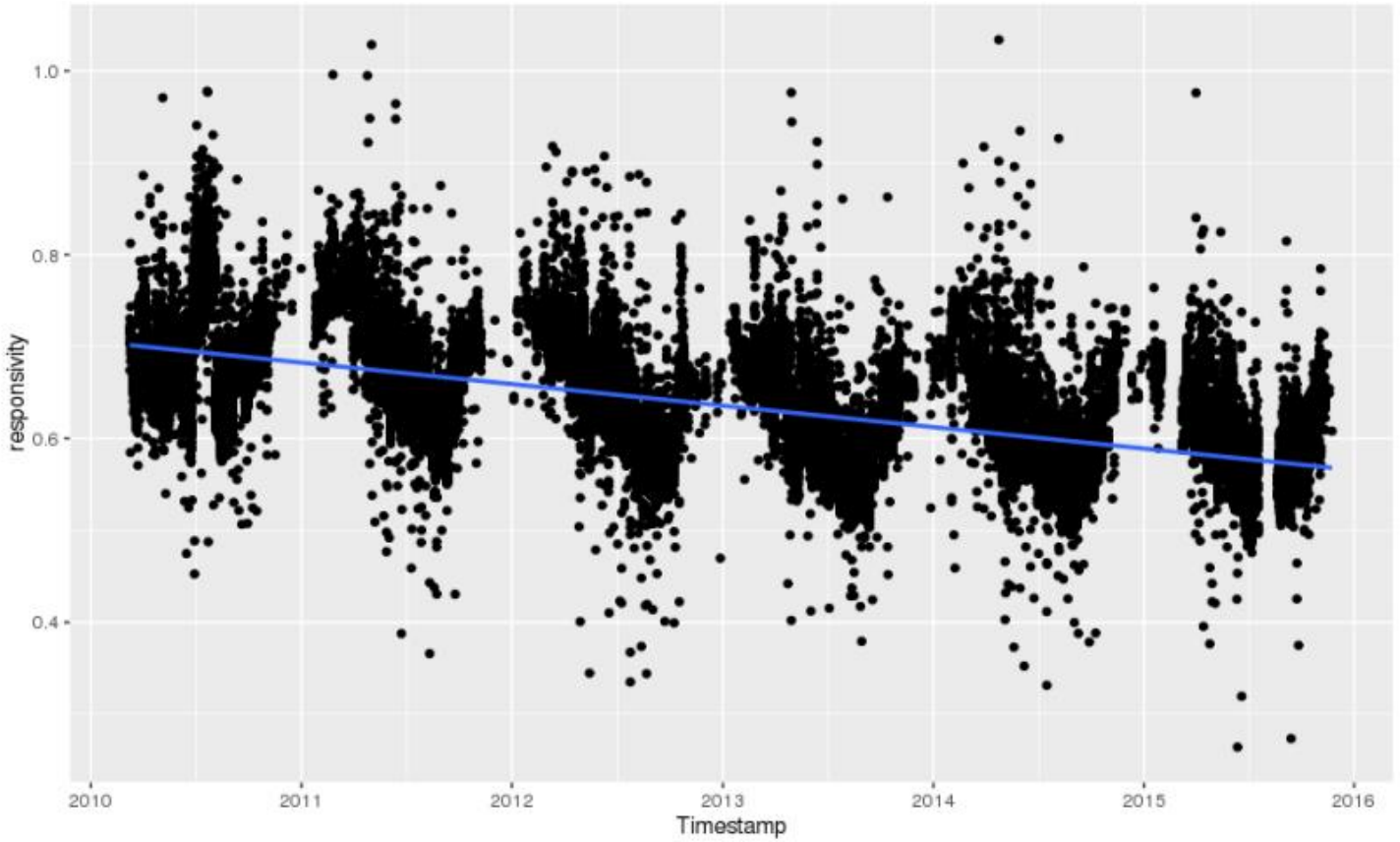
Data is subset and reduced to performance ratios
And data filters are applied

- Convert data to a given irradiance and temperature usually STC
- Most notably a deviation filter removes data points that lie outside a given standard deviation
- Assumes a linear trend

Results can be influenced
by selection of data filtering thresholds

Our MbM method aims to reduce
the amount of human interference in the data

- Automated process
- Removes far less data



An example of a Responsivity method fit

The data shows high variance,
leading to a less robust regression

Other ROC Methods: Year-on-year Degradation

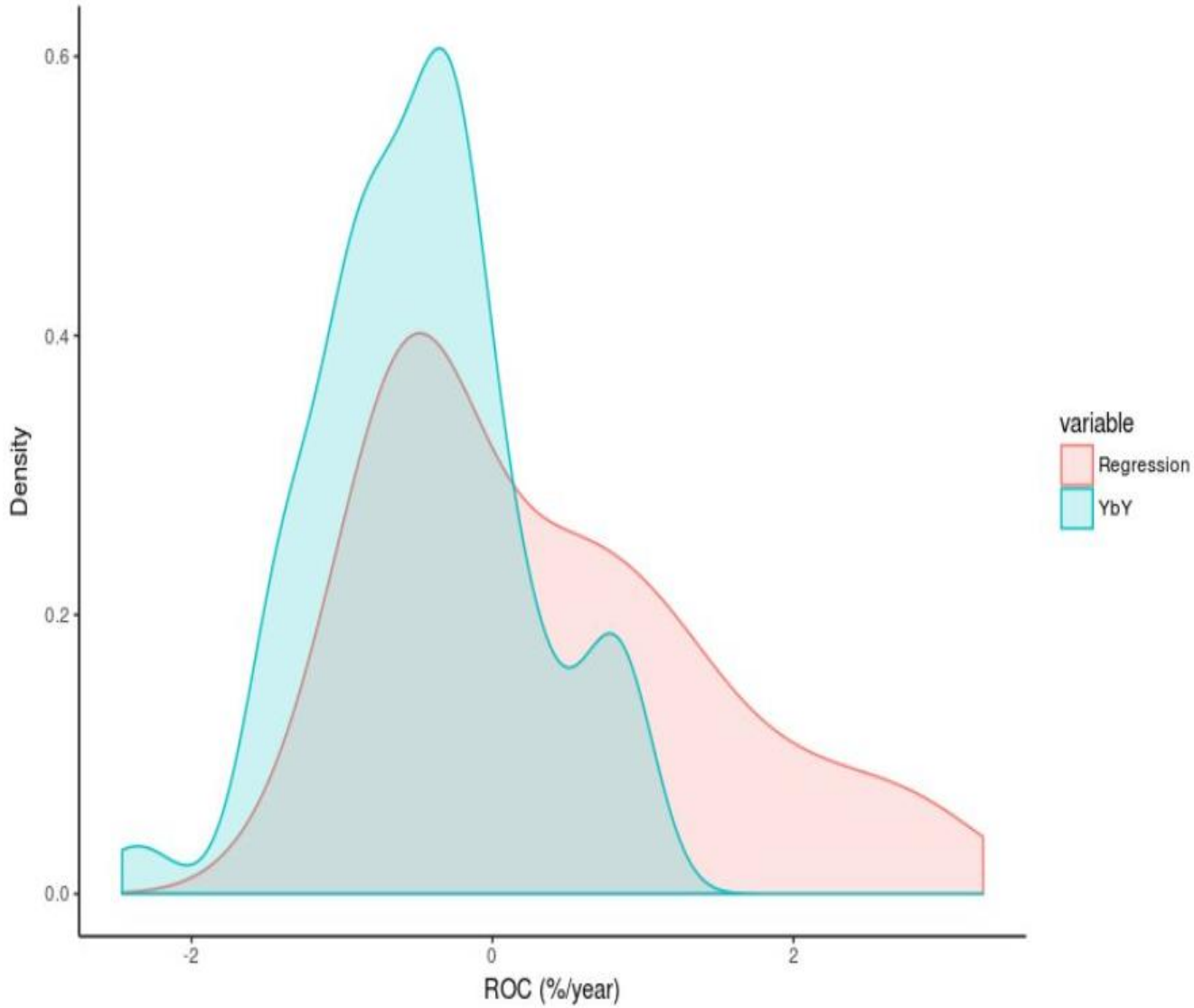
PVLife Model developed by SunPower[1]

Tracks the slope between data exactly one year apart

- Large distribution of slopes
- Median of distribution gives good estimate of the system ROC
- Highly robust to missing points or outliers
  Which can influence traditional regression

Can be used with the MbM method to track differences between each month by year

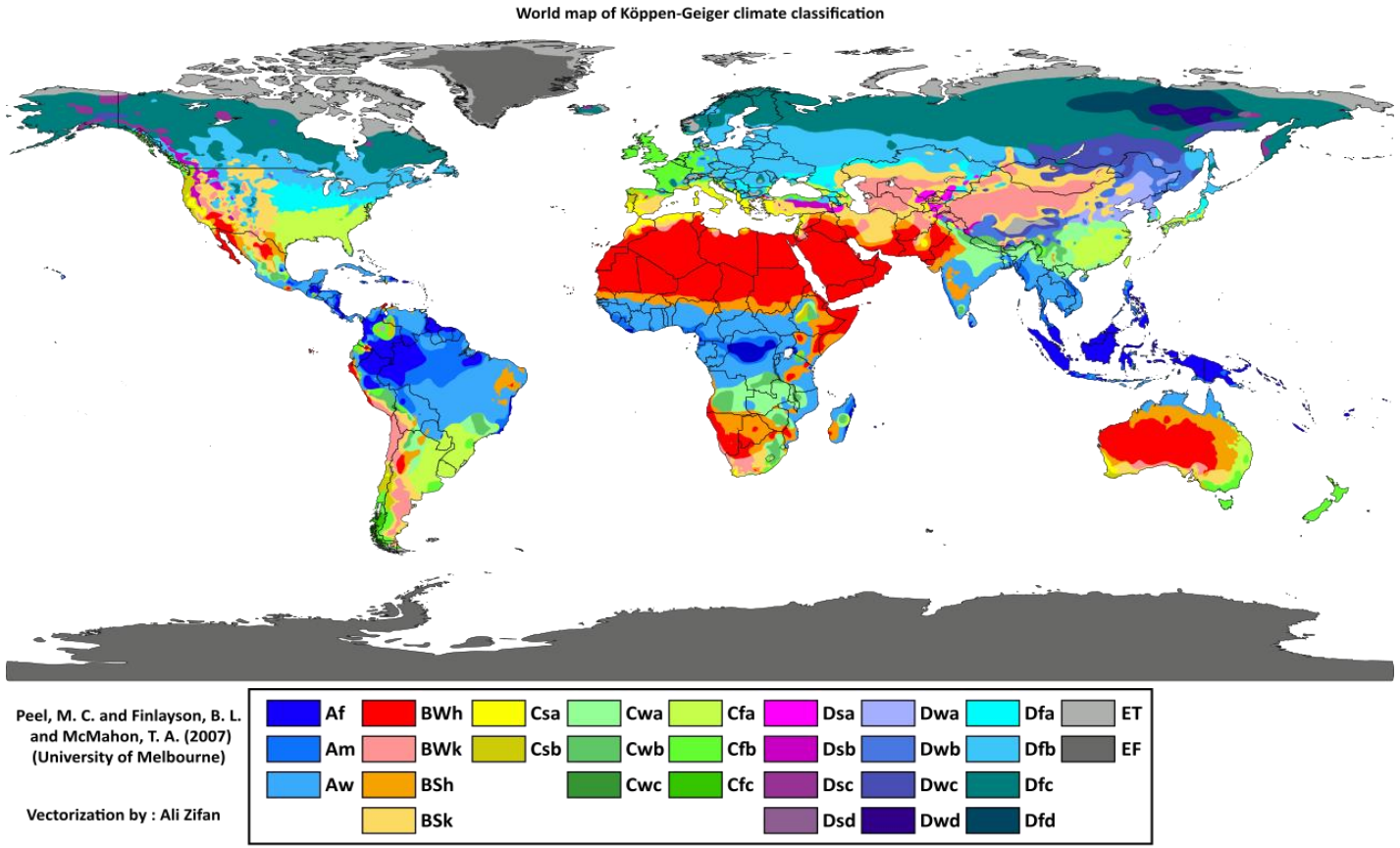
- More robust ROC determination
  If data is messy or contains outlier months
- Especially useful in areas with harsh winters
  Lots of snow and cloud cover
- YbY analysis gives a much narrower and reasonable ROC
  Shown for 100 inverters in example to right



Ernest Hasselbrink et al., "Validation of the PVLife Model Using 3 Million Module-Years of Live Site Data," in IEEE 39th Photovoltaic Specialists Conference (PVSC), 2013, pp. 0007–0012.

Köppen-Geiger Climatic Zones

Type = Humidity/Wetness
Subtype = Temperature/Temperature Range



Group	Type	SubType	Description	Criterion
A-Tropical $T_{min} \geq +18℃$	f		Rainforest	Pmin ≥ 60mm
	m		Monsoon	Pann≥ 25(100-Pmm) mm
	w		Savanna	Pmin < 60 mm in winter
B-Arid $P_{ann} < 10\ P_{th}$	w		Desert	Pann ≤ 5 Pth
	s		Steppe	Pann > 5 Pth
		h	Hot	$T_{ann} \geq +18$ °C
		k	Cold	$T_{ann} < +18$ °C
C-Temperate -3 °C < T_{min} < +18 °C	s		Dry Summer	
	w		Dry Winter	$P_{smax} > 10\ P_{wmin}$, $P_{wmin} < P_{smin}$
	f		Without dry season	Not Cs or Cw
		a	Hot Summer	$T_{max} \geq +22$ °C
		b	Warm Summer	$T_{max} < +22$ °C, 4 $T_{mon} \geq +10$ °C
		c	Cold Summer	$T_{max} < +22$ °C, 4 $T_{mon} < +10$ °C, $T_{min} > -38$ °C
D-Cold(Continental) $T_{min} \leq -3$ °C	s		Dry Summer	$P_{smin} < P_{wmin}$, $P_{wmax} > 3\ P_{smin}$, $P_{smin} < 40$ mm
	w		Dry Winter	$P_{smax} > 10\ P_{wmin}$, $P_{wmin} < P_{smin}$
	f		Without dry season	Not Ds or Dw
		a	Hot Summer	$T_{max} \geq +22$ °C
		b	Warm Summer	$T_{max} < +22$ °C, 4 $T_{mon} \geq +10$ °C
		c	Cold Summer	$T_{max} < +22$ °C, 4 $T_{mon} < +10$ °C, $T_{min} > -38$ °C
		d	Very cold Winter	$T_{max} < +22$ °C, 4 $T_{mon} < +10$ °C, $T_{min} \leq -38$ °C
E-Polar $T_{max} < +10$ °C	T		Tundra	$T_{max} \geq 0$ °C
	F		Frost(Ice cap)	$T_{max} < 0$ °C

Generated based on precipitation and temperature

- Begun in 1884, further classified 1954
- Consistent and comprehensive climatic zones

29 total K-G Climatic Zones defined

- Understand environmental stressors

KGC R package

Published on CRAN

Peel, M. C.; Finlayson, B. L.; McMahon, T. A. (2007). "Updated world map of the Köppen–Geiger climate classification".Hydrol. Earth Syst. Sci. 11: 1633–1644. doi:10.5194/hess-11-1633-2007. ISSN 1027-5606

Month-by-Month ROC Method

Rate of Change: ROC

Underlying assumption:

- Train an un-biased regression model
- System performance change is a long-term phenomena
  No obvious degradation within 30 days

Data analytic procedure

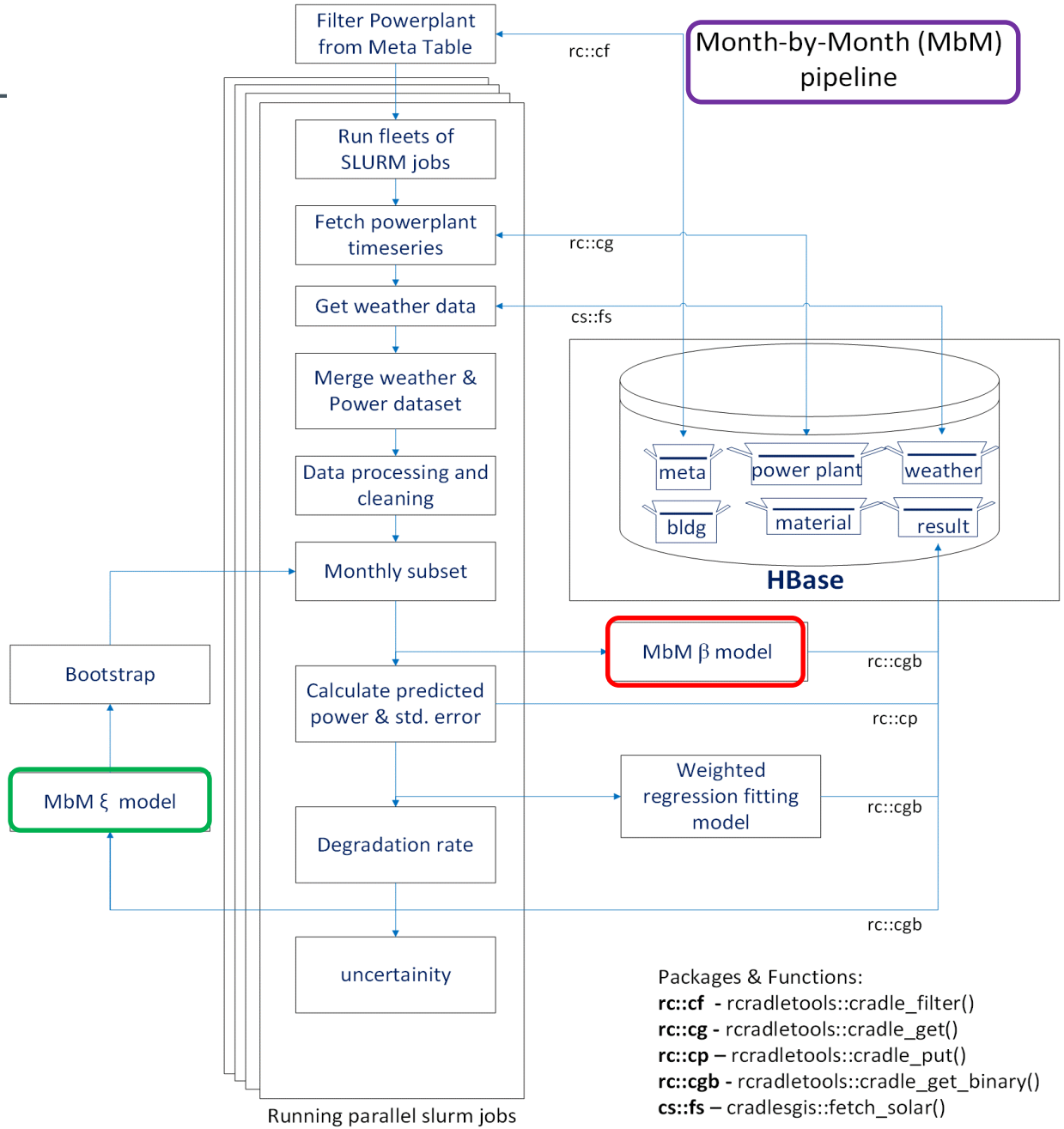
- Use all data, not excessive filtering
- Categorize data by age
  Every 30 days considered a pseudo-month

Psuedo-Month Predictive model (β model)

- Use monthly regression models
- A snapshot of the system status
- Predict system performance each Month
  To same climate condition

Longitudinal Regression Model (ξ model)

- Don't assume linear degradation rate
  Enable Piece-wise Regression Models of Change Rate
- Use bootstrap approach to estimate the uncertainty



Power Plant m4jmg2n: 15 years, BSk Arid-Steppe-Cold

Each data point is predicted output in that month Normalized to a standard environmental condition, Error bars shows predictive error.

ROC = 0.67%/year

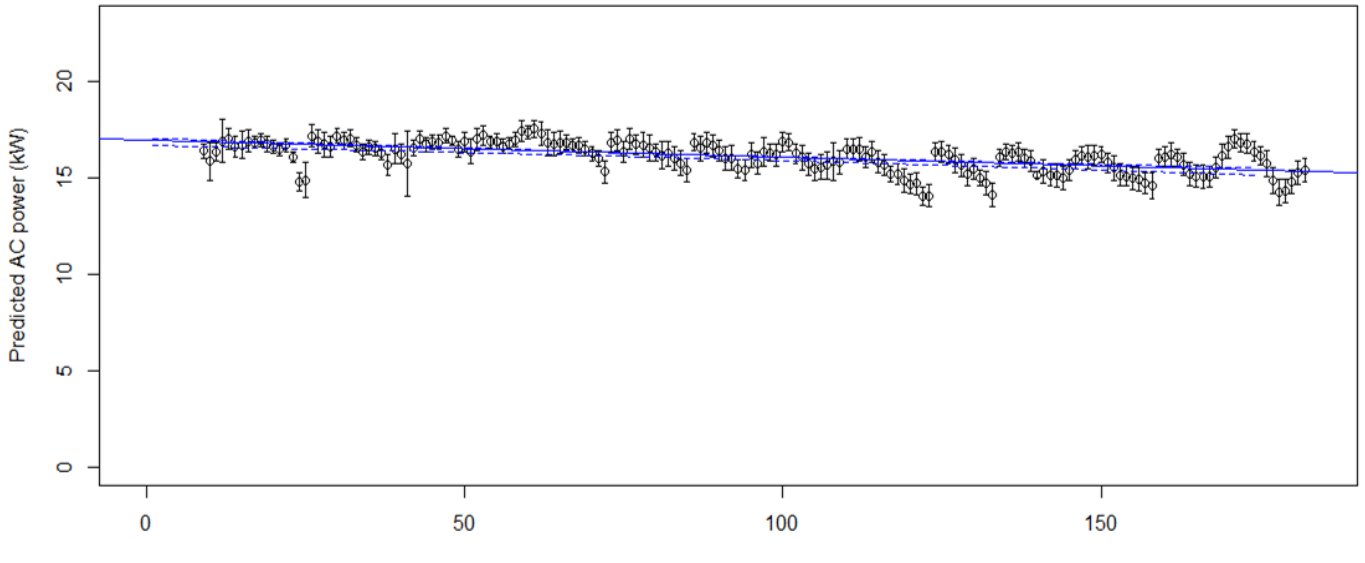
(statistically significant)

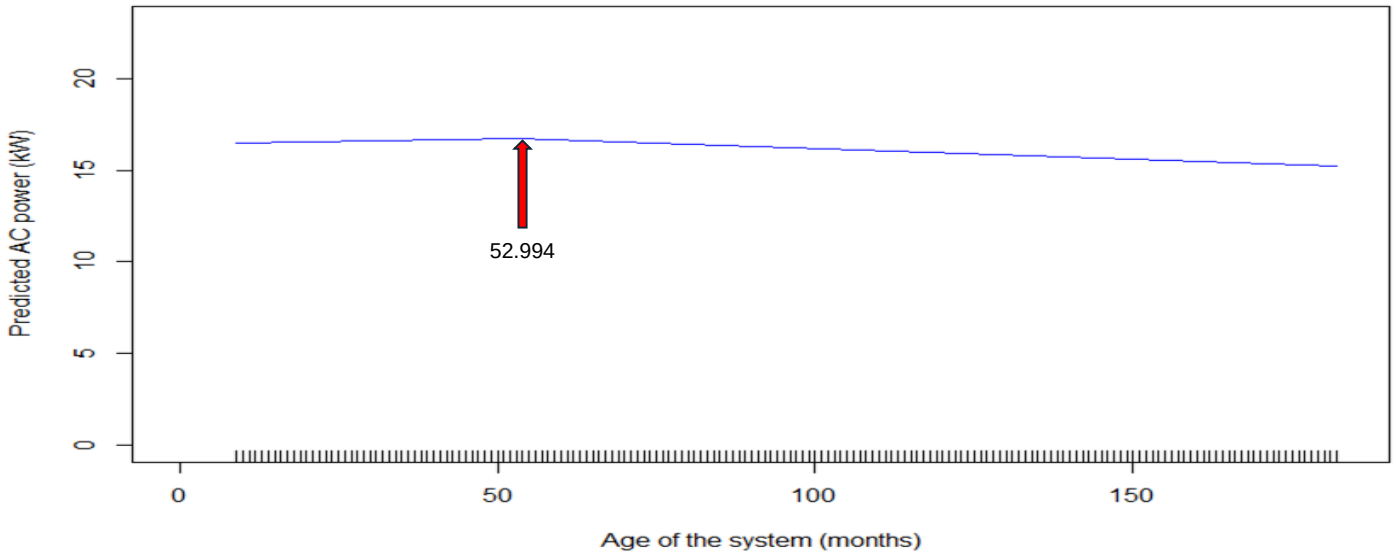
- Predicted MbM value exhibits seasonality
- Seasonality get stronger
  after 50 months

Change Point: 53rd month

- 4 years and 5 months
- "Segmented" change-point R package

Horizontal Irradiance 850 w/m^2 Ambient Temperature 25 C Windspeed 1m/s

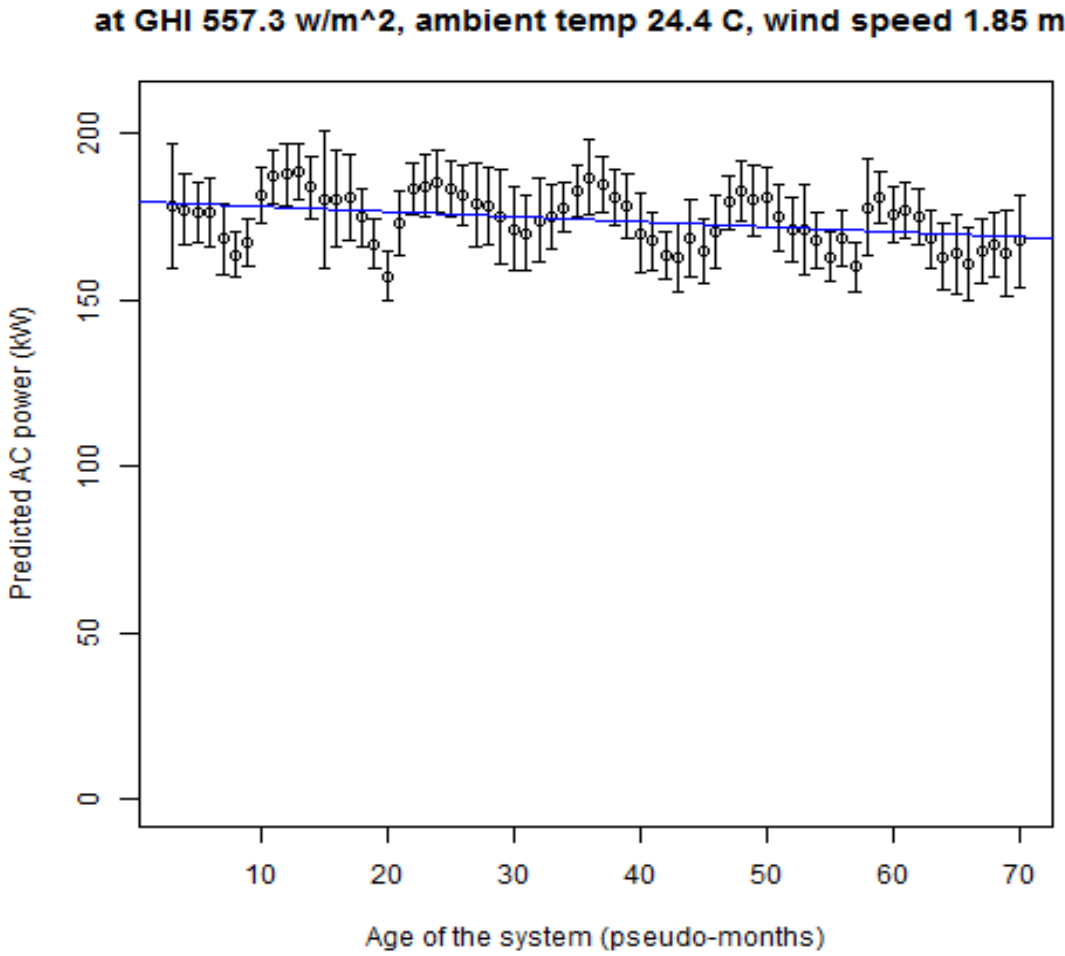


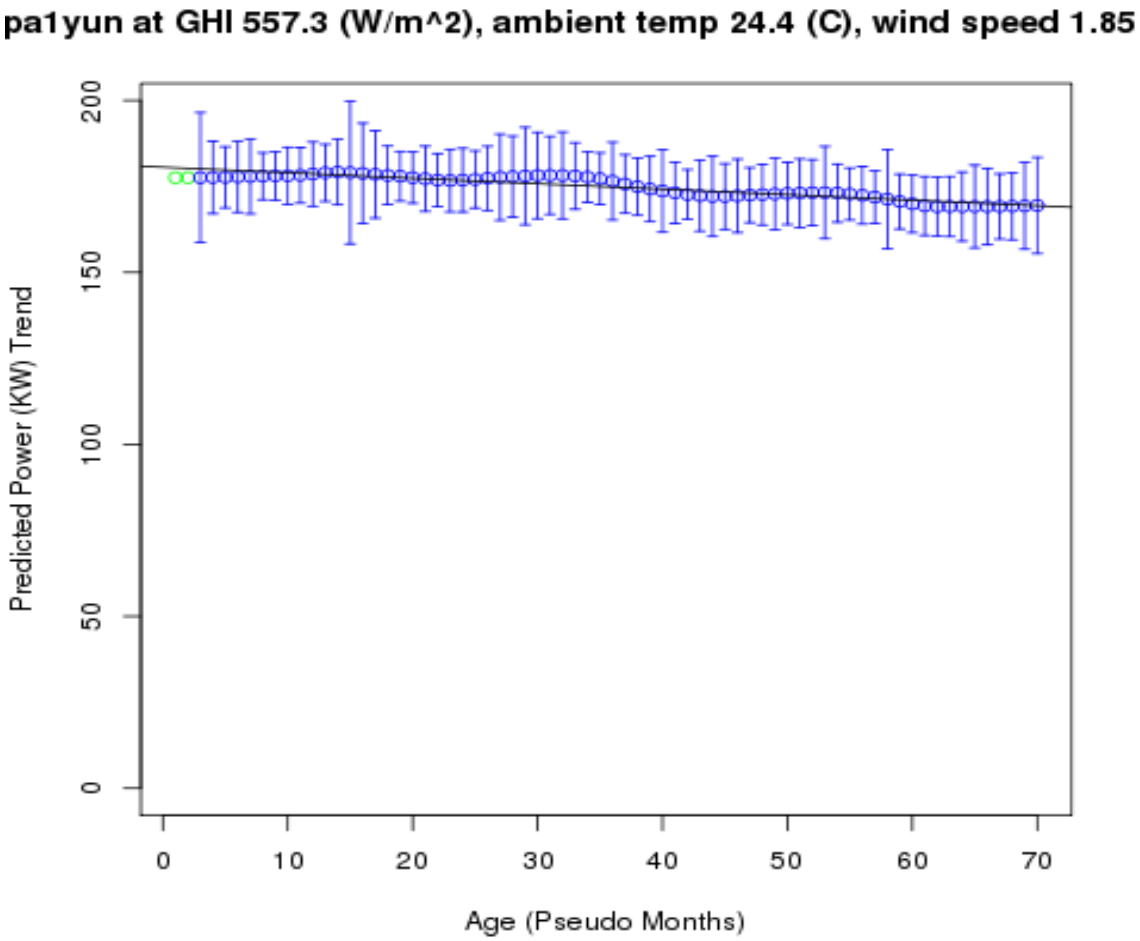


Seasonality

Strong seasonality can be seen in the trends

Time-series classical decomposition can remove this





Snow Detection

Snow coverage can cause anomalies

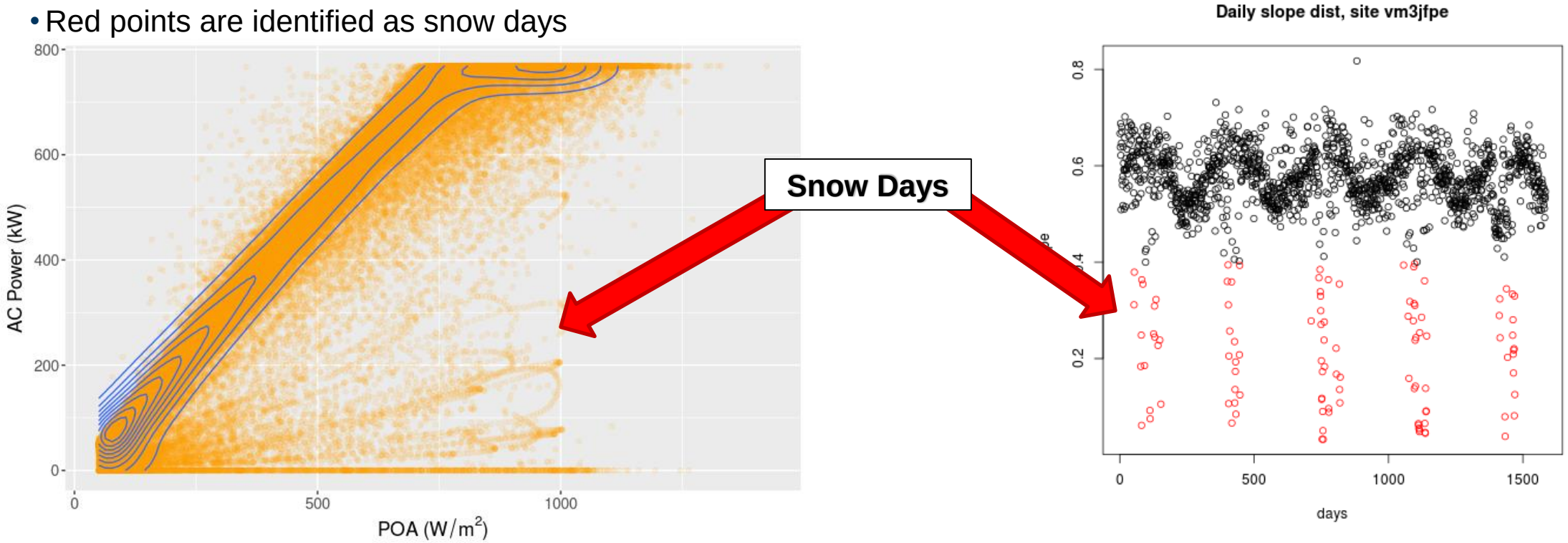
- By preventing power output but not affecting irradiance detection
- Logs are sometimes available that identify snow days, these are stored as metadata

Snow is tracked by looking at the slope between power and irradiance for every day

- The slope distribution is not normal,
- As snow strongly negatively affects the slope

K means clustering with 2 clusters can be used to separate most of the snow days

- Red points are identified as snow days



Clear Sky Identification

Clear sky correction reduces noise in data

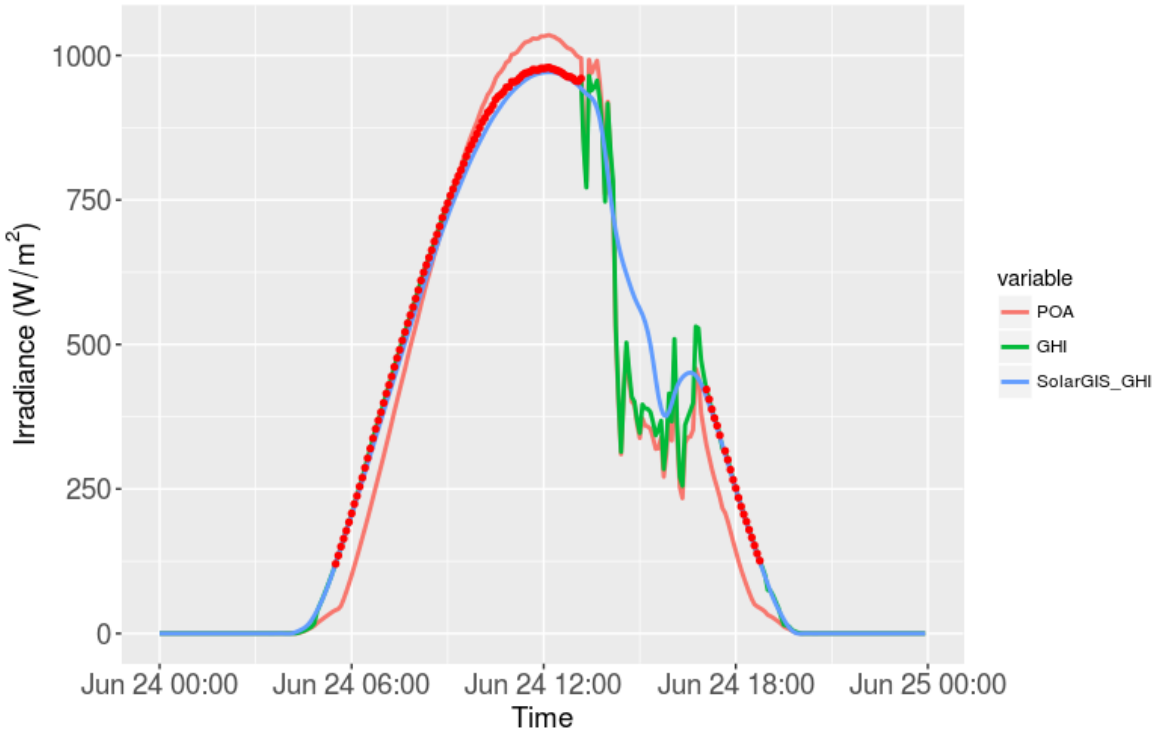
- And is robust against sensor drifting
- SolarGIS weather data does not drift over time
  as a ground sensor might

Clear sky points are detected using PVlib-Python[1]

- Clear sky points shown with red dots on the plot
- Clear sky points show less noise
- Sensor GHI can be replaced with SolarGIS GHI
  to prevent the influence of sensor drifting

SolarGIS data is automatically queried and stored

- In the Hbase weather table
- For a given latitude, longitude, and time interval
- Allows for easy integration into MbM pipeline



1. W. F. Holmgren, R. W. Andrews, A. T. Lorenzo, and J. S. Stein, “PVLib-Python 2015,” in 2015 IEEE 42nd PVSC, 2015, pp. 1–5.

Interpreting ROC Results

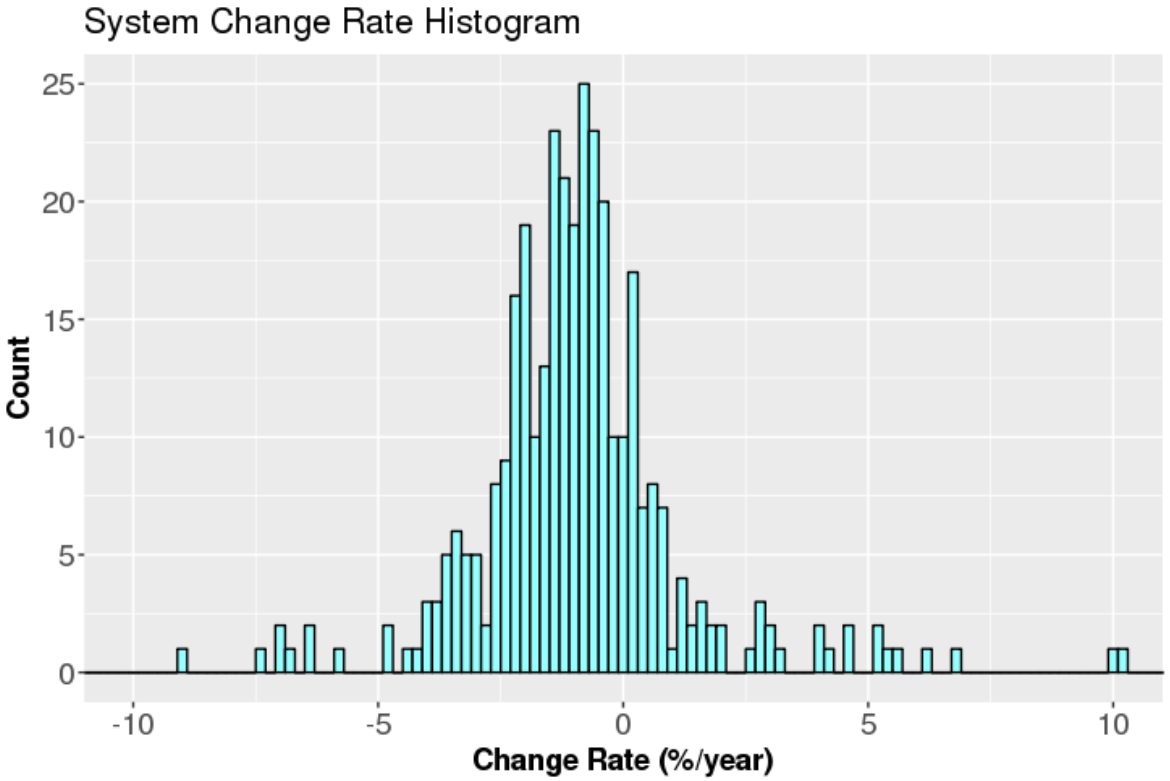
Pipelining allows the automatic analysis

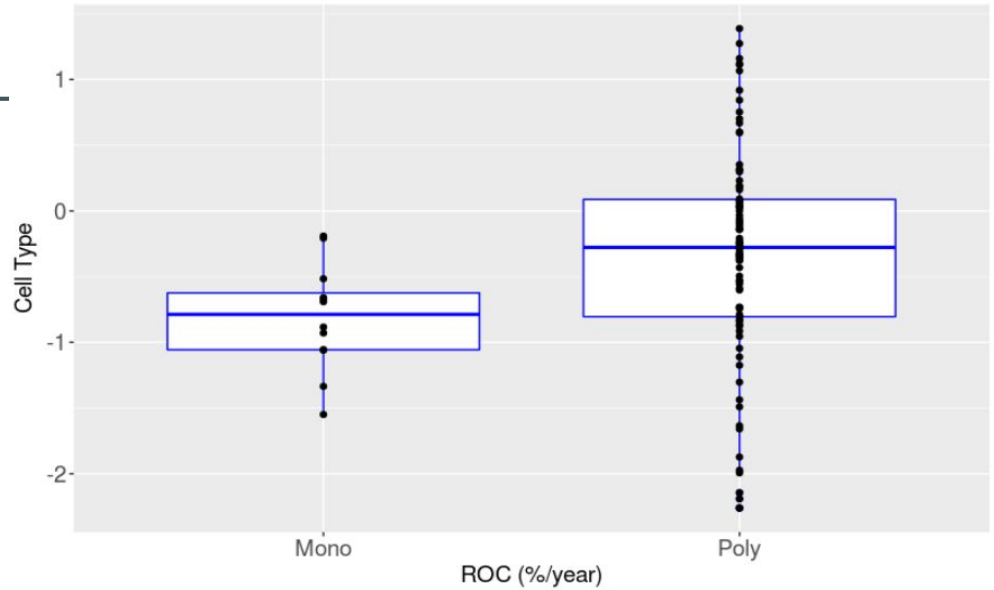
- Of massive numbers of PV systems
- Basis for statistically significant findings

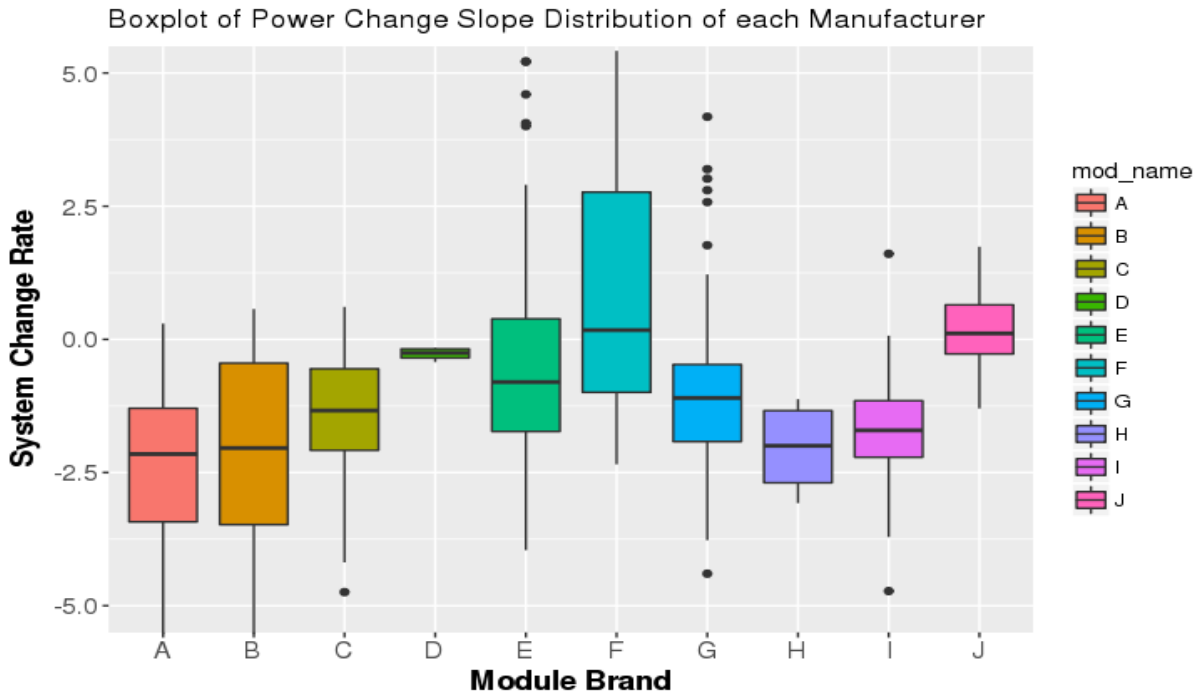
  Instead of observational reports

Rate of Change distributions can be used

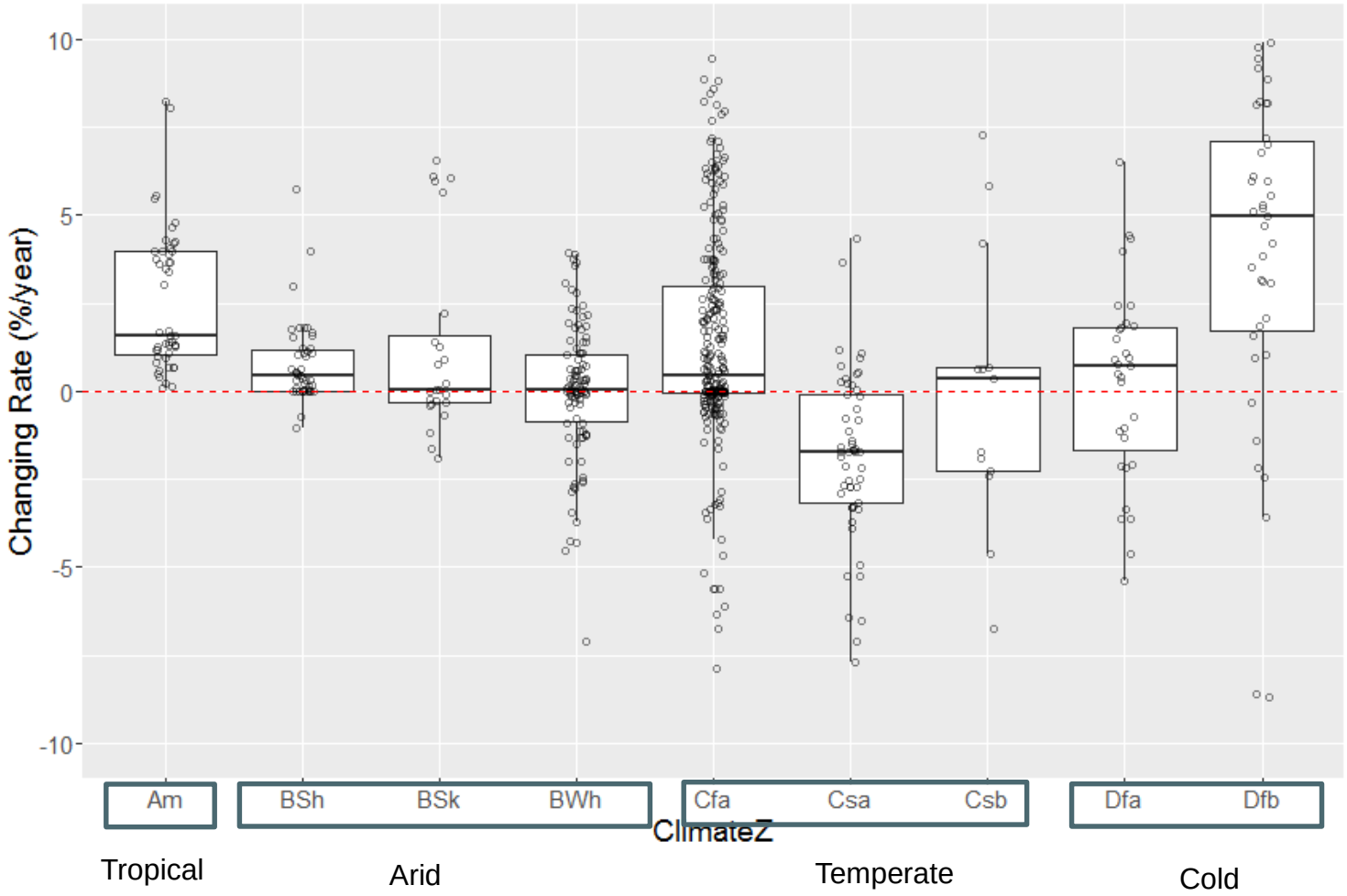
- To identify performance variations as function the predictors
- Such as module brand, climate zone, or cell type







Analysis of Variance Across 9 Climate Zones



Mean of ROC in each Climate Zone show large variance

- Dfb mean is over +5%/year
- Csa mean is about -2%/year

There may be confounding variables That influence the change rate

Develop statistical models To help solve the problem.

Energy & Materials Data Science: Encompassing Broader Opportunities

Where we started: Lifetime and Degradation Science

- Focusing on PV Modules Degradation Over 25 year
  - Now Shifting Focus to 50 years
  - And To High Efficiency c-Si PERC Modules

Expanding Across All Data Types

- Time Series Analysis of Power/Energy Data: Power Plants, Building Energy Efficiency
- Spectral Analysis: Materials Degradation and Mechanistic Identification
- Image Processing: Electroluminescent, Thermographic, Optical, Video Images

Expanding Beyond Time-series Analysis and Network Modeling

- Machine Learning
- Ensemble Modeling
- Deep Learning
- From NoSQL Databases, to NoSQL Document Databases

Expanding Beyond Long Term Degradation, Into Data-driven Analysis & Modeling

- Solar Irradiance Forecasting
- Research and Data Text Mining
- Information Security and CyberSecurity (VerisDB)

CASE SCHOOL
OF ENGINEERING

CASE WESTERN RESERVE
UNIVERSITY

Abstract

As solar power becomes a larger source of electricity and power for locations, it becomes increasingly important to fully understand and predict the power output of solar modules over their entire lifetime. Traditional solar module degradation tests are done under accelerated exposure environments, where the conditions are more aggressive than an outdoor environment, with the intent of testing the lifetime performance of a module within a more reasonable time scale. While these tests are certainly important, they can be either over or inadequately aggressive; therefore it is also critical to monitor real-world, outdoor power plants degrading under actual real-world exposure conditions. A combination of the two methods provides the best rate of change (ROC) or lifetime performance prediction of PV power plants, with indoor exposures degrading modules in a shorter time span, and outdoor modeling giving insight into the actual degradation patterns of systems and providing a comparison, by cross-correlation, of accelerated and real-world degradation.

With this in mind, the SDLE Research Center is developing data-driven modeling of ROC for PV systems based on a massive collection of time series data from numerous PV systems, both research and commercially fielded, including a variety of ages, brands, module types, and climate zones. To analyze and manage data from diverse PV plants, we have developed Energy- CRADLE, an automated data acquisition, management and analytics pipeline. The Energy- CRADLE is built in a high performance computing (HPC) environment which leverages distributed computing features of HBase/Hadoop and Spark cluster for distributed storage and parallel computing. We have also developed R and Python packages for integrating with HBase tables. For cross-sectional study of running on 100s of PV systems, we use fleets of parallel jobs via the SLURM workload manager.

While commercially fielded PV power plant data sources may be of a lower quality than research focused PV sites, being able to use data from commercial plants greatly increases the length of time series datasets available for analysis, making it a unique, at-scale resource for cross-sectional studies of thousands of PV systems. This large scale data collection is used to determine what the degradation patterns of real world systems are as a function of location, climatic zone, PV module and inverter brands and what factors might affect the behavior of PV modules over time. The current scope of the data available includes thousands of PV system inverters located across hundreds of sites with power capacities from single modules to hundreds of megawatt plants, located across many different climate zones.

Given the large scale, heterogeneity and diversity of the data between the PV systems, a method had to be developed to determine the rate of change, or the rate at which the power output changes over time, for each PV system consisting of PV modules and their inverter. As this data comes from many sources, there are inconsistencies between datasets, such as different available variables, data quality, or the data capture interval, that the method had to be able to accommodate. The Month-by-Month (MbM) method was developed at the SDLE Research Center with these problems in mind, being able to handle various intervals of data, as well as different variables, the most common of which being different irradiance measurements. The MbM method consists of three models, the **β** Pseudo-month Predictive Model divides the data into 30 day long “pseudo-months” where it is assumed that negligible degradation occurs over the 30 day time period. A multiple linear regression model is built for each pseudo month based on the given environmental variables, such as irradiance, temperature, and wind speed. Once a model has been built for each month, representative weather conditions are determined for the given PV system which are the average temperature, the average wind speed, and the minimum value of all the peak irradiances for each pseudo-month. The representative weather conditions are applied to each **β** model and predicted power outputs for each month are determined. Once the predicted power for each month has been determined, the **ξ** Piecewise Regression Model uses a weighted regression to calculate the rate of change of the system (%/year) from the slope and y-intercept of the predicted power over time. The **ξ** model is weighted to the standard errors for each **β** model, improving the robustness of the method by reducing the influence or noisy or less precise pseudo-months. Once the rate of change for each system is determined, a **γ** Cross-Sectional model of the rate of change as a function of the metadata for the PV systems, such as module brand or climate zone, providing insights into the factors causing more or less severe power loss in these outdoor PV systems.

Seasonal decomposition is used to reduce the impact of seasonality on the calculated rate of change. Fluctuations in power can be seen as a yearly cycle with the seasons, potentially influencing the calculated rate of change. Performing time series seasonal decomposition to isolate the seasonal and trend components of the power time series so as to reduce the influence of seasonality on the ROC results. Clear sky identification is the latest addition to the MbM analysis pipeline. Modeled weather data, derived from satellite imagery combined with an empirical atmospheric model, is pulled from SolarGIS for each PV system as supplemental weather data. By comparing the modeled weather conditions from SolarGIS and the measured weather conditions from the system, the clear sky, or points at which there was no cloud cover, can be identified. This identification is done using the PVLib-Python open source library. Clear sky identification has many benefits. Isolating clear sky points can reduce the noise of the data and ensure that the conditions are similar between two given points. Most importantly, however, is it can be used to track sensor drifting which can be highly problematic in long term time series. The SolarGIS data can also be used as a supplement if sensor drifting is observed, as the SolarGIS data will not drift over time.