A Novel Machine Learning Framework for Anomaly Detection and Data Quality Assessment

SHAOCHENG XIE AND XIAO CHEN

LAWRENCE LIVERMORE NATIONAL LABORATORY

Cloud Processes Research and Modeling/Lawrence Livermore National Laboratory

2018 Joint ARM/ASR User Facility and PI Meeting, Vienna, VA. March 19 – 23, 2018

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

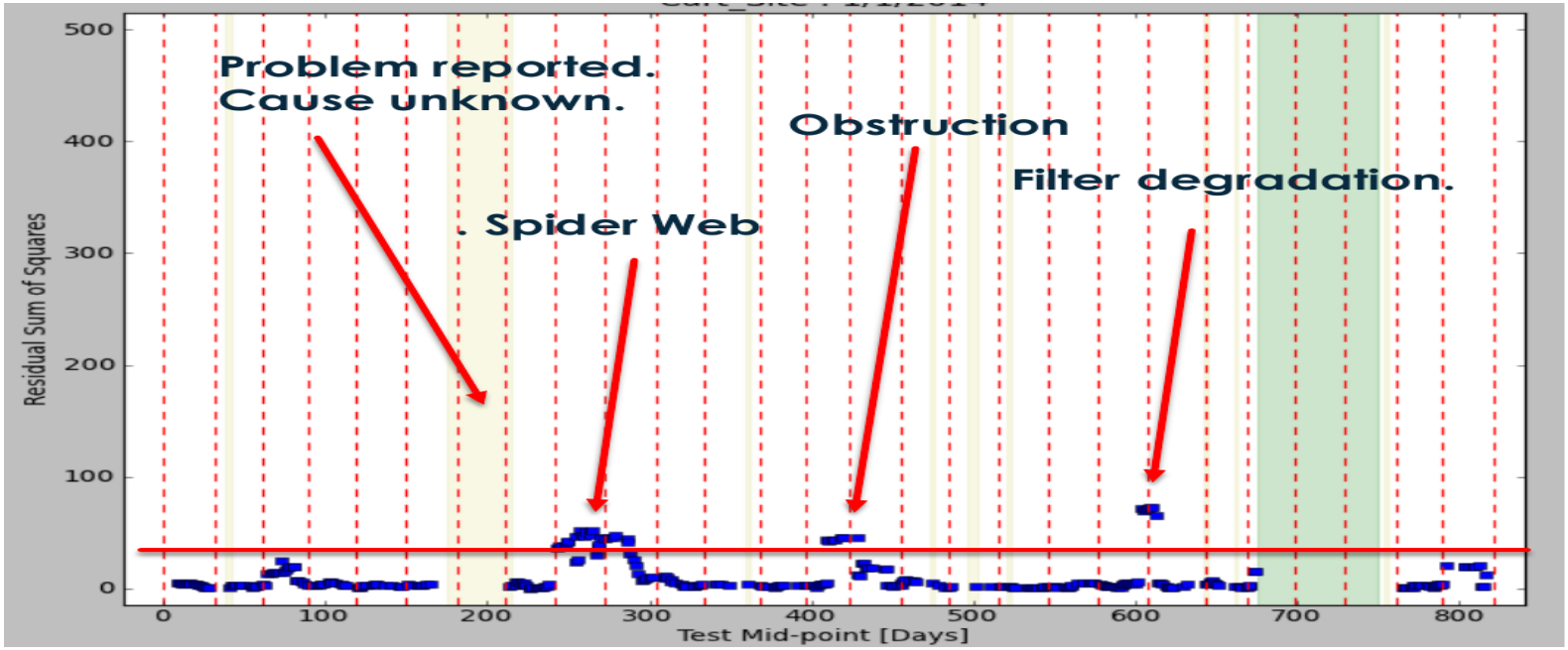
Goals

- Develop a novel Machine Learning (ML) framework to improve current ML algorithms used in ARM
- Effectively identify ARM data anomalies and automate data quality assessment

Issues with the Current ARM ML Model – ADMLA

ARM has recently used the **Anomaly Detection ML Algorithm (ADMLA)** to address data quality issues in measurements made by CSPHOT, MFRSR, and AOS.

Issues with CSPHOT

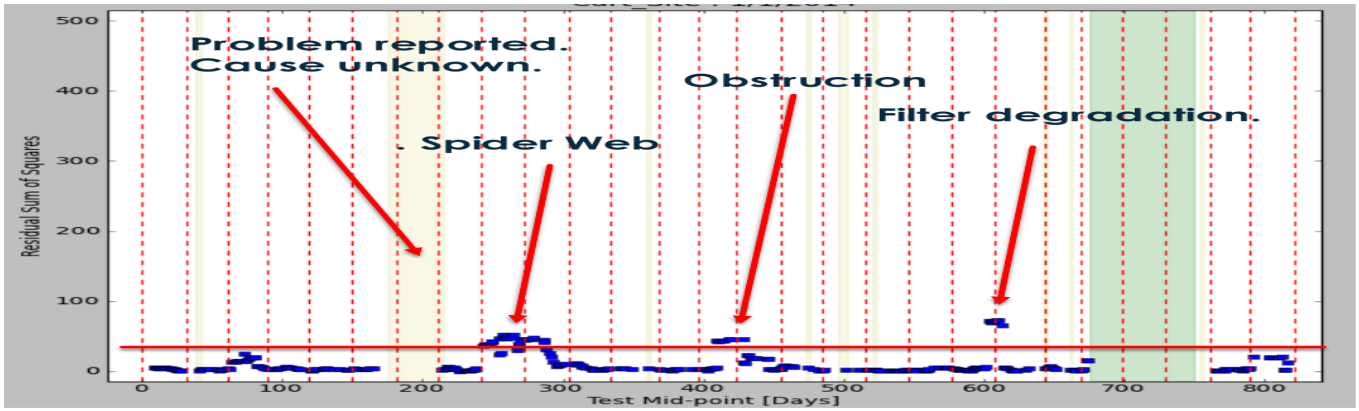


Mitchell, Gregory et al. (2016)

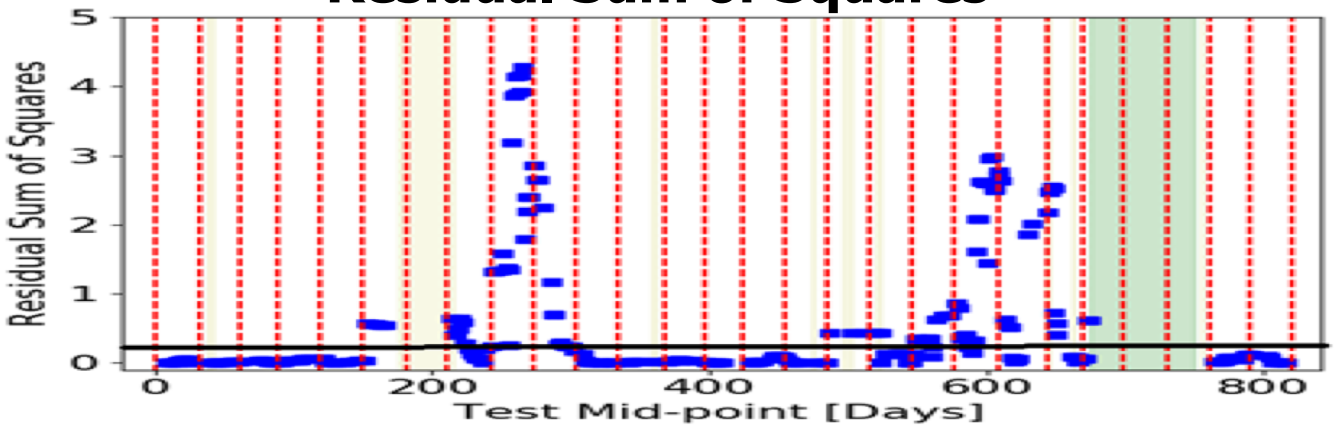
CSPHOT: the Cimel Sunphotometer (Solar irrdiance and sky radiance)
MFRSR: Multi-filter Rotating Shadowband Radiometer
AOS: the Aerosol Observing System (Aerosol optical properties)

Issues with the Current ARM ML Model - ADMLA

Issues with CSPHOT



Residual Sum of Squares



ADMLA – Anomaly Detection ML Algorithm

Many false positive

Many false positive caused by fitting errors due to large number of features being selected (57) and the small number of training samples

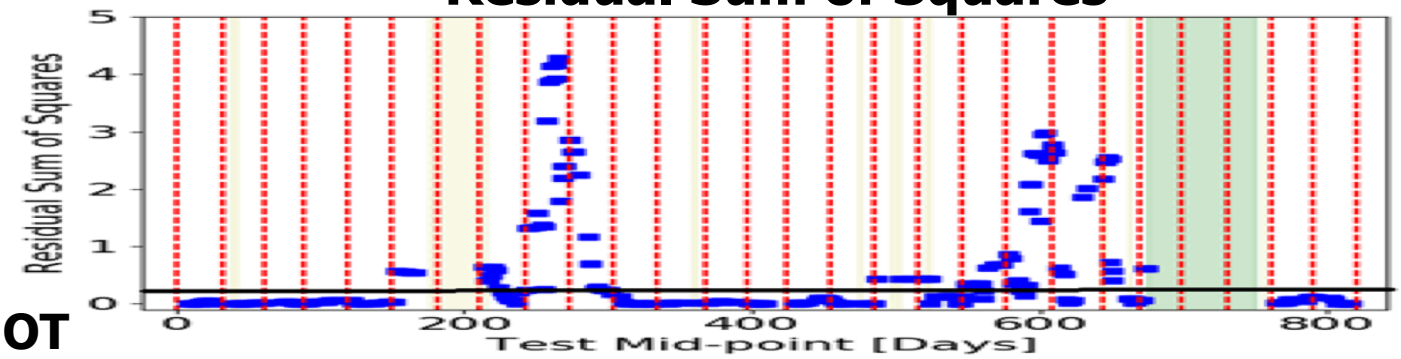
Improved ML Algorithms - ARD

ARD – Automatic Relevance Determination

- identifying the relevant features using a Bayesian feature selection – used for reduce ML fitting errors

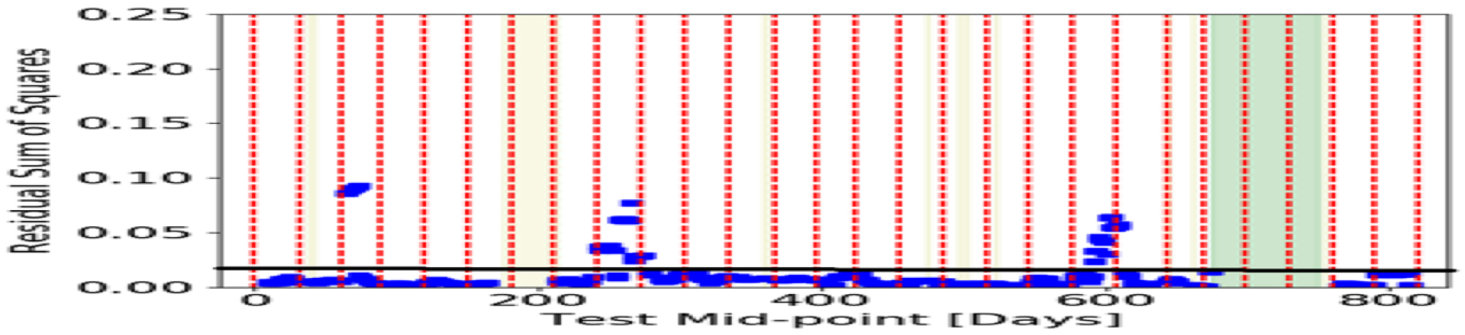
Residual Sum of Squares

ADMLA



CSPHOT

ARD

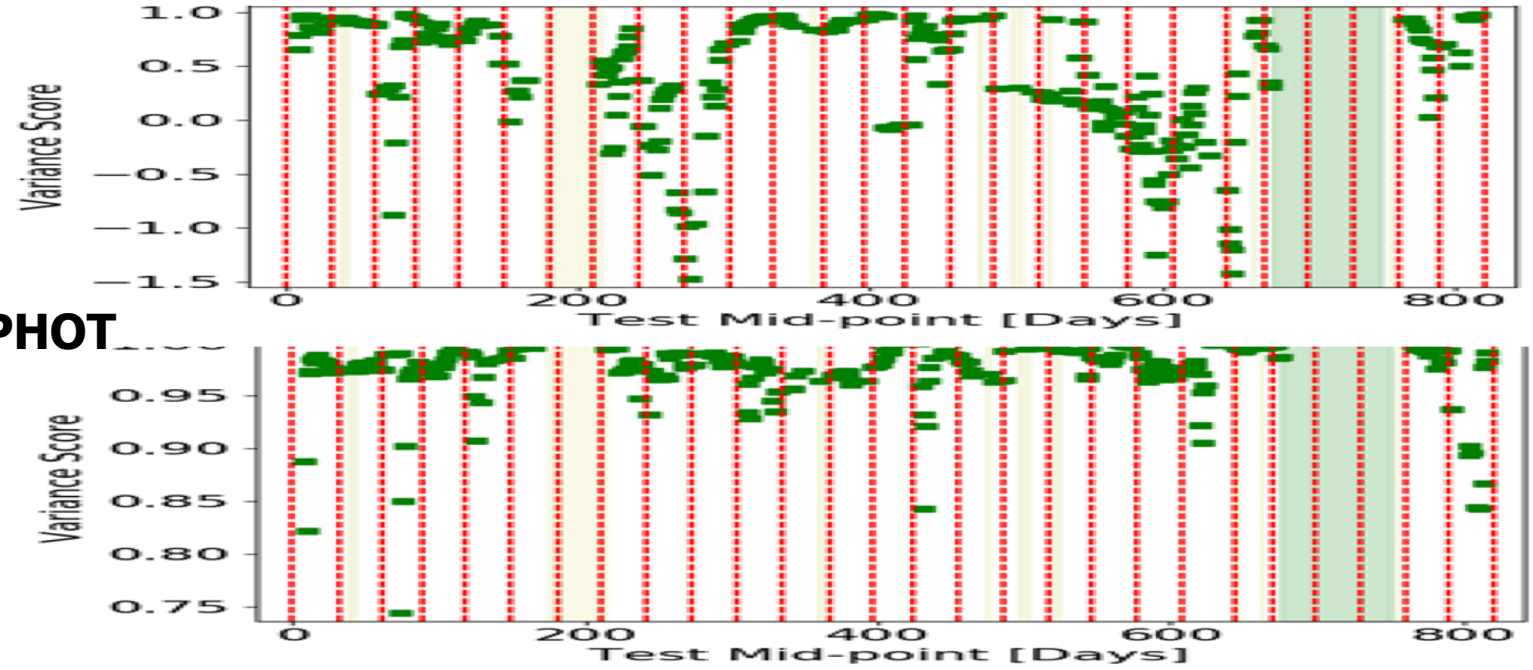


ARD: Less false positives for known problems compared to traditional ARM ML models

Improved ML Algorithm - ARD

ARD – Automatic Relevance Determination

Variance Score



DMLA

CSPHOT

ARD

ARD: Majority of days are close to **one** using ARD ML model: good prediction accuracy of the estimate using CSPHOT_ARD

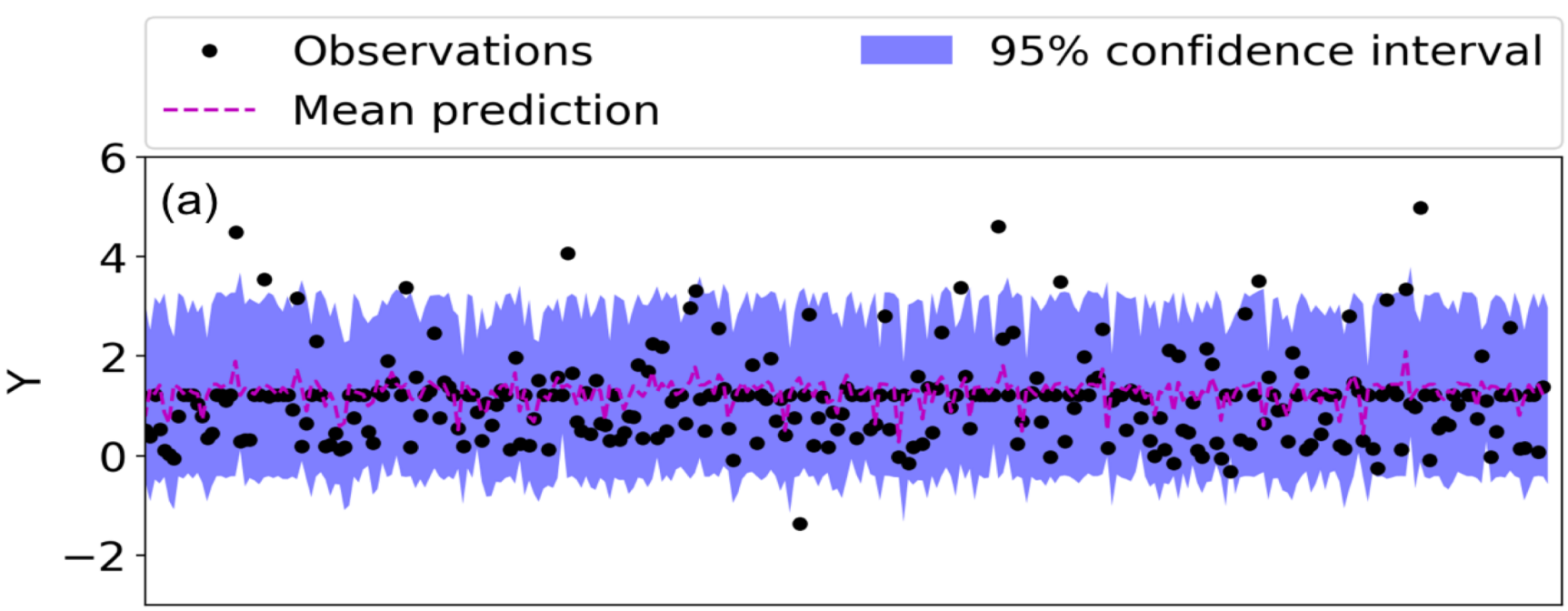
Estimate Data Error for the Period between the Observational Data Points - GPR

ARM CLIMATE RESEARCH FACILITY

GPR: Gaussian Process Regression

- Automatically provide some error bar
- No need for an arbitrary threshold

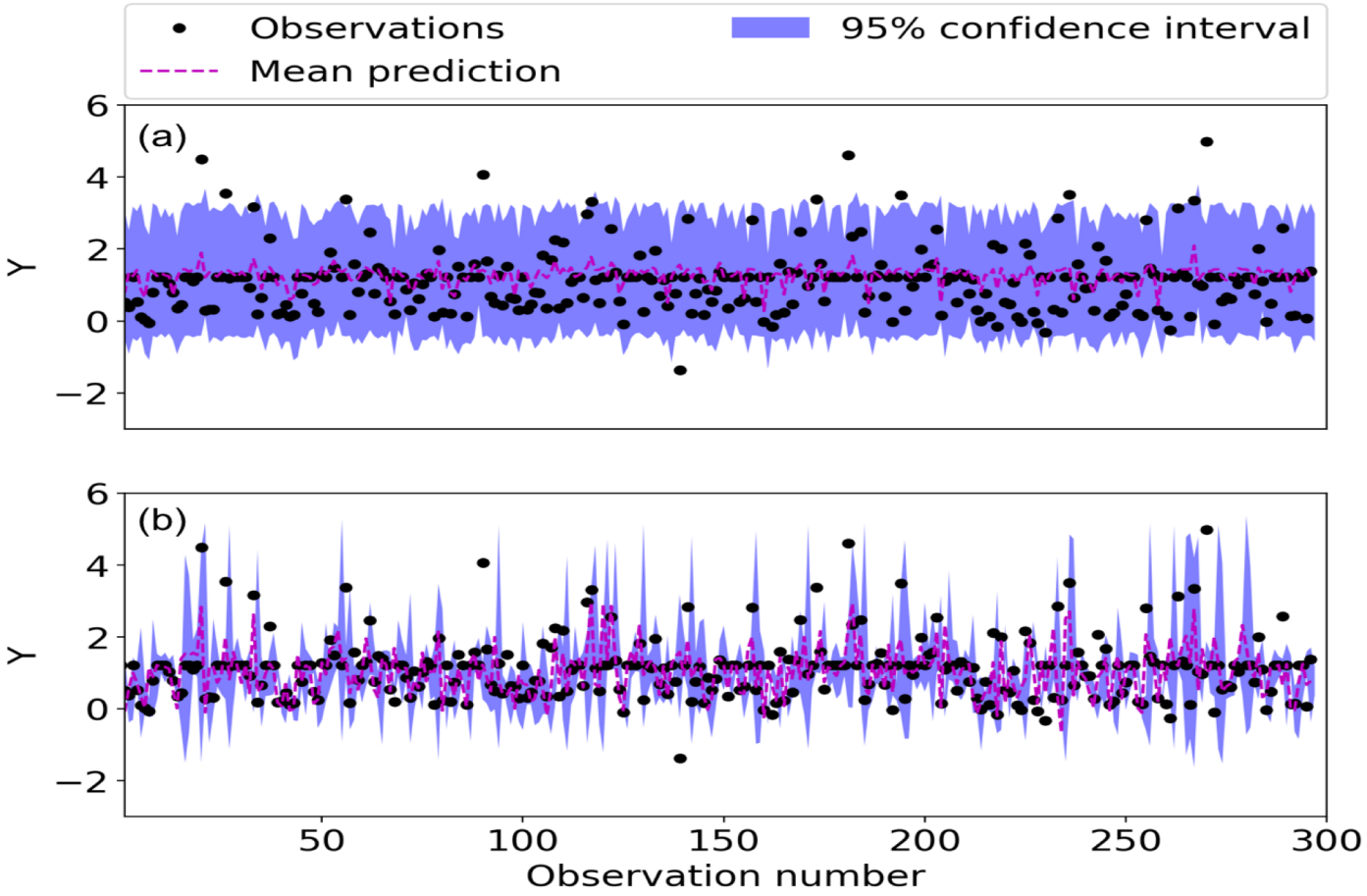
Error bar from GPR based on 57 dimensional feature spaces that the current CSPHOT-ADMLA is built on



Enhance GPR ML with DASSI

- DASSI – Data Assimilation for Stochastic Source Inversion (developed through a LLNL LDRD project led by Co-PI Xiao Chen)
  - Nonlinear dimension reduction on the ML feature vector
  - Additional training-data generation through advanced ML techniques for the reduced order feature space to enable more accurate ML based data quality control
    - To address the fundamental challenge encountered in the application of any kind of ML algorithms to ARM data, that is, the sample size of training dataset

Enhance GPR ML with DASSI



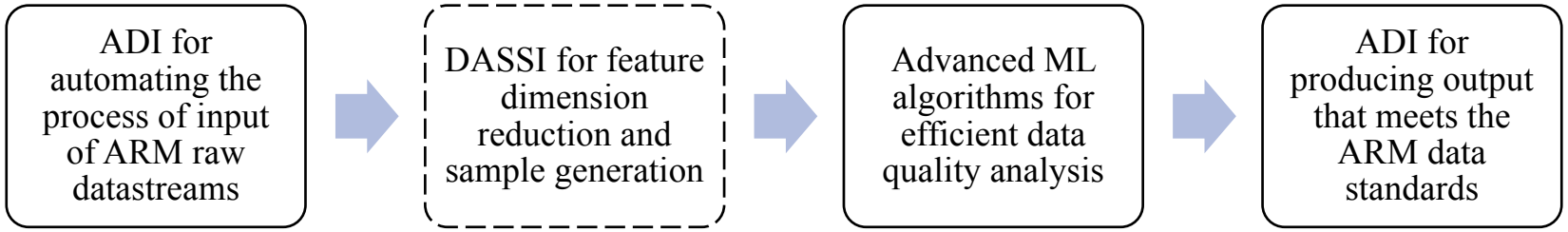
GPR: Gaussian Process Regression

57 dimensional feature spaces

GPR ML model enhanced by DASSI : much narrower error bar based on the reduced dimensional feature spaces (17)

Y: a quantity of interest that will be predicted on the test data set based on GPR ML algorithm

Proposed A Novel Machine Learning Framework



Our framework

- ARM Data Integrator (ADI) for automation of input ARM raw datastreams
- DASSI for feature dimension reduction and sample generation
- Advanced ML algorithms for efficient data quality analysis based on the low-dimensional reduced-order feature space obtained from DASSI
- ADI for producing output that meets ARM data standards
- User-friendly interface that allows users to implement with specific needs

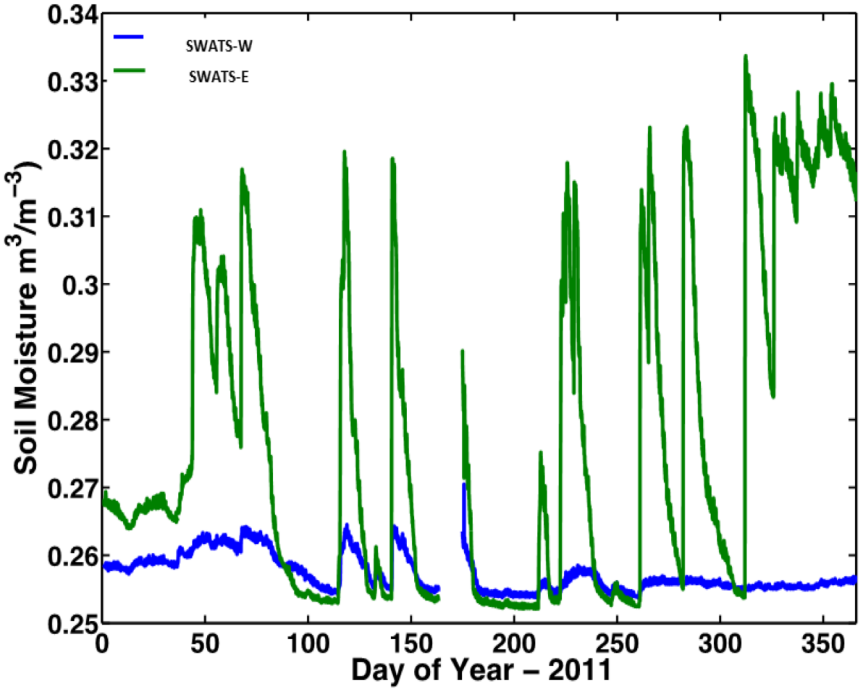
Explore Three Advanced Machine Learning Algorithms

ARM CLIMATE RESEARCH FACILITY

- Automatic relevance determination (**ARD**): identifying the relevant features using a Bayesian feature selection – used for reduce ML fitting errors
- Gaussian process regression (**GPR**):
  - Assume Gaussian distribution of the training samples and build a probabilistic ML
  - Automatically produce error bar for the neighboring data points based on the previous ARD data error estimation from the observational data points.
  - Efficient data quality analysis based on the low-dimensional reduced-order feature space obtained from DASSI
- Autoregressive Integrated Moving Average (**ARIMA**): removes the seasonal and trend components

Application to MWR, SWATS, and STAMP

- **MWR**: **Microwave Radiometers**
  - **Problem**: the presence of water on the MWR Radome – one of the most difficult QC issues the ARM is facing.
- **SWATS: Soil Water and Temperature Systems**
  - **Problem**: Failed sensors (primarily due to aging) and the lack of sensitivity to low soil moisture.
- **STAMP: Soil Temperature and Moisture Profiles**
  - **Problem**: sensor failure as a result of the intrusion of lightning energy



Soil moisture measured by two co-located SWATS sensors apart 1 meter (SWATS-W and SWATS-E), demonstrating that the west sensor was not working properly

A Workflow to Address Water Contamination Issues with MWR

ADI for reading the *Dual Microwave Radiometer Experiment (DMRE) field campaign* training data (11-month long dataset (1/1/2016-12/1/2016)

DASSI for reducing number of features and generate more additional free samples to improve prediction reliabilities and accuracies

ARIMA ML model removes the seasonality by computing the difference of the datastream and applying regression on the deseasonalized data

ARD ML model provides less false positives and better variances by Bayesian feature selection

GPR ML model incorporate measurement errors at the test data points and produces probabilistic estimates of the predication at neighboring points

THE END