SLAC-R-1132

Visualization and Analytics of Distribution Systems with Deep Penetration of Distributed Energy Resources (VADER)

SLAC National Accelerator Laboratory, Stanford University, Stanford, CA 94309

This material is based upon work supported by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy (EERE) under the SunLamp Award Number 31003.

Final Technical Report (FTR)

a. Federal Agency	Department of Energy	
b. Award Number	31003	
c. Project Title	Visualization and Analytics of Distribution Systems with Deep Penetration of Distributed Energy Resources (VADER)	
d. Principal Investigator	Laura Schelhas, (Sila Kiliccote) Staff Scientist Email: schelhas@slac.stanford.edu Phone: 650-926-2059	
e. Business Contact	Nani Sarosa Financial Analyst Email: nanis@slac.stanford.edu Phone: 650-926-3668	
f. Submission Date	07/01/2019	
g. DUNS Number	929 824 415	
h. Recipient Organization	SLAC National Accelerator Laboratory	
i. Project Period	**Start:** 02/01/16	**End:** 06/30/19
j. Submitting Official Signature		

Acknowledgement: This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the SunLamp Award Number 31003.

Executive Summary:

In its current state, the distribution system is incapable of handling small to moderate amounts of photovoltaic (PV) penetration. This is because it was initially designed for handling passive loads which, at the level of a substation, have low variability and are forecastable with high accuracy. It has been an open loop system with little monitoring and control. With the addition of PV energy sources, the overall scenario will change dramatically due to (1) two way power flow on network; and (2) high aggregate variability. Additionally, changes on the consumption side lead to a number of smart loads, Electric Vehicles (EVs), and Demand Response.

These fundamental changes in the characteristics of the generation and consumption of power will lead to a number of practical engineering problems which must be overcome to allow increased penetration of Distributed PV. Solving the unique engineering challenges which arise at moderate levels of PV penetration requires closed loop integration of data from (1) PV sources; (2) customer load data from smart meters; (3) EV charging data; and (4) local and line mounted precision instruments.

These data are not traditionally used by utilities in operations since they are ”non-SCADA” and the current grid does not require such levels of control. To integrate this data and provide real time intelligence from these non-SCADA data, we created the Visualization and Analytics of Distribution Systems with Deep Penetration of Distributed Energy Resources (VADER) platform. VADER is a collection of analytics enabled by integration of massive and heterogeneous data streams for granular real-time monitoring, visualization and control of Distributed Energy Resources (DER) in distribution networks. VADER analytics enable utilities to have greater visibility into distributed energy resources. We built several batch- and stream-analytics in VADER which help operators better understand the impact of distributed energy resources on the grid.

Background:

With increasing penetration of DERs and other grid-edge devices, modeling of the distribution grid is becoming a crucial aspect of grid planning and operations. Grids with high penetration of DERs experience issues with voltage violations, stability, and congestion, and often respond with curtailment of these resources.[1] Utilities need more sophisticated modeling tools for their distribution grids to better forecast and control grid-edge resources and avoid curtailment. Other work has designed methods for these controls, notably for using solar inverters to implement volt-var control,[2,3] but these approaches must adopt either simple control laws or very complex agent-based methods to work around the missing grid model.

While utility grid models are often out of date or non-existent at the edge of the grid, more and more utilities are installing advanced metering devices in the distribution grid to capture data on power and voltage levels. The projects in VADER leverage these new streams of data, applying state-of-the-art data analysis and machine learning techniques to build flexible, new, data-informed models.

One main focus of the project is on grid topology and parameter estimation, as discussed further in the Results section of this report. Several studies have addressed this problem but few explicitly account for the impact of measurement errors[4–9], which are often very significant in distribution grid devices. One study which did consider measurement errors took a similar approach of minimizing total least squares, but only used a linear model of the power flow and did not build the statistical model of the method.[6] The work in this project aimed to fill that gap.

Project Objectives:

The major research effort in the VADER project focused on developing high resolution temporal models for EV, PV and loads which are data driven and can be integrated into the VADER system. The integrated data gives researchers an unprecedented opportunity to develop more accurate state estimation techniques. Integration of the new data-driven models of load and distributed energy resources enhances the capability and accuracy of power flow solver for analyzing sensitivities of various key system variables (e.g. distribution feeder bus voltages) with respect to different DER (in particular, PV) penetration levels.

In summary, the VADER project objectives were to:

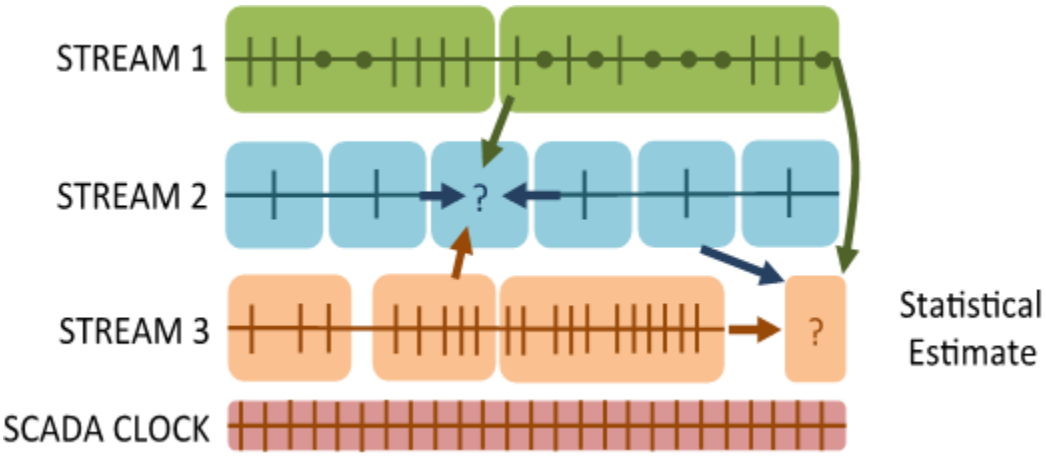
1. Build a set of tools, using data science techniques such as machine learning, to integrate and model a large number of disparate sensor sources to enable distribution system planning and control;
2. Verify the tools utilizing data from industry and utility partners; and
3. Validate the platform in a pilot testbed combining hardware-in-the-loop simulations and real-time data from deployed hardware in the field.

The project was completed over the course of three years. The first year of the project was spent in defining specific problems of value to industry stakeholders, integrating initial datasets, and developing software for streaming data integration. By specifying the problems of interest, we defined the algorithms that needed implementation and test criteria for the VADER system. In the second year we began the joint tasks of VADER platform development effort as well as Distribution System Tomography and What Now Analytics development. Finally, while the first two years dealt with data collection, virtual-SCADA generation, development and testing of VADER platform and DS Tomography pieces, the third and final year of the project contained tasks for integration of the developed analytics with modeling tools and visualization software. Additionally, there was a focus on the algorithm development and testing for What If and Network Analytics. The goal of the DS SEER work was to provide all “what if” functionality using historical data and commercial power flow software with a scenario generation module providing a standard set of planning exercises which can be mapped to changes in topology, sensing, loads, device behavior, etc. The output goal was a modified power flow problem with data-driven solutions providing new insights and filling in various real-world modeling gaps.

Planning and operations of a reliable, stable and efficient distribution system with high PV penetration (>100% of peak load) requires adequate monitoring and accurate prediction capability that allows scenario analysis and closed-loop control of the distribution system. Furthermore, a unified data analytics platform that integrates massive and heterogeneous data streams for planning and granular real-time monitoring with analytics, visualization and control of distributed energy resources is required for modern energy management systems.

The main innovations of the project include: ingesting, combining, and using multiple data streams of different data types together for grid analytics; developing a sophisticated platform to manage real-time data and computations; designing a set of data-driven algorithms to provide “what now” and “what if” analytics on grid operations; and publishing a large number of research papers on how innovative techniques from data science can be applied to these crucial problems for the grid.

Virtual SCADA: The data cleaning methodology, called the Virtual SCADA system, is illustrated in the following figure.



The methodology is based on the idea that data-derived relationships in data streams can be used to fill in holes, both within a single stream and between different streams. Many of the "what now" analytics address this: for example, topology estimation uses metered data to recreate grid parameters, and machine learning based power flow can be used to fill in holes in voltage measurements from power measurements or vice versa.

Data Plug: The "Data Plug" was designed as a tool to interface with partner database APIs to stream data and apply advanced data management and validation tools to process the data received.

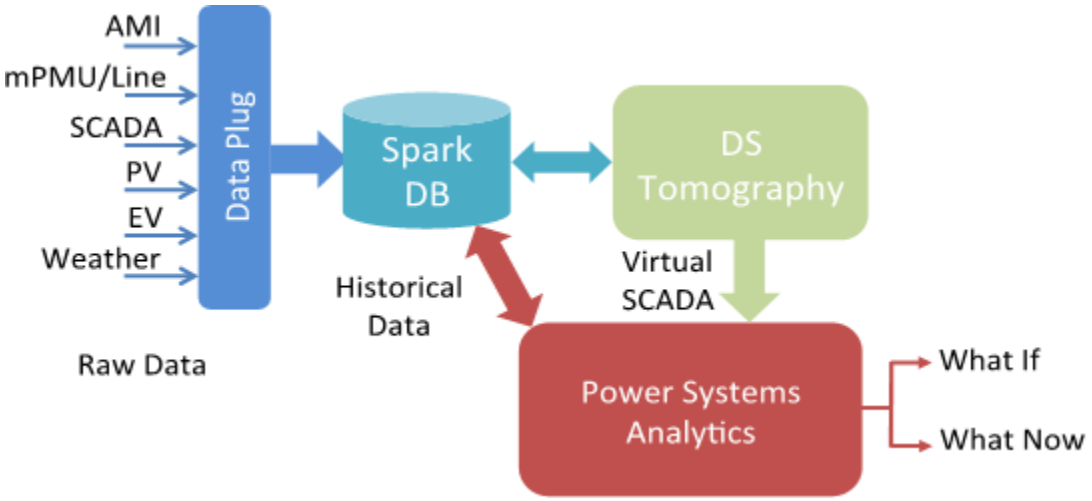
DS tomography module: This was designed to integrate disparate, unreliable data into "virtual SCADA" data streams for power system analysis, as described above.

"What now" Analytics: A crucial component to understanding today's grid, these analytics include advanced state estimation, topology estimation, outage detection, switch configuration detection, data-driven power flow models, quantification of EV flexibility, and disaggregation to study distributed PV generation. These all provide situational awareness based on the wide range of data inputs.

"What if" Analytics: The "what now" tools applied to analyze different scenarios of PV integration for planning, time-space analysis, and location benefits define the "what if" analytics.

Platform: To integrate all these components together a data pipeline was designed. The original high-level design is depicted in the figure below. After conducting user workshops,

interviews, and Technical Advisory Group meetings, the data pipeline design was modified to support extensibility and ease of development. These modifications and changes are discussed in more detail in the following section.



The project was organized into 7 tasks summarized below:

Task 1. Project Management: This was not a technical task but an important one for the success of the project. The project required heavy coordination with industry and all the research partners. It required that all the milestones were met on time and that the deliverables were high quality and on time. This task focused on project coordination and communication ensuring the team was available and responsive to funder's needs as well as being well coordinated with the evolving needs of the industry.

Task 2. Strategic Planning: While there is a growing number of systems (Advanced Distributed System Management, Demand Response Management Systems, GIS, etc.) with increasing capabilities, there currently is no software platform for integrating the ubiquitous but non-SCADA measurements that are and will be on the distribution system. Therefore, building the set of functional requirements for system operators and data providers and developing the key algorithms was crucial in developing the VADER system. The team and the technical advisors were able to evaluate the state-of-the art systems and decide on the key capabilities and analytics which would be useful for the industry to increase PV adoption.

Task 3 Data Collection and Integration: This task defines all data generation, collection, processing, and validation and spanned the three years of the project. To begin development of the VADER system, test data and APIs were collected from partners. This data was used in the initial development and testing of the methods which comprise the

VADER platform. The Data Plug module would handle access and entry of data from the various partners involved in the project. The real-time API access to partner systems would be custom for every data integration partner. The generation of database tables was to be input into the database system and passed through the data validation and initial processing step. Streaming access to the data was required to be developed and tested. The Virtual SCADA module provided the initial validation and alignment of the data. The goal of this module would be to create multiple sets of data that are time aligned with missing values imputed using purely statistical methods. The streams would be outputted to visualization dashboards.

Moving into year two, there was continued effort to define all data generation, collection, processing, and validation. While most of the work was to be competed in year 1, the task left open the possibility of iterations as needed as the project progresses if we found different, more innovative ways of solving the same problems or expanding the scope of problems due to TAG participation and feedback.

Finally, in year three, the task work was to refine, and complete validation of Data Plug with existing (or potentially growing) data sets.

Task 4 VADER Big Data: In this task, we designed and developed the core big data system for VADER. The data engine was to consist of the Virtual SCADA and Database component. The activities included finalizing the system architecture (Figure 1), database selection, and implementation. Reference Architecture for Amazon Web Services-based public cloud implementation is included in Project Results and Discussion section of this document. In year two, this task focused on the design and implementation of VADER's dashboard metrics, alarms and visualization. Finally in year three, work would focus on architecting and developing the core big data system for VADER, we were to integrate VADER with a power system modeling and simulation tool (GridLAB-D), and integrate it with an open-source visualization tool. This would allow us to explore using data for model validation and calibration as well as evaluate the connectivity with a commercially available visualization tool.

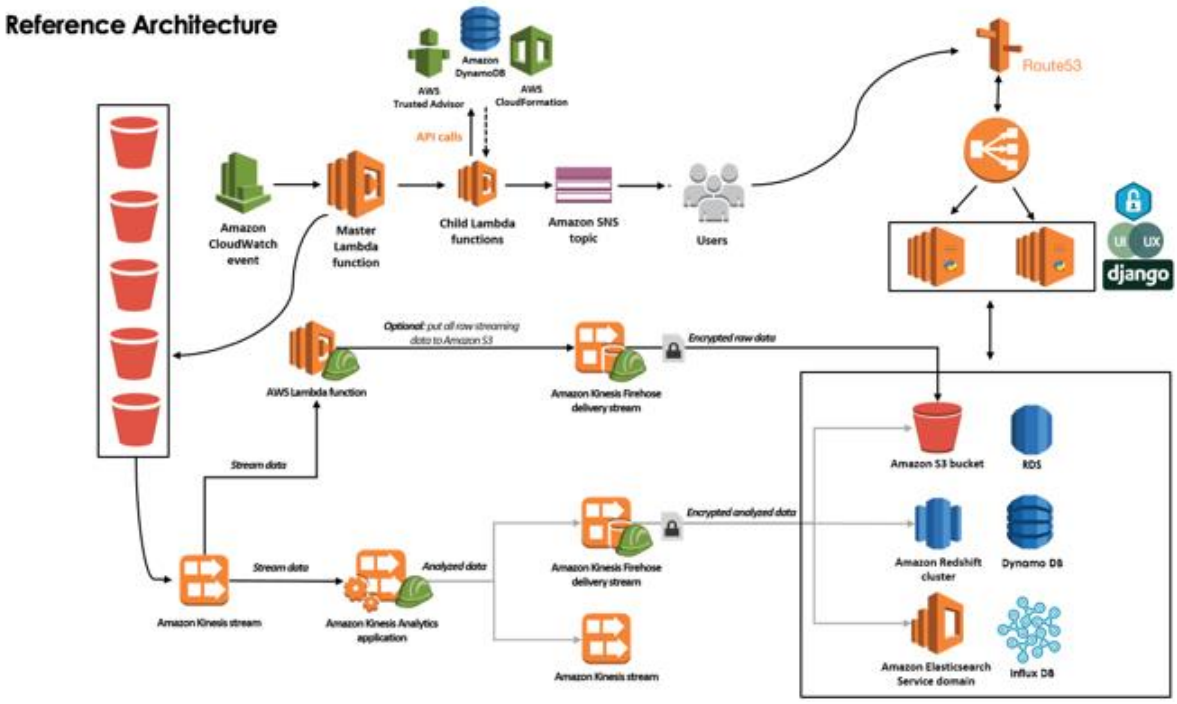


Figure 1. Schematic of the reference architecture for the VADER Data Platform built in AWS Public Cloud

Task 5. DS Tomography: What Now Analytics: In this task, our aim was to research representative traditional state estimation software; design data inputs and design of workflow, and research machine learning capabilities for state estimation for AMI data, PV data model (inverters, etc.), μPMUs, EV data (flexibility). In year two of this task, we would select representative traditional state estimation software; identified data inputs and design of workflow; prepared and connected the Virtual SCADA with traditional state estimation software, developed machine learning capabilities for state estimation for AMI data, PV data model (inverters, etc.), uPMUs, EV data (flexibility). Finally in year 3 our goal was to test our development efforts using traditional state estimation software using the data sets we to which we had access.

Task 6. DS Seer: This task, which was scheduled to begin in year two, was focused on selecting representative traditional power flow software; identifying data inputs and designing workflow; preparing and connecting the Virtual SCADA with power flow software, and developing machine learning capabilities for power flow for AMI data, PV data model (inverters, etc.), uPMUs, EV data (flexibility). Finally in year three, we aimed to expand the capabilities developed in year two to include additional data streams and use the selected representative traditional power flow software; identify data inputs and

design of workflow; develop machine learning capabilities for power flow for AMI data, PV data model (inverters, etc.), metering dataEV data (flexibility).

Task 7. Network Analytics: This task was to start in year two and leverage much of the previous task's work. Here we aimed to research and develop the topology identification algorithm based on Bayesian algorithms, sensor placement algorithm based on DS Seer models, outage detection algorithm based on Bayesian algorithms, state estimator for sub components of system and test the algorithms with datasets in the project for various scenarios. In year three, most of the effort would focus on wrapping up all the tasks started in previous years and completing this task.

Final Phase Milestone Deliverables:

VADER made significant contributions to reducing interconnection study time and increasing PV penetration. While achieving these two significant goals, the project also contributed to the use of data in DMS solutions. VADER team:

1. Published in peer-reviewed journals all the ML algorithms developed for modeling SE and all the other analytics and share the results of the analytics in peer-reviewed journals.
2. Published mature VADER analytics as an open source library of software tools. The code will be object-oriented, modular, and well documented.
3. Containerized and published VADER analytics as an open source reference model (including its architecture and modules) of how to deploy and use individual analytics modules and integrate with proprietary datasets using standard schema defined for each analytic module.
4. Reported on decisions made and implementations of data processing techniques, databases and streaming
5. The entire IP developed through this effort is open source, mature code posted on GitHub, and is available to the industry and vendor community.
6. Published workshop results: create an online video/webinar on how to install and use VADER Analytics including demo of individual analytics and associated documentation.

Project Results and Discussion:

The outcomes of the VADER project include the development of several analytics and also the data platform. In the following sections we describe the major results of these accomplishments.

Solar Disaggregation:

Disaggregation of behind the meter (BTM) solar power generation from net load measurements is quickly becoming a critical issue for grid operators. For BTM systems the power system operator (PSO) does not have access to the solar generation directly, but can only measure the net between the electrical load and PV output. At the end of June 2018 the number of non-utility BTM rooftop solar installations reached 6,200 MW in the Californian ISO's balancing area, with over 2,500 MW installed since 2016.[10] Fundamental operations such as switching, state-estimation, voltage management, and black start procedures need this information in order to be performed reliably and effectively, as evidenced by the inclusion of this topic area in the most recent Grid Modernization Lab Consortium funding opportunity announcement.

Data-driven approaches to estimating BTM solar generation is currently being researched by the Grid Integration Group at Lawrence Berkeley National Laboratory,[11] the Sensing and Predictive Analytics Group at National Renewable Energy Laboratory (led by Y.C. Zhang), and Viktor K. Prasanna's group at the University of Southern California.[12] All teams show certain similarities in their approaches:

- Unsupervised rather than supervised statistical learning methods
- A convex optimization formulation that is related to the concept of conceptually supervised source separation (CSSS)[13]

The different teams have all tackled different specific use cases and approaches including:

- AMI-level versus aggregate measurements
- Ex post versus real time analysis

The work on solar disaggregation supports Task 5: "DS Tomography – What Now Analytics". Building on the theoretical work presented in,[13] the VADER team formulated a domain-specific application of CSSS for disaggregating solar production from house-level AMI data, and validated the work on data proved by Riverside Public Utility.[14] In this work, we extended Wytock and Kolter's original analysis to show that, for the special case L-2 norm objective functions used here, the optimal solution for model coefficients are the

same as those found by a linear estimator, regardless of how each model's prediction errors are weighted in the objective function. We then proposed to estimate optimal weights for the problem objective function by comparing the variance of the linear model's predictions during daytime hours versus nighttime hours. The variance of the aggregate model's errors during nighttime hours reflects only load predictions, while the variance during daytime hours contains errors from both the PV and load predictions. Assuming the errors from each model are independent, we use this information to estimate variances for prediction errors from the load and PV models separately.

The VADER team further extended this work to include the use case of real-time disaggregation of solar production from streaming data, while doing further validation on the ex post use case.[15] The historical problem relies on data from advanced metering infrastructure (AMI), and data from a proxy signal that is contemporaneously related to solar generation behind the meters. We used generation from one or more nearby PV systems as solar proxies. On a set load and generation data from 52 homes within the Pecan Street dataset, we find the historical method is able to accurately predict which homes have solar in over 90% of cases. The historical problem is able to recover the 15-min resolution PV generation signals with root mean square errors between 20% and 50% of average daily PV generation. An example of these results are shown below in Fig. 1. A sensitivity analysis shows the method to be robust to the number of buildings and time span of data used to fit. However, including more than three solar proxies can cause false positive solar generation due to overfitting of solar proxies to unrelated noise in consumption data. Also, the method works better in homes which export electricity to the grid more often. We show that once the historical problem is fit to a set of homes, it can be applied in real time relying only on aggregate net load data from a substation, instead of net load measurements from individual buildings' AMI. We find the streaming problem performs with the same accuracy as the historical data problem on simplified test data explicitly constructed to model this problem, but we were unable to test the methods on real SCADA data from a distribution system operator.

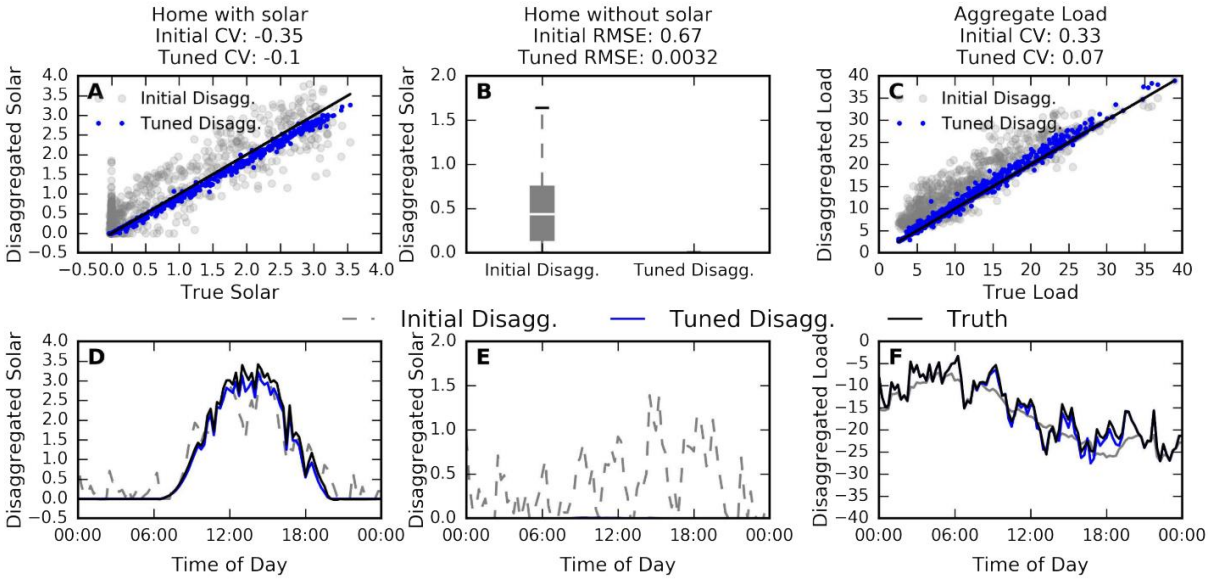


Figure 2: Descriptive results from model fitting. (Top row) Performance of the model before and after tuning alpha for (left) a home with solar, where points along the y=x are preferred because they indicate that the results of disaggregation match actual solar production; (center) a home without solar, where a lower RMSE is preferred; (right) aggregate load where, again, points along the y=x line are preferred. (Bottom row) Trace plots of disaggregated and actual signals for one day.

The work on solar disaggregation was further extended by a visiting master's student from Denmark Technical University, who deeply investigated sources of error in the previously proposed approach, methods for improving accuracy, and the best procedures implementing the proposed approach.[16] Important results from this research include:

- The tuning alpha routine showed to successfully remove nighttime generation which was assigned to some solar houses and significantly improve the performance of the algorithm at non solar houses.
- The estimated SNR (solar to noise ratio) proved to be an accurate proxy to estimate the quality of the signal separation after having performed it. Sites with lower SNR experienced a higher error. Intuitively, it is harder to extract the true solar signal when its energy is lower than that of the load.
- Fifteen minute data resolution proved to convey the same performance as the hourly set at houses with solar, while it outperformed the latter at houses without solar.

- The model proved to perform much better at the feeder level than at individual customers, as expected. The average CV dropped from 37 % at the individual level to 21.7 % at the feeder level.

Finally, the team developed open-source Python software, implementing the solar disaggregation algorithm, which is available in the project-level GitHub repository.

Machine Learning Based Power Flow:
Rather than trying to fill in the missing pieces of a traditional power flow model for the grid, we replaced the power flow equations with a purely data-driven set of models. The method proposed was a kernel-based Support Vector Regression (SVR).[17] SVR is a tunable regression method designed to be robust to outliers and overfitting. This was chosen for its flexibility to different observability and noise levels in the data for different networks. The kernel of the SVR model was quadratic to match the form of the physics-based power flow equations, so training the SVR model was equivalent to estimating line parameters. This connection to the physics based model helps insure against overfitting.

The results of the study were very positive. Power flow estimation error was very small, and was shown to be robust to networks with limited observability, devices with unknown droop control rules, outliers, and noisy data.[17]

Further work on this topic focused on the Inverse Power Flow Mapping, also called voltage estimation. Simple mappings for voltage estimation are very useful in designing voltage control laws, and research reported through this project demonstrated the success of this idea for small networks. For the inverse mapping, simple linear regression was found to match the performance of an SVR model. Other work in this area has implemented more complex, neural network based models[18] (Figures 3-5), but using a simple approach was found to be more than sufficient for small low voltage grids. These models are published in the VADER-Analytics Github along with sample notebooks showing how they can be implemented for different network datasets.

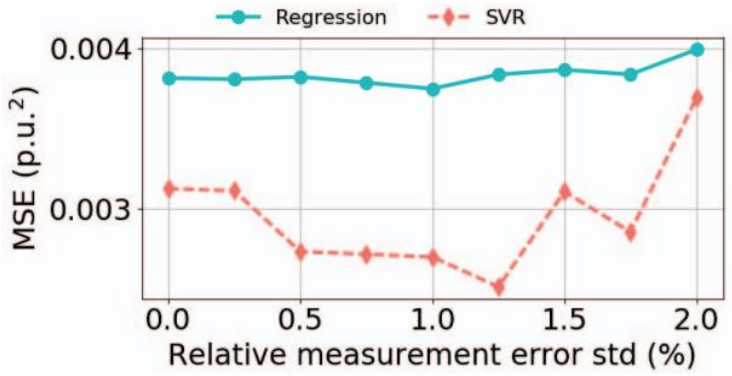


Figure 3. Mean squared error of power flow estimation on an 8 bus test system showing the improvement of the SVR approach over a traditional regression method under different measurement error levels.

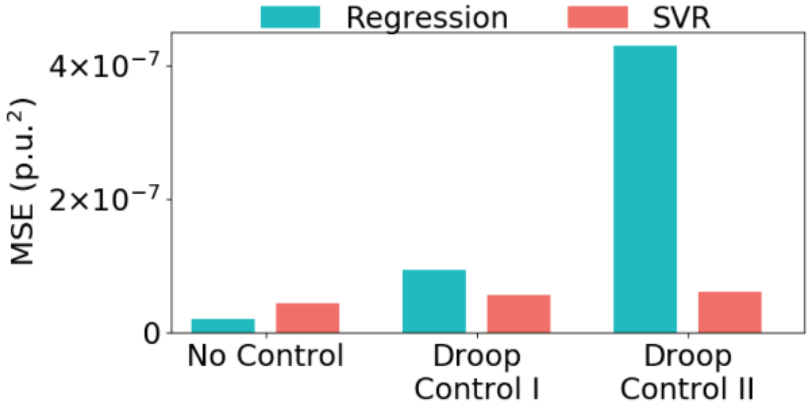


Figure 4. Mean squared error results for the same system showing the impact of having unobservable controllers in the system.

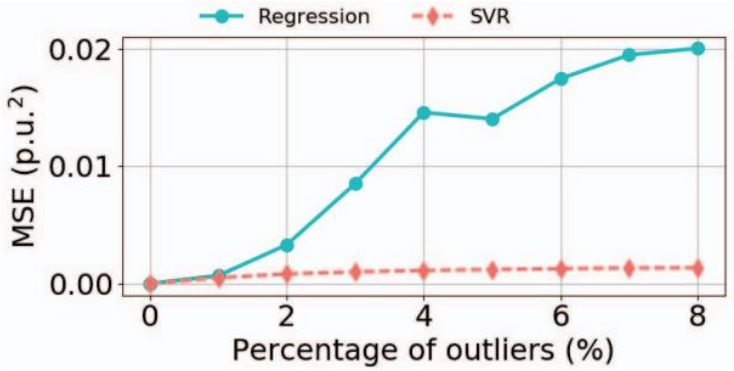


Figure 5. Results for the same system showing the robustness to outliers in the measurement data.

Application of ML Based Power Flow to Virtual SCADA:

Considering a dataset where a small percentage of entries have holes or missing measurements, the power flow models described above can be learned on the complete portion of the data and applied to the holes to fill in the data set. This application was coded and found to perform very well, and code to implement this method is included in the Github repository.

Topology Estimation:
To estimate grid parameters, one method developed in this project built a graphical model to assign connections between nodes based on estimates of the probabilistic relationships between voltage measurements.[19] A second study used Lasso regression to make the method robust to meshed grid topologies, and found very positive results for estimating grid topologies in standard IEEE test networks, including in the presence of loops.[20]

Parameter Estimation:
Building on the topology estimation work, two main studies were conducted on the coupling of data-driven topology and parameter estimation.

The first study focused on a three-step method where parameter estimation feeds into topology estimation which feeds into a second round of parameter estimation, “PaToPa”. The regression was hardened to the impact of measurement error in variables by implementing an error-in-variables model for the maximum likelihood estimation problem. A low-rank approximation was used to make the problem tractable, and results were found using smart meter data and the model of the real SCE feeder topology.[21,22]

The second study extended the “PaToPa” method to handle systems with multiple states and state changes. State changes in historical data were treated as an unobserved latent variable and an expectation-maximization algorithm was used to recover the hidden states. This combined framework was called “PaToPaEM” and showed very strong results on a range of IEEE test feeders.[23] More detail on the real-time IEEE 123 bus network feeder follows.

Real-time IEEE 123 bus network feeder implementation:

The IEEE 123 bus standard feeder network was implemented in the GridLAB-D modeling environment and augmented with 344 typical residential houses customized to replace the original constant current, power and impedance spot loads. In addition, the circuit accommodates four voltage regulators, overhead and underground lines, shunt capacitor banks, and a number of switching elements for controlling two-phase and three-phase

internal and lateral feeder components. The model is able to run into two modes: simulation and real-time. The real-time simulation mode has a front-end which has been categorized into six components: home, control, weather, feeder, meter and map.

Home:
The Home page allows the user to set up the model to their unique use-case by defining the server specifications, location-based information, load characteristics, data collection methods (e.g., SCADA, AMI), and output location.

Control:
The Control page allows for simulation output customization options, where the user can toggle options such as debug, warning, and other relevant messages.

Weather:
The Weather page displays the weather file information based on the input specified in Home. The environment can handle TMY3 and CSV weather file formats.

Feeder:
The Feeder page displays the status of the electrical feeder and allows to specify capacitor bank, internal switch and lateral switch configuration in real-time mode. The implementation utilizes standard voltage control through end-of-line voltage measurements.

Map:
The Map page displays the geo-coordinate mapping of the model onto Bakersfield, CA grid network using Google Earth tools and KML files generated by the network.

Several uses and new features were developed based on the tool described above. The real-time IEEE 123 model allows for validation of topology detection methodology and direct line measurement object development using DS system topology tool in conjunction with line flow measurements.[24] Additionally, a new switch coordination controller for DMS applications was developed using this platform.

Load Forecasting:
The work on load forecasting supports Task 6: “DS Seer–What If Analytics”. Recently, massive amounts of detailed individual electricity consumption data has been collected by newly deployed smart meters in households. A key technical challenge is to analyze such data to better predict the electricity generation and demand. The VADER team focused efforts in this area on developing methods for statistical forecasting of data

collected from advanced metering infrastructure (AMI) or “smart meters.” Work in this area focused on two fronts:

1. The forecasting of time-series load data that have hierarchical structure[25]
2. Developing a method for transfer learning for deep “long short term memory” (LSTM) neural networks, trained for time-series forecasting[26]

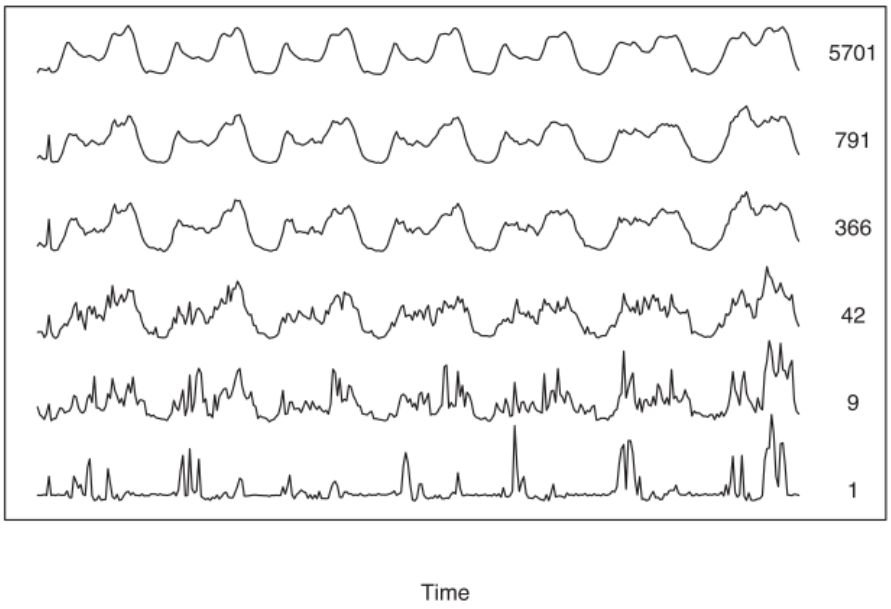


Figure 6: One week of electricity demand for different levels of aggregation, with the number of aggregated meters from the AECOM 2011 data set.

The first study showed that, by applying adjustments to the individual forecasts of a hierarchical time series (see Fig. 6), it is possible to obtain revised forecasts which satisfy hierarchical aggregation constraints. In existing approaches, the computation of these adjustments involve a high-dimensional unpenalized regression and the estimation of a high-dimensional covariance matrix. As a result, the existing forecasting methods can suffer from extensively adjusted base forecasts with poor prediction accuracy. We overcame this challenge by proposing a new forecasting algorithm that adds a sparsity constraint to the adjustments, while still preserving the aggregation constraints. The proposed method was validated on data collected during a smart metering trial conducted across Great Britain by four energy supply companies (AECOM 2011). The data set contains half-hourly measurements of electricity consumption gathered from over 14,000 households between January 2008 and September 2010, along with some geographic and demographic data. The experimental setup focused on generating one-day ahead hierarchal electricity demand forecasting at 30-minute intervals. The experiments

performed using hierarchical electricity demand data showed the effectiveness of our approach compared with the state-of-the art methods. In particular, the revised forecasts have a high sparsity in the adjustments during night hours, and reduce to the “best linear unbiased” revised forecasts during peak hours.

The second study builds on the well-researched area of transfer learning in the area of image classification to the domain of time-series data. We introduced a new loss function that aims to provide both regression loss, which is important for our forecasting objective, and a reconstruction loss, which is important for generalization and transferability. We have shown that our approach outperforms the baseline deep learning methods used for forecasting. More specifically, we have shown a dramatic forecasting accuracy improvement with transfer learning under small to medium training data size conditions, as shown in figure 7.

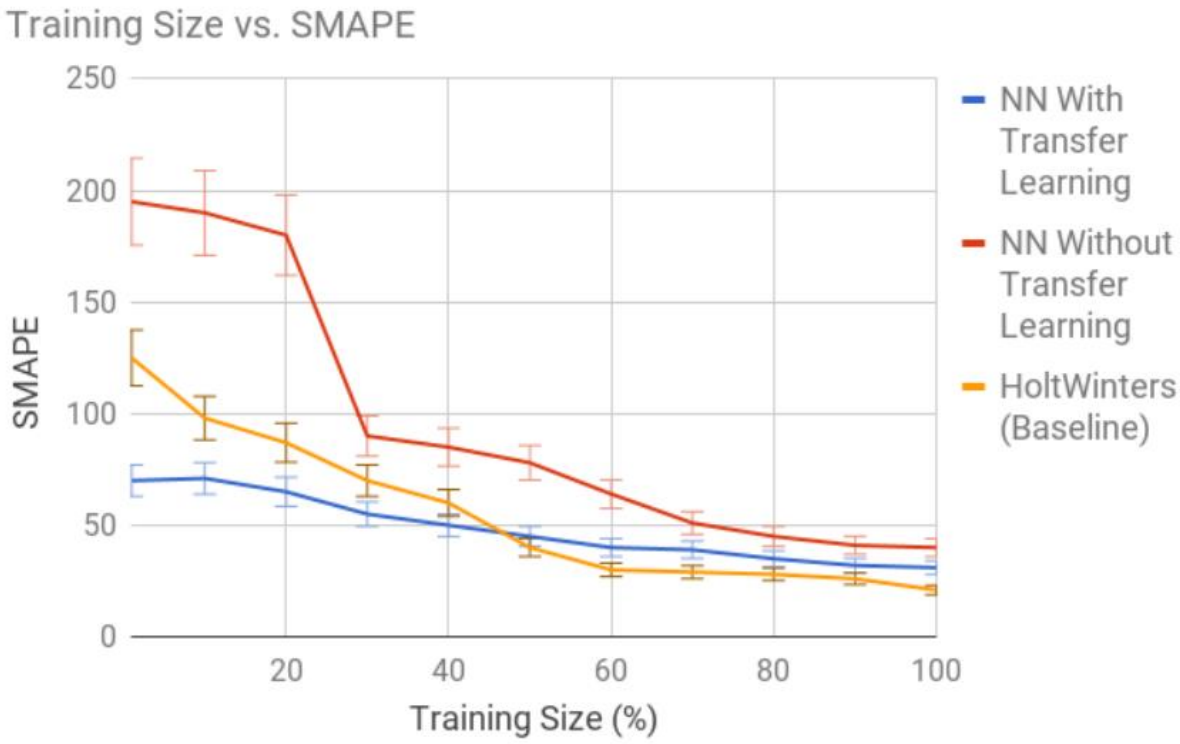


Figure 7: Performance comparison of deep LSTM models between training with single time series and training with transfer learning. The x-axis is the training size; the y-axis is symmetric mean absolute percentage error (SMAPE). The red curve represents single time series-trained model; the blue curve represents model using transfer learning. The performance gap is huge for short training sizes. When training size increases, the performance difference shrinks. Each round dot represents the mean SMAPE of all

customers from B2, and the error bar illustrates the standard deviation of SMAPE over 58,000 customers.

Data Platform:
The VADER Data Platform is a cloud-based big-data analytics platform which enables integration of massive and heterogeneous data streams for granular real-time monitoring, visualization, and control of Distributed Energy Resources (DER) in distribution networks. The cloud-based platform provides data- and app- hosting infrastructure. In addition to hosting analytics and visualizations developed in-house by the team, VADER Data Platform can be used to host 3rd party analytics and visualizations.

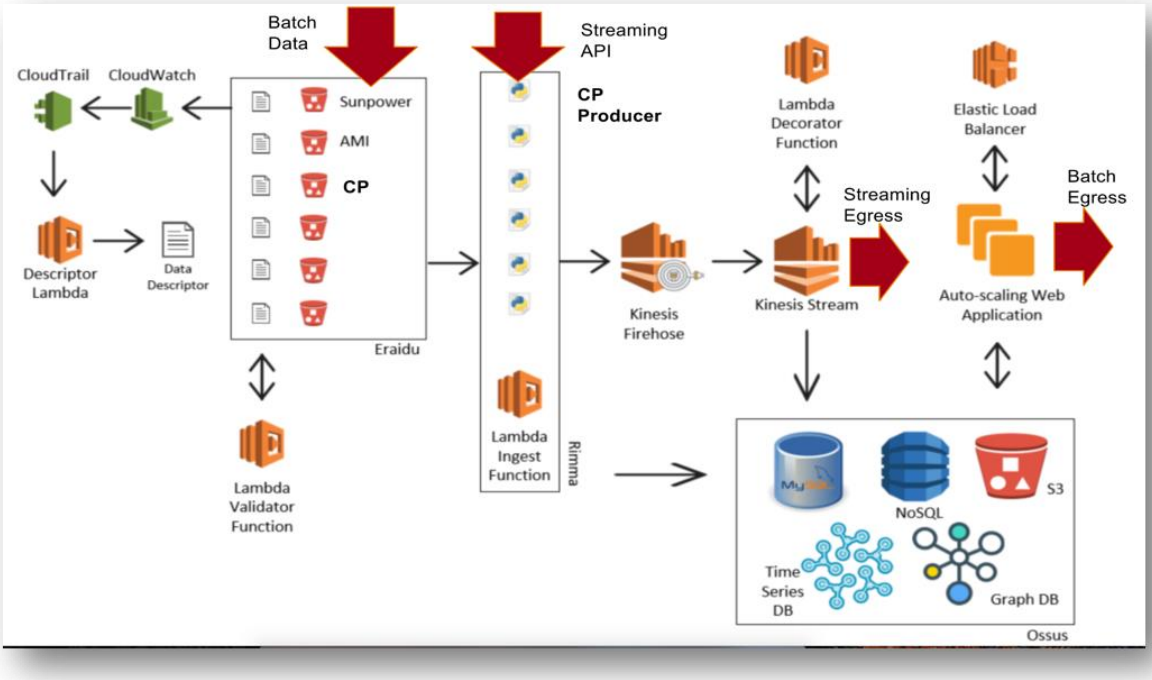


Figure 8. Final architecture (system diagram) of the VADER platform.

Figure 8 above shows the final architecture of VADER's underlying data management pipeline. The pipeline is developed using a combination of open-source software and managed services in Amazon Web Services cloud. Pipeline infrastructure supports both batch and API-based data ingestion. Ingested data may be pre-processed or transformed. Raw and/or pre-processed data is eventually persisted in one or more database formats allowing the option for polyglot persistence. Web applications and Jupyter notebooks, can serve as application-level interfaces with access to persistent storage and may be used to perform queries, develop analytics, and render plots.

This final architecture differs from the high-level diagram proposed early in the project. During the first year of VADER, we established a Technical Advisory Group and hosted user workshops, conducted 1:1 interviews, and group discussions to better understand the needs of utility operators. One feedback we received repeatedly was about building a flexible data pipeline that can support needs of smaller as well as large utilities. Building such a system using on-premise computing infrastructure can be technically challenging and cost-prohibitive. This was the motivation behind our biggest architecture-level change -- to use cloud-based infrastructure as opposed to on-premise technology. Figure 1 shows the reference architecture we developed for a cloud-based data platform. VADER reference architecture can be used as a blueprint for developing a highly flexible, scalable, and extensible data platform. Figure 8 shows the actual system diagram for VADER data pipeline developed based upon the reference architecture laid out in Figure 1.

Significant Accomplishments and Conclusions:

To assess and track quality of our software delivery, we defined a maturity model for our code-based research deliverables. The figure below list code maturity levels as defined by the VADER team.

Repeat

Reuse

Level	Description	Deliverable
1	Desktop under a researcher's desk; local processing of data	Research Paper
2	All code checked in to code sharing repository like github	Code Repository Research Paper
3	Coding standards followed; well documented installation notes; developer manual	Docs Code Repository Research Paper
4	Loose coupling; Target runtimes not OS; Autodeploy scripts	Deployable Packages Docs Code Repository Research Paper
5+	Robust authentication and authorization framework; service-oriented architecture; integration services; built-in security	Tools delivered via browser Deployable Packages Docs Code Repository Research Paper

The table below lists our analytic deliverables alongside their code maturity level.

Analytic	Code Maturity Level

Solar Disaggregation	4
ML based Power Flow - Forward	3
ML based Power Flow - Reverse	3
Topology Estimation	1
Switch State Detection	2
Virtual SCADA Demonstration	2
Load Forecasting	1

One major accomplishment for this project was the installation of VADER analytics at Southern California Edison. (SCE). The Advanced Technology Group (ATG) at SCE hosted us on their site to install VADER algorithms behind their computing firewall. We successfully deployed Solar Disaggregation and ML-based Power Flow (MLPF) inside the SCE computing environment, enabling their data science team to connect their internal datasets and test our algorithms with field data. We provided SCE staff with analytic code, installation environment, and deployment scripts. In addition to that, we provided onsite training to SCE data scientists, showing them how to test, and use VADER analytics. The analytics integration and deployment was successfully tested on a 6-node cluster in SCE's computing environment.

The VADER project exposed some major challenges associated with acquiring the massive data required for machine learning based algorithms.

For data science projects, like VADER, access to real world data is critical to properly validate the analytics developed under the scope of the project. However, gaining access to utility data can be challenging for both the researchers and the utility partners. The numerous sets of data, varying systems of records, incomplete of mapping between these data sets can result in an iterative process of generating a usable data set. This iterative process may not be easily automated and could be labor intensive. Therefore, it is highly important to communicate very clearly on what the available data will look like, to the level of headers, and fake sample data, to make sure it will match with the analytics once the legal agreements are in place and the data can be transferred. Additionally, having multiple sources for data can provide flexibility if significant delays are encountered.

Inventions, Patents, Publications, and Other Results:

- Kara et al. “Estimating Behind-the-meter Solar Generation with Existing Measurement Infrastructure (Short Paper)” , Buildsys’16 ACM International Conference on Systems for Energy-Efficient Built Environments (2016)
- Raffi Sevlian and Ram Rajagopal, "Distribution System Topology Detection Using Consumer Load and Line Flow Measurements", IEEE Transactions on Control of Network (to be submitted)
- Yizheng Liao, Yang Weng, and Ram Rajagopal, “Urban Distribution Grid Topology Reconstruction via Lasso”, IEEE Power & Energy Society General Meeting, 17-21 July, 2016.
- Yizheng Liao, Yang Weng, Chin-Woo Tan, and Ram Rajagopal, “Urban Distribution Grid Line Outage Identification”, IEEE International Conference on Probabilistic Methods Applied to Power Systems, 17-20 October, 2016.
- Jiafan Yu, Junjie Qin, and Ram Rajagopal, “On Certainty Equivalence of Demand Charge Reduction Using Storage”, Proceedings of American Control Conference, Seattle, WA, 24-26 May, 2017.
- Bennet Meyers and Mark Mikofski, “Accurate Modeling of Partially Shaded PV Arrays”, Proceedings of Photovoltaic Specialists Conference (PVSC-44), Washington, DC, 25-30 June, 2017.
- Jiafan Yu, Yang Weng, and Ram Rajagopal, “Data-Driven Joint Topology and Line Parameter Estimation for Renewable Integration”, Proceedings of IEEE Power and Energy Society General Meeting, Chicago, IL, 16-20 July, 2017.
- Jiafan Yu, Yang Weng, and Ram Rajagopal, “Robust Mapping Rule Estimation for Power Flow Analysis in Distribution Grids”, North American Power Symposium, Morgantown, WV, 17-19 September, 2017.
- Yu, Jiafan, Yang Weng, and Ram Rajagopal. "Mapping Rule Estimation for Power Flow Analysis in Distribution Grids." arXiv preprint arXiv:1702.07948(2017).
- Yizheng Liao, Yang Weng, and Ram Rajagopal, “Distributed Energy Resources Topology Identification via Graphical modeling”, IEEE Transactions on Power Systems, 2017
- S. Ben Taieb, R. Rajagopal, S. Ben Taieb, J. Yu, M. Neves Barreto, and R. Rajagopal, “Regularization in Hierarchical Time Series Forecasting With Application to Electricity Smart Meter Data,” Proc. Thirty-First AAAI Conf. Artif. Intell., no. 2011, pp. 4474–4480, 2017.
- M. Malik et al. “A Common Data Architecture for Energy Data Analytics”, IEEE SmartGridComm 2018
- Jiafan Yu, Yang Weng, and Ram Rajagopal, “PaToPa: A Data-Driven Parameter and Topology Joint Estimation Framework in Distribution Grids”, IEEE Transactions on Power Systems 2018

- Bennet Meyers, Michaelangelo Tabone, and Emre Kara, “Statistical Clear Sky Fitting Algorithm”, World Conference on Photovoltaic Energy Conversion, 2018.
- N. Laptev, J. Yu, and R. Rajagopal, “Reconstruction and Regression Loss for Time-Series Transfer Learning,” Proc. SIGKDD 2018, 2018.

Path Forward:

The work accomplished over the course of the VADER project has provided a foundation for many other efforts at SLAC. VADER’s data processing pipeline was initially deployed on Amazon AWS and has been used as a basis for the data processing pipelines in new projects that have started in FY19 at SLAC. Most notable are the OpenFIDO platform funded by the California Energy Commission under the EPIC program from FY18 through F22, as well as the LoadInsight, funded by DOE’s Advanced Grid Modeling Program in FY18 and DOE’s Technology Commercialization Fund in FY19. The data pipeline has also been used for the SETO-funded PVInsight project that kicked of FY19.

Many of the analytics that were first developed for the VADER project are being further developed in new projects. The statistical clear sky fitting (SCSF) work served as the basis for the PV-Insight project. In addition the ML-based powerflow analytics are being further developed for use in California Energy Commision projects HiPAS and SCRIPT. Looking further the solar disaggregation work has served as the basis of a number of new proposals submitted this fiscal year that still under review, most notably would be GMLC.

Another continuation of the VADER platform is the Grid Resilience & Intelligence Project (GRIP). This GMLC funded project is borrowing several architecture references developed in VADER. Although, GRIP is being developed on Google Cloud Platform (not Amazon Web Services, on which VADER was built) architectural references are transferable, particularly in the areas of data management, data pipeline, and serverless functions.

The code for the project is open sourced and available for further development (currently hosted at: github.com/malikmayank/VADER-Analytics/).

References:

1 L. Bird, D. Lew, M. Milligan, E. M. Carlini, A. Estanqueiro, D. Flynn, E. Gomez-Lazaro, H. Holttinen, N. Menemenlis, A. Orths, P. B. Eriksen, J. C. Smith, L. Soder, P. Sorensen, A. Altiparmakis, Y. Yasuda and J. Miller, *Renew. Sustain. Energy Rev.*, 2016, **65**, 577–586.

2 X. Zhang, A. J. Flueck and C. P. Nguyen, *IEEE Trans. Smart Grid*, 2016, **7**, 600–607.

3 P. Jahangiri and D. C. Aliprantis, *IEEE Trans. Power Syst.*, 2013, **28**, 3429–3439.

4 Y. Yuan, O. Ardakanian, S. Low and C. Tomlin, *arXiv:1610.06631*.

5 C. S. Indulkar and K. Ramalingam, *Int. J. Electr. Power Energy Syst.*, 2008, **30**, 337–342.

6 K. Dasgupta and S. A. Soman, *IEEE Power Energy Soc. Gen. Meet.*, 2013, 1–5.

7 S. Bolognani, N. Bof, D. Michelotti, R. Muraro and L. Schenato, *Proc. IEEE Conf. Decis. Control*, 2013, 1659–1664.

8 O. Ardakanian, Y. Yuan, R. Dobbe, A. Von Meier, S. Low and C. Tomlin, *IEEE Power Energy Soc. Gen. Meet.*, 2018, **2018**–**January**, 1–5.

9 L. Ding, T. S. Bi and D. N. Zhang, *APAP 2011 - Proc. 2011 Int. Conf. Adv. Power Syst. Autom. Prot.*, 2011, **3**, 2187–2191.

10 *Calif. ISO*.

11 E. Vrettos, E. C. Kara, E. M. Stewart and C. Roberts, *J. Renew. Sustain. Energy*, , DOI:10.1063/1.5094161.

12 C. M. Cheung, W. Zhong, C. Xiong, A. Srivastava, R. Kannan and V. K. Prasanna, *2018 IEEE Int. Conf. Commun. Control. Comput. Technol. Smart Grids, SmartGridComm 2018*, 2018, 1–6.

13 M. Wytock and J. Z. Kolter, in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI Press, 2014, pp. 486–492.

14 E. C. Kara, C. M. Roberts, M. Tabone, L. Alvarez, D. S. Callaway and E. M. Stewart, *arXiv:1607.02919*, 2016, 1–32.

15 M. Tabone, S. Kiliccote and E. C. Kara, in *Proceedings of the 5th Conference on Systems for Built Environments*, ACM, New York, NY, USA, 2018, pp. 43–52.

16 D. Innocenti, *Master's Thesis*.

17 J. Yu, Y. Weng and R. Rajagopal, in *2017 North American Power Symposium (NAPS)*, 2017, pp. 1–6.

18 M. Pertl, K. Heussen, O. Gehrke and M. Rezkalla, *IEEE Power Energy Soc. Gen. Meet.*, 2016, **2016**–**November**, 1–5.

19 Y. Weng, Y. Liao and R. Rajagopal, *IEEE Trans. Power Syst.*, 2017, **32**, 2682–2694.

20 Y. Liao, Y. Weng and R. Rajagopal, *IEEE Power Energy Soc. Gen. Meet.*, 2016, **2016**–**November**, 1–5.

21 J. Yu, Y. Weng and R. Rajagopal, *IEEE Power Energy Soc. Gen. Meet.*, 2018, **2018**–**January**, 1–5.

22 J. Yu, Y. Weng and R. Rajagopal, *IEEE Trans. Power Syst.*, 2018, **33**, 4335–4347.

23 J. Yu, Y. Weng and R. Rajagopal, *IEEE Trans. Power Syst.*, 2019, **34**, 1682–1692.

24 R. A. Sevlian and R. Rajagopal, *arXiv:1503.07224*, 2015, 1–16.

25 S. Ben Taieb, J. Yu, M. N. Barreto and R. Rajagopal, in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI Press, 2017, pp. 4474–4480.

26 N. Laptev, J. Yu and R. Rajagopal, *Proc. SIGKDD 2018*.