Applying Data Science to Improve Solar Power Production and Reliability

Mauricio Soto Karen Smiley Xiao Qu Travis Galoppo Rohini Kapoor
Alok Kucheria Melwin Jose

{mauricio.soto, karen.smiley, xiao.qu, travis.galoppo,
rohini.kapoor, alok.kucheria, melwin.jose}@us.abb.com

Abstract

Worldwide demand for a reliable and sustainable supply of renewable energy, including solar, is growing. Accurate estimates of solar energy production and insights into solar equipment performance help solar plant owners and operators optimize inspections, schedule maintenance, improve the operational performance of their equipment, and maximize the environmental benefit of their investments in renewable energy. However, due to the uncertainties inherent in the unpredictable nature of this renewable resource, many challenges are associated with estimation of solar power production and detection of performance issues.

In this study, our goal is to explore how predictions of solar inverter and plant production can be improved by applying data science techniques, and how machine learning models can be applied to correctly classify malfunction causes for solar inverters. Our results show that regional weather data can be used to estimate (and potentially predict) solar energy production for some applications; that a hybrid machine learning model based on historical data, temperature, and information from physical models outperforms predictions from state-of-the-art physical models; and that environmental factors such as lightning and ambient temperature, as well as grid operating conditions, can influence device reliability.

1 Introduction

In asset-intensive industries and operations, efficiently keeping critical long-lived operational equipment healthy and performant is essential to mission success. Active asset performance management (APM) enables customers to increase operational awareness of the health and performance of enterprise assets [10, 25]. Heightened health awareness empowers customers to move from costly reactive maintenance towards risk-based management techniques that optimize performance and maximize Return On Net Assets ('RONA') [32]. In this context, analytics have become essential for understanding and optimizing asset health and performance.

With rising demand for a reliable and sustainable energy supply [29], solar energy production is playing a crucial role in the residential, commercial, and industrial segments as well as for electric utilities. Penetration of renewable energy is increasing around the world. As one example, in the first quarter of 2016, the added generation capacity from solar to the U.S. grid represented 64% of the newly added generation capacity [17]. Optimization of solar power converters is critical in solar energy production: their failures account for 51% of maintenance tickets in solar plants [30], and their performance is essential for maximizing solar production under conditions which inherently have high variability [13].

An initiative was launched in early 2017 by a major manufacturer of solar power converters and related equipment for generation of renewable energy. The goal of the initiative is to improve reliability, production, and forecasting accuracy for solar production facilities. The manufacturer provided the authors with access to over 20TB of real-world solar monitoring data, including metadata, telemetry (periodic measurements), and events (machine and system states) associated with over 250,000 devices worldwide, as well as failure data.

In this paper, we describe how we applied data science along with physics-based models to this real-world data on solar production assets. First, we provide some background on the study content with respect to asset health analytics and the solar industry, as well as an overview of related work. Three aspects of our study are then described: better understanding and prediction of production of renewable energy (Section 2), improved accuracy and lead time for detection of degradation and diagnosis of failures (Section 3), and making these new analytics and the associated data readily accessible to end users and to internal customers who are not data science experts themselves (Section 4). For each aspect, we identify the primary research questions, our approach to the data science, and the results we have achieved to date. The paper concludes with a recap of lessons learned and business benefits achieved, plus a view towards future work and related challenges.

Asset Health Key Performance Indicators (KPIs): To maximize the value of an asset, it is helpful to predict as accurately as possible its behavior, production, events, risk of failure, and remaining lifetime. While there are many ways to quantify health of an asset [27], most algorithms reflect on some level the Risk of Failure (RoF) and the Remaining Useful Life (RUL). However, these health KPIs alone may not reflect performance degradation, which can reduce production and RONA long before actual failure or end of life. Accordingly, comprehensively managing asset performance requires analytics that quantify asset degradation as well as production quantity and quality.

RoF is the statistical probability of failure of an asset at a point in time or over a time period. Understanding RoF (or more generally, the risk of an adverse event impacting performance) is highly important. End of useful life is often aligned with a specific critical event of high interest. However, useful life may be deemed 'over' prior to failure, e.g. if a device's efficiency has dropped below an acceptable level. Estimation of degradation, RoF, and RUL can enhance condition-based maintenance, prognostics, and health management (Figure 1). However, these indicators are highly challenging to estimate or predict. Failures can be identified as functional, design, process, or random, and may be temporary or permanent. Determining useful values for degradation, RoF, and RUL is a complex task involving various failure scenarios, asset health, maintenance history, etc. A first estimate of RUL can often be derived from manufacturer-provided guidelines, then improved by using information obtained via condition and health monitoring. Algorithms are typically device-specific and incorporate expert domain knowledge, as well as statistical and/or machine learning techniques. Analytics may be applied predictively to drive proactive preventive maintenance decisions, or retrospectively to gain a better understanding of failures that have already occurred.

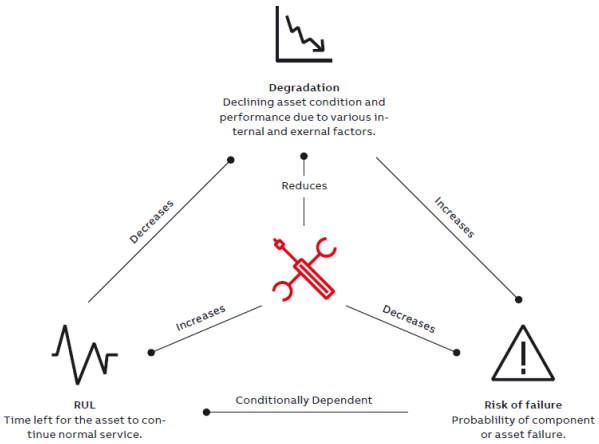

Figure 1: Relationships among Degradation, RUL, and RoF

Approach: One of the main challenges in solar forecasting methods is developing new tools and practices that manage the variability and uncertainty of solar power [29]. Our study leverages off-the-shelf machine learning (ML) mechanisms applied to our corpus of data on real-world solar production equipment together with physics-based models. Achievements from these solar analytics collaborations included novel algorithms for benchmarking and forecasting solar inverter performance and reliability; algorithms for real-time estimates of AC output and DC input power; automated diagnostic tools for service engineers for analyzing events and telemetry; and new visualizations to help customers better understand (and gain more value from) their solar equipment. The research also demonstrated how environmental data can augment the business value of the analytics. These capabilities are now being integrated into the company's portfolio of solar monitoring and asset performance solutions.

Related Work: Due to the importance of reliability in energy supply, analysis of improvement in production and reliability in power plants has been studied in the past [4, 12]. To mitigate the potential risks of imbalances between supply and demand, the high variances in power generation in solar and other renewable sources demand major attention to forecasting methods. Unpredictability in solar production can be driven by many time-variant causes (e.g., changes in behind-the-meter self-consumption, the positions of

clouds, air pressure, or changes in temperature [31]). Accuracy of prediction for renewable energy is known to vary according to its granularity, or time resolution. Previous studies [9, 14, 20, 29] have demonstrated that it is easier to achieve high accuracy in coarser granularity (e.g., day-ahead total daily production) than in finer granularity (e.g., 5 minutes or 1 hour ahead). Data science applied to solar forecasting models [2, 3, 5–8, 11, 15, 16, 18, 21–24] is proving to be a powerful tool for analyzing asset data and characterizing asset production, condition, and degradation; estimating RUL; quantifying RoF; and assessing the potential impact of maintenance actions. As one example, Alanazi et al. [2] proposed a two-stage hybrid day-ahead solar forecasting mechanism that introduces linear and nonlinear forecasting, therefore improving the accuracy of the obtained results. Other previous efforts have analyzed photovoltaic and solar thermal electricity generation from solar energy. Hoff et al. [13] presented a rigorous method to quantify power output variability from a set of photovoltaic systems, and Martin et al. [20] produced statistical models based on time series applied to predict half-daily values of solar irradiance by using auto-regressive neural networks and fuzzy logic models.

To aid in selecting analysis approaches depending on the available data, previous studies [26] have classified algorithms relevant to the asset health concepts of degradation, RUL, and RoF. This work also developed an assessment tool for characterizing "data readiness" for an analytics application, to provide valuable insight for choosing which algorithms to apply, and catalogued various data imputation strategies for handling missing values. These approaches were applied in this study.

2 Predicting Power Production

Models for solar power production[1] may be used in many scenarios. In this paper, we discuss two principal instances:

1) *Estimation* of power production, based on factors such as seasonality and current irradiance, may be used for detecting defects in the solar plant. Estimated DC or AC energy can be compared to the actual values measured by inverters[2] during the same time period. Significant deviations or patterns may indicate possible defects in the solar plant (e.g., degradation or sudden failure in one or more pieces of equipment, from the panels to inverters).

2) *Forecasts* of power production may be used to reduce inherent uncertainties associated with variable renewable energy generation. Grid operators today rely upon forecasts of both load [28] and generation to balance electricity supply and demand. Accurate forecasts not only support the safe and reliable operation of the grid, but also encourage cost-effective operations by improving the scheduling of generation and reducing the use of costly 'spinning reserves'.

Towards improving the accuracy of power production estimation based on physical models, we identified the following research questions:

- RQ1: How do model results differ when using regional weather data, with lower temporal resolution, vs. using data from hyperlocal weather stations with higher temporal resolution? (Section 2.1)
- RQ2: Does combining physical and Machine Learning (ML) models improve accuracy for estimating and predicting power production? (Section 2.2)

2.1 Impact of Weather Data on Prediction Accuracy

To address RQ1, we initially targeted all available solar plants which have one or more active inverters, in-plant weather stations (with high geographic and temporal resolution), and for which we could obtain similar weather data (lower geographic and temporal resolution) from a third-party source. In the manufacturer's monitoring system, data is collected from the in-plant weather stations, called *Environment Units (EUs)*, along with the inverters in solar plants. For this part of our study, we used two sets of historical weather values: hyperlocal *EU* data from our corpus, and regional data from a third-party weather data source (referred to as I). The third-party data was acquired through a limited quota of API calls that take the

[1] Our algorithms and models estimate and predict solar production in terms of DC *energy* for RQ1 and *power* for RQ2.

[2] The inverter measures DC and AC power. To be consistent with the outputs of our models for RQ1, we calculated values for energy based on equation 7 from IEC 61724-1:2017, Section 9.4.2. [1]

coordinates of a plant as input. The I API calls return data from the weather station which is closest to the plant location.

For each plant, we collected telemetry data for the inverters in the plants, which provide the actual measurement of DC power as 'ground truth'. We also collected the *Global Horizontal Irradiance (GHI)*, and the *Ambient Temperature (AT)* from both weather data sources (i.e., in-plant EU and the closest I weather station to the plant), in order to predict the power production based on various models. In this analysis, we implemented the GHI-based physical model defined in IEC 61724-1:2017 [1]. Similar physical models and a ML model were also introduced and studied for RQ2, as will be described in Section 2.2.

While the manufacturer's monitoring of the inverters, and in-plant EUs typically captures data every 5-15 minutes, the I data source currently provides regional weather data on at most an hourly basis. Therefore, we used *linear interpolation* to fill in missing data points for the regional data. Figure 2 compares

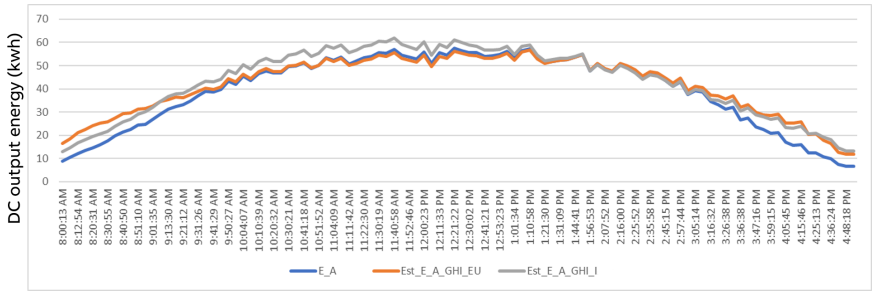

Figure 2: Example of Actual DC Energy vs. Estimated (one day, one plant, one inverter)

actual DC input energy (E_A) with estimated DC energy ($Est\_E\_A\_GHI\_EU$ and $Est\_E\_A\_GHI\_I$) for a single solar plant and a single day. E_A is the 'ground truth' read and calculated from a single inverter. $Est\_E\_A\_GHI\_EU$ is the estimated DC energy based on a GHI model using irradiance from the in-plant EUs. $Est\_E\_A\_GHI\_I$ is the estimated DC energy based on the same GHI model, where the input irradiance data is from weather source I. The distance between the plant and the weather station of I is 6.8 km. This single-plant example with one day data illustrates the potential effectiveness of the GHI model for estimating DC energy using either weather data source.

To answer RQ1, we repeated this analysis on multiple plants for a wide set of date ranges. We used the mean absolute percentage error ($MAPE$) prediction accuracy method as the key metric, given its intuitive interpretation in terms of relative error. After excluding inverters with invalid values for DC input signal (i.e., missing E_A) and EUs with invalid irradiance data (i.e., missing input for $Est\_E\_A\_GHI\_EU$) for the period 2016-01-01 to 2016-05-30, 53 plants were studied. The distances between plants and the closest I weather stations were in the range of (2.36km, 11.87km).

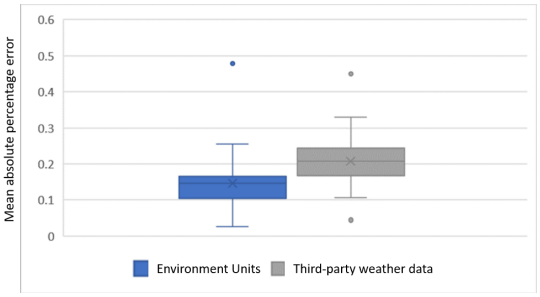

Figure 3: MAPE distribution of DC energy estimated by data from EU and I

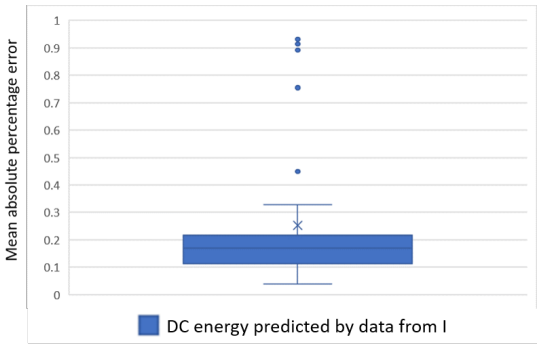

Figure 4: MAPE distribution of DC energy predicted by data from I

As shown in Figure 3, the median MAPE_EU (actual DC energy vs. estimated DC energy based on GHI measured by in-plant weather station EU, for each single inverter, averaged by all measured time) is

14.6%, and the median MAPE_I (actual DC energy vs. estimated DC energy based on GHI provided by weather station of I) is 20.7%, which is comparable to the error rate with in-plant weather stations. These results indicate that, for some applications, predictions based on regional weather data may acceptably represent daily power production in lieu of predictions based on in-plant weather data. This conclusion was verified with domain experts in solar plant service. We also confirmed our results using *BIAS*, an alternative prediction error metric. Currently, the prediction is applied on a single inverter. This meets the needs of residential or small commercial applications with a single inverter, and also provides the granularity needed to identify potential anomalies or degradation.

To explore the second scenario (*forecasting power production*, as described in the introduction of Section 2), we performed an additional quick experiment with the I weather data. Due to the limited quota of API calls, we were only able to collect 7-day GHI forecasting from 2018-07-02 to 2018-07-08 for 14 plants and studied 66 inverters in those plants. Distances between plants and the closest I weather stations were in the range of (2.27km, 16.38km). As shown in Figure 4, the median MAPE (actual DC energy vs. predicted DC energy based on GHI forecast provided by I) is ∼16%. Although this was a highly constrained, short experiment which precludes drawing any conclusion, the results are promising. Further exploration and more data would be required to determine if power production can be forecasted accurately enough for a given application by using third-party weather data.

2.2 Physics-based and Machine Learning models to estimate DC power

To address RQ2, we used 5-minute telemetry data for 2013-2017 from one inverter and its corresponding *EU* to estimate DC power with physical and ML models. As introduced in the previous section, IEC 61724-1:2017 [1] defines models that estimate the DC power produced by an inverter using the irradiance. Three different physics-based models are commonly used to estimate DC power in watts. The first ('Clear Sky') model assumes that the sky is clear, and calculates irradiance based on latitude, longitude, time of year, and other parameters. The second is based on GHI. The third and most accurate method is based on In-plane Irradiance, or Plane of Array Irradiance (POA). We tested if ML can outperform the physical model by leveraging time-series aspects of the data. Also, from the subset of data we examined, we noticed that irradiance is often not measured on-site in solar plants. For those plants, we can only use the Clear Sky model to estimate the DC power. As seen from Table 1, Clear Sky performs the worst out of the physical models and is associated with very high errors.

Next, we trained a Long Short Term Memory (LSTM) network which takes advantage of the time series aspect of our data. The input to this model consists of the 100 preceding DC power values and the output of this network is an estimate for the current DC power output.

The inputs to this model can be seen in Figure 5. In this figure, $Pin(t)$ refers to the DC power predicted at time t; $Pin(t-1)$ refers to the measured DC power at time $(t-1)$; and so on.

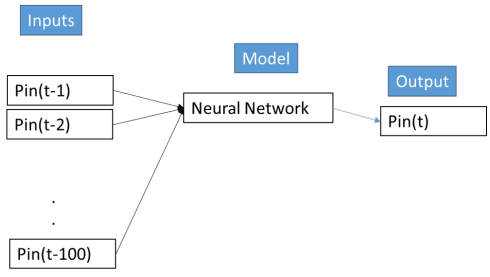

Figure 5: Inputs to the Machine Learning Model

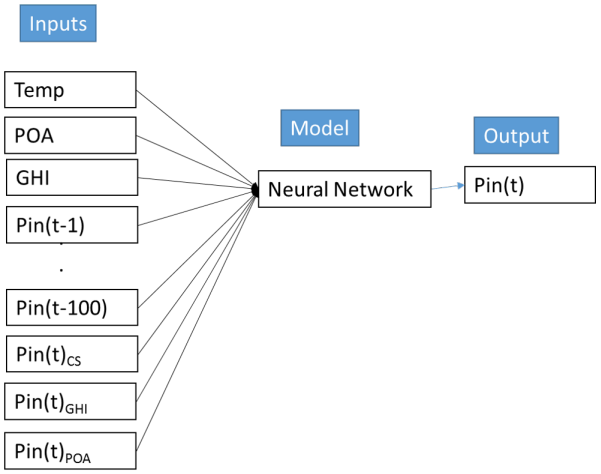

Figure 6: Inputs to the Hybrid Model

We also built a hybrid model which combines ML with all 3 physical models. This model extends the NN model shown in Figure 5 by using the current DC power estimates from the Clear Sky, GHI, and POA models as additional inputs. The input parameters for the hybrid model are shown in Figure 6.

For all of the above experiments in section 2.2, we used data from one inverter of power 55KW. The data was preprocessed as per guidelines introduced in IEC 61724-1:2017. We used a $75-25$ split for our train and test sets so that we could test the model on substantial data volumes, and also compare with the physical models. Accordingly, the NN models were trained using the initial 323K data points, and were tested with 107K data points. The same test set was used for the physical models.

To assess the quality of the models under inspection, we used *Root Mean Squared Error (RMSE)*. This metric tells us how far the estimates are from the actual values, and penalizes estimates which are further away from the actual values. Lower values for RMSE correspond to a better fit of the model. The RMSE values obtained on our test set are shown in Table 1.

	Clear Sky Model	GHI Model	POA Model	ML Model	Hybrid Model
RMSE	14,581	6,655	6,354	5,370	4,590

Table 1: RMSE of various models

These results show how hybrid ML models based on physical models and additional features could be used to estimate DC power more accurately than current state-of-the-art physical models. The results also show that ML models can potentially provide much better estimation than the Clear Sky Model in the absence of irradiance data. However, this is only a preliminary result: further validation with many more inverters and plants is needed before any conclusions can be drawn.

3 Failure Diagnosis for Inverters

The previous section answered research questions on predicting solar power production and contrasting current vs. predicted behavior of solar plants and inverters. For inverters which have already been determined to not be working properly, we examined failure diagnosis by using off-the-shelf ML algorithms on a portion of our corpus reflecting malfunctions in the inverters. In this section, we explore the following research questions:

- RQ3: How does lightning-related data correlate to inverter failure? (Section 3.1)
- RQ4: Can historical telemetry data on inverters be used to diagnose failures? (Section 3.2)

3.1 Impact of Weather and Lightning on Failures

Since photovoltaic equipment (often including the inverter) is constantly exposed to atmospheric conditions, identifying relationships between weather exposure and device failure is a critical step in developing more resilient equipment and for understanding degradation, performance, or failures. For instance, field engineers servicing solar inverters have noted that device failures seem to occur more frequently after intense lightning storms. Based on this anecdotal evidence, we analyzed device failure rates over a 28-month period in an effort to assess correlations between ambient weather conditions and inverter failure rates.

The data selected for this study included device age, geolocation, and communication timestamps from over 100,000 monitored, globally-deployed solar inverters, along with hourly weather measurements in the vicinity of each inverter (temperature, humidity, and precipitation), and global lightning strike data (timestamp, position, magnitude). For each device, the data was summarized as average monthly temperature, humidity, and precipitation along with the number and intensity of lightning strikes occurring within a 15-km radius for each month. In total, over 2.7 million device months and over 750 million cloud-to-ground lightning flashes were available for analysis in our corpus. (The lightning data was graciously provided pro bono by Earth Networks for this research.)

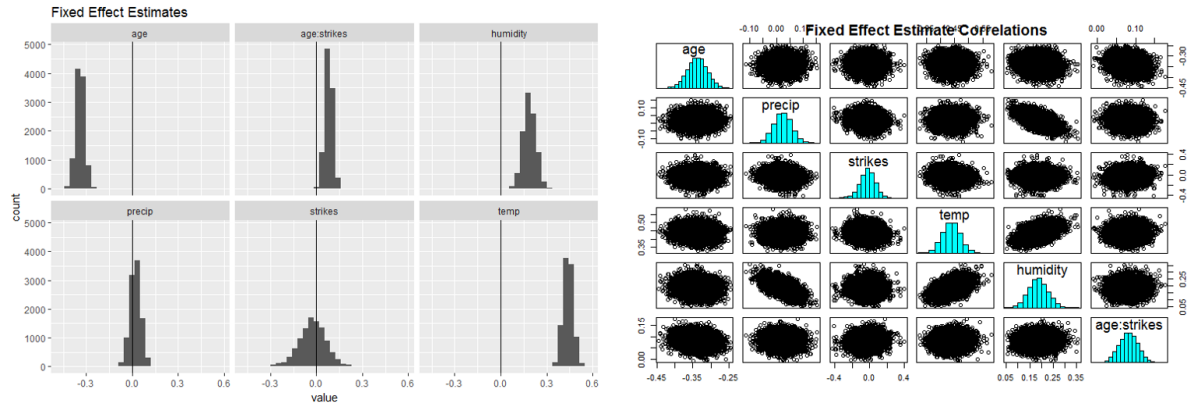

Figure 7: **Left:** Distributions of coefficient estimates resulting from Markov Chain Monte Carlo (MCMC) sampling. The vertical line is at 0, showing *temperature*, *humidity*, and *age * strike-count* having significant positive correlation with failure. **Right:** Pairwise scatter plots between coefficient estimates suggesting collinearity between {*temperature*, *precipitation*} and *humidity*.

To investigate possible correlations between weather exposure and device failure, we employed a mixed-effects logistic regression with fixed effects {age, temperature, humidity, precipitation, strike-count, age * strike-count}, and per-model intercept and strike-count random effects. The regression was performed in an objective Bayesian framework with coefficient estimation achieved via Monte-Carlo sampling. Figure 7 shows the distributions of the resulting coefficient estimates for each of the fixed effects, along with pairwise scatter plots between coefficient estimates to help identify collinearity.

Our analysis shows temperature, humidity, and age * strike-count all having statistically significant positive correlation with device failure rates. With respect to age * strike-count, this implies that as devices age, the correlation between lightning exposure and failure rate intensifies, suggesting, potentially, that resilience to lightning-related failure declines with age (or prolonged lightning exposure).

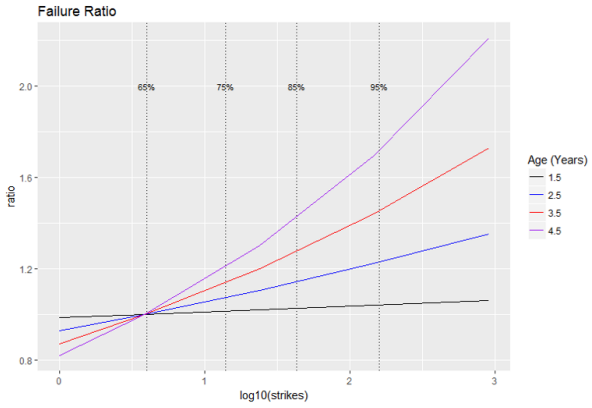

Figure 8: Y-axis shows the ratio of failure rates for devices exposed to increasing lightning counts, by age, with relation to similar age devices with average lightning exposure (ratio 1 is average exposure).

Figure 8 shows the correlation between lightning exposure and failure rates by device age. The graph shows the *ratio* of failure rates with respect to average lightning exposure; for instance, 3.5-year-old devices exposed to 100 (10^2) lightning strikes within a 15km radius in a month fail at a rate approximately 40% higher than similar-age devices exposed to average (~ 4) strikes in a month.

Figure 9 shows similar plots for temperature and humidity, where average daily temperature is very strongly correlated with increased failure rates. During any given month, devices exposed to an average daily temperature slightly over 100°F fail at twice the rate of those exposed to an average daily temperature of $\sim 77^o$F. Unfortunately, irradiance data (which could help identify if this is merely a function of increased operation, or truly a relationship between ambient temperature and failure) was unavailable for the solar plants used in this portion of the study. Nevertheless, this type of analysis can provide potential evidence of weather-induced device failure which can be further investigated.

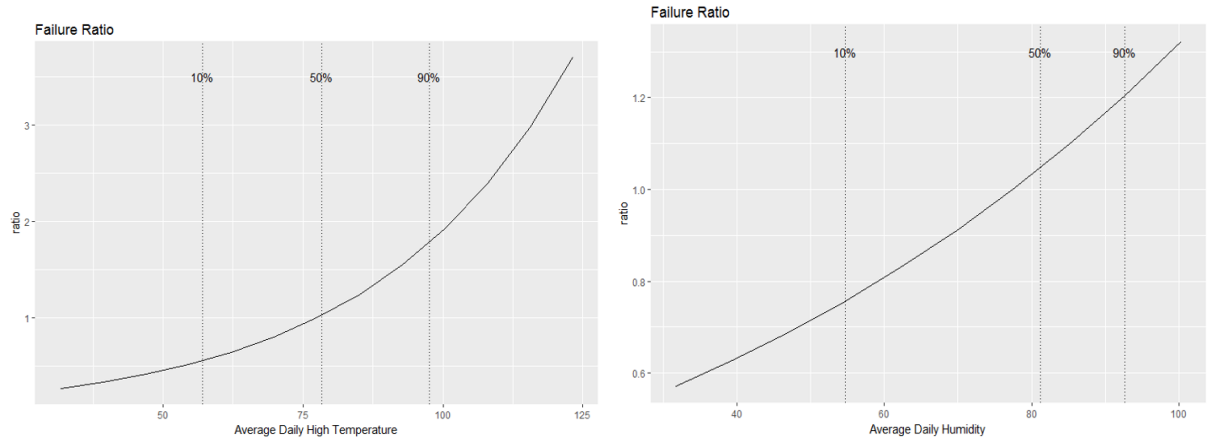

Figure 9: Failure ratios correlated with temperature (left) and humidity (right)

3.2 Impact of Non-Weather Conditions on Failures

Diagnosing causes of device failure is a complex and time-consuming task which often requires involvement of a domain expert, which increases costs for the manufacturer or servicer. Grid inrush is one infrequent, but highly damaging, cause of premature failure which is related to the environment in which an inverter is operated. In this part of our study, we evaluated whether grid inrush failures could be diagnosed from monitoring data. We classified a subset of failures from our corpus for a single inverter type by leveraging data science, historical failure data, and the telemetry data generated by the device under consideration. The monitoring data we used included the booster temperature (TempBst), inverter temperature (TempInv), current (Igrid), voltage (Vgrid), and frequency (Fgrid) readings from our monitoring data. We used the results from failure diagnoses by service center experts to label our inverters according to three categories: *Healthy inverters (H)*; *inverters that failed due to Booster error (B)*; and *inverters that failed due to grid INRush (F_IN_R)*. For the inverter subset we selected for this study, we had 147, 40 and 24 inverters respectively in these classes.

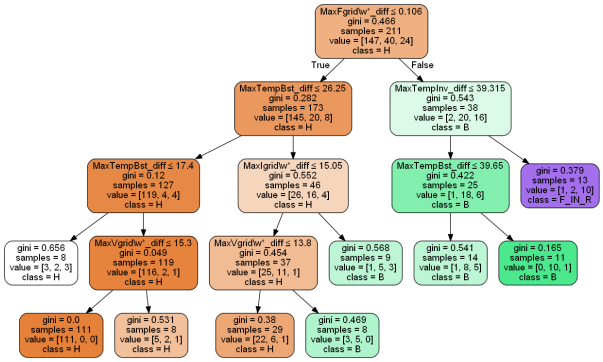

Figure 10: Decision Tree Classifier. Max⟨*metric*⟩_diff refers to the maximum fluctuation in the specified metric. H, B and F_IN_R indicates the class of inverters: healthy, failure due to booster error, and failure due to grid inrush. 'value' indicates the numbers of data points in these three classes.

Figure 10 shows how a decision tree classified these inverters based on the monitoring (time series) data captured prior to the failure. The hyper-parameters of the tree were selected by performing a grid search. The model with the highest accuracy used the following hyper-parameter values: maximum depth=4, minimum samples at leaf node=8, and minimum samples for a node to be considered for splitting=5. Using these values, we obtained an accuracy of 84% in a 4-fold cross-validation, as an initial proof of concept. We then used Random Forest to overcome the limitations of a single tree, which improved accuracy to 95%. Precision and recall for the H, B and F_IN_R classes with the Random Forest model were (0.941, 0.986), (0.886, 0.775) and (0.941, 0.986) respectively.

Even though this is a relatively small dataset, we consider that overall the analysis is potentially valuable and might be generalizable to a broader dataset. This could enable creation of a service tool to diagnose the cause for failure of a solar inverter by analyzing its telemetry data.

4 Visual Self-Service Analytics

In visual analytics, self-service Business Intelligence (BI) tools are employed to illustrate various data types — including metadata, telemetry data, and event data — using figures, maps and other charts. The BI tools can automatically fuse multiple data sources and types via common fields. Information is easily tailored for different user requirements and interests through interactive filters and selections, and all customizations made to one view are automatically applied to other views in real time. These features can help owners and operators of solar plants and equipment to visually identify anomalies and obtain other insights efficiently and effectively [26]. In this section, we highlight some ways we applied visual analytics tools[3] and techniques to empower our users to benefit from the new solar analytics described in this paper.

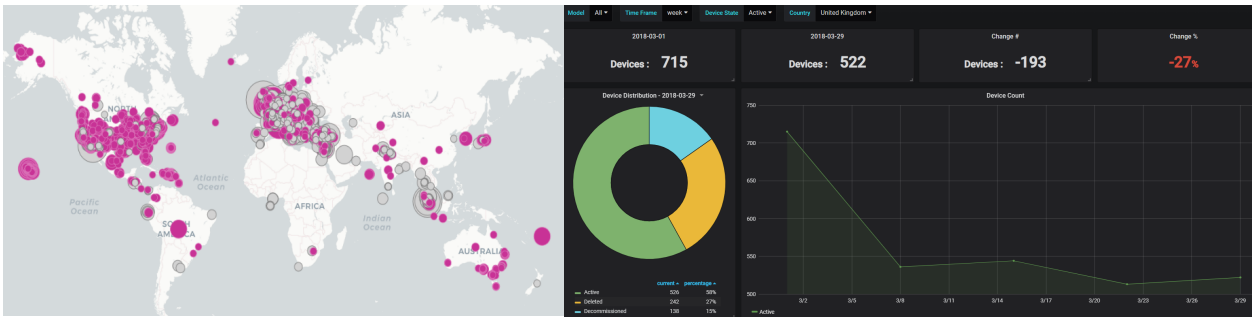

Figure 11: Geographic distribution of solar plants and KPI visualization

Examples of such features are detailed in Figure 11. The left graph shows the distribution of a subset of solar plants all over the world. Each circle represents a plant, and its size represents the number of a certain type of devices (e.g., inverters) in the plant. Filters such as device model type, device state, and region, are available for users. For all inverters in a selected group, several main KPIs and their changes are calculated and illustrated, including *Devices Under Management* and *Power Under Management*. Each KPI is visualized separately to show more details (Figure 11 right).

In addition to providing an overview of a group of devices filtered by different user requirements, visual analytics are also useful for understanding a single device. For example, the concept of an *asset timeline* [19] is to show a mixture of metadata, event data, and telemetry data in a single composite visualization to make it easier to see correlations. Figure 12 shows three days of telemetry and event data for one solar inverter on the same timeline. We can easily see that one event (ID: 72549264) occurred right after one telemetry data value (parsed ID: 6) was captured. This indicates that it may be valuable to investigate how suspicious values for some signals (captured in the telemetry data) may correlate with a subsequent error event.

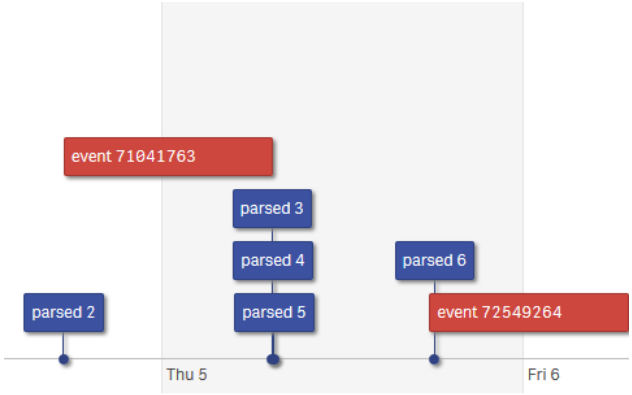

Figure 12: Asset timeline of telemetry and event data for one inverter

Finally, we also provided visual tools to pinpoint sub-optimally performing devices. For each device for which historical telemetry data is available, we generated an ideal performance curve by averaging the device's best historical performance. These ideal curves can be overlaid on the DC/AC power generated, or any other metric of interest, to check the device performance in real time. Figure 13 shows how an ideal curve (orange) for power generated is overlaid on the power produced by the inverter (blue) in real-time. Such a tool can enable the plant owner or servicer to investigate under-performing devices and make adjustments to achieve optimal power production.

[3]Qlik Sense (https://www.qlik.com/us/products/qlik-sense) and Grafana (https://grafana.com/)

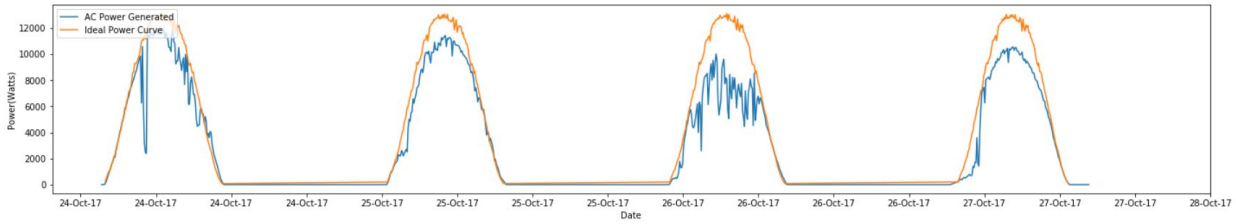

Figure 13: Ideal curve overlaid on the actual AC power produced

5 Discussion

5.1 Conclusion

Our goal for these studies was to explore how predictions of solar inverter and plant health and production can be improved by applying data science techniques. To date, our results have shown that predicting solar AC production based on regional weather, rather than in-plant weather, can be sufficiently accurate for some applications, and that blending physics-based models with ML and additional features can increase accuracy of estimations of DC power production. These analytics enable detection of degradation and improvements in production of solar energy. Similarly, our study shows that we can leverage Machine Learning and data science to better understand and classify potential root causes of device malfunction such as lightning, booster error, or grid inrush.

5.2 Business Value

Solar professionals are highly interested in minimizing Total Costs of Ownership (TCO). Inverter manufacturers can help meet their needs by offering analytics that provide early detection of degradation and early prediction of failure. These analytics can provide valuable lead time for performing maintenance and, if necessary, for acquiring and installing a replacement device or the right parts, to minimize maintenance costs and prevent days of lost solar energy production. Actual monetary benefits are situational and can be calculated based on factors such as days of lead time, whether hot spares are available, and avoidance of multiple service trips by bringing replacements for the most-likely components. Other factors include regulatory penalties and the value of electricity in the region (which ranges widely in the US and worldwide, e.g. under $0.10/kWh in India or for industrial customers in North Carolina, or $0.33/kWh in Germany or for residential customers in Hawaii).

5.3 Active Challenges and Future Work

In our ongoing work with solar analytics, we validate our assumptions and preliminary results, and consider other types of solar analytics which may be of high value to solar customers and service providers. For instance, so far we have assumed that the population of monitored inverters is representative of the population of non-monitored inverters. That assumption can be systematically tested, and we intend to do so. In some of the described use cases, a preliminary proof-of-concept was developed for a single inverter model or family. Our future work includes expanding the analyses to multiple inverter models.

Going forward, we anticipate further incorporation of new modes of data. For instance, some work was begun in 2017 and 2018 on natural language processing, using textual data which may be automatically captured or manually entered in various languages (e.g. technicians' diagnosis or repair notes); work on leveraging this data is expected to continue. Use of other data modes, such as images from various sources, is also being analyzed. We continue to leverage 'intracloud' data collection for solar inverters and plants to create new analytics for benchmarking, forecasting, event and weather correlations, and self-service visual BI. As they are developed, the analytics and tools described herein are being productized and deployed for internal business and customer use in managing solar asset health and performance.

References

[1] IEC 61724-1:2017. https://webstore.iec.ch/publication/33622.

[2] Mohana Alanazi, Mohsen Mahoor, and Amin Khodae. Two-stage hybrid day-ahead solar forecasting. In *North American Power Symposium (NAPS)*, 2017.

[3] Peder Bacher, Henrik Madsen, and Henrik Aalborg Nielsen. Online short-term solar power forecasting. *Solar Energy*, 83:1772–1783, 2009.

[4] Hector Beltran, Endika Bilbao, Enrique Belenguer, Ion Etxeberria-Otadui, and Pedro Rodriguez. Evaluation of storage energy requirements for constant production in PV power plants. *IEEE Transactions on Industrial Electronics*, 60:1225 – 1234, 2012.

[5] Songjian Chai, Zhao Xu, and Wai Kin Wong. Optimal granule-based PIs construction for solar irradiance forecast. *IEEE Transactions on Power Systems*, 31:3332–3333, 2016.

[6] Chi Wai Chow, Bryan Urquhart, Matthew Lave, Anthony Dominguez, Jan Kleissl, Janet Shields, and Byron Washom. Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed. *Solar Energy*, 85:2881–2893, 2011.

[7] Chi Wai Chow, Bryan Urquhart, Matthew Lave, Anthony Dominguez, Jan Kleissl, Janet Shields, and Byron Washom. Online 24-h solar power forecasting based on weather type classification using artificial neural network. *Solar Energy*, 85:2856–2870, 2011.

[8] Yang Dazhi, Panida Jirutitijaroen, and Wilfred M. Walsh. Hourly solar irradiance time series forecasting using cloud cover index. *Solar Energy*, 86:3531–3543, 2012.

[9] Alex Dobbs, Tarek Elgindy, Bri-Mathias Hodge, and Anthony Florita. Short-term solar forecasting performance of popular machine learning algorithms. In *International Workshop on the Integration of Solar Power into Power Systems*, 2017.

[10] Eric Fidler. Asset performance management helps oil and gas companies increase asset availability, improve uptime and empower more intelligent decision making. In *Offshore Technology Conference*, 2009.

[11] Jorge M. Filipe, Ricardo Bessa, Jean Sumaili, R. Tome, and J. N. Sousa. A hybrid short-term solar power forecasting tool. pages 1–6, 2015.

[12] Cody A. Hill, Matthew Clayton Such, Dongmei Chen, Juan Gonzalez, and W. Mack Grady. Battery energy storage for enabling integration of distributed solar power generation. *IEEE Transactions on Smart Grid*, 3, 2012.

[13] Thomas E. Hoff and Richard Perez. Quantifying PV power output variability. *Solar Energy*, 84:1782–1793, 2010.

[14] Chiou-Jye Huang and Ping-Huan Kuo. A short-term wind speed forecasting model by using artificial neural networks with stochastic optimization for renewable energy systems. *Energies*, 11, 2018.

[15] Rich H. Inman, Hugo T. C. Pedro, and Carlos F. M. Coimbra. Solar forecasting methods for renewable energy integration. *Progress in Energy and Combustion Science*, 39:535–576, 2013.

[16] Han Seung Jang, Kuk Yeol Bae, Hong-Shik Park, and Dan Keun Sung. Solar power prediction based on satellite images and support vector machine. *IEEE Transactions on Sustainable Energy*, 7:1255–1263, 2016.

[17] Shayle Kann, Justin Baca, MJ Shiao, Cory Honeyman, Austin Perea, and Shawn Rumery. US solar market insight - Q3 2016 - executive summary. Technical report, Wood Mackenzie Business and the Solar Energy Industries Association, 2016.

[18] Elke Lorenz, Johannes Hurka, Detlev Heinemann, and Hans Georg Beyer. Irradiance forecasting for the power prediction of grid-connected photovoltaic systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2:2–10, 2009.

[19] Shakeel M. Mahate, Karen J Smiley, and Paul F. Wood. U.S. 9,547,695 - Industrial Asset Event Chronology.

[20] Luis Martín, Luis F. Zarzalejo, Jesús Polo, Ana Navarro, Ruth Marchante, and Marco Cony. Prediction of global solar irradiance based on time series analysis: Application to solar thermal power plants energy production planning. *Solar Energy*, 84:1772–1781, 2010.

[21] Adel Mellit, Mohamed Benghanem, and Soteris Kalogirou. An adaptive wavelet-network model for forecasting daily total solar-radiation. *Applied Energy*, 83:705–722, 2006.

[22] Ricardo Correa Márquez and Carlos F. M. Coimbra. Intra-hour DNI forecasting based on cloud tracking image analysis. *Solar Energy*, 91:327–336, 2013.

[23] Christophe Paoli, Cyril Voyant, Marc Muselli, and Marie-Laure Nivet. Forecasting of preprocessed daily solar radiation time series using neural networks. *Solar Energy*, 84:2146–2160, 2010.

[24] Hugo Pedro and Carlos F.M. Coimbra. Assessment of forecasting techniques for solar power production with no exogenous inputs. *Solar Energy*, 86:2017–2028, 2012.

[25] Mehul Shah and Matthew Littlefield. *Asset Performance Management: Aligning the Goals of CFO's and Maintenance Managers*. Aberdeen Group, 2009.

[26] Karen Smiley, Xiao Qu, Travis Galoppo, Eric K. Harper, Alok Kucheria, Mithun P. Acharya, and Frank Tarzanin. Managing solar asset performance with connected analytics. *ABB Review*, pages 34–41, 2019.

[27] Karen J Smiley, Shakeel M. Mahate, Chihhung Hou, and Paul F. Wood. U.S. 9,665,843 - Industrial Asset Health Profile.

[28] Mauro Tucci, Emanuele Crisostomi, Giuseppe Giunta, and Marco Raugi. A multi-objective method for short-term load forecasting in European countries. *IEEE Transactions on Power Systems*, 31:3537–3547, 2016.

[29] Aidan Tuohy, John Zack, Sue Ellen Haupt, Justin Sharp, Mark Ahlstrom, Skip Dise, Eric Grimit, Corinna Mohrlen, Matthias Lange, Mayte Garcia Casado, Jon Black, Melinda Marquis, and Craig Collier. Solar forecasting: Methods, challenges, and performance. *IEEE Power and Energy Magazine*, 13:50–59, 2015.

[30] Sonia Vohnout, Patrick Edwards, and Neil Kunst. Uptime improvements for photovoltaic power inverters. Technical report, 2011.

[31] S. Watetakarn and S. Premrudeepreechacharn. Forecasting of solar irradiance for solar power plants by artificial neural network. In *IEEE Innovative Smart Grid Technologies - Asia (ISGT ASIA)*, 2015.

[32] Mike K. Williams, Ananth Seshan, and Karen Smiley. Asset Performance Management (APM) 2.0 - guidelines for goal setting and implementation planning. Technical report, 2017.