Synthesis of multi-year PV production data using generative adversarial networks

Gregory M. Kimball, Camille M. Pauchet, Rasoul Ghadami, Alberto Fonts Zaragoza
SunPower Corporation, Richmond, CA 94804, USA

***Abstract*—Multi-year forecasts of PV production are important for economic assessment of behind-the-meter PV+BES (photovoltaic plus battery energy storage) systems. Historical solar resource data is available for many locations in the United States, but these data are limited and must be converted from solar resource to PV production data before they can be used in BES control simulations. We propose both rule-based and generative adversarial network methods for synthesizing multi-year PV production forecasts. These methods use reference PV production and latitude-longitude inputs to generate hundreds of PV production scenarios which enable detailed simulation of behind-the-meter demand charge management.**

***Index Terms*—solar resource variability, energy storage, demand charge management, generative adversarial networks**

I. Introduction

Variability in solar resource has an important influence on the economics of behind-the-meter PV+BES (photovoltaic plus battery energy storage) systems. Many large retail electricity customers pay demand charges based on their peak monthly electricity consumption, and PV+BES systems work to reduce peak electricity consumption and lower monthly demand charges [1]. PV production offsets the gross load and supports solar-only charging of the BES, leading to an impact of solar resource variability on customer savings. Monte Carlo simulations of behind-the-meter PV+BES demand savings take advantage of synthetic PV production data to explore hundreds of upside and downside scenarios.

Satellite-based insolation data is available from 1998 onward in the United States [2], but on an annual basis this data only includes a few examples of upside and downside cases. Previous studies using historical solar resource data have been limited to observed scenarios rather than the wider set of likely scenarios accessible using generative methods [3]. Several studies have generated maps of interannual variability in solar resource [4]–[6] which provide starting point for studying variability in PV production. Although many PV production forecasting methods focus on day-ahead or month-ahead time horizons [7], PV+BES systems that are financed over 10-20 year terms require assessment over longer time horizons.

Rule-based algorithms make a good starting point for generating variants based on input data and can be tailored to study the different types of variability. Recent advances in generative adversarial networks (GANs) have been shown to accurately produce new samples with high diversity and complexity for many types of time series data [8]–[10]. In this study we propose both rule-based and generative adversarial network methods for synthesizing multi-year PV production forecasts (Fig. 1) and analyze the resulting data using statistical approaches as well as demand charge management simulations.

II. Experimental

A. Data sources

NSRDB, the National Solar Radiation Database, is a public data source for hourly global horizontal irradiance (GHI) data across the United States from 1962 to present. In the continental United States, the WBAN data set includes 239 sites from 1962 to 1990 and the USAF data set includes 1020 sites from 1991 to 2005. Combined this data set gives us 19690 site-years of hourly irradiance data. In this work we developed generators for multi-year forecasting based on a reference one year 15-min data file. The algorithms can generate either GHI or PV production data depending on the reference time series data provided. The primary application is generating PV production data that reflects the solar resource variability for locations across the United States.

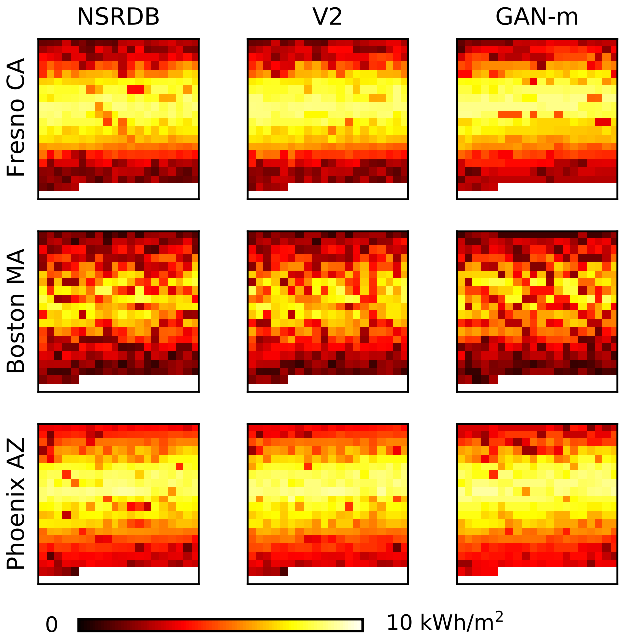

Fig. 1. Images showing daily insolation for one year from three locations and three sources, where **NSRDB** refers to ground station measured data, **V2** refers to a data from rule-based algorithm, and **GAN-m** refers to data from a generator trained in an adversarial network. The images contain 20x20 pixels with each pixel representing a daily insolation value.

B. Rule-based generators

The rule-based generators receive one year of 15-min PV production data plus latitude-longitude as input and generate a new variant of one year 15-min PV production data as output. The method ’V1’ modifies the input PV production data to meet a higher or lower annual production target. First, the method selects an annual target based on regional normal distributions of insolation and project latitude-longitude [5], [6]. Then the method modifies the 15-min PV production interval data iteratively until the target is achieved. For each iteration the highest (lowest) production days are copied over the lowest (highest) production days within random 14-day intervals until the annual sum reaches the target. The method ’V2’ adds an additional step to reorder portions of the data before running the target-searching step of V1. For method V2 this additional step modifies the PV production data by applying 30 operations as follows: select three days in a random 14-day window and swap them with another three days.

C. Generators from adversarial networks

The methods ’GAN’ and ’GAN-m’ use a generator model trained as part of a generative adversarial network. The ’GAN’ method combines an array of random numbers conditioned on latitude-longitude as input and generates 365 normalized daily production values as output. The raw 365-day output of the generator is then de-normalized based on the highest daily sum in the reference data set. The ’GAN-m’ method uses the output of the ’GAN’ method and further modifies the data by matching each daily production value to the day with closest production within ±10 days in a reference data set. The ’GAN’ method only provides daily interval data, while the ’GAN-m’ method provides both daily and 15-min interval data.

The generative adversarial network consists of a discriminator model that is trained to determine whether input data is from a real or synthetic data source, and a generator model that is trained to generate data that the discriminator will mark as real. The generator and discriminator models use a conditioned GAN (cGAN) design where the input data are conditioned by latitude-longitude inputs. The discriminator receives 365 daily production values conditioned on latitude-longitude and is trained to output 1 for real and 0 for synthetic data. The generator receives 50 random-normal values conditioned on latitude-longitude and outputs 365 daily production values. The random input values represent a point in the ”latent” space of possible output patterns.

The discriminator consists of two 1D convolutional neural network layers with a fully-connected layer to a single sigmoid output. The generator consists of two 1D convolutional transpose layers with a 1D convolutional layer that forms the 365-value output. The training process used normalized NSRDB GHI data as input, applying ±4-day roll to the data for each epoch. The latitude and longitude data were augmented with a ±1.2 degree random-uniform jitter. The training showed unstable behavior for the first 300-400 epochs, followed by

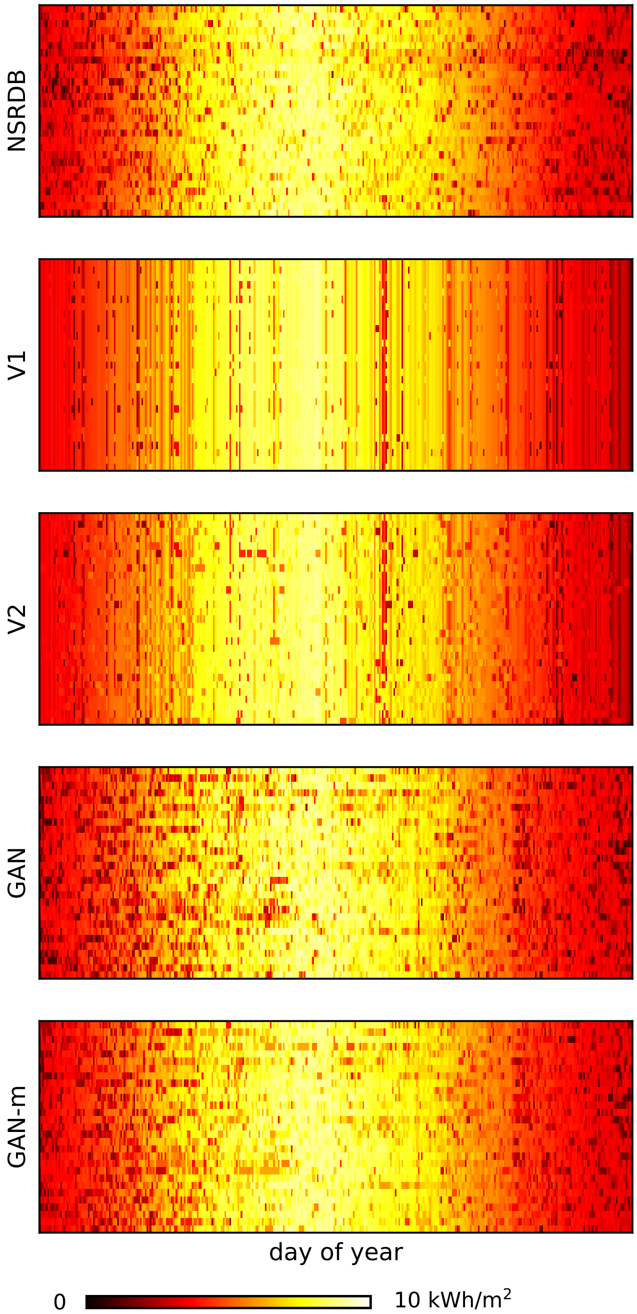

Fig. 2. Images showing 29 years of daily insolation data from Phoenix AZ. **NSRDB** shows ground station measured data, **V1** shows data from a generator targeting annual insolation, **V2** shows data from a generator that reorders parts of the data and then targets a new annual insolation value, **GAN** shows the output of a generator trained in an adversarial network, and **GAN-m** shows the GAN generator output matched to the nearest ±10 days of a reference year.

stabilization around 80-90% accuracy for real and synthetic samples, and training was terminated at 3000 epochs.

D. Model validation

Figure 2 shows 29 years of measured daily insolation data from NSRDB ground stations, as well as 29 examples of generated data by the ’V1’ and ’V2’ rule-based models and the ’GAN’ and ’GAN-m’ machine learning models. For each location, the generators were run using GHI data from the NSRDB median annual insolation year. In this mode the generators produce GHI data rather than PV production data.

For each location, t-Distributed Neighbor Embedding (t-SNE) [11] was applied to the daily insolation data from the 5 data sources. The embedding was performed with 2 components to facilitate visualization of clustering between the measured and generated data. Two-sigma confidence ellipses were also computed as a qualitative measure of diversity for each data source.

E. Demand savings simulations

Multi-year PV production data was also used as input for demand charge management simulations of a behind-the-meter PV+BES system. The simulated site was a commercial location in Pacific Gas & Electric territory using a tariff based on PG&E E19 Secondary Option R. For the validation exercise, the simulations used the same gross load profile, incorporated static and dynamic BES losses, and optimized the BES dispatch commands (Fig. 5). The demand savings simulation upscales the input gross load and PV production data from 15-min time series data to 2-sec interval data using a Markov chain method. The 2-sec data is used as input to a realtime economic dispatch optimizer which provides a sequence of BES dispatch commands while complying with constraints to charge from PV production and limit discharge to the gross load. The gross load, PV production and BES dispatch data is combined to compute the net load for the customer. Demand savings from PV+BES is defined as the difference between the demand charges for the gross load and the demand charges for the net load, and is aggregated over 12 billing cycles making up one year of operations.

For validating the data from each data source, we used insolation data as inputs to the methods, and processed rescaled output data as PV production. Even though there are different tariff constraints and typical load profiles in the three locations, we chose to only adjust the PV production by location to focus on the impact of solar resource variability on demand savings. Each demand charge management simulation provided a single PV+BES demand savings value for each year of 15-min PV production data.

III. Results

Figure 3 shows several statistical metrics used to compare measured versus generated insolation. The top pane shows annual insolation data from 29 years for 5 data sources for Phoenix AZ, and shows a good matches between NSRDB and the rule-based models (V1, V2). The V1 and V2 models target annual insolation values from a normal distribution of NSRDB sites within 100 km of the site location, so their correspondence with NSRDB is a key feature of the algorithm. The ’GAN’ model output shows a median annual insolation value about 2.3% lower than the NSRDB median, due to a greater incidence of heavy clouds in the generated data. After applying the matching correction, the ’GAN-m’ method shows an median annual insolation consistent with the NSRDB median. The middle pane of Fig. 3 shows mean monthly insolation values for the 5 data sources, showing that correspondence with NSRDB is preserved even as monthly insolation values change seasonally from 100 to 250 kWh/m2/yr. The bottom pane of Fig. 3 shows the standard deviation across 29 samples of monthly insolation values. For standard deviation in monthly insolation, the rule-based models V1 and V2 show lower variability than the NSRDB and GAN-based models.

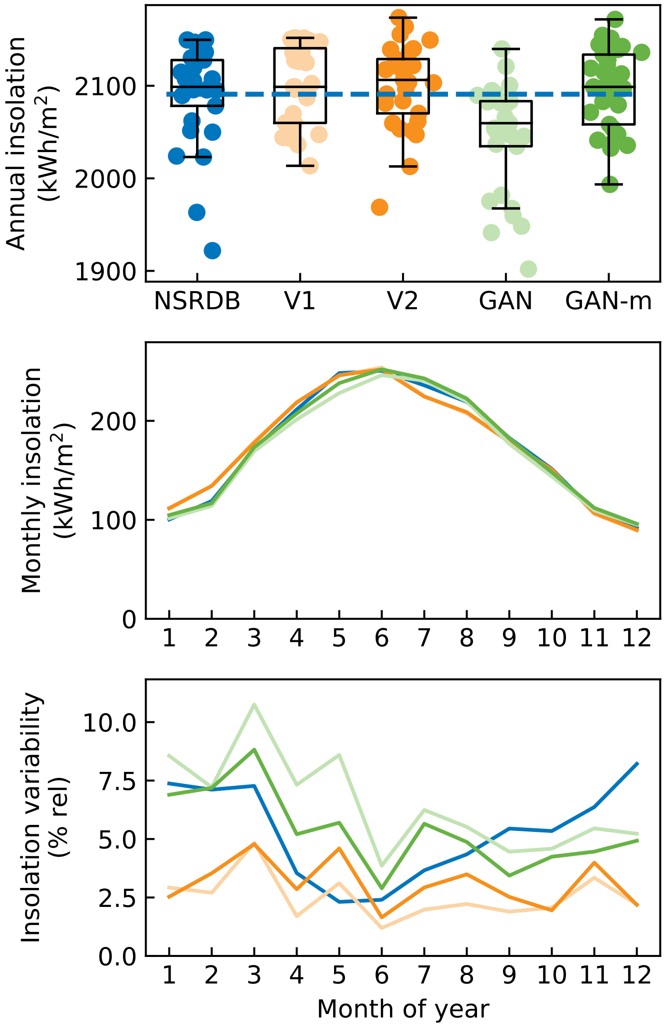

Fig. 3. Solar insolation statistics from measured and generated data for Phoenix AZ. The top pane shows the distribution of annual insolation values for each data source. The middle pane shows the average monthly insolation for each data source. The bottom pane shows the standard deviation of monthly insolation for each data source.

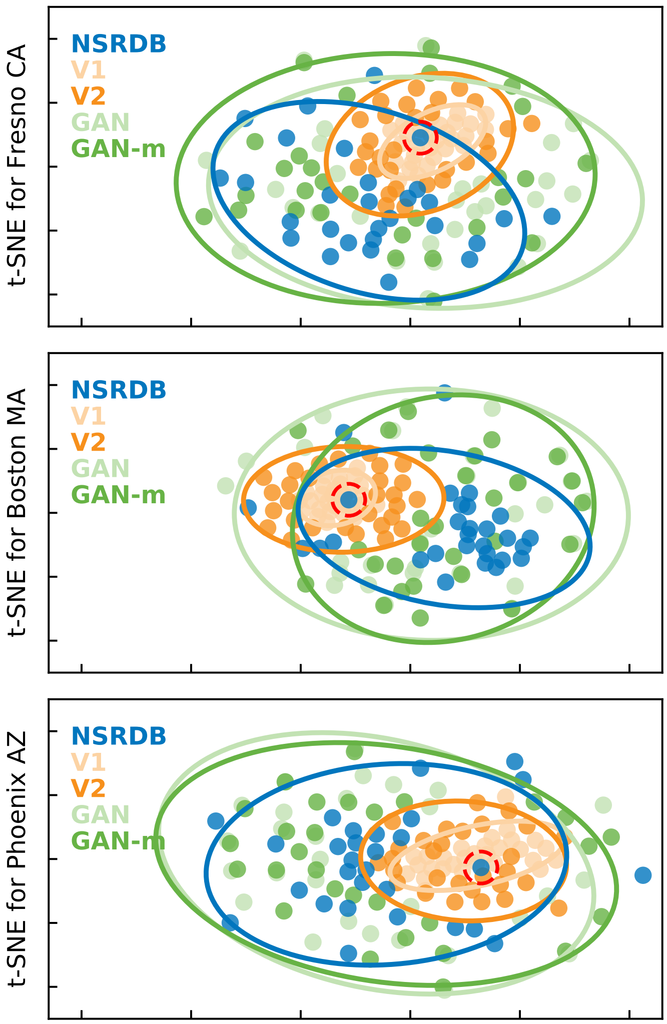

Fig. 4. Analysis of several measured and generated data sources based on the t-Distributed Neighbor Embedding (t-SNE) method for 3 locations and 29 samples per source. Each point represents 365 daily insolation values and the ellipses show the covariance confidence at 2σ. The red circle denotes the NSRDB median insolation data point used for the V1 and V2 models.

For each location, t-SNE analysis showed that the generated data largely overlaps with the measured data (Fig. 4). For the V1 method the results were tightly clustered around the median annual insolation year that had been used as reference data during data generation. The additional modifications of the V2 method expanded the cluster relative to V1. By contrast, the 'GAN' and 'GAN-m' clusters showed larger confidence ellipses that expanded beyond the NSRDB cluster to encompass some additional variability.

The measured and generated data served as PV production input to PV+BES demand charge management simulations. Figure 5 shows six days of a demand charge management simulation, highlighting the daily patterns in state-of-charge (SoC), BES dispatch and net load. Demand savings analysis for 29 years of 4 data sources and 3 locations were compiled to assess the variability provided by each data source. Figure 6 shows that the measured and generated data sources follow the same trends across the 3 sample locations.

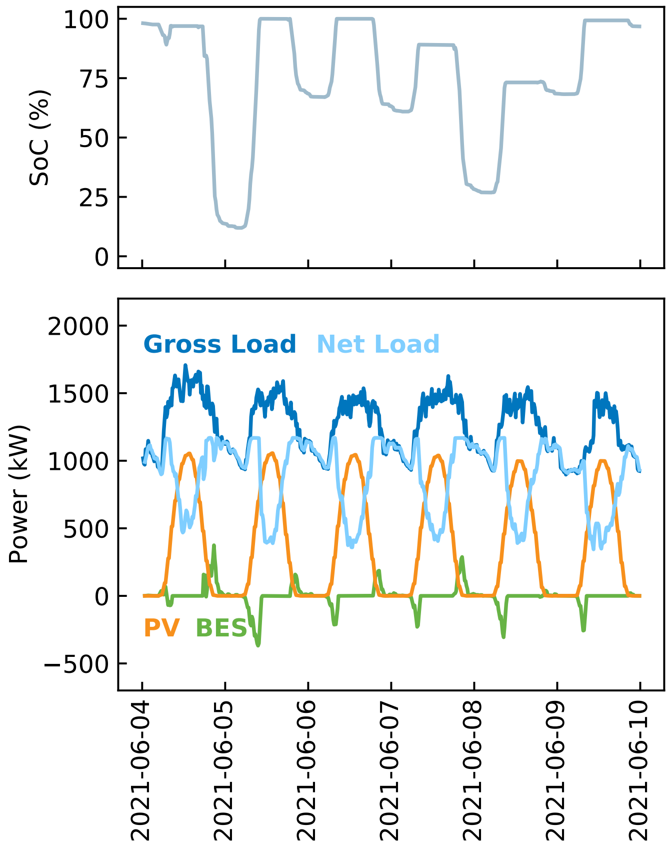

Fig. 5. Time series data showing the PV production, gross load, BES dispatch, net load and BES state of charge (SoC) for six days of a one year simulation.

Table I summarizes the results of the basic statistical metrics, the t-SNE qualitative analysis, and the demand charge management simulations.

IV. Discussion

The generator methods in this manuscript synthesize either GHI or PV production data based on the type of reference data used. The methods treat days as independent units, so any time-of-day details in the reference data set will be preserved in the output data. To facilitate comparison with NSRDB data, we ran the generator models with GHI data inputs from the NSRDB median annual insolation year for each location. Although we treated these GHI outputs as PV production in the demand savings simulations (Fig. 5), a more rigorous approach would decompose the GHI data into direct normal and diffuse horizontal irradiance, transpose to plane-of-array, combine with appropriate temperature and windspeed data, and process using a PV power model.

The rule-based methods V1 and V2 showed the expected distribution of annual and monthly production values with somewhat lower variability metrics for locations with more stable daily insolation. For the V1 method, standard deviation of monthly production was lower than the NSRDB reference and the t-SNE cluster also showed limited diversity. However, the V2 method improved the variability metrics relative to

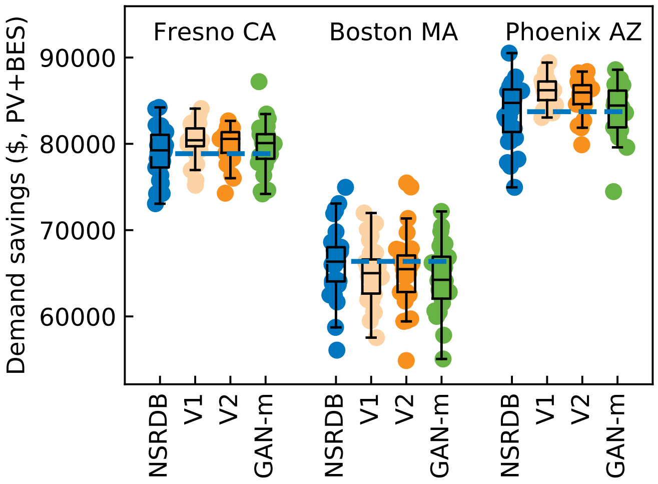

Fig. 6. Demand charge management simulation results showing PV+BES demand savings for a range of PV production scenarios based on 4 data sources and 3 locations. The horizontal blue line shows the mean demand savings for simulations using PV production based on the NSRDB reference.

V1 and increased the standard deviation of demand savings. Although we found an improvement in the sample diversity from V1 to V2, further increasing the swap operations did not bring variability metrics closer to the NSRDB reference. The V2 method has the benefit of being easily audited, capable of producing realistic variability in PV production, and good for generating a reasonable distribution of demand savings.

The ’GAN-m’ method generally showed accurate annual and monthly production values with some differences caused by low geographic detail in the model. Although it did not impact the demand savings distribution, the 3.8% higher annual insolation for Fresno from ’GAN-m’ was caused by greater incidence of clear periods during winter months that is more characteristic of desert southwest locations. The matching process from methods ’GAN’ to ’GAN-m’ acted as a rolling min-max filter on the raw outputs and improved the correspondence with annual and monthly sums versus measured data. The t-SNE analysis showed diversity and variety in the outputs as high or greater than the NSRDB reference. When used in demand charge management simulations, the ’GAN-m’ method generated the expected distribution of demand savings.

Initial designs of the GAN model were conditioned on 365 daily production values, but failed by mode collapse and decreasing weights for latent space inputs. Subsequent designs of the GAN model were conditioned on latitude-longitude and showed stable convergence with good diversity in outputs. Models based on 1-D convolutional layers outperformed models based on dense layers, with dense layers resulting in 3-day autocorrelation values lower than the NSRDB reference.

Demand savings based on generated data showed similar median values within each location, differentiating between Phoenix with the highest savings, Boston the lowest, and Fresno intermediate. This trend reflects generally that higher PV production leads to higher demand savings, all else being equal. The V2 method showed demand savings standard deviation within 2% absolute of the NSRDB reference, and the ’GAN-m’ method showed demand savings standard deviation within 0.8% absolute of the NSRDB reference.

TABLE I
METRICS FOR MEASURED AND GENERATED DATA*

Metric	Location	NSRDB	V1	V2	GAN	GAN-m
$\overline{G}_{yr}$	Fresno	1884	1887	1887	1936	1957
	Boston	1429	1421	1422	1372	1416
	Phoenix	2090	2099	2100	2046	2096
$\overline{\sigma}_{G_{mo}}$	Fresno	7.5	2.4	3.9	7.4	5.9
	Boston	7.1	3.3	5.2	6.8	6.4
	Phoenix	5.2	2.5	3.1	6.5	5.4
A_{tSNE}	Fresno	657	75	290	1123	1173
	Boston	507	48	237	1110	833
	Phoenix	811	119	276	1235	1227
σ_{sav}	Fresno	3.7	2.5	2.5	-	3.4
	Boston	6.2	5.2	6.5	-	5.9
	Phoenix	4.3	1.7	2.3	-	5.1

*where $\overline{G}_{yr}$ is average annual insolation in (kWh/m^2), $\overline{\sigma}_{G_{mo}}$ is average monthly standard deviation of daily insolation in relative percent, A_{tSNE} is the area of the 2σ covariance confidence ellipse after 2-dimensional t-SNE, and σ_{sav} is the standard deviation of simulated PV+BES demand savings in relative percent.

V. Conclusion

Multi-year forecasts of PV production are an important component of economic assessment for behind-the-meter PV+BES (photovoltaics plus battery energy storage) systems. Using a reference PV production time series and latitude-longitude, we present both rule-based and generative adversarial network (GAN) methods to generate PV production variants covering the range of likely scenarios over a 10-20 year planning horizon. The method development was based on NSRDB global horizontal insolation data for sites across the continental United States. The resulting diversity, variability, and distribution of demand savings from synthetic PV production data matches the expected characteristics of historical weather patterns.

References

[1] B. P. Bhattarai, K. S. Myers, and J. W. Bush, “Reducing demand charges and onsite generation variability using behind-the-meter energy storage,” in *2016 IEEE Conference on Technologies for Sustainability (SusTech)*, pp. 140–146, Oct. 2016.

[2] A. Habte, M. Sengupta, and A. Lopez, “Evaluation of the National Solar Radiation Database (NSRDB): 1998-2015,” tech. rep., NREL (National Renewable Energy Laboratory (NREL), Golden, CO (United States)), 2017.

[3] N. R. Darghouth, G. Barbose, J. Zuboy, P. J. Gagnon, A. D. Mills, and L. Bird, “Demand charge savings from solar PV and energy storage,” *Energy Policy*, vol. 146, p. 111766, Nov. 2020.

[4] C. A. Gueymard and S. M. Wilcox, “Assessment of spatial and temporal variability in the US solar resource from radiometric measurements and predictions from models using ground-based or satellite data,” *Solar Energy*, vol. 85, pp. 1068–1084, May 2011.

[5] A. Habte, M. Sengupta, C. Gueymard, A. Golnas, and Y. Xie, “Long-term spatial and temporal solar resource variability over America using the NSRDB version 3 (1998–2017),” *Renewable and Sustainable Energy Reviews*, vol. 134, p. 110285, Dec. 2020.

[6] G. Kimball, C. Chaudhari, P. Keelin, J. Dise, M. Grammatico, and B. Bourne, “Improved model of solar resource variability based on aggregation by region and climate zone,” in *IEEE Photovoltaics Specialists Conference*, pp. 2571–2574, June 2018.

[7] Y. Chen, Y. Wang, D. Kirschen, and B. Zhang, “Model-Free Renewable Scenario Generation Using Generative Adversarial Networks,” *IEEE Transactions on Power Systems*, vol. 33, pp. 3265–3275, May 2018. Conference Name: IEEE Transactions on Power Systems.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” 2014.

[9] J. Yoon, D. Jarrett, and M. v. d. Schaar, “Time-series Generative Adversarial Networks,” *NeurIPS Proceedings*, Sept. 2019.

[10] Z. Lin, A. Jain, C. Wang, G. Fanti, and V. Sekar, “Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions,” *Proceedings of the ACM Internet Measurement Conference*, pp. 464–483, Oct. 2020. arXiv: 1909.13403.

[11] L. v. d. Maaten and G. Hinton, “Visualizing Data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.