Machine Learning for Global Residential-Commercial-Industrial Load Decomposition

William R. Wade*, Richard Asiamah*, Jean-Paul Watson†, and Daniel K. Molzahn*
*School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
†Lawrence Livermore National Laboratory, Livermore, CA, USA

***Abstract*—Synthetic power grids are impactful public tools for advancing research, informing policy, and facilitating education. A critical aspect of creating realistic synthetic grid models is modeling electricity demand across various regions. Electricity demand and its characteristics can vary greatly depending on which economic sector is dominant in the area: residential, commercial, or industrial (RCI). This type of data, however, is generally not publicly available: it is either not published by electric utilities around the world or not recorded in many regions. This paper presents a method for estimating RCI load-share percentages using a deep neural network model trained on several available datasets. Our results indicate that the model is, on average, accurate to within 6.51% mean absolute error for counties in the United States, with some significant outliers in the industrial share predictions. We discuss improvements in future work to enhance the model's ability to accurately predict RCI composition in electric demand.**

***Index Terms*—Synthetic Power Grids, RCI Decomposition, Neural Network, Global Load Modeling**

I. Introduction

Modern electric power grids are large networks that span thousands of miles, connecting millions of people. They are the source of power for most equipment and devices used in our daily lives and are, as such, considered critical infrastructure. A necessary result of this is that much of the data about them is kept confidential [1]. To advance research despite the challenge of obtaining real-world grid models, researchers have instead created synthetic but realistic models of existing power grids based upon publicly available data. These models enable researchers to evaluate optimal power flow solutions, load growth, interconnection locations, resilience studies, etc., all while mitigating security and confidentiality concerns [2].

Synthetic grid creation has advanced significantly over the years, with major repositories providing datasets of synthetic grid models for most parts of the United States and Europe [3], [4]. Despite these advancements, one major challenge in creating effective synthetic grid models is estimating the distribution of electricity demand types across the network. Early synthetic grid work in [5] estimates this demand by using population data only. The major implicit assumption here is that electricity demand is directly correlated to the number of people living in a particular area. However, in real power grids, electric demand in a given location is dependent not only on population, but also on the economic sector contributions, i.e., residential, commercial, or industrial, that are prevalent in that area. References [6] and [7] provide a more accurate method of modeling the nodal loads in a synthetic grid network by using the residential, commercial, and industrial (RCI) composition at the individual nodes. Furthermore, residential, commercial, and industrial loads each have distinct daily load profiles, as shown in Figure 1, which is critical for some applications.

Although this detailed modeling method works well, this information is seldom recorded or made publicly available, and disproportionately so for non-Western power grids, leaving researchers and policymakers with only historical or population-based estimates to inform future demand projections. Past analyses of these population-only methods have shown they are inadequate for developing synthetic grids [8].

This paper addresses these limitations by developing a method for predicting the residential, commercial, and industrial composition of electricity demand across various parts of an electricity network. We train a deep neural network on counties in the United States using publicly available RCI data, to be used in other regions where this data may not be available. In addition to population data, we use other publicly available parameters and metrics that could indicate industrial and commercial activity. More specifically, we train the model using satellite-derived proxies that capture key physical indicators of economic activity, such as nighttime light intensity, atmospheric concentrations of SO_2, CH_4, and NO_2, land surface temperatures, vegetation indices, and measures of built-up land. We introduce composite predictors to improve the model's predictive performance. Our results show that the neural network prediction performed fairly well, achieving a mean absolute error of 6.5% across the testing set. The model struggled most with predicting the industrial sector's share, resulting in a few outliers.

The remainder of the paper is organized as follows. Section II outlines the methods used in obtaining our training data and describes the structure of our neural network. Section III presents an evaluation of the model, and Section IV offers a detailed discussion of the results and applications of this model Finally, Section V concludes the paper, outlining future work.

II. Methodology

This section outlines the steps for predicting the RCI composition across regions. This includes the approach to

This work was performed in part under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DEAC52-07NA27344 and was supported by the LLNL LDRD Program under Project 25-SI-007.

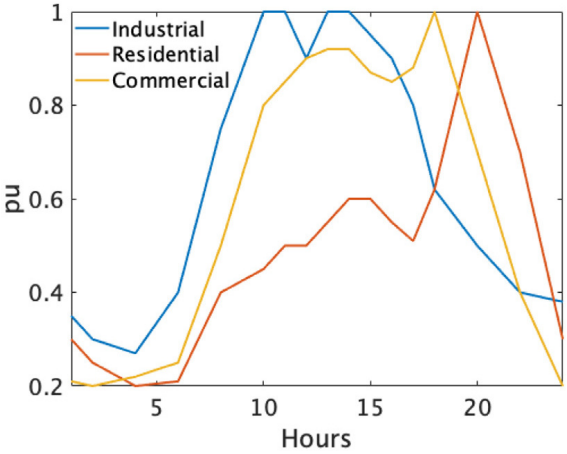


Fig. 1: Template load curves for residential, commercial, and industrial electric load demands over a 24-hour period in Switzerland [7].

obtaining the training dataset, descriptions of the global input datasets and their sources, and ends with a description of the neural network model that predicts global RCI compositions.

A. Compiling RCI Data

The RCI training data for this model was obtained from the United States Energy Information Administration (EIA), using the 2020 version of the annual EIA-861 report [9]. The year 2020 was used for all datasets in this paper because it is the most recent year for which all datasets were available. This EIA-861 report presents a wide range of data from all electric utilities in the United States, including service territories, customer sales, mergers, reliability, operational data, efficiency, and more. The data we focus on are the annual energy demand share (in MWh) supplied to each residential, commercial, and industrial customer, as provided in the `Sales_Ult_Cust.xlsx` file, and the utility service territory by county, as provided in the `Service_Territory.xlsx` file. Figure 2 shows an example of this utility service area data overlapped with county boundaries for electric co-ops in the US state of North Carolina.

The data from the EIA report forms the basis for developing RCI data for each county across the United States. We do this using the steps below:

1) We first validate the provided list of utilities. This is done by cross-referencing the list of utilities across the two files, i.e., the energy demand file and the service territory file. By using only utilities listed in both files, we eliminate utilities that served only specific industries or had dedicated loads.
2) Calculate utility populations by summing the populations of all counties [11] listed in the utility's service area.
3) Determine the county weights using the equation below. If a county is served by multiple utilities, it will have a different weight for each.

$$\text{County weight} = \frac{\text{County population}}{\text{Utility population}}.$$

4) Multiply county weights by the respective residential, commercial, and industrial energy demand in MWh provided by each utility.
5) For counties served by multiple utilities, sum these results to aggregate all utilities serving that county.

The result of the steps outlined above is the energy demand (in MWh) supplied to residential, commercial, and industrial loads at the county level. Completing this process provides us with RCI data for 3022 counties in 2020. Notably, 3143 counties make up the United States; however, some did not meet the requirement in Step 1 above and were therefore excluded from the training dataset.

B. Global Data Sets and Input Parameters

A suitable factor for predicting RCI composition should not only be a good indicator of RCI composition but also be publicly available globally to enable its application across many regions. This section describes these quantities and provides details on their sources and applications in predicting the RCI composition across various regions. We extracted the quantities from gridded global datasets derived using satellite observations produced by space agencies, with the data processed into a latitude-longitude based grid. Each parameter was chosen because it demonstrates physical differences across regions dominated by residential, commercial, or industrial loads. This gives us 11 primary quantities that form the basis of our RCI prediction.

To improve model prediction, some parameters were log-transformed to compress the wide range of data. This was only necessary for data sets with wide ranges, such as population, while others with smaller ranges, such as land surface temperature, were passed directly into the model. Table I provides the full details for all of these parameters.

The RCI data were transformed into county-level data; as such, it was necessary to do the same for the gridded global datasets in Table I. To do this, shapefiles from the US Census Bureau [23] were used to map grid cells at multiple resolutions to US counties. After this mapping process was completed, accuracy was verified by comparing the datasets with readily available public data. For example, derived county populations were checked to ensure they closely matched the official US census values [11].

C. Composite Inputs

To further boost our model accuracy, we applied some feature engineering techniques [24]. Our primary parameters were used to construct seven additional "composite" parameters. These composite parameters were determined intuitively as combinations of variables that indicate a high presence of residential, commercial, or industrial load in a given area. Each proposed composite parameter was added to the model and retained only if it improved performance.

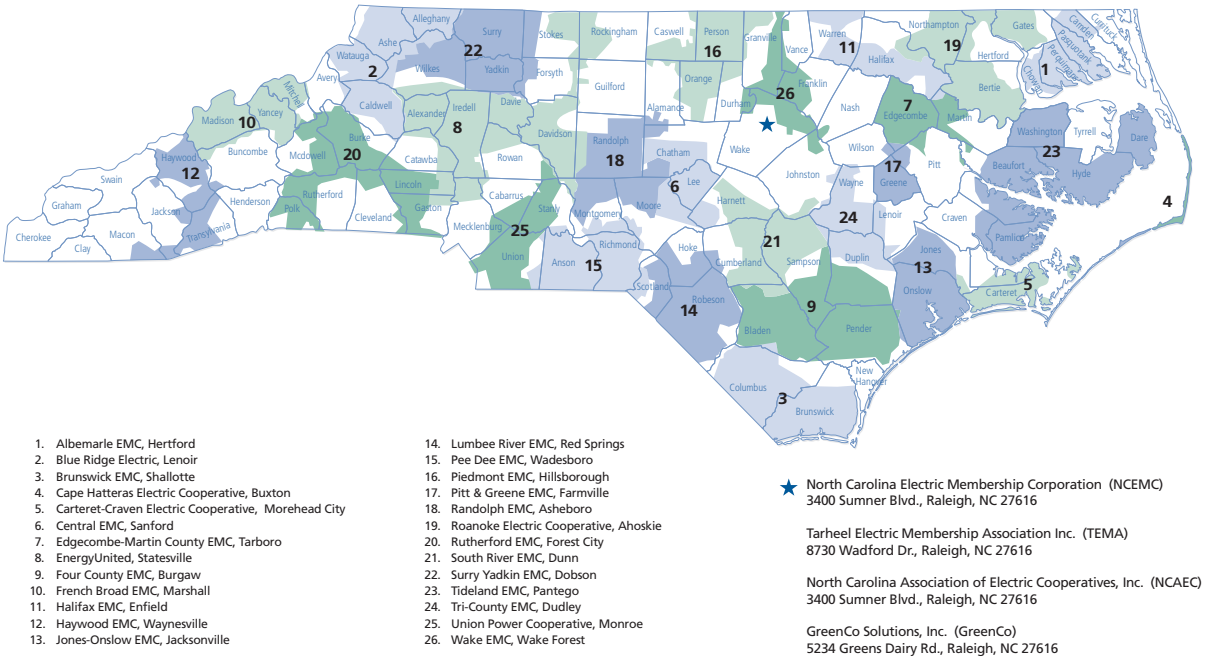


Fig. 2: Electric co-op utility area overlapped with county boundaries in North Carolina [10]. Although only electric co-ops are pictured, the analysis in this paper considers all electric utility providers.

TABLE I: Description of all global datasets used in predicting RCI composition. These parameters serve as the primary inputs to the neural networks.

Primary Parameter	# of Inputs	Source	Resolution (km)	Description [Units]
(1) Population	1	Copernicus GHS [12]	1	Log of Population per 1 km grid cell [Persons]
(2) Built up surface area	1	Copernicus GHS [13]	0.1	log of Built up surface area per grid cell, $[m]$
(3)Degree of Urbanization	1	Copernicus GHS [14]	1	Urban / Rural Classification from [10 to 30] (Normalized 0 to 1)
(4) Nighttime light intensity	1	NASA- VIIRS [15]	0.5	Average & total nighttime light intensity, [nW/cm^2 $\cdot$ sr] y
(5) Land Surface Temperature	3	NASA - MODIS [16]	1	Day, night, and diurnal land surface temperature, $[K]$
(6) Vegetation Index	1	NASA - MODIS [17]	1	Average of the measure of canopy greenness from [-0.2 to 1]
(7) Land Cover Class	5	NASA - MODIS [18]	0.5	Binary land cover classifications (Using 4,8,9,10,17)
(8) Atmospheric NO_2 concentration	1	Sentinel-5p [19]	3.5	Log of average NO_2 atmospheric concentration $[mol/m^2]$)
(9) Atmospheric SO_2 concentration	1	Sentinel-5p [20]	3.5	Log of average SO_2 atmospheric concentration $[mol/m^2]$)
(10) Atmospheric CH_4 concentration	1	Sentinel-5p [21]	7	Log of average CH_4 atmospheric concentration $[mol/m^2]$)
(11) GDP	1	Kummu GDP [22]	1	Log of total GDP per cell [USD]

For example, one of these parameters was rural classification multiplied by nighttime light intensity, which could be a stronger indicator of an industrial area than nighttime light intensity alone, since it can also reflect commercial activity. Similarly, NO_2 concentration multiplied by rural classification was used to better differentiate between industrial and residential areas. High NO_2 concentrations are caused by burning fuel, a common occurrence in both industrial and residential areas around cities where traffic builds up. The introduction of these composite parameters mitigates potential confusion between high-traffic residential areas and industrial areas in rural cases. Including composite parameters introduces additional computational overhead and can pose challenges such as overfitting, but provides better overall performance. More details on this are provided in Section III. All composite input parameters are shown in Table II.

TABLE II: Final list of composite parameters

Composite Parameter Description	Primary Parameters
Log(Avg. Night Light per capita)	(4), (1)
Log(GDP per capita)	(11), (1)
NO_2 $\times$ (1 - Degree of Urbanization)	(8), (3)
SO_2 $\times$ (1 - Degree of Urbanization)	(9), (3)
CH_4 $\times$ (1 - Degree of Urbanization)	(10), (3)
Nighttime Light Intensity $\times$ (1 - Degree of Urbanization)	(4), (3)
Vegetation Index $\times$ Degree of Urbanization	(6),(3)

D. Neural Network Model

As shown in Figure 3, the deep neural network utilized in this model has six hidden layers with 256, 128, 128, 64, 64, and 48 neurons, trained using the Adam [25] optimizer set at a learning rate of 0.001. We apply early stopping if the loss does not improve after 80 training epochs, with a maximum of 400 epochs. The batch size used is 32, and L2 normalization [26] is applied to all layers within the neural network.

The input layer of this neural network has 24 neurons

to account for the 17 primary and 7 composite features as outlined in Tables I and II. This input layer is normalized to have a mean of zero and unit variance before it is passed to a dense 256-dimensional layer, whose size was chosen to balance accuracy and limit overfitting on the small training set of 2417 counties. After each layer of this neural network, batch normalization, ReLU, and a dropout of 30% are applied. We apply these at each layer to minimize noise and prevent overfitting by setting many perceptron outputs to zero during training.

The output layer of this model contains three neurons, one dedicated to each economic sector, R, C, and I, predicting the RCI load share percentage for its sector. This output head was trained using Kullback–Leibler divergence [27], with weights of 1.025 for residential and commercial and 1 for industrial. Softmax was applied to ensure that the three output load-share percentages summed to 100%. We split the data of 3022 samples, allocating 80% to training and 20% to testing, yielding 2417 counties in the training set and 605 in the test set.

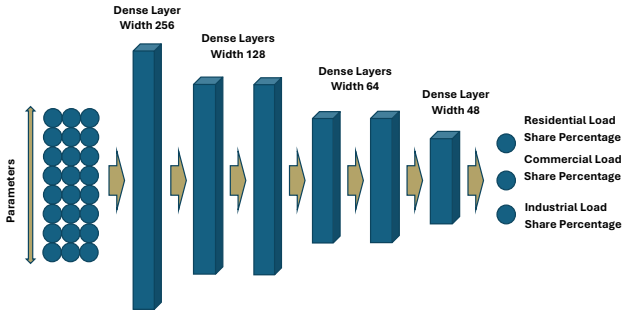


Fig. 3: Elementary representation of the deep neural network model architecture.

III. Model Evaluation

This section outlines the evaluation of the deep neural network model across key performance metrics for training and testing.

A. Setup

This neural network model was created in Python. Keras [28], and Scikit-Learn [29] were both used in the model. Rasterio [30], and Rasterstats [31] were used to extract points from the gridded global data sets. Computing was performed on a 64-bit Windows 11 operating system with 16 GB of RAM and a 12th Gen Intel Core (1.70 GHz).

B. Key Metrics

The load-share predictions produced by our model are evaluated using two key metrics. The first is Mean Absolute Error measured in percentage points (MAE (pp)), which is the mean of the absolute differences between predicted and actual percentage values. Formally, this metric is:

$$\text{MAE (pp)} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|\,, \tag{1}$$

where n is the number of data points, y_i is the actual value, and $\hat{y}_i$ is the predicted value. Second, R^2 values were used to measure the correlation of the predictions with the actual values.

C. Training and Testing Results

In this section, we present our neural network's performance on our dataset. The model's early stopping was triggered at epoch 280, and weights were reverted to epoch 200. Results of the application of the model to the 605 county training set are given in Table III. As shown in Table IV, the increase in error from training to testing is moderate, indicating some overfitting in this model. The test set R^2 value of around 0.5 is reasonable, given the noise inherent in the data, suggesting that the model is likely memorizing patterns during training. This issue has been difficult to address, especially in the industrial sector, as many industrial facilities have on-site generation [32].

TABLE III: Test set performance metrics

Metric	Residential	Commercial	Industrial	Total
R^2	0.599	0.377	0.530	0.503
MAE (pp)	6.25	5.26	8.02	6.51

TABLE IV: Generalization gap between training and test sets

Metric	Residential	Commercial	Industrial	Total
ΔR^2	0.278	0.384	0.323	0.331
ΔMAE (pp)	2.39	1.80	3.57	2.59

To further dissect the results, we present the histogram of the error range shown in Figure 4, which indicates that 78.7% of predictions have errors less than 10%. This suggests that significant outliers are driving these metrics. These outliers largely stem from the industrial sector, highlighting the diverse environmental impacts across industries that pose challenges for this proxy-based model.

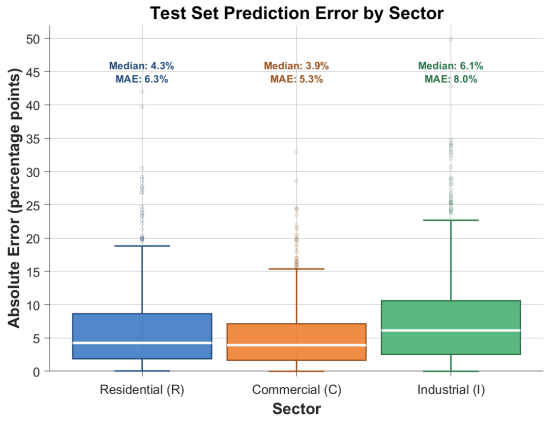
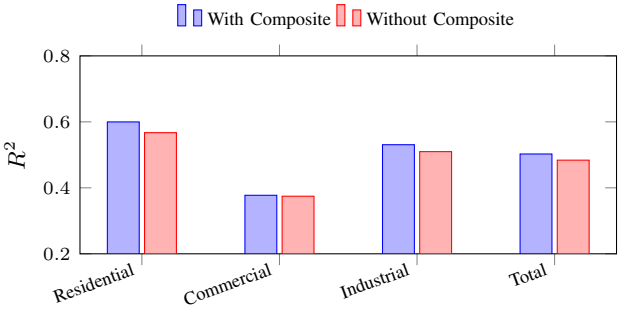
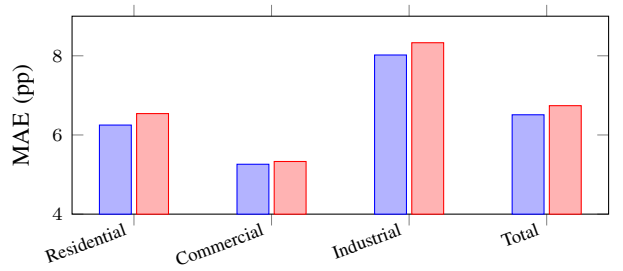


Fig. 4: Box plot showing the lower quartile, upper quartile, mean, maximum, and outliers for residential, commercial, and industrial, for the test set predictions



(a) R^2 comparison



(b) MAE (pp) comparison

Fig. 5: Neural network model performance comparison with and without composite features.

D. Importance of Composite Parameters

The addition of composite parameters as inputs to the model could have adverse effects, such as slower convergence and increased computing burdens. To verify whether these extra inputs improve model performance, they were removed, and performance was evaluated. As demonstrated in Figure 5, when composite parameters are removed from the model, R^2 values decrease and $\mathrm{MAE}\,(\mathrm{pp})$ values increase. This signals that these composite parameters do improve model performance and confirms that they will be kept as inputs.

E. Feature Importance Metric

Finally, we investigate the importance of each of the 24 inputs we have to the neural network. Unlike classic regression, where the criticality of input parameters can be determined by their coefficients, we employ a technique called permutation feature importance [33] in neural networks. For this model, we randomize one feature, effectively turning it into noise, and then pass the input parameters through the trained network to see the increase in MAE (pp). This process is repeated for each input, and the resulting measure of importance is unique to this model. Removal of even a seemingly unimportant feature can change how the model learns, thereby drastically altering the importance measured by this method.

As shown in Figure 6, the most important features are: Log(Built Surface Area), Average Vegetation Index, and Nighttime Surface Temperature. The least important are: Degree of Urbanization, CH_4 Concentration $\times$ Degree of Urbanization, and various land cover classes. Population importance is not among the top 10 using this evaluation method; however, its impact remains large, as two other composite parameters

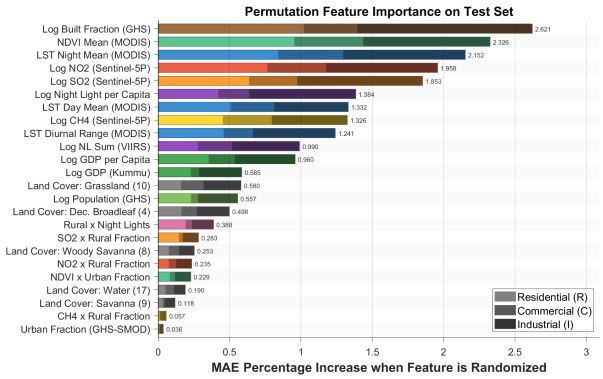


Fig. 6: Importance of each input feature using a permutation-based evaluation method.

depend on it. This highlights the complexity of RCI load decomposition and demonstrates the importance of population as a factor in determining the scale of each sector.

This feature-importance method was also used in developing the model and determining its input parameters. Many possible proxies tested using this method were found to be counterproductive to model performance and were, as such, removed. All proposed composite features were tested using this method, and only those listed in Table II had a positive impact on the model.

IV. Discussion

The neural network model developed in this paper demonstrates good performance in predicting the RCI composition of various counties within the United States. The biggest challenge came from predicting the composition of industrial demand. The large training-to-test gap for this sector's MAE (pp) indicates the model is likely overfitting, memorizing sector-specific patterns that do not fully generalize, and this is likely driven by incomplete and noisy data, especially concerning on-site generation for industrial facilities. Additionally, the parameters used to characterize industrial activity (NO_2, SO_2, CH_4) can vary widely between facilities depending on the industrial process, fuel source, and pollution control technology. This is important when comparing developed and developing nations as differences in these factors can result in different emission patterns not directly correlated to energy usage.

With reference to the model's global applications, there is much room for improvement. Currently, to the best of our knowledge, there are no publicly available sources that report the second-level subdivision RCI composition for any other countries, so we cannot directly compare the neural network's performance. This absence creates further issues in understanding the effects of on-site generation for this model globally, but also prevents the training set from expanding internationally, thereby tying the model's predictions to the economic and social patterns of the United States. Other developing countries would not have levels of industrial or commercial activity comparable to those in the United States.

Furthermore, this neural network model cannot account for differences in electrification rates and energy consumption across regions. These can vary greatly across countries and are critical for demand prediction. Using these two characteristics as input parameters for this model will not only improve the neural network's accuracy but also reduce the effort required to apply it to any country.

V. Conclusions and Future Work

This paper investigates the ability of satellite-derived proxies to accurately estimate the load share percentages of the real power grid across residential, commercial, and industrial economic sectors. Using gridded global datasets as input to a deep neural network, we trained the model to predict RCI load share percentages that utilities often do not track or keep confidential.

After training on 2417 US counties, the model was tested on another 605. In this test, the neural network model predicted load-share percentages with an average error of 6.51% and achieved an overall R^2 of 0.503. Although the model performed fairly well, it struggled with predicting the industrial load share across multiple samples. We also identified the most important features for determining the RCI composition of any region in the United States as the Built Surface Area, the Vegetation Index, and the Nighttime Surface Temperature.

Our future work will focus on obtaining or creating realistic RCI data worldwide to further assess the accuracy of this model's global application. With the expansion of global RCI data, the input parameter set can be extended to include per-capita energy usage, electrification rate, nodal degree of the transmission network, and other parameters that cannot currently be used, as the United States is the sole source of training data. This expansion will allow for more accurate model projections across a wider range of nations and electric grids.

References

[1] Federal Energy Regulatory Commission, "Critical Energy/Electric Infrastructure Information." https://www.ferc.gov/ceii, May 2025.

[2] M. Hampson, "Synthetic Power Grid Is Safer Than the Real Thing." https://spectrum.ieee.org/power-grid, 2023. IEEE Spectrum, Accessed: 2026-04-24.

[3] Texas A&M University, "Electric Grid Test Case Repository." Online, 2024. https://electricgrids.engr.tamu.edu, Accessed Sep. 8, 2024.

[4] IEEE PES Task Force on Benchmarks for Validation of Emerging Power System Algorithms, "The Power Grid Library for Benchmarking AC Optimal Power Flow Algorithms," *arXiv:1908.02788*, 2021.

[5] K. M. Gegner, A. B. Birchfield, T. Xu, K. S. Shetye, and T. J. Overbye, "A Methodology for the Creation of Geographically Realistic Synthetic Power Flow Models," in *7th IEEE Power and Energy Conference at Illinois (PECI)*, 2016.

[6] H. Li, A. L. Bornsheuer, T. Xu, A. B. Birchfield, and T. J. Overbye, "Load Modeling in Synthetic Electric Grids," in *2nd IEEE Texas Power and Energy Conference (TPEC)*, 2018.

[7] R. Gupta, F. Sossan, and M. Paolone, "Countrywide PV Hosting Capacity and Energy Storage Requirements for Distribution Networks: The case of Switzerland," *Applied Energy*, vol. 281, October 2020.

[8] R. Asiamah, K. Ji, J.-P. Watson, and D. K. Molzahn, "Sensitivity Analyses of Synthetic Power Grid Modeling Techniques in a Global Context: A Case Study in Ghana," *59th Hawaii International Conference on System Sciences (HICSS)*, January 2026.

[9] U.S. Energy Information Administration, "Form EIA-861 Annual Electric Power Industry Report, 2020." https://www.eia.gov/electricity/data/eia861/, 2020. Accessed: 2026-04-01.

[10] North Carolina Electric Membership Corporation, "North Carolina's Electric Cooperatives Service Territory Map." https://web.njuns.com/wp-content/uploads/2014/08/NC-Co-ops-Territory-Map-1.pdf, 2014. Accessed: 2025.

[11] U.S. Census Bureau, "County Intercensal Population Totals: 2010–2020." https://www.census.gov/data/tables/time-series/demo/popest/intercensal-2010-2020-counties.html, 2024. Accessed: 2026-05-03.

[12] European Commission, Joint Research Centre (JRC), "Global Human Settlement Layer – Population Grid (GHS-POP) Dataset, 2020." https://human-settlement.emergency.copernicus.eu/download.php?ds=pop, 2020. Accessed: 2026-04-01.

[13] European Commission, Joint Research Center (JRC), "Global Human Settlement Layer – Built-Up Surface Grid (GHS-BUILT-S), 2020, 100 m Resolution." https://human-settlement.emergency.copernicus.eu/download.php?ds=bu, 2020. Accessed: 2026-04-01.

[14] European Commission, Joint Research Centre (JRC), "Global Human Settlement Layer – Settlement Model Grid (GHS-SMOD), 2020, 1 km Resolution." https://human-settlement.emergency.copernicus.eu/download.php?ds=smod, 2020. Accessed: 2026-04-01.

[15] Earth Observation Group, Colorado School of Mines, "VIIRS Nighttime Lights Annual Composite, Version 2.0 (2020)." https://eogdata.mines.edu/nighttime_light/annual/v20/2020/, 2020. Accessed: 2026-04-01.

[16] Z. Wan, S. Hook, and G. Hulley, "MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V061," 2021. Accessed via Google Earth Engine, 2026-04-01.

[17] K. Didan, "MODIS/Terra Vegetation Indices Monthly L3 Global 1km SIN Grid V061," 2021. Accessed via Google Earth Engine, 2026-04-01.

[18] M. Friedl and D. Sulla-Menashe, "MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061," 2022. Accessed via Google Earth Engine, 2026-04-01.

[19] European Space Agency (ESA) and Copernicus Sentinel-5P Mission, "Sentinel-5P TROPOMI Tropospheric NO2 Column Density (L3), 2020." https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_NRTI_L3_NO2, 2020. Accessed via Google Earth Engine, 2026-04-01.

[20] European Space Agency (ESA) and Copernicus Sentinel-5P Mission, "Sentinel-5P TROPOMI Tropospheric SO2 Column Density (L3), 2020,."

[21] European Space Agency (ESA) and Copernicus Sentinel-5P Mission, "Sentinel-5P TROPOMI Methane (CH4) Column Mixing Ratio (L3 OFFL), 2020." https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_CH4, 2020. Accessed via Google Earth Engine, 2026-04-01.

[22] M. Kummu, M. Taka, and J. H. A. Guillaume, "Data From: Gridded Global Datasets for Gross Domestic Product and Human Development Index over 1990–2015," 2020.

[23] U.S. Census Bureau, "TIGER/Line Shapefiles: County and Equivalent Entities, 2025." https://www2.census.gov/geo/tiger/TIGER2025/COUNTY/, 2025. Accessed: 2026-04-01.

[24] J. Murel and E. Kavlakoglu, "What is Feature Engineering?." https://www.ibm.com/think/topics/feature-engineering, 2024. Accessed: 2025.

[25] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980*, 2017.

[26] Google for Developers, "Overfitting: L2 Regularization." https://developers.google.com/machine-learning/crash-course/overfitting/regularization. Accessed: 2026-05-25.

[27] DataCamp, "KL-Divergence Explained: Intuition, Formula, and Examples." https://www.datacamp.com/tutorial/kl-divergence. Accessed: 2026-05-25.

[28] Keras Team, "Keras." https://keras.io, 2015. Accessed: 2026-04-24.

[29] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] Rasterio Developers, "Rasterio: Geospatial Raster I/O for Python." https://rasterio.readthedocs.io, 2024. Accessed: 2026-04-24.

[31] Matthew Perry, "Rasterstats: Zonal Statistics for Raster Data." https://pythonhosted.org/rasterstats, 2024. Accessed: 2026-04-24.

[32] U.S. Energy Information Administration, "Many Industrial Electricity Customers are Farmers." https://www.eia.gov/todayinenergy/detail.php?id=16331, 2014. Accessed: 2026-04-17.

[33] scikit-learn developers, "5.2. Permutation Feature Importance — scikit-learn 1.8.0 Documentation." https://scikit-learn.org/stable/modules/permutation_importance.html, 2025. Accessed: 2025.