Open-Access Data and Toolbox for Tracking COVID-19 Impact on Power Systems

Guangchun Ruan, Zekuan Yu, Shutong Pu, Songtao Zhou,
Haiwang Zhong, Le Xie, Qing Xia, Chongqing Kang,

***Abstract*—Intervention policies against COVID-19 have caused large-scale disruptions globally, and led to a series of pattern changes in the power system operation. Analyzing these pandemic-induced patterns is imperative to identify the potential risks and impacts of this extreme event. With this purpose, we developed an open-access data hub (COVID-EMDA+), an open-source toolbox (CoVEMDA), and a few evaluation methods to explore what the U.S. power systems are experiencing during COVID-19. These resources could be broadly used for research, policy making, or educational purposes. Technically, our data hub harmonizes a variety of raw data such as generation mix, demand profiles, electricity price, weather observations, mobility, confirmed cases and deaths. Several support methods and metrics are then implemented in our toolbox, including baseline estimation, regression analysis, and scientific visualization. Based on these, we conduct three empirical studies on the U.S. power systems and markets to introduce some new solutions and unexpected findings. This conveys a more complete picture of the pandemic's impacts, and also opens up several attractive topics for future work. Python, Matlab source codes, and user manuals are all publicly shared on a Github repository.**

***Index Terms*—Extreme event, data-driven assessment, power system operation, electricity market, open-source**

I. Introduction

A. Background

THE COVID-19 pandemic is a once-in-a-century crisis for the globe, causing 181.5 million infections and nearly 4 million deaths until the first half of 2021 [1]. Governments worldwide took a wide range of non-pharmaceutical interventions in response to the pandemic [2], and as a result, these restrictions and lockdowns have significantly changed the electricity consumption patterns, and had a domino effect through the entire power systems. Although the power sector has long prepared against a few predictable threats [3], such kind of large-scale, long-term, and high-intensity interference is still quite unique.

A systematic perspective [4] and the empirical studies [5] are both critical to understand the pandemic's impacts on power systems. In fact, COVID-19 has opened up an opportunity for power system operators in assessing the abnormal operation patterns, and identifying the future pathways for sustainable recovery [6]. Existing works have discussed part of these topics, but the complete picture is still unclear.

G. Ruan, Z. Yu, S. Pu, S. Zhou, H. Zhong, Q. Xia, and C. Kang are with the State Key Lab of Power Systems, Department of Electrical Engineering, Tsinghua University, Beijing 100084, China.

L. Xie is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA.

This motivates us to record the potential data resources during COVID-19, and develop a support toolbox with adequate built-in methods and metrics for different groups of people, such as scholars (with diverse backgrounds), policy makers, educators, students, and the general public.

Such an idea came true as a joint project initiated in May 2020 by Texas A&M University and Tsinghua University. We soon received constructive feedback from the power community, and until now, our work has been successfully applied in several research works and university courses.

B. Literature Review

Recent advances in the literature have increased our understanding of COVID-19 with some empirical studies worldwide. Typical examples include the observations in France [7], Italy [8], Great Britain [9], the U.S. [10], and Canada [11]. There are extensive works to evaluate the potential impacts in different energy topics, such as power system operation [12], household electricity consumption [13], gasoline demand [14], energy security [15], green recovery [16], and climate change [17].

But according to current progress, the energy community has made very limited efforts to standardize the pandemic-related data and models. While those findings on a case-by-case basis, e.g., [7]–[9], are still informative, it remains highly complicated to reproduce a published work, or make meaningful comparisons among different results (even for the same country). Below are some more illustrations.

1) Data Issue: Many data resources are not available for the public, especially the cleaned or fine-tuned data. In this case, the similar but tedious data prepossessing could repeatedly dominate the research time of everyone, or even worse, academic communication might be interrupted due to data availability. Reference [7] collected the power consumption and meteorological data from the French system operator RTE and Meteo-France, but their cleaned dataset was not shared to the public. Similarly, reference [8], [9], [11] did not directly share their datasets either. Up to now, two of the most popular data sources are the U.S. Energy Information Administration (EIA) [12], and the European Network of Transmission System Operators for Electricity (ENTSO-E) [18]. But users are still required to get familiar with the complex data category and storage rules, and implement all the data preprocessing steps by themselves. For comparison, it would be tougher to get access to some other rare data sources, e.g., the Indian electricity market data [19], or the Swedish building standards and statistics [20].

Another finding is that most scholars (including the above) have rarely expanded their data category to consider some cross-domain data which may inspire interdisciplinary studies.

2) Model Issue: Quite different methods, models, and criteria are applied in different publications, but very few of them provides an open-source license. Benchmarking is so challenging in this condition that one may take a long time to realize even a basic function. This is, of course, not friendly to the public, students, and scholars in other fields. For example, an ordinary least squares model was used in [13] to analyze the online survey data in California. Since a few household characteristics and respondent demographics were mentioned, it would need extra efforts to specify the detailed expressions. Then a join-point regression was applied in [21] to assess the electricity load trends in Brazil and its geographic regions, but the discussion about model details was somehow limited. Reference [22] developed three time-series models to determine the impacts on the Spain electricity market, and reference [23] used a five-year moving average method to establish a non-pandemic scenario. It is a pity that both works [22] and [23] didn't share the codes for public use.

In addition, machine learning approaches become increasingly popular in analyzing the potential impacts on the operation or resilience of power systems [24]. Reference [25] used five classical machine learning approaches for electric load forecast in India. Reference [26] established a random-forest-bagging and board learning system for estimating the daily confirmed cases. Many other learning models were also found to be effective, including deep learning models [27], capsule networks [28], and domain adaptation [29]. Although powerful, these models made it tougher to reproduce or benchmark because of their increasing complexity [30].

3) Open Source Efforts: Open source community has actively involved in combating COVID-19 [31]. Perhaps the most prominent efforts in tracking the pandemic's impacts and sharing open data are made by Johns Hopkins University [1] and Oxford University [2]. In reference [1], an interactive dashboard was developed for all affected countries in real time. And an Oxford COVID-19 Government Response Tracker (OxCGRT) was established in [2] to assess the policy responses of over 180 countries and subnational jurisdictions.

Other efforts include CovidCounties (a public health data tracker at the level of U.S. counties) [32], COVID-ResNet (a radiography scanner) [33], and OpenABM (an agent-based model for non-pharmaceutical interventions) [34]. All these works, however, are mainly conducted in the public health field, with special focuses on the confirmed cases, deaths, government responses, and so on.

One of the few examples from the energy community is reference [35], where the authors have made both their data and codes available on Github. This is a positive step forward, but these resources only cover five months and lack frequent updates. Dynamic data aggregation is thus needed, but different from the NRGStream (a charged service) in [36], the resources are preferred to be fully free for use.

To the best of our knowledge, we are the unique team that develops and constantly upgrades the open-source resources (both data and toolbox) to track the pandemic's impacts on power systems. Not to mention that we have extensively collected the cross-domain data for interdisciplinary studies.

C. Contributions and Paper Structure

This paper has made a special effort to evaluate the potential COVID-19 impacts on power system operations. Here, the major contributions of our work are summarized as follows:

- The proposed data hub and toolbox have unique values for data-driven analysis on power system operations. A variety of (cross-domain) data from power system operation to public health are collected, dynamically updated, and quality-controlled by a support team. The toolbox is built on Python and Matlab to cover most users.
- Several novel evaluation methods and metrics are proposed to adapt the classical power system analysis to a pandemic case. Other typical and popular methods are involved in the toolbox for comparison as well.
- Three empirical studies are conducted on U.S. power systems to introduce some new perspectives, solutions, and unexpected findings. This shows the high potential, sufficient flexibility, and great convenience of the proposed resources.

Notice that we have established a Github repository [37] to launch our data hub and toolbox online before finalizing this paper. The open repository has attracted a special attention from the power community, and supported over 40 research groups or individuals up to now, e.g., a team from Florida State University and New York University [38]. It has also been successfully applied in two graduate courses at Texas A&M University and Tsinghua University.

The remainder of this paper is organized as follows: Section II introduces the overall framework, main features, and several quick start guides. Section III demonstrates the details of data, models, and algorithms, then Section IV discusses the implementation issues in Python and Matlab. Three empirical studies are conducted in Section V. At last, Section VI gives the concluding remarks.

II. Framework

A. Overall Workflow

This paper creates a Github repository [37] that consists of an open-access data hub (COVID-EMDA+) and an open-source toolbox (CoVEMDA). One can access these resources from the directories of "data_release/" and "toolbox/" respectively. Note that COVID-EMDA+ is the abbreviation for "Coronavirus Disease – Electricity Market Data Aggregation+", and CoVEMDA for "CoronaVirus – Electricity Market Data Analyzer".

Fig. 1 demonstrates the latest framework and workflow of the proposed data hub and toolbox.

As shown, the backend system will routinely run the data formatter and quality controller to update the data hub. Outliers and missing data are largely handled with backup data or historical trends, while we also prepare a data quality report to record those highly problematic data.

Fig. 1 has listed out three widely used functions in the toolbox: baseline estimation, regression analysis, and scientific

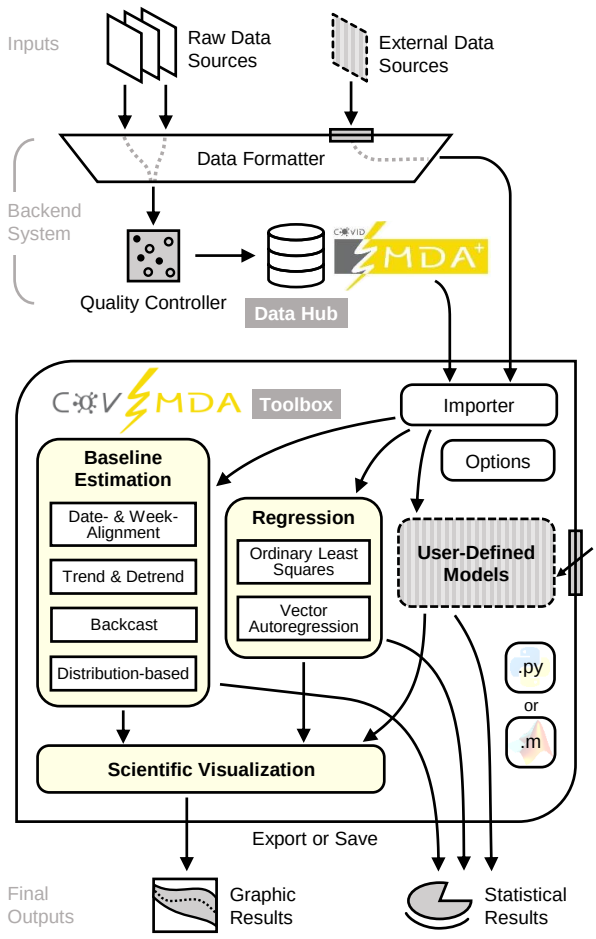

Fig. 1. Overall workflow for the proposed data hub and toolbox. All the processing steps from inputs to outputs are shown, and the main functions and extensions are demonstrated as well.

visualization. Users are allowed to run this toolbox with Python or Matlab consoles, and generate a variety of graphic and statistical outputs if needed.

In addition, external data and user-defined models are all supported, and this provides great flexibility for special or advanced extensions.

The whole system, including the data hub and toolbox, is maintained by a support team from Texas A&M University and Tsinghua University. The routine maintenance includes making regular backups, fixing bugs, handling feedback, upgrading online systems, logging, and so on.

B. Main Features

We summarize the main features of the data hub (COVID-EMDA+) and toolbox (CoVEMDA) as follows:

- Data Resources: Broad data categories to support the classical or cross-domain analysis on power systems.
- Baseline Estimation: With a comprehensive collection of the most typical and popular methods. Rigorous comparisons among different baselines are allowed for different power system measurements.
- Regression Analysis: Two typical and powerful models with built-in statistical tests. Flexible to support multiple kinds of model extensions.
- Scientific Visualization: Tailored designs for various power system applications. Intuitive, convenient and powerful for different users.

C. Getting Started

1) Data and Toolbox Downloads: Users may choose to download or clone the data and toolbox from our Github repository, where all built-in methods are flexible for extensions. Online data entry is allowed when the internet is working.

2) Toolbox Installation: We provide an install script in the root directory to automate the entire installation process. Readers may refer to Section IV and the toolbox manuals for more details.

3) Illustrative Examples: One typical example to show the usage of our data, method, and toolbox is the baseline load estimation in New York City, i.e., the possible electricity consumption without the impacts of COVID-19.

Here is a Python command to make this estimation:

```
City("nyc").cal_demand_baseline()
```

By default, this will return a table with the baselines of power consumption before mid-2020, calculated by the method of date alignment. Find more technical details in Subsection III-C.

III. Data, Models, and Algorithms

A. Data Sources

Our data hub collects raw data from multiple sources: (i) electricity data from all regional system operators (e.g., CAISO for California, NYISO for New York) along with backup data from EIA and EnergyOnline company, (ii) public health data from Johns Hopkins University, (iii) meteorological data from Iowa State University, (iv) mobile device location data (mobility data) from Safegraph company, and (v) satellite image data from NASA (for visualization only). Readers may find all the detailed links for these sources on Github [37].

Most data records in our data hub could be expressed by X_{ymdt}. Here, X is a placeholder for some variable, and the indices collectively specify a time—year y, month m, day d, and hour t. We often use X_{ymd} or X_{ym} to represent different kinds of mean values, for example:

$$X_{ym} = \frac{1}{N_{\{d,t\}}} \sum_{\forall d,t} X_{ymdt} \tag{1}$$

where X_{ym} denotes the mean value of month m in year y. It is derived by averaging X_{ymdt} along the axes d and t, and $N_{\{d,t\}}$ denotes an auxiliary number.

B. Data Structure and Preprocessing

One barrier for merging multiple data sources is the inconsistent data structures. This motivates us to standardize and convert those messy formats to a better one, i.e., the wide data frame.

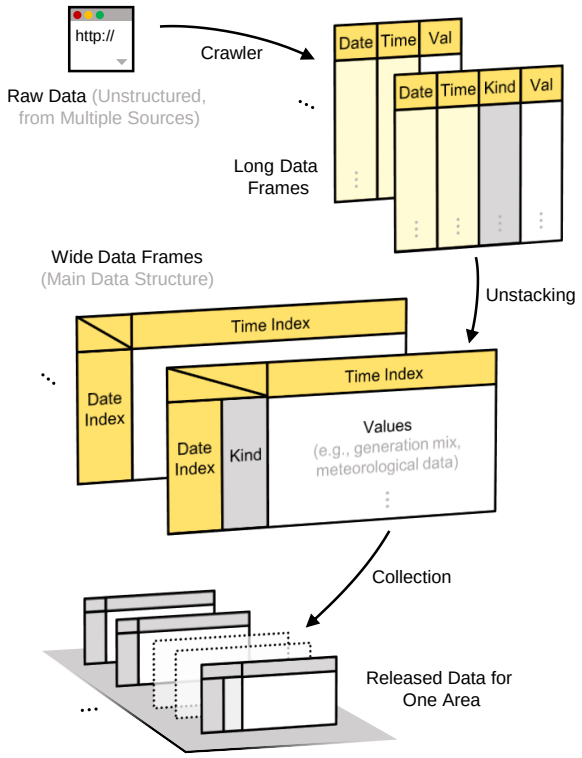

Fig. 2. Demonstration of the proposed data structure and preprocessing steps. This procedure is already automated and executed by the backend system.

Fig. 2 shows the proposed data structure with details. Here, a wide data frame refers to a kind of unstacked table that has more columns than a long frame. This structure enables a more compact way to store data, and both the row-wise and column-wise operations have clear physical meanings. Besides, a variety of basic operations (e.g., filtering, resampling, and statistical computing) have been developed in Python and Matlab to handle such a matrix-like structure.

We store the raw data (e.g., $\{X_{ymdt}\ \forall y, m, d, t\}$) as a wide data frame by assigning a date index (combining axes y, m, and d) to the rows and an hour index (axis t) to the columns.

Fig. 2 also demonstrates how to finalize the released data after several preprocessing steps. These clean and regularly-updated data can be found from the Github repository [37]. Although local and offline data reading is allowed, we strongly recommend online data retrieval by the toolbox (more elegant, no data update concerns).

In fact, all the preprocessing steps have been automated by our backend system which consists of a few web crawlers, a set of automation and management modules, the workflow controller and quality controller, and a logging module. This backend system is scheduled to run periodically, and for each run, 31 raw data files from 25 sources will be extracted and cleaned to update 73 spreadsheets. Here, outliers and missing data are efficiently detected and handled by analyzing the historical trend or backup data—different rules are specialized for different variables. We further record some problematic data (very rare) in a quality control report for ease of reference.

C. Baseline Estimation

Baselines refer to the reference points for comparison. We are focused on estimating a counterfactual situation that assumes the absence of COVID-19. The difference between a counterfactual outcome and an actual observation will naturally substantiate the pandemic's impacts.

Baseline estimation is recognized as the first-and-foremost step for any impact assessments, and a bad baseline may distort our judgment on the impacts' intensity and duration. We next summarize the existing practice to select four popular methods which may hopefully cover most applications.

1) Date- and Week-Aligned Estimation: This method is simple but effective for many use cases, and the main idea is choosing the proper historical records to be the baselines.

A date-aligned estimator selects the same date last year or several years before, shown as:

$$X_{ymdt} \xrightarrow{\text{baseline}} X_{y'mdt} \tag{2}$$

where $y' \leq y-1$, and the annotated arrow links an observation (left) with its baseline (right).

A week-aligned estimator selects another historical date which shares the same week-weekday index as the current date. This method is technically formulated as follows:

$$X_{ymdt} \xrightarrow{\text{baseline}} X_{y'm'd't} \tag{3}$$

where the above two dates should satisfy:

$$f_{\text{d2w}}(y, m, d) = f_{\text{d2w}}(y', m', d') \tag{4}$$

In equation (4), the function $f_{\text{d2w}}(\cdot)$ calculates the week number and weekday for a specific date. For example, $f_{\text{d2w}}(2020, 6, 1) = f_{\text{d2w}}(2019, 6, 3)$ because they are both Mondays of the 22nd week.

2) Trend and Detrend Estimation: This method is designed to extract or eliminate the trends' impacts, and thus leads to a better estimation result. Here, the trend can be estimated by either of the following formulas:

$$T_{ymdt} = f_w^{\text{trend}}(X_{ymdt}, \cdots) \tag{5}$$

$$T_{ymdt} = \hat{f}_w^{\text{trend}}(X_{ymdt}, \cdots; \theta^{\text{trend}}) \tag{6}$$

where T_{ymdt} is the trend series, $f_w^{\text{trend}}(\cdot)$ and $\hat{f}_w^{\text{trend}}(\cdot)$ are two estimation functions, w is a given length of the sliding window, and θ^{trend} denotes the model parameters to be calibrated. For illustration, weekly moving average is an instance of (5), and other advanced models may follow the format of (6).

A trend and a detrend estimator calculate the baselines differently, shown as follows:

$$T_{ymdt} \xrightarrow{\text{baseline}} T_{y'mdt} \tag{7}$$

$$X_{ymdt} \xrightarrow{\text{baseline}} T_{ymdt} \tag{8}$$

The baselines in (7) use the trend to remove potential noises, while the baselines in (8) detrend the original data to find any additional changes, e.g., extra increments.

3) Backcast Estimation: This method has a complicated expression based on machine learning, so more data and computations are required to calibrate the unknown parameters. This method is originally used to analyze the electricity consumption with great improvement in accuracy. Here, a backcast estimation can be described as follows:

$$B_{ymdt} = \hat{f}_w^{\text{back}}(X_{y'mdt}, Y_{y'mdt}, \cdots; \theta^{\text{back}}) \tag{9}$$

where B_{ymdt} is the backcast outcome calculated by a machine learning model $\hat{f}_w^{\text{back}}(\cdot)$, and θ^{back} denotes the corresponding model parameters (often high-dimensional). In addition, X and Y are both placeholders for some variables, and the ellipsis mark represents other possible inputs.

It is simple to extend (9) to an ensemble backcast model by averaging the outputs of multiple base models (indexed by i):

$$\hat{f}_w^{\text{back}}(\cdot) = \frac{1}{N_{\{i\}}} \sum_{\forall i} \hat{f}_{w,i}^{\text{back}}(\cdot) \tag{10}$$

Often, a backcast estimation can largely mitigate the adverse impacts of non-pandemic factors to establish a reliable baseline, shown as follows:

$$X_{ymdt} \xrightarrow{\text{baseline}} B_{ymdt} \tag{11}$$

Note that one distinct advantage of this method is the flexibility because there are so many possible options and combinations for the base models.

4) Distribution-based Estimation: This method provides a new perspective of the data distribution to understand the underlying patterns. The key point is turning to monitor the distributions of those fluctuating variables, e.g., electricity price. This could be surprisingly effective if the sliding window is well configured.

Technically, we develop a novel fluctuation index to evaluate the possibility that an observation might be abnormal. The following expression gives more details:

$$I_{ymdt} = f_w^{\text{fluc}}(X_{ymdt}) = \left|1 - 2F_w(X_{ymdt})\right| \tag{12}$$

where I_{ymdt} is the proposed fluctuation index, $f_w^{\text{fluc}}(\cdot)$ is an estimation function, $F_w(\cdot)$ is the cumulative distribution function, and the sliding windows for both functions have a length of w.

Fig. 3 offers a graphic illustration of the fluctuation index from two aspects, i.e., the highlighted distance and the shaded area. By definition, $0 \leq I_{ymdt} \leq 1$, and $I_{ymdt} \geq 0.5$ may rarely happen. It is thus possible to evaluate the abnormal dynamics by monitoring this fluctuation index.

A distribution-based estimator is able to offer a baseline in either of the following way:

$$I_{ymdt} \xrightarrow{\text{baseline}} I_{y'mdt} \tag{13}$$

$$I_{ym} \xrightarrow{\text{baseline}} I_{y'm} \tag{14}$$

where I_{ym} is derived similarly as (1), and could be used to analyze the electricity prices because the monthly price distributions are roughly stable.

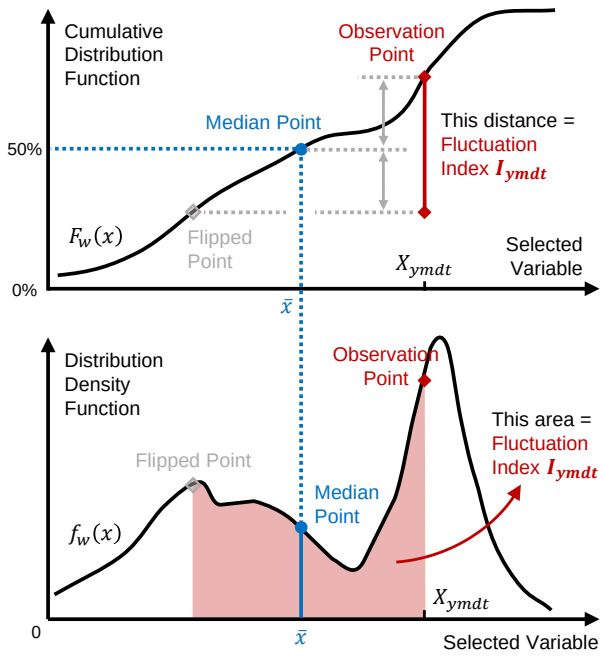

Fig. 3. Illustration of the proposed fluctuation index. This index can be physically explained by a highlighted distance in the cumulative distribution curve or a shaded area in the probability density curve.

Calculating the difference between two cumulative distribution functions is another option to study the distributions. A distance metric to quantify the difference is given as follows:

$$S_y = \left\| F_{w,y} - F_{w,y'} \right\| \tag{15}$$

where S_y is the propsoed distribution distance, while $F_{w,y}$ and $F_{w,y'}$ describes the cumulative distribution for year y and y'.

D. Regression Analysis

Regression is widely used in empirical analysis to explore the potential relationship between different factors. In particular, regression allows us to answer a few questions on correlation or causality during COVID-19. We have collected two popular regression models, along with several useful statistical tests.

1) Ordinary Least Squares Regression (OLS): This method offers multiple expressions to check the underlying correlation or casuality. This method allows linear expressions as well as a few nonlinear expressions (with quadratic, interaction, or logarithms terms).

An OLS model can be formulated as follows:

$$Z_{ymdt} = \theta_1^{\text{ols}} X_{ymdt} + \theta_2^{\text{ols}} Y_{ymdt} + \cdots + \epsilon_{ymdt}^{\text{ols}} \tag{16}$$

where X, Y, Z are all placeholders for some variables, θ_1^{ols}, θ_2^{ols} denote the regression coefficients, and $\epsilon_{ymdt}^{\text{ols}}$ represents the error term. The ellipsis mark indicates that other regression terms (linear or nonlinear) are fully allowed.

We calibrate an OLS model by determining a set of regression coefficients to minimize the regression residuals. Here is the related optimization problem:

$$\min \sum_{\forall y,m,d,t} \left(Z_{ymdt} - \theta_1^{\text{ols}} X_{ymdt} - \theta_2^{\text{ols}} Y_{ymdt} - \cdots \right)^2 \tag{17}$$

In addition, an OLS model can be further validated by running a few statistical tests, including t-test, F-test, and normality-test. R-squared and adjusted R-squared are also informative to evaluate the goodness of fit.

2) Vector Autoregression (VAR): This method is specialized to capture the complicated correlation between multiple time-series data. One can extend this method to restricted vector autoregression when some regression coefficients are imposed to be zeros. Both models are powerful and widely adopted in empirical studies.

A VAR model combines all the variables together and uses the following formula to model the evolution over time:

$$X_{ymdt} = \sum_{i=1}^{p} \theta_i^{\text{var}} X_{y,m,d,t-i} + \theta_0^{\text{var}} + \epsilon_{ymdt}^{\text{var}} \tag{18}$$

where X_{ymdt} should be interpreted as some variable or a concatenation of several variables. p is called the order of this VAR model, and the lag terms for the last p periods are considered above. Besides, $\theta_0^{\text{var}}, \cdots, \theta_p^{\text{var}}$ are regression coefficients, and $\epsilon_{ymdt}^{\text{var}}$ denotes the error term. p is called the order of this VAR model, and the lag terms for the last p periods are considered above.

The flowchart for establishing a VAR model can be divided into four steps: pre-estimation preparation, model calibration, model verification, and post-estimation analysis.

First, we need to conduct an Augmented Dickey-Fuller (ADF) test, a cointegration test, and a Granger causality test to analyze the situations of stationarity, cointegration, and potential causality respecitvely.

Second, the regression coefficients can be determined by a series of minimization problems, each of which is similar as (17). For a p-order VAR model (18), one should run a total number of p optimizations.

Third, another ADF test is used to test if the residual series is stationary, while a Ljung-Box test and a Durbin-Watson test are used to inspect the underlying endogeneity and autocorrelation. A robustness test is also preferred to demonstrate the model performance against coefficient perturbations.

Finally, the calibrated VAR model can provide further insights by running the impulse response analysis and forecast error variance decomposition.

E. Scientific Visualization

Scientific visualization is one of the most intuitive way to exhibit empirical findings, but the methods turn out to be highly diverse in different applications. We thus specialize the methods for several classical use cases.

A line chart is probably the most simplest way to show a series of changing data. It is useful to visualize the raw data X_{ymdt}, any aggregated data like $\sum_t X_{ymdt}$, and any filtered data like $X_{ymdt}(m \leq 6)$. When the x-axis represents dates, our toolbox further supports labeling the dates of some selected big events of COVID-19.

A stacked bar chart is able to compare between different categories. Visually, different bars (representing those categories) are stacked end-to-end and assigned different colors for distinction. Assume the raw data X_{ymdt} can be divided into several sub-categories $X_{ymdt}^k \ \forall k$, then the corresponding proportion for X_{ymdt}^k is calculated as:

$$X_{ymdt}^{k\%} = \frac{X_{ymdt}^k}{X_{ymdt}} \times 100\% \tag{19}$$

A histogram describes the distribution or frequency features of a group of fluctuating data. This is helpful to handle a large amount of observations and detect any possible outliers. To be specific, our toolbox supports visualizing the cumulative distribution function and the probability density.

A box plot is designed to graphically display groups of data through their quantiles. It can effectively handle a data matrix by calculating the quantiles for each column and visualizing these quantiles with box labels or color bands. Let $F(\cdot)$ denotes the cumulative distribution function for one column, the toolbox will calculate the following five quantiles:

$$Q_i = F^{-1}(q_i), \quad i = 1, \cdots, 5 \tag{20}$$

where $q_1 = 0.1$, $q_2 = 0.25$, $q_3 = 0.5$ (mean value), $q_4 = 0.75$, and $q_5 = 0.9$.

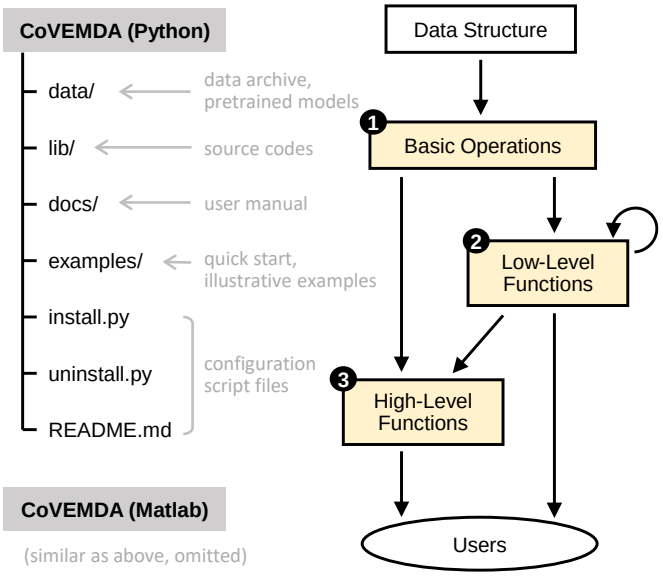

Fig. 4. Folder structure (left) and three-level programming architecture (right) for the proposed toolbox.

IV. Python and Matlab Implementations

A. Architecture Design

We next focus on the programming details to implement the models and algorithms in Section III. It is necessary to start a discussion about the high-level architecture before diving into the project details of Python or Matlab.

1) Folder Structure: Fig. 4 shows a concise folder structure in the left part. Note that all the archived data and pretrained model are located in the "data/" folder, and the source codes can be found in the "lib/" folder. Beginners may get started by reading the user manual or quick start examples.

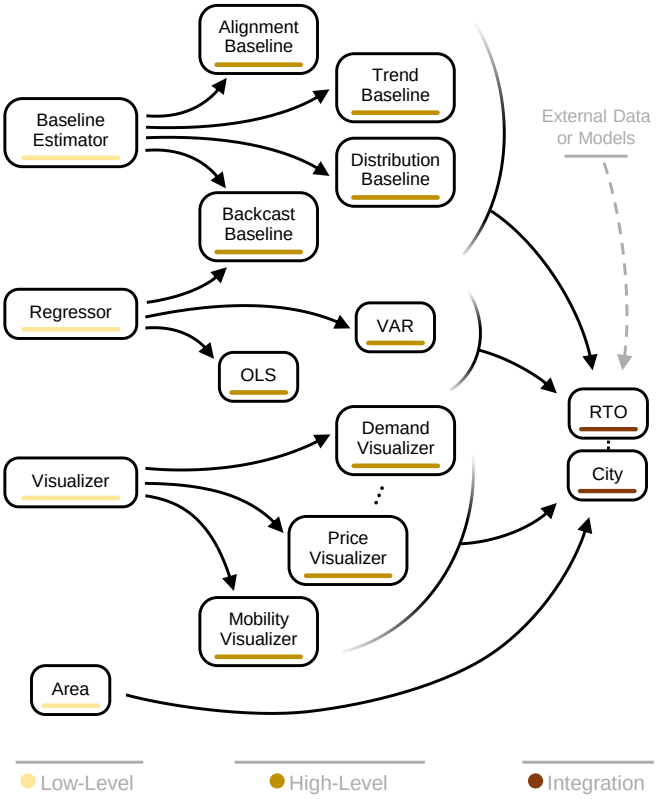

Fig. 5. Class inheritance map of the proposed toolbox (Python version). Different classes are designed with functions of different levels. External data and models are also supported for further extensions.

2) Programming Structure: Fig. 4 also illustrates a programming structure that classifies the entire function family into three levels: basic operations, low-level functions, and high-level functions. This structure breaks down large tasks (user-oriented) into small activities (data-oriented), and helps clarify the calling relationships and dependencies between different functions.

B. Python Implementation

1) Data Structure: The toolbox establishes a new DataFormatter class to realize the wide data frame structure mentioned in Subsection III-B. This class wraps the popular DataFrame class from Pandas package, and extends the built-in function family with a lot of specialized functions.

2) Object-Oriented Design: Fig. 5 elaborates how to organize the major classes and their inheritance relationship to realize the proposed methods. There are four base classes—a baseline estimator class, a regressor class, a visualizer calss, and an area class—they mainly build up the fundamental properties and some key components. A few high-level classes are then established to specify the method details, and integrated to construct the RTO and City classes at last. These classes provide concise and powerful interfaces for ease of use.

As for extensions, users are allowed to develop their own class based on the predefined classes. External data sources, special parsers, and user-defined functions could be included in this new class to support further development.

We follow the folder structure in Subsection IV-A to organize the Python script files (.py files). The relevant classes and functions are collected in the same file with increased readability. We make efforts to keep clean logic so that our codes can be easily reused or extended.

C. Matlab Implementation

1) Data Structure: The toolbox constructs a new data structure based on the built-in table array in Matlab. A lot of efforts are made to simplify and robustify the syntax system, so that all the basic operations can run smoothly as planned.

2) Functional Design: Functions are carefully assigned to different abstraction levels (Fig. 4), and the calling relations remain clean and efficient. Note that many functions share the same or similar names as those in the Python version, e.g., both versions have developed `cal_demand_baseline(·)`.

Using the folder structure in Subsection IV-A, the Matlab script files (.m files) are collected in three different folders according to the function level. Clear logic and explanatory comments are useful to increase the readability.

V. Empirical Studies

Among all possible use cases, this section will select three of them to demonstrate our findings in several questions of public concerns.

A. Pandemic Impact on Steady State of Power Systems

The very first question for most studies is how much and how long COVID-19 has influenced the operation of U.S. power systems. We will next conduct a few use cases to answer these questions from different perspectives.

1) Peak Demand Changes among Different Regions: For a given region, the reduction of peak demand is assessed for each day and averaged for the whole month:

$$\alpha_{ym} = \frac{1}{N_{\{d\}}} \sum_{\forall d} \left(\frac{B_{ymd} - D^{\text{peak}}_{ymd}}{B_{ymd}} \right) \times 100\% \qquad (21)$$

where the peak demand D^{peak}_{ymd} has a baseline B_{ymd}, which can be derived by running the pretrained backcast model in the toolbox.

Table I collects the estimation results for seven U.S. marketplaces. MISO (Midcontinent area) and NYISO (New York) are the top two markets that have experienced more than 10% drop in both April and May. According to the average reduction rates, the situations in June were largely alleviated for all seven regions, but NYISO appeared to recover much more slowly.

2) Price Distribution Shift in Chicago: We then apply the fluctuation index to evaluate the price distributions in Chicago.

Results show that the monthly index values in 2020 are 0.80 (March), 0.85 (April), 0.62 (May), and 0.63 (June). The largest difference between 2019 and 2020 lies in April when the index value grew up by 90.18%. On average, a 65.70% increase could be observed, from 0.44 in 2019 to 0.72 in 2020. All these results validate that Chicago had truly experienced a period of severe price changes during COVID-19.

TABLE I
REDUCTION RATES OF PEAK DEMAND IN DIFFERENT REGIONS

Region	March	April	May	June
CAISO	2.90%	9.28%	6.23%	3.56%
ERCOT	-0.85%	3.36%	2.52%	2.70%
ISO-NE	3.14%	6.76%	9.07%	2.33%
MISO	2.57%	**10.23%**	**10.70%**	2.49%
NYISO	**4.38%**	10.21%	10.46%	**7.06%**
PJM	1.71%	9.52%	9.08%	1.14%
SPP	0.91%	7.16%	7.08%	1.43%
Average	2.11%	8.07%	7.88%	2.96%

Note: the largest changes among all the regions are highlighted above.

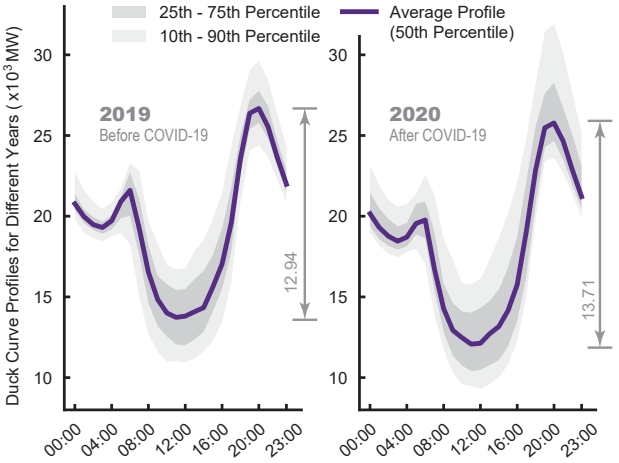

Fig. 6. California's duck curves in 2019 and 2020. The ramping is highlighted and labeled with specific numbers.

3) Duck Curves and Renewable Energy Share in California: A duck curve, also known as the residual demand, is derived by calculating the difference between electricity consumption and the solar generation.

$$R_{ymdt} = D_{ymdt} - G^{\text{solar}}_{ymdt} \tag{22}$$

Fig. 6 compares the duck curves in California. Comparing with 2019, the average duck curve in 2020 has a higher ramping requirement of 761.90 MW, and a larger fluctuation range of 3923.24 MW. As shown, the increased peak-valley difference or peak-valley ratio will call for more flexible resources for power system balancing.

The share of renewable energy is calculated as follows:

$$\beta_{ym} = G^{\text{hydro\%}}_{ym} + G^{\text{solar\%}}_{ym} + G^{\text{wind\%}}_{ym} \tag{23}$$

We typically consider the monthly proportions in California, and apply an ARIMA model for trend estimation. This model is configured by grid searching the best hyperparameters, and the final setting turns out to be ARIMA(2,0,1). Results show that the observed share of renewable energy during March–June is 34.88% on average, while the ARIMA model estimates a slightly larger baseline of 34.90%. This tiny difference, much less than the demand drop, is clearly against the statement that renewables might enjoy extra benefits during COVID-19 because of their low marginal costs. A possible explanation for this finding is the conservative dispatch strategies that take the system safety into consideration.

TABLE II
REGRESSION RESULTS OF EQUATION (24) AND (25)

Parameter	Coeff	Std	t-Test	p-Value
$\boldsymbol{\theta_1}$	**-2.8715**	**1.083**	**-2.652**	**0.009**
θ_2	0.8714	0.845	1.031	0.304
$\boldsymbol{\theta_3}$	**5.4063**	**2.016**	**2.681**	**0.008**
θ_4	-0.6941	1.624	-0.428	0.669
θ_5	4.4138	2.562	1.723	0.087
$\boldsymbol{\theta_6}$	**2.8960**	**0.996**	**2.907**	**0.004**
$\boldsymbol{\theta_7}$	**-1.4503**	**0.560**	**-2.591**	**0.011**
θ_8	-8.0344	4.548	-1.766	0.079

Note: "Coeff" is the coefficient value, "Std" is the standard deviation. The top part shows the results for (24), and the bottom part for (25). In addition, we highlight the rows when the corresponding coefficients are statistically significant.

B. Factor Analysis on Electricity Price Changes

There is an open debate on how the electricity prices were influenced by COVID-19 and the gas price collapse in 2020, because both events have a time overlap before mid-2020. We take Boston as an example, and conduct regression analysis to demonstrate our findings for this debate.

The first step is selecting proper variables and data for the prices and the pandemic situations. We calculate the logit value of the fluctuation index, denoted by LoI_{ymd}, to describe the abnormality of electricity price observations. Often, $LoI_{ymd} > 3$ is highly unusual. We also need to construct a gas price variable $\lambda^{\text{gas}}_{ymd}$ by importing and organizing the data from an external source. As for the pandemic modeling, we come up with two ways: one is the daily confirmed cases C_{ymd}, and the other is a binary dummy variable δ_{ymd} that indicates the absence ($\delta_{ymd} = 0$) or presence ($\delta_{ymd} = 1$) of the pandemic.

With the above variables, two OLS regression models are designed as follows:

$$LoI_{ymd} = (\theta_1\delta_{ymd} + \theta_2)\lambda^{\text{gas}}_{ymd} + (\theta_3\delta_{ymd} + \theta_4) \tag{24}$$

$$LoI_{ymd} = \theta_5\lambda^{\text{gas}}_{ymd} + \theta_6 C_{ymd} + \theta_7\lambda^{\text{gas}}_{ymd}C_{ymd} + \theta_8 \tag{25}$$

The basic idea for (24) and (25) is controlling the effects of gas prices when assessing the pandemic's impacts. We are also curious about the interaction between these two factors.

Table II illustrates the results of model calibration and statistical tests. We highlight four coefficients that are statistically significant: θ_1, θ_3, θ_6, and θ_7.

Here, the pandemic's impact is validated to exist according to a strong statistical evidence that θ_1 and θ_3 in (24) are nonzero—no one could deny that LoI_{ymd} is dependent on δ_{ymd}.

Another finding is that there may exist an offset relationship between the impacts of COVID-19 and gas prices. One supporting evidence is the negative sign of θ_1. This is further validated by (25) with a negative θ_7. While the impacts of both factors are synergistic rather than additive (because $\theta_7 \neq 0$), it is at least statically clear that COVID-19 have truly caused more abnormal electricity prices (because $\theta_6 > 0$).

C. Improved Load Forecast Using Mobility Data

One severe outcome of COVID-19 is the rapid drop of electricity consumption. Even worse, most load forecast models may perform poorly because they can hardly capture this sudden break caused by the lockdown policy. This calls for an improved forecasting strategy that could quickly adapt to the new situation and make more accurate predictions. We will next show that using mobility data to enhance the load forecast models might be an effective solution.

This case considers the day-ahead hourly load prediction tasks. Three typical models are tested here: neural network (NN), random forest (RF), and support vector machine (SVM). The inputs for these models include calendar variables, meteorological variables, and the previous load. We also grid search the hyperparameters for each kind of model carefully.

The above models cannot capture the novel load pattern during COVID-19, so we improve them in the following two ways. One is fine-tuning the model with new observations, the other is using mobility data to enhance the results.

Technically, the latter idea can be described as follows:

$$\begin{aligned}\hat{D}_{ymdt} =& \hat{f}^{\text{pred}}(D_{y,m,d-1,t}, \cdots; \theta^{\text{pred}}) \\ &+ \Delta\hat{f}^{\text{enh}}(M_{y,m,d-1,t}; \theta^{\text{enh}})\end{aligned} \tag{26}$$

where the improved result $\hat{D}_{ymdt}$ has an enhanced item $\Delta\hat{f}^{\text{enh}}(\cdot)$ that takes the previous mobility data as its input. $\hat{f}^{\text{pred}}(\cdot)$ is exactly the same as the original model, but we avoid listing all inputs here by an ellipsis mark.

For simplification, we only consider a linear regression formula for $\hat{f}^{\text{enh}}(\cdot)$, and we calibrate its parameter θ^{enh} by the residual error series during COVID-19 (very few are needed).

We pay attention to the forecast task in New York City on March 21, two weeks after the state-of-emergence order on March 7, 2020. The main focus is on the prediction performance of different models in the remaining days before mid-2020. We also validate the performance gaps between the normal period (January 1–March 21) and the lockdown period (March 21–June 30).

Table III gives the comparison results for different models, whose performances are measured by mean average percentage errors (MAPE).

It may not be surprising that the performance gaps between the normal and lockdown periods are exceeding 5%, and some errors are almost tripled. Also, there is nearly no difference when fine-tuning these models with new observations, e.g., RF and its updated model RF-Updated has the same error estimation of 8.20%.

The major message from Table III is that using mobility data might improve the forecast performance with an accuracy increase of nearly 25–40% or 2–4 percentage points. This result can be further improved when obtaining more abnormal observations (only 14 days in this case) or considering better enhancement models (only linear regression in this case).

TABLE III
PERFORMANCE OF DIFFERENT LOAD FORECAST MODELS

Models	Normal Period	Lockdown Period
NN	3.10%	8.63%
NN–Updated	—	8.73%
NN–Mobility	—	**5.25%**
RF	2.84%	8.20%
RF–Updated	—	8.20%
RF–Mobility	—	**6.15%**
SVM	4.54%	10.79%
SVM–Updated	—	10.69%
SVM–Mobility	—	**6.84%**

Note: "NN" is neural network, "RF" is random forest, and "SVM" is support vector machine. "*–Updated" denotes a updated model, while "*–Mobility" denotes an improved model that uses mobility data. We highlight the best estimators of each kind for the lockdown period.

VI. CONCLUSION

Evaluating how COVID-19 has influenced the real-world power systems is critical to understand the risky conditions as well as the abnormal operation patterns. But up to now, there is still a lack of reliable and ready-to-use data, methods, or toolboxes for empirical studies.

This paper overcomes the above difficulty by developing an open-access data hub, an open-source toolbox, and several powerful methods for users with diverse backgrounds, such as researches, policy makers, and educators. The toolbox is implemented in Python and Matlab with three key functions: baseline estimation, regression analysis, and scientific visualization. Further extensions are allowed to handle more complicated applications.

REFERENCES

[1] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet Infectious Diseases*, vol. 20, no. 5, pp. 533–534, 2020.

[2] T. Hale, N. Angrist, R. Goldszmidt, B. Kira, A. Petherick, T. Phillips, S. Webster, E. Cameron-Blake, L. Hallas, S. Majumdar, and H. Tatlow, "A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)," *Nature Human Behaviour*, vol. 5, no. 4, pp. 529–538, 2021.

[3] B. Wormuth, S. Wang, P. Dehghanian, M. Barati, A. Estebsari, T. P. Filomena, M. H. Kapourchali, and M. A. Lejeune, "Electric power grids under high-absenteeism pandemics: History, context, response, and opportunities," *IEEE Access*, vol. 8, pp. 215727–215747, 2020.

[4] P. Jiang, Y. V. Fan, and J. J. Klemes, "Impacts of COVID-19 on energy demand and consumption: Challenges, lessons and emerging opportunities," *Applied Energy*, vol. 285, no. January, 2021.

[5] G. Ruan, J. Wu, H. Zhong, Q. Xia, and L. Xie, "Quantitative assessment of U.S. bulk power systems and market operations during the COVID-19 pandemic," *Applied Energy*, vol. 286, p. 116354, 2021.

[6] W. Kanda and P. Kivimaa, "What opportunities could the COVID-19 outbreak offer for sustainability transitions research on electricity and mobility?," *Energy Research and Social Science*, vol. 68, no. June, p. 101666, 2020.

[7] D. Obst, J. De Vilmarest, and Y. Goude, "Adaptive methods for short-term electricity load forecasting during COVID-19 lockdown in France," *IEEE Transactions on Power Systems*, vol. 36, no. 5, pp. 4754–4763, 2021.

[8] P. Gallo, J. M. Guerrero, R. Musca, E. Riva Sanseverino, J. C. Vasquez Quintero, and G. Zizzo, "Effects of COVID-19 pandemic on the Italian power system and possible countermeasures," *Electric Power Systems Research*, vol. 201, no. December 2020, p. 107514, 2021.

[9] L. Badesa, G. Strbac, M. Magill, and B. Stojkovska, "Ancillary services in Great Britain during the COVID-19 lockdown: A glimpse of the carbon-free future," *Applied Energy*, vol. 285, no. January, p. 116500, 2021.

[10] G. Ruan, D. Wu, X. Zheng, H. Zhong, C. Kang, M. A. Dahleh, S. Sivaranjani, and L. Xie, "A cross-domain approach to analyzing the short-run impact of COVID-19 on the US electricity sector," *Joule*, vol. 4, no. 11, pp. 2322–2337, 2020.

[11] A. Abu-Rayash and I. Dincer, "Analysis of the electricity demand trends amidst the COVID-19 coronavirus pandemic," *Energy Research and Social Science*, vol. 68, no. July, p. 101682, 2020.

[12] D. Agdas and P. Barooah, "Impact of the COVID-19 pandemic on the U.S. electricity demand and supply: An early view from data," *IEEE Access*, vol. 8, pp. 151523–151534, 2020.

[13] C. Zanocco, J. Flora, R. Rajagopal, and H. Boudet, "Exploring the effects of California's COVID-19 shelter-in-place order on household energy practices and intention to adopt smart home technologies," *Renewable and Sustainable Energy Reviews*, vol. 139, no. July 2020, p. 110578, 2021.

[14] S. Ou, X. He, W. Ji, W. Chen, L. Sui, Y. Gan, Z. Lu, Z. Lin, S. Deng, S. Przesmitzki, and J. Bouchard, "Machine learning model to project the impact of COVID-19 on US motor gasoline demand," *Nature Energy*, vol. 5, no. 9, pp. 666–673, 2020.

[15] M. Graff and S. Carley, "COVID-19 assistance needs to target energy insecurity," *Nature Energy*, vol. 5, no. 5, pp. 352–354, 2020.

[16] B. Steffen, F. Egli, M. Pahle, and T. S. Schmidt, "Navigating the clean energy transition in the COVID-19 crisis," *Joule*, vol. 4, no. 6, pp. 1137–1141, 2020.

[17] K. T. Gillingham, C. R. Knittel, J. Li, M. Ovaere, and M. Reguant, "The short-run and long-run effects of Covid-19 on energy and the environment," *Joule*, vol. 4, no. 7, pp. 1337–1341, 2020.

[18] S. Halbrugge, P. Schott, M. Weibelzahl, H. U. Buhl, G. Fridgen, and M. Schopf, "How did the German and other European electricity systems react to the COVID-19 pandemic?," *Applied Energy*, vol. 285, no. August 2020, p. 116370, 2021.

[19] K. V. Kumar, A. Kumar, G. Verma, S. Machal, S. C. Saxena, D. De, S. S. Barpanda, and K. V. Baba, "Experience of Indian electricity market operation and other events during COVID-19 pandemic," *2020 21st National Power Systems Conference, NPSC 2020*, 2020.

[20] X. Zhang, F. Pellegrino, J. Shen, B. Copertaro, P. Huang, P. Kumar Saini, and M. Lovati, "A preliminary simulation study about the impact of COVID-19 crisis on energy demand of a building mix at a district in Sweden," *Applied Energy*, vol. 280, no. September, p. 115954, 2020.

[21] D. B. d. M. Delgado, K. M. de Lima, M. d. C. Cancela, C. A. d. S. Siqueira, M. Carvalho, and D. L. B. de Souza, "Trend analyses of electricity load changes in Brazil due to COVID-19 shutdowns," *Electric Power Systems Research*, vol. 193, no. July 2020, p. 107009, 2021.

[22] N. Norouzi, G. Z. Zarazua de Rubens, P. Enevoldsen, and A. Behzadi Forough, "The impact of COVID-19 on the electricity sector in Spain: An econometric approach based on prices," *International Journal of Energy Research*, vol. 45, no. 4, pp. 6320–6332, 2021.

[23] A. Werth, P. Gravino, and G. Prevedello, "Impact analysis of COVID-19 responses on energy grid dynamics in Europe," *Applied Energy*, vol. 281, no. August 2020, p. 116045, 2021.

[24] G. Ruan, H. Zhong, G. Zhang, Y. He, X. Wang, and T. Pu, "Review of learning-assisted power system optimization," *CSEE Journal of Power and Energy Systems*, vol. 7, no. 2, pp. 221–231, 2021.

[25] P. Gulati, A. Kumar, and R. Bhardwaj, "Impact of Covid-19 on electricity load in Haryana (India)," *International Journal of Energy Research*, vol. 45, no. 2, pp. 3397–3409, 2021.

[26] C. Zhan, Y. Zheng, H. Zhang, and Q. Wen, "Random-Forest-Bagging broad learning system with applications for COVID-19 pandemic," *IEEE Internet of Things Journal*, vol. X, pp. 1–14, 2021.

[27] M. Jamshidi, A. Lalbakhsh, J. Talla, Z. Peroutka, F. Hadjilooei, P. Lalbakhsh, M. Jamshidi, L. L. Spada, M. Mirmozafari, M. Dehghani, A. Sabet, S. Roshani, S. Roshani, N. Bayat-Makou, B. Mohamadzade, Z. Malek, A. Jamshidi, S. Kiani, H. Hashemi-Dezaki, and W. Mohyuddin, "Artificial intelligence and COVID-19: Deep learning approaches for diagnosis and treatment," *IEEE Access*, vol. 8, no. December 2019, pp. 109581–109595, 2020.

[28] A. Saif, T. Imtiaz, S. Rifat, C. Shahnaz, M. O. Ahmad, and W.-P. Zhu, "CapsCovNet: A modified capsule network to diagnose Covid-19 from multimodal medical imaging," *IEEE Transactions on Artificial Intelligence*, pp. 1–1, 2021.

[29] G. X. Xu, C. Liu, J. Liu, Z. Ding, F. Shi, M. Guo, W. Zhao, X. Li, Y. Wei, Y. Gao, C. X. Ren, and D. Shen, "Cross-site severity assessment of COVID-19 from CT Images via domain adaptation," *IEEE Transactions on Medical Imaging*, vol. XX, no. XX, pp. 1–15, 2021.

[30] M. A. Mohammed, K. H. Abdulkareem, A. S. Al-Waisy, S. A. Mostafa, S. Al-Fahdawi, A. M. Dinar, W. Alhakami, A. Baz, M. N. Al-Mhiqani, H. Alhakami, N. Arbaiy, M. S. Maashi, A. A. Mutlag, B. Garcia-Zapirain, and I. De La Torre Diez, "Benchmarking methodology for selection of optimal COVID-19 diagnostic model based on Entropy and TOPSIS methods," *IEEE Access*, vol. 8, pp. 99115–99131, 2020.

[31] J. S. Frazer, A. Shard, and J. Herdman, "Involvement of the open-source community in combating the worldwide COVID-19 pandemic: A review," *Journal of Medical Engineering and Technology*, vol. 44, no. 4, pp. 169–176, 2020.

[32] D. Arneson, M. Elliott, A. Mosenia, B. Oskotsky, S. Solodar, R. Vashisht, T. Zack, P. Bleicher, A. J. Butte, and V. A. Rudrapatna, "CovidCounties is an interactive real time tracker of the COVID19 pandemic at the level of US counties," *Scientific Data*, vol. 7, no. 1, pp. 1–10, 2020.

[33] M. Farooq and A. Hafeez, "COVID-ResNet: A deep learning framework for screening of COVID-19 from radiographs," *arXiv*, 2020.

[34] R. Hinch, W. J. Probert, A. Nurtay, M. Kendall, C. Wymant, M. Hall, K. Lythgoe, A. Bulas Cruz, L. Zhao, A. Stewart, L. Ferretti, D. Montero, J. Warren, N. Mather, M. Abueg, N. Wu, O. Legat, K. Bentley, T. Mead, K. Van-Vuuren, D. Feldner-Busztin, T. Ristori, A. Finkelstein, D. G. Bonsall, L. Abeler-Dorner, and C. Fraser, "OpenABM-Covid19-An agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing," *PLoS Computational Biology*, vol. 17, no. 7, pp. 1–26, 2021.

[35] J. Lopez Prol and S. O, "Impact of COVID-19 measures on short-term electricity consumption in the most affected EU countries and USA states," *iScience*, vol. 23, no. 10, 2020.

[36] A. Leach, N. Rivers, and B. Shaffer, "Canadian electricity markets during the COVID-19 pandemic: An initial assessment," *Canadian Public Policy*, vol. 46, pp. S145–S159, 2020.

[37] Support Team, "The home page for COVID-EMDA+ and CoVEMDA," Available: https://github.com/tamu-engineering-research/COVID-EMDA, 2020.

[38] J. Ospina, X. Liu, C. Konstantinou, and Y. Dvorkin, "On the feasibility of load-changing attacks in power systems during the COVID-19 pandemic," *IEEE Access*, vol. 9, pp. 2545–2563, 2021.