IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 20, 2021, accepted February 5, 2021, date of publication February 11, 2021, date of current version March 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3058887

An Empirical Study of Environmental Data Prediction in the United States Energy-Water Nexus

YING JIN[1], (Member, IEEE), EMILY J. YANG[2], AND JULIAN FULTON[3]
[1]Computer Science Department, California State University, Sacramento, CA 95819-6021, USA
[2]Folsom High School, Folsom, CA 95630-3053, USA
[3]Environmental Studies Department, California State University, Sacramento, CA 95819-6001, USA

Corresponding author: Ying Jin (jiny@csus.edu)

This work was supported by the United States Environmental Protection Agency Exchange Network Grant Program under Grant OS-83923301.

ABSTRACT Electricity generation systems are dependent on water availability and planning for future water scarcity is currently hindered by limited data and predictive models. The Energy-Water-Emissions Dashboard (EWED) is a novel environmental data management system that integrates multiple heterogeneous data sources and provides information for nearly 10,000 individual power plants across the United States. This article describes our empirical research of using machine learning models for electricity prediction and water usage in the context of water availability constraints. We evaluate the use of linear regression, decision tree regression, random forest regression, eXtreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN). Based on the performance evaluation of each model, we use ANN for generation and water consumption and XGBoost for water withdrawal prediction in the production environment. Model performance evaluation is based on statistical measures including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (R^2), Willmott's Index of Agreement (WIA), RMSE-observations to Standard deviation Ratio (RSR), Nash–Sutcliffe model Efficiency Coefficient (NSEC), and Percent Bias (PBIAS). This article presents performance improvements of our machine learning approach compared to the conventional coefficient method used by EWED, for example, RMSE decreased 8.1% in generation, 59% in water consumption, and 53% in water withdrawal prediction. The significance of this research is that it covers a wide variety of power plant types, it uses consistent methods across energy and water systems, and provides predictions at multiple management scales across the United States to assist with future planning at the energy-water nexus.

INDEX TERMS Power systems modeling, electricity generation prediction, water consumption prediction, water withdrawal prediction, machine learning, energy-water nexus.

I. INTRODUCTION

The association between energy and water systems is an important factor to consider in environmental management, and the term *energy-water nexus* has been used to draw attention to these connections. Water plays an important role in energy production, such as supplying cooling systems in various types of power plants. The availability of water is determined by a range of environmental and societal factors including climate, water management, and competing uses in the agricultural, municipal, and other sectors. In addition, electricity generation produces emissions that contribute to climate change, which is expected to reduce water availability in many parts of the world. Meanwhile, demand for electricity is expected to increase, which will, in turn, likely place more water demands on increasingly scarce water resources. Planning for such future constraints in the energy-water nexus is thus critical to energy reliability and managing environmental impacts. However, there is not enough attention paid to building models based on existing data and at resolutions that can be used for planning and management purposes. In addition, data on energy generation, emission, water

The associate editor coordinating the review of this manuscript and approving it for publication was Tallha Akram.

withdrawal, and water consumption in the United States are scattered in disparate data sources with lag time and are often incomplete. Thus, there is a need for an integrated system that allows users to visualize and analyze data at a finer resolution on both temporal and spatial levels. The Energy-Water-Emissions Dashboard (EWED) project designed and implemented such a software system through a collaboration of computer and environmental scientists, and was supported by a United States Environmental Protection Agency (EPA) Exchange Network Grant.

The EWED project integrated federal data sources from EPA, the Energy Information Administration (EIA), and others to support visualization and planning with both graphical web-based user interfaces and web services. EWED provides the services for querying and visualizing a historic system and a projected system covering the entire United States. The historic system integrates power plant generation, greenhouse gas emissions, water usage, and hydrologic unit water availability data at the monthly time step from 2003 to the most-recent available. The projected system is based on the EIA's regional electricity generation projections out to 2050. The current implementation of EWED produces the prediction by disaggregating these regional electricity projections to the power plant scale and multiplying that generation by coefficients based on past environmental performance, as described in detail in the subsequent section. As a parallel approach, we also predicted generation, water withdrawal, and water consumption at the power plant level using machine learning models.

This article describes our machine learning approach of prediction, including data cleaning and pre-processing, feature selection, and training. We evaluate the use of linear regression, decision tree regression, random forest regression, eXtreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN) [1]. We implement the machine learning part of the project using Python because ample libraries are available. Scikit-learn [2] is used for linear regression, decision tree regression, and random forest regression. XGBoost is based on the implementation provided by [3]. Our ANN is built based on the Keras sequential model [4]. Other models and tools exist for machine learning. However, we did not perform an exhaustive search to evaluate all options in this version of the project because it is widely accepted that ANN and the ensemble learning approach, such as XGBoost, are very powerful compared to other models. In addition, the performance of ANN and XGBoost are excellent in this project in the context of a wide variety of power plants, which satisfied the purpose of this empirical study. We use various statistical measures to evaluate the performance, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (R^2), Willmott's Index of Agreement (WIA), RMSE-observations to Standard deviation Ratio (RSR), Nash–Sutcliffe model Efficiency Coefficient (NSEC), and Percent Bias (PBIAS). Based on the performance evaluation results, we use ANN for the prediction of generation and water consumption, and XGBoost for water withdrawal prediction.

In the related work section, we discuss related machine learning research for generation and water usage predictions. As detailed in Section VI, existing power generation prediction studies focus on specific types of power plants, such as solar [5], [6], [7], [8], [9], [10], wind [11] [12] [13], and thermal [14]. Existing machine learning research on the prediction of water use is limited, for example, for building water consumption [15], for water consumption of a city [16], for water treatment [17], and for water resource management [18].

Compared to the related research and the coefficient approach, the main contributions of this research are:

- Our machine learning approach has a better performance than the conventional coefficient approach to prediction. RMSE decreased 8.1% in electricity generation, 59% in water consumption, and 53% in water withdrawal.
- This project covers power generation prediction of nearly 10,000 power plants with a wide range of generation capacities and power plant types across the United States, compared to existing machine learning research that focused on a specific type of power plant generation prediction.
- We provide water consumption and withdrawal prediction for power plants. Power plant water usage prediction lacks attention and study in the existing literature.

The rest of the article is organized as follows. Section II provides an overview of the EWED project. Section III describes our machine learning approach to the prediction of future power plant generation, while Section IV presents water consumption and withdrawal prediction. Section V discusses issues and solutions in the implementation process. Section VI describes related work. Section VII summarizes and concludes the article.

II. THE EWED PROJECT

To identify our set of power plants, we first integrated data from the EPA Facility Registry Service (FRS) web service [19] through the filter of "EIA860." This dataset contains power plants reported from the U.S. Energy Information Administration (EIA) program to the EPA FRS. The registration information includes registry identifier, plant code, facility name, address, latitude, longitude etc. Most of the data from FRS are accurate, but they contain inaccurate attribute values for some plants, such as latitude, longitude, and address. They also contain outdated attribute values, such as county and watershed (eight-digit hydrologic unit code, or HUC-8), that need to be replaced. We used Geographical Information System (GIS) software, Google Maps API, customized Python scripts, and GeoJSON files retrieved from United State Census [20] and USDA Water Supply Stress Index Model (WaSSI) [21] in order to produce up-to-date information and add additional attributes

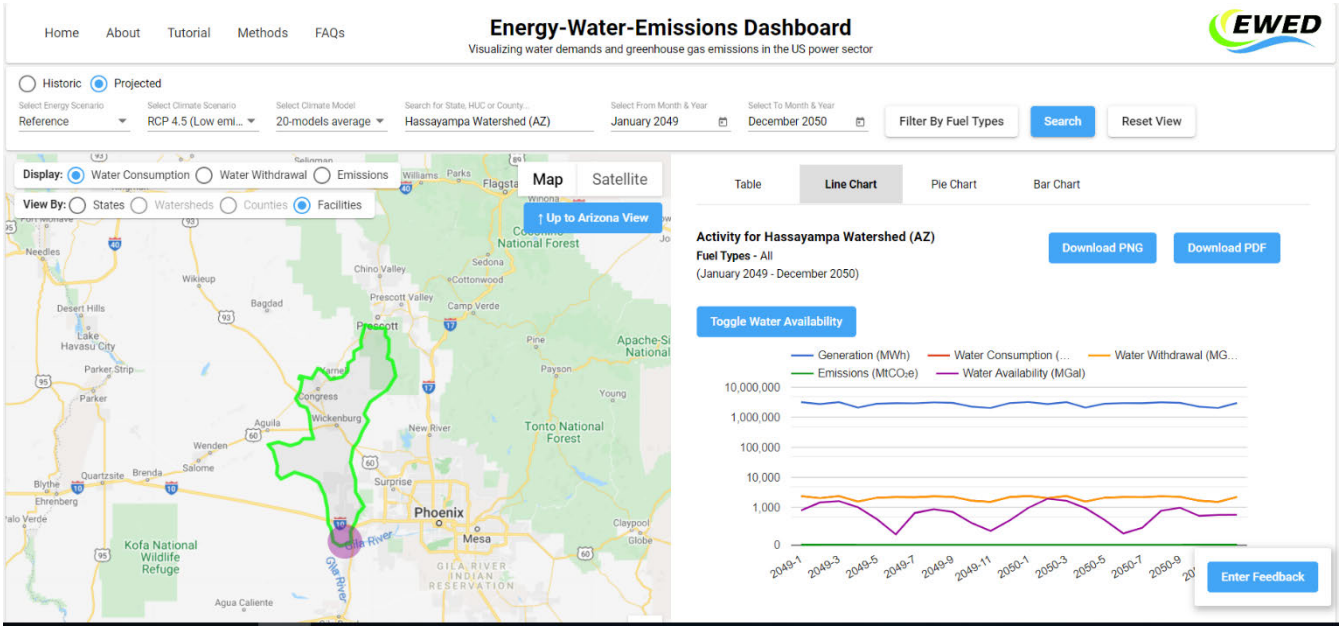


FIGURE 1. **EWED Interface displaying watershed-scale prediction.**

to power plants. Other data sources that we integrated are EPA GHGRP (EnviroFacts) greenhouse emissions [22] and the U.S. Energy Information Administration (EIA) Plant-level Generation [23]. Both are web services with Restful APIs.

We retrieved the information of water withdrawal and consumption per power plant from the EIA's thermoelectric cooling water data [24], which is in the format of excel files. Historic water availability data were retrieved from USFS WaSSI [21] in the format of CSV files. Future generation prediction was provided by the EIA Annual Energy Outlook 2019 (AEO) [25] through Restful APIs, providing data only at the level of regions. Future water availability predictions from twenty climate models, each using two different climate scenarios, are obtained from WaSSI [26] in the format of CSV files. Those heterogeneous sources have different formats with different qualities, present incomplete data, have different unique identifiers, and are challenging to reference each other. Eventually, the data were cleaned thoroughly and stored in Microsoft SQL Server database with 108 tables. In addition to providing web-based graphical interface to registered users, EWED also offers Web Services through EPA Virtual Node Services.

EWED aggregates power plant data to different scales, such as state, county, and hydrologic unit levels, providing information about future usage and alerting users to possible water constraints. Users can select a future year and month to compare the predicted water availability and the predicted water withdrawal or consumption for a hydrologic unit, which can be used to support planning and decision making for this hydrologic area. For example, if the predicted water use is constantly more than water availability for a region in consecutive time periods, the types of plants that consume more water may need to be replaced. Figure 1 is a snapshot of the EWED interface, which presents the projected data in 2049 and 2050 using the conventional coefficient-based approach for Hassayampa Watershed (alternative term for HUC in this project), in which projected water consumption and withdrawal values are more than the water that is available.

III. POWER PLANT GENERATION PREDICTION

Most of the existing, related research predicts specific types of power generation based on factors such as humidity, wind speed, and temperature, for example, for solar power systems [5], [6], [7], [8]. EWED has nearly 10,000 power plants across the United States with various energy technologies and operating under a range of environmental and operational conditions. To cover the large variety of power plants, the practical solution for EWED's projection system was to use data published by the EIA AEO 2019 [25]. The AEO is based on modeling from the fusion of domain experts that consider not only physical climate conditions, but other factors such as economic growth, energy prices, and technological development. As a result, complicated factors, including both climate and non-climate concerns, are incorporated in our prediction.

EIA AEO 2019 provides yearly generation data to 2050 under different energy sector scenarios (called "cases") for each Electric Market Module (EMM) region, per fuel type and prime mover (together called "fuel-mover" herein). There are eight cases representing variability in energy markets due to ranges of potential economic growth, oil prices, and technology development. There are 22 EMM regions throughout contiguous United States, which correspond to the North American Electric Reliability

Corporation (NERC) and Independent System Operator (ISO) regions. For example, EMM code TRE corresponds to the Texas Reliability Entity. There are 13 fuel types – for example NG is natural gas–and two prime movers used only to differentiate solar photovoltaic (SLR-PHTVL) from solar thermal (SLR-THERM), for a total of 14 fuel-mover types. We stored each case in a relational table with EMM code, fuel-mover, generation year, and generation value.

Because the EIA AEO data are yearly and at the regional scale, they cannot satisfy the need to show monthly data for each power plant or other scales such as county, state, or watershed (HUC-8). To produce plant-level monthly data, we used the coefficient approach of disaggregating the AEO raw data by calculating the percent contribution to the total EMM-regional generation by a power plant for each month in the historic data for year 2018. We then applied that percentage to AEO's projected regional generation to produce monthly data at the plant level for future years. Note that the most recent complete calendar year for raw data of generation with dominant types is 2018, due to lag time of reporting to the EIA. At the time of this writing in 2020, only 19% of the plants reported a dominant type in 2019 and zero had been reported for 2020. The subsequent sub-sections present an alternative way of using machine learning to achieve the same goal of predicting monthly generation, but with better performance.

A. DATA PRE-PROCESSING AND FEATURE SELECTION

Data in our database have been cleaned both by our customized computer programs and manually by environmental scientists in our team to ensure data consistency and to enable correct cross-referencing among different database tables.

The target in this supervised learning is monthly generation per power plant. Table 1 show the selected features, which are from different data sources and stored in different tables in the database. The regional data correspond to EMM region. Power plants within different regions, even with the same fuel-mover, may have different generation, since different regions have different climate conditions, efficiencies, energy policies, and other operational factors that affect the makeup of generation technologies. Plant Code uniquely identifies a power plant with a specific generation pattern. Even with the same fuel-mover in the same EMM region, the power generation of each plant can be affected by conditions such as electricity demand. To use this data as our training data, we need to further transform our database data according to these features. Each power plant is associated with one EMM region, which is located using a GIS software, recorded in our database table, and reflected in our *plantEmmMapping* database virtual table (view).

Although fuel-mover is not an attribute of the FRS data in the historic system, dominant type is retrievable from EIA. The EIA annual data web service (historic data) provides information on the combination of fuel type and prime mover for each power plant annually, which we be retrieve through REST APIs. More than one combination of fuel type and prime mover can be used in one power plant, so we use a dominant type to present the combination that generates the maximum amount of electricity within a power plant in a given year. Table 2 shows the examples of relationships between dominant type and fuel-mover, which are stored in our database table. There are 75 dominant types and 13 fuel-mover types.

TABLE 1. Features for Generation Prediction.

Features	Description	Type
Plant Code	Unique identifier of a power plant	Discrete
EMM Code	Unique name for each EMM region	Categorical
Fuel-mover	Prime mover such as steam turbine, gas turbine	Categorical
Regional data	Aggregated yearly data per EMM region per fuel-mover type	Continuous
Month	Month	Discrete
Recent generation	Last year's generation in the given month	Continuous

TABLE 2. Samples of Relationships Between Dominant Type and Fuel-Mover.

Dominant Type	Fuel-Mover
DFO-CA	PET_NA
DFO-CT	PET_NA
JF-GT	PET_NA
LFG-CT	BGM_NA
NG-CA	NG_NA
OG-GT	NG_NA
SUN-CP	SLR_THERM
SUN-PV	SLR_PHTVL

One feature in Table 1 is the yearly data of EMM regions per fuel type. In the production environment, we retrieve EMM regional data from EIA AEO 2019 to fulfil this feature. Our training set is based on historic data retrieved from EIA monthly data web services, which does not have EMM regional data. To prepare the training data with this feature, we use aggregation functions within database queries. Different from popular approaches of using Pandas data frames for data manipulation, we perform a majority of the work at the database level using database views and queries. The following are the code snippets of the views we created to prepare the training data. The view of *generation_structure* is queried directly to produce the data frame to perform training, validation, and testing.

```
create or alter view dbo.generationAll as
select g.plantCode as plantCode, g.genYear as genYear,
      g.genMonth as genMonth,g.genData as genPerPlant,
      p.FuelMover, e.emmCode as EMM, d.dominantType
from generation as g join dominantPlantType as d on
      (g.plantCode = d.plantCode and
      g.genYear = d.genYear)
      join PlantTypeToFuelMover as p
      on (d.dominantType = p.PlantType)
      join plantEmmMapping as e on
      (g.plantCode = e.plantCode)
```

```
create view dbo.genPerEmmFuel_hist as
select EMM as emmCode, fuelMover, genYear,
      CASE when sum(genPerPlant) is NULL then 0
            else sum(genPerPlant)/1000000
      END as genData
from dbo.generationAll
group by EMM,fuelMover,genYear
create or alter view generation_structure as
select cur.plantCode, cur.emmCode,cur.fuelMover,
       cur.genData, cur.genMonth,cur.genYear,
       recent.genData as recGenPerPlant,
       cur.genPerPlant
from (select t.emmCode, t.fuelMover, t.genData,
            p.plantCode, p.genMonth, p.genYear,
            p.genPerPlant
      from genPerEmmFuel_hist as t join
            generationAll as p on
            (t.emmCode = p.EMM and t.fuelMover
            = p.fuelMover and t.genYear
            = p.genYear)) as cur,
            generation as recent
where cur.genYear = (recent.genYear + 1) and
      cur.plantCode = recent.plantCode and
      cur.genMonth = recent.genMonth
```

The data frame contains 15 years of monthly data from 9,961 power plants. We use one-hot encoding for the EMM code, fuel-mover, and month. Although plant code is numeric, this numeric value only serves as a unique identifier. It is not realistic to use one-hot encoding for more than 9,000 power plants because it will increase the dimensions to more than 9,000. Instead, we use Hashing Encoder of Scikit Learn [27] with hyperparameter of "n_component = 16" which produces 16 dimensions. The process of data preparation before training is shown in Figure 2, illustrating that after data are retrieved from heterogeneous data sources, database operations such as join and aggregate are performed to produce the initial data frame. Feature scaling and different encoding techniques are used before finally splitting the data into training set and testing set. The validation set is within the training set, which is explained in the subsequent paragraphs.

Training data are based on the historic data using the data structures presented above. We summarize the generation values of a power plant in a month and year in the training set as follows, where $\vec{G}$ is the generation values in the *generationAll* view. Υ is the symbol of grouping and aggregation function in relational algebra.

∀ p ∈ {Plant Code}, i ∈[2004], [2018], j ∈ [1], [12], m ∈ {linear regression, decision tree regression, random forest regression, XGBoost, ANN}

$$\vec{G}_{\text{year(i),p,month(j)}} \leftarrow \psi_{\text{m}}(\vec{G}_{\text{year(i}-1)}, \vec{emm}, \vec{mover}, \vec{month}_{\text{(j)}}, \vec{p}, \vec{\Upsilon}_{\text{emm, year(i),fuel\_mover,sum(Gyear(i))}} \text{(generationAll)})$$

In the production environment, we use different cases, such as HighMacro, from EIA AEO, which can be summarized as follows:

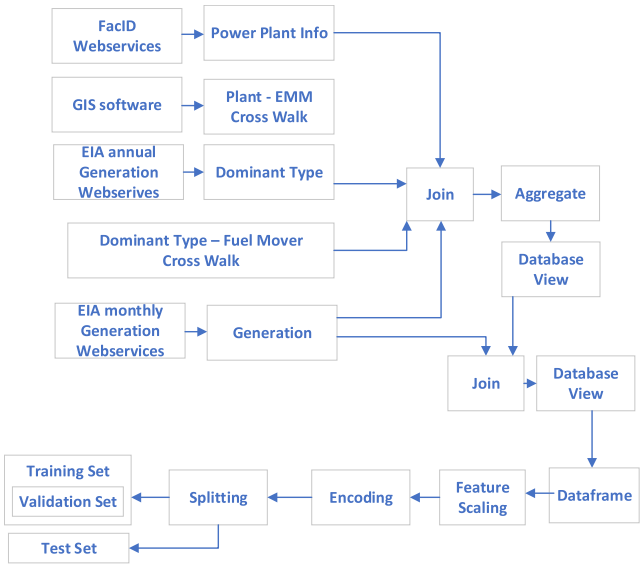


FIGURE 2. Data processing process.

∀ p ∈ {Plant Code}, i ∈[2021], [2050], j ∈ [1], [12], m ∈ {linear regression, decision tree regression, random forest regression, XGBoot, ANN}, k ∈ {HighMacro, AEO19_NO, HighPrice, HighRT, LowMacro, LowPrice, LowRT, REF19}

$$\vec{G}_{\text{year(i),p,month(j)}} \leftarrow \psi_{\text{m}}(\vec{G}_{\text{year(i}-1)}, \vec{emm}, \vec{mover}, \vec{month}(j), \vec{p}, \vec{AEO}(K)_{\text{year(i)}})$$

B. TRAINING

We measure the performance of a model by both RMSE and R2. Root Mean Square Error (RMSE) is used to measure the errors made in the predictions, which is defined as (1), where m is the number of instances. For the i^{th} instance, $yp^{(i)}$ is the predicted value, and $y^{(i)}$ is the label.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(yp^{(i)} - y^{(i)}\right)^2} \tag{1}$$

A coefficient of determination (R^2) is defined by [28] as "Given the variance of the data generating process *pdata*, this metric is proportional to the probability of predicting new samples that actually belong to *pdata*." In definition (2) [28], r_i is the residual and $\bar{y}$ is the average. In other words, R^2 compares the model with the simple model that uses the average of target values. In general, an R^2 score that is close to 0 or negative means unsatisfactory prediction, while close to 1 means almost perfect prediction (it can have exceptions). In all of our prediction results, the decrease of RMSE and the increase of R^2 are consistent.

$$R^2 = 1 - \frac{\Sigma_i \gamma_i^2}{\Sigma_i (y_i - \bar{y})^2} \tag{2}$$

The research in [29], [30], [31], and [32] use various additional measures for performance evaluation. Motivated by their evaluation approaches, our research used the following additional statistical measures: Mean Absolute Error (MAE), Willmott's Index of Agreement (WIA), RMSE-observations

to Standard deviation Ratio (RSR), Nash–Sutcliffe model Efficiency Coefficient (NSEC), and Percent Bias (PBIAS).

The research in [29], [30], [31], and [32] used Mean Absolute Percentage Error (MAPE). MAPE is defined as

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{yp^{(i)} - y^{(i)}}{y^{(i)}}\right| \tag{3}$$

In our database, 5.3% of the records (rows) have 0 generation value. To avoid ''divided by 0'' problem in the MAPE calculation, we add a very small Epsilon (say 0.0000000001) to the denominator. However, if the predicted value is not equal to 0, say 0.01, then after the division of

$$\begin{aligned}|0.01 - -0|/|0 + \text{Epsilon}| &= 0.01/0.0000000001\\ &= 100,000,000\end{aligned}$$

the result will be a very large number. Notice that zero generation typically is not part of a constant pattern for a power plant. A power plant can have zero generation if the plant does not generate energy in the given month or year. However, the same plant may generate a huge amount in any other months or years. This type of prediction error, even only a few, will make MAPE unreasonably large. We feel that MAE, as defined in (4), is better than MAPE to measure the performance in this specific application.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|yp^{(i)} - y^{(i)}\right| \tag{4}$$

Among different approaches to normalize RMSE based on the scale of the data, for the same reason, we believe the following is the most suitable way in our application, as defined in (5). Y is the set of actual values.

$$NRMSE = \frac{RMSE}{\max(Y) - min(Y)} \tag{5}$$

NRMSE provides a clear presentation of errors over the range of values in the data. In this project, NRMSE is relatively small, as presented in the subsequent sections. To show the difference in models more perceptibly, we compare the performance using RMSE. We use NRMSE to show the errors over the value ranges in different models.

The models covered in this research are linear regression, decision tree regression, random forest regression, eXtreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN) [1]. The goal of this project is to use a good model that has a better performance than the coefficient model. Our focus is not an extensive evaluation of different machine learning models. To achieve our goal, we follow the typical way of selecting a model in machine learning, which has been described in [1]. Firstly, we divide the data into a training set and a test set using Scikit-learn ''train_test_split'' method, in which 80% is the training set and 20% is the test set. The test set is not touched until a model is selected. We use the training set to do a preliminary selection of the preferred models. K-fold cross validation is used for the selection, specifically, 10-fold cross validation. K-fold cross validation divides a set into k subsets, which are called folds. This approach uses one fold for evaluation and the other k-1 folds for training. Each time, a different fold is picked for evaluation and the other k-1 folds are for training. We use the Scikit-Learn cross-validation feature for linear regression, decision tree regression, random forest regression, and XGBoost. A common practice is to try out different models without focusing on tuning the hyperparameters, in order to shorten the list of promising models [1]. Following this practice, we use the default hyperparameters for each model and the results show that XGBoost is the best. The average RMSE cross validation scores of linear regressions, decision tree, random forest, and XGBoost are 136,409, 69,622, 56,328, 50,814, respectively. ANN is also well known for its performance, so we also consider ANN as one promising model in our preliminary selection.

After the preliminary selection, we use the training set to train of all models. Although models such as the linear model are not in the short list, we still perform training on them, as discussed in the subsequent paragraphs. Our work on water consumption and withdrawal also follows the same methodology.

Since XGBoost and ANN are in the short list, these two models are presented in detail here. Gradient Boosting is an ensemble learning approach, which sequentially adds predictors into an ensemble. The successor corrects its predecessor to improve the learning, specifically by considering the residual errors made by the predecessor. eXtreme Gradient Boosting (XGBoost) is an optimized implementation of Gradient Boosting, as provided by [3]. In XGBoost, a validation set is mandatory. We split the training set into a validation set (20% of the training set), and a reduced training set (80% of the training set), using the Scikit-learn ''train_test_split'' method. We use the validation set for the parameter of ''eval_set'' in the regressor's ''fit'' function during training. We set 10 as the value of the ''max_depth'' hyperparameter and use default values for all other hyperparameters. Hyperparameter tuning was performed manually. We are interested in exploring non-manual ways in our future research.

Similarly, ANN also needs a validation set. The validation set is provided to the ''validation_data'' parameter during training. We split the training set into a validation set (20% of the training set), and a reduced training set (80% of the training set), using the Scikit-learn ''train_test_split'' method. ANN simulates the way a neuron works in a human brain. ANN consists of an input layer, one or more hidden layers, and an output layer. Figure 3 is a sketch of our ANN model for generation prediction. In the input layer, the number of nodes is the same as the dimensions of the training data. The 6 features in Table 1 produce 64 dimensions after one-hot and hashing encoding. We use Keras sequential mode [4] to add two hidden layers; each layer has 253 neurons. The output layer has only one node in this regression problem, which corresponds to plant-level generation. We tried different quantities of hidden layers, number of neurons, and learning rates. The architecture was designed with the assistance of *RandomizedSearchCV* of *sklearn* [2]. The optimizer is ''nadam''. The

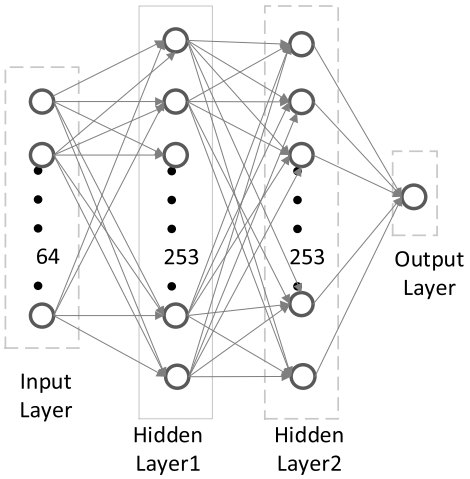


FIGURE 3. ANN for generation prediction.

output of each layer is processed by the activation function of *ReLU* [33], which is defined as $r(z) = max(0, z)$. The structure is summarized as follows [33]. The input layer is $\vec{X} = [x_0, x_{1,\ldots}, x_{63}]$. In general, the layer can be presented as $\left[N_0^{(j)}, N_1^{(j)}, \cdots N_{252}^{(j)}\right]$, where $j = 1, 2, 3$. The first hidden layer is $\left[N_0^{(1)}, N_1^{(1)}, \cdots N_{252}^{(1)}\right]$. In the first hidden layer, a neuron is $N_i^{(1)} = r(\sum_{k=0}^{63} w_{ki}^{(0)} x_k)$, in which we use the input and the associated weights. The second hidden layer is $\left[N_0^{(2)}, N_1^{(2)}, \cdots N_{252}^{(2)}\right]$, where $N_i^{(2)} = r(\sum_{k=0}^{252} Output^{(1)} w_{ki}^{(1)})$, which uses the output from the first hidden layer and associated weights. The third layer produces the output, which has one output for this regression problem: $N_0^{(3)} = r(\sum_{k=0}^{252} Output^{(2)} w_{k0}^{(2)})$.

After a model is trained, we perform testing using the test set for each model. To ensure that our choices of XGBoost and ANN are correct, we use the same test set to test each model. This approach is applied to water withdrawal and consumption prediction too. Default values of the hyperparameters are used except with XGBoost and ANN.

The results of testing different models using the test set are shown in Table 3. In all the tables of this article, we use bold fonts to highlight the best results. Specifically, the following are bolded: the lowest RMSE, NRMSE, MAE, and RSR value; PBIAS with the smallest absolute value; the highest R^2, WIA, and NSEC value; the name of the best model. As shown in Table 3, XGBoost and ANN have the smallest RMSE, NRMSE, MAE, and RSR values. XGBoost and ANN also have the highest R^2 score, WIA, and NSEC. As shown in Table 3, XGBoost outperforms ANN slightly. Random Forest is better than decision tree, and decision tree is better than the linear model in most of the statistical measures. For example, the RMSE of linear regression is about three times that of ANN. RMSE of decision tree and random forest is about 34% and 9% more, respectively, than XGBoost. The

TABLE 3. Performance Comparison Using Different Machine Learning Models.

	Linear Regression	Decision Tree	Random Forest	*XGBoost*	ANN
RMSE	135,713	68,938	55,861	***51,311***	52,353
NRMSE	0.0433	0.022	0.0178	**0.0159**	0.0163
R^2	0.5088	0.8732	0.9168	**0.9311**	0.9283
MAE	63,756	19,008	16,743	**14,607**	16,203
RSR	0.701	0.356	0.288	**0.262**	0.268
PBIAS	**0.0145**	-0.6789	-0.3144	0.0146	3.2691
WIA	0.8095	0.9672	0.9777	**0.9819**	0.9812
NSEC	0.5088	0.8732	0.9168	**0.9311**	0.9283

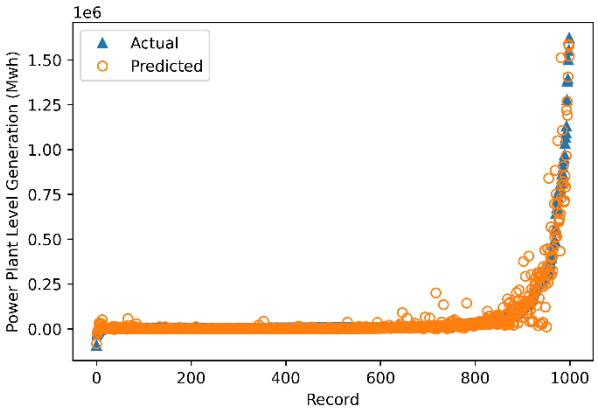


FIGURE 4. Performance of XGBoost (Using ordered records).

PBIAS values of XGBoost and Linear regression show the smallest overestimation of the two models. Overall, the best models are XGBoost and ANN. We did further evaluation using AEO data that is presented in the next subsection. Based on the results of both evaluations, we select ANN for our production environment.

The plotted chart of XGBoost and ANN for 1000 ordered rows/records are shown in Figure 4 and Figure 5, respectively. We order the records in ascending order of the generation values (label) in order to show the relationship between the predicted values and the actual values. Figure 4 and Figure 5 visually show that the predicted values are close to the actual values. Using unordered records, we randomly select 1,000 records, plot the results with the X-axis as the predicted values and the Y-axis as the actual values, as shown in Figure 6 and Figure 7. If a predicted value is the same as the actual value, the dot falls on the blue line. The figures illustrate that a majority of dots are on or close to the line, which indicates that our predictions have relatively small deviations from the actual values in both XGBoost and ANN.

C. COMPARISON WITH THE COEFFICIENT APPROACH

As described above, the coefficient approach considers the historic percent-contribution of a plant over each fuel-mover-EMM region combination to disaggregate the future generation given by the EIA AEO 2019. One main limitation of this approach is that if a plant has a negative generation value in 2018, we have to set it to zero to avoid offsetting the total. As a result, this plant will be excluded from any future

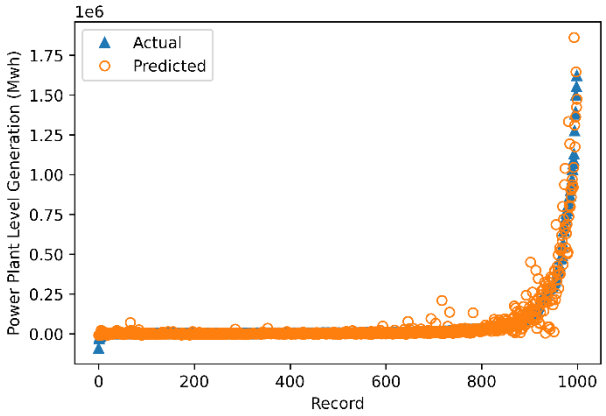


FIGURE 5. Performance of ANN (Using ordered records).

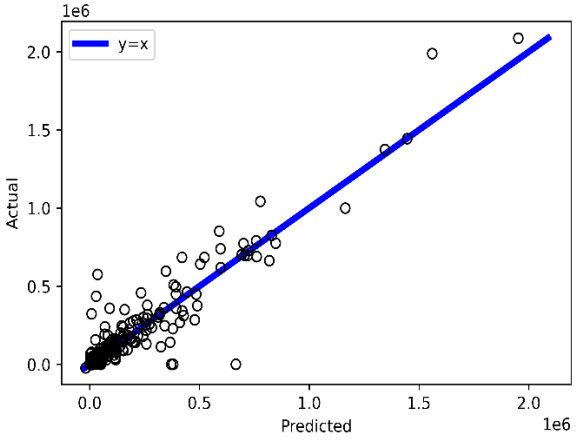


FIGURE 6. Comparison of predicted and actual values with random selected records (XGBoost).

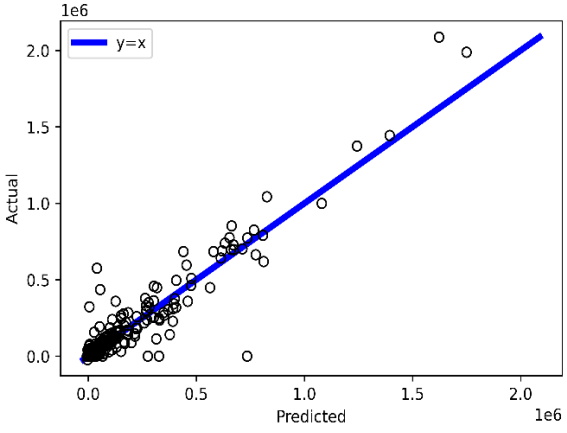


FIGURE 7. Comparison of predicted and actual values with random selected records (ANN).

prediction. This approach also assumes that a plant's future percent-contribution to its fuel-mover-EMM region quantity will remain the same in any future year and month as it did in 2018. Compared to the coefficient approach, the machine learning approach considers the history of generation since 2003, accommodates negative values, and has better results.

To compare the performance of the two approaches, we apply the coefficient method to 2017 data in order to predict the data in 2018. In the machine learning approach, we use 2003-2017 data as the training data to predict 2018. EIA AEO has the prediction for 2018, which of course varies from the actual generation reported. It is expected that this difference will be propagated to our prediction and will cause larger RMSE in both the coefficient and machine learning approach. However, because we are using AEO for our future prediction, it is reasonable to use this AEO predicted data instead of the aggregated actual data for this evaluation.

TABLE 4. Comparison of Coefficient With XGBoost and ANN.

	Coefficient	XGBoost	*ANN*
RMSE	66,307	65,385	***60,956***
NRMSE	0.0213	0.021	**0.0195**
R^2	0.8340	0.8386	**0.8597**
MAE	***15,312***	20,710	19,664
RSR	0.407	0.402	**0.375**
PBIAS	**-1.0116**	37.9179	31.465
WIA	0.958	0.9653	**0.9681**
NSEC	0.834	0.8386	**0.8597**

EWED initially had 7,271 power plants, but recently added another 2,700 plants creating a total of 9,961 plants. Most of these new plants have a scattered pattern; many of them only have a few months of generation reported instead of the whole year, and some only have data for the most recent few years. We therefore evaluated the cases for both 7,271 and 9,961 plants. The performance of both XGBoost and ANN is better without the 2,700 power plants due to the scattered nature of their data. In the case of 7,271 plants, the performance difference between the coefficient approach and ANN is more outstanding; RMSE decreased 27% and R^2 score increased from 0.828 to 0.907.

Table 4 shows the performance of 9,961 power plants based on the REF19 case model, in which ANN outperforms XGBoost. Both ANN and XGBoost have a better performance than the coefficient approach. As shown in Table 4, compared to the coefficient approach, the RMSE of ANN decreased 8.1%; R^2 increased 0.0257; RSR decreased 0.032; WIA increased 0.01; NSEC increased 0.0257. The MAE and PBIAS of the coefficient model are better. When training ANN, a "loss" parameter, such as "mean_squared_error" or "mean_absolute_error", needs to be specified as the goal of a regression problem. The environmental and computer scientists in our team collectively decided that Mean Squared Error is the best way to evaluate both generation and water use, so we used it as the training goal in the current version of the project. As a result, the trained ANN is optimized for RMSE, not for other statistical measures such as MAE (only one value such as MSE can be chosen for the loss parameter). We are interested in exploring other regression "loss" values in our future work.

Figure 8 shows the RMSE comparisons of ANN with the Coefficient approach using 9,961 plants, based on each of

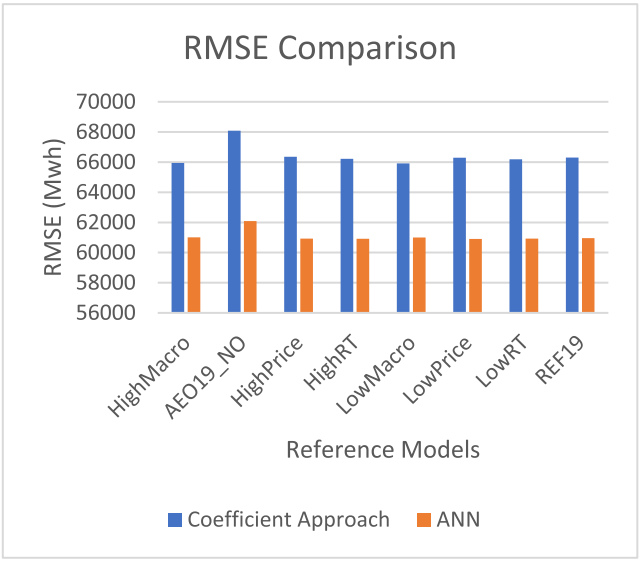


FIGURE 8. RMSE comparisons in different case models.

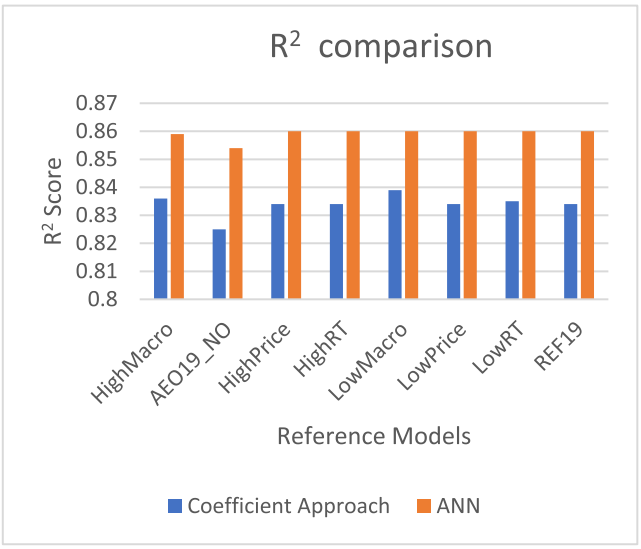


FIGURE 9. R^2 comparisons in different case models.

the eight AEO 2019 case models. Similarly, Figure 9 shows the comparisons of R^2 scores. The X-axis depicts the eight case models. The Y-axis values correspond to RMSE in Figure 8 and R^2 score in Figure 9. Visually, we can see that ANN outperformed the coefficient approach in all cases.

IV. PREDICTION OF WATER CONSUMPTION AND WITHDRAWAL

Cooling water data retrieved from EIA [24] include raw values for water withdrawal and consumption self-reported by power plants to the EIA in years 2014-2018. Power plants typically withdraw water from the nearby water sources such as lakes, river, and reservoirs. Water consumption is the amount that is withdrawn but does not return to the water source. In the coefficient approach used by EWED, for each month of a power plant's operation (from 2014-2018), we calculate the ratio of total water consumption divided by total generation.

TABLE 5. Features of Water Consumption.

Features	Description	Type
Plant Code	Unique identifier of a power plant	Discrete
EMM Code	Unique name for each EMM region	Categorical
Dominant type	Dominant combination of fuel type and prime mover	Categorical
Generation	Power generation of a power plant	Continuous
Month	Month	Discreate
Recent consumption	Last year's water consumption in the given month	Continuous

TABLE 6. Features of Water Withdrawal.

Features	Description	Type
Plant Code	Unique identifier of a power plant	Discreate
EMM Code	Unique name for each EMM region	Categorical
Dominant type	Dominant combination of fuel type and prime mover	Categorical
Generation	Power generation of a power plant	Continuous
Month	Month	Discrete
Recent withdrawal	Last year's water withdrawal in the given month	Continuous

Next, using the predicted monthly generation multiplied by this ratio, we calculate predicted water consumption. Water withdrawal can be calculated similarly. This article presents an alternative approach using machine learning models to achieve better performance.

A. FEATURE SELECTION

The amount of water withdrawal and consumption is related to the generation amount and the fuel-mover type of the power plant. The amount of withdrawal can be different in different months, since warmer weather may require more cooling water. Accordingly, since weather differs from region to region, water use rates can vary widely across the United States. However, individual power plants generally follow a similar pattern of water use from year to year. The highest correlation feature which predicts future consumption/withdrawal is recent consumption/withdrawal. The features for water consumption and withdrawal are listed in Table 5 and Table 6, respectively. These features are from different data sources and are stored in different tables in the database. We use complex queries and views to join and aggregate data to produce the features.

B. TRAINING AND COMPARISON

We did training, validation, and testing using Linear Regression, Decision Tree, Random Forest, XGBoost, and ANN. As discussed in Section III (B), we use default hyperparameters for Linear Regression, Decision Tree, and Random Forest. In XGB, we use "max_depth = 12" and use default for other hyperparameters. With the assistance of *RandomizedSearchCV* of *sklearn* [2], the ANN model for water consumption consists of 2 hidden layers, with each hidden layer

TABLE 7. Different Models of Water Consumption.

	Linear Regression	Decision Tree	Random Forest	XGBoost	*ANN*
RMSE	167.62	143.20	114.91	107.25	***95.87***
NRMSE	0.0643	0.0549	0.0441	0.0411	**0.0368**
R^2	0.4586	0.6048	0.7455	0.7783	**0.8229**
MAE	92.92	43.71	48.66	**38.20**	38.23
RSR	0.736	0.629	0.504	0.471	**0.421**
PBIAS	1.5038	***0.2445***	1.268	-8.1555	5.6962
WIA	0.7836	0.8946	0.9222	0.9314	**0.9494**
NSEC	0.4586	0.6048	0.7455	0.7783	0.8229

TABLE 8. Different Models of Water Withdrawal.

	Linear Regression	Decision Tree	Random Forest	*XGBoost*	ANN
RMSE	9438.11	4035.11	4148.19	***3206.03***	3670.32
NRMSE	0.0604	0.0258	0.0265	**0.0205**	0.0235
R^2	0.4409	0.8978	0.8920	**0.9355**	0.9155
MAE	5873.89	1313.00	1498.13	**1092.40**	1256.10
RSR	0.748	0.32	0.329	**0.254**	0.291
PBIAS	6.5587	**-0.3174**	0.9898	-2.0064	1.4519
WIA	0.788	0.9741	0.971	**0.9829**	0.9781
NSEC	0.4409	0.8978	0.892	**0.9355**	0.9155

containing 344 neurons. The input layer has 78 nodes, due to one-hot and hash encoding of features. ''Adam'' is the optimizer and the learning rate is 0.00032983006724298584. The activation function is *ReLU*. The ANN model for water withdrawal consists of one input layer with 78 nodes, 2 hidden layers with 253 neurons in each layer, and one node in the output layer. The optimizer is ''Adam'' and the learning rate is 0.0012178834831452913. The activation function is *ReLU*.

Table 7 and Table 8 present the evaluation results for different models. The best model for water consumption is ANN. The best model for water withdrawal is XGBoost. As shown in Table 7 for water consumption, ANN outperforms all other models in RMSE, NRMSE, R^2, RSR, WIA, and NSEC. ANN is only 0.08% more than XGBoost in MAE. Decision tree has the best performance in PBIAS, but is worse than XGBoost and ANN in other measures. As shown in Table 8 for water withdrawal, XGBoost outperforms others on RMSE, R^2, MAE, RSR, WIA, and NSEC. Decision tree is the best regarding PBIAS, but is worse than ANN in other measures. Based on the results, we selected to use ANN for water consumption and XGBoost for water withdrawal.

To evaluate the coefficient approach, we used 2014 to 2017 EIA cooling water data to predict 2018 data based on the coefficient method, then further compared the results with the actual data in 2018. Table 9 presents a comparison of the coefficient approach with the machine learning approach. Table 9 shows the decrease of RMSE, NRMSE, MAE, and RSR and the increase of R^2, MAE, WIA, NSEC using the machine learning models. The machine learning approach demonstrates dramatic improvement in comparison to the coefficient approach in water usage, with an RMSE decrease of 59% and 53% in water consumption and withdrawal, respectively. For water consumption, ANN outperforms the coefficient approach in all measures; for example, MAE decreased 16.8%, R^2 increased 0.41, RSR decreased 0.345, WIA increased 0.1756, and NSEC increased 0.4168. For water withdrawal, XGBoost outperforms the coefficient approach in RMSE, R^2, MAE, RSE, WIA, and NSEC. For example, R^2 increased 0.1778; MAE decreased 33%; RSR increased 0.238; WIA increased 0.048; NSEC increased 0.1778.

TABLE 9. Comparison in Water Consumption and Withdrawal.

	Water Consumption		Water Withdrawal	
	ANN	Coefficient	XGBoost	Coefficient
RMSE	***95.87***	232.45	***3206.03***	6831.25
NRMSE	**0.0368**	0.0890	**0.0205**	0.0430
R^2	**0.8229**	0.4131	**0.9355**	0.7577
MAE	**38.23**	45.95	**1092.40**	1630.61
RSR	**0.421**	0.766	**0.254**	0.492
PBIAS	**5.6962**	-7.0364	-2.0064	**-1.9006**
WIA	**0.9494**	0.7729	**0.9829**	0.9349
NSEC	**0.8229**	0.4131	**0.9355**	0.7577

V. DISCUSSION

This empirical study uses datasets from real-world data sources, and the results are applicable to a concrete application in planning for water scarcity in the U.S. electricity sector. During this process, the collaboration of domain experts in environmental science and data engineering contributes greatly to the learning process. The features listed in the previous sections come from heterogeneous data sources with varying structures. After data cleaning, data are stored in 108 tables and features are located in different tables. We note that the design of the table schema is for the purpose of serving the EWED application, rather than for machine learning only. We used various SQL queries and views to link data from different tables and to perform aggregations, so features are ready to be presented in one data frame before training. In comparison to querying multiple data frames, database queries are especially convenient for complex queries and nested queries. Compared to learning based on simple structures, such as a few tables (or data frames), handling complex systems like EWED is more challenging and requires more in-depth knowledge across different domains.

The generation prediction system was initially trained based on the features of plant code, EMM region, fuel-mover, regional data, and month. This is a typical set of features that a data scientist would produce based on the semantic.

However, this set of features did not yield anything better than an RMSE of 111,498 and an R^2 of 0.654991. Is there anything we can learn from the conventional coefficient approach? The coefficient approach used the generation value of the previous year (most recent available year) to calculate the percentage of contribution to future values. This motivated us to investigate the previous year's generation data in our feature selection process for the machine learning approach.

We observed that the generation of each power plant in consecutive years is related, as shown in Figure 10. The X-axis is generation per plant, and the Y-axis is the plant's generation in the previous year. Adding the recent generation feature ended up making the prediction much more accurate.

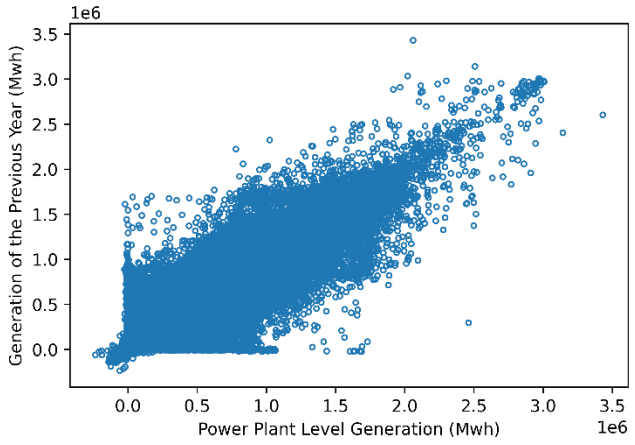


FIGURE 10. Generation data in two consecutive years.

We have a similar observation and solution for water consumption and water withdrawal. Water consumption and withdrawal are related to the type of the power plant, the generation value, month, and the region. However, using these features yielded poor results ($R^2 \leq 0.29$). A possible modification from a data scientist's point of view is to use a finer regional granularity to better reflect environmental and management conditions, with the processing cost of increased dimensions. Instead, the environmental science approach of using coefficients again motivated our feature selection process by considering the previous year's water use. This ''thinking out of the box'' way of feature selection improved the performance significantly, in which R^2 increases from 0.29 to 0.82.

On the other hand, not everything originating from the domain knowledge of environmental science contributes to a better result. One feature that seems useful is *cooling system type*, which classifies how power plants withdraw and consume water. For example, an ''open-loop'' cooling system releases almost all water to the water source, which results in a higher value of withdrawal with a lower value of consumption. Two other features that may not be as useful as cooling system type, but possibly helpful are *water type*, such as fresh, saline, and mix, and *water source*, such as ground, surface, and reclaimed. This combination of features produces 16 more dimensions after one-hot encoding. We trained different models by adding the above three features, as well as adding only cooling system type.

TABLE 10. Additional Features for Water Consumption.

Model	Measures	*None of the three features*	All three features	Cooling system type only
XGBoost	RMSE	107.25	130.35	116.14
	NRMSE	0.0411	0.062	0.0455
	R^2	0.7783	0.4237	0.7470
	MAE	38.20	38.08	34.24
	RSR	0.471	0.535	0.503
	PBIAS	-8.1555	-0.9789	-3.5175
	WIA	0.9314	0.9149	0.9245
	NSEC	0.7783	0.7142	0.747
ANN	RMSE	***95.87***	118.60	106.42
	NRMSE	**0.0368**	0.0397	0.0373
	R^2	**0.8229**	0.7634	0.7876
	MAE	**38.23**	41.48	38.87
	RSR	**0.421**	0.486	0.461
	PBIAS	5.6962	-4.1339	**1.3946**
	WIA	**0.9494**	0.9246	0.9374
	NSEC	**0.8229**	0.7634	0.7876

Table 10 presents the evaluations of all three features, only one feature, and none of the features. Similar to other evaluations presented in other tables, XGBoost and ANN are consistently better than other models, which is also the case in this comparison. To make the comparison focus on the features instead of model selection, we only present XGBoost and ANN in Table 10.

Table 10 shows that best RMSE value is 95.87, which is ANN without any of the three features. All other measures except PBIAS are also slightly better without the three features. Comparing ''none of the three features'' with ''all three features'' and ''cooling system type only'', RMSE increased 23.7% and 11%; R2 decreased 0.059 and 0.0353; MAE increased 8.5% and 1.67%; RSR increased 0.065 and 0.04; WIA decreased 0.025 and 0.012; and NSEC decreased 0.0595 and 0.0353, respectively. The conclusion is that adding the features slightly decreased the performance and the best result is without any of the three features.

VI. RELATED WORK

There are many research projects on power generation prediction, most of which are for solar power. Many of the solar power predictions are based on environmental factors, such as [5], [6], [7], [8], as mentioned in Section III. Similarly, the research in [9] used ensembled decision trees for photovoltaic power generation, based on the analysis of environmental data. The research in [10] used particle swarm optimization algorithm to optimize the weight and

threshold of the neural network to predict photovoltaic power generation.

Reported in [11], the researchers used a model-driven and data-driven approach for the prediction of wind power systems with doubly fed induction generators. Another wind power generation prediction is the research in [12], which is for small-wind turbines. The prediction is based on the atmospheric variables of the locations of wind power farms. The research in [13] predicted wind power generation using Echo State Networks, which consist of randomly connected neural networks.

The research in [14] is the prediction of thermal power generation. It provided an optimized method based on the neural network model and tailored it for multimedia applications.

Although these existing research projects are related to our research, the focus is completely different. They targeted a specific type of power generation prediction. In contrast, our research covers a wide variety of power plants with various fuel types and fuel-movers.

There is limited research on the prediction of water usage using machine learning. The research in [15] used deep learning models to predict future water consumption based on the data collected from multiple buildings in a university campus. The research in [16] modeled the dynamics of water consumption time series based on the data from smart meters by a water utility in France to predict future water consumption behaviors. It used a mixture of non-homogeneous hidden Markov models for the consumption behavior series. In [17], the researchers used fractal theory for the prediction of urban hourly water consumption in cities. The work in [18] performed anomaly detection in water treatment facilities. Genetic algorithms are used in [18] for water resource management, such as optimization of water distribution system and operation of reservoirs.

To the best of our knowledge, there is no other research using the same or similar data sets to compare our results with. Our research focused on power plant level generation, water consumption and withdrawal and as such, represents a novel synthesis of predictions in energy and water systems. Different levels of aggregations, such as HUC, county, and state, are also provided in EWED to facilitate decision making in accordance with water availability constraints.

VII. SUMMARY

The interdependency among electricity generation, emissions, water withdrawal, water consumption, and water availability has become increasingly important challenge in our modern society. Awareness of water constraints is critical for decision making in future energy management. Unfortunately, there is not enough attention nor studies on power plant water use and prediction in the existing machine learning literature, which makes our unique empirical research significant. The EWED project is a comprehensive information system that calculates and presents integrated data in a temporal and geographical environment. This article described our approach of using machine learning for the prediction of generation, water consumption, and water withdrawal, which can be used to predict water use and potential conflicts with water supplies. This approach improved on the conventional approach of EWED to use coefficients to proportionally predict results. Compared to the conventional approach, the machine learning approach produced better predictions. The performance enhancement is especially significant for water consumption and water withdrawal, with a decrease of 59% and 53% of RMSE, respectively. In addition, this article presented the practical issues raised and solved in this real-world project, including data preparation using a database-oriented method and feature selection.

To the best of our knowledge, this is the first machine learning project that has covered such a wide variety of power plants across United States with both energy and water prediction in a cohesive environment, allowing users to retrieve data at different geographic scales and perform data analysis at the finest possible granularity, as well as supporting timely decision making with rich visualization. The current version of EWED uses the EIA AEO 2019 projection. A future research direction is to retrieve the new AEO data and apply our research results to the new data for the production environment. EIA releases AEO data annually with updated and refined projections. AEO data is an important feature in our prediction, so we expect our results to be improved accordingly. In addition, we will retrieve more data from different data sources upon their new release and add them to our training data. More data for the training can also potentially improve the predictions.

ACKNOWLEDGMENT

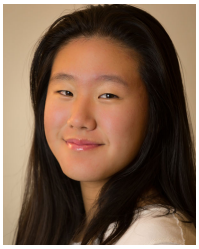
The authors would like to thank the student assistants' contributions to different phases of the EWED project (in alphabetical order): Tejaswini Bhorkar, Gaurav Bora, Trent Buchanan, Aaron Enberg, Jasmie Guan, Priya Gundlupet, Khoi Hoang, Priyanka Makwana, and Karan Mitra.

REFERENCES

[1] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. Newton, MA, USA: O'Reilly Media, 2019.
[2] *Scikit-Learn*. Accessed: Feb. 11, 2021. [Online]. Available: https://scikit-learn.org/stable/
[3] XGBoost. *XGBoost Documentation*. Accessed: Feb. 11, 2021. [Online]. Available: https://xgboost.readthedocs.io/en/latest/
[4] *Keras*. Accessed: Feb. 11, 2021. [Online]. Available: https://keras.io/
[5] S. Al-Dahidi, M. Louzazni, and N. Omran, "A local training strategy-based artificial neural network for predicting the power production of solar photovoltaic systems," *IEEE Access*, vol. 8, pp. 150262–150281, 2020.
[6] S. Al-Dahidi, O. Ayadi, M. Alrbai, and J. Adeeb, "Ensemble approach of optimized artificial neural networks for solar photovoltaic power prediction," *IEEE Access*, vol. 7, pp. 81741–81758, 2019.
[7] A. Dahl and E. V. Bonilla, "Grouped Gaussian processes for solar power prediction," *Mach. Learn.*, vol. 108, nos. 8–9, pp. 1287–1306, Sep. 2019.
[8] W. Cabrera, D. Benhaddou, and C. Ordonez, "Solar power prediction for smart community microgrid," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, St. Louis, MO, USA, May 2016, pp. 1–6.
[9] S. Zhang, H. Dai, A. Yang, and Z. Shi, "Environmental parameters analysis and power prediction for photovoltaic power generation based on ensembles of decision trees," *Advances in Information and Communication Technology*, vol. 581. Cham, Switzerland: Springer. 2020, pp. 78–85.

[10] Y. Li, Y. Wan, J. Xiao, and Y. Zhu, "Prediction of photovoltaic power generation based on POS-BP neural network," in *Communications in Computer and Information Science*, vol. 1160. Singapore: Springer, 2020, pp. 443–453.
[11] J. Yi, W. Lin, J. Hu, J. Dai, X. Zhou, and Y. Tang, "An integrated model-driven and data-driven method for on-line prediction of transient stability of power system with wind power generation," *IEEE Access*, vol. 8, pp. 83472–83482, 2020.
[12] B. Baruque, E. Jove., S. Porras, J.L.Calvo-Rolle, "Small-wind turbine power generation prediction from atmospheric variables based on intelligent techniques," in *Advances in Intelligent Systems and Computing*, vol. 1268. Cham, Switzerland: Springer, 2021, pp. 33–43.
[13] R. R. B. de Aquino, O. N. Neto, R. B. Souza, M. M. S. Lira, M. A. Carvalho, T. B. Ludermir, and A. A. Ferreira, "Investigating the use of echo state networks for prediction of wind power generation," in *Proc. IEEE Symp. Comput. Intell. Eng. Solutions (CIES)*, Orlando, FL, USA, Dec. 2014, pp. 148–154.
[14] F. Chen, Z. Fu, and L. Zhen, "Thermal power generation fault diagnosis and prediction model based on deep learning and multimedia systems," *Multimedia Tools Appl.*, vol. 78, no. 4, pp. 4673–4692, Feb. 2019.
[15] G. Bejarano, A. Kulkarni, R. Raushan, A. Seetharam, and A. Ramesh, "SWaP: Probabilistic graphical and deep learning models for water consumption prediction," in *Proc. 6th ACM Int. Conf. Syst. Energy-Efficient Buildings, Cities, Transp.*, Nov. 2019, pp. 233–242.
[16] N. Zhi-guang, C. Fa, and L. Ren-qiang, "Study on fractal prediction model of urban hourly water consumption," in *Proc. 5th Int. Conf. Natural Comput.*, Tianjin, China, 2009, pp. 154–158.
[17] H. Haimi, M. Mulas, F. Corona, S. Marsili-Libelli, P. Lindell, M. Heinonen, and R. Vahala, "Adaptive data-derived anomaly detection in the activated sludge process of a large-scale wastewater treatment plant," *Eng. Appl. Artif. Intell.*, vol. 52, pp. 65–80, Jun. 2016.
[18] D. Rani, S. K. Jain, D. K. Srivastava, and M. Perumal, "Genetic algorithms and their applications to water resources systems," in *Metaheuristics in Water, Geotechnical and Transport Engineering*, Amsterdam, The Netherlands: Elsevier, 2013, pp. 43–78.
[19] US Environmental Protection Agency. *Facility Registry Service*. Accessed: Feb. 11, 2021. [Online]. Available: https://www.epa.gov/frs/frs-data-resources
[20] *TIGER/Line Shapefiles*. Accessed: Feb. 11, 2021. [Online]. Available: https://www.census.gov/cgi-bin/geo/shapefiles/index.php
[21] United State Department of Agriculture. *WaSSI Ecosystem Services Model*. Accessed: Feb. 11, 2021. [Online]. Available: https://web.wassiweb.fs.usda.gov/
[22] US Environmental Protection Agency. *Greenhouse Gas Reporting Program (GHGRP)*. Accessed: Feb. 11, 2021. [Online]. Available: https://www.epa.gov/ghgreporting/ghgreporting-program-data-sets
[23] Energy Information Administration (EIA). *Open Data*. Accessed: Feb. 11, 2021. [Online]. Available: https://www.eia.gov/opendata/qb.php?category=1017
[24] Energy Information Administration (EIA). *Thermoelectric Cooling Water Data*. Accessed: Feb. 11, 2021. [Online]. Available: https://www.eia.gov/electricity/data/water/
[25] Energy Information Administration (EIA). *Annual Energy Outlook 2019*. Accessed: Feb. 11, 2021. [Online]. Available: https://www.eia.gov/outlooks/archive/aeo19/
[26] *United State Department of Agriculture, Prediction of Water Availability*. Accessed: Feb. 11, 2021. [Online]. Available: https://web.wassiweb.fs.usda.gov/s/profile
[27] S. Learn. *Category Encoders, Hashing*. Accessed: Feb. 11, 2021. [Online]. Available: http://contrib.scikit-learn.org/category_encoders/hashing.html
[28] G. Bonaccorso, *Machine Learning Algorithms*. Birmingham, U.K.: Packt, 2018.
[29] M. Shafiullah, M. Abido, and T. Abdel-Fattah, "Distribution grids fault location employing ST based optimized machine learning approach," *Energies*, vol. 11, no. 9, p. 2328, Sep. 2018.
[30] Y. D. Mamuya, Y.-D. Lee, J.-W. Shen, M. Shafiullah, and C.-C. Kuo, "Application of machine learning for fault classification and location in a radial distribution grid," *Appl. Sci.*, vol. 10, no. 14, p. 4965, Jul. 2020.
[31] M. S. Shahriar, M. Shafiullah, M. J. Rana, A. Ali, A. Ahmed, and S. M. Rahman, "Neurogenetic approach for real-time damping of low-frequency oscillations in electric networks," *Comput. Electr. Eng.*, vol. 83, May 2020, Art. no. 106600.
[32] S. A. Razzak, S. A. Hossain, S. M. Rahman, M. M. Hossain, and J. Zhu, "A multigene genetic programming approach for modeling effect of particle size in a liquid–solid circulating fluidized bed reactor," *Chem. Eng. Res. Des.*, vol. 134, pp. 370–381, Jun. 2018.
[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

YING JIN (Member, IEEE) received the Ph.D. degree in computer science and engineering from Arizona State University, in 2004. She is currently a Professor with the Department of Computer Science, California State University, Sacramento, where she was an Associate Professor from 2010 to 2015, and an Assistant Professor from 2004 to 2010. Her research interests include enterprise application integration, database systems, active rule systems, data science, data security, and fuzzy information systems. She is a member of the Association for Computing Machinery (ACM). She is the co-editor of two conference proceedings.



EMILY J. YANG is currently pursuing the degree with the Folsom High School. She is taking computer science courses through the Accelerated College Entrance Program at California State University, Sacramento. She is participating in the research of the Energy-Water-Emissions Dashboard Project under the supervision of Dr. Jin and Dr. Fulton, focusing on applying machine learning models to the prediction of future energy generation and water usage.

JULIAN FULTON received the Ph.D. degree in energy and resources from the University of California, Berkeley. He is currently an Assistant Professor with the Department of Environmental Studies, California State University, Sacramento, where he teaches and conducts research on issues related to water sustainability. He has authored numerous publications on water footprint, global trade in virtual water, and the water-energy-climate nexus. His research interests include urban water resources management, flood management, and stormwater management.