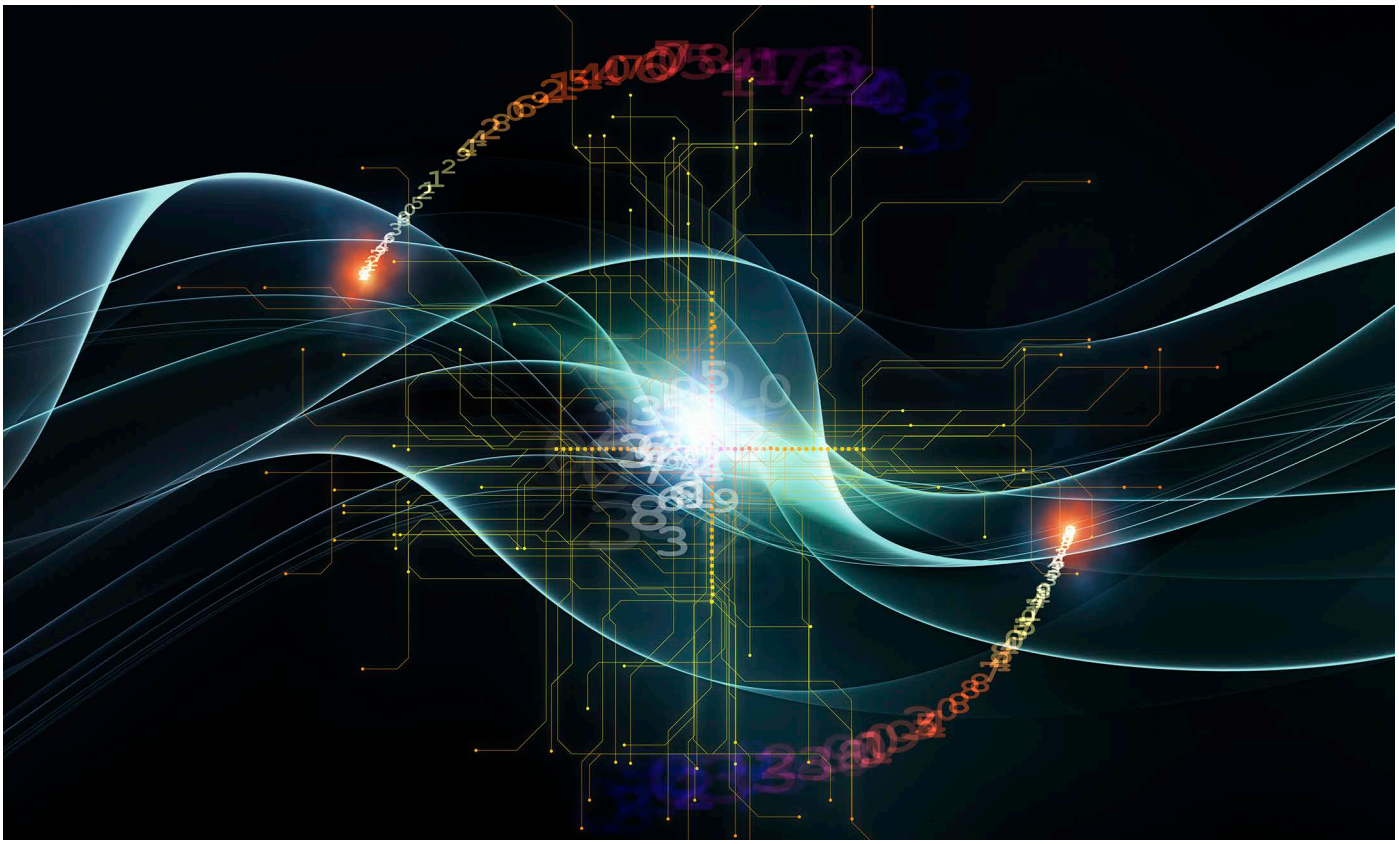
EPRI

MACHINE LEARNING FOR ANALYSIS OF POWER PLANT ENVIRONMENTAL CONTROLS

Novel Random Forest Approach Development



December 2022

DISCLAIMER OF WARRANTIES AND LIMITATION OF LIABILITIES

THIS DOCUMENT WAS PREPARED BY THE ORGANIZATION(S) NAMED BELOW AS AN ACCOUNT OF WORK SPONSORED OR COSPONSORED BY THE ELECTRIC POWER RESEARCH INSTITUTE, INC. (EPRI). NEITHER EPRI, ANY MEMBER OF EPRI, ANY COSPONSOR, THE ORGANIZATION(S) BELOW, NOR ANY PERSON ACTING ON BEHALF OF ANY OF THEM:

(A) MAKES ANY WARRANTY OR REPRESENTATION WHATSOEVER, EXPRESS OR IMPLIED, (I) WITH RESPECT TO THE USE OF ANY INFORMATION, APPARATUS, METHOD, PROCESS, OR SIMILAR ITEM DISCLOSED IN THIS DOCUMENT, INCLUDING MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, OR (II) THAT SUCH USE DOES NOT INFRINGE ON OR INTERFERE WITH PRIVATELY OWNED RIGHTS, INCLUDING ANY PARTY'S INTELLECTUAL PROPERTY, OR (III) THAT THIS DOCUMENT IS SUITABLE TO ANY PARTICULAR USER'S CIRCUMSTANCE; OR

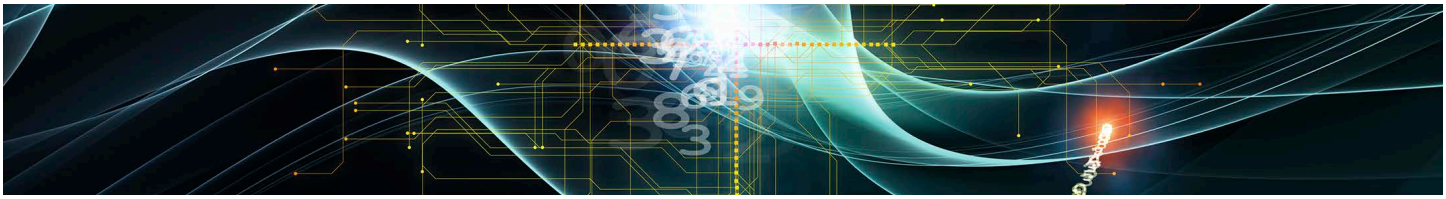
(B) ASSUMES RESPONSIBILITY FOR ANY DAMAGES OR OTHER LIABILITY WHATSOEVER (INCLUDING ANY CONSEQUENTIAL DAMAGES, EVEN IF EPRI OR ANY EPRI REPRESENTATIVE HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES) RESULTING FROM YOUR SELECTION OR USE OF THIS DOCUMENT OR ANY INFORMATION, APPARATUS, METHOD, PROCESS, OR SIMILAR ITEM DISCLOSED IN THIS DOCUMENT.

REFERENCE HEREIN TO ANY SPECIFIC COMMERCIAL PRODUCT, PROCESS, OR SERVICE BY ITS TRADE NAME, TRADEMARK, MANUFACTURER, OR OTHERWISE, DOES NOT NECESSARILY CONSTITUTE OR IMPLY ITS ENDORSEMENT, RECOMMENDATION, OR FAVORING BY EPRI.

EPRI PREPARED THIS REPORT.

NOTE

For further information about EPRI, call the EPRI Customer Assistance Center at 800.313.3774 or e-mail askepri@epri.com.



Introduction

Overview

Several industries are working to harness the power of Artificial Intelligence (AI) to improve safety, quality control, increase productivity, performance, efficiency, and customer satisfaction. AI is transforming businesses in a similar way to how electricity transformed industries over 100 years ago. AI is used when traditional programming is not applicable to model high-volume data sources and complex processes that require constant human interaction. The technology works by employing computer science, data science, and engineering knowledge to create an intelligent machine or an expert system framework that helps with decision making or simply for knowledge archiving. In summary, AI works to augment human expertise and capture knowledge about events, incidents, and faults in a system or an industrial process.

Past Applications in the power industry

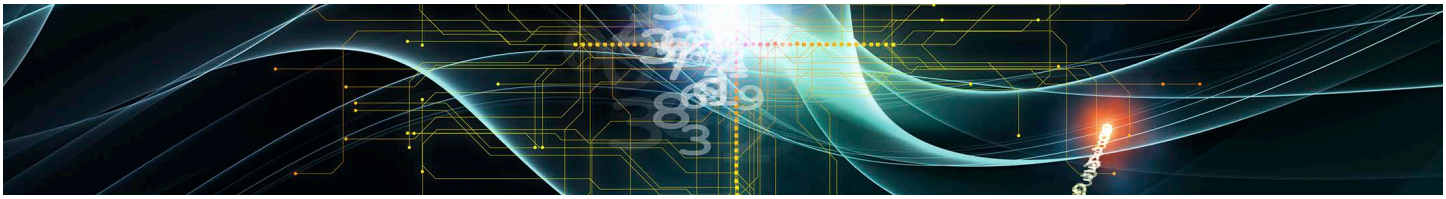
As AI applications continue to expand in the power industry, some implementations were used to predict gaseous emissions from electrical generating units, such as NOx, SOx, and mercury. Methods for globally enhanced general regression neural networks (GE-GRNN), as presented in [Song 2017], were applied. The input space to carry out that testing included five parameters (features or predictors) such as boiler generating load, burner tilt angles, secondary air flows, over fire air damper positions and boiler exit oxygen concentration. Their effort focused on modeling the boiler output NOx emissions and fuel burning efficiency. AI models such as Principal Component Analysis (PCA), K-Means clustering, Artificial Neural Networks (ANNs), and Support Vector Machines (SVMs) have been applied to model Ammonium Bisulfate (ABS) formation temperature on air heaters as a byproduct of the Selective Catalytic Reduction reactor operation in [Nie 2017]. Forty-nine features in a dataset set comprised of 14,230 observations (samples) were employed to develop the predictive models. In addition, a PCA algorithm was used to reduce the number of model inputs (features) and K-Means Clustering was applied to reduce their sample space. The latter study involved using an Ammonium Bisulfate (ABS) fouling probe that operates by recording the temperature at which its thermally controlled probe tip detects condensation – this temperature point is recorded as the ABS formation temperature. Their best model included the use of sensitivity analysis to reduce the number of process features (inputs) from 25 to 4, K-Means Clustering to reduce the number of samples, or observations, from 12,717 to 369. SVM was used for analytical modeling. In another study [Xu 2013], NOx emissions and boiler efficiency were modeled simultaneously on a 600 MW tangential fired pulverized boiler that utilized a back-propagation neural network and genetic algorithms. The purpose of employing the genetic algorithm was to identify optimized minimum and maximum operating points for both NOx emissions and boiler efficiency, respectively. In this effort, one variation of a machine learning algorithm has been developed and tested for ranking and identifying features or predictors that affect one dependent variable.

Random Forest Background

Within the large umbrella encompassed by Artificial Intelligence, Random Forests (RF) are classified as Ensemble Machine Learning-Supervised algorithms that can be applied to solve classification or regression type problems. RFs are relatively fast to train and require little tuning. In addition, RFs are useful for analyzing datasets with large number of features, also called predictors. RFs utilize decision

Table of Contents

This white paper was prepared by EPRI.



trees which are simplified algorithms where significant research has been abundantly conducted in terms of measuring variable importance.

Decision Trees and Random Forests

Decision trees possess advantages over other machine learning algorithms when it comes to modeling complex, multi-dimensional interactions. The decision tree algorithms possess lower sensitivity and robustness to outliers as opposed to other deterministic models. On the other hand, decision trees tend to be inaccurate compared to other algorithms such as deep learning convolution and recurrent neural networks. Although decision trees, when used alone, may have inaccuracies, they are the foundation of the RF algorithms. In a RF, the trees are grown from an independent bootstrap sample of the data set and the best split point is determined based on a random number of predictors or feature variables subset m of p, where p is the total number of predictors (features) in the model. The individual trees are grown to their terminal node without pruning thereby creating high-bias models of the ensemble. The basic idea behind the ensemble is to grow multiple trees which are then combined to yield a single prediction through a bagging technique. Statistics suggest that averaging several observations yields a reduction of variance. For example, consider n observations, each with a variance of σ^2 and a variance of the mean represented by σ^2/n. In the case where the model is constrained to a small original dataset, independent bootstrap samples are drawn to conserve the concept behind using the RF algorithm. In the case of a regression fit, bagging involves training the model on different bootstrap samples, constructing many trees, and averaging their output to obtain a single bagged prediction. The pseudo-code to construct a regression decision tree is depicted in Algorithm 1. One of the primary advantages of RFs over parametric models is that learned rules are intuitive which leads to simple rules for determining feature importance. Researchers have performed analysis of bias techniques for assessing variable importance measures in RFs [Strobl 2007]. In this study, the maximum reduction in variance and the input perturbation for variable importance methods, as explained by [Estes 2016] and [Stobl 2007], were explored to determine feature importance. Testing the accuracy of the model is accomplished through splitting the original database into training and testing datasets, this is referred to as cross-validation. The split ratio of test vs. training data is a tunable parameter that is pre-determined by the user to prevent the scenarios of overfitting and underfitting. Overfitting data generates a model with high variance, whereas underfitting generates a model with high bias. The split ratio is arbitrary and subject to investigation, a typical split ratio is 80% training and 20% testing. A RF process diagram is illustrated in Figure 1 and a simple RF pseudocode is presented in Algorithm 2. Since the random forest pulls a bootstrap sample from the original dataset, it is important to highlight the inner workings of this methodology. Considering that the original dataset contains N observations and p features, the bootstrapping is the probability of selecting a single observation which is $1/N$.

Algorithm 1: Pseudo-code for the construction of a regression decision tree

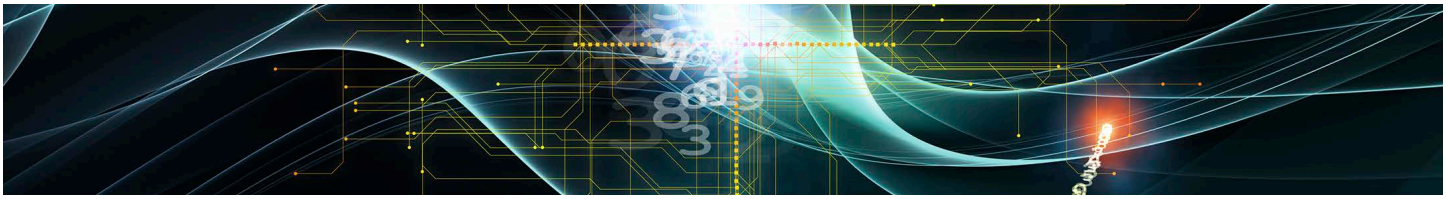
Model:
Given p predictors (features) and N observations (samples) assuming a continuous dependent variable

While "Stopping Criteria" is not met
  Search for the best split amongst all predictor variables per the variance reduction formula in equation 1
  Split the node into a left and right descendant nodes per the best split value
End While

Prediction:
To make a prediction based on an array of predictors values x, pass x down the tree until K terminal nodes are reached
Record the values observed at all K terminal nodes
Average out the response values according to the formula below

$$F(x) = \frac{1}{K}\sum_{k=1}^{k=K} h_k(x)$$

Performing n selections with replacement for a large dataset (allowing for duplicates) yields, on average, in 36.8% of the observations not selected. The latter indicates that a bootstrap sample contains approximately 63% of the total number of observations.



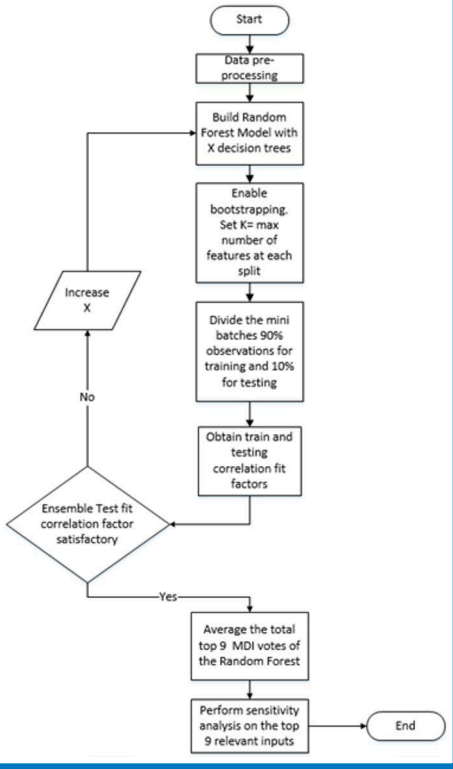


Figure 1. Flow chart depicting the standard Random Forest Model

Algorithm 2: Pseudo-code for the construction of a random forest

Model:
For h=1.....J trees
Given *p* predictors and *N* bootstrap from the original dataset
Fit a tree using the bootstrap sample
While "Stopping Criteria" is not met
Search for the best split amongst random *m* predictor variables per the variance reduction formula in equation 1
Split the node into a left and right descendant nodes per the best split value
End While
Next h

Prediction:
To make a prediction based on a new data point *x*, where $\hat{h}$ represents a single tree prediction.

$$F(x) = \frac{1}{J}\sum_{h=1}^{h=j} \hat{h}_j(x)$$

Forest of Forests

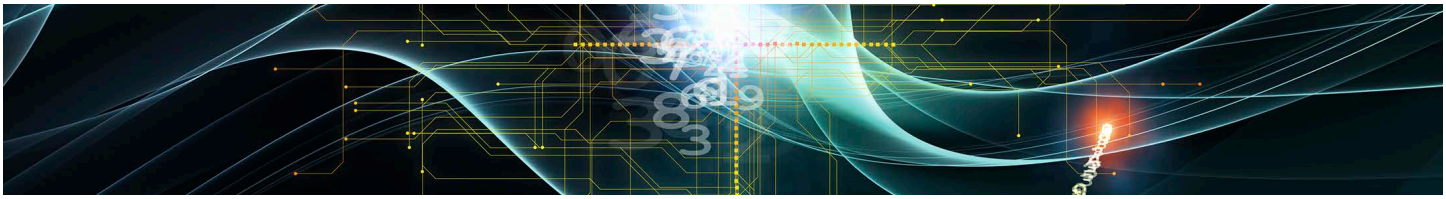
The novel Forests of Forests (FOF) model consists of aggregating a pre-specified number of RFs with the goal of finding an optimum balance between the bias and variance of the predictions. This modeling technique was first implemented for analysis of an air quality control system [Bazzi 2018]. The concept of this approach is 'divide and conquer', that is, the dataset is shuffled randomly and divided into mini batches. The total number of mini batches governs the number of random forests that are utilized. The feature importance algorithms, such as mean decrease in impurity or the input perturbation, can be applied to each forest and then aggregated across the total number of random forests. The simple pseudo-code for the FOF is explained in Algorithm 3. Aggregating multiple random forests produces a robust model that optimizes the bias versus variance tradeoff when making predictions. Combining multiple machine learning algorithms reduces the likelihood of over-fitting thereby preventing the user from utilizing additional algorithms for regularization purposes to produce a low variance and a low bias model.

Algorithm 3: Pseudo-code for the construction of an FOF model

Model:
Shuffle data randomly
Divide the training set into B mini batches
While (b is less or equal to max(B))
For h=1.....J trees
Given *p* predictors and *N* bootstrap from the original dataset
Fit a tree using the bootstrap sample
While "Stopping Criteria" is not met
Search for the best split amongst random *m* predictor variables per the variance reduction formula in equation 1
Split the node into a left and right descendant nodes per the best split value
End While
Next h
End While

Prediction:
To make a prediction based on a new data point *x*, where $\hat{h}$ represents a single tree prediction.

$$F(x) = \frac{1}{J}\sum_{h=1}^{h=j} \hat{h}_j(x)$$



Application Description

The current analytical capability of the regression RF and FoF algorithms has been integrated into a desktop application called Forest Ensemble Machine Learning (ForEMaL) Explorer. The Windows-based prototype software incorporates data preprocessing, feature and observation filtering, model parameter adjustment, code execution and postprocessing of results. The software analyzes data from a ".csv" file obtained directly from a DCS system. Upon loading, the software performs data validation and cleanup. Additional filtering for specific user-selected ranges can then be performed. For initial model development, the creation of an RF model utilizing the software requires the contributions of a data science engineer and a process subject matter expert. In the latest version, the user can adjust model hyperparameters to optimize the model. Figure 2 illustrates a high-level workflow for the application.

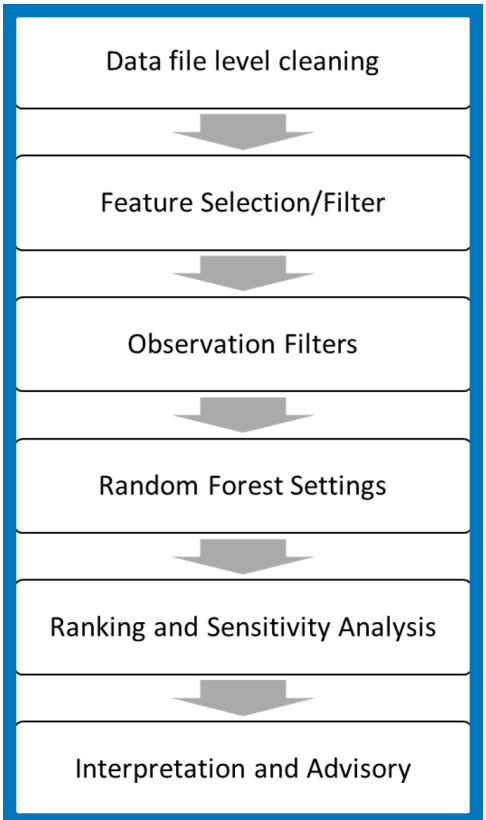


Figure 2. Overview of FoREMaL Explorer Workflow

The FoREMaL Explorer prototype has been tested with various case studies within the realm of power plant process and environmental controls. The software applicability is not limited to a specific process. Rather, the limitations are dependent on the data quality, availability over a periodic timeline, and the inclusion of a dependent variable of interest.

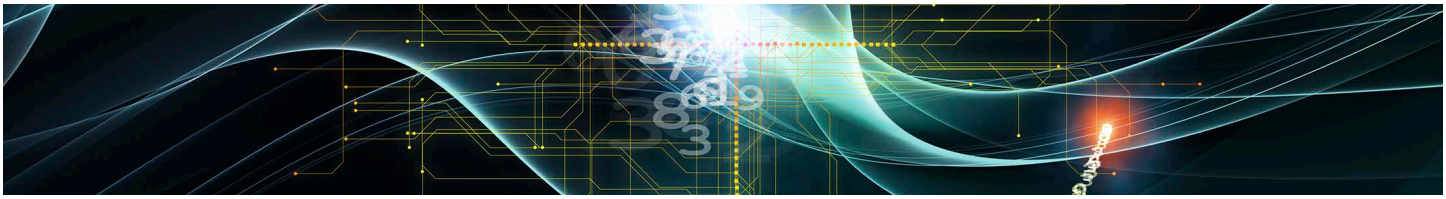
Application Case Studies

Case I: Electrostatic Precipitator Opacity Excursions

The FoREMaL Explorer prototype was utilized to analyze a known performance issue of the particulate matter control system at a 640 MWg coal-fired electrical generating unit (EGU.) Although the RF analysis was conducted after a solution was identified and implemented in the field, the data set provides a good opportunity for a validation test case on the application of the RF approach. FoREMaL Explorer was utilized to troubleshoot continuous opacity excursions. These opacity excursions required that the EGU limit their generating capacity during those time periods in order to maintain environmental compliance.

The EGU furnace design is opposed-wall burner configuration with balanced-draft control. Various blends of bituminous and subbituminous coals fueled this plant. Under full-load operation, coal is supplied to the furnace by six pulverizers through 48 burners. The flue gas particulate control system is an electrostatic precipitator (ESP) with four chambers. Each ESP chamber contains six collector electrical fields in the flue gas flow direction. The gas passages within the chambers are spaced 18" apart and their respective height is 41 feet. The ESP total specific collector area (SCA) is 669 ft^2/kacfm of flue gas which is properly sized for the particulate loading to maintain opacity within the required limits. The ESP electrical fields incorporate 48 transformer-rectifier (TR) sets that are managed by a Forry control system. The TR-set system design secondary voltage and output current are 50 kV and 2500 mA, respectively. Four induced draft fans (IDF) are located downstream of each ESP chamber and pulled flue gas through a common cross-balancing duct before release via a stack into the atmosphere. Louver dampers at each IDF inlet are modulated to control flue gas flow. The equipment layout is presented as a top view schematic in Figure 3.

The EGU's operating data was retrieved from an OSIsoft process historian. FoREMaL Explorer was utilized to troubleshoot opacity excursions observed over a 10-day period. During this timeframe, the opacity baseline shift increased from an average of 1% to 8.5%. The EGU opacity is targeted for less than 20%. The periods when the opacity exceedances occurred are shown in Figure 4.



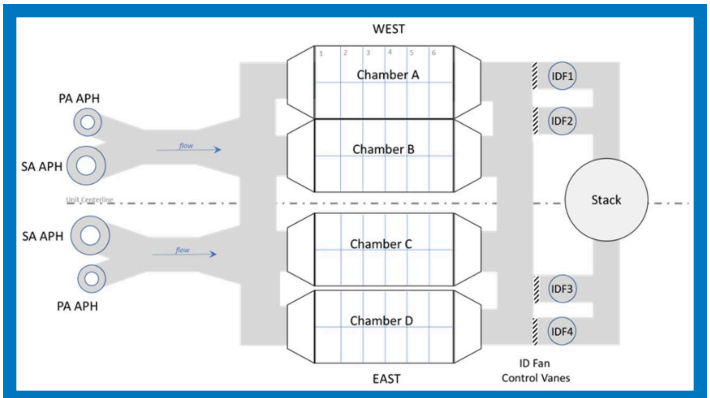


Figure 3. Flue gas pathway from air heaters, through ESP chambers and out the stack

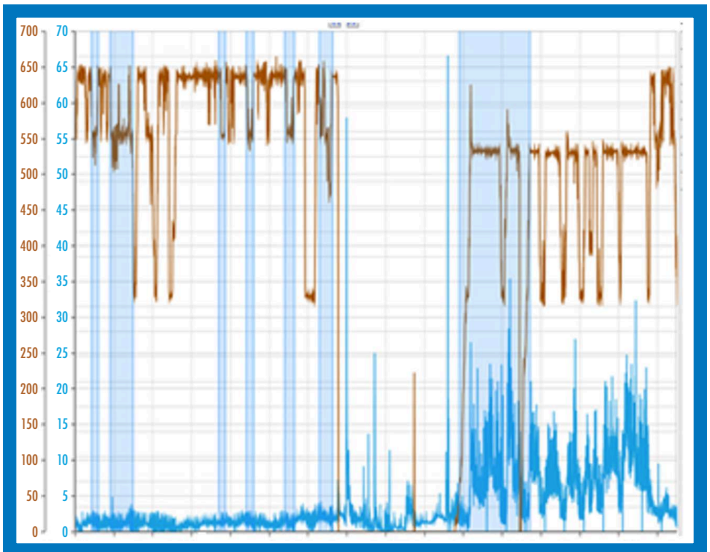


Figure 4. Opacity % (blue line) and generating load in MW (brown line) trends over time showing reduced generation load during times when opacity spikes occurred.

Approach

To construct the RF model, 300 process DCS tags (features) were selected and included: Unit gross load, TR-Set voltages, TR-set currents, spark rates, firing angles, IDF motor current, IDF control vane positions and opacity. To incorporate time periods with and without opacity excursions, 10-min rolling averages were extracted from the historian to construct an input CSV file. The timeline interval included over 13,000 observations. Once the CSV file was loaded into the software, hyperparameters for the RF model were entered via the GUI. The target dependent variable of interest was the opacity measurement at the stack. The analytical modeling stage is divided into three phases: Data pre-processing, model development, and post-processing. In the pre-processing phase, the features were filtered based on their relative standard deviation. Other filtering options were available to further reduce large datasets to one or more specific criteria at the discretion of the user, this is the equivalent of a single or multi-variable row filter. Additional RF parameter settings can be entered by the user to test and validate the model. These include RF or FOF approach, use of specific random number generator seed, number of trees, leaves, and several other RF hyperparameters [MS 2021]. Model development phase involved testing the outcome for differences on the rankings and on the obtained root mean squared error for each model. The calculation time for each single run was less than one minute. After several iterations, the predicted model accuracy was 0.86.

Post-Processing Model Analysis and Results

The results from the RF analysis are presented as ranked features in order of correlation agreement with opacity according to R^2 coefficient. In addition, scatter plots for opacity as a function of each ranked variable and a combined histogram and sensitivity plot are also generated. Based on these plots, a list of action items and respective feature thresholds to either minimize or maximize the dependent variable are produced. The user can then display an 'optimized' data trend that meets the list of action items thresholds. The following sections describe each of these workflow steps.

Top Rankings

Upon completion of the RF analysis, the user-specified top 15 features were ranked in descending order of importance. The feature ranking is displayed in Figure 5. Many features ranked by the analysis were located on the Chamber C west electrical sections. In

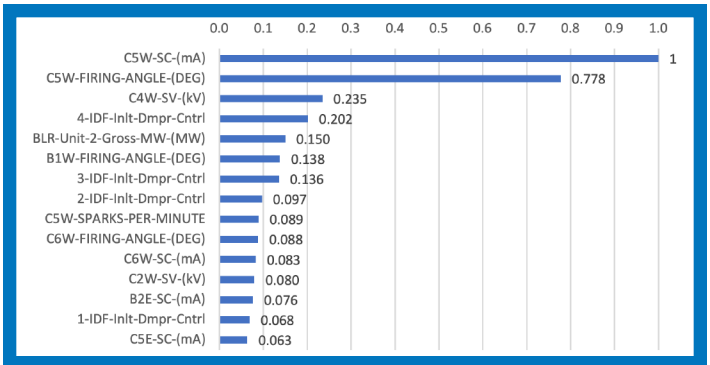


Figure 5. Bar chart describing the top 15 relevant features in the machine learning model with respect to opacity

addition, all four IDF damper controls were also ranked within the top 15 features although their order of importance differed significantly. For instance, IDF damper #4 was ranked 4th whereas the other dampers were ranked much lower. Two electrical sections were also identified on Chamber B within the top 15.

In consultation with the process engineer, these trends indicate that, granted the known ESP conditions, some chambers were working harder than others and that some IDFs were pulling more gas flow than typical at the same respective load. Unfortunately, gas flow meters were not available at these locations. A sketch of the sections identified by the RF analysis is depicted in Figure 6.

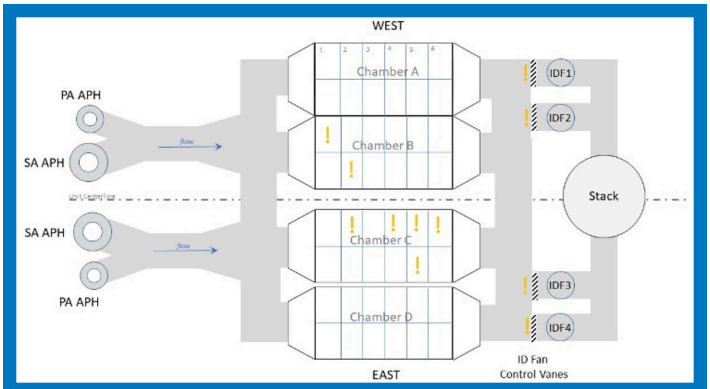


Figure 6. Sketch identifying the locations of top ranked features (!) for case study

Scatter Plots, Sensitivity Analysis and Action Lists

Scatter plot trends associated to each of the ranked features were generated by the software and are displayed in Figure 7. The scatter plots provide the user with an indication of the relationship between each feature and the dependent variable 'Stack opacity'. Also shown in Figure 8 are corresponding combination charts for each feature and these are displayed on the right side of the window. Each combo chart shows two data trends for each ranked feature: 1.) a histogram across the data range (gray bars) and a sensitivity trend (black line). The histogram frequency scale resides on the right axes whereas the average sensitivity difference is displayed on the left axes. The sensitivity line is calculated by propagating that specific feature input data through the RF model using a constant value, in a stepwise format, from its minimum to maximum range while predicting the model average opacity at each step. The calculated opacity value is then compared to the average opacity from the actual data set and the difference between these averages is plotted as the line. This process is repeated for each feature to generate each

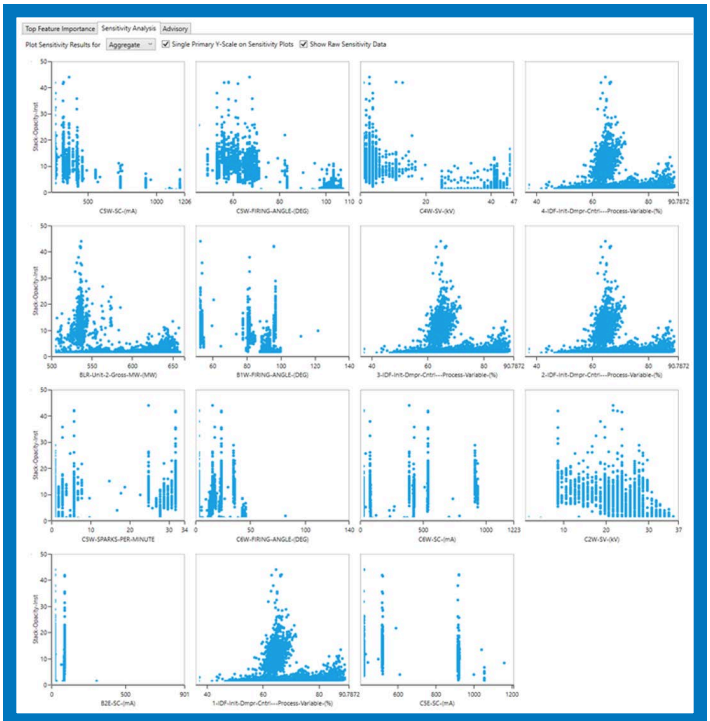


Figure 7. Dataset scatter plots of opacity vs. top 15 features identified by FoREMaL Explorer

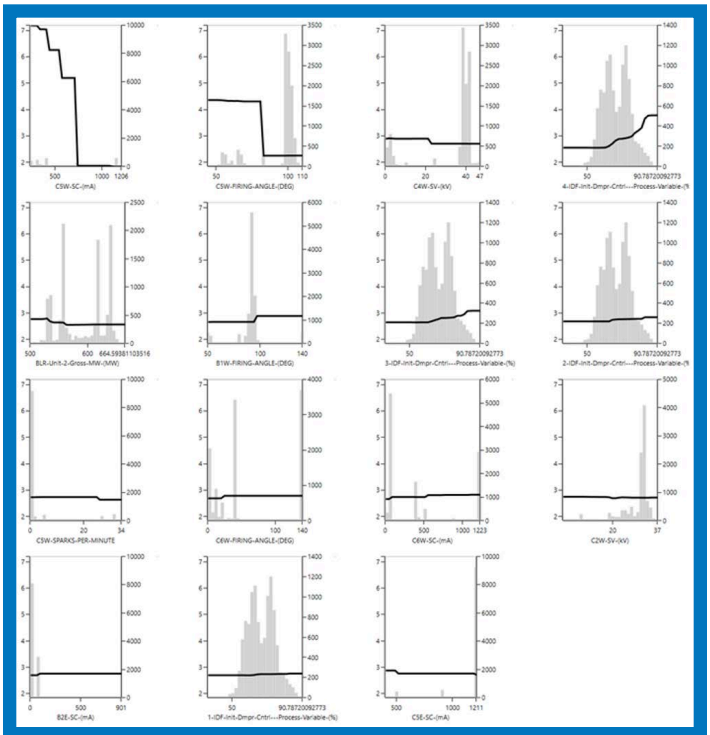
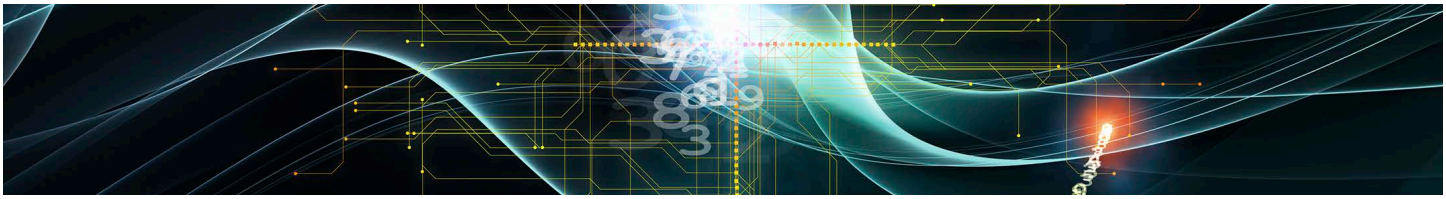


Figure 8. Sensitivity analysis trend lines and data histograms for the respective top 15 features



respective sensitivity line. The sensitivity line provides the prediction of the model based on the analyzed data set and indicates if that feature exhibits a noticeable step change with respect to dependent variable of interest – which in this case is opacity.

Interpretation

As an example, the leftmost plot for feature C5W-SC-(mA) refers to TR-Set secondary current for Chamber C Field 5 West. The calculated sensitivity line trend indicates that at a value of ~740mA, a significant step change decline in average opacity difference was observed and the opacity remains minimized at values greater than 740mA. On the same feature, the histogram bar plots show where the actual data set lies with respect to the sensitivity line. From these two information pieces, the analyst can observe that most of the actual data resides in an area with low average opacity. The second rank feature was also associated with that electrical field. Similar analyses can be conducted to establish the trends. In this case, the information may prompt a subject matter expert to take a closer look at the conditions of this TR-Set, of that C5 field or of Chamber C. As a rule of thumb, for most ESPs in good working condition, the expected current and voltage operating trends conform to lower current densities and higher secondary voltages at the inlet. As the dust laden flue gas enters the downstream collection fields, the dust concentration is gradually reduced, less particulate mass is collected on the downstream collector plates and the current density should increase while the secondary voltage should decrease. Deviation from this trend may indicate some operational, mechanical, or electrical anomaly. Other features identified by the RF analysis were the IDF damper process variables with identification #4, #3, #2, and #1. These were ranked 6th, 9th, 12th, and 13th, respectively. The histogram data distribution trends look similar for the four IDFs, as one may expect, but the sensitivity line trends showed some differences. For instance, IDF #4 vane position shows a greater sensitivity to opacity increase than the other IDF process variables when the vanes are > 57% open. This may prompt an examination of the flue gas path into the respective IDFs and the equalizing cross-over duct that connects all four ESP chambers. Although two features associated with the Chamber B inlet fields were ranked 8th and 15th, the respective trend lines indicated lower sensitivities to opacity than other ranked features.

Development of an Action List

It is evident from these results that the interpretation of the model output, at this stage, requires the input from a data science engineer and from an ESP subject matter expert. The goal of the project is to develop a tool that captures this level of expertise during model development for that specific process and allows deployment of a case-specific solution for quicker interpretation by field personnel such as engineers and operators. At the current time, FoREMaL Explorer provides a strategy to either minimize or maximize the dependent variable. For this example, the objective was to minimize the effect on opacity. FoREMaL prepares an Action List for the top 15 features which, in an ideal case, should be maintained to minimize the opacity. These results are displayed in an advisory screen within ForEMaL Explorer. Table 1 shows the proposed Action List of threshold values for each flagged parameters and Figure 9 displays graphical trends for opacity. The blue line displays the actual 'raw data' opacity trend for the modeling period while the brown trend demonstrates the 'optimized' predicted opacity if all the predictor variables are maintained at the respective Action List values. Note that at this time FoREMaL Explorer does not discern between control variables or adjustable variables. This prompts the need for subject matter expertise with result interpretation.

Table 1. Advisory action list to minimize opacity for case study

Process Parameter	Threshold	Value
C5W-SC-(mA)	>	1206
C5W-FIRING-ANGLE-(DEG)	>	98.6
C4W-SV-(kV)	>	42.1
A2E-SC-(mA)	>	1219
C4W-FIRING-ANGLE-(DEG)	>	105.2
4-IDF-Inlt-Dmpr-Cntrl (%)	<	56.7
A2E-FIRING-ANGLE-(DEG)	<	0
B6W-SV-(kV)	>	25.5
3-IDF-Inlt-Dmpr-Cntrl (%)	<	58.6
BLR-Unit-2-Gross-MW-(MW)	>	579.4
B1W-FIRING-ANGLE-(DEG)	<	90.3
2-IDF-Inlt-Dmpr-Cntrl (%)	<	58.6
1-IDF-Inlt-Dmpr-Cntrl (%)	<	58.6
B1E-FIRING-ANGLE-(DEG)	<	52
C6W-FIRING-ANGLE-(DEG)	<	9.7

The model parameter settings and respective results can be saved and repurposed for analysis of other time periods. During this initial stage, arriving at these results required the input from a data science engineer and a process (ESP) subject matter expert. Upon investiga-

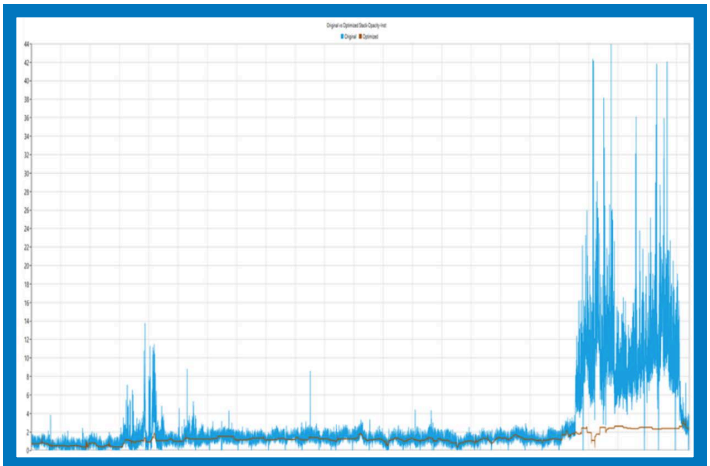


Figure 9. Original opacity trend (blue) and model predicted opacity trend if Action List is maintained (brown)

tion by plant engineers, it was found that ID fan #2 drive shaft was decoupled from the inlet louvers which introduced a flow imbalance to the ESP multi-chamber gas flow system. The model predictions did not explicitly provide this exact information. Instead, the predictions indicated traceability to abnormal vane positions on the ID fans and indication that Chamber C was strongly correlated with the opacity along with Chamber B. The model, when applied, could save the plant operations and engineers troubleshooting time which, in turn, may reduce the risk of unexpected derates on the unit. This is beneficial from the standpoint of increasing the operating margins and reducing the generating unit random outage factor while maintaining unit reliability and minimizing unwanted pollutant emission excursions.

Case II: Correlating Wet FGD Process Variables

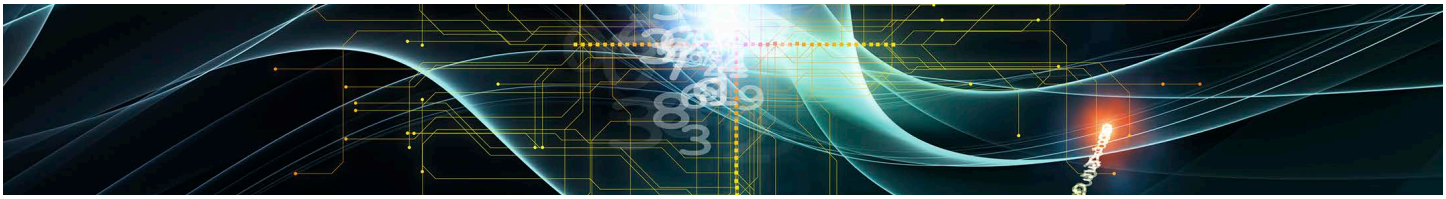
Another proof-of-concept application of the RF analysis program involved the analysis of process variables from a wet flue gas desulfurization unit. The objective of this analysis was to identify which of the process variables correlated with real-time gaseous Hg emissions measured downstream of the FGD system, at the stack. The data derives from a case study that used another nonlinear modeling technique, called Least Squares Support Vector Machines, to develop a process model from its variables. In this instance, the RF software was utilized to identify the process variables that correlate strongly with gaseous mercury (Hg) emissions from a coal-fired unit. Hg found in coal is volatilized during combustion in the form of elemental mercury (Hg^0) which is difficult to remove from the flue gas and is insoluble in water. Hg^0 can be converted to an oxidized form Hg^{2+} in upstream air pollution control devices such as the Selective Catalytic Reduction (SCR) reactor. Hg^{2+} is highly soluble in water can be efficiently removed in a wet FGD system. However, the WFGD chemistry is affected by many process variables which sometimes interfere with efficient Hg^{2+} removal by reversing the speciation to elemental form Hg0. This change often results in mercury re-emission which is not desirable. Therefore, coal-fired units that operate with WFGD must optimize and monitor their processes to reduce the chance of Hg re-emissions. Process variables that have been correlated with mercury in WFGD include absorber liquor pH, and oxidation-reduction potential (ORP).

Approach

Three individual EGUs feed two WFGD absorbers towers for flue gas clean up. The flue gas, which is mostly free from solid particulate fly ash, is conveyed by induced draft fans into the WFGD absorber towers. There the gas is quenched with a limestone slurry via five levels of spray nozzles, each fed by a dedicated slurry pump. The sulfur species in the flue gas react with the limestone to form solid calcium sulfite particles which are later oxidized in the reaction tank to calcium sulfate. The pH of the liquor can be controlled by modulating the input of limestone slurry into the system. The cleaned flue gas exits the top of the tower towards the stack and the atmosphere while the slurry liquor, which now contains captured SO_2 and Hg, is recirculated and processed for phase separation. For instance, the solid by-product, which is mostly gypsum, is periodically removed from the slurry by dewatering. Most of the oxidized mercury is removed with the liquor and sent to a wastewater treatment plant. Monitored process variables in the wet FGD system are flue gas flow, flue gas inlet and outlet temperatures, inlet and outlet SO_2 concentrations, pH and ORP. Control variables are generally, oxidation air flow, limestone slurry flow and recycle pump flows. All the information is generally stored in a historian. The units and wet FGD system analyzed in this case study are detailed elsewhere [Lv 2019]. The RF software was applied to a 30-day period data set that included full-load constant operation, low-load constant operation and transient load periods in between these states.

Results and Post-processing

The dataset contained 20 process variables and 4,400 process data records. Incomplete records were excluded from the analysis as were load points lower than 750 MWeq. The model was setup to identify



and rank the top nine variables correlating to total Hg emissions at the stack. Results of the analysis showed strong correlation between the absorber flue gas inlet temperatures, followed by load equivalent flue gas flow rate, outlet temperature and absorber inlet SO_2. Other variables identified, although not as strongly as the first four variables, were oxidizing air flow, bleed density, recycle pump 2 flow and ORP. Figure 10 summarizes the results along with respective R^2 correlations.

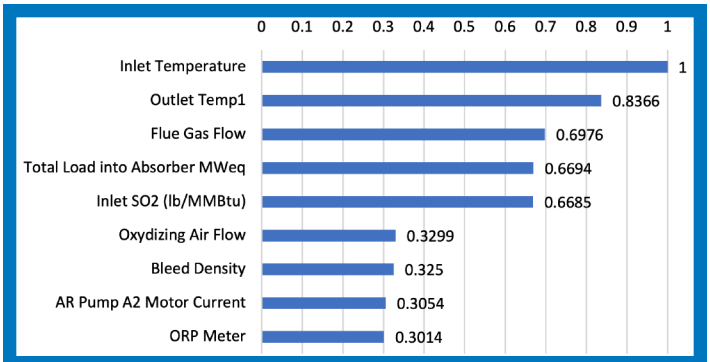


Figure 10. Top nine ranked variables that correlate with Hg emissions for a WFGD system

Figure 11 shows the scatter plots for each respective feature. Mostly positive correlations are observed for the first five variables from upper left to bottom right. Figure 12 shows the equivalent sensitivity plots for each variable along with the corresponding data frequency histograms.

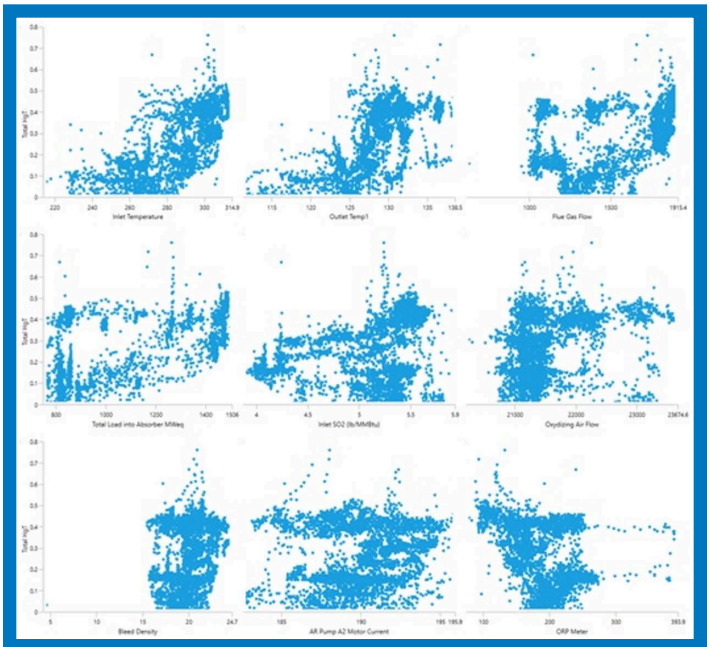


Figure 11. Scatter plots for the nine variables identified by the RF software

Action List Findings

The action list is presented in Table 2 and indicates the threshold values identified to minimize the Hg emissions. Although the program provides advisory information about the process variables, it does not discern whether a system operator can adjust or maintain the minimization threshold values. That decision will require the input of a process subject matter expert or the careful selection control variables for the system during initial case study setup. The predicted optimized trend, if all the values are maintained over time, is shown in Figure 13. The blue line represents the original data trends whereas the brown line is the minimized mercury emission trend if the action list can be maintained. One feature of the RF software is the ability to discern the sensitivity each action list item by varying the optimized value and observing its impact on the predicted trend.

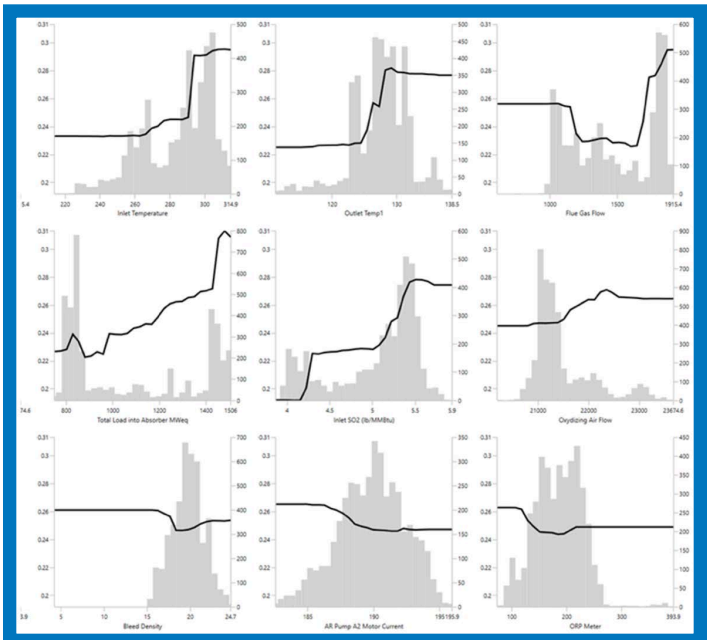


Figure 12. Corresponding sensitivity plots for the nine variables identified by the RF software

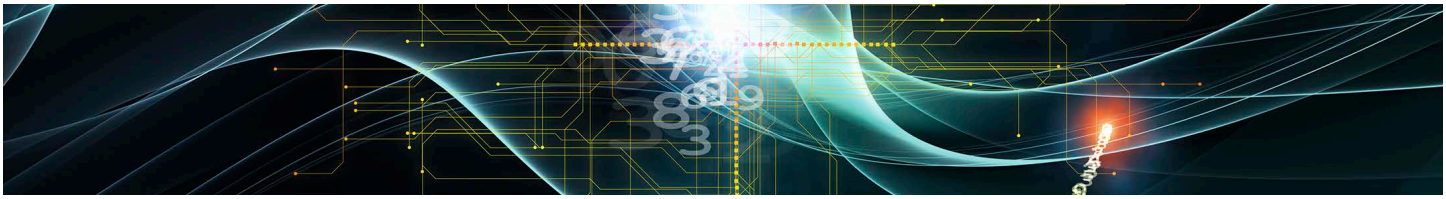


Table 2. Action items list to minimize Hg emissions for the wet FGD case study

Process Parameter	Threshold	Value
Inlet Temperature	<	241.7
Outlet Temp	<	111.3
Flue Gas Flow	<	1600
Total Load into Absorber MWeq	<	880
Inlet SO_2 (lb/MMBtu)	<	4.08
Oxidizing Air Flow	<	20,995
Bleed Density	>	19.05
Ab Pump 2	>	191.8
ORP Meter	>	184.1

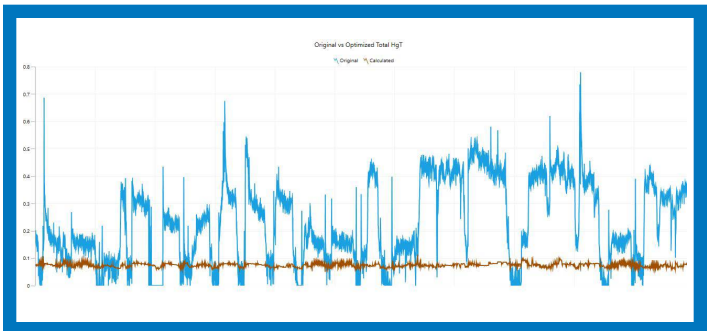


Figure 13. Actual (blue) and minimized (brown) mercury emission trend based on recommendations from the random forest and sensitivity analysis software

Conclusions

An application that incorporates a random forest and forest of forest approach for data curation, filtering, and data analysis has been developed and packaged into a prototype Windows desktop application called FoREMaL Explorer. The program is able to identify, rank and minimize or maximize a user-specified dependent variable. Two applications of this analytical tool were presented as examples. The first case study involved the analysis of opacity upset events and process variables at one coal-fired power plant equipped with an ESP and no other downstream environmental control. The second case study was applied to process data, at another power plant designed with an ESP and a multi-unit fed wet FGD system, to identify process variables that correlated with gaseous mercury emissions. In both case studies, ForEMaL Explorer provided a ranking, identified the level of correlation for each ranked variable, conducted a sensitivity analysis, and proposed an action list of threshold values to minimize the dependent variable. This list could also be utilized by an analyst to maximize the target variable, if so desired. The case studies are presented as examples of applications for this newly developed random forest tool for regression type analysis. The application is well suited for large data sets that involve several features or predictors. The ongoing development is focused on enhancing the tool to create simple independent robust models that do not require any additional data input. In addition, future efforts may also concentrate on incorporating a predictive capability for use as open-loop advisory tool. These tools may help operators to understand the interactions between complex systems such as those experienced by many modern power generation systems.

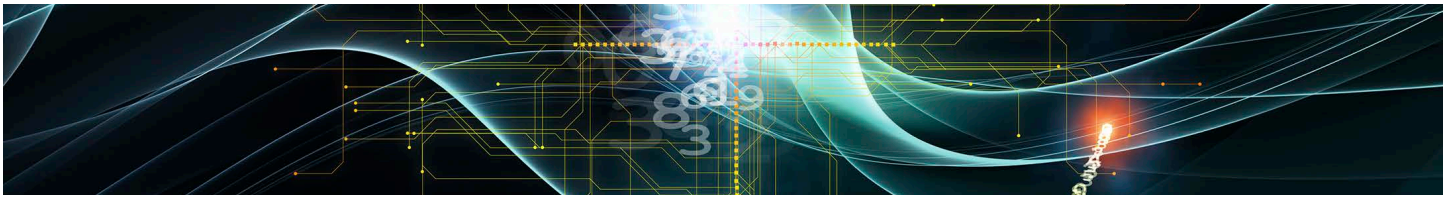
References

1. Application of Novel Random Forest Approach for Environmental Controls: Development Status. EPRI, Palo Alto, CA: 2021. 3002021062

2. [Song 2017] Song, Jingge et al. "A Globally Enhanced General Regression Neural Network for on-Line Multiple Emissions Prediction of Utility Boiler." Knowledge-Based Systems 118 (2017): 4–14. Web.Information Collection Request (ICR) Data Analysis to Meet Mercury and Air Toxics Standards (MATS) Requirements: Particulate Matter, Mercury, and HCl Investigation. EPRI, Palo Alto, CA: 2012.

3. [Nie 2017] Nie, Fengxiang, and Oztekin, Alparslan. "A Study of the Prediction of Ammonium Bisulfate Formation Temperature by Artificial Intelligence." ProQuest Dissertations Publishing, 2017. Web.Maninder, T., Tessum, C., Azevedo, I., Marshall, J., Fine Particulate Air Pollution from Electricity Generation in the U.S.: Health Impacts by Race, Income and Geography. Environmental Science and Technology, 2019, 53, 1410 – 1419

4. [Xu 2013] Wang Xu, and Chang Taihua. "The Balanced Model and Optimization of NOx Emission and Boiler Efficiency at a Coal-Fired Utility Boiler." IEEE Conference Anthology. IEEE, 2013. 1–4. Web.2017 National Emissions Inventory (NEI) Data; US Environmental Protection Agency, Washington DC, 2020. https://www.epa.gov/air-emissions-inventories/2017-national-emissions-inventory-nei-data



5. [Bazzi 2018] Bazzi, T., & Zohdy, M. (2018, December). Artificial Intelligence for Air Quality Control Systems: A Holistic Approach. In 2018 Twentieth International Middle East Power Systems Conference (MEPCON) (pp. 25-32). IEEE.

6. [Stobl 2007] C. Strobl, A.-L. Buolesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8(1):1{21, 2007.

7. [Estes 2016] Estes, M., Artificial Intelligence for Precipitator Diagnostics, Presented at The A&WMA Mega Symposium, Baltimore, MD. August 2016.

8. [EPRI 2020] FoREMaL Explorer User Manual, beta-version 1.0.1 EPRI, Palo Alto, CA: December 2020 (unpublished).

9. [Microsoft 2021] Microsoft.ML Trainers. Accessed 9/2021, Microsoft.ML.Trainers.FastTree Namespace | Microsoft

10. Lv, Y., Romero, C., Pauvlich, K., Charles, J., Watkins, R., Developing steady and dynamic ORP models for mercury emissions control in power plants using WFGD operating data, Fuel 235 (2019) 54–62

EPRI RESOURCES

Jose Sanchez, *Principal Technical Leader, Environmental Controls*
650.855.2143, josanche@epri.com

Technology Innovation
Artificial Intelligence Initiative

Program 210 Combustion, Emissions, and Carbon Control for All Fuels

About EPRI

Founded in 1972, EPRI is the world's preeminent independent, non-profit energy research and development organization, with offices around the world. EPRI's trusted experts collaborate with more than 450 companies in 45 countries, driving innovation to ensure the public has clean, safe, reliable, affordable, and equitable access to electricity across the globe. Together, we are shaping the future of energy.

3002023918 December 2022

EPRI
3420 Hillview Avenue, Palo Alto, California 94304-1338 USA • 800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com