npj | materials degradation

Article

Published in partnership with CSCP and USTB

https://doi.org/10.1038/s41529-024-00545-8

Assessing the feasibility of using a data-driven corrosion rate model for optimizing dosages of corrosion inhibitors

Chamanthi Denisha Jayaweera[1,2,3] ✉, David Fernandes del Pozo[1,2], Ivaylo Plamenov Hitsov[1,2,3], Maxime Van Haeverbeke[4], Thomas Diekow[5], Arne Verliefde[1,3] & Ingmar Nopens[1,2]

Optimizing dosages of corrosion inhibitors requires experimental data gathered from time-consuming methods. The current study examines the feasibility of optimizing inhibitor dosages using a model trained for predicting corrosion rates more easily measured using linear polarization resistance in a full-scale cooling water system. A comprehensive study on variable selection showed that linearly correlated variables are necessary to predict corrosion trends. The Sobol sensitivity of inhibitors is trivialized by variables linearly correlated to the corrosion rate. The study highlights the importance of achieving high model prediction accuracy and high Sobol sensitivity of inhibitors to the corrosion rate, for using the model for inhibitor dosage optimization.

Corrosion can result in damage to equipment and long downtime in cooling water systems. Corrosion mitigation strategies commonly adopted are maintaining a small scaling deposit on the metal surface, usage of corrosion inhibitors, and treatment of the cooling water to remove corrosive constituents such as chlorides and sulphates[1–8]. Corrosion inhibitors slow the rates of both cathodic and anodic reactions by reducing the active surface or changing the activation energy of the oxidation or reduction process[9]. The development of corrosion inhibitors and investigation of their inhibition mechanisms have been carried out with the aid of molecular modeling and quantum chemical calculations using density functional theory[10]. The latter is based on electron density, which carries information related to atoms and molecules. Investigating micro-mechanisms requires combining first-principles techniques based on fundamental theory, such as the density functional theory with molecular dynamics, peridynamic theory and finite-element methods[10]. Although these methods provide comprehensive information about the system under consideration, they are highly computationally intensive. However, their prevailing interest in the field of corrosion is apparent in obtaining insights into micro-mechanisms and interactions among components in a water matrix and the metal surface.

The addition of more than one corrosion inhibitor in cooling systems is commonly applied with the hope of synergistic effects among inhibitor compounds improving the overall inhibition efficiency. Studies have demonstrated the synergism of Zn^{2+} ions with organic corrosion inhibitors[11]. Synergism between corrosion inhibitors has been investigated using electrochemical impedance spectroscopy by Marin-Cruz et al.[8] and Touir et al.[5] in cooling water systems. However, antagonistic effects among inhibitors have also been shown when water qualities change[12]. Thus, the inhibitive effectiveness of corrosion inhibitors depends on complex interactions between background ions and other inhibitors[2–5,11,12], as well as the metal under test[6]. These effects are very difficult to predict a priori. Therefore, the determination of corrosion rates, as well as inhibition efficiency in real-life aqueous environments, is commonly done via pilot tests and time-consuming and costly experiments (e.g. using electrochemical impedance spectroscopy)[7,8].

Model development is an alternative method for capturing the synergism among multiple corrosion inhibitors and ions present in the system. Research reported on modeling corrosion inhibition using data-driven models has focused on predicting an aspect related to inhibitors, such as the inhibition efficiency. Corrosion inhibition of mild steel in sulfuric acid has been modeled by Edoziuno et al.[13]. They analyzed corrosion inhibition-related process parameters and their relationships to obtain optimal inhibitor concentration, immersion time, and acid concentration that maximized inhibitor efficiency. Omran et al.[14] conducted a factorial experimental design to maximize the inhibition efficiency in a system of mild steel and water-containing plant extracts. Ansari et al.[15] optimized the interactive effects of temperature, the concentration of inhibitor and

[1]CAPTURE - Centre for advanced process technology for urban resource recovery, Frieda Saeysstraat 1, 9052 Gent, Belgium. [2]BIOMATH - Model-based analysis and optimisation of bioprocesses, Department of data analysis and mathematical modeling, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000 Gent, Belgium. [3]PAINT – Particle and interfacial technology, Department of green chemistry and technology, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000 Gent, Belgium. [4]KERMIT – Knowledge-based systems, Department of data analysis and mathematical modeling, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000 Gent, Belgium. [5]Dow Olefinverbund GmbH, Bohlen, Germany.
✉e-mail: chamanthidenisha.jayaweera@ugent.be; chamanthidj@gmail.com

immersion time for a maximum response of inhibition efficiency using the Response Surface Method (RSM) in a C38 steel/H_2SO_4 solution. Commercial software such as French-Creek models[16] can optimize multiple inhibitors, but only sequentially. For instance, Ferguson et al.[16] demonstrate how the orthophosphate dosage is first optimized, and subsequently, the copolymer dosage is optimized. These studies have specifically measured the corrosion inhibition efficiency and modeled it with respect to concentrations of corrosion inhibitors and ions. Corrosion inhibition efficiency can be measured using complex and time-consuming techniques such as electrochemical impedance spectroscopy and potentiodynamic polarization techniques.

Corrosion rates of cooling water are conveniently measured in large-scale plants using linear polarization resistance (LPR) in an hourly or daily basis. In the meantime, concentrations of several ions are also regularly measured and recorded. It would be advantageous to use these measurements to replace, or at least minimize the use of costly and time-consuming methods such as potentiodynamic polarization techniques for optimizing dosages of corrosion inhibitors. Due to the complexity in the corrosion inhibition process, data-driven models such as neural networks are most appropriate. These models aim to counter the complexity and the time-intensive character of pilot-scale, lab-scale or molecular modelling experiments.

Authors of the current study identify the following essential properties when using a model developed for predicting the corrosion rate for optimizing inhibitor dosages:

- The prediction accuracy of the model should be satisfactory.
- The model should identify the relationship between the corrosion rate and corrosion inhibitors sufficiently.

Data-driven modeling studies carried out in the literature have demonstrated the corrosion rate can be predicted with adequate prediction accuracy. Aghaaminiha et al.[17] developed a random forest model to predict the corrosion rate of mild steel in CO_2 aqueous solutions with the mean squared error ranging from 0.005 to 0.093 as a function of time over 160 hours. Coelho et al.[18] comprehensively compares the prediction accuracies of different types of data-driven models. Machine learning models that have been reported in corrosion literature are kernel-based methods such as support vector regression, back propagation neural network, deep neural network, tree-based models such as random forest, and gradient-boosted decision trees. The study by Coelho et al.[18] reveals that kernel-based methods such as support vector regression have higher prediction accuracy across different corrosion topics. Although these studies have reported high testing performances, the methods employed in dividing the data set into training and test sets for a proper model evaluation were not clearly described. It is important to demonstrate that the model is capable of predicting events ahead of time. Such events can be defined as changes in corrosion rate in response to varying operational and environmental factors. A study by Zhi et al.[19] demonstrates the difficulty of predicting corrosion rates in diverse conditions.

The relationship between inhibitors and corrosion rate is not as direct as that between inhibitors and inhibition efficiency. Other factors, such as pH have a higher impact on the corrosion rate than inhibitor dosages. During the training process of a data driven model such as a neural network, the model assigns weights to connections between the input variables and the output. More often than not, the model assigns high weights to a handful of variables while the significance of the remainder is diminished[20]. Therefore, it is likely that the weights assigned to corrosion inhibitors is dependent on other variables used as inputs to the model.

The taxonomy of variable selection for model development has been discussed in several studies[20–22]. Selection methods are most commonly classified into linear and nonlinear filter, wrapper, and embedded methods. Studies have been carried out to identify suitable variable selection methodologies per field of study[23,24]. A comparison of how commonly used variable selection methods affect the corrosion rate prediction accuracy and the significance assigned to corrosion inhibitors has not been carried out to the best of the authors' knowledge.

Therefore, the current study investigates the feasibility of reliably developing and using a data-driven model trained to predict the corrosion rate for optimizing multiple corrosion inhibitors simultaneously. In view of the stated gaps noted in the literature and challenges experienced when developing a predictive model for corrosion using cooling water data sets, the current study investigates the impact of common variable selection methods on the prediction accuracy of corrosion rate and sensitivity of corrosion inhibitors to the predicted corrosion rate.

Methodology

Figure 1 presents an overview of the methodology followed for demonstrating how combinations of input variables affect the prediction accuracy of corrosion rate as well as the significance assigned by the model to corrosion inhibitors. The software used in this study are mentioned in Supplementary Table 1.

Data set

A data set containing daily measurements of water qualities, chemical concentrations, and corrosion rates (mm/year) of mild steel pertaining to the years 2020 to 2023 was obtained from a cooling water circuit of a chemical plant. The corrosion rates were measured using a mild steel LPR probe (linear polarization resistance) embedded in an epoxy resin (to minimize crevice corrosion) inserted into the piping of a cooling water circuit. The sensor is mounted horizontally in the side branch of a tee, with the flow entering the tee through the top branch and flowing away from the base of the sensor, towards the tips of the electrodes.

Hourly corrosion rates were recorded. As only daily concentrations of ions and inhibitors were available, the corrosion rates were averaged to daily values. Variables available in the data set are shown in Table 1. All variables in Table 1 were not used for model development. Value ranges of each variable shown in Table 1 are given in Supplementary Table 2.

The mean, standard deviation, median, coefficient of variation, skewness and kurtosis of each variable shown in Table 1 is given in Supplementary Table 3. Physical properties and concentrations in the data set had been maintained between practical limits throughout 3 years. Among the corrosion inhibitors analysed in this study, benzotriazole has the highest coefficient of variation (0.38). A kurtosis of 3.26 indicates that the standard deviation is due to frequent modestly sized deviations. Lowest coefficients of variations among the inhibitors can be observed in inhibitor A and orthophosphate (0.1 and 0.12, respectively). However, their coefficient of variation is similar to that of CWFR (0.11). The higher kurtoses of orthophosphate (5.7) and zinc (7.04) indicate that most of their values are closer to the mean. The coefficient of variation of zinc (0.24) is higher than inhibitor A, orthophosphate and CWFR. The correlation coefficients and the p-values for testing the hypothesis that there is no relationship between the variables and the corrosion rate (null hypothesis) are mentioned in Supplementary Table 4. The relationship between the corrosion rate and the variables are nonlinear as the corrosion process is a complex phenomenon. Therefore, the values of the correlation coefficients are low. However, as evident in Supplementary Table 4, the low p values of the variables indicate that the null hypothesis (that no relationship exists between the corrosion rate and the variables) can be rejected. Therefore, the data is suitable for model development.

All variables shown in Table 1 affect the corrosion rate. Orthophosphate is a widely known corrosion inhibitor[25]. Benzotriazole is commonly used to mitigate corrosion of copper[26]. However, it has also been proven effective for mild steel[27]. Inhibitor A is a proprietary chemical tailored to minimize corrosion. Zinc acts as a cathodic corrosion inhibitor by forming complexes with hydroxide ions and precipitating on metal surfaces[28]. Phosphates and phosphorous based compounds provide anodic as well as cathodic protection to metals[29,30]. However, phosphorous-based compounds and TOC could contribute to microbial corrosion[31]. Microbial corrosion is an unavoidable phenomenon in cooling systems[32]. Therefore, the bacterial count is as an important factor affecting corrosion rate. Ca^{2+} and Mg^{2+} contribute to scaling deposits that often act as a protective

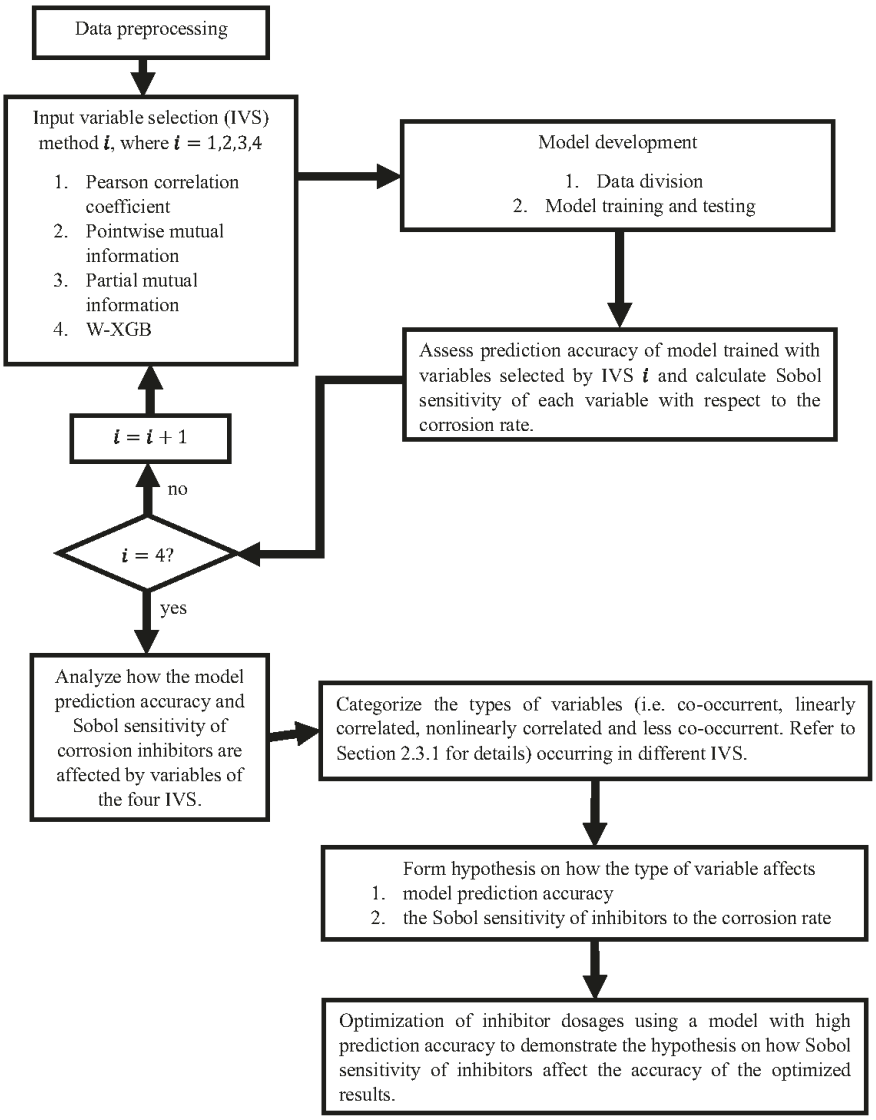

Fig. 1 | Research methodology. An overview of how input variable selection methods were used to assess the feasibility of optimizing inhibitor dosages.

Table 1 | Input variables available in the data set

Variables	Variables	Variables	Variables
pH	NO^-_2	AOX	Inhibitor A
KS8,2 and KS4,3	Ca/Mg	Total Fe	Benzotriazole
Temperature (°C)	Cl	Dissolved Fe	Zinc
Turbidity	ClO_2^-	TOC	Orthophosphate
COD	Ca	Silicates	Cooling water flow rate (CWFR) – kg/h
Total inorganic Nitrogen (TIN)	NO_3^-	Cu	Filtered substances
Conductivity	Mn	Total chlorine	Corrosion rate
Total Carbon	Na	Dissolved zinc	Thickening agent
Total inorganic Carbon (TIC)	NH_4^+	Organic phosphorous	HCO_3^-
Free chlorine	SO_4^{3-}	Bacterial count	Total phosphate

Concentrations are given in mg/L.

covering preventing corrosion of metal surfaces[33]. Therefore, antiscalants that are added to regulate scaling deposits affect corrosion. Minute amounts of copper dispensed from copper-containing parts of the cooling system could plate on steel surfaces and induce rapid galvanic effect on them as the distance between steel and copper is high in the galvanic series[34]. Effect of Nitrate ions on corrosion varies depending on water chemistry, type of metal and physical parameters such as temperature. Nitrates could aggravate corrosion through adsorption and reduction to ammonium[35,36] while also reducing the chances of corrosion by passivation[37]. Halogenated organic compounds (AOX) are harmful to the environment. The addition of benzotriazole contributes to the AOX concentration[26]. Variables such as free chlorine, chloride ions, pH, conductivity and temperature are widely known to affect corrosion.

In addition to variables shown in Table 1, historical values of benzotriazole, inhibitor A, orthophosphate, zinc, and flow rate were also considered. Historical values refer to values of a variable from a previous time period. For example, zinc(t-1) refers to the zinc concentration from a day prior to the considered day. Historical values of the corrosion rate were not considered in the first part of this study, as the significantly higher sensitivity of historical corrosion rates makes it difficult to assess the diminished impact of other variables on the predicted corrosion rate. However, they are considered when the final model is presented with an example-result on how inhibitors are optimized at the end of this study. Historical values for inhibitors (i.e. benzotriazole, inhibitor A, orthophosphate and zinc) are considered up to three days (i.e. t-1, t-2, t-3). Those of CWFR was considered up to 4 days as the corrosion rate was visually highly correlated to CWFR, as shown in Fig. 2. It is assumed that high flow rates contribute to

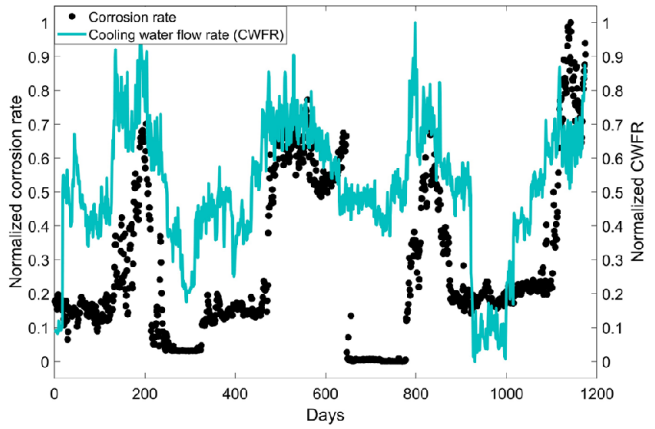

Fig. 2 | Corrosion rate versus CWFR as a function of time. Filled black circles represent measured corrosion rate over time. The light blue-green continuous line represents the corresponding cooling water flow rate maintained in the circuit.

greater shearing of the corrosion fouling layer[37]. As the development of the corrosion fouling layer is time-dependent, longer historical CWFR could be relevant to the prediction of the corrosion rate.

Data preprocessing

All variables were normalized between 0 and 1 using Eq. 1 as

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

Outlier removal was carried out using a method known as the 'local outlier factor' (LOF). This method is based on computing the local density deviation of a given data point with respect to its neighbors and was implemented using the sci-kit learn python library[38]. The LOF method was found to be superior to Z-score method. The Z-score method is highly reliant on the mean of the data. As pointed out by May et al.[20], the mean will be affected if a large number of outliers is present in the data set.

Variable selection

Five variable selection methodologies were applied in this study.

1. Pearson correlation coefficient (PCC)
   PCC is determined by

$$\mathrm{PCC} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{2}$$

   where, x_i and y_i are i^{th} input vector and output value in the sample; $\bar{x}$ and $\bar{y}$ mean values of x and y.
   PCC is a popular linear method of examining the correlation between two variables. The magnitude of the correlation suggests the degree of the linear correlation existing between two variables, while the sign of the value obtained suggests whether the variable is enhancing or inhibiting to its counterpart.

2. Point-wise mutual information (point-wise MI)
   Point-wise MI is determined by

$$\mathrm{Point-wiseMI}(x, y) = \log \frac{P(x, y)}{P(x)P(y)}, \tag{3}$$

   where $P(x, y)$ is the joint probability distribution of x (an input variable) and y (the output variable). Reader is referred to May et al.[39] for further details on point-wise mutual information. A Gaussian kernel was used to estimate the probability distributions. MI is a nonlinear filter used to quantify the correlation between two variables. Variables with high positive pointwise MI values are ranked high.

3. Partial mutual information (Partial MI)
   Partial MI uses the same concept as mutual information; however, it enables the reduction in uncertainty in predicting the output while quantifying the additional mutual observation gained by adding another variable (Z) into a set of already established variables (X). The algorithm employed for implementing partial MI is shown in Supplementary Figure 1. Further details of the partial MI algorithm have been presented by May et al.[39]. The last step of the algorithm involves selecting a variable that maximizes the pointwise mutual information (I(v; u)). It was noted that the values of I(v; u) were negative. Therefore, selecting the highest value of I(v; u) could be done as per the magnitude of I(v; u) or based on the more positive value of I(v; u). Thus, partial MI was implemented in two ways[1]: positive partial MI - selection of variables giving priority to more positive pointwise mutual information (I(v; u)) and[2] magnitude-based partial MI - selection of variables giving priority to the magnitude of pointwise mutual information (I(v; u)).

4. Model-embedded weights-based variable selection using the XGBoost implementation of gradient-boosted decision trees
   XGBoost regression has been successfully used in a multitude of applications including corrosion for prediction and computing feature importance[40]. Feature importances were extracted from an XGBoost model trained with data. The *xgboost* python package was used for this purpose. Readers are referred to Pedregosa et al.[38] for more details of the method.

The methods discussed from 1 to 4 provide a ranking of variables in the order of importance. The selection of the optimal number of variables for model development was done as displayed in Supplementary Figure 2.

Linearly correlated, co-occurrent, and less co-occurrent variables. Three terms are used to categorize variables encountered in this study: linearly correlated, co-occurrent, and less co-occurrent variables. The term 'linearly correlated' refers to variables that are considered most relevant by PCC. Co-occurrent variables are considered most relevant by positive point-wise MI. Less co-occurrent variables are high in the magnitude of point-wise mutual information and top-ranked by magnitude-based partial MI, yet not by positive partial MI.

Sensitivity analysis of variables. Sobol sensitivity analysis of input variables was carried out to evaluate the impact corrosion inhibitors had on the predicted corrosion rate. The main purpose of the analysis is to identify how different types of variables affect the Sobol sensitivity of corrosion inhibitors. Most modeling studies that had been carried out on optimizing corrosion inhibitors had focused on predicting quantities such as corrosion inhibition efficiency[13–15], which is a more direct consequence of the inhibitors than the corrosion rate. Therefore, it is important to get a clear understanding of the sensitivity of corrosion inhibitors in a model trained to predict the corrosion rate.

Data division

Definition of the test data set to evaluate the model's ability to predict future events. According to a review by Bowden et al.[41], optimal data division for applications in water resources requires ensuring similar statistical properties among training, validation, and test data sets. While paying due consideration to statistical properties warrants adequate training, defining a test data set based on statistical properties does not guarantee that corrosion events will be well predicted. Therefore, cross-validation was carried out by dividing the time-series data, as shown in Fig. 3.

The corrosion profile was divided into divisions of approximately 300 consecutive days where each division defines an event. The model was

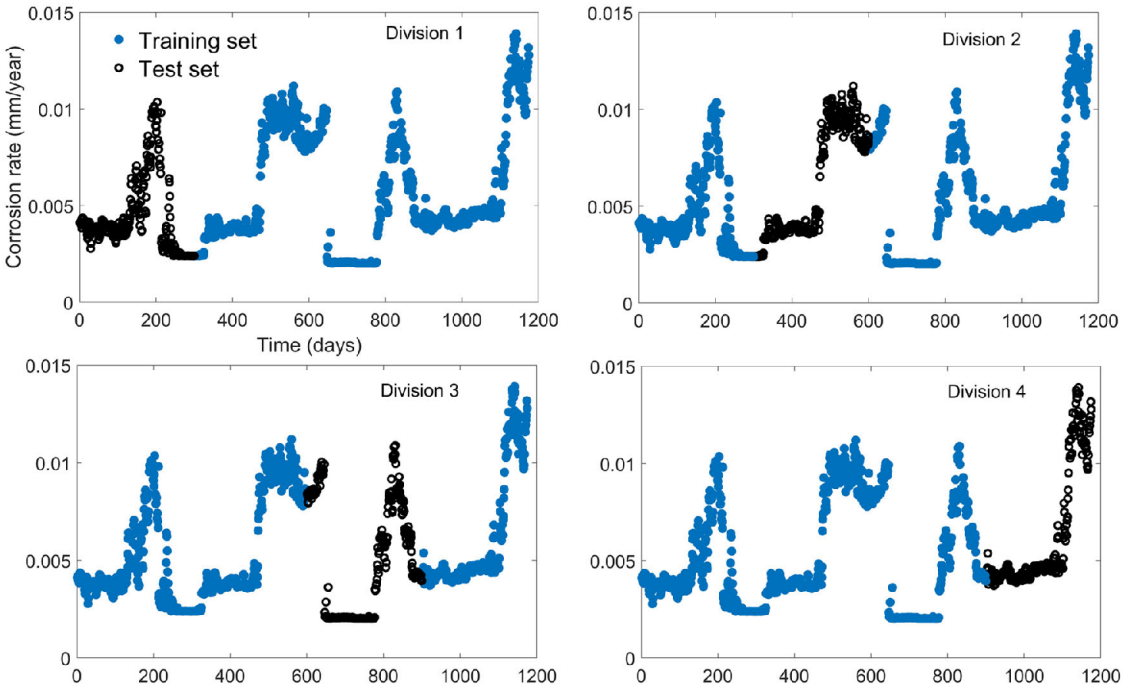

Fig. 3 | Division of the corrosion profile for cross-validation. Filled blue circles represent data points allocated to the training data set. Black clear circles represent data points allocated to the test data set.

trained with nearly 900 daily measurements, shown by blue markers in Fig. 3, and tested with the remaining 300 consecutive daily measurements represented by the black continuous line in Fig. 3. For example, when the first 300 data points were defined as the test data set, the remaining 900 were used for training. Thus, a total of four models were trained, one per each division using data points represented by filled blue circles. Each model was tested with data points indicated by black clear circles.

The corrosion profile shown in Fig. 3 pertains to 4 years of data. Periodical increases in corrosion rate noted in Fig. 3 correspond to summer months when the cooling water flow rate is increased to meet the required cooling capacity. As the temperature of cooling water, which is generally drawn from a nearby water body, is high during summer, CWFR is increased to enhance the circuit's heat capacity. As previously mentioned, increasing the flow rate may result in the shearing of the barrier layer between the metal surface and the bulk fluid. As the barrier layer is formed by inhibitors and corrosion fouling deposits, the removal of the barrier layer enables easier access for corrosive ions towards the metal surface.

The events denoted by the four divisions of Fig. 3 are unique. The test data set of Division 1 demonstrates corrosion rates well within the total range of the data set without any special occurrences. Division 2 represents corrosion rates that are influenced by the occurrence of filtered substances, likely due to an operational event that was only noted in this time period. Division 3 consists of lowest corrosion rates available in the data set. Division 4 consists of highest corrosion rates available in the data set. Therefore, the model's generalization ability, extrapolation ability and its capability of handling responses to operational changes are tested through the division shown in Fig. 3.

Data division in input variable selection. Input variable selection was carried out using the entire data set. Training and test data sets used for input variable selection were defined based on statistical properties, i.e. the mean and the standard deviation of the training data set were made to be approximately equal to those of the test data set[41]. Such an equitable distribution was ensured by first condensing all variables to one variable using principal component analysis and it was used to generate a normal distribution curve as shown in Fig. 4. The distribution curve was partitioned into sections of constant width (e.g. $w = 0.1$) where two consecutive partitions were allocated to the training data set and the third consecutive partition was allocated to the test data set. The process was repeated from the negative-most Z-score to the positive-most Z-score.

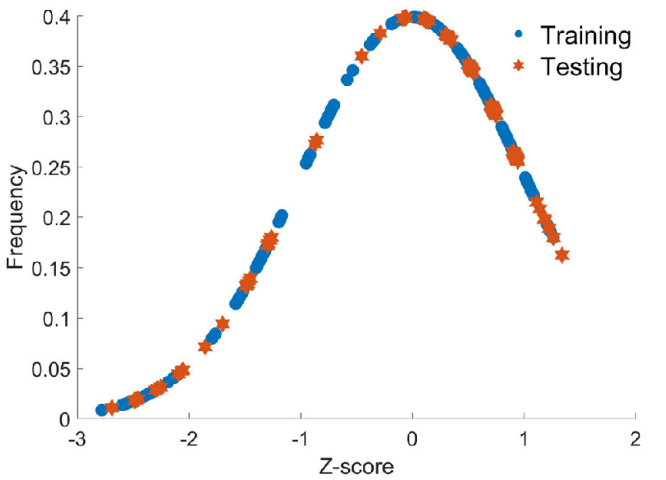

Fig. 4 | Data division according to statistical properties. Blue circles represent data points allocated to the training data set. Orange stars represent data points allocated to the test data set.

The Z-score was computed as

$$\mathrm{Z-score} = \frac{x - \mu}{\sigma} \quad (4)$$

where μ is the mean and σ the standard deviation of the data set.

The width w is varied from 0.1 to 0.4 at intervals of 0.1, and the algorithm in Fig. 4 is repeated each time. The width w is changed to make sure that the performance of the selected variables remains high through varied apportioning of data points in the test data set in an informed manner.

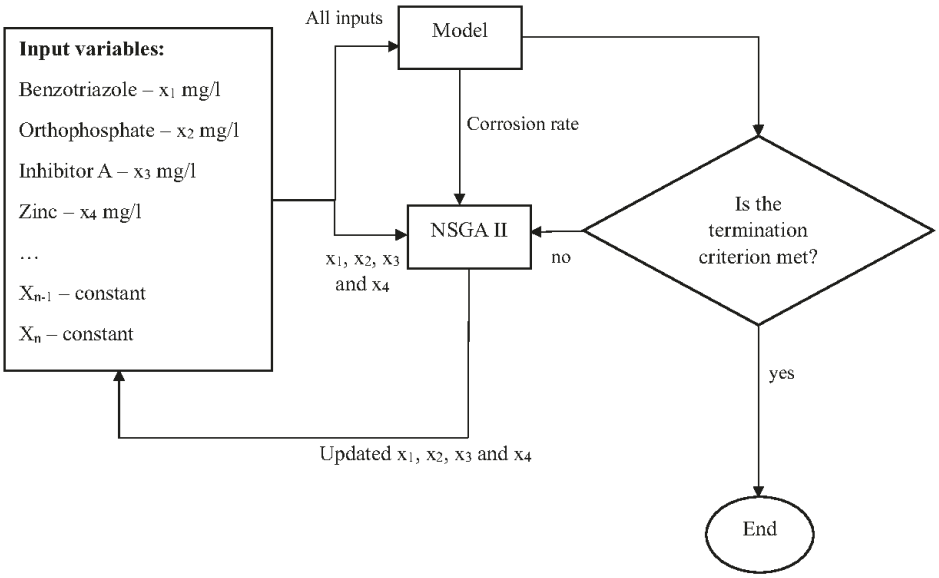

Fig. 5 | Method for optimizing inhibitors using a data-driven model and NSGA II. Concentrations of benzotriazole, orthophosphate, inhibitor A and zinc are varied such that the required dosages of inhibitors are minimized, while maintaining the corrosion rate at a minimum possible value. The termination criterion used in this study is the number of iterations of the algorithm.

Support vector regression

As per a comprehensive review on the use of machine learning in corrosion by Coelho et al.[18], support vector regression has demonstrated high generalization ability and prediction accuracy across several corrosion topics.

The mathematical model of SVR is given by Eq. 5.

$$y = \sum_{i=1}^{k} \beta_i K(x_i, x) + b$$

$$\beta = \alpha^* - \alpha \tag{5}$$

where 'k' is the number of support vectors; α, α^* – Lagrange multipliers; x_i – input vector of support vector; b – bias; x – input vector; y – output, $K(x_i, x)$ is the kernel, for which the radial basis function was used.

More information on the support vector regression model can be read in Welling[42].

Parameters that need to be tuned in a SVR are the box constraint (C, which was set to 1), error sensitivity parameter (ε, set to 0.005) and the smoothing parameter of the radial basis function (γ, set to 0.7). Values of the parameters were determined using a global exhaustive search method.

Optimization

The purpose of presenting the results of optimization in this paper is to demonstrate the implications of using a data-driven model that assigns low weight and sensitivity to optimization. Other aspects of optimization are not within the scope of this study. In order to ensure high prediction accuracy, historical values of the corrosion rate up to 7 days (i.e. CR(t-7)) was used as an input variable. Therefore, in practice, the model developed can only be used for optimizing dosages of the following 7 days. The reason for choosing a week's time for the historical values of corrosion is to ensure that all allowances are made for the response time required for a change in the inhibitor dosage to take effect. The remaining variables required for developing the model used in this section were selected using the most effective input variable selection method out of Section Variable selection. The inhibitors (i.e. benzotriazole, orthophosphate, zinc and inhibitor A) were also included as input variables.

The NSGA II evolutionary algorithm[43] was used for optimization of inhibitor dosages. NSGA II facilitates multi-objective optimization with faster convergence and evaluation of solutions over a larger search space than methods such as gradient descent or particle swarm optimization. The method of optimizing inhibitor dosages using NSGA II is demonstrated in Fig. 5. It should be noted that the inhibitors were varied to determine the optimum values per data point while the remaining input variables of the SVR were fixed as constants (at recorded values), per data point.

The objective functions were defined as stated in Eq. 6

$$f_1 = 0.5 \times [\text{benzotriazole}] + 0.167 \times [\text{orthophosphate}] + 0.167 \times [\text{zinc}] + 0.167 \times [\text{inhibitorA}]$$

$$f_2 = \text{corrosion rate} \tag{6}$$

Equation 6 represents two objective functions. The first objective function f_1 demonstrates how the concentrations of the inhibitors are weighted. Benzotriazole is weighted higher than the remaining inhibitors. Therefore, the NSGA II algorithm gives higher priority to reducing the concentration of benzotriazole over others, due to its harmful environmental impact. In the meantime, the algorithm also minimizes the corrosion rate that is resultant from minimized concentrations of inhibitors. The constraints imposed on this algorithm are that the concentrations of the inhibitors are always maintained between the minimum and maximum values in the available data set.

Results and discussion

Variable selection

The current section presents results of variables selected for model development by the four input variable selection methods discussed in Section Variable selection. The variables are listed in the descending order of importance as specified by each method. The number of variables per method was selected according to the algorithm shown in Supplementary Figure 2. The results obtained are presented in Table 2.

A similarity can be noted among the variables chosen by PCC, positive point-wise MI, and positive partial MI. W-XGB ranks variables based on the overall weights assigned by a trained XGBoost implementation of gradient boosted decision trees to each variable. Therefore, the method captures the phenomenological relationship between the corrosion rate and the input variables than PCC, pointwise MI or partial MI. Therefore, W-XGB identifies the importance of nonlinear and less co-occurrent variables such as Nitrate. Partial MI enables identifying variables that further reduce the uncertainty surrounding the corrosion rate that is gained by the additional mutual observation of a variable[39]. Therefore, variables selected by partial MI complement the information already embedded in the first set of variables. The first set of variables were set as corrosion inhibitors in the system.

Table 2 | Results of five variable selection methodologies

Rank	Model-embedded weights-based variable selection using XGBoost (W-XGB)	PCC	Positive point-wise MI	Positive partial MI
1	CWFR	CWFR	Filtered substances	Benzotriazole
2	CWFR (t-1)	CWFR (t-1)	CWFR	Inh. A
3	KS4.3	CWFR (t -2)	CWFR (t-1)	Orthophosphate
4	Nitrate	CWFR (t-4)	CWFR (t-2)	Zinc
5	CWFR (t-3)	CWFR (t-3)	CWFR (t-3)	CWFR (t-4)
6	CWFR (t-4)	Filtered substances	CWFR (t-4)	CWFR (t-1)
7	Filtered substances	Benzotriazole	Bacterial count	CWFR (t-2)
8	TOC	Ca^{2+}	Manganese	CWFR
9	Manganese	Manganese	Thickening agent	CWFR (t-3)
10	Chloride	pH	Ca^{2+}	Filtered substances
11	Turbidity	Bacterial count	Benzotriazole	Manganese
12	CWFR(t-2)	Dissolved Zinc	Chloride	
13	HCO_3^-	Chloride	pH	
14	Orthophosphate (t-1)	Benzotriazole (t-1)	Dissolved Zinc	
15	Inh. A (t-1)	Benzotriazole(t-2)	Ca^{2+}/Mg^{2+}	
16	Benzotriazole	Ca^{2+}/Mg^{2+}		
17	TIC	Benzotriazole(t-3)		

Thus, the remainder of the variables are expected to complement the information embedded in the data pertaining to the inhibitors. As noted in Table 2, CWFR is considered an important variable by all variable selection methods. Additionally, the variable named *filtered substances* is considered important by all methods shown in Table 2, even though it corresponds to a one-time event that occurred in Division 2 of Fig. 3. W-XGB recognizes N-containing compounds as well as historical values of Inhibitor A and orthophosphate as important variables as opposed to PCC, point-wise MI, and partial MI.

Test results from predicting events demonstrated in Fig. 3 using the trained SVR with input variables in Table 2 are illustrated in Supplementary Figs. 3 to 6. The models developed by all variable selection methods have predicted the rate of corrosion in Division 1 with reasonable accuracy. PCC and MI have best predicted the corrosion rate in Division 2. None of the models were able to successfully extrapolate less than the minimum corrosion rate or higher than the maximum corrosion rate in the training data set, as evident from the predicted corrosion profiles in divisions 3 and 4. The minimum corrosion rate occurs in Division 3, while the maximum corrosion rate occurs in Division 4.

As stated in Section on Variable selection in the Methodology, partial MI in Table 2 was implemented such that variables with high positive point-wise mutual information were given priority. Positive point-wise MI suggests that variables co-occur with the output, i.e. variables that respond at the same time as corrosion with a high probability. According to Table 2, there is a significant similarity among variables selected by partial MI and PCC. It appears that the most linearly correlated variables are similar to the most co-occurrent variables. However, variables with negative point-wise MI cannot be deemed irrelevant as only a value close to 0 is considered irrelevant[44].

Partial MI determines water quality measurements that support the prediction of the corrosion rate, in addition to the starting fixed set of predetermined variables (i.e. benzotriazole, orthophosphate, inhibitor A, and zinc). In doing so, most variables were noted to have high negative point-wise MI. High negative point-wise information indicates that a variable does not co-occur well with the output variable. In other words, such a feature is of a probability distribution complementary to that of the output variable. A comparison of features selected by magnitude-based partial MI and positive partial MI, as well as magnitude-based point-wise MI and positive point-wise MI is given in Table 3. Nitrate was ignored from variables under magnitude-based point-wise MI as it is highly correlated to TIN. Similarly, HCO_3^- was ignored as it is highly correlated to TIC.

Variables selected by magnitude-based partial MI shown in Table 3 include those that co-occurs less with the corrosion rate, such as organic phosphorous, total phosphate, and AOX. KS4.3, HCO_3^-, inhibitor A (t-1) and TIC are common with magnitude-based partial MI and W-XGB. The main difference between the sets of variables selected by positive point-wise MI and magnitude-based point-wise MI is that the latter identifies nitrates and TIN as important variables. As observed in Table 2, Nitrate is also identified as important by W-XGB. As W-XGB is capable of identifying nonlinear variables that significantly affect corrosion, it appears that magnitude-based partial MI ensures the inclusion of variables that are non-linearly correlated and less co-occurrent with the corrosion rate. However, due to the lack of linearly correlated/co-occurrent variables among magnitude-based partial MI variables, the prediction accuracy has declined, as shown in Supplementary Table 5.

None of the variables selected by magnitude-based partial MI are sufficiently co-occurrent and/or linearly correlated to the corrosion rate. Therefore, the prediction accuracy of a model trained with these variables is not adequate. In order to facilitate predicting the corrosion rate using data-driven models, the presence of one or more co-occurrent/linearly correlated variables seem essential. This was demonstrated by the addition of two co-occurrent variables: CWFR and $[Ca^{2+}]$. As shown in Fig. 2, CWFR is highly correlated to the corrosion rate. As scaling contributes to the formation of a barrier layer between the metal surface and the solution, Ca^{2+} ions contribute to the inhibition of the corrosion rate. Ca^{2+} concentration has also been listed as a highly linearly correlated variable under PCC in Table 2. Therefore, CWFR and Ca^{2+} were included among the top-ranked variables along with benzotriazole, inhibitor A, orthophosphate and zinc. The remaining variables were selected by re-implementing the partial MI algorithm in May et al. (2008). The number of variables was determined with the algorithm shown in Supplementary Figure 2. The variables resulting from repeating the magnitude-based partial MI algorithm were Ca^{2+}, CWFR, benzotriazole, inhibitor A, orthophosphate, zinc, COD, AOX, zinc(t-1), total phosphate, zinc(t-3), zinc(t-2), and organic phosphates. A comparison of prediction accuracies of models trained with positive partial MI, magnitude-based partial MI and the latter with Ca^{2+} and CWFR is given in Supplementary Table 6. It can be observed that the addition of linearly correlated and co-occurrent variables improved the predictive ability of the model trained with magnitude-based partial MI variables.

Table 3 | Comparison of ranking generated by magnitude-based and positive partial MI

Ranking	Magnitude-based partial MI	Positive partial MI	Magnitude based point-wise MI	Positive point-wise MI
1	Benzotriazole	Benzotriazole	Filtered substances	Filtered substances
2	Inhibitor A	Inhibitor A	CWFR	CWFR
3	Orthophosphate	Orthophosphate	CWFR (t-1)	CWFR (t-1)
4	Zinc	Zinc	CWFR (t-2)	CWFR (t-2)
5	COD	CWFR (t-4)	CWFR (t-3)	CWFR (t-3)
6	AOX	CWFR (t-1)	CWFR (t-4)	CWFR (t-4)
7	Zinc (t-1)	CWFR (t-2)	Bacterial count	Bacterial count
8	Total phosphates	CWFR	Manganese	Manganese
9	Zinc (t-3)	CWFR (t-3)	TIN	Thickening agent
10	Zinc (t-2)	Filtered substances	Thickening agent	Ca^{2+}
11	Ca^{2+}/Mg^{2+}	Manganese	Ca^{2+}	Benzotriazole
12	Organic phosphate		Benzotriazole	Chloride
13	TIC		Chloride	pH
14	KS4,3		pH	Dissolved Zinc
15	Inhibitor A (t-3)		TIC	Ca^{2+}/Mg^{2+}
16	Inhibitor A (t-2)			
17	Sulfate			

Sensitivity analysis

Sobol sensitivity analysis[45] was carried out based on SVR models developed using input variables selected by PCC and magnitude-based partial MI by calculating first-order and total Sobol indices. The purpose of the analysis is to understand the difference between the impact of linearly and nonlinearly correlated variables on the sensitivity of inhibitors. Dissolved zinc was omitted as a variable for model comparison as it is correlated to the total amount of zinc. Filtered substances was also omitted as a variable, as its values correspond to a single operational change. The main objective of the model is to optimize dosages of benzotriazole, inhibitor A, orthophosphate and zinc. Therefore, including variables highly correlated to the four inhibitors will provide redundant information and inconvenience optimization of the inhibitors from having to account for the correlations with the input variables. Orthophosphate and inhibitor A are slightly correlated to total phosphates. However, the model prediction accuracy demonstrated in Figs. A3 to A6 are also not perfect. In the meantime, total phosphates are highly influenced by antiscalants added to cooling water. As antiscalants are not accounted for by the variables included in this study, total phosphates were not excluded. The results of the Sobol sensitivity analysis are given in Table 4.

When comparing 1st order Sobol sensitivities of magnitude-based partial MI with and without CWFR and Ca^{2+}, the relative sensitivities assigned to corrosion inhibitors, especially benzotriazole, decline notably upon the addition of variables more linearly correlated to the corrosion rate. As observed in Fig. 6, the Sobol sensitivity of zinc is low in both instances when magnitude-based partial MI was used. Based on Table 4, historical values of zinc appear to have a higher influence on the corrosion rate than the present value. Among variables considered by PCC pH, benzotriazole and CWFR, have the highest 1st order Sobol indices. Therefore, they have the highest sensitivity to corrosion rate.

Total Sobol index is an overall sensitivity accounting for sensitivity of a variable to the output as well as interaction effects among other variables. The highest total Sobol indices of PCC variables can be observed in pH, benzotriazole, and zinc.

The 1st-order Sobol sensitivity of benzotriazole among magnitude-based partial MI (without CWFR and Ca^{2+}) and PCC variables are similar. The sensitivity of benzotriazole is low when the model is trained with partial MI variables with CWFR and Ca^{2+}. Therefore, unique combinations of variables have an impact on the Sobol sensitivity of an inhibitor. It appears that linearly correlated and co-occurrent variables are suitable for predicting trends of corrosion rate, whereas, magnitude-based partial MI variables ensure a high Sobol sensitivity of corrosion inhibitors to the corrosion rate.

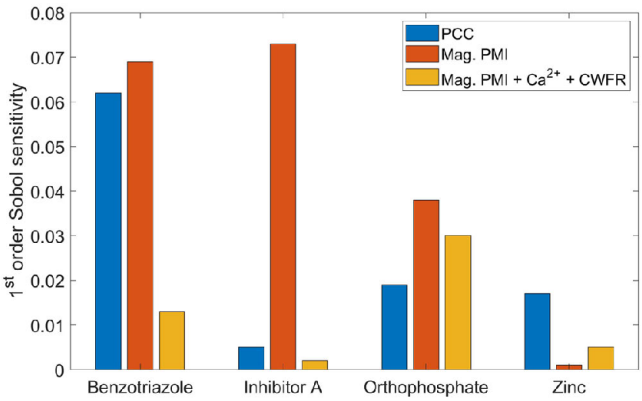

Fig. 6 | Bar chart demonstrating that inhibitors are given higher priority when magnitude-based partial MI is used for variable selection. The 1st order Sobol sensitivity of benzotriazole, inhibitor A and orthophosphate are highest among variables chosen by magnitude-based partial MI. The decline in the sensitivity of zinc is due to zinc (t-2) and zinc(t-3) being correlated to zinc, which have higher sensitivities as presented in Table 4.

Optimization

Model prediction accuracy and sensitivity to corrosion inhibitors were considered vital for optimization of corrosion inhibitors in this study. Although PCC variables demonstrated the highest prediction accuracy, it is apparent from Supplementary Fig. 6 that the predictions largely fluctuate from the measured rates. Therefore, historical values of corrosion rate were included as an input variable. As described in Section on Optimization in the Methodology, the historical value of the corrosion rate 7 days prior to the considered time 't', CR (t-7), were used as historical values. The remainder of the input variables (in addition to the inhibitors) were selected using PCC. Variables used for model development are benzotriazole, inhibitor A, orthophosphate, zinc, CR (t-7), CWFR, CWFR (t-7) and benzotriazole (t-7). Sobol sensitivity indices with respect to the variables are given in Table 5.

It is clear from Table 5 that the sensitivity of CR(t-7) is significantly higher than other variables. The Sobol sensitivities of benzotriazole,

Table 4 | Sobol sensitivity analysis of PCC and partial MI variables

PCC			Magnitude-based partial MI			Magnitude-based partial MI + CWFR + Ca^{2+}		
Variables	1st order Sobol index	Total Sobol index	Variables	1st order Sobol index	Total Sobol index	Variables	1st order Sobol index	Total Sobol index
Benzo	0.062	0.204	Benzo	0.069	0.160	Ca^{2+}	0.114	0.177
Inh. A	0.005	0.067	Inh. A	0.073	0.122	CWFR	0.181	0.402
Ortho	0.019	0.193	Ortho	0.038	0.131	Benzo	0.013	0.231
Zinc	0.017	0.096	Zinc	0.001	0.105	Inh. A	0.002	0.056
CWFR	0.052	0.111	COD	0.006	0.070	Ortho.	0.030	0.130
CWFR (t-1)	0.006	0.057	AOX	0.020	0.080	Zinc	0.005	0.070
CWFR (t -2)	0.011	0.041	Zinc (t-1)	0.005	0.029	COD	0.056	0.160
CWFR (t-4)	0.019	0.053	Total phosphate	0.075	0.120	AOX	0.012	0.067
CWFR (t-3)	0.013	0.049	Zinc (t-3)	0.037	0.060	Zinc(t-1)	0.0002	0.015
Ca^{2+}	0.012	0.108	Zinc (t-2)	0.024	0.047	Total phosphate	0.040	0.110
Manganese	0.019	0.082	Ca^{2+}/Mg^{2+}	0.043	0.131	Zinc(t-3)	0.017	0.041
pH	0.177	0.339	Organic phosphate	0.021	0.059	Zinc(t-2)	0.003	0.019
Bacterial count	0.001	0.063	TIC	0.006	0.086	Organic phosphate	0.020	0.093
Chloride	5.2E-05	0.061	KS4,3	0.102	0.179			
Benzotriazole (t-1)	0.016	0.095	Inh. A (t-3)	0.001	0.037			
			Inh. A (t-2)	0.003	0.042			
			Sulfate	0.015	0.087			

Table 5 | Sobol sensitivity of input variables

Variables	1st order Sobol index	Total Sobol index
Benzotriazole	0.015	0.023
Inhibitor A	0.002	0.011
Orthophosphate	0.002	0.007
Zinc	0.0006	0.012
CR (t-7)	0.773	0.817
CWFR	0.082	0.113
CWFR (t -7)	0.025	0.060
Benzotriazole (t-7)	0.003	0.024
Manganese	0.0006	0.002
Nitrite	0.0007	0.001

inhibitor A, orthophosphate and zinc are notably lower than the Sobol sensitivity of CR(t-7). Prediction of events indicated in Fig. 3 is shown in Fig. 7.

A significant improvement of the prediction accuracy can be observed upon the addition of CR(t-7) as a variable. In order to demonstrate how model structure affects prediction accuracy as well as Sobol sensitivities of inhibitors, the SVR was compared with two other models that have frequently demonstrated high predictive performance in the corrosion literature: XGBoost implementation of gradient based decision trees (XGB) and Gaussian process regression (GPR). Apart from the fact that they are all regression models SVR, GPR and XGB differ from each other. SVR is deterministic (each input always provides the same output), GPR is probabilistic based on Bayesian inference (provides a distribution over functions that fit the data) with uncertainty quantification. GPR is non-parametric with complexity adapting to the data while SVR is parametric and assumes a specific form of the function it fits (e.g., polynomial). XGB builds an ensemble of decision trees sequentially, where each new tree corrects errors made by the previous trees, while SVR uses mathematical optimisation to try and find the hyperplane that best fits the data (with kernel functions to deal with non-linearity in the data). Input variables indicated in Table 5 were used for the model comparison. As demonstrated in Supplementary Figure 7, the prediction accuracies of GPR and XGB are similar to SVR. As shown in Supplementary Table 7, Sobol sensitivity analysis reveals that GPR and XGB models also assign the highest sensitivity to the most linearly correlated variable (CR(t-7)).

The model was used to optimize inhibitor dosages and minimize the corrosion rate as described in the section on Optimization under the Methodology. It should be noted that the inhibitors were varied to determine the optimum values per data point while the remaining input variables of the SVR shown in Table 3 were fixed as constants (at recorded values), per data point.

According to Fig. 8 the optimized concentrations of benzotriazole is set to nearly zero throughout the entire time period. Optimized concentrations of orthophosphate, inhibitor A and zinc are also set to minimum values for 1100 days. However, the optimized corrosion rate follows the trend of measured rates despite the significant decreases in the inhibitor concentrations. The model appears to depend on CR(t-7), which is the most linearly correlated to the corrosion rate, to make predictions. The lack of sensitivity to inhibitors has resulted in the model barely accounting for the large decreases in inhibitor dosages.

The optimized corrosion rate is slightly higher than measured rates at low corrosion rates. The increase in corrosion rate in response to decreases in inhibitor dosages can be expected. The optimized corrosion rates appear to be less than measured rates at high corrosion rates. However, the inhibitor concentration transported to the surface of the metal decreases as the inhibitor dosages decreases as drastically as shown in Fig. 8. This should result in an increase in the corrosion rate. Studies carried out by Barmatov et al.[46] and Khan et al.[47] demonstrate how the corrosion rate increases when the inhibitor dosages are decreased at high flow rates.

Therefore, regardless of the high prediction accuracy, inhibitors cannot be optimized using a data-driven model if it overlooks the impact of inhibitors by assigning insignificant Sobol sensitivities. It is apparent that in any data-driven model trained to improve the prediction accuracy of the corrosion rate, linearly correlated or co-occurrent variables are given the highest priority.

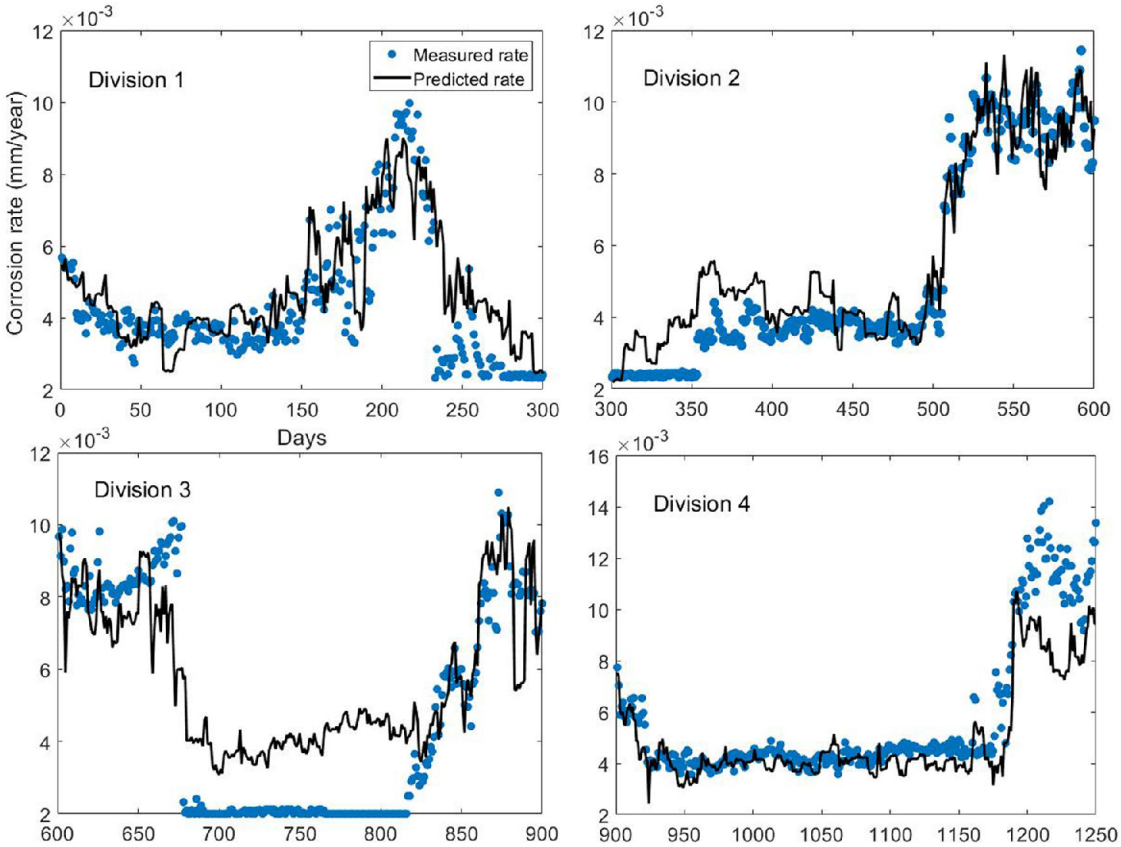

Fig. 7 | Model prediction accuracy for the four divisions shown in Fig. 3 The R^2 and RMSE of the predictions of the four divisions are given in Table 6.

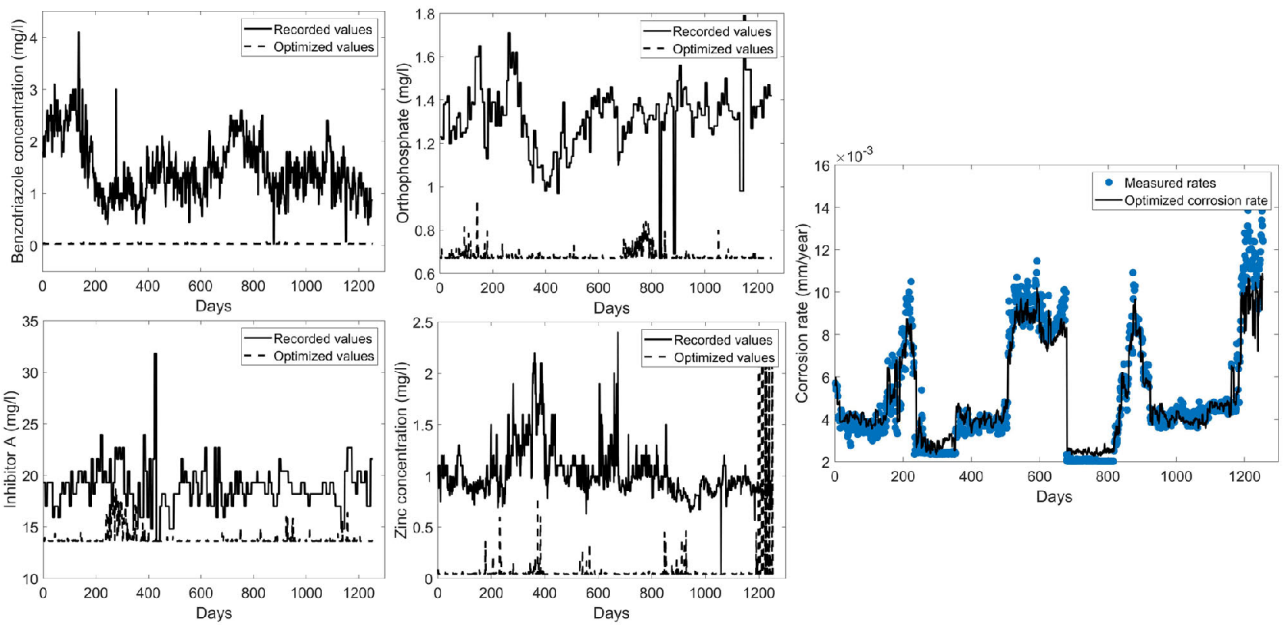

Fig. 8 | **Results of optimization.** Figure 8 demonstrates how the corrosion rate responds to the optimized concentrations of benzotriazole, inhibitor A, orthophosphate and zinc.

As discussed in the Section on Variable selection in the Methodology, variables selected using magnitude-based partial mutual information give higher priority to inhibitors. The shortcoming of magnitude-based partial mutual information was that the model prediction accuracy was not sufficient. However, it is possible to use deep neural networks (DNNs) for improving the prediction accuracy of corrosion rate based on these variables. Although, DNNs can capture nonlinearities well, they are extremely data-hungry models. Therefore, the only remaining alternative is to integrate mechanistic aspects of corrosion with a data-driven component. As corrosion is a complex process involving several reactions, it is not possible

Table 6 | Model performance

Division	RMSE (mm/year)	R^2
1	0.0012	0.80
2	8.67×10^{-4}	0.93
3	0.0018	0.81
4	0.0015	0.88

to rely on a mechanistic model alone. Therefore, a combination of mechanistic as well as data-driven, also known as a hybrid modelling approach must be employed to facilitate optimizing inhibitor dosages.

Data availability

Data used to conduct the study was obtained from a large-scale chemical plant. The authors are not at the liberty to share the data set, as it is confidential. The prime objective of this study is to demonstrate the feasibility of using data driven models trained to predict the corrosion rate for optimizing the dosages of corrosion inhibitors. Therefore, the conclusions of the study can be replicated using any data set large enough to train a data driven model containing water qualities and the respective corrosion rate measured using linear polarization resistance.

Received: 12 June 2024; Accepted: 7 December 2024;
Published online: 19 December 2024

References

1. *Water handbook - Cooling water corrosion control | Veolia*. (n.d.). Industrial Water & Process Treatment Technologies & Solutions | Veolia. https://www.watertechnologies.com/handbook/chapter-24-corrosion-control-cooling-systems.
2. Chirkunov, A., & Kuznetsov, Y. Corrosion inhibitors in cooling water systems. Mineral Scales and Deposits, 85-105, https://doi.org/10.1016/b978-0-444-63228-9.00004-8 (2015).
3. Finšgar, M. & Milošev, I. Inhibition of copper corrosion by 1,2,3-benzotriazole: A review. *Corros. Sci.* **52**, 2737–2749 (2010).
4. Opel, O., Wiegand, M., Neumann, K., Zargari, M. & Plesser, S. Corrosion in heating and cooling water circuits - A Field study. *Energy Procedia* **155**, 359–366 (2018).
5. Touir, R. et al. Corrosion and scale processes and their inhibition in simulated cooling water systems by monosaccharides derivatives. *Desalination* **249**, 922–928 (2009).
6. Mahgoub, F., Abdel-Nabey, B. & El-Samadisy, Y. Adopting a multipurpose inhibitor to control corrosion of ferrous alloys in cooling water systems. *Mater. Chem. Phys.* **120**, 104–108 (2010).
7. Rahmani, K., Jadidian, R. & Haghtalab, S. Evaluation of inhibitors and biocides on the corrosion, scaling and biofouling control of carbon steel and copper–nickel alloys in a power plant cooling water system. *Desalination* **393**, 174–185 (2016).
8. Marín-Cruz, J., Cabrera-Sierra, R., Pech-Canul, M. & González, I. EIS study on corrosion and scale processes and their inhibition in cooling system media. *Electrochim. Acta* **51**, 1847–1854 (2006).
9. Monticelli, C. Corrosion inhibitors. *Encyclopedia Interfacial Chem.* 164–171. https://doi.org/10.1016/b978-0-12-409547-2.13443-2 (2018).
10. Verma, D. K. et al. Computational modeling: Theoretical predictive tools for designing of potential organic corrosion inhibitors. *J. Mol. Struct.* **1236**, 130294 (2021).
11. Verma, C., Ebenso, E. E. & Quraishi, M. Corrosion inhibitors for ferrous and non-ferrous metals and alloys in Ionic sodium chloride solutions: A review. *J. Mol. Liq.* **248**, 927–942 (2017).
12. Ou, H., Tran, Q. T. & Lin, P. A synergistic effect between gluconate and molybdate on corrosion inhibition of recirculating cooling water systems. *Corros. Sci.* **133**, 231–239 (2018).
13. Edoziuno, F. O. et al. Optimization and development of predictive models for the corrosion inhibition of mild steel in sulphuric acid by methyl-5-benzoyl-2-benzimidazole carbamate (mebendazole). *Cogent. Eng.* **7**, 1714100 (2020).
14. Omran, M. A., Fawzy, M., Mahmoud, A. E. & Abdullatef, O. A. Optimization of mild steel corrosion inhibition by water hyacinth and common reed extracts in acid media using factorial experimental design. *Green. Chem. Lett. Rev.* **15**, 216–232 (2022).
15. Ansari, A. et al. Experimental, theoretical modeling and optimization of inhibitive action of ocimum basilicum essential oil as green corrosion inhibitor for C38 steel in 0.5 M H2SO4 medium. *Chem. Afr.* **5**, 37–55 (2021).
16. Ferguson, R. J. OPTIMIZING INHIBITOR BLENDS USING COMPUTER MODELING. French-Creek software. https://www.frenchcreeksoftware.com/online-library/ (2007).
17. Aghaaminiha, M. et al. Machine learning modeling of time-dependent corrosion rates of carbon steel in presence of corrosion inhibitors. *Corros. Sci.* **193**, 109904 (2021).
18. Coelho, L. B. et al. Reviewing machine learning of corrosion prediction in a data-oriented perspective. *npj Mater. Degradat.*, 6. https://doi.org/10.1038/s41529-022-00218-4 (2022).
19. Zhi, Y., Yang, T. & Fu, D. An improved deep forest model for forecast the outdoor atmospheric corrosion rate of low-alloy steels. *J. Mater. Sci. Technol.* **49**, 202–210 (2020).
20. May, R., Dandy, G., & Maier, H. Review of input variable selection methods for artificial neural networks. Artificial Neural Networks - Methodological Advances and Biomedical Applications. https://doi.org/10.5772/16004 (2011).
21. Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. A review of feature selection methods on synthetic data. *Knowl. Inf. Syst.* **34**, 483–519 (2012).
22. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **40**, 16–28 (2014).
23. Ang, J. C., Mirzal, A., Haron, H. & Hamed, H. N. Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **13**, 971–989 (2016).
24. Remeseiro, B. & Bolon-Canedo, V. A review of feature selection methods in medical applications. *Comput. Biol. Med.* **112**, 103375 (2019).
25. The effect of chloride and orthophosphate on the release of iron from a drinking water distribution system cast iron pipe | Science Inventory | Us EPA. (n.d.). Retrieved from https://cfpub.epa.gov/si/si_public_record_Report.cfm?Lab=NRMRL&dirEntryId=96762.
26. Walker, R. The use of benzotriazole as a corrosion inhibitor for copper. *Anti-Corros. Methods Mater.* **17**, 9–15 (1970).
27. Gopi, D. et al. Corrosion and corrosion inhibition of mild steel in groundwater at different temperatures by newly synthesized Benzotriazole and Phosphono derivatives. *Ind. Eng. Chem. Res.* **53**, 4286–4294 (2014).
28. Young, T. J. Association of Water Technologies, Inc. Spring Convention & Exposition. Retrieved from https://www.lubrizol.com/-/media/9357AA2C9139437DA193F430618B4654.pdf (2010).
29. Weimer, P. J., Van Kavelaar, M. J., Michel, C. B. & Ng, T. K. Effect of phosphate on the corrosion of carbon steel and on the composition of corrosion products in two-stage continuous cultures of Desulfovibrio desulfuricans. *Appl. Environ. Microbiol.* **54**, 386–396 (1988).
30. Anaee, R. A. Sodium silicate and phosphate as corrosion inhibitors for mild steel in simulated cooling water system. *Arab. J. Sci. Eng.* **39**, 153–162 (2013).
31. Szklarska-Smialowska, Z. & Mańkowski, J. Cathodic inhibition of the corrosion of mild steel in phosphate, tungstate, arsenate and silicate solutions containing Ca2+ions. *Br. Corros. J.* **4**, 271–275 (1969).
32. Water handbook - Cooling system Microbiological control | Veolia. (n.d.). Retrieved from https://www.watertechnologies.com/handbook/chapter-26-microbiological-control-cooling-system.

33. Osorio-Celestino, G. R. et al. Influence of calcium scaling on corrosion behavior of steel and aluminum alloys. *ACS Omega* **5**, 17304–17313 (2020).
34. Xu, W., Zhang, B., Deng, Y., Yang, L. & Zhang, J. Nitrate on localized corrosion of carbon steel and stainless steel in aqueous solutions. *Electrochim. Acta* **369**, 137660 (2021).
35. Vargas, S. M., Woollam, R., Durnie, W., Hodges, M., & Betancourt, D. Effect of nitrate on carbon steel corrosion. *Day 1 Mon, April 03, 2017*. https://doi.org/10.2118/184512-ms (2017).
36. Inhibitor ECP *yellow metal corrosion control technology put to the test at a West Coast refinery*. (n.d.). Industrial Water & Process Treatment Technologies & Solutions | Veolia. https://www.watertechnologies.com/case-study/inhibitor-ecp-yellow-metal-corrosion-control-technology-put-test-west-coast-refinery-2.
37. Somerscales, E. Fundamentals of corrosion fouling. *Br. Corros. J.* **34**, 109–124 (1999).
38. Pedregosa et al. Scikit-learn: Machine Learning in Python. *JMLR.* **12**, 2825-2830 (2011).
39. May, R. J., Maier, H. R., Dandy, G. C. & Fernando, T. G. Non-linear variable selection for artificial neural networks using partial mutual information. *Environ. Model. Softw.* **23**, 1312–1326 (2008).
40. Yan, L., Diao, Y., Lang, Z. & Gao, K. Corrosion rate prediction and influencing factors evaluation of low-alloy steels in marine atmosphere using machine learning approach. *Sci. Technol. Adv. Mater.* **21**, 359–370 (2020).
41. Bowden, G. J., Maier, H. R., & Dandy, G. C. Optimal division of data for neural network models in water resources applications. *Water Resources Res.* **38**, https://doi.org/10.1029/2001wr000266 (2002).
42. Welling, M. (2004). Support vector regression. Department of Computer Science, University of Toronto, Toronto (Canada).
43. Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evolut. Comput.* **6**, 182–197 (2002).
44. Xu, Y., Jones, G., Li, J., Wang, B. & Sun, C. A Study on Mutual Information-based Feature Selection for Text Categorization. *J. Comput. Inform. Syst.* **3**, 1007–1012 (2007).
45. Zhang, X., Trame, M., Lesko, L. & Schmidt, S. Sobol sensitivity analysis: A tool to guide the development and evaluation of systems pharmacology models. *CPT: Pharmacomet. Syst. Pharmacol.* **4**, 69–79 (2015).
46. Barmatov, E., Hughes, T. & Nagl, M. Efficiency of film-forming corrosion inhibitors in strong hydrochloric acid under laminar and turbulent flow conditions. *Corros. Sci.* **92**, 85–94 (2015).
47. Khan, P. F., Shanthi, V., Babu, R. K., Muralidharan, S. & Barik, R. C. Effect of benzotriazole on corrosion inhibition of copper under flow conditions. *J. Environ. Chem. Eng.* **3**, 10–19 (2015).

Acknowledgements

The authors acknowledge the 'EU Horizon 2020' grant agreement No. 958396 for funding the research as part of the AQUASPICE project undertaken for advancing sustainability of process industries through digital and circular water use innovations. The authors also appreciate the support given by members of BIOMATH and PaInT research groups of Ghent University.

Author contributions

Conception of study by C.D.J. Methodology by C.D.J., D.F.P., M.V.H and I.N. Writing by C.D.J. Revision by all authors. Supervision by I.P.H., D.F.P., A.V. and I.N. Project management by I.N. Provision of data and industry knowledge by T.D.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41529-024-00545-8.

Correspondence and requests for materials should be addressed to Chamanthi Denisha Jayaweera.