Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2023.30212

316 (2023) 744–754
December

Prediction of calcium concentration in circulating seawater in a closed-cycle seawater cooling system using machine learning models

Zhijie Li[a], Zhuoxiao Li[b], Lianqiang Zhang[a], Chong Chen[a,*], Mingming Hu[a], Xue Li[a], Kai Xu[c]

[a]*The Institute of Seawater Desalination and Multipurpose Utilization, Ministry of Natural Resources, No. 55 HangHai Road, Nankai District, Tianjin, China, Tel./Fax: +86 22 87891247; emails: qdchenchong@163.com (C. Chen), lizhijie@tju.edu.cn (Z. Li), 13920935129@163.com (L. Zhang), mmhu_sdmu@126.com (M. Hu), xuefenfei1123@163.com (X. Li)*
[b]*Tianjin Zhong Hai Water Treatment Technology Co., Ltd., No. 55 HangHai Road, Nankai District, Tianjin, China, email: glimon729@hotmail.com*
[c]*Shandong Lubei Bishuiyuan Seawater Desalination Co., Ltd., No. 55 HangHai Road, Nankai District, Tianjin, China, email: bsyxukai@163.com*

Received 15 June 2023; Accepted 5 December 2023

ABSTRACT

The objective of this investigation is to establish and evaluate the efficacy of two machine learning models, namely random forests (RFs) and support vector machines, in forecasting the calcium concentration within the circulating seawater of a closed-cycle seawater cooling system, thereby replacing conventional and time-consuming laboratory testing. These models were constructed based on daily seawater quality data, and their predictive capabilities were evaluated utilizing metrics such as the coefficient of determination (R^2) and root mean square error. Additionally, a sensitivity analysis employing the Sobol sensitivity analysis technique was performed. The findings indicated that both models effectively forecasted the calcium concentration in the circulating seawater within 1-d intervals. The RF model displayed superior prediction accuracy during the training phase, and it yielded comparable results during the validation phase. Moreover, the sensitivity analysis revealed that the RF model outperformed other models in capturing the causal relationship between calcium concentration and the input variables associated with the closed-cycle seawater cooling system.

Keywords: Seawater cooling; Random forest; Support vector machine; Calcium concentration; Prediction accuracy; Sensitivity analysis

1. Introduction

The distribution of global water resources is highly imbalanced, with freshwater constituting a mere 2.5% and seawater comprising the remaining 97.5% [1]. As economic development progresses, the scarcity of water resources has become increasingly severe. In coastal regions, the utilization of seawater as a cooling source has been employed since the 1970s, either through once-through or closed-cycle methods [2]. However, the once-through seawater cooling approach poses significant environmental challenges, particularly regarding the discharge of heated water back into the sea. Consequently, governments have enacted regulations governing the disposal of high-temperature wastewater into marine environments. To comply with these regulations, the adoption of closed-cycle cooling technology, which incorporates the use of seawater cooling towers, has proven to be highly effective. By the end of 2021, China had established a total of 22 closed-cycle seawater cooling systems, with a collective circulation capacity of 1,934,800 tons/h [3]. Nevertheless, the presence of salts in seawater gives rise to several complex engineering challenges for the cooling system, including corrosion, microbial attachment [4], and scaling [5]. Of particular concern is the formation of calcium carbonate scaling in the cooling devices, as it can lead to blockages and operational

* Corresponding author.

shutdown incidents. Thus, the implementation of robust chemical treatment and monitoring technologies for seawater is imperative to inhibit scale formation and ensure the stable operation of closed-cycle cooling systems.

Seawater quality indicators play a vital role in the implementation of optimization strategies in seawater cooling systems. The chemical treatment of seawater scaling in cooling systems presents a significant challenge, primarily due to the complexities associated with monitoring calcium concentration. Despite the challenges in measuring key indicators like calcium concentration, most utilities rely on laboratory tests using offline measurements to monitor chemical dosing and predict calcium carbonate deposition tendencies. Istepanian [6] argue that the current reliance on experiential and offline water quality testing conducted by laboratory analysts for system control in water cooling systems often leads to excessive dosing to mitigate scaling risks. To significantly improve the efficiency of monitoring and controlling chemical inventory and residues in seawater cooling systems, precise and reliable instruments, along with advanced control methods for continuous real-time online monitoring, are essential. The concentration of calcium ions in closed-cycle seawater is a critical parameter for assessing scaling tendencies and serves as fundamental data for developing scale inhibitor dosing strategies. Currently, the detection of calcium ions in seawater circulation primarily involves chemical techniques such as ethylenediaminetetraacetic acid (EDTA) titration [7,8], emerging detection methods like ion chromatography [9,10], and the selective electrode method [11]. Chemical methods remain the most widely employed approaches in closed-cycle seawater cooling systems due to their advantages in terms of accuracy, reliability, and maturity. However, these methods require skilled personnel, involve time-consuming processes, and are unable to meet the demands of real-time seawater quality analysis. Emerging detection methods, although reliant on specialized instruments, suffer from drawbacks such as high costs and limited stability. Additionally, sensors used for calcium carbonate scaling issues and restricted detection ranges pose common challenges, limiting their ability to provide real-time monitoring within short timeframes for seawater closed-cycle cooling systems. Consequently, the acquisition of fast and real-time online data on calcium concentration has become a paramount priority in the management of seawater closed-cycle cooling systems.

The emergence of machine learning (ML) technology has profoundly transformed diverse domains by facilitating predictive capabilities, identification of significant features, and detection of anomalies. ML has found wide-ranging applications in prediction tasks, employing regression or classification modeling techniques. At the core of this process lies the assumption that the training examples supplied to the algorithm are representative of the examples encountered by the model during prediction, irrespective of any temporal dependencies. Supervised learning entails utilizing labeled sample outputs to facilitate model training. ML methods have been extensively applied to forecast variations in numerous wastewater variables, such as nitrogen, phosphorus, solids, chemical oxygen demand, biochemical oxygen demand, and future flow rate [12–17]. These applications effectively showcase the efficacy of ML algorithms in automatically classifying and elucidating chemical patterns within the water environment that would otherwise be arduous to discern manually.

This research focuses on the development of two machine learning models, namely random forests (RFs) and support vector machines (SVMs), to predict the calcium concentration in seawater within a closed-cycle seawater cooling system installed in a power plant located on the east coast of China. The modeling and validation processes employ 7 y worth of seawater quality analysis data. The model parameters are optimized using an optimization method to enhance their performance. Additionally, sensitivity analysis is conducted to investigate the cause-and-effect relationships between input and output values, thereby aiding in future process control and the selection of appropriate machine learning models for seawater cooling applications. These models utilize online seawater quality parameters, including pH and conductivity, to forecast the calcium concentration, enabling real-time online monitoring of crucial indicators for precise closed-cycle cooling seawater control. The continuous and real-time monitoring, facilitated by accurate and dependable instrumentation, along with advanced control methodologies, enhances the effectiveness of monitoring and controlling chemical inventory and chemical residuals within seawater cooling systems.

2. Methods

2.1. Field sampling

The current research collected seawater quality analysis data from a closed-cycle seawater cooling power plant situated in the eastern region of China. The power plant operates with a circulating seawater flow rate of 100,000 tons/h, as depicted in Fig. 1. With a successful operational history of over 10 y, the system was designed with seawater concentration cycles set at 2.0. Through the analysis of the gathered data, it can be inferred that the seawater exhibits a relatively stable behavior with periodic variations, displaying characteristics typical of regular seawater. Daily sampling was conducted on both the makeup and circulating seawater, directly collected from the pipelines. These samples were transported to an on-site laboratory for subsequent analysis, focusing on parameters such as pH, calcium concentration, chloride concentration, conductivity, and other relevant variables.

2.2. Sample analysis

The pH and conductivity measurements were performed using a pH meter (GB 6920-1986) and a conductivity tester (GB 11007-1989), respectively, following the prescribed instrument calibration procedures. On the other hand, the determination of calcium concentration employed the pH method (GB/T15452-2009). For the measurement of calcium concentration in a seawater sample, a 50 mL sample was subjected to filtration, and the calcium ion content was determined through titration using an EDTA standard titration solution. The titration process was carried out within a pH range of 12–13, utilizing calcium-carboxylic acid as the indicator. During titration, EDTA formed a complex

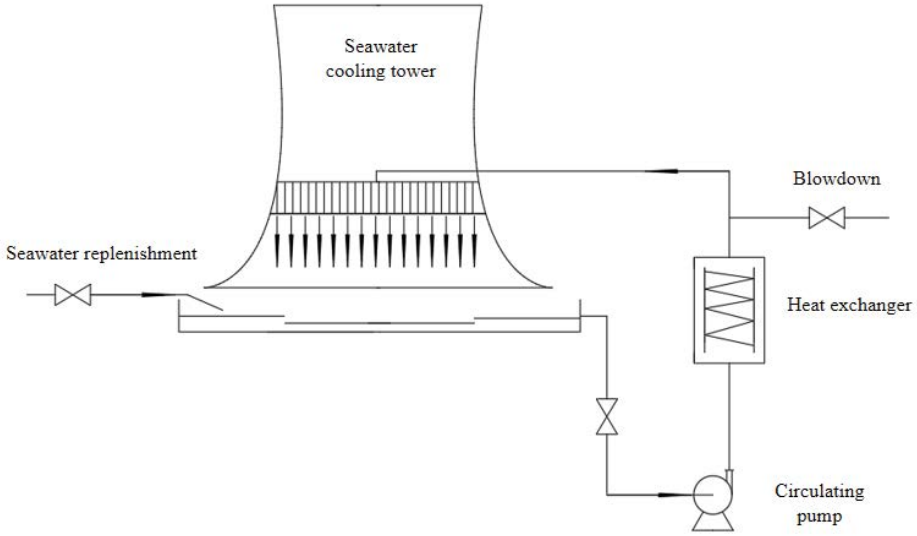

Fig. 1. Schematic diagram of the closed-cycle seawater cooling system.

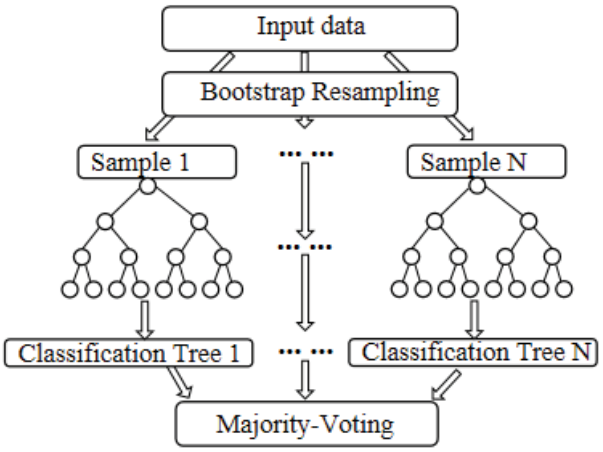

Fig. 2. Illustration of general conceptual model structure for Random Forest algorithm.

with the unbound calcium ions present in the solution, and the endpoint was attained when the color of the solution transitioned from purplish-red to bright blue.

2.3. Modeling approaches

2.3.1. Random Forest algorithm

The Random Forest algorithm, depicted in Fig. 2, is a widely utilized and robust supervised machine learning technique proficient in addressing both regression and classification problems. It leverages ensemble learning by combining multiple decision trees, which serve as the fundamental units of the algorithm [18]. Decision trees are non-linear and non-parametric supervised classification algorithms, where non-terminal nodes represent features and terminal nodes correspond to outcomes. As a meta-classifier, the random forest relies on an ensemble of unpruned trees [18]. These trees are generated by randomly selecting N features, where N is the square root of the total number of features. The construction of the random forest involves bootstrapping, where training samples are selected with replacement from the original dataset. On average, each tree is trained on approximately two-thirds of the dataset, while the out-of-bag (OOB) samples are used for evaluating tree performance. The OOB evaluation also provides insights into feature importance. Feature importance is assessed by permuting each feature across the OOB observations for every tree and estimating the resulting changes in prediction error. If the accuracy of the new model significantly deviates from the original model, it indicates the importance of the corresponding feature. To obtain a normalized measure of variable importance, the ensemble average of this measure is divided by the overall standard deviation of the ensemble. Classification is accomplished by aggregating the majority votes from the ensemble of generated trees. The ensemble nature of the random forest method mitigates the risk of overfitting training datasets, which is a notable drawback of single decision trees. Random forest demonstrates superior performance compared to individual tree algorithms [19].

2.3.2. Support Vector Machine algorithm

The Support Vector Machine algorithm, illustrated in Fig. 3, is a prominent and powerful supervised machine learning approach widely employed for regression tasks. SVMs are data-driven models rooted in the concept of structural risk minimization (SRM) [20]. SRM aims to simultaneously minimize empirical error and model complexity, thereby enhancing the generalization capability of classification and regression problems. SVMs have been extensively validated in various environmental research domains. For instance, Khalil et al. [21] utilized SVM to analyze the spatial distribution characteristics of groundwater in an agriculture-dominated watershed. SVMs have also found application in fields such as streamflow forecasting, water level prediction in lakes, and soil moisture prediction [22,23]. SVM models can be categorized into two

types linear support vector regression and nonlinear support vector regression. In this research, the nonlinear support vector regression mathematical model was adopted for model development. Mathematically, it can be expressed by Eq. (1):

$$f(X_i) = \sum_{i}^{N} W_i \varphi(X_i) + b \quad (1)$$

where W_i and b are the parameters of the linear support vector regression function, and $\phi(X_i)$ represents the nonlinear mapping function. Various kernel functions, including linear, polynomial, sigmoid, and radial basis function, were tested, and it was determined that the radial basis function yielded the optimal fit for predicting circulating water quality in this research. In addition, for the key model parameters, the optimal parameter sets of the cost constant (C), the radius of insensitive tube (ε), and the scale parameter for stable performance of model (σ) were determined by the optimization algorithm.

2.4. Modeling construction

2.4.1. Input data preparation

The current research employed the Random Forest algorithm and Support Vector Machine algorithm models to forecast the calcium concentration in the circulating seawater flow. The architectural diagrams of the RF and SVM models are depicted in Fig. 4. The entire dataset was partitioned into two distinct subsets the training dataset and the validation dataset. For model training, 80% of the data was allocated to the training dataset, while the remaining 20% was reserved for validation. Prior to training and validation, all data underwent normalization using the MinMaxScaler method to ensure that they ranged from 0 to 1, except for the date parameter. Specific processing techniques were applied to the date parameter to enhance its representation of seasonality. Subsequently, the normalized data were employed as both input and output data for the RF and SVM models. The optimal model parameters for these two models were determined using a global optimization algorithm

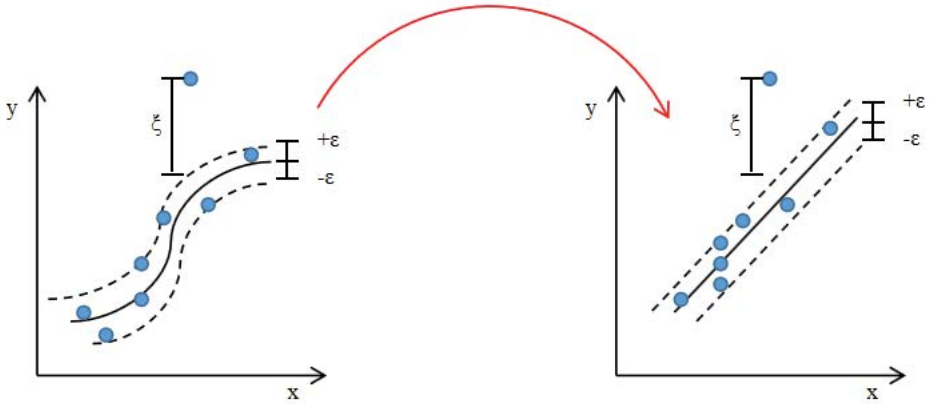

Fig. 3. Illustration of general conceptual model structure for Support Vector Machine algorithm.

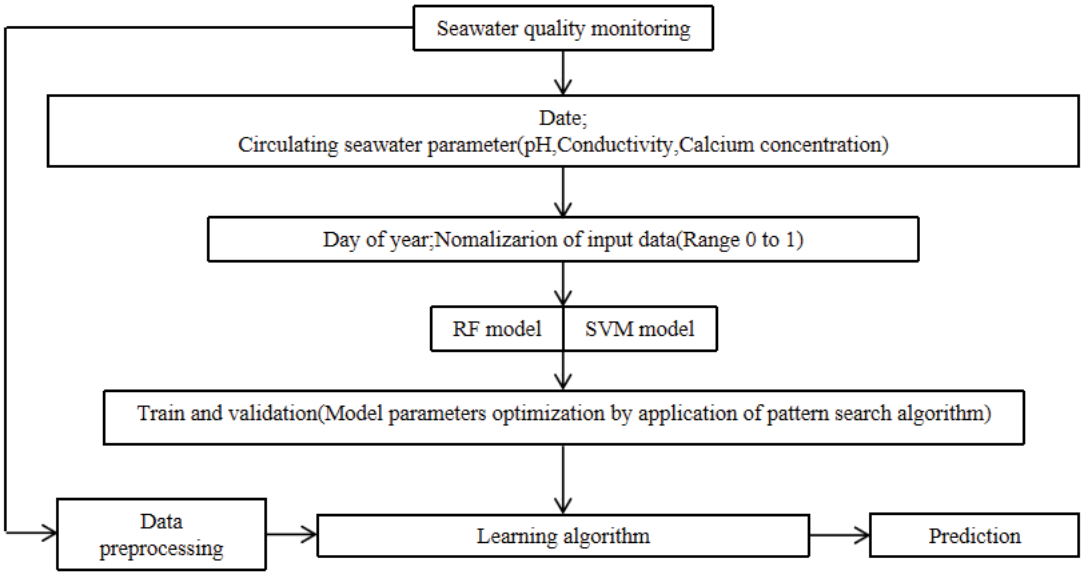

Fig. 4. Logical flow for two machine learning models.

tailored to each specific model. Following the determination of the model parameters, the calcium concentration was predicted using the RF and SVM models, and the predicted values were compared against the measured values to assess the prediction performance.

2.4.2. Model parameter optimization

The selection of appropriate parameter values for the Random Forest algorithm and Support Vector Machine algorithm models is essential for achieving optimal learning and prediction accuracy. Traditionally, these values are determined through trial and error or by referencing prior studies. In this research, we utilized the grid search algorithm to identify the optimal parameter values for the RF and SVM models. The initial ranges for each parameter were carefully chosen based on pre-training procedures, enabling a comprehensive exploration of the parameter space during the optimization process [24–26].

2.4.3. Assessment of model performance

The objective of this research was to forecast the calcium concentration in a closed-cycle seawater cooling system, a task that presents a time-dependent challenge owing to the fluctuating calcium concentration and concentration rate of the make-up seawater across different seasons. To tackle this issue, a regression-based approach was adopted for time series prediction. The conceptual model utilized in this research is illustrated in Fig. 4. For model training and evaluation, the loss function was employed:

$$\text{LOSS} = \sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \tag{2}$$

where y_i is measured value at time i, and $\hat{y}_i$ is predicted value at time i.

All code implemented in this research was written in the Python programming language. The computational experiments were conducted on a computing system equipped with an Intel Core i7-6700HQ CPU and 8GB of RAM. The selection of appropriate criteria to assess the performance of the machine learning models (RF and SVM) is vital for validating their effectiveness.

In this research, two commonly used metrics, namely the root mean square error (RMSE) and the coefficient of determination (R^2), were employed to evaluate the performance of the models by comparing the predicted values to the measured values. The RMSE provides comprehensive information about the predictive capabilities of the models by quantifying the accuracy of the predictions. It is computed by squaring the errors and taking the square root of the average. This process yields a non-negative, real-valued measure that inherently expresses the average magnitude of prediction errors. The utilization of RMSE offers several distinct advantages. Firstly, it provides a more robust evaluation of model performance than simpler metrics like the mean absolute error (MAE) due to its sensitivity to outliers, squared errors amplify the impact of larger deviations, providing a balanced perspective on the model's performance. Additionally, RMSE is readily interpretable as it is measured in the same units as the dependent variable, facilitating a more intuitive comprehension of the magnitude of prediction errors. Furthermore, RMSE is conducive to mathematical operations, making it suitable for analytical comparisons and optimizations. Therefore, the RMSE stands as an indispensable measure in the field of predictive modeling, enabling the quantification of predictive accuracy and aiding in the selection and refinement of models for a diverse range of applications. The RMSE is defined by Eq. (3):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{N}} \tag{3}$$

R-squared (R^2), also known as the coefficient of determination, is a statistical measure used in machine learning to evaluate the quality of a regression model. This principle plays a pivotal role in quantifying the proportion of variance in the dependent variable that can be explained by the independent variables included in the model. R^2 is typically calculated as the ratio of the explained variance to the total variance, with values ranging from 0 to 1. In the context of evaluating model bias, the R^2 principle serves as a valuable tool to measure the degree to which the model captures the underlying relationships between variables and to assess the presence of systematic errors or biases. A higher R^2 value indicates that the model accounts for a larger portion of the variance in the dependent variable, suggesting a better fit, while a lower R^2 value may signal potential bias, as it implies that the model inadequately explains the variance. The advantages of employing the R^2 principle in model bias evaluation are twofold. First, R^2 provides a quantitative and interpretable metric for assessing model performance, enabling researchers to compare different models and determine whether any observed biases are statistically significant. Second, R^2 facilitates the identification of potential areas for model improvement, as a lower R^2 may highlight the need for additional independent variables or model refinement to mitigate bias. The R^2 principle is a vital method for evaluating model bias due to its ability to quantitatively measure model fit and its capacity to inform model refinement and enhancement as Eq. (4):

$$R^2 = \frac{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2 - \sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2} \tag{4}$$

where y_i is measured value at time i, $\hat{y}_i$ is mean of y_i, (i 1, …, N) and $\overline{y}_i$ is predicted value at time i.

2.5. Variable sensitivity analysis

Sensitivity analysis is a valuable tool utilized in modeling to investigate the influence of varying independent variables on a specific dependent variable under predefined conditions. It has proven particularly useful in researching and analyzing "Black Box Processes," where the relationship between inputs and outputs is not readily transparent. Sensitivity analysis serves multiple purposes, including

(i) understanding the input–output relationship, (ii) assessing the contribution of uncertainties in structural model parameters to overall output variability, (iii) identifying influential parameters that significantly affect output magnitudes, and (iv) guiding future experimental designs [27,28]. For model developers, sensitivity analysis provides insights into the model structure and uncertainties associated with input parameters, enabling model refinement and increasing confidence. Particularly in complex models, sensitivity analysis allows focusing on critical parameters that drive model outputs. In this research, the sensitivity of each input parameter was quantified using sensitivity indices. The Sobol sensitivity analysis function implemented in the Python software was employed to conduct the sensitivity analysis. By sampling all parameters using this method, any observed changes in the output values can be clearly attributed to the modified inputs. The Sobol sensitivity indices are defined by Eq. (5):

$$S_{i_1,\ldots,i_s} = \frac{D_{i_1,\ldots,i_s}}{D} \tag{5}$$

where D is the variance of model function output whose sensitivity to the input parameters. S_{ij} is used to compute the second-order contribution from interaction between i-th and j-th parameters. $D_{i_1\ldots i_s}$ is the partial variance corresponding to that subset of parameters.

3. Results and discussion

3.1. Seawater quality monitoring

The dataset utilized in this research comprises daily measurements of three seawater quality parameters, namely pH, conductivity, and calcium concentration, collected over a duration of 7 y. Table 1 presents the daily data for these parameters in the closed-cycle seawater system. The calcium concentration in the closed-cycle seawater serves as a critical parameter for evaluating seawater quality. Alongside the measured pH and conductivity data, the date variable was included as an input parameter, accessible online, to construct the machine learning model.

Seawater, a complex mixture of various dissolved salts and minerals such as calcium (Ca^{2+}), magnesium (Mg^{2+}) ions and so on. Hardness in seawater refers to the concentration of these divalent cations, predominantly Ca^{2+} and Mg^{2+}, and is often expressed in calcium carbonate equivalent units ($CaCO_3$). Seawater scaling involve interconnected nature of seawater quality parameters. Fig. 5 shows cases box plots, which offer statistical summaries of the

Table 1
7 y measured data of circulating seawater quality variables in the closed-cycle seawater cooling system

	pH	Conductivity (mS/cm)	Calcium (mg/L)	Magnesium (mg/L)	Alkalinity (mg/L)
Max.	8.78	64.00	651.15	2,250.87	245.45
Min.	7.85	39.80	366.49	849.78	82.35
Mean	8.27	51.96	508.80	1,567.86	139.40
Standard deviation	0.16	4.70	49.35	216.51	32.95

*Alkalinity (calculated as $CaCO_3$) M alkali (mg/L).

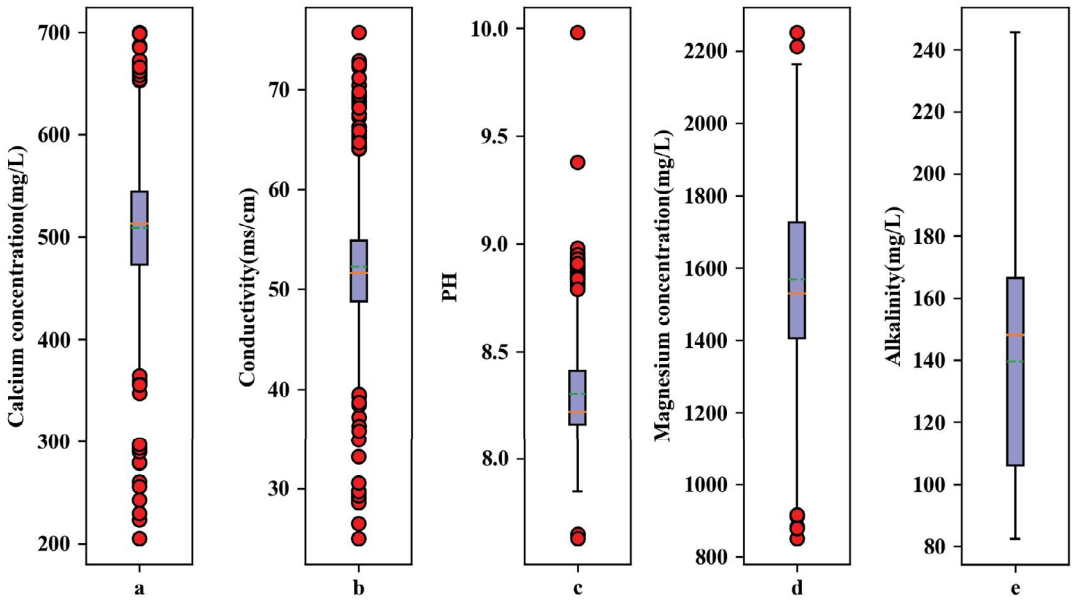

Fig. 5. Basic statistics analysis of the measured circulating seawater quality data.

measured circulating seawater quality variables obtained from Table 1. To ensure data integrity, the five-number summary method was employed to detect and eliminate outliers. The analysis reveals that the calcium concentration in the closed-cycle seawater ranges from 366.49 to 651.15 mg/L, with a mean value of 508.80 mg/L and a standard deviation of 49.35 mg/L. This demonstrates significant fluctuations in the system during the operation of the seawater circulating cooling system, attributable to variations in makeup water quality, sewage discharge, and system evaporation, with the maximum value being 1.78 times the minimum value. Consequently, if a fixed-scale inhibitor dosing scheme is employed, it becomes challenging to adapt to the dynamic operating conditions of seawater circulating water, leading to excessive chemical dosing and increased operational costs. Fig. 6 shows the monthly sampling total hardness variation curve of circulating seawater over a year. The total hardness values in the circulating seawater of the project range from 65.60 to 92.24 mmol/L, which varies accordingly with month. Understanding the complex relationship between these ions in seawater is essential for effective scaling management in various applications. Limited by current technological conditions, real-time online detection of magnesium and total hardness is still quite difficult. Moreover, this article focus on adopting a relatively simple method to predict circulating seawater key scaling parameter calcium concentration. Due to the difficulty in achieving online detection of magnesium ions and total hardness, they are not included in the models features. The effects of magnesium ions and hardness on the models are summarized in other features.

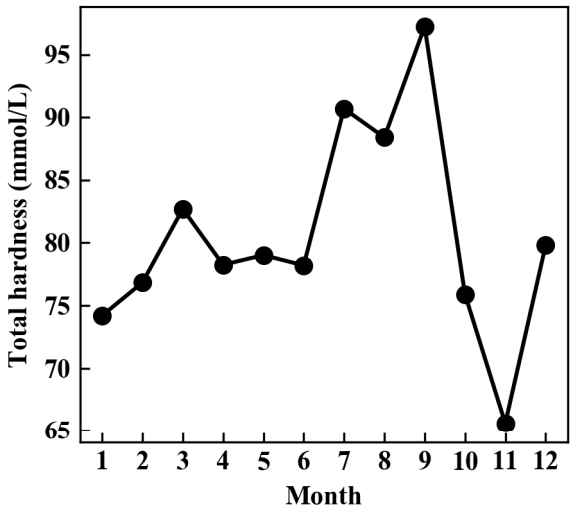

Fig. 6. Monthly sampling total hardness variation curve of circulating seawater over a year.

3.2. Training and validation of models

In order to ascertain the optimal model for predicting circulating seawater calcium concentration in this research, a range of random forest models and support vector machine models were constructed and validated. For the RF models, the maximum depth of the decision tree was adjusted to prevent overfitting, while the number of estimators played a crucial role in controlling the model's performance. In the case of SVM, the radial basis function kernel function was utilized in the transformation layer. Moreover, selecting an appropriate number of nodes for the hidden layers of SVM was essential to avoid overfitting. The grid search algorithm was employed to determine the optimal parameters for both the RF and SVM models. The optimal parameters for the RF and SVM models, obtained through the grid search algorithm, are presented in Table 2.

3.3. Model test

The calcium concentration values of the circulating from the closed-cycle seawater cooling system, obtained from both observed measurements and the predictions generated by the machine learning models (RF and SVM), were compared. The regression model plots, illustrating the training and validation datasets for both RF and SVM, are presented in Fig. 7a and b. It is evident that both models demonstrate a strong fit with the observed data. According to the results presented in Table 2, the RF model achieved high coefficient of determination (R^2) values of 0.94 and 0.93 for the training and validation datasets, respectively. The corresponding RMSE values for the RF model were 11.90 and 12.64. Conversely, the SVM model exhibited R^2 values of 0.82 and 0.80 for the training and validation datasets, respectively, with RMSE values of 21.41 and 20.18. In terms of R^2 and RMSE, the RF model outperformed the SVM model slightly, demonstrating superior predictive performance.

In order to thoroughly evaluate the adequacy of the RF and SVM models, we conducted a comprehensive analysis of their fitness by examining the relative error, as illustrated in Fig. 8a and b. The relative error plots for both the RF and SVM models during the training and validation datasets indicate that the relative error remains consistently below 10%. However, it is important to note

Table 2
Comparison of the Random Forest algorithm model and Support Vector Machine algorithm performances for prediction of calcium concentration

Model	Model parameters	R^2		RMSE	
		Tr	Va	Tr	Va
RF	{'min_samples_split': 2}{'min_samples_leaf': 1}{'max_depth': 17}{'n_estimators': 900}	0.94	0.93	11.90	12.64
SVM	Svr_c:1,svr_gamma:1	0.80	0.82	21.41	20.18

Tr: Training dataset; Va: Validation dataset.

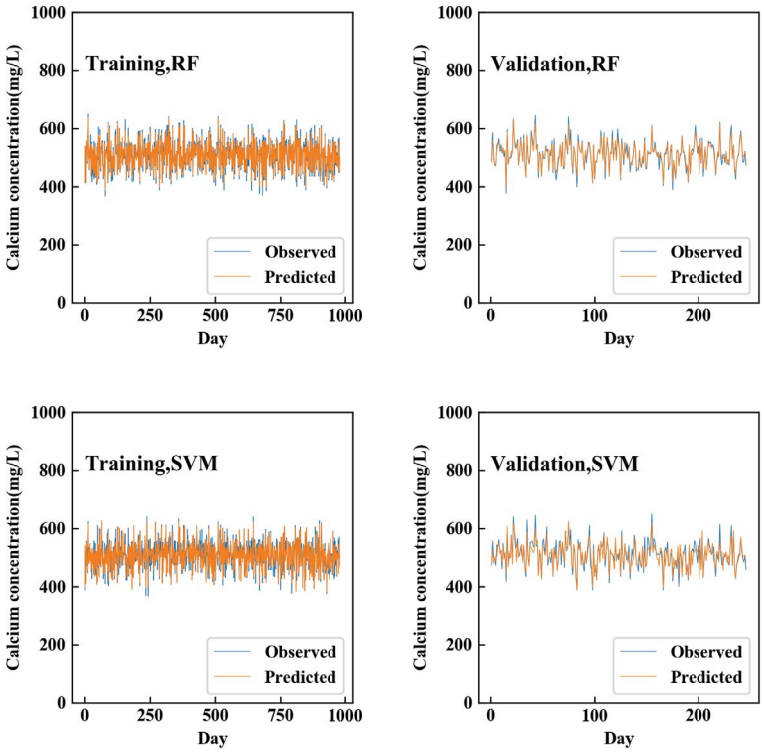

Fig. 7. Comparison of the modeled and observed calcium concentration from the closed-cycle seawater cooling system training and validation datasets using RF model and SVM model.

that certain data points exhibit notable prediction errors, reaching approximately 10%, particularly within the calcium ion concentration range of 300–400 mg/L, as evident in both the training and validation datasets. This observation can be attributed to several underlying factors. Firstly, when the calcium ion concentration falls within this range, the system's concentration cycles tend to be low, introducing complexities in the behavior of calcium ions. Moreover, unconventional influencing factors, such as precipitation, further contribute to the nonlinearity among various water quality parameters. Balabin and Lomakina [29] have previously emphasized the potential of higher nonlinear disturbances to compromise model accuracy in certain machine learning models. Thus, for future modeling endeavors, it is recommended to consider the incorporation of climate factors, including precipitation, as additional features to enhance the predictive accuracy of the model. Despite these challenges, the RF and SVM models exhibited satisfactory modeling accuracy, highlighting their potential for future predictions of circulating seawater calcium concentration.

3.4. Variable sensitivity analysis

Table 3 provides a comprehensive summary of the sensitivity rankings for the input parameters influencing the prediction of circulating seawater calcium concentration using Sobol sensitivity analysis method. The results highlight the significant role played by both spatial and temporal variables in shaping the model's predictive performance. In the case of the RF model, conductivity emerges as the most influential parameter, followed by date and pH. Conductivity exhibits a first-order sensitivity, indicating its direct impact on the prediction. However, date and pH do not exhibit a first-order effect. Notably, the total order index of conductivity and pH surpasses the first-order index, suggesting the possibility of higher-order interactions. Furthermore, the second-order index indicates a weak interactivity between conductivity, pH, and date. On the other hand, for the SVM model, pH and conductivity emerge as the two most significant parameters. Similarly, the total order index exceeds the first-order index, indicating the potential for higher-order interactions. Specifically, high-order interactions are likely to occur between pH-conductivity and date-conductivity. These findings shed light on the intricate relationships between the input parameters and the prediction of circulating seawater calcium concentration, emphasizing the need to consider both spatial and temporal factors in future modeling efforts. In the context of a closed-cycle seawater cooling system, conductivity emerges as the most crucial parameter for predicting calcium ion concentration, serving as an indicator of the relative stability of seawater quality. Therefore, it is reasonable to consider conductivity as the most significant parameter for the machine learning models employed in this research to predict circulating seawater calcium concentration. Moreover, date also plays a significant role as an input parameter, directly influencing the seasonal variations in seawater quality. Based on the characteristics of the closed-cycle seawater cooling system, the RF model appears to offer a more suitable approach compared to SVM. This is because the RF model captures the physical relations more effectively, making it a more reliable choice for managing the impact

Table 3
Sensitivity rank of input variables in RF model and SVM model using Sobol sensitivity analysis method

	Variable	ST	ST_conf	S1	S1_conf		S2	S2_conf
RF	Conductivity	0.879241	0.072107	0.716750	0.062699	pH-Conductivity	0.097877	0.053854
	pH	0.017515	0.023172	0.021943	0.037859	Date-Conductivity	0.062365	0.043922
	Date	0.153901	0.017515	0.019231	0.045539	Date-pH	0.056753	0.045539
SVM	Conductivity	0.874688	0.278289	0.009485	0.101294	pH-Conductivity	0.325847	0.335684
	pH	0.926215	0.301642	0.052405	0.121983	Date-Conductivity	0.076679	0.107361
	Date	0.520664	0.210845	0.019458	0.049106	Date-pH	–0.007330	0.098905

*First-order index: measure the contribution of single model input to output variance.
Second-order index: measure the contribution of the interaction between the inputs of two models to the output variance.
Total order index: measures the contribution of model input to output variance, including first order and higher order.
_conf: corresponding confidence interval, the confidence level is 95%.

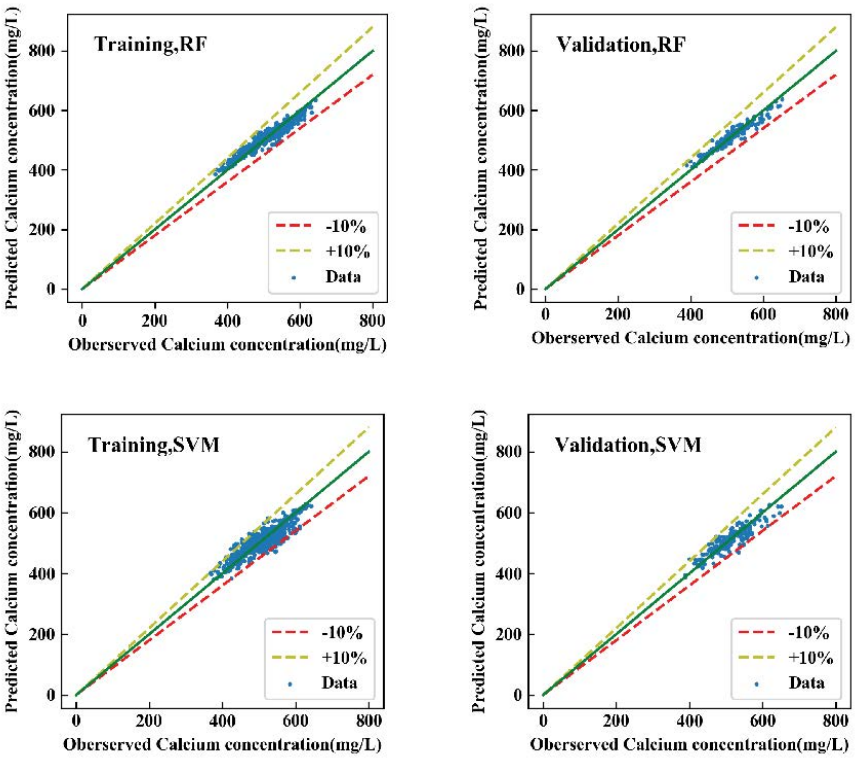

Fig. 8. Plot of the predicted calcium concentration relative error of Random Forest algorithm model and Support Vector Machine algorithm from the closed-cycle seawater cooling system training and validation datasets.

of high calcium concentration by adjusting the parameters that are closely associated with the system's physical characteristics. Although machine learning models do not need to represent the complete physical meaning behind the input and output variables, the sensitivity analysis results reaffirm the importance of conductivity, as it consistently ranked highest among the parameters in the RF model. In contrast, the SVM model exhibited pH as the highest-ranking parameter, which may not possess a direct physical interpretation. On the other hand, the RF model provided more acceptable results due to its consideration of the relationship between conductivity and date, as well as the additional influence of factors such as ionic strength and flocs based on seasonal variations [30]. Furthermore, the impact values presented in Table 3 demonstrated minimal variation across all variables in the SVM model. From a process control perspective, the RF model demonstrated greater reliability and reasonableness compared to the SVM model.

4. Conclusion

The primary objective of this research is to develop two robust machine learning models, specifically random forest and support vector machine, for accurate prediction of calcium concentration in the circulating seawater

of a closed-cycle seawater cooling system. The prediction results aim to establish a basis for enhanced management of system operations. The models utilize daily data on seawater quality, encompassing parameters such as pH, conductivity, and date, as inputs. Both models effectively forecasted the calcium concentration in the circulating seawater within 1-d intervals. Key performance metrics, including the coefficient of determination (R^2) and RMSE, indicate slightly superior performance of the RF model compared to the SVM model. In addition, the RF model emerges as a more reasonable and dependable option in constructing decision-making models and facilitating process control in closed-cycle seawater cooling systems. The predictive model presented herein offers an invaluable tool for the accurate estimation of calcium ion concentrations, contributing profoundly to the enhanced comprehension and control of critical parameters within these cooling systems. This precision is instrumental in mitigating operational inefficiencies, reducing maintenance costs, and ensuring the optimal functioning of such systems. In essence, this research underscores the meaningful role of machine learning in advancing the efficacy and sustainability of seawater cooling systems. It further highlights the vital impact of real-time, data-driven decision support systems, which have the potential to revolutionize the management of these systems, ultimately promoting environmental stewardship and resource conservation. Thus, the outcomes of this investigation signify a pioneering step towards the practical application of machine learning in the optimization of seawater cooling systems, underlining the research's tangible implications and its promise in the ongoing quest for sustainability and efficiency in industrial cooling operations.

As a step forward, future directions should encompass a comprehensive comparative analysis with moderately time-consuming machine learning methods. This comparative evaluation is poised to furnish essential insights into the efficacy and efficiency of the machine learning models deployed in this study. Such an approach can facilitate a deeper understanding of the predictive capabilities and computational demands of various models, ultimately guiding the selection of the most suitable approach for seawater parameter modeling. The investigation's outcomes not only underscore its immediate implications but also establish a promising trajectory for further research and application in the domain of seawater quality monitoring and industrial process optimization.

Acknowledgements

This research was supported financially by the special funds for foundation research of central public welfare research institutes (K-JBYWF-2021-QR02 and K-JBYWF-2021-T03).

References

[1] T. Oki, S. Kanae, Global hydrological cycles and world water resources retentate, Science, 313 (2006) 1068–1072.
[2] D.M. Nester, Salt water cooling tower, Chem. Eng. Prog., 67 (1971) 7.
[3] National Seawater Utilization Report, Ministry of Natural Resources, Beijing, 2021.
[4] R. Daniel, J.F. Casanueva, N. Enrique, Assessment of the antifouling effect of five different treatment strategies on a seawater cooling system, Appl. Therm. Eng., 85 (2015) 124–134.
[5] L. Zhang, D.A. Dzombak, Challenges and Strategies for the Use of Saline Water as Cooling Water in Power Plant Cooling Systems, Carnegie Mellon University & National Energy Technology Laboratory, Pennsylvania, 2010.
[6] H. Istepanian, Monitoring of sea water chemical treatment for cooling system in power utilities–the challenges, Meas. Control, 41 (2008) 54–58.
[7] R. Stoodley, R. Jose, R. Nuñez, T. Bartz, Field and in-lab determination of Ca^{2+} in seawater, J. Chem. Educ., 91 (2014) 1954–1957.
[8] C.J. Zhu, H.Q. Shao, B.Q. Ma, H. Li, L. Yang, GB/T15452-2009 Industrial Closed-Cycle Cooling Water-Determination of Calcium and Magnesium-EDTA Titration Method, General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, China National Standardization Administration, Beijing, 2009.
[9] H. He, Y. Li, S. Wang, Q. Ma, Y. Pan, A high precision method for calcium determination in seawater using ion chromatography, Front. Mar. Sci., 7 (2020) 1–11, doi: 10.3389/fmars.2020.00231.
[10] M. Meléndez, E.P. Nesterenko, P.N. Nesterenko, J.E. Corredor, Direct chromatographic separation and determination of calcium and magnesium in seawater and sediment porewaters, Limnol. Oceanogr. Methods, 11 (2013) 466–474.
[11] M. Whitfield, J.V. Leyendekkers, Liquid ion-exchange electrodes as end-point detectors in complexiметric titrations. determination of calcium and magnesium in the presence of sodium: theoretical considerations, Anal. Chim. Acta, 45 (1969) 383–398.
[12] J.-J. Zhu, P.R. Anderson, Performance evaluation of the ISMLR package for predicting the next day's influent wastewater flowrate at Kirie WRP, Water Sci. Technol., 80 (2019) 695–706.
[13] H. Haimi, M. Mulas, F. Corona, R. Vahala, Data-derived soft-sensors for biological wastewater treatment plants: an overview, Environ. Modell. Software, 47 (2013) 88–107.
[14] J.-J. Zhu, L. Kang, P.R. Anderson, Predicting influent biochemical oxygen demand: balancing energy demand and risk management, Water Res., 128 (2018) 304–313.
[15] Z.F. Wang, Y. Man, Y.S. Hu, J.G. Li, M.N. Hong, P. Cui, A deep learning based dynamic COD prediction model for urban sewage, Environ. Sci. Water Res. Technol., 5 (2019) 2210–2218.
[16] K.B. Newhart, R.W. Holloway, A.S. Hering, T.Y. Cath, Data-driven performance analyses of wastewater treatment plants: a review, Water Res., 157 (2019) 498–513.
[17] P. Agrawal, A. Sinha, S. Kumar, A. Agarwal, A. Banerjee, V. Govind Kumar Villuri, C.S. Rao Annavarapu, R. Dwivedi, V. Vardhan Reddy Dera, J. Sinha, S. Pasupuleti, Exploring artificial intelligence techniques for groundwater quality assessment, Water, 13 (2021) 1172, doi: 10.3390/w13091172.
[18] L. Breiman, Random forests, Mach. Learn, 45 (2001) 5–32.
[19] V.N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Networks, 10 (1999) 988–999.
[20] H. Yoon, S.-C. Jun, Y. Hyun, G.-O. Bae, K.-K. Lee, A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer, J. Hydrol., 396 (2011) 128–138.
[21] A. Khalil, M.N. Almasri, M. McKee, J.J. Kaluarachchi, Applicability of statistical learning algorithms in groundwater quality modeling, Water Resour. Res., 41 (2005) W05010, doi: 10.1029/2004WR003608.
[22] M.S. Khan, P. Coulibaly, Application of support vector machine in lake water level prediction, J. Hydrol. Eng., 11 (2006) 199–205.
[23] S.Y. Liong, C. Sivapragasam, Flood stage forecasting with support vector machines, J. Am. Water Resour. Assoc., 38 (2002) 173–186.
[24] R.M. Lewis, V. Torczon, A globally convergent augmented Lagrangian grid search algorithm for optimization with general constraints and simple bounds, Siam J. Optim., 12 (2002) 1075–1089.

[25] K.H. Cho, J.-H. Kang, S.J. Ki, Y. Park, S.M. Cha, J.H. Kim, Determination of the optimal parameters in regression models for the prediction of chlorophyll-a: a case study of the Yeongsan Reservoir, Korea., Sci. Total Environ., 407 (2009) 2536–2545.
[26] W. Wang, Z. Xu, W. Lu, X. Zhang, Determination of the spread parameter in the Gaussian kernel for classification and regression, Neurocomputing, 55 (2003) 643–663.
[27] A. Kiparissides, S.S. Kucherenko, A. Mantalaris, E.N. Pistikopoulos, Global sensitivity analysis challenges in biological systems modeling, Ind. Eng. Chem. Res., 48 (2009) 7168–7180.
[28] A. Mokhtari, H.C. Frey, Sensitivity analysis of a two-dimensional probabilistic risk assessment model using analysis of variance, Risk Anal.: An Int. J., Off. Publ. Soc. Risk Anal., 25 (2010) 1511–1529.
[29] R.M. Balabin, E.I. Lomakina, Support vector machine regression (LS-SVM)—an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data?, Phys. Chem. Chem. Phys., 13 (2011) 11710–11718.
[30] A. Zita, M. Hermansson, Effects of ionic strength on bacterial adhesion and stability of flocs in a wastewater activated sludge system, Appl. Environ. Microbiol., 60 (1994) 3041–3048.