Journal of Computational Design and Engineering, 2025, **12**, 184–203

DOI: 10.1093/jcde/qwaf022
Advance access publication date: 13 March 2025
Research Article

Prediction of train aerodynamic coefficients under diverse shape parameters and yaw angles

Xiaoshuai Huo[1,2,3], Tanghong Liu[1,2,3], Xiaodong Chen[1,2,3], Zhengwei Chen[4,*] and Xinran Wang[1,2,3]

[1]Key Laboratory of Traffic Safety on Track, Ministry of Education, School of Traffic & Transportation Engineering, Central South University, Changsha 410075, China
[2]Joint International Research Laboratory of Key Technology for Rail Traffic Safety, School of Traffic & Transportation Engineering, Central South University, Changsha 410075, China
[3]National & Local Joint Engineering Research Center of Safety Technology for Rail Vehicle, School of Traffic & Transportation Engineering, Central South University, Changsha 410075, China
[4]Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong 999077, China
*Correspondence: zhengwei.chen@polyu.edu.hk

Abstract

Acquiring aerodynamic coefficients of a high-speed train considering its shape parameters and environmental yaw angles typically requires resource-intensive model tests or numerical simulations. To address this issue, this paper proposes an innovative surrogate model approach to cost-efficiently predict the aerodynamic coefficients. Six critical shape variables are chosen to construct a parametric train model, concurrently integrating the yaw angle (0–90°) to generate a sample space using optimal Latin hypercube design. Then, four original regression algorithms [polynomial regression, support vector regression (SVR), least square support vector regression (LSSVR), and Kriging] and three improved regression algorithms (IPSO-SVR, IPSO-LSSVR, and IPSO-Kriging), incorporating an improved particle swarm optimization (IPSO) algorithm with SVR, LSSVR, and Kriging, are introduced to construct surrogate models. Finally, the prediction accuracy, prediction uncertainty and generalization potential of each surrogate model are compared in terms of the side force coefficient (C_s), lift force coefficient (C_l) and rolling moment coefficient (C_m). The results show that the IPSO-Kriging model outperforms the other surrogate models by exhibiting higher prediction accuracy and generalization performance, although the IPSO-LSSVR model provides a better assessment of the prediction uncertainty in the C_l. The absolute percentage error of IPSO-Kriging is within 5% for all test samples, which implies that this model can provide an effective and economical alternative for model tests or computational fluid dynamic simulations to acquire aerodynamic coefficients.

Keywords: Train shape design, aerodynamic coefficients, polynomial regression, support vector regression, Kriging regression

1. Introduction

High-speed trains, as an important form of transportation, have been widely used throughout the world due to their fast, efficient, comfortable, and environmental friendly characteristics (Tian, 2019). When a high-speed train encounters a strong crosswind, its aerodynamic performance and operational stability will deteriorate dramatically, leading to a significant increase in the risk of overturning or derailment (Chen *et al.*, 2022b; Guo *et al.*, 2024; Huo *et al.*, 2023a; Mohebbi & Rezvani, 2018a). Relevant researchers have performed extensive exploratory studies on the above issue (Heleno *et al.*, 2021; Mohebbi *et al.*, 2023; Montenegro *et al.*, 2022; Zeng *et al.*, 2024). They found that the intrinsic geometric features of the train, including the head shape (Hemida & Krajnović, 2010), streamlined head length (Chen *et al.*, 2018), train height (Wang *et al.*, 2023), and bogie structure (Guo *et al.*, 2020), have notable influences on the aerodynamic response of a train subjected to crosswinds. In addition, the external environmental information, such as the incoming flow at different yaw angles, also has an important effect on the aerodynamic properties of trains (Baker, 2013). Therefore, it is essential to comprehend the aerodynamic characteristics of trains resulting from the combination of diverse train shapes and varying yaw angles. This understanding is critical for ensuring the safe operation of trains in daily scenarios.

In recent years, the aerodynamic performances of trains under crosswinds have been studied experimentally by wind tunnel tests (Brambilla *et al.*, 2022; Mohebbi & Rezvani, 2018b) and full-scale measurements (Gao *et al.*, 2021; Liu *et al.*, 2022), and analyzed systematically by computational fluid dynamics (CFD; Chen *et al.*, 2019; Niu *et al.*, 2022). In general, full-scale measurements lack reproducibility and are challenging to implement (Chen *et al.*, 2022a, 2023); wind tunnel tests are expensive and complex to set up; while CFD (Mohebbi *et al.*, 2024; Mohebbi & Rezvani, 2019, 2021), despite its relative simplicity, remains time-consuming and resource-intensive. A substantial effort has been made to obtain the aerodynamic loads of trains with diverse shape dimensions under different incoming flow conditions, for a particular train shape dimension at a specific yaw angle, however, wind tunnel tests or CFD simulations are still required to determine the specific aerodynamic loads.

Given the development of machine learning techniques, proposing surrogate models to predict the aerodynamic loads on trains under different parameters is becoming possible. Regression-type algorithms in machine learning techniques, including polynomial regression (PR), support vector regression (SVR), random forest regression (RFF), Kriging regression, and neural network (NN), are capable of representing complex nonlinear relationships between multiple variables through available

Received: November 19, 2024. **Revised:** February 10, 2025. **Accepted:** February 12, 2025

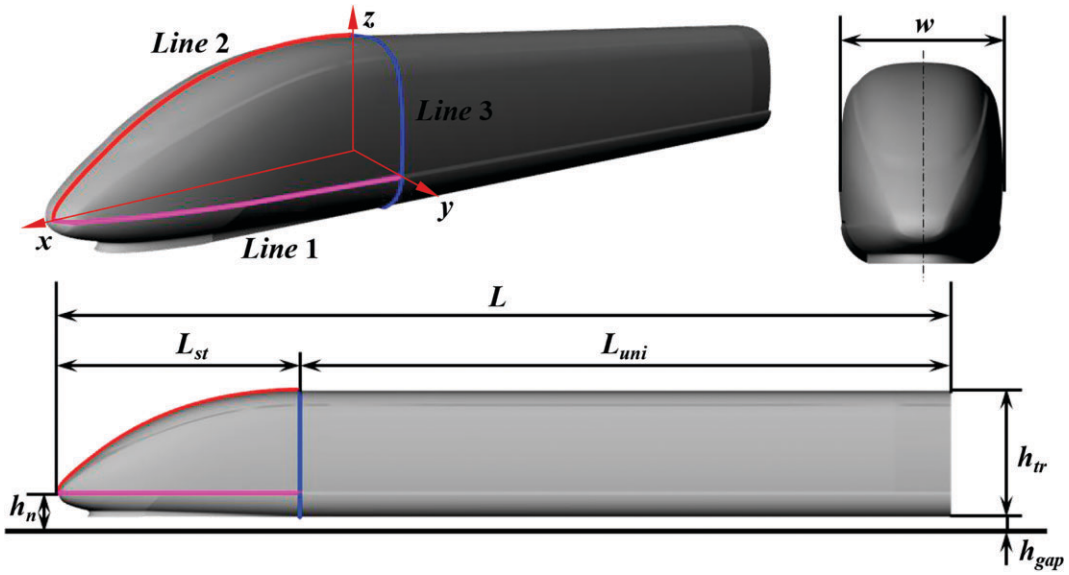

Figure 1: Prototype train model.

data, and they have been successfully applied in several engineering fields (Keane & Voutchkov, 2020; Lee & Kang, 2023; Wu *et al.*, 2023). Nevertheless, the application of regression-type algorithms is still in its preliminary stages in terms of train aerodynamics. To shorten the design cycle of train nose shape, Yao *et al.* (2016) obtained the nonlinear relationship between design variables and aerodynamic drag based on SVR model. Muñoz-Paniagua & García (2019) employed a feedforward NN model combined with genetic algorithm (GA) to optimize the train nose shape in terms of the crosswind stability of a train under crosswind and passing-by scenarios. Kriging surrogate models were constructed by Zhang *et al.* (2018) and Xu *et al.* (2017) to hasten the optimization efficiency of train head design under crosswinds and the streamlined shape of trains without crosswinds, respectively. He *et al.* (2022b) hybridized polynomial response surface (PRS) with radial basis function (RBF) to build a surrogate model, which improves the accuracy of predicting the aerodynamic coefficients of trains running in an open air. To examine the prediction performance of the surrogate model, Zhang *et al.* (2021) constructed different Kriging and NN models using low-dimensional and high-dimensional design variables of train shape at 90° yaw angle. They found that the NN model had higher prediction accuracy compared to the Kriging model under appropriate parameters, especially for high-dimensional variables.

The prediction accuracy of the surrogate model will directly affect the reliability in determining train aerodynamic results. Although a few enhancements have been made to the surrogate model, most of the studies mentioned above have essentially used the original surrogate model to predict train aerodynamic coefficients. Moreover, these researches mainly focus on the aerodynamic loads of trains designed with different nose shape at specific yaw angles, while the aerodynamic loads of trains with the simultaneous changes of global shape parameters and yaw angles are rarely studied. Therefore, this study systematically investigates the effects of diverse shape parameters and varying yaw angles on aerodynamic forces, with the objective of developing appropriate surrogate models for rapidly and accurately predicting train aerodynamic coefficients under multivariate combinations. It is anticipated that the proposed surrogate models can provide an efficient and economical alternative to the traditional model test or CFD simulation for acquiring train aerodynamic coefficients, thereby establishing a robust framework for practical applications in shape design and aerodynamic analysis. The research motivations outlined above determine the scope of this paper, focusing on the following objectives:

- Optimization of hyperparameters for surrogate models to enhance their performance.
- Development of surrogate models to efficiently predict aerodynamic coefficients across varying parameters.

The paper is organized as follows. Section 2 describes the methodology, including parametric geometry model, sampling design, surrogate model algorithm, optimization algorithm, and performance evaluation metrics. CFD simulation is presented in detail in Section 3. This section emphasizes the reliability of the numerical methods. The results and analyses are shown in Section 4, which primarily discusses the data distribution characteristics of train aerodynamic coefficients, performance comparisons of surrogate models, and model generalization capabilities. Section 5 summarizes the main conclusions of this study.

2. Methodology

2.1. Geometric model

To construct a predictive model for aerodynamic coefficients, a parametric train model is required to obtain aerodynamic data. A simplified model of the ICE 3 leading car is employed as the base model due to its popularity worldwide. All attached structures on the train surface, such as pantographs and doorknobs, are ignored to reduce the computational expenses of numerical simulation, and the bogie position is directly replaced by a flat surface, as shown in Figure 1. Considering primary parameters of a train such as nose shape, streamline length, width, and height significantly affect train's aerodynamic characteristics, and the nose tip height and the gap below the train body's also affect the generation of the train's exterior shape, six design parameters are chosen to construct the train profile, including streamlined head length (L_{st}), uniform section length (L_{uni}), nose height (h_n), train height (h_{tr}), the gap between the bottom of the train and the top of the rail (h_{gap}), and train width (w). The value range of train design

Table 1: Variation ranges of train shape parameters.

Design variable	L_{st}	L_{uni}	h_n	h_{tr}	h_{gap}	w
Min (m)	4.0	4.0	0.5	3.0	0.1	2.8
Max (m)	20.0	25.0	1.5	5.0	0.3	4.8

parameters is listed in Table 1, and each parameter has a considerable variation range to obtain a more widely series of train shapes. All models using parametric modeling in this study are three-car-group trains, namely the head vehicle, middle vehicle, and tail vehicle. Three vehicles of one train have identical lengths, varying simultaneously between 20 and 30 m in different cases. The head and tail vehicles have identical streamlined noses and uniform cross-sectional bodies, while the middle vehicle consists solely of a uniform cross-sectional body.

Inspired by the principle of idealized train generation (Chiu & Squire, 1992), an improved function control method is proposed to construct train profiles in terms of train shape parameters. Since the shape of the uniform section remains unchanged, only the outline of the streamlined head is demonstrated in Figure 2. First, the equation of the cross-sectional profile of the idealized train can be written as

$$|y_i|^n + |z_i|^n = c^n, \tag{1}$$

where n is equal to five and uniformly reduced to two towards the nose tip; c follows a semi-elliptical curve with a large diameter of $2L_{st}$ and a small diameter of w; y_i and z_i represent the y-coordinate and z-coordinate on the ith cross-section of idealized train (Chiu & Squire, 1992), respectively.

Second, the contour equations of *Line* 1 and *Line* 2 are obtained by fitting a series of coordinate points, which can be represented by Equations 2 and 3, respectively:

$$y_{max,i} = \frac{0.5w \times (e^{-0.4795x_i} + e^{-1.984x_i} - 2)}{e^{-0.4795L_{st}} + e^{-1.984L_{st}} - 2}, \tag{2}$$

$$z_{max,i} = \frac{(h_{tr} + h_{gap} - h_n) \times (e^{-0.3921x_i} - 1)}{e^{-0.3921L_{st}} - 1}, \tag{3}$$

where, x_i, $y_{max,i}$, and $z_{max,i}$ are the x-coordinate, y-coordinate, and z-coordinate of the ith cross-section, respectively. Equation 2 defines the maximum half-width variation of the streamlined head. Due to the symmetry of the train, the expressions for positive and negative $y_{max,i}$ are identical. Equation 3 determines the change in maximum height above the nose tip. Because the train bottom can be approximated as a plane, the height variation below the nose tip can be considered as a constant equal to h_n–h_{gap}.

Finally, according to the relationship between the y-coordinate (y_i) and z-coordinate (z_i) obtained by Equation 1 and the actual y-coordinate and z-coordinate of the ICE3 leading vehicle and using $y_{max,i}$ and $z_{max,i}$ to adjust, the final y-coordinate and z-coordinate of the ith cross-section can be expressed as

$$y^{\pm}_{real,i} = \pm \left| \frac{y_{max,i} \times \log(|y_i| + 1)}{\log(c_i + 1)} \right|, \tag{4}$$

$$z^{+}_{real,i} = \frac{z_{max,i} \times \log(z_i + 1)}{\log(c_i + 1)}, \tag{5}$$

$$z^{-}_{real,i} = -\frac{h_{gap} - h_n}{{c_i}^2} \times z_i^2 + \frac{2(h_{gap} - h_n)}{c_i} \times z_i, \tag{6}$$

where $y^{\pm}_{real,i}$ and $z^{\pm}_{real,i}$ are the positive or negative y-coordinate and z-coordinate on the ith cross-section of the ICE3 leading vehicle.

A geometric sample is generated by the above functional control equation. Figure 3 compares the shape of a functional train and a prototype train. It can be seen that two models have a very high degree of similarity, in spite of certain deviations in some details. Thus, the improved functional equations enable a construction of train shapes in a design space.

2.2. Sampling design

The spatial sampling is critical to design of experiments. The selection of sample points needs to meet spatial sampling requirements to ensure that the samples are representative of the entire parameter space. If the sample space distribution is inadequate or nonuniform, the surrogate model may not accurately reflect the behavior of the complex system. In this study, a sampling strategy called optimal Latin hypercube design (OLHD) is adopted to generate the sample design points (Jin *et al.*, 2005). In addition to the six train shape variables mentioned above, the yaw angle (β) is chosen as the seventh design variable to construct the sample space. Besides, the yaw angle varies from 0° to 90°. A training dataset consisting of 100 samples is generated through OLHD, and 10 additional different samples are randomly generated as the test dataset. All data are normalized in Figure 4 to better reflect the spatial distribution of design points. Design points numbered 1 to 100 are training samples, while those numbered 101 to 100 are test samples.

2.3. Basic theory of algorithms

In this paper, four regression algorithms including PR, SVR, least square support vector regression (LSSVR), and Kriging are used as surrogate models to predict aerodynamic force coefficients. Additionally, an improved particle swarm optimization (IPSO) algorithm is introduced to seek the optimal parameters of SVR (IPSO-SVR), LSSVR (IPSO-LSSVR), and Kriging (IPSO-Kriging). To simplify the algorithm description, the detailed mathematical derivations of the PR, SVR, LSSVR, and Kriging algorithms are provided in the Appendix, while only the construction of the IPSO is presented.

Particle swarm optimization (PSO) is an evolutionary computational technique that simulates the behavior of a flock of birds foraging for food (Banks *et al.*, 2007). In PSO, each solution is called

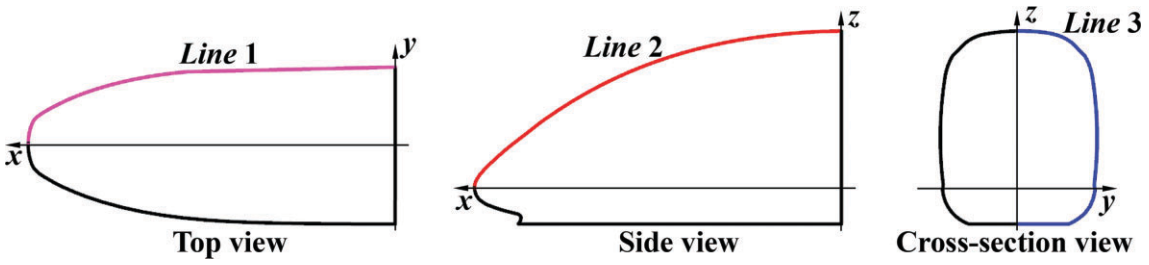

Figure 2: Shape outline curves of streamlined train head.

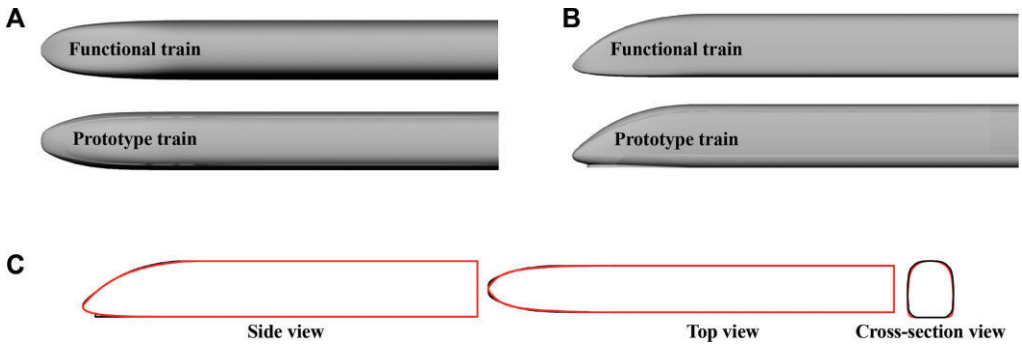

Figure 3: Shape comparison of the functional and prototype trains: (A) top view, (B) side view, and (C) outline curves of both models (The black curve represents the realistic model and the red curve represents the functional model.).

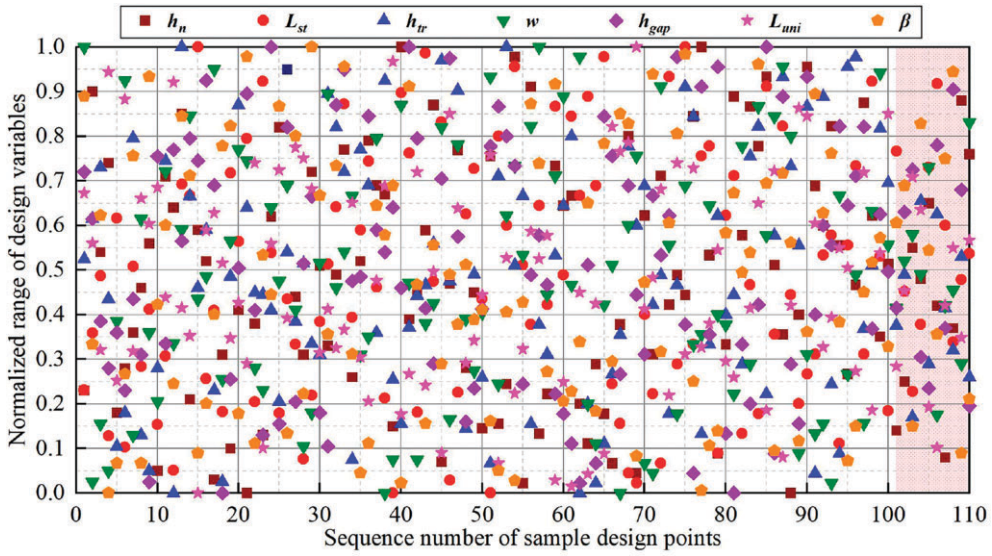

Figure 4: Spatial distribution of design points.

a "particle," and each particle has a position and a velocity. The particle updates its velocity and position by following the best particle in the current population, thus finding the optimal solution in the search space. The optimal solutions found for each particle and all particles in the history are called the individual best position ($P_i^{(t)}$) and the global best position ($P_g^{(t)}$), respectively. For the d-dimensional search space, the position and velocity of the ith particle in generation t can be represented by $x_i^{(t)} = (x_{i1}^{(t)}, x_{i2}^{(t)}, \ldots, x_{id}^{(t)})$ and $v_i^{(t)} = (v_{i1}^{(t)}, v_{i2}^{(t)}, \ldots, v_{id}^{(t)})$. The information update on the velocity and position of each particle in generation $t + 1$ can be expressed as the following equations:

$$v_{id}^{(t+1)} = wv_{id}^{(t)} + c_1 r_1 (P_{id}^{(t)} - x_{id}^{(t)}) + c_2 r_2 (P_{gd}^{(t)} - x_{id}^{(t)}), \tag{7}$$

$$x_{id}^{(t+1)} = x_{id}^{(t)} + v_{id}^{(t+1)}, \tag{8}$$

where w denotes the inertia weight, and a linearly decreasing inertia weight is used here (i.e., $w = w_{max} - (w_{max} - w_{min})t/t_{max}$, $w_{max} = 0.9$, $w_{min} = 0.4$); c_1 and c_2 represent the cognitive and social factors, and $c_1 = c_2 = 2$; r_1 and r_2 are random values ranging from 0 to 1. A detailed explanation of parameters w, c_1, and c_2 can be found in Kennedy (2010).

Traditional PSO tends to fall into local optima resulting in poor solution quality (Naka *et al.*, 2003). Therefore, chaos mapping embedded PSO (called IPSO) is proposed in this study to increase the diversity and randomness of the population, thereby improving the global search capability of the algorithm. The classical and convenient logistic equation is chosen to construct IPSO with the following expression:

$$x_{n+1} = \mu x_n (1 - x_n), 0 < x_0 < 1, \tag{9}$$

where μ ($0 < \mu \leq 4$) represents the control variable, and the system is in a fully chaotic state when $\mu = 4$; x_n represents the chaotic variable, and $x_n \notin \{0.25, 0.5, 0.75\}$. Although Equation 9 has a deterministic representation, its sensitivity to initial values allows chaos mapping to produce chaos sequences with pseudorandomness and unpredictability.

Chaos mapping should be executed when the algorithm falls into a local optimum during iteration. Thus, a benchmark value is needed to make a judgment. In this paper, we choose the mean particle spacing (D_{mean}) as this evaluation index, which is defined as follows:

$$D_{mean} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{d} (p_{ij} - \bar{p}_j)^2}, \tag{10}$$

where N denotes the swarm size; d denotes the dimensionality of the search space; p_{ij} denotes the jth value of the ith particle; and $\bar{p}_j$ denotes the average value of all particles in dimensionality j. A smaller D_{mean} indicates a more concentrated population, and vice versa.

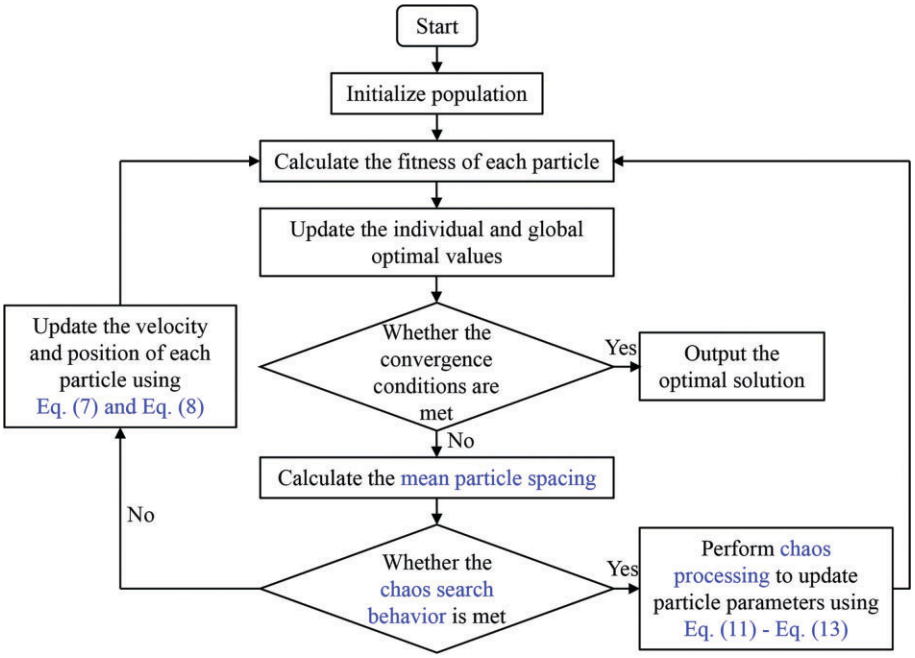

Figure 5: Flow diagram of the IPSO algorithm.

Given a threshold δ, the chaotic search behavior will be performed if D_{mean} is smaller than δ. First, the decision variable $x_{ij}^{(0)}$ of the jth dimension of the ith particle is mapped to chaotic initial variable $cx_{ij}^{(0)}$, and the transformation takes the following form:

$$cx_{ij}^{(0)} = \frac{x_{ij}^{(0)} - x_{min,ij}}{x_{max,ij} - x_{min,ij}}, \tag{11}$$

where $x_{max,ij}$ and $x_{min,ij}$ are the maximum and minimum values of the particle iterations. Then, the next chaotic variable in the chaotic search process can be obtained by Equation 9 in the following form:

$$cx_{ij}^{(k+1)} = 4cx_{ij}^{(k)}(1 - cx_{ij}^{(k)}). \tag{12}$$

Further, the chaotic variable $cx_{ij}^{(k+1)}$ is transformed into the corresponding decision variable $x_{ij}^{(k+1)}$ through the following formula:

$$x_{ij}^{(k+1)} = x_{min,ij} + cx_{ij}^{(k+1)}(x_{max,ij} - x_{min,ij}). \tag{13}$$

Finally, the fitness values of the newly generated particle $x_i^{(k+1)}$ and the current particle are evaluated and compared. If the new solution is better than the current particle, the current particle is replaced by the new solution. Otherwise, it goes directly to the next iteration (i.e., let $k = k + 1$) until the condition terminates.

The flow diagram of the IPSO algorithm is shown in Figure 5. The detailed steps are described below:

Step 1: Initialize the algorithm parameters, including the population size N, the maximum number of iterations t_{max}, the space search dimensionality d, the variable range, the maximum number of chaos search k_{max}, the threshold δ, etc. Randomly generate the initial population.

Step 2: Calculate the fitness value of each particle, and update individual and global best solutions.

Step 3: Determine whether the termination condition is reached, and if so, output the best solution. Otherwise, execute the next step.

Step 4: Judge whether the chaotic search behavior is satisfied, i.e., $D_{mean} < \delta$? If it is satisfied, the position of each particle is updated by chaos mapping based on Equations 11–13. Otherwise, Equations 7 and 8 are used to update the velocity and position of each particle. Then, return to Step 2.

2.4. Performance evaluation of surrogate models

Each surrogate model is compared not only for its accuracy in predicting aerodynamic coefficients, but also for its ability to quantify the prediction uncertainty. The k-fold cross-validation technique is employed to evaluate the model performance. Numerous studies (Hu & Kwok, 2020) have shown that k-fold cross-validation has reasonable variability and helps to avoid overfitting. In addition, k is taken 10 in this paper based on the suggestion of Refaeilzadeh *et al.* (2009). Some classical validation metrics, including mean absolute error (MAE), mean square error (MSE), and model efficiency coefficient (MEC; Wadoux *et al.*, 2018), are used to assess model prediction accuracy. The absolute percentage error (APE) of a single test point and the mean absolute percentage error (MAPE) of all test points are also adopted to evaluate the potential generalization ability of the surrogate model. Moreover, for the performance evaluation of the prediction uncertainty, the prediction interval coverage probability (PICP) is used in this study (Malone *et al.*, 2011). A 95% confidence level is specified to define the prediction bounds encompassing the true but unknown values of times on average. Confidence intervals (CIs) ranging from 5% to 95% are selected to evaluate the model sensitivity by sequentially decreasing the confidence limits. In general, the PICP value close to the corresponding confidence level implies a better performance in terms of the prediction uncertainty.

3. CFD Simulation

3.1. Computational domain and boundary conditions

The size of the computational domain for all numerical cases is the same as 90 H_m × 60 H_m × 20 H_m in Figure 6. Here, H_m is the maximum height in the range of train shape variables. To meet the grid need of the turbulence model, a 1/8 scaled model is used for the simulation analysis. Resultant velocity method is introduced to

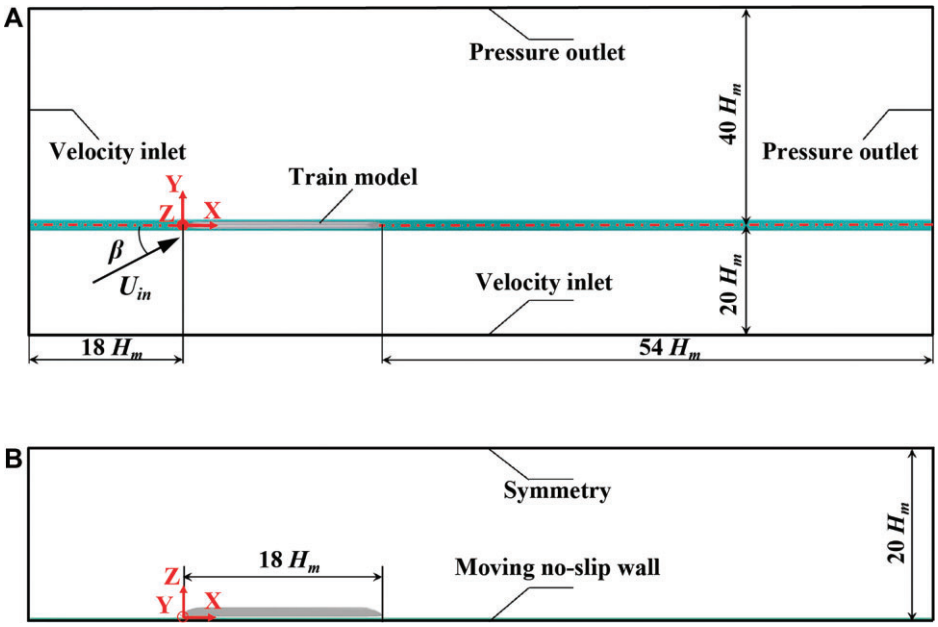

Figure 6: Computational domain: (A) top view and (B) side view.

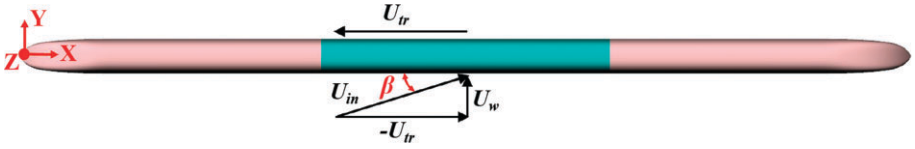

Figure 7: Diagram of resultant wind speed.

achieve different yaw angles in Figure 7. The resultant wind speed (U_{in}) is composed of train speed (U_{tr}) and crosswind speed (U_w), and the desired β can be obtained by changing the magnitude of U_{tr} and U_w. The incoming velocity (i.e., the resultant wind speed) is a fixed value of 60 m/s, and $U_{tr} = U_{in}\cos\beta$ and $U_w = U_{in}\sin\beta$. The upstream boundary and the windward side (WW) boundary are set as velocity inlets with a uniform velocity U_{in}, and the downstream boundary and the leeward side (LW) boundary are defined as pressure outlets with a zero static pressure. The symmetry condition is imposed to the top surface of the computational domain, and no-slip wall condition is applied to the train surface. To reflect the relative motion of the train and the ground, the floor and rail are treated as moving no-slip wall conditions with the train speed. The Reynolds number is equal to 2.57×10^6 considering H_m and U_{in}.

3.2. Numerical method

The aerodynamic forces acting on the train are not completely constant resulting from the inherent instability of turbulent flow. Nevertheless, the time-averaged aerodynamic forces experienced by the train are our primary interest in this study. To balance solution accuracy and computational cost, we use the steady Reynolds-averaged Navier-Stokes (RANS) method, which is commonly employed to solve the train aerodynamic problems (Huo *et al.*, 2023b; Li *et al.*, 2019; Morden *et al.*, 2015; Premoli *et al.*, 2016;). In addition, the shear-stress transport (SST) k-ω turbulence model, which has superior reproducibility for complicated flows with strong separations (Catanzaro *et al.*, 2016; Huo *et al.*, 2023a; Li *et al.*, 2022b; Zamiri & Chung, 2017), is chosen to represent the Reynolds stress. All simulations are conducted on the basis of the pressure-based solver in ANSYS Fluent. The convective and diffusive terms are treated using the second-order upwind scheme in spatial discretization approaches, and the SIMPLE algorithm is adopted to couple the pressure-velocity field. All simulations are conducted for 10 000 iterations, ensuring that the residuals of each equation drop below 10^{-5} to satisfy the convergence criterion.

Dimensionless side force coefficient (C_s), lift force coefficient (C_l) and rolling moment coefficient (C_m) are used for the analysis, and they are defined as follows:

$$C_s = \frac{F_s}{0.5\rho U_{in}^2 S_y}, \tag{14}$$

$$C_l = \frac{F_l}{0.5\rho U_{in}^2 S_z}, \tag{15}$$

$$C_m = \frac{M_x}{0.5\rho U_{in}^2 S_y h_{tr}}, \tag{16}$$

where, F_s, F_l, and M_x represent the side force, lift force, and rolling moment, respectively; ρ means the air density equal to 1.225 kg/m^3; S_y and S_z are the projected areas of the head vehicle in the y- and z-directions, respectively.

3.3. Mesh strategy

The SnappyHexMesh in OpenFOAM is employed to discretize spatial grids, and the majority of the grids follows a 2-by-2 growth pattern based on the defined zone divisions. Four different densities of grids, including coarse grid (1.27×10^7 cells), medium grid (2.58×10^7 cells), fine grid (4.56×10^7 cells), and extra-fine grid (7.25×10^7 cells), are used for grid-independent validation. The C_s and C_l values of head vehicle at 30° yaw angle are compared for the four grids in Figure 8. The results for both C_s and C_l in-

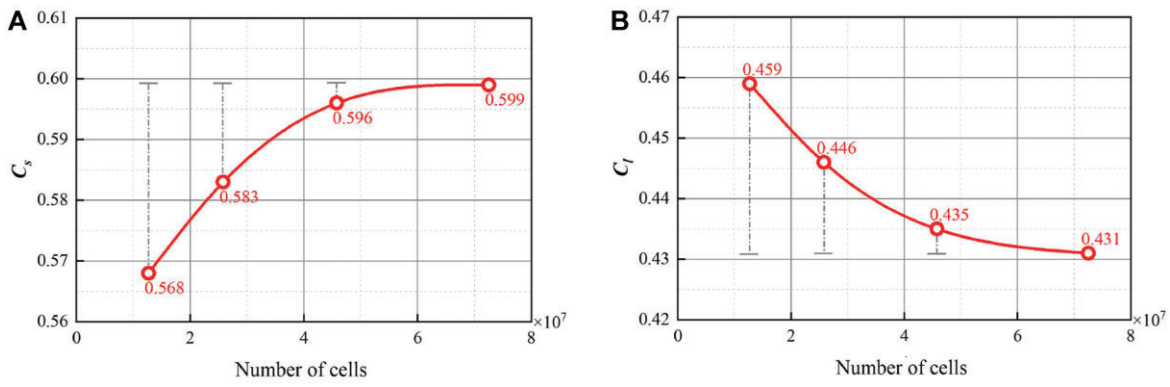

Figure 8: Grid independence test at 30°: (A) C_s of the head vehicle and (B) C_l of the head vehicle.

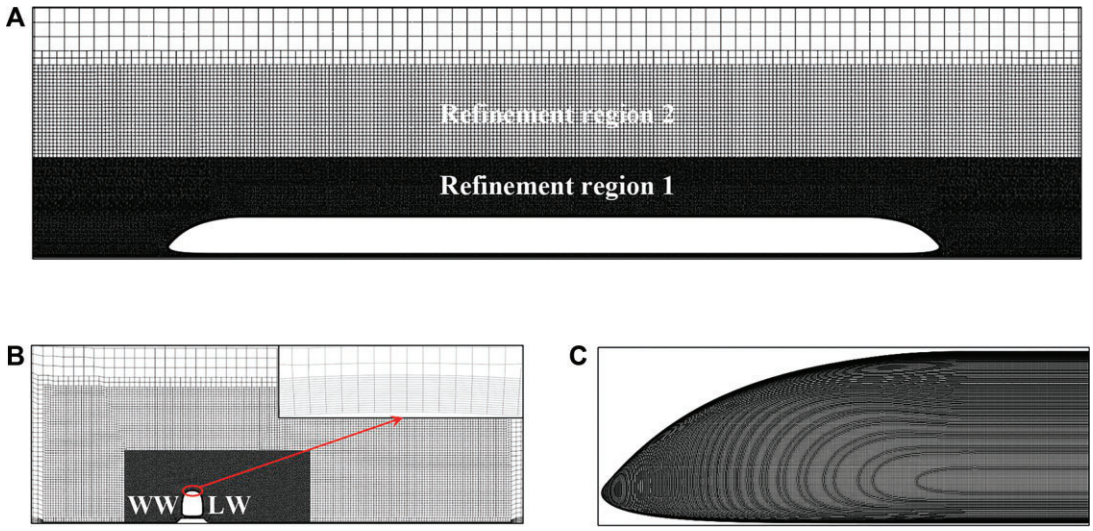

Figure 9: Fine grid distribution: (A) grid on the symmetrical plane along the longitudinal direction of the train, (B) grid on the cross-section of the head vehicle, and (C) grid on the train head surface.

dicate convergence to the fine grid, yielding differences of 0.50% and 0.93% from those of the extra-fine grid, respectively, which means that the fine grid is sufficient to obtain accurate aerodynamic forces. Thus, the fine-grid strategy is chosen for all numerical simulations. The fine grid distribution for the functional model of the original train is shown in Figure 9. Two refinement regions are constructed to better capture the turbulent vortex structures around the train, and the range of refinement on the LW of the train is larger than that on the WW of the train. In addition, sixteen inflation layers are developed from all wall surfaces. Each prism layer has a uniform grid thickness of 0.04 mm, and the value of y^+ ranges from about 1 to 5.

3.4. Numerical verification

In the absence of experimental data for the functional model we constructed, wind tunnel test results of an idealized train (Copley, 1987) are adopted to verify the accuracy of the numerical algorithm. Yaw angles ranging from 20° to 35° are measured in the experiments (Copley, 1987), and we use the results at 35° for validation. In Hemida & Krajnović (2010), the large eddy simulation (LES) of this idealized train is conducted in terms of 35°. Additionally, the Reynolds number for both the test and the LES is 3.7×10^5. Thus, same computational domain and boundary conditions are set up following their methodology, and detailed descriptions can be referred to Hemida & Krajnović (2010). The cross-sectional profile of the idealized train is defined by Equation 1, where $c = 62.5$ mm and $n = 5$. Moreover, n reduced uniformly from 5 to 2 towards the nose tip, and c follows a semi-elliptical curve with a large diameter of 160 mm and a small diameter of 125 mm. The length, width, and height of this idealized train model are 10 D, D, and D ($D = 125$ mm), respectively, as shown in Figure 10A.

The pressure coefficient ($C_p = \frac{P-P_0}{0.5\rho U_{in}^2}$) distributions at the cross-sections $x/D = 2.5$ and $x/D = 6.5$ are presented in Figure 10B and C. The LES results at the cross-section $x/D = 6.5$ reported by Hemida & Krajnović (2010) are also added in Figure 10C for comparison. Overall, the simulation results for both cross-sections exhibit a similar tendency to the experimental data. Moreover, well approximated results are observed for our simulation and LES. However, there is a significant overestimation of the pressure from the simulation underneath the train compared to the test. The reason for this deviation may be that the floor setup in the simulation is different from that in the test. In the test, two identical models are mounted symmetrically on both sides of the floor, whereas in the simulation, only one model is installed on the sidewall. As a result, a stagnation region is formed between the train and the sidewall. This stagnation region induces a larger pressure increase on the streamwise side compared to that of the test. Thus, the numeri-

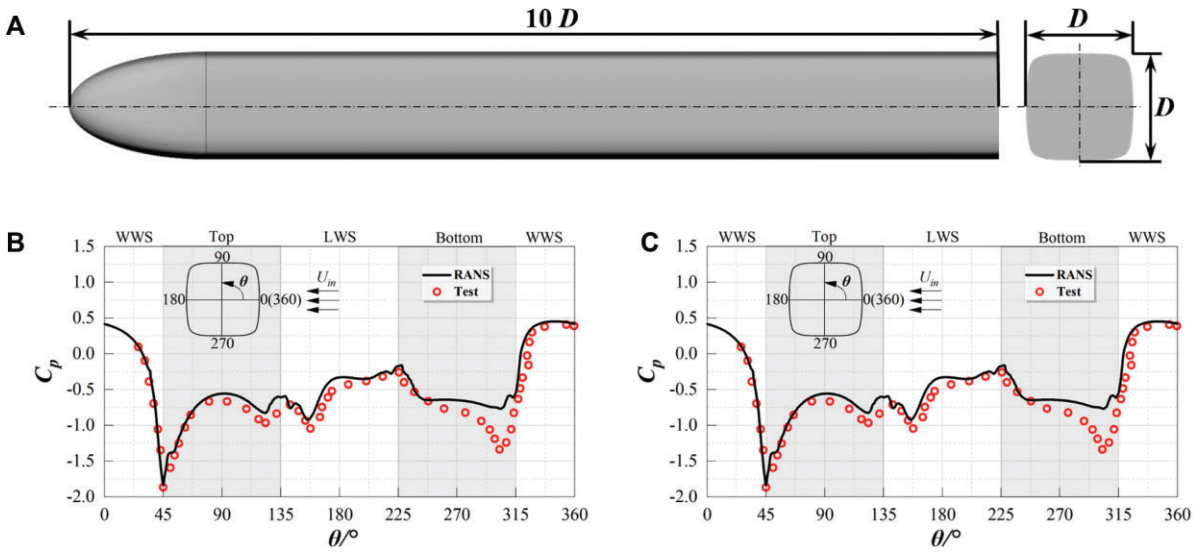

Figure 10: Numerical validation: (A) simulation model of the idealized train; comparison of pressure coefficients from numerical simulations and experimental results at (B) x/D = 2.5 and (C) x/D = 6.5.

cal algorithm in this paper can be considered as an appropriate choice to obtain accurate CFD results.

4. Results and Analyses

4.1. Data distribution characteristics of aerodynamic coefficients

In general, the leading car of a train running under crosswinds is identified as the most critical case with a risk of derailment or overturning (Li *et al.*, 2022a; Huo *et al.*, 2021). Therefore, aerodynamic coefficients of the leading car are solely investigated in this study. To demonstrate the sample data distribution from CFD simulations, a series of scatter plots for the C_s, C_l, and C_m with different train external dimensions of h_n, L_{st}, h_{tr}, w, h_{gap}, L_{uni}, and yaw angle β are presented in Figure 11. All shape design variables and yaw angle are normalized to a range of 0 to 1 for convenient analysis. Blue dots correspond to the training data set in the design of experiments, while the red pentagram markers represent the test data set used to evaluate the final surrogate models in Section 4.3. The data points from Figure 11A to F exhibit relatively random distributions without apparent patterns. However, the aerodynamic coefficients (C_s, C_l, and C_m) increase with the yaw angle in the range from 0 to 0.6 for the normalized β in Figure 11G, while an opposite tendency appears in the remaining range.

Estimated Pearson correlation coefficients (PCCs) are illustrated to further explore the relationship between each variable and the aerodynamic coefficients in Figure 12. Red and blue colors indicate positive and negative correlations between any two variables, respectively. The values in the lower triangular matrix and the color shades of the ellipses in the upper triangular matrix are symmetrically distributed about the diagonal and correspond to each other, and both of them represent the magnitude of the specific correlation coefficients. The correlations between C_s and h_n, L_{st}, h_{gap}, and L_{uni} are extremely weak with the absolute values of the PCCs of less than 0.1, while the h_{tr} and w show relatively bigger correlations with C_s with the absolute values of the PCCs of more than 0.1. Moreover, C_s has positive and negative correlations with h_{tr} and w, respectively. A negative correlation exists between C_l and h_n, h_{gap}, and L_{uni}, while the remaining shape parameters exhibit a remarkably weak positive correlation with C_l due to the PCCs being less than 0.1. The absolute values of the PCCs between C_m and all shape variables are less than 0.1, which implies that the C_m is very weakly correlated with the shape variables. For the yaw angle β, the aerodynamic coefficients are observed to have different degrees of positive correlation with it. Among them, C_s and C_m demonstrate a stronger correlation with the β than C_l. On the whole, although the yaw angle presents a stronger correlation with the aerodynamic coefficients, the impact of the train shape parameters on the aerodynamic coefficients is still not negligible.

Descriptive statistics are performed in terms of all data sets, including both training and test data sets, as shown in Table 2. The minimum values of C_s and C_m are close to zero, and their maximum and mean values are 1.8889, 0.7662 and 0.7741, 0.3502, respectively. C_l ranges from −0.1810 to 1.3345 with a mean value of 0.5355. The variability in C_s is the largest than the remaining aerodynamic coefficients given the standard deviations (SDs). However, C_l has been exhibited the greatest variability considering the coefficient of variation (CV), which is a relative measure of variability and accounts for the disparate mean values. Although different aerodynamic coefficients demonstrate either positive or negative skewness, the absolute values of these skewnesses do not exceed 0.2. As a result, all data distributions are not significantly skewed in any one side, which indicates that the surrogate model will be more accurate and stable in processing these data.

4.2. Performance comparisons of different surrogate models

4.2.1. *Model parameter selection and calibration*

A total of 100 training samples are used to construct the surrogate model. Moreover, an important preprocessing step—normalization—has been firstly conducted to improve the performance and stability of the surrogate model. For PR, all processes are performed using the Multiple linear regression in the Statistics and Machine Learning Toolbox of MATLAB. PR models of different orders are compared for C_s, C_l, and C_m, and the final orders chosen

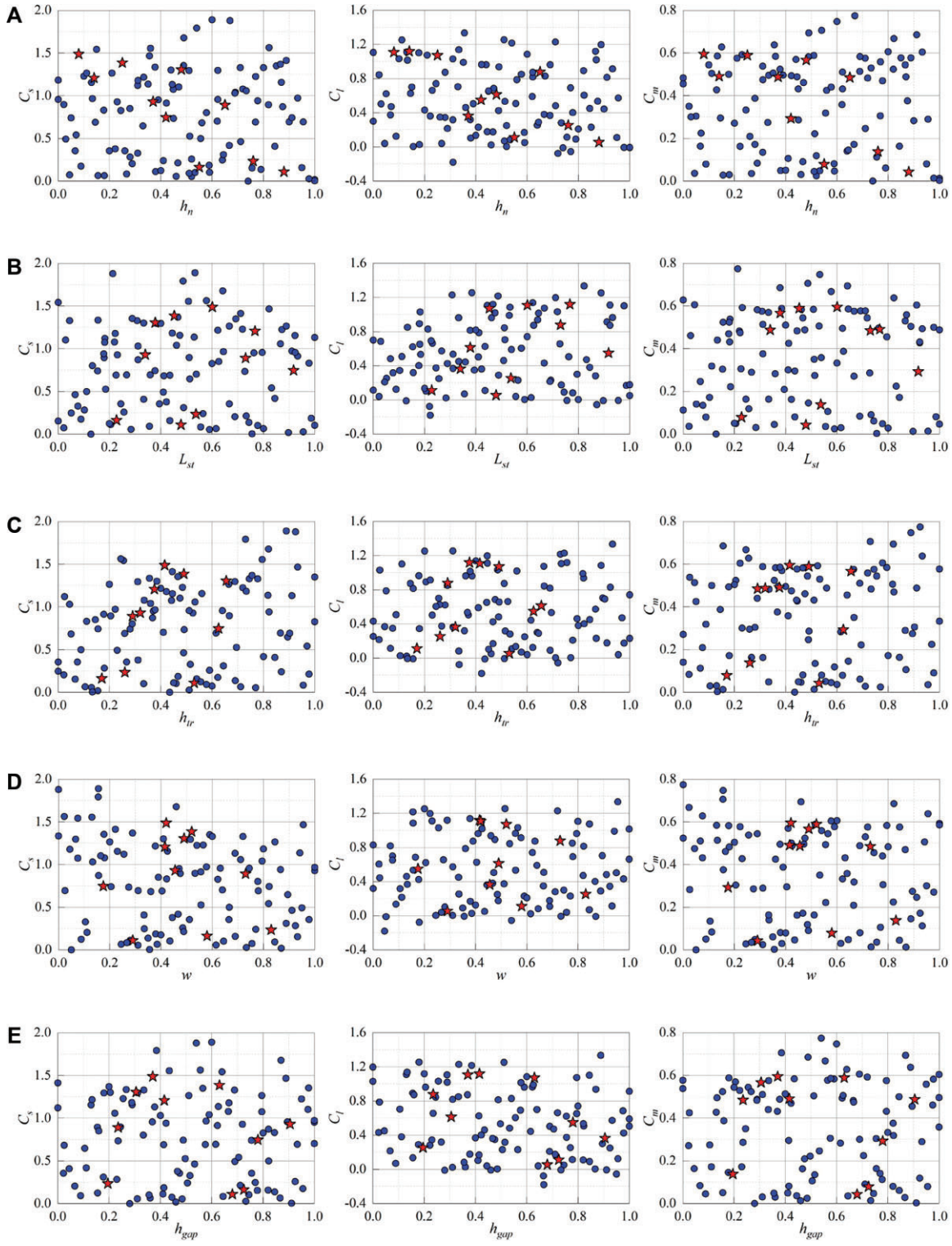

Figure 11: Effect of each variable on aerodynamic coefficients: from left to right, C_s, C_l, and C_m; from top to bottom, h_n, L_{st}, h_{tr}, w, h_{gap}, L_{uni}, and β.

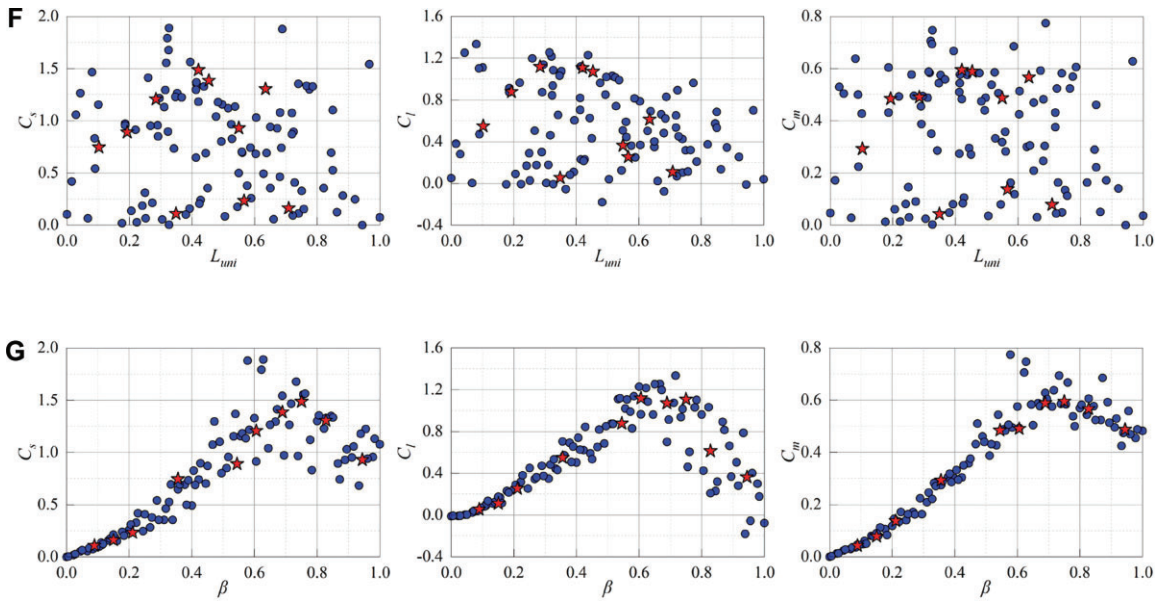

Figure 11: Continued.

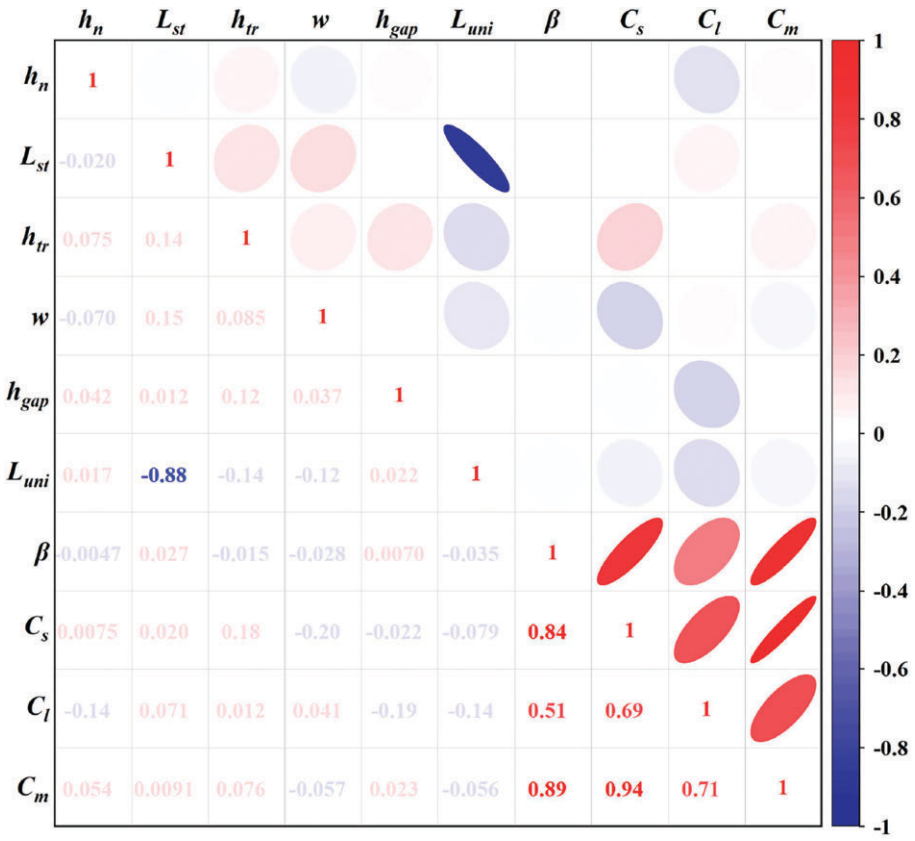

Figure 12: Correlation between all design variables and aerodynamic coefficients.

Table 2: Statistics analysis of aerodynamic coefficients.

Aerodynamic coefficients	Min.	Mean	Max.	SD	CV (%)	Skewness
C_s	−9.31e−05	0.7662	1.8889	0.5086	66.38	0.1279
C_l	−1.81e−01	0.5355	1.3345	0.3972	74.18	0.1994
C_m	−3.73e−05	0.3502	0.7741	0.2206	63.00	−0.1181

Table 3: Comparison of performance metrics for each surrogate model.

Surrogate models	C_s			C_l			C_m		
	MAE	MSE	MEC	MAE	MSE	MEC	MAE	MSE	MEC
PR	8.46E−02	7.89E−02	0.9195	1.98E−01	2.92E−01	0.8794	5.86E−02	5.56E−02	0.9119
SVR	6.40E−02	1.03E−02	0.9465	5.77E−02	1.12E−02	0.9309	2.14E−02	1.11E−03	0.9484
IPSO-SVR	2.96E−02	2.02E−03	0.9814	3.24E−02	3.70E−03	0.9741	1.06E−02	1.36E−04	0.9919
LSSVR	4.91E−02	6.26E−03	0.9529	4.36E−02	8.12E−03	0.9473	2.12E−02	1.06E−03	0.9538
IPSO-LSSVR	2.04E−02	8.73E−04	0.9887	2.43E−02	2.06E−03	0.9877	1.02E−02	1.29E−04	0.9946
Kriging	2.25E−02	1.97E−03	0.9809	3.57E−02	2.50E−03	0.9816	1.85E−02	7.53E−04	0.9821
IPSO-Kriging	1.02E−02	1.99E−04	0.9994	9.23E−03	1.62E−04	0.9988	3.76E−03	2.81E−05	0.9996

are 4, 6, and 5, with the least squares parameter estimates of 29, 43, and 36 dimensional vectors, respectively.

For the SVR, LSSVR, and Kriging models, the IPSO algorithm is employed to optimize the hyperparameters of these models, thereby developing the corresponding IPSO-SVR, IPSO-LSSVR, and IPSO-Kriging models. In the IPSO algorithm, the particle swarm size is 80, and the maximum number of iterations is 300. The maximum number of chaos search is 100, and the threshold of particle spacing is set to 0.01. The identical parameters are set in the IPSO algorithm for all developed surrogate models. The hyperparameters of the SVR, LSSVR, and Kriging models are treated as the input variables for optimization, and the minimization of MSE is selected as the optimization objective here. Moreover, we divide the training samples into 10 subsets, then each subset contains 10 groups of data. Consequently, the surrogate model is trained based on a 10-fold cross-validation technique. The penalty factor C and kernel parameter γ are crucial for the SVR and LSSVR models (Bi *et al.*, 2023; Pham *et al.*, 2020), and thus these two parameters are searched by the IPSO algorithm for the optimal combination. The initial combinations of parameters for the C_s, C_l, and C_m are $(C, \gamma) = (0.83, 0.35)$, (0.12, 0.46), and (1.41, 0.77) in the SVR models, respectively; while the optimal combinations of parameters for the above-mentioned coefficients are $(C, \gamma) = (2.59, 0.89)$, (0.39, 2.75), and (4.94, 1.01) in IPSO-SVR models, respectively. Similarly, the initial and optimal combinations of parameters for the C_s, C_l, and C_m are $(C, \gamma) = \{(12.07, 5.62); (8.14, 4.51); (11.98, 6.63)\}$ and {(24.81, 7.51); (37.31, 5.04); (48.94, 9.29)} in LSSVR and IPSO-LSSVR models, respectively.

It has been proved that if the optimal correlation parameter $\boldsymbol{\theta}^*$ in the maximum likelihood sense can be assured under any initial conditions (Wang *et al.*, 2022; Kaymaz, 2005), the optimal unbiased property of Kriging prediction can also be guaranteed. Therefore, the IPSO algorithm is adopted to seek the optimal correlation parameter $\boldsymbol{\theta}^*$. The initial values of parameter $\boldsymbol{\theta}^*$ for the C_s, C_l and C_m are (0.0315, 0.0050, 0.1714, 0.1470, 0.0002, 0.0043, 1.4932), and (0.0068, 0.0167, 0.1782, 0.0981, 0.0180, 0.0242, 1.4531), and (0.2160, 0.0028, 0.1080, 0.0303, 0.0006, 0.0317, 0.9051) in the Kriging models, respectively; and the optimal values of parameter $\boldsymbol{\theta}^*$ for the C_s, C_l, and C_m are (0.0035, 0.0198, 0.1133, 0.1160, 0.0735, 0.0501, 1.9240), (0.0182, 0.0112, 0.1330, 0.0638, 0.1064, 0.0373, 1.3172), and (0.0130, 0.0038, 0.1204, 0.0893, 0.0296, 0.0044, 1.4641) in the IPSO-Kriging models, respectively.

4.2.2. *Analysis of prediction accuracy*

To evaluate the prediction performance of different surrogate models, three accuracy metrics, including MAE, MSE, and MEC, are employed for comparison in Table 3. The PR model underperforms the other models by presenting higher MAE and MSE but lower MEC values for all three aerodynamic coefficients. Moreover, the prediction accuracies of the IPSO-SVR, IPSO-LSSVR and IPSO-Kriging models are significantly better than those of the simple SVR, LSSVR and Kriging models. Especially for the MSE values, almost all developed models are enhanced by an order of magnitude. This phenomenon suggests that the optimal hyperparameters searched by the IPSO algorithm are beneficial in improving the prediction accuracy of the surrogate model. The (IPSO-) SVR and (IPSO-) LSSVR models exhibit similar prediction tendencies due to their identical kernel functions and similar model structures. However, the (IPSO-) LSSVR model has a slightly higher prediction accuracy than the (IPSO-) SVR model for each aerodynamic coefficient. In essence, the LSSVR algorithm solves quadratic programming problems under equality constraints rather than inequality constraints of the SVR algorithm, which reduces the complexity of solving the problem and improves the computational speed and convergence accuracy of the algorithm (He *et al.*, 2022a; Wang *et al.*, 2021). The (IPSO-) Kriging model is remarkably superior to the other models. Generally, a better sampling design will have a favorable impact on the Kriging algorithm for predicting data trends, fitting the variogram function and kriging interpolation, thus improving the prediction accuracy of the Kriging algorithm (Ma *et al.*, 2020; Mulder *et al.*, 2013). The OLHD strategy is chosen to generate the training samples in this study, which provides a relatively preferable data set for Kriging modeling. In addition, the MEC values of all aerodynamic coefficients in the IPSO-Kriging models are more than 0.99, which indicates that the predictions of the IPSO-Kriging models are extremely close to the actual observations.

Based on the aforementioned analysis, the prediction values from only the PR model and the other models optimized by the IPSO algorithm are further compared with the CFD results from the training set in Figure 13, along with the APE for the aerodynamic coefficients predicted by each model. The predictions obtained by the IPSO-Kriging model closely align with the CFD results, with the APE for each training sample remaining below 5%. The IPSO-SVR and IPSO-LSSVR models can also predict the real values relatively well, although there are some deviations at individual locations. However, the predictions from the PR model demonstrate considerable deviations in several locations, especially for the C_l, which has the highest MAPE of 17.62%. Under extreme yaw angles, the IPSO-Kriging model achieves a maximum APE of 3.48% for aerodynamic coefficients at the yaw angle of 0° and 4.35% at the yaw angle of 90°, outperforming other models whose APE values exceed 5%. These results underscore the IPSO-Kriging model's superior predictive accuracy and robustnes, making it highly suitable for aerodynamic coefficient prediction in complex and extreme scenarios.

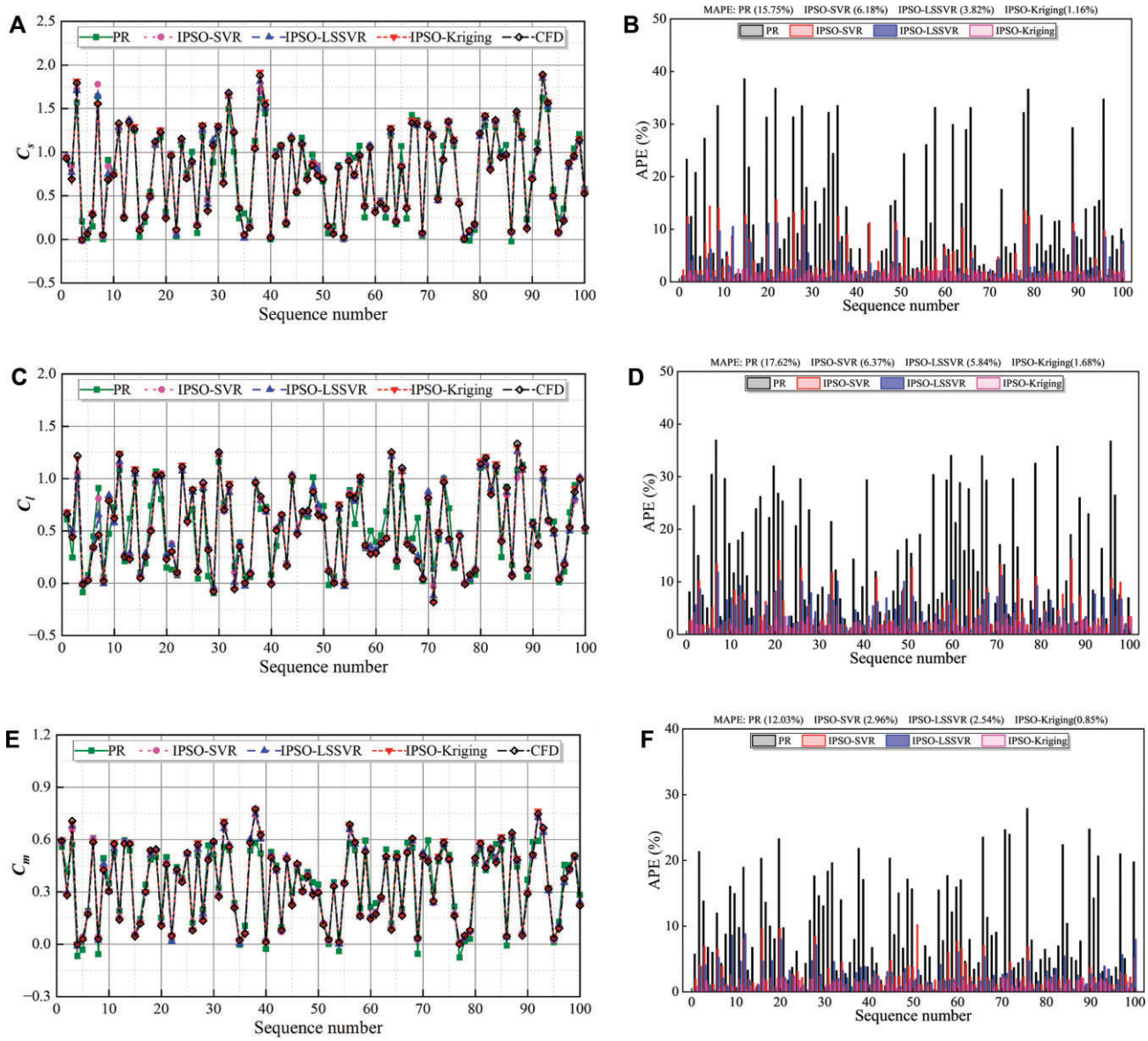

Figure 13: Comparison of aerodynamic coefficients from numerical simulations and predictions obtained by different surrogate models, along with the APE for the predicted aerodynamic coefficients from each model: (A) C_s, (B) APE of C_s, (C) C_l, (D) APE of C_l, (E) C_m, and (F) APE of C_m.

The regression plots of prediction values against CFD results are visualized in Figure 14. More scatter exists around the 1:1 line for the PR model, thus resulting in a lowest correlation coefficient (R^2). The IPSO-SVR and IPSO-LSSVR models show a similar pattern, with both overestimating lower values and underestimating higher values. Moreover, the R^2 values of the scatters for the C_m in both models are greater than 0.99, while the R^2 value of the scatter for the C_l in the IPSO-SVR model is no more than 0.98. The scatter is concentrated on the 1:1 line for the IPSO-Kriging model, which yields a R^2 value of more than 0.99 for each aerodynamic coefficient. This finding corroborates well with the lower MSE and higher MEC values of the IPSO-Kriging model for the three aerodynamic coefficients in Table 3.

4.2.3. *Quantification of prediction uncertainty*

Since the above analyses of prediction accuracy do not provide information related to prediction uncertainty, the PICPs for different confidence levels ranging from 5% to 95%, in terms of each surrogate model for the C_s, C_l, and C_m, are computed and displayed in Figure 15. In general, the ideal case is that all points fall on the gray dotted line. PICP curves for each model are consistently higher than the ideal values for the C_s and C_m, which means that all models overestimate the prediction of uncertainty. However, IPSO-SVR and IPSO-LSSVR models exhibit relatively lower values of PICPs at lower confidence levels (below 40%) for the C_l. Overall, the larger deviations from the 1:1 line occur in the PR model for each aerodynamic coefficient. This result implies that the PR model not only has a poor prediction accuracy, but also a poor prediction uncertainty. For the C_s, the overestimation of the uncertainty for the IPSO-Kriging model is lower than that of the other models, and the same tendencies are identified for the IPSO-SVR and IPSO-LSSVR models. Compared to the other models, the observed PICPs are closer to the ideal values for the IPSO-LSSVR model regarding the C_l. In other words, the IPSO-LSSVR model allows for a better assessment of the prediction uncertainty, although the IPSO-

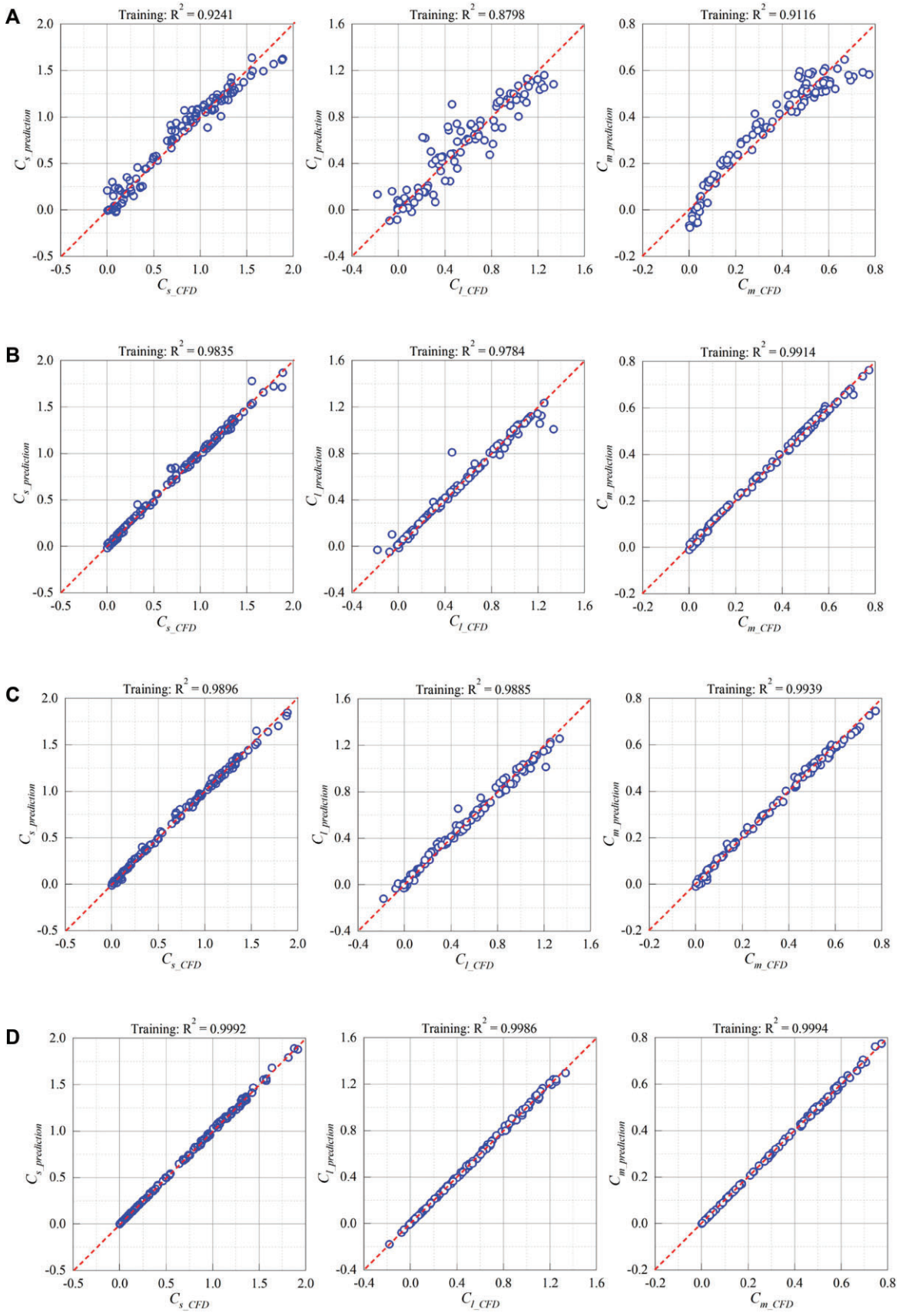

Figure 14: Correlation curves of CFD and predicted results (training data set): from left to right, C_s, C_l, and C_m; from top to bottom, PR, IPSO-SVR, IPSO-LSSVR, and IPSO-Kriging.

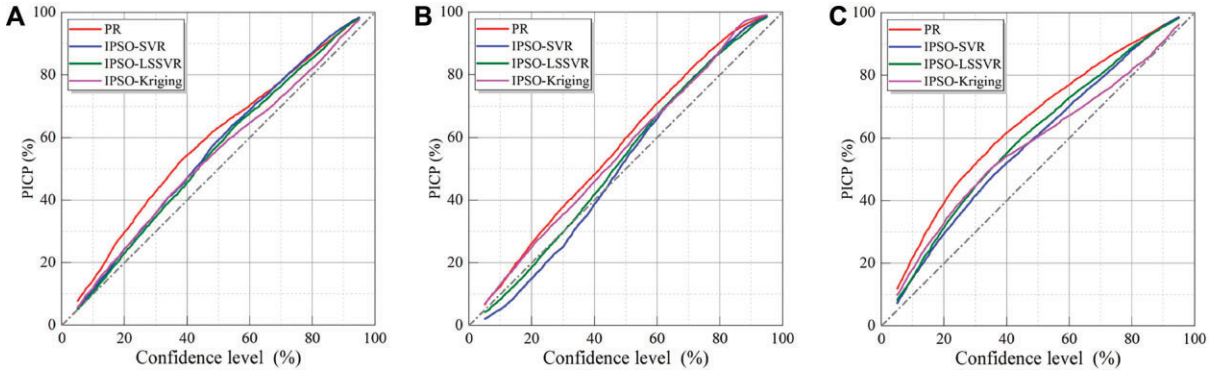

Figure 15: PICPs for different confidence levels: (A) C_s, (B) C_l, and (C) C_m.

Kriging model demonstrates a greater accuracy in predicting the C_l. For the C_m, the IPSO-SVR model delivers better uncertainty predictions for confidence levels below 50%, while the IPSO-Kriging model presents better uncertainty predictions for confidence levels above 50%.

4.3. Evaluation of generalization potential of surrogate models

To evaluate the generalization potentials of each surrogate model, ten random test samples have been generated in Figure 4 at the beginning in Section 2.2. The scatter plots of CFD data versus predicted values are shown in Figure 16 for the test data set. It is not surprising that the correlation coefficients of the test data set in Figure 16 are generally smaller than those of the corresponding training data set in Figure 14, due to the fact that the optimal hyperparameters of the surrogate model are tuned with respect to the training data. Similarly, the PR model exhibits the lowest R^2 values for all the aerodynamic coefficients of the test data set, and the highest R^2 value among them is only approximately 0.91 for the C_s. The R^2 values of the IPSO-SVR and IPSO-LSSVR models are relatively comparable for each aerodynamic coefficient. Moreover, although the R^2 value of the IPSO-LSSVR model is slightly higher, it is still below 0.99. The IPSO-Kriging model presents the highest R^2 values of above 0.99 for all the aerodynamic coefficients, which means that this model has the best performance in predicting the C_s, C_l, and C_m.

The APE of each test sample point and the MAPE of all test sample points for all surrogate models are illustrated in Figure 17. The maximum APE and MAPE values are both found in the PR model for all these aerodynamic coefficients, approaching 45% and 24%, respectively. Although the maximum APE and MAPE values in the prediction results of the IPSO-LSSVR model for the test samples are slightly lower than those of the IPSO-SVR model, they are still up to 26.36% and 10.64%, respectively. However, the IPSO-Kriging model has an APE of less than 5% in predicting all the aerodynamic coefficients for each test sample, and the MAPE value across all test samples is within 3%. Therefore, it could be concluded that the prediction performance of the IPSO-Kriging model is much better than the other models, indicating a superiority of the IPSO-Kriging model over the other models in terms of the generalization potential.

4.4. General discussion

Surrogate model techniques have been proven to be successful in predicting train aerodynamic coefficients under the combination of diverse train shape parameters and varying yaw angles. Moreover, the surrogate model demonstrates better prediction capability for unknown combinations of input parameters, especially for the developed IPSO-Kriging model. Nevertheless, constructing the surrogate model still requires considerable computational effort, particularly when dealing with a high-dimensional design space. The IPSO algorithm, used to optimize the surrogate model, introduces an additional layer of complexity due to its iterative nature and the need for meticulous parameter tuning to achieve optimal convergence during model training. Despite these challenges, the IPSO-Kriging framework offers a balance of computational efficiency and predictive accuracy. As a result, using surrogate models instead of model tests or CFD simulations to obtain train aerodynamic coefficients will be very promising in practice.

In this study, the training data for all surrogate models are derived from CFD simulations. To save computational resources, a relatively simplified train model is chosen to study the aerodynamic effects of changes in train shape parameters. While six design variables are representative of the most influential shape characteristics, other factors such as bogies, windshields, pantographs, and various attached structures may also play a role in train aerodynamic performance. These factors could be overlooked in the current model, potentially limiting its applicability in more detailed or real-world scenarios. Thus, the inclusion of these elements in the surrogate model is essential for attaining a higher degree of scientific accuracy and practical relevance. In addition, railway infrastructure scenarios, including the type of railway (Schober *et al.*, 2010) and the form of windbreaks (Mohebbi & Rezvani, 2018c; Mohebbi & Safaee, 2022; Xia *et al.*, 2022), also have a nonnegligible impact on the aerodynamic performance of trains under crosswinds. The above makes acquiring sufficient relevant data from model tests or CFD simulations still a hindrance in training proper surrogate models. Therefore, future research efforts will consider the above information under adequate resources.

5. Conclusions

This paper focuses on predicting the aerodynamic coefficients resulting from the combination of diverse train shapes and varying yaw angles based on surrogate models that combine regression algorithms and CFD simulations.

CFD results reveal the relationships between input and output variables, and provide statistical insights into the aerodynamic coefficients. Compared to the train shape parameters, the yaw angle exhibits a stronger correlation with the C_s, C_l, and C_m, with the PCCs of 0.84, 0.51, 0.89, respectively. The absolute skewness of

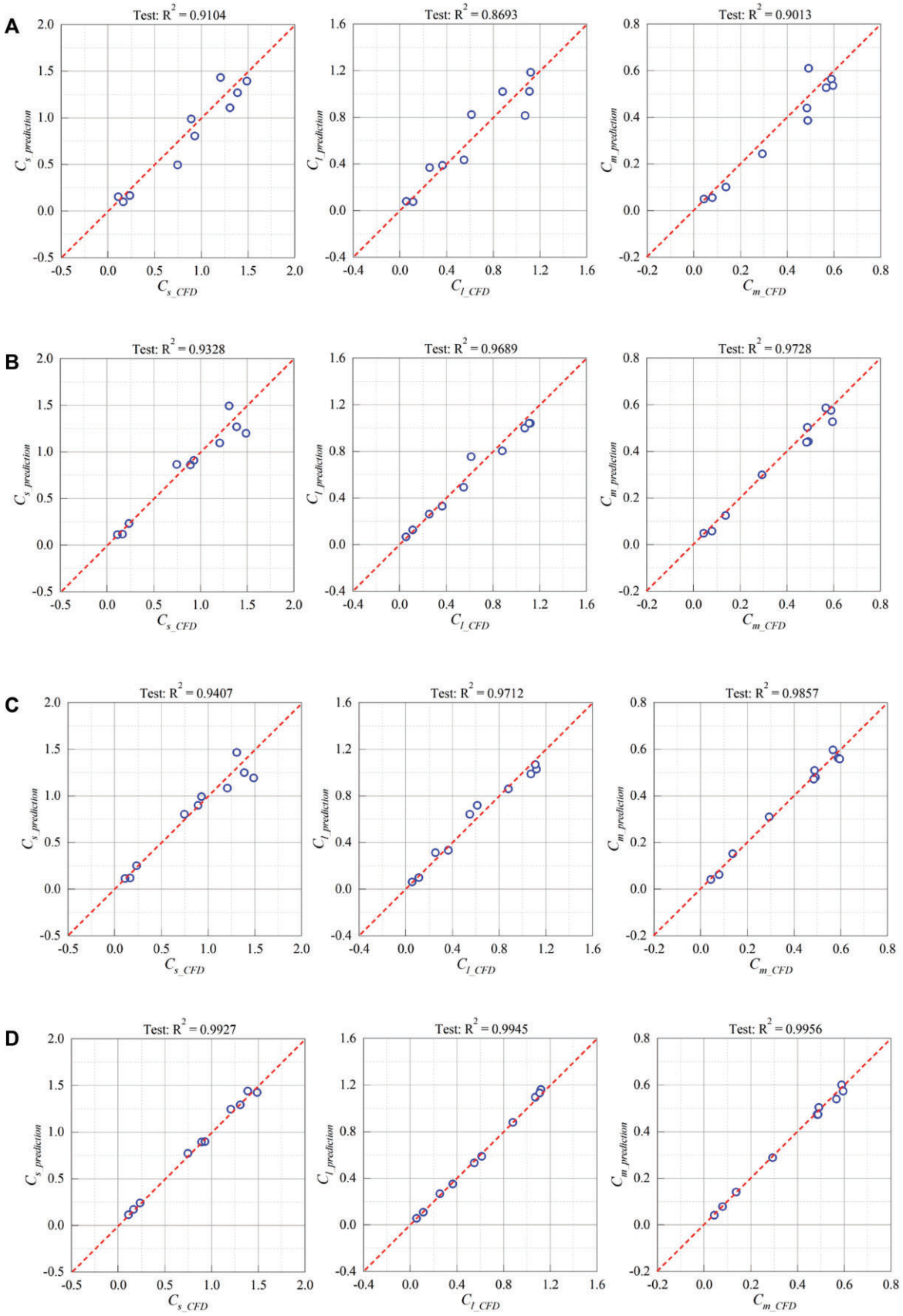

Figure 16: Correlation curves of CFD and predicted results (test data set): from left to right, C_s, C_l, and C_m; from top to bottom, PR, IPSO-SVR, IPSO-LSSVR, and IPSO-Kriging.

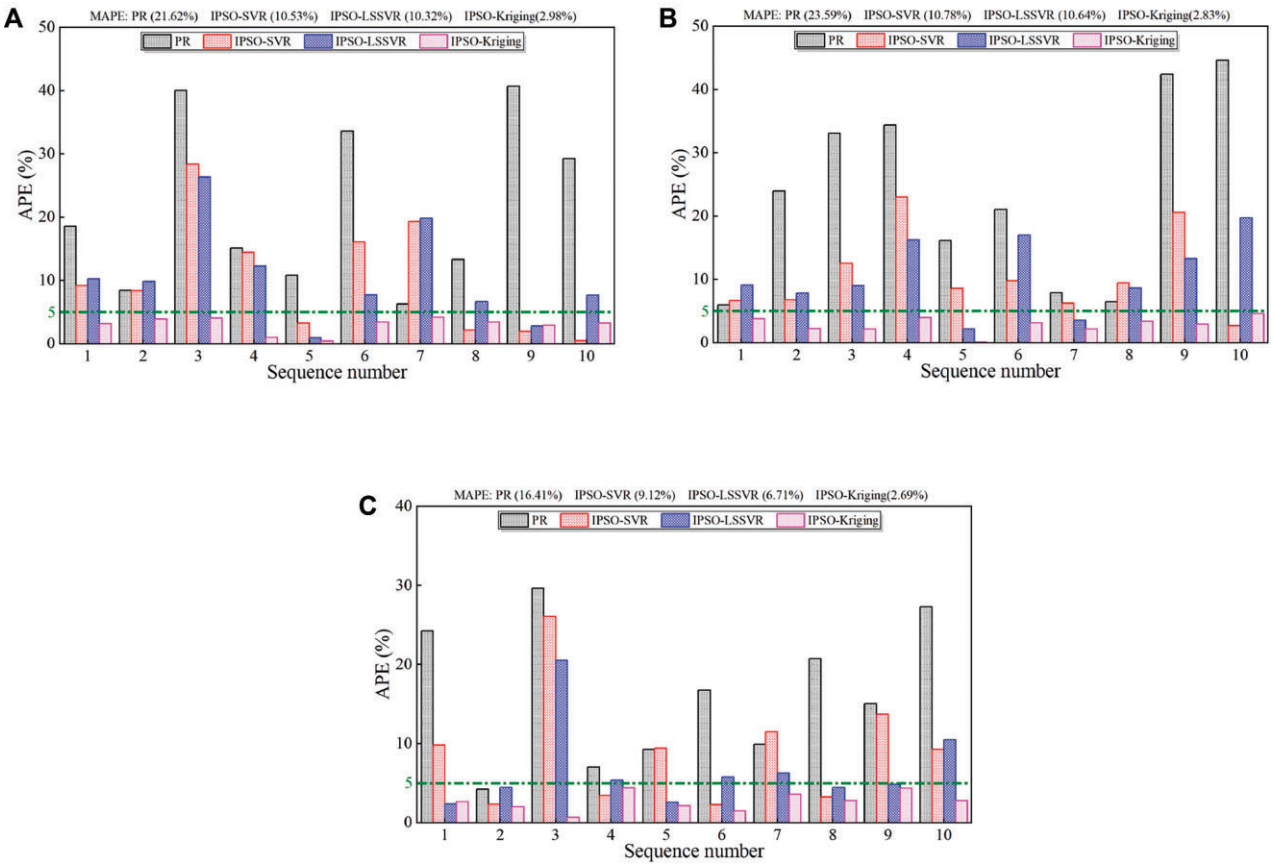

Figure 17: APE in predicting aerodynamic coefficients by each surrogate model: (A) C_s, (B) C_l, and (C) C_m.

all aerodynamic coefficients does not exceed 0.2, indicating stable and accurate surrogate model performance.

Seven surrogate models, including four original models and three IPSO-optimized versions, are developed to predict the aerodynamic coefficients. The application of IPSO significantly enhances the prediction performance of the original models. Among them, the IPSO-Kriging model outperforms other surrogate models in terms of prediction accuracy. Moreover, the IPSO-Kriging model demonstrates better generalization ability by providing an APE of less than 5% for each test sample. Although the prediction uncertainty for the C_s is better quantified by the IPSO-Kriging model, the prediction uncertainty for the C_l is better evaluated by the IPSO-LSSVR model. Overall, the IPSO-Kriging model could be considered an effective and economical alternative to wind tunnel tests and CFD simulations for quickly acquiring the aerodynamic coefficients under the combination of various train shape parameters and different yaw angles.

Nevertheless, this study has effectively employed surrogate models to predict aerodynamic coefficients across various combinations of train shape parameters and yaw angles. This success encourages future research to delve deeper into the application of surrogate models in exploring aerodynamic coefficients under more complex conditions, including realistic train geometries, diverse railway types, and various windshield designs. By incorporating additional shape parameters that reflect practical design considerations, the train model can be refined to more accurately represent actual high-speed train configurations. Furthermore, extending the current framework to incorporate environmental factors would significantly enhance its applicability in real-world railway operations.

Conflicts of Interest

The authors declare that no relevant financial or nonfinancial conflicts of interest exist in this paper.

Author Contributions

Xiaoshuai Huo: Conceptualization, Methodology, Software, Data curation, Writing—original draft. **Tanghong Liu**: Writing—original draft. **Xiaodong Chen**: Methodology, Writing—review & editing. **Zhengwei Chen**: Supervision, Investigation, Writing—review & editing. **Xinran Wang**: Investigation.

Funding

This work is supported by the National Natural Science Foundation of China (Grant Number 52202426), and grants from the Research Grants Council of the Hong Kong Special Administrative Region (SAR), China (Grants Numbers 15 205 723 and 15 226 424).

Data availability

Any relevant source code and study data can be available from the corresponding author by appropriate requests.

Acknowledgments

The authors are grateful for resources from the High Performance Computing Center of Central South University.

References

Baker, C. (2013). A framework for the consideration of the effects of crosswinds on trains. *Journal of Wind Engineering and Industrial Aerodynamics*, **123**, 130–142. https://doi.org/10.1016/j.jweia.2013.09.015.

Banks, A., Vincent, J., & Anyakoha, C. (2007). A review of particle swarm optimization. Part I: Background and development. *Natural Computing*, **6**, 467–484. https://doi.org/10.1007/s11047-007-9049-5.

Bi, S., Shao, L., Qi, Z., Wang, Y., & Lai, W. (2023). Prediction of coal mine gas emission based on hybrid machine learning model. *Earth Science Informatics*, **16**, 501–513. https://doi.org/10.21203/rs.3.rs-2080112/v1.

Brambilla, E., Giappino, S., & Tomasini, G. (2022). Wind tunnel tests on railway vehicles in the presence of windbreaks: Influence of flow and geometric parameters on aerodynamic coefficients. *Journal of Wind Engineering and Industrial Aerodynamics*, **220**, 104838. https://doi.org/10.1016/j.jweia.2021.104838.

Catanzaro, C., Cheli, F., Rocchi, D., Schito, P., & Tomasini, G. (2016). High-speed train crosswind analysis: CFD study and validation with wind-tunnel tests. In *The Aerodynamics of Heavy Vehicles III: Trucks, Buses and Trains* (pp. 99–112). Springer. https://doi.org/10.1007/978-3-319-20122-1_6.

Chen, X., Zhong, S., Ozer, O., & Weightman, A. (2022a). Control of afterbody vortices from a slanted-base cylinder using sweeping jets. *Physics of Fluids*, **34**, 075115. https://doi.org/10.1063/5.0094565.

Chen, X., Zhong, S., Ozer, O., Weightman, A., & Gao, G. (2023). On the unsteady interactions between a sweeping jet and afterbody vortices. *Physics of Fluids*, **35**, 105153. https://doi.org/10.1063/5.0167467.

Chen, Z., Liu, T., Jiang, Z., Guo, Z., & Zhang, J. (2018). Comparative analysis of the effect of different nose lengths on train aerodynamic performance under crosswind. *Journal of Fluids and Structures*, **78**, 69–85. https://doi.org/10.1016/j.jfluidstructs.2017.12.016.

Chen, Z.-W., Liu, T.-H., Yan, C.-G., Yu, M., Guo, Z.-J., & Wang, T.-T. (2019). Numerical simulation and comparison of the slipstreams of trains with different nose lengths under crosswind. *Journal of Wind Engineering and Industrial Aerodynamics*, **190**, 256–272. https://doi.org/10.1016/j.jweia.2019.05.005.

Chen, Z.-W., Ni, Y.-Q., Wang, Y.-W., Wang, S.-M., & Liu, T.-H. (2022b). Mitigating crosswind effect on high-speed trains by active blowing method: A comparative study. *Engineering Applications of Computational Fluid Mechanics*, **16**, 1064–1081. https://doi.org/10.1080/19942060.2022.2064921.

Chiu, T., & Squire, L. (1992). An experimental study of the flow over a train in a crosswind at large yaw angles up to 90. *Journal of Wind Engineering and Industrial Aerodynamics*, **45**, 47–74. https://doi.org/10.1016/0167-6105(92)90005-u.

Copley, J. (1987). The three-dimensional flow around railway trains. *Journal of Wind Engineering and Industrial Aerodynamics*, **26**, 21–52. https://doi.org/10.1016/0167-6105(87)90034-1.

Gao, H., Liu, T., Gu, H., Jiang, Z., Huo, X., Xia, Y., & Chen, Z. (2021). Full-scale tests of unsteady aerodynamic loads and pressure distribution on fast trains in crosswinds. *Measurement*, **186**, 110152. https://doi.org/10.1016/j.measurement.2021.110152.

Guo, Z., Liu, T., Chen, Z., Xia, Y., Li, W., & Li, L. (2020). Aerodynamic influences of bogie's geometric complexity on high-speed trains under crosswind. *Journal of Wind Engineering and Industrial Aerodynamics*, **196**, 104053. https://doi.org/10.1016/j.jweia.2019.104053.

Guo, Z.-J., Guo, Z.-H., Chen, Z.-W., Zeng, G.-Z., & Xu, J.-Q. (2024). On the active flow control in maglev train safety under crosswinds: Analysis of leeward suction and blowing action. *Physics of Fluids*, **36**, 095130. https://doi.org/10.1063/5.0224211.

He, Z., Liu, T., & Liu, H. (2022a). Improved particle swarm optimization algorithms for aerodynamic shape optimization of high-speed train. *Advances in Engineering Software*, **173**, 103242. https://doi.org/10.1016/j.advengsoft.2022.103242.

He, Z., Xiong, X., Yang, B., & Li, H. (2022b). Aerodynamic optimisation of a high-speed train head shape using an advanced hybrid surrogate-based nonlinear model representation method. *Optimization and Engineering*, **23**, 59–84. https://doi.org/10.1007/s11081-020-09554-3.

Heleno, R., Montenegro, P., Carvalho, H., Ribeiro, D., Calcada, R., & Baker, C. (2021). Influence of the railway vehicle properties in the running safety against crosswinds. *Journal of Wind Engineering and Industrial Aerodynamics*, **217**, 104732. https://doi.org/10.1016/j.jweia.2021.104732.

Hemida, H., & Krajnović, S. (2010). LES study of the influence of the nose shape and yaw angles on flow structures around trains. *Journal of Wind Engineering and Industrial Aerodynamics*, **98**, 34–46. https://doi.org/10.1016/j.jweia.2009.08.012.

Hu, G., & Kwok, K. C. S. (2020). Predicting wind pressures around circular cylinders using machine learning techniques. *Journal of Wind Engineering and Industrial Aerodynamics*, **198**, 104099. https://doi.org/10.1016/j.jweia.2020.104099.

Huo, X., Liu, T., Yu, M., Chen, Z., Guo, Z., Li, W., & Wang, T. (2021). Impact of the trailing edge shape of a downstream dummy vehicle on train aerodynamics subjected to crosswind. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, **235**, 201–214. https://doi.org/10.1177/0954409720915039.

Huo, X. S., Liu, T. H., Chen, Z. W., Li, W. H., Gao, H. R., & Xu, B. (2023b). Comparison of RANS, URANS, SAS and IDDES for the prediction of train crosswind characteristics. *Wind and Structures, An International Journal*, **37**, 303–314. https://doi.org/10.12989/was.2023.37.4.303

Huo, X.-S., Liu, T.-H., Chen, Z.-W., Li, W.-H., Niu, J.-Q., & Gao, H.-R. (2023a). Aerodynamic characteristics of double-connected train groups composed of different kinds of high-speed trains under crosswinds: A comparison study. *Alexandria Engineering Journal*, **64**, 465–481. https://doi.org/10.1016/j.aej.2022.09.011.

Jin, R., Chen, W., & Sudjianto, A. (2005). An efficient algorithm for constructing optimal design of computer experiments. *Journal of statistical planning and inference*, **134**, 268–287. https://doi.org/10.1016/j.jspi.2004.02.014.

Kaymaz, I. (2005). Application of kriging method to structural reliability problems. *Structural Safety*, **27**, 133–151. https://doi.org/10.1016/j.strusafe.2004.09.001.

Keane, A. J., & Voutchkov, I. I. (2020). Robust design optimization using surrogate models. *Journal of Computational Design and Engineering*, **7**, 44–55. https://doi.org/10.1093/jcde/qwaa005.

Kennedy, J. (2010). *Particle Swarm Optimization. Encyclopedia of Machine Learning* (pp. 760–766) . Springer US.

Lee, U., & Kang, N. (2023). Adaptive neural network ensemble using prediction frequency. *Journal of Computational Design and Engineering*, **10**, 1547–1560. https://doi.org/10.1093/jcde/qwad071.

Li, T., Qin, D., & Zhang, J. (2019). Effect of RANS turbulence model on aerodynamic behavior of trains in crosswind. *Chinese Journal of Mechanical Engineering - English Edition*, **32**, 1–12. https://doi.org/10.1186/s10033-019-0402-2.

Li, W., Liu, T., Martinez-Vazquez, P., Chen, Z., Huo, X., Liu, D., & Xia, Y. (2022a). Correlation tests on train aerodynamics between multiple wind tunnels. *Journal of Wind Engineering and Industrial Aerodynamics*, **229**, 105137. https://doi.org/10.1016/j.jweia.2022.105137.

Li, W., Liu, T., Martinez-Vazquez, P., Guo, Z., Huo, X., Xia, Y., & Chen, Z. (2022b). Effects of embankment layouts on train aerodynamics in a wind tunnel configuration. *Journal of Wind Engineering and Industrial Aerodynamics*, **220**, 104830. https://doi.org/10.1016/j.jweia.2021.104830.

Liu, T.-H., Wang, L., Chen, Z.-W., Gao, H.-R., Li, W.-H., Guo, Z.-j., Xia, Y.-T., Huo, X.-S., & Wang, Y.-W. (2022). Study on the pressure pipe length in train aerodynamic tests and its applications in crosswinds. *Journal of Wind Engineering and Industrial Aerodynamics*, **220**, 104880. https://doi.org/10.1016/j.jweia.2021.104880.

Ma, T., Brus, D. J., Zhu, A.-X., Zhang, L., & Scholten, T. (2020). Comparison of conditioned Latin hypercube and feature space coverage sampling for predicting soil classes using simulation from soil maps. *Geoderma*, **370**, 114366. https://doi.org/10.1016/j.geoderma.2020.114366.

Malone, B., McBratney, A., & Minasny, B. (2011). Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. *Geoderma*, **160**, 614–626. https://doi.org/10.1016/j.geoderma.2010.11.013.

Mohebbi, M., Ma, Y., & Mohebbi, R. (2023). The influence of inclined barriers on airflow over a high speed train under crosswind condition. In Hessami A. G., & Muttram R., *Book: New Research on Railway Engineering and Transportation*, IntechOpen, ISBN 978-1-83768-620-9. https://doi.org/10.5772/intechopen.112751.

Mohebbi, M., Ma, Y., & Mohebbi, R. (2024). The analysis of utilizing multiple fences in high-speed tracks on the aerodynamic characteristics of a high-speed train model. *Iranian Journal of Science and Technology, Transactions of Mechanical Engineering*, **48**, 847–863. https://doi.org/10.1007/s40997-023-00702-5.

Mohebbi, M., & Rezvani, M. (2018a). The impact of air fences geometry on air flow around an ICE3 high speed train on a double line railway track with exposure to crosswinds. *Journal of Applied Fluid Mechanics*, **11**, 743–754. https://doi.org/10.29252/jafm.11.03.27862.

Mohebbi, M., & Rezvani, M. A. (2018b). Multi objective optimization of aerodynamic design of high speed railway windbreaks using Lattice Boltzmann Method and wind tunnel test results. *International Journal of Rail Transportation*, **6**, 183–201. https://doi.org/10.1080/23248378.2018.1463873.

Mohebbi, M., & Rezvani, M. A. (2018c). Two-dimensional analysis of the influence of windbreaks on airflow over a high-speed train under crosswind using lattice Boltzmann method. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, **232**, 863–872. https://doi.org/10.1177/0954409717699502.

Mohebbi, M., & Rezvani, M. A. (2019). Analysis of the effects of lateral wind on a high speed train on a double routed railway track with porous shelters. *Journal of Wind Engineering and Industrial Aerodynamics*, **184**, 116–127. https://doi.org/10.1016/j.jweia.2018.11.011.

Mohebbi, M., & Rezvani, M. A. (2021). 2D and 3D numerical and experimental analyses of the aerodynamic effects of air fences on a high-speed train. *Wind and Structures*, **32**, 539–550. https://doi.org/10.12989/was.2021.32.6.539

Mohebbi, M., & Safaee, A. M. (2022). The optimum model determination of porous barriers in high-speed tracks. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, **236**, 15–25. https://doi.org/10.1177/0954409721995323.

Montenegro, P., Carvalho, H., Ortega, M., Millanes, F., Goicolea, J., Zhai, W., & Calçada, R. (2022). Impact of the train-track-bridge system characteristics in the runnability of high-speed trains against crosswinds-part I: Running safety. *Journal of Wind Engineering and Industrial Aerodynamics*, **224**, 104974. https://doi.org/10.1016/j.jweia.2022.104974.

Morden, J. A., Hemida, H., & Baker, C. J. (2015). Comparison of RANS and detached eddy simulation results to wind-tunnel data for the surface pressures upon a class 43 high-speed train. *Journal of Fluids Engineering*, **137**, 041108. https://doi.org/10.1115/1.4029261.

Mulder, V., de Bruin, S., & Schaepman, M. E. (2013). Representing major soil variability at regional scale by constrained Latin Hypercube sampling of remote sensing data. *International Journal of Applied Earth Observation and Geoinformation*, **21**, 301–310. https://doi.org/10.1016/j.jag.2012.07.004.

Muñoz-Paniagua, J., & García, J. (2019). Aerodynamic surrogate-based optimization of the nose shape of a high-speed train for crosswind and passing-by scenarios. *Journal of Wind Engineering and Industrial Aerodynamics*, **184**, 139–152. https://doi.org/10.1016/j.jweia.2018.11.014.

Muthukrishnan, S., Krishnaswamy, H., Thanikodi, S., Sundaresan, D., & Venkatraman, V. (2020). Support vector machine for modelling and simulation of heat exchangers. *Thermal Science*, **24**, 499–503. https://doi.org/10.2298/TSCI190419398M.

Naka, S., Genji, T., Yura, T., & Fukuyama, Y. (2003). A hybrid particle swarm optimization for distribution state estimation. *IEEE Transactions on Power Systems*, **18**, 60–68. https://doi.org/10.1109/TPWRS.2002.807051.

Niu, J., Zhang, Y., Li, R., Chen, Z., Yao, H., & Wang, Y. (2022). Aerodynamic simulation of effects of one-and two-side windbreak walls on a moving train running on a double track railway line subjected to strong crosswind. *Journal of Wind Engineering and Industrial Aerodynamics*, **221**, 104912. https://doi.org/10.1016/j.jweia.2022.104912.

Pham, A.-D., Ngo, N.-T., Nguyen, Q.-T., & Truong, N.-S. (2020). Hybrid machine learning for predicting strength of sustainable concrete. *Soft Computing*, **24**, 14965–14980. https://doi.org/10.1007/s00500-020-04848-1.

Premoli, A., Rocchi, D., Schito, P., & Tomasini, G. (2016). Comparison between steady and moving railway vehicles subjected to crosswind by CFD analysis. *Journal of Wind Engineering and Industrial Aerodynamics*, **156**, 29–40. https://doi.org/10.1016/j.jweia.2016.07.006.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). *Cross-validation. Encyclopedia of Database Systems* (pp. 532–538). Springer US

Schober, M., Weise, M., Orellano, A., Deeg, P., & Wetzel, W. (2010). Wind tunnel investigation of an ICE 3 endcar on three standard ground scenarios. *Journal of Wind Engineering and Industrial Aerodynamics*, **98**, 345–352. https://doi.org/10.1016/j.jweia.2009.12.004.

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, **9**, 293–300. https://doi.org/10.1023/A:1018628609742.

Tian, H. (2019). Review of research on high-speed railway aerodynamics in China. *Transportation Safety and Environment*, **1**, 1–21. https://doi.org/10.1093/tse/tdz014.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, **10**, 988–999. https://doi.org/10.1109/72.788640.

Wadoux, A. M.-C., Brus, D. J., & Heuvelink, G. B. (2018). Accounting for non-stationary variance in geostatistical mapping of soil properties. *Geoderma*, **324**, 138–147. https://doi.org/10.1016/j.geoderma.2018.03.010.

Wang, K., Xiong, X., Wen, C., Li, X., Chen, G., Chen, Z., & Tang, M. (2023). Impact of the train heights on the aerodynamic behaviour of a high-speed train. *Engineering Applications of Computational Fluid Mechanics*, **17**, 2233614. https://doi.org/10.1080/19942060.2023.2233614.

Wang, S., Wu, K., Zhao, Q., Wang, S., Feng, L., Zheng, Z., & Wang, G. (2021). Multienergy load forecasting for regional Integrated Energy systems considering Multienergy coupling of variation characteristic curves. *Frontiers in Energy Research*, **9**, 635234. https://doi.org/10.3389/fenrg.2021.635234.

Wang, X., Zhang, J., Sun, Y., Wu, Z., Tchuente, N. F. C., & Yang, F. (2022). Stiffness identification of deteriorated PC bridges by a FEMU method based on the LM-assisted PSO-Kriging model. *Structures*, **43**, 374–387. https://doi.org/10.1016/j.istruc.2022.06.060.

Wu, H., Wei, P., Hu, R., Liu, H., Du, X., Zhou, P., & Zhu, C. (2023). Study on the relationship between machining errors and transmission accuracy of planetary roller screw mechanism using analytical calculations and machine-learning model. *Journal of Computational Design and Engineering*, **10**, 398–413. https://doi.org/10.1093/jcde/qwad003.

Xia, Y., Liu, T., Su, X., Jiang, Z., Chen, Z., & Guo, Z. (2022). Aerodynamic influences of typical windbreak wall types on a high-speed train under crosswinds. *Journal of Wind Engineering and Industrial Aerodynamics*, **231**, 105203. https://doi.org/10.1016/j.jweia.2022.105203.

Xu, G., Liang, X., Yao, S., Chen, D., & Li, Z. (2017). Multi-objective aerodynamic optimization of the streamlined shape of high-speed trains based on the Kriging model. *PLoS ONE*, **12**, e0170803. https://doi.org/10.1371/journal.pone.0170803.

Yao, S., Guo, D., Sun, Z., Chen, D., & Yang, G. (2016). Parametric design and optimization of high speed train nose. *Optimization and Engineering*, **17**, 605–630. https://doi.org/10.1007/s11081-015-9298-6.

Zamiri, A., & Chung, J. T. (2017). Ability of URANS approach in prediction of unsteady turbulent flows in an unbaffled stirred tank. *International Journal of Mechanical Sciences*, **133**, 178–187. https://doi.org/10.1016/j.ijmecsci.2017.08.008.

Zeng, G.-Z., Chen, Z.-W., Ni, Y.-Q., & Rui, E.-Z. (2024). Investigating embedded data distribution strategy on reconstruction accuracy of flow field around the crosswind-affected train based on physics-informed neural networks. *International Journal of Numerical Methods for Heat & Fluid Flow*, **34**, 2963–2985. https://doi.org/10.1108/HFF-11-2023-0709

Zhang, L., Li, T., & Zhang, J. (2021). Research on aerodynamic shape optimization of trains with different dimensional design variables. *International Journal of Rail Transportation*, **9**, 479–501. https://doi.org/10.1080/23248378.2020.1817803.

Zhang, L., Zhang, J., Li, T., & Zhang, Y. (2018). A multiobjective aerodynamic optimization design of a high-speed train head under crosswinds. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, **232**, 895–912. https://doi.org/10.1177/0954409717701784.

Appendix

A. Mathematical Derivations of Algorithms

A.1. Polynomial regression

Multiple PR is a form of regression analysis that models the relationship between one or more independent variable and a dependent variable as an nth-order polynomial. Assume that the dataset D contains m records: $t_1, t_2, \ldots, t_m$, and has d independent variables $X_1, X_2, \ldots, X_d$ and a dependent variable Y. Then the d-dimensional nth-order PR model can be expressed as follows:

$$y = \sum_{i=1}^{n}\sum_{j=1}^{d} b_{ij}x_j^i + b_0, \tag{A1}$$

where b_{ij} is the coefficient of x_j^i and b_0 is constant term. The fitting of PR can be converted to multiple linear regression by variable substitution. Each x_j^i can be equated to a new independent variable z_i with the following transformation equation:

$$\begin{cases} z_1 = x_1, z_2 = x_2, \cdots, z_d = x_d \\ z_{d+1} = x_1^2, z_{d+2} = x_2^2, \cdots, z_{2d} = x_d^2 \\ \vdots \quad\quad \vdots \quad\quad \vdots \\ z_{(n-1)d+1} = x_1^n, z_{(n-1)d+2} = x_2^n, \cdots, z_{nd} = x_d^n \end{cases} \tag{A2}$$

After the mapping of Equation A2 , the PR can be converted into a multiple linear regression on z_i, as shown in the following equation:

$$y = zw + w_0 \tag{A3}$$

where $\boldsymbol{w}$ is the coefficient vector of $\boldsymbol{z}$, corresponding to the coefficient b_{ij} of x_j^i in PR. This new linear model can then be fitted using the least square method, which minimizes the residual sum of squares to estimate the model parameters. Thus, the parameter vector of the regression model can be solved using matrix operations according to the least square method, with the following results:

$$w = (Z^TZ)^{-1}Z^TY, \tag{A4}$$

where $\boldsymbol{Z}$ is the matrix of independent variables, $\boldsymbol{Y}$ is the vector of dependent variable, and $\boldsymbol{w}$ is the vector of model parameters.

A.2. Support vector regression

SVR is an application model of support vector machine (SVM) to regression problems (Muthukrishnan *et al.*, 2020). Considering the training dataset $\{(\boldsymbol{x}_i, y_i),\ i = 1, 2, \ldots, n,\ x_i \in R^n,\ y_i \in R\}$, the input vector $\boldsymbol{x}$ is initially transformed from a low-dimensional space to a high-dimensional (m-dimensional) space, and the regression function is given by

$$f(x, w) = \sum_{j=1}^{m} w_j g_j(x) + b \tag{A5}$$

where w_j represents the 'weight' coefficient, $g_j(\boldsymbol{x})$ indicates a set of nonlinear transformation functions, and b denotes the 'bias' term. In general, the absolute or square error is employed as a loss function to minimize the risk of prediction. However, the ε-insensitive loss function (Vapnik, 1999) is introduced in the SVR, which is formulated as

$$L_\varepsilon(y_i, f(x_i, w)) = \begin{cases} 0, |y_i - f(x_i, w)| \le \varepsilon \\ |y_i - f(x_i, w)| - \varepsilon, \text{otherwise} \end{cases}. \tag{A6}$$

Meanwhile, the slack factors ξ_i and ξ_i^* are proposed to monitor the training sample deviations outside the ε-insensitive region. Thus, the optimization problem of the SVR is defined by minimizing the corresponding cost function as follows:

$$\begin{aligned} &\text{minimize} \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &\text{subject to} \begin{cases} y_i - f(x_i, w) \le \varepsilon + \xi_i \\ f(x_i, w) - y_i \le \varepsilon + \xi_i^* \\ \xi_i \ge 0, \xi_i^* \ge 0 \end{cases} \end{aligned} \tag{A7}$$

Here, C denotes a positive constant, which controls the trade-off between approximation inaccuracy and model complexity. Transforming the above optimization into a dual problem to solve, the

final functional expression of SVR is given as

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b. \tag{A8}$$

Here, α_i and α_i^* are the dual factors, having values in the range from zero to C; and $K(\boldsymbol{x}_i, \boldsymbol{x})$ denotes the kernel function, which is capable of converting sample data to a high-dimensional feature space. In this study, the most widely used RBF is chosen as a kernel function, and its expression is as follows:

$$K(x_i, x) = \exp(-\gamma \|x_i - x\|^2), \tag{A9}$$

where γ is the kernel parameter, which is used to control the locality of the kernel function.

A.3. Least square support vector regression

As a deformation algorithm of SVR, LSSVR (Suykens & Vandewalle, 1999) transforms the inequality constraint into an equation constraint. In addition, the error sum of squares is used as a loss function, and the solution algorithm is transformed from solving a convex quadratic optimization problem to solving a system of linear equations. Thus, the optimization problem can be formulated as

$$\begin{aligned} &\text{minimize } J(w, b, e) = \tfrac{1}{2}\|w\|^2 + \tfrac{1}{2} C \sum_{i=1}^{n} e_i^2, \\ &\text{subject to } y_i = w g(x_i) + b + e_i, \; i = 1, 2, \cdots, n, \end{aligned} \tag{A10}$$

where C is the penalty parameter, $e_i \in R$ is the error variable, $\boldsymbol{w}$ represents the weight vector, and $g(\boldsymbol{x}_i)$ indicates the nonlinear transformation function. The Lagrange equation can be established by introducing a Lagrange multiplier α into Equation A10 as follows:

$$L(w, b, e, \alpha) = J(w, b, e) - \sum_{i=1}^{n} \alpha_i [w g(x_i) + b + e_i - y_i]. \tag{A11}$$

By the partial derivatives of Equation A11 for different variables, the following equations can be obtained:

$$\begin{cases} \frac{\partial L(w,b,e,\alpha)}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i g(x_i) \\ \frac{\partial L(w,b,e,\alpha)}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i = 0 \\ \frac{\partial L(w,b,e,\alpha)}{\partial e_i} = 0 \Rightarrow \alpha_i = C e_i \\ \frac{\partial L(w,b,e,\alpha)}{\partial \alpha_i} = 0 \Rightarrow w g(x_i) + b + e_i = y_i \end{cases}. \tag{A12}$$

Equation A12 can be further simplified to the following formula by eliminating $\boldsymbol{w}$ and e_i:

$$\begin{cases} \sum_{i=1}^{n} \alpha_i (g(x_i)^T g(x_j)) + b + C^{-1} \alpha_i = y_i \\ \sum_{i=1}^{n} \alpha_i = 0 \end{cases}. \tag{A13}$$

Kernel function $K(\boldsymbol{x}_i, \boldsymbol{x})$ is introduced to Equation A13, thus the final functional expression of LSSVR is constructed as

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b^*. \tag{A14}$$

Similarly, the RBF is used here as a kernel function.

A.4. Kriging

The Kriging regression model consists of two components: a deterministic regression model and a stochastic process. Deterministic regression model is used to fit the overall trend of the data, while stochastic process is used to describe the random noise in the data. Given a set of inputs $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$ with $x_i \in R^n$ and outputs $\boldsymbol{Y} = [y_1, y_2, \ldots, y_n]$ with $y_i \in R$, their approximate relationship can be defined by the Kriging model as follows:

$$f(x) = g^T(x) w + z(x), \tag{A15}$$

where $\boldsymbol{g}(\boldsymbol{x}) = [g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), \ldots, g_m(\boldsymbol{x})]^T$ is an optional regression function and 0th order (constant) polynomial model is chosen in this study; $\boldsymbol{w} = [w_1, w_1, \ldots, w_m]^T$ denotes the regression parameters; $\boldsymbol{z}(\boldsymbol{x})$ is a Gaussian stochastic process and can be expressed as follows:

$$E(z(x)) = 0, \tag{A16}$$

$$\mathrm{Var}(z(x)) = \sigma^2, \tag{A17}$$

$$Cov[Z(x_i), Z(x_j)] = \sigma^2 R(x_i, x_j), \tag{A18}$$

where σ^2 represents the process variance and $R(\boldsymbol{x}_i, \boldsymbol{x}_j)$ represents the correlation function. The most commonly used Gaussian correlation function is employed here (Kaymaz, 2005), which is formulated as

$$R(x_i, x_j) = \exp \sum_{k=1}^{p} [-\theta_k (x_i^{(k)} - x_j^{(k)})^2], \tag{A19}$$

where p is the dimensionality of the input variable and θ_k is the correlation parameter.

Suppose $\boldsymbol{R} = (R_{ij})_{n \times n}$, $R_{ij} = R(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\boldsymbol{G} = (G_{ij})_{n \times m}$, $G_{ij} = g_i(\boldsymbol{x}_j)$, then $\boldsymbol{w}$ and σ^2 can be calculated as

$$\hat{w} = (G^T R^{-1} G)^{-1} G^T R^{-1} Y, \tag{A20}$$

$$\hat{\sigma}^2 = \frac{1}{n} (Y - G\hat{w})^T R^{-1} (Y - G\hat{w}). \tag{A21}$$

Using the maximum likelihood estimation, the optimal correlation parameter $\boldsymbol{\theta}^*$ can be obtained as

$$\theta^* = \arg\min_{\theta} \frac{1}{2} (n \ln \hat{\sigma}^2 + \ln(\det(R))). \tag{A22}$$

Once $\boldsymbol{w}$ and $\boldsymbol{\theta}^*$ are obtained, then the final Kriging model can be expressed as

$$\hat{f}(x) = g^T(x)\hat{w} + r^T(x) R^{-1} (Y - G\hat{w}), \tag{A23}$$

where $\boldsymbol{r}(\boldsymbol{x}) = [R(\boldsymbol{x}, \boldsymbol{x}_1), R(\boldsymbol{x}, \boldsymbol{x}_2), \ldots, R(\boldsymbol{x}, \boldsymbol{x}_n)]^T$ represents the correlation vector between the test point $\boldsymbol{x}$ and all training points.

Received: November 19, 2024. **Revised:** February 10, 2025. **Accepted:** February 12, 2025