PHYSICS-GUIDED FEATURE SELECTION FOR TEMPERATURE PREDICTION OF STATOR WINDING'S HOLLOW CONDUCTORS IN EVAPORATIVE COOLING HYDROGENERATORS

***Yiwei DING*[1,2], *Jian AI*[3], *Jinxiu CHEN*[1], *Lin RUAN*[1,2,*]**

[1]State Key Laboratory of High Density Electromagnetic Power and Systems, Institute of Electrical Engineering, Chinese Academy of Sciences, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Institute of Electrical Engineering and Advanced Electromagnetic Drive Technology, QiluZhongke, Jinan 250101, China

[*]Corresponding author: rosaline@mail.iee.ac.cn

Abstract: *Accurate temperature prediction for hollow conductors in evaporative cooling hydrogenerators is critical for design optimization but hampered by the high cost of experiments and simulations. To address this, this paper proposes a physics-guided surrogate modeling framework using experimental data from 12 conductor designs. An optimized Gaussian Process Regression (GPR) model is shown to outperform Random Forest and XGBoost, reducing RMSE by 16.2% and 9.7%, respectively. The framework identifies two physically distinct feature subsets for complementary use cases. A six-feature (6D) monitoring model that achieves* $R^2 = 0.893$ *and RMSE* $= 1.770\,^{\circ}C$ *under random 5-fold cross-validation, and a four-feature (4D) design model that obtains pooled* $R^2 = 0.616$ *and RMSE* $= 3.200\,^{\circ}C$ *under rigorous Leave-One-Group-Out (LOGO) validation, sufficient for ranking candidate designs. The analysis further identifies outlet measurements as information-leaking features that inflate within-design accuracy but degrade extrapolation to unseen geometries, highlighting the importance of causal feature selection for robust design-stage surrogate models.*

Keywords: *Hydrogenerator, Evaporative cooling, Surrogate modeling, Feature selection, Gaussian process regression, Particle swarm optimization*

1. Introduction

Cooling system performance in large hydrogenerators directly affects operational efficiency and service life in hydropower stations. Evaporative cooling technology has been successfully applied in large units such as Three Gorges and LiJiaXia projects due to its high heat transfer efficiency [1–3]. However, accurate conductor temperature distribution remains critical for performance optimization. During design, large-scale parametric studies spanning hollow conductor dimensions, coolant flow rate and pressure, and load conditions are required to identify optimal configurations. Methods for temperature field prediction thus face dual challenges of computational efficiency and accuracy.

Traditional methods cannot adequately meet these demands. CFD, while capable of describing flow-thermal processes, encounters generalizability challenges in boiling scenarios due to empirical

parameter sensitivity and prohibitive computational costs [4–7]. Empirical correlations based on lumped thermal networks suffer from accuracy degradation when operating conditions deviate from calibration boundaries [8]. Experimental testing, though direct, incurs substantial time and resource investments limiting design space exploration. These limitations motivate the development of efficient and accurate temperature prediction methods.

Surrogate modeling offers a promising alternative for reducing computational costs in engineering thermal management [9]. Physics-informed approaches exhibit better generalization than purely data-driven methods [10, 11]. GPR demonstrates robust performance under small-sample conditions and has been successfully applied in thermal systems.

However, existing surrogate models generally lack physical interpretability in feature selection. Common practices either use black-box screening or include all measurable quantities, increasing model complexity and overfitting risk while obscuring causal relationships. With highly correlated variables, statistical methods like LASSO may produce unstable or physically irrelevant selections. Evaporative cooling couples flow, heat transfer, and phase change, making physics-guided feature selection crucial for accuracy, extrapolation, and credibility. Physics-informed approaches using governing equations demonstrate advantages in extrapolation and generalization [12].

Implementing such physics-informed selection requires an algorithmic framework capable of evaluating feature subsets in context. For problems with moderate dimensions, limited samples, feature correlation, and interpretability needs, wrapper-based selection offers distinct advantages over filter and embedded methods [13, 14]. By coupling feature evaluation with model performance via cross-validation, wrapper methods systematically search combinations while accounting for dependencies [15]. This capability is particularly valuable when guided by physical constraints, enhancing both prediction accuracy and mechanistic interpretability.

Accordingly, this study proposes a physics-guided feature selection approach based on governing equations to construct a high-precision surrogate model for hollow conductor temperature prediction. Starting from energy, momentum, and mass conservation, the approach systematically identifies candidate features while ensuring physical completeness [16, 17]. Sobol sensitivity analysis and correlation analysis serve as interpretability tools to quantify feature contributions and interactions. Sequential Floating Forward Selection (SFFS) with cross-validation then optimizes the feature subset for prediction accuracy. The SFFS-optimized selection is compared with three conventional approaches: LASSO, Mutual Information, and Recursive Feature Elimination (RFE). Particle swarm optimization tunes hyperparameters of GPR, random forest, and XGBoost, enabling robust model comparison and selection.

2. Physical Modeling of Hollow Conductors

2.1. System Description and Thermal Mechanisms

Hollow conductors consist of highly conductive copper conductors with embedded rectangular cooling channels. Resistive losses form a linear heat source $q'(z)$ along the length, with heat conducting radially through the conductor to the inner wall and transferring to the fluid through convection or boiling at the inner surface. When the fluid temperature T_f is below the saturation temperature $T_{sat}(P)$, the system operates in single-phase forced convection mode. When the wall temperature T_w exceeds the

onset superheat for nucleate boiling, the system enters the nucleate boiling regime, generating two-phase flow. The flow pattern along the channel evolves through bubble flow, slug flow, annular flow, and other morphologies, leading to significant variations in heat transfer and pressure drop characteristics.

To reflect the coupling relationship between wall heat transfer and fluid heat exchange, a linear network of thermal resistances is employed. The heat transfer per unit length is expressed as:

$$q' = \frac{T_s - T_f}{R'_{\mathrm{cond}} + R'_{\mathrm{conv}}} \tag{1}$$

where T_s represents the equivalent heat source temperature or average solid temperature, $R'_{\mathrm{conv}} = 1/(h \cdot P_{\mathrm{wet}})$ represents the convective thermal resistance where P_{wet} denotes the wetted perimeter, and h is the single-phase or two-phase heat transfer coefficient.

2.2. Governing Equations

The thermal-hydraulic behavior of hollow conductors is governed by steady-state conservation of mass, momentum, and energy [18]. Under steady-state conditions, mass conservation requires a constant mass flow rate $\dot{m}$ along the channel, while the energy conservation relates heat input $q'(z)$ to the enthalpy rise:

$$\dot{m} \cdot \frac{dh}{dz} = q'(z) \tag{2}$$

In single-phase regions, this reduces to $\dot{m}c_{p,l}(dT_f/dz) = q'(z)$, while in two-phase regions the enthalpy change couples with pressure-dependent saturation properties. The pressure distribution is governed by momentum conservation, where the total pressure drop ΔP integrates frictional, accelerational, and gravitational components:

$$\Delta P = P_{\mathrm{in}} - P_{\mathrm{out}} = \int_0^L \left[\left(\frac{dP}{dz}\right)_f + \left(\frac{dP}{dz}\right)_a + \rho_m g \right] dz \tag{3}$$

These governing equations establish the dependencies between boundary conditions (P_{in}, P_{out}, T_{in}), flow parameters ($\dot{m}$), geometric parameters (A_c, D_h, P_{wet}), and thermal load (q') that collectively determine the temperature field. The systematic identification of the minimal independent parameter set based on these conservation laws is presented in Section 3.

3. Physics-Guided Feature Engineering

3.1. Candidate Parameter Identification

The temperature distribution in hollow conductors in evaporative cooling hydrogenerators is governed by the coupled action of three fundamental conservation equations: mass, momentum, and energy. Traditional physics-based methods solve these partial differential equations numerically (e.g., CFD) to obtain the full spatiotemporal temperature field. This work adopts a different strategy. Rather than solving these equations directly, the conservation laws are used to identify the minimal set of measurable macroscopic parameters that completely specify the thermal state.

From the conservation-equation analysis in Section II, this minimal parameter set comprises three physical categories:

The energy conservation equation requires specification of the heat source along the channel $q'(z)$ and mass flow rate $\dot{m}$. The momentum conservation equation, through coupling between pressure drop ΔP and flow rate $\dot{m}$, jointly determines the system's flow state.

The cross-sectional shape and dimensions of the channel determine the flow area A_c, wetted perimeter P_{wet}, and characteristic length D_h. These parameters appear directly in the convective heat transfer term of the energy conservation equation, the friction pressure drop term of the momentum conservation equation, and the velocity calculation in the mass conservation equation.

Integration of the energy conservation equation requires specification of the inlet temperature T_{in} as an initial condition. The momentum conservation equation requires inlet pressure P_{in} and outlet pressure P_{out} or equivalently the pressure drop ΔP to determine the pressure distribution along the channel, thereby affecting saturation temperature and phase change characteristics.

Based on this physical analysis, a candidate feature space containing 13 physical variables is constructed (Tab. 1). These parameters fall into five categories. Geometric parameters include inner channel width, height, area, aspect ratio, hydraulic diameter, and perimeter. The remaining categories are thermal load (power loss), flow (mass flow rate), boundary conditions (inlet/outlet temperature and pressure), and pressure characteristics (pressure drop). Here $T_{W_{20}}$ denotes the conductor surface temperature at the outlet thermocouple; its role in defining the two use-case subsets is detailed in Section 4.1. These 13 parameters ensure physical completeness from the perspective of conservation equations. The parameter set is intentionally over-redundant, including both basic geometric parameters (width, height) and derived quantities (hydraulic diameter, perimeter, area). Although derivable from basic parameters, these combinations play different roles in physical mechanisms and may encode nonlinear interactions.

Tab. 1. Candidate physical parameters, definitions, and ML feature mapping

Category	Symbol	Definition	Units	ML Feature
Geometric	w	Inner width of rectangular channel	m	inner_width
Geometric	H	Inner height of rectangular channel	m	inner_height
Geometric	A_c	$w \times H$	m^2	inner_area
Geometric	w/H	Width-to-height ratio	–	inner_aspect_ratio
Geometric	D_h	$2wH/(w+H)$	m	hydraulic_diameter
Geometric	P_{wet}	$2(w+H)$	m	perimeter
Thermal	P_{loss}	Total power loss ($= q' \cdot L$, $L = 5$ m)	W	loss
Flow	$\dot{m}$	Coolant mass flow rate	kg/s	coolant_flow
Boundary	T_{in}	Coolant temperature at inlet	°C	Tin
Boundary	$T_{W_{20}}$	Surface temp. at outlet thermocouple W20	°C	Tout
Boundary	P_{in}	Coolant pressure at inlet	kPa	P_in
Boundary	P_{out}	Coolant pressure at outlet	kPa	P_out
Pressure	ΔP	$P_{\text{in}} - P_{\text{out}}$	kPa	pressure_drop

3.2. Experimental Design and Dataset Construction

Twelve rectangular hollow conductor specifications with a length of 5 m are selected for experimental investigation using Latin Hypercube Sampling (LHS) [19] adapted to discrete space.

Because the hydrogenerator conductor design follows the GB/T 7672.1-2008 national standard [20] with discrete dimension specifications, a sliced super Latin sampling method is employed to map continuous LHS samples to the discrete design space.

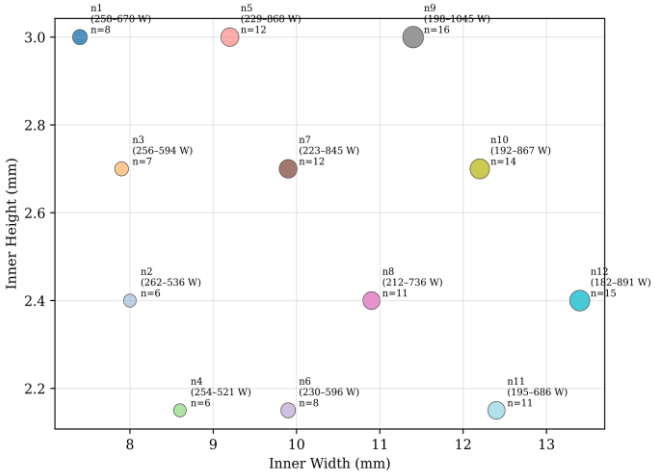


Fig. 1. Sampling distribution and power range of the 12 selected hollow conductor specifications. Each point represents a conductor specification, with annotations showing the power range (W) and number of valid operating conditions.

Experiments were conducted on a dedicated evaporative cooling platform. As shown in Fig. 2, the experimental platform is equipped with temperature, pressure, and mass flow rate sensors to detect bar inlet and outlet pressures, temperatures, and return liquid mass flow rates. Twenty T-type thermocouples are uniformly installed on the outer surface of each conductor to monitor temperature distributions along the conductor, with measurement accuracy of $\pm 0.5°C$. Experimental current increases in 25 A increments until the hollow conductor reaches its limiting thermal load, ultimately completing 153 experimental operating conditions.

Through establishing state validity criteria, 126 valid steady-state operating conditions were selected from the original 153 conditions for subsequent modeling. The filtering criterion enforces a strictly monotonic decreasing temperature profile at the outlet-region thermocouples:

$$T(W_i) > T(W_{i+1}), \quad i = 15,16,17,18,19 \tag{4}$$

Under steady-state evaporative cooling, the wall temperature decreases monotonically as the coolant approaches saturation; violations indicate transient boiling or measurement anomalies. Tab. 2 presents the per-design breakdown, with an overall exclusion rate of 17.6%

Tab. 2. Per-design breakdown of steady-state filtering

Design	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10	n11	n12	Total
Total	10	8	10	8	14	10	15	14	19	17	13	15	153
Valid	8	6	7	6	12	8	12	11	16	14	11	15	126
Excluded	2	2	3	2	2	2	3	3	3	3	2	0	27

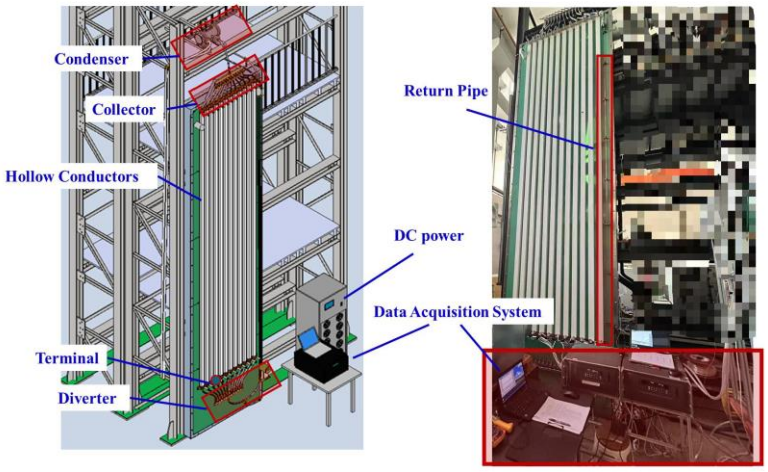


Fig. 2. The experimental platform for evaporative cooling.

3.3. Feature Selection

3.3.1 Sensitivity and Correlation Analysis

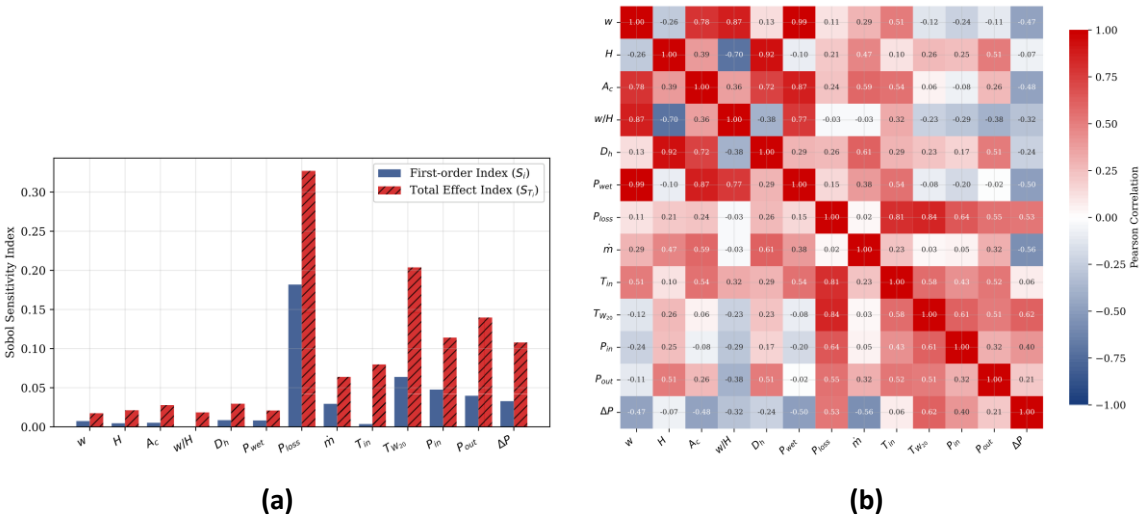
To provide physical interpretability for the feature selection process, Sobol global sensitivity analysis and Pearson correlation analysis are employed. Sobol analysis quantifies feature importance through variance decomposition, computing both first-order indices S_i for direct contributions and total-effect indices S_{T_i} that include interaction effects for all 13 candidate features. The difference $S_{T_i} - S_i$ reveals the extent of feature interactions.

Sobol indices were estimated using the Saltelli scheme with $M = 13$ features and $N = 1024$ base samples, with bounds set to the 1st–99th percentiles of experimental data. To bypass expensive direct evaluations, a GPR surrogate trained on 126 operating conditions was employed, achieving 5-fold cross-validation $R^2 = 0.858$, sufficient for ranking feature importance.

The preliminary GPR model outputs 20 temperature targets. To obtain a unified feature importance ranking, a median aggregation strategy across all targets is employed, which exhibits stronger robustness compared to arithmetic averaging.

Fig. 5(a) presents the Sobol sensitivity index results. Power loss exhibits the highest total-effect index, nearly an order of magnitude higher than other features, reflecting its dominant influence on temperature distribution. Outlet temperature and inlet pressure rank second and third in importance. The gap between first-order and total-effect indices for power loss indicates substantial coupling effects with other features.

Fig. 5(b) presents the Pearson correlation matrix among 13 candidate features. Multiple highly correlated pairs with $|r| > 0.85$ exist due to mathematical definitional relationships among geometric parameters and physical coupling between thermal load and temperature variables. These correlations indicate information redundancy that would reduce model stability and generalization if features with high correlation are included simultaneously.



(a) (b)

Fig. 5. Feature importance and correlation analysis: (a) Sobol first-order S_i and total-effect S_{T_i} sensitivity indices for the 13 candidate features; (b) Pearson correlation matrix showing multicollinearity among features.

3.3.2 Wrapper-Based Optimization

While sensitivity and correlation analyses provide physical interpretability, directly selecting the top k features according to Sobol ranking presents two limitations. First, this approach does not account for multicollinearity among features, potentially leading to information redundancy. Second, it overlooks synergistic effects of feature combinations, potentially missing complementary features. These limitations motivate a wrapper-based optimization strategy using Sequential Floating Forward Selection (SFFS) .

To address these limitations, a wrapper-based feature exchange strategy adapted from SFFS refines the initial subset. For a given k, the algorithm initializes with the top k Sobol-ranked features. It then iteratively evaluates all possible exchanges between selected and unselected features, accepting whichever exchange maximizes the 5-fold cross-validation R^2. The procedure terminates when no exchange yields improvement or after 20 iterations.

The optimization method described above is applied for $k \in \{3,4,5,6,7,8,9,10\}$, systematically evaluating the influence of different numbers of features on model performance. For each k value, the top k features are first selected in descending order of S_{T_i} as the initial subset, then the feature exchange optimization algorithm is applied to refine the initial subset.

Fig. 8 shows 5-fold cross-validation R^2 and RMSE versus feature count. Feature exchange optimization yields substantial improvements for small k, with gains gradually saturating as dimensionality increases. For $k \geq 6$, optimization gains are less than 0.05.As detailed in Tab. 3, the SFFS-optimized 6D subset achieves the best trade-off between accuracy and dimensionality, outperforming all baseline methods including Sobol-Top6, LASSO, MI, and RFE. The optimization replaces pressure-related parameters (P_{in}, ΔP, T_{in}) with geometric features ($\dot{m}$, w/H, w), which capture channel geometry effects on convective heat transfer while avoiding multicollinearity. Increasing to 7 or 8 features provides only marginal improvement of less than 0.2%, confirming $k = 6$ as optimal.

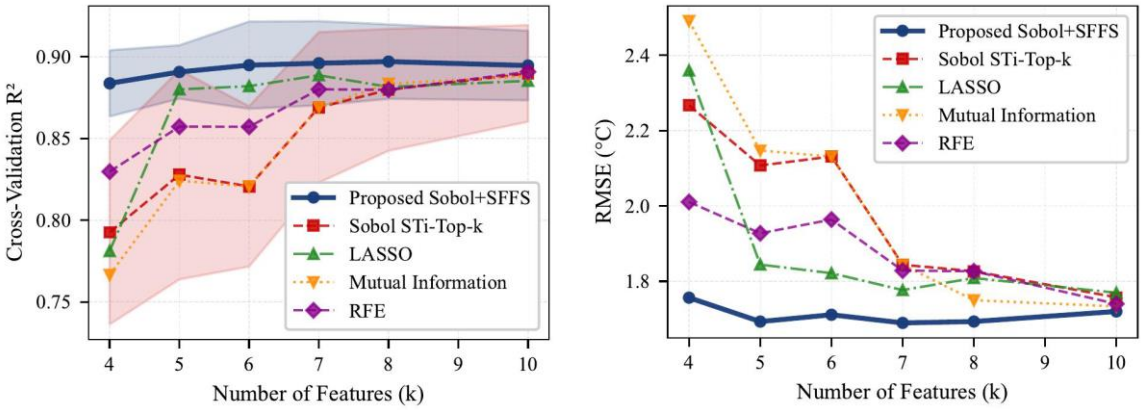


Fig. 8. Model performance (R^2 and RMSE) versus number of features for the initial (Sobol-ranked) and refined (SFFS-optimized) subsets.

The SFFS-4D design subset $\{P_{\text{loss}}, \dot{m}, w/H, w\}$ is physically grounded in the conservation equations. P_{loss}, equivalently $q' = P_{\text{loss}}/L$, drives the energy equation and $\dot{m}$ governs convective transport and Reynolds number. w determines the flow cross-section and w/H controls the hydraulic diameter and Nusselt number. These four features represent causal design inputs, whereas pressure drop is a downstream consequence of geometry and flow rate as expressed by the Darcy–Weisbach equation:

$$\Delta P = f(\text{Re}) \cdot \frac{L}{D_h} \cdot \frac{\dot{m}^2}{2\rho A_c^2} \tag{5}$$

where D_h and A_c are functions of w and H. The Pearson correlation between P_{in} and ΔP reaches $r = 0.87$ (Fig. 4), confirming this redundancy. Selecting causal inputs over downstream responses yields models that generalize better across unseen designs, as demonstrated by the LOGO cross-validation in Section 4.3.

4. Surrogate Modeling and Validation

4.1. Use-Case Scope and Feature Subset Definition

The SFFS results in Tab. 3 yield two physically distinct feature subsets. The four-feature design-stage subset (SFFS-4D: $\{P_{\text{loss}}, \dot{m}, w/H, w\}$) contains only parameters known a priori and constitutes the minimal physically complete input set from the governing equations. The six-feature monitoring subset (SFFS-6D: $\{P_{\text{loss}}, T_{W_{20}}, P_{\text{out}}, \dot{m}, w/H, w\}$) additionally incorporates outlet sensor measurements available in instrumented systems. The two subsets are distinguished by whether they include $T_{W_{20}}$, the conductor surface temperature at the outlet thermocouple. This quantity depends on energy-equation variables (T_{in}, q', $\dot{m}$, geometry) and conjugate heat transfer, making it unavailable at the design stage. Because $T_{W_{20}}$ is one of the 20 thermocouple targets $[T_{W_1}, \dots, T_{W_{20}}]$, using it as an input creates information leakage. The monitoring model therefore predicts only the remaining 19 targets $[T_{W_1}, \dots, T_{W_{19}}]$.

With the feature subsets and target definitions established for both use cases, an appropriate surrogate model must be selected. Three representative algorithms are compared: GPR [21], Random Forest (RF) [22], and Extreme Gradient Boosting (XGBoost) [23]. Each model utilizes the physics-optimized feature subsets identified in Section 3. To ensure fair comparison, Particle Swarm

Optimization (PSO) systematically tunes the hyperparameters of all three models under identical optimization criteria, enabling robust evaluation of their predictive capabilities.

Tab. 3. Comparison of feature selection strategies and selected feature subsets

Method	k	Pooled R^2	RMSE (°C)	Selected Features
SFFS-4D	4	0.878	1.809	$P_{\text{loss}}, \dot{m}, w/H, w$
SFFS-5D	5	0.891	1.717	$P_{\text{loss}}, P_{\text{out}}, \dot{m}, w/H, P_{\text{in}}$
SFFS-6D*	6	**0.893**	**1.770**	$P_{\text{loss}}, T_{W_{20}}, P_{\text{out}}, \dot{m}, w/H, w$
SFFS-7D	7	0.894	1.757	$P_{\text{loss}}, T_{W_{20}}, P_{\text{out}}, \dot{m}, w, w/H, T_{\text{in}}$
SFFS-8D	8	0.893	1.773	$P_{\text{loss}}, T_{W_{20}}, P_{\text{out}}, \dot{m}, w, w/H, P_{\text{wet}}, T_{\text{in}}$
Sobol-Top6	6	0.854	2.061	$P_{\text{loss}}, T_{W_{20}}, P_{\text{out}}, P_{\text{in}}, \Delta P, T_{\text{in}}$
LASSO	6	0.882	1.897	$P_{\text{loss}}, H, \dot{m}, T_{W_{20}}, w/H, T_{\text{in}}$
MI	6	0.854	2.061	$P_{\text{loss}}, T_{W_{20}}, P_{\text{in}}, T_{\text{in}}, P_{\text{out}}, \Delta P$
RFE	6	0.854	2.140	$D_h, P_{\text{loss}}, \dot{m}, T_{W_{20}}, P_{\text{in}}, \Delta P$

4.2. Model Architecture Selection

To ensure fair comparison across different surrogate model architectures, the hyperparameters of GPR, RF, and XGBoost were systematically optimized using PSO. Because PSO is formulated as a minimization problem, the optimization objective is defined as the negative of the average 5-fold cross-validation R^2 score:

$$f(\theta) = -\frac{1}{K}\sum_{k=1}^{K} R^2\left(Y_{\text{val}}^{(k)}, \hat{Y}_{\text{val}}^{(k)}\left(\theta, \mathcal{D}_{\text{train}}^{(k)}\right)\right) \tag{6}$$

where $K = 5$ is the number of folds, and θ represents the hyperparameter vector.All three surrogate models use the same evaluation procedure. Model inputs consist of the 6-dimensional physically optimized feature subset selected in Section 3, processed through Z-score standardization. The output target is a vector containing temperatures at 19 measurement points T_{W_1} through $T_{W_{19}}$, also standardized. $T_{W_{20}}$ is excluded from the target vector because it serves as an input feature.

The temperature targets along the conductor (19 for the monitoring model, 20 for the design model) are each modeled by an independent GPR with a Matérn kernel and individually optimized hyperparameters. This independent modeling approach allows hyperparameters to adapt to spatially varying heat transfer regimes, from single-phase convection near the inlet to two-phase boiling near the outlet, while maintaining $O(n^3)$ complexity per target rather than per joint output. Performance metrics (R^2, RMSE, MAE) are aggregated uniformly across all targets per cross-validation fold.

During the initial model screening phase, baseline hyperparameters are configured for each model. GPR employs an RBF kernel with length scale and noise variance automatically determined through maximum likelihood estimation. RF is set with 100 trees, unlimited maximum depth, and minimum split samples of 2. XGBoost is configured with 100 trees, learning rate 0.1, and maximum depth 6.

Model generalization performance is evaluated through a rigorous 5-fold cross-validation scheme on the complete 126 sample dataset. Evaluation metrics include three indicators: R^2, RMSE, and mean

absolute error (MAE). Tab. 4 presents the performance metrics of the three models after PSO optimization.

Tab. 4. Performance metrics after PSO optimization (random 5-fold CV, SFFS-6D monitoring subset, 19 targets excluding $T_{W_{20}}$)

Model	Pooled R^2	RMSE (°C)	MAE (°C)
GPR	0.893	1.770	0.914
Random Forest	0.845	2.113	1.242
XGBoost	0.871	1.960	1.103

Fig. 9 presents the predicted versus actual temperature scatter plots for the three surrogate models. The GPR model (left panel) exhibits the tightest clustering around the ideal prediction line across the entire 30–75°C temperature range. The Random Forest model (middle panel) shows notably larger scatter, particularly above 60°C, while XGBoost (right panel) demonstrates intermediate performance with visible deviation at temperature extremes. Quantitative metrics are summarized in Tab. 4.

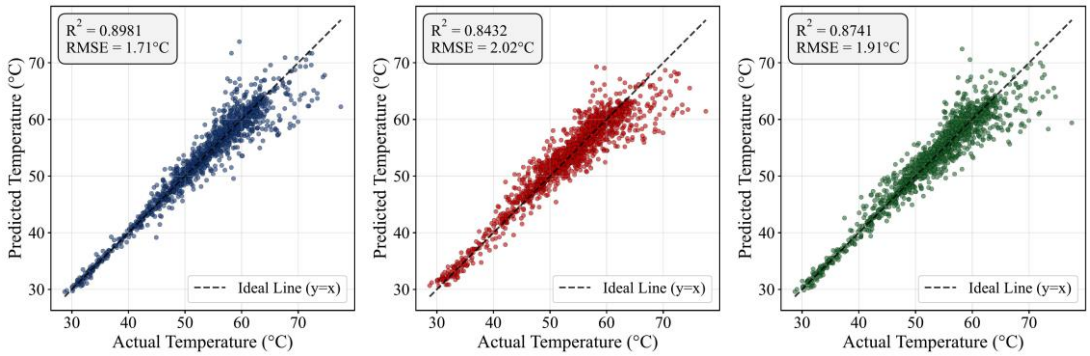


Fig. 9. Predicted versus actual temperatures for GPR, Random Forest, and XGBoost models. Results are aggregated from 5-fold cross-validation.

As shown in Tab. 4, the optimized GPR model achieves the best performance with substantial RMSE reductions compared to RF and XGBoost. This accuracy is sufficient for thermal design and online monitoring applications.

Beyond aggregate metrics, Fig. 12 illustrates the spatial prediction quality for two representative test samples. GPR accurately captures both peak temperature magnitude and location under high (n11) and low (n12) thermal loads, while RF and XGBoost exhibit noticeable deviations in the inlet heating and mid-section plateau regions.

The above results confirm GPR as the optimal surrogate model under random 5-fold CV, where training and test data share the same conductor designs. A more demanding question is whether the model can extrapolate to entirely unseen geometries, which is the design-stage use case defined in Section 4.1.

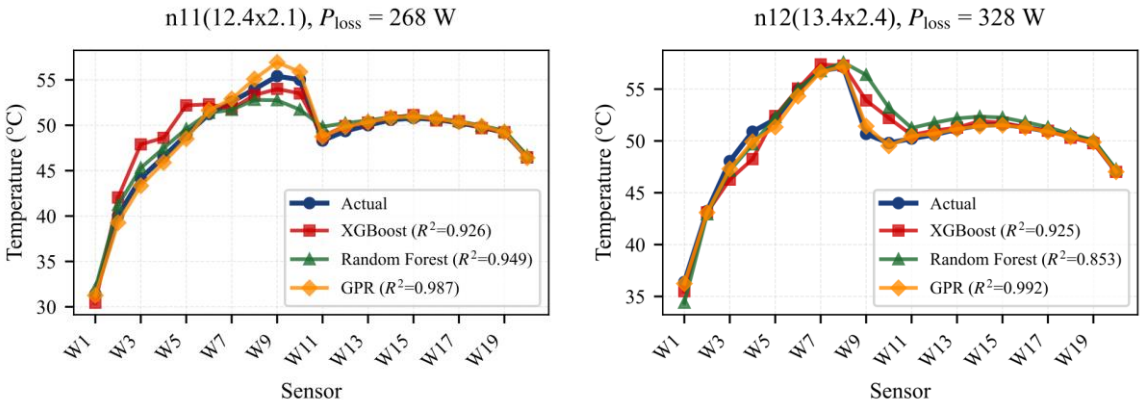


Fig. 12. Temperature distributions along the conductor predictions from GPR, RF, and XGBoost compared to ground truth for two test samples.

4.3. Cross-Design Generalization

To assess the design-stage use case defined in Section 4.1, namely generalization to unseen conductor geometries, Leave-One-Group-Out (LOGO) cross-validation is performed, where each of the 12 designs is held out in turn while the model trains on the remaining 11. Tab. 5 compares the validation protocols.

Tab. 5. Comparison of validation protocols for the optimized GPR model

Validation Protocol	Feature Set	Pooled R^2	RMSE (°C)
Random 5-fold CV	SFFS-6D (Monitoring, 19 targets)	0.893	1.770
LOGO CV	SFFS-6D (Monitoring, 19 targets)	0.602	3.219
LOGO CV	SFFS-4D (Design)	0.616	3.200

The LOGO validation reveals a substantial pooled R^2 drop from 0.893 under random 5-fold CV to 0.616 for the 4D design model, confirming that random CV inflates generalization estimates through design-specific information leakage. Under LOGO, the 6D monitoring model (Section 4.1) achieves a comparable pooled $R^2 = 0.602$, indicating that outlet measurements $T_{W_{20}}$ and P_{out} improve within-design interpolation but encode design-specific signatures that degrade extrapolation to unseen geometries.

The per-fold results in Tab. 6 corroborate this finding. The 4D model yields positive R^2 in 11 of 12 folds, ranging from 0.268 to 0.843, with only design n12 marginally negative at $R^2 = -0.037$. In contrast, the 6D model produces $R^2 = -0.294$ for design n1, confirming that outlet measurements mislead generalization to certain geometries. Design n11 exhibits the largest prediction errors for both models, with RMSE of 5.009 and 6.042 °C respectively. Its extreme aspect ratio $w/H = 5.77$, the largest in the dataset, places it at the boundary of the training distribution.

Overall, the 4D design model exhibits lower per-fold variance of ± 0.239 than the 6D model's ± 0.325. Its worst-case fold n12 at $R^2 = -0.037$ is also less severe than that of the 6D model's n1 at $R^2 = -0.294$. These results indicate that causal physical variables yield more robust cross-design transfer than downstream sensor measurements.

Despite the moderate pooled R^2, the design-stage model serves a practical role as a screening tool for ranking candidate geometries rather than providing absolute temperature predictions. With a LOGO-validated RMSE of $3.20\,^{\circ}\mathrm{C}$, this precision is useful relative to the 20–$40\,^{\circ}\mathrm{C}$ operational temperature gradients and $\pm 0.5\,^{\circ}\mathrm{C}$ thermocouple measurement uncertainty. The model can reliably identify promising designs, substantially reducing the need for expensive CFD simulations. The deployed model, trained on all 12 designs, is expected to outperform this leave-one-out estimate.

Tab. 6. Per-fold LOGO cross-validation results across 12 conductor designs

Design	n	SFFS-4D (Design)		SFFS-6D (Monitoring)	
		R^2	RMSE (°C)	R^2	RMSE (°C)
n1	8	0.268	1.717	-0.294	2.315
n2	6	0.381	1.547	0.611	1.146
n3	7	0.718	1.325	0.841	1.434
n4	6	0.502	2.101	0.553	2.212
n5	12	0.675	2.456	0.607	2.510
n6	8	0.843	1.406	0.759	1.630
n7	12	0.703	2.615	0.665	2.792
n8	11	0.334	2.569	0.138	2.938
n9	16	0.523	3.326	0.590	3.112
n10	14	0.728	2.876	0.778	2.668
n11	11	0.403	5.009	0.112	6.042
n12	15	-0.037	5.176	0.297	4.238
Per-fold mean		0.503	–	0.471	–
Per-fold std		± 0.239	–	± 0.325	–
Pooled		0.616	3.200	0.602	3.219

5. Conclusions

This paper proposes a physics-guided surrogate modeling framework for predicting hollow conductor temperatures in evaporative cooling hydrogenerators. The main findings are summarized as follows.

First, conservation-equation analysis identifies two physically distinct feature subsets for complementary use cases. The four-feature (4D) design subset, containing only causal parameters $\{P_{\mathrm{loss}}, \dot{m}, w/H, w\}$, achieves a pooled LOGO CV $R^2 = 0.616$ and RMSE $= 3.200\,^{\circ}\mathrm{C}$, sufficient for ranking candidate geometries. The six-feature (6D) monitoring subset, incorporating outlet sensor data, achieves $R^2 = 0.893$ and RMSE $= 1.770\,^{\circ}\mathrm{C}$ under 5-fold cross-validation.

Second, outlet measurements are identified as information-leaking features. While they improve within-design interpolation, they encode design-specific signatures that degrade generalization to unseen geometries. Causal design parameters generalize more robustly than downstream response measurements, a principle applicable to surrogate modeling in broader multi-physics contexts.

Third, among three candidate models, the optimized GPR outperforms Random Forest and XGBoost in RMSE by 16.2% and 9.7%, respectively. This confirms its suitability for small-sample, high-dimensional thermal modeling.

Future work should extend this framework to larger design spaces with nested cross-validation and investigate transfer learning strategies for geometrically novel designs at the boundary of the training distribution.

6. Nomenclature

A_c	Area	[m^2]	R^2	Determination coeff.	[–]
$c_{p,l}$	Liquid specific heat	[J/(kg·K)]	S_i	First-order Sobol index	[–]
D_h	Hydraulic diameter	[m]	S_{T_i}	Total-effect Sobol index	[–]
g	Gravity	[m/s^2]	T_f	Fluid temperature	[°C]
h	Heat transfer coeff.	[W/(m^2 ·K)]	T_{in}	Coolant inlet temp.	[°C]
H	Channel height	[m]	$T_{W_{20}}$	Outlet surface temp. (W20)	[°C]
K	CV folds	[–]	T_s	Heat source temp.	[°C]
L	Channel length	[m]	T_{sat}, T_w	Saturation/wall temp.	[°C]
$\dot{m}$	Mass flow rate	[kg/s]	w	Channel width	[m]
M, N	Features/samples	[–]	z	Axial position	[m]
P_{in}, P_{out}	Inlet/outlet pressure	[kPa]	ΔP	Pressure drop	[kPa]
P_{wet}	Wetted perimeter	[m]	ρ_m	Mixture density	[kg/m^3]
P_{loss}	Total power loss	[W]	θ	Hyperparameters	[–]
q'	Linear heat source	[W/m]	T_{W_i}	Thermocouple i temp.	[°C]
R'_{cond}, R'_{conv}	Thermal resistances	[K·m/W]			

References

[1] Gu, G., Ruan, L., Applications And Developments Of The Evaporative Cooling Technology In The Field Of Hydrogenerators, *Proc. CSEE*, *34* (2014), 29, pp. 5112-5119

[2] Yuan, J., et al., Applied engineering design of three gorges' 840MVA hydro-generator with close loop self circulating evaporative cooling system, *Proceedings*, Proceedings of the 11th international conference on electrical machines and systems (ICEMS), 2008, pp. 1-6

[3] Lin, R., et al., The evaluation of the design and operation of the 400MW evaporative cooling hydrogenerator in lijiaxia hydropower station, *Proceedings*, Proceedings of the 2015 18th international conference on electrical machines and systems (ICEMS), 2015, pp. 1-5

[4] Kurul, N., Podowski, M.Z., MULTIDIMENSIONAL EFFECTS IN FORCED CONVECTION SUBCOOLED BOILING, *Proceedings*, Proceeding of International Heat Transfer Conference 9, Jerusalem, Israel, 1990, pp. 21-26

[5] Krepper, E., et al., CFD Modelling Of Subcooled Boiling—Concept, Validation And Application To Fuel Assembly Design, *Nucl. Eng. Des.*, *237* (2007), 7, pp. 716-731

[6] Krepper, E., Rzehak, R., CFD For Subcooled Flow Boiling: Simulation Of DEBORA Experiments, *Nucl. Eng. Des.*, *241* (2011), 9, pp. 3851-3866

[7] Yang, J., Wang, Z., A Simple And Accurate Method For Estimating The Stator Winding Real-Time Temperature Of Air-Cooled Hydrogenerator, *Therm. Sci.*, *27* (2023), 1A, pp. 167-177

[8] Ruan, L., et al., Different Influence Of Cooling Method To Stator Bar Insulation Characteristics In Pumped Storage Units, *Trans. China Electrotech. Soc.*, *32* (2017), 14, pp. 246-251

[9] Ebbs-Picken, T., et al., Hierarchical Thermal Modeling And Surrogate-Model-Based Design Optimization Framework For Cold Plates Used In Battery Thermal Management Systems, *Appl. Therm. Eng.*, *253* (2024), pp. 123599

[10] Santiago Galicia, E., et al., Machine Learning-Driven Prediction Of Heat Transfer Coefficients For Pure Refrigerants In Diverse Heat Exchangers Types, *J. Exp. Theor. Anal.*, *3* (2025), 4, pp. 32

[11] Qiu, Y., et al., An Artificial Neural Network Model To Predict Mini/Micro-Channels Saturated Flow Boiling Heat Transfer Coefficient Based On Universal Consolidated Data, *Int. J. Heat Mass Transf.*, *149* (2020), pp. 119211

[12] Hughes, M.T., et al., Status, Challenges, And Potential For Machine Learning In Understanding And Applying Heat Transfer Phenomena, *J. Heat Transf.*, *143* (2021), 12, pp. 120802

[13] Saltelli, A., et al., Global Sensitivity Analysis. The Primer, *Proceedings*, December 18, 2007

[14] Sobol′, I.M., Global Sensitivity Indices For Nonlinear Mathematical Models And Their Monte Carlo Estimates, *Math. Comput. Simul.*, *55* (2001), 1-3, pp. 271-280

[15] Pudil, P., et al., Floating Search Methods In Feature Selection, *Pattern Recognit. Lett.*, *15* (1994), 11, pp. 1119-1125

[16] Wei, P., et al., Variable Importance Analysis: A Comprehensive Review, *Reliab. Eng. Syst. Saf.*, *142* (2015), pp. 399-432

[17] Sudret, B., Global Sensitivity Analysis Using Polynomial Chaos Expansions, *Reliab. Eng. Syst. Saf.*, *93* (2008), 7, pp. 964-979

[18] Rohsenow, W.M., et al., *Handbook Of Heat Transfer*, McGraw-Hill, New York, 1998

[19] Mckay, M.D., et al., A Comparison Of Three Methods For Selecting Values Of Input Variables In The Analysis Of Output From A Computer Code, *Technometrics*, *42* (2000), 1, pp. 55-61

[20] Standardization Administration of China, Glass-fiber wound winding wires. Part 1: Glass-fiber wound rectangular copper wire. General requirements, Report No. GB/T 7672.1-2008, Standards Press of China, 2008

[21] Rasmussen, C.E., Williams, C.K.I., *Gaussian Processes For Machine Learning*, MIT Press, Cambridge, MA, 2006

[22] Breiman, L., Random Forests, *Mach. Learn.*, *45* (2001), 1, pp. 5-32

[23] Chen, T., Guestrin, C., XGBoost: a scalable tree boosting system, *Proceedings*, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 785-794

Paper submitted: 26.01.2026

Paper revised: 27.03.2026

Paper accepted: 03.04.2026