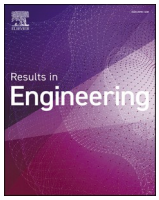
Results in Engineering 32 (2026) 113183



Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering



Research paper

Integrating dimensionless transport features with machine learning for evaporation rate prediction under mixed and forced convection

Ismail Masalha [a], Hamza Abu-Zienah [b], Omer A. Alawi [c,d], Omar Badran [e], Tibor Poós [b,*]

[a] Mechanical and Industrial Engineering Department, Applied Science Private University, Amman, 11937, Jordan
[b] Department of Building Services and Process Engineering, Faculty of Mechanical Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111, Budapest, Hungary
[c] Department of Thermofluids, Faculty of Mechanical Engineering, Universiti of Teknologi Malaysia, UTM Skudai, Johor Bahru, 81310, Malaysia
[d] Department of Power Mechanics Engineering Techniques, Technical Engineering College, Al-Bayan University, Baghdad, 10011, Iraq
[e] Department of Mechanical Engineering, Faculty of Engineering Technology, Al–Balqa Applied University, Amman, Jordan

ARTICLE INFO

Keywords:
Evaporation rate
Mixed-forced convection
Heat and mass transfer
Dimensionless transport features
Physics-informed machine learning
Held-out testing

ABSTRACT

Accurately predicting evaporation rate under mixed and forced convection remains challenging because of the interaction among heat, mass transfer, and flow dynamics. Conventional correlations are limited to specific conditions, while purely data-driven models often lack Physically Consistent Trends. This study proposes a Dimensionless Transport Feature-Based Machine Learning (DTF-ML) framework that integrates dimensionless parameters such as Reynolds, Richardson, and Schmidt numbers with data-driven models. The two utilized datasets are a limited dataset of 32 observations and a multi-source database containing 487 observations for model development, of which 462 remained after outliers were eliminated. An additional 32 observations were reserved exclusively for independent external testing. While, for the limited 32-point dataset, the Hybrid–Linear Regression model achieved $R^2 = 0.977$, MAE = 0.037, and RMSE = 0.043 using Leave-One-Out Cross-Validation. On the internal held-out test set of the extended database, CatBoost achieved $R^2 \approx 0.98$, $MAE \approx 0.02$, and $RMSE \approx 0.06$. On the independent 32-point external dataset, the Physics–CatBoost configuration achieved $R^2 = 0.983$, $MAE = 0.006$, and $RMSE = 0.021$. The PINN-like model shows competitive performance, while maintaining Physically Consistent Trends. The results demonstrate that incorporating dimensionless transport features improves prediction accuracy and model stability across different operating conditions. The proposed framework provides a data-supported approach to evaporation-rate prediction within the investigated operating-condition ranges.

1. Introduction

Evaporation with a free water surface is a simultaneous heat- and mass-transfer process including phase change at the liquid–air interface. The driving forces depend on the concentration difference between the liquid water and the bulk air; the evaporation rate is also governed by boundary layers and momentum, energy, and species transfer. As a result, evaporation cannot be described by vapor-pressure difference alone; gas-phase transport processes must also be considered [1–3]. The previous studies of forced-convection conditions illustrated that evaporation is affected by the vapor-pressure difference between liquid and bulk air, as well as air velocities [4–7]:

$$N = (a_1 + a_2 v_G)\left(p_{v,f} - p_{v,G}\right)^m \quad (1)$$

Where the exponent (m) represents the evaporation rate affected by air velocity and the driving force of vapor-pressure difference [8,9].

Other experimental studies have examined evaporation under both laboratory and experimentally operating circumstances. Evaporation can be represented in many applications, such as indoor swimming pools, where water temperature, air temperature, relative humidity, and airflow velocity strongly affect the evaporation rate [10,11]. Other experiments used wind tunnels and ducts to clarify the significance of airflow conditions and categorize convection types, such as natural, mixed, and forced [7,10,11]. Other studies have identified additional parameters that affect evaporation. Salinity can affect evaporation by driving the vapor pressure gradient, which can be decreased, while surface turbulence, such as waves, can change the interfacial transport

* Corresponding author.
E-mail address: poos.tibor@gpk.bme.hu (T. Poós).

https://doi.org/10.1016/j.rineng.2026.113183
Received 17 August 2026; Received in revised form 19 September 2026; Accepted 23 September 2026
Available online 24 September 2026
2590-1230/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Nomenclature

Symbol	Description
A	Surface [m^2]
c	Molar concentration [mol/m^3]
D	Diffusion coefficient [m^2/s]
Fr	Froude number [-]
g	Gravitational acceleration [m/s^2]
Gr	Grashof number for mass transfer [-]
Gu	Guckman number [-]
h	Step distance [-]
k	Mass transfer coefficient [m/s]
L	Characteristic length [m]
$\dot{m}$	Mass flow rate [kg/s]
M	Molar mass [g/mol]
MAE	Mean absolute error [-]
N	Evaporation rate [$kg/(m^2h)$]
p	Pressure [Pa]
r	Latent heat of evaporation [J/kg]
R	Ideal gas constant [8.314 J/(mol·K)]
R^2	Coefficient of determination [-]
Symbol	Description
RMSE	root mean square error [-]
Ra	Rayleigh number for mass transfer [-]
Re	Reynolds number [-]
Ri	Richardson number [-]
Sh	Sherwood number [-]
Sc	Schmidt number [-]
T	Temperature [°C]
v	Velocity [m/s]
x	Salt concentration [ppm]
y	Mass fraction [$kg_{moisture}/kg_{wet\ gas}$]
Y	Absolute humidity [$kg_{moisture}/kg_{dry\ gas}$]
δ	Relative error [-]
μ	Dynamic viscosity [Pa·s]
ρ	Density [kg/m^3]
ν	Kinematic viscosity [m^2/s]
φ	Relative humidity [-]
Φ_T	Temperature correction term [-]
Φ_P	Vapor pressure correction term [-]

and affect the air velocity [12–14]. These results showed how sensitive evaporation is to air velocity and the surrounding environmental factors.

Many studies have also investigated correlations to predict evaporation rate using experimental data, such as the heat and mass transfer analogy, the Dalton model using a driving force based on vapor pressure, and Sherwood number correlations expressed in terms of Reynolds, Schmidt, and buoyancy-related dimensionless parameters [15,16].

Many applications have used different relations, such as HVAC and energy calculations for indoor swimming pools [17–20]. However, some comparative studies have shown differences in evaporation-rate predictions across convection regimes and operating conditions, with relatively few correlations maintaining satisfactory accuracy across experimental databases [21,22]. Therefore, the extent to which the relevant geometry, operating conditions, convection regime, and governing transport processes are significantly affect reliability.

Numerical modeling and computational fluid dynamics (CFD) are another approach to investigate evaporation by solving the transport phenomena above the liquid surface. This approach can provide parameters above the surface, like velocity, temperature, and concentration distributions, based on boundary layer theory, but this approach is very sensitive and can be affected by turbulent flow and how interfacial momentum, heat, and mass transfer are represented [16,18,21]. These parameters are particularly significant in forced and mixed convection, where evaporation is dependent not only on bulk airflow velocity but also on the boundary layers developed and the near-interface flow structure. As a result, CFD can provide more detail than empirical correlations; however, its computational complexity and dependence on modeling assumptions may limit its use for rapid prediction across large operating ranges.

In engineering systems, machine learning (ML) provides a novel way to express nonlinear interactions. When the evaporation-rate relationship cannot be described by a suitable correlation formula, a data-driven approach allows interpretation from previous data without a predetermined functional form. Recent works in complex Multiphysics systems have shown that advanced data-driven approaches can determine coupled nonlinear relationships [23]. However, purely data-driven models may create predictions with known physical behavior, especially outside well-represented parts of the training domain, and they rely heavily on the amount and breadth of available training data.

These limitations raised interest in building physical knowledge into data-driven models. Physics-informed and physics-driven approaches introduce governing relationships, physical constraints, or trend predictions into the learning process. For example, mixed-loss modular regression has incorporated physical constraints, including boundary and monotonicity requirements, directly into model training [24]. Physics-informed learning architectures have similarly shown how to integrate a physics-based approach with nonlinear predictive models [25]. More recently, data–physics co-driven approaches have desired to integrate physical findings and observational data within a common framework. Generative data–physics fusion, for example, was applied to develop physically meaningful predictions under limited-data conditions [26]. Taken together, these advancements show a broader shift away from purely data-driven prediction and toward learning frameworks that incorporate physical knowledge through training constraints, input representation, or model structure.

Despite these improvements, the application of physics-guided machine learning to evaporation from open water surfaces under mixed and forced convection regimes is still comparatively limited. Previous evaporation correlations rely on mathematical formulations or coefficients specific to convection regimes. While purely data-driven approaches can determine non-linear relations for that purpose, this leaves a methodological gap between traditional mathematical correlations based on dimensionless numbers and data-driven approaches. In particular, the role of dimensionless parameters representing momentum, buoyancy, and species transport as inputs to machine-learning models, together with the preservation of physically consistent trends, requires further investigation.

This study addresses this problem by using a literature-based experimental dataset and an extended multi-source database. The dataset was derived from previously published studies that collected evaporation measurements, involving input data such as air temperature, surface temperature, humidity, and evaporation rate, followed by systematic screening and harmonization. The previous dataset yielded a general Sherwood correlation under mixed and forced convection [19]. Thus, this study performed the primary source-specific consolidation of the database before using it. For the machine-learning analysis, we represented the harmonized data using uniform transport variables and dimensionless features.

Therefore, this study aims to develop a transport-informed machine learning framework to predict evaporation rate under mixed and forced convection conditions. The methodology investigates how the physical representation of evaporation affects machine learning model performance. The unified modeling framework evaluates three input

representations: raw data, transport-based dimensionless features, and dimensionless parameters. Both linear and nonlinear machine learning models to investigate the interaction between input representation and model complexity. Furthermore, we develop a neural network with monotonicity constraints to incorporate expected physical relationships into the model without explicitly determining the governing transport equations. The predictive performance of the proposed approach is evaluated using both limited datasets and larger datasets compiled from multiple sources.

This study presents a framework that combines physical modeling and a machine learning approach for predicting evaporation rate under mixed and forced convection. The novelty lies in how the model represents physical variables and how this influences performance. The main contributions are:

- Use of dimensionless numbers and groups such as Reynolds, Richardson, and Schmidt numbers as model inputs instead of relying only on raw variables such as temperatures, velocities, humidities, and pressures.
- Systematic comparison between three types of inputs: raw variables, dimensionless transport features, and dimensionless parameters within the same modeling framework.
- Apply both linear and nonlinear machine learning models, with analysis of the performance between the models.
- Inclusion of a PINN-like model that incorporates physical relationships within the model structure without directly solving governing equations.
- Evaluation of model performance using two datasets of different sizes, with validation approaches suited to each case.

Fig. 1 summarizes the overall process flow and provides an overview of the Dimensionless Transport Feature-Based ML Framework used in this study to predict evaporation rate. The process starts with the experimental variables consisting of air velocity, air temperature, relative humidity, total pressure, and water surface temperature, along with the target output, evaporation rate, which is measured at the beginning. The raw data then undergo preprocessing. This is followed by thermophysical computations and the creation of dimensionless transport features that describe heat and mass transfer processes. These features are employed as input to three machine learning models, i.e., Linear Regression, Random Forest, and CatBoost, in order to make comparison of the performance of these models for predicting evaporation rate. The framework concludes by comparing the evaluation metrics and feature significance gained from these models using dimensionless transport features.

2. Methodology

In this work, a proposed hybrid model that combines a physical approach with machine learning models to forecast the evaporation rate under mixed and forced convection. In this way, the model is constructed using variables derived from heat and mass transfer, and machine learning methods may correct empirical errors.

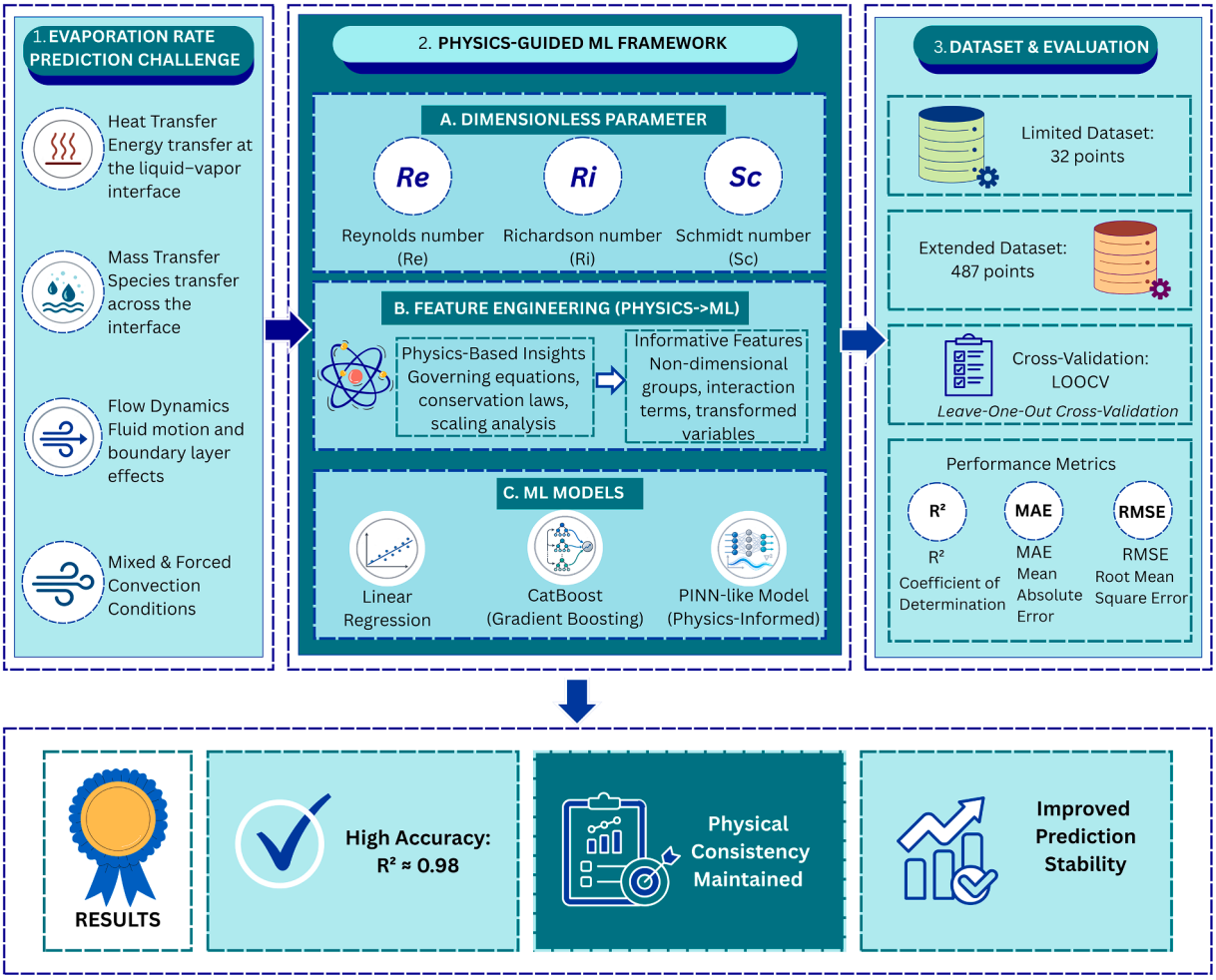


Fig. 1. Overview of the DTF-ML framework and complete research workflow.

2.1. Structure of the study and database design

This work comprises two parts that differ in aim and database size. The first part used a constrained dataset of 32 points from Varju and Poós [19]. Using a constrained dataset helped to study model behavior with limited data, which can be typical in mass transfer studies. We constructed a simple linear regression with raw operating variables, as well as a linear regression with dimensionless transport features, using governing physical equations derived from the dimensionless mass transfer analysis. We also coded a physics-based, correlation-based mass transfer model for direct comparison. The objective was to study the simple linear model and the linear model using dimensionless transport feature's ability to utilize limited data for learning their performance relative to a physical model with constant empirical coefficients. Moreover, we applied two additional nonlinear models, Random Forest and CatBoost, using dimensionless transport features to test nonlinear learners in a limited-data setting and to assess whether they can match linear performance. Leave-One-Out Cross-Validation was used to (a) distinguish the models in a fair manner and (b) avoid overfitting.

In the second part of this study, the multi-source database initially contained 519 observations. Then, the 32 observations reported by Gallero et al. were separated before preprocessing and reserved exclusively for independent external testing. The remaining 487 observations constituted the model-development database. After eliminate 25 outliers, 462 observations remained for model training and internal evaluation. Table 1

These studies were combined to develop the present extended database used to construct dimensionless transport features and improve the machine-learning model.

2.2. Data preprocessing and outlier handling

To guarantee the accuracy and validity of the data before analysis and simulation, several steps were implemented. Firstly, the integrity of the data was established utilizing the [isnull, sum] functions in the panda's library in all the columns in order to make sure there were no missing values. That's mean the data was organized for direct use without requiring imputation [19]. The second step entailed the removal of outliers that could have an adverse effect on the models' performance by employing the Interquartile Range (IQR) approach. Any values falling outside the statistical range $(Q1 - 1.5 \times IQR)$ or $(Q3 + 1.5 \times IQR)$were considered outliers and removed to decreased their adverse effect on the model. This approach assists to ensure a high level of predictive accuracy during model training. Data points lying outside the statistical bounds defined by [Q1-1.5×IQR, Q3+1.5×IQR] were categorised as outliers. These points were eliminated prior to model training to minimize noise and keep model stable and physically interpretable learning behavior. No outliers were detected in the limited dataset (32 points). For the expanded dataset, outlier detection was performed using the IQR method on the training data only. A total of 25 outliers were removed, reducing the dataset from 487to 462 points used for model training and evaluation. The IQR thresholds were calculated using the training data only. Twenty-five training observations were flagged, including 14 based on gas temperature and 11 based on liquid-surface temperature, while the testing data remained unchanged. A sensitivity analyses with and without filtering showed lower MAE and RMSE after filtering, without changing the model ranking.

Table 1
Number of data points and literature database.

Study	Number of Data Points	Ref.
Blázquez et al.	26	[20]
Asdrubali	33	[10]
Yen and Landvatter	59	[27]
Pauken	36 (47 total series available)	[7]
Raimundo et al.	24	[18]
Inan and Atayilmaz	9	[28]
Gallero et al.	32 for test	[29]
Evelin Varju et al.	32	[19]
Rohwer	268	[30]

2.3. Optimal hyperparameters and model performance comparison

Experiments assessed the predictive power of the best machine learning models for the evaporation rate. Bayesian Optimization optimized the most influential hyperparameters of Random Forest and CatBoost. Table 2 reports the improved hyperparameters and predictive performance on the test set.

CatBoost appears to perform better than Random Forest across all metrics (MAE, RMSE, R^2), providing more accurate predictions and generalizing well to the present dataset. We optimized hyperparameters such as decision-tree depth, learning rate, and number of iterations to obtain an optimal model for the present dataset.

2.4. Experimental data

The modeling framework is based on an experimental dataset compiled from multiple studies, consisting of 32 and 487 measured evaporation-rate data points. The dataset includes air velocity, air temperature, relative humidity, total pressure, liquid surface temperature, and measured evaporation rate. These variables define the operational domain of the open surface water evaporation process and form the basis for all subsequent physical and statistical analyses.

2.5. Evaluation of thermophysical properties and driving forces

The saturation vapor pressure at the liquid surface and the vapor in the ambient air are obtained from the Antoine equation. The vapor partial pressure in the bulk air depends on relative humidity and temperature, while at the liquid surface depends on surface temperature. These two satisfy the definition of the concentration-driving force for the evaporation rate. The transport properties and thermophysical properties of moist air were evaluated using correlations obtained experimentally [31], which compute the effects of temperature and relative humidity on the transport properties.

2.6. Derivation of governing dimensionless groups

The measured thermophysical properties were used to determine the essential parameter groupings that describe mixed convection mass transfer. These were the non-dimensional groups of Reynolds number, Grashof number, the Richardson number, and the Schmidt number, as well as non-dimensional parameters for the driving potentials of temperature and vapor pressure. These govern the regimes of forced convection, natural convection, and diffusion.

The Richardson number *Ri* can be defined as follows:

$$Ri = \frac{Gr}{Re^2} \quad . \tag{2}$$

The Grashof number *Gr* number in the *Ri* number can be calculated with the following relationship:

Table 2
Optimized hyperparameters and test set performance.

Model	Hyperparameters	MAE	RMSE	R^2
Random Forest	$max_depth = 12$, $max_features = 1.0$, $min_samples_leaf = 1$, $min_samples_split = 2$, $n_estimators = 1000$	0.0318	0.0502	0.9036
CatBoost	$border_count = 32$, $depth = 3$, $iterations = 1000$, $l_2_leaf_reg = 10.0$, $learning_rate = 0.3$	0.0226	0.0367	0.9486

$$Gr = \left(\rho_G^G - \rho_G^f\right)\frac{L_e^3 g \overline{\rho}_G}{\overline{\mu}_G{}^2} \ , \tag{3}$$

while the Reynolds number ***Re*** number can be determined with the following equation:

$$Re = \frac{v_G L_e}{\overline{\nu}_G} \tag{4}$$

The Schmidt number *Sc* number is a quantity expressing the ratio of the kinematic viscosity and the diffusion factor:

$$Sc = \frac{\overline{\nu}_G}{\overline{D}_{LG}} \ . \tag{5}$$

After determining the *Sc* number, the calculation of the correction terms resulting from the difference in temperature and humidity follows [19]:

$$\Phi_T = \frac{T_G^G + 273.15}{T_L^f + 273.15} \ , \tag{6}$$

$$\Phi_P = \frac{\Delta p_v}{p_{tot,G}} \ . \tag{7}$$

Finally, the Sherwood number *Sh* number can be calculated with the help of the mass transfer coefficient which was interpreted in terms of molar concentrations:

$$Sh = \frac{k_c L_e}{\overline{D}_{LG}} \ . \tag{8}$$

2.7. Assessment of the physical correlation

The Sherwood-number correlation for evaporation under mixed and forced convection is applied to evaluate the mass transfer coefficient and the corresponding theoretical evaporation rate. The theoretical predictions are quantitatively compared with experimental measurements using statistical error indicators. This step establishes a direct physical benchmark for subsequent machine learning model evaluation [19].

$$Sh_{corr} \ = \ 0.24\ Ri^{0.03}\ Re^{0.7}\ Sc^{\frac{1}{3}}\ {\Phi_T}^{-2.0}\ {\Phi_P}^{0.1} \tag{9}$$

In Eq. (9), the coefficient is 0.24, while the exponents of *Ri, Re, Sc,* Φ_T, and Φ_P are 0.03, 0.7, 1/3, −2.0, and 0.1, respectively. This fixed correlation was used without machine-learning-based recalibration. The resulting Sherwood number was converted to the mass-transfer coefficient using Eq. (10), and the evaporation rate was then calculated using Equations (11) and (12).

Within the development database used in this study, the correlation valid for these ranges [$1400 \le Re \le 1.85 \times 10^5$], [$2.34 \times 10^{-6} \le Ri < 7.4$], [$0.58 \le Sc \le 0.7$], [$0.88 \le \Phi T \le 1.154$], and [$3.0 \times 10^{-4} \le \Phi P \le 0.1$]. Predictions outside these ranges should be considered extrapolations.

$$k_c = \frac{Sh\overline{D}_{LG}}{L_e} \ . \tag{10}$$

In the evaporation process, uni-molar diffusion typically occurs in a two-component system, consisting of the liquid-vapor and the bulk air. When expressed in terms of molar concentration, the molar flux is given by:

$$N_c = k_c\left(c_{v,f} - c_{v,G}\right) \ , \tag{11}$$

where k_c is the mass transfer coefficient interpreted in terms of molar concentrations. Molar concentrations can be obtained using the ideal gas law, based on the local vapor pressure, the universal gas constant, and the air temperature. To convert the molar flux to the mass flux, the physics-based evaporation rate is expressed as:

$$N = M_L k_c\left(\frac{p_{v,f}}{R\left(T_L^f + 273.15\right)} - \frac{p_{v,G}}{R\left(T_G^G + 273.15\right)}\right) \tag{12}$$

Where M_L is the molar mass of the water and its 18.015 g/mol

2.8. Development of the conventional machine learning model

A preliminary machine learning model based on linear regression is built using raw operating parameters as input features. Given the small data set, Leave-One-Out Cross-Validation is used to test the model. This model serves as a point of comparison to quantify the value added by physics modeling. Under forced convection, the evaporation rate depends on the competing effects of the convective transport and vapor pressure. Gas velocity and temperature, liquid surface temperature, relative humidity, partial pressure, and total pressure are the main factors affecting the evaporation process and the evaporation rate. Therefore, the evaporation rate can be described as follows:

$$N = f(v_G, T_G, T_L, \phi_G, p_v, P_{tot}). \tag{13}$$

This functional relationship shows that many factors influence the evaporation rate, and no single parameter defines it. It depends on the interaction of the flow conditions. Convective transport depends on gas velocity, while the driving force is the water-vapor concentration difference between the gas and liquid surface. The vapor pressure gradient, which directly affects the rate of mass transfer, is defined by the total pressure and the relative humidity. Together, these factors determine the intensity of the convective mass transfer and the evaporation rate.

2.9. DTF-ML integration

To reformulate the physical correlation into a dimensionless transport feature representation for machine learning, the dimensionless groups derived from transport and thermophysical properties are used directly as input features to the machine learning model. In this structure, the model learns the relationship between physically meaningful dimensionless parameters and evaporation rate, rather than relying solely on raw variables. This strategy incorporates transport-theory information into the input representation while allowing experimental data to adapt empirical coefficients.

Dimensionless groups derived from transport theory such as Reynolds (*Re*), Richardson (*Ri*), and Schmidt (*Sc*) numbers, along with a temperature correction term (Φ_T) and vapor pressure correction term (Φ_P) are used as input features to the machine learning model. Following that, the model learns the relationship between these physically meaningful parameters and the evaporation rate (*N*), effectively adjusting empirical coefficients based on experimental data while retaining physical interpretability through dimensionless parameters

$$N = f\left(Re, Ri, Sc, \Phi_T, \Phi_p, T_L, T_G, v_G; \text{coefficients learned from data}\right). \tag{14}$$

where $f(\cdot)$ is learned using linear regression with Leave-One-Out Cross-Validation. This approach combines the interpretability of physics-based correlations with the adaptability of data-driven learning.

2.10. PINN-like model formulation

The proposed model is a PINN-like model in which the laws of physics behind heat and mass transfer are included in the learning process. Unlike classical PINN-like models, which solve the governing partial differential equations, the proposed model incorporates physical constraints through dimensionless parameters and physical relations implemented in the model. This approach ensures agreement with physical expectations, such as correct dependencies on the Reynolds number and driving forces, while also leveraging the flexibility of data-

driven models. The results show that the proposed model delivers strong predictive performance with physically consistent outputs, justifying comparison with purely data-driven models over the investigated data range.

In the process of developing the model, the basic physical relations of mass transfer were maintained, but instead of empirical Sherwood correlation, a data-driven version was used. In this way, the physical model allows for keeping physically expected trends of the system while having full flexibility to capture nonlinearity with machine learning.

The model does not include an empirical expression for the Sherwood number; instead, it predicts it from a neural network using a set of dimensionless parameters.

$$\log(Sh) = f_\theta(Re, Ri, Sc, \Phi_T, \Phi_p) \ . \tag{15}$$

where f_θ represents a nonlinear mapping learned by the neural network with trainable parameters theta. The logarithmic form develops numerical stability and reflects the multiplicative nature of transport correlations. The Sherwood number is then recovered as:

$$Sh = \exp\left[f_\theta(Re, Ri, Sc, \Phi_T, \Phi_p)\right]. \tag{16}$$

Once Sh is obtained, the mass transfer coefficient is computed as:

$$k_c = \frac{Sh\overline{D}_{LG}}{L_e} \ , \tag{17}$$

where $\overline{D}_{LG}$ is the diffusion coefficient and L_e is the characteristic length. The molar mass flux is expressed as:

$$N_{mol} = k_c(c_{v,f} - c_{v,G}) \ , \tag{18}$$

where $c_{v,f}$ and $c_{v,G}$ are the vapor concentrations at the surface and bulk air. These concentrations are calculated using the ideal gas law:

$$c_{v,f} = \frac{p_{v,f}}{R\left(T_L^f + 273.15\right)}, \ c_{v,G} = \frac{p_{v,G}}{R\left(T_G^G + 273.15\right)} \ . \tag{19}$$

The evaporation rate is then obtained by converting molar flux to mass flux:

$$N = M_L \, N_{mol} \ , \tag{20}$$

which leads to the final expression:

$$N = M_L k_c \left(\frac{p_{v,f}}{R\left(T_L^f + 273.15\right)} - \frac{p_{v,G}}{R\left(T_G^G + 273.15\right)} \right) \ . \tag{21}$$

The model is trained using a composite loss function:

$$l = l_{data} + \lambda_{mono} l_{mono} \ . \tag{22}$$

The data loss is defined as:

$$l_{data} = \frac{1}{n} \sum_{i=1}^{n} \left(N_{pred,i} - N_{exp,i}\right)^2 \ . \tag{23}$$

To implement physical behavior, monotonicity constraints are applied:

$$\frac{\partial N}{\partial Re} \geq 0 \ , \ \frac{\partial N}{\partial \Phi_p} \geq 0 \ . \tag{24}$$

These constraints are implemented as:

$$l_{mono} = ReLU\left(-\frac{\partial N}{\partial Re} \right)^2 + ReLU\left(-\frac{\partial N}{\partial \Phi_p} \right)^2 \ . \tag{25}$$

The prescribed monotonicity constraints are applied to trends that occur conditionally within the bounds of the observed data. An increase in Re generally promotes the forced-convective mass transfer process, and an increase in Φ_P promotes the vapor-pressure gradient driving force. Yet because Re and Ri are related as $Ri = Gr/Re^2$, varying Re and Ri independently is not always a feasible physical path in mixed convection. Thus, the constraints are imposed as soft penalties weighted by the parameter λ_{mono}. A high penalty can unrealistically suppress expected buoyancy-driven behavior and bias the prediction. The constrained model is therefore considered to favor expected local trends within the considered range.

Table 3 illustrates the input variables, pre-processing operations, derived quantities, and the training target for each model variant.

2.11. Internal and external validation

We validated the model separately on the two datasets. We applied leave-one-out cross-validation to the limited 32-point dataset because of its small sample size. For the extended analysis, we excluded the complete Gallero et al. dataset (32 observations) before preprocessing, feature construction, hyperparameter optimization, and model training. Hence, the remaining multi-source data were used for model development and internal evaluation, whereas the excluded study was used only for final external testing. This design prevents observations from the same experimental source from appearing in both the model-development and external-test datasets.

2.12. Performance metrics

The model performance is measured by mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2). These metrics, mentioned in Table 4, were used uniformly to evaluate the physical correlation of different studies, traditional machine learning, and the DTF-ML model. The percentage residual is defined as $r_i = 100(\hat{y}_i - y_i)/y_i$. MAPE represents the mean absolute percentage prediction error. The approximate 95% residual-error half-width is calculated as $U_{95} = 1.96\sigma_r$, where σ_r is the sample standard deviation of the percentage residuals. This calculation assumes an approximately normal residual distribution. Therefore, U_{95} describes prediction-error dispersion and should not be interpreted as experimental measurement uncertainty. We did not quantify experimental uncertainty because consistent uncertainty estimates were unavailable across the source studies.

The above-derived physical correlation and DTF-ML models are valid within the operating bounds defined by the dataset ranges (Re, Ri, Sc, temperature, humidity, and pressure conditions). These bounds encompass the experimental investigation of flow, airflow conditions, thermodynamic conditions, and the evaporation system's geometrical parameters. The bounds for validity refer to Reynolds number from 1400 *to* 1.85×10^5, Richardson number from 2.34×10^{-6} *to* 7.4, and

Table 3
Definition of the input configurations and preprocessing procedures used in the models.

Configuration	Input features	Normalization	Derived quantities and output
Raw	v_G, T_G, T_L, φ_G, p_{tot}	Standard Scaler for Linear Regression only. No scaling for Random Forest or CatBoost.	None.
Physics	Re, Ri, Sc, Φ_T, Φ_P	Standard Scaler for Linear Regression only. No scaling for Random Forest or CatBoost.	Calculated from operating variables and thermophysical properties.
Hybrid	v_G, T_G, T_L, φ_G, p_{tot}, Re, Ri, Sc, Φ_T, Φ_P	Standard Scaler for Linear Regression only. No scaling for Random Forest or CatBoost.	Combination of raw and dimensionless transport features.
PINN-like	Re, Ri, Sc, Φ_T, Φ_P	StandardScaler fitted only on the training inputs.	Predicts $log(Sh)$, then calculates Sh, k_c, and N.

Table 4
Statistical error evaluation metrics [32–34].

Stat. Param.	Statistical Criteria	Formula
RMSE	Root Mean Square Error	$\sqrt{\sum_{j=1}^{N}\frac{\left(y_j-\hat{y}_j\right)^2}{N}}$
MAE	Mean Absolute Error	$\sum_{j=1}^{N}\frac{\lvert y_j-\hat{y}_j\rvert}{N}$
R^2	Coefficient of determination	$1-\frac{\sum_{j=1}^{N}\left(y_j-\hat{y}_j\right)^2}{\sum_{j=1}^{N}\left(y_j-\bar{y}\right)^2}$
PCC	Pearson Correlation Coefficient	$\frac{\sum_{j=1}^{N}\left(y_j-\overline{y}_j\right)\left(\hat{y}_j-\overline{\hat{y}}_j\right)}{\sqrt{\sum_{1}^{N}\left(y_j-\overline{y}_j\right)^2}\sqrt{\sum_{1}^{N}\left(\hat{y}_j-\overline{\hat{y}}_j\right)^2}}$
U_{95}	Approximate 95% prediction -error half-width	$1.96\sigma_r$
e	Residual Error	$\sum_{j=1}^{N}\frac{\lvert y_j-\hat{y}_j\rvert}{y_j}$

Schmidt number from 0.58 *to* 0.70, including laminar to transitional mixed-convection regimes. Gas-phase conditions encompass gas velocities from 0.17 m/s *to* 5.7 m/s, gas temperatures from $-19\ ^{\circ}C$ *to* $79\ ^{\circ}C$, and relative humidity levels from 4% *to* 99%. Liquid surface temperatures encompass $1\ ^{\circ}C$ *to* $61\ ^{\circ}C$, and absolute total pressure 84.3 *kPa to* 101.4 *kPa*. Vapor pressure difference encompasses 25 *Pa to* 10,000 *Pa*, representing low and high evaporation drives. The evaporation surface area encompasses 0.015 m^2 *to* 1.09 m^2, and the characteristic length 0.12 *m to* 1.04 *m* [19]. The machine learning models using dimensionless transport features can't be valued when predictions extend outside these bounds.

2.13. Evaluation of nonlinear learning algorithms

To examine whether more advanced non-linear approaches can perform better than the well-known method, linear regression, in predicting the rate of evaporation, this research used Random Forest and CatBoost. Indeed, it's known that the process of evaporation is greatly dependent on many different parameters, such as flow speed, temperatures, humidity, and even pressures, where the interrelation between most parameters prevails non-linear behavior, and many thermal-fluid computations are non-linear originally.

The present study utilized the Random Forest algorithm, as it is a robust algorithm and is able to pick up complex nonlinear relations between variables. In addition, it is less prone to overfitting with sufficient data, which is why it serves as a reference point in many physical regression problems. CatBoost is a top-tier gradient boosting model chosen because it handles complex nonlinear features and performs well on small to medium-sized datasets. It provides stable training and is less picky about hyperparameters, making it a tough contender against traditional algorithms.

These two algorithms will serve two primary purposes. Firstly, these will be used to validate that within an evaporation process under these operating conditions, nonlinearity is prominent, hence making these complex learning approaches worthwhile. Second, they will be used to confirm that any discernible advantages of linear regression over other models are not coincidental and to assess them within the linear relationships evident in these variables under experimental conditions.

A consistent tuning protocol was applied to Random Forest, CatBoost, and the PINN-like model. We evaluated each tunable model using 30 Bayesian-optimization trials with five-fold cross-validation applied exclusively to the development training data. We used mean validation RMSE as the common optimization objective, and we did not use the held-out test data during hyperparameter selection. For the PINN-like model, the search covered 2-4 hidden layers, 16-64 neurons per layer, learning rates from 10^{-4} to 10^{-2}, and λ_{mono} values from 10^{-4} to 10^{-1}. The selected configuration contained two hidden layers with 64 neurons per layer, Tanh activation, a learning rate of 0.01, and $\lambda_{mono} = 2.3474 \times 10^{-4}$. Training used the Adam optimizer with a maximum of 5000 epochs and early-stopping patience of 300 epochs. We standardized the inputs using training-data statistics, but did not normalize the target evaporation rate.

3. Results and decisions

3.1. Pearson correlation matrix

The Pearson correlation matrix in Fig. 2 illustrates how each operating variable aligns with the rate of evaporation, *N*. For instance, *N* has a considerable tendency to vary with increases in air velocity, v_G (0.64), air temperatures, T_G (0.72), and surface temperature, T_L (0.68). These variables explain the expected impact on evaporation rates, as faster rates and increasing temperatures will lead to higher rates of mass transfer and augmented vapor pressure at surfaces, facilitating faster rates of evaporation. The rate of evaporation has a tendency to lower in cases where relative humidity, φ_G, is on the higher end (–0.62). The variables also tend to lower with increases in total pressure, P_{tot} (–0.58). All these effects will tend to lower rates of evaporation by suppressing the pressure driving force and raising rates of mass transfer. What is interesting in this data set is that surface temperature, T_L, and air temperature, T_G, tend to vary strongly (0.94). A strong anti-correlation between T_G and relative humidity, φ_G, (–0.90) is recorded, which implies that temperature increases co-occurred with low humidity in their experiments. Therefore, based on a correlation matrix , it can be confirmed that they chose good input variables in regard to expected effects on rates of evaporation, *N*.

3.2. Principal component (PC)

Based on Fig. 3, PC1 dominates, capturing 61.2% of the total variance, and *N* appears to follow this pattern of main directions. This demonstrates that *N* is subject to control by the main underlying physical phenomena identified by the first principal component. The direction of the vectors of air velocity v_G is very similar in direction to *N*, showing a strong positive correlation, which is logical since a faster air velocity means a larger mass transfer at the surface of the fluid, which enhances evaporation. Both the air temperature T_G and the surface temperature of the fluid T_L point in the same direction as *N*, showing a similar correlation, since an increase in these temperatures enhances evaporation, given by an increase in surface vapor pressure. However, the vector of total pressure, φ_G, which is directed in an opposite direction to *N*, demonstrates a similar correlation, showing a reduction in evaporation, given by a reduction in vapor pressure. Hence, this first component can be seen as a linear combination of an evaporation index, since in experiments characterized by a high PC1, there is a larger value of *N*, which is characterized by high air velocities, high temperatures, low humidities, and low pressures, which in turn enhance evaporation.

3.3. OAT sensitivity analysis

Fig. 4 provides a One-At-a-Time (OAT (sensitivity analysis, which studies how the evaporation rate *N* changes when you vary each operating variable in isolation from the others. Numbers indicate the relative change in *N* for a unit change in each input; hence, they explain the ranking of the importance of each variable to a given model.

In Linear Regression, the total pressure P_{tot} stands out with a sensitivity of 4.1, meaning small pressure shifts cause big swings in predicted evaporation. The following are the liquid surface temperatures T_L with 1.0076, air temperature T_G with 0.7587, and air velocity v_G with 0.6708. Relative humidity φ_G is much less influential (0.0060), suggesting that within the linear framework and the data range studied, humidity barely

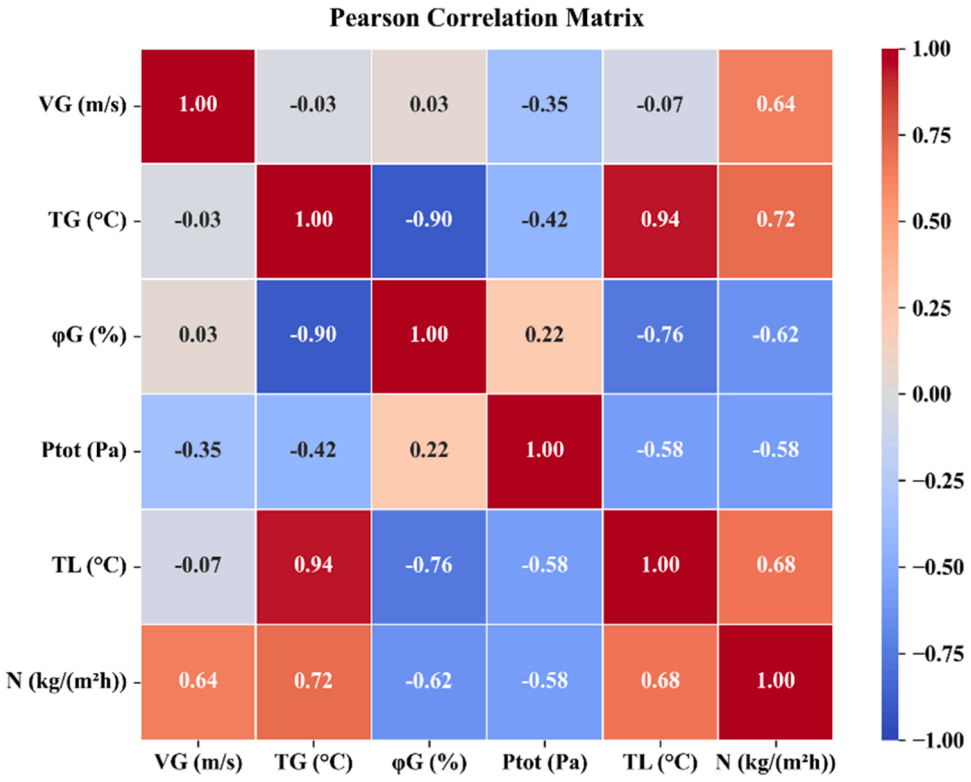


Fig. 2. Pearson correlation matrix of operating parameters.

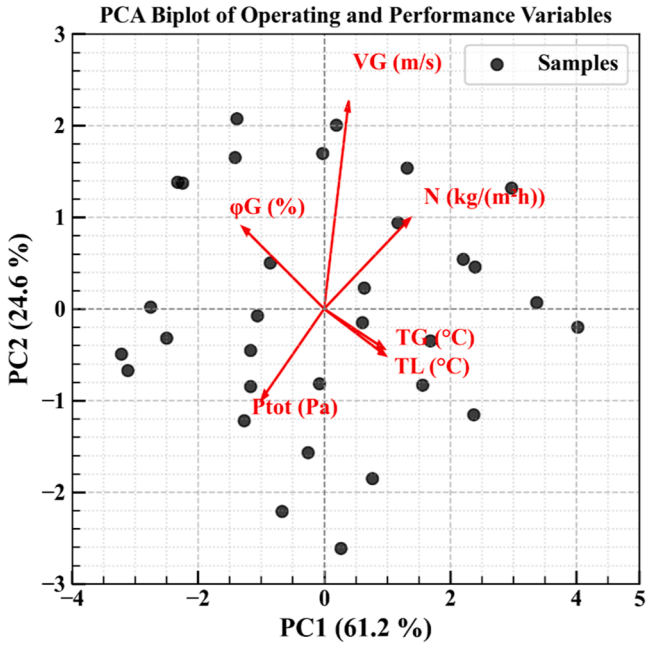


Fig. 3. Principal component analysis of operating parameters influencing the evaporation rate.

affects the evaporation rate. Overall, this pattern shows the linear model mainly links evaporation behavior to thermal and pressure-driven effects. For RF, the ranking order is different; P_{tot} becomes the most important input with 2.1958, followed by T_L at 1.4112, φ_G at 0.3064. While ν_G and T_G have lower sensitivities: 0.2019 and 0.2380, respectively. This reflects the tree-based model's ability to capture nonlinear relationships and emphasize parameters most directly linked to interfacial evaporation.

In CatBoost, the lead is taken by P_{tot} with 1.5917, then T_L with 1.1966, and ν_G with 0.8970 . The value of φ_G indicates a medium effect, 0.4023, whereas for T_G, it is low, 0.2805. This suggests that CatBoost balances convective transport against thermodynamic driving forces. As expected, sensitivity analysis of each OAT model highlights the consistency of the three models in terms of the key drivers: total pressure, liquid surface temperature, and air velocity. In contrast, humidity and air temperature rank differently across models, reflecting each algorithm's representation of linear and nonlinear relationships in the data and further indicating that sensitivity analyses are important for interpreting ML models of physical evaporation processes.

3.4. *Performance on limited dataset*

Table (5) presents model performance on a limited dataset of 32 points, which clearly represents a data-scarce scenario. Under these conditions, simple models such as Linear Regression perform better when combined with physical information. The Hybrid | Linear Regression model achieved the highest accuracy with (R^2 = 0.977) and low errors (MAE = 0.037, RMSE = 0.043). This indicates that dimensionless transport features improved predictive performance for the evaluated limited dataset.

In the case of more complicated models such as Random Forest and CatBoost, the drop in results compared to linear models is quite noticeable, with the (R^2) score falling within the range of 0.73 – 0.91. This suggests that these models need large amounts of data to capture non-linear features.

The PINN-like model showed acceptable performance ($R^2 \approx 0.917$), which is lower than the hybrid linear model; however, it remains acceptable because it includes implicit physical constraints. This shows that physics-informed models require large datasets to achieve their full potential.

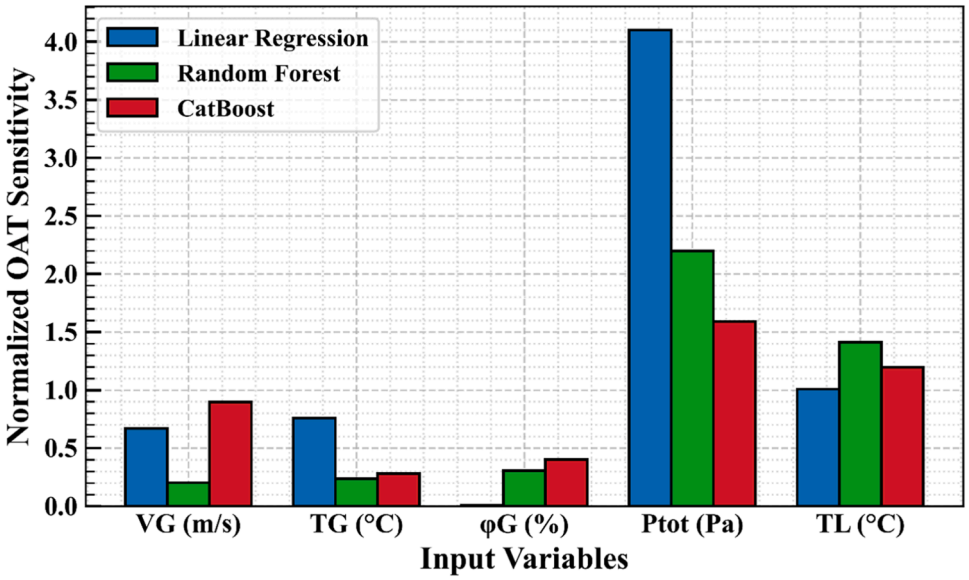


Fig. 4. Normalized OAT sensitivity of input variables for linear regression, random forest, and catboost models.

Table 5
The performance of the models using a limited dataset of 32 points.

Feature_Set	Model	MAE	RMSE	R^2
Hybrid	Linear Regression	0.037	0.043	0.977
Hybrid	PINN-LIKE	0.062	0.083	0.917
Raw	Linear Regression	0.063	0.079	0.926
Physics	Linear Regression	0.071	0.090	0.904
Hybrid	CatBoost	0.068	0.090	0.902
Raw	CatBoost	0.078	0.101	0.878
Physics	CatBoost	0.078	0.102	0.874
Physics	Random Forest	0.101	0.131	0.794
Raw	Random Forest	0.117	0.149	0.734
Hybrid	Random Forest	0.121	0.150	0.728

3.5. Performance and scatter plots on expanded dataset

Based on the different study results (Table 6 and Fig. 5), input representation and model type notably affect predictive power. In particular, the use of dimensionless transport features in combination with either raw inputs or without them demonstrates the best accuracy of predictions when CatBoost model is used. The gap between the Physics and Raw + Physics setups is insignificant, meaning dimensionless features effectively capture the system's physical behavior and including raw data does not provide a significant performance gain.

Raw inputs alone also show lower prediction accuracy, showing that embedding physical knowledge into the input features is crucial for better results. The same trend holds for the Random Forest model, although it performs worse than CatBoost because the models capture nonlinearity differently.

In linear regression, performance is considerably lower in all cases because it cannot capture the nonlinearity of the mass transfer process, despite using dimensionless transport features.

The PINN-like model with a monotonic constraint is able to provide competitive results ($R^2 \approx 0.974$), but it still underperforms compared to the top-performing CatBoost models. These results arise from imposing physical constraints, which make the model less flexible and lead to more physically consistent trends in the predictions.

Overall, the ablation study (Table 6) suggests that using physics as features has a larger effect on performance than using it as an in-model constraint. Although physics-constrained models improve Physically Consistent Trends, hybrid models using dimensionless transport features and machine learning models provide the best results.

Fig. 6 shows the Taylor diagrams of model performance in predicting evaporation rate for the training and test sets. The diagrams use three metrics: standard deviation, correlation, and distance from the reference point. A model close to the reference point best agrees with the experimental data and makes the most accurate predictions.

In the training-set diagram, Linear Regression already shows a significant distance from the reference point and a higher standard deviation. This shows a weaker ability to reproduce the data dispersion, as well as a weaker correlation with the measured values. The other ML models show greater similarity to the point, with stronger correlation with the experimental values. Most notable is the CatBoost model, which is closest to the point and shows the highest correlation between the measured and experimental values.

The testing set diagram shows the same results as the previous diagram, indicating the model's generality and how well it generalizes to new, unseen data. For the testing data, the linear regression model is far

Table 6
Ablation study results for feature representation and model comparison without data leakage.

Feature_Set	Model	Train_MAE	Train_RMSE	Train R^2	Test_MAE	Test_RMSE	Test R^2
Raw + Physics	CatBoost	0.01	0.01	0.99	0.02	0.06	0.98
Physics	CatBoost	0.01	0.01	0.99	0.03	0.06	0.98
Raw	CatBoost	0.01	0.01	0.99	0.04	0.08	0.97
Physics	Random Forest	0.01	0.02	0.99	0.04	0.09	0.97
Raw + Physics	Random Forest	0.01	0.02	0.99	0.04	0.09	0.97
Raw	Random Forest	0.01	0.03	0.99	0.04	0.09	0.97
Raw	Linear Regression	0.14	0.19	0.72	0.16	0.23	0.77
Raw + Physics	Linear Regression	0.05	0.08	0.95	0.09	0.29	0.61
Physics	Linear Regression	0.06	0.09	0.94	0.11	0.32	0.53
PINN-like	PINN-like	0.01	0.02	0.99	0.02	0.08	0.97

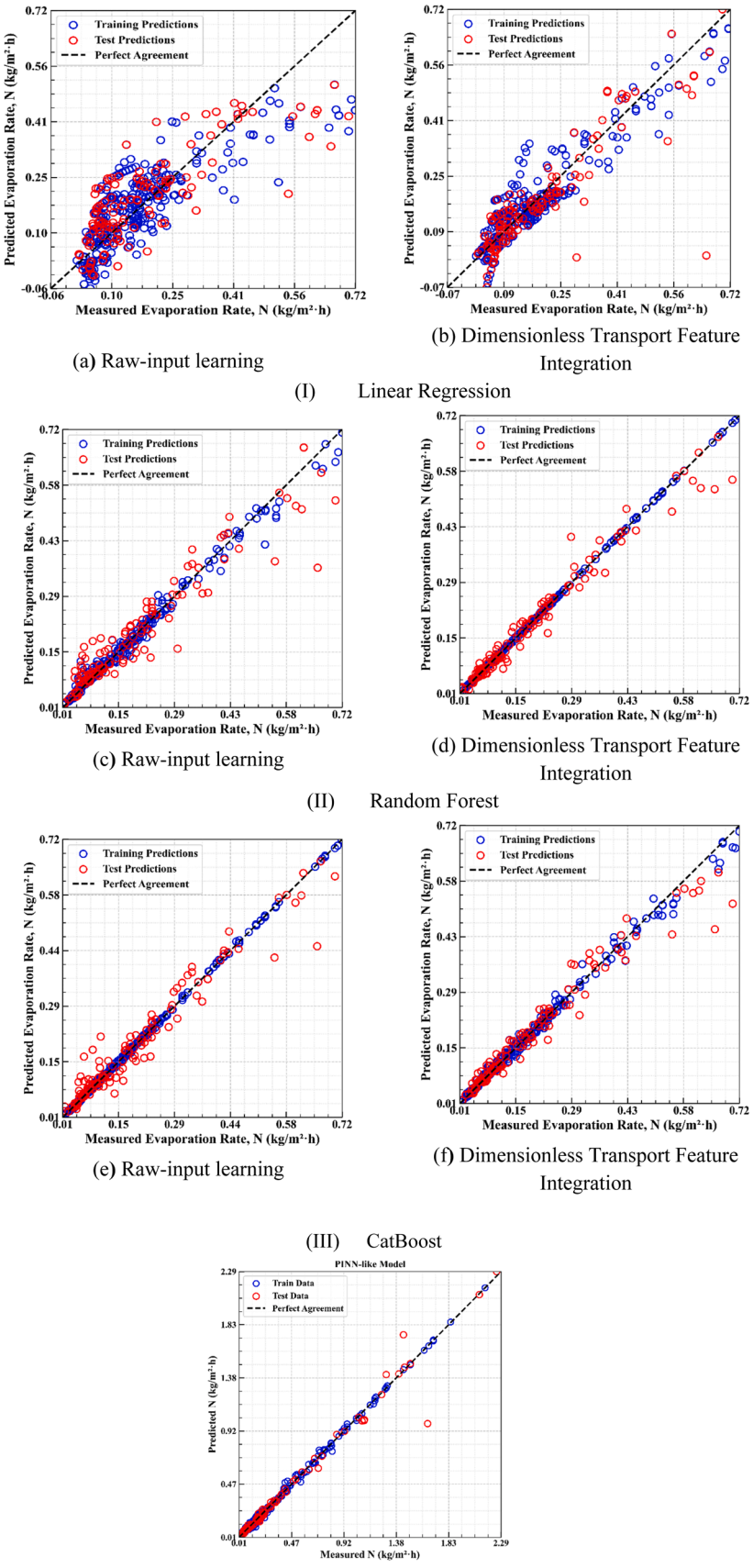


Fig. 5. Comparison of model performance without physics integration (raw-input learning) and with dimensionless transport feature integration (dimensionless transport features).

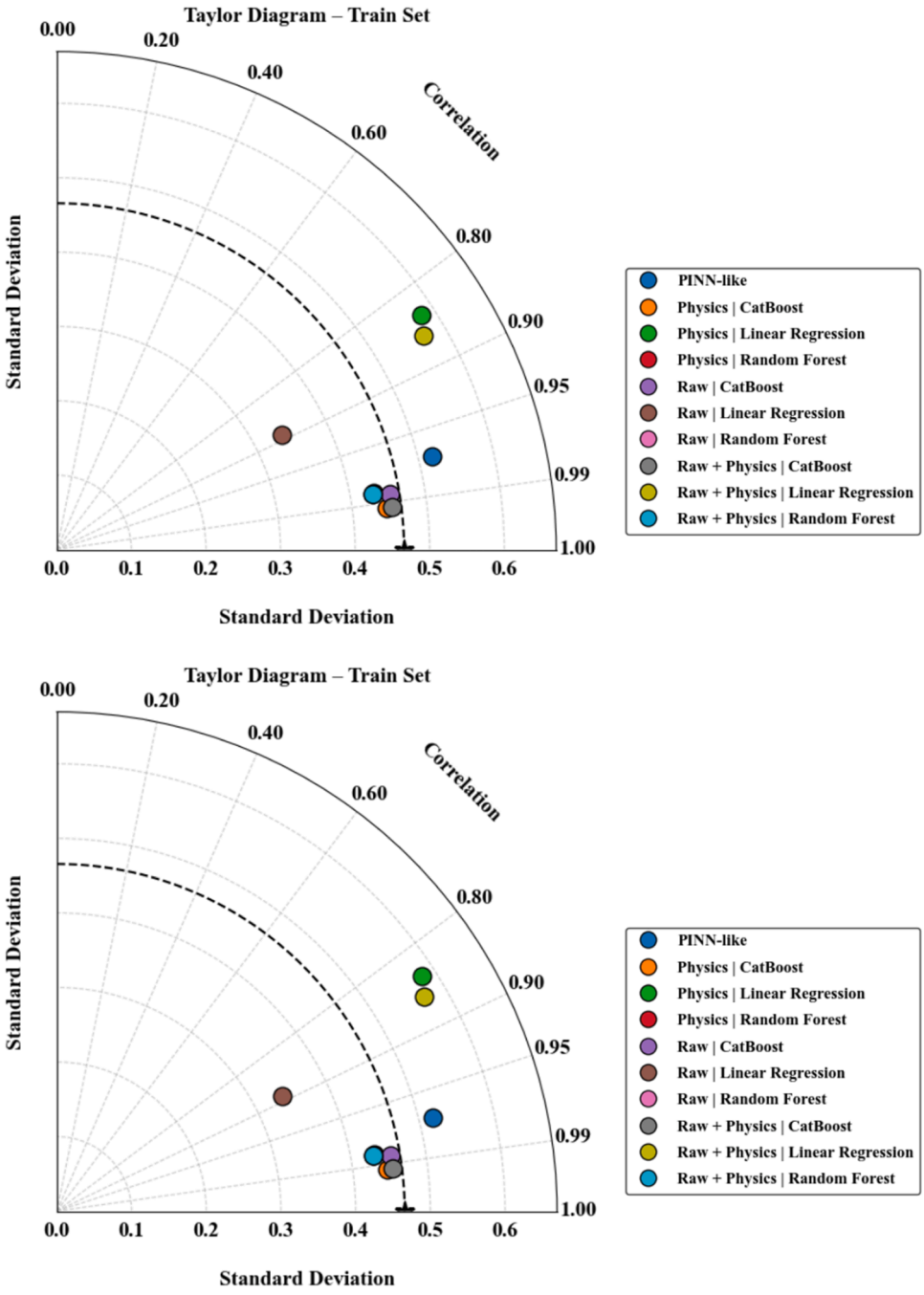


Fig. 6. Taylor diagram comparing the performance of evaporation rate prediction models.

from the reference point, indicating that the model prediction is unreliable. The Random Forest model remains close to the reference point, as it was on the training set, indicating a stable, consistent model with generalization capability. CatBoost is again closest to the reference point, indicating it can stay consistent and accurate when predicting new experimental data not present in the training dataset.

In summary, Fig. 6 shows that, compared to linear regression, the machine learning models provided the best variation and correlation with the experimental data. The CatBoost model gives the best compromise between standard deviation and correlation to give the most reliable and accurate predictions of evaporation rate for both training and testing conditions.

Fig. 7 illustrates the distribution of predicted values compared to measured values for both the training and testing datasets. In the training set, most models show clear agreement with the measured data, as indicated by similar median positions and data spread. This suggests the model can learn system behavior well. The best-performing models are CatBoost with physics-based or hybrid features, and they show a narrow, centered distribution compared with linear regression, which has a wide distribution and indicates its inability to capture nonlinear dependencies.

In contrast, models show high variance on the test set. CatBoost and Random Forest models still show a narrower distribution than the measured one, while other models, including linear models, show a wide distribution, indicating poor generalization. The PINN-LIKE model also shows a relatively narrow distribution, but its accuracy is lower than that of the top-performing tree-based models.

Overall, it can be demonstrated that the inclusion of dimensionless transport features in the prediction increased the stability and decreased dispersion of the results, especially for machine learning models. In

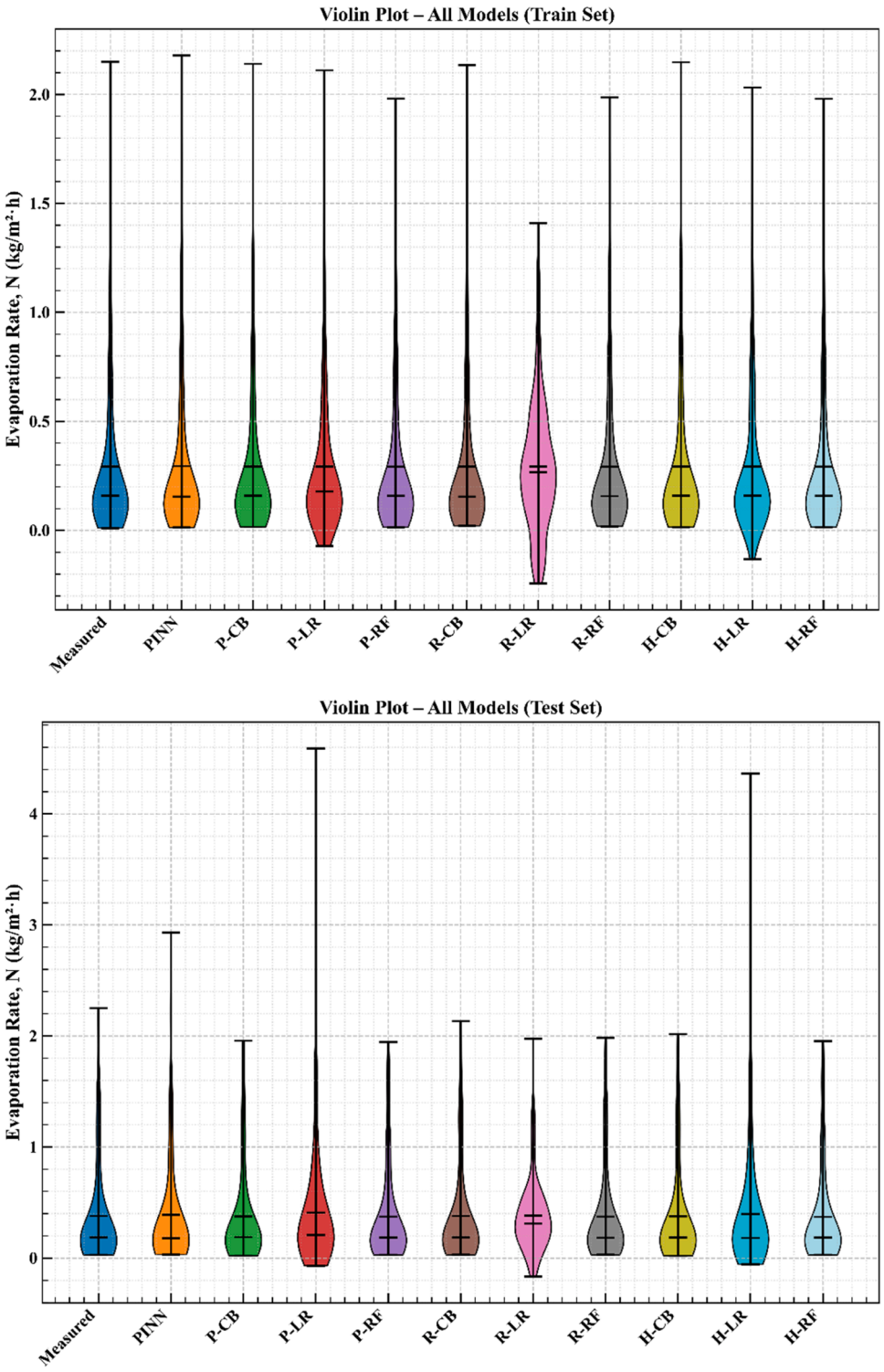
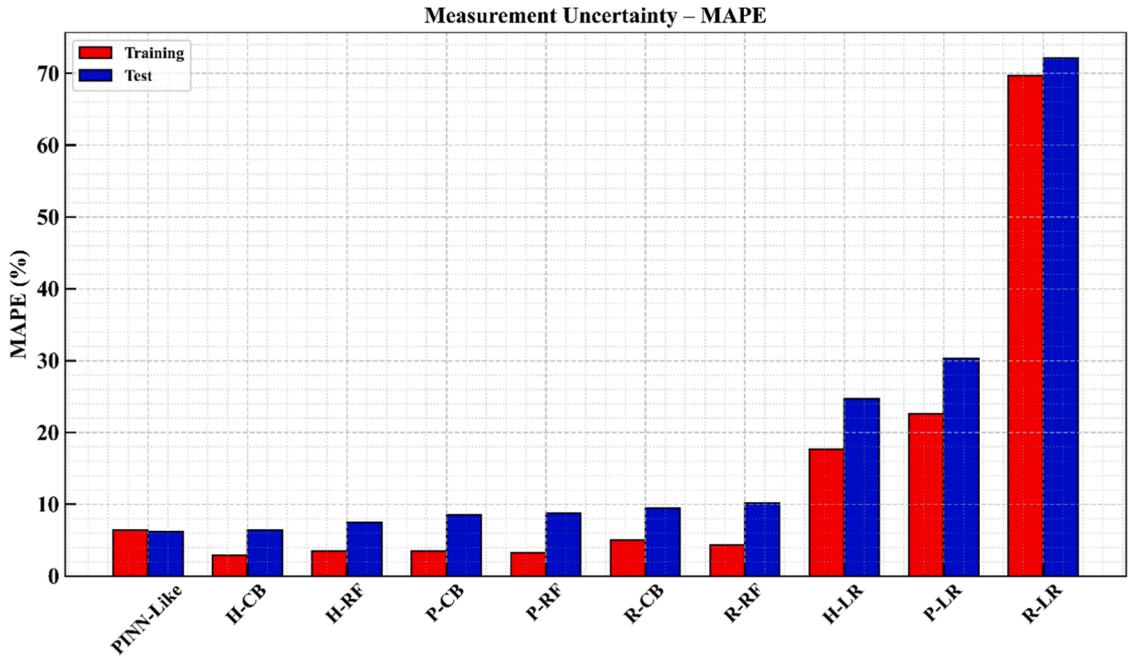


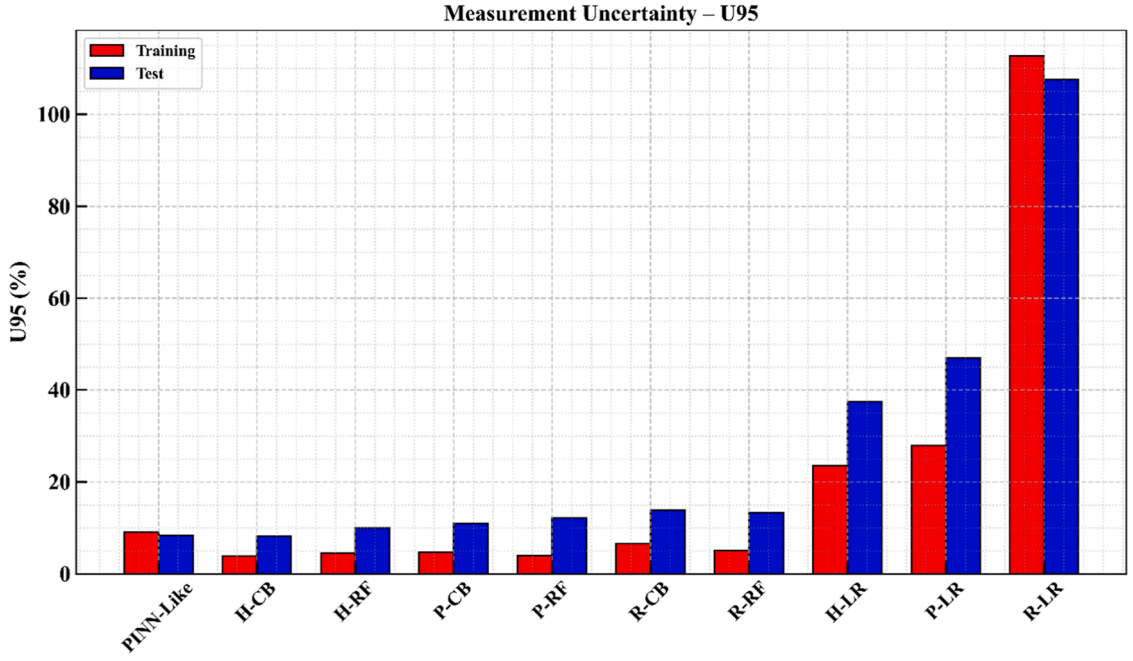
Fig. 7. Violin plot comparison of measured and predicted evaporation rate distributions.

addition, the real differences between the models are evident in the test dataset.

As shown in Fig. 8, nonlinear models clearly outperform linear models in uncertainty reduction and prediction stability. In particular, the PINN-like model provided the best results on the test data sets, with a Test MAPE of 6.19% and a U_{95} of 8.34%, indicating well generalization and consistent predictive performance. Similarly, CatBoost and Random Forest models predicted the test dataset using a combination of Raw and Physics features, with Test MAPE ranging from 6.3% to 7.5%. These results show that combining physically relevant features with raw data improves prediction accuracy and reduces uncertainty. At the same time, the usage of either Physics or Raw inputs only resulted in an increase in error because the Test MAPE exceeded [8–10] %, which meant that some information was lost. Linear regression models showed significant prediction limitations because the Test MAPE exceeded 70%. Overall, these findings show that nonlinear machine learning models supported by appropriate physical representations, particularly hybrid approaches, provide the best balance between predictive accuracy and



(a) Prediction Error – MAPE



(b) 95% Residual-Error Half-Width – U_{95}

Fig. 8. Comparison of MAPE and the approximate 95% residual-error half-width for the training and test predictions.

uncertainty reduction for such applications.

Fig. 9 shows the normalized residual histograms of three predictive models. Relative residuals give a more thorough picture of a model's reliability than measures of correlated predictions, such as the RMS error statistic. This is because residuals show the distribution of model prediction errors. Ideally, they will be symmetrically distributed about zero, with narrow distributions in both the training and testing sets. The residuals for the linear regression model appear broadly distributed about zero, with considerable skewness evident in the long tails extending to large negative and positive residuals. The linear model clearly estimates variability poorly. However, because of the residuals spread for the training and testing data is similar, this model does not overfit much. While the residuals are broadly distributed, the linear model is an ineffective predictor of the nonlinear interactions between the variables governing evaporation.

The Random Forest model shows clear improvement. The training residuals are tightly clustered near zero, fitting the training data well. However, the test-data residuals are slightly broader, indicating a slight drop-off in generalization on unseen data. This suggests moderate overfitting: the Random Forest can extract information from the training data but struggles to transfer it to new operating conditions. The Cat-Boost and PINN-LIKE residuals show the best behavior: the training

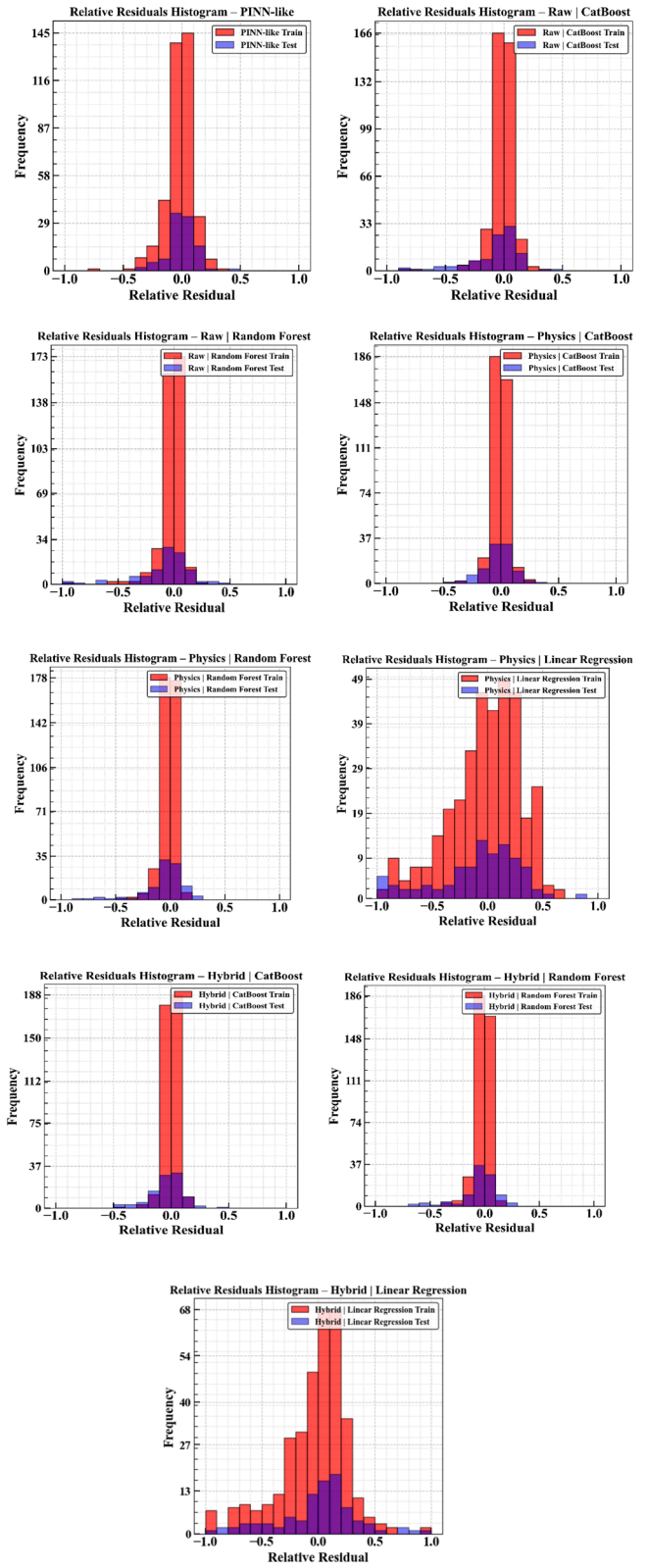


Fig. 9. Relative residual distributions for training and test sets using linear regression, Random Forest, CatBoost, and PINN-LIKE Models.

residuals are tightly peaked close to zero, while the residual distribution of the test data is equally tightly peaked with a nearly symmetric shape. This robust residual distribution indicates strong generalization, with little systematic bias. The close match between the training and test residual distributions indicates that CatBoost can model complex nonlinear relationships while remaining stable on unseen data.

Overall, the residual plots in Fig. 9 show that CatBoost and PINN-LIKE provide the most balanced residual structure, avoiding the extremes of high accuracy and high generalization seen in Random Forest and Linear Regression, respectively.

Table (7) and Fig. (10) present the performance of the models when evaluated on new experimental data that were not used during training, thus providing an independent assessment on observations excluded from model development. The results show that CatBoost-based models achieve the highest accuracy, where the Physics CatBoost model provides the best performance with ($R^2 \approx 0.983$) and a very low error ($RMSE \approx 0.021$), indicating strong agreement between predicted values and the new experimental measurements. Similarly, the Raw | CatBoost and Hybrid | CatBoost models perform very similarly, confirming that CatBoost can capture complex relationships even without fully embedded physical features.

The PINN-like model performs very well ($R^2 = 0.982$) with a low RMSE, nearly matching the best-performing model. Thus, we conclude that it performs competitively on the tested independent dataset with an incorporated monotonicity penalty term, making the model more physically realistic than data-driven models. Although its error is slightly higher than CatBoost, but it performs better from a physical perspective.

Random Forest models perform well, with (R^2) values from 0.96 to 0.97, indicating strong predictive performance and accurate predictions; however, compared with CatBoost models, they are less accurate. In this case, using dimensionless transport features does not improve performance; thus, Random Forest appears less efficient at using physical information.

Compared with previous model types, linear regression models perform poorly; their R^2 values are below 0.7 in most cases, and they have relatively high MAE and RMSE. Thus, this model fails to describe the nonlinearity of the evaporation process.

In summary, the CatBoost models show the best correlation with the new experiment, while the PINN-like approach balances prediction performance with physical meaning. This comparison shows that incorporating physics knowledge improves prediction performance.

4. Potential application scope and limitations

Within the investigated operating ranges, the DTF-ML framework may support evaporation-rate estimation while retaining transport-based information. However, its performance should not be interpreted as evidence of applicability to systems or conditions not represented in the database. Future studies could examine the framework in evaporative cooling systems, cooling towers, solar desalination systems, and evaporation basins. Such applications require validation using data that represent their specific geometries, fluids, boundary conditions, and operating ranges. The present findings should not be directly extended to lakes, irrigation channels, reservoirs, industrial drying, or process-control systems because the evaluated database did not include these applications. Additional experimental data and application-specific

Table 7
Model performance on the independent 32-point external test dataset.

Feature_Set	Model	MAE	RMSE	R^2
Physics	CatBoost	0.006	0.021	0.983
PINN-like	PINN-like	0.010	0.021	0.982
Raw	CatBoost	0.007	0.021	0.982
Hybrid	CatBoost	0.006	0.022	0.981
Hybrid	Random Forest	0.015	0.029	0.967
Physics	Random Forest	0.015	0.029	0.966
Raw	Random Forest	0.017	0.031	0.963
Hybrid	Linear Regression	0.057	0.078	0.759
Physics	Linear Regression	0.065	0.089	0.681
Raw	Linear Regression	0.074	0.107	0.542

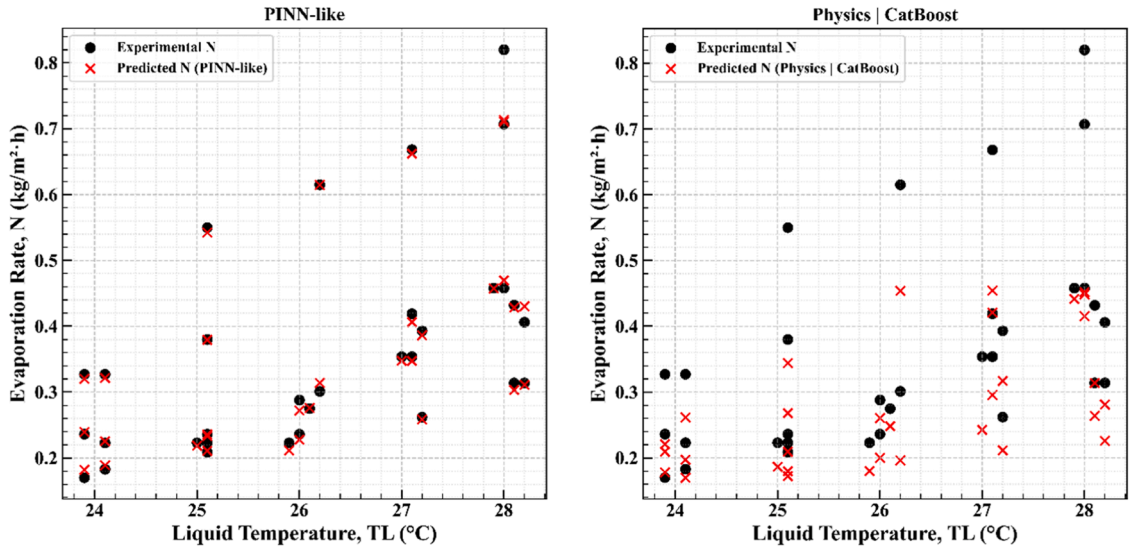


Fig. 10. Combined comparison of experimental and predicted evaporation rates versus liquid temperature for the two best models.

validation are required before engineering implementation.

This study shows that dimensionless transport features can be incorporated into machine-learning models for evaporation-rate prediction. Predictions outside the investigated ranges constitute extrapolation and should be independently validated before practical use.

5. Conclusion and limitations

In this research, we developed and verified a DTF-ML technique to estimate the evaporation rate of fluids under mixed and forced convective regimes. In the proposed method, mass-transfer correlations, originally expressed as empirical equations, were converted into dimensionless transport features for use in learning algorithms.

We considered two data regimes to assess predictive performance and generalization. For the limited 32-point dataset, the Hybrid Linear Regression model achieved the best performance, with an *MAE* of 0.037 $kg/(m^2{\cdot}h)$, an *RMSE* of 0.043 kg/(m²·h), and an $R^2$of 0.977, compared with an *MAE* of 0.063 kg/(m²·h), an *RMSE* of 0.079 kg/(m²·h), and an $R^2$of 0.926 for the Raw Linear Regression model. For the expanded multi-source dataset, nonlinear models substantially outperformed Linear Regression on the held-out test data. CatBoost models using dimensionless transport features achieved the highest predictive performance, with test R^2 values of approximately 0.98, while the PINN-like model achieved a competitive test R^2 of approximately 0.97. These results indicate that the effectiveness of the input representation depends on both the available data regime and the learning algorithm.

This clearly shows how classical evaporation relations can be transformed into smart predictive tools through incorporating mass transport computations with machine learning algorithms. The suggested framework offers an interpretable methodology that minimizes the need for expensive experiments without compromising physical consistency and can be extended to other coupled phenomena with new representative data and independent verification.

This approach does not impose the governing equations during training and should thus be considered a physics-guided but not physics-constrained model. The suggested methodology is limited by the range of conditions present in the dataset.

Reproducibility checklist

Code availability: All source code used to generate the data, evaluate all physical properties, perform dimensionless analysis, train the models, and evaluate their performance will be made available on an open-access repository with the submission of the present manuscript. https://github.com/Masalha-84/Physics-Guided-Evaporation-ML/tree/main

Availability of data sets: The experimental database assembled from published literature will be published before the publication of this thesis. It was used for model development, validation, and external prediction.

Hyperparameter disclosure: We enumerate all optimized hyperparameters and model parameters in the manuscript to facilitate reproducible implementation of the proposed framework.

Fundings

This work was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (BO/00059/23/6) and by the Hungarian Scientific Research Fund (NKFIH FK-142204).

CRediT authorship contribution statement

Ismail Masalha: Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Hamza Abu-Zienah:** Investigation, Formal analysis, Data curation. **Omer A. Alawi:** Writing – review & editing, Validation, Supervision. **Omar Badran:** Supervision, Resources, Project administration. **Tibor Poós:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] W.K. Lewis, The evaporation of a liquid into a gas, Trans. Am. Soc. Mech. Eng. 44 (Jan. 1922) 325–332, https://doi.org/10.1115/1.4058175.

[2] B.M. Smolsky, G.T. Sergeyev, Heat and mass transfer with liquid evaporation, Int. J. Heat Mass Transf. 5 (10) (Oct. 1962) 1011–1021, https://doi.org/10.1016/0017-9310(62)90081-9.
[3] W. Brutsaert, S.L. Yu, Mass transfer aspects of pan evaporation, J. Appl. Meteorol. 7 (4) (Aug. 1968) 563–566, https://doi.org/10.1175/1520-0450.
[4] J. Dalton, "Experimental essays on the constitution of mixed gases; on the force of steam or vapour from water and other liquids in different temperatures, both in a Torricellian vacuum and in air; on evaporation; and on the expansion of gases by heat", vol. 5. Manchester: R. & W. Dean, 1802.
[5] M. Moghiman, A. Jodat, Effect of air velocity on water evaporation rate in indoor swimming pools, Iran. Soc. Mech. Eng. 8 (Jul. 2007).
[6] A. Jodat, M. Moghiman, M. Anbarsooz, Experimental comparison of the ability of Dalton based and similarity theory correlations to predict water evaporation rate in different convection regimes, Heat Mass Transf. 48 (8) (Aug. 2012) 1397–1406, https://doi.org/10.1007/s00231-012-0984-z.
[7] M.T. Pauken, An experimental investigation of combined turbulent free and forced evaporation, Exp. Therm. Fluid Sci. 18 (4) (Dec. 1998) 334–340, https://doi.org/10.1016/S0894-1777(98)10038-9.
[8] G.W. Himus, J.W. Hinchley, The effect of a current of air on the rate of evaporation of water below the boiling point, J. Soc. Chem. Ind. 43 (34) (Aug. 1924) 840–845, https://doi.org/10.1002/jctb.5000433402.
[9] M.M. Shah, Evaluation of methods for prediction of evaporation from water pools, J. Build. Phys. 45 (5) (Mar. 2022) 629–648, https://doi.org/10.1177/17442591211034193.
[10] F. Asdrubali, A scale model to evaluate water evaporation from indoor swimming pools, Energy Build. 41 (3) (Mar. 2009) 311–319, https://doi.org/10.1016/j.enbuild.2008.10.001.
[11] C.E. Hyldgaard, "Water evaporation in swimming baths," Aalborg Universitet, Institut for Bygningsteknik, Aalborg, Denmark, 1990.
[12] S.M. Aldarabseh, "Evaporation rate from free water surface," PhD thesis, Western Michigan University, Kalamazoo, Michigan, USA, 2020.
[13] D. Obracaj, M. Korzec, K. Zwolinska, Experimental investigation of the forced evaporation coefficient from the disturbed surface of heated water, Environ. Eng. Manag. J. 20 (2) (2021) 267–282, https://doi.org/10.30638/eemj.2021.027.
[14] M. Al-Shammiri, Evaporation rate as a function of water salinity, Desalination 150 (2) (Nov. 2002) 189–203, https://doi.org/10.1016/S0011-9164(02)00943-8.
[15] J.S. Goodling, B.L. Sill, W.J. McCabe, An evaporation equation for an open body of water exposed to the atmosphere, JAWRA J. Am. Water Resour. Assoc. 12 (4) (Aug. 1976) 843–854, https://doi.org/10.1111/j.1752-1688.1976.tb02750.x.
[16] M.M. Shah, Evaluation of available correlations for rate of evaporation from undisturbed water pools to quiet air, HVACR Res. 8 (1) (Jan. 2002) 125–131, https://doi.org/10.1080/10789669.2002.10391292.
[17] M.M. Shah, Improved method for calculating evaporation from indoor water pools, Energy Build. 49 (Jun. 2012) 306–309, https://doi.org/10.1016/j.enbuild.2012.02.026.
[18] A.M. Raimundo, A.R. Gaspar, A.V.M. Oliveira, D.A. Quintela, Wind tunnel measurements and numerical simulations of water evaporation in forced convection airflow, Int. J. Therm. Sci. 86 (Dec. 2014) 28–40, https://doi.org/10.1016/j.ijthermalsci.2014.06.026.
[19] E. Varju, T. Poós, New dimensionless equation for mass transfer at evaporation of open liquid surface in mixed or forced convection, Int. J. Heat Mass Transf. 227 (Aug. 2024) 125522, https://doi.org/10.1016/j.ijheatmasstransfer.2024.125522.
[20] J.L. Foncubierta Blázquez, I.R. Maestre, F.J. González Gallero, P. Álvarez Gómez, Experimental test for the estimation of the evaporation rate in indoor swimming pools: validation of a new CFD-based simulation methodology, Build. Environ. 138 (Jun. 2018) 293–299, https://doi.org/10.1016/j.buildenv.2018.05.008.
[21] T. Poós, E. Varju, Review for prediction of evaporation rate at natural convection, Heat Mass Transf. 55 (6) (Jun. 2019) 1651–1660, https://doi.org/10.1007/s00231-018-02535-4.
[22] E. Sartori, A critical review on equations employed for the calculation of the evaporation rate from free water surfaces, Sol. Energy 68 (1) (Jan. 2000) 77–89, https://doi.org/10.1016/S0038-092X(99)00054-7.
[23] L.-K. Song, F. Tao, Y. Choy, L.-C. Yang, Y.-F. Wei, Cascade sampling-driven block mapping for coupled reliability evaluation of turbine cooling systems, Mech. Syst. Signal Process. 236 (Aug. 2025) 113008, https://doi.org/10.1016/j.ymssp.2025.113008.
[24] L.-K. Song, F. Tao, G.-Z. Peng, Mixed loss-guided modular regression for dependent system reliability, Reliab. Eng. Syst. Saf. 267 (Mar. 2026) 111898, https://doi.org/10.1016/j.ress.2025.111898.
[25] X.-Q. Li, Q.-S. Li, M.G.R. Faes, M. Beer, Physics-informed synchronous transformer for fatigue reliability analysis, Reliab. Eng. Syst. Saf. 277 (Jan. 2027) 113316, https://doi.org/10.1016/j.ress.2026.113316.
[26] L.-K. Song, M.G.R. Faes, F. Tao, Generative data-physics fusion for hierarchical reliability assessment, Reliab. Eng. Syst. Saf. 272 (Aug. 2026) 112642, https://doi.org/10.1016/j.ress.2026.112642.
[27] Y. Yen, G.R. Landvatter, Evaporation of water into a sub-zero air stream, Water Resour. Res. 6 (2) (Apr. 1970) 430–439, https://doi.org/10.1029/WR006i002p00430.
[28] M. İnan, Ş.Ö. Atayılmaz, Experimental investigation of evaporation from a horizontal free water surface, Sigma J. Eng. Nat. Sci. 35 (1) (2017) 119–131.
[29] F.J.G. Gallero, I.R. Maestre, J.L. Foncubierta Blázquez, J.D. Mena Baladés, Enhanced CFD-based approach to calculate the evaporation rate in swimming pools, Sci. Technol. Built Environ. 27 (4) (Apr. 2021) 524–532, https://doi.org/10.1080/23744731.2020.1868219.
[30] C. Rohwer, "Evaporation from free water surfaces," U.S. Department of Agriculture, Washington, DC, Technical Bulletin No. 271, 1931.
[31] P.T. Tsilingiris, Review and critical comparative evaluation of moist air thermophysical properties at the temperature range between 0 and 100°C for engineering calculations, Renew. Sustain. Energy Rev. 83 (Mar. 2018) 50–63, https://doi.org/10.1016/j.rser.2017.10.072.
[32] I. Masalha, A. Alahmer, A.N. Olimat, M.A. Al-Khawaldeh, O. Badran, M.Y. Mustafa, Experimental data-driven prediction of multi-point temperature fields in parabolic trough solar collectors, Results Eng. 32 (2026) 111805, https://doi.org/10.1016/j.rineng.2026.111805.
[33] E. VARJU, T. POÓS, New dimensionless equation for mass transfer at evaporation of open liquid surface in mixed or forced convection, Int. J. Heat Mass Transf. 227 (2024) 125522, https://doi.org/10.1016/j.ijheatmasstransfer.2024.125522.
[34] Z. Al Sudani, G. Salem, M. Awwad, S.U. Masuri, H. Maaitah, Evaporation rate prediction using advanced machine learning models, Adv. Meteorol. 2022 (2022) 1–13, https://doi.org/10.1155/2022/1433835.