Received: July 1, 2026 | Received in revised form: August July 15, 2026 | Accepted: September 3, 2026 | Available: September 25, 2026

JOURNAL OF APPLIED RESEARCH ON SCIENCE AND TECHNOLOGY
JARST
INSTITUTE OF RESEARCH AND DEVELOPMENT
RAJAMANGALA UNIVERSITY OF TECHNOLOGY THANYABURI
ISSN ONLINE: 2773-9473
https://ph01.tci-thaijo.org/index.php/rmutt-journal

Benchmarking tree-based machine learning surrogates for the coupled Luikov heat and mass transfer system

Pakorn Thongkhumsan[1] and Somboon Sukpancharoen[2*]

[1]Department of Mechanical Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, 40002, THAILAND
[2]Department of Agricultural Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, 40002, THAILAND
[*]Corresponding author: sombsuk@kku.ac.th

ABSTRACT

Replacing repeated high-quality simulations with a fast surrogate is central to parametric studies, sensitivity analyses, and inverse design of coupled transport systems, yet the relative merits of tree-based ensembles for this task remain insufficiently characterized. A purely data-driven benchmark is therefore presented, in which five regressors of increasing flexibility — ridge regression, random forest, extremely randomized trees, histogram-based gradient boosting, and XGBoost — are compared as surrogates for the coupled Luikov system of heat and mass transfer, a stiff multiphysics problem whose bidirectional coupling and multi-scale behavior make repeated numerical solution costly. The surrogates map the five governing dimensionless groups (the thermal and mass Biot numbers, the Luikov number, the Kossovitch number, and the Posnov number) together with the spatio-temporal coordinates directly to the dimensionless temperature and moisture fields, with no physics embedded in the model. A dataset of 120 cases was produced with a fully implicit finite-difference solver, yielding 243,000 pointwise samples. The five candidates were then compared on a case-level 80/20 split with five-fold grouped cross-validation, so that accuracy reflects generalization to unseen parameter sets. The extremely randomized trees model attained the highest cross-validated accuracy ($T^*\,R^2 = 0.971, M^*\,R^2 = 0.952$), with the lowest fold-to-fold variance for the temperature field and generalized to 24 held-out cases ($T^*\,R^2 = 0.980, M^*\,R^2 = 0.984$; per-case median R^2 of 0.990 and 0.984). Averaged over ten independent case-level splits, the held-out R^2 was 0.978 ± 0.011 (T^*) and 0.971 ± 0.010 (M^*), the more stable estimate of generalization. Permutation importance singled out dimensionless time as the strongest predictor in both fields, with the thermal Biot number ranking next for temperature and the Luikov number and spatial position next for moisture—an ordering consistent with the underlying transport physics. Extremely randomized trees, and tree-based ensembles more broadly, thus emerged as accurate and interpretable surrogates for the coupled Luikov system that remain inexpensive to evaluate, at least across the five-parameter envelope sampled here.

Keywords: Data-driven modeling, Convective drying, Porous media, Predictive modeling, Numerical simulation

INTRODUCTION

Surrogate modeling has become a standard strategy for accelerating the many-query analysis of expensive physics simulations, in which the same governing equations must be solved repeatedly across a parameter set for design optimization, sensitivity analysis, and inverse identification [1]. Once trained on a set of high-accuracy solutions, a surrogate predicts the quantities of interest at negligible cost, so the key practical question is which learning method best balances accuracy, generalization, and computational economy for a given class of problems. For tabular input–output data produced by numerical solvers, this question is rarely settled in advance, and a controlled, head-to-head comparison of competing regressors is required before recommending any single method for routine deployment.

Tree-based ensembles are particularly well matched to this kind of tabular data. They handle strongly nonlinear interactions without any feature scaling and cope easily with inputs that span very different ranges, while remaining quick to train and offering a natural measure of variable importance. The most widely used are random forests [2], extremely randomized trees [3], and gradient-boosted trees [4], all of which have a strong track record on structured regression problems. A contrasting strategy embeds the governing equations directly in the training objective, as in physics-informed neural networks, which can parameterize the solution over operating conditions [5] but require explicit formulation of the partial-differential residuals and can be sensitive to

©2026 Institute of Research and Development, RMUTT, Thailand

DOI: 10.60101/jarst.2026.268876

JOURNAL OF APPLIED RESEARCH ON SCIENCE AND TECHNOLOGY

the multi-scale stiffness of the underlying system. When a high-fidelity dataset is available, a purely data-driven surrogate imposes no constraints on the form of the governing equations and offers a simpler, computationally lighter alternative, making thorough benchmarking of such regressors particularly worthwhile.

The coupled Luikov system of heat and mass transfer in capillary-porous bodies provides a demanding testbed for such a benchmark [6]. It governs the simultaneous transport of heat and moisture during convective drying — an operation central to food processing, agricultural-residue valorization, and bioenergy production [7] — and links the temperature and moisture fields through phase change and a thermo-gradient mechanism, a coupling absent from the uncoupled thin-layer diffusion models usually adopted in drying analyses [8]. Mechanistic and computational-fluid-dynamics models resolve these coupled fields accurately but become costly when many operating conditions must be explored [9]. In dimensionless form, the system is parameterized by five groups — the thermal and mass Biot numbers, the Luikov number, the Kossovitch number, and the Posnov number — together with the spatial and temporal coordinates, and it is numerically stiff: at the small Luikov numbers characteristic of porous media, the moisture field evolves far more slowly than the temperature field, so unconditionally stable implicit schemes with fine temporal resolution are required [10]. Solving the system once is manageable, but tasks such as parametric sweeps, sensitivity analysis, and inverse identification require repeated solutions, and this repeated cost is what makes a fast surrogate worthwhile.

Despite the demonstrated strength of tree-based ensembles on tabular problems, their systematic evaluation as surrogates for the full five-parameter Luikov system has received little attention, and the generalization of such surrogates to parameter sets not seen during training remains unestablished. This gap is addressed here through a controlled benchmark of five regressors of increasing flexibility, all trained on the same 120-case finite-difference dataset. Because the folds are grouped by case, holding out entire parameter combinations rather than interpolating between points on one trajectory, the comparison reports how each surrogate behaves on conditions it has never seen. The most accurate model is then characterized in three respects central to surrogate deployment: its predictive accuracy on held-out cases, the relative influence of each input as determined by permutation importance, and its ability to reconstruct the complete spatiotemporal fields. Among these candidates, the four nonlinear models span the bagging and boosting families, while ridge regression provides a linear reference. Deep neural networks, neural operators, and physics-informed neural networks were not benchmarked here, as they typically require substantially more tuning or data, or the explicit imposition of the governing equations, and LightGBM was omitted because histogram-based gradient boosting already represents that class. This work therefore provides a controlled, case-level generalization benchmark of standard regressors for the coupled Luikov system rather than a new learning algorithm, and the parametric-sweep, sensitivity-analysis, and inverse-identification applications noted above are motivating use cases rather than tasks demonstrated here.

MATERIALS AND METHODS

Benchmark problem and reference data

The data analyzed in this work are numerical solutions of the coupled Luikov system for a one-dimensional porous slab [6]. No public benchmark dataset exists for the full five-parameter dimensionless Luikov system with paired temperature and moisture fields, so the reference data were generated in-house under controlled conditions. In dimensionless form, the temperature field T^* and the moisture field M^* evolve according to Eqs. (1) and (2):

$$\frac{\partial T*}{\partial \tau} = \frac{\partial^2 T*}{\partial X^2} + \varepsilon\, Ko\, \frac{\partial M*}{\partial \tau} \tag{1}$$

$$\frac{\partial M*}{\partial \tau} = Lu\, \frac{\partial^2 M*}{\partial X^2} + Lu\, Pn\, \frac{\partial^2 T*}{\partial X^2} \tag{2}$$

where $\varepsilon \in [0, 1]$ is the phase-change coefficient. In the present computations, the phase-change coefficient was set to ε = 1 (fully coupled phase change), so the coupling term in Eq. (1) reduces to the Ko term. The dimensionless coordinates and field variables are defined in Eq. (3) as

$$X = \frac{x}{L}, \tau = \frac{\alpha t}{L^2}, T* = \frac{T - T_0}{T_a - T_0}, M* = \frac{M}{M_0} \tag{3}$$

in which x is the spatial coordinate, t is time, L is the slab thickness, α is the effective thermal diffusivity, and the subscripts 0 and a denote the initial state and the drying-air condition. The system is parameterized by five dimensionless groups defined in Eq. (4), namely the Luikov number Lu, the Kossovitch number Ko, and the Posnov number Pn,

$$Lu = \frac{D}{\alpha}, Ko = \frac{L_v M_0}{C_p \Delta T}, Pn = \frac{\delta \Delta T}{M_0} \tag{4}$$

together with the thermal and mass Biot numbers in Eq. (5):

$$Bi_t = \frac{hL}{k}, Bi_m = \frac{h_m L}{D} \tag{5}$$

where D is the effective moisture diffusivity, L_v the latent heat of vaporization, C_p the specific heat, δ the thermo-gradient coefficient, $\Delta T = T_a - T_0$ the reference temperature difference, k the thermal conductivity, h the convective heat-transfer coefficient, and h_m the convective mass-transfer coefficient. The Luikov

©2026 Institute of Research and Development, RMUTT, Thailand

DOI: 10.60101/jarst.2026.268876

JOURNAL OF APPLIED RESEARCH ON SCIENCE AND TECHNOLOGY

number, which is the ratio of moisture to thermal diffusivity, is small ($Lu \ll 1$) for porous media; the slow moisture mode that results is the origin of its multi-scale stiffness.

The reference solutions assume a uniform initial state($T^* = 0, M^* = 1$), a symmetry condition at the centerline, and Robin-type convective exchange of heat and moisture at the exposed surface, with the surface moisture flux additionally driven by the local thermal gradient. These conditions act only within the finite-difference solver that generates the data and are never presented to the surrogate, which learns the input-to-field mapping directly from the tabulated solutions.

The solver applies second-order central differences in space and backward-Euler integration in time on a uniform mesh of 81 nodes; its absolute stability accommodates the slow moisture mode that dominates at small Lu. Coverage of the five-dimensional parameter envelope was obtained with a 120-case Latin Hypercube design [11] spanning the ranges in Table 1, selected to represent convective drying conditions [8]. Each converged case was archived on a fixed schedule of 25 dimensionless time levels between $\tau = 0.002$ and $\tau = 300$, and the resulting field snapshots form the ground-truth pool from which the learning dataset is assembled. The finite-difference solution was verified to be grid- and time-step-independent. Improving the mesh from 41 to 81 to 161 nodes reduced the relative L2 deviation from the finest (321-node) solution from 0.010% to 0.006% to 0.003% for T* and from 0.105% to 0.072% to 0.030% for M*, confirming that the 81-node mesh is well converged; reducing the time step to 0.002 lowered the deviation below 0.01% (T*) and 0.04% (M*). A complete field of 2,025 points was obtained in about 1.5 s per case.

Table 1 Ranges of the five dimensionless groups sampled by the 120-case Latin Hypercube design.

Dimensionless group	Symbol	Range
Thermal Biot number	Bi_t	0.3 – 1.5
Mass Biot number	Bi_m	1.0 – 12.0
Luikov number	Lu	0.001 – 0.05
Kossovitch number	Ko	3.0 – 35.0
Posnov number	Pn	0.01 – 10.0

Surrogate modeling

The dataset was organized as a pointwise supervised-learning problem, in which each archived field value constitutes a single sample. With 25 time levels and 81 spatial nodes, each case contributes 2,025 field values per output, for 243,000 samples across the 120 cases. Each sample is described by a seven-dimensional input vector comprising the five dimensionless groups (Bi_t, Bi_m, Lu, Ko, Pn), which are constant within a case, together with the spatio-temporal coordinates (X, τ), which vary within a case. The dimensionless temperature $T^*$and moisture M^* are the two regression targets, each modeled by an independent regressor to avoid imposing a shared error structure on the two physically distinct fields. The 120 cases were partitioned into 96 training cases and 24 held-out test cases (80/20 ratio, random seed 42). Partitioning was performed at the case level rather than the sample level, because the 2,025 samples from a single case are strongly autocorrelated; a sample-level split would place near-duplicate points in both partitions and inflate apparent accuracy. The case-level protocol instead measures generalization to parameter combinations not seen during training.

On this pointwise representation, five regressors of increasing flexibility were evaluated: an L2-regularized linear model (ridge regression) as a lower-bound reference; two bagging tree ensembles, the random forest [2] and the extremely randomized trees (extra-trees) method [3]; and two gradient-boosted tree ensembles, the histogram-based gradient-boosting machine and XGBoost [4]. Let $z \in \mathbb{R}^7$ denote the input vector and y the corresponding target. Ridge regression estimates the weight vector w and bias b by minimizing the L2-regularized least-squares objective in Eq. (6):

$$min_{w,b} \sum_{i=1}^{N} (y_i - w^T z_i - b)^2 + \alpha \|w\|^2 \quad (6)$$

where α controls the regularization strength. The bagging ensembles average the predictions of B regression trees h_b grown on randomized subsets of the training data, as in Eq. (7):

$$\hat{y}(z) = \frac{1}{B} \sum_{b=1}^{B} h_b(z) \quad (7)$$

whereas the gradient-boosted ensemble is constructed, as in Eq. (8), as an additive model of K regression trees f_k drawn from the function space $\mathcal{F}$,

$$\hat{y}(z) = \sum_{k=1}^{K} f_k(z), f_k \in \mathcal{F} \quad (8)$$

fitted stagewise via minimizing the regularized objective in Eq. (9):

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (9)$$

with the per-tree complexity penalty in Eq. (10):

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (10)$$

where l denotes a differentiable convex loss, T the number of leaves, w the vector of leaf weights, and γ and λ penalize model intricacy [4]. The principal hyperparameters are listed in Table 2; all remaining settings followed the library defaults, and a fixed

©2026 Institute of Research and Development, RMUTT, Thailand

DOI: 10.60101/jarst.2026.268876

JOURNAL OF APPLIED RESEARCH ON SCIENCE AND TECHNOLOGY

random seed was used throughout to guarantee reproducibility. No per-model hyperparameter search was carried out; because every model used a single fixed configuration, the reported comparison is conservative, and the boosting baselines in particular could improve with dedicated tuning, a limitation of the present benchmark.

Table 2 Hyperparameter settings of the candidate surrogate models.

Model	Hyperparameters
Ridge regression	Regularization strength $\alpha = 1.0$; input features standardized
Random forest	60 trees; maximum depth 18; minimum 5 samples per leaf
Extra-trees	60 trees; maximum depth 18; minimum 5 samples per leaf
Histogram gradient boosting	300 boosting iterations; learning rate 0.08
XGBoost	400 trees; maximum depth 7; learning rate 0.08; subsample 0.9; column subsample 0.9

Note. The selected model (extra-trees) was retrained on the complete training set with 120 trees, a maximum depth of 20, and a minimum of 3 samples per leaf. The bagging ensembles (random forest and extra-trees) were fitted on a random subsample of 45,000 points per cross-validation fold. Unless otherwise stated, library default values were used, and the random seed was fixed at 42.

Model selection relied on five-fold grouped cross-validation over the 96 training cases. Folds were defined at the case level (GroupKFold), so the samples from any one case fell entirely on either the fitting or the validation side of a fold, never both. Each candidate was ranked by its coefficient of determination average across the five folds. To keep the search affordable, the two bagging ensembles, the random forest and extremely randomized trees, were fitted on a random 45,000-point subsample within each fold, while the ridge and gradient-boosting models were trained on the full fold. The model with the highest mean cross-validated R^2 across both targets was then retrained on the complete training set and evaluated once on the held-out test cases. Because this subsampling reduces the data seen by the bagging ensembles relative to the boosting models fitted on the full fold, the comparison is conservative for the selected extra-trees model rather than biased in its favor.

Model evaluation

Predictive accuracy on the N held-out samples was quantified by four complementary measures, reported separately for T^* and M^*. With y_i the reference value, $\hat{y}_i$ the prediction, and $\bar{y}$ the sample mean, the coefficient of determination, the root-mean-square error, the mean absolute error, and the maximum absolute error are defined in Eqs. (11)–(14) as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (11)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (12)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (13)$$

$$E_{max} = max_i\,|y_i - \hat{y}_i| \quad (14)$$

Generalization was further characterized by the per-case distribution of R^2 over the 24 test cases. The relative contribution of each input was assessed by permutation feature importance [2], computed on the test set as the mean decrease in R^2 over $K = 30$ random permutations of each feature, according to Eq. (15):

$$I_j = R^2 - \frac{1}{K}\sum_{k=1}^{K} R^2_{j,k} \quad (15)$$

where $R^2_{j,k}$ is the coefficient of determination obtained after the $k-th$ random permutation of feature j, with all other features held fixed. Finally, all computations were carried out in Python 3.12, with the surrogates implemented using scikit-learn [12] for the linear, bagging, and histogram-based models and the XGBoost library [4] for gradient boosting. The Latin Hypercube design was generated following [11]. All computations used scikit-learn (version 1.8.0) and XGBoost (version 3.2.0); the tidy 243,000-point dataset and the analysis notebook are available from the authors upon reasonable request.

RESULTS AND DISCUSSION

Correlation structure and model comparison

The Pearson correlation structure of the dataset is shown in Figure 1. Because each case was drawn by Latin Hypercube sampling, the five dimensionless parameters vary almost independently of one another — no pairwise correlation exceeds 0.13 in magnitude. This near-orthogonality lets the feature-importance rankings below be read directly, without the ambiguity correlated inputs would introduce. The dimensionless time τ exhibits the strongest linear association with the outputs($\tau - T^* = 0.54, \tau - M^* = -0.63$), indicating the progressive heating and drying of the slab, while the moisture field is additionally correlated with the Luikov number ($Lu - M^* = -0.41$) and the spatial coordinate ($X - M^* = -0.31$). The two fields are themselves anticorrelated ($T^* - M^* = -0.47$), consistent with the coupled nature of the Luikov system. The generally modest magnitude of these linear correlations indicates a strongly nonlinear input–output mapping and motivates the use of nonlinear regressors.

Cross-validation results are summarized in Figure 2 and Table 3. The L2-regularized linear model performs poorly ($T^*\,R^2 = 0.335, M^*\,R^2 = 0.720$),

©2026 Institute of Research and Development, RMUTT, Thailand

DOI: 10.60101/jarst.2026.268876

confirming the nonlinearity of the mapping. All four tree ensembles exceed $R^2 = 0.90$ on both targets. Among them, the extra-trees model attains the highest cross-validated accuracy together with the smallest fold-to-fold variance ($T^* R^2 = 0.971 \pm 0.004$; $M^* R^2 = 0.952 \pm 0.021$) and the lowest RMSE ($T^* 0.036, M^* 0.069$); it was therefore selected as the final surrogate.

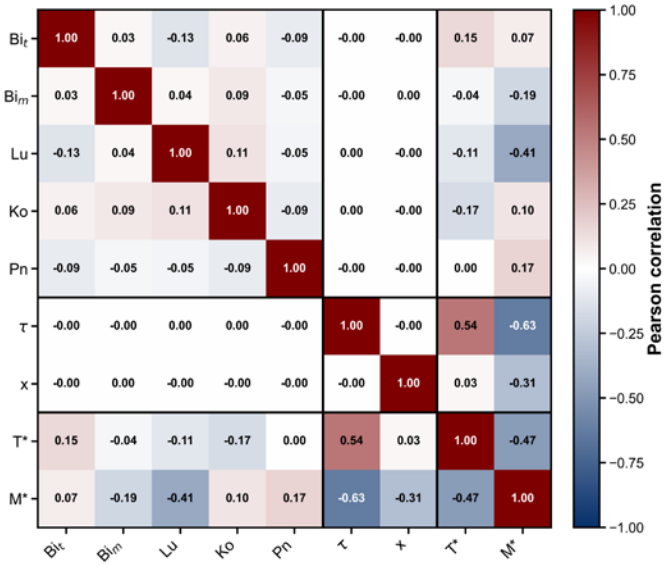


Figure 1 Pearson correlation among the five dimensionless input parameters, the two coordinates (X,τ), and the two outputs (T^*,M^*) over the full 243,000-point dataset. The parameter–parameter block reflects the 120 independent Latin Hypercube design points (each group is constant within a case), whereas the coordinate and output rows are evaluated point-wise over all 243,000 samples.

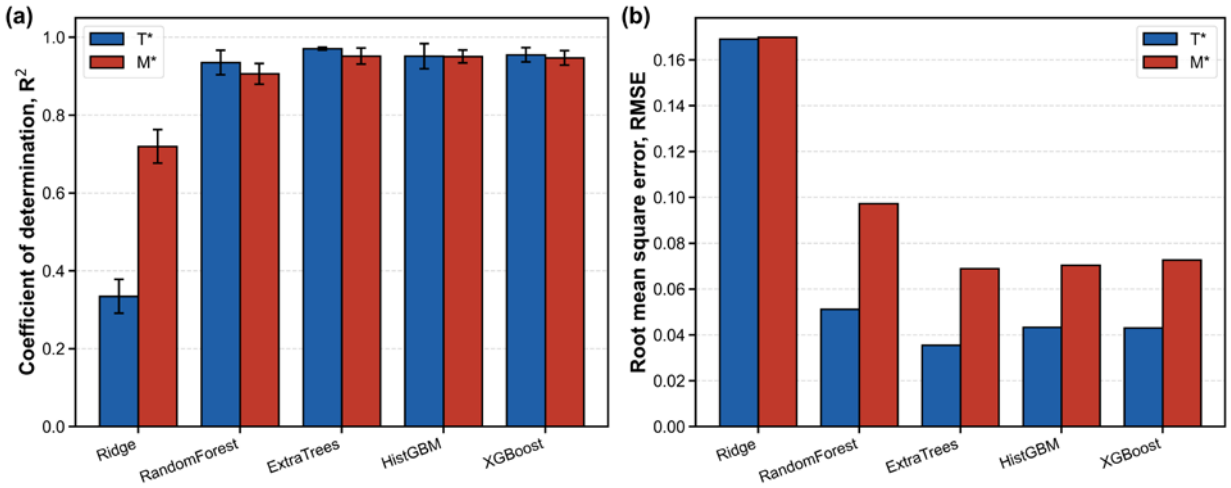


Figure 2 Five-fold GroupKFold (case-grouped) cross-validation performance of the five candidate models: (a) coefficient of determination R^2; (b) root-mean-square error. Error bars in (a) denote the standard deviation across folds.

Table 3 Model comparison via five-fold GroupKFold cross-validation (grouped by case). Values for R^2 are mean ± standard deviation across folds; RMSE and MAE are fold means.

Model	T^* R^2	T^* RMSE	T^* MAE	M^* R^2	M^* RMSE	M^* MAE
Ridge	0.335 ± 0.043	0.169	0.101	0.720 ± 0.043	0.170	0.136
RandomForest	0.935 ± 0.031	0.051	0.024	0.906 ± 0.027	0.097	0.066
Extra-trees	0.971 ± 0.004	0.036	0.018	0.952 ± 0.021	0.069	0.048
HistGBM	0.952 ± 0.032	0.043	0.020	0.951 ± 0.016	0.070	0.048
XGBoost	0.952 ± 0.022	0.045	0.022	0.945 ± 0.020	0.075	0.051

©2026 Institute of Research and Development, RMUTT, Thailand

DOI: 10.60101/jarst.2026.268876

JOURNAL OF APPLIED RESEARCH ON SCIENCE AND TECHNOLOGY

Table 4 Generalization of the selected extra-trees surrogate to the 24 unseen test cases.

Output	Test R^2	Test RMSE	Test MAE	Max |error|	Per-case R^2 (min / median / max)
T^*	0.9801	0.0296	0.0162	0.2649	0.9245 / 0.9904 / 0.9995
M^*	0.9844	0.0401	0.0297	0.1855	0.9174 / 0.9836 / 0.9988

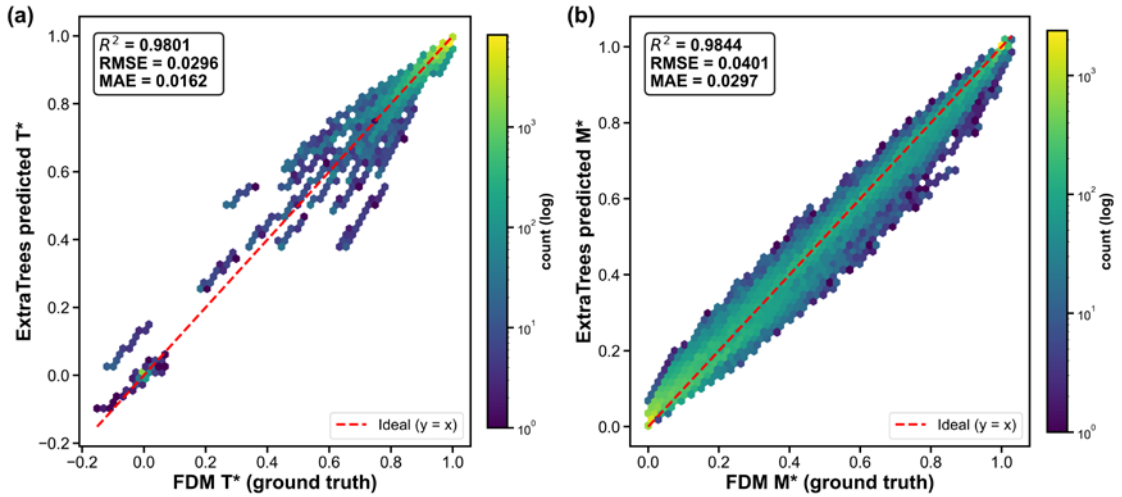


Figure 3 Predicted-versus-reference parity (logarithmic hexbin density) of the extra-trees surrogate on the 24 unseen test cases (48,600 points per output): (a) T^*; (b) M^*. The dashed line denotes the ideal $y = x$.

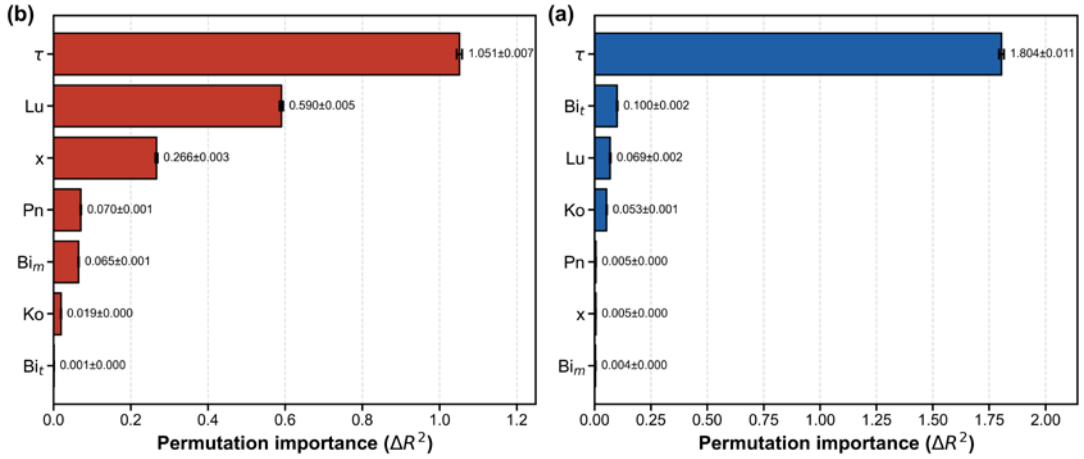


Figure 4 Permutation feature importance (mean decrease in R^2 over 30 permutations; error bars denote one standard deviation across the 30 permutations) of the extra-trees surrogate on the test set: (a) T^*; (b) M^*.

Table 5 Training and inference cost of the five candidate surrogates (both fields combined) and speedup relative to the finite-difference reference solver.

Model	Train (s)	Infer (ms/case)	Speedup
Ridge	0.3	0.8	1900×
RandomForest	40	39	38×
Extra-trees	20	42	35×
HistGBM	6.7	60	25×

Note. Training and inference times cover both the temperature and moisture models. The finite-difference solver requires about 1.47 s per case with reused LU factorization (about 12.6 s with a per-step sparse solve). The deployed extra-trees model (120 trees) trains in about 49 s and predicts a full field in about 89 ms per case, an approximately 17-fold speedup; speedups are reported relative to the reused-factorization solver.

Generalization and feature importance

The generalization of the selected extra-trees surrogate to the 24 held-out test cases is reported in Figure 3 and Table 4. The predicted fields agree closely with the FDM reference, yielding R^2 = 0.980 (RMSE = 0.030, MAE = 0.016) for T^* and R^2= 0.984 (RMSE = 0.040, MAE = 0.030) for M^*. The parity densities

©2026 Institute of Research and Development, RMUTT, Thailand

DOI: 10.60101/jarst.2026.268876

JOURNAL OF APPLIED RESEARCH ON SCIENCE AND TECHNOLOGY

cluster tightly along the ideal line; the residual scatter for T^* is concentrated at intermediate values, where the transient gradients are steepest. The per-case coefficient for determination remains high across the entire test set, with median values of 0.990 (T^*) and 0.984 (M^*) and minima of 0.925 and 0.917, respectively, indicating that no individual parameter combination is reproduced poorly. The maximum absolute errors (0.265 for T^* and 0.186 for M^*) are localized rather than systematic. Retraining with the cross-validation configuration (60 trees, depth 18, 5 samples per leaf) gave held-out R^2 of 0.977 (T*) and 0.983 (M*), so the deployed configuration (120 trees, depth 20, 3 samples per leaf; R^2 0.980 and 0.984) improves accuracy only marginally without changing the model ranking. Across ten independent case-level splits, held-out accuracy was $R^2 = 0.978 \pm 0.011$ (T*) and 0.971 ± 0.010 (M*), which provides a more robust estimate of generalization than the single split. Relative to the finite-difference solver (about 1.5 s per case), the trained surrogate predicts a complete field in roughly 90 ms, a speedup of one to two orders of magnitude (Table 5).

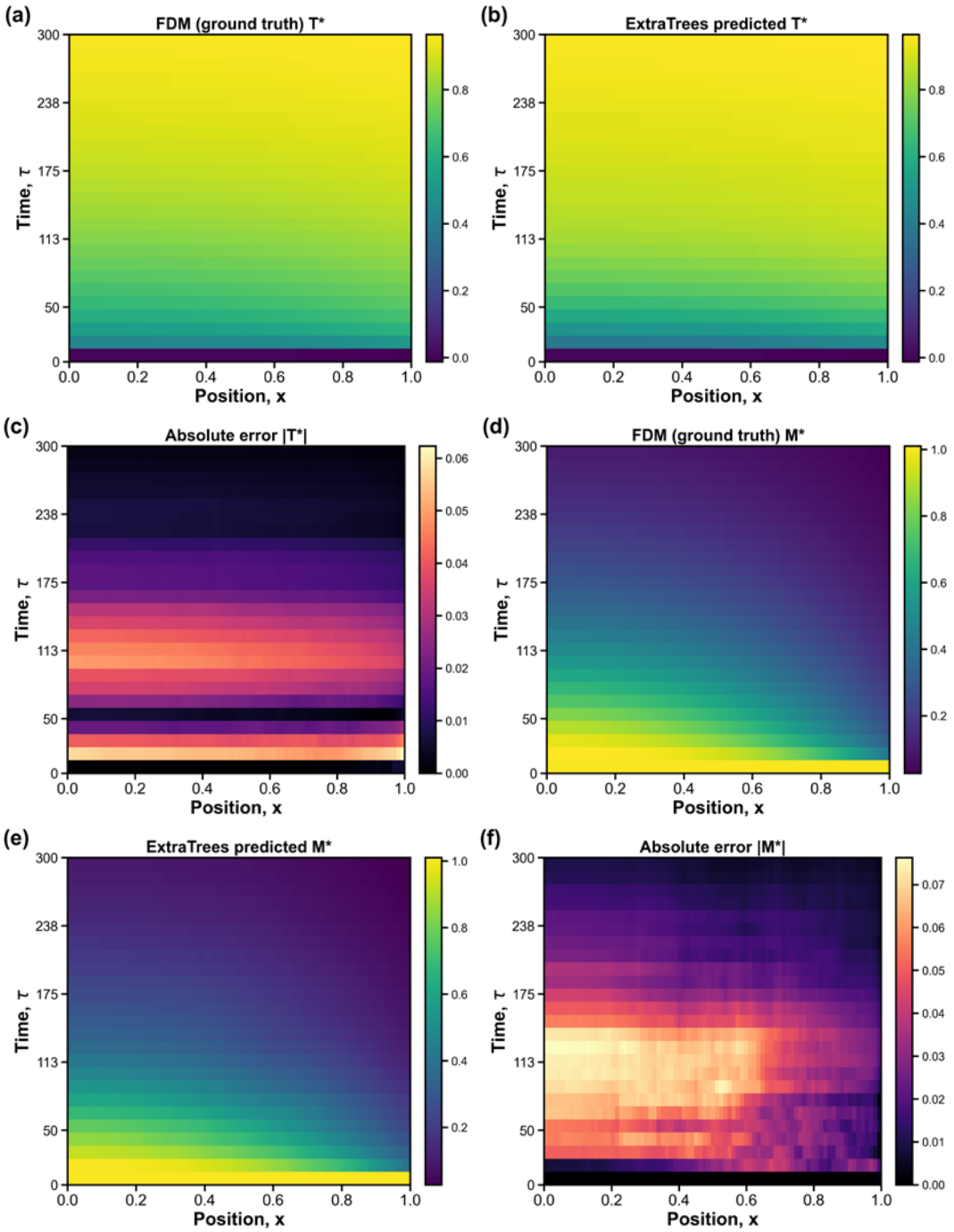


Figure 5 Spatiotemporal reconstruction for a representative unseen case (Case 091): FDM ground truth, extra-trees prediction, and absolute error for T^*(a–c) and M^* (d–f).

©2026 Institute of Research and Development, RMUTT, Thailand

DOI: 10.60101/jarst.2026.268876

JOURNAL OF APPLIED RESEARCH ON SCIENCE AND TECHNOLOGY

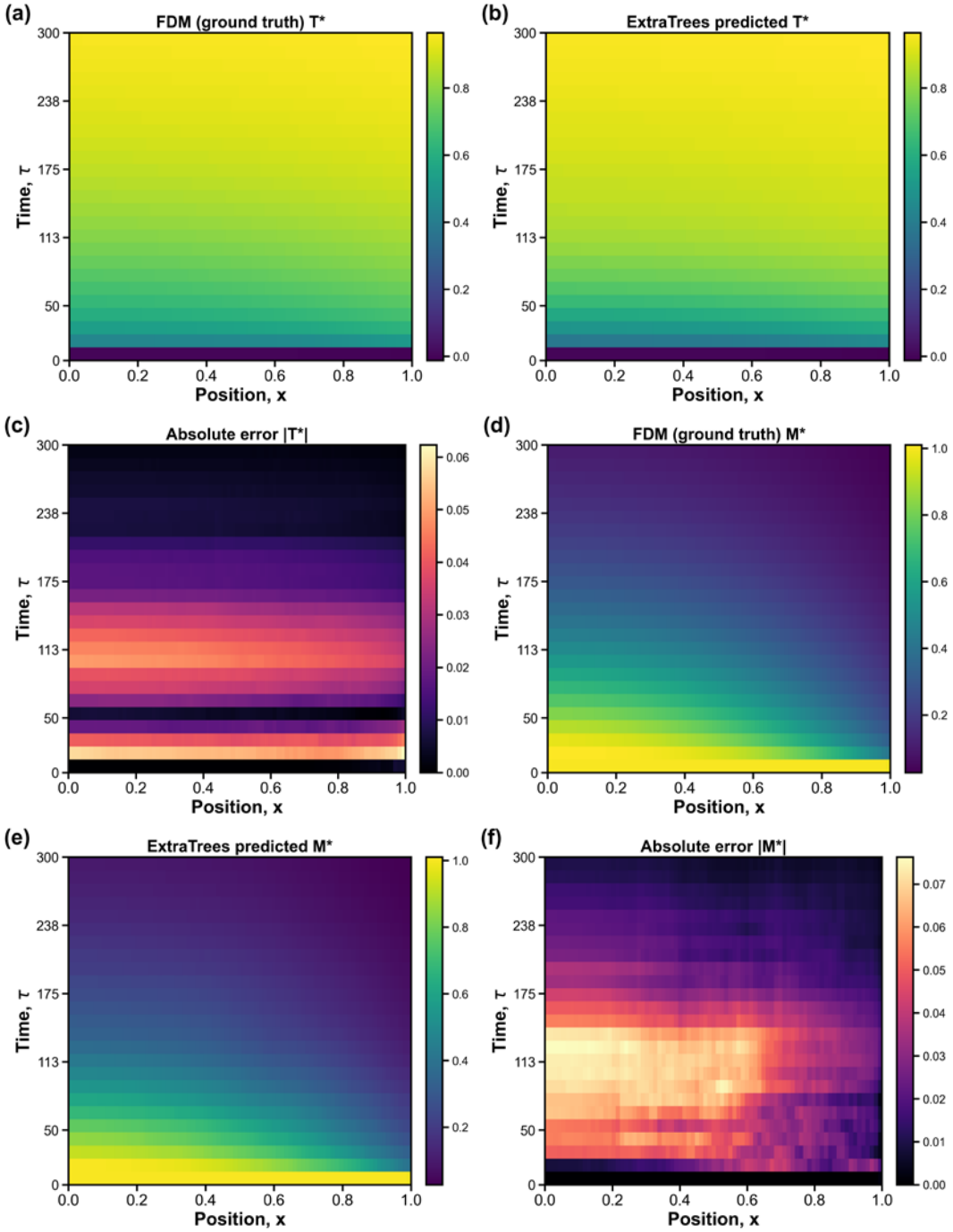


Figure 5 Spatiotemporal reconstruction for a representative unseen case (Case 091): FDM ground truth, extra-trees prediction, and absolute error for T^*(a–c) and M^* (d–f).

The permutation importances of the seven inputs are shown in Figure 4. For both outputs the dimensionless time τ is by far the most influential feature ($\Delta R^2 \approx 1.8$ for T^* and ≈ 1.05 for M^*), as expected for a transient process. For T^*, the next-ranked features are the thermal Biot number Bi_t, the Luikov number, and the Kossovitch number, whereas the mass Biot number, Posnov number, and spatial coordinate contribute negligibly. For M^*, the Luikov number ($\Delta R^2 \approx 0.59$) and the spatial coordinate ($\Delta R^2 \approx 0.27$) are the dominant secondary features, followed by the Posnov and mass Biot numbers. This attribution is qualitatively consistent with the expected transport behavior: the temperature field is governed primarily by surface heat exchange (Bi_t), while the moisture field is controlled by the moisture-to-thermal diffusivity ratio (Lu), the through-thickness position, and the thermo-moisture coupling (Pn). The ranking also corresponds to the linear correlation structure of Figure 1.

Spatiotemporal field reconstruction

A representative reconstruction for an unseen case (Case 091; Bi_t = 0.50, Bi_m = 4.21, Lu = 0.027, Ko = 26.5, Pn = 6.03) is shown in Figure 5. The extra-trees surrogate reproduces the full spatiotemporal evolution of both fields: the monotonic heating of T^*

©2026 Institute of Research and Development, RMUTT, Thailand

DOI: 10.60101/jarst.2026.268876

and the progressive drying front of M^* from the surface inward are recovered without visible distortion. The absolute-error maps show that the residuals are concentrated in the early-to-intermediate transient, where the fields change most rapidly, and decay toward zero as the steady state is approached. The peak local errors (about 0.06 for T^* and 0.07 for M^*) are consistent with the maximum errors and parity scatter reported above and indicate that the principal limitation of the surrogate is the resolution of the sharpest transient gradients rather than any systematic bias. Case 091 was selected as the case whose average per-case R^2 is closest to the median over the 24 test cases; the worst-performing case (Case 062, average $R^2 = 0.941$; Figure 6) shows the same localization of error in the early-to-intermediate transient, confirming that the reported behavior is representative.

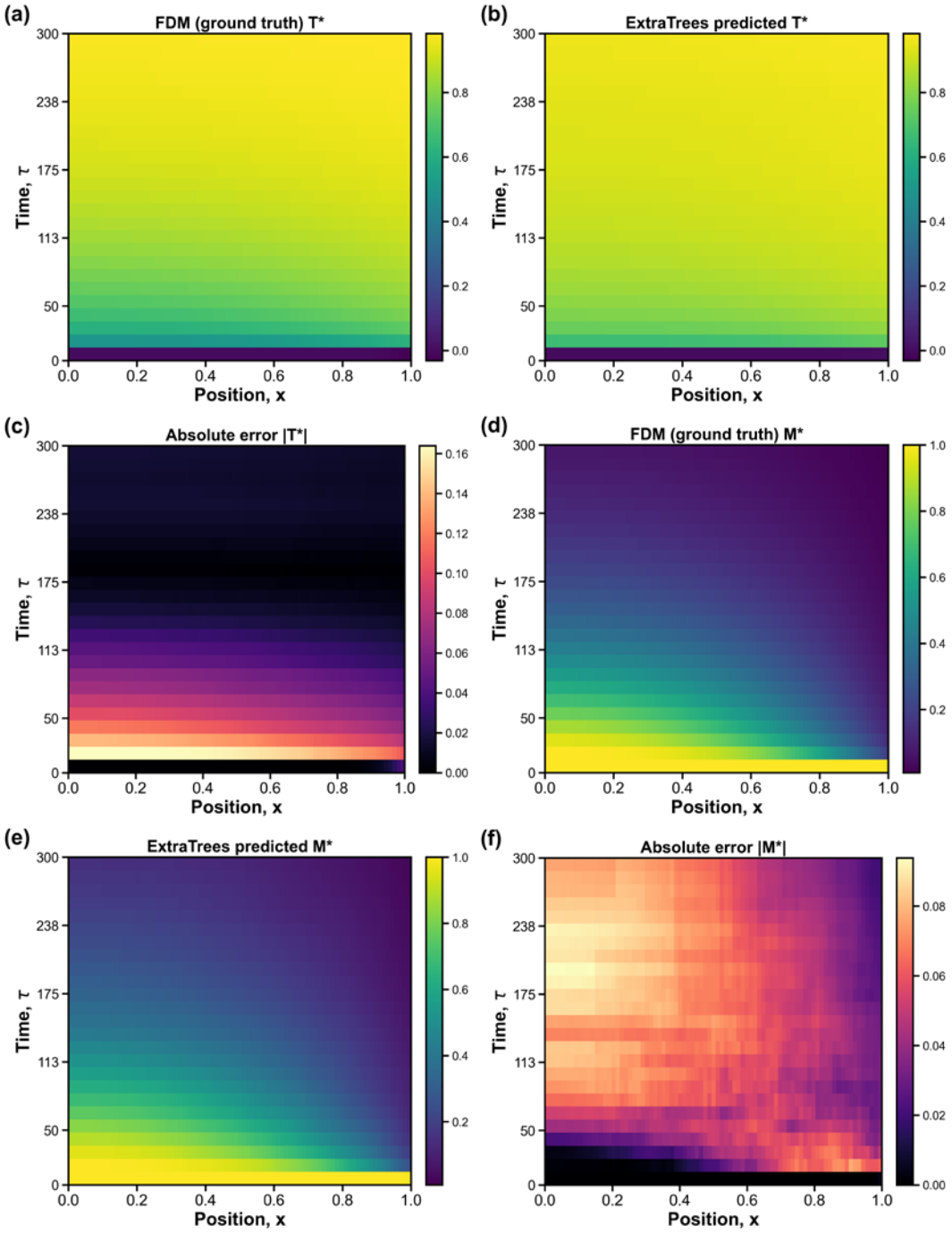


Figure 6 Spatiotemporal reconstruction for the worst-performing unseen case (Case 062, average $R^2 = 0.941$): FDM ground truth, extra-trees prediction, and absolute error for T* (a–c) and M* (d–f).

The overall ranking of the candidate models matches the complexity of the underlying mapping. The L2-regularized linear model fails markedly for the temperature field (T^* $R^2 = 0.335$) and only moderately for the moisture field (M^* $R^2 = 0.720$), indicating that the temperature response depends more strongly on

©2026 Institute of Research and Development, RMUTT, Thailand

DOI: 10.60101/jarst.2026.268876

nonlinear interactions among the dimensionless time and the Biot and coupling groups than the moisture response does so. All four tree ensembles capture these interactions without feature scaling and exceed R^2 = 0.90 on both targets. That the extra-trees model attains both the highest cross-validated accuracy and the smallest fold-to-fold variance for temperature (± 0.004 for T^*) indicates that additional randomization of split thresholds is especially effective at reducing variance on a dataset in which samples drawn from a single case are strongly autocorrelated. The bagging ensembles (random forest and extra-trees) and the boosting ensembles (histogram-based gradient boosting and XGBoost) reach equivalent accuracy, with the extra-trees model retaining a small but consistent advantage in both accuracy and fold stability; this dependability across model parameter combinations is the decisive property for a surrogate intended to generalize.

The explainability of the selected surrogate follows from the agreement among its permutation importances, the linear correlation structure, and the governing physics. The dimensionless time dominates both fields, as expected for a transient process. At the same time, the thermal Biot number governs the temperature field through surface heat exchange, and the Luikov number together with the spatial coordinate governs the moisture field through the moisture-to-thermal diffusivity ratio and the through-thickness position. Because these rankings reproduce known transport mechanisms rather than spurious associations, the surrogate can be regarded as producing feature attributions qualitatively consistent with expected transport behavior rather than as a black box. The uniformly high per-case coefficients of determination obtained under the strict case-level partition (median 0.990 for T^* and 0.984 for M^*, with minima of 0.925 and 0.917) further confirm that the model applies to unseen parameter combinations rather than memorizing autocorrelated samples.

The principal limitation is that the largest residuals are confined to the early-to-intermediate transient, where the spatial gradients are steepest, rather than being distributed systematically; the surrogate therefore reproduces the quasi-steady regime almost exactly while slightly under-resolving the sharpest transient features. This behavior also delimits the range of valid application, since the accuracy reported here is established only within the sampled five-parameter envelope and predictions outside that envelope are not warranted. Relative to physics-informed approaches, which embed the governing equations directly in the training objective but require the partial-differential residuals to be formulated explicitly and can be sensitive to the multi-scale stiffness of the Luikov system [5], the present purely data-driven surrogate imposes no constraint on the form of the equations and trains rapidly, at the cost of requiring a precomputed high-fidelity dataset. The two strategies are therefore complementary rather than competing. Because the extra-trees ensemble averages many trees, it yields a native prediction interval at no additional cost: the mean tree-to-tree standard deviation is 0.035 (T^*) and 0.084 (M^*), and $\pm 2\sigma$ intervals contain about 99.9% of held-out points. The ensemble variance is therefore a usable, if conservative, uncertainty proxy for the inverse-identification and sensitivity tasks that motivate a surrogate, rather than a rigorously calibrated interval.

These findings match a growing body of data-driven surrogate studies in drying research: gradient-boosted-tree models have reproduced solar-drying kinetics with high accuracy [13], machine-learning models have predicted both the mass and the temperature of produce during greenhouse solar drying [14], and physics-informed neural networks have been extended from the coupled fields to derived quantities such as shrinkage [15]. This benchmark complements these largely experiment-fitted or physics-embedded studies by evaluating standard tree ensembles on a controlled, high-fidelity dataset spanning the full five-parameter dimensionless Luikov system.

Beyond drying, tree-based ensembles have served as accurate surrogates for other simulation-based engineering problems, including building-energy and cost prediction [16] and latent-heat thermal storage [17]. The case-level partitioning adopted here also reflects the recognition that random splits of strongly autocorrelated samples inflate apparent accuracy, so hold-out sets are better formed by grouping [18], and the permutation-importance rankings reported above correspond to the model-agnostic permutation-importance measure [19].

CONCLUSIONS

A purely data-driven tree-ensemble surrogate was developed and evaluated for the coupled Luikov system of heat and mass transfer during convective drying of porous biomass. Five regressors were compared on a 120-case finite-difference dataset spanning the five-dimensional dimensionless parameter envelope, using a case-level partition and grouped cross-validation that measures generalization to unseen parameter combinations rather than interpolated points within a single trajectory.

The extremely randomized trees model proved the most accurate and the most stable across folds, reaching cross-validated coefficients of determination of 0.971 for temperature and 0.952 for moisture, and generalizing to the 24 held-out cases with R^2 of 0.980 and 0.984, respectively. Permutation importance attributed the predictions primarily to dimensionless time, with the thermal Biot number governing the temperature field and the Luikov number together

©2026 Institute of Research and Development, RMUTT, Thailand

DOI: 10.60101/jarst.2026.268876

with the spatial coordinate governing the moisture field, consistent with the underlying transport physics. The surrogate reproduced the complete spatiotemporal evolution of both fields, with residuals confined to the early-to-intermediate transient where the gradients are steepest. The horizontal banding visible in the error maps reflects the fixed archival schedule of twenty-five discrete dimensionless-time levels at which each case was stored, rather than any surrogate artifact; the brightest bands coincide with this early-to-intermediate window where the moisture gradients, and hence the interpolation error, are largest.

These results show that a fast and interpretable tree-based surrogate, trained on a moderate number of high-accuracy solutions, can replace repeated numerical solutions of the Luikov system across its full five-parameter envelope at negligible evaluation cost. Two limitations should be noted. First, the surrogate is interpolative: its accuracy is established only within the sampled parameter ranges, and predictions outside the trained envelope are not guaranteed. Second, the largest errors occur in the sharpest transient region, which suggests that denser temporal sampling or transient-focused training could reduce them further. Subsequent research might extend the parameter envelope to encompass specific biomass feedstocks, integrate the surrogate into parametric optimization and inverse-identification loops where its low evaluation cost is most valuable, and examine the transferability of the approach to higher-dimensional or geometry-dependent drying configurations.

DECLARATION OF GENERATIVE AI IN SCIENTIFIC WRITING

During the preparation of this work, the authors used Claude Opus 4.8 to improve the readability and language of the manuscript. After using this tool/service, the authors carefully reviewed and edited the content as necessary and certified that the technology was used under full human management. The authors take full responsibility for the content of the published article.

ACKNOWLEDGEMENT

The authors would like to thank the Faculty of Engineering, Khon Kaen University, for their support and facilities provided throughout this research.

REFERENCES

1. Wijerathne ADHT, Joardder MUH, Welsh ZG, Nayak R, Sablani SS, Karim A. Recent advances in food drying modeling: empirical to multiscale physics-informed neural networks. Compr Rev Food Sci Food Saf [internet]. 2025;24:e70194. Available from: https://doi.org/10.1111/1541-4337.70194.
2. Breiman L. Random forests. Mach Learn [internet]. 2001;45(1):5–32. Available from: https://doi.org/10.1023/A:1010933404324.
3. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn [internet]. 2006;63(1):3–42. Available from: https://doi.org/10.1007/s10994-006-6226-1.
4. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. [internet]. New York: Association for Computing Machinery; 2016. p. 785–94. Available from: https://doi.org/10.1145/2939672.2939785.
5. Manavi S, Fattahi E, Becker T. A parameterized physics-informed machine learning approach for solving heat and mass transfer equations in the drying process. Int Commun Heat Mass Transf [internet]. 2024;158:107897. Available from: https://doi.org/10.1016/j.icheatmasstransfer.2024.107897.
6. Luikov AV. Systems of differential equations of heat and mass transfer in capillary-porous bodies (review). Int J Heat Mass Transf [internet]. 1975;18(1):1–14. Available from: https://doi.org/10.1016/0017-9310(75)90002-2.
7. Datta AK. Toward computer-aided food engineering: mechanistic frameworks for evolution of product, quality and safety during processing. J Food Eng [internet]. 2016;176:9-27. Available from: https://doi.org/10.1016/j.jfoodeng.2015.10.010.
8. Koukouch A, Bakhattar I, Asbik M, Idlimam A, Zeghmati B, Aharoune A. Analytical solution of coupled heat and mass transfer equations during convective drying of biomass: experimental validation. Heat Mass Transf [internet]. 2020;56(6):1971–83. Available from: https://doi.org/10.1007/s00231-020-02817-w.
9. Defraeye T, Blocken B, Carmeliet J. Analysis of convective heat and mass transfer coefficients for convective drying of a porous flat plate by conjugate modelling. Int J Heat Mass Transf [internet]. 2012;55(1–3):112–24. Available from: https://doi.org/10.1016/j.ijheatmasstransfer.2011.08.047.
10. LeVeque RJ. Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems. Philadelphia: SIAM; 2007. Available from: https://doi.org/10.1137/1.9780898717839.
11. McKay MD, Beckman RJ, Conover WJ. Comparison of three methods for selecting values of input variables in the analysis of output from a computer

©2026 Institute of Research and Development, RMUTT, Thailand

DOI: 10.60101/jarst.2026.268876

code. Technometrics [internet]. 1979;21(2):239-45. Available from: https://doi.org/10.1080/00401706.1979.10489755.

12. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011; 12:2825-30.

13. Tagnamas Z, Idlimam A, Lamharrar A. Predictive models of beetroot solar drying process through machine learning algorithms. Renew Energy [internet]. 2023;219:119522. Available from: https://doi.org/10.1016/j.renene.2023.119522.

14. Daliran A, Taki M, Marzban A, Rahnama M, Farhadi R. Experimental evaluation and modeling the mass and temperature of dried mint in greenhouse solar dryer: application of machine learning method. Case Stud Therm Eng [internet]. 2023;47:103048. Available from: https://doi.org/10.1016/j.csite.2023.103048.

15. Manavi S, Fattahi E, Becker T. Surrogate modeling of shrinkage due to drying using physics-informed machine learning method. Phys Fluids [internet]. 2025;37(7):073134. Available from: https://doi.org/10.1063/5.0274073.

16. Shirzadi N, Lau D, Stylianou M. Surrogate modeling for building design: energy and cost prediction compared to simulation-based methods. Buildings [internet]. 2025;15(13):2361. Available from: https://doi.org/10.3390/buildings15132361.

17. Çolak AB. Extreme gradient boosting (XGBoost) prediction of latent heat storage in conical-fin solar thermal units with graphene-alumina hybrid-nano-additives. Heat Mass Transf [internet]. 2026;62:11. Available from: https://doi.org/10.1007/s00231-026-03644-1.

18. Ploton P, Mortier F, Réjou-Méchain M, Barbier N, Picard N, Rossi V, et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat Commun [internet]. 2020;11:4540. Available from: https://doi.org/10.1038/s41467-020-18321-y.

19. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res. 2019;20(177):1-81.

©2026 Institute of Research and Development, RMUTT, Thailand

DOI: 10.60101/jarst.2026.268876