energies

MDPI

Article

Machine-Learning-Based Parameterisation of Soil Thermal Conductivity for Shallow Geothermal and Ground Heat Exchanger Modelling

Mateusz Żeruń *, Ewa Jagoda and Edyta Majer

Department of Engineering Geology and Shallow Geothermal Energy, Polish Geological Institute—National Research Institute, 76 Jagiellońska St., 03-301 Warsaw, Poland; ewa.jagoda@pgi.gov.pl (E.J.); edyta.majer@pgi.gov.pl (E.M.)

* Correspondence: mateusz.zerun@pgi.gov.pl; Tel.: +48-536-938-964

Abstract

Thermal conductivity is a key input parameter in geotechnical and shallow geothermal engineering, directly influencing the design, efficiency, and long-term performance of ground heat exchangers, energy piles, and ground-source heat pump systems. Reliable parameterisation of this property in sandy soils remains challenging due to nonlinear interactions between water content, bulk density, and soil structure. This study develops a machine-learning-based workflow for robust parameterisation of thermal conductivity in quartz-rich sands using a large, internally consistent laboratory dataset comprising 1716 samples, including 1455 moist measurements used for modelling, obtained from nationwide site investigations. Air-dry specimens were identified as laboratory-induced drying states and excluded to restrict the analysis to hydro-mechanical conditions representative of typical shallow subsurface environments. Several regression algorithms representing different modelling strategies were evaluated within a unified and reproducible framework and benchmarked against selected classical empirical formulations. Model performance was assessed using standard accuracy metrics together with diagnostics describing the functional stability of predicted thermal-conductivity surfaces. The results reveal a systematic trade-off between predictive accuracy and functional consistency, indicating that models optimised for accuracy may produce functionally unstable and less suitable parameterisations for engineering applications. Accuracy-optimised models frequently produce locally irregular parameter fields, whereas more strongly regularised models yield smoother and physically more coherent response surfaces. The proposed workflow supports reliable thermal-property parameterisation for geotechnical design and shallow geothermal modelling.

Keywords: soil thermal conductivity; shallow geothermal systems; heat transfer in porous media; machine learning; ground-source heat pumps; parameterisation; thermal modelling

Academic Editors: Christopher Micallef and Charles Yousif

Received: 12 March 2026
Revised: 30 March 2026
Accepted: 7 April 2026
Published: 8 April 2026

1. Introduction

Thermal conductivity (λ) is a fundamental parameter governing heat transfer in near-surface porous materials and a key input in geotechnical and shallow geothermal engineering [1]. It directly influences the design, sizing, and long-term performance of ground-source heat pump systems, energy piles, and other thermo-active underground structures, where heat exchange between engineered systems and surrounding soils controls thermal efficiency and operational stability [2–5].

In engineering practice, λ is routinely used in numerical simulations and design calculations. It is treated not merely as a measured quantity but as a spatially and state-dependent material parameter subject to interpolation, mapping, and reuse. Uncertainty or instability in its parameterisation may therefore lead to systematic design errors and increased operational risk. In particular, non-smooth or physically inconsistent parameter fields may introduce artificial gradients into numerical models, leading to biassed predictions of heat-exchanger performance and long-term thermal behaviour. Reliable parameterisation of soil thermal properties is therefore essential for accurate geothermal resource assessment, ground heat exchanger design, and performance simulations.

In sandy soils, thermal conductivity is controlled primarily by water content (w) and bulk density (ρ_b), which regulate pore-scale thermal connectivity and grain-to-grain heat transfer within the solid skeleton [6–10]. Increasing water content replaces air—an effective thermal insulator—with liquid water, enhancing conductive pathways, while increasing bulk density improves grain contacts and reduces porosity. Under partially saturated conditions, these effects interact nonlinearly, making reliable estimation difficult without extensive laboratory testing [6,8,9].

In practice, thermal conductivity is commonly estimated using empirical or semi-empirical formulations derived from simplified physical assumptions and limited calibration datasets [3,5,7,9,11]. Although physically motivated, such models often exhibit limited transferability outside their original domains, particularly for partially saturated sands and site-specific conditions [7,9,11]. As a result, their use in mapping, site characterisation, geothermal simulations, and engineering design remains associated with substantial uncertainty.

Recent studies show that machine-learning models can capture continuous and potentially nonlinear relationships between thermal conductivity and soil-state variables and often outperform empirical formulations in numerical accuracy [10,12,13]. Consequently, data-driven approaches are increasingly applied to large geotechnical and geological archives where water content and bulk density are available, but laboratory measurements of λ are sparse. In such settings, trained models function as continuous mappings across the predictor domain rather than as pointwise predictors. However, model evaluation is typically limited to standard error metrics, providing little insight into whether the learned relationships are stable, physically interpretable, and suitable for reuse in engineering workflows [14–16].

This limitation is particularly relevant in geotechnical and geothermal applications, where thermal conductivity is used as a continuous material parameter for interpolation, spatial mapping, and numerical modelling. Models that reproduce measured values with low error but encode unstable or physically implausible relationships may introduce artificial gradients and discontinuities into reconstructed parameter fields. When such parameterisations are used in subsurface heat-transfer simulations, these artefacts may lead to unrealistic thermal gradients or biassed performance predictions, even when conventional validation metrics indicate high accuracy.

The present study focuses on quartz-rich sands, a lithologically coherent material class in which mineralogical variability has only a secondary influence on effective thermal conductivity [4,6,8–10,17]. In this setting, saturation state and packing density are the dominant controls on heat transfer, allowing thermal conductivity to be parameterised using water content and bulk density without additional compositional predictors. To avoid laboratory-induced artefacts, air-dry specimens with negligible moisture variability were excluded, restricting the analysis to hydro-mechanical states representative of engineering conditions. A detailed justification of this subset definition is provided in the Supplementary Materials.

Rather than proposing a new predictive formulation, this study develops and evaluates a machine-learning-based workflow for thermal conductivity parameterisation that explicitly considers both predictive accuracy and functional plausibility. Multiple regression models representing contrasting modelling strategies are benchmarked against selected physical formulations within a unified and reproducible framework. Model performance is assessed using conventional accuracy metrics supplemented by diagnostics of functional stability relevant for interpolation, spatial mapping, and reuse in engineering applications.

The main contributions of this study are threefold. First, a systematic evaluation framework is introduced that extends beyond pointwise accuracy and explicitly assesses the physical plausibility of modelled relationships. Second, a quantitative diagnostic measure—the Physical Consistency Index (PCI)—is used to characterise the monotonicity, smoothness, and admissibility of model-predicted thermal conductivity fields. Third, the study demonstrates a consistent trade-off between predictive accuracy and functional consistency across different machine-learning model families.

By distinguishing between numerical accuracy and functional reliability, the study supports the informed selection of data-driven models for geotechnical and shallow geothermal applications and provides a transparent basis for generating stable and physically interpretable thermal-property fields from archival datasets. In this context, the study demonstrates that maximising predictive accuracy alone does not guarantee the suitability of a model for engineering parameterisation, as the stability and physical coherence of the inferred functional relationships are equally critical.

2. Materials and Methods

2.1. Dataset Origin, Geological Context, and Sampling Strategy

The dataset analysed in this study originates from the national thermal-property testing programme conducted at the Polish Geological Institute–National Research Institute (PGI–NRI). Within routine geotechnical and shallow geothermal investigations, soils and rocks collected during site characterisation campaigns are systematically tested for thermal conductivity using standardised laboratory procedures. The resulting database constitutes a large, internally consistent archive that is routinely used to support engineering design, subsurface heat-transfer modelling, and regional geothermal assessments [5,17].

Samples submitted for testing were obtained during standard site investigations and were not pre-selected based on lithology. For the purposes of this study, the analysis was restricted to materials classified macroscopically as sands or sand-dominated soils in accordance with PN-EN ISO 14688-1:2018-05 [18], ensuring consistency with established geotechnical practice. Lithological classification was performed by trained personnel using macroscopic criteria routinely applied in engineering investigations.

The analysed subset is further restricted to quartz-rich sands, which constitute a lithologically coherent material class in which mineralogical variability exerts only a secondary influence on effective thermal conductivity [4,6,8–10,17]. In such materials, saturation state and packing density represent the dominant controls on heat transfer, allowing thermal conductivity to be parameterised using water content and bulk density without introducing additional compositional predictors. This restriction enhances the reliability and interpretability of derived parameterisations in engineering and geothermal applications.

All samples entering the PGI–NRI laboratory were processed using unified preparation and testing workflows independent of lithology. Bulk material was homogenised and, where necessary, lightly screened to remove oversized particles that could interfere with probe insertion. Specimens were prepared in rigid cylindrical moulds of known volume, enabling controlled generation of multiple hydro-mechanical states through systematic adjustment of water content and compaction level.

Moisture content was increased stepwise through controlled water addition combined with thorough mixing, while bulk density was modified using vibratory compaction, mechanical compaction, or Proctor-type procedures following standard geotechnical laboratory practices for non-cohesive soils [17,18]. This preparation strategy enabled efficient and reproducible coverage of the moisture–density domain relevant for shallow subsurface heat-transfer conditions.

As a result, the dataset represents a lithologically coherent, physically controlled, and engineering-relevant parameter space suitable for evaluating machine-learning-based thermal conductivity models intended for interpolation, spatial mapping, and reuse in geotechnical design, geothermal simulations, and subsurface heat-transfer modelling.

2.2. Laboratory Measurements

Water content, bulk density, and thermal conductivity (λ) were determined using standardised laboratory procedures routinely applied in the PGI–NRI thermal-property laboratory. All measurements followed established international and European standards to ensure consistency with nationwide geotechnical and geothermal datasets and to provide reliable input data for engineering analysis, geothermal modelling, and subsurface heat-transfer studies.

Water content was determined gravimetrically in accordance with PN-EN ISO 17892-1:2015-02 [19]. Following thermal-conductivity testing, representative subsamples were weighed and oven-dried at 105 °C until constant mass was achieved. Water content was calculated as the ratio of evaporable water mass to dry soil mass and expressed as a gravimetric percentage. Drying was performed only after λ measurements and therefore did not affect recorded thermal-conductivity values.

Bulk density was measured in rigid cylindrical moulds of known volume following PN-EN ISO 17892-2:2015-02 [18]. Prepared specimens were trimmed flush with the mould and weighed using an analytical balance with a precision of 0.01 g. Bulk density ρ_b [g/cm^3] was calculated as the ratio of specimen mass to mould volume. Replicate specimens were prepared at each target density to verify measurement reproducibility and to limit procedural variability relevant for engineering and geothermal applications.

Thermal conductivity was determined using transient probe methods in accordance with ASTM D5334 [20] and IEEE 442 [21], employing Tempos (METER Group) and TK04 (Lippmann) instruments. During each test, a controlled heat pulse was applied and temperature evolution was recorded as a function of time. Under the line-source assumption specified in the standards, λ was derived from the slope of the temperature–log(time) relationship. Measurements were conducted under controlled laboratory temperatures ranging from 8 to 20 °C.

For each hydro-mechanical state, at least three independent thermal-conductivity measurements were performed. If discrepancies between replicate measurements exceeded approximately 5%, additional tests were conducted and the most consistent values were retained. This quality-control procedure ensured robust estimation of λ while maintaining throughput suitable for large-scale engineering laboratory programmes.

The applied laboratory workflow generated specimens spanning a wide range of water content and bulk density conditions, including both moist ($w \geq 1\%$) and air-dry or oven-dry ($w < 1\%$) states. Exploratory analysis (Section 2.4) indicates that the dry subset exhibits negligible moisture variability and forms a compact cluster in multivariate space, reflecting laboratory drying procedures rather than hydro-mechanical states representative of shallow subsurface conditions. Consequently, only moist samples were retained for modelling and interpretation in order to preserve physically meaningful relationships relevant to subsurface heat-transfer processes.

The final dataset therefore consists exclusively of moist sands (w $\geq$ 1%), each characterised by controlled water content, bulk density, and replicated thermal-conductivity measurements obtained under rigorously standardised laboratory conditions. This provides a high-quality and reliable basis for data-driven thermal conductivity parameterisation intended for geotechnical design, geothermal modelling, and heat-transfer applications in near-surface soils.

2.3. Data Preprocessing and Subset Definition

The raw dataset comprised laboratory measurements of thermal conductivity (λ), water content, and bulk density collected between 2018 and 2025 for sand samples submitted to the PGI–NRI laboratory as part of nationwide geotechnical and shallow geothermal investigations. Prior to modelling and interpretation, the data were subjected to a structured quality-control and qualification workflow designed to ensure physical plausibility, internal consistency, and suitability for subsequent engineering and geothermal parameterisation.

Records with missing values, duplicated identifiers, or values outside realistic ranges for non-cohesive sands were removed. Applied thresholds included water content between 0 and 40%, bulk density between 1.2 and 2.2 g/cm^3, and thermal conductivity between 0 and 7.6 W/m·K. These limits reflect documented ranges for unconsolidated sandy soils and operational limits of laboratory testing. After quality control, the cleaned dataset comprised 1716 specimens classified macroscopically as sands or sand-dominated soils according to PN-EN ISO 14688-1:2018-05 [22].

Inspection of the water content distribution revealed a pronounced bimodal structure, with one population exhibiting measurable moisture contents and a second population characterised by values close to zero. Based on this pattern and laboratory preparation procedures, two operational subsets were defined using a water content threshold of 1%: air-dry sands (w < 1%, N = 261) and moist sands (w $\geq$ 1%, N = 1455).

Descriptive statistics for the full ("global") dataset and for both subsets are summarised in Table 1. Statistics for the global dataset are reported for completeness, reflecting the structure of the national laboratory archive, whereas subsequent analyses focus on the moist-sand subset. The air-dry subset is characterised by negligible moisture variability, extreme right-skewness, and narrow interquartile ranges. In multivariate space, these samples form a compact and clearly separated cluster (Figure S1), indicating that they represent laboratory drying conditions rather than hydro-mechanical states representative of shallow subsurface environments.

Table 1. Descriptive statistics of water content, bulk density and thermal conductivity for the global dataset and for the air-dry (w < 1%) and moist (w $\geq$ 1%)-sand subsets. Reported metrics include sample size (N), mean value (Mean), standard deviation (Std), minimum (Min), maximum (Max), skewness (Skew), kurtosis (Kurt), coefficient of variation (CV) and interquartile range (IQR).

Variable	Subset	N	Mean	Std	Min	Max	Skew	Kurt	CV	IQR
Water content [%]	Global	1716	8.58	7.19	0.00	28.94	0.58	−0.95	0.84	12.48
	Air-dry	261	0.09	0.22	0.00	0.76	2.64	5.24	2.34	0.00
	Moist	1455	10.11	6.76	1.00	28.94	0.49	−1.10	0.67	12.31
Bulk density [g/cm^3]	Global	1716	1.77	0.19	1.21	2.18	−0.11	−0.38	0.11	0.26
	Air-dry	261	1.65	0.15	1.23	2.14	−0.37	1.66	0.09	0.12
	Moist	1455	1.79	0.19	1.21	2.18	−0.18	−0.53	0.11	0.28
Thermal conductivity [W/m·K]	Global	1716	1.63	0.77	0.06	3.15	−0.34	−0.75	0.47	1.03
	Air-dry	261	0.37	0.30	0.20	1.73	3.01	0.82	0.17	0.22
	Moist	1455	1.86	0.59	0.06	3.15	−0.10	−0.57	0.32	0.88

In contrast, the moist-sand subset spans a broad and continuous range of water content, bulk density, and thermal conductivity values (Table 1; Figure 1). Bivariate relationships within this subset are monotonic and physically interpretable, with thermal conductivity increasing systematically with both water content and bulk density. This subset therefore defines a physically meaningful hydro-mechanical domain suitable for interpolation, spatial mapping, and thermal parameterisation of sandy soils.

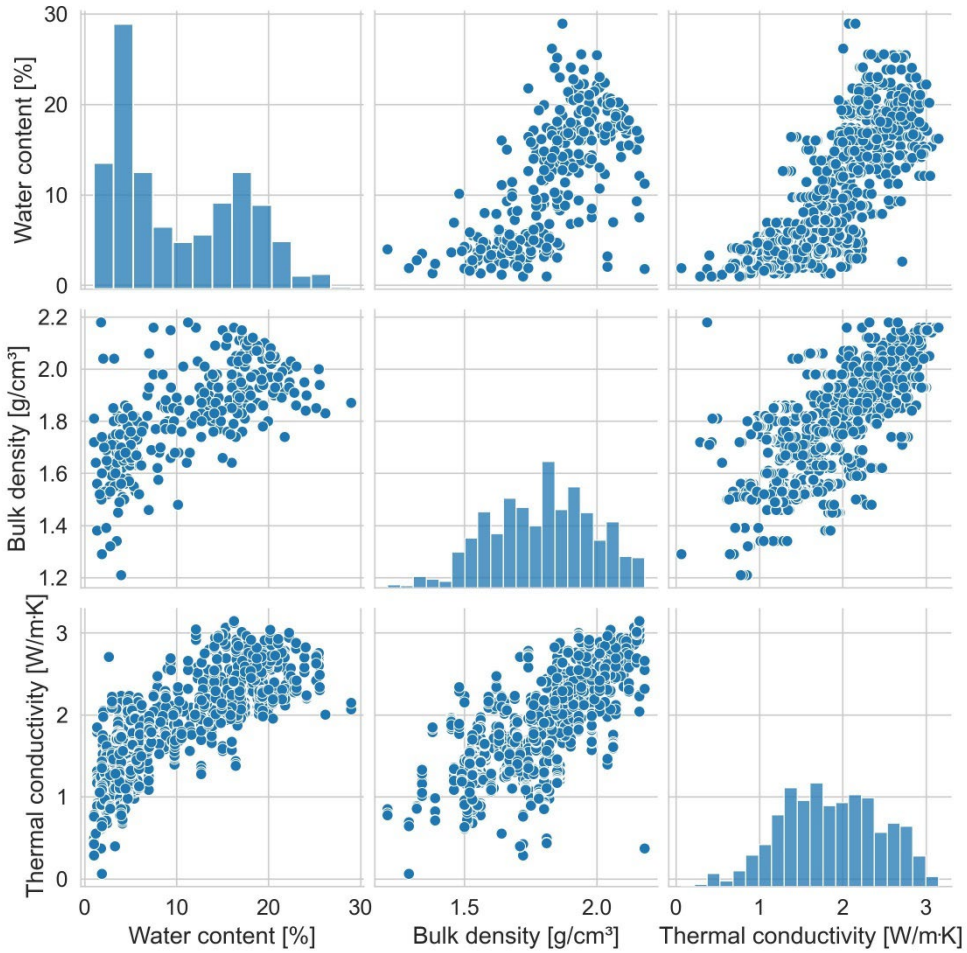

Figure 1. Pairplot illustrating marginal distributions and bivariate relationships among water content, bulk density and thermal conductivity λ for the moist-sand subset (w ≥ 1%).

Accordingly, all subsequent analyses were performed exclusively on the moist-sand subset (w ≥ 1%), while air-dry samples were retained only for diagnostic and descriptive purposes. Including these states would require separate physical treatment and dedicated modelling frameworks beyond the scope of the present study. This subset definition (w ≥ 1%) ensures that machine-learning models are trained and evaluated on data representing realistic hydro-mechanical conditions relevant for geotechnical design and shallow geothermal heat-transfer applications.

2.4. Exploratory Data Analysis (EDA)

Exploratory data analysis was conducted to verify the internal structure and physical coherence of the moist-sand subset (w ≥ 1%) prior to predictive modelling. Following the subset definition in Section 2.3, all analyses presented here refer exclusively to hydro-mechanical states representative of engineering and shallow geothermal conditions.

Pairwise relationships and marginal distributions (Figure 1) indicate that the dataset occupies a broad and continuous domain of water content, bulk density, and thermal

conductivity. All three variables exhibit unimodal distributions, and no evidence of discrete material classes, phase transitions, or abrupt behavioural changes is observed within the analysed range. Thermal conductivity increases progressively with both predictors across the domain, consistent with established physical expectations for partially saturated sands and heat transfer in porous granular materials. The distribution of thermal conductivity across the predictor space does not indicate strong imbalance or sparsely populated regions, supporting stable model training within the analysed domain.

Correlation analysis was used to assess linear dependence between variables and to exclude problematic collinearity prior to regression modelling [23]. The correlation matrix (Figure 2) allows direct quantitative comparison of pairwise relationships independently of distributional effects visible in the pairplot and confirms that thermal conductivity is strongly coupled to water content and moderately correlated with bulk density. Correlations between predictors remain within acceptable limits, indicating that water content and bulk density provide complementary and non-redundant descriptions of the hydro-mechanical state. Figure 2 provides a quantitative complement to the graphical relationships shown in Figure 1.

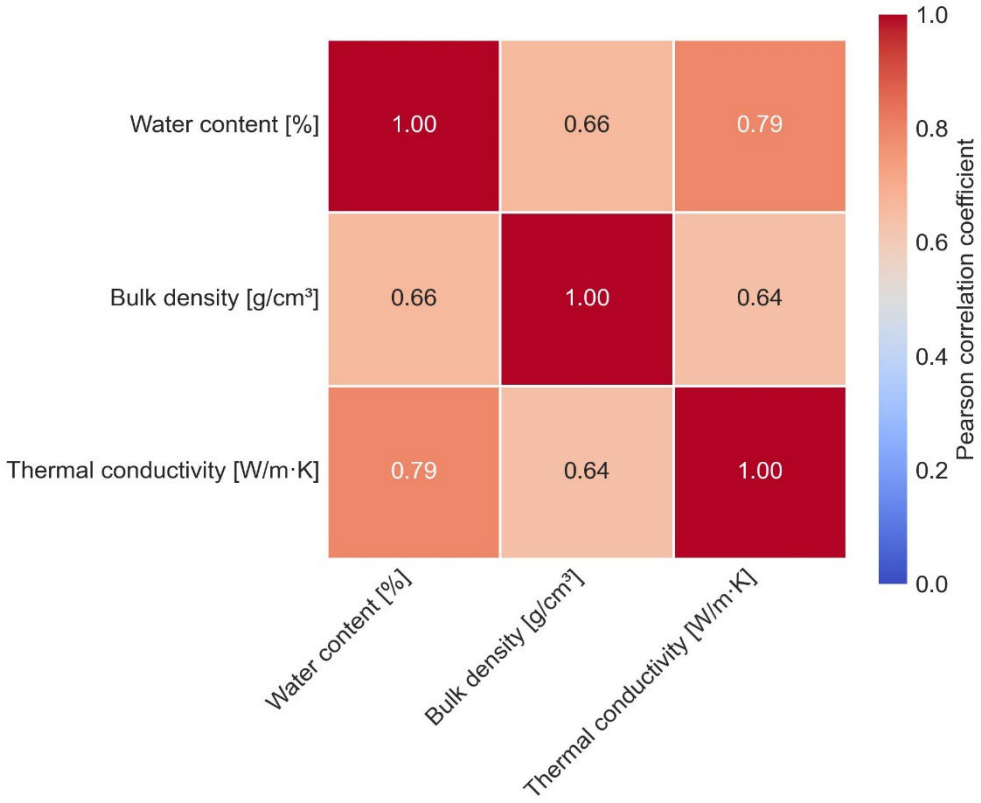

Figure 2. Pearson correlation matrix for water content, bulk density and thermal conductivity λ in the moist-sand subset ($w \geq 1\%$).

Principal Component Analysis (PCA) was applied as an exploratory tool to obtain a complementary multivariate representation of the dataset structure [24]. In the space defined by the first two principal components, the samples form a continuous distribution without evidence of discrete subpopulations (Figure 3). The first principal component (PC1) captures the dominant combined variance of water content, bulk density, and thermal conductivity and represents a progressive hydro-mechanical gradient, whereas the second component (PC2) reflects secondary variability without a clear independent physical interpretation. This behaviour supports interpretation of the analysed parameter space as a continuous physical domain rather than a mixture of distinct material classes.

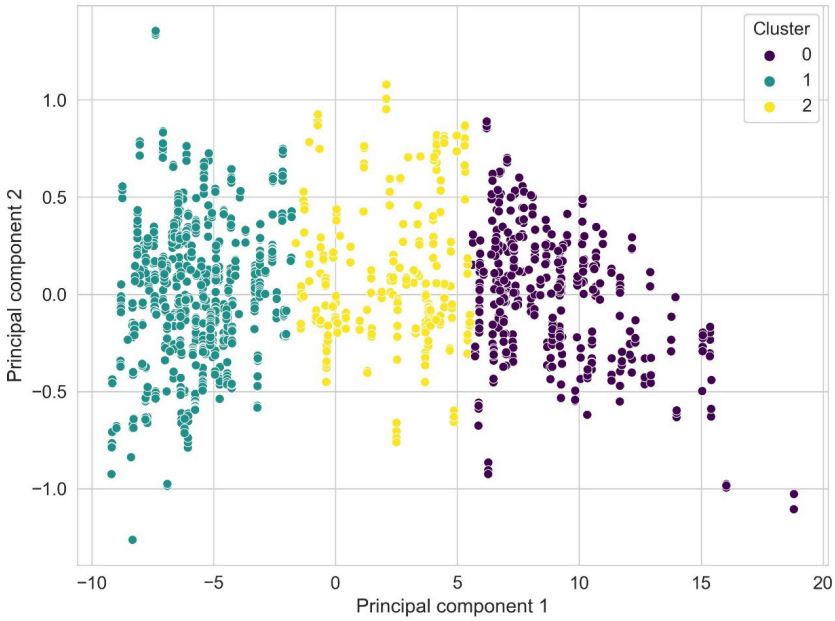

Figure 3. Principal Component Analysis (PCA) of the moist-sand subset (w ≥ 1%)—projection of hydro-mechanical variables. Points are coloured by k-means clusters (k = 3) in the principal component 1–principal component 2 space to visualise the continuous structure of the dataset and broad hydro-mechanical regimes.

To support visual interpretation of this continuous structure, k-means clustering was applied in the PCA space [25] (Figure 3). The resulting clusters partition the dataset into broad segments corresponding to gradual variations in hydro-mechanical conditions. These clusters do not represent discrete geological or lithological units but instead reflect continuous transitions within the moisture–density domain. The numerical ranges of principal components are dataset-specific and do not have direct physical meaning. Clustering is used here solely as a visualisation aid and does not imply material classification or influence subsequent modelling. PCA is used here as an auxiliary exploratory tool and does not influence the subsequent modelling workflow.

Overall, the exploratory analyses demonstrate that thermal conductivity within the moist-sand subset varies continuously and approximately monotonically as a function of water content and bulk density. This structure confirms the suitability of the dataset for parameterisation, interpolation, and mapping of thermal conductivity as a continuous engineering material property and provides a physically justified basis for subsequent machine-learning modelling.

2.5. Machine-Learning Workflow

Data-driven modelling was performed exclusively on the moist-sand subset (w ≥ 1%), representing hydro-mechanical states relevant for engineering and geothermal applications (Sections 2.3 and 2.4). Water content [%] and bulk density [g/cm^3] were used as predictor variables, while thermal conductivity λ [W/m·K] served as the response variable. The overall modelling workflow is summarised in Figure 4.

Prior to model training, the cleaned dataset was divided into training and test subsets using an 80/20 split with a fixed random seed (random_state = 42). All model calibration, hyperparameter optimisation, and cross-validation procedures were conducted exclusively on the training subset to prevent information leakage. The independent test subset was reserved for final performance assessment. Hyperparameter tuning was performed using k-fold cross-validation within the training dataset.

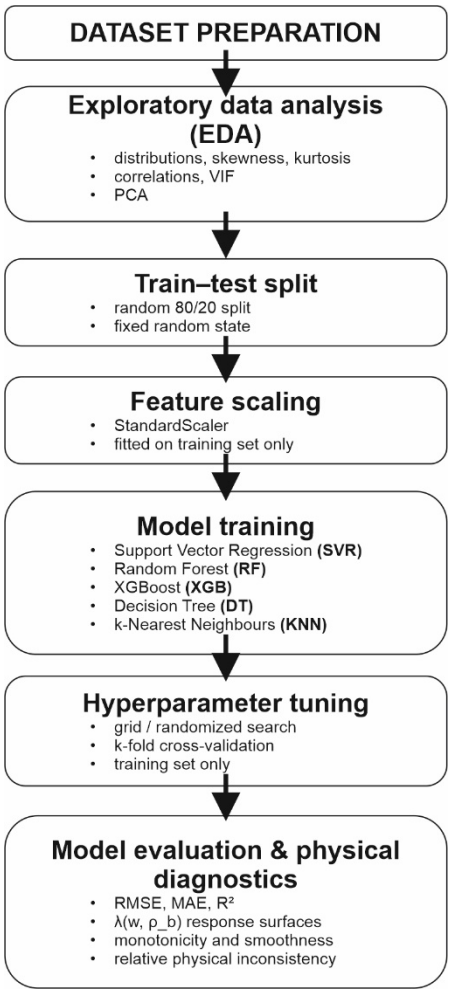

Figure 4. Machine-learning workflow used to model thermal conductivity of moist sands, including data preparation, model training, hyperparameter tuning, performance evaluation and physically motivated diagnostics.

Five established regression algorithms were selected to represent complementary modelling strategies commonly applied in geotechnical and environmental engineering: Support Vector Regression (SVR), Decision Tree regression (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost, XGB), and k-Nearest Neighbours (KNN). Together, these models represent a broad range of global and local approximation behaviours, enabling systematic evaluation of their suitability for thermal conductivity parameterisation. The selection of these models was guided by the objective of analysing differences in inductive bias and functional behaviour rather than maximising predictive performance alone. Deep neural network (DNN) approaches were not considered in this study due to the moderate dataset size and the focus on interpretable model behaviour, as highly flexible models are expected to exhibit response-surface characteristics similar to other high-capacity methods (e.g., ensemble models) without providing additional insight into the accuracy–consistency trade-off. Furthermore, the objective of this study is not to maximise predictive performance but to analyse model behaviour, for which interpretable model families are sufficient. Moreover, the inclusion of DNN models would primarily increase model complexity without providing additional interpretability within the scope of this study.

For each algorithm, several model configurations were examined, including baseline, cross-validated, and manually regularised variants. This approach enabled assessment of the influence of model flexibility on predictive performance and functional stability.

All configurations were trained using the same training–testing split and included in the comparative analysis. Model naming conventions (e.g., SVR_2, RF_T) follow the definitions provided in the Supplementary Materials (Section S3). Detailed hyperparameter search spaces and final model configurations are reported in the Supplementary Materials (Section S3; Table S1).

Hyperparameter tuning was performed using cross-validation on the training subset only. RandomizedSearchCV was applied for SVR, RF, and XGB, while GridSearchCV was used for DT and KNN, reflecting differences in the structure of their parameter spaces. All implementations were based on the scikit-learn machine-learning library [26], a widely used open-source Python (v. 3.11.5) framework for statistical learning and data analysis. The coefficient of determination (R^2) was used as the primary optimisation criterion, as it provides a scale-independent measure of explained variance and is commonly used in regression benchmarking.

Feature scaling was applied only where required by the modelling approach. SVR and KNN were implemented within scikit-learn Pipeline objects, which ensure that preprocessing steps (e.g., standardisation) are consistently applied within cross-validation folds and fitted exclusively on training data, preventing data leakage. In contrast, tree-based models (DT, RF, and XGB), which are invariant to monotonic transformations of predictor variables, were trained on unscaled inputs.

Model performance was evaluated on the independent test set using complementary accuracy metrics, including R^2, mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). These indicators quantify numerical agreement between predicted and measured values and provide a baseline for comparison of alternative modelling strategies. Agreement between predicted and measured values is evaluated quantitatively using these metrics and residual diagnostics, which provide a comprehensive assessment without introducing redundant visualisations.

Because thermal conductivity is treated here as a continuous engineering material parameter, model behaviour was additionally examined beyond pointwise accuracy. For each trained model, two-dimensional response surfaces $\lambda(w, \rho_b)$ were generated over a regular grid spanning the observed water content–bulk density domain. These surfaces enable direct assessment of smoothness, continuity, and monotonic behaviour relevant for interpolation and spatial mapping of thermal properties. Direct overlay of measured data points on response surfaces was not applied in order to preserve the clarity of the global surface structure and avoid visual clutter, which would hinder interpretation of model behaviour.

Figure 5 presents illustrative examples of characteristic response-surface geometries associated with different regression model families. The schematic visualises the general effects of model structure on global continuity and local variability. Axes are abstract and do not correspond to physical variables or value ranges used elsewhere in the study. As the figure is conceptual, it is provided to support interpretation of model behaviour rather than to present empirical results.

2.6. Physical Consistency Assessment and Definition of the Physical Consistency Index (PCI)

Conventional performance metrics such as R^2, RMSE, MAE, and MAPE quantify the numerical agreement between predicted and measured values but do not indicate whether a model encodes physically interpretable and stable relationships between thermal conductivity, water content, and bulk density. In geotechnical and shallow geothermal engineering, thermal conductivity is typically treated as a continuous material parameter that is subsequently interpolated, mapped, and reused in design calculations. For such

applications, numerical accuracy alone is insufficient if the inferred functional relationships exhibit local instabilities, segmentation, or physically implausible behaviour.

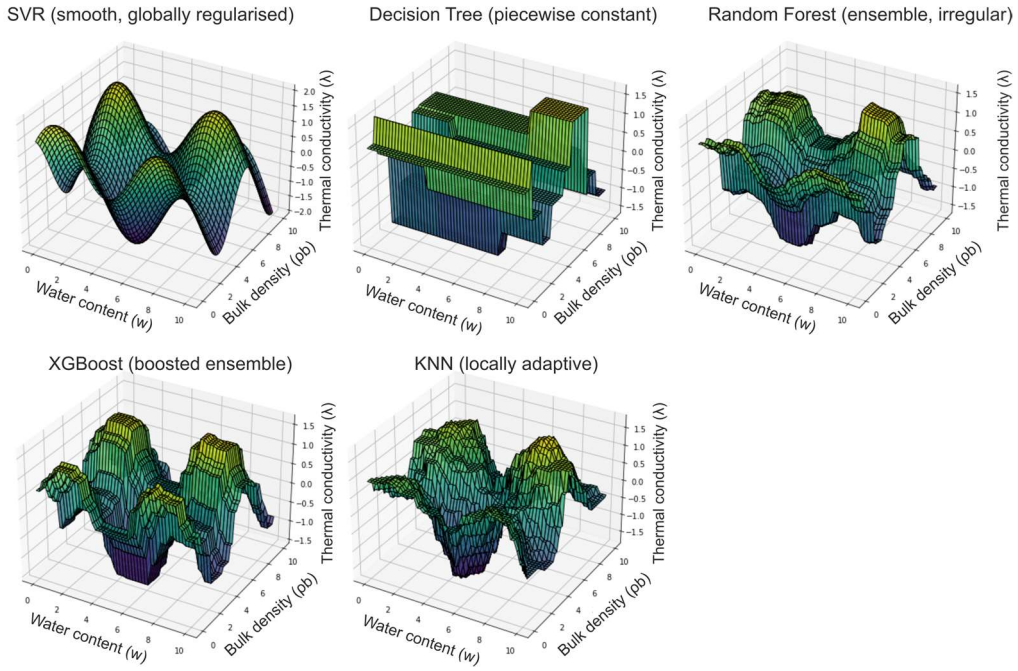

Figure 5. Conceptual illustration of characteristic response-surface geometries associated with different regression model families. Axes represent generic predictor variables (e.g., water content and bulk density) and the corresponding response (thermal conductivity). The colors represent a scale for thermal conductivity values: darker shades show the lowest values, while lighter shades show the highest values. The figure highlights differences in smoothness, continuity, and local variability arising from model structure. This schematic does not represent actual model predictions.

For unconsolidated quartz-rich sands tested under controlled laboratory conditions, several qualitative expectations can be formulated when thermal conductivity is interpreted as an engineering material parameter. Thermal conductivity is expected to increase monotonically with increasing water content and bulk density, reflecting improved thermal connectivity within the pore space and enhanced grain-to-grain heat transfer within the solid skeleton [6–9]. In addition, its dependence on these variables is expected to vary smoothly across the hydro-mechanical domain, since abrupt local oscillations or discontinuities would imply unrealistically sharp transitions in material behaviour lacking geological or laboratory justification.

To complement conventional accuracy metrics, model behaviour was therefore evaluated using a diagnostic framework based on the Physical Consistency Index (PCI). The term "physical consistency" refers to qualitative structural expectations rather than direct validation against physical laws. PCI is a fully quantitative composite metric computed exclusively from model-predicted response surfaces $\lambda(w, \rho_b)$ evaluated over a regular grid spanning the observed water content–bulk density domain. The index provides a quantitative summary of selected qualitative properties relevant for engineering interpretation of thermal-conductivity parameterisations and enables consistent comparison of models with different structural characteristics.

PCI integrates three classes of diagnostics:

- Violations of monotonicity with respect to water content and bulk density;
- Surface roughness reflecting local irregularities of the predicted response surface;
- Violations of physically admissible thermal-conductivity bounds.

These diagnostics reflect basic expectations regarding directional coherence, smoothness, and admissible value ranges of thermal conductivity fields used in engineering applications.

The Physical Consistency Index is defined as

$$\mathrm{PCI} = \alpha \mathrm{MVR} + \beta \mathrm{Rough}_{\mathrm{norm}} + \gamma \mathrm{PBV} \quad (1)$$

where MVR denotes the monotonicity violation rate of the predicted response surface, $Rough_{norm}$ represents a normalised measure of surface roughness derived from first-order gradients of the response surface, and PBV quantifies violations of physically admissible thermal-conductivity bounds. The coefficients α, β, and γ are weighting factors satisfying $\alpha + \beta + \gamma = 1$.

In this study, equal weights were assigned to monotonicity and smoothness diagnostics ($\alpha = \beta = 0.4$), while a lower weight was applied to physical-bound violations ($\gamma = 0.2$), which occur infrequently within the analysed domain. This weighting scheme is adopted as a transparent heuristic rather than an optimal formulation and does not affect the relative ranking of model families, as demonstrated in the robustness analysis provided in the Supplementary Materials (Section S2).

Lower PCI values correspond to response surfaces that more consistently satisfy the adopted structural diagnostics relative to alternative modelling approaches. PCI values are interpreted strictly in a relative sense and do not represent an absolute measure of physical correctness. The index does not constitute physical validation of the underlying models but rather provides a practical indicator supporting interpretation of surrogate behaviour when model outputs are reused as continuous engineering parameterisations.

Importantly, PCI is applied strictly as a post hoc diagnostic. No physical constraints are imposed during model training, and the index is not used for model calibration or optimisation. Consequently, PCI reflects emergent properties of trained models resulting from their structure and inductive bias rather than compliance with imposed physical assumptions. The weighting scheme adopted in this study is not intended as a universal formulation but as a transparent and reproducible reference for comparative evaluation within the analysed dataset.

The detailed mathematical formulation of the individual PCI components, including numerical approximations of response-surface gradients and sensitivity analyses of the weighting scheme, is provided in the Supplementary Materials (Section S2).

3. Results

3.1. Trade-Off Between Predictive Accuracy and Physical Consistency of Modelled Thermal Conductivity

Figure 6 summarises the joint evaluation of all analysed models in terms of predictive accuracy and relative physical consistency. Predictive accuracy is expressed by the root mean squared error (RMSE) evaluated on the independent test set, while physical consistency is quantified using the Physical Consistency Index (PCI) defined in Section 2.6. PCI is interpreted strictly as a relative behavioural diagnostic rather than an absolute physical metric and is used here solely to compare surrogate models trained and evaluated within the same predictor domain. Lower PCI values indicate higher physical consistency of the predicted response surfaces. Each point in Figure 6 represents one of the five evaluated variants for a given model family, as defined in Section 2.5 and detailed in Table S1.

In addition to machine-learning models, Figure 6 includes selected classical physical formulations of thermal conductivity (Kersten [6], Johansen [7], and Lu et al. [9]) used as reference benchmarks. These models were applied using water content and bulk density as inputs without local recalibration and evaluated on the same test dataset. Their per-

formance therefore reflects the behaviour typically observed when physically constrained formulations are used as general engineering parameterisations.

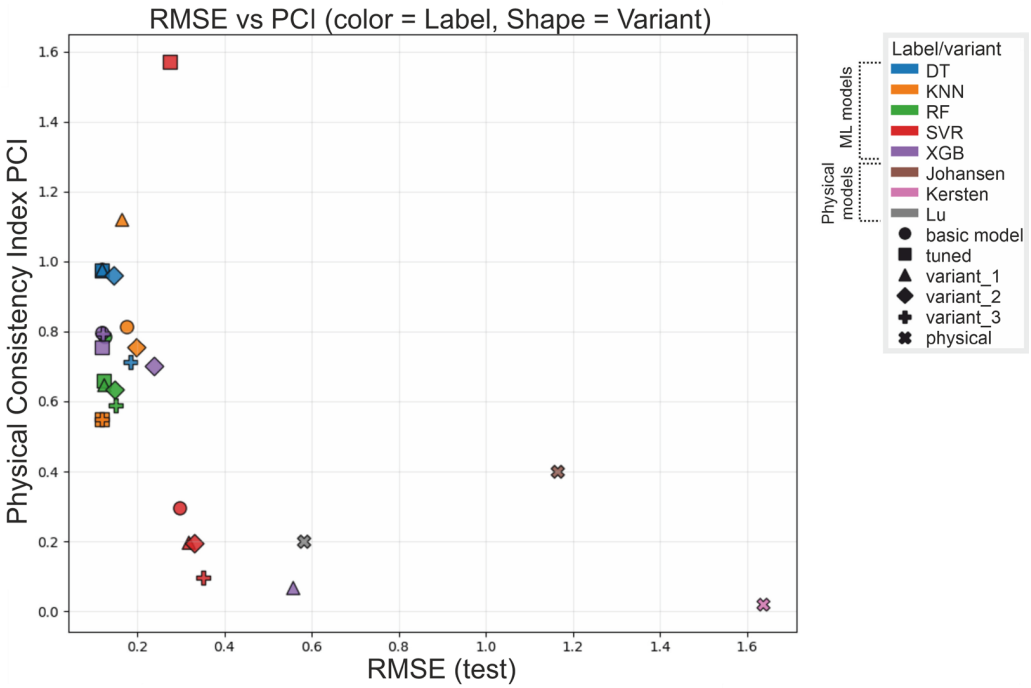

Figure 6. Trade-off between predictive accuracy and physical consistency for all evaluated machine-learning models. Each point represents a single model variant plotted as test-set RMSE versus the Physical Consistency Index (PCI). Distinct clustering of model families illustrates systematic differences in the balance between numerical accuracy and physical consistency.

The distribution of the model results in the RMSE–PCI space reveals a systematic trade-off between predictive accuracy and functional consistency within the analysed dataset. Tree-based and instance-based models, including Decision Trees, Random Forests, XGBoost, and k-Nearest Neighbours, cluster in regions characterised by low RMSE values, typically between approximately 0.12 and 0.15 W/m·K, combined with comparatively higher PCI values. These models therefore achieve high numerical accuracy but tend to generate response surfaces exhibiting local irregularities or segmentation when evaluated using the adopted diagnostics.

Support Vector Regression (SVR) models occupy a distinct region of the RMSE–PCI space. Compared with locally adaptive approaches, SVR variants exhibit higher RMSE values, on the order of approximately 0.30–0.33 W/m·K for the most physically consistent configurations, but substantially lower PCI values. This behaviour reflects stronger global regularisation and smoother, more coherent functional relationships at the cost of reduced local flexibility.

One SVR configuration forms an extreme point in the RMSE–PCI space, exhibiting the lowest PCI value among all evaluated models but markedly elevated RMSE. This configuration represents the limit of functional regularity attainable within the analysed modelling framework and illustrates the inherent trade-off between numerical accuracy and structural smoothness.

Classical physical models occupy a separate region of the RMSE–PCI space characterised by very low PCI values and substantially higher RMSE compared with all data-driven approaches. Their position reflects the strong physical constraints embedded in their functional forms, which promote monotonicity and smoothness but limit adaptability to site-specific variability when applied without recalibration.

Overall, Figure 6 indicates that no single model configuration simultaneously minimises prediction error and physical inconsistency within the analysed domain. Instead, the evaluated models define a Pareto-like distribution in the RMSE–PCI space, representing alternative compromises between predictive accuracy and functional consistency. This pattern reflects the systematic differences in model structure and inductive bias rather than the effects of individual hyperparameter choices.

Supplementary sensitivity and ranking analyses indicate that the relative positioning of model families in the RMSE–PCI space remains stable under alternative formulations of physical-consistency diagnostics (Supplementary Section S2; Figure S2). This confirms that the observed trade-off reflects systematic properties of the modelling approaches rather than artefacts of a specific diagnostic definition.

From an engineering perspective, these results indicate that model selection for thermal-conductivity parameterisation involves an explicit balance between numerical accuracy and functional coherence. In applications prioritising stable interpolation and spatial mapping of thermal properties, moderately less accurate but structurally smoother models may therefore be preferable, whereas locally adaptive models may be advantageous where pointwise prediction accuracy is the primary objective. These recommendations should be interpreted within the scope of the analysed dataset and modelling framework.

3.2. Comparative Analysis of Two-Dimensional Response Surfaces

Figure 7 compares two-dimensional response surfaces $\lambda(w, \rho_b)$ generated for three representative model variants selected from the full set of evaluated configurations (Table S1): the physically regularised SVR_2 variant, the tuned Random Forest model (RF_T), and the tuned k-Nearest Neighbours model (KNN_T). These variants were selected to represent characteristic regions of the RMSE–PCI space identified in Section 3.1. All surfaces were evaluated over the same water content–density domain defined by the observed data and visualised using an identical colour scale to enable direct structural comparison.

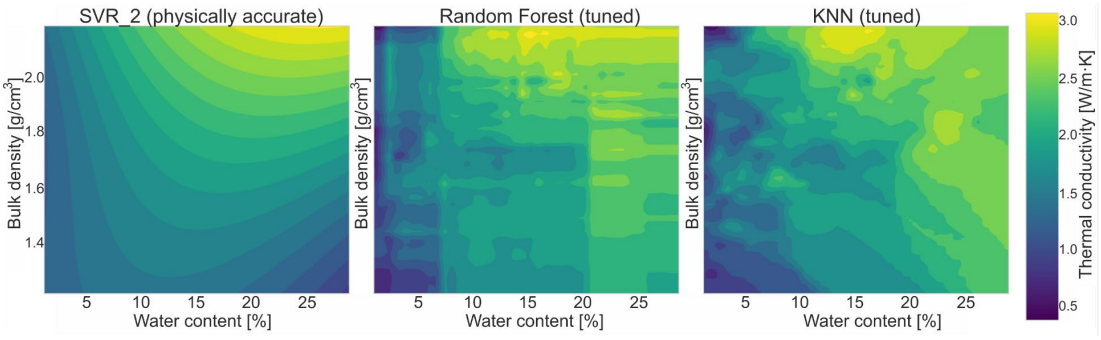

Figure 7. Comparative two-dimensional response surfaces $\lambda(w, \rho_b)$ for three representative models: SVR_2, tuned Random Forest (RF_T), and tuned k-Nearest Neighbours (KNN_T). All surfaces are evaluated over the same water content–density domain and visualised using an identical colour scale, highlighting differences in smoothness, continuity, and global structure.

The response surface produced by the SVR_2 model is smooth and globally coherent across the analysed domain. Thermal conductivity increases monotonically with both water content and bulk density, and iso-conductivity contours are continuous and gently curved. No abrupt transitions, local extrema, or segmented regions are apparent within the evaluated predictor range. This behaviour reflects strong global regularisation and results in a stable functional representation suitable for interpolation and spatial mapping applications.

In contrast, the tuned Random Forest model produces a response surface composed of locally homogeneous regions separated by relatively sharp transitions. While the overall increase in λ with water content and bulk density is preserved at a broad scale, local plateaus

and step-like changes are visible. These features result in discontinuous iso-conductivity contours and variations in local surface gradients within the water content–density domain. Such behaviour is consistent with recursive partitioning and ensemble aggregation and reflects the strong local adaptivity of tree-based models.

The response surface generated by the tuned k-Nearest Neighbours model exhibits locally smooth behaviour but reduced global regularity. Undulating patterns, isolated anomalies, and locally distorted contours are visible, particularly in regions of lower data density. Although λ generally increases with both predictors, the surface contains irregular features that vary in magnitude and spatial extent across the domain. This reflects the strong dependence of instance-based models on local neighbourhood structure and sampling density.

Table 2 summarises predictive performance and selected surface-based consistency metrics for representative model variants evaluated on the independent test dataset. Reported metrics include the coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), together with monotonicity indices with respect to water content (w) and bulk density (ρ_b), and a surface smoothness metric derived from first-order gradients. These metrics are reported to support the transparent comparison of model behaviour and are not used as optimisation criteria. The composite Physical Consistency Index (PCI) is used exclusively for comparative assessment in the Pareto analysis presented in Section 3.1.

Table 2. Predictive performance and selected surface-based consistency metrics for representative model variants evaluated on the independent test dataset. Reported metrics include R^2, RMSE, MAE, MAPE, monotonicity indices and surface smoothness. PCI is used exclusively for comparative assessment in Section 3.1.

Model Family	Variant	R^2 [-]	RMSE [W/m·K]	MAE [W/m·K]	MAPE [%]	Monotonicity (W)	Monotonicity (ρ_b)	Smoothness
KNN	tuned	0.9594	0.1188	0.0690	4.84	0.812	0.711	0.107
Random Forest	tuned	0.9567	0.1226	0.0756	5.30	0.790	0.725	0.106
XGBoost	tuned	0.9594	0.1188	0.0703	4.94	0.752	0.432	0.113
Decision Tree	tuned	0.9596	0.1184	0.0689	4.85	0.813	0.290	0.122
SVR	variant 2	0.6845	0.3310	0.2505	19.24	0.431	0.960	0.044

Despite broadly similar predictive accuracy according to standard error metrics, the analysed models encode different functional representations of the λ–water content–bulk density relationship. These differences are expressed primarily in surface smoothness, continuity, and the presence or absence of local segmentation rather than in overall trend direction. In particular, models achieving similar RMSE values may exhibit different degrees of local stability and structural coherence.

To complement the two-dimensional analysis, Figure 8 presents three-dimensional visualisations of the same response surfaces. The three-dimensional perspective highlights differences in global surface curvature and continuity that may be less apparent in contour plots alone. In particular, the low-curvature geometry of the SVR_2 surface contrasts with the stepped morphology of the Random Forest model and the locally irregular geometry of the KNN_T surface.

The differences observed in Figures 7 and 8 are consistent with the quantitative monotonicity, smoothness, and PCI diagnostics reported in Sections 2.6 and 3.1 and remain stable under alternative diagnostic formulations presented in the Supplementary Materials (Section S2). This consistency suggests that the observed surface characteristics

reflect the systematic properties of the models within the analysed framework rather than visualisation artefacts.

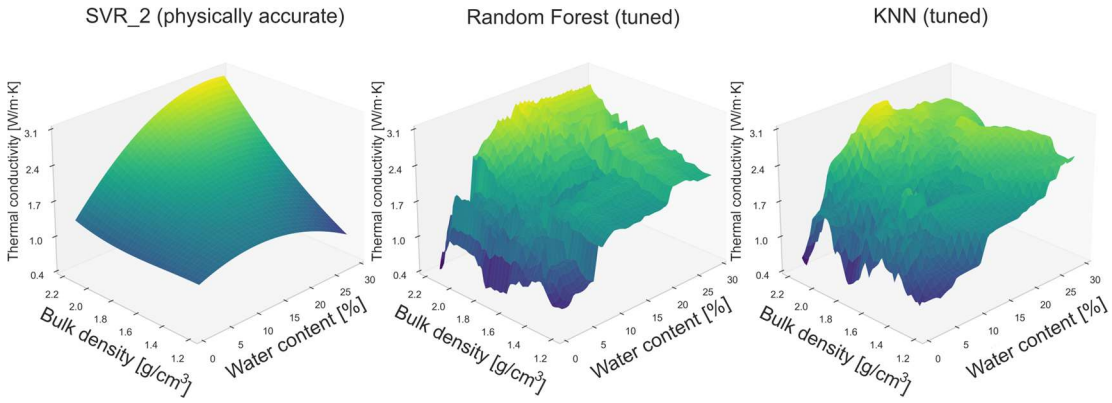

Figure 8. Three-dimensional response surfaces of thermal conductivity λ as a function of water content (w) and bulk density (ρ_b) for three representative models: SVR_2, tuned Random Forest (RF_T), and tuned k-Nearest Neighbours (KNN_T). The colors represent a scale for thermal conductivity values: darker shades show the lowest values, while lighter shades show the highest values.The three-dimensional view illustrates differences in surface curvature, continuity, and local irregularities and is intended for qualitative comparison only.

From an engineering perspective, these results indicate that similar levels of pointwise predictive accuracy may correspond to different spatial representations of thermal conductivity. Locally irregular response surfaces may introduce discontinuities or gradients when used for interpolation or mapping, whereas smoother and more coherent surfaces provide more stable parameter fields for subsequent numerical analyses and geothermal modelling. These observations are intended as qualitative guidance and should be interpreted within the scope of the analysed dataset and modelling framework.

3.3. Local Effects, Interaction Structure and Error Behaviour

While response-surface analysis (Section 3.2) provides a global representation of model behaviour, it does not resolve how predictor effects vary locally across the hydro-mechanical domain. To examine local response stability and interaction structure, model behaviour was analysed using Partial Dependence Plots (PDPs), Individual Conditional Expectation (ICE) curves, and residual diagnostics. These tools provide insight into how predicted thermal conductivity responds to local variations in water content and bulk density and how prediction errors are distributed across the parameter space.

Figure 9 presents combined PDP and ICE analyses for water content and bulk density for three representative modelling approaches selected to illustrate contrasting functional behaviours: the physically regularised SVR_2 model, the tuned Random Forest model (RF_T), and the tuned k-Nearest Neighbours model (KNN_T). All models reproduce the overall increasing trend of thermal conductivity with respect to both predictors; however, differences are observed in local response stability and interaction structure.

For the SVR_2 model, PDP curves for both predictors are smooth and monotonic. Corresponding ICE trajectories form relatively narrow and well-ordered bands with limited curve crossing, indicating consistent local responses across the dataset. This behaviour reflects strong global regularisation and is consistent with the coherent response-surface geometry observed in Figures 7 and 8. From an engineering perspective, such stability supports reliable interpolation and spatial mapping of thermal-conductivity fields.

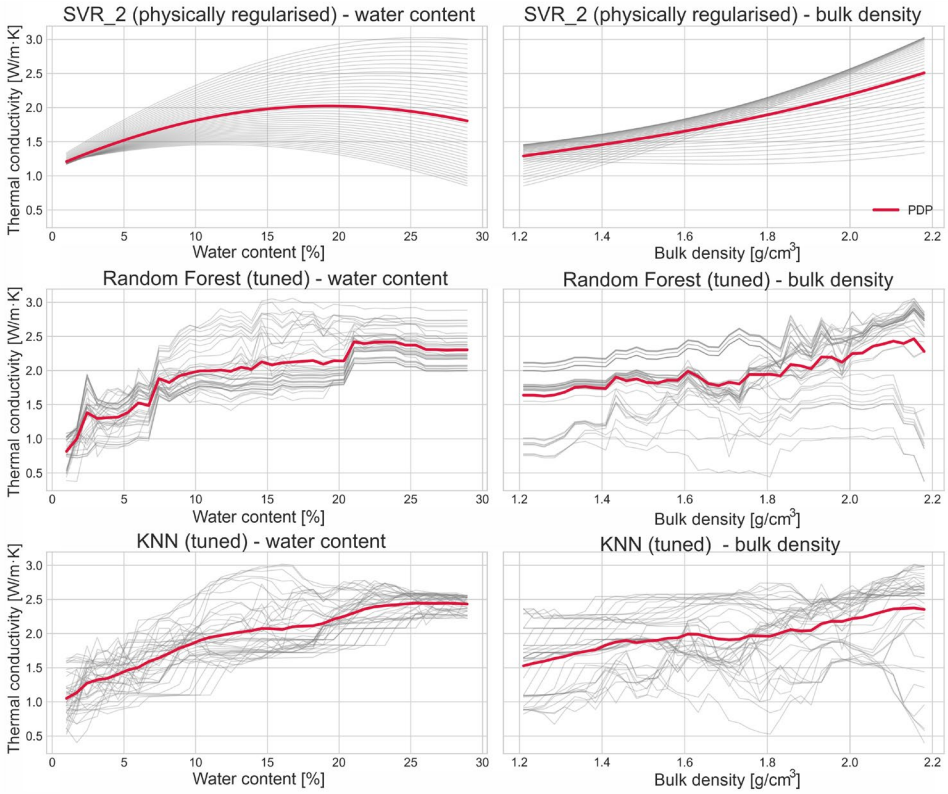

Figure 9. Partial Dependence Plots (PDPs) with Individual Conditional Expectation (ICE) curves illustrating the effect of water content (**left column**) and bulk density (**right column**) on predicted thermal conductivity for three representative models: SVR_2 (physically regularised), tuned Random Forest (RF_T) and tuned k-Nearest Neighbours (KNN_T). Grey lines represent individual conditional responses, while red curves show averaged partial dependence.

The tuned Random Forest model preserves the overall increasing trend of thermal conductivity with respect to both predictors but exhibits stepwise structures in the PDPs, particularly for water content. ICE curves display increased dispersion and frequent crossings, indicating greater local variability compared with the SVR model. This behaviour is consistent with local partitioning and ensemble aggregation and aligns with the response-surface characteristics identified in Section 3.2.

For the KNN_T model, PDPs recover the general direction of predictor effects; however, ICE curves are highly irregular and densely intersecting. Local response patterns vary across the predictor space, indicating sensitivity to neighbourhood structure and sampling density. As a result, the functional representation shows reduced global regularity relative to the other modelling approaches.

To complement the analysis of local response structure, prediction residuals were examined for the SVR_2 and Random Forest models (Figure 10). Residuals are defined as the difference between model-predicted and experimentally measured thermal conductivity values (λ_predicted $-$ λ_measured) and are evaluated on the independent test dataset as a function of water content and bulk density.

For the Random Forest model, residuals display regime-dependent patterns and spatial clustering within specific ranges of water content and bulk density. These patterns suggest localised deviations associated with segmented response behaviour. In contrast, residuals for the SVR_2 model are more uniformly distributed across the predictor space despite larger overall error magnitudes, indicating more homogeneous approximation errors at the scale of the analysed domain.

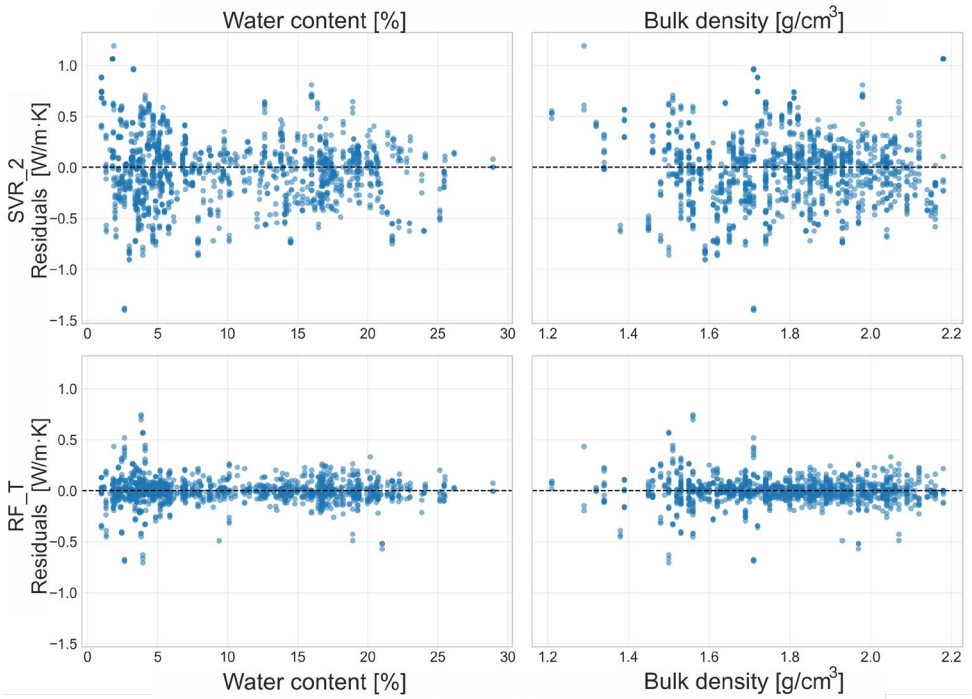

Figure 10. Residuals of predicted thermal conductivity plotted against water content and bulk density for the tuned Random Forest and physically regularised SVR_2 models. The Random Forest exhibits regime-dependent and clustered residuals, whereas SVR_2 shows more uniformly distributed errors across the hydro-mechanical domain.

Residual diagnostics were not repeated for the KNN_T model, as the pronounced irregularity and dense crossing of ICE trajectories already indicate strong neighbourhood-driven variability. In this context, additional residual maps would be unlikely to provide substantially different insights beyond those obtained from the PDP–ICE analysis.

Overall, the combined PDP–ICE and residual analyses indicate that models with comparable predictive accuracy may encode different local response patterns and error structures. This behaviour is consistent with the separation of model families observed in the RMSE–PCI space (Figure 6). These differences are expressed primarily in the stability of local predictor effects and the spatial distribution of prediction errors across the hydro-mechanical domain. When thermal conductivity is used as a reusable engineering material parameter, such local stability becomes an important consideration for ensuring robust interpolation, spatial mapping, and subsequent numerical modelling. These observations should be interpreted as qualitative characteristics of model behaviour within the analysed dataset.

3.4. Summary of Results

The results demonstrate a systematic trade-off between predictive accuracy and surface-based physical-consistency diagnostics across the evaluated modelling approaches. No single model configuration simultaneously minimises prediction error and relative physical inconsistency within the analysed hydro-mechanical domain. Instead, model configurations form a Pareto-like frontier in the RMSE–PCI space, representing alternative compromises between predictive accuracy and functional consistency.

Tree-based and instance-based models achieve the lowest prediction errors according to conventional accuracy metrics; however, their predicted response surfaces exhibit increased segmentation, local irregularities, and reduced stability of local responses. In contrast, Support Vector Regression models produce smoother and more monotonic $\lambda(w, \rho_b)$ relationships, accompanied by higher prediction errors.

Comparative analysis of two- and three-dimensional response surfaces shows that models with similar global accuracy can encode fundamentally different functional representations of the relationship between thermal conductivity, water content, and bulk density. These differences are expressed primarily in surface smoothness, continuity, and the degree of local segmentation rather than in overall trend direction.

Local diagnostic analyses further reveal marked differences in how models encode predictor effects and interaction structure. Partial Dependence and Individual Conditional Expectation analyses indicate stable and ordered local responses for strongly regularised models, whereas locally adaptive models display pronounced local variability and interaction effects. Residual diagnostics show that prediction errors are distributed differently across the hydro-mechanical domain, with some models exhibiting spatially clustered error regimes and others showing more homogeneous error patterns.

Importantly, the separation between model families observed across accuracy, response-surface, and local-diagnostic analyses is consistent across all evaluated variants of each algorithm and does not depend on a particular hyperparameter configuration, as documented in the Supplementary Materials. This consistency indicates that the observed trade-offs reflect intrinsic properties of the modelling approaches rather than artefacts of specific tuning choices.

From an engineering perspective, these findings indicate that model selection for thermal-conductivity parameterisation requires explicit consideration of both numerical accuracy and functional stability. Depending on whether priority is given to pointwise prediction or to reliable interpolation and spatial mapping, different modelling strategies may therefore be appropriate.

4. Discussion

4.1. Accuracy Versus Physical Consistency in Thermal Conductivity Modelling

The results demonstrate that modelling thermal conductivity in moist, quartz-rich sands involves an inherent trade-off between numerical accuracy and physical consistency when model outputs are interpreted as continuous material-property relationships. Models that achieve very low prediction errors according to conventional accuracy metrics do not necessarily provide coherent functional representations of the λ–water content–bulk density relationship when examined beyond pointwise predictions. This distinction is particularly important in engineering applications, where model outputs are reused as continuous parameter fields rather than as isolated predictions.

This trade-off reflects fundamental differences in model structure and approximation strategy rather than sensitivity to a particular hyperparameter configuration. As shown in the RMSE–PCI space (Figure 6), tree-based and instance-based models systematically prioritise local numerical agreement with laboratory measurements. Their high flexibility enables excellent predictive performance but also permits abrupt local changes, segmented response surfaces, and unstable local sensitivities that are difficult to reconcile with the interpretation of thermal conductivity as a smooth engineering material parameter. When such models are used in interpolation or spatial mapping, these artefacts may introduce artificial gradients into reconstructed thermal-property fields.

In contrast, kernel-based Support Vector Regression models impose strong global regularisation, favouring smooth and monotonic functional relationships between thermal conductivity, water content, and bulk density. This structural constraint suppresses local adaptivity and limits the model's ability to reproduce fine-scale variability in the data, resulting in higher prediction errors. However, the resulting response surfaces remain globally coherent and physically interpretable, providing stable parameter fields suitable

for interpolation and spatial mapping. This behaviour highlights the fundamental balance between predictive precision and representational stability.

Classical physically based models form a third, distinct regime within this trade-off. As shown in Table 3, formulations such as those proposed by Kersten [6], Johansen [7], and Lu et al. [9] exhibit minimal physical inconsistency by construction, reflecting explicitly imposed monotonic and smooth functional forms derived from simplified physical assumptions. When applied to the present dataset without site-specific recalibration, these models yield substantially higher prediction errors than all data-driven approaches. This behaviour reflects limitations in transferability and calibration scope rather than deficiencies in their underlying physical concepts. In practice, this suggests that their performance may be improved through local calibration, although this requires site-specific data that are often unavailable in large archival datasets.

Table 3. Predictive performance of selected classical physical models of thermal conductivity (λ) evaluated on the independent test dataset for moist non-cohesive soils. Models were applied without site-specific calibration using water content and bulk density as inputs.

Model	R^2 [-]	RMSE [W/m·K]	MAE [W/m·K]	MAPE [%]
Kersten [6]	−7.02	1.68	1.59	82.3
Johansen [7]	−2.96	1.18	1.02	53.2
Lu et al. [9]	0.09	0.57	0.47	29.8

Taken together, these results indicate that numerical accuracy and physical consistency represent competing but complementary objectives in data-driven parameterisation of subsurface thermal properties. The appropriate balance between them depends on the intended use of model outputs. Applications focused on pointwise prediction may benefit from locally adaptive models, whereas applications requiring interpolation, parameterisation, or reuse of thermal conductivity as a material property benefit from globally regularised formulations that prioritise coherent functional structure and stability over minimal prediction error.

From an engineering perspective, this trade-off should therefore be considered explicitly during model selection. For applications focused on pointwise estimation (e.g., laboratory-scale prediction), high-capacity ensemble models may be preferred. In contrast, for applications involving interpolation, spatial mapping, or numerical modelling of heat transfer, models exhibiting smoother and more physically coherent response surfaces provide more reliable parameter fields. Classical physically based models may serve as a physically constrained reference, particularly when data availability is limited or when interpretability is prioritised over predictive accuracy.

4.2. Model-Dependent Inductive Bias and Its Physical Implications

The observed differences between model families reflect fundamental differences in model structure and approximation strategy, which govern how functional relationships are inferred from finite laboratory datasets. These structural characteristics determine whether a model favours global smoothness or local adaptability and therefore directly influence how thermal conductivity is represented across the hydro-mechanical domain when model outputs are interpreted as continuous material properties rather than as isolated pointwise predictions [27,28]. In this context, model behaviour is primarily controlled by inductive bias, i.e., the inherent assumptions about function shape and smoothness imposed by a given modelling approach.

Tree-based ensembles and instance-based methods approximate complex relationships through local partitioning or neighbourhood-based interpolation [29–31]. This modelling strategy enables excellent predictive accuracy but commonly results in segmented, piecewise, or locally irregular response surfaces when applied to inherently continuous physical systems. Such behaviour has increasingly been recognised as a limitation of accuracy-oriented data-driven models in Earth and environmental sciences, particularly when model outputs are reused as physical surrogates in engineering analyses rather than as purely statistical predictors [14–16,32]. High-capacity models, including ensemble methods and more complex architectures, tend to share this behaviour due to their strong local adaptivity.

In contrast, kernel-based methods such as Support Vector Regression impose strong global regularisation through kernel smoothing and margin maximisation, favouring smooth and globally coherent functional relationships at the expense of local adaptability [27,33]. This behaviour aligns more closely with engineering expectations that thermal conductivity varies continuously and predominantly monotonically with water content and bulk density when treated as a material parameter. As a result, such models provide more stable parameter fields for interpolation and spatial mapping, despite reduced ability to reproduce fine-scale experimental variability.

Classical physically based models represent an extreme form of structural constraint in which monotonicity and smoothness are imposed a priori through fixed functional assumptions rather than learned from data [3,6,7]. While such formulations guarantee physically coherent behaviour by construction, they also limit flexibility and explain the deterioration in predictive performance observed in Table 3 when these models are transferred beyond their original calibration domains.

Independent support for this interpretation is provided by the re-examination of the de Vries [3] thermal conductivity model by Tarnawski et al. [34]. Based on extensive validation against laboratory datasets, the authors demonstrate that classical formulations retain conceptual validity and global physical coherence but exhibit systematic quantitative deviations when applied outside their original calibration domains. These discrepancies are attributed primarily to parameterisation and transferability limitations rather than to deficiencies in the underlying physical representation.

Recent studies on interpretable and explainable machine learning in the natural sciences further emphasise that predictive accuracy alone is insufficient when data-driven models are intended to serve as physically meaningful surrogates rather than as black-box predictors [16]. In such contexts, global functional coherence and consistency with domain knowledge constitute essential criteria for scientific and engineering validity. Post hoc interpretability tools cannot compensate for fundamentally non-physical relationships encoded in the learned response surface.

The present results are consistent with this perspective, demonstrating that models optimised primarily for numerical accuracy may reproduce laboratory measurements very closely while yielding response-surface geometries that are difficult to interpret or justify physically when thermal conductivity is treated as a reusable engineering material parameter. This highlights that model selection in such contexts should be guided not only by predictive performance but also by the structural properties of the inferred functional relationship.

4.3. Physical Interpretability of Response Surfaces and PDP/ICE Diagnostics

Response-surface analysis provides a direct and physically intuitive means of assessing whether a regression model encodes plausible relationships beyond pointwise predictive accuracy. As demonstrated in Section 3.2, models characterised by similar RMSE values may generate fundamentally different $\lambda(w, \rho_b)$ surfaces, ranging from smooth and

globally coherent to segmented or locally distorted. These structural differences have direct implications for the suitability of model outputs for interpolation, spatial mapping and engineering parameterisation.

Local diagnostic tools further clarify these differences by revealing how predictor effects vary across the hydro-mechanical domain. Partial Dependence and Individual Conditional Expectation analyses provide a semi-quantitative description of local model behaviour and illustrate whether changes in water content and bulk density produce stable and consistent responses in predicted thermal conductivity or whether local sensitivities vary irregularly across the dataset [35,36]. Narrow, ordered ICE trajectories indicate stable local behaviour compatible with continuous material-property interpretation, whereas dense crossing of ICE curves reflects strong local interactions and unstable response patterns that limit practical interpretability and controlled extrapolation in engineering applications [37].

Residual diagnostics complement response-surface and PDP/ICE analyses by characterising how prediction errors are distributed within the predictor domain. Spatial clustering of residuals and regime-dependent error patterns indicate systematic model limitations within specific hydro-mechanical states, whereas more uniform residual distributions suggest greater robustness of the learned relationship across the domain. From an engineering perspective, such patterns are particularly relevant because they indicate where model-derived thermal-conductivity fields may introduce local bias or uncertainty in subsequent simulations.

Taken together, global response-surface geometry, local diagnostic behaviour and residual structure provide a consistent and application-oriented framework for evaluating whether data-driven thermal-conductivity models are suitable for engineering parameterisation rather than solely for pointwise prediction. Although these diagnostics are partly qualitative, their combined interpretation provides a robust basis for assessing the functional stability and physical interpretability of surrogate models used in geothermal and geotechnical applications.

4.4. Residual Structure and Hydro-Mechanical Regimes

Across all evaluated models, increased residual variability is consistently observed at low water contents and low bulk densities. These conditions correspond to transitional hydro-mechanical states in which heat transfer is governed by discontinuous grain contacts and partially air-filled pore space, resulting in greater internal variability of effective thermal conductivity [4,6,8,9]. Under such conditions, small variations in saturation or packing structure may produce relatively large changes in thermal conductivity due to rapid changes in solid–water–air connectivity within the pore network.

Different modelling strategies respond to this intrinsic variability in distinct ways. Accuracy-optimised models tend to concentrate prediction errors within sparsely sampled transitional regimes, reflecting strong local adaptivity and sensitivity to neighbourhood structure. In contrast, globally regularised models distribute errors more uniformly across the hydro-mechanical domain, indicating reduced sensitivity to local fluctuations but increased overall systematic bias. These patterns are consistent with the residual distributions shown in Figure 10 and reflect fundamental differences in model structure and inductive bias.

From an engineering perspective, these differences are practically significant. Transitional low-saturation and low-density regimes commonly occur during site preparation, seasonal moisture fluctuations, and early operational phases of shallow geothermal systems. The concentration of prediction errors within such regimes may therefore introduce systematic uncertainty into thermal design calculations and long-term performance assess-

ments. In contrast, models exhibiting more uniform residual behaviour provide greater robustness for applications requiring spatial interpolation and stable parameter fields, even at the expense of slightly reduced pointwise accuracy.

Similar patterns have been reported in other geophysical and environmental machine-learning studies, where locally adaptive models amplify regime-dependent uncertainty, particularly in sparsely sampled or physically transitional domains [14,15,32]. The present results indicate that this mechanism is also relevant for data-driven thermal-conductivity parameterisation of sandy soils and should therefore be considered explicitly in engineering-oriented model selection.

4.5. Implications for Geothermal and Engineering Applications

The results emphasise that high predictive accuracy alone is insufficient when thermal conductivity is treated as a physically meaningful engineering material parameter. Models that reproduce laboratory measurements with minimal numerical error may nevertheless introduce non-physical oscillations, segmentation or discontinuities that compromise interpretability, spatial interpolation and controlled extrapolation in practical applications.

In engineering workflows, thermal conductivity is routinely propagated into numerical simulations of heat exchangers, ground-source heat pump systems, energy piles and other thermo-active underground structures. In such contexts, locally unstable surrogate relationships may translate into unrealistic thermal gradients, artefacts in temperature fields and increased uncertainty in performance predictions [38]. Consequently, model selection based exclusively on pointwise error metrics may lead to systematically biassed or unstable design inputs.

In light of the inductive-bias analysis presented above, Support Vector Regression—particularly the SVR_2 configuration—represents a balanced compromise for applications requiring stable and physically coherent parameterisation [33]. Although its prediction errors are higher than those of ensemble or instance-based models, its smooth and largely monotonic behaviour supports reliable interpolation and reconstruction of thermal-conductivity fields from incomplete geotechnical archives, as increasingly required in shallow geothermal assessments and regional subsurface modelling [17,39].

Accuracy-optimised ensemble and instance-based models remain suitable for applications focused on pointwise estimation or local anomaly detection, where spatial coherence and long-range extrapolation are of secondary importance. However, their use for large-scale parameterisation and long-term thermal modelling requires additional validation and post-processing to mitigate the effects of surface segmentation and local instability.

In practical terms, model selection should therefore be guided by the intended application. For pointwise prediction and local estimation tasks, high-capacity ensemble models (e.g., Random Forest or XGBoost) provide superior numerical accuracy. For applications involving interpolation, spatial mapping, or use of thermal conductivity as an input parameter in numerical simulations, models exhibiting smoother and more physically consistent response surfaces (e.g., SVR) provide more reliable and stable parameter fields. Classical physically based models may serve as a physically constrained reference, particularly when data availability is limited or when interpretability is prioritised over predictive accuracy.

In this context, classical physically based models (Table 3) remain valuable as conceptual references and lower bounds of physical inconsistency, even if their quantitative performance is insufficient for direct parameterisation of large archival datasets [6,7,9]. Rather than replacing physically based modelling, the proposed framework complements it by enabling transparent, application-specific selection of data-driven surrogates that balance numerical accuracy with functional coherence.

Overall, the presented results support a shift from accuracy-driven model selection toward integrated evaluation strategies that explicitly consider how surrogate models behave when reused as engineering material parameters in design, mapping and numerical simulation workflows.

5. Conclusions

This study demonstrates that modelling the thermal conductivity of moist, quartz-rich sands using data-driven approaches involves a fundamental and practically relevant trade-off between predictive accuracy and physical consistency. Models that achieve very low prediction errors according to conventional accuracy metrics do not necessarily provide physically coherent representations of the λ–water content–bulk density relationship when interpreted as continuous material-property fields. In the analysed dataset, accuracy-optimised models achieved RMSE values of approximately 0.12–0.15 W/m·K, whereas the most physically consistent SVR configurations exhibited RMSE values of approximately 0.30–0.33 W/m·K but substantially lower PCI values.

By integrating conventional accuracy metrics with diagnostics derived from the response-surface structure, local predictor effects and residual behaviour—summarised through the Physical Consistency Index (PCI)—the proposed framework enables transparent discrimination between numerically accurate models and those suitable for physically interpretable parameterisation. Even within a lithologically coherent domain, different model families encode fundamentally different functional representations of thermal conductivity. Sensitivity analyses confirm that the relative ranking of model families remains stable under alternative weighting schemes.

Tree-based and instance-based models prioritise predictive accuracy but tend to produce segmented or locally irregular response surfaces, limiting their suitability for controlled interpolation and spatial mapping. In contrast, kernel-based Support Vector Regression models impose stronger global regularisation and generate smoother and largely monotonic $\lambda(w, \rho_b)$ relationships, albeit at the cost of increased prediction error. Classical physically based models represent the opposite extreme of this spectrum, exhibiting strong physical coherence but limited quantitative performance when applied without site-specific recalibration.

The results indicate that no single modelling approach is universally optimal. Instead, model selection should be guided by the intended application. Models optimised for pointwise prediction may favour locally adaptive algorithms, whereas applications requiring interpolation, parameterisation or reuse of thermal conductivity as an engineering material property benefit from globally coherent functional representations.

The proposed framework is particularly suited for large geotechnical and geological archives in which water content and bulk density are available while laboratory measurements of thermal conductivity remain sparse. In such datasets, physically consistent data-driven models provide a transparent basis for reconstructing continuous thermal-property fields for shallow geothermal assessment, infrastructure design and subsurface heat-transfer modelling.

The present analysis is intentionally restricted to quartz-rich sands with limited mineralogical variability, where saturation state and packing density dominate thermal behaviour. Extending the framework to cohesive soils and mixed sediments, where mineralogy, bound water and fabric introduce additional controls, represents a natural direction for future research. The results are inherently dataset-specific and reflect the structure and variability of the analysed laboratory archive; therefore, direct transfer to other soil types or datasets should be approached with caution.

Overall, the results demonstrate that the evaluation of data-driven thermal-property models should not rely solely on predictive accuracy but should also explicitly account for the physical coherence of inferred relationships. Such considerations are essential when machine-learning models are reused as engineering material parameters in geothermal and geotechnical applications. Although demonstrated for quartz-rich sands, the proposed evaluation framework is transferable to other geotechnical and geological datasets where thermal properties must be reconstructed from incomplete laboratory observations.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/en19081827/s1, Figure S1: Pairplot of the air-dry sand subset ($w < 1\%$); Figure S2: Robustness of the accuracy–physical consistency trade-off to alternative formulations of the Physical Consistency Index (PCI); Table S1: Hyperparameter configurations and performance metrics for evaluated model variants; Table S2: Final hyperparameter configurations of representative models used for detailed physical interpretation.

Author Contributions: Conceptualization, M.Ż.; methodology, M.Ż.; software, M.Ż.; validation, M.Ż., E.J. and E.M.; formal analysis, M.Ż.; investigation, M.Ż. and E.J.; data curation, M.Ż. and E.J.; writing—original draft preparation, M.Ż.; writing—review and editing, M.Ż. and E.J.; visualisation, M.Ż. and E.J.; supervision, E.M. All authors have read and agreed to the published version of the manuscript.

Funding: The article processing charge was funded by the National Fund for Environmental Protection and Water Management.

Data Availability Statement: The laboratory data and computational workflows used in this study are part of institutional databases maintained by the Polish Geological Institute–National Research Institute. Due to legal and administrative restrictions, the underlying datasets are not publicly available. Access to the data may be granted for scientific purposes upon reasonable request and subject to applicable institutional regulations.

Acknowledgments: Generative artificial intelligence (AI) tools were used solely to assist in the development of computer code for data processing and machine-learning model implementation. All analyses, modelling decisions, data interpretation, and figure generation were performed and verified by the authors using their own data and scientific software.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DT	Decision Tree
EDA	Exploratory Data Analysis
GSHP	Ground-Source Heat Pump
ICE	Individual Conditional Expectation
KNN	k-Nearest Neighbours
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MSE	Mean Squared Error
PCI	Physical Consistency Index
PCA	Principal Component Analysis
PDP	Partial Dependence Plot
RF	Random Forest
RMSE	Root Mean Squared Error
SVR	Support Vector Regression
XGB	Extreme Gradient Boosting

References

1. Aresti, L.; Alvi, M.R.; Cecinato, F.; Fan, T.; Halaj, E.; Li, Z.; Loveridge, F.; McCartney, J.S.; Bourne-Webb, P.; Christodoulides, P. Energy geo-structures: A review of their integration with other sources and its limitations. *Renew. Energy* **2024**, *230*, 120835. [CrossRef]
2. López-Acosta, N.P.; Zaragoza-Cardiel, A.I.; Barba-Galdámez, D.F. Determination of Thermal Conductivity Properties of Coastal Soils for GSHPs and Energy Geostructure Applications in Mexico. *Energies* **2021**, *14*, 5479. [CrossRef]
3. de Vries, D.A. Thermal properties of soils. In *Physics of Plant Environment*; van Wijk, W.R., Ed.; North-Holland Publishing Company: Amsterdam, The Netherlands, 1963; pp. 210–235.
4. Farouki, O.T. *Thermal Properties of Soils*; Trans Tech Publications: Clausthal-Zellerfeld, Germany, 1986.
5. Wu, Y.H.; Wu, Y.F.; Fan, L.W.; Yu, Z.T.; Khodadadi, J.M. Thermal conductivity of soil: A review on the vast experimental data and predictive models. *Int. J. Therm. Sci.* **2025**, *208*, 109486. [CrossRef]
6. Kersten, M.S. *Thermal Properties of Soils*; University of Minnesota Engineering Experiment Station Bulletin: Minneapolis, MN, USA, 1949.
7. Johansen, O. Thermal Conductivity of Soils. Ph.D. Thesis, University of Trondheim, Trondheim, Norway, 1975.
8. Abu-Hamdeh, N.H. Thermal properties of soils as affected by density and water content. *Biosyst. Eng.* **2003**, *86*, 97–102. [CrossRef]
9. Lu, S.; Ren, T.; Gong, Y.; Horton, R. An improved model for predicting soil thermal conductivity from water content at room temperature. *Soil Sci. Soc. Am. J.* **2007**, *71*, 8–14. [CrossRef]
10. Zhao, T.; Liu, S.; Xu, J.; He, H.; Wang, D.; Horton, R.; Liu, G. Comparative analysis of seven machine learning algorithms and five empirical models to estimate soil thermal conductivity. *Agric. For. Meteorol.* **2022**, *323*, 109080. [CrossRef]
11. Adrinek, S.; Singh, R.M.; Janža, M.; Żeruń, M.; Ryżyński, G. Evaluation of thermal conductivity estimation models with laboratory-measured thermal conductivities of sediments. *Environ. Earth Sci.* **2022**, *81*, 380. [CrossRef]
12. Li, K.; Horton, R.; He, H. Application of machine learning algorithms to model soil thermal diffusivity. *Int. Commun. Heat Mass Transf.* **2023**, *149*, 107092. [CrossRef]
13. Wang, X.; Gao, Y.; Hou, J.; Yang, J.; Smits, K.; He, H. Machine learning facilitates connections between soil thermal conductivity, soil water content, and soil matric potential. *J. Hydrol.* **2024**, *633*, 130950. [CrossRef]
14. Karpatne, A.; Atluri, G.; Faghmous, J.H.; Steinbach, M.; Kumar, V. Theory-guided data science: A new paradigm for scientific discovery. *Science* **2017**, *357*, eaam9780. [CrossRef]
15. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat, F. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef]
16. Roscher, R.; Bohn, B.; Duarte, M.F.; Garcke, J. Explainable machine learning for scientific insights and discoveries. *IEEE Access* **2020**, *8*, 42200–42216. [CrossRef]
17. Łukawska, A.; Ryżyński, G.; Żeruń, M. Serial laboratory effective thermal conductivity measurements of cohesive and non-cohesive soils for the purpose of shallow geothermal potential mapping and databases—Methodology and testing procedure recommendations. *Energies* **2020**, *13*, 914. [CrossRef]
18. *PN-EN ISO 17892-2:2015-02*; Geotechnical Investigation and Testing—Laboratory Testing of Soil—Part 2: Determination of Bulk Density. Polish Committee for Standardization: Warszawa, Poland, 2015.
19. *PN-EN ISO 17892-1:2015-02*; Geotechnical Investigation and Testing—Laboratory Testing of Soil—Part 1: Determination of Water Content. Polish Committee for Standardization: Warszawa, Poland, 2015.
20. *ASTM D5334-22*; Standard Test Method for Determination of Thermal Conductivity of Soil and Soft Rock by Thermal Needle Probe Procedure. ASTM International: West Conshohocken, PA, USA, 2022.
21. *IEEE 442-2017*; Guide for Thermal Resistivity Measurements of Soils and Backfill Materials. IEEE Standards Association: Piscataway, NJ, USA, 2017.
22. *PN-EN ISO 14688-1:2018-05*; Geotechnical Investigation and Testing—Identification and Classification of Soil—Part 1: Identification and Description. Polish Committee for Standardization: Warszawa, Poland, 2018.
23. Belsley, D.A.; Kuh, E.; Welsch, R.E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; Wiley: New York, NY, USA, 1980.
24. Jolliffe, I.T. *Principal Component Analysis*, 2nd ed.; Springer: New York, NY, USA, 2002.
25. MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, USA, 1967; pp. 281–297.
26. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubour, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
27. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
28. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.

29. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Wadsworth International Group: Belmont, CA, USA, 1984.
30. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
31. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016. [CrossRef]
32. Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; Kumar, V. Integrating physics-based modeling with machine learning: A survey. *arXiv* **2003**, arXiv:2003.04919.
33. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
34. Tarnawski, V.R.; Wagner, B.; Leong, W.H.; McCombie, M.; Coppa, P.; Bovesecchi, G. Soil thermal conductivity model by de Vries: Re-examination and validation analysis. *Eur. J. Soil Sci.* **2021**, *72*, 1940–1953. [CrossRef]
35. Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.* **2015**, *24*, 44–65. [CrossRef]
36. Molnar, C. *Interpretable Machine Learning*, 2nd ed.; Lulu: Durham, NC, USA, 2022. Available online: https://christophm.github.io/interpretable-ml-book (accessed on 14 January 2026).
37. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
38. Salhein, K.; Kobus, C.J.; Zohdy, M.; Annekaa, A.M.; Alhawsawi, E.Y.; Salheen, S.A. Heat Transfer Performance Factors in a Vertical Ground Heat Exchanger for a Geothermal Heat Pump System. *Energies* **2024**, *17*, 5003. [CrossRef]
39. Żeruń, M.; Jagoda, E.; Majer, E.; Majer, K. Very shallow geothermal potential in urban areas: Parameterization algorithm and mapping tool for thermal conductivity and thermal resistivity in Warsaw (Poland). *Geol. Q.* **2025**, *69*, 55. [CrossRef]