Jurnal Polimesin
Department of Mechanical Engineering
State Polytechnic of Lhokseumawe
http://e-jurnal.pnl.ac.id/polimesin
e-ISSN : 2549-1999 No. : 3 Month : June
p-ISSN : 1693-5462 Volume : 24 Year : 2026

Processing dates: received on 2026-4-24, reviewed on 2026-05-18, accepted on 2026-05-31 and online availability on 2026-06-30

Remaining useful life prediction of railway wheelsets using a composite wear index under sparse monitoring data

Agustinus Winarno[1,*], Ahmad Fauzan Karnadi[1], Herjuno Rizki Priatomo[1], Slamet Afif Mansuri[2], Rioko Aji[2], Sudianto[2], Miming Kuncoro[2]

[1]Department of Mechanical Engineering, Gadjah Mada University, Sleman 55281, Indonesia

[2]PT. Kereta Api Indonesia (Persero), Kota Bandung 40117, Indonesia

*Corresponding author: winarno_agustinus@ugm.ac.id

Abstract

Accurate prediction of the Remaining Useful Life (RUL) of railway wheelsets is important for operational safety and efficient maintenance planning. This study proposes a physics-informed composite wear index integrating wheel diameter and flange wear and validates it against field operational data. The composite index, $W = f(D) \circ f(L)$, combines diameter wear and flange wear through a reprofiling coefficient k, representing diameter reduction per unit flange restoration. Using machining records from a 60-wagon freight train operated by PT Kereta Api Indonesia (1,791 monthly records over 3 years), the field-based median k was estimated at 2.75 mm/mm for the train set and 3.00 mm/mm fleet-wide. The selected modelling value of k = 3.2 mm/mm lies near the upper range of field observations and provides a conservative approximation. Applied to the operational dataset, the index successfully tracked coupled wear progression, showing that 62% of wagons had exceeded the midlife threshold (W ≥ 0.50). A deterministic benchmark using 201 observations across five reprofiling cycles was used to compare five machine-learning models under 30–60% monitoring densities. Linear regression achieved the lowest error ($R^2 \approx 1.000$), while gradient boosting showed the most reliable non-linear performance. The results support the proposed composite wear index as a practical basis for wheelset RUL estimation under limited monitoring data.

Keywords:
Railway wheel wear, remaining useful life, composite wear model, flange wear, predictive maintenance.

1 Introduction

Railway wheel maintenance is a safety-critical and cost-intensive activity in rail operations. Wheelset degradation occurs through two interacting mechanisms: tread diameter reduction caused by rolling-contact fatigue and operational wear, and flange-profile deterioration driven by lateral wheel–rail interaction forces [11], [18]. The interaction between these two mechanisms is often overlooked in conventional maintenance scheduling and in data-driven predictive models, even though both processes evolve simultaneously and influence each other [16].

In current railway industry practice, wheel condition is monitored through periodic measurements. In Indonesia, monthly inspection cycles are the standard interval for rolling stock operated by PT Kereta Api Indonesia, including freight wagons, mass-rapid-transit vehicles, and commuter trains. These measurements are used to track dimensional changes and to determine reprofiling schedules. However, such monitoring activities require significant resources, and in practice many operators collect fewer observations than theoretically required. This raises an important question: what is the minimum data density required for reliable RUL prediction?

The problem is further complicated by the multi-cycle nature of wheel life. A wheel typically undergoes several reprofiling cycles before reaching its discard diameter. Each reprofiling restores the flange profile but permanently reduces the wheel diameter. Accumulated flange wear therefore determines the amount of diameter material that must be removed during each subsequent maintenance event, a reprofiling coupling that is a physical characteristic of the wheel–rail tribological system and that directly influences the total number of achievable reprofiling cycles within the allowable diameter range.

Previous research on the RUL prediction of railway wheels generally follows two main approaches. Physics-based models, including those derived from the Archard wear model and multi-body simulation, provide detailed representation of wear mechanisms but require extensive computational effort and detailed input parameters [1], [2]. In contrast, data-driven approaches such as neural networks, support vector regression, and ensemble learning offer flexibility and efficiency but typically treat flange wear and diameter reduction as independent variables without incorporating their physical relationship [3], [4], [13]. As a result, existing models may fail to capture the cumulative impact of flange wear on long-term diameter loss.

The research gap addressed by this paper is the absence of a composite wear-state formulation that explicitly encodes both the direct diameter wear and the deferred diameter cost committed by flange accumulation, together with its evaluation across multiple machine-learning algorithms and sparse-data densities. The main contributions of this study are: (1) A composite wear index $W = f(D) \circ f(L)$ derived from the coupling ratio k = 3.2 mm of diameter reduction per mm of flange restoration, cited from the empirical range 2.95–3.31 mm/mm reported by Karnadi [7] and physically grounded in the wheelset reprofiling process; (2) Field validation of the coupling coefficient using the real reprofiling log of a 60-wagon PT Kereta Api Indonesia freight train set, showing that the modelling value k = 3.2 (midpoint of the Karnadi [7] range 2.95 to 3.31 mm/mm) sits at the upper edge of the observed field distribution, and that the composite index tracks coupled wear across a real three-year fleet record; (3) A systematic comparison of five machine-learning algorithms (LR, RF, GB, SVR-RBF, KNN) under multiple sparse-data densities, trained against a 201-observation simulated ground-truth dataset; (4) Identification of the minimum data-density threshold for reliable RUL prediction using physics-informed composite features; (5) Practical maintenance-scheduling guidance based on the composite-wear-index thresholds corresponding to reprofiling-cycle boundaries.

The research question is formulated as: how can a physics-informed representation of coupled wear improve the accuracy and reliability of RUL prediction under varying data availability? The objective of this study is to develop a composite wear index that captures the coupling mechanism between wear components and to evaluate its effectiveness in supporting accurate prediction and practical maintenance decision-making, under simulated deterministic conditions that serve as a controlled initial validation prior to field deployment.

2 Literature review

2.1 Physics based wheel wear prediction

The prediction of railway wheel-profile wear through physics-based simulation has a well-established tradition spanning several decades. The foundational methodology implemented by Jendel [1] at KTH uses the GENSYS multi-body simulation software and a load-collective approach in which measured track data, validated rail profiles, and operating conditions drive time-domain dynamic simulations [17]. Wheel–rail contact is modelled using Hertzian theory and Kalker's FASTSIM algorithm, while wear is computed

using the Archard model calibrated through tribology laboratory tests. Validation against 200 000 km of consecutive wheel-profile measurements on the Stockholm commuter network showed very good agreement across four scalar wear measures: flange thickness, flange height, flange inclination, and area worn off [1], [15]. This work forms the direct methodological foundation of the simulation framework used in the present study.

The broader simulation workflow has been widely discussed in the literature. The standard iterative procedure runs dynamic simulations in incremental wear steps, updating the worn profile in each step until a target mileage or profile limit is reached. Among the available wear laws, the Archard model, which relates wear volume to the product of normal force and sliding distance divided by material hardness, and the USFD wear model, which expresses wear rate as a function of $\frac{T\gamma}{A}$, where A is the contact area and $T\gamma$ is the product of tangential force and creepage, are the most widely applied in railway studies [2], [12], [14]. Song *et al.* [3] demonstrated that the Archard model, when calibrated using real measurement data, can reproduce high-speed-train tread-wear profiles within 5% relative error after 80 000 km under braking conditions. However, these physics-based methods do not produce a unified scalar wear state suitable for direct ML regression, a gap addressed by the composite index proposed here.

A significant contribution that directly motivates the composite model proposed in this study is the physics-informed, data-driven framework by Zeng, *et al*. [4]. Using high-speed trains, the authors showed that tread wear and flange wear must be modelled jointly because their rates are interdependent: the current diameter influences flange contact conditions, while flange thickness affects lateral contact position. Their reprofiling model demonstrated that the cutting amount at each reprofiling event is governed by flange thickness prior to maintenance, and the relationship between diameter reduction and flange restoration was represented using Markov state-transition matrices.

Their Monte Carlo–based RUL simulation produced multimodal distributions with discrete peaks corresponding to reprofiling-cycle boundaries, which explains the periodic behavior observed in the composite wear index developed in this study [4]. Braghin, *et al*. [5] further showed that the ratio between diameter wear and flange wear is not constant but varies depending on dynamic contact conditions. In curved track sections, flange contact forces become dominant and increase flange wear relative to tread wear. This variable relationship supports the need to treat the coupling coefficient between diameter and flange wear as an empirically determined parameter rather than a fixed universal constant.

2.2 Data-driven and machine-learning approaches for wheelset RUL

Data-driven approaches for RUL prediction of railway wheelsets have emerged as a complementary direction to physics-based simulation. Wang *et al.* [6] proposed a multitask-learning framework that simultaneously predicts RUL as a regression problem and failure type as a classification problem for freight wheelsets using wayside-detector data. Their formulation combines least-squares loss and logistic-regression likelihood with a regularization term that enforces shared feature selection across tasks. This approach improves prediction accuracy and reduces overfitting to periodic maintenance patterns, which is relevant to the cyclic wear behavior observed in wheel data [6].

A relevant study in the Indonesian railway context is presented by Karnadi [7], who analyzed the service life of coal-freight-wagon wheels operated by PT Kereta Api Indonesia. Using historical measurements of diameter and flange wear combined with reprofiling records, the study applied a composite-function approach in which flange-wear data were transformed into a cumulative trend and modelled using linear regression. The empirically derived ratio of diameter reduction per unit flange restoration ranged between 2.95 and 3.31 mm per mm, with coefficient-of-determination values between 0.96 and 0.98 and a predicted average service life of 10.6 years. The finding that flange-wear rate dominates diameter-wear rate is consistent with the role of lateral forces as the primary degradation mechanism [7]. The present study adopts the midpoint of this empirical range, k = 3.2 mm/mm, as a practical fixed value for the composite formulation; sensitivity to variation within the empirical range is discussed in §5.4. The present study goes further by re-estimating the coupling coefficient directly from the real PT Kereta Api Indonesia reprofiling log analysed here (§3.8, §4.6), thereby validating the adopted value rather than only citing it.

The broader application of machine learning for industrial RUL prediction has been reviewed by Zonta, *et al.* [8], who identified ensemble-based methods as particularly robust for irregular time-series data. Techniques such as random forest and gradient boosting consistently outperform single-model approaches when data are sparse or unevenly sampled, which reflects the characteristics of the maintenance datasets used in this study. However, existing ML studies have generally not encoded the diameter–flange coupling as a single composite feature, the methodological extension proposed by the present work.

2.3 Sparse data and the role of physics-informed features

A key challenge in railway-wheel RUL prediction is the limited availability of monitoring data. In practice, measurements are collected at discrete intervals due to operational constraints, resulting in incomplete representation of wear progression. The impact of such sparsity on prediction accuracy has not been fully addressed, particularly for models that incorporate multiple wear mechanisms.

Jendel [1] showed that simulation-based predictions can still achieve good agreement with field measurements even when input data are limited, suggesting that physically grounded models can compensate for data sparsity. Zeng, *et al*. [4] also reported that a physics-informed approach requires relatively few data points to achieve convergence because wear behavior is constrained by underlying physical laws. These findings support the hypothesis that incorporating physics-based features can improve prediction reliability under limited-data conditions.

The broader literature on physics-informed machine learning has consolidated this insight: Karpatne, *et al*. [19] and Willard, *et al*. [20] demonstrate that physics-informed feature engineering, augmenting purely empirical models with variables derived from domain knowledge, improves data efficiency, generalization, and interpretability for engineering prognostics. The prognostics' literature represented by Saxena, *et al*. C-MAPSS benchmark [21] further confirms that compact engineered features can match or exceed the performance of high-capacity deep models when the underlying degradation physics is well understood. Emzain, *et al*. [9] complement this from a reliability-engineering perspective: their FMEA-based analysis of centrifugal-pump failure modes shows that interactions between failure mechanisms are often underestimated when analyzed independently, and that accounting for these interactions yields more accurate maintenance-interval estimation, a principle that directly motivates the coupled (rather than independent) treatment of diameter and flange wear adopted here. Similarly, Ruspendi, *et al*. [10] showed that integrated performance metrics provide a more accurate representation of system condition compared with single-variable monitoring, reinforcing the importance of composite-modelling approaches. The impact of sparse sampling on coupled-wear models across multiple reprofiling cycles, however, has not been systematically reported, the empirical gap addressed by the present study.

2.4 Research gap

A synthesis of the reviewed literature reveals three converging gaps: (1) No existing RUL-prediction framework explicitly represents the coupling between flange wear and diameter reduction as a unified composite variable for machine-learning applications [1], [4]; (2) Empirical studies have identified variation

in the coupling ratio between diameter reduction and flange restoration, but its effectiveness as a fixed modelling parameter across multiple predictive algorithms has not been evaluated [5], [7]; (3) The effect of limited monitoring data on prediction accuracy has not been systematically analyzed for models that incorporate coupled wear behavior across multiple reprofiling cycles [8].

This study addresses these gaps by constructing a composite wear index that integrates diameter and flange wear through an empirically derived coupling coefficient, evaluating multiple machine-learning algorithms using this representation, and assessing prediction performance under varying data-availability conditions in a simulated deterministic setting that serves as a controlled initial validation. The coupling coefficient is additionally validated against the real PT Kereta Api Indonesia field log (§4.6).

3 Method

This study draws on two complementary datasets. The primary empirical dataset is the real wheel-measurement and reprofiling log of a 60-wagon freight train set operated by PT Kereta Api Indonesia, used to estimate the coupling coefficient k and to verify that the composite index tracks real degradation (§3.8, §4.6). A deterministic simulation, parameterised by the same physical constants, serves as a controlled, noise-free benchmark to compare the five machine-learning algorithms in isolation from field noise (§3.1 to §3.7, §4.1 to §4.5), an analytical companion to the field data rather than the sole basis of the study.

3.1 Vehicle and operating parameters

The dataset used in this study was generated from a simulation of wheelset wear for a 50-ton freight wagon. The simulation parameters represent typical Indonesian freight-rail operating conditions, including standard wheel-material properties and rail geometry. All simulation parameters are summarized in Table 1. The initial wheel diameter is set to 850 mm with a retirement limit of 750 mm, resulting in a total diameter budget of 100 mm. Two flange thresholds are used: the reprofiling trigger at −7.8 mm (operational maintenance trigger) and the absolute safety limit at −8.0 mm (normalization reference; see §3.4). Each reprofiling event restores the flange profile while reducing the wheel diameter according to the deterministic coupling rule described in §3.4.

Table 1. Operational wear parameters used in the simulation

Parameter	Value	Unit
Initial diameter D_0	850	mm
Retirement limit D_{min}	750	mm
D_{budget}	100	mm
Flange reprofiling trigger	−7.8	mm
Flange absolute safety limit	−8.0	mm
Diameter wear rate (operational)	0.020	mm/month
Flange wear rate	0.200	mm/month
Duration per reprofiling cycle	1170	days (≈3.2 yr)
Reprofiling coefficient k	3.2	mm ΔD / mm Δf
Diameter removed per reprofiling	24.96	mm
Number of reprofiling cycles	5	-

The operational wear rates used in the simulation are 0.020 mm/month for diameter reduction and 0.2 mm/month for flange wear. These rates are engineering estimates consistent with the regime reported for Indonesian freight-wagon operations [7] and are treated as representative deterministic values; the relative model comparison reported in §4 is independent of the absolute rate values, while the absolute prediction errors scale linearly with them. Based on these rates, the duration of each reprofiling cycle is approximately 1170 days, or about 3.2 years. The coupling coefficient k = 3.2 mm of diameter reduction per mm of flange restoration is adopted as the midpoint of the empirical range 2.95–3.31 mm/mm reported by Karnadi [7]. This leads to an average diameter reduction of 24.96 mm per reprofiling event and allows a total of five reprofiling cycles before the retirement limit is reached.

3.2 Dataset generation procedure

The simulated dataset is generated under a deterministic model parameterized by the values in Table 1. Monthly observations are produced from January 2022 to June 2038, resulting in 201 records over approximately 16.4 years and capturing five complete reprofiling cycles followed by final wheel retirement. The dataset is produced by the open Jupyter notebook the simulation script, ensuring reproducibility.

For each monthly time step: (1) The diameter value of each of the eight wheel positions decreases by 0.020 mm; (2) The flange wear value of each of the eight positions decreases by 0.2 mm (i.e. its magnitude increases by 0.2 mm); (3) The reprofiling trigger is evaluated: when the average flange wear reaches −7.8 mm, a reprofiling event is executed in the next time step; (4) At a reprofiling event, the magnitude of the average flange wear at that moment is multiplied by k = 3.2 to determine the diameter that is removed from the running diameter, and the flange wear is reset to its nominal post-machining condition (≈0 mm). This idealized reset represents complete profile restoration; residual wear effects in real reprofiling operations are acknowledged as a limitation in §5.4.

Reprofiling is modelled as deterministic and identical across all five cycles, with the same material removal per event (24.96 mm of diameter) and the same complete restoration of the flange profile. Cycle-to-cycle reprofiling variability that would arise in real operations from variations in machining quality and operator practice is not represented in the simulation; it is acknowledged as a limitation in §5.4 (item c).

Because the model is deterministic, the simulation is run a single time and produces a unique ground-truth trajectory. Stochastic extensions (e.g. Monte Carlo realizations with rate variability) are flagged as future work in §5.4.

3.3 Measurement columns and feature engineering

Each monthly observation contains diameter measurements for eight wheels: four positions in bogie one (BG 1 Ø1 to Ø4) and four positions in bogie two (BG 2 Ø1 to Ø4). The corresponding flange-wear measurements are recorded for the same eight positions (BG 1 F1 to F4; BG 2 F1 to F4). Under the assumption of uniform wear distribution across all wheel positions within each bogie, averaged values represent the overall wheelset condition: the average diameter is the mean of the eight diameter measurements, and the average flange wear is the mean of the eight flange measurements.

Within each reprofiling cycle, the average diameter decreases monotonically because of continuous material removal during operation. In contrast, flange wear follows a cyclic sawtooth pattern, starting at zero, increasing in magnitude until the reprofiling trigger of −7.8 mm is reached, and resetting to zero at each reprofiling event. The combined evolution is illustrated in Fig. 1. Fig. 1(a). Wheel diameter degradation, with reference levels marked at the four reprofiling targets (825, 800, 775 mm) and the retirement limit (750 mm) and Fig. 1(b). Flange wear, with the reprofiling trigger at −7.8 mm and the absolute safety limit at −8.0 mm. Quantitatively, the wear progression shown in Fig. 1 follows the deterministic schedule of Table 1. Each reprofiling cycle spans approximately 1170 days (≈3.2 years, or about 39 months), during which the average diameter decreases monotonically by ≈0.020 mm per month and the average flange wear accumulates by ≈0.200 mm per month. The flange-wear trace reaches the −7.8 mm reprofiling trigger after approximately 39 months and then resets to zero at the reprofiling event, while the diameter decreases by an additional 24.96 mm at that event ($= k \times 7.8$). The repeated sawtooth pattern thus reflects five complete operational cycles followed by the final retirement event at 750 mm.

A total of nine features were constructed to represent the temporal progression, physical wear state, and lifecycle position of the wheelset, as summarized in Table 2. These features combine direct measurements with derived variables that encode the physical relationship between diameter wear and flange wear.

Fig. 1. Wheel diameter degradation (a) and flange wear per cycle (b) across the full 16.4-year life.

Table 2. Features used in the RUL-prediction models

Feature	Description	Physical meaning
$days_{elapsed}$	Days since first observation	Global time index
$avg_{diameter}$	Mean of 8-wheel diameter (mm)	Current wheel size
avg_{flange}	Mean of 8-wheel flange wear (mm)	Current flange state
$f(D)$	$\frac{(850 - avg_{diameter})}{100}$	Normalized diameter wear ∈ [0,1]
$f(L)_{raw}$	$\frac{\lvert avg_{flange} \rvert}{8.0}$	Normalized flange wear per cycle ∈ [0,1]
$D_{pending\ reprofilin}$	$\lvert avg_{flange} \rvert \times 3.2\ mm$	Pending diameter cost of current flange
$D_{committed}$	$(850 - avg_{diameter}) + D_{pending\ reprofiling}$	Total diameter budget committed
$W_{composite}$	$\frac{D_{committed}}{100}$	Composite wear index ∈ [0,1]
$n_{reprofiling}$	Count of reprofiling events so far	Life-cycle stage identifier

The composite feature $W_{composite}$ integrates past diameter loss and future diameter reduction associated with the current flange condition, providing a more complete representation of the wear state than individual variables. The discrete feature $n_{reprofiling}$ is critical for distinguishing observations that share similar wear magnitudes but belong to different reprofiling cycles. Without this feature, distance-based learners such as KNN and SVR cannot differentiate identical composite states corresponding to different RUL values, a limitation explored in §4.4.

3.4 Composite wear model

The central contribution of this study is the formulation of a composite wear-state index W, integrating the diameter-degradation function $f(D)$ and the flange-wear function $f(L)$ through the physically defined reprofiling coefficient k. The composite index is defined in Eq. (1), where $D_{committed} = (D_{NEW} - D_{current}) + |f_{current}| \cdot k$, D_{NEW} is 850 mm (new wheel diameter), $D_{current}$ is current average diameter (mm), $f_{current}$ is current average flange wear (mm, negative), $k = 3.2 \frac{mm\ \Delta D}{mm\ \Delta f}$ (reprofiling ratio, from [7]), and D_{budget} is 100 mm $(D_{NEW} - D_{limit})$.

$$W = \frac{D_{committed}}{D_{budget}} \quad (1)$$

The term $D_{committed}$ has two physically meaningful components. The first, $(D_{NEW} - D_{current})$, is the material that has already been removed by operational wear and previous reprofiling events. The second, $|f_{current}| \cdot k$, is the deferred diameter reduction that will occur during the next reprofiling event, the forward-looking term that captures the future maintenance cost embedded in the current flange condition. This deferred-cost component is a key distinguishing feature of the proposed model.

To ensure comparability and interpretability, the diameter and flange contributions are normalized individually as in Eq. (2) and Eq. (3), where $|f_{limit}| = 8$ mm is the absolute safety limit (Table 1). Two thresholds are used because they serve distinct physical roles:

−7.8 mm is the operational reprofiling trigger, while 8.0 mm is the safety bound used as the normalization denominator so that $f(L)$ saturates only at the hard safety boundary, not at the routine maintenance trigger.

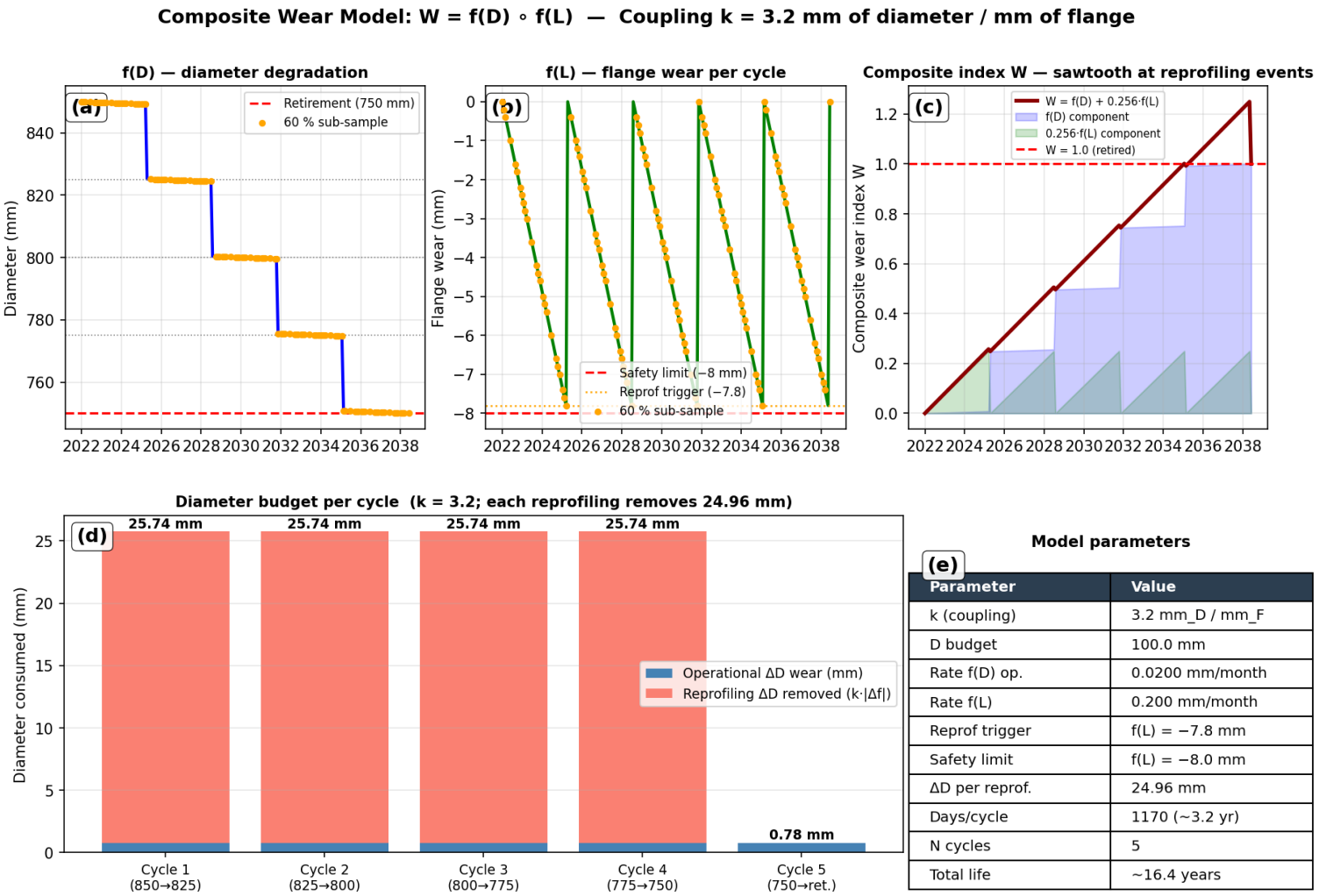
$$f(D) = \frac{D_{NEW} - D_{current}}{D_{budget}} \quad (2)$$

$$f(L) = \frac{|f_{current}|}{|f_{limit}|} \quad (3)$$

The relationship between Eq. (1), Eq. (2), and Eq. (3) can be expressed as Eq. (4). With the multiplier $\frac{k \cdot |f_{limit}|}{D_{budget}} = 3.2 \times 8.0 / 100 = 0.256$ in Eq. (4) makes explicit that W is a weighted sum of the normalized diameter wear and the normalized flange wear, with the flange-weight (0.256) representing the deferred-cost coupling.

$$W = f(D) + \frac{k \cdot |f_{limit}|}{D_{budget}} \cdot f(L) = f(D) + 0.256 \cdot f(L) \quad (4)$$

The composite index W evolves dynamically over time. Within each reprofiling cycle, W increases as the diameter gradually decreases and flange wear accumulates. Immediately before reprofiling, W reaches its local maximum because both components contribute fully. After reprofiling, W partially resets: the flange-wear term returns to zero while the accumulated diameter loss remains. This produces the cyclic sawtooth pattern that reflects the multi-cycle degradation behavior of the wheelset, illustrated in Fig. 2.



Parameter	Value
k (coupling)	3.2 mm_D / mm_F
D budget	100.0 mm
Rate f(D) op.	0.0200 mm/month
Rate f(L)	0.200 mm/month
Reprof trigger	f(L) = −7.8 mm
Safety limit	f(L) = −8.0 mm
ΔD per reprof.	24.96 mm
Days/cycle	1170 (~3.2 yr)
N cycles	5
Total life	~16.4 years

Fig. 2. Composite wear-model visualization.

3.5 Sparse sampling strategy

To evaluate model performance under realistic conditions of limited monitoring data, four reduced datasets were derived from the full 201-observation baseline. The reduced datasets contain 30% (66 records), 40% (82 records), 50% (99 records), and 60% (119 records) of the full data and are stored as Excel files alongside the source code.

The reduced datasets are deterministic subsamples of the full reference rather than drawings from a randomized sampling procedure. Each subsamples preserves chronological order and is constructed to maintain approximately equal representation across the five reprofiling cycles, as documented in Table 3. The retirement event (the 201st observation) is included in the 50% and 60% subsamples and omitted from the 30% and 40% subsamples. The observed gap distribution between consecutive observations in each subsamples is a multiple of 30 days (30, 60, 90, 120, 150, or 240 days), which approximates the practical pattern of an operator missing one or more scheduled monthly inspections.

Because the subsamples are deterministic single realizations rather than random draws, the metrics reported in Table 5 (§4.1) reflect a single subsampling configuration per data density and do not characterize the variability that would arise under different random subsampling. A multi-seed randomized evaluation reporting mean ± standard-deviation across draws is identified as priority future work in §5.4 (item d).

The complete 201-observation dataset is the reference benchmark for ground-truth behavior. All models are trained only on the reduced datasets, and the reported metrics are computed on the held-out observations, the subset of the 201-record reference that does not appear in the training subsamples. This evaluation ensures that predictions are assessed on unobserved intermediate states across the lifecycle.

Table 3. Per-reprofiling-cycle observation count in each subsample (out of 40 records per cycle in the full reference).

Sub-sample	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Retirement	Total
Full reference	40	40	40	40	40	1	201
60%	25	23	23	24	23	1	119
50%	20	19	20	20	19	1	99
40%	16	16	17	17	16	0	82
30%	13	12	14	14	13	0	66

3.6 Algorithm descriptions

Five regression algorithms were evaluated, spanning linear, ensemble, kernel-based, and distance-based learning paradigms: (1) Linear Regression (LR), ordinary least-squares applied to the nine features defined in Table 2. Selected to evaluate the extent to which the inherent linearity of the composite-wear representation can be captured without additional model complexity. No hyperparameters required; (2) Random Forest (RF), an ensemble of 200 decision trees with bootstrap sampling and mean-squared-error splitting criterion. Improves prediction stability by averaging multiple trees and is robust to noise; captures non-linear feature interactions without explicit specification; (3) Gradient Boosting (GB), a sequential ensemble of 200 shallow decision trees, where each tree corrects residual errors of previous iterations. A learning rate of 0.1 controls the contribution of each tree, reducing the risk of overfitting under limited data; (4) Support Vector Regression (SVR), Radial-Basis-Function (RBF) kernel with regularization parameter $C = 500$ and kernel parameter γ using the scale formulation based on feature variance. All features are standardized using z-score normalization prior to training; (5) K-Nearest Neighbor (KNN), predicts RUL as the average value of the five nearest observations in the standardized feature space. Particularly sensitive to the cyclic behavior of the composite wear index, as observations with similar feature values may belong to different reprofiling cycles.

3.7 Experimental framework and evaluation protocol

The complete dataset of 201 observations is treated as the ground-truth benchmark and is not used during training. The four reduced training sets (30%, 40%, 50%, 60%) defined in §3.5 are used for fitting; for each training configuration, performance metrics are computed on the held-out observations, the subset of the 201-record reference is not present in the training subsample. This setup captures both interpolation capability and full-horizon prediction performance, both of which are critical for RUL estimation.

Model performance is quantified using four standard regression metrics: Mean Absolute Error in days (MAE), Root-Mean-Square Error in days (RMSE), coefficient of determination R^2, and Mean Absolute Percentage Error in percent (MAPE). All models are implemented using scikit-learn 1.4 in Python 3.11. The dependence of these metrics on training data size is summarized in Fig. 3.

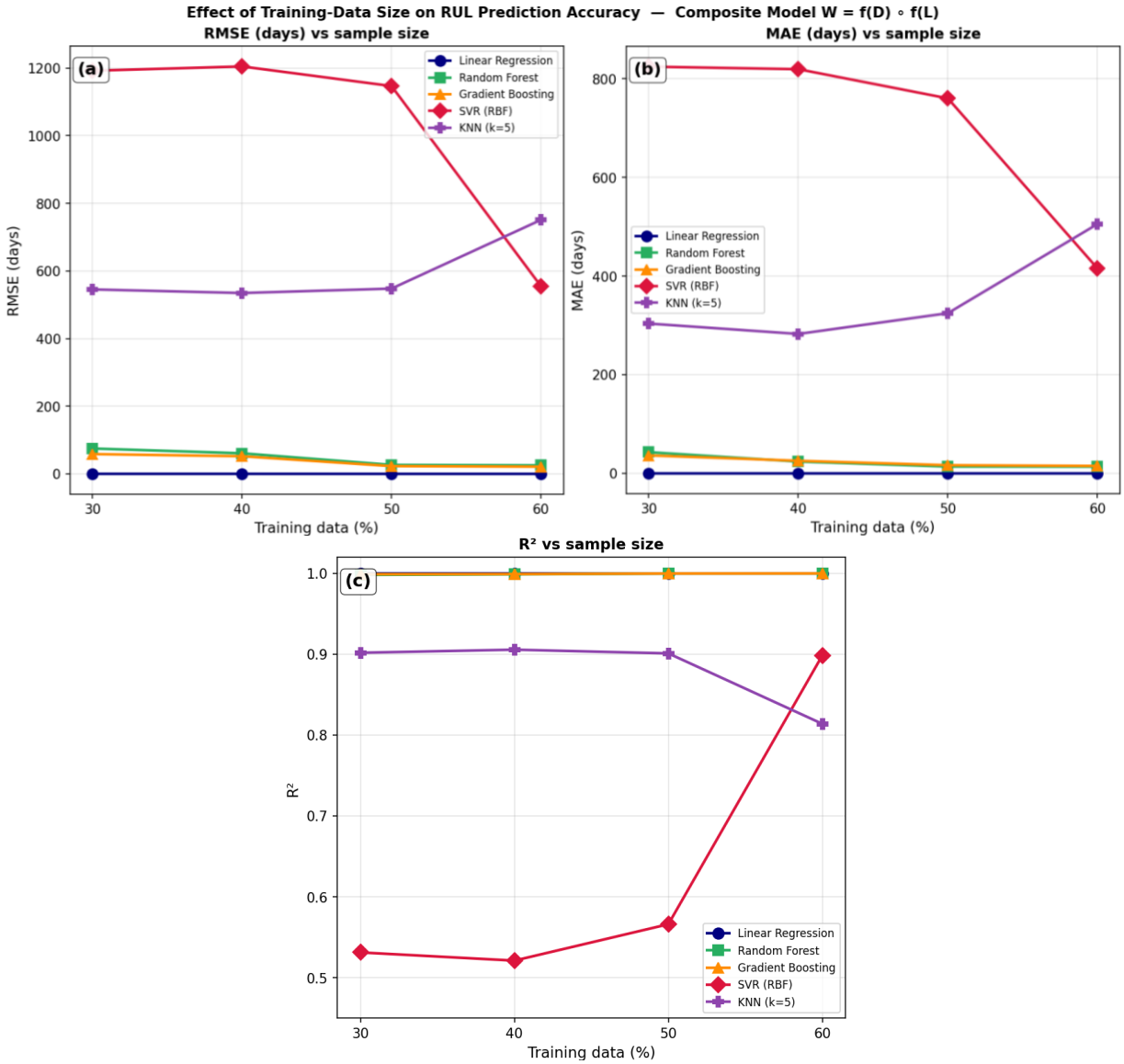


Fig. 3. Effect of training-data size on RUL-prediction accuracy: (a) RMSE; (b) MAE; (c) R^2. Each curve corresponds to one of the five evaluated algorithms (LR, RF, GB, SVR, KNN).

3.8 Field dataset and empirical estimation of the coupling coefficient

To validate the composite model against real behaviour, this study uses the actual maintenance dataset of a single freight train set, denoted TS1, operated by PT Kereta Api Indonesia: 60 wagons, eight measured wheel positions each (four per bogie), monitored monthly over 3 years. After removing one physically implausible record (average diameter outside 745 to 852 mm or average flange wear outside 0 to 8.0 mm), 1,791 monthly wheel-state records remain. The wagons span three batches of the same 50-ton KKBW design. Summary characteristics are given in Table 4.

Table 4. Characteristics of the real field dataset (train set TS1, PT Kereta Api Indonesia).

Wagons	60
Wheel positions per wagon	8 (4 per bogie)
Monthly records (after cleaning)	1,791
Observation window	30 months
Avg diameter (median; range)	829.5 mm (766.8 – 850.0)
Avg flange wear (median)	2.69 mm
Composite index W (median, all records)	0.29

The field log records two derived quantities at each reprofiling (*pembubutan*) event: the diameter removed (*selisih diameter*) and the flange restored (*selisih flens*) at each wheel position. The empirical coupling coefficient is computed per position as k_obs = ΔD / Δf, keeping only positions with a positive diameter cut and positive flange restoration and trimming equalising over-cuts (k > 10) as machining artefacts. Three views are reported: the pooled ratio Σ(ΔD)/Σ(Δf), the per-position median, and the per-event ratio. The estimator is applied to flange-driven events of TS1 (primary scope) and the whole KKBW fleet (broader reference), and compared against k = 3.2 and the Karnadi [7] range in §4.6.

4 Results

4.1 Overall performance comparison

Table 5 presents the complete performance comparison across all regression models and training-data proportions, computed on the held-out portion of the full reference (see §3.5 and §3.7).

Table 5. RMSE (days) and R² across all model and sample-size combinations (single deterministic subsample per density).

Model	30%	40%	50%	60%
Linear regression	RMSE = 0.1, R² = 1.000	RMSE = 0.1, R² = 1.000	RMSE = 0.0, R² = 1.000	RMSE = 0.0, R² = 1.000
Gradient boosting	RMSE = 58.5, R² = 0.9989	RMSE = 52.1, R² = 0.9991	RMSE = 22.9, R² = 0.9998	RMSE = 21.2, R² = 0.9999
Random forest	RMSE = 75.0, R² = 0.9981	RMSE = 60.8, R² = 0.9988	RMSE = 26.5, R² = 0.9998	RMSE = 25.2, R² = 0.9998
KNN ($k = 5$)	RMSE = 545, R² = 0.902	RMSE = 535, R² = 0.906	RMSE = 548, R² = 0.901	RMSE = 751, R² = 0.814
SVR (RBF)	RMSE = 1192, R² = 0.531	RMSE = 1205, R² = 0.521	RMSE = 1147, R² = 0.566	RMSE = 555, R² = 0.898

The ranking of algorithms is consistent across the four data densities: linear regression < gradient boosting < random forest < KNN < SVR (in order of increasing RMSE). The composite-wear features yield three distinct performance tiers. The first tier, linear Rregression, achieves $RMSE < 0.1$ days under all densities, an outcome interpreted in §4.2 as an analytical property of the composite-feature space rather than as a claim of operational predictive accuracy. The second tier, gradient boosting and random forest, exhibits a clear improvement with increasing data density: RMSE drops by approximately 60% from 30% to 60% data for both methods, with the largest gain occurring between 40% and 50% data. The third tier, KNN and SVR, fails to track the cyclic composite-wear behavior and produces errors one to two orders of magnitude larger than the ensemble methods.

A particularly notable feature of Table 5 is the small improvement between 50% and 60% data for the tree-based ensembles (gradient boosting improves from 22.9 to 21.2 days; random forest from 26.5 to 25.2 days), in contrast with the substantial improvement between 30% and 50%. This diminishing-returns pattern defines a practical sparsity threshold for ensemble methods, discussed in §5.1.

4.2 Linear regression: an analytical interpretation

Linear regression achieves near-zero prediction error ($RMSE \approx 0.1\,days$, $R^2 \approx 1.000$) across all training-data proportions, including the 30% subset. This result is not caused by overfitting in the conventional sense: model evaluation is conducted against the full 201-observation reference rather than the training subset, so a model that overfit the training set would not score well against the held-out portion. Instead, the outcome reflects the intrinsic linear structure of the composite-wear formulation under deterministic wear rates.

Within each reprofiling cycle, the composite index W is a piecewise linear function of $days_{elapsed}$ because the underlying wear rates (0.020 mm/month for diameter, 0.200 mm/month for flange) are constant in the simulation. RUL, by construction, is also a linear function of $days_{elapsed}$. Consequently, the engineered features in Table 2, particularly $days_{elapsed}$, $D_{commited}$, and $W_{composite}$, span the same analytical functional space as the target variable. A linear model that has access to these features can therefore reconstruct the mapping between features and RUL essentially without residual error, regardless of how many training points are observed within each cycle.

This finding has two interpretations that must be carefully separated: (1) Methodological validation (what the result does show). Linear regression's $R^2 \approx 1$ confirms that the composite-wear features (Eq. 1–4) are internally consistent with the deterministic data-generating process. The composite index successfully encodes the time-decay structure of RUL: there is no missing degree of freedom in the proposed feature set; (2) Operational predictive accuracy (what the result does not show). $R^2 \approx 1$ does not imply that linear regression will achieve operational $R^2 \approx 1$ on field data. Real wheelset wear is subject to stochastic effects, non-uniform load distribution, track-curvature variability, lubrication state, residual flange wear after imperfect reprofiling, none of which are present in the deterministic simulation. Under such conditions, the linear feature–target relationship will be perturbed, and the analytical exactness of linear regression will degrade towards the performance of gradient boosting or below.

For this reason, the practical algorithm recommended for field deployment is gradient boosting, not linear regression. Linear regression serves as an analytical baseline that validates the composite features; gradient boosting provides the robustness needed under the noise, non-uniform rates, and non-linear regime changes that field data exhibit. The role of linear regression in this study is therefore evidentiary (the composite formulation is mathematically complete) rather than prescriptive.

4.3 Gradient boosting and random forest: robust non-linear prediction

For scenarios in which real measurement data contain noise, non-uniform wear rates, or unknown non-linear behavior, tree-based ensemble methods provide a robust alternative to linear modelling. Gradient boosting demonstrates the strongest performance, achieving RMSE of 21.2 days ($R^2 = 0.9999$) at 60% data and RMSE of 58.5 days ($R^2 = 0.9989$) at 30% data, a prediction-error increase of 37.3 days when data density is reduced from 60% to 30%. Random forest exhibits a comparable trend, with RMSE rising from 25.2 to 75.0 days over the same range.

At 50% data density, both tree-based ensembles achieve RMSE values below 27 days over an operational horizon of ≈16 years. This corresponds to a relative prediction error of approximately 0.45%, which is sufficiently accurate for maintenance-planning

applications under realistic operating conditions. Combined with the diminishing-returns behavior above 50% data, this defines a practical minimum monitoring level of around half the maximum achievable measurement frequency (Fig. 4).

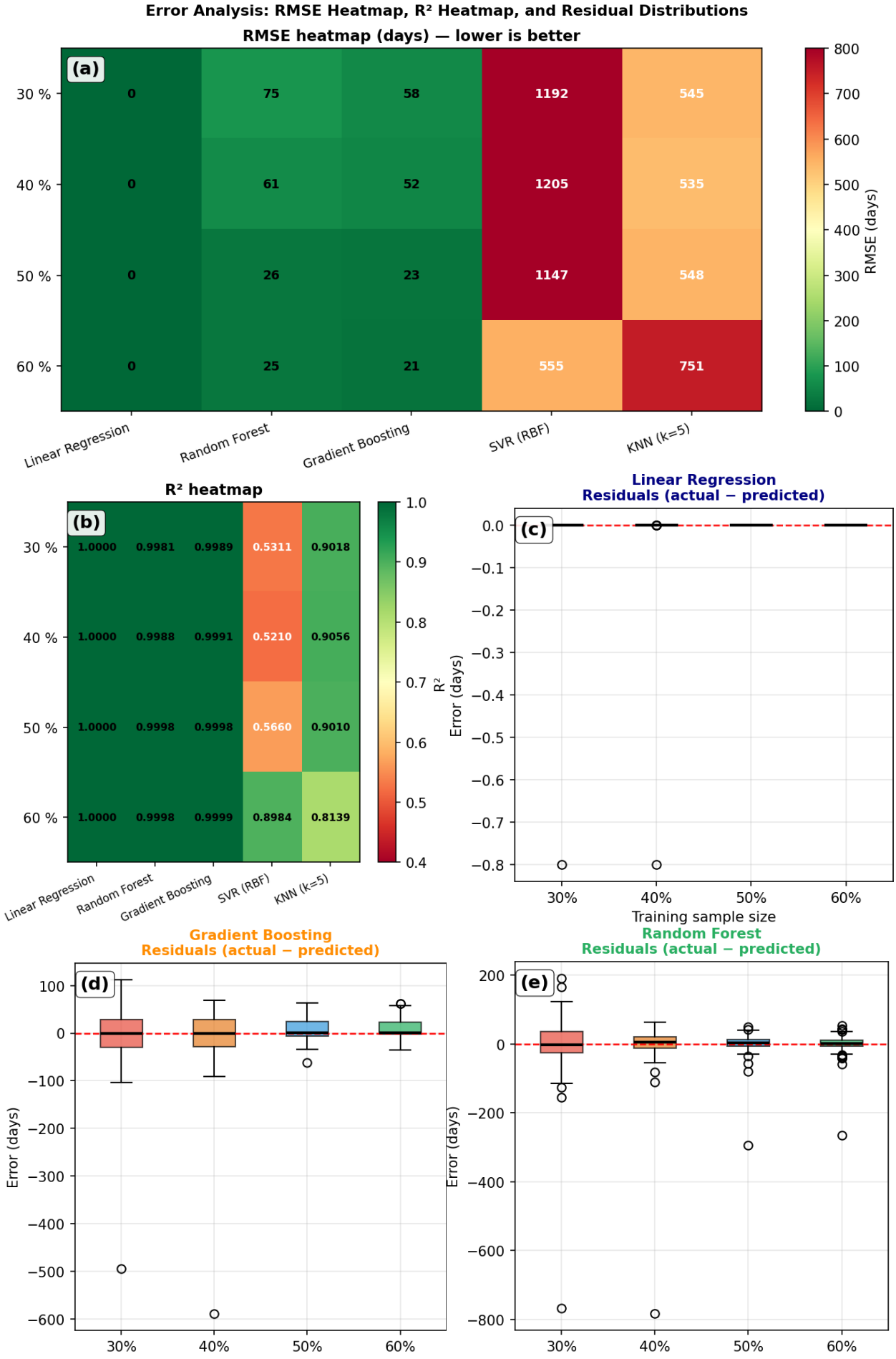
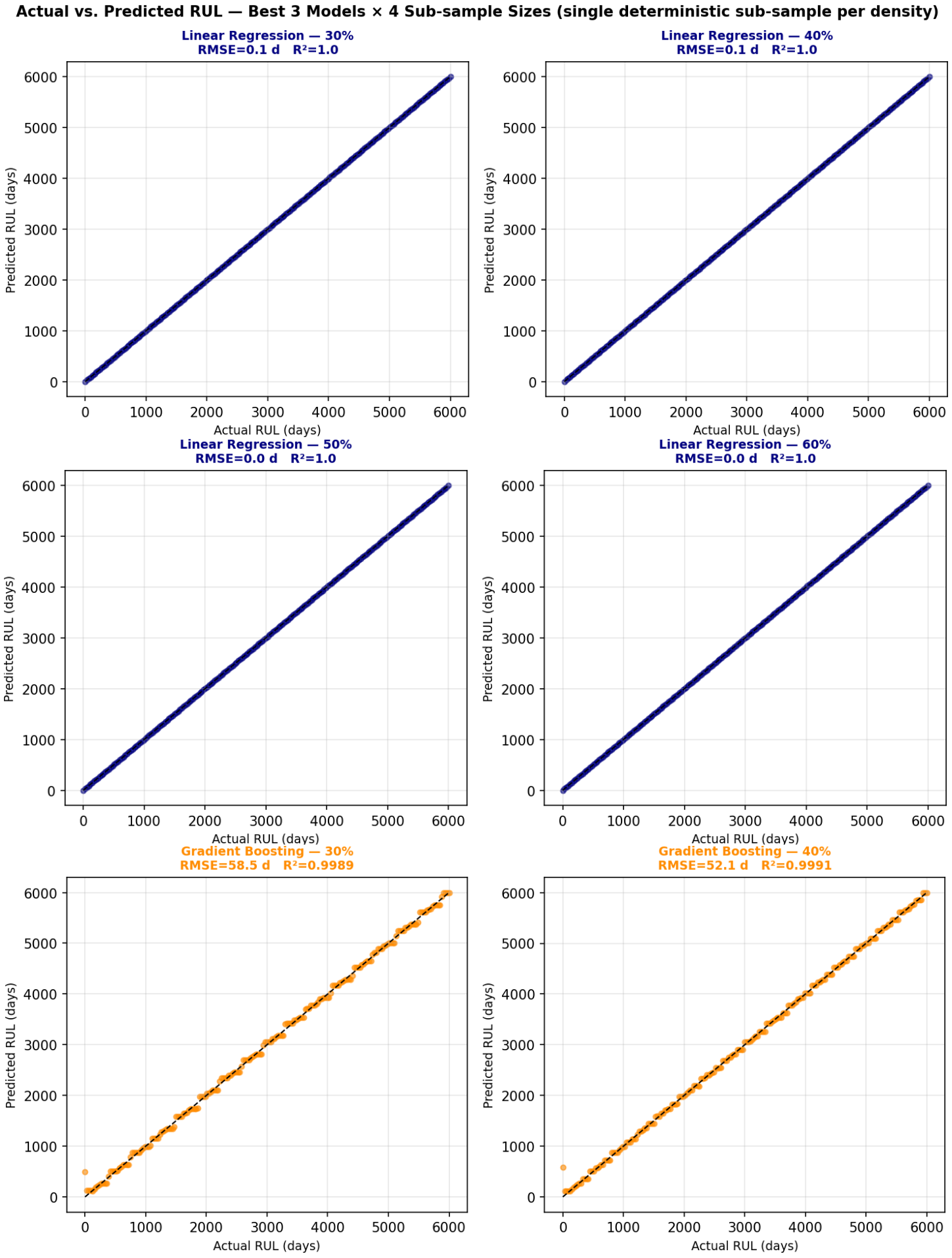


Fig. 4. Error analysis: (a) RMSE heatmap across the five models and four subsample sizes (lower is better); (b) R^2 heatmap; (c) residual boxplot for linear regression; (d) residual boxplot for gradient boosting; (e) residual boxplot for random forest. Residuals are computed as actual−predicted RUL on the held-out portion of the reference dataset.

4.4 Failure of distance-based methods (KNN and SVR)

KNN and SVR show consistently poor performance across all training-data proportions. SVR in particular exhibits very high prediction error, with RMSE ranging from 555 to 1,205 days. The primary cause is the cyclic sawtooth behavior of the composite wear index W (Fig. 2 and Fig. 5).

Because W returns approximately to $W_{cycle\ start}$ after each reprofiling event, the same value of W recurs across different reprofiling cycles. However, each occurrence corresponds to a different RUL value, with cycle-to-cycle RUL differences of approximately 1200 days. Observations that are close in feature space are therefore far apart in target space.



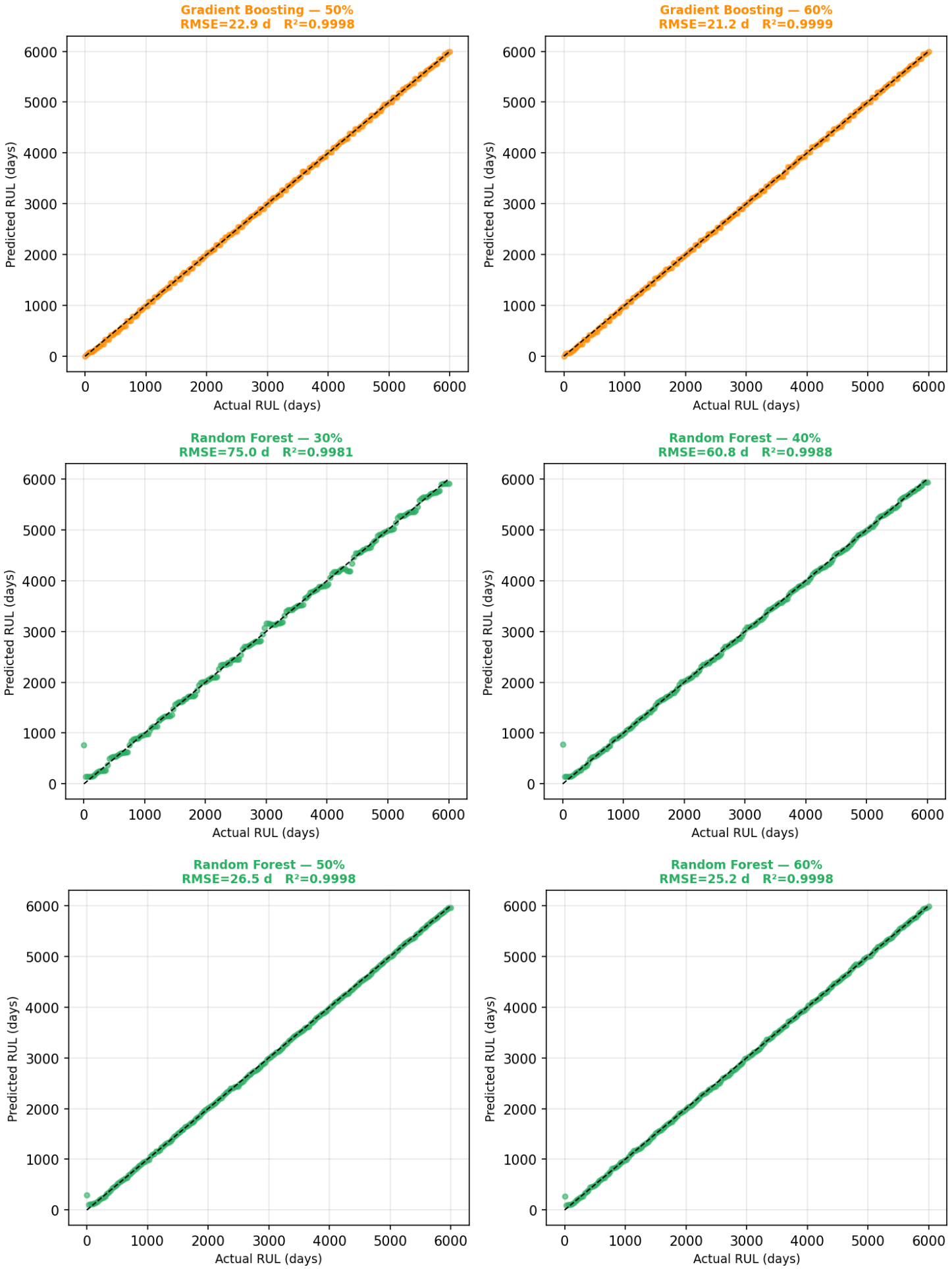


Fig. 5. Actual-vs-predicted RUL scatter plots for the three best-performing models (linear regression, gradient boosting, random forest; rows) across all four subsample sizes (30%, 40%, 50%, 60%; columns).

Distance-based and kernel-based methods, which rely on geometric proximity (KNN) or kernel similarity (SVR), cannot distinguish between these cyclic repetitions. The predicted RUL tends to be an average of multiple lifecycle stages, leading to large systematic errors. The discrete feature $n_{reprofiling}$ partially reduces this ambiguity by providing explicit lifecycle-stage information,

but the interaction between this discrete feature and the continuous wear variables requires a complex weighting that radial-basis-function kernels and simple neighbor averaging cannot efficiently capture. This limitation is consistent with findings reported in prior studies on regime-change degradation [3].

Each panel in Fig. 5. reports RMSE (days) and R^2 in its title. Perfect predictions lie on the black dashed diagonal. A complementary visualization of the full predictive behavior over the operational horizon is provided in Fig. 6. Fig. 6 shows the actual RUL trajectory (solid blue line), the model prediction (colored dashed line, one color per algorithm), and the training observations (orange dots) for all twenty combinations of five algorithms and four subsample sizes. None of the twenty panels represent missing data or model failures: every panel contains a complete prediction over the 16.4-year horizon. The visually larger excursions in the SVR and KNN rows reflect the systematic prediction errors of those algorithms (analyzed quantitatively in Table 5 and discussed above), not invalid or missing observations.

RUL prediction trained on each of the four subsample sizes (rows: 30%, 40%, 50%, 60%) as illustrated in Fig. 6. Solid blue line: actual RUL trajectory. Colored dashed line: model prediction over the full 16.4-year horizon. Orange dots: training observations included in the subsample. Panel titles report RMSE (days) and R^2 for each model–density combination. All twenty panels contain a complete prediction; none represent missing data or model failures.

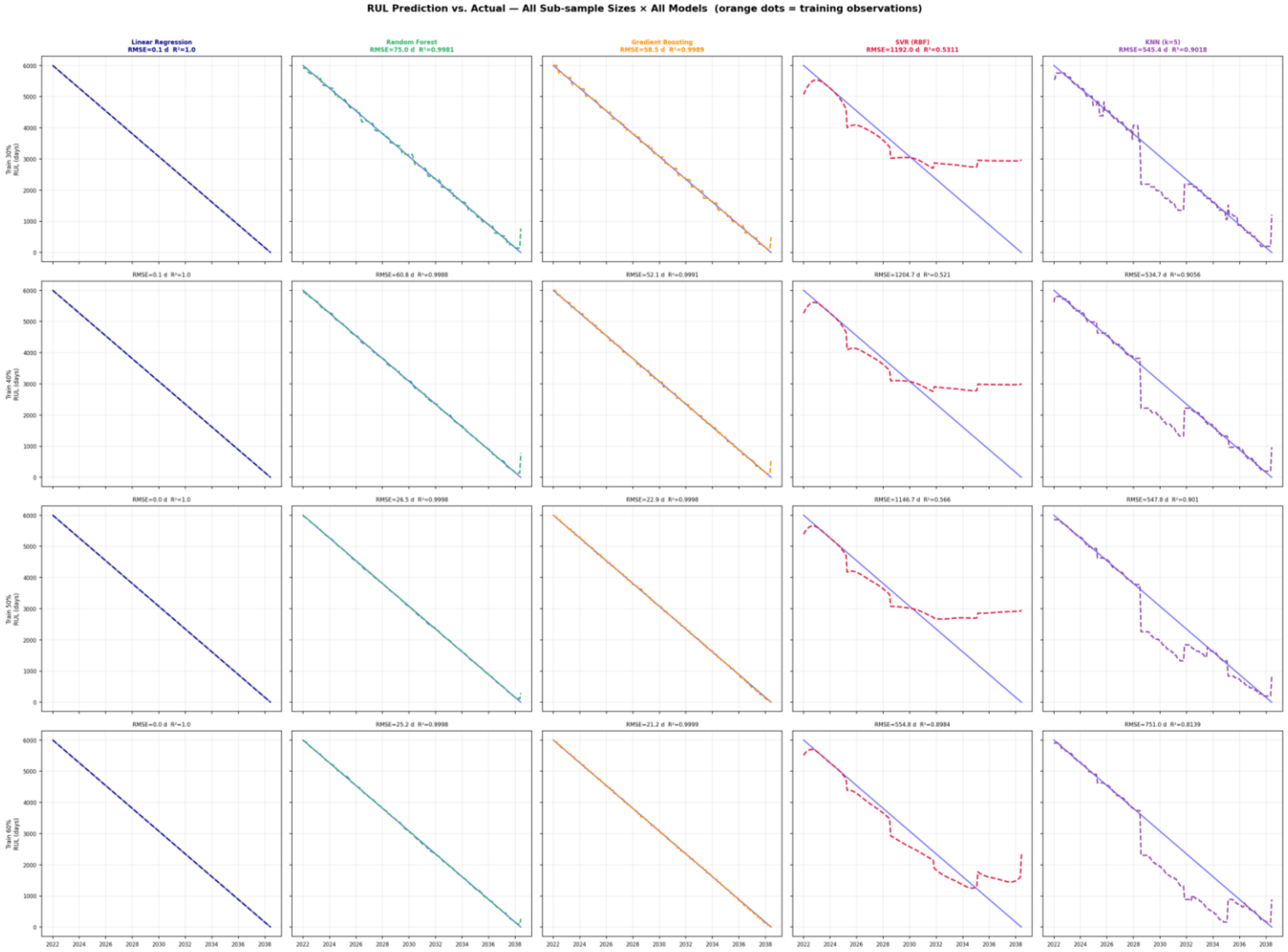


Fig. 6. RUL prediction curves for all five algorithms (columns: LR, RF, GB, SVR, KNN).

4.5 Composite-index thresholds and maintenance actions

The composite wear index W provides a physically interpretable framework for maintenance decision-making by defining threshold values that correspond to reprofiling-cycle boundaries. These thresholds, summarized in Table 6, are not arbitrary: they are direct geometric consequences of the deterministic model. Each reprofiling event removes 24.96 mm of diameter, which is approximately one quarter of the 100 mm diameter budget; hence the W trajectory naturally crosses 0.25, 0.50, 0.75, and 1.00 at the boundaries of successive cycles.

Table 6. Composite wear-index thresholds and corresponding maintenance actions, derived from the deterministic coupling $k = 3.2$.

W threshold	Diameter (mm)	Cycle	Maintenance action
0.00 – 0.25	850 → 825 (post-reprofiling.)	Cycle 1	Normal operation; monthly monitoring
≈ 0.25	825	Reprofiling 1	Schedule 1st reprofiling; budget 24.96 mm
0.25 – 0.50	825 → 800 (post-reprofiling.)	Cycle 2	Normal operation; tighten inspection interval
≈ 0.50	800	Reprofiling 2	Schedule 2nd reprofiling; confirm remaining budget
0.50 – 0.75	800 → 775 (post-reprofiling)	Cycle 3	Normal operation
≈ 0.75	775	Reprofiling 3	3rd reprofiling; assess retirement timeline
0.75 – 1.00	775 → 750 (post-reprofiling)	Cycle 4	Continued operation; near retirement
≈ 1.00	750	Reprofiling 4 / Cycle 5	4th reprofiling; final cycle before retirement
≥ 1.00	≤ 750	Final	Retirement or final reprofiling before withdrawal

These thresholds are physical boundaries that arise directly from the model parameters: they answer the question "at which W value does the next reprofiling become geometrically necessary?" Operationally and economically optimal thresholds, which would balance inspection cost, reprofiling cost, and the cost of unplanned downtime, are a complementary empirical-optimization problem that requires cost-model inputs from a specific operator and is flagged as future work in §5.4.

The thresholds in Table 6 can be implemented within an automated maintenance-alert system: when W exceeds a predefined level, the corresponding maintenance action is triggered. Because both wear effects are integrated within the composite representation, this decision is made independently of whether the limiting condition arises from diameter reduction or flange wear.

4.6 Field validation on real fleet data

The preceding results establish the composite formulation and the algorithmic benchmark on controlled deterministic data. This subsection validates the two central modelling choices, the coupling coefficient k and the composite index W, against the real PT Kereta Api Indonesia field log of TS1 (§3.8).

Empirical coupling coefficient. Applying the per-position estimator of §3.8 to the flange-driven reprofiling events of TS1 yields 179 valid position samples pooled ratio 2.84 mm/mm, per-position median 2.75 mm/mm (interquartile range 2.25 to 3.67), per-event median 2.71 mm/mm. The broader KKBW fleet (5,560 samples across 744 events) gives a pooled ratio of 2.93 mm/mm and a median of 3.00 mm/mm (Table 7, Fig. 7). The modelling value k = 3.2 from Karnadi [7] lies at the upper edge of the field distribution and within, or immediately adjacent to, the Karnadi range 2.95 to 3.31 mm/mm. The field evidence therefore confirms k = 3.2 as reasonable and slightly conservative: it marginally overwhelms the diameter cost of flange wear, biasing the composite index towards earlier, safer maintenance recommendations.

Composite index on the real fleet. Applying W (Eq. 1 to 4) to the 1,791 cleaned monthly records of TS1 confirms that it captures coupled degradation on operational data. Fig. 8 shows the monthly fleet-median trajectory with the interquartile spread across the 60 wagons: average diameter declines from about 845 to about 800 mm over three years, flange wear follows the expected reset-and-rise around reprofiling, and the composite median W rises monotonically from about 0.10 to 0.58. On the most recent record of each wagon, 62% of the fleet has crossed $W \geq 0.50$ and 15% has reached $W \geq 0.75$, demonstrating that the index discriminates wear stage across a heterogeneous real fleet, not only the idealised single-wagon simulation.

Taken together, the field results validate the two benchmark assumptions: the coupling coefficient is empirically grounded, and the composite index behaves as intended on noisy operational data. Full machine-learning deployment on multi-fleet field data with quantified uncertainty is future work (§5.4).

Table 7. Empirically observed coupling coefficient $k = \Delta D/\Delta f$ from real reprofiling events, compared with the modelling value and the Karnadi [7] range.

Source	Position-samples	Pooled k	Median k	p25 – p75
Train set TS1 (flange-driven)	179	2.84	2.75	2.25 – 3.67
Whole KKBW fleet (flange-driven)	5,560	2.93	3.00	2.20 – 4.00
Karnadi [7] empirical range	–	–	–	2.95 – 3.31
Modelling value (this study)	–	–	3.20	–

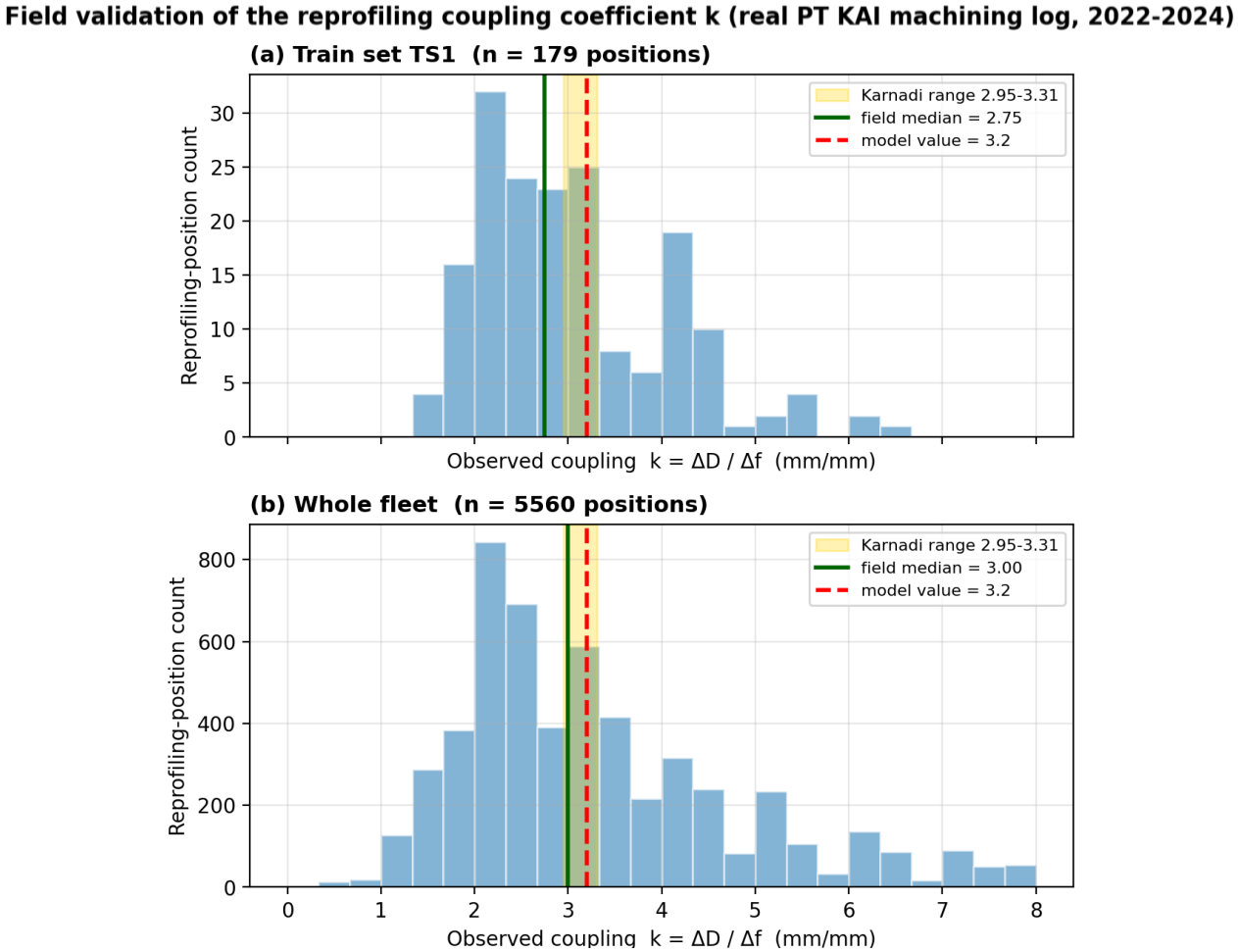


Fig. 7. Field validation of the reprofiling coupling coefficient k from the real PT Kereta Api Indonesia machining log: (a) train set TS1; (b) whole KKBW fleet. Histograms show the observed per-position distribution of k = ΔD/Δf; the gold band marks the Karnadi [7] empirical range (2.95 to 3.31), the solid green line the field median, and the red dashed line the modelling value k = 3.2.

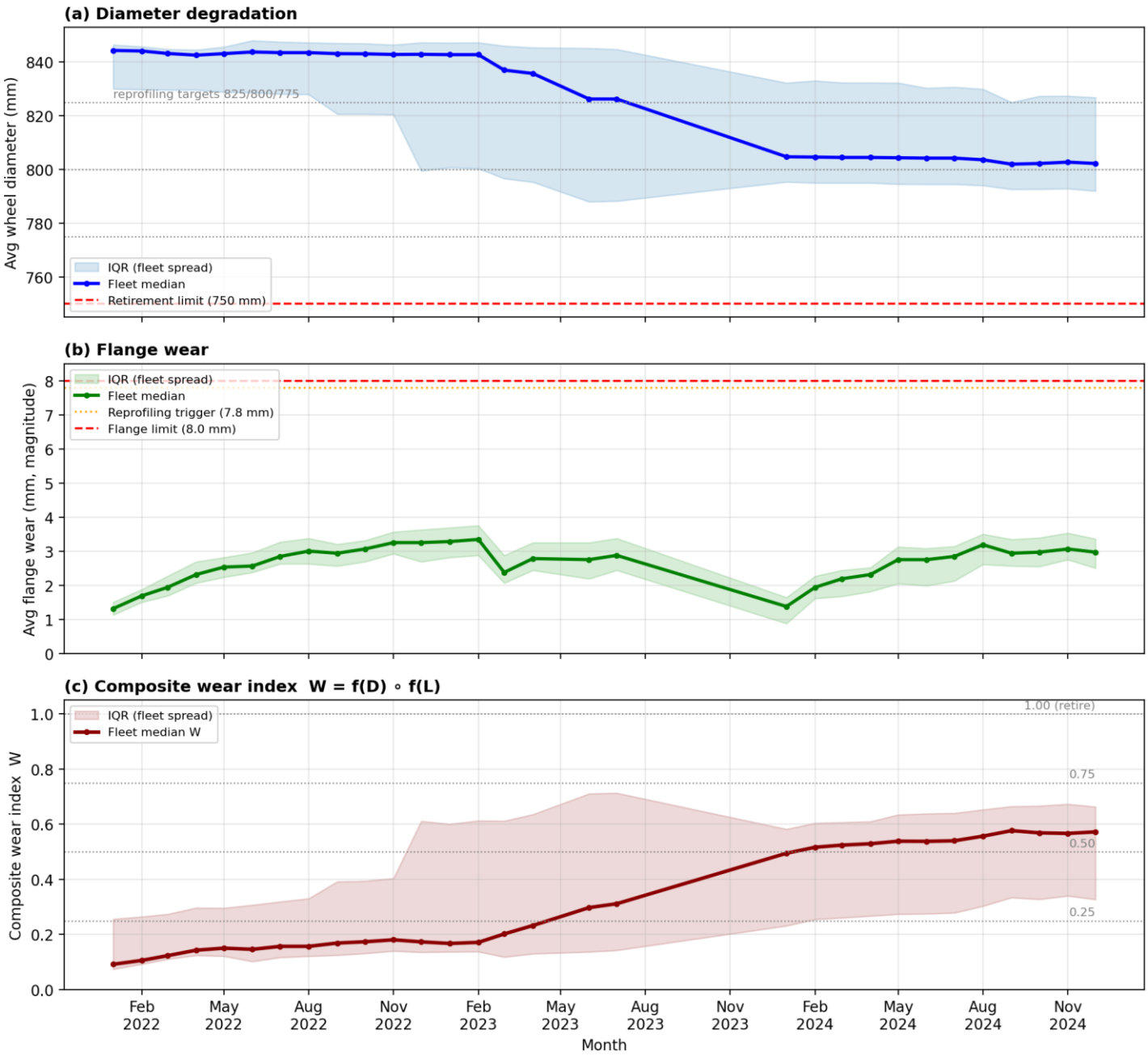


Fig. 8. Composite wear index on the real fleet (train set TS1, 60 wagons).

5 Discussion

5.1 Practical implications for railways maintenance

Third, the performance behavior of gradient boosting and random the findings of this study provide several practical implications for railway maintenance under simulated deterministic conditions. First, the composite wear index W offers a unified scalar representation that integrates both direct diameter loss and the deferred diameter reduction associated with flange restoration. This enables maintenance planners to evaluate the true proportion of the wheel-life budget that has been consumed at any given time. The advantage of this representation becomes particularly evident during the intermediate phase of a reprofiling cycle, where flange wear may have progressed significantly while the remaining diameter still appears sufficient when assessed independently.

Second, linear regression's $R^2 \approx 1$ across all data densities, interpreted analytically in §4.2, confirms that the composite-feature set is internally consistent with the deterministic data-generating process. This methodological evidence supports the use of physics-informed composite features as a foundation for prognostic modelling, although the operational predictive accuracy under field conditions must still be established empirically.

Forest shows that prediction accuracy improves substantially as data availability increases up to approximately 50% of the full dataset. Beyond this level, additional data provides only marginal improvement. For ensemble-based methods, the practical minimum monitoring level is therefore around half of the maximum achievable measurement frequency. Monitoring intervals may be extended (for example to every two months instead of every month) without compromising prediction performance, provided that physics-informed composite features are used. This has direct economic implications for large fleets where inspection activities contribute significantly to operational cost.

5.2 Why linear regression achieves a near-perfect fit: risk of target leakage and mitigations

The analytical interpretation of linear regression's $R^2 \approx 1$ was presented in §4.2; this subsection examines the implications and the steps required to ensure the result is reported honestly. Linear regression's performance must not be interpreted as evidence of operational predictive superiority. Instead, two methodological observations should be carried forward: (1) The composite-feature set is structurally complete for deterministic wear. No latent

variable is missing from the proposed encoding. This is a positive finding for the formulation of the composite wear index but says nothing about robustness under stochastic perturbations; (2) The feature $days_{elapsed}$ represents a time-related feature that, in combination with deterministic wear rates, creates a near-direct analytical mapping to RUL. In real operational data, where wear rates fluctuate due to load, curvature, weather, and crew practice, $days_{elapsed}$ is no longer a strict proxy for RUL, and the analytical exactness of linear regression will degrade.

For honest reporting, the mitigations are adopted in the present paper: (1) The wording of the abstract, results, and conclusion explicitly qualifies "highly accurate" with "under the simulated deterministic conditions investigated"; absolute claims such as perfect, exact, and fully reliable are avoided; (2) Gradient boosting, not linear regression, is positioned as the algorithm recommended for field deployment, on the grounds that it generalizes better to noisy, non-deterministic data than a strictly linear mapping; (3) A sensitivity analysis evaluating model performance after removal of the $days_{elapsed}$ feature, or under additive Gaussian noise on the wear rates, is identified as priority future work (§5.4).

These safeguards address the concern that the reported linear regression performance reflects the structure of the simulated data rather than true predictive capability.

5.3 Realism of the sparsity pattern

Random observation removal approximates inspection skips due to scheduling or budget constraints, in which a measurement is simply not taken at an otherwise-scheduled time. Real maintenance datasets, however, also exhibit systematic gaps, for example, seasonal shutdowns, audit periods, scheduled overhauls, or fleet rotations, that produce structured rather than random missingness. The present study does not reproduce such structured patterns. Future work should evaluate the composite-wear model under realistic structured-missingness regimes derived from actual maintenance logs of Indonesian rolling stock.

5.4 Limitations of the current study and future work

The principal limitations of the present work, and the corresponding directions for future research, are: (1) Scope of field validation. The coupling coefficient k and the composite index W are validated against real PT Kereta Api Indonesia field data (§4.6), which addresses the central concern that the formulation might hold only under idealised conditions. The algorithmic comparison (Table 5), however, is still computed on the controlled deterministic benchmark, which abstracts away asymmetric load distribution, track curvature, vehicle dynamics, lubrication condition, and local contact effects. Running the full machine-learning comparison directly on multi-fleet operational measurements, where the near-perfect Linear Regression fit is expected to degrade, is the next step; (2) Fixed coupling coefficient. k = 3.2 mm/mm is treated as a fixed parameter. The field analysis of §4.6 validates this choice (field median 2.75 to 3.00 mm/mm; k = 3.2 at the upper edge of the observed distribution), but the spread is substantial (interquartile range 2.2 to 3.7), confirming that k varies across events with wheel diameter, rail profile, contact condition, and lubrication. A sensitivity analysis sweeping k across the observed field range, and ultimately an adaptive composite index in which k is updated per reprofiling event from the live machining log, are recommended as follow-up studies. $k = 3.2\frac{mm\ \Delta D}{mm\ \Delta f} k = 2.95 - 3.31\frac{mm\ \Delta D}{mm\ \Delta f}$; (3) Idealized reprofiling reset. Flange wear is reset to ≈0 at each reprofiling event, assuming complete profile restoration. Real reprofiling leaves residual wear that depends on machining quality. Cycle-to-cycle reprofiling variability should be modelled in extended studies; (4) Single deterministic subsample per density. The reported sparse-data metrics in Table 5 are based on a single fixed (non-random) subsample per data density, supplied as the pre-computed Excel files referenced in §3.5. A randomised multi-seed evaluation reporting mean ± standard deviation across, e.g., 30 random draws per density would provide tighter confidence bounds and is recommended for future revisions; (5) Single vehicle, single scenario. Results are based on a single freight-wagon configuration. Fleet-level validation across diverse vehicle types, operating regions, and maintenance histories is required for generalization; (6) Static regression formulation. The current models are static regressors that ignore temporal sequence. Integration of the composite features into sequence-based models such as long-short-term memory networks or temporal convolutional networks may capture temporal dependencies missed by static regression; (7) Deterministic-only thresholds. The W thresholds in Table 6 are physical boundaries derived from the deterministic coupling. Operationally and economically optimal thresholds, which would integrate inspection cost, reprofiling cost, and unplanned downtime cost, require operator-specific cost models and are deferred to future work; (8) No uncertainty quantification. The current framework produces point estimates only. Extensions to probabilistic prediction (Gaussian-process regression, Bayesian neural networks, conformal prediction) would provide confidence intervals and support risk-informed maintenance decisions; (9) Integration with real-time monitoring. Integration of the composite-wear model with onboard sensors and wayside monitoring systems, enabling continuous estimation of wear state between scheduled inspections, is a natural deployment direction.

6 Conclusion

The main conclusions of this study, qualified by the deterministic simulation conditions investigated, are: (1) Linear regression achieves a near-perfect fit across all training-data proportions. This outcome is interpreted as an analytical consequence of the composite-feature space matching the linear time-decay of RUL under deterministic wear, providing methodological validation that the proposed feature set is structurally complete; (2) Gradient boosting provides the most reliable performance among non-linear models (RMSE 21.2 days at 60% data, RMSE 58.5 days at 30% data) and is recommended for practical deployment under variable operating conditions due to its robustness to noise and non-linear regime changes. Performance gains diminish beyond 50% data availability; (3) Distance-based methods (KNN, SVR) are not suitable for multi-cycle RUL prediction without explicit identification of the lifecycle stage, because the cyclic sawtooth behavior of the composite wear index causes the same feature-space location to correspond to different RUL values across cycles; (4) The coupling coefficient serves as a physically interpretable parameter that defines maintenance decision thresholds at 0.25, 0.50, 0.75, and 1.00, corresponding to successive reprofiling events. These thresholds are physical boundaries derived from the deterministic model rather than arbitrary choices. k = 3.2 is also validated against the real field log (field median 2.75 to 3.00 mm/mm), with 62% of the real fleet past midlife W ≥ 0.50); (5) Reduced monitoring frequency can deliver reliable RUL estimation under the deterministic simulation conditions investigated, when physics-informed composite features are used. Although the composite wear model was validated using field data, broader multi-fleet validation under realistic operational variability remains necessary for full deployment.

References

[1] T. Jendel, "Prediction of wheel profile wear, comparisons with field measurements," Wear, vol. 253, no. 1–2, pp. 89–99, 2002. https://doi.org/10.1016/S0043-1648(02)00087-X

[2] S. Bruni et al., "Simulation of wheel and rail profile wear: a review of numerical models," Railway Engineering Science, vol. 30, no. 4, pp. 362–385, 2022. https://doi.org/10.1007/s40534-022-00279-w

[3] Y. Song et al., "Analysis of wheel wear and wheel–rail dynamic characteristics of high-speed trains under braking

conditions," Shock and Vibration, vol. 2024, art. 9618500, 2024. https://doi.org/10.1155/2024/9618500
[4] Y. Zeng, D. Song, W. Zhang, B. Zhou, M. Xie, and X. Tang, "A new physics-based data-driven guideline for wear modelling and prediction of train wheels," Wear, vol. 456–457, art. 203355, 2020. https://doi.org/10.1016/j.wear.2020.203355
[5] F. Braghin, R. Lewis, R. S. Dwyer-Joyce, and S. Bruni, "A mathematical model to predict railway wheel profile evolution due to wear," Wear, vol. 261, no. 11–12, pp. 1253–1264, 2006. https://doi.org/10.1016/j.wear.2006.03.025
[6] W. Wang, "Joint prediction of remaining useful life and failure type of train wheelsets: a multi-task learning approach," arXiv preprint arXiv:2101.03497, 2021. https://doi.org/10.48550/arXiv.2101.03497
[7] A. F. Karnadi, "Prediksi Umur Pakai Roda Gerbong Batu Bara Sumatera Selatan Menggunakan Metode Regresi," M.Eng. thesis, Universitas Gadjah Mada, Yogyakarta, Indonesia, 2025. [Online]. Available: https://etd.repository.ugm.ac.id/penelitian/detail/260415
[8] T. Zonta, C. A. Da Costa, R. Da Rosa Righi, M. J. de Lima, E. S. Da Trindade, and G. P. Li, "Predictive maintenance in the Industry 4.0: A systematic literature review," Computers & Industrial Engineering, vol. 150, art. 106889, 2020. https://doi.org/10.1016/j.cie.2020.106889
[9] M. Emzain et al., "Implementation of failure mode and effect analysis (FMEA) for centrifugal pump maintenance in water supply distribution system," Jurnal Polimesin, vol. 22, no. 3, 2024. https://doi.org/10.30811/jpl.v22i3.4739
[10] Ruspendi et al., "Mitigating operational risks and enhancing machine performance through total productive maintenance and OEE: a case study on packaging equipment," Jurnal Polimesin, vol. 22, no. 5, 2024. [Online]. Available: https://e-jurnal.pnl.ac.id/polimesin/article/view/7469
[11] J. F. Archard, "Contact and rubbing of flat surfaces," Journal of Applied Physics, vol. 24, no. 8, pp. 981–988, 1953. https://doi.org/10.1063/1.1721448
[12] T. G. Pearce and N. D. Sherratt, "Prediction of wheel profile wear," Wear, vol. 144, no. 1–2, pp. 343–351, 1991. https://doi.org/10.1016/0043-1648(91)90025-P
[13] A. Shebani and S. Iwnicki, "Prediction of wheel and rail wear under different contact conditions using artificial neural networks," Wear, vol. 406–407, pp. 173–184, 2018. https://doi.org/10.1016/j.wear.2018.01.007
[14] Y. Ye, C. Huang, J. Zeng, S. Wang, C. Liu, and F. Li, "Predicting railway wheel wear by calibrating existing wear models: principle and application," Reliability Engineering & System Safety, vol. 238, art. 109462, 2023. https://doi.org/10.1016/j.ress.2023.109462
[15] M. E. Lutema and T. Edison, "Wear analysis of freight train within different curve parameters," International Journal of Industrial and Manufacturing Systems Engineering, vol. 10, no. 1, 2025. https://doi.org/10.11648/j.ijimse.20251001.11
[16] P. Mallioris, E. Aivazidou, and D. Bechtsis, "Predictive maintenance in Industry 4.0: A systematic multi-sector mapping," CIRP Journal of Manufacturing Science and Technology, vol. 50, pp. 80–103, June 2024.
[17] B. An et al., "A wheel-wear prediction model of non-Hertzian wheel–rail contact considering wheelset yaw," Wear, vol. 474–475, art. 203736, 2021. https://doi.org/10.1016/j.wear.2021.203736
[18] Q. Wu et al., "Heavy-haul rail/wheel wear and RCF assessments using 3-D train models and a new wear map," Wear, vol. 538–539, art. 205226, 2024. https://doi.org/10.1016/j.wear.2023.205226
[19] A. Karpatne, R. Kannan, and V. Kumar, Eds., Knowledge Guided Machine Learning: Accelerating Discovery Using Scientific Knowledge and Data. Boca Raton, FL, USA: Chapman & Hall/CRC, 2022. https://doi.org/10.1201/9781003143376
[20] J. Willard, X. Jia, S. Xu, M. Steinbach, and V. Kumar, "Integrating scientific knowledge with machine learning for engineering and environmental systems," ACM Computing Surveys, vol. 55, no. 4, art. 66, pp. 1–37, 2022. https://doi.org/10.1145/3514228
[21] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in Proc. 2008 International Conference on Prognostics and Health Management (PHM), Denver, CO, USA, 2008, pp. 1–9. https://doi.org/10.1109/PHM.2008.4711414