

Classifying climate-related failures for regional-national railway networks

Abdolamir Karbalaie^{a,b}, Khosro Soleimani-Chamkhorami^{a,c},
Stephen M. Famurewa^{a,d}, A.H.S Garmabaki^{a,*}

^a Department of Civil, Environmental, and Natural Resources Engineering, Division of Operation and Maintenance Engineering, Luleå University of Technology, Luleå 97187, Sweden

^b Department of Community Medicine and Rehabilitation, Physiotherapy, Umeå University, Umeå 901 87, Sweden

^c Department of Computer Engineering, Na.C., Islamic Azad University, Najafabad, Iran

^d Swedish Transport Administration, Luleå, Sweden

ARTICLE INFO

Keywords:

Railway infrastructure failures
Climate change impacts
Region-specific models
Explainable AI
Infrastructure resilience
Machine learning

ABSTRACT

The frequency and intensity of extreme weather events pose various types of hazards and vulnerabilities to the railway system. This article aims to challenge the assumption that a single, nationally and internationally aggregated model is sufficient to capture region-specific failure patterns. While useful, these national models are often unable to account for linkages and feedback between heterogeneous regions and associated railway failure mechanisms and processes. This study develops and evaluates regional-national specific machine learning (ML) framework to enhance the classification of climate-related railway failures utilizing a synchronized five-fold cross-validation strategy.

We illustrate the approach through a case study of the Switches and Crossings (S&C) assets across five railway operational regions in Sweden, over a 24-year period, to demonstrate meaningful climate variability. Our methodology provides a consistent procedure that consistently compares national models with region-specific models. Results show that the localized model outperforms the national baseline, with improvements of up to 2.0% in Matthews Correlation Coefficient and 1.2% in F1-score at the 24-hour window preceding failure. Shapley Additive Explanations (SHAP) analysis revealed regionally distinct predictors, such as snow accumulation in the North and weather data coverage issues in the West. The results provide a basis for further process refinement, offering interpretable insights for operation and maintenance planning, risk assessment and supporting climate adaptation actions to achieve a climate-proof and resilient railway system.

1. Introduction

Railway networks, as spatially distributed infrastructures, span diverse geographical and environmental regions. Extreme weather events, such as heavy snowfall, heatwaves, sea level rise, and intense rainfall have become more frequent, making the railway system more susceptible to a wide range of failures caused by environmental and extreme weather events (Andersson et al., 2015; Garmabaki

* Corresponding author: Luleå University of Technology, Sweden.

E-mail address: amir.garmabaki@ltu.se (A.H.S Garmabaki).

et al., 2021; IPCC6 et al., 2022; Koks et al., 2019; Misnevs et al., 2015; Nemry & Demirel, 2012; Thaduri et al., 2021).

As climate variability and extremes intensify, maintaining dependable rail service becomes increasingly challenging across diverse geographic regions, climatic zones and heterogeneous environments. Given this geographic heterogeneity, a uniform approach for failure prediction may fail to capture localized risk factors. Recent investigations have revealed the need for various interventions to address emerging threats, including the slow pace of climate change, the increasing frequency and intensity of extreme events, and shifting weather patterns. Comprehensive classification of climatic impacts, including physical and operational impacts on different railway infrastructure assets, have been reviewed by (Argyroudis et al., 2019; Haghighi et al., 2025) and (Forzieri et al., 2018; Ochsner et al., 2024; Palin et al., 2021). Table 1 summarizes the climate impacts based on the above-mentioned studies. These studies consistently indicate that railway systems will face heightened exposure to climate-induced disruptions in the future.

In addition to the above-mentioned climate stressors reported by various scholars, Zhao et al. (2020) analyse railway disruptions in China between 2000 and 2016 caused by gales (13 %), heavy rainfall (72 %), and snow (5 %). Compound climatic hazards—such as snow combined with freezing rain or strong winds—can exacerbate risks by creating stressors that manifest as infrastructure failures, including slippery tracks, damaged catenary systems, and broken power lines. These.

S&Cs = Switches and Crossings.

failures, in turn, lead to prolonged service disruptions, as observed in China during 2008, 2015, and 2018 (Adams & Heidarzadeh, 2021; Gao, 2016; Wang et al., 2020; Zhou et al., 2017).

The railway system is facing new climate-induced failures due to various extreme climatic impacts across the network (see Fig. 1). These failures include collapsing snow galleries due to excessive snow loads, passenger train derailments caused by heavy precipitation, as well as rail buckling,

switches and crossings (S&Cs) issues, and overhead line failures resulting from thermal expansion. In addition, two derailments of LKAB iron ore trains on the Malmbanan line in Sweden, in 2023 and 2024, were caused by extreme cold and hard-packed snow, resulting from strong winds and large temperature fluctuations, respectively (Bačić, 2024; Nyström, 2025).

Artificial Intelligence (AI) and machine learning (ML) and digital technologies are proven tools for integrating diverse datasets, infrastructure risk modeling, identifying correlations among variables, and developing predictive models that adapt to new data over time (Argyroudis et al., 2022; Climate-ADAPT; Kristina Gustafsson, 2021; Tang et al., 2022). Garmabaki et al. (2023) developed a Climate-ID feature for railway assets by integrating silos databases and creating correlation between meteorological parameters and infrastructure failure using ML.

Tian et al. (2024) studied the correlation of delay in cold regions under various Representative Concentration Pathways (RCPs) using Deep Learning. They reveal that the number of short delays decreased due to changes in winter temperatures, and long delays were exacerbated as the frequency of extreme weather events increased.

Wu et al. (2025) propose an optimization model for finding an optimal timetable to minimize the carbon emissions resulting from its operation while the intermittency brought by natural weather conditions imposes difficulties on their utilization. Lessan et al. (2019) proposed a Bayesian network (BN)-based model with an accuracy of 80 % for predicting train delays, capturing the complex dependencies in train operations. Chen et al. (2021) incorporated weather data into ML models to forecast delays in high-speed rail networks.

Table 1
Effects of climate change on the railway infrastructure network .

Climate Hazards	Impacts on Vulnerable Asset	Supporting References
Extreme rainfall including flooding	• Landslides • Track erosion & wash-out • Slippery tracks • Flooded signalling/electrical systems (e.g., S&Cs) • Access prevention to tracks/stations • Reduced speed & capacity	(Manning et al., 2008), (Jiang et al., 2015), (Binti Sa'adin et al., 2016), (Kellermann et al., 2016), (Kalantari et al., 2017), (Matthias & Christoph, 2018), (Bubeck et al., 2019), (Liu et al., 2022), (Kasraei et al., 2024), (Garmabaki et al., 2023), (Liu et al., 2021)
Extreme high temperature air & ground	• Track buckling • Structural thermal stress (rails, bridges, catenary) • Line closures • Safety risks for maintenance crews • Signal & equipment malfunction • Vegetation fires • Soil shrinkage, bridge scour	(Dobney et al., 2010), (Kasraei et al., 2024), (Nguyen et al., 2012), (Palin et al., 2013), (Davies & Andrews, 2021), (Li et al., 2023), (Garmabaki et al., 2021), (Mulholland & Feyen, 2021)
High winds	• Rolling stock instability • Overhead/catenary system overload • Structural load on bridges • Falling trees • Snow galleries damage	(Nezval et al., 2023), (Saback et al., 2024), (Nasr et al., 2020), (Imam, 2019), (Fu & Easton, 2018), (Szymczak et al., 2022),
Snow and ice	• Clogged areas from trees • Blockage in switches and crossings due to snow/ice; snow cleaning; reduced speed & capacity • Frozen overhead lines • Increased aerodynamic load (with wind) • Slippery tracks	(Roghani, 2021), (Soleimani, Karbalaie, et al., 2024), (Kasraei et al., 2024), (Soleimani, Garmabaki, et al., 2024), (Ochsner et al., 2024), (Haghighi et al., 2025)
Sea level rise	• Coastal erosion • Erosion of coastal defenses, saltwater corrosion	(Snoussi et al., 2009), (Thaduri et al., 2021)

adapted from (Haghighi et al., 2025)



Fig. 1. Examples of climate-related stressors and resulting failures in railway infrastructure. Panels illustrate how climate hazards (stressors) such as snow accumulation (a, e), flooding (b,d), extreme temperature (c,f), and extreme snow and wind (a, e) can lead to infrastructure failures including overload, erosion, rail buckling, and derailments. These examples highlight the vulnerability of railway systems to evolving climatic extremes. Photos

adapted from Sources (Bačić, 2024; Nyström, 2025).

Liu et al. (2018) used Random Forests to estimate rainfall-induced risks and failure probabilities of track components. Kasraei et al. (2024) investigated the impact of variations in climate variables—such as temperature, precipitation, humidity, and wind speed on railway infrastructure assets failures. Hong et al. (2021) further demonstrated the use of an ML-based model to predict rail temperatures across networks, thereby helping to mitigate rail buckling, using weather and solar radiation data. Recent frameworks have further emphasized the need for explainability, developing online machine learning pipelines that predict failures from sensor data streams and provide real-time, natural language explanations for maintenance decisions (García-Méndez et al., 2025).

While prior work has shown the effectiveness of machine learning in classifying climate- and non-climate-related railway failures at the national level (Soleimani, Karbalaie, et al., 2024), it relied on the assumption of spatial homogeneity. However, climate-related failures are strongly correlated with geographic location and associated features. This highlights a gap in existing research for further improvement of railway resilience and mitigation actions.

To address this, we test the following hypotheses:

H1: Although national models provide valuable insights, region-specific machine learning models provide more accurate and operationally meaningful predictions of climate-related railway failures.

H2: Different regions exhibit structurally distinct predictors for failure classification, which vary geographically in ways that only disaggregated modeling can reveal.

H1 is tested by training synchronized national and regional models on identical folds and comparing $\Delta = \text{Local} - \text{National}$ for Accuracy, F1-score, and MCC across time windows. H2 is tested by computing SHAP values within each fold and aggregating them to compare feature rankings across regions.

1.1. Proposed framework

The proposed framework focuses on establishment a connection between railway infrastructure failures and Climate key performance indicators (KPIs) at both national and regional levels utilizing ML models. We evaluate national and regional models on a 24-year Swedish dataset combining failures, asset inventory, inspections, maintenance, and meteorology across five regions. A synchronized five-fold cross-validation (CV) ensures identical train/test partitions for national and regional models. Class imbalance was handled using fold-specific class weights computed from the training split only; no over-/under-sampling was applied. Furthermore, models are assessed using Accuracy, F1-score, and Matthews Correlation Coefficient (MCC); we also report $\Delta = \text{Local} - \text{National}$ by region and time window (12, 24, 48, 72 h). For interpretability, Shapley Additive Explanations (SHAP) are computed within each fold and aggregated to identify region-specific drivers. A compact overview appears in Fig. 5, and a formal definition of the CV scheme is provided in Appendix A.

Our proposed framework has a portability to other critical infrastructure network. Applying the framework requires (i) mapping local failure categories (ii) climate failure labeling (iii) alignment of weather inputs to different time windows considering geographical location and homogeneity. Through these processes, the synchronized CV, Δ reporting, and SHAP steps will remain unchanged.

The main contributions of this paper are as follows:

Developing a framework to evaluate national and regional models through synchronized, identical splits to perform a fair comparison across five Swedish regions (24-year dataset).

- Quantifying gains from regionalization by reporting Δ (Local – National) per region and horizon (12, 24, 48, 72 h).
- Analyzing regional drivers with SHAP to reveal how predictive features differ geographically.

The rest of the paper is structured as follows: Section 2 investigates climate change projections and failure trends for railway infrastructure assets in Sweden. Section 3 outlines the dataset, feature engineering, and modeling approach. Section 4 presents model performance and interpretability analyses. Section 5 discusses results and describes implications, limitations, and future research

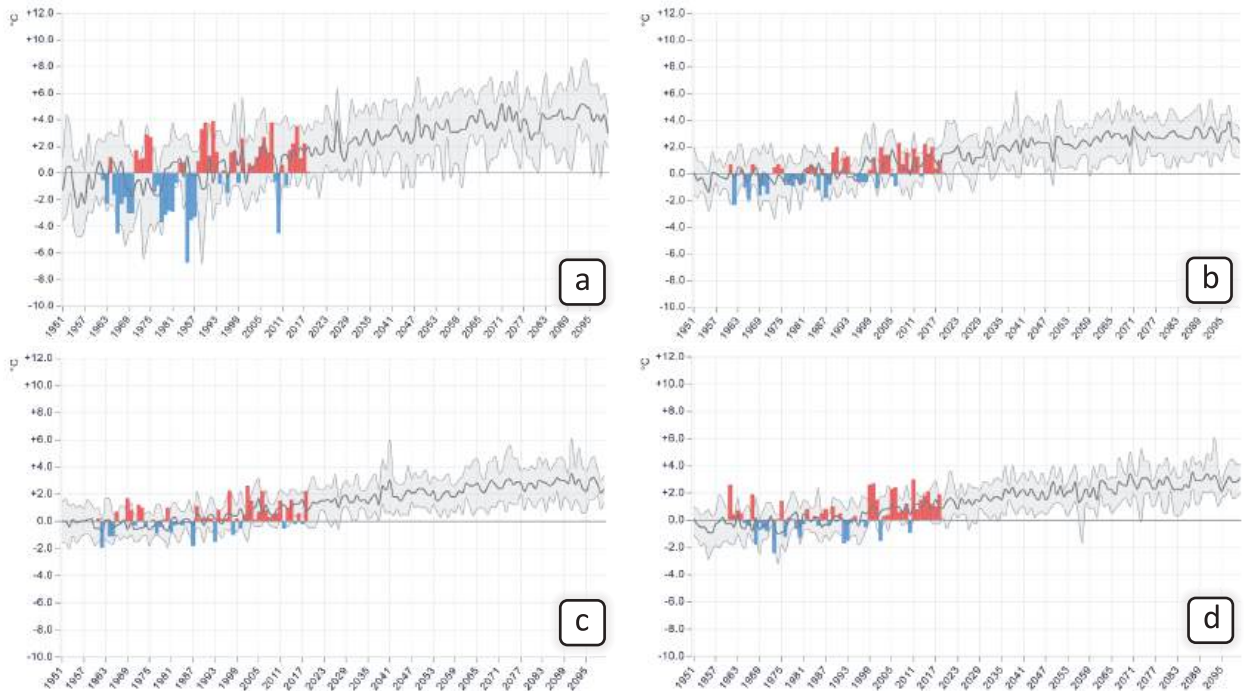


Fig. 2. : Projected seasonal temperature changes (°C) for Sweden under RCP4.5, based on SMHI ensemble data. Each panel shows the projected mean and range for the periods 1971–2100, with the historical reference mean (1971–2000) annotated. **a) DJF (Dec–Feb):** Mean reference: -7.3 °C; projected increase: $+4.1$ °C; **(b) MAM (Mar–May):** Mean reference: 1.3 °C; projected increase: $+2.9$ °C; **(c) JJA (Jun–Aug):** Mean reference: 12.6 °C; projected increase: $+2.8$ °C; **(d) SON (Sep–Nov):** Mean reference: 2.6 °C; projected increase: $+2.8$ °C. Blue/red bars (blue = colder, red = warmer) show historical data derived from observations (1961–2017), the upward red bars show temperatures higher than the reference period and the descending blue bars show temperatures lower than the reference period. The grey curve shows an average of several climate models for the RCP4.5 emission scenario. The grey bar shows the range of variation between the 10th and 90th percentiles. **RCP** = Representative Concentration Pathway; **SMHI** = Swedish Meteorological and Hydrological Institute. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

directions. The paper concludes by summarizing the key findings in [Section 6](#).

2. Climate change projections and railway failure trends

To assess the long-term vulnerability of railway infrastructure to climate change, it is critical to examine both projected climatic shifts and observed patterns in failure data. Because our verification case is Sweden, this section focuses on Sweden's geography and railway network.

According to Swedish Meteorological and Hydrological Institute (SMHI)'s projections under the Representative Concentration Pathway (RCP) 4.5 scenario, Sweden is expected to experience the following changes by the period 2071–2100 relative to the 1971–2000 baseline:

1. Hot days ($>25^{\circ}\text{C}$) will increase across the country, with the greatest temperature rise observed in northern regions despite the highest frequency occurring in the south.
2. Zero-crossing events—temperatures fluctuating around freezing—will become more frequent in the north and less frequent in the south, leading to increased freeze–thaw cycles in northern zones.
3. Total snowfall is expected to decrease overall, though intensity and accumulation may rise in the north due to colder winters.
4. Annual precipitation will increase throughout the country, heightening the risk of track flooding and soil saturation.

These projected changes directly affect railway functionality, particularly in thermal expansion, snow accumulation, subgrade instability, and switch malfunctions.

[Fig. 2](#) illustrates the seasonal temperature changes for Sweden, based on the RCP 4.5 scenario. The most pronounced warming is expected during winter (December–February), with average temperatures projected to increase by up to 4.1°C . The projected temperature increments for spring, summer, and autumn are approximately 2.8°C , which still represents substantial seasonal warming.

To contextualize the projected climatic shifts, we analyzed empirical failure data from the Swedish Transport Administration (Trafikverket) covering 2001–2023. [Fig. 3](#) presents the temporal distribution of climatic and non-climatic railway failures. Climate-related failures show distinct peaks during or immediately following extreme weather years, notably in 2009–2010, 2018–2021, and 2023 which coincide with periods of snowstorms, heatwaves, and flooding. The temporal alignment between these events and failure counts supports the inclusion of weather anomalies as predictive features in modeling frameworks.

For some years, the peaks in non-climatic failures have coincided with the peaks related to climate failure ([Fig. 3](#)). This can be explained in several ways. Some failures recorded as non-climatic are still indirectly influenced by weather; for example, heavy snow or rain can lead to signaling faults or mechanical breakdowns coded as technical rather than climatic. In addition, extreme events place extra strain on the whole system, increasing the number of failures. In addition, Climate-ID labeling processes pose some degrees of subjectivity due to the lack of proper failure reporting system. These combined effects explain why both climatic and non-climate-related failures show similar spikes in some years.

Complementing the national view, [Fig. 4](#) illustrates the distribution of failure modes in urban areas specifically Stockholm, Gothenburg, and Skåne over 2010–2023. The figure shows a steady and substantial increase in failure frequency and variety across various asset types (e.g. switches, contact lines, and signaling). The distribution of all the failures across different regions in Sweden is discussed and presented by (Soleimani, Garmabaki, et al., 2024).

The urban failure escalation observed since 2010 mirrors the intensification of climate-related incidents and reflects growing operational challenges in densely trafficked regions. Relief in recent years, as seen in [Fig. 4](#), can be attributed to elevated awareness of

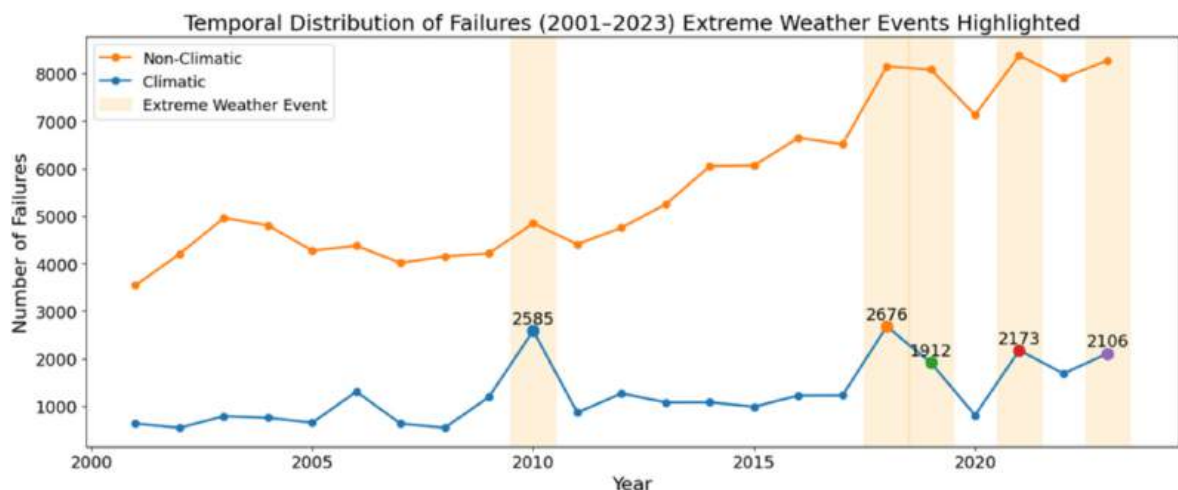


Fig. 3. Temporal distribution (2001–2023) of climate-related vs non-climatic failures. Shaded bands mark known extreme-weather years.

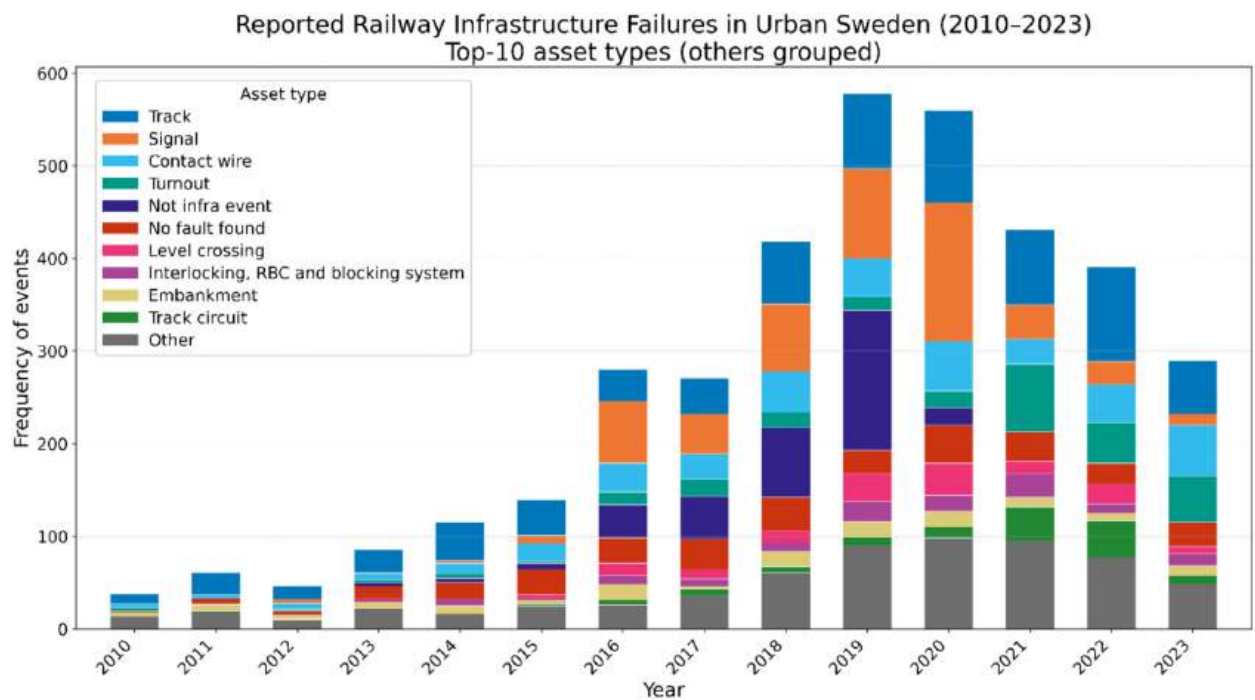


Fig. 4. Urban subset of reported railway infrastructure failures for recent years (2010–2023). Annual frequencies for the 10 most common asset types (all others grouped as “Other”). Scope is limited to Stockholm, Gothenburg, and Skåne, where urban tagging is consistently available from 2010 onward. *Bro* = Bridges; *S&Cs* = Switches and Crossings; *Balise group* = track transponder equipment; *Anläggningstyp.BVS* = asset type category (BVS standard); *Infra event* = non-infrastructure incident (e.g., vehicle-system fault or traffic-control priority); *No fault found* = cause could not be established.

the challenge or temporary climate adaptation actions, such as improved maintenance readiness.

3. Materials and methods

This section outlines the methodology used to classify climate and non-climate-related failures. The process included data acquisition and preparation, training and evaluation of regional and national models using a synchronized framework, and an interpretable machine learning approach. The entire workflow is visually summarized in Fig. 5. This section describes the data sources, the steps used to merge them into a unified dataset, the preprocessing procedures, and the quality assurance protocols applied.

3.1. Data acquisition and preparation

3.1.1. Data sources

Failure event data were obtained from the Swedish Transport Administration (Trafikverket) through the Ofelia failure reporting system (Trafikverket, 2000–2024). These sources provided time-stamped records of infrastructure failures between 2001 and 2023, including metadata on asset type, location, failure cause, and corrective actions. The BanInformation System (BIS) asset registry was used to extract asset characteristics, including installation year, asset category (e.g., track, switches, overhead line), traffic density class, and other technical features of the asset. Weather data were sourced from the SMHI, which provides daily gridded observations across Sweden. Climate key performance indicators (KPIs), including minimum and maximum temperatures, cumulative precipitation (rain and snow), wind speed, humidity, and zero-crossing indicators, were extracted from SMHI to construct weather-related factors for failure classification.

3.1.2. Dataset integration

The dataset integration process involved aligning failure reports, infrastructure metadata, and meteorological data into a single dataset for the model. In the first phase, failure events related to switching and crossing (S&C) assets from Ofelia failure reporting were integrated with infrastructure metadata from the BIS asset registry. Some records lacked regional identification for the events; therefore, we addressed this limitation by using geospatial coordinates for the regions and the locations of the S&C assets (see Fig. 6).

In the second phase, meteorological data from the SMHI were matched by location and time to each failure instance. These parameters were aggregated into four fixed prediction windows: 12, 24, 48, and 72 h before each failure. These windows capture short-term and intermediate atmospheric conditions leading up to infrastructure failures. In the third phase, after data cleaning and filtering, the final dataset included approximately 159,669 records with complete infrastructure and weather attributes.

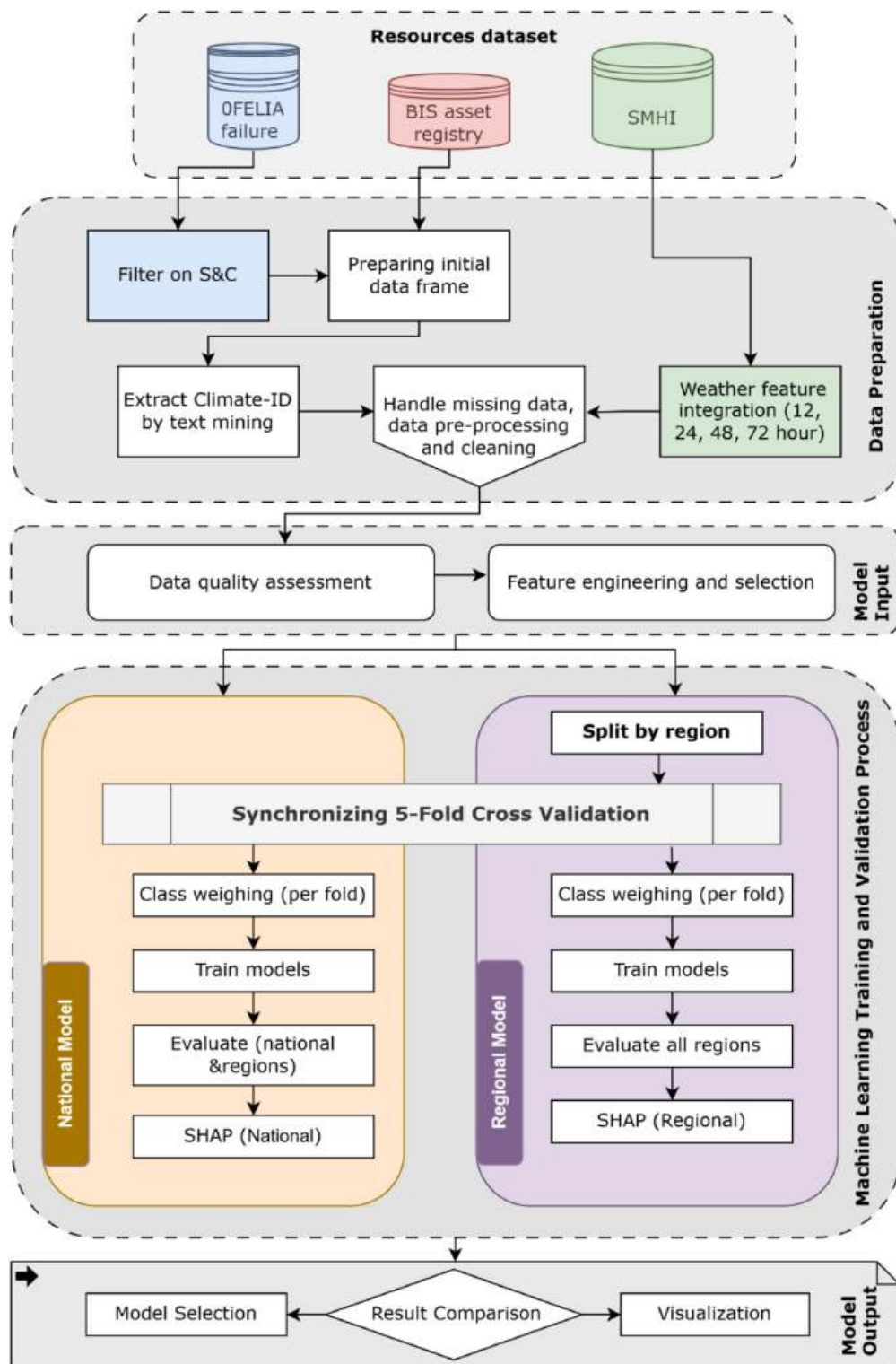


Fig. 5. Workflow for Climate-Aware Railway Failure Classification Using Regional Machine Learning Models.

In addition, the target variable (Climate-ID) was then constructed using failure codes, text mining, and expert review. Because failure and inspection databases are not designed to capture climate-related failures, expert input and free-text descriptions were analyzed to identify relevant keywords. This process resulted in a fully labeled dataset structured for the machine learning models.

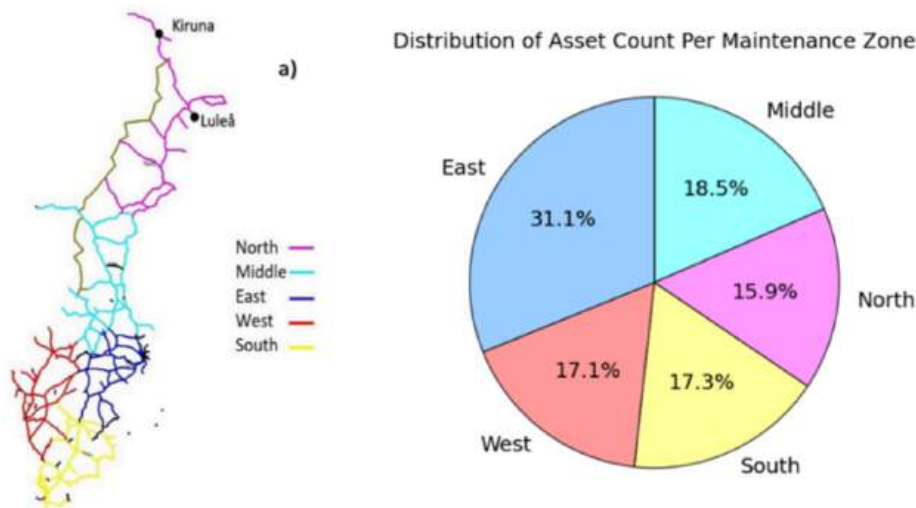


Fig. 6. Spatial division of the Swedish railway network into five operational zones used for region-specific modeling (left), and distribution of railway assets across these zones (right). Color coding aligns with regional asset distributions and is used consistently throughout the manuscript in downstream visualizations and performance analysis.

3.1.3. Data preprocessing

A structured preprocessing pipeline was applied to enhance data quality and model readiness. Incomplete records—such as failures lacking timestamps, geospatial coordinates, or infrastructure identifiers—were excluded. Text fields with limited predictive value (e.g., free-text notes) and attributes with excessive missing data were also removed.

Weather metrics (mean, minimum, maximum, and sums) were aggregated to reflect short-term climatic variability and climate KPIs. Categorical features with no inherent order, such as *Asset_type*, were one-hot encoded, while ordinal features with a natural ranking, such as *Railway_class*, were label encoded to keep the ranked order. Continuous variables, including Temperature, *Asset_age*, and Distance (km) were standardized using z-score normalization. To prevent data leakage, the z-score normalization scaler was fitted exclusively on the training data within each CV fold. Fig. 5 illustrates the full modeling workflow, from data acquisition and preprocessing to regional model training and evaluation.

Table 2 summarizes the complete set of input features, which includes raw meteorological inputs, derived weather indicators, categorical infrastructure attributes, and domain-informed variables such as asset age and temperature change. These features were selected to optimize interpretability and performance within tree-based classification models across retrospective time windows.

Table 2

Summary of Input and Engineered Features for Model Training.

Feature Name	Type	Source	Description	Engineering Formula
<i>Temp_min</i>	Meteorological	SMHI	Minimum temperature prior to failure	Raw value
<i>Temp_max</i>	Meteorological	SMHI	Maximum temperature prior to failure	Raw value
<i>Temp_change</i>	Engineered	<i>temp_max</i> , <i>temp_min</i>	Temperature fluctuation as failure indicator	$temp_max - temp_min$
<i>Snow_sum</i>	Meteorological	SMHI	Total snow accumulation	If $temp < 0$, $precip = snow_sum$
<i>Rain_sum</i>	Engineered	precipitation, <i>temp</i>	Cumulative rainfall only (excludes snow)	If $temp > 0$, $precip = rain_sum$
<i>Humidity_max</i>	Meteorological	SMHI	Max humidity before failure	Raw value
<i>Windspeed_max</i>	Meteorological	SMHI	Max wind speed	Raw value
<i>Asset_age</i>	Engineered	BIS	Age of asset at time of failure	$failure_year - install_year$
<i>Asset_type</i>	Categorical	BIS	Track type or asset classification	One-hot or label-encoded
<i>Region</i>	Categorical	Ofelia /derived	Operational region of the asset	Encoded ordinal
<i>Railway_class</i>	Categorical	BIS	Operational classification (e.g., mainline vs. local)	Encoded numerically
<i>Repair</i>	Numeric	Ofelia	Duration between failure report and fix	Used for filtering and prioritization
<i>Climate_id</i>	Target	Derived	Binary label for failure type (climatic vs. non-climatic)	Text mining + failure codes + expert annotation
<i>Distance(km)</i>	Engineered	Derived	Distance (km) between the railway asset and the nearest weather station.	Geospatial calculation

Note: Meteorological features were sourced from SMHI; infrastructure metadata was obtained from BIS and Ofelia. Engineered features such as *Temp_change*, *Repair_time*, and *Climate_id* were computed using raw measurements and domain-informed logic, as shown in the last column. Categorical features (*Asset_type*, *Region*, *Railway_class*) were encoded for use in tree-based models. The target variable *Climate_id* distinguishes between climate-related and non-climate-related failures based on text mining and manual annotation. All meteorological features were created using the event's failure timestamp as a reference point.

To ensure the integrity of the final dataset, additional quality assurance steps were applied, including the removal of duplicates, validation of timestamp-asset mappings, manual inspection of spatial joints between infrastructure and SMHI grid data, and targeted review of Climate-ID labels for borderline cases.

3.2. Model Development and training

3.2.1. Data Splitting strategy

To evaluate model generalization at both the national and regional levels, we implemented a synchronized five-fold CV strategy. In each iteration, one-fold per region was used for testing, while the remaining folds were used for training. The national model was trained on the combined training sets from all five regions and tested on the combined test sets, whereas each regional model was trained and evaluated exclusively on its own partitions. This synchronized approach ensured that the performance of national and regional models could be directly compared under identical conditions. We selected five folds as a balance between computational efficiency and robust estimation of generalization performance, following standard practice in machine learning (Hastie et al., 2009; Kohavi, 1995). For the formal notation of this partitioning, see Appendix A.

3.2.2. Performance metrics

To evaluate classification performance, we used five standard metrics derived from the confusion matrix: Accuracy, Precision, Recall, F1-score, and MCC (Matthews, 1975). Accuracy captures overall correctness; Precision reflects how many predicted climate-related failures were correct; Recall measures how many actual failures were identified; F1-score balances Precision and Recall; and MCC provides a single, robust score accounting for all four confusion matrix elements, particularly suitable for imbalanced data. To mitigate class imbalance, inverse class frequency weighting was applied during model training in each CV fold.

Delta Performance Scores (Δ): To compare model effectiveness across geographic levels, we computed Δ as the difference between each regional model's performance and the national model's performance on the same regional test data:

$$\Delta_{\text{metric}} = \text{Metric}_{\text{regional}} - \text{Metric}_{\text{national}}$$

These values were calculated per region and per fold and then averaged across folds. A positive Δ indicates that the regional model outperformed the national model on that region's data, while a negative delta suggests the opposite. This analysis enabled us to quantify the relative benefits of regional specialization across various performance dimensions.

3.2.3. Model training and hyperparameter optimization

To classify railway infrastructure failures as climatic or non-climatic, eight machine learning algorithms were evaluated: Extreme Gradient Boosting (XGB), Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), AdaBoost (ADA), LightGBM (LGB), and Gradient Boosting Classifier (GBC). These models span a range of complexity, interpretability, and scalability, making them suitable for evaluating structured datasets with class imbalance.

Model training and hyperparameter optimization were embedded within the five-fold CV strategy described in Section 3.2.1. To prevent information leakage, tuning was performed only on the training portion of each fold. A single representative training fold selected based on its class distribution closely mirroring the full dataset was used for hyperparameter tuning per model type. The optimized configuration was then fixed and reused across all CV folds for both national and regional models. This strategy balanced computational efficiency with consistency evaluation across spatial domains.

For ensemble models (XGB, RF, LGB, and GBC), Bayesian optimization was employed using the Tree-structured Parzen Estimator via the Hyperopt framework. This approach was preferred over grid or random search due to its efficiency in finding the best settings, even with limited computing time. A composite objective with equal weight to Accuracy, F1-score, and MCC guided the search (50 iterations per model). The hyperparameter search space included:

- **XGBoost:** n_estimators (100–1000), max_depth (3–10), learning_rate (0.01–0.3), subsample (0.5–1.0)
- **Random Forest:** n_estimators (100–1000), max_features ('sqrt', 'log2'), min_samples_leaf (1–10)

For baseline models with few tunable parameters (NB, LR), default scikit-learn configurations were adopted. DT and ADA underwent limited tuning on tree depth and boosting iterations. All experiments were tracked using MLflow, ensuring full reproducibility of parameter search and CV results. Model training was conducted independently for both national and regional configurations, with performance metrics computed on the held-out test folds in each case.

3.2.4. Handling class imbalance

To address class imbalance while preventing test data leakage, a fold-specific class weighting strategy was employed within the five-fold CV framework. Preliminary tests showed that Synthetic Minority Over-sampling Technique (SMOTE), achieved higher recall but lower precision and MCC on a held-out validation set; consequently, class weighting was selected as the preferred approach. This enabled robust evaluation without leakage across folds and regions.

3.2.5. Feature importance analysis

To ensure model transparency and understand the factors driving failure classifications, methods from Explainable AI (XAI) were

used. Specifically, SHAP was used, a leading XAI technique that enables the interpretation of the predictions made by the trained models by quantifying the contribution of each input feature. Based on cooperative game theory, a SHAP value is assigned to each feature to quantify its contribution to the model's output for a given prediction (Lundberg & Lee, 2017).

The SHAP framework explains a model's output $f(x)$ by distributing the prediction difference from the baseline $E[f(X)]$ across all input features:

$$f(x) = E[f(X)] + \sum_{i=1}^M \phi_i$$

where ϕ_i is the SHAP value of feature i , quantifying its influence on the prediction for a given instance. To measure global importance, we aggregated SHAP values across all test samples using the mean of the absolute values:

$$I_j = \frac{1}{n} \sum_{k=1}^n |\phi_j^{(k)}|$$

Here, I_j represents the overall importance of feature j , and $\phi_j^{(k)}$ is the SHAP value of feature j for the k -th instance.

To ensure the SHAP explanations were robust, their calculation was integrated into the five-fold CV process. In each fold, SHAP values were generated for the test instances using the corresponding trained model. The values from all five folds were then aggregated to create a complete and stable interpretation set for both the national and each regional model configuration. In each case, two types of visualizations were produced:

SHAP bar plots: displaying the mean absolute SHAP value for each feature, indicating its overall impact on model output.

SHAP beeswarm plots: showing the distribution of SHAP values across instances, which highlights how both the magnitude and direction of feature effects vary depending on the feature's value.

3.2.6. Model selection strategy and evaluation Criteria

To determine the optimal classification model and forecasting horizon, we employed a structured selection strategy based on three principles was employed: (i) predictive performance, (ii) stability across different time periods and regions, and (iii) compatibility with interpretability methods.

Predictive Performance: Each candidate model $m \in \mathcal{M}$ was evaluated across four time windows $h \in \{12, 24, 48, 72\}$ using three core classification metrics: Accuracy, F1-score, and MCC. For regional models, performance was computed independently within each region $r \in \{\text{North, Middle, East, West, South}\}$. A composite performance score for a given model, region, and horizon was defined as:

$$S_{m,r,h} = \frac{1}{3} (\text{Accuracy}_{m,r,h} + \text{F1}_{m,r,h} + \text{MCC}_{m,r,h})$$

To summarize performance across regions, an average composite score for each model and horizon was computed as:

$$\bar{S}_{m,h} = \frac{1}{5} \sum_r S_{m,r,h}$$

For national models, which are not region-specific $\bar{S}_{m,h}$ was calculated directly over the full dataset.

Stability Across Regions and Horizons: In addition to high average performance, robustness was assessed by examining the variance of metric values across both regions and horizons. Models exhibiting low variability in Accuracy, F1-score, and MCC were considered more stable and thus more reliable for deployment in varied operational environments. This aligns with our goal of ensuring temporal and geographic generalizability, especially under changing climate conditions.

Interpretability: Only models compatible with post-hoc, model-agnostic attribution techniques specifically SHAP, as described in Section 3.2.5, were considered eligible. This constraint ensures that model predictions are explainable, allowing infrastructure managers to understand the influence of individual features and support transparent decision-making in environments where outcomes are critical.

Final Model Selection: The final model-horizon pair (m^*, h^*) was chosen according to the following maximization criterion:

$$(m^*, h^*) = \arg \max_{m \in \mathcal{M}_{\text{SHAP}}, h} \bar{S}_{m,h}$$

where $\mathcal{M}_{\text{SHAP}}$ denotes the set of models that support SHAP-based interpretability. This multi-objective selection approach ensured that the final configuration exhibited high predictive accuracy, consistency across settings, and interpretability aligned with domain needs.

3.2.7. Implementation and computational environment

All analyses were conducted in Python (3.11.4); Scikit-learn (1.2.2) for baseline, XGBoost (1.7.6) and LightGBM (3.3.5) for ensembles, Hyperopt (0.2.7) for optimization, SHAP (0.41.0) for interpretability and MLflow (2.10.2) for experiment tracking. Computations ran on a Linux workstation (Intel Core i7-1265U; 32 GB RAM). The end-to-end pipeline (preprocessing, training, validation, interpretability, logging) was modular and version-tracked to ensure reproducibility.

4. Results

4.1. Model performance across time horizons (Global Trends)

Across all time windows 12, 24, 48, and 72 h—ensemble-based classifiers consistently achieved the highest classification performance across both Local and National configurations (Fig. 7). XGB, RF, and GBC showed average scores (mean of Accuracy, F1-score, and MCC) exceeding 0.73 across both Local and National configurations. XGB yielded the best results at the 24-hour and 48-hour windows (peak scores 0.75 and 0.74, respectively). Model rankings remained stable across horizons: ensemble models outperformed simpler baselines such as NB (below 0.42), while LR, DT, ADA, and LGB achieved mid-range performance, showing limited variation across regions or time windows. Performance stabilized at 24 h and beyond, suggesting that mid-range windows maintain predictive accuracy and offer a practical balance between accuracy and timeliness.

4.2. XGB classification performance across regions

Fig. 8 summarizes XGB performance across the five operational regions and four time windows (12, 24, 48, 72 h). The left panel's report absolute mean Accuracy, F1-score, and MCC (averaged over five synchronized CV folds). The right panel's report Δ , enabling a direct read of regional gains over the national baseline. Across all horizons, the Middle region exhibited the highest composite scores (≈ 0.75 – 0.78), followed by North and South. Local and National models produced nearly identical values in the West and East, with minimal variation across metrics. Scores were highest at 24 and 48 h and slightly lower at 12 and 72 h. The Δ panels indicate positive regional gains in most settings, with small negative Δ limited to West in the 12 h and 24 h windows. The largest Δ values occurred in Middle and North (up to $+0.03$ in Accuracy and F1-score at 24–48 h). MCC trends mirrored Accuracy and F1-score, with Middle and North showed the highest absolute values and the largest positive Δ ; West exhibited the least variation across windows.



Fig. 7. Model-wise average performance (mean of Accuracy, F1-score, and MCC) across different horizons (12 h, 24 h, 48 h, and 72 h) for Local and National models. **Model abbreviations:** RF = Random Forest, XGB = Extreme Gradient Boosting, GBC = Gradient Boosting Classifier, ADA = AdaBoost, DT = Decision Tree, LGB = LightGBM, NB = Naive Bayes, LR = Logistic Regression.



Fig. 8. Performance of regional and national XGBoost models across regions and four time windows (12 h, 24 h, 48 h, 72 h). Left panels: absolute metrics (Accuracy, F1-score, MCC). Right panels: differences Δ = Local – National for the same metrics. Bars indicate means across. **Note:** A positive Δ indicates the Local model performed better than the National model. L = local; N = National; MCC = Matthews Correlation Coefficient.

4.3. Detailed model performance by region and time window

Table 3 reports Local / National / Δ (Local – National) for Accuracy, Precision, Recall, F1-score, and MCC across 12, 24, 48, 72 h and all five regions. Precision, Recall, and F1-score are reported for the positive class (climatic failures); Accuracy and MCC are computed across all classes. Across most configurations, Local models achieved higher or equal F1-scores than their National counterparts, with the advantage most pronounced at the 24 h and 48 h. For the Middle region, F1-score was 63.0 vs 61.9 at 24 h and 63.8 vs 62.1 at 48 h (Local vs National). These gains correspond to Δ F1 of + 1.1 (24 h) and + 1.7 (48 h). Similar patterns were observed in the North and West regions. These results motivated the use of the 24 h horizon for the later SHAP analysis.

4.4. Model interpretability via SHAP analysis

To examine the contribution, SHAP was applied to the trained XGB classifiers using the 24 h inputs. SHAP values were computed within each CV fold and aggregated to summarize feature importance for national and regional models. Top features were visualized using beeswarm plots (distribution across instances) and bar plots (mean absolute contribution). **Error! Reference source not found.** Fig. 9 presents SHAP results for the national model (panels a–b) and each regional model (panels c–l). In the national model, the features with the highest mean SHAP values were Repair, Temperature min, Month, Temperature Change, and Snow sum. The regional models showed overlap with the national configuration but also revealed differences in ranking and magnitude. Across all models, Repair and Temperature min were consistently among the top two features while Temperature Change, Snow sum, Month, and Railway class varied by region.

5. Discussion

This study examined whether regionalized ML models improve the classification of climate-related railway failures compared to a single national model. The results support both hypotheses as region-specific models consistently achieved better classification

Table 3

XGBoost performance by region and time window (12 h, 24 h, 48 h, 72 h). Metrics are reported as Local / National with Δ = Local – National. Accuracy and MCC are computed over both classes; Precision, Recall, and F1-score refer to the climate-related class. Values are means across synchronized 5-fold CV. Boldface (or shading) indicates $\Delta > 0$.

Region	Time (h)	Accuracy		Precision		Recall		F1-score		MCC	
		(L/N)	Δ	(L/N)	Δ	(L/N)	Δ	(L/N)	Δ	(L/N)	Δ
Middle	12	79.3/77.5	1.8	54.6/51.5	3.1	71.5/77.4	-5.9	61.9/61.8	0.1	48.8/48.7	0.1
	24	79.9/77.6	2.3	55.7/51.6	4.1	72.5/77.6	-5.1	63.0/61.9	1.1	50.4/48.9	1.5
	48	80.5/77.9	2.6	56.7/52.5	4.2	72.8/76.9	-4.1	63.8/62.1	1.7	50.8/47.7	3.1
	72	80.4/78.2	2.2	56.7/52.5	4.2	71.6/77.2	-5.6	63.3/62.5	0.8	50.8/49.7	1.1
North	12	75.9/73.8	2.1	60.0/57.1	2.9	82.4/85.1	-2.7	69.4/68.4	1.0	52.1/50.3	1.8
	24	76.1/74.0	2.1	60.2/57.2	3.0	83.1/85.3	-2.2	69.7/68.6	1.1	52.6/50.8	1.8
	48	76.2/74.1	2.1	60.2/57.3	2.9	83.5/85.4	-1.9	69.9/68.7	1.2	52.9/50.9	2.0
	72	76.6/74.2	2.4	60.7/57.4	3.3	83.8/85.6	-1.8	70.0/68.7	1.3	53.3/50.9	2.4
South	12	88.9/88.3	0.6	54.1/51.8	2.3	71.3/78.9	-7.6	56.1/54.9	1.2	49.8/48.3	1.5
	24	89.1/88.8	0.3	55.0/53.7	1.3	71.5/79.1	-7.6	56.3/56.3	0.0	50.1/50.0	0.1
	48	89.5/89.2	0.3	56.5/55.0	1.5	72.1/79.7	-7.6	57.4/56.2	1.2	51.4/50.0	1.4
	72	89.2/88.8	0.4	55.7/54.2	1.5	72.1/79.7	-7.6	56.6/55.1	1.5	50.4/48.7	1.7
West	12	87.3/86.5	0.8	51.7/49.0	2.7	68.9/77.7	-8.8	53.5/54.8	-1.3	46.2/47.4	-1.2
	24	87.3/86.1	1.2	53.6/50.5	3.1	69.8/78.6	-8.8	55.2/55.7	-0.5	48.3/48.5	-0.2
	48	87.9/87.0	0.9	56.4/50.3	6.1	70.2/79.0	-8.8	55.0/55.4	-0.4	48.1/48.2	-0.1
	72	88.0/87.0	1.0	56.5/50.2	6.3	70.2/79.0	-8.8	55.2/55.0	0.2	48.2/47.7	0.5
East	12	87.7/85.8	1.9	62.3/46.1	16.2	71.5/77.4	-5.9	56.5/55.1	1.4	49.7/48.3	1.4
	24	87.8/86.2	1.6	52.0/47.4	4.6	71.5/79.1	-7.6	56.6/56.3	0.3	49.9/49.7	0.2
	48	88.3/86.7	1.6	53.4/48.1	5.3	72.1/79.7	-7.6	57.9/56.5	1.4	51.4/49.9	1.5
	72	88.1/86.5	1.6	53.0/48.0	5.0	72.1/79.7	-7.6	57.3/55.8	1.5	50.7/49.1	1.6

L=local; N= National; MCC= Matthews Correlation Coefficient.

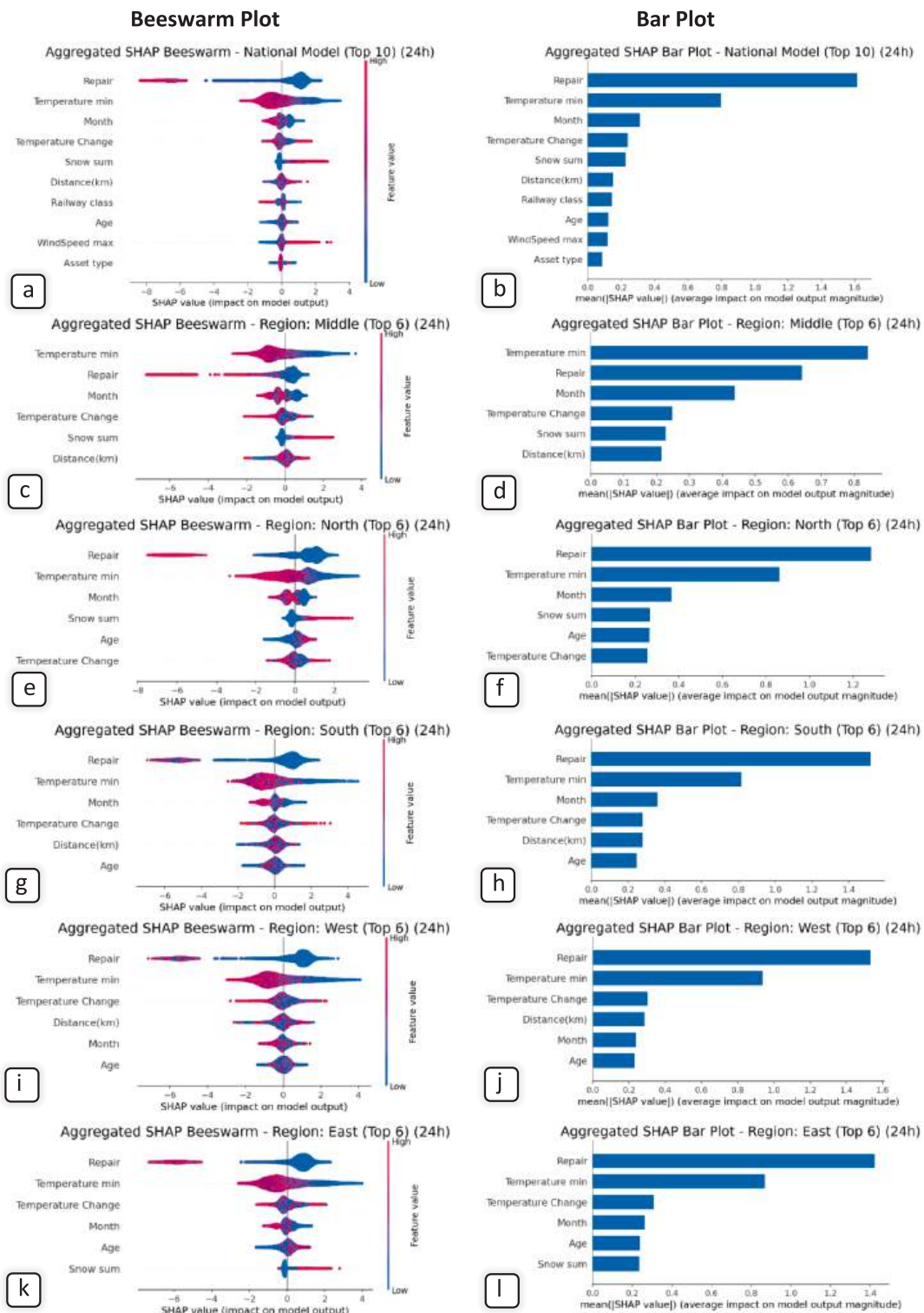


Fig. 9. SHAP Summary Plots for National and Regional Models (24 h Forecast). **Note:** Panels (a–b): national model (top 10 features). Panels (c–l): regional models (Middle, North, South, West, East; top 6 features each). Beeswarm plots (left column) show the distribution and direction of individual SHAP values across instances (blue = low, pink = high). Bar plots (right column) summarize the mean absolute SHAP value per feature across five CV folds. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

performance and revealed geographically distinct patterns of predictive features.

5.1. Region-Specific models outperform national baseline (H1)

To examine H1, we assessed whether ML models designed for individual regions perform better than a single national model trained on aggregated data across all regions. This was confirmed across selected regions and time windows. For instance, within the 24 h time window the best balance between data reliability and sample coverage models achieved improvements of around 2 % in Accuracy, F1-score, and MCC. Although these numerical margins are small, they represent meaningful improvements given the scale and structure of the dataset. On this dataset, a 1 % gain corresponds to $\sim 1,500$ additional correctly classified failures. This impact is magnified when considering the distribution of climate-related failures. For example, the North and Middle regions show notably high proportions of climate-related failures 33 % and 24 %, respectively compared to less than 13 % in other regions. In these zones, improvements of 1.5–2.3 % in Accuracy and MCC translate to several hundred additional correctly classified climate-related disruptions. The consistent improvement in MCC across regional models confirms that a trade-off between precision and recall did not accompany performance gains. These results suggest that regionally adapted models provide practical benefits in terms of classification performance and operational relevance.

On computing, each regional model trains on fewer records than the national model, so single runs are lighter. The total runtime is higher only because we train several regions and run five-fold CV and SHAP for each region. With our automated pipeline and saved models, this added cost is modest. SHAP also points to where and why failures occur (e.g., snow in the North; sparser weather coverage in the West), which helps target maintenance and data improvements. A practical rule is to use regional models where the weather risk is high and keep the national model where gains are small.

5.2. SHAP analysis Reveals distinct feature contributions across regions (H2)

The SHAP analysis provided strong support for H2 by revealing that the contributing factors to failure vary considerably across the regions. While the national model emphasized general factors influencing failure across the entire network such as Repair and Minimum Temperature the regional models highlighted region-specific features that reflect local climatic and operational conditions.

For example, Snow sum was most influential in the North, consistent with winter-related disruptions; in the South and East, Temperature change and Month were more dominant, indicating sensitivity to seasonal transitions; and in the West, the prominence of Distance (km) suggests weather-station coverage effects rather than a direct environmental mechanism. In the East and West, regional and national models perform similarly, likely because conditions are close to the network average and a few general predictors dominate.

In the Middle region, feature rankings closely resembled the national model with Repair, Temperature min, and Month among the top features which underscores the connection between national and regional patterns where conditions align. We note that SHAP explains associations, not causality; operational actions should be paired with domain evidence and monitoring.

Unlike prior modeling efforts that relied on aggregated or national-level ML without explainability (Chen et al., 2021; Lessan et al., 2019), a few recent studies have employed XAI in transport contexts (e.g., (Sariyer et al., 2024)). In contrast, our framework systematically compares feature importance across the national model and five regional models, highlighting location-specific drivers such as snow accumulation in the North, seasonal variability in the South and East, and data coverage issues in the West. Additionally, SHAP is embedded within a synchronized CV pipeline, supporting robust and reproducible interpretation with operational value for infrastructure planning and climate adaptation.

5.3. Implications for infrastructure management

The findings of this study suggest that national agencies and infrastructure managers would benefit from adopting regional modeling strategies to enhance maintenance planning, operational response, and long-term adaptation efforts. Regional classification enables more accurate identification of climate-related failures, supporting inspections and maintenance scheduling while improving preparedness in terms of resources, equipment, and costs. This, in turn, provides valuable input for predictive maintenance, thereby enhancing the resilience and safety of the railway infrastructure network. In addition, such classification provides infrastructure managers with a clear understanding of how different types of climate-related failures impact infrastructure and to what extent, thereby supporting robust climate adaptation actions and helping to avoid maladaptation. These operational benefits follow from the observed performance gains (Fig. 8; Table 3), enabling more efficient deployment of inspection resources, tailored maintenance scheduling, and targeted infrastructure upgrades.

In particular, the use of XAI (e.g., SHAP) enhances transparency in model outputs and supports data-informed decision-making. Rather than relying solely on predictive accuracy, infrastructure managers can gain insight into the underlying drivers of failure within each region. This supports more customized interventions whether that involves bolstering winter operations in the North, monitoring thermal variability in the South and East, or improving sensor coverage in the West.

Predictions of climate-related failures can also feed downstream response planning, such as passenger-flow control and bus-bridging when disruptions occur. These outcomes reinforce the practical value of regional modeling frameworks for transportation systems operating under varied environmental conditions. Beyond predictive improvement, this approach supports evidence-based adaptation strategies aligned with geographic priorities and resource constraints.

Beyond Sweden, the framework is general and can be applied in other railway systems if a few steps are taken. First, map local

failure codes to a common set (asset class and a binary climate/non-climate label). Second, align weather inputs to fixed horizons and units and match them in space and time to assets. Third, retrain and validate with the same CV on local data. Where reporting schemes differ, a simple code crosswalk helps; where weather stations are sparse, gridded products can be used with care. Because maintenance practice varies, action thresholds should be set with local rules in mind.

5.4. Limitations and Future work

This study validates the proposed method using railway infrastructure failure data and does not capture rolling-stock effects directly. Future work should incorporate covariates such as track-bed condition, train-passage frequency, axle load, and wheel–rail interaction forces so that asset condition and traffic loads are represented alongside weather.

Model output depends on reasonable feature-engineering choices. We used consistent time-windows (12, 24, 48, 72 h) and fitted all transforms within training folds to avoid leakage, yet feature attributions may vary with different window definitions, aggregation rules, or preprocessing steps. Future work should assess the stability of feature effects under small, controlled variations of these settings while keeping the comparison rules unchanged.

The labeling of Climate-ID combined expert review with linked databases. Although this procedure reduced noise, some subjectivity may remain. Future iterations should include auditable labeling protocols and inter-rater agreement checks to strengthen traceability and reproducibility.

Weather station coverage is another potential limitation. In some regions, sparse or uneven station distribution may reduce how well local weather conditions are represented and influence model outputs. Although this was most evident in the West, similar effects could occur elsewhere. Future work should examine alternative data sources (e.g., higher-resolution gridded climate data) to mitigate this issue.

Applying the framework in other countries need minor adjustments in failure codes and maintenance actions and their classification. These gaps can lower accuracy if used without local checks. To reduce risk, build a simple code crosswalk, run an external or cross-site validation step, and use a small local calibration set to adjust thresholds and confirm feature effects before deployment.

Disaggregating the national dataset to regions introduces smaller sample sizes and a higher risk of overfitting. We therefore used a synchronized five-fold CV framework, which preserves class ratios within each region and evaluates national and regional models on identical partitions; all preprocessing (scaling, imputation, encoding) was fitted on the training data within each fold. Leave-one-year-out and single holdout-years were considered in pilots, but several regions year subsets contained too few climatic-failure for stable estimates. Targeted assessments of unseen years or external regions are appropriate when fold-level positive counts are sufficient or when suitable external periods are available.

Class imbalance remains a challenge, especially in regions and years with few climate-related failures. We used class weighting during training to maintain the empirical distribution and enable fair national–regional comparison. SMOTE was not adopted because some training folds had too few positive cases for reliable synthetic sampling. Future work in data-rich settings could test ensemble balancing or anomaly-detection approaches.

Finally, the analysis is retrospective. To support adaptation planning, the framework should be coupled with RCP climate projections by constructing scenario-conditioned inputs and refitting over extended periods. This would enable estimation of shifts in failure probabilities for planning horizons (e.g., 2030, 2050, 2100) and provide forward-looking evidence for infrastructure management.

6. Conclusion

This study challenges the assumption that nationally aggregated machine learning models are sufficient for classifying climate-related railway failures. By implementing a synchronized five-fold CV framework and applying SHAP-based interpretability within XAI, we demonstrate that region-specific models consistently outperform a national baseline in both predictive performance and interpretability.

These improvements, although modest in percentage terms, scale to system level, translating into thousands of additional correct classifications when applied to large-scale infrastructure datasets. More importantly, the SHAP analysis uncovered distinct regional patterns such as the role of snow accumulation in the North and data coverage limitations in the West that are obscured when using a single national model.

The findings suggest that region-adapted modeling is a necessary step toward actionable and locally responsive infrastructure planning. This perspective supports a transition in transportation system management from generalized, reactive strategies to tailored, proactive, and climate-aware interventions that enhance operational resilience.

By providing a clear regional modeling framework, the study offers tools for geographically aware planning and more precise responses to local failure conditions, improving network management and service continuity and contributing to the long-term resilience of railway systems. The integration of ML and SHAP supports climate-risk assessment and adaptation planning under evolving conditions.

CRedit authorship contribution statement

Abdolamir Karbalaie: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Khosro Soleimani-Chamkhorami:** Writing – review & editing, Methodology, Formal

analysis, Data curation, Conceptualization. **Stephen M Famurewa**: Writing – review & editing, Validation, Data curation. **A.H.S Garmabaki**: Writing – review & editing, Writing – original draft, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: A.H.S Garmabaki reports financial support was provided by Swedish Research Council for Sustainable Development, FORMAS. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge the funding provided by the Swedish Research Council for Sustainable Development, FORMAS, to the project titled “Climate Adaptation and Risk Mitigation of Swedish Railway Infrastructure (AdaptRail with grant number, Garmabaki- 2022-00835).

Appendix A. . Formal definition of the synchronized Five-Fold Cross-Validation strategy

This appendix provides the formal mathematical description of the synchronized five-fold cross-validation strategy applied in this study. The formalism complements the conceptual explanation in [Section 3.2.1](#) of the main text.

A.1 Dataset partitioning

Let the complete dataset be denoted as:

$$D = \bigcup_{r=1}^5 D_r$$

where D_r represents the dataset from region $r \in \{\text{North, Middle, East, West, South}\}$.

Each regional dataset D_r was partitioned into five mutually exclusive and collectively exhaustive folds:

$$D_r = \bigcup_{i=1}^5 D_r^{(i)}, D_r^{(i)} \cap D_r^{(j)} = \emptyset \text{ for } i \neq j$$

A.2 Training and test sets per fold

At cross-validation iteration $j \in \{1, 2, 3, 4, 5\}$:

- The test set for region r is:

$$T_r^{(j)} = D_r^{(j)}$$

- The training set for region r is:

$$S_r^{(j)} = \bigcup_{i \neq j} D_r^{(i)}$$

A.3 National model splits

For the national model, the training and test sets were constructed by pooling across all regions:

$$S_{\text{national}}^{(j)} = \bigcup_{r=1}^5 S_r^{(j)}, T_{\text{national}}^{(j)} = \bigcup_{r=1}^5 T_r^{(j)}$$

This synchronized structure ensured that the national and regional models were always trained and tested on identical data partitions, enabling direct and fair comparisons.

A.4 Stratification

To address the inherent class imbalance between climatic and non-climatic failures, folds were generated using stratified sampling,

preserving class proportions within each fold for every region. This design ensured balanced representation of both classes across training and test data.

Data availability

The authors do not have permission to share data.

References

- Adams, K., Heidarzadeh, M., 2021. A multi-hazard risk model with cascading failure pathways for the Dawlish (UK) railway using historical and contemporary data. *International Journal of Disaster Risk Reduction* 56, 102082.
- Andersson, Y., Thorsson, S., Rayner, D., Lindberg, F., Janhäll, S., Jonsson, A., Moback, U., Bergman, R., Granberg, M., 2015. An integrated method for assessing climate-related risks and adaptation alternatives in urban areas. *Climate Risk Management* 7, 31–50.
- Argyroudis, S.A., Mitoulis, S.A., Chatzi, E., Baker, J.W., Brilakis, I., Gkoumas, K., Voudoukas, M., Hynes, W., Carluccio, S., Keou, O., Frangopol, D.M., Linkov, I., 2022. Digital technologies can enhance climate resilience of critical infrastructure. *Climate Risk Management* 35, 100387. <https://doi.org/10.1016/j.crm.2021.100387>.
- Argyroudis, S.A., Mitoulis, S.A., Winter, M.G., Kaynia, A.M., 2019. Fragility of transport assets exposed to multiple hazards: State-of-the-art review toward infrastructure resilience. *Reliability Engineering & System Safety* 191, 106567. <https://doi.org/10.1016/j.res.2019.106567>.
- Bacić, T. (2024). *Malmöbanan iron ore railway reopens after derailments*. Retrieved 13 March 2024 from <https://www.railwaygazette.com/infrastructure/malmbanan-iron-ore-railwayreopens-after-derailments/66095.article>.
- [AuthorError] et al., 2016. Binti Sa'adin, S. L., Kaewunruen, S., & Jaroszweski, D., 2016. Heavy rainfall and flood vulnerability of Singapore-Malaysia high speed rail system [Article]. *Australian Journal of Civil Engineering* 14 (2), 123–131. <https://doi.org/10.1080/14488353.2017.1336895>.
- Bubeck, P., Dillenardt, L., Alfieri, L., Feyen, L., Thieken, A.H., Kellermann, P., 2019. Global warming to increase flood risk on European railways [Article]. *Climatic Change* 155 (1), 19–36. <https://doi.org/10.1007/s10584-019-02434-5>.
- Chen, Z., Wang, Y., Zhou, L., 2021. Predicting weather-induced delays of high-speed rail and aviation in China. *Transport policy* 101, 1–13. *Climate-ADAPT*. <https://climate-adapt.eea.europa.eu/>.
- Davies, B., Andrews, J., 2021. The impact of summer heatwaves on railway track geometry maintenance [Article]. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit* 235 (9), 1158–1171. <https://doi.org/10.1177/0954409720984292>.
- Dobney, K., Baker, C.J., Chapman, L., Quinn, A.D., 2010. The future cost to the United Kingdom's railway network of heat-related delays and buckles caused by the predicted increase in high summer temperatures owing to climate change [Article]. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit* 224 (1), 25–34. <https://doi.org/10.1243/09544097JRR292>.
- Forzieri, G., Bianchi, A., e Silva, F. B., Herrera, M. A. M., Leblois, A., Lavalle, C., Aerts, J. C., & Feyen, L., 2018. Escalating impacts of climate extremes on critical infrastructures in Europe. *Global environmental change* 48, 97–107.
- Fu, Q., Easton, J.M., 2018. Prediction of weather-related incidents on the rail network: prototype data model for wind-related delays in Great Britain. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* 4 (3), 04018027.
- Gao, J., 2016. Analysis and assessment of the risk of snow and freezing disaster in China. *International Journal of Disaster Risk Reduction* 19, 334–340.
- García-Méndez, S., de Arriba-Pérez, F., Leal, F., Veloso, B., Malheiro, B., Burguillo-Rial, J.C., 2025. An explainable machine learning framework for railway predictive maintenance using data streams from the metro operator of Portugal. *Scientific Reports* 15 (1), 27495.
- Garmabaki, A., Thaduri, A., Famurewa, S., Kumar, U., 2021. Adapting Railway Maintenance to Climate Change. *Sustainability* 13 (24), 13856.
- Garmabaki, A.H.S., Masoud, N., Odelius, J., Asplund, M., Famurewa, S.M., Barabady, J., Strandberg, G., 2023. *Risk assessment of climate change on Railway infrastructure asset*. 7th International Congress and Workshop on Industrial AI and eMaintenance.
- Haghighi, E., Kasraei, A., Famurewa, S., Strandberg, G., Sas, G., Garmabaki, A.H.S., 2025. Climate change risks on railway infrastructure: A systematic review and analysis. *Sustainable Cities and Society* 129, 106504. <https://doi.org/10.1016/j.scs.2025.106504>.
- Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2. Springer.
- Hong, S., Park, C., Cho, S., 2021. A rail-temperature-prediction model based on machine learning: warning of train-speed restrictions using weather forecasting. *Sensors* 21 (13), 4606.
- Imam, B., 2019. Climate change impact for bridges subjected to scour and corrosion. In: *Climate adaptation engineering*. Elsevier, pp. 165–206.
- IPCC6, Pörtner, H. O., Roberts, D. C., Adams, H., Adler, C., Aldunce, P., Ali, E., Begum, R. A., Betts, R., Kerr, R. B., & Biesbroek, R. (2022). *Climate change 2022: impacts, adaptation and vulnerability*. <https://www.ipcc.ch/report/ar6/wg2/>.
- Jiang, H., Bian, X., Chen, Y., Han, J., 2015. Impact of water level rise on the behaviors of railway track structure and substructure: Full-scale experimental investigation [Article]. *Transportation Research Record* 2476, 15–22. <https://doi.org/10.3141/2476-03>.
- Kalantari, Z., Cavalli, M., Cantone, C., Crema, S., Destouni, G., 2017. Flood probability quantification for road infrastructure: Data-driven spatial-statistical approach and case study applications [Article]. *Science of the Total Environment* 581–582, 386–398. <https://doi.org/10.1016/j.scitotenv.2016.12.147>.
- Kasraei, A., Garmabaki, A.H.S., Odelius, J., Famurewa, S.M., Chamkhorami, K.S., Strandberg, G., 2024. Climate change impacts assessment on railway infrastructure in urban environments. *Sustainable Cities and Society* 101, 105084.
- Kellermann, P., Schönberger, C., Thieken, A.H., 2016. Large-scale application of the flood damage model Railway Infrastructure Loss (RAIL) [Article]. *Natural Hazards and Earth System Sciences* 16 (11), 2357–2371. <https://doi.org/10.5194/nhess-16-2357-2016>.
- Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*.
- Koks, E.E., Rozenberg, J., Zorn, C., Tariverdi, M., Voudoukas, M., Fraser, S.A., Hall, J., Hallegatte, S., 2019. A global multi-hazard risk analysis of road and railway infrastructure assets. *Nature communications* 10 (1), 2677.
- Kristina Gustafsson, R. J. o. M. S. (2021). *Trafikverkets forsknings- och innovationsplan För åren 2021-2026*. Trafikverket Retrieved from <http://trafikverket.diva-portal.org/smash/record.jsf?pid=diva2%3A1516095&dsid=8343>.
- Lessan, J., Fu, L., Wen, C., 2019. A hybrid Bayesian network model for predicting delays in train operations. *Computers & Industrial Engineering* 127, 1214–1222.
- Li, Q., Punzo, G., Robson, C., Arbabi, H., Mayfield, M., 2023. A Systematic Approach to Climate Resilience Assessment of Infrastructure Networks [Article]. *IEEE Systems Journal* 1–12. <https://doi.org/10.1109/JSYST.2023.3329765>.
- Liu, H., Jiang, H., Zhao, C., Bian, X., 2022. Long-term responses of high-speed railway subjected to extreme precipitation events [Article]. *Transportation Geotechnics* 37, 100852. <https://doi.org/10.1016/j.trge.2022.100852>.
- Liu, K., Wang, M., Cao, Y., Zhu, W., Wu, J., Yan, X., 2018. A comprehensive risk analysis of transportation networks affected by rainfall-induced multihazards. *Risk Analysis* 38 (8), 1618–1633.
- Liu, K., Wang, M., Zhou, T., 2021. Increasing costs to Chinese railway infrastructure by extreme precipitation in a warmer world. *Transportation Research Part D: Transport and Environment* 93, 102797.
- Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: *Advances in neural information processing systems*, p. 30.
- Manning, L.J., Hall, J.W., Kilsby, C.G., Glendinning, S., Anderson, M.G., 2008. Spatial analysis of the reliability of transport networks subject to rainfall-induced landslides [Article]. *Hydrological Processes* 22 (17), 3349–3360. <https://doi.org/10.1002/hyp.6927>.

- Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein. Structure* 405 (2), 442–451. [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9).
- Matthias, S., Christoph, M., 2018. Potential future exposure of European land transport infrastructure to rainfall-induced landslides throughout the 21st century [Article]. *Natural Hazards and Earth System Sciences* 18 (4), 1121–1132. <https://doi.org/10.5194/nhess-18-1121-2018>.
- Misnev, B., Melikyan, A., Bazaras, D., 2015. Hazard Assessment of Weather Factors for the Occurrence of an Emergency on the Railway. *Procedia Computer Science* 77, 40–47.
- Mulholland, E., Feyen, L., 2021. Increased risk of extreme heat to European roads and railways with global warming. *Climate Risk Management* 34, 100365.
- Nasr, A., Kjellström, E., Björnsson, I., Honfi, D., Ivanov, O.L., Johansson, J., 2020. Bridges in a changing climate: A study of the potential impacts of climate change on bridges and their possible adaptations. *Structure and Infrastructure Engineering* 16 (4), 738–749.
- Nemry, F., & Demirel, H. (2012). Impacts of Climate Change on Transport: A focus on road and rail transport infrastructures. *European Commission, Joint Research Centre (JRC), Institute for Prospective Technological Studies (IPTS)*.
- Nezval, V., Andrášik, R., Bíl, M., 2023. Impact of storms on rail transport: a case study from Czechia [Article]. *Natural Hazards*. <https://doi.org/10.1007/s11069-023-06321-2>.
- Nguyen, M.N., Wang, X., Wang, C.H., 2012. A reliability assessment of railway track buckling during an extreme heatwave [Article]. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit* 226 (5), 513–517. <https://doi.org/10.1177/0954409712441743>.
- Nyström, U. (2025). *Snödriva orsakade urspårning på Malmaban.* <https://jarnvagar.nu/>. Retrieved 2025-02-07 from Snödriva orsakade urspårning på Malmaban – Järnvägar.nu.
- Ochsner, M., Fisher, R., Palmqvist, C.-W., 2024. The impacts of weather on railway infrastructure in Sweden. *Sustainable and Resilient Infrastructure* 9 (6), 582–598.
- Palin, E.J., Stipanovic Oslakovic, I., Gavin, K., Quinn, A., 2021. Implications of climate change for railway infrastructure. *Wiley Interdisciplinary Reviews: Climate Change* 12 (5), e728.
- Palin, E.J., Thornton, H.E., Mathison, C.T., McCarthy, R.E., Clark, R.T., Dora, J., 2013. Future projections of temperature-related climate change impacts on the railway network of Great Britain [Article]. *Climatic Change* 120 (1–2), 71–93. <https://doi.org/10.1007/s10584-013-0810-8>.
- Roghani, A., 2021. Quantifying the effect of freeze-thaw cycles on track surface deformation and degradation of railway track geometry; Case study [Article]. *Transportation Geotechnics* 30, 100601. <https://doi.org/10.1016/j.trgeo.2021.100601>.
- Saback, V., Gonzalez-Libreros, J., Daescu, C., Popescu, C., Garmabaki, A., Sas, G., 2024. Adapting to climate change: snow load assessment of snow galleries on the Iron Ore Line in Northern Sweden. *Frontiers in Built Environment* 9, 1308401.
- Sariyer, G., Mangla, S.K., Sozen, M.E., Li, G., Kazancoglu, Y., 2024. Leveraging explainable artificial intelligence in understanding public transportation usage rates for sustainable development. *Omega* 127, 103105. <https://doi.org/10.1016/j.omega.2024.103105>.
- Snoussi, M., Ouchani, T., Khouakhi, A., Niang-Diop, I., 2009. Impacts of sea-level rise on the Moroccan coastal zone: Quantifying coastal erosion and flooding in the Tangier Bay [Article]. *Geomorphology* 107 (1–2), 32–40. <https://doi.org/10.1016/j.geomorph.2006.07.043>.
- Soleimani, K., Garmabaki, A., Kasraei, A., Famurewa, S.M., Odelius, J., Strandberg, G., 2024a. Life cycle cost assessment of railways infrastructure asset under climate change impacts. *Transportation Research Part D: Transport and Environment* 127, 104072.
- Soleimani, K., Karbalaie, A., Kasraei, A., Haghighi, E., Famurewa, S.M., Garmabaki, A.H.S., 2024b. Identifying climate-related failures in railway infrastructure using machine learning. *Transportation Research Part D: Transport and Environment* 135, 104371. <https://doi.org/10.1016/j.trd.2024.104371>.
- Szymczak, S., Bott, F., Babeck, P., Frick, A., Stöckigt, B., Wagner, K., 2022. Estimating the hazard of tree fall along railway lines: a new GIS tool. *Natural hazards* 112 (3), 2237–2258.
- Tang, R., De Donato, L., Besinović, N., Flammini, F., Goverde, R.M., Lin, Z., Liu, R., Tang, T., Vittorini, V., Wang, Z., 2022. A literature review of Artificial Intelligence applications in railway systems. *Transportation Research Part C: Emerging Technologies* 140, 103679.
- Thaduri, A., Garmabaki, A., & Kumar, U. (2021). Impact of climate change on railway operation and maintenance in Sweden: A State-of-the-art review. *Maintenance, Reliability and Condition Monitoring (MRCM)*.
- Tian, X., Lu, C., Song, Z., An, C., Wan, S., Peng, H., Feng, Q., Chen, Z., 2024. Quantifying weather-induced unreliable public transportation service in cold regions under future climate model scenarios. *Sustainable Cities and Society* 113, 105660.
- Trafikverket. (2000-2024). *Ofelia-Failure reporting system* (<https://bransch.trafikverket.se/tjanster/system-och-verktyg/forvaltning-och-underhall/Ofelia/>).
- Wang, Z., Ding, Y., Zhou, B., Chen, L., 2020. Comparison of two severe low-temperature snowstorm and ice freezing events in China: Role of Eurasian mid-high latitude circulation patterns. *International Journal of Climatology* 40 (7), 3436–3450.
- Wu, C., Ochieng, W., Pien, K.-C., Shang, W.-L., 2025. Carbon-efficient timetable optimization for urban railway systems considering wind power consumption. *Applied Energy* 388, 125593.
- Zhao, J., Liu, K., Wang, M., 2020. Exposure analysis of Chinese railways to multihazards based on datasets from 2000 to 2016. *Geomatics, Natural Hazards and Risk* 11 (1), 272–287.
- Zhou, Y., Yue, Y., Gao, Z., & Zhou, Y. (2017). Climatic characteristics and determination method for freezing rain in China. *Advances in Meteorology*, 2017.