PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Research

Article submitted to journal

Subject Areas:
machine learning, transportation engineering, predictive modeling, applied statistics

Keywords:
train delay prediction, uncertainty quantification, conformal prediction

Author for correspondence:
Rui Luo
e-mail: ruiluo@cityu.edu.hk

Uncertainty Quantification in Train Delay Prediction with Conformal Prediction

Rui Luo[1*], Xiaoyi Su[1*] and Khuong An Nguyen[2]

[1]Department of Systems Engineering, City University of Hong Kong, Hong Kong SAR, China
[2]Computer Science Department, Royal Holloway University of London, Surrey, United Kingdom
∗ Rui Luo and Xiaoyi Su are joint first authors on this paper.

Predicting train delays is crucial for railway operations and passenger experience, but point predictions fail to capture uncertainty, limiting their use in risk-aware decisions. Existing uncertainty quantification (UQ) methods often rely on unverifiable assumptions and produce miscalibrated intervals. This paper evaluates conformal prediction (CP) as a distribution-free framework for generating prediction intervals with rigorous coverage guarantees. Using a large-scale dataset from Southeastern Railway in the UK, we show that UQ methods such as quantile regression, Monte Carlo Dropout, and Deep Ensembles frequently under- or over-cover nominal levels. In contrast, CP corrects miscalibration across models, ensuring valid marginal coverage. Conformalized quantile regression (CQR) achieves the best efficiency by producing adaptively sized intervals while maintaining calibration. To address heterogeneity across stations, we apply Mondrian conformal prediction (MCP), which enforces conditional coverage within strata. Empirical results confirm MCP delivers reliable intervals for each subgroup. Our work demonstrates that CP is a model-agnostic, robust framework for trustworthy uncertainty quantification, offering a practical pathway to reliable risk assessment in transportation and other high-stakes domains.

1. Introduction

The punctuality of public transit is a cornerstone of urban efficiency and passenger satisfaction. For

THE ROYAL SOCIETY PUBLISHING

railway networks as the backbone of transportation systems, adherence to meticulously planned schedules is paramount for operational integrity and service quality [1,2]. Deviations from these schedules, manifesting as delays, can cascade through the network, disrupting passenger journeys, eroding public trust, and even incurring significant economic costs for operators [3,4]. The accurate prediction of train delays is therefore not merely a matter of convenience but a critical function for dynamic resource allocation, proactive incident management, and transparent passenger communication. However, the inherent stochasticity of railway operations which is driven by various factors such as adverse weather, infrastructure faults, and fluctuating passenger loads, renders delay prediction a formidable and complex scientific challenge [5–7].

While numerous studies have developed models for predicting train arrival times, often yielding a single point estimate (e.g., [8–10]), such deterministic forecasts are fundamentally misaligned with the stochastic nature of delays. A single value, however accurate on average, fails to convey the range of plausible outcomes and the associated risks. For both passengers and operators, decision-making is often more informed by an understanding of this uncertainty. A passenger might alter their travel plans if there is a high probability of a significant delay, while an operator might preemptively re-route services based on a predicted interval of disruption. Consequently, the field is progressively moving beyond point prediction towards uncertainty quantification (UQ), which aims to provide a prediction interval, a range of likely delay times that is of far greater practical value for risk-aware decision-making [11,12].

Existing approaches to UQ in transportation science, however, often suffer from a critical limitation: they lack formal, distribution-free guarantees on their performance. Methods ranging from quantile regression (QR) to Bayesian deep learning produce intervals whose validity is contingent upon unverifiable assumptions about the data distribution or model structure. Consequently, their empirical coverage can silently deviate from the nominal level, rendering them untrustworthy for high-stakes operational decisions. This highlights a foundational gap between current UQ practices and the need for truly reliable risk assessment tools.

This paper systematically introduces and evaluates the conformal prediction (CP) framework as a robust, general-purpose, and distribution-free solution to this challenge [13–16]. Unlike conventional UQ techniques, CP provides rigorous, non-asymptotic guarantees on prediction interval coverage, regardless of the underlying data distribution or the complexity of the predictive model. It operates as a wrapper that transforms the outputs of other machine learning models into theoretically sound prediction intervals. This study highlights the power of CP as a framework that offers two principal capabilities. First, it can augment any point prediction model, from classic regression to advanced deep neural networks (DNNs) or gradient boosting machines, to produce reliable interval predictions. Second, it can perform *post-hoc calibration* on the outputs of existing interval prediction methods, such as conformalized quantile regression (CQR) [17], correcting their biases and enforcing the desired coverage rate. Furthermore, we address a limitation of standard CP that its coverage guarantee is marginal (i.e., averaged over all data points) leveraging Mondrian conformal prediction (MCP) [13]. This extension enables us to enforce valid coverage conditionally, ensuring that the prediction intervals maintain their stated reliability across predefined strata of the data, such as different railway stations, routes, or times of day. By demonstrating these capabilities on a comprehensive dataset of train operations from Southeastern Railway in the UK, our research provides a pathway for transport system operators to develop a truly trustworthy risk assessment tool.

The remainder of this paper is organized as follows. Section 2 reviews prior research on train delay prediction and the evolution of UQ techniques. Section 3 provides a detailed description of the Southeastern Railway dataset and our feature engineering process for the delay prediction task. In Section 4, we first formally define the prediction problem, then introduce the underlying point and interval prediction models that form the basis of our study, and finally provide a detailed exposition of the conformal prediction framework and its key variants. The experimental design and comparative results are presented in Section 5, followed by a discussion of their implications and directions for future work in Section 6. Finally, Section 7 concludes the paper.

2. Related Work

The challenge of predicting train delays is intrinsically linked to the broader scientific pursuit of modeling complex, dynamic systems. The evolution of methodologies in this domain reflects a clear trajectory: a progressive journey from deterministic point forecasting towards a more nuanced, probabilistic understanding of operational uncertainty. This section traces this evolution, charting the path from foundational statistical models to the frontiers of trustworthy uncertainty quantification, situating the contribution of this paper within its scientific context.

(a) The Pursuit of Accuracy: From Statistical Models to Deep Learning

Initial forays into travel time and delay prediction established the viability of data-driven approaches, supplanting purely theoretical or simulation-based models. This first wave of research predominantly employed classical machine learning and statistical techniques. For instance, support vector regression (SVR) was shown to be effective for general traffic time prediction [18], while various forms of artificial neural networks (ANNs) were adapted for the specific task of railway delay forecasting [19]. Comparative studies from this period sought to identify the optimal algorithm for short-term predictions, with findings often highlighting the superior performance of shallow ANNs and their variants, such as the nonlinear autoregressive model with external inputs (NARX), over traditional linear regression or instance-based methods like k-nearest neighbours (k-NN) [20,21]. Even k-NN, when enhanced, demonstrated competitive performance for bus arrival time [12] and train delay prediction [22]. More recently, ensemble methods, especially gradient boosting algorithms like XGBoost [23] and LightGBM [24], have become standard tools for train delay prediction, offering high accuracy and robustness on structured transit data [8,25–27].

While these foundational models demonstrated that delays were, to a significant extent, predictable, they often reached a performance plateau. Their capacity was limited in capturing the intricate, non-linear, and long-range dependencies inherent in large-scale railway networks. The advent of deep learning marked a paradigm shift, offering a powerful toolkit to model the complex spatiotemporal dynamics that govern network-wide delay propagation. Researchers began to construct sophisticated architectures capable of learning from the confluence of causal text information [28], open data sources [27], network topology, temporal patterns, and exogenous factors. For example, deep learning models have been integrated with interaction networks to explicitly model how delays cascade between connected services [10], and have been designed to leverage rich spatiotemporal features for improved accuracy [11,29]. Innovations such as modular deep neural networks tailored to specific operational clusters [9], fully connected architectures for arrival time estimation [30], and even neural time point processes to model the stochastic timing of events [31], have collectively pushed the boundaries of predictive accuracy. A comprehensive review by Tiong et al. [32] reinforces this trajectory, emphasizing the increasing importance of multi-source data fusion and dynamic, multi-station prediction models as the field matures.

(b) The Paradigm Shift: From Point Predictions to Quantifying Uncertainty

The relentless pursuit of higher accuracy in point predictions, while valuable, obscures a fundamental limitation: a single-value forecast is an incomplete representation of a stochastic future. For operational decision-making under risk, knowing the most likely delay is less useful than knowing the range of plausible delays. This recognition has catalyzed a paradigm shift in transportation science, moving the focus from mere prediction to the more challenging task of uncertainty quantification. The need for robust prediction intervals, even when dealing with multi-source or complex datasets, has become increasingly apparent [33].

Early and important work in this area centred on constructing prediction intervals around the outputs of neural networks. Methodologies such as the delta and Bayesian techniques were among the first to be rigorously explored for quantifying uncertainty in travel time

predictions, providing a probabilistic envelope around the deterministic output [34,35]. This line of inquiry has since diversified, with modern approaches leveraging more sophisticated probabilistic machine learning [36]. Bayesian deep learning, for example through encoder-decoder architectures, has been used to capture model uncertainty [37,38], while econometric models like the generalized autoregressive conditional heteroskedasticity (GARCH) model have been employed to account for the time-varying volatility in delay patterns [39].

However, a critical weakness pervades many of these UQ techniques: their reliability is contingent upon strong, and often unverifiable, assumptions about the data-generating distribution or the correctness of the model specification. Bayesian methods, for instance, require a correctly specified prior and likelihood, and their posterior distributions are only as valid as these assumptions. An incorrect assumption can lead to prediction intervals that are silently miscalibrated, systematically over- or under-confident, rendering them untrustworthy for the high-stakes decisions common in railway operations.

It is in response to this fundamental need for trustworthy, assumption-free uncertainty estimates that conformal prediction has emerged as a powerful alternative [13–15]. Various univariate conformal regression methods have been reviewed and comparatively analyzed, providing a solid foundation for the field [40]. As a distribution-free framework, CP transforms the outputs of any point prediction algorithm into prediction intervals with finite-sample, marginal coverage guarantees, without making any assumptions about the underlying data distribution beyond exchangeability [16]. Recognizing that the intervals produced by standard CP can be inefficiently wide, recent research has focused on refining the methodology. This includes developing conformal prediction specifically for models like random forests [41] and exploring methods for multi-output regression, both exact and approximate, along with unified comparative studies of new conformity scores [42,43]. Additionally, techniques for approximating score-based explanation within conformal regression have been developed [44]. Conformalized quantile regression, for example, elegantly combines the efficiency of quantile regression with the rigorous guarantees of CP to produce narrower, more informative intervals [17]. Further advancements, such as the QUANTRAFFIC framework, demonstrate how post-hoc calibration techniques can be applied to enhance the uncertainty outputs of sophisticated deep learning models in traffic forecasting [45]. This study builds directly upon this lineage, systematically applying and extending the CP framework to provide a truly reliable risk assessment tool for railway operations, addressing the critical need for trustworthy UQ that prior methodologies have struggled to meet.

3. Data and Feature Engineering

(a) Data Description

This study utilizes a comprehensive dataset of train operations from the Southeastern Railway network in the United Kingdom, spanning the period from March 2022 to October 2023. The raw dataset comprises 4,280,181 records, documenting 271,928 unique journey instances across 7,486 distinct service lines (*Headcode*) and 173 stations (*Tiploc*). Each record corresponds to a train's scheduled and actual performance at a specific station, with consecutive records detailing a train's movement between adjacent stations. The dataset contains 15 primary variables, including:

— *Headcode*: A nationally standardized alphanumeric identifier for a specific train service.
— *UnitNumber*: An identifier for the specific trainset (engine) assigned to the service.
— *Tiploc*: A unique national identifier for a specific station.
— *BookedDeparture* and *ActualDeparture*: The scheduled and actual departure times.
— *BookedArrival* and *ActualArrival*: The scheduled and actual arrival times.
— *DwellBooked* and *DwellActual*: The scheduled and actual time a train spends stationary at a station.
— *UntilNextLocationBookedTime* and *UntilNextLocationActualTime*: The scheduled and actual travel times to the subsequent station.

— A set of difference variables (*DepartureDiff*, *ArrivalDiff*, etc.), calculated as the actual time minus the booked time.

All time-related measurements in the dataset are recorded in seconds. To characterize the dataset, we first analyzed its network structure and delay patterns. A directed graph representing the entire network topology was constructed, where nodes are stations and edges represent direct train movements (Figure 1). The distribution of arrival delays across all services reveals a long-tailed pattern, characteristic of complex transportation systems where small delays are common but large disruptions, though infrequent, are significant (Figure 2).

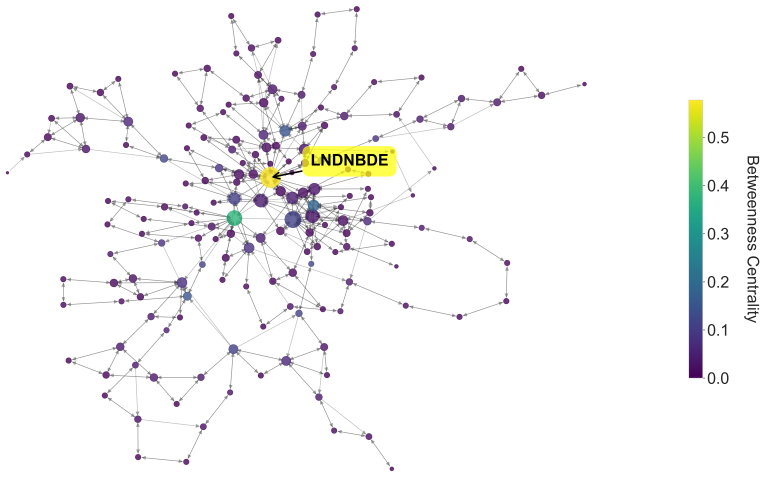

Figure 1. The topological structure of the Southeastern Railway network derived from the dataset. Nodes represent stations (*Tiploc*), and directed edges represent observed train movements between them. Node size reflects degree centrality, while color indicates betweenness centrality (scale on the right). The station *LNDNBDE* (London Bridge), highlighted in yellow, stands out as a critical hub with high both degree and betweenness centrality, serving as a key connector in the network.

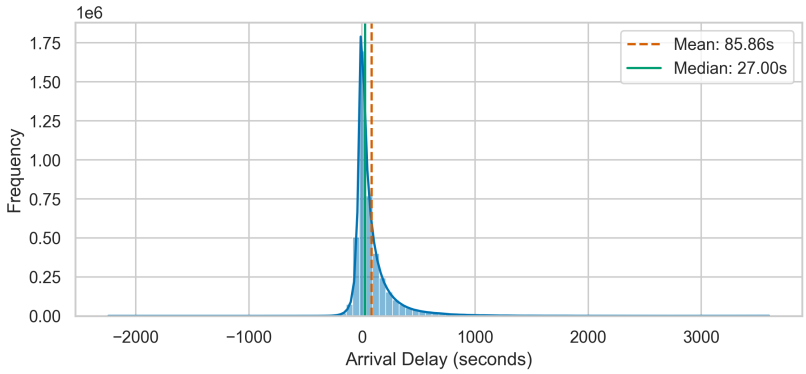

Figure 2. Histogram of arrival delays (*ArrivalDiff*) for all station stops in the dataset. The distribution is right-skewed, indicating that while most trains arrive close to their scheduled time, a lot of services experience substantial delays.

For this study, we adopted a systematic approach to construct a representative cross-sectional dataset. We identified the five busiest stations based on their total traffic frequency (count of stops). These stations are *LNDNBDE* (London Bridge), *CHRX* (London Charing Cross), *WLOE* (London Waterloo), *TONBDG* (Tonbridge), and *SVNOAKS* (Sevenoaks). We then extracted all journey segments that terminate at one of these five key stations. A journey segment is defined as the portion of a train's journey from any upstream station to one of these five terminal stations. This process yielded a rich dataset of 4,310,125 unique segments, each serving as a sample for our prediction task.

(b) Feature Engineering

To develop a robust predictive model, we engineered a comprehensive set of features for each journey segment. These features are designed to capture the multi-faceted dynamics of train operations by transforming raw data into meaningful numerical representations.

The primary objective is to predict the final arrival delay of a train at a target station, given its state at a current, upstream station. The target variable, *ArrivalDiff*, is the difference between the train's actual arrival time and its booked arrival time. The engineered features were systematically constructed and can be grouped into the following categories:

Dynamic and Temporal Features. This group of features quantifies the journey's real-time state and its temporal context. To capture cyclical patterns, we encoded the scheduled departure time into time-of-day categories (e.g., morning, afternoon), day of the week, month, and a weekend flag. The train's immediate performance is described by its arrival delay, departure delay, and dwell time deviation at the current station. To model delay propagation, we also engineered features summarizing the journey's history, including the current stop number, trip completion percentage, the mean and maximum arrival delays experienced so far, the mean dwell time deviation, and the short-term delay trend between the last two stops.

Structural and Inherent Features. Complementing the dynamic attributes, this set of features encodes the static characteristics of the journey's route, network topology, and the trainset itself. High-cardinality identifiers were not used directly but were transformed to extract their underlying information. The journey segment from the current to the target station is characterized by the number of intermediate stops and its total scheduled travel and dwell times. The current station's role in the network was quantified by its in-degree and out-degree from the network graph (Figure 1). Similarly, the service route (*Headcode*) was represented by the total number of stations on its path and the aggregated degrees of these stations. Finally, the trainset (*UnitNumber*) was characterized by extracting its three-character prefix (often denoting train class) and calculating its operational frequency across the dataset as a proxy for its usage pattern.

Finally, all engineered features were transformed into a purely numerical format. One-hot encoding was applied to categorical features with a manageable number of classes: the target station, the time-of-day category, and the extracted *UnitNumber* prefix. To preserve the cyclical nature of temporal features, the day of the week and month were transformed using sine and cosine functions. Specifically, for a cyclic feature x with period T, the transformation is defined as:

$$\text{sin\_encoded}(x) = \sin\left(2\pi \cdot \frac{x}{T}\right), \quad \text{cos\_encoded}(x) = \cos\left(2\pi \cdot \frac{x}{T}\right). \tag{3.1}$$

For the day of the week, $x \in \{1, 2, \ldots, 7\}$ and $T = 7$; for the month, $x \in \{1, 2, \ldots, 12\}$ and $T = 12$. This comprehensive feature engineering and encoding process, summarized in Table 1, resulted in a final feature matrix of size 4,310,125 $\times$ 34, ready for model training and evaluation.

4. Methodology

This section details the methodological framework for generating train delay predictions with rigorous uncertainty guarantees. We first formally define the prediction tasks, then introduce the

Table 1. Summary of engineered features for train delay prediction.

Category	Feature Name	Description
Dynamic & Temporal	Time of Day	One-hot encoded departure time categories (e.g., Morning).
	Day/Month	Cyclic encoding (sin/cos) of day of week and month.
	Weekend Flag	Binary indicator for weekend services.
	Current Delays	Arrival and departure delays at the current station (s).
	Dwell Deviation	Difference between actual and booked dwell time (s).
	Trip Progress	Percentage of journey completed; Current stop number.
	Delay History	Mean/Max arrival delays and mean dwell deviation so far.
	Delay Trend	Change in delay between the last two stops.
Structural & Inherent	Segment Info	Number of intermediate stops; Scheduled travel/dwell times.
	Station Centrality	In-degree and Out-degree of the current station.
	Route Complexity	Total stops on the route; Aggregated route degree.
	Train Unit	Extracted unit class prefix (One-hot); Usage frequency.
	Target Station	One-hot encoded identifier of the destination station.

underlying prediction models, and finally present the conformal prediction framework, including its extensions for enhanced efficiency and conditional validity.

(a) Problem Formulation

Let the data consist of N samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ drawn from an unknown distribution, where $\mathbf{x} \in \mathbb{R}^{34}$ is the feature vector that encapsulates the dynamic, temporal, and structural characteristics detailed in Section 3 and $y \in \mathbb{R}$ is the target arrival delay. Our objective is framed through three progressively sophisticated tasks.

Point Prediction. This task is to learn a mapping function $f : \mathbb{R}^{34} \to \mathbb{R}$ that generates a point estimate $\hat{y} = f(\mathbf{x})$ for the true delay y. The objective is to train a model that approximates the best mapping by minimizing a given loss function (e.g., mean squared error), such that for a new, unseen feature vector $\mathbf{x}$, the prediction $\hat{y}$ is as close as possible to the true outcome y.

Interval Prediction. For risk-aware decision-making, this task is to construct a prediction interval, $C(\mathbf{x})$, that contains the true delay y with a high, user-specified probability. Formally, given a desired nominal coverage level $1 - \alpha$, where $\alpha \in (0, 1)$ is the tolerable miscoverage rate, the primary goal is to construct a function $C(\cdot)$ that generates intervals satisfying the marginal coverage guarantee, i.e., for a new sample pair $(\mathbf{x}, y)$,

$$\mathbb{P}\left(y \in C(\mathbf{x})\right) \geq 1 - \alpha. \tag{4.1}$$

A secondary objective is efficiency, which seeks to minimize the interval length $\mathbb{E}[|C(\mathbf{x})|]$ while maintaining the validity guarantee.

Conditional Coverage. To ensure reliability across specific operational contexts, we aim for a stronger guarantee. Let $A : \mathbb{R}^{34} \to \mathcal{K}$ be an attribute function that maps a feature vector $\mathbf{x}$ to a specific category k within a finite set of strata $\mathcal{K}$ (e.g., a specific target station or time period). The goal is to ensure the intervals satisfy conditional validity for each stratum $k \in \mathcal{K}$:

$$\mathbb{P}\left(y \in C(\mathbf{x}) \mid A(\mathbf{x}) = k\right) \geq 1 - \alpha. \tag{4.2}$$

By achieving this stronger guarantee, the prediction intervals are trustworthy across different operational contexts defined by A.

(b) Underlying Prediction Models

This study evaluates a spectrum of models, ranging from simple baselines to sophisticated deep learning architectures. Both the point prediction and interval prediction models provide the

underlying forecasts that can be subsequently wrapped by the conformal prediction framework to generate statistically guaranteed prediction intervals.

(i) Point Prediction Models

The following models are employed to generate single-value forecasts of train arrival delays.

Naive Forecast (Naive). It assumes that the delay observed at the current point in a journey will carry forward to the final destination. The prediction for the final arrival delay is therefore set equal to the train's departure delay from its current station.

LASSO [46]. We employ LASSO as a linear baseline. Its ℓ_1-norm penalty performs automated feature selection, yielding a sparse and interpretable linear model useful for benchmarking against non-linear approaches.

Gradient Boosting Machines. We employ three state-of-the-art gradient boosting frameworks: **XGBoost** [23], **LightGBM** [24], and **CatBoost** [47], which construct strong predictive models by sequentially adding weak learners (decision trees) to correct predecessor errors.

Deep Neural Network (DNN) [48]. A feed-forward neural network is used to capture complex, non-linear interactions. The architecture consists of fully connected layers with ReLU activation, trained via backpropagation to minimize the mean squared error.

(ii) Baseline Interval Prediction Models

Our study considers the following six interval prediction models. These models are trained to directly output a prediction interval $C(\mathbf{x}) = [\hat{y}_{\text{low}}, \hat{y}_{\text{high}}]$. They represent common approaches to uncertainty quantification in the machine learning literature.

Empirical Prediction Interval (EPI). This method provides a non-conditional prediction interval based on the empirical distribution of historical delays. The interval is constructed by computing the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of all arrival delays in the training dataset. The resulting interval is identical for all predictions.

Quantile Regression (QR) [49]. Unlike standard regression which models the conditional mean, QR models the conditional quantiles of the target variable. To construct a $1-\alpha$ prediction interval, two separate models are trained to estimate the conditional $\alpha/2$ and $1-\alpha/2$ quantiles, denoted $\hat{q}_{\alpha/2}(\mathbf{x})$ and $\hat{q}_{1-\alpha/2}(\mathbf{x})$. Each model is trained by minimizing the pinball loss function,

$$\mathcal{L}_\tau(y, \hat{q}_\tau) = (y - \hat{q}_\tau)\tau \cdot \mathbb{I}(y > \hat{q}_\tau) + (\hat{q}_\tau - y)(1-\tau) \cdot \mathbb{I}(y \leq \hat{q}_\tau), \tag{4.3}$$

for the corresponding quantile level τ.

Quantile Random Forest (QRF) [50,51]. A non-parametric and robust method that extends the random forest algorithm to estimate conditional quantiles. For a new input $\mathbf{x}$, each tree in the forest provides a prediction. Instead of averaging these predictions, QRF retains the full set of training labels (y_i) present in the terminal leaves where $\mathbf{x}$ lands. The conditional quantiles are then estimated from the empirical distribution of this collection of values.

Quantile Deep Neural Network (QuantDNN). Built upon the idea in [45], QuantDNN is constructed by attaching linear layers to the last layer of the DNN architecture as quantile functions for the lower and upper quantiles of the prediction interval. These quantile functions are trained using pinball loss.

Monte Carlo Dropout (Dropout) [52]. It is a technique to approximate Bayesian inference in deep neural networks and estimate model uncertainty. A standard DNN with dropout layers is trained as usual. However, at inference time, dropout is kept active. The same input $\mathbf{x}$ is passed through the network multiple times (M stochastic forward passes), yielding a distribution of predictions $\{\hat{y}^{(t)}\}_{m=1}^{M}$. The prediction interval is then formed by taking the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of this output distribution.

Deep Ensembles (DE) [53]. This method quantifies uncertainty by training an ensemble of K identical neural networks, each initialized with different random weights and trained on shuffled versions of the same data. For a new input $\mathbf{x}$, each model k predicts the parameters of a Gaussian distribution, (μ_k, σ_k^2), where σ_k^2 captures the aleatoric uncertainty. The final predictive

distribution is a Gaussian mixture $p(y|\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(y|\mu_k, \sigma_k^2)$. To construct the prediction interval, this mixture is often approximated by a single Gaussian distribution. The total predictive variance σ_{pred}^2 is calculated as the sum of the average aleatoric uncertainty ($\frac{1}{K}\sum\sigma_k^2$) and the epistemic uncertainty ($\frac{1}{K}\sum(\mu_k - \bar{\mu})^2$, where $\bar{\mu}$ is the mean of μ_k).

(c) The Conformal Prediction Framework

Unlike methods that rely on distributional or model-specific assumptions, CP's guarantees hold under the sole, mild assumption of exchangeability—that the data points $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are drawn from a joint distribution that is invariant to permutation. This makes it an ideal tool for high-stakes applications like railway operations, where the cost of untrustworthy uncertainty estimates is high. The core idea of CP is to quantify the "strangeness" of a new data point relative to a set of calibration data using a *nonconformity score*. By calibrating these scores, we can construct an interval that is guaranteed to contain the true outcome with a user-specified probability $1-\alpha$. In this study, we leverage three key variants of the CP framework.

(i) Split Conformal Prediction

Split conformal prediction (SCP), also known as inductive conformal prediction, is the most common and computationally efficient implementation of the CP framework [15]. The procedure begins by partitioning the available training data into two disjoint sets: a *proper training set* $\mathcal{D}_{\text{train}}$ and a *calibration set* $\mathcal{D}_{\text{cal}}$. Let the calibration set be of size n, i.e., $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. An arbitrary black-box predictor $\hat{f}$ is trained exclusively on $\mathcal{D}_{\text{train}}$. This trained model is then applied to the calibration set to compute *nonconformity scores*, which measure the discrepancy between predictions and true outcomes. For regression tasks, a standard choice is the absolute residual:

$$s_i = |y_i - \hat{f}(\mathbf{x}_i)|. \tag{4.4}$$

The collection of these scores $\{s_i\}_{i=1}^n$ empirically represents the distribution of model errors on unseen data. To achieve the target coverage $1-\alpha$, we then determine a calibration term $\hat{q}$ by calculating the $\lceil(1-\alpha)(n+1)\rceil/(n+1)$ empirical quantile of $\{s_i\}_{i=1}^n$. This adjustment ensures finite-sample marginal coverage validity [16].

Given a new test point $\mathbf{x}_{n+1}$, the prediction interval is constructed symmetrically around the point prediction $\hat{f}(\mathbf{x}_{n+1})$:

$$C(\mathbf{x}_{n+1}) = [\hat{f}(\mathbf{x}_{n+1}) - \hat{q}, \hat{f}(\mathbf{x}_{n+1}) + \hat{q}]. \tag{4.5}$$

By construction, this interval is guaranteed to satisfy the marginal coverage property:

$$\mathbb{P}\left(y_{n+1} \in C(\mathbf{x}_{n+1})\right) \geq 1 - \alpha. \tag{4.6}$$

A key characteristic of SCP is that the resulting interval width, $2\hat{q}$, is constant across all test points, reflecting an implicit homoscedasticity assumption on the prediction errors.

(ii) Conformalized Quantile Regression

To address the limitation of fixed-width intervals and improve efficiency, conformalized quantile regression (CQR) [17] produces intervals whose widths are adaptive to the local difficulty of the prediction, as captured by the input features $\mathbf{x}$. It elegantly combines the predictive power of quantile regression with the rigorous guarantees of the conformal framework. The process mirrors that of SCP, beginning with a train/calibration split. On the proper training set $\mathcal{D}_{\text{train}}$, two quantile regression models, $\hat{q}_{\text{low}}(\mathbf{x})$ and $\hat{q}_{\text{high}}(\mathbf{x})$, are trained to estimate the conditional $\alpha/2$ and $1-\alpha/2$ quantiles, respectively. This yields an initial, uncalibrated interval $[\hat{q}_{\text{low}}(\mathbf{x}), \hat{q}_{\text{high}}(\mathbf{x})]$. Calibration is then performed on $\mathcal{D}_{\text{cal}}$ by computing a nonconformity score for each sample, defined as the signed distance of the true value y_i from the nearest boundary of its predicted

interval:

$$s_i = \max(\hat{q}_{\text{low}}(\mathbf{x}_i) - y_i, y_i - \hat{q}_{\text{high}}(\mathbf{x}_i)). \tag{4.7}$$

A score $s_i \leq 0$ indicates that the true value is contained within the initial interval, whereas $s_i > 0$ quantifies the magnitude of the miscoverage. As before, the $\lceil (1-\alpha)(n+1) \rceil / (n+1)$ empirical quantile $\hat{q}$ of these scores is calculated.

For a new test point $\mathbf{x}_{n+1}$, the prediction interval is formed by uniformly expanding the initial quantile interval by this calibration term $\hat{q}$:

$$C(\mathbf{x}_{n+1}) = [\hat{q}_{\text{low}}(\mathbf{x}_{n+1}) - \hat{q}, \hat{q}_{\text{high}}(\mathbf{x}_{n+1}) + \hat{q}]. \tag{4.8}$$

CQR thus retains the adaptivity of quantile regression while enforcing the same formal coverage guarantee as SCP, leading to more efficient and informative intervals.

(iii) Mondrian Conformal Prediction

The guarantee provided by standard CP is marginal (Equation 4.1), meaning it holds on average over the entire data distribution. For operational use, however, it is often critical to ensure reliability across specific subgroups. For instance, prediction intervals should be equally valid for trains arriving at a major hub like *LNDNBDE* as for those at a relatively smaller station like *SVNOAKS*. Mondrian conformal prediction provides a direct path to achieving this stronger conditional coverage guarantee (Equation 4.2) [13]. The MCP framework achieves this by stratifying the data and applying the conformal procedure independently within each stratum.

A partitioning function $A(\mathbf{x})$ is defined to map each data point to a specific category k. The calibration set $\mathcal{D}_{\text{cal}}$ is then partitioned into disjoint subsets $\{\mathcal{D}_{\text{cal},k}\}$ according to these strata. The calibration process, whether using the point or interval error, is then performed separately for each stratum. This results in a unique, stratum-specific quantile, $\hat{q}_k$, calculated only from the nonconformity scores within that group. When a new test point $\mathbf{x}_{n+1}$ is presented, its stratum $k_{n+1} = A(\mathbf{x}_{n+1})$ is identified, and the prediction interval is constructed using the corresponding quantile $\hat{q}_{k_{n+1}}$. For instance, using the SCP approach, the interval would be:

$$C(\mathbf{x}_{n+1}) = [\hat{f}(\mathbf{x}_{n+1}) - \hat{q}_{k_{n+1}}, \hat{f}(\mathbf{x}_{n+1}) + \hat{q}_{k_{n+1}}]. \tag{4.9}$$

By enforcing the coverage guarantee within each defined stratum, MCP ensures that the prediction intervals are trustworthy not just on average, but also conditionally for predefined operational contexts, which provides a more robust and reliable risk assessment tool.

5. Experiments and Results

To empirically validate the theoretical properties of the discussed uncertainty quantification methods, we conduct a comprehensive set of experiments on the Southeastern Railway dataset with five busiest target stations. The primary objectives of this experimental evaluation are threefold: first, to establish a performance baseline by comparing a range of conventional point prediction models; second, to assess the quality of prediction intervals generated by various native UQ techniques, including Bayesian-inspired neural networks and quantile regression methods; and third, to systematically demonstrate the effectiveness and versatility of the CP framework in enhancing the reliability of predictions from both point and interval predictors.

(a) Experimental Setup

To obtain robust and statistically significant results, we employ a repeated random splitting procedure across 10 independent runs. In each run, the dataset is partitioned into four distinct, non-overlapping sets: a training set (50% of the data) for model fitting; a validation set (10%) for hyperparameter tuning and early stopping; a dedicated calibration set (20%) used exclusively by the conformal prediction algorithms; and a final test set (20%) for performance evaluation.

A unified hyperparameter optimization (HPO) strategy is adopted for all trainable models using the Optuna framework with a tree-structured Parzen estimator (TPE) sampler. To ensure fairness and computational efficiency, HPO is performed once for each model architecture. This is done on a 10% subset of the training and validation data from the first experimental split, with a budget of 50 trials. The resulting optimal hyperparameters are then saved and consistently applied to that model across all 10 experimental runs.

The CP framework is systematically applied as a post-processing calibration step. Standard point prediction models are calibrated using SCP, while the baseline interval predictors are adjusted using the interval calibration technique of CQR. Furthermore, to investigate conditional coverage guarantees, we apply the MCP variant. For this, the calibration and prediction sets are stratified based on the five target stations encoded in the feature set, allowing for the computation of separate calibration terms for each station group.

(i) Evaluation Metrics

Model performance is evaluated using a comprehensive suite of metrics. Point prediction accuracy is assessed using the **mean absolute error (MAE)**, **root mean squared error (RMSE)**, and the **coefficient of determination (R^2)**, defined as follows:

$$\mathrm{MAE} = \frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{i \in \mathcal{I}_{\text{test}}} |y_i - \hat{y}_i|, \tag{5.1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{i \in \mathcal{I}_{\text{test}}} (y_i - \hat{y}_i)^2}, \tag{5.2}$$

$$R^2 = 1 - \frac{\sum_{i \in \mathcal{I}_{\text{test}}} (y_i - \hat{y}_i)^2}{\sum_{i \in \mathcal{I}_{\text{test}}} (y_i - \bar{y})^2}, \tag{5.3}$$

where y_i denotes the true value, $\hat{y}_i$ the predicted value, $\bar{y}$ the mean of the true values, and $\mathcal{I}_{\text{test}}$ the index of test set.

Given a coverage level $1-\alpha$, let $C(\mathbf{x}_i) = [\hat{y}_{\text{low},i}, \hat{y}_{\text{high},i}]$ denote the predicted interval for sample i. The quality of prediction intervals is judged by three key metrics: the **coverage rate (CR)**, **mean width (MW)**, and the **Winkler score (Winkler)** [54], defined as follows:

$$\mathrm{CR} = \frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{i \in \mathcal{I}_{\text{test}}} \mathbb{I}\left(\hat{y}_{\text{low},i} \le y_i \le \hat{y}_{\text{high},i}\right), \tag{5.4}$$

$$\mathrm{MW} = \frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{i \in \mathcal{I}_{\text{test}}} \left(\hat{y}_{\text{high},i} - \hat{y}_{\text{low},i}\right), \tag{5.5}$$

$$\begin{aligned} \mathrm{Winkler} = \frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{i \in \mathcal{I}_{\text{test}}} \Big[& (\hat{y}_{\text{high},i} - \hat{y}_{\text{low},i}) \\ & + \frac{2}{\alpha}\left((y_i - \hat{y}_{\text{high},i})_+ + (\hat{y}_{\text{low},i} - y_i)_+\right)\Big]. \end{aligned} \tag{5.6}$$

Coverage rate measures the empirical frequency with which the true value falls within the predicted interval, which should be close to the nominal level $1-\alpha$. Mean width quantifies the average size of the prediction intervals, reflecting the precision of uncertainty estimates. A lower MW indicates narrower intervals and thus higher precision, but should be balanced against sufficient coverage. Finally, the Winkler score (also known as the interval score) provides a combined evaluation of both coverage and width, penalizing both under-coverage and excessive interval width. A lower Winkler score indicates better overall interval performance. These interval metrics are calculated for a range of nominal coverage levels: $1-\alpha \in \{0.75, 0.80, 0.85, 0.90, 0.95, 0.99\}$. For the Mondrian methods, we also report the conditional coverage rate for each station-based subgroup to explicitly verify the effectiveness of the conditional calibration.

(ii) Implementation Details

Following the HPO process, all point and interval prediction models were configured with their determined optimal parameters. These specific configurations, which form the basis of our comparative analysis, are comprehensively detailed in Table 2. For interval prediction, some models are directly derived from these point predictors. Specifically, Dropout and QuantDNN are both built upon the optimized DNN architecture, thereby inheriting its structural parameters. Similarly, the QR model is based on the tuned XGBoost and inherits its parameters. The DE method is composed of an ensemble of five neural networks, where each network shares an identical, independently optimized architecture. All neural network models were trained using the Adam optimizer with a batch size of 1024 and early stopping on validation loss with a patience of 20 epochs (maximum 200 epochs). Tree-based models applied early stopping with 50 rounds of patience.

(b) Results

This section presents the empirical results of our comprehensive experiments. We first evaluate the performance of the point prediction models, then conduct a detailed analysis of the interval prediction methods, focusing on coverage validity and efficiency, and finally, we assess the effectiveness of Mondrian conformal prediction in providing conditional coverage guarantees.

(i) Point Prediction Performance

The performance of the six point prediction models is summarized in Table 3. The results clearly indicate that the gradient boosting models, particularly XGBoost, significantly outperform the other methods. XGBoost emerges as the superior model across all metrics, achieving the lowest MAE of 61.7422 seconds and RMSE of 96.8771 seconds, along with the highest R^2 of 0.7824. This suggests that XGBoost is highly effective at capturing the complex, non-linear relationships within the feature set. The other gradient boosting models, CatBoost and LightGBM, also demonstrate strong predictive power, substantially outperforming the linear LASSO model and the DNN. The Naive forecast, which simply projects the current delay forward, serves as a performance baseline and, as expected, yields the highest errors. The robust performance of XGBoost makes it an excellent candidate for generating the point estimates that support the standard CP approach.

(ii) Interval Prediction Performance

Moving from point estimates to interval predictions, we first analyze performance at a nominal coverage level of $1 - \alpha = 0.90$, as detailed in Table 4. A primary and striking finding is the perfect calibration achieved by all conformalized methods. Whether applied to point predictors (e.g., C-XGBoost) or as a post-hoc correction for interval predictors (e.g., C-QR), the conformal framework ensures that the empirical coverage rates are almost exactly the desired 0.90. In stark contrast, most baseline interval predictors fail to achieve the nominal coverage. For instance, QR is under-confident with a coverage of 0.8265, while QRF and DE are over-confident (0.9273 and 0.9352, respectively). Dropout and QuantDNN exhibit severe under-coverage, rendering their intervals unreliable. While the simple EPI method achieves the target coverage by design, it does so with an impractically large mean width (579.5s) and the highest Winkler score among valid methods, highlighting its inefficiency.

This pattern of miscalibration among baseline methods and the corrective power of CP is consistent across all studied coverage levels, as illustrated in Figures 3 and 4. Figure 3 demonstrates that applying standard CP to point predictors results in prediction intervals whose empirical coverage rates align perfectly with the desired nominal levels, tracing the diagonal line of perfect calibration. Similarly, Figure 4 shows the dramatic effect of conformalization on baseline interval methods. The dashed lines, representing the uncalibrated models, frequently

Table 2. Optimal hyperparameters for underlying prediction models.

Model	Hyperparameter	Search Range	Value
LASSO	alpha	[1e-5, 1e-1]	0.00018
XGBoost	n_estimators	(100, 1000)	944
	max_depth	(4, 16)	13
	learning_rate	(0.01, 0.2)	0.0549
	subsample	(0.6, 1.0)	0.8746
	colsample_bytree	(0.6, 1.0)	0.7750
	reg_alpha	(0.0, 10.0)	2.5061
	reg_lambda	(0.0, 10.0)	9.1311
LightGBM	n_estimators	(100, 1000)	998
	max_depth	(4, 16)	11
	num_leaves	(50, 150)	83
	learning_rate	(0.01, 0.2)	0.1918
	feature_fraction	(0.6, 1.0)	0.8142
	bagging_fraction	(0.6, 1.0)	0.9986
	bagging_freq	(1, 7)	6
	reg_alpha	(0.0, 10.0)	5.8440
	reg_lambda	(0.0, 10.0)	9.9892
CatBoost	iterations	(100, 1000)	959
	depth	(4, 16)	13
	learning_rate	(0.01, 0.2)	0.1484
	l2_leaf_reg	(1.0, 10.0)	3.0076
	bagging_temperature	(0.0, 1.0)	0.0409
DNN	Hidden Layers	[(64,), (128,), (64, 32), (128, 64)]	[128, 64]
	learning_rate	(1e-4, 1e-2)	0.00073
	dropout_rate	(0.1, 0.5)	0.2165
	weight_decay	(0.0, 0.01)	0.00612
Dropout	Built upon the optimized point-prediction DNN architecture.		
QuantDNN	Built upon the optimized point-prediction DNN architecture.		
QR	Based on the tuned XGBoost model and inherits its hyperparameters.		
QRF	n_estimators	(100, 1000)	685
	max_depth	(4, 16)	16
	min_samples_split	(20, 200)	46
	min_samples_leaf	(10, 100)	17
	max_features	(0.6, 1.0)	0.9415
DE (5 NNs)	Hidden Layers	[(64,), (128,), (64, 32), (128, 64)]	[128, 64]
	learning_rate	(1e-5, 1e-3)	0.00054
	dropout_rate	(0.1, 0.5)	0.1053
	weight_decay	(0.0, 0.01)	6.199e-05

Table 3. Performance comparison of point prediction models on the test set. Values are reported as mean ± standard deviation over 10 runs. MAE and RMSE are in seconds.

Method	MAE (s) ↓	RMSE (s) ↓	R^2 ↑
Naive	104.3392 ± 0.1197	167.5438 ± 0.4090	0.3490 ± 0.0021
LASSO	101.3731 ± 0.1128	157.4337 ± 0.4045	0.4252 ± 0.0019
XGBoost	**61.7422 ± 0.2315**	**96.8771 ± 0.3539**	**0.7824 ± 0.0016**
LightGBM	78.0200 ± 0.1852	120.5043 ± 0.5266	0.6633 ± 0.0027
CatBoost	70.2972 ± 0.1442	108.2061 ± 0.3213	0.7285 ± 0.0022
DNN	93.9260 ± 0.9931	149.6675 ± 0.6536	0.4805 ± 0.0044

Table 4. Performance comparison of prediction intervals at coverage level of $1 - \alpha = 0.90$. Values are mean $\pm$ standard deviation over 10 runs. Mean Width is in seconds.

Method	CR	MW (s)	Winkler ↓
Conformalized Point Predictors (C-)			
C-Naive	0.9004 ± 0.0007	469.2000 ± 1.0328	797.7541 ± 1.1304
C-LASSO	0.8999 ± 0.0006	433.3018 ± 0.5854	730.6524 ± 1.2292
C-XGBoost	0.9000 ± 0.0005	268.9706 ± 1.1674	450.1775 ± 1.6660
C-LightGBM	0.8999 ± 0.0006	336.8313 ± 0.6628	559.8310 ± 2.0779
C-CatBoost	0.9000 ± 0.0004	304.2067 ± 1.0284	502.9287 ± 1.2048
Baseline Interval Predictors			
EPI	0.9003 ± 0.0003	579.5000 ± 0.5270	900.5474 ± 1.9724
QR	0.8265 ± 0.0020	275.3361 ± 2.4578	415.1835 ± 2.7452
QRF	0.9273 ± 0.0004	399.9141 ± 1.1197	543.4603 ± 1.0234
QuantDNN	0.7284 ± 0.0103	399.4149 ± 4.1427	848.6902 ± 17.7194
Dropout	0.2725 ± 0.0077	68.9254 ± 2.8033	1365.7148 ± 17.1403
DE	0.9352 ± 0.0012	416.6096 ± 3.0212	581.0682 ± 1.1054
Conformalized Interval Predictors (C-)			
C-QR	0.8999 ± 0.0004	299.2960 ± 1.9299	$\mathbf{407.6149 \pm 3.0386}$
C-QRF	0.9002 ± 0.0005	379.7141 ± 0.7975	541.2892 ± 1.0006
C-QuantDNN	0.9000 ± 0.0002	530.6516 ± 3.6485	756.5759 ± 9.4672
C-Dropout	0.9000 ± 0.0004	389.2066 ± 2.1145	676.4284 ± 7.1303
C-DE	0.9000 ± 0.0002	357.0705 ± 1.3940	568.8646 ± 1.5883

and significantly deviate from the perfect calibration diagonal, confirming their unreliability. The solid lines, representing their conformalized counterparts, align almost perfectly with the diagonal, showcasing the framework's ability to enforce coverage guarantees regardless of the underlying model's initial bias.

While coverage is a prerequisite for trustworthy intervals, their utility is also determined by their width (efficiency), a trade-off formally captured by the Winkler score. Figure 5 plots the log-transformed Winkler score for all interval methods across the range of coverage levels. At lower coverage levels, C-QR, uncalibrated QR, and C-XGBoost achieve the lowest scores. As coverage increases, their performance diverges: C-QR consistently achieves the lowest Winkler scores, with uncalibrated QR closely behind, while C-XGBoost gradually falls behind and is surpassed by C-QRF at $1 - \alpha = 0.99$. This is particularly noteworthy because the best interval predictor (C-QR) is not built upon the best point predictor (XGBoost). Although C-XGBoost produces valid intervals, their width is static, determined by a single calibration term. In contrast, QR implicitly learns the data's heteroscedasticity, producing wider intervals for more uncertain predictions and narrower ones for more certain predictions. CQR leverages this adaptivity, making a small adjustment to guarantee coverage while largely preserving the efficiency of the adaptive-width intervals. This makes C-QR more efficient and, therefore, more useful than the fixed-width C-XGBoost intervals.

The benefit of conformalization is further quantified in Table 5, which compares the Winkler scores of baseline interval models against their conformalized versions. Conformalization consistently reduces the Winkler score for every method at every coverage level. This improvement is driven by the correction of miscoverage: for under-confident models like QR, QuantDNN, and Dropout, the score is reduced by expanding intervals to meet the coverage target, thereby avoiding large penalties for uncovered points; for over-confident models like QRF and DE, the score is reduced by slightly shrinking the intervals while still maintaining the target coverage. A Wilcoxon signed-rank test [55], with Holm-Bonferroni correction [56] for multiple comparisons, confirms that the reduction in Winkler score achieved by conformalization is statistically significant for all baseline methods ($p_{\text{corrected}} = 0.0293$).

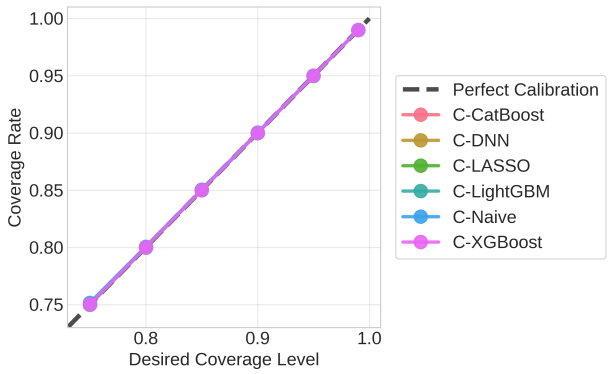

Figure 3. Empirical coverage rates of prediction intervals from SCP applied to various point prediction models, which validates the theoretical guarantee of CP: all methods achieve coverage rates nearly identical to the target levels.

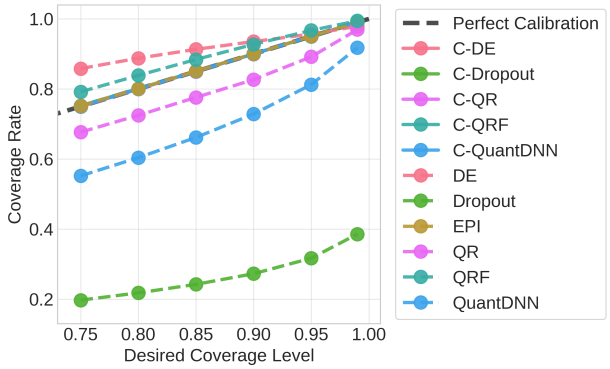

Figure 4. Effect of post-hoc conformal calibration on coverage rates of baseline interval methods. While baseline methods are often miscalibrated, conformalization corrects their biases, aligning empirical coverage precisely with target levels.

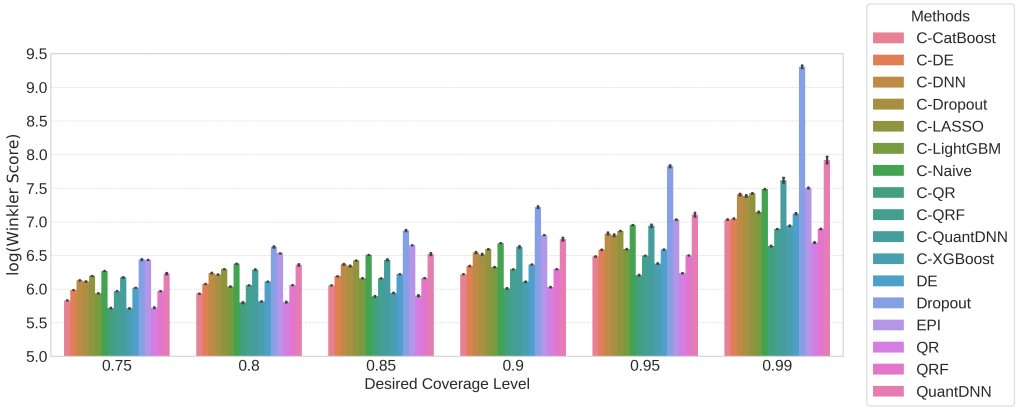

Figure 5. Comparison of interval prediction performance using the log-transformed Winkler score across all methods and coverage levels. Each bar shows the mean score for a method at a given level, with error bars indicating standard deviation over 10 runs. The figure highlights that C-QR is the most efficient, demonstrating the utility of its adaptive intervals.

Table 5. Comparison of average Winkler score between conformalized (C-) and Baseline interval prediction models across all studied coverage levels ($1-\alpha$) over 10 runs. The Winkler score is consistently reduced by conformalization, and all reductions are statistically significant (Wilcoxon signed-rank test, $N = 10$, $p_{\text{corrected}} = 0.0293$).

Model	Coverage Level	Base (Mean Winkler)	C- (Mean Winkler)	Reduction
QR	0.75	306.0959	304.3624	1.7334
	0.80	332.2852	329.6732	2.6120
	0.85	365.9900	361.9521	4.0380
	0.90	415.1835	407.6149	7.5685
	0.95	511.3870	495.8110	15.5760
	0.99	805.2701	764.2641	41.0059
QRF	0.75	391.9429	391.1223	0.8206
	0.80	427.9471	426.8203	1.1268
	0.85	475.1236	473.5705	1.5531
	0.90	543.4603	541.2892	2.1711
	0.95	666.3547	663.3557	2.9990
	0.99	987.7230	986.1496	1.5734
QuantDNN	0.75	508.0059	480.0046	28.0013
	0.80	578.5441	539.3152	39.2289
	0.85	680.4976	623.0272	57.4704
	0.90	848.6902	756.5759	92.1143
	0.95	1222.1984	1035.4191	186.7793
	0.99	2757.8633	2038.4157	719.4476
Dropout	0.75	625.5436	451.6930	173.8506
	0.80	755.0034	500.6153	254.3881
	0.85	963.7182	568.7107	395.0075
	0.90	1365.7148	676.4284	689.2864
	0.95	2509.5138	899.1505	1610.3633
	0.99	11005.8435	1612.6648	9393.1786
DE	0.75	411.5701	397.6669	13.9032
	0.80	451.4767	436.0844	15.3922
	0.85	504.0446	488.5239	15.5207
	0.90	581.0682	568.8646	12.2036
	0.95	726.8549	724.9134	1.9415
	0.99	1239.9027	1150.6383	89.2644

(iii) Conditional Coverage with Mondrian Conformal Prediction

Finally, we assess the stronger guarantee of conditional coverage using Mondrian CP, with results presented in Figure 6. This figure compares the station-specific coverage rates of standard CP methods (Standard CP, represented by bars) and their Mondrian counterparts (Mondrian CP, represented by points) across the five busiest target stations. The results reveal a critical limitation of the marginal guarantee provided by standard CP. While these methods (bars) maintain the nominal coverage on average across all stations, their conditional coverage can be unreliable. For several station groups and desired coverage levels, the empirical coverage rate dips below the target level (the dashed red line), meaning the intervals are not as trustworthy for those specific stations. For example, at the 0.95 desired coverage level, the standard CP method for several models provides coverage below 0.94 for the London Charing Cross station (*CHRX*).

In contrast, the Mondrian CP variants (points) consistently achieve the target coverage rate within each station-specific stratum. By calculating separate calibration terms for each station, MCP enforces the conditional coverage guarantee, ensuring that the prediction intervals are equally reliable across these different operational contexts.

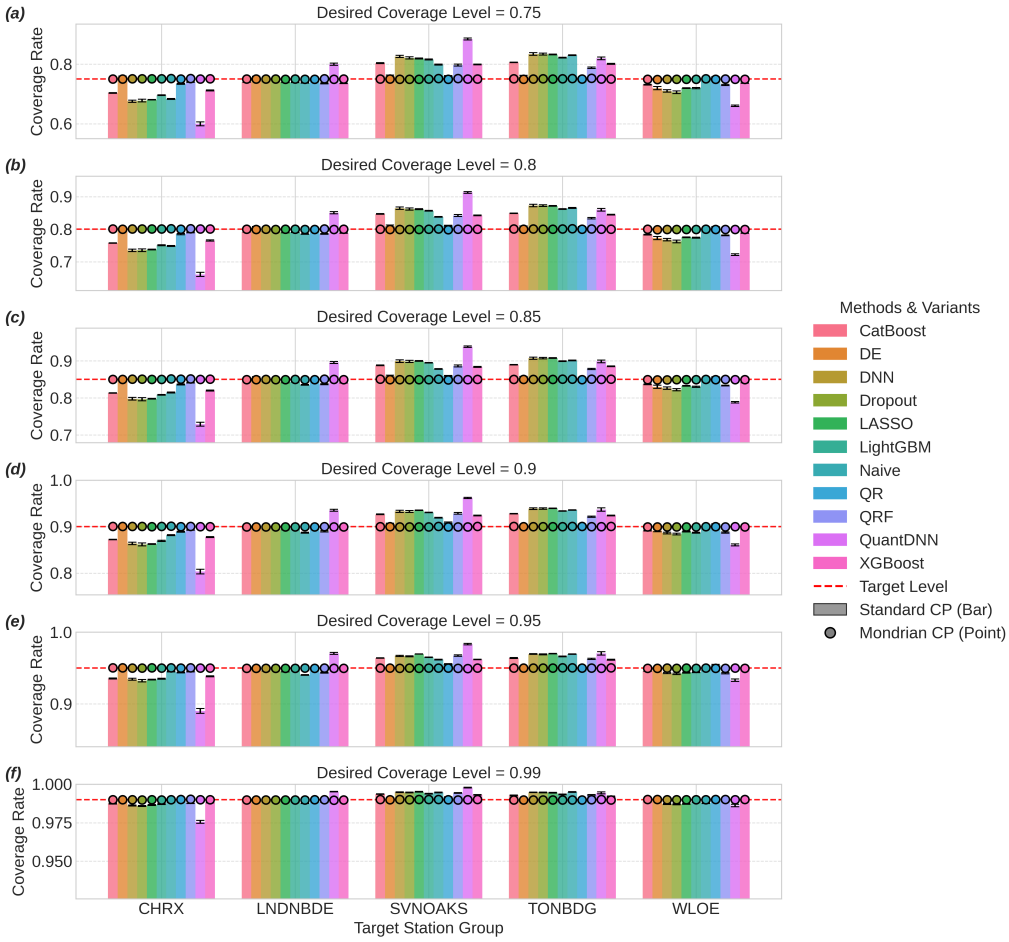

Figure 6. Comparison of conditional coverage rates for standard (marginal) and Mondrian conformal prediction across the five busiest target stations. Each subplot (a–f) corresponds to a desired coverage level $(1 - \alpha)$. For each station group, bars with error bars show the mean and standard deviation of standard CP over 10 runs; black-outlined points show Mondrian CP results. The figure shows that while standard CP may fail to meet nominal coverage for subgroups, MCP enforces conditional coverage, ensuring reliable performance across all stations.

6. Discussion

Our empirical results highlight a critical gap in uncertainty quantification for train delay prediction: widely used methods such as quantile regression, MC Dropout, and Deep Ensembles consistently fail to achieve their nominal coverage on real-world data. Despite theoretical grounding, these approaches produce miscalibrated intervals that may appear reasonable but are operationally unreliable, posing risks to scheduling, resource allocation, and passenger trust. This fragility underscores the need for robust, assumption-free calibration in high-stakes domains.

Conformal prediction addresses this challenge effectively as a model-agnostic, post-hoc framework that guarantees marginal coverage regardless of the underlying model. It corrects miscalibration across all underlying predictors, from simple linear regression to complex deep neural networks, establishing CP not merely as another UQ method, but as a foundational tool for reliable uncertainty reporting. Notably, the efficiency of the resulting intervals depends on the quality of the base model's uncertainty structure. In our study, conformalized quantile regression (CQR) excels in this regard, consistently producing the most useful intervals (lowest Winkler

score). Since quantile regression is inherently designed to model the conditional distribution of the data, it produces adaptive, heteroscedastic intervals that, even if miscalibrated, provide a far better starting point for calibration than the static-width intervals derived from a point predictor like XGBoost. This highlights that for building the best interval predictors, the goal should be to select a base model that best captures the shape of the predictive uncertainty, rather than one that simply minimizes point-wise error.

Beyond marginal guarantees, operational reliability requires conditional validity, particularly across heterogeneous stations or routes. Standard CP can exhibit significant coverage variability across subgroups, undermining trust in specific contexts. Mondrian conformal prediction resolves this by enforcing coverage within predefined strata, ensuring consistent performance across the network. This shift from average to conditional reliability is not theoretical nuance; it is essential for fair and trustworthy deployment in practice.

For railway operators, our findings provide a clear pathway to risk-aware decision-making. CQR emerges as a particularly strong candidate, balancing guaranteed coverage with interval efficiency. Deploying such calibrated systems enables proactive actions, such as preemptively adjusting crew schedules, holding connecting services, or providing passengers with reliable delay windows–all backed by statistically rigorous guarantees.

While our findings are based on a large-scale dataset from a single operator, they motivate several important directions for future research. First, extending this evaluation to other networks and longer timeframes would further test the generality of our findings, and the feature space, though comprehensive, could be enriched with external data sources like real-time weather information or network-wide passenger flow data. More importantly, railway operations are inherently multivariate and spatiotemporally interconnected. A natural extension of this work is the transition from predicting delays at a single terminal node to forecasting the entire delay trajectory of a train across a sequence of future stations. In such a multivariate setting, the target variable becomes a vector $\mathbf{y} \in \mathbb{R}^k$, representing delays at the next k stations. For practical railway operations, prediction sets shaped as hyperrectangles are particularly helpful, as they facilitate marginal interpretation [57]. Relevant approaches include independent modeling with Bonferroni or similar correction [58,59], CQR-based multi-output regression with max-score aggregation [43,60], and Copula-based CP [61]. Future research can explore multi-output CP techniques that generate such interpretable sets while accounting for dependencies to maintain efficiency.

Furthermore, if real-time data streams are available, the problem can be framed as a sequential learning task to model the inherent time-varying relationship between the operational state and delay outcomes. Adaptive CP methods for time series [62–64] offer promising mechanisms to handle distribution shifts by dynamically adjusting interval widths in real-time. Finally, since train delays are not isolated events but propagate through the physical network topology, integrating Graph Neural Networks (GNNs) to model the spatiotemporal state of the entire railway network would be a logical step to capture these complex interactions.

7. Conclusion

This paper presents a comprehensive evaluation of conformal prediction (CP) as a robust framework for uncertainty quantification in train delay prediction, a high-stakes domain where inaccurate risk assessment can have significant operational consequences. We show that widely used interval methods often fail to achieve nominal coverage, while standard CP and, particularly, conformalized quantile regression (CQR) effectively calibrate predictions to provide rigorous, distribution-free marginal coverage guarantees. CQR emerges as the most efficient approach in our study, preserving the adaptive interval widths of quantile regression while ensuring validity. Furthermore, we demonstrate that Mondrian conformal prediction (MCP) is essential for practical deployment, delivering conditional coverage across heterogeneous subgroups such as railway stations. Together, these results establish CP as a powerful, practical tool for transforming machine learning models into reliable, decision-ready systems, with broad implications for transportation and other safety-critical domains.

Data Accessibility. The Southeastern Railway dataset is confidential as this is Southeastern Railway's actual production data. The code used in this study is publicly available in the GitHub repository [65].

Competing Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Funding. The work described in this paper was supported by the National Natural Science Foundation of China (Project No. 62506315) and City University of Hong Kong (Project No. 9610639 and No. 7020161).

Acknowledgements.

References

1. Mouwen A. 2015 Drivers of customer satisfaction with public transport services. *Transportation Research Part A: Policy and Practice* **78**, 1–20. (https://doi.org/10.1016/j.tra.2015.05.005)
2. Juan de Oña, Rocio de Oña LE, Mazzulla G. 2015 Heterogeneity in Perceptions of Service Quality among Groups of Railway Passengers. *International Journal of Sustainable Transportation* **9**, 612–626. (10.1080/15568318.2013.849318)
3. Keaveney SM. 1995 Customer Switching Behavior in Service Industries: An Exploratory Study. *Journal of Marketing* **59**, 71–82.
4. Huang Z, Loo BP. 2023 Vulnerability assessment of urban rail transit in face of disruptions: A framework and some lessons from Hong Kong. *Sustainable Cities and Society* **98**, 104858. (https://doi.org/10.1016/j.scs.2023.104858)
5. Ferranti E, Chapman L, Lowe C, McCulloch S, Jaroszweski D, Quinn A. 2016 Heat-Related Failures on Southeast England's Railway Network: Insights and Implications for Heat Risk Management. *Weather, Climate, and Society* **8**, 177 – 191. (10.1175/WCAS-D-15-0068.1)
6. Shi J, Wen S, Zhao X, Wu G. 2019 Sustainable Development of Urban Rail Transit Networks: A Vulnerability Perspective. *Sustainability* **11**. (10.3390/su11051335)
7. Wang J, Mantas-Nakhai R, Yu J. 2023 Statistical learning for train delays and influence of winter climate and atmospheric icing. *Journal of Rail Transport Planning & Management* **26**, 100388. (https://doi.org/10.1016/j.jrtpm.2023.100388)
8. Shanthi S, Maruthu Kannan B, S G, Babuji R, Sivakumar S, Malathi N. 2024 Optimizing City Transit: IoT and Gradient Boosting Algorithms for Accurate Bus Arrival Predictions. In *2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)* pp. 1–5. (10.1109/IITCEE59897.2024.10467430)
9. Deng W, Li K, Zhao H. 2024 A Flight Arrival Time Prediction Method Based on Cluster Clustering-Based Modular With Deep Neural Network. *IEEE Transactions on Intelligent Transportation Systems* **25**, 6238–6247. (10.1109/TITS.2023.3338251)
10. Li X, Cottam A, Wu YJ. 2023 Transit Arrival Time Prediction Using Interaction Networks. *IEEE Transactions on Intelligent Transportation Systems* **24**, 3833–3844. (10.1109/TITS.2023.3238289)
11. Chen MY, Chiang HS, Yang KJ. 2022 Constructing Cooperative Intelligent Transport Systems for Travel Time Prediction With Deep Learning Approaches. *IEEE Transactions on Intelligent Transportation Systems* **23**, 16590–16599. (10.1109/TITS.2022.3148269)
12. Liu T, Ma J, Guan W, Song Y, Niu H. 2012 Bus Arrival Time Prediction Based on the k-Nearest Neighbor Method. In *2012 Fifth International Joint Conference on Computational Sciences and Optimization* pp. 480–483. (10.1109/CSO.2012.111)
13. Vovk V, Gammerman AA, Shafer G. 2005 *Algorithmic learning in a random world*. New York: Springer.
14. Gammerman A, Vovk V, Vapnik V. 1998 Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* UAI'98 p. 148–155 San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
15. Papadopoulos H, Proedrou K, Vovk V, Gammerman A. 2002 Inductive Confidence Machines for Regression. In Elomaa T, Mannila H, Toivonen H, editors, *Machine Learning: ECML 2002* pp. 345–356 Berlin, Heidelberg. Springer Berlin Heidelberg.
16. Lei J, Wasserman L. 2013 Distribution-free Prediction Bands for Non-parametric Regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **76**, 71–96. (10.1111/rssb.12021)
17. Romano Y, Patterson E, Candès EJ. 2019 In *Conformalized quantile regression*,. Red Hook, NY, USA: Curran Associates Inc.

18. Wu CH, Ho JM, Lee D. 2004 Travel-time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems* **5**, 276–281. (10.1109/TITS.2004.837813)
19. Yaghini M, Khoshraftar MM, Seyedabadi M. 2013 Railway passenger train delay prediction via neural network model. *Journal of Advanced Transportation* **47**, 355–368. (https://doi.org/10.1002/atr.193)
20. Goudarzi F. 2018 Travel Time Prediction: Comparison of Machine Learning Algorithms in a Case Study . In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)* pp. 1404–1407 Los Alamitos, CA, USA. IEEE Computer Society. (10.1109/HPCC/SmartCity/DSS.2018.00232)
21. Mane AS, Pulugurtha SS. 2018 Link-level Travel Time Prediction Using Artificial Neural Network Models. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* pp. 1487–1492. (10.1109/ITSC.2018.8569731)
22. Wang H. 2022 A Two-Stage Train Delay Prediction Method Based on Data Smoothing and Multimodel Fusion Using Asymmetry Features in Urban Rail Systems. *Wireless Communications and Mobile Computing* **2022**, 5188105. (https://doi.org/10.1155/2022/5188105)
23. Chen T, Guestrin C. 2016 XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '16 p. 785–794. ACM. (10.1145/2939672.2939785)
24. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. 2017 Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **30**.
25. Barbour W, Samal C, Kuppa S, Dubey A, Work DB. 2018 On the Data-Driven Prediction of Arrival Times for Freight Trains on U.S. Railroads. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* pp. 2289–2296. (10.1109/ITSC.2018.8569406)
26. Kankanamge KD, Witharanage YR, Withanage CS, Hansini M, Lakmal D, Thayasivam U. 2019 Taxi Trip Travel Time Prediction with Isolated XGBoost Regression. In *2019 Moratuwa Engineering Research Conference (MERCon)* pp. 54–59. (10.1109/MERCon.2019.8818915)
27. Sarhani M, Voß S. 2024 Prediction of rail transit delays with machine learning: How to exploit open data sources. *Multimodal Transportation* **3**, 100120. (https://doi.org/10.1016/j.multra.2024.100120)
28. Liu Q, Wang S, Li Z, Li L, Zhang J, Wen C. 2023 Prediction of high-speed train delay propagation based on causal text information. *Railway Engineering Science* **31**, 89–106.
29. Buijse BJ, Reshadat V, Enzing OW. 2021 A Deep Learning-Based Approach for Train Arrival Time Prediction. In *Intelligent Data Engineering and Automated Learning – IDEAL 2021: 22nd International Conference, IDEAL 2021, Manchester, UK, November 25–27, 2021, Proceedings* p. 213–222 Berlin, Heidelberg. Springer-Verlag. (10.1007/978-3-030-91608-4_22)
30. Rashvand N, Hosseini SS, Azarbayjani M, Tabkhi H. 2024 Real-Time Bus Arrival Prediction: A Deep Learning Approach for Enhanced Urban Mobility. .
31. Zhang D, Du C, Peng Y, Liu J, Mohammed S, Calvi A. 2024 A Multi-Source Dynamic Temporal Point Process Model for Train Delay Prediction. *IEEE Transactions on Intelligent Transportation Systems* **25**, 17865–17877. (10.1109/TITS.2024.3430031)
32. Tiong KY, Ma Z, Palmqvist CW. 2023 A review of data-driven approaches to predict train delays. *Transportation Research Part C: Emerging Technologies* **148**, 104027. (https://doi.org/10.1016/j.trc.2023.104027)
33. Spjuth O, Carrión Brännström R, Carlsson L, Gauraha N. 2019 Combining Prediction Intervals on Multi-Source Non-Disclosed Regression Datasets. In *Proceedings of the 8th Symposium on Conformal and Probabilistic Prediction with Applications* pp. 53–65. PMLR.
34. Mazloumi E, Rose G, Currie G, Moridpour S. 2011 Prediction intervals to account for uncertainties in neural network predictions: Methodology and application in bus travel time prediction. *Engineering Applications of Artificial Intelligence* **24**, 534–542. (https://doi.org/10.1016/j.engappai.2010.11.004)
35. Khosravi A, Mazloumi E, Nahavandi S, Creighton D, van Lint JWC. 2011 Prediction Intervals to Account for Uncertainties in Travel Time Prediction. *IEEE Transactions on Intelligent Transportation Systems* **12**, 537–547. (10.1109/TITS.2011.2106209)
36. Spanninger T, Wiedemann N, Corman F. 2024 Quantifying the dynamic predictability of train delay with uncertainty-aware neural networks. *Transportation Research Part C: Emerging Technologies* **162**, 104563. (https://doi.org/10.1016/j.trc.2024.104563)

37. Zhao W, Wang G, Wang Z, Liu L, Wei X, Wu Y. 2022 A uncertainty visual analytics approach for bus travel time. *Visual Informatics* **6**, 1–11. (https://doi.org/10.1016/j.visinf.2022.06.002)
38. Huang P, Corman F. 2024 Probabilistic Modeling of Train Operations for Uncertainty Quantification: A Context-Aware Bayesian Network Approach. *IEEE Transactions on Intelligent Transportation Systems* **25**, 21117–21128. (10.1109/TITS.2024.3477424)
39. Zhang Y, Haghani A, Zeng X. 2015 Component GARCH Models to Account for Seasonal Patterns and Uncertainties in Travel-Time Prediction. *IEEE Transactions on Intelligent Transportation Systems* **16**, 719–729. (10.1109/TITS.2014.2339097)
40. Bao J, Colombo N, Manokhin V, Cao S, Luo R. 2025 A Review and Comparative Analysis of Univariate Conformal Regression Methods. In *Proceedings of the 14th Symposium on Conformal and Probabilistic Prediction with Applications (COPA)* vol. 266 pp. 282–304. PMLR.
41. Johansson U, Boström H, Löfström T, Linusson H. 2014 Regression conformal prediction with random forests. *Machine Learning* **97**, 155–176.
42. Johnstone C, Ndiaye E. 2025 Exact and Approximate Conformal Inference for Multi-Output Regression. In *Proceedings of the 14th Symposium on Conformal and Probabilistic Prediction with Applications (COPA)* vol. 266 pp. 153–172. PMLR.
43. Dheur V, Fontana M, Estievenart Y, Desobry N, Taieb SB. 2025 A Unified Comparative Study with Generalized Conformity Scores for Multi-Output Conformal Regression. In *Forty-second International Conference on Machine Learning*.
44. Alkhatib A, Boström H, Ennadir S, Johansson U. 2023 Approximating Score-based Explanation Techniques Using Conformal Regression. In *Proceedings of the 12th Symposium on Conformal and Probabilistic Prediction with Applications (COPA)* vol. 204 pp. 450–469. PMLR.
45. Wu Y, Ye Y, Zeb A, Yu JJ, Wang Z. 2024 Adaptive Modeling of Uncertainties for Traffic Forecasting. *IEEE Transactions on Intelligent Transportation Systems* **25**, 4427–4442. (10.1109/TITS.2023.3327100)
46. Tibshirani R. 1996 Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288.
47. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. 2018 CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* **31**.
48. Hinton GE, Osindero S, Teh YW. 2006 A fast learning algorithm for deep belief nets. *Neural Comput.* **18**, 1527–1554. (10.1162/neco.2006.18.7.1527)
49. Koenker R, Bassett Jr G. 1978 Regression quantiles. *Econometrica: journal of the Econometric Society* pp. 33–50.
50. Meinshausen N. 2006 Quantile Regression Forests. *Journal of Machine Learning Research* **7**, 983–999.
51. Johnson RA. 2024 quantile-forest: A Python Package for Quantile Regression Forests. *Journal of Open Source Software* **9**, 5976. (10.21105/joss.05976)
52. Gal Y, Ghahramani Z. 2016 Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Balcan MF, Weinberger KQ, editors, *Proceedings of The 33rd International Conference on Machine Learning* vol. 48*Proceedings of Machine Learning Research* pp. 1050–1059 New York, New York, USA. PMLR.
53. Lakshminarayanan B, Pritzel A, Blundell C. 2017 Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* NIPS'17 p. 6405–6416 Red Hook, NY, USA. Curran Associates Inc.
54. Winkler RL. 1972 A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association* **67**, 187–191.
55. Wilcoxon F. 1992 Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution* , pp. 196–202. Springer.
56. Holm S. 1979 A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* pp. 65–70.
57. Sampson M, Chan KS. 2024 Conformal Multi-Target Hyperrectangles. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **17**, e11710.
58. Messoudi S, Destercke S, Rousseau S. 2020 Conformal multi-target regression using neural networks. In *Conformal and Probabilistic Prediction and Applications* pp. 65–83. PMLR.
59. Neeven J, Smirnov E. 2018 Conformal stacked weather forecasting. In *Conformal and Probabilistic Prediction and Applications* pp. 220–233. PMLR.
60. Zhou Y, Lindemann L, Sesia M. 2024 Conformalized Adaptive Forecasting of Heterogeneous Trajectories. In *Forty-first International Conference on Machine Learning*.

61. Sun SH, Yu R. 2024 Copula Conformal prediction for multi-step time series prediction. In *The Twelfth International Conference on Learning Representations*.
62. Gibbs I, Candes E. 2021 Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems* **34**, 1660–1672.
63. Zaffran M, Féron O, Goude Y, Josse J, Dieuleveut A. 2022 Adaptive conformal predictions for time series. In *International Conference on Machine Learning* pp. 25834–25866. PMLR.
64. Gibbs I, Candès EJ. 2024 Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research* **25**, 1–36.
65. Su X. 2025 Code and Experiments for Uncertainty Quantification in Train Delay Prediction with Conformal Prediction. Source code available at: https://github.com/jjoeysu/conformal-train-delay-uncertainty.