Robust Conformal Prediction under Joint Distribution Shift

Anonymous Author(s)
Affiliation
Address
email

Abstract

Uncertainty prevails due to the lack of knowledge about data or model, and conformal prediction (CP) predicts multiple potential targets, hoping to cover the true target with a high probability. Regarding CP robustness, importance weighting can address covariate shifts, but CP under joint distribution shifts remains more challenging. Prior attempts addressing joint shift via f-divergence ignores the nuance of calibration and test distributions that are critical for coverage guarantees. More generally, with multiple test distributions shifted from the calibration distribution, simultaneous coverage guarantees for *all* test domains requires a new paradigm. We design *Multi-domain Robust Conformal Prediction (mRCP)* that first formulates the coverage difference that importance weighting fails to capture under any joint shift. To squeeze such coverage difference and guarantee the $(1-\alpha)$ coverage in all test domains, we propose *Normalized Truncated Wasserstein distance (NTW)* to comprehensively capture the nuance of any test and calibration conformal score distributions, and design an end-to-end training algorithm incorporating NTW to provide elasticity for simultaneous coverage guarantee over distinct test domains. With diverse tasks (seven datasets) and architectures (black-box and physics-informed models), NTW strongly correlates (Pearson coefficient=0.905) with coverage differences beyond covariate shifts, while mRCP reduces coverage gap by 50% on average robustly over multiple distinct test domains.

1 Introduction

The growing data volume, enhanced computation capability, and advanced models significantly improve machine learning predictive accuracy. Nevertheless, noises, unobservable factors, and the lack of knowledge lead to uncertainty that stakeholders should ponder along model predictions when making decisions particularly in areas such as fintech [25], autonomous driving [2], traffic forecasting [4], and epidemiology [32, 27]. Conformal Prediction (CP) addresses uncertainty by predicting a set of possible target(s) rather than a single guess [31]. Specifically, CP computes conformal scores (residuals between predicted and true targets for regression tasks) of a trained model f on a calibration set, and calculates the $1-\alpha$ quantile q of these scores. For any input x, CP produces the smallest prediction set $C(x)$ consisting of target values whose conformal scores are less than q. Assuming that the test and calibration data are exchangeable (including i.i.d.), the true target y is guaranteed to be covered by $C(x)$ with at least $1-\alpha$ probability.

In practice, calibration distribution P_{XY} and test distribution Q_{XY} may differ thus $P_{XY} \neq Q_{XY}$, termed as **joint distribution shift** and violate the exchangeability assumption. Joint shift can occur with either covariate shift ($P_X \neq Q_X$) or concept shift ($P_{Y|X} \neq Q_{Y|X}$), though what causes a joint shift is difficult to infer from the observed data only. With importance weighting, covariate shift is shown not to affect the coverage confidence guarantee [29]. To address CP under joint

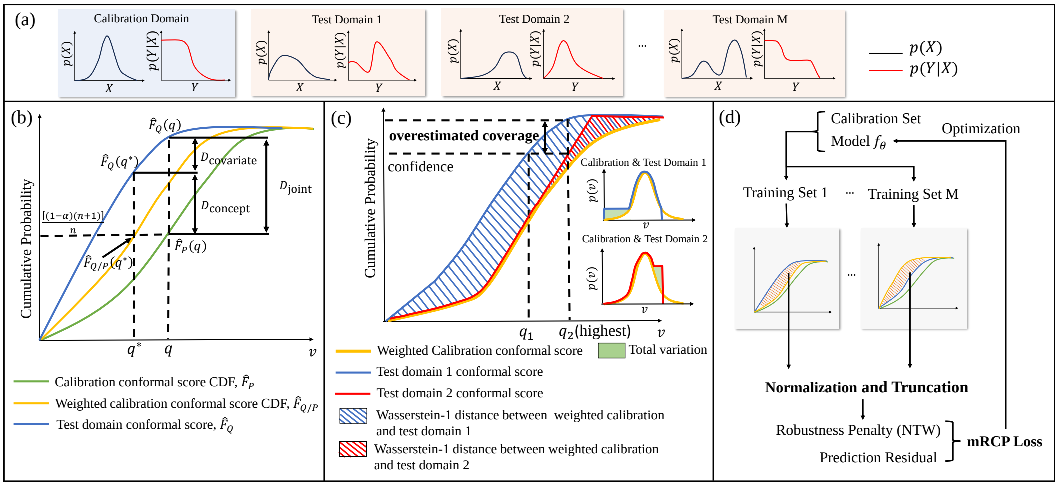

Figure 1: (a) Multiple test domains $\mathcal{E} = \{e_1, ..., e_M\}$ with joint shifts ($Q^{(e)}_{XY} \neq P_{XY}$); (b) coverage difference $D_{\text{joint}} = \hat{F}_Q(q) - \hat{F}_P(q)$ (Eq. (5)) due to $Q^{(e)}_{XY} \neq P_{XY}$ is decomposed into the $D_{\text{coviriate}}$ (Eq. (7)) caused by covariate shift ($Q^{(e)}_{X} \neq P_X$) and the remaining D_{concept} (Eq. (8)) due to concept shift ($Q^{(e)}_{Y|X} \neq P_{Y|X}$); (c) Wasserstein-1 (W-) distances (Eq. (11)) between test and weighted calibration conformal score CDFs capture the expected D_{concept} (Eq. (10)). However, f-divergence (e.g., total variation, KL divergence) does not compare two CDFs pointwisely and fails to capture such an expectation. Test domains 1 and 2 both have identical total variations to the calibration domain but different W-distances. With multiple test domains, using a single $1-\alpha$ quantile [33, 6] lead to the dilemma of CP coverage efficiency and confidence guarantee; (d) Solution: Normalized Truncated W-distance (Eq. (16)) is robust to outlier scores and different score scales across test domains, and the mRCP algorithm reduces NTW on all test domains can elastically train a model to guarantee conformal coverage for $Q^{(e)}_{Y|X} \neq P_{Y|X} \forall e \in \mathcal{E} = \{e_1, ..., e_M\}$.

shift, f-divergence is adopted in [33, 6] to measure the difference between P_{XY} and Q_{XY} or the corresponding conformal scores distributions. However, f-divergence ignores where the two distributions differ, which quantiles and coverage guarantees depend on (Figure 1, (c)). When test data are sampled from multiple distinct test distributions $Q^{(e)}_{XY}, e \in \mathcal{E} = \{e_1, ..., e_M\}$, it is desired to ensure simultaneous $1-\alpha$ coverage for all test distributions. Previous work selects the highest $1-\alpha$ quantile from all test distributions and constructs $C(x)$ for $x \in Q^{(e)}_X, \forall e \in \mathcal{E} = \{e_1, ..., e_M\}$, producing excessively large set $C(x)$. Selecting other quantiles may lead to smaller coverage on a test domain than was expected during calibration, leading to prediction overconfidence. Without a new paradigm to guarantee coverage under multiple shifted test distributions, the dilemma between CP coverage efficiency and confidence guarantee seems unavoidable.

We first decompose the coverage difference under any joint distribution shift to a component due to covariate shift ($P_X \neq Q^{(e)}_X$, addressed by importance weighting [29]) and that due to concept shift ($P_{Y|X} \neq Q^{(e)}_{Y|X}$). We propose *Normalized Truncated Wasserstein distance (NTW)* to robustly capture where the test and importance-weighted calibration conformal score cumulative density function (CDF) deviate (Figure 1, (b)). We design *Multi-domain Robust Conformal Prediction (mRCP)* by minimizing all NTW terms over $\mathcal{E} = \{e_1, ..., e_M\}$ during model training (Figure 1, (d)) to elastically guarantee coverage confidence for all test domains. Experiments on regression tasks on seven datasets demonstrate that: *1)* NTW well-correlates with the coverage difference after importance weighting (Pearson coefficient 0.905); *2)* mRCP provides conformal predictions that reduce average coverage difference by 50% compared to baselines under multiple joint shifts; *3)* mRCP is sufficiently general to address joint distribution shifts even after incorporating domain knowledge when available.

2 Background and related work

2.1 Conformal prediction

Let $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ denote the input and output random variable, respectively, where $\mathcal{X}$ and $\mathcal{Y} \subseteq \mathbb{R}$ is the input and output space, respectively. On $\mathcal{X} \times \mathcal{Y}$, the calibration domain is defined by a joint distribution P_{XY}, and we consider a calibration set $S_c = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ are drawn *i.i.d.* from P_{XY}. Similarly, a test set $S_t = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ is drawn *i.i.d.* from test domain, which is defined by a joint distribution Q_{XY}.

With a trained regression model f, the conformal score $v_i = v(x_i, y_i) = |f(x_i) - y_i|$ is the residual between the predicted target $f(x_i)$ and the true target y_i. The set of calibration conformal scores is denoted as $V_c = \{v(x_i, y_i)|(x_i, y_i) \in S_c\}$. Let q be the $\lceil(1-\alpha)(n+1)\rceil/n$ quantile of V_c:

$$q = \text{Quantile}\left(\frac{\lceil(1-\alpha)(n+1)\rceil}{n}, \frac{1}{n}\sum\nolimits_{v_i \in V_c} \delta_{v_i}\right), \tag{1}$$

where δ_{v_i} represents the point mass at v_i (i.e., the distribution placing all mass at the value v_i). Quantile$(1-\alpha, F) := \inf\{z|\Pr(Z < z) \geq 1-\alpha\}$ and F is the CDF of Z. With the quantile q, the CP prediction set of an input x from S_t is

$$C(x) = \{\hat{y} \in \mathbb{R}||f(x) - \hat{y}| \leq \ q, (x, y) \in S_t\}\,. \tag{2}$$

Most CP methods, such as[22, 23], rely on the assumption of exchangeability, which is relaxed from the i.i.d. assumption [31]. In our scenario, if the calibration and test samples are drawn from the identical joint probability distribution ($P_{XY} = Q_{XY}$), these calibration and test samples are i.i.d. Under this assumption, the probability that the true target y is included in $C(x)$ is at least $1-\alpha$, which is called **coverage guanrantee**, or more formally,

$$\Pr\left(y \in C(x)\right) \geq 1 - \alpha. \tag{3}$$

2.2 Conformal prediction under domain shift

Covariate shift ($P_X \neq Q_X$) means marginal distributions between the calibration and test domains are different. CP under covariate shift is addressed using importance weighting [29]. Under a probabilistic view, [14] defined the covariate shift as a bounded perturbation on any test input and developed adaptive probabilistically robust CP. The condition of multiple test domains is discussed in [15], and similar topics include coverages under feature-stratification [7, 11].

Joint distribution shift ($P_{XY} \neq Q_{XY}$) indicates at least one of covariate shift ($P_X \neq Q_X$) and concept shift (different conditional distributions, $P_{Y|X} \neq Q_{Y|X}$) will occur [17]. This shift is more general and the importance weighting method cannot address changes in conditional distribution. With M test domains $\mathcal{E} = \{e_1, ..., e_M\}$, each $e \in \mathcal{E}$ is defined by a joint distribution $Q^{(e)}_{XY}$ and holds a joint shift with calibration domain P_{XY} (i.e., $P_{XY} \neq Q^{(e)}_{XY}$). Considering this condition, previous works, such as [6, 33], presume all test domains fall in a predefined f-divergence range, calculate confidence-specified quantile of each test domain, and apply the highest quantile to all domains. This method causes excessively high coverages and thus overlarge prediction sets, which reduces prediction efficiency because smaller prediction sets can help locate true targets better.

3 Conformal prediction under joint distribution shift

3.1 Decomposition of coverage difference

We decompose the coverage difference between a calibration domain P_{XY} and a test domain Q_{XY} under **joint distribution shift** at a user-specified confidence $(1-\alpha)$.

Similar to V_c, we define the test conformal score set $V_t = \{v(x_i, y_i)|(x_i, y_i) \in S_t\}$. With the indicator function $\mathbb{1}$, empirical CDFs of calibration and test conformal scores are

$$\hat{F}_P(v) = \frac{1}{n}\sum\nolimits_{v_i \in V_c} \delta_v \mathbb{1}_{v_i < v}, \quad \hat{F}_Q(v) = \frac{1}{m}\sum\nolimits_{v_i \in V_t} \delta_v \mathbb{1}_{v_i < v}. \tag{4}$$

With given $1-\alpha$ confidence, quantile q is calculated in Eq. (1), and the coverage difference under a joint distribution shift can be quantified as

$$D_{\text{joint}}(q) = \hat{F}_Q(q) - \hat{F}_P(q). \tag{5}$$

[29] employs importance weighting for CP under covariate shift. Specifically, if the ratio of test to calibration covariate likelihoods, Q_X/P_X, is known, a calibration conformal score $v_i \in V_c$ is

weighted by $p_i = w(x_i)/\sum_{j=1}^{n} w(x_j)$, where $w(x_i) = Q_X(x_i)/P_X(x_i)$. Therefore, the empirical CDF of weighted empirical calibration scores is given by

$$\hat{F}_{Q/P}(v) = \sum_{i=1}^{n} p_i \delta_{v_i} \mathbb{1}_{v_i < v},$$

where the subscript Q/P indicates conformal scores of calibration domain P is weighted by conformal scores of test domain Q. The confidence-specified quantile of the weighted calibration conformal scores is

$$q^* = \text{Quantile}\left(\lceil (1-\alpha)(n+1) \rceil / n, \sum_{i=1}^{n} p_i \delta_{v_i}\right). \tag{6}$$

As importance weighting ensures the $1-\alpha$ coverage as though covariate shift were absent, coverage difference $D_{\text{covariate}}$ caused by covariate shift is the gap between the coverages under test conformal score CDF using quantiles on unweighted and weighted calibration conformal score distributions.

$$D_{\text{covariate}}(q, q^*) = \hat{F}_Q(q) - \hat{F}_Q(q^*). \tag{7}$$

Importance weighting can not address CP under joint shift as it fails to capture changes in conditional probability distribution caused by concept shift, thus we present the coverage difference caused by concept shift is

$$D_{\text{concept}}(q, q^*) = D_{\text{joint}}(q) - D_{\text{covariate}}(q, q^*) = \hat{F}_Q(q^*) - \hat{F}_P(q), \tag{8}$$

which is remaining coverage difference after applying importance weighting. Here we assume $\hat{F}_P(q) = \hat{F}_{Q/P}(q^*)$, so we can rewrite D_{concept} by

$$D_{\text{concept}}(q^*) = \hat{F}_Q(q^*) - \hat{F}_{Q/P}(q^*). \tag{9}$$

The error bound for the assumption is quite small especially when the calibration set size n is large. The detailed proof is provided in Appendix B. We denote D_{concept} as D for simplification.

3.2 Normalized Truncated Wasserstein distance

To develop a metric that is independent of confidence level and can quantify the overall closeness between weight calibration and test conformal scores, we estimate the expected coverage difference under concept shift as

$$\mathbb{E}[D] = \frac{1}{n} \sum_{v_i \in V_c} \left| \hat{F}_Q(v_i) - \hat{F}_{Q/P}(v_i) \right|, \tag{10}$$

based on the approximation in Eq. (9), where $\mathbb{E}$ indicates the expectation function.

Definition 1 (Wasserstein-1 Distance)**.** *If F_1 and F_2 are two cumulative distribution functions (CDFs), the Wasserstein-1 distance, d_{W}, is quantified by the area between F_1 and F_2.*

$$d_{\mathrm{W}}(F_1, F_2) = \int_{\mathbb{R}} |F_1(v) - F_2(v)| dx. \tag{11}$$

Applying Wasserstein-1 distance (W-distance) in Eq. (11) to $\hat{F}_Q$ and $\hat{F}_{Q/P}$, we get

$$d_{\mathrm{W}}(\hat{F}_Q, \hat{F}_{Q/P}) = \int_0^{\infty} |\hat{F}_Q(v) - \hat{F}_{Q/P}(v)| dv. \tag{12}$$

As we define conformal scores as the residuals between predicted and true targets, they are always positive, so we only need to integral from 0 to ∞ in Eq. (12).

We assume V_c is sorted. As both $\hat{F}_Q$ and $\hat{F}_{Q/P}$ are empirical CDFs, we can approximately represent $d_{\mathrm{W}}(\hat{F}_Q, \hat{F}_{Q/P})$ in a discrete form as

$$d_{\mathrm{W}}(\hat{F}_Q, \hat{F}_{Q/P}) \approx \sum_{i=1}^{n-1} \left| \hat{F}_Q(v_i) - \hat{F}_{Q/P}(v_i) \right| (v_{i+1} - v_i), \;\; v_i \in V_c. \tag{13}$$

Eq. (13) shows $d_{\mathrm{W}}(\hat{F}_Q, \hat{F}_{Q/P})$ can be estimated as a weighted summation of $|\hat{F}_Q(v_i) - \hat{F}_{Q/P}(v_i)|$ for $v_i \in V_c \backslash \{v_n\}$ with the corresponding weight $v_{i+1} - v_i$. Also, Eq. (10) indicates that $\mathbb{E}[D]$ can be regarded as the weighted summation of $|\hat{F}_Q(v_i) - \hat{F}_{Q/P}(v_i)|$ for $v_i \in V_c$ with weight $1/n$. The similarity between Eq. (13) and Eq. (10) allows us to apply the W-distance between the test and weighted calibration conformal score to capture expected coverage difference under concept shift.

Care needs to be taken for Eq. (13) to make this metric more robust. At first, we expect the weights $v_{i+1} - v_i$ to be approximately equal, as weights in Eq. (10) are constants $1/n$. However, some outlier calibration conformal scores have large distances from their neighbors, causing involved weights much higher than $1/n$. These outlier scores are represented as a long tail of $\hat{F}_{Q/P}$ when it converges to 1. Therefore, it is necessary to establish a partition threshold to truncate the long tail. We calculate the partition threshold

$$v_\sigma = \inf \left\{ v_i | \hat{F}_{Q/P}(v_i) \geq 1 - \sigma, v_i \in V_c \right\}, \tag{14}$$

which is the smallest calibration conformal score whose coverage is greater or equal to a user-defined value $1 - \sigma$. In contrast to the original $d_{\mathrm{W}}(\hat{F}_Q, \hat{F}_{Q/P})$ integrated on the set of real numbers, the truncated form is integrated from 0 to v_σ as

$$d_{\mathrm{TW}}(\hat{F}_Q, \hat{F}_{Q/P}) = \int_0^{v_\sigma} |\hat{F}_Q(v) - \hat{F}_{Q/P}(v)| dv. \tag{15}$$

Secondly, as the summation of weights in Eq. (10) is 1, we also need to divide each $v_{i+1} - v_i$ by $v_\sigma - v_1$. When the calibration set is large enough, it is plausible to assume the existence of a calibration sample fitting the trained model f very well, causing the smallest calibration conformal score $v_1 \approx 0$. Therefore, this normalized can be formulated as

$$d_{\mathrm{NTW}}(\hat{F}_Q, \hat{F}_{Q/P}) = \frac{1}{v_\sigma} \int_0^{v_\sigma} |\hat{F}_Q(v) - \hat{F}_{Q/P}(v)| dv. \tag{16}$$

A lower d_{NTW} indicates more similarity between $\hat{F}_{Q/P}$ and $\hat{F}_Q$, thus leading to more robust conformal prediction in the test domain. As a result, NTW enables us to assess the expected coverage difference due to concept shift in Eq. (10). Experiment results in Section 5 and Appendix E show the necessity of truncation and normalization. We also prove that the W-distance between the test and weighted calibration conformal score population CDF can establish an upper bound for coverage difference under concept shift in Appendix C.

4 Multi-domain robust conformal prediction

If a calibration set S_c, and a test set S_t are drawn from a domain P_{XY}, the i.i.d. assumption is satisfied, and the coverage guarantee in Eq. (3) holds for $(x, y) \in S_t$.

The domain P_{XY} can be decomposed into M multiple domains, denoted as $\mathcal{E} = \{e_1, ..., e_M\}$.

$$P_{XY}(x, y) = \frac{1}{M} \sum\nolimits_{e \in \mathcal{E}} Q_{XY}^{(e)}(x, y) \tag{17}$$

However, for $e \in \mathcal{E}$, denote $S_t^{(e)}$ a test set drawn from $Q_{XY}^{(e)}$, then the coverage guarantee may no longer hold for $(x, y) \in S_t^{(e)}$, because joint distribution shift may occur between P_{XY} and $Q_{XY}^{(e)}$. It indicates CP can be overconfident and underconfident for samples from different $Q_{XY}^{(e)}$, resulting in prediction biases.

Inspired by the works of multi-domain generalization [26, 18, 19, 1], we propose **Multi-domain Robust Conformal Prediction (mRCP)** to make the coverage approach confidence in all domains, using a training set $S^{(e)}$ from the data distribution $Q_{XY}^{(e)}$ for $e \in \mathcal{E}$ and a calibration set S_c from P_{XY}.

The objective function of mRCP includes two components. First, for the minimization of prediction residuals, denoting l a loss function, Empirical Risk Minimization (ERM) [30] is incorporated as

$$\mathcal{L}_{\mathrm{ERM}}(\theta) = \sum\nolimits_{e \in \mathcal{E}} \mathcal{L}^{(e)}(\theta) = \sum\nolimits_{e \in \mathcal{E}} \mathbb{E}_{(x_i, y_i) \sim S^{(e)}} \left[l(f_\theta(x_i), y_i) \right]. \tag{18}$$

Secondly, we aim for robust conformal prediction on each domain during testing, seeking a low value of $\mathbb{E}[D]$ in Eq. (10) across test domains, so mRCP needs to address coverage differences due to covariate and concept shifts simultaneously. To remove coverage differences due to covariate shifts, it applies importance weighting to each domain $e \in \mathcal{E}$ during training and obtains $\hat{F}_{Q^{(e)}/P}$, which is the calibration conformal score CDF weighted by $Q_{XY}^{(e)}$.

Besides, as we have a training set $S^{(e)}$ from domain $Q_{XY}^{(e)}$, an empirical CDF of conformal scores in $Q_{XY}^{(e)}$ can be computed, denoted as $\hat{F}^{tr}_{Q^{(e)}}$. NTW quantifies the expected coverage difference caused by concept shift between $\hat{F}_{Q^{(e)}/P}$ and training conformal score CDF $\hat{F}^{tr}_{Q^{(e)}}$. Combining these two components, the objective function of mRCP is

$$\mathcal{L}_{\text{mRCP}}(\theta) = \sum\nolimits_{e\in\mathcal{E}} \mathcal{L}^{(e)}(\theta) + \beta \sum\nolimits_{e\in\mathcal{E}} d_{\text{NTW}}(\hat{F}^{tr}_{Q^{(e)}}, \hat{F}_{Q^{(e)}/P}), \tag{19}$$

where β is a hyperparameter balancing these two parts. mRCP algorithm is shown in Algorithm 1.

Algorithm 1 Multi-domain Robust Conformal Prediction

Require: M training sets $S^{(e)}$, $e \in \mathcal{E}$; one calibration set S_c; N training epochs; model f_θ; partition value σ; loss function l; penalty hyperparameter β.

for $e \in \mathcal{E}$ **do**
  for $(x_i, y_i) \in S_c$ **do**
    $w(x_i) = \frac{Q_X^{(e)}(x_i)}{P_X(x_i)}$, $p_{(i,e)} = \frac{w(x_i)}{\sum_{j=1}^{n} w(x_j)}$ ▷ Covariate shift between $Q_{XY}^{(e)}$ and P_{XY}
  end for
end for

for $i = 1$ to N **do**
  $V_c = \{v(x_i, y_i) | (x_i, y_i) \in S_c\}$ ▷ Calibration score set
  for $e \in \mathcal{E}$ **do**
    $\mathcal{L}^{(e)}(\theta) = \mathbb{E}_{(x_i,y_i)\sim S^{(e)}}\left[l(f_\theta(x_i), y_i)\right]$ ▷ ERM loss of domain e
    $V^{(e)} = \left\{v(x_i, y_i) | (x_i, y_i) \in S^{(e)}\right\}$ ▷ Training score set of domain e
    $\hat{F}^{tr}_{Q^{(e)}} = \sum_{v_i \in V^{(e)}} \delta_{v_i} \mathbb{1}_{v_i \le v}$ ▷ Training score CDF of domain e
    $\hat{F}_{Q^{(e)}/P}(v) = \sum_{v_i \in V_c} p_{(i,e)} \delta_{v_i} \mathbb{1}_{v_i \le v}$ ▷ Calibration score CDF weighted by $Q_{XY}^{(e)}$
    $v_\sigma = \inf\left\{\hat{F}_{Q^{(e)}/P}(v_i) \ge 1 - \sigma, v_i \in V_c\right\}$ ▷ Truncation threshold
    $d_{\text{NTW}}\left(\hat{F}^{tr}_{Q^{(e)}}, \hat{F}_{Q^{(e)}/P}\right) = \frac{1}{v_\sigma}\int_0^{v_\sigma} \left|\hat{F}^{tr}_{Q^{(e)}}(v) - \hat{F}_{Q^{(e)}/P}(v)\right| dv$ ▷ NTW calculation
  end for
  Optimize f_θ based on $\mathcal{L}_{\text{mRCP}}(\theta) = \sum_{e\in\mathcal{E}} \mathcal{L}^{(e)}(\theta) + \beta \sum_{e\in\mathcal{E}} d_{\text{NTW}}\left(\hat{F}^{tr}_{Q^{(e)}}, \hat{F}_{Q^{(e)}/P}\right)$
end for

5 Experiment

In this section, we validate NTW in Eq. (16) as a good indicator of expected coverage difference due to concept shift and demonstrate the effectiveness of mRCP in obtaining coverage robustness across different test domains.

5.1 Datasets and models

We conducted experiments across various datasets: (a) the airfoil self-noise dataset [5]; (b) Seattle-loop [9], PeMSD4, PeMSD8 [16] for traffic speed prediction; (c) US-Regions, US-States, and Japan-Prefectures [10] for epidemic spread forecasting. The airfoil dataset was manually altered to create three subsets demonstrating covariate and concept shifts. 24 domains for the traffic datasets were designated based on data generation hours, while epidemic dataset instances were categorized into four domains reflecting different pandemic stages. A multilayer perceptron (MLP) with a (input dimension, 64, 64, 1) architecture was utilized for all datasets. Traffic and epidemic prediction tasks were also trained on corresponding physics-informed partial differential equations (PDEs), which are the Susceptible-Infected-Recovered (SIR) model and the Reaction-Diffusion (RD) model respectively. We refer to Appendix D for detailed experiment setups.

5.2 Experiments of NTW

For each of the experiment setups, a training set, a validation set, and a test set were sampled from each $Q_{XY}^{(e)}$ for $e \in \mathcal{E}$. One calibration set was sampled from P_{XY} which is a mixture probability distribution of $Q_{XY}^{(e)}$ for $e \in \mathcal{E}$, as shown in Eq. (17). To validate NTW is a good indicator of $\mathbb{E}[D]$, we only need to use ERM in Eq. (18) to train the model f_θ, which can be an MLP or a PDE. The loss function l is the ℓ_1 norm, as same as how we compute conformal scores.

After training, for $e \in \mathcal{E}$, we first calculated the NTW between the calibration conformal score CDF weighted by $Q_X^{(e)}/P_X$, and validation conformal score CDF of $Q_X^{(e)}$. Denote the NTW of domain e as $d_{\text{NTW}}^{(e)}$. Then, we estimated the expected coverage difference caused by concept shift on a test domain e, denoted as $\mathbb{E}_\alpha[D^{(e)}]$, using the coverage difference expectation between the test and weighted calibration conformal score CDFs on a $1-\alpha$ confidence set $\{0.1, ..., 0.9\}$.

$\mathbb{E}_\alpha[D^{(e)}]$ and $d_{\text{NTW}}^{(e)}$should have a positive correlation for $e \in \mathcal{E}$, proving NTW can capture the expected coverage difference caused by concept shift.

Baselines: We select six baseline metrics to validate the effectiveness of NTW. Total variation d_{TV} [13], and Kullback-Leibler (KL) divergence d_{KL} [21] are chosen as two typical f-divergence metrics. Expectation difference $\Delta\mathbb{E}$ [19] is selected since it is a widely applied generalization metric. We also measure standard, normalized, and truncated W-distance, denoted as d_{W}, d_{NW}, and d_{TW} respectively, to demonstrate applying normalization and truncation together is necessary.

Metric: We apply the **Pearson coefficient** to quantify the correlations between metrics and the coverage difference expectation. It measures the linear correlation between two values by giving a value between -1 and 1 inclusive. 1,0, and -1 indicate perfect positive linear, no linear, and negative linear correlations, respectively. Therefore, if the Pearson coefficient of a metric is **higher**, this metric can indicate the expected coverage difference **better**. We provide a detailed definition of the Pearson coefficient in Appendix E.

Table 1: Pearson coefficients between metrics and coverage difference expectation under concept shift

Dataset	Model	d_{NTW}	d_{TV}	d_{KL}	$\Delta\mathbb{E}$	d_{W}	d_{NW}	d_{TW}
Airfoil	MLP	**1.000**	-0.356	-0.545	0.891	0.878	0.951	0.967
Seattle-loop	MLP	**0.971**	0.461	0.054	0.781	0.759	0.762	0.765
	PDE	**0.996**	0.890	0.058	0.897	0.893	0.909	0.921
PeMSD4	MLP	**0.992**	0.846	-0.390	0.926	0.915	0.964	0.941
	PDE	**0.986**	0.682	-0.068	0.858	0.872	0.928	0.858
PeMSD8	MLP	**0.905**	0.397	-0.089	0.333	0.267	0.371	0.529
	PDE	**0.827**	0.129	-0.114	0.253	0.118	0.141	0.527
US-States	MLP	**0.999**	0.966	0.965	0.872	0.885	0.912	0.931
	PDE	**0.999**	0.966	0.964	0.817	0.848	0.890	0.899
US-Regions	MLP	**0.636**	-0.530	-0.338	-0.205	-0.308	-0.352	-0.405
	PDE	**0.709**	0.308	0.350	0.484	0.355	0.322	0.137
Japan-Prefectures	MLP	**0.996**	0.986	0.988	0.943	0.948	0.954	0.950
	PDE	**0.997**	0.983	0.981	0.907	0.918	0.935	0.924
Average		**0.905**	0.574	0.325	0.619	0.583	0.607	0.629
Standard Deviation		**0.128**	0.474	0.562	0.368	0.420	0.437	0.428

Results: Table 1 illustrates the Pearson coefficients between NTW and the coverage difference expectation among seven datasets and different models, compared with the other six baseline metrics. We highlight that NTW keeps holding the largest Pearson coefficient among all experiment setups, which means the proposed metric can keep indicating the coverage difference expectation. Specifically, the coefficients of total variation d_{TV} and KL divergence d_{KL} fluctuate along experiments, meaning that they can not truly indicate the coverage difference expectation. $\Delta\mathbb{E}$ can not capture the coverage difference expectation either. Lastly, due to the lack of robustness to score scales and outliers, standard, normalized, and truncated W-distance, denoted as d_{W}, d_{NW}, and d_{TW} respectively, can not indicate the coverage difference expectation as well as d_{NTW}. It also displays the average and standard deviation of the Pearson coefficient of the proposed NTW and six baselines. NTW not only has the highest average Pearson coefficient but also has the lowest standard deviation, which means the correlation between NTW and the coverage difference expectation caused by concept shift is very stable. In Figure 3 and Figure 4, we also visually show the correlation between the expected coverage difference under concept shift and each metric. We refer to Appendix E for detailed analysis. This observation suggests the potential of incorporating NTW in the training process, leading to the development of the mRCP approach. By applying the NTW metric, mRCP aims to enhance coverage robustness in test domains.

5.3 Experiments of mRCP

Since we prove NTW can assess expected coverage difference under concept shift effectively, mRCP is designed to minimize it during training. In this case, validation sets are unnecessary, and we only draw training, and test sets from $Q_{XY}^{(e)}$. Again, we draw one calibration set from P_{XY}. The model f_θ can also be an MLP or PDE based on different experiment setups. The loss function l is the ℓ_1 norm. We implement mRCP according to Algorithm 1.

Baselines: Two methods of optimization with out-of-distribution data are selected as baselines. **DRO** in Eq. (20) by [26] follows the minimax principle to reduce the highest $\mathcal{L}^{(e)}$ to obtain fair prediction among test distributions. On the other hand, **V-REx** in Eq. (21), introduced by [18], focuses on reducing the variance of $\mathcal{L}^{(e)}$ to obtain fairness. As we include importance weighting in mRCP, we do not take it as a baseline, and the effectiveness of importance weighting is discussed in Section 6.

$$\mathcal{L}_{\text{DRO}}(\theta) = \max_{e\in\mathcal{E}} \mathcal{L}^{(e)}. \tag{20}$$

$$\mathcal{L}_{\text{V-REx}}(\theta) = \sum\nolimits_{e\in\mathcal{E}} \mathcal{L}^{(e)} + \beta \operatorname{Var}(\mathcal{L}^{(e)} \mid e \in \mathcal{E}). \tag{21}$$

Metric: Denote $\mathbb{E}'_e[\mathbb{E}_\alpha[D^{(e)}]]$ the **expectation of coverage difference** over confidence levels and test domains and $\mathbb{E}'_e[\mathcal{L}^{(e)}]$ the **expectation of prediction residual** over test domains. The two expectations become **smaller** means the algorithm's performance is **better**. Both values are normalized by the corresponding results from the same experiment setup trained by ERM. Changing the weight β in Eq. (19) will draw a Pareto front, thus we want the Pareto front closer to the origin. Since V-REX is also controlled by a hyperparameter, we draw Pareto fronts for it as well.

Result: Figure 2 displays the Pareto fronts for mRCP, DRO, and V-REx, highlighting the trade-offs between prediction residual and coverage difference expectation across different models and datasets. Figure 2, (a) shows the results for the airfoil self-noise dataset when trained with a Multilayer Perceptron (MLP) model. The mRCP method achieves a more favorable Pareto front compared to V-REx, indicating a better balance between prediction residual and coverage difference expectation. Additionally, mRCP attains a lower normalized coverage difference expectation than DRO at a comparable level of the prediction residual. In Figure 2, (b), we observe the experiment results on the epidemic spread prediction task using three epidemic datasets. With the same MLP architecture, mRCP delivers superior Pareto fronts relative to the baselines. When employing the epidemic PDE, the SIR model only has two trainable parameters, so their data points can not compose Pareto curves due to the model's limited flexibility. Thus, we show the average of these points. Despite this limitation, mRCP maintains its advantage over the baseline methods. Figure 2, (c) and (d) present results from the traffic prediction task on three different traffic datasets. Here, the Pareto curves for both the MLP and the reaction-diffusion (RD) PDE model are well-defined, because RD model with six parameters, offers greater adaptability, allowing for clearer Pareto fronts. Overall, Figure 2 collectively indicates that mRCP consistently achieves lower coverage difference expectations without compromising prediction residual as significantly as DRO and V-REx in different tasks and datasets.

6 Discussion

mRCP can distinguish coverage differences under concept shift and covariate shift. A notable feature of the mRCP Pareto curves depicted in Figure 2 is their results when β is small, which are not at $\mathbb{E}'_e[\mathbb{E}_\alpha[D^{(e)}]] = 1$, unlike the Pareto curves of V-REx. This is because, during training, mRCP

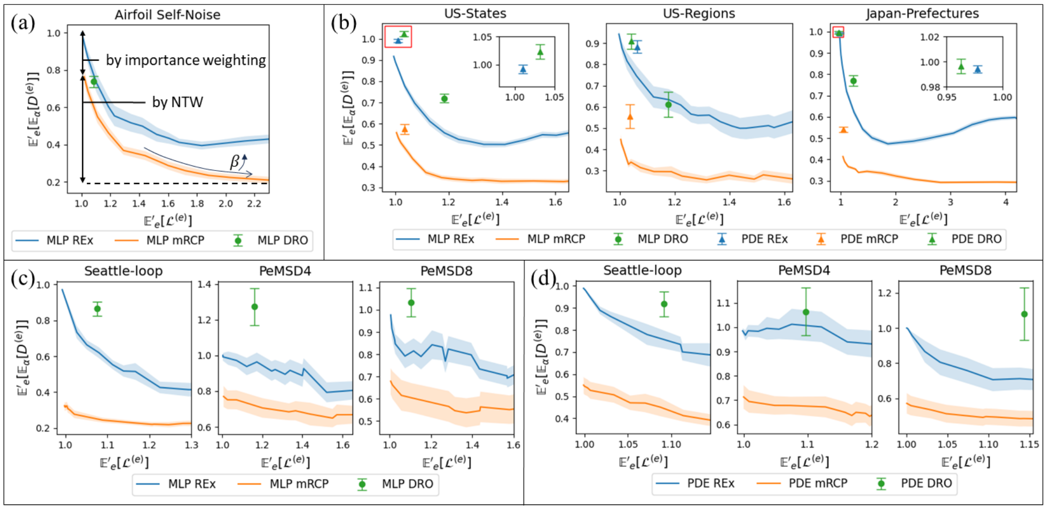

Figure 2: **Pareto fronts of Multi-domain Robust Conformal Prediction(mRCP), compared with DRO and V-REx:** Experimental results of (a) airfoil self-noise example, (b) epidemic spread prediction, and (c) (d) traffic speed prediction. mRCP always reaches a smaller coverage difference expectation than DRO and V-REx with less increase in prediction residual. Red boxes in (b) are zoomed-in areas. Shadow areas and error bars indicate the standard error.

has considered the coverage difference under covariate shift by applying importance weighting to calibration conformal score CDF. Consequently, as β in Eq. (19) increases, the NTW term is only trained to mitigate the coverage difference under the concept shift, as shown in Figure 2,(a).

DRO and V-REx are defeated because of improper selection of optimization metrics. Examining Eq. (20) and Eq. (21), we can see both baselines aim to promote fairness by equalizing the expected losses across different domains. As the loss function is ℓ_1 norm, which is identical to how conformal scores are calculated, the experiment results of $\Delta\mathbb{E}$ in the last row of Figure 3 show this metric is ineffective in capturing the coverage difference due to concept shift.

Nonetheless, mRCP's limitations arise from the inherent challenges associated with penalty-based optimization algorithms. Whether it is mRCP or V-REx, penalty-based optimization algorithms necessitate a model with a high capacity for fitting complex patterns. For instance, in Figure 2, (b), the Pareto curves are not discernible when predictions are derived from an epidemic PDE (SIR model) with only two adjustable parameters. In contrast, as shown in Figure 2, (d), the traffic PDE (RD model) demonstrates greater flexibility and adaptability with six tunable parameters, exhibiting distinct Pareto curves.

7 Conclusion

This study begins by decomposing the coverage difference caused by covariate and concept shifts. We then introduce the Normalized Truncated Wasserstein distance (NTW) as a metric for capturing coverage difference expectation under concept shift by comparing the test and weighted calibration conformal score CDFs. This metric can indicate the discrepancy position in calibration and test score distributions. Normalization and truncation make the metric score scales and outliers. Finally, we develop an end-to-end algorithm called Multi-domain Robust Conformal Prediction (mRCP) that incorporates NTW during training, allowing coverage to approach confidence in all test domains.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Mrinal R Bachute and Javed M Subhedar. Autonomous driving architectures: insights of machine learning and deep learning algorithms. *Machine Learning with Applications*, 6:100164, 2021.

[3] Leonardo Bellocchi and Nikolas Geroliminis. Unraveling reaction-diffusion-like dynamics in urban congestion propagation: Insights from a large-scale road network. *Scientific reports*, 10(1):4876, 2020.

[4] Azzedine Boukerche and Jiahao Wang. Machine learning-based traffic prediction models for intelligent transportation systems. *Computer Networks*, 181:107530, 2020.

[5] Pope D. Brooks, Thomas and Michael Marcolini. Airfoil Self-Noise. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5VW2C.

[6] Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, pages 1–66, 2024.

[7] Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of machine learning research*, 22(81):1–42, 2021.

[8] Ian Cooper, Argha Mondal, and Chris G Antonopoulos. A sir model assumption for the spread of covid-19 in different communities. *Chaos, Solitons & Fractals*, 139:110057, 2020.

[9] Zhiyong Cui, Kristian Henrickson, Ruimin Ke, and Yinhai Wang. Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[10] Songgaojun Deng, Shusen Wang, Huzefa Rangwala, Lijing Wang, and Yue Ning. Cola-gnn: Cross-location attention based graph neural networks for long-term ili prediction. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 245–254, 2020.

[11] Shai Feldman, Stephen Bates, and Yaniv Romano. Improving conditional coverage via orthogonal quantile regression. *Advances in neural information processing systems*, 34:2060–2071, 2021.

[12] Robert E. Gaunt and Siqi Li. Bounding kolmogorov distances through wasserstein and related integral probability metrics, 2022.

[13] Amanda Gentzel, Dan Garant, and David Jensen. The case for evaluating causal models using interventional measures and empirical data. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Subhankar Ghosh, Yuanjie Shi, Taha Belkhouja, Yan Yan, Jana Doppa, and Brian Jones. Probabilistically robust conformal prediction. In *Uncertainty in Artificial Intelligence*, pages 681–690. PMLR, 2023.

[15] Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

[16] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 922–929, 2019.

[17] Wouter M Kouw and Marco Loog. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018.

[18] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[19] Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.

[20] Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[21] Harsh Parikh, Carlos Varjao, Louise Xu, and Eric Tchetgen Tchetgen. Validating causal inference methods. In *International conference on machine learning*, pages 17346–17358. PMLR, 2022.

[22] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Neural Information Processing Systems*, 2019.

[23] Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. *arXiv: Methodology*, 2020.

[24] Nathan Ross. Fundamentals of stein's method. 2011.

[25] Hyun-Sun Ryu and Kwang Sun Ko. Sustainable development of fintech: Focused on uncertainty and perceived quality issues. *Sustainability*, 12(18):7669, 2020.

[26] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[27] Silvia Seoni, Vicnesh Jahmunah, Massimo Salvi, Prabal Datta Barua, Filippo Molinari, and U Rajendra Acharya. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). *Computers in Biology and Medicine*, page 107441, 2023.

[28] Yue Sun, Chao Chen, Yuesheng Xu, Sihong Xie, Rick S Blum, and Parv Venkitasubramaniam. Reaction-diffusion graph ordinary differential equation networks: Traffic-law-informed speed prediction under mismatched data. The 12th International Workshop on Urban Computing, held in conjunction with . . . , 2023.

[29] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

[30] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

[31] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[32] Timothy L Wiemken and Robert R Kelley. Machine learning in epidemiology and health outcomes research. *Annu Rev Public Health*, 41(1):21–36, 2020.

[33] Xin Zou and Weiwei Liu. Coverage-guaranteed prediction sets for out-of-distribution data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17263–17270, 2024.

A Related work

Table 2: Related works and mRCP

Task	Number of Test Domains	Test Domain Property	Work
Adaptive Conformal Prediction under Exchangeability	1	Identical to Calibration Domain	[22, 23]
Conformal Prediction under Covariate Shift	1	Covariate Shift	[29, 14]
Multi-Domain Conformal Prediction	Multiple	Feature-stratified	[7, 11]
		Covariate Shift	[15]
		Joint Distribution Shift in Certain F-divergence Range	[33, 6]
		Joint Distribution Shift	mRCP

B Error bound for the assumption of identical coverages

According to the computation of q and q^* in Eq. (1) and Eq. (6), respectively, we can define the coverages in unweighted and weighted calibration score distributions as

$$\hat{F}_P(q) = \inf\left\{\hat{F}_P(v_i)|\hat{F}_P(v_i) \geq \lceil(1-\alpha)(n+1)\rceil/n, v_i \in V_c\right\},$$
$$\hat{F}_{Q/P}(q^*) = \inf\left\{\hat{F}_{Q/P}(v_i)|\hat{F}_{Q/P}(v_i) \geq \lceil(1-\alpha)(n+1)\rceil/n, v_i \in V_c\right\}.$$

Denoting $q_+ = \inf\{v_i|v_i \in V_c, v_i > q\}$ and $q_+^* = \inf\{v_i|v_i \in V_c, v_i > q^*\}$, we can bound $\hat{F}_P(q)$ and $\hat{F}_{Q/P}(q^*)$ as

$$\hat{F}_P(q) \in \left[\lceil(1-\alpha)(n+1)\rceil/n, \hat{F}_P(q_+)\right), \ \ \hat{F}_{Q/P}(q^*) \in \left[\lceil(1-\alpha)(n+1)\rceil/n, \hat{F}_{Q/P}(q_+^*)\right).$$

Therefore, the absolute difference between $\hat{F}^*(q^*)$ and $\hat{F}(q)$ is bounded by

$$|\hat{F}_{Q/P}(q^*) - \hat{F}_P(q)| < \max\left(\hat{F}_{Q/P}(q_+^*) - \lceil(1-\alpha)(n+1)\rceil/n, \hat{F}_P(q_+) - \lceil(1-\alpha)(n+1)\rceil/n\right).$$

Especially, when the calibration set size n is large enough (like having thousands of samples), $\hat{F}_{Q/P}$ and $\hat{F}_P$ will be quite smooth, the upper above will be even negligible, allowing us to assume $\hat{F}_{Q/P}(q^*) = \hat{F}_P(q)$.

C Upper bound of coverage difference under concept shift

In this section, we prove that the W-distance between a test and weighted calibration conformal score population CDF can establish an upper bound for coverage difference under concept shift.

As D quantifies the absolute difference between $\hat{F}_{Q/P}$ and $\hat{F}_Q$ at a calibration conformal score, it can be constrained by an upper bound given by the Kolmogorov distance [12] defined as follows.

Definition 2 (Kolmogorov Distance)**.** *If F_1 and F_2 are two cumulative distribution functions (CDFs), the Kolmogorov distance, d_{K}, is defined as the maximum absolute difference between the CDFs.*

$$d_{\mathrm{K}}(F_1, F_2) = \sup_{v \in \mathbb{R}} |F_1(v) - F_2(v)|.$$

As $\hat{F}_{Q/P}$ and $\hat{F}_Q$ are empirical (not population) CDFs of weighted calibration and test conformal scores, the bounding relationship can be reformulated as

$$d_{\mathrm{K}}(\hat{F}_Q, \hat{F}_{Q/P}) = \sup_{v \in V_c \cup V_t} |\hat{F}_Q(v) - \hat{F}_{Q/P}(v)| \geq \sup_{v \in V_c} |\hat{F}_Q(v) - \hat{F}_{Q/P}(v)| = \sup_{v \in V_c} |D(v)|. \quad (22)$$

The upper bound $d_{\mathrm{K}}(\hat{F}_Q, \hat{F}_{Q/P})$ depends on the two conformal score sets V_c and V_t, indicating that the inclusion of samples in S_c and S_t is likely to introduce variability in $d_{\mathrm{K}}(\hat{F}_Q, \hat{F}_{Q/P})$. Nevertheless, we aim for an upper bound that is not reliant on specific samples and relies on the calibration and test conformal score **population** CDFs, F_P and F_Q.

Firstly, we convert the upper limit in Eq. (22) into terms of F_P and F_Q. Denoting the joint probability density function (PDF) of features and score in the calibration and test domain as $\mathbf{p}_{XV}$ and $\mathbf{q}_{XV}$ respectively, the corresponding continuous CDFs of conformal scores are illustrated as

$$F_P(v) = \int_0^v \int_{\mathcal{X}} \mathbf{p}_{XV}(u,t)dudt, \;\; F_Q(v) = \int_0^v \int_{\mathcal{X}} \mathbf{q}_{XV}(u,t)dudt, \tag{23}$$

where $\mathcal{X}$ is the space of the feature variable X.

PDFs of features in calibration and test domains, denoted as $\mathbf{p}_X$ and $\mathbf{q}_X$ respectively, are defined as

$$\mathbf{p}_X = \int_{\mathbb{R}} \mathbf{p}_{XV}(u,t)dt, \;\; \mathbf{q}_X = \int_{\mathbb{R}} \mathbf{q}_{XV}(u,t)dt. \tag{24}$$

To address the coverage difference due to covariate shift, importance weighting from [29] is rewritten as $w = \frac{\mathbf{q}_X}{\mathbf{p}_X}$. Also, normalization is unnecessary, because w here is a correction function to transform the marginal distribution of $\mathbf{p}_X$ into $\mathbf{q}_X$. The weighted version of $\mathbf{p}_{XV}$ is denoted as $\mathbf{p}'_{XV} = w\mathbf{p}_{XV} = \mathbf{q}_X\mathbf{p}_{V|X}$, which can be applied to derive the weighted continuous CDF of calibration conformal score by

$$F_{Q/P}(v) = \int_0^v \int_{\mathcal{X}} \mathbf{p}'_{XV}(u,t)dudt = \int_0^v \int_{\mathcal{X}} \mathbf{q}_X(u)\mathbf{p}_{V|X}(u,t)dudt. \tag{25}$$

The Kolmogorov distance between $F_{Q/P}$ and F_Q is $d_{\mathrm{K}}(F_Q, F_{Q/P}) = \sup_{v\in\mathbb{R}} |F_Q(v) - F_{Q/P}(v)|$.

Theorem 1 (Triangular Inequality for Kolmogorov Distance)**.** *If F_1, F_2, and F_3 are three cumulative distribution functions (CDFs), their Kolmogorov distances follow this inequality:*

$$d_{\mathrm{K}}(F_1, F_3) \leq d_{\mathrm{K}}(F_1, F_2) + d_{\mathrm{K}}(F_2, F_3).$$

Proof. Consider any point $x \in \mathbb{R}$, then we have $|F_1(x) - F_3(x)| \leq |F_1(x) - F_2(x)| + |F_2(x) - F_3(x)|$.This inequality holds due to the triangle inequality for absolute values. Now, taking the supremum over all x, we have $\sup_{x\in\mathbb{R}} |F_1(x) - F_3(x)| \leq \sup_{x\in\mathbb{R}} (|F_1(x) - F_2(x)| + |F_2(x) - F_3(x)|)$. Note that the right-hand side is not necessarily equal to the sum of the suprema of the individual terms, because the points at which the suprema of $|F_1(x) - F_2(x)|$ and $|F_2(x) - F_3(x)|$ are attained may be different. However, we know that for any x, $|F_1(x) - F_2(x)|$ is at most $d_K(F_1, F_2)$ and $|F_2(x) - F_3(x)|$ is at most $d_K(F_2, F_3)$. Therefore, $\sup_{x\in\mathbb{R}} |F_1(x) - F_3(x)| \leq d_K(F_1, F_2) + d_K(F_2, F_3)$. Since the left-hand side is the definition of $d_K(F_1, F_3)$, we can demonstrate that $d_{\mathrm{K}}(F_1, F_3) \leq d_{\mathrm{K}}(F_1, F_2) + d_{\mathrm{K}}(F_2, F_3)$. □

As Kolmogorov distance satisfies the triangular inequality theorem, as shown and proved in Theorem 1, the triangular inequality relationship can be expanded to

$$d_{\mathrm{K}}(\hat{F}_Q, \hat{F}_{Q/P}) \leq d_{\mathrm{K}}(F_{Q/P}, \hat{F}_{Q/P}) + d_{\mathrm{K}}(F_Q, F_{Q/P}) + d_{\mathrm{K}}(\hat{F}_Q, F_Q). \tag{26}$$

Secondly, the Kolmogorov distance between an empirical CDF and its corresponding population CDF can be constrained by Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [20], defined in Definition 3.

Definition 3 (Dvoretzky–Kiefer–Wolfowitz (DKW) Inequality)**.** *If F is a population cumulative distribution function (CDF), and $\hat{F}$ is an empirical CDF with n samples of a random variable X, then for any $\epsilon \geq \sqrt{\frac{1}{2n}\ln 2}$, the following inequality holds.*

$$\Pr(d_{\mathrm{K}}(\hat{F}, F) > \epsilon) \leq e^{-2n\epsilon^2}.$$

Based on Definition 3, saying $|V_c| = n$ and $|V_t| = m$, we can apply DKW inequality to $d_{\mathrm{K}}(\hat{F}_{Q/P}, F_{Q/P})$ and $d_{\mathrm{K}}(\hat{F}_Q, F_Q)$ as follows, for $\epsilon \geq \sqrt{\frac{1}{2n}\ln 2}$ and $\rho \geq \sqrt{\frac{1}{2m}\ln 2}$.

$$\Pr(d_{\mathrm{K}}(\hat{F}_{Q/P}, F_{Q/P}) \leq \epsilon) > e^{-2n\epsilon^2}, \;\; \Pr(d_{\mathrm{K}}(\hat{F}_Q, F_Q) \leq \rho) > e^{-2m\rho^2}.$$

If the two events $d_{\mathrm{K}}(\hat{F}_{Q/P}, F_{Q/P}) < \epsilon$ and $d_{\mathrm{K}}(\hat{F}_Q, F_Q) < \rho$ are independent, the inequality in Eq. (26) can be expanded in Eq. (27), which holds with at least probability $e^{-2(n\epsilon^2+m\rho^2)}$. By applying DKW inequality, we successfully quantify the variability of $d_{\mathrm{K}}(\hat{F}_Q, \hat{F}_{Q/P})$ in Eq. (22) as a form of a probable event, and use the population conformal score CDFs to limit the worst-case of coverage difference under concept shift.

$$d_{\mathrm{K}}(\hat{F}_Q, \hat{F}_{Q/P}) \leq d_{\mathrm{K}}(F_Q, F_{Q/P}) + \rho + \epsilon. \tag{27}$$

Finally, having established in Eq. (13) that the W-distance can serve as an estimator for coverage difference expectation, we explore whether Eq. (27) may similarly be bounded by this metric. The W-distance of the two population conformal score CDFs are explicitly shown as

$$\begin{aligned} d_{\mathrm{W}}(F_Q, F_{Q/P}) &= \int_{\mathbb{R}} \left|F_Q(v) - F_{Q/P}(v)\right| dv = \int_{\mathbb{R}} \left| \int_0^v \int_{\mathbb{R}} \mathbf{q}_{XV}(u,t)dudt - \int_0^v \int_{\mathbb{R}} \mathbf{p}'_{XV}(u,t)dudt \right| dv \\ &= \int_{\mathbb{R}} \left| \int_0^v \int_{\mathbb{R}} \mathbf{q}_{XV}(u,t)dudt - \int_0^v \int_{\mathbb{R}} \mathbf{q}_X(u)\mathbf{p}_{V|X}(u,t)dudt \right| dv \end{aligned} \tag{28}$$

According to [24], if the weighted calibration conformal score probability density function (PDF) has Lebesgue density bounded by $\mathcal{C}$, which means $\mathbf{p}'_V$ does not exceed $\mathcal{C}$, then for any test conformal score PDF $\mathbf{q}_V$, $d_{\mathrm{K}}(F_Q, F_{Q/P})$ can be bounded as

$$d_{\mathrm{K}}(F_Q, F_{Q/P}) \leq \sqrt{2\mathcal{C} d_{\mathrm{W}}(F_Q, F_{Q/P})} \tag{29}$$

Finally, we can derive the upper limit of coverage difference under concept shift, $\sup_{v \in V_c} |D(v)|$, in Eq. (30) at least probability $e^{-2(n\epsilon^2+m\rho^2)}$.

$$\sup_{v \in V_c} |D(v)| \leq d_{\mathrm{K}}(\hat{F}_Q, \hat{F}_{Q/P}) \leq \sqrt{2\mathcal{C} d_{\mathrm{W}}(F_Q, F_{Q/P})} + \epsilon + \rho \tag{30}$$

This property is attractive in that the maximum difference in coverage due to concept shift can also be constrained in relation to the W-distance of population score CDFs, denoted as $d_{\mathrm{W}}(F_Q, F_{Q/P})$. Despite the unobservability of $d_{\mathrm{W}}(F_Q, F_{Q/P})$, we can still estimate it using its empirical form, $d_{\mathrm{W}}(\hat{F}_Q, \hat{F}_{Q/P})$.

Even though coverage guarantee on an arbitrary joint shift is almost impossible, Eq. (28) demonstrates robust conformal prediction is attainable if we can train a function reducing the discrepancy between calibration and test conformal score distributions. To be specific, $d_{\mathrm{W}}(F_Q, F_{Q/P})$ can be reduced to zero as far as $\mathbf{p}_{V|X} = \mathbf{q}_{V|X}$. In other words, if we regard $\mathbf{p}_{XV}$ and $\mathbf{q}_{XV}$ as push-forward probability distribution of P_{XV} and Q_{XV} by the trained model f, making the concept shift between $\mathbf{p}_{V|X}$ and $\mathbf{q}_{V|X}$ smaller will reduce coverage difference expectation on test domain.

D Datasets, models, and experiment setups

Extensive experiments are conducted under 3 tasks with 7 datasets. Some tasks involve both black-box and physics-informed models to demonstrate the generalizability of NTW and mRCP.

D.1 Airfoil self-noise example

The airfoil dataset from the UCI Machine Learning Repository [5] consists of 1503 instances of 1-dimensional target Y and 5-dimensional feature $X = (X_1, X_2, X_3, X_4, X_5)$. This dataset is manually separated and modified to create three different domains.

Domain separation:

Step 1. Covariate Shift by Data Separation. The original dataset is initially segmented into three primary subsets A, B, C based on the 33% and 66% quantiles of the first dimension X_1. Subsequently, each of these subsets is further divided into three smaller portions at a 7:2:1 ratio, denoted like $A_{0.7}, A_{0.2}, A_{0.1}$ from A. Finally, we assemble three new datasets with covariate shift as $S^{(e_1)} = A_{0.7} \cup B_{0.2} \cup C_{0.1}$, $S^{(e_2)} = A_{0.2} \cup B_{0.1} \cup C_{0.7}$, $S^{(e_3)} = A_{0.2} \cup B_{0.1} \cup C_{0.2}$.

Step 2. Concept Shift by Target Modification. Differently distributed random noises are added to target values to cause concept shifts. For y_i from $S^{(e_1)}$, $y_i += y_i/1000 * \tau$; for y_i from $S^{(e_2)}$, $y_i += y_i/\tau$; for y_i from $S^{(e_3)}$, $y_i += \tau$. τ follows a normal distribution $N(0, 10^2)$. Since we obtain three subsets in the end, $|\mathcal{E}| = 3$.

Model selection:

We utilize a straightforward multilayer perceptron (MLP) as a trainable model, with an architecture of (input dimension, 64, 64, 1) tailored for the regression task.

D.2 Traffic speed prediction

The Seattle-loop [9], and PeMSD4, PeMSED8 datasets [16] contain sensor-observed traffic volume and speed data collected in Seattle, San Francisco, and San Bernardino. The snapshots from sensors are taken at 5-minute intervals. This task aims to predict the traffic speed of the local road segment in the next time step, using the traffic data from local and neighboring segments collected currently.

Domain separation:

Naturally, instances can be categorized into 24 subsets, $|\mathcal{E} = 24|$, based on the hour they are obtained. It is anticipated that there are joint shifts between the data distribution of every single hour (test domains) and the data distribution of the whole day (calibration domain), as traffic patterns vary over time, making it unnecessary to modify any data. We select the workday data from the three datasets.

Model selection:

(a) MLP with the same structure (input dimension, 64, 64, 1) is applied to the traffic prediction task.

(b) The Reaction-Diffusion (RD) model is selected as the physics-informed Partial differential equation (PDE) for traffic speed prediction. Reaction-diffusion mechanism, originally formulated for chemical systems to describe particle dynamics, has been adapted for traffic analysis by [3] to uncover traffic patterns on different road segments, offering an alternative to purely data-driven models like long-short-term memory. [28] further advanced this approach by integrating the RD model into graphical neural networks to capture traffic state interactions among adjacent road segments, with the reaction term accounting for influences against traffic flow and the diffusion term for influences along it. To be specific, for a given sensor i, with N^d upstream and N^r downstream neighboring sensors, the traffic states from these sensors impact sensor i after δt time through diffusion and reaction effects, respectively. We expand the original RD model in [28] to Eq. (31), where the traffic speed and volume at sensor i at time t is $u_i(t)$ and $q_i(t)$, respectively. The parameters $\rho_{(i,j)}$ and $\sigma_{(i,j)}$ represent the diffusion and reaction strengths between sensor i and sensor j, while their superscripts indicate if they serve for speed or volume. Also, d_i and r_i are bias terms for the two components.

$$\begin{aligned} u_i(t+\delta t) - u_i(t) = &\sum_{j \in N^d} (\rho^u_{(i,j)}(u_i(t) - u_j(t)) + \rho^q_{(i,j)}(q_i(t) - q_j(t)) + d_i \\ &+ \tanh(\sum_{j \in N^r} \sigma^u_{(i,j)}(u_i(t) - u_j(t)) + \sigma^q_{(i,j)}(q_i(t) - q_j(t)) + r_i). \end{aligned} \tag{31}$$

D.3 Epidemic spread prediction

Three epidemic datasets, US-Regions, US-States, and Japan-Prefectures [10] include the number of patients infected by influenza-like illness (ILI) recorded by U.S. Department of Health and Human Services, Center for Disease Control and Prevention (CDC), and Japan Infectious Diseases Weekly Report. We aim to use the local population, the rise in the number of infected patients observed this week, and the cumulative total of infections as predictive features of the increase in infections for the upcoming week.

Domain separation: According to the Pandemic Intervals Framework (PIF) by CDC, samples are divided by four pandemic intervals, Initiation, Acceleration, Declaration, and Subsidence, so $|\mathcal{E}| = 4$. We establish the interval endpoints based on specific percentages of the total infected patient count, specifically at the 15%, 50%, and 85% thresholds.

Model selection:

(a) MLP with the same architecture is utilized for the epidemic spread forecasting task as well.

(b) PDE for this task is the SIR model that categorizes the population into three groups: those susceptible to the disease S, those infectious I, and those who have recovered and gained immunity R. It outlines the temporal changes in their populations, as described by [8]. The governing differential equations can be expressed as Eq. 32, where N, λ, and γ represent the total population, infection rate, and recovery rate, respectively.

$$\begin{cases} \frac{dS(t)}{dt} = \frac{-\lambda S(t) I(t)}{N}, \\ \frac{dI(t)}{dt} = \frac{\lambda S(t) I(t)}{N} - \gamma I(t) = (\frac{\lambda S(t)}{N} - \gamma) I(t), \\ \frac{dR(t)}{dt} = \gamma I(t). \end{cases} \tag{32}$$

We make the assumption that the location is isolated, hence $N = S(t) + I(t) + R(t)$. Additionally, the population of recovered individuals is represented by $R(t) = \gamma \int_0^t I(t) dt$. Given this, if t_o signifies the initial time of the current epidemic and δt denotes the time step, which is a week in the three datasets, we can express the dynamic change of infectious individuals discretely as Eq. (33).

$$I(t + \delta t) - I(t) = \left(\frac{\lambda (N - I(t) - \gamma \sum_{t_o}^{t} I(t))}{N} - \gamma \right) I(t). \tag{33}$$

D.4 Experiment setups for NTW and baseline metrics

As we only need to validate the positive correlation between NTW and coverage difference expectation, all models are trained by ERM. In the airfoil self-noise example, 100 trials are carried out. For the traffic task, 61 locations from the Seattle-loop, 59 locations from PeMSD4, and 33 locations from PeMSD8 are chosen, with 10 trials conducted at each location. For simplicity in the calculation, all selected locations have just one segment upstream and one segment downstream. For epidemic datasets, all locations from US-Regions, US-States, and Japan-Prefectures (49 locations in US-States, 10 locations in US-Regions, and 46 locations in Japan-Prefectures) are encompassed in the experiments, with 10 trials implemented on each location. The same experiment setups are operated on all baseline metrics and NTW. σ values for MLP and PDE are 0.8 and 0.95, respectively. The ratio of training, calibration, validation, and testing data on airfoil self-noise datasets, three traffic datasets, and three epidemic datasets are 1:1:1:1, 3:2:2:3, and 1:2:1:1, respectively. Data separation was conducted randomly. Adam optimizer with a learning rate of 0.001 was applied for all experiments. On average, one trial requires one hours on a workstation with double NVIDIA RTX 3090 GPU.

D.5 Experiment setups for mRCP, V-REx, and DRO

We define 1 trial as running a series of experiments of all predefined β values once, except for DRO. For the airfoil self-noise example, 100 trials with random data preprocessing are conducted. For the traffic speed prediction task, we randomly select 10 locations from each of the three traffic datasets and operate one trial on all selected locations. In the epidemic spread prediction task, all locations of the three datasets are included and we operate one trial on each of them. All combinations of models (MLP and PDE) and algorithms (mRCP, DRO, V-REx) share the same experiment setups mentioned above. σ values for MLP and PDE are 0.8 and 0.95, respectively. β values for mRCP and V-REx in different experiment setups are shown in Table 3. Each Pareto curve consists of at least 10 β values. For airfoil self-noise datasets and three traffic datasets, the original data is evenly and randomly split for training, calibration, and testing. For three epidemic datasets, we randomly split the original data for training, calibration, and testing with a ratio of 2:1:2. Adam optimizer with a learning rate of 0.001 was applied for all experiments. On average, one trial requires 12 hours on a workstation with double NVIDIA RTX 3090 GPU.

Table 3: β values for mRCP and V-REx in experiment setups

Dataset	Model	Algorithm	β Values
Airfoil Self-Noise	MLP	mRCP	0.1, 0.2, 0.5, 1, 2, 5, 10, 15, 20, 30, 50, 80, 100.
		V-REx	0.1, 1, 2, 2.5, 3, 3.5, 4, 4.5, 5, 6, 7, 8, 9, 10, 15, 20.
Japan-Prefectures	MLP	mRCP	0.1, 0.2, 0.4, 0.8, 1, 2, 5, 10, 20, 40, 100, 200, 500.
		V-REx	0.1, 1, 1.5, 2, 3, 4, 5, 7.5, 10, 20, 40, 100, 200, 500.
	PDE	mRCP	0.1, 0.2, 0.4, 0.6, 0.8, 1, 2, 5, 7, 10.
		V-REx	0.1, 0.2, 0.4, 0.6, 0.8, 1, 2, 5, 7, 10.
US-Regions	MLP	mRCP	0.1, 0.2, 0.4, 0.8, 1, 2, 5, 10, 20, 40, 100, 200, 500.
		V-REx	0.1, 1, 2, 3, 4, 5, 6, 7, 8, 10, 15, 20, 30, 40, 100.
	PDE	mRCP	0.1, 0.2, 0.4, 0.6, 0.8, 1, 2, 5, 7, 10.
		V-REx	0.1, 0.2, 0.4, 0.6, 0.8, 1, 2, 5, 7, 10.
US-States	MLP	mRCP	0.1, 0.2, 0.4, 0.8, 1, 2, 5, 10, 20, 40, 100, 200, 500.
		V-REx	0.1, 1, 1.2, 1.7, 2, 2.5, 3, 3.5, 4, 5, 7, 10, 15.
	PDE	mRCP	0.1, 0.2, 0.4, 0.6, 0.8, 1, 2, 5, 7, 10.
		V-REx	0.1, 0.2, 0.4, 0.6, 0.8, 1, 2, 5, 7, 10.
Seattle-loop	MLP	mRCP	0.1, 1, 2, 3, 4, 5, 10, 25, 100, 200, 400, 700, 1000.
		V-REx	0.1, 1, 1.5, 2, 2.5, 3, 4, 5, 10, 50.
	PDE	mRCP	0.1, 1, 5, 10, 20, 40, 80, 160, 320, 640.
		V-REx	0.1, 0.2, 0.5, 0.8, 1, 1.5, 2, 3, 4, 5.
PeMSD4	MLP	mRCP	0.1, 1, 2, 5, 10, 50, 100, 150, 200, 300, 400, 500.
		V-REx	0.1, 1, 2, 3, 4, 5, 7.5, 10, 13, 16, 19, 22, 25.
	PDE	mRCP	0.1, 1, 5, 10, 50, 100, 200, 500, 1000, 5000, 10000.
		V-REx	0.1, 1, 2, 3, 4, 5, 7, 8, 10, 12, 15.
PeMSD8	MLP	mRCP	0.1, 1, 2, 5, 10, 50, 100, 150, 250, 300, 400, 500.
		V-REx	0.1, 1, 2, 3, 4, 5, 7.5, 10, 20, 30, 40, 75, 80, 150.
	PDE	mRCP	0.1, 1, 5, 10, 50, 100, 200, 500, 1000, 2000.
		V-REx	0.1, 1, 2, 3, 5, 7, 10, 15, 20, 30.

E Additional experiment results

E.1 Pearson coefficient definition

Here we provide a detailed definition of the Pearson coefficient as follows.

Definition 4 (Pearson coefficient)**.** *The Pearson correlation coefficient, denoted as r, is calculated as the covariance of the two variables divided by the product of their standard deviations, as follows.*

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}. \tag{34}$$

where x_i and y_i are the individual sample points of random variables X and Y indexed with i and $\overline{x}$ and $\overline{y}$ are the means of their samples, respectively.

The Pearson correlation coefficient measures the linear correlation between two variables. It gives a value between -1 and 1 inclusive, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear correlation.

E.2 Correlation visualization

Figure 3 shows the experimental results of the correlation between NTW and coverage difference expectation, compared with three baselines: total variation, KL divergence, and expectation difference. It is organized into a matrix of subplots, with each column corresponding to a specific dataset and each row depicting the performance of a metric. Within these subplots, individual points represent the conjunction of a metric's value with the associated coverage difference expectation for a given test domain. A positive trend between NTW and the coverage difference expectation is shown in the top row, showcasing NTW's strong correlation. In contrast, the other metrics exhibit inconsistent correlations across the varied datasets and models, as seen in the lower three rows of subplots. Figure 4 also illustrates the expected coverage difference's correlation to NTW, standard W-distance, normalized W-distance, and truncated W-distance, proving that normalization and truncation are equally important for robust correlations.

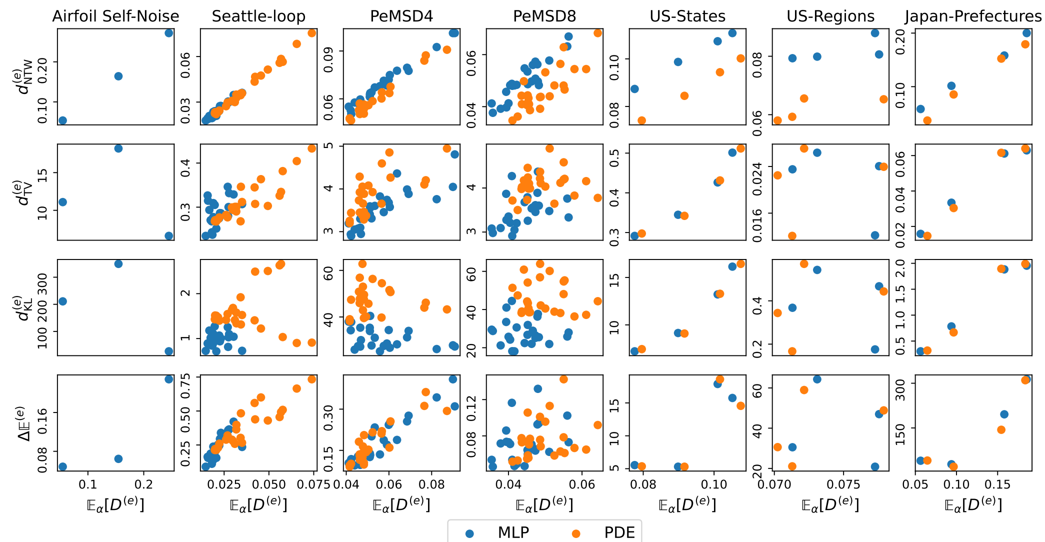

Figure 3: **Experimental results of the correlation between Normalized Truncated Wasserstein distance (NTW) and coverage difference expectation, compared with total variation, KL divergence, and expectation difference**. Each point represents a pair of metric value and coverage difference expectation for a test domain. The first row of the subplots demonstrates NTW indicates the expectation across different datasets and models, whereas other baseline metrics, represented in the other three rows, can not consistently capture it.

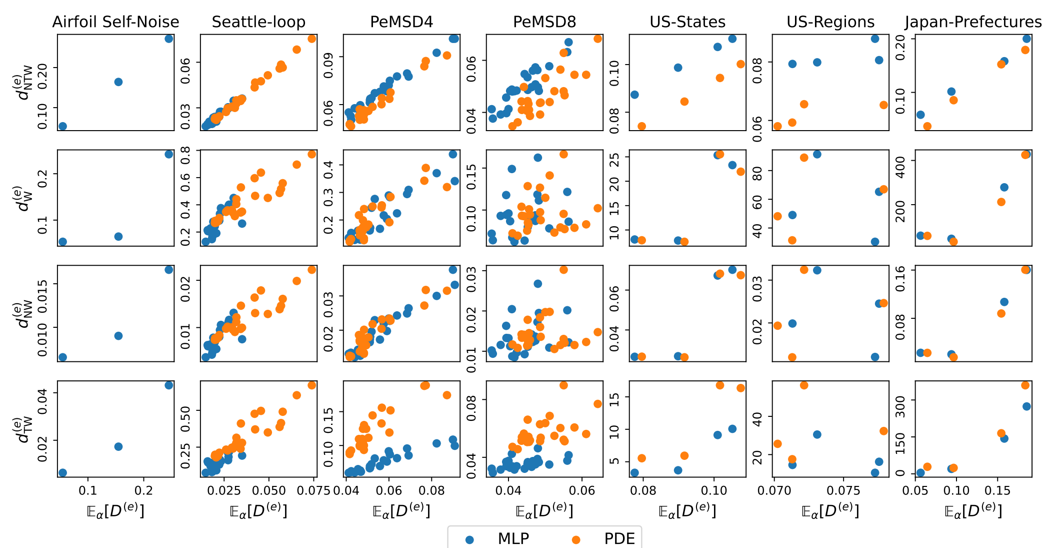

Figure 4: **Experimental results of the correlation between Normalized Truncated Wasserstein distance (NTW) and coverage difference expectation of concept shift, compared with standard, normalized, and truncated Wasserstein distance**. Each point represents a pair of metric value and coverage difference expectation of a test domain. By comparing the first row with the rest three rows, we validate the necessity of applying normalization and truncation together.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We focus on robust conformal prediction under joint distribution shift. The abstract and introduction accurately state our contributions. We propose Normalized Wasserstein distance to quantify the coverage difference caused by concept shift and develop multi-domain robust conformal prediction to make coverage approach confidence when multiple test domains hold joint shifts with the calibration domain.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation of our proposed method in Section 6. The proposed method requires the training model's enough capacity to fit complex patterns. Also, as it needs to approximate the ratio of covariate likelihood between calibration and test domains, it requires enough training and calibration samples to conduct accurate kernel density estimation.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We refer to Section 3, Appendix B, Appendix C for detailed theoretical work.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We show detailed experiment setups, including models, datasets, and algorithms, in Section 5 and Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We would like to provide open access to the data and code if this submission is accepted.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed information about data splits and preprocessing, hyperparameters, optimizer, black-box model architecture, and physics-informed equations in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Statistical measures of experiment results are shown in Figure 2 and Table **??**.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the information about our workstation and computation time for one trial in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research follows the NeurPIS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We mentioned the positive impact of robust conformal prediction in the introduction. We do not see negative societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our submission does not pose such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Datasets and models are properly cited in the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not introduce new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This research does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This research does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.