Wasserstein Distributional Learning via Majorization-Minimization

Chengliang Tang[1] **Nathan Lenssen**[2] **Ying Wei**[3] **Tian Zheng**[1]
[1]Department of Statistics, Columbia University
[2]Department of Atmospheric and Oceanic Sciences (ATOC), University of Colorado Boulder
[3]Department of Biostatistics, Columbia University

Abstract

Learning function-on-scalar predictive models for conditional densities and identifying factors that influence the entire probability distribution are vital tasks in many data-driven applications. We present an efficient Majorization-Minimization optimization algorithm, Wasserstein Distributional Learning (WDL), that trains Semi-parametric Conditional Gaussian Mixture Models (SCGMM) for conditional density functions and uses the Wasserstein distance W_2 as a proper metric for the space of density outcomes. We further provide theoretical convergence guarantees and illustrate the algorithm using boosted machines. Experiments on the synthetic data and real-world applications demonstrate the effectiveness of the proposed WDL algorithm.

1 INTRODUCTION

In scientific fields such as economics, biology, and climate science, examining the drivers of distributional heterogeneity is a powerful way for knowledge discovery. For example, climate change has profoundly impacted multiple aspects of a climate outcome's distribution, including its mean, overall variability, and the frequency of extreme values (Field et al., 2012; Reich, 2012). Figure 1 displays annual distributions of daily temperature anomalies from 1880 to 2012. These distributions exhibit a shift in the mean and a substantial increase in tail behavior heterogeneity. Identifying drivers of such distributional shifts and characterizing their effects are active areas of research (Lewis and King, 2017; Fahey et al., 2017).

Traditional models that focus on conditional mean or other summary statistics are limited in studying complex distributional heterogeneity. In this paper, we focus on predictive learning of conditional distributions (density functions), i.e., *Distributional Learning*, as a more direct, comprehensive and unified approach to discovering complex associations between a distributional outcome and a set of drivers. One major challenge for distributional learning lies in the unique features of the density functional space, i.e., non-negative, Borel measurable, and integrate to one. Functional regression (e.g. Reiss et al., 2010; Wang et al., 2016) has been applied to model probability density functions (PDFs) outcomes, under the commonly used L_2 distance as the measure of discrepancies. These methods are not structured to address the necessary constraints of PDFs and lead to problematic results (Delicado, 2011). Recent developments (Arata, 2017; Boogaart et al., 2010; Egozcue et al., 2006; Talská et al., 2018; Van den Boogaart et al., 2014) resort to centered log-ratio (CLR) transformations to map PDFs onto zero-integral elements of the space of square-integrable real functions. Due to the transformations, these approaches could be over-restrictive, hard to assess and interpret, and as the data dimension grows, lead to substantial bias. In particular, they do not adopt the Wasserstein geometry, a well-defined metric for the density space.

In this paper, we directly address the needs of *Distri-*

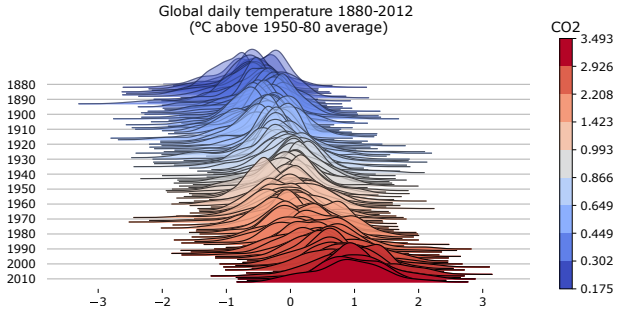

Figure 1: Annual distributions of daily land-surface average temperature. Temperatures are anomalies (in Celsius) relative to the Jan. 1951 – Dec. 1980 average. The color of each density curve shows the annual radiative effect of increased atmospheric CO2.

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

butional Learning by structuring a theoretically proper function-on-scalar regression framework for PDFs that (1) is defined on a sufficiently large and flexible functional model space of PDFs; (2) uses the Wasserstein loss function (Panaretos and Zemel, 2019; Villani, 2008), a proper measure of discrepancy for PDFs; and (3) enables computationally tractable optimization algorithms for carrying out the model training. Our contributions are as follows:

1. We propose a comprehensive *Wasserstein Distributional Learning* (WDL) framework, based on the global approximation property of finite mixture models. WDL satisfies the inherent constraints of density functions and offers flexibility and expressiveness in modeling complicated density outputs.

2. We derive theoretical guarantees for universal approximation and consistency of the WDL framework.

3. We develop an efficient majorization-minimization optimization algorithm that strongly resembles expectation–maximization (EM). Optimization under the Wasserstein loss is notoriously challenging, which has been the obstacle for distributional learning under the Wasserstein geometry. Our proposed algorithm is, to the best of our knowledge, the first majorization-based solution to the estimation problem associated with the Wasserstein geometry. It achieves good convergence performance both theoretically and empirically.

4. We demonstrate the excellent modeling performance of WDL using boosted machines as an example use.

2 BACKGROUND AND DEFINITIONS

Data Setup and Notations. In this paper, we consider distributional outcomes over $\mathbb{R}$ and define $\mathcal{P}_2(\mathbb{R})$ as the set of all Borel probability measures on $\mathbb{R}$ with a finite 2nd moment. Let $\mathbf{X}$ be a p-dimensional random covariate vector with a support $\mathcal{X} \subset \mathbb{R}^p$ and a probability density function (PDF) $f_{\mathbf{X}}$; and $\mathcal{G}$ is a distributional response representing the distribution for an outcome $Y \in \mathbb{R}$ and $\mathcal{G} \in \mathcal{P}_2(\mathbb{R})$. Assume a random sample of n i.i.d. draws from the joint distribution of the random process $(\mathbf{X}, \mathcal{G})$ on the product space $\mathcal{X} \times \mathcal{P}_2(\mathbb{R})$: $\mathcal{D} = \{(x_i, g_i)\}_{i=1}^n$. For temperature distributions in Figure 1, for the i-th year, g_i is the observed empirical distribution of daily temperatures, while x_i consists of the values of potential drivers for that year. Our goal is to model the expected conditional distribution $\mathbb{E}(\mathcal{G}|\mathbf{X} = x)$ from the random sample $\mathcal{D} = \{(x_i, g_i)\}_{i=1}^n$.

Wasserstein Distance. The Wasserstein distance measures the aggregated discrepancies between two distributions. It offers excellent convergence properties in the distributional function space (Villani, 2008) and has gained popularity for its intuitive interpretation as the optimal transport costs (Bernton, 2019; Panaretos and Zemel, 2019; Villani, 2003), in addition to its utility in real-world applications (Arjovsky et al., 2017; Duy and Takeuchi, 2022; Sgouropoulos et al., 2015). Compared with other commonly used distributional loss functions such as the Kullback–Leibler divergence and L^2 functional distance, the Wasserstein distance does not require the distributions to have a common support or rely on specific transformations. It is, hence, more suitable for modeling highly heterogeneous distributions, and enjoys a more straightforward interpretation (Mueller et al., 2018; Pegoraro and Beraha, 2022; Sharma and Gerig, 2020; Zhang et al., 2022). In this paper, we focus on the one-dimensional 2-Wasserstein distance $W_2(f_1, f_2)$ for continuous PDFs, f_1 and f_2 from $\mathcal{P}_2(\mathbb{R})$, defined as

$$W_2(f_1, f_2) = \Big[\int_0^1 \big(F_1^{-1}(s) - F_2^{-1}(s)\big)^2 ds \Big]^{\frac{1}{2}},$$

where F_1^{-1} and F_2^{-1} are the quantile functions derived from f_1 and f_2, respectively.

3 WASSERSTEIN DISTRIBUTIONAL LEARNING

3.1 The overview of the WDL framework

Functional mapping from the scalar covariates to the density outputs is challenging due to the infinite dimensions of the output space. To circumvent this difficulty, we propose a semi-parametric conditional distribution family,

$$\begin{aligned} \mathfrak{F} \otimes \mathcal{T} = \{ f_{\boldsymbol{\theta}} \circ \boldsymbol{\tau}(x) \mid f_{\boldsymbol{\theta}} \in \mathfrak{F},\, \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^q; \\ \boldsymbol{\tau}(\cdot) \in \mathcal{T}(\mathcal{X}, \boldsymbol{\Theta});\; x \in \mathcal{X} \subset \mathbb{R}^p \}, \end{aligned} \tag{1}$$

where $\mathfrak{F} = \{f_{\boldsymbol{\theta}} | \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^q\}$ is a parametric distribution family, and $\mathcal{T}(\mathcal{X}, \boldsymbol{\Theta})$ is a non-parametric functional family of mappings from the covariate space $\mathcal{X}$ to the distribution parameter space $\boldsymbol{\Theta}$. This semi-parametric conditional distribution family should be sufficiently large and flexible such that the expected conditional distributions $\mathbb{E}(\mathcal{G}|\mathbf{X} = x)$ can be well approximated by its elements.

For a given set of observations $\mathcal{D} = \{(x_i, g_i)\}_{i=1}^n$, the goal of WDL is then to identify the optimal mapping $\widehat{\boldsymbol{\tau}}(\cdot)$ in a specified functional space $\mathcal{T}(\mathcal{X}, \boldsymbol{\Theta})$ and minimize the Wasserstein loss that is evaluated at $\mathcal{D} = \{(x_i, g_i)\}_{i=1}^n$:

$$\begin{aligned} \widehat{\boldsymbol{\tau}}(\cdot) &= \underset{\boldsymbol{\tau}(\cdot) \in \mathcal{T}(\mathcal{X}, \boldsymbol{\Theta})}{\arg\min} \sum_{i=1}^n W_2^2(g_i, f_{\boldsymbol{\theta} = \boldsymbol{\tau}(x_i)}) \\ &= \underset{\boldsymbol{\tau}(\cdot) \in \mathcal{T}(\mathcal{X}, \boldsymbol{\Theta})}{\arg\min} \sum_{i=1}^n \int_0^1 \big(G_i^{-1}(s) - F_{\boldsymbol{\theta} = \boldsymbol{\tau}(x_i)}^{-1}(s)\big)^2 ds, \end{aligned} \tag{2}$$

where $G_i^{-1}(s)$ is the quantile function derived from g_i and $F_{\boldsymbol{\theta} = \boldsymbol{\tau}(x_i)}^{-1}(s)$ is that derived from $f_{\boldsymbol{\theta} = \boldsymbol{\tau}(x_i)}$. For simplicity,

we refer to $F^{-1}_{\boldsymbol{\theta}=\boldsymbol{\tau}(x_i)}(s)$ and $f_{\boldsymbol{\theta}=\boldsymbol{\tau}(x_i)}$ as $F^{-1}_{\boldsymbol{\tau}(x_i)}$ and $f_{\boldsymbol{\tau}(x_i)}$ for the rest of the paper.

3.2 Semi-parametric Conditional Gaussian Mixture Model (SCGMM) as $\mathfrak{F} \otimes \mathcal{T}$

We propose, for $\mathfrak{F} \otimes \mathcal{T}$, a class of Semi-parametric Conditional Gaussian Mixture Model (SCGMM),

$$f_{\boldsymbol{\tau}(x)} = \sum_{k=1}^{K} \pi_k(x) \mathrm{N}\{\mu_k(x), \sigma_k^2(x)\}, \tag{3}$$

where N represents a Gaussian distribution, $\mu_k(x)$ and $\sigma_k^2(x)$ are the mean and variance of the k-th component, and $\pi_k(x)$ is the weight of the k-th component. We assume that all the parameters are unknown functions of the covariate x, and denote $\boldsymbol{\tau}(x) = \{\pi_k(x), \mu_k(x), \sigma_k^2(x)\}_{k=1}^K$ as the collection of all *parameter functions* of model (3). By definition, the SCGMM functional space automatically satisfies the non-negativity and unit-integral constraint of density functions. We offer below two novel theoretical guarantees for establishing that the SCGMM model space, under the 2-Wasserstein distance metric, supports a valid WDL framework as defined in Section 3.1. Proofs of these guarantees can be found in the Appendix.

The *Universal Approximation* guarantee. We establish that SCGMM is dense in $\mathcal{P}_2(\mathbb{R})$, under the Wasserstein geometry, with all the parameters being step functions of $\mathbf{X}$. We denote the distribution function of $\mathbf{X}$ by $P_{\mathbf{X}}$. For simplicity, we denote $H(x) \triangleq \mathbb{E}(\mathcal{G}|\mathbf{X} = x) \in \mathcal{P}_2(\mathbb{R})$ as the expected conditional density of Y given $\mathbf{X} = x$, and denote both $\boldsymbol{\tau}$ and $\boldsymbol{\tau}(\cdot)$ as the mapping from $\mathcal{X}$ to $\boldsymbol{\Theta}$ without distinction. Furthermore, for a given mapping $\boldsymbol{\tau}$, we denote $\widetilde{\boldsymbol{\tau}}$ as its equivalence class. Specifically, we say $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ belong to the same equivalence class $\widetilde{\boldsymbol{\tau}}$ if

$$\begin{aligned}&\int_{x\in\mathcal{X}} W_2(f_{\boldsymbol{\tau}_1(x)}, f_{\boldsymbol{\tau}(x)}) dP_{\mathbf{X}}(x)\\ =&\int_{x\in\mathcal{X}} W_2(f_{\boldsymbol{\tau}_2(x)}, f_{\boldsymbol{\tau}(x)}) dP_{\mathbf{X}}(x)\\ =&0,\end{aligned}$$

which also means $\widetilde{\boldsymbol{\tau}}_1 = \widetilde{\boldsymbol{\tau}}_2 = \widetilde{\boldsymbol{\tau}}$.

With the above definitions in place, we first introduce the following assumptions.

Assumption 1. *(The speed of decay of the covariates.) The covariate $\mathbf{X}$ follows a light-tailed distribution, i.e., there exist positive constants λ and M_0, such that $P_{\mathbf{X}}(\|\mathbf{X}\|_2 > M) < \exp(-\lambda M)$ for any $M > M_0$.*

Assumption 2. *(Continuity of $H(x)$.) $H(x) : \mathcal{X} \longrightarrow \mathcal{P}(\mathbb{R})$ is Lipschitz continuous, i.e., there exists a real constant $L > 0$ such that, for all x_1 and x_2 in $\mathcal{X}$,*

$$W_2(H(x_1), H(x_2)) \leq L\|x_1 - x_1\|_2.$$

Theorem 1. *(Universal approximation of SCGMM.) Under Assumptions 1 and 2, for any $\varepsilon > 0$, there exists a positive integer $K > 0$, and corresponding Gaussian mixture regression $f_{\boldsymbol{\tau}(x)} = \sum_{k=1}^K \pi_k(x) N\left(\mu_k(x), \sigma_k^2(x)\right)$, such that*

$$\int_{x\in\mathcal{X}} W_2(H(x), f_{\boldsymbol{\tau}(x)}) dP_{\mathbf{X}}(x) < \varepsilon,$$

where $\boldsymbol{\tau}(x) = \{\pi_k(x), \mu_k(x), \sigma_k^2(x)\}$ are all scalar-valued step functions of x.

The *Consistency* guarantee. Here we establish that, the optimizer from the Wasserstein regression (2), $f_{\widehat{\boldsymbol{\tau}}(x)}$, is uniformly consistent in estimating $\mathbb{E}(\mathcal{G}|\mathbf{X})$, the true conditional distribution over $\mathcal{X}$. We denote

$$\mathcal{M}_n(\boldsymbol{\tau}) = \frac{1}{n}\sum_{i=1}^{n} W_2^2(f_{\boldsymbol{\tau}(x_i)}, g_i)$$
$$\text{and } \mathcal{M}(\boldsymbol{\tau}) = \mathbb{E}\Big(W_2^2(f_{\boldsymbol{\tau}(x)}, g)\Big).$$

Here, $\mathcal{M}_n, \mathcal{M} : \mathcal{T}(\mathcal{X}, \boldsymbol{\Theta}) \to \mathbb{R}_+$ are non-negative functions defined over the functional space $\mathcal{T}(\mathcal{X}, \boldsymbol{\Theta})$. Further, we introduce metric $d(\cdot, \cdot)$ over $\mathcal{T}(\mathcal{X}, \boldsymbol{\Theta}) \times \mathcal{T}(\mathcal{X}, \boldsymbol{\Theta})$ by

$$d(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2) \triangleq \sup_{x\in\mathcal{X}} \|\boldsymbol{\tau}_1(x) - \boldsymbol{\tau}_2(x)\|.$$

We first introduce the following assumptions.

Assumption 3. *(Continuity of $f_{\boldsymbol{\theta}}$ over $\boldsymbol{\theta}$.) The map $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}$ is continuous in the sense that for any sequence $\{\boldsymbol{\theta}_n\} \subset \boldsymbol{\Theta}$ and point $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\| \to 0$ implies $W_2(f_{\boldsymbol{\theta}_n}, f_{\boldsymbol{\theta}_0}) \to 0$.*

Assumption 4. *(Uniqueness of the minimizer.) For any $g \in \mathcal{P}_2(\mathbb{R})$, the minimizer set $\arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} W_2(g, f_{\boldsymbol{\theta}})$ is non-empty and belong to the same equivalence class.*

Theorem 2. *(The consistency of $\widehat{\boldsymbol{\tau}}_n$.) Under Assumptions 3 and 4, suppose $\arg\min_{\boldsymbol{\tau}\in\mathcal{T}(\mathcal{X},\boldsymbol{\Theta})} \mathcal{M}(\boldsymbol{\tau}) \subseteq \widetilde{\boldsymbol{\tau}}_0$, the equivalence class of $\boldsymbol{\tau}_0$, and the SCGMM estimators $\widehat{\boldsymbol{\tau}}_n = \arg\min_{\boldsymbol{\tau}\in\mathcal{T}(\mathcal{X},\boldsymbol{\Theta})} \mathcal{M}_n(\boldsymbol{\tau})$ all lie in a compact set $S \subset \mathcal{T}(\mathcal{X}, \boldsymbol{\Theta})$, then $\widehat{\boldsymbol{\tau}}_n$ are consistent in the sense that for every $\varepsilon > 0$,*

$$\mathbb{P}(d(\widehat{\boldsymbol{\tau}}_n, \widetilde{\boldsymbol{\tau}}_0) \geq \varepsilon) \to 0.$$

Identifiability constraint. In practice, to ensure the model identifiability of $f_{\boldsymbol{\tau}(x)}$ using finite data, we add an order constraint to the component means throughout the paper, stating that $\mu_1(x) \leq \mu_2(x) \leq ... \leq \mu_K(x)$.

3.3 Majorization-Minimization Optimization

Optimizing under the Wasserstein distance has been known to be computationally challenging (Bernton et al., 2019; Kolouri et al., 2017). Following our notations, the gradient of the 2-Wasserstein loss takes the following form:

$$\frac{\partial W_2^2(f_\theta, g)^2}{\partial \theta} = 2\int_{-\infty}^{\infty} \phi_\theta(t) \cdot \frac{\partial f_\theta}{\partial \theta}(t) dt,$$

where $\phi_\theta(t) = \int_{-\infty}^{t} \big(G^{-1} \circ F_\theta(x) - x\big)dx$ is the displacement potential for optimal transport plan with parameter θ. To compute the above gradient, multiple steps of numerical differentiation, integration and function inverse are needed. This leads to both unstable results and high computational costs.

For our proposed WDL framework, we derive a novel and efficient EM-like majorization-minimization optimization algorithm. Without loss of generality, we introduce the algorithm without the covariate $\mathcal{X}$. We demonstrate in Section 3.4 how the algorithm enables distributional learning between $\mathcal{G}$ and $\mathcal{X}$ in $\mathfrak{F} \otimes \mathcal{T}$ when integrated with boosting machines. Proofs of the theoretical results in this section can be found in the Appendix.

Let $F_\mathcal{D}$ be the observed empirical distribution and $\{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be a target parametric distribution family. We aim to find the optimal parameters $\widehat{\boldsymbol{\theta}}$ that minimize $L(\boldsymbol{\theta}) = W_2^2(F_\mathcal{D}, F_{\boldsymbol{\theta}})$ over all the possible $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We begin with a simple case where $K = 1$ and $F_{\boldsymbol{\theta}}$ belongs to a location-scale family, i.e., $\mathfrak{F} = \{F_{\boldsymbol{\theta}}(y) = F_0(\frac{y-\mu}{\sigma}) : \boldsymbol{\theta} = (\mu, \sigma), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$, expanding from a standard distribution $F_0(Y)$. The minimum Wasserstein estimations of (μ, σ) can be derived as

$$\begin{aligned} \widehat{\mu} &= \int F_\mathcal{D}^{-1} ds - \widehat{\sigma} \cdot \int F_0^{-1} ds, \\ \widehat{\sigma} &= \frac{\int F_0^{-1} F_\mathcal{D}^{-1} ds - \int F_0^{-1} ds \cdot \int F_\mathcal{D}^{-1} ds}{\int (F_0^{-1})^2 ds - (\int F_0^{-1} ds)^2}. \end{aligned} \tag{4}$$

The major convenience offered by the location-scale family is that their quantile functions can be written as linear functions of the location and scale parameters. In practice, the integrals in Equation (4) can be approximated numerically by summation over a sequence of discrete quantile levels $0 < q_1 < ... < q_M < 1$ or via Monte Carlo integration. This estimation naturally holds for Gaussian distributions, a location-scale distribution family.

When the target distribution $F_{\boldsymbol{\theta}}$ is a Gaussian mixture, however, the quantile functions $F_{\boldsymbol{\theta}}^{-1}(s)$ become complicated nonlinear functions of the model parameters $\boldsymbol{\theta} = \{(\pi_k, \mu_k, \sigma_k)\}_{k=1}^K$. The optimization over $L(\boldsymbol{\theta})$ is then much more challenging. Moreover, the Wasserstein loss function is not convex, which leads to instability and slow convergence for gradient-based optimization algorithms. In this paper, we derive a Majorization-Minimization (MM) algorithm built on the following *Majorization* theorem that provides an alternative loss function as a tight upper bound of the Wasserstein loss.

Theorem 3. *(Majorization.) For any two continuous PDFs $f \in \mathcal{P}(\mathbb{R})$ and $g \in \mathcal{P}(\mathbb{R})$, along with any mixture decomposition of $f = \sum_{k=1}^K \pi_k \cdot f_k$ and $g = \sum_{k=1}^K \pi_k \cdot g_k$, the following inequality holds*

$$W_2^2(f, g) \le \sum_{k=1}^K \pi_k W_2^2(g_k, f_k), \tag{5}$$

and the equality holds when

$$\begin{aligned} g_k(x) = g(x) \cdot \frac{f_k \circ F^{-1} \circ G(x)}{\sum_{j=1}^K \pi_j f_j \circ F^{-1} \circ G(x)}, \\ \forall x \in \mathbb{R}, \quad for \quad k = 1, ..., K. \end{aligned} \tag{6}$$

Here, F^{-1} is the QF of f, and G is the CDF of g.

Let $g = f_\mathcal{D}$, $f = f_{\boldsymbol{\theta}}$ and $f_{\boldsymbol{\theta}}$ is a Gaussian mixture with $\boldsymbol{\theta} = \{(\pi_k, \mu_k, \sigma_k)\}_{k=1}^K$. The left side of Equation (5) in Theorem 3 is the target loss function of our proposed WDL with SCGMM. The right side of Equation (5) provides a surrogate loss that *majorizes* the original objective function. Based on these results of Theorem 3, we derive the following algorithm that minimize $L(\boldsymbol{\theta}) = W_2^2(f_\mathcal{D}, f_{\boldsymbol{\theta}})$ by iteratively updating the decomposition of the empirical distribution $f_\mathcal{D}$ and model distribution $f_{\boldsymbol{\theta}}$.

Majorization-Minimization algorithm. At the end of the $(m-1)$-th iteration, let $f_k^{(m-1)} = \mathrm{N}(\mu_k^{(m-1)}, \sigma_k^{(m-1)})$ be the components of the model fit $f_{\boldsymbol{\theta}^{(m-1)}}$, and $\pi^{(m-1)}$ be the component weights. The m-th iteration consists of the following three sub-steps.

1. **[Calculate $g_k^{(m-1)}$.]** Given $\pi_k^{(m-1)}$ and $f_k^{(m-1)}$, the goal is to find the optimal mixture decomposition for the empirical distribution $g = f_\mathcal{D} = \sum_{k=1}^K \pi_k^{(m-1)} g_k$, such that the surrogate loss is minimized. Since the surrogate function has a lower bound as the original loss function, by the equality conditions in Theorem 3, we update $g_k^{(m-1)}$ using Equation (6). Note that this step is quite similar with the E-step in the conventional EM algorithm, and the only difference is that x is replaced by $F_{\boldsymbol{\theta}^{(m-1)}}^{-1} \circ G(x)$.

2. **[Update $f_k^{(m-1)}$.]** Given $\pi_k^{(m-1)}$ and $g_k^{(m-1)}$, find the optimal $f_k^{(m)} = \mathrm{N}(\mu_k^{(m)}, \sigma_k^{(m)})$ for minimizing $\sum_{k=1}^K \pi_k^{(m-1)} W_2^2(f_k, g_k^{(m-1)})$ using Equation (4).

3. **[Update $\pi_k^{(m-1)}$.]** Given $f_k^{(m)}$, find the optimal $\pi_k^{(m)}$ for minimizing the target loss $L(\boldsymbol{\theta}) = W_2^2(f_{\boldsymbol{\theta}}, f_\mathcal{D})$ with $f_{\boldsymbol{\theta}} = \sum_{k=1}^K \pi_k f_k^{(m)}$.

   This step is the most challenging part in the optimization as there is no explicit formula for the optimal π_k. Here, we provide two solutions. The first solution is based on gradient descent. The derivative of loss function $L(\boldsymbol{\theta})$ with respect to π_k is

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \pi_k} = \int_\mathbb{R} (G^{-1} \circ F_{\boldsymbol{\theta}}(t) - t) \cdot F_k(t) dt.$$

In this case, F_k is the CDF of $f_k^{(m)}$. The convergence of this solution is guaranteed by the convexity of $L(\boldsymbol{\theta})$ over π. The second solution is to use the Maximization step in EM algorithm for approximating the optimal π, which is expressed as

$$\pi_k^{(m)} = \int_{\mathbb{R}} g_k^{(m-1)}(x) \cdot \frac{\pi_k^{(m-1)} f_k^{(m)}(x)}{\sum_{j=1}^K \pi_j^{(m-1)} f_j^{(m)}(x)} dx.$$

Since the Wasserstein distance characterizes the weak topology in the density distribution space, it is dominated by KL divergence (strong topology) in the limiting case. This solution does not find the optimal π under the Wasserstein loss, but works well in practice and is much easier to implement.

We show, in the Appendix, that the original loss function $L(\boldsymbol{\theta})$ decreases during the optimization of the upper bound by each iteration, i.e., $L(\boldsymbol{\theta}^{(m)}) \leq L(\boldsymbol{\theta}^{(m-1)})$. The algorithm hence converges under the original loss function.

3.4 Boosted Wasserstein Distributional Learning

In this section, we implement a full WDL framework, $\mathfrak{F} \otimes \mathcal{T}$, as defined in Equation (1). For a sample of data, $\mathcal{D} = \{(x_i, g_i)\}_{i=1}^n$, we use SCGMM as $\mathfrak{F}$, and boosted regression trees (Friedman, 2001) as $\mathcal{T}$, with optimization facilitated by the Majorization-Minimization algorithm in Section 3.3.

To model the SCGMM parameters with regression trees, we apply the following transformations. For the mixing weights $\pi_k(x)$, $\sum_{k=1}^K \pi_k(x) = 1$ and $\pi_k(x) \geq 0\ \forall k$, we introduce reparameterization through the softmax function:

$$\pi_k(x) = \frac{\exp(\alpha_k(x))}{\sum_{k=1}^K \exp(\alpha_k(x))},$$

where $\alpha_k(x)$'s are outputs of boosted trees and take values in $\mathbb{R}$. Similarly, the scale parameters $\sigma_k(x)$ are represented in terms of the exponential of the boosted tree outputs $\sigma_k(x) = \exp(z_k(x)) > 0$. The mean components, $\mu_k(x)$'s, are not transformed.

We fit the mixture regression model via the following iterative boosting Algorithm 1 with the Majorization-Minimization algorithm in Section 3.3 at its core. The computation complexity of each step is $O(nKq^2 + nmd)$. Here, n is the sample size, K is the number of components, q is the number of quantile levels, m is the dimension of $\boldsymbol{\theta}$, and d is the regression tree depth. The first part represents the complexity of the Majorization-Minimization algorithm, and the second part represents the complexity of fitting regression trees.

Besides using a fixed number of iterations, we could also use early stopping (Yao et al., 2007) to avoid overfitting and accelerate iterative optimization. Specifically, we calculate the Wasserstein loss $\sum_{i=1}^n W^2(g_i, f_{\boldsymbol{\tau}(x_i)})$ on a validation set along the training process and stop when it is no longer decreasing. To satisfy the identifiability constraint (Section 3.2), in each optimization step, the updated components will be sorted before feeding to the boosting machine. Further algorithmic details can be found in the Appendix.

Algorithm 1 Wasserstein Distribution Learning with Boosted Machines

Data: $\mathcal{D} = \{(x_i, g_i)\}_{i=1}^n$
Training Controls: Set learning rate $\eta > 0$, maximum of iterations, $M > 0$;
Random Initialization: Randomly sample $\tilde{\boldsymbol{\theta}}^{(0)} \overset{\text{iid}}{\sim} U(-0.5, 0.5)$, and fit regression tree $\boldsymbol{T}^{(0)}(x)$ to $\tilde{\boldsymbol{\theta}}^{(0)}$ as initialization $\widehat{\boldsymbol{\tau}}^{(0)}(x)$.
for $1 \leq m \leq M$ **do**
  Based on $\widehat{\boldsymbol{\tau}}^{(m-1)}(x)$, for $i = 1, \ldots, n$,
  $\widehat{\boldsymbol{\theta}}^{(m-1)}(x_i) \leftarrow \widehat{\boldsymbol{\tau}}^{(m-1)}(x_i) =$
    $\{\hat{\pi}_k^{(m-1)}(x_i), \hat{\mu}_k^{(m-1)}(x_i), \hat{\sigma}_k^{(m-1)}(x_i)\}_{k=1}^K$;
  procedure MAJORIZATION-MINIMIZATION
    for all $i = 1, \ldots, n$ **do**
      Based on $\widehat{\boldsymbol{\theta}}^{(m-1)}(x_i)$
      Derive MM estimate, $\tilde{\boldsymbol{\theta}}^{(m)}(x_i)$, against g_i.
    end for
  end procedure
  Fit regression tree $\boldsymbol{T}^{(m)}(x)$ to $\left(\tilde{\boldsymbol{\theta}}^{(m)} - \widehat{\boldsymbol{\theta}}^{(m-1)}\right)$
  Update $\widehat{\boldsymbol{\tau}}^{(m)}(x) \leftarrow \widehat{\boldsymbol{\tau}}^{(m-1)}(x) + \eta \boldsymbol{T}^{(m)}(x)$.
end for
  return $\widehat{\boldsymbol{\tau}}^{(M)}(x)$.

As opposed to the gradient-based optimization algorithm introduced in Bishop (1994) and Rothfuss et al. (2019), our model training framework utilizes the additive structure of tree ensembles and the upper bound of the Wasserstein loss to achieve a more stable and efficient solution path. In actuality, any machine learning algorithms can be used to represent and estimate the nonparametric coefficient functions. Compared with other models, such as polynomial functions and neural networks, the boosted decision trees achieve a balance between the expressive power and the generalization ability. Moreover, the tree structure greatly improves the model transparency and interpretability.

4 EXPERIMENTS

In this section, we demonstrate the estimation performance and prediction accuracy of the proposed WDL framework using simulations and two real-world applications. In all the experiments, we compare the WDL framework with two existing density regression methods: the global Fréchet

regression (Petersen et al., 2019), a generalization of linear regression in the quantile functional space under the Wasserstein loss, and the B-spline smoothed density regression with a centered log-ratio (CLR) transformation (Talská et al., 2018), which, for reasons of simplicity, is referred to as the CLR regression for the rest of the paper. The main ideas of these methods are provided in Appendix. Reproducible codes for generating all results are available on GitHub (`https://github.com/ChengliangTang/WDL_MM`).

4.1 Simulation Study

In this experiment, we consider multivariate covariates $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3})$ that are mutually independent and follow the uniform distribution on $[-1, 1]$. The conditional density outcomes $\mathcal{G}$ are generated as follows.

$$\mathcal{F}(\mathcal{G}|\mathbf{X} = x) = \pi_1(x) \cdot f_1(x) + \pi_2(x) \cdot f_2(x), \qquad (7)$$

where

$$\pi_1(x) = \frac{1}{1+\exp(x_3)}, \quad \pi_2(x) = \frac{\exp(x_3)}{1+\exp(x_3)},$$
$$f_1(x) = \mathrm{N}\big(x_1 + \varepsilon, (|x_2| + 0.5)^2\big),$$
$$f_2(x) = \mathrm{N}\big(2x_2^2 + 2 + \varepsilon, (|x_1| + 0.5)^2\big),$$

with independent random noise variable $\varepsilon \sim \mathrm{N}(0, \omega^2)$.

By design, we assume that the conditional distribution $\mathcal{F}(\mathcal{G}|\mathbf{X} = x)$ is a Gaussian mixture. The component-wise means and variances are functions of $\mathbf{x_1}$ and $\mathbf{x_2}$, while the component weights, $\pi_1(x)$ and $\pi_2(x)$, are governed by $\mathbf{x_3}$. In addition, we let $\mu_1(x) \leq \mu_2(x), \forall x \in \mathcal{X}$ to avoid identifiability issues. We also incorporate an additive random noise ε to the component-wise means $\mu_1(x)$ and $\mu_2(x)$, which is independent of all $\mathbf{X}$ variables. The additive random noise follows a zero-mean Gaussian distribution $\varepsilon \sim \mathrm{N}(0, \omega^2)$.

Given a noise level ω, we generate, from the model specified in (7), $N = 200$ random samples $(x_i, g_i) \sim \mathcal{F}(\mathbf{X}, \mathcal{G})$. More simulation details can be found in the Appendix. We apply the proposed WDL to the generated $\{(x_i, g_i)\}_{i=1}^{N=200}$ to estimate the parameter functions $\boldsymbol{\tau}(x) = \{\pi(x), \mu(x), \sigma^2(x)\}$, and derive the expected conditional distributions $\mathbb{E}(\mathcal{G}|\mathbf{X})$.

Using this simulation study with known ground truth parameter functions, we also demonstrate, in the Appendix, that the proposed WDL framework accurately recovers the parameter functions. Here, we evaluate the performance using two different measures: accuracy in estimating functional dependence of $\mathcal{G}$ on $\mathbf{X}$, and accuracy in predicting the functional outputs. The first measure focuses on the estimation performance of each method using the training set, and the second measure evaluates their generalization abilities from a training set to an independent test set.

For the first measure, we evaluate the average performance over 500 Monte Carlo replications. On each Monte Carlo replication, we begin by randomly splitting the data into the training set (80%) and the validation set (20%), along with choosing the best tuning parameter based on validation results. Then, we refit each model with the best tuning parameter over the entire data set. For the second measure, following what we would use in a real data scenario, we evaluate the performance using a nested five-fold cross validation. In order to minimize the optimism bias in performance evaluation, hyper-parameter selection and model training were performed using another layer of train-valid split over the training folds, at which point we evaluated the prediction loss on the held-out test fold. We applied parameter tuning to WDL and CLR regression. No tuning step was needed for the Fréchet regression.

Accuracy in estimating functional dependence of $\mathcal{G}$ on $\mathbf{X}$. Here we compare how well the three methods estimate the conditional quantile functions of $\mathcal{G}$ given $\mathbf{X}$. We generalize partial dependence plot (PDP) to the quantile functional space to measure the estimation accuracy. At a given quantile level $0 < \rho < 1$, let the target covariate be $\mathbf{X_s}$, and the set of all other covariates be $\mathbf{X_c}$. We define the corresponding functional PDP at point value $\mathbf{X_s} = x_s$ as $PD_{\mathbf{X_s}}(x_s; \rho) = \int_{x_c \in \mathcal{X}_c} F^{-1}_{\boldsymbol{\tau}(x_s, x_c)}(\rho) f_{\mathbf{X_c}}(x_c) dx_c$, where $f_{\mathbf{X_c}}$ is the marginal density of $\mathbf{X_c}$. Figure 2 compares the functional PDPs estimated by the three methods with the ground truth at various quantile levels in $\rho \in \{10\%, 30\%, 50\%, 70\%, 90\%\}$. The functional PDPs of the ground truth correspond to the conditional expectation of the functional outputs $\mathbb{E}(\mathcal{G}|\mathbf{X})$. The functional PDPs of the three methods were calculated using 500 Monte Carlo replications with noise level $\omega = 0.1$. As shown in Figure 2, WDL is capable of capturing the heterogeneity in the partial dependence curves and is closest to the ground truth. The partial dependence curves for Fréchet regression are all approximately straight lines of different slopes, due to its linearity assumption. As for the CLR regression, their fitted functional partial dependence curves are also constrained by a linearity assumption after the centered log-ratio transformation.

Table 1: Predictive performance comparison at different noise levels: Wasserstein loss and $\widehat{R}^2$ (in bracket).

ω	0.1	0.2	0.5	1	2
WDL	**0.05 (0.9)**	**0.1 (0.9)**	**0.3 (0.7)**	**1.1 (0.3)**	3.9 (0.02)
Fréchet	0.3 (0.5)	0.3 (0.5)	0.5 (0.4)	1.2 (0.3)	**3.9 (0.03)**
CLR	0.3 (0.5)	0.3 (0.5)	0.5 (0.4)	1.2 (0.3)	4.0 (0.00)

Accuracy in predicting g_i's. We used nested five-fold cross validations to evaluate the predictive performance of WDL with comparison to the other methods. To nu-

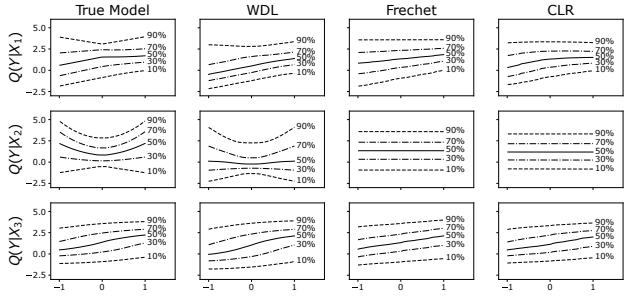

Figure 2: Functional partial dependence plots (PDP) for predicted conditional quantiles versus the input scalar variables. The results are averaged over 500 Monte Carlo replications with noise level $\omega = 0.1$.

merically measure the discrepancy between the observed quantile functions and their model predictions from the test samples, we defined an approximated Wasserstein distance from a dense array of equally spaced quantile levels $\{0.01, 0.02, ..., 0.99\}$,

$$W_i^2 = \int_0^1 \left(Q_i(s) - \widehat{Q}_i^{(cv)}(s)\right)^2 ds \approx \frac{1}{100}\sum_{i=1}^{99}\left(Q_i(\frac{i}{100}) - \widehat{Q}_i^{(cv)}(\frac{i}{100})\right)^2,$$

where Q_i and $\widehat{Q}_i^{(cv)}$ are the observed and predicted quantile functions for a test sample in cross validation, and W_i^2 is the Wasserstein prediction loss of the i-th observation. We also define an R^2-like statistic:

$$\widehat{R}^2 = 1 - \frac{\sum_{i=1}^N W_i^2 \Big/ N}{Var(\mathcal{G})},$$

where the variance of $\mathcal{G}$ is approximated by

$$Var(\mathcal{G}) \approx \frac{1}{N}\sum_{i=1}^N \int_0^1 \left(Q_i(s) - \bar{Q}(s)\right)^2 ds$$

and $\bar{Q}(s) = \frac{1}{N}\sum_{i=1}^N Q_i(s)$ for any $s \in [0,1]$. Table 1 summarizes the average Wasserstein loss and R-square at different noise levels ($\omega = 0.1, 0.2, 0.5, 1$ and 2) using a nested five-fold cross validation. The prediction accuracy (measured by average Wasserstein loss) and power (R-square) decline as the noise level increases. At most noise levels ($\omega = 0.1, 0.2, 0.5$ and 1), WDL delivers the best prediction due to its ability to model complicated density output. When the noise level is high ($\omega = 2$), the Fréchet regression performs slightly better than the others due to the robustness of its linear model assumption.

4.2 Modeling Annual Temperature Distributions

A fundamental step in climate research is to identify the factors that impact the radiative balance of the planet and are expected to change the temperature distribution. In this section, we apply the proposed Wasserstein distributional learning to understand how the radiative effects, or "radiative forcings" of solar irradiance, volcanic eruptions, and CO2, as well as natural climate variability through the El-Niño Southern Oscillation (ENSO) are associated with annual temperature distributions, using data from 1880-2012. See the Appendix for a detailed description of the data set.

Here, we set the number of mixture components as three, which correspond to: low temperatures (Component I), medium temperatures (Component II), and high temperatures (Component III). We fit the proposed Wasserstein distributional learning between the annual temperature quantile functions and the four environmental drivers. To avoid overfitting, we run a nested five-fold cross validation with hyper-parameter selection (learning rate and number of iterations) and calculated the predicted density function for each year when it was in the *test* fold. Figure 3 is the histogram of daily temperatures on selected years overlain with the model estimated temperature density curve. The results clearly demonstrate that our method effectively captures the heterogeneity in the functional outputs. Results for each year in the data set can be found in the Appendix.

One advantage of distributional learning is that it provides the utility to predict any distributional features of interest such as the center, spread, and tail behaviors of the distributional outcome. Here, we evaluate the three methods' performance in predicting extreme temperatures. For each observed year, we calculate the number of days above the 90th percentile daily threshold (high temperatures) and the number of days below the 10th percentile daily threshold (low temperatures). The 10th and 90th percentiles are derived from a 30-year climatological baseline period (1981-2010). In Figure 4, we visualize the ground truth and the test-fold predictions from each method. WDL and Fréchet regression achieve the best prediction performance in terms of R-squared. WDL is the only method that is able to capture the "plateau" of cold days and achieves positive R-squared within the 1925 - 1975 time window. This is due to WDL's better characterization of the nonlinear dependence of conditional quantiles. See more discussion in Appendix.

In Figure 5, we visualize the predicted density curves of each component versus CO2 and ENSO radiative forcings. For each given value of a physical driver $\mathbf{X_s} = x_s$, we compute its marginal prediction of the distribution parameters by averaging over all other covariates. We further visualize the density curve of each component using different colors according to the value of $\mathbf{X_s} = x_s$.

As shown in the figure, as the CO2 radiative forcing increases, the mean temperature of all the three components slowly increases. CO2 also substantially rearranges the weights among the three components. Different from CO2, ENSO primarily influences the weight and variance of each

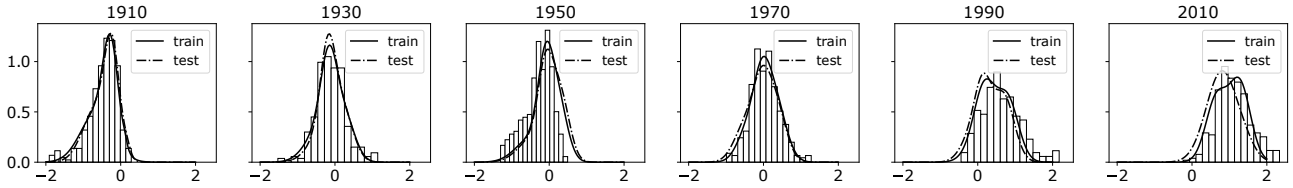

Figure 3: Selected predictions of annual temperature distributions.

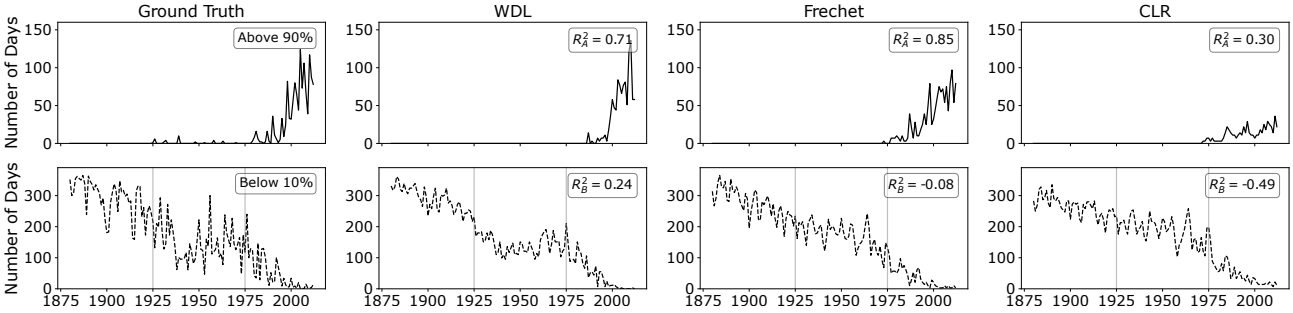

Figure 4: Predicted numbers of days with extreme temperatures. R^2_A: Overall test R-squared for predicting the number of days above 90th percentile. R^2_B: Local test R-squared for predicting the number of days below 10th percentile in the 1925 - 1975 time window.

component, which results in more frequent instances of extreme weather.

4.3 Modeling Regional Income Distributions

Modeling income distribution is a central topic in macroeconomic studies. Several indices, including Gini index, are widely used to characterize the income distributions. In this section, we apply WDL to model the regional income distribution of the 167 counties in New York, California and Michigan, from which one could derive multiple indices simultaneously and explicitly study their joint distributions. See the Appendix for a detailed data description.

We use WDL to model the association between the regional income distributions against the scalar county health factors. The income data were log-transformed, which is a common practice for highly skewed distributions.

As with the temperature distribution modeling example, WDL showed similar excellent predictive performance in modeling the income density curve for each county when it was in the *test* fold. See the Appendix for the full results. Here, we focus on demonstrating WDL's performance in predicting distributional features such as derived statistics and their inter-dependence. From the predicted distributions when the counties were in the test fold, we calculated three commonly used indices by the economists – Gini index, median income, and poverty rate, and then compared them with the true values.

In Table 2, we compared the proposed WDL method with the other methods in terms of estimating individual indices, using RMSE and the conventional R^2. In addition to the two comparison methods, we also implemented two methods that directly model the indices: lasso regression and tree regression, which corresponds to the conventional approach of modeling summary statistics in macroeconomics. As shown in Table 2, these index-based methods adequately model the observed indices individually. Our WDL algorithm offered the best performance for all the indices, even outperforming the index-based methods (lasso regression and tree regression). In Figure 6, we evaluate the estimated joint distribution of median income versus poverty rate calculated from the predicted income distributions for the counties under study. As shown in Figure 6, WDL is able to accurately capture the relationship between indices (summary statistics) without directly modeling them. In particular, predictions based on WDL preserve the true nonlinear association between the

Table 2: Performance comparison in terms of RMSE (and R^2) of different indices of income distributions. Results are evaluated on the test folds.

Method	Gini Index	Median Income	Poverty Rate
WDL	**0.029 (0.2)**	**4017.4 (0.4)**	**0.037 (0.3)**
Fréchet	0.052 (-1.6)	5113.1 (-0.0)	0.054 (-0.5)
CLR	0.030 (0.2)	11065.0 (-3.8)	0.040 (0.2)
Lasso Reg.	0.030 (0.2)	4433.3 (0.2)	0.041 (0.1)
Tree Reg.	0.032 (0.1)	4536.0 (0.2)	0.039 (0.2)

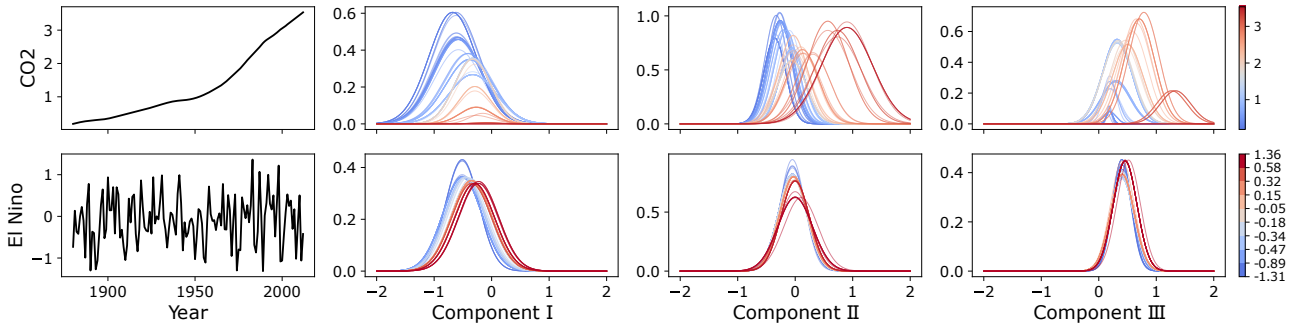

Figure 5: Density curves versus physical drivers. Column 1: temporal trends of physical driver values. Column 2-4: Density curves corresponding to each of three components in the fitted Gaussian mixture, with the color of curves representing the value of the corresponding physical driver.

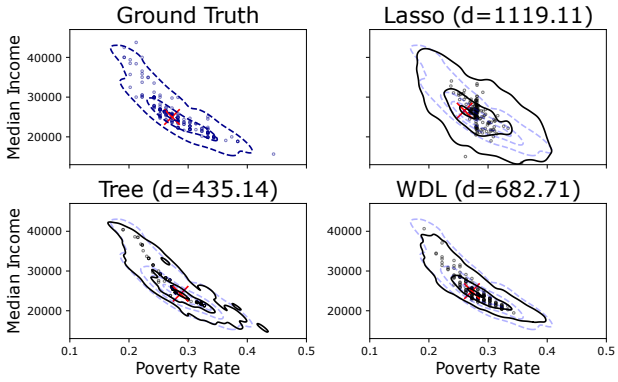

Figure 6: Joint distributions of median income versus poverty rate. Scatterplots overlaid with contour lines. Contour plots of ground truths are visualized as transparent dashed contour curves in each subplot. The modes of the contour plots are marked by red crosses. Wasserstein distances to the ground truth are shown in the sub-figure titles.

two indices without overfitting, offering both stability and flexibility in estimation. In comparison, non-Wasserstein-distance-based methods such as CLR methods would fail on these tasks. In particular, in Table 2, CLR methods have the worst performance in modeling median income. See the Appendix for detailed computation details and more results on these indices.

5 CONCLUSION

Predictive distributional learning is important in data-driven discoveries. The main contribution of our paper is a novel and efficient function-on-scalar regression framework for modeling distributional outputs. By definition, our framework satisfies the inherent constraints of density functions, and is capable of modeling highly heterogeneous outputs. We offer theoretical guarantees for the convergence of the proposed algorithm. Compared with other methods in the literature, our proposed WDL framework better captures the nonlinear dependence of the density functions over the covariates. Moreover, this framework produces more convenient and accurate predictions for derived density summary statistics of interest.

References

American Community Survey (2014). Public Use Microdata Sample (PUMS). `https://www2.census.gov/programs-surveys/acs/data/pums/`. [Online; accessed 16-March-2021].

Arata, Y. (2017). A functional linear regression model in the space of probability density functions. Technical report, Research Institute of Economy, Trade and Industry (RIETI).

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Berkeley Earth (2021). Daily Land Temperature. `http://berkeleyearth.lbl.gov/auto/Global/Complete_TAVG_daily.txt`. [Online; accessed 16-March-2021].

Bernton, E. (2019). *Optimal Transport in Statistical Inference and Computation*. PhD thesis, Harvard University.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). On parameter estimation with the wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676.

Bishop, C. M. (1994). Mixture density networks.

Boogaart, K. G. v. d., Egozcue, J. J., and Pawlowsky-Glahn, V. (2010). Bayes linear spaces. *SORT: statistics and operations research transactions, 2010, vol. 34, núm. 4, p. 201-222*.

Delicado, P. (2011). Dimensionality reduction when data are density functions. *Computational Statistics & Data Analysis*, 55(1):401–420.

Duy, V. N. L. and Takeuchi, I. (2022). Exact statistical inference for the wasserstein distance by selective inference: Selective inference for the wasserstein distance. *Annals of the Institute of Statistical Mathematics*, pages 1–31.

Egozcue, J. J., Díaz-Barrero, J. L., and Pawlowsky-Glahn, V. (2006). Hilbert space of probability density functions based on aitchison geometry. *Acta Mathematica Sinica*, 22(4):1175–1182.

Fahey, D., Doherty, S., Hibbard, K. A., Romanou, A., and Taylor, P. (2017). Physical drivers of climate change. *Climate Science Special Report: Fourth National Climate Assessment*, 1:73–113.

Field, C. B., Barros, V., Stocker, T. F., and Dahe, Q. (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge University Press.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59.

Lewis, S. C. and King, A. D. (2017). Evolution of mean, variance and extremes in 21st century temperatures. *Weather and climate extremes*, 15:1–10.

Martín-Fernández, J.-A., Hron, K., Templ, M., Filzmoser, P., and Palarea-Albaladejo, J. (2015). Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2):134–158.

McPhaden, M. J., Santoso, A., and Cai, W. (2020). *El Niño Southern Oscillation in a changing climate*, volume 253. John Wiley & Sons.

Miller, R. L., Schmidt, G. A., Nazarenko, L. S., Tausnev, N., Bauer, S. E., DelGenio, A. D., Kelley, M., Lo, K. K., Ruedy, R., Shindell, D. T., et al. (2014). Cmip5 historical simulations (1850–2012) with giss modele2. *Journal of Advances in Modeling Earth Systems*, 6(2):441–478.

Mueller, J., Jaakkola, T., and Gifford, D. (2018). Modeling persistent trends in distributions. *Journal of the American Statistical Association*, 113(523):1296–1310.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6:405–431.

Pegoraro, M. and Beraha, M. (2022). Projected statistical methods for distributional data on the real line with the wasserstein metric. *J. Mach. Learn. Res.*, 23:37–1.

Petersen, A., Liu, X., and Divani, A. A. (2021). Wasserstein f-tests and confidence bands for the fréchet regression of density response curves. *The Annals of Statistics*, 49(1):590–611.

Petersen, A., Müller, H.-G., et al. (2019). Fréchet regression for random objects with euclidean predictors. *The Annals of Statistics*, 47(2):691–719.

Reich, B. J. (2012). Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(4):535–553.

Reiss, P. T., Huang, L., and Mennes, M. (2010). Fast function-on-scalar regression with penalized basis expansions. *The international journal of biostatistics*, 6(1).

Rothfuss, J., Ferreira, F., Walther, S., and Ulrich, M. (2019). Conditional density estimation with neural networks: Best practices and benchmarks. *arXiv preprint arXiv:1903.00954*.

Sgouropoulos, N., Yao, Q., and Yastremiz, C. (2015). Matching a distribution by matching quantiles estimation. *Journal of the American Statistical Association*, 110(510):742–759.

Sharma, A. and Gerig, G. (2020). Trajectories from distribution-valued functional curves: A unified wasserstein framework. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23*, pages 343–353. Springer.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the american statistical association*, 21(153):65–66.

Suckling, E. B., van Oldenborgh, G. J., Eden, J. M., and Hawkins, E. (2017). An empirical model for probabilistic decadal prediction: global attribution and regional hindcasts. *Climate Dynamics*, 48(9):3115–3138.

Talská, R., Menafoglio, A., Machalová, J., Hron, K., and Fišerová, E. (2018). Compositional regression with functional response. *Computational Statistics & Data Analysis*, 123:66–85.

The County Health Rankings & Roadmaps (2014). County Health Rankings. `https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/2021-measures`. [Online; accessed 16-March-2021].

Van den Boogaart, K. G., Egozcue, J. J., and Pawlowsky-Glahn, V. (2014). Bayes hilbert spaces. *Australian & New Zealand Journal of Statistics*, 56(2):171–194.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Villani, C. (2003). *Topics in optimal transportation*. Number 58. American Mathematical Soc.

Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295.

Yao, Y., Rosasco, L., and Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315.

Zhang, C., Kokoszka, P., and Petersen, A. (2022). Wasserstein autoregressive models for density time series. *Journal of Time Series Analysis*, 43(1):30–52.

A NOTATIONAL TABLE

In this section, we list all the main notations in the following table

Table A1: Main notations in the paper.

Team sheet	
$\mathbb{R}$	The 1-dimensional space of real line.
$\mathbb{R}^p$	The p-dimensional Euclidean space.
$\mathbf{X}$	Random covariate vector.
$\mathcal{X}$	The support of $\mathbf{X}$. $\mathcal{X} \subset \mathbb{R}^p$
$f_{\mathbf{X}}$	The probability density function (PDF) of $\mathbf{X}$.
$\mathcal{G}$	A distributional response representing the distribution of a random variable.
$\mathcal{P}_2(\mathbb{R})$	The set of all Borel probability measures on $\mathbb{R}$ with a finite 2nd moment.
$f_{\boldsymbol{\theta}}$	Distribution function parameterized by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.
$\mathcal{T}(\mathcal{X}, \boldsymbol{\Theta})$	Non-parametric functional family of mappings from $\mathcal{X}$ to $\boldsymbol{\Theta}$.

B THEOREM PROOFS

B.1 Proof of Theorem 1

Proof. To prove this theorem, we first introduce a lemma for the dense property of Gaussian mixture models in the functional space $\big(\mathcal{P}_2(\mathbb{R}), W_2\big)$.

Lemma 1. *Let $\mathfrak{F}_G \subset \mathcal{P}_2(\mathbb{R})$ be the family of finite Gaussian mixture distributions over the real line $\mathbb{R}$, i.e.,*

$$\mathfrak{F}_G = \left\{ \sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2) \middle| K \in \mathbb{N}_+ \right\}.$$

Then the family $\mathfrak{F}_G$ is dense in $\big(\mathcal{P}_2(\mathbb{R}), W_2\big)$, i.e., the set of probability measures of a finite second moment with the 2-Wasserstein distance.

The proof of this lemma follows the idea of Theorem 6.18. in Villani (2003). It suffices to show that for any probability distribution $g \in \mathcal{P}_2(\mathbb{R})$ and any given constant $\varepsilon > 0$, there exists a finite Gaussian mixture distribution $f = \sum_{k=1}^{K} \pi_k f_k$, where $f_k = \mathrm{N}(\mu_k, \sigma_k^2)$, such that $W_2(f, g) < \varepsilon$.

We prove the above claim in two steps.

First, since $g \in \mathcal{P}_2(\mathbb{R})$, it has a finite second moment, which means $\mathbb{E}_g(X^2) < \infty$. Then, there exists a constant $M > 0$ large enough, such that

$$\mathbb{E}_g[X^2 \mathbf{1}_{\{|X|>M\}}] < \frac{\varepsilon^2}{9}.$$

Cover the compact set $[-M, M]$ by a finite family of balls $\{B(x_k, \varepsilon/3)\}_{1 \le k \le K}$, with centers $x_k \in [-M, M]$, and define

$$B'_k = \begin{cases} B(x_1, \varepsilon/3) & \text{if } k = 1 \\ B(x_k, \varepsilon/3) \backslash \bigcup_{j<k} B(x_j, \varepsilon/3) & \text{if } k = 2, ..., K \end{cases}$$

Then all B'_k are disjoint and still cover $[-M, M]$.

Define function J on $\mathbb{R}$ by

$$J(x) = \begin{cases} x_k & \text{if } x \in B'_k \cap [-M, M] \text{ for some } k \\ 0 & \text{if } x \in \mathbb{R} \backslash [-M, M] \end{cases}$$

Then, for any $x \in [-M, M]$, we have $|x - J(x)| < \varepsilon/3$, which leads to the following inequality

$$\begin{aligned}\mathbb{E}_g[(X - J(X))^2] &= \mathbb{E}_g[(X - J(X))^2 \mathbf{1}_{\{|X|>M\}}] + \mathbb{E}_g[(X - J(X))^2 \mathbf{1}_{\{|X|\le M\}}] \\ &\le \mathbb{E}_g[(X - 0)^2 \mathbf{1}_{\{|X|>M\}}] + \frac{\varepsilon^2}{9} \cdot \mathbb{E}_g[\mathbf{1}_{\{|X|\le M\}}] \\ &< \frac{\varepsilon^2}{9} + \frac{\varepsilon^2}{9} < \frac{\varepsilon^2}{4}.\end{aligned} \tag{A1}$$

Suppose random variable $X \sim g$, we denote $\tilde{g}$ as the distribution of $J(X)$, saying $J(X) \sim \tilde{g}$. Then, by the construction of J, the distribution $\tilde{g}$ can be written as $\tilde{g} = \sum_{k=1}^{K} \pi_k \delta_{x_k}$, where δ_{x_k} is the point mass at x_k. Moreover, using the definition of the Wasserstein distance, from Equation (A1) we have

$$W_2(g, \tilde{g}) \le \sqrt{\mathbb{E}_g[(X - J(X))^2]} < \frac{\varepsilon}{2}. \tag{A2}$$

Second, we approximate each point mass δ_{x_k} by a Gaussian distribution $\mathrm{N}(x_k, \sigma_k^2)$. Let $f = \sum_{k=1}^{K} \pi_k \mathrm{N}(x_k, \varepsilon^2/4)$, then using Theorem 3 (Majorization property) we can have

$$\begin{aligned}W_2^2(f, \tilde{g}) &= W_2^2\Big(\sum_{k=1}^{K} \pi_k \delta_{x_k}, \sum_{k=1}^{K} \pi_k \mathrm{N}(x_k, \varepsilon^2/4)\Big) \\ &\le \sum_{k=1}^{K} \pi_k W_2^2\Big(\delta_{x_k}, \mathrm{N}(x_k, \varepsilon^2/4)\Big) \\ &= \varepsilon^2/4.\end{aligned}$$

As a result, we have

$$W_2(f, \tilde{g}) \le \frac{\varepsilon}{2}. \tag{A3}$$

In conclusion, with $f = \sum_{k=1}^{K} \pi_k \mathrm{N}(x_k, \varepsilon^2/4)$ as defined above, combining Equation (A2) and Equation (A3) we have

$$W_2(f, g) \le W_2(f, \tilde{g}) + W_2(\tilde{g}, g) < \varepsilon,$$

which means the family $\mathfrak{F}_G$ is dense in $(\mathcal{P}_2(\mathbb{R}), W_2)$. Thus Lemma 1 is proved.

Now, back to Theorem 1, we prove it in three steps. First, we prove the special case of compact support $\mathcal{X}$. Second, we extend the proof to the case of closed support $\mathcal{X}$. Finally, we prove the theorem for a general $\mathcal{X} \subset \mathbb{R}^p$.

Step 1. First, suppose $\mathcal{X}$ is a compact set in $\mathbb{R}^p$, we can prove a stronger version of the theorem, i.e., there exists a step function $\boldsymbol{\tau}(\cdot) \in \mathcal{T}(\mathcal{X}, \boldsymbol{\Theta})$ such that

$$W_2\big(H(x), f_{\boldsymbol{\tau}(x)}\big) < \varepsilon, \quad \text{for } \forall x \in \mathcal{X}.$$

In fact, by Lipschitz continuity assumption, for $\forall x_1, x_2 \in \mathcal{X}$ and $\|x_1 - x_2\|_2 < \frac{\varepsilon}{3L}$, we have

$$W_2\big(H(x_1), H(x_2)\big) \le L \cdot \frac{\varepsilon}{3L} = \varepsilon/3.$$

Let $\delta = \frac{\varepsilon}{6L\sqrt{p}} > 0$, we define $\delta-$box each $x \in \mathcal{X}$ as

$$B(x, \delta) = \bigotimes_{i=1}^{p} (x^{(i)} - \delta, x^{(i)} + \delta),$$

which is an open square-shaped neighbourhood covering $x \in \mathcal{X} \subset \mathbb{R}^p$. Also, we have $diam(B(x, \delta)) < \frac{\varepsilon}{3L}$ for each x under the Euclidean distance. Since $\mathcal{X} \subset \bigcup_{x \in \mathcal{X}} B(x, \delta)$, and $\mathcal{X}$ is compact, there exists a finite set $\{x_i\}_{i=1}^{N} \subset \mathcal{X}$, such that

$$\mathcal{X} \subset \bigcup_{i=1}^{N} B(x_i, \delta).$$

Define

$$\widetilde{B}_i = B(x_i, \delta) \backslash \bigcup_{j<i} B(x_j, \delta),$$

then all $\widetilde{B}_i$ are disjoint and still cover $\mathcal{X}$. With the constructed finite set $\{\widetilde{B}_i\}_{i=1}^N$, we define a function $\widetilde{H}(\cdot)$, such that

$$\widetilde{H}(x) = H(x_i), \quad \text{if } x \in \widetilde{B}_i.$$

By the definition of δ, we have $W_2\big(H(x), \widetilde{H}(x)\big) < \varepsilon/3$ for any $x \in \mathcal{X}$.

By Lemma 1, for for each $H(x_i)$, there exists a Gaussian mixture distribution $f_i = \sum_k^{K_i} \pi_{(k;i)} \mathrm{N}(\mu_{(k;i)}, \sigma^2_{(k;i)})$ such that $W_2(f_i, H(x_i)) < \varepsilon/3, i = 1, ..., N$. Let $K = \max_i K_i$, and further decompose each Gaussian mixture distribution into K components. Specifically, for the Gaussian mixture distribution f_i with $K_i < K$ components, we create another Gaussian mixture distribution $\tilde{f}_i$ by equally dividing weight of the last component into $(K - K_i + 1)$ components. Therefore, without loss of generality, here we simply assume each mixture distribution f_i has the same number of components as K, and the components are following the increasing order of their means, saying $\mu_{(1;i)} \leq \mu_{(2;i)} \leq ... \leq \mu_{(K;i)}$.

With the Gaussian mixture distributions $f_i = \sum_k^K \pi_{(k;i)} \mathrm{N}(\mu_{(k;i)}, \sigma^2_{(k;i)})$ in place, we construct the following step $\boldsymbol{\tau}(x) = \{\pi_k(x), \mu_k(x), \sigma_k^2(x)\}_{k=1}^K$. For each $x \in \mathcal{X}$, it belongs to one and only one $\widetilde{B}_i$. Then, for $k = 1, ..., K$, we let

$$\pi_k(x) = \pi_{(k;i)}, \quad \mu_k(x) = \mu_{(k;i)}, \quad \sigma_k(x) = \sigma_{(k;i)}, \quad \text{if } x \in \widetilde{B}_i.$$

By definition, each $\widetilde{B}_i$ is the difference between a series of $\delta-$boxes, which makes their boundaries piecewise axis-parallel. Therefore, the above step function construction $\boldsymbol{\tau}(x)$ is feasible. Let $f_{\boldsymbol{\tau}(x)} = \sum_k^K \pi_k(x) \cdot \mathrm{N}\big(\mu_k(x), \sigma_k^2(x)\big)$, we have

$$W_2\big(f_{\boldsymbol{\tau}(x)}, \widetilde{H}(x)\big) < \varepsilon/3, \quad \text{for } \forall x \in \mathcal{X}.$$

In conclusion, by combining the two parts, we have

$$\begin{aligned} W_2\big(H(x), f_{\boldsymbol{\tau}(x)}\big) \leq & W_2\big(H(x), \widetilde{H}(x)\big) + W_2\big(\widetilde{H}(x), f_{\boldsymbol{\tau}(x)}\big) \\ < & \varepsilon/3 + \varepsilon/3 < \varepsilon, \quad \text{for } \forall x \in \mathcal{X}. \end{aligned}$$

Step 2. Second, we prove the theorem for closed support $\mathcal{X} \subset \mathbb{R}^p$. We choose M large enough, then the integration can be decomposed as

$$\begin{aligned} \int_{x \in \mathcal{X}} W_2(H(x), f_{\boldsymbol{\tau}(x)}) dP_{\mathbf{X}}(x) = & \int_{\{x \in \mathcal{X} | \max_i |x_i| \leq M\}} W_2(H(x), f_{\boldsymbol{\tau}(x)}) dP_{\mathbf{X}}(x) \\ & + \int_{\{x \in \mathcal{X} | \max_i |x_i| > M\}} W_2(H(x), f_{\boldsymbol{\tau}(x)}) dP_{\mathbf{X}}(x). \end{aligned}$$

Because $\mathcal{X} \subset \mathbb{R}^p$ is closed, $S = \{x \in \mathcal{X} | \max_i |x_i| \leq M\}$ is compact for any finite $M > 0$. As proved in Step 1, there exist tree models $\boldsymbol{\tau}(x)$ defined over S such that

$$W_2(H(x), f_{\boldsymbol{\tau}(x)}) < \varepsilon/3, \ \forall x \in S = \{x \in \mathcal{X} | \max_i |x_i| \leq M\}.$$

Thus, we have

$$\int_{\{x \in \mathcal{X} | \max_i |x_i| \leq M\}} W_2(H(x), f_{\boldsymbol{\tau}(x)}) dP_{\mathbf{X}}(x) < \varepsilon/3.$$

Further, we choose an arbitrary fixed point $x_0 \in S = \{x \in \mathcal{X} | \max_i |x_i| \leq M\}$, and extend $\boldsymbol{\tau}(x)$ to the entire $\mathcal{X}$ by letting $\boldsymbol{\tau}(x) = \boldsymbol{\tau}(x_0)$, for $\forall x \in S = \{x \in \mathcal{X} | \max_i |x_i| > M\}$.

By the Lipschitz continuity assumption, we have

$$\begin{aligned}
&\int_{S=\{x\in\mathcal{X}|\max_i |x_i|>M\}} W_2(H(x), f_{\boldsymbol{\tau}(x)})dP_{\mathbf{X}}(x)\\
\leq &\int_S W_2(H(x), H(x_0))dP_{\mathbf{X}}(x) + \int_S W_2(H(x_0), f_{\boldsymbol{\tau}(x)})dP_{\mathbf{X}}(x)\\
< &\int_S L\|x-x_0\|_2 dP_{\mathbf{X}}(x) + \int_S W_2(H(x_0), f_{\boldsymbol{\tau}(x_0)})dP_{\mathbf{X}}(x)\\
< &\varepsilon/3 + \varepsilon/3, \text{ as } M\longrightarrow\infty.
\end{aligned}$$

The first term is because of light tail assumption of $P_{\mathbf{X}}$, and the second term is by the definition of $\boldsymbol{\tau}(x)$.

In conclusion, we have

$$\begin{aligned}
\int_{x\in\mathcal{X}} W_2(H(x), f_{\boldsymbol{\tau}(x)})dP_{\mathbf{X}}(x) = &\int_{\{x\in\mathcal{X}|\max_i |x_i|\leq M\}} W_2(H(x), f_{\boldsymbol{\tau}(x)})dP_{\mathbf{X}}(x)\\
&+\int_{\{x\in\mathcal{X}|\max_i |x_i|>M\}} W_2(H(x), f_{\boldsymbol{\tau}(x)})dP_{\mathbf{X}}(x)\\
&<\frac{\varepsilon}{3}+\frac{\varepsilon}{3}+\frac{\varepsilon}{3}=\varepsilon.
\end{aligned}$$

Step 3. Finally, we prove the theorem for general $\mathcal{X}\subset\mathbb{R}^p$. In fact, due to the continuity assumption, we can extend $H(x):\mathcal{X}\longrightarrow\mathcal{P}(\mathbb{R})$ to $\overline{\mathcal{X}}$, the closure of $\mathcal{X}$. We define $\overline{H}(x):\overline{\mathcal{X}}\longrightarrow\mathcal{P}(\mathbb{R})$ as follows

$$\overline{H}(x)=\begin{cases} H(x) & \text{if } x\in\mathcal{X}\\ \lim_{y\to x,\ y\in\mathcal{X}} H(y) & \text{if } x\in\overline{\mathcal{X}}\setminus\mathcal{X}\end{cases}$$

Moreover, we can generalize $P_{\mathbf{X}}$ to $\overline{P}_{\overline{\mathbf{X}}}$ by letting $\overline{P}_{\overline{\mathbf{X}}}(A)=P_{\mathbf{X}}(A)$ for any $A\subset\mathcal{X}$ and $\overline{P}_{\overline{\mathbf{X}}}(\overline{\mathcal{X}}\setminus\mathcal{X})=0$.

By the result of Step 2, we can find $f_{\boldsymbol{\tau}(x)}$ such that the condition is satisfied. Therefore, over $\mathcal{X}$, we have

$$\begin{aligned}
&\int_{\mathcal{X}} W_2(H(x), f_{\boldsymbol{\tau}(x)})dP_{\mathbf{X}}(x)\\
=&\int_{\overline{\mathcal{X}}} W_2(\overline{H}(x), f_{\boldsymbol{\tau}(x)})d\overline{P}_{\overline{\mathbf{X}}}(x)<\varepsilon.
\end{aligned}$$

Thus, Theorem 1 is proved. □

B.2 Proof of Theorem 2

Proof. This theorem is an extension of Theorem 5.14 in Van der Vaart (2000). The referenced theorem proves the consistency of M-estimators under regularity assumptions. In our case, the SCGMM estimators are a special case of M-estimators if we generalize the model parameter space $\boldsymbol{\Theta}$ from a Euclidean space into the functional space $\mathcal{T}(\mathcal{X},\boldsymbol{\Theta})$.

For any $\widehat{\boldsymbol{\tau}}\in\mathcal{T}(\mathcal{X},\boldsymbol{\Theta})$ such that $\mathcal{M}(\widehat{\boldsymbol{\tau}})<\infty$, let $U_l\downarrow\widehat{\boldsymbol{\tau}}$ be a decreasing sequence of open balls covering $\widehat{\boldsymbol{\tau}}$ of diameter converging to zero. For any $(x,g)\in\mathcal{X}\times\mathcal{P}_2(\mathbb{R})$ and any open ball $U\subset\mathcal{T}(\mathcal{X},\boldsymbol{\Theta})$, define $\boldsymbol{m}_U(x,g)=\inf_{\boldsymbol{\tau}\in U} W_2^2(f_{\boldsymbol{\tau}(x)},g)$ and $\mathcal{M}(U)=\int_{\mathcal{X}\times\mathcal{P}_2(\mathbb{R})}\boldsymbol{m}_U(x,g)d\mathcal{F}(x,g)$. Then, by the construction of $\{U_l\}$ and the continuity of $f_{\boldsymbol{\theta}}$, the sequence $\boldsymbol{m}_{U_l}(x,g)\uparrow W_2^2(f_{\widehat{\boldsymbol{\tau}}(x)},g)$ for $(x,g)\in\mathcal{X}\times\mathcal{P}_2(\mathbb{R})$ almost surely. Further, by the dominated convergence theorem and the finite assumption of $W_2^2(f_{\widehat{\boldsymbol{\tau}}(x)},g)$, we have $\int_{\mathcal{X}\times\mathcal{P}_2(\mathbb{R})}\boldsymbol{m}_{U_l}(x,g)d\mathcal{F}(x,g)\uparrow\mathcal{M}(\widehat{\boldsymbol{\tau}}(x))=\int_{\mathcal{X}\times\mathcal{P}_2(\mathbb{R})} W_2^2(f_{\widehat{\boldsymbol{\tau}}(x)},g)d\mathcal{F}(x,g)<\infty$.

By definition, $\arg\min_{\boldsymbol{\tau}\in\mathcal{T}(\mathcal{X},\boldsymbol{\Theta})}\mathcal{M}(\boldsymbol{\tau})\subseteq\widetilde{\boldsymbol{\tau}}_0$. For any $\boldsymbol{\tau}\notin\widetilde{\boldsymbol{\tau}}_0$, due to the uniqueness assumption we have $\mathcal{M}(\boldsymbol{\tau})>\mathcal{M}(\boldsymbol{\tau}_0)$. Combine this with the preceding paragraph to see that for every $\boldsymbol{\tau}\neq\boldsymbol{\tau}_0$, there exists an open ball $U_{\boldsymbol{\tau}}$ around $\boldsymbol{\tau}$ with $\mathcal{M}(U_{\boldsymbol{\tau}})>\mathcal{M}(\boldsymbol{\tau}_0)$. The set $B=\{\boldsymbol{\tau}\in S: d(\boldsymbol{\tau},\boldsymbol{\tau}_0)\geq\varepsilon\}$ is compact and covered by the balls $\{U_{\boldsymbol{\tau}}:\boldsymbol{\tau}\in B\}$. Let $U_{\boldsymbol{\tau}_1},...,U_{\boldsymbol{\tau}_p}$ be a finite sub-cover of B, then by the law of large numbers,

$$\min_{\boldsymbol{\tau}\in B}\mathcal{M}_n(\boldsymbol{\tau})\geq\min_{j=1,...,p}\mathcal{M}_n(U_{\boldsymbol{\tau}_j})\xrightarrow{a.s.}\min_{j=1,...,p}\mathcal{M}(U_{\boldsymbol{\tau}_j})>\mathcal{M}(\boldsymbol{\tau}_0). \tag{A4}$$

Therefore, we have

$$\liminf_n \min_{\boldsymbol{\tau}\in B} \mathcal{M}_n(\boldsymbol{\tau}) > \mathcal{M}(\boldsymbol{\tau}_0) \quad \text{almost surely,}$$

which means

$$\mathbb{P}\big(\liminf_n \min_{\boldsymbol{\tau}\in B} \mathcal{M}_n(\boldsymbol{\tau}) > \mathcal{M}(\boldsymbol{\tau}_0)\big) = 1. \tag{A5}$$

If $\widehat{\boldsymbol{\tau}}_n \in B$, then $\mathcal{M}_n(\widehat{\boldsymbol{\tau}}_n) = \min_{\boldsymbol{\tau}\in B} \mathcal{M}_n(\boldsymbol{\tau})$, which by definition of $\widehat{\boldsymbol{\tau}}_n$ is no larger than $\mathcal{M}_n(\boldsymbol{\tau}_0)$. Thus, for any $n >= 1$, we have

$$\{\widehat{\boldsymbol{\tau}}_n \in B\} \subset \{\min_{\boldsymbol{\tau}\in B} \mathcal{M}_n(\boldsymbol{\tau}) \le \mathcal{M}_n(\boldsymbol{\tau}_0)\}.$$

On the other hand, we have the following inequality chain for the RHS term,

$$\begin{aligned}
&\limsup_n \mathbb{P}\big\{\min_{\boldsymbol{\tau}\in B} \mathcal{M}_n(\boldsymbol{\tau}) \le \mathcal{M}_n(\boldsymbol{\tau}_0)\big\} \\
&\le \mathbb{P}\big(\limsup_n \{\min_{\boldsymbol{\tau}\in B} \mathcal{M}_n(\boldsymbol{\tau}) \le \mathcal{M}_n(\boldsymbol{\tau}_0)\}\big) \\
&\le \mathbb{P}\big(\liminf_n \min_{\boldsymbol{\tau}\in B} \mathcal{M}_n(\boldsymbol{\tau}) \le \liminf_n \mathcal{M}_n(\boldsymbol{\tau}_0)\big) \\
&= \mathbb{P}\big(\liminf_n \min_{\boldsymbol{\tau}\in B} \mathcal{M}_n(\boldsymbol{\tau}) \le \mathcal{M}(\boldsymbol{\tau}_0)\big) \quad \text{(law of large numbers)} \\
&= 1 - \mathbb{P}\big(\liminf_n \min_{\boldsymbol{\tau}\in B} \mathcal{M}_n(\boldsymbol{\tau}) > \mathcal{M}(\boldsymbol{\tau}_0)\big) \\
&= 0.
\end{aligned}$$

Therefore, the LHS term $\mathbb{P}(d(\widehat{\boldsymbol{\tau}}_n, \boldsymbol{\tau}_0) \ge \varepsilon) = \mathbb{P}(\widehat{\boldsymbol{\tau}}_n \in B) \to 0$. Since $\mathbb{P}(d(\widehat{\boldsymbol{\tau}}_n, \widetilde{\boldsymbol{\tau}}_0) \ge \varepsilon) \le \mathbb{P}(d(\widehat{\boldsymbol{\tau}}_n, \boldsymbol{\tau}_0) \ge \varepsilon)$, we also have $\mathbb{P}(d(\widehat{\boldsymbol{\tau}}_n, \widetilde{\boldsymbol{\tau}}_0) \ge \varepsilon) \to 0$, which concludes the consistency proof.

□

B.3 Proof of Theorem 3

Proof. To prove this theorem, we need to use the alternate definition of the Wasserstein distance. For any two distribution densities $f, g \in \mathcal{P}_2(\mathbb{R})$, the 2-Wasserstein distance $W_2(f, g)$ between them can also be defined as

$$W_2^2(f, g) = \inf_{\gamma\in\Pi(f,g)} \int_{\mathbb{R}\times\mathbb{R}} (x_1 - x_2)^2 \gamma(x_1, x_2) dx_1 dx_2,$$

where $\Pi(f, g)$ is the set of joint distributions $\gamma \in \mathcal{P}_2(\mathbb{R} \times \mathbb{R})$ such that for any $(x_1, x_2) \in \mathbb{R} \times \mathbb{R}$ the marginal distributions satisfy $\int_{\mathbb{R}} \gamma(x_1, s)ds = f(x_1)$ and $\int_{\mathbb{R}} \gamma(s, x_2)ds = g(x_2)$. More important, it can be proved the above definition is equivalent with our previous definition of the Wasserstein distance in Section 2, and the γ^* achieving the infimum is called the optimal coupling. In our case, since $\mathbb{R}$ is a Polish space, the optimal coupling exists. More details can be found in Villani (2008).

By the existence of optimal coupling, there are $\{\gamma_k^* \in \Pi(f_k, g_k)\}_{k=1}^K$, such that for $k = 1, ..., K$, the lower bound of the Wasserstein distance is achieved

$$W_2^2(f_k, g_k) = \int_{\mathbb{R}\times\mathbb{R}} (x_1 - x_2)^2 \gamma_k^*(x_1, x_2) dx_1 dx_2.$$

Define $\gamma^* = \sum_{k=1}^K \pi_k \gamma_k^*$, then $\gamma^* \in \Pi(f, g)$ since for any $(x_1, x_2) \in \mathbb{R} \times \mathbb{R}$ the marginal distributions satisfy $\int_{\mathbb{R}} \gamma^*(x_1, s)ds = f(x_1)$ and $\int_{\mathbb{R}} \gamma^*(s, x_2)ds = g(x_2)$. By definition,

$$\begin{aligned}
W_2^2(f, g) &\le \int_{\mathbb{R}\times\mathbb{R}} (x_1 - x_2)^2 \gamma^*(x_1, x_2) dx_1 dx_2 \\
&= \sum_{k=1}^K \pi_k \int_{\mathbb{R}\times\mathbb{R}} (x_1 - x_2)^2 \gamma_k^*(x_1, x_2) dx_1 dx_2 \\
&= \sum_{k=1}^K \pi_k W_2^2(g_k, f_k).
\end{aligned}$$

Thus, it remains to prove the equality condition. Define function $\mathbf{t}_f^g(x) = G^{-1} \circ F(x)$, where G^{-1} is the QF of g, F is the CDF of f. The optimal coupling γ^* can be expressed as the joint distribution of $\big(X, \mathbf{t}_f^g(X)\big)$ with random variable $X \sim f$ (Villani, 2008). Further, we let g_k be the distribution of $\mathbf{t}_f^g(X_k)$ with random variable $X_k \sim f_k$, then g_k are in the form of Equation (6) in Section 3. Moreover, the joint distribution γ_k^* of $(X_k, \mathbf{t}_f^g(X_k))$ is the optimal coupling such that the lower bound of the Wasserstein distance is achieved

$$W_2^2(f_k, g_k) = \int_{\mathbb{R}\times\mathbb{R}} (x_1 - x_2)^2 \gamma_k^*(x_1, x_2) dx_1 dx_2.$$

Finally, the following equations hold

$$\begin{aligned} W_2^2(f, g) &= \int_{\mathbb{R}\times\mathbb{R}} (x_1 - x_2)^2 \gamma^*(x_1, x_2) dx_1 dx_2 \\ &= \int_{\mathbb{R}} (x - \mathbf{t}_f^g(x))^2 f(x) dx \\ &= \sum_{k=1}^{K} \pi_k \int_{\mathbb{R}} (x - \mathbf{t}_f^g(x))^2 f_k(x) dx \\ &= \sum_{k=1}^{K} \pi_k W_2^2(g_k, f_k), \end{aligned}$$

and $g = \sum_{k=1}^{K} \pi_k \cdot g_k$ is a valid mixture decomposition. □

C OPTIMIZATION FRAMEWORK

In this section, we provide more details for the optimization framework.

First, we show the target loss function is guaranteed to decrease in our optimization framework. In fact, by plugging in $g = f_{\mathcal{D}}, f = f_{\boldsymbol{\theta}}$, in Theorem 3 and assuming $f_{\boldsymbol{\theta}}$ belongs to the family of Gaussian mixture distributions with $\boldsymbol{\theta} = \{(\pi_k, \mu_k, \sigma_k)\}_{k=1}^K$, then the left side of Equation (5) in Theorem 3 becomes the target loss function $L(\boldsymbol{\theta}) = W_2^2(f_{\mathcal{D}}, f_{\boldsymbol{\theta}})$. Moreover, if we treat the mixture decomposition component $\{g_k\}_{k=1}^K$ of the empirical distribution as the latent parameters, the right side of Equation (5) provides a natural upper bound of $L(\boldsymbol{\theta})$. This can serve as a surrogate function that *majorizes* the original objective function. Denoting the right side of Equation (5) by $R(\boldsymbol{\nu}, \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k W_2^2(g_k, f_k)$ with $\boldsymbol{\nu} = \{g_k\}_{k=1}^K$, $\boldsymbol{\theta} = \{(\pi_k, \mu_k, \sigma_k)\}_{k=1}^K$ and $f_k = \mathrm{N}(\mu_k, \sigma_k^2)$, we have $L(\boldsymbol{\theta}) \le R(\boldsymbol{\nu}, \boldsymbol{\theta})$ by Theorem 3. Therefore, we have the following inequality chain

$$\begin{aligned} L(\boldsymbol{\theta}^{(m)}) = L(\pi^{(m)}, \mu^{(m)}, \sigma^{(m)}) &\le L(\pi^{(m-1)}, \mu^{(m)}, \sigma^{(m)}) \le R(\{g_k\}^{(m-1)}, \pi^{(m-1)}, \mu^{(m)}, \sigma^{(m)}) \\ &\le R(\{g_k\}^{(m-1)}, \pi^{(m-1)}, \mu^{(m-1)}, \sigma^{(m-1)}) = L(\pi^{(m-1)}, \mu^{(m-1)}, \sigma^{(m-1)}) = L(\boldsymbol{\theta}^{(m-1)}). \end{aligned}$$

It indicates that the original loss function $L(\boldsymbol{\theta})$ decreases during the optimization of the upper bound $R(\boldsymbol{\nu}, \boldsymbol{\theta}) = R\big(\{g_k\}_{k=1}^K, \pi, \mu, \sigma\big)$ in each iteration step. The outline algorithm hence converges to the minimum of the original loss function.

Then, we provide more details for our boosted MM optimization framework in Figure A1.

Moreover, using the example of fitting a single three-component Gaussian mixture model $f_{\boldsymbol{\theta}}$ to empirical distribution g derived from data sample $\mathcal{D}$ under the Wasserstein distance loss, we show our proposed WDL framework can achieve fast convergence and jump out of local optima compared with the vanilla EM algorithm. In the experiment, we started from the same random initialization for EM and WDL, and then calculate the Wasserstein distance loss along the optimization process. In Video S1, we visualize the input data $\mathcal{D}$ or g as the histogram, represent the fitted components using three curves of different colors, and also utilize colorful bars for the decomposition g_k of the empirical distribution g. As we can see in the video, the vanilla EM algorithm easily got stuck in the local optima. For WDL, the process identified one component first, and then pushed the other two components away from the first one. Every time of such a push created a sharper drop in the loss. In Figure A2, we provide a screenshot of Video S1 after 100 optimization steps.

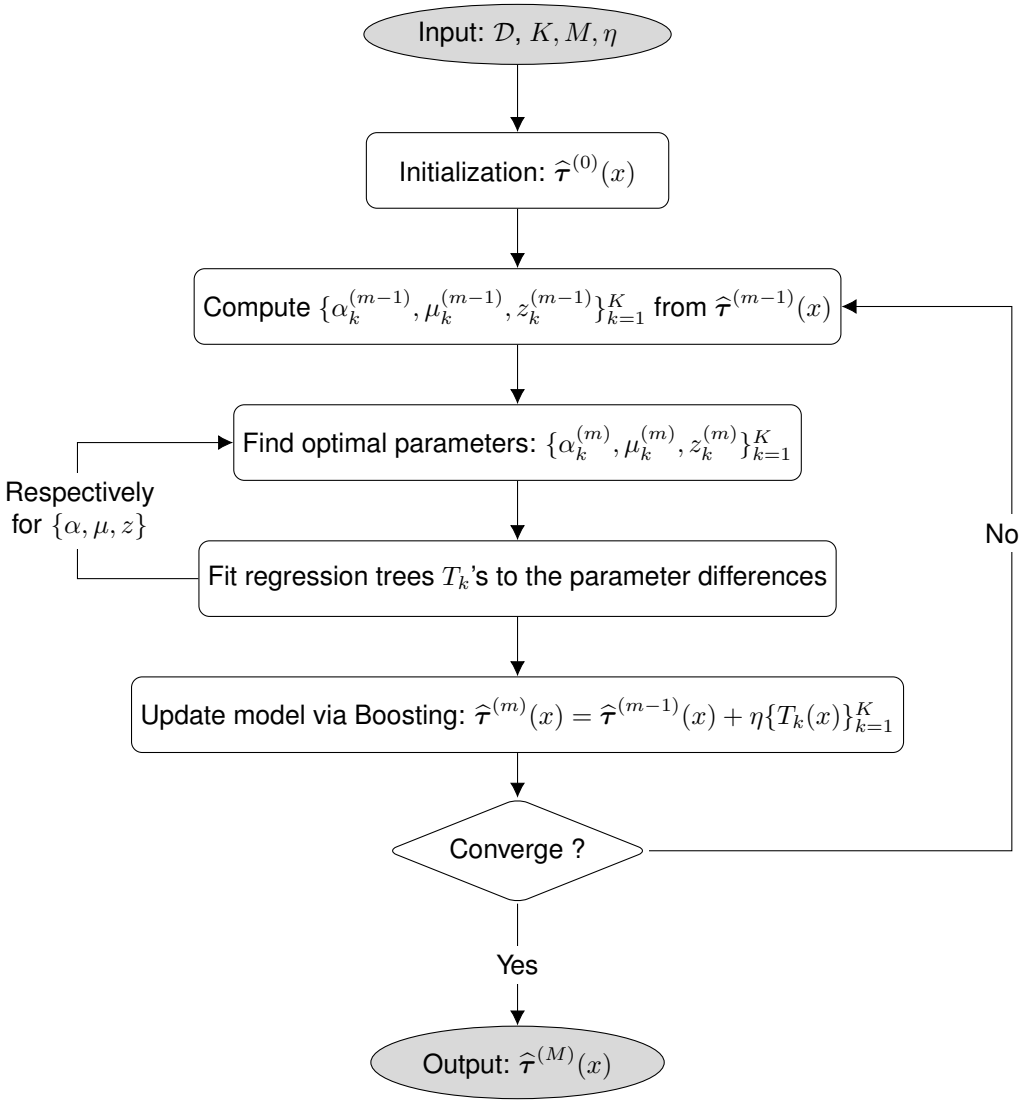

Figure A1: Diagram for fitting SCGMM

D SIMULATION DETAILS

In this part, we introduce the simulation details for reproducibility. All the simulations were implemented with `Python` version 3.6 and `R` version 4.0.3.

D.1 Simulation Setup

The simulation mechanism is illustrated in Equation (7) of Section 4. In the experiment, with predefined parameters (N, n, ω), density-on-scalar data $\mathcal{D} = \{(x_i, \widehat{g}_i)\}_{i=1}^N$ were simulated as follows

D.2 Fréchet Regression

The implementation of global Fréchet regression was following the algorithm introduced in the reference paper (Petersen et al., 2019, 2021). All the simulations were coded in `R` using the package `frechet` [1] developed by the author. In model training, we first calculated the empirical quantile function $\widehat{G}_i^{-1}$ from the random points $(y_{i,1}, ..., y_{i,n})$, and then fed them to the function `GloDenReg`. There is no tuning parameter in global Fréchet regression.

[1] `https://cran.r-project.org/web/packages/frechet/index.html`

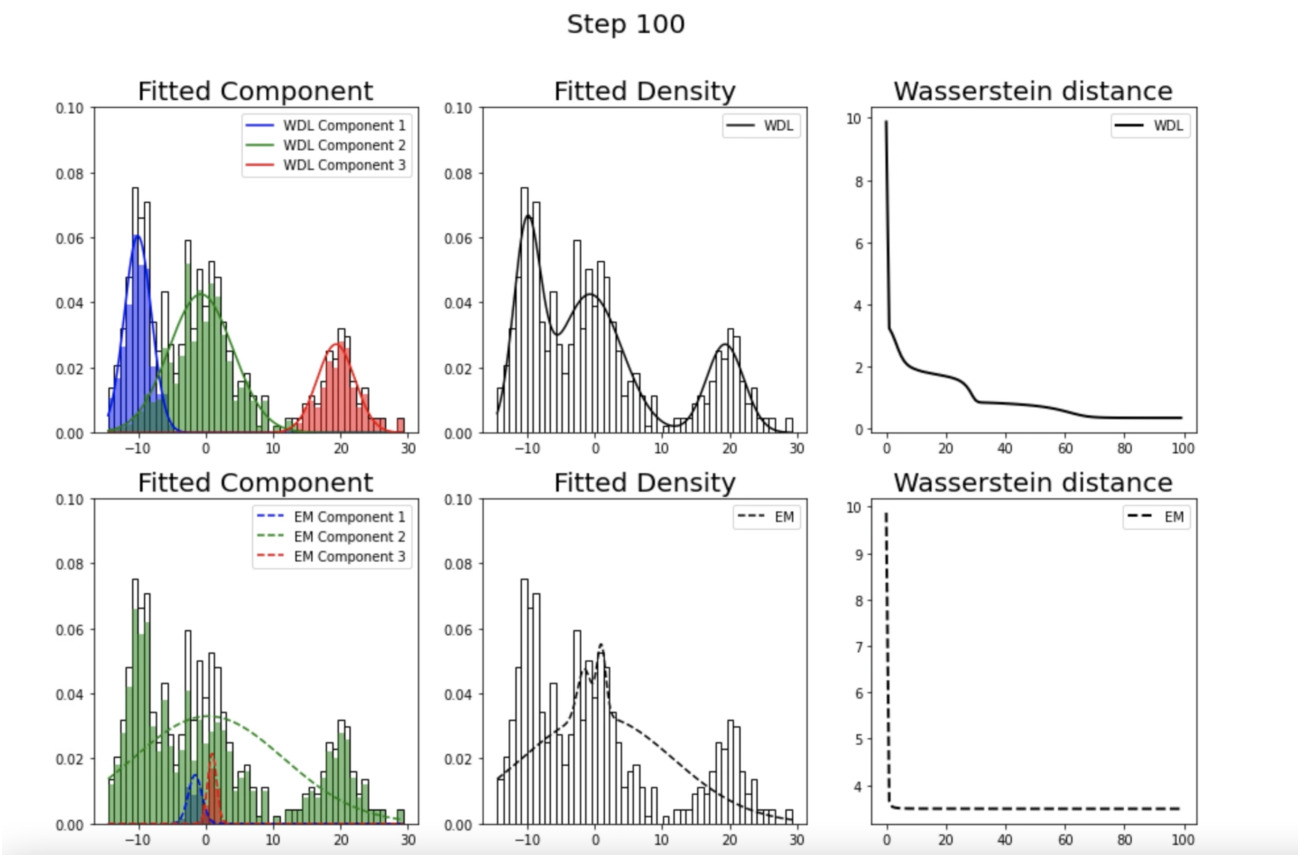

Figure A2: A screenshot of Video S1 after 100 optimization steps. First row: WDL framework; Second row: vanilla EM algorithm.

D.3 CLR Regression

The implementation of the B-spline smoothed density regression with centered log-ratio transformation was based on the sample codes [2] from the reference paper (Talská et al., 2018). Specifically, for a PDF $f \in \mathcal{B}^2(I)$, the centered log-ratio transformation is defined as

$$CLR[f](t) = \log f(t) - \frac{1}{\gamma}\int_I \log f(s)ds, \quad \forall t \in I,$$

where γ is the normalization constant such that $\int_I CLR[f](t)dt = 0$. Actually, the centered log-ratio transformation defines an isometric isomorphism between the Bayes space $\mathcal{B}^2(I)$ and the Hilbert space $L^2(I)$.

The hyper-parameters of CLR regression were chosen following the reference paper (Talská et al., 2018). With randomly sampled data points $(y_{i,1}, ..., y_{i,n})$, we first built the histogram of each functional output, of which the optimal number of classes were decided by the Sturges' rule (Sturges, 1926). Because of the output heterogeneity, possible count zeros were replaced by positive posterior expectations with Perks prior using methods from Martín-Fernández et al. (2015). Afterwards, centered log-ratio transformation was applied to map the density estimations (histograms) into the Bayes space $\mathcal{B}^2(\mathbb{R})$, and B-spline polynomials with equally spaced knots were utilized to smooth the log density curve. To calculate the Wasserstein loss, we transformed the estimated density function $\widehat{f}_i$ back into a quantile function $\widehat{F}_i^{-1}$ at the given 99 equally spaced quantile levels $\{\rho_l = \frac{l}{100}\}_{l=1}^{99}$ using linear interpolation, and then numerically calculated the Wasserstein loss.

In the sample codes from the reference paper, quadratic B-splines with five equally spaced knots were used. In our implementation, we fine-tuned these hyper-parameters (degree $\in \{2, 3, 4\}$, number of knots $\in \{5, 8, 10\}$) using cross-validation.

[2] https://www.sciencedirect.com/science/article/abs/pii/S0167947318300276

Algorithm A1 Data Simulation

Input: N-number of samples , n-number of data points in each density, ω-noise level.
for $i = 1$ **to** N **do**
[Inputs] Randomly sample covariate vectors $x_i \sim U[-1, 1]$ and random noises $\varepsilon_i \sim N(0, \omega^2)$.
[Outputs] Randomly sample i.i.d. $(y_{i,1}, ..., y_{i,n})$ from the conditional density $p(Y|X = x_i, \varepsilon_i)$, and construct empirical $\widehat{g}_i$.
end for
Output: Density-on-scalar data $\mathcal{D} = \{(x_i, \widehat{g}_i)\}_{i=1}^N$.

Table A2: Performance comparison in terms of Wasserstein loss and R-squared (bracket) when the quantiles are sparse.

$\omega =$	0.1	0.3	0.5
WDL	**0.072 (0.84)**	**0.132 (0.64)**	**0.216 (0.36)**
Fréchet	0.226 (0.51)	0.247 (0.35)	0.292 (0.16)

D.4 Regression with Sparse Quantiles

In real-world applications, a common scenario is that the conditional quantiles $G_i^{-1}(\rho)$ of the functional outputs $\{g_i\}_{i=1}^N$ are only available at a series of sparse quantile values, for instance, $\rho \in \{0, 0.1, ..., 0.9, 1\}$ in the UK biobank data [3]. To apply the Wasserstein distributional learning framework to these scenarios, additional treatments are essential due to the definition of the Wasserstein loss. To be more specific, as introduced in Section 2, the Wasserstein distance between the two density functions is the integral of their quantile differences. When the dense quantiles are available, e.g. $\rho \in \{0.01, 0.02, ..., 0.99\}$, the Wasserstein distance can be numerically approximated by the average of all quantile differences, as show below.

$$W_2^2(g_1, g_2) = \int_0^1 \left(G_1^{-1}(s) - G_2^{-1}(s)\right)^2 ds \approx \frac{1}{100} \sum_{i=1}^{99} \left(G_1^{-1}(\frac{i}{100}) - G_2^{-1}(\frac{i}{100})\right)^2.$$

While, such approximation is far from accurate when the quantiles are sparse, e.g. $\rho \in \{0.1, 0.2, ..., 0.9\}$. Similarly, to measure the discrepancy between functional outputs, we can still define the quasi-Wasserstein loss as

$$\widetilde{W}_2^2(g_1, g_2) = \frac{1}{10} \sum_{i=1}^{9} \left(G_1^{-1}(\frac{i}{10}) - G_2^{-1}(\frac{i}{10})\right)^2.$$

However, to the best of our knowledge, there is no efficient algorithm for solving this optimization problem when the model family is Semi-parametric Conditional Gaussian Mixture Models (SCGMM) and the quantiles levels are sparse.

Practically, a solution to address this issue is to apply a linear interpolation. Over the training set, we can augment the sparse quantiles into a dense array using linear interpolation. It should be noted that the augmented quantiles naturally satisfy the non-crossing constraints since linear interpolation keeps the monotonicity of the quantile function. Then, the WDL framework can be fitted over the training set using the augmented dense quantiles. Optimization convergence is theoretically guaranteed in Section 3. Last, but not least, we can take the predicted sparse quantiles as the functional output, and the test error can be evaluated over the test set at the given quantile levels.

Using the same simulation setup as in Section 4, we rerun the experiments with sparse quantiles $\rho \in \{0, 0.1, ..., 0.9, 1\}$, and compare the performance of WDL with Fréchet regression. CLR regression is ignored here because under the sparse quantiles, the estimated density would be highly unstable. As shown in Table S1, the WDL performance is similar with that in the case of dense quantiles (Table 1), and is significantly better than Fréchet regression.

D.5 Simulations in The Linear Case

In this part, we present simulation results in a simpler setup, and prove our proposed method is flexible under different settings. The simulation setup of this section follows the experiment in Petersen et al. (2019). In this case, the quantile values of the functional output are linear functions of the input variables, which is the underlying assumption of global

[3] https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=23000

Fréchet regression. Results show our proposed Wasserstein distributional learning framework is able to achieve comparable performance even when the data are simulated in a different way from the model assumption.

To simulate the functional regression data, the responses Y are distributions represented by quantile functions $Q(Y)$ and the predictors are random vectors $X \in \mathbb{R}^3$. For any given quantile level $0 < \rho < 1$, the regression function is

$$Q_Y^{-1}(\rho) = (\mu_0 + \beta^\top x) + (\sigma_0 + \gamma^\top x) \cdot \Phi^{-1}(\rho),$$

where Φ is the standard normal distribution function. $\mu_0, \sigma_0 \in \mathbb{R}$, $\beta, \gamma \in \mathbb{R}^3$, satisfy $\sigma_0 + \gamma^\top x > 0$ for all x. In fact, this simulation scenario corresponds to cases in which the response functions are normal distributions with linear parameters on average.

In the simulation, the functional response Y is generated conditional on X by adding noise to the quantile functions. For each input $X = x$, the distribution parameters (μ, σ) are independently sampled from $p(\mu|X = x) = \mathrm{N}(\mu_0 + \beta^\top x, v_1)$ and $p(\sigma|X = x) = Gam((\sigma_0 + \gamma^\top x)^2/v_2, v_2/(\sigma_0 + \gamma^\top x))$, and the corresponding functional output is $Y = \mu + \sigma\Phi^{-1}$. Specifically, we set the parameters as $\mu_0 = 0, \sigma_0 = 3, v_1 = 0.25, v_2 = 1, \beta = (1, -1, 3)^\top, \gamma = (0.1, 0.2, 0.3)$, and simulated $N = 200$ random distributions, which are each represented by $n = 300$ random data points.

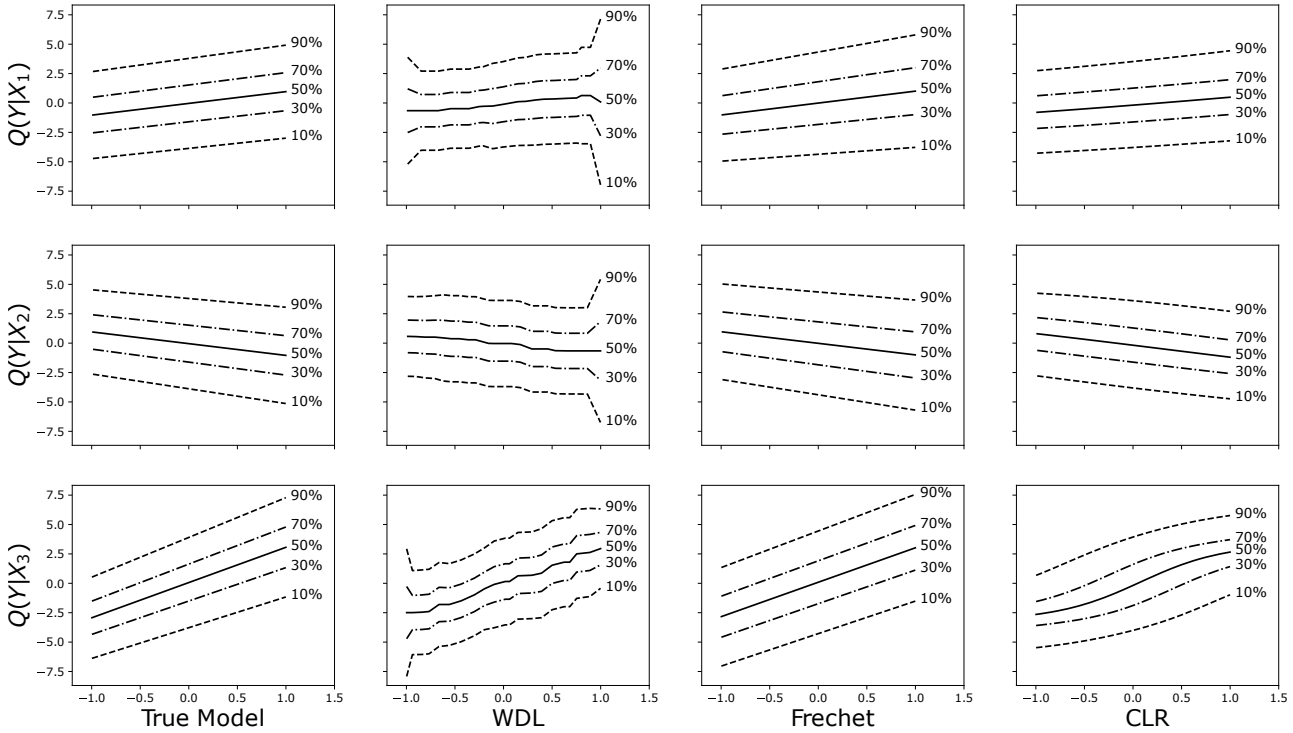

Figure A3: Functional partial dependence plot for the three methods.

The functional partial dependence plots on test sets are observed in Figure A3. Results show that the proposed Wasserstein distributional learning has stable performance under different simulation settings.

E REAL-WORLD APPLICATIONS

In this appendix section, we provide more details for the real-world applications.

E.1 Climate Modeling

E.1.1 Data Collection

In this appendix section, we apply the proposed Wasserstein distributional learning to understand how the radiative effect of solar irradiance, volcanic eruptions, and CO2, as well as natural climate variability through the El-Niño Southern

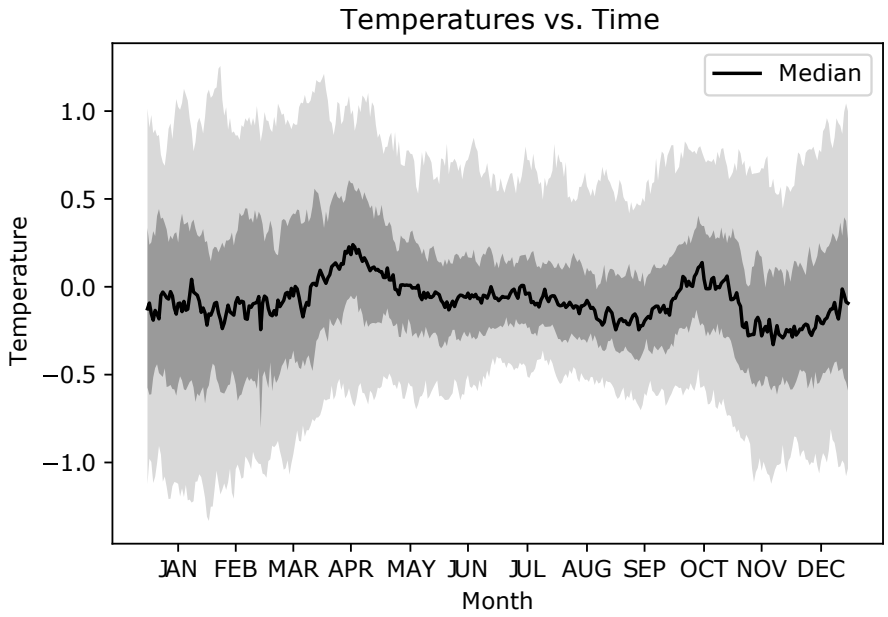

Figure A4: Daily trend of global average temperatures. The solid curve represents the median temperature of each day. The dark gray band represents the 30% and 70% quantiles. The light gray band represents the 10% and 90% quantiles.

Oscillation (ENSO) are associated with annual temperature distributions. These factors have been suggested in climate change literature (Fahey et al., 2017; Lewis and King, 2017), and represent natural and human drivers for climate variability and change. We obtain the daily land-surface average temperatures from Berkeley Earth daily TAVG full dataset (Berkeley Earth, 2021), where temperatures are reported as daily anomalies relative to the Jan 1951 $\sim$ Dec 1980 average. We calculate the empirical quantile functions of daily average temperature anomalies for each year between 1880 and 2011 as functional outputs. The global radiative effects, or "radiative forcings" used in this example have units of Wm^{-2} and represent the global average energy balance that arises due to changes in atmospheric composition. Here, radiative forcings of solar irradiance, volcanic eruptions, and CO2 as as calculated by the NASA Goddard Institute for Space Studies (GISS) analysis checking the the historical (1850-2012) simulation of their dynamical climate model GISS Model E2 Miller et al. (2014). In addition to the three radiative forcing predictors, year-to-year climate variability is summarized through the Niño3.4 index, a sea surface temperature index that captures the oscillatory of the ENSO system between warm El Niño events and cool La Niña events McPhaden et al. (2020). Together, these four predictors have been shown to be highly predictive of global annual mean temperature Suckling et al. (2017) and are therefore expected to be predictive of the distribution of daily global mean temperature.

E.1.2 Additional Evaluations

In Figure A4, we visualize the daily trend of global average temperatures. Using data from 1880 to 2012, we calculate the temperature quantiles (10%, 30%, 50%, 70%, 90%) for each day. An interesting finding is that spring and autumn have a higher temperature in general than the other two seasons. Also, the temperature variability in summer is smaller than the other seasons. A potential explanation would be that the temperatures were calculated by averaging the records from multiple weather stations both from the north and south hemisphere. As a result, the averaged temperatures would display a more complicated trend since the north and south hemispheres always have different seasons.

We also visualize the average component weights of each day in Figure A5. Specifically, for each day from 1880 and 2012, we calculate the weight of each component using the predicted WDL model, and then average them across years. Finally, we plot them using a calendar heatmap with each grid representing a day in the year 2020 (we chose 2020 because it is a leap year with 366 days).

In Figure A6, we make the Individual Conditional Expectation (ICE) plots for each method. From the figure, WDL and Fréchet regression are the only two methods that can give unbiased estimations of conditional quantiles due to their choices of Wasserstein loss. Compared with Fréchet regression, WDL performs better when there exists nonlinearity in the

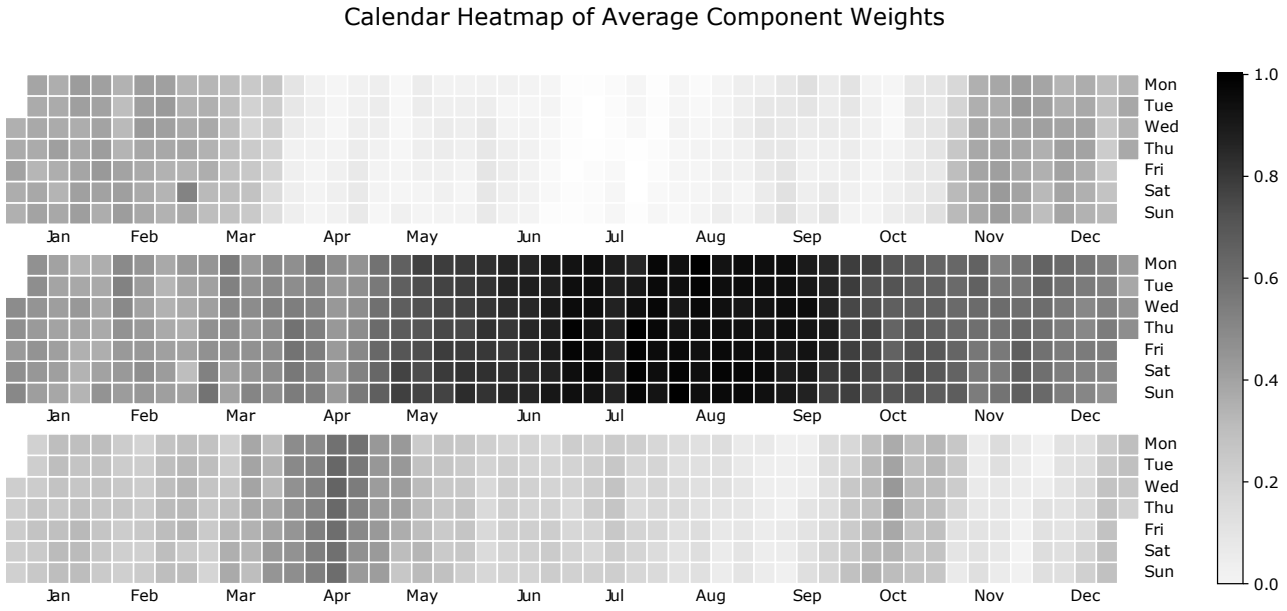

Figure A5: Average component weights of each day. First row: Component I; Second row: Component II; Third row: Component III. Component weights are represented using different colors.

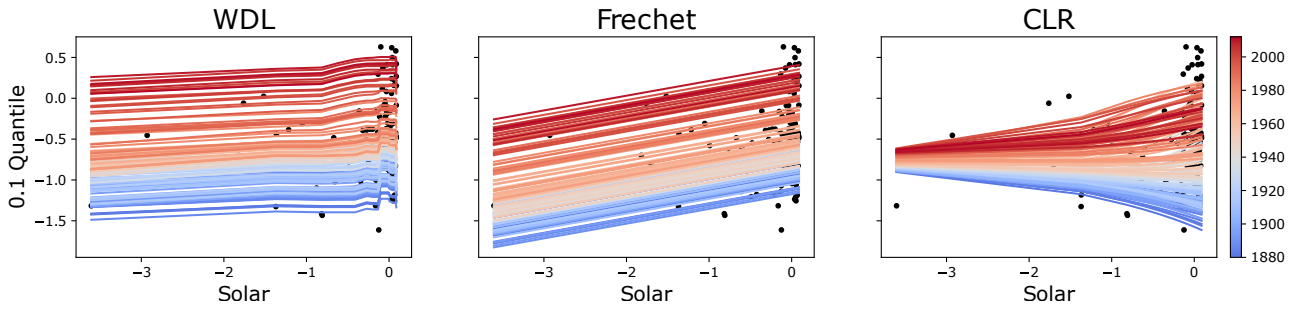

Figure A6: ICE plots of conditional temperature quantiles (10%) by solar irradiance. True conditional quantiles vs. solar irradiance from raw data are represented as black dots.

conditional dependence. These findings explain the phenomenon that WDL is the only method that can predict the "cold temperature plateau" between 1940 and 1960 in Figure 4. In Figure A7 to Figure A9, we visualize the predicted annual temperature distributions for each method from 1880 to 2012. Also in those figures, WDL captures the tail behavior more accurately than the others.

E.2 Income Modeling

E.2.1 Data Collection

In this experiment, we apply Wasserstein distributional learning to model the regional income distribution of the 167 counties in New York, California and Michigan, from which one could derive multiple indices simultaneously and explicitly study their joint distributions. The income distribution data are from American Community Survey (ACS), which we used with survey weights from 2014 ACS Public Use Microdata Sample (PUMS) (American Community Survey, 2014) to produce the county-level income distributions. We also collected scalar county-level health indices of the same year (2014) from County Health Rankings & Roadmaps 2014. Seven important variables were selected for our analysis: Education, Environment, Population, Crime, GDP Per Capita, Diabetes, and Unemployment rate. With all the data in place, the functional regression was conducted at the county level, which means each county served as an independent data point in

the regression.

E.2.2 Additional Evaluations

In this part, similar with weather distribution modeling, we illustrate the predicted density for each county when it was in the *test* fold. As shown in Figure A10, the income distributions vary across counties, and our method is able to capture the distribution accurately. For example, the income distribution in New York has a larger variance, and the income distribution in Orleans has a much higher peak in mode. The WDL framework is able to capture the distinctive features in these two distributions. Also, we make the functional partial dependence plot for predicted conditional quantiles versus the input scalar covariates in Figure A11. The abbreviation information are as follows, EDU: Education, ENV: Environment, PPL: Population, GDP: GDP Per Capita, CRM: Crime Rate, DBT: Diabetes, EMP: Unemployment. Since there is no ground truth in the real dataset, this figure only explains how each method interprets the functional dependence between the density output and scalar covariates in different ways.

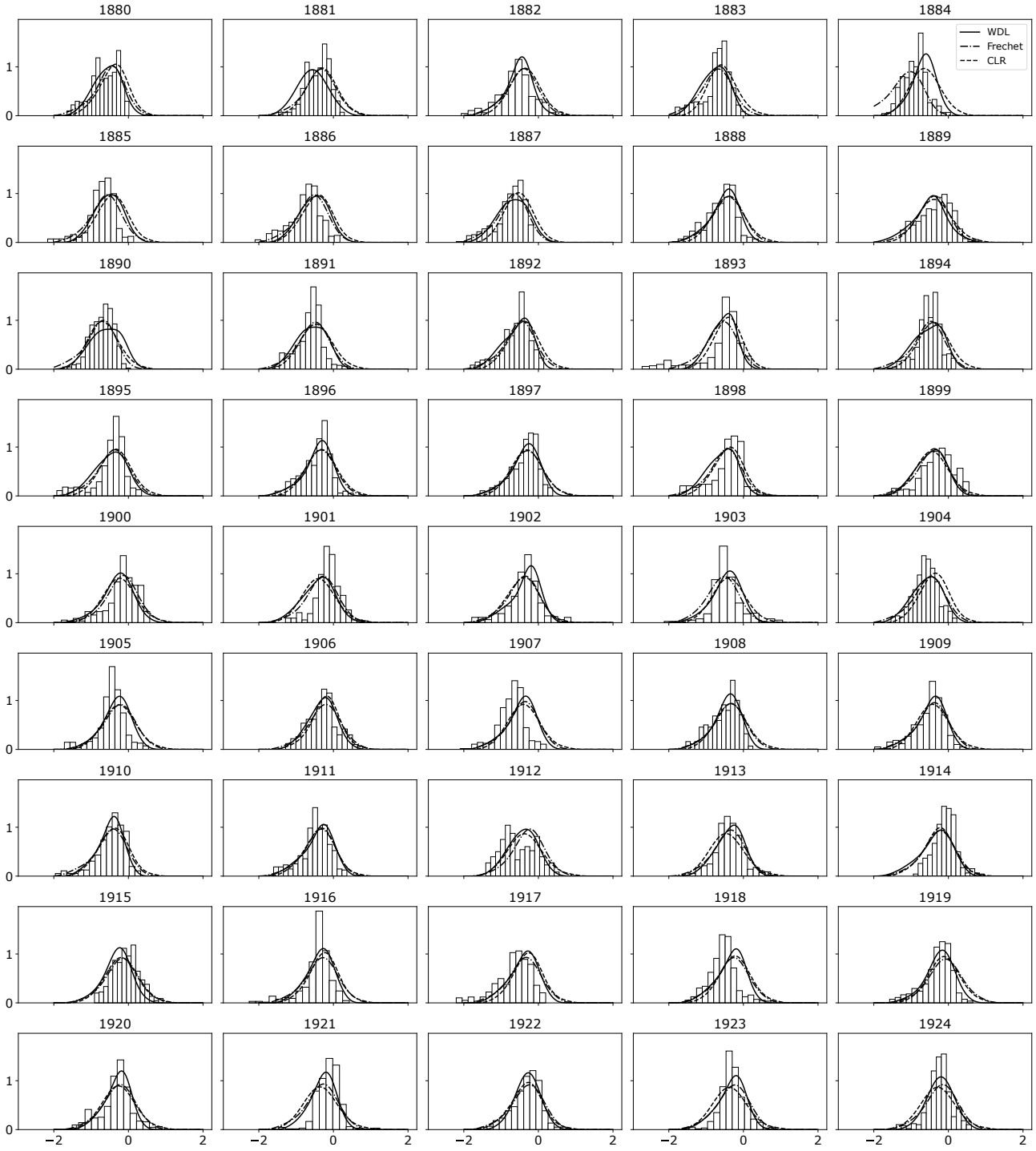

Figure A7: Predictions of annual temperature distributions (Part I).

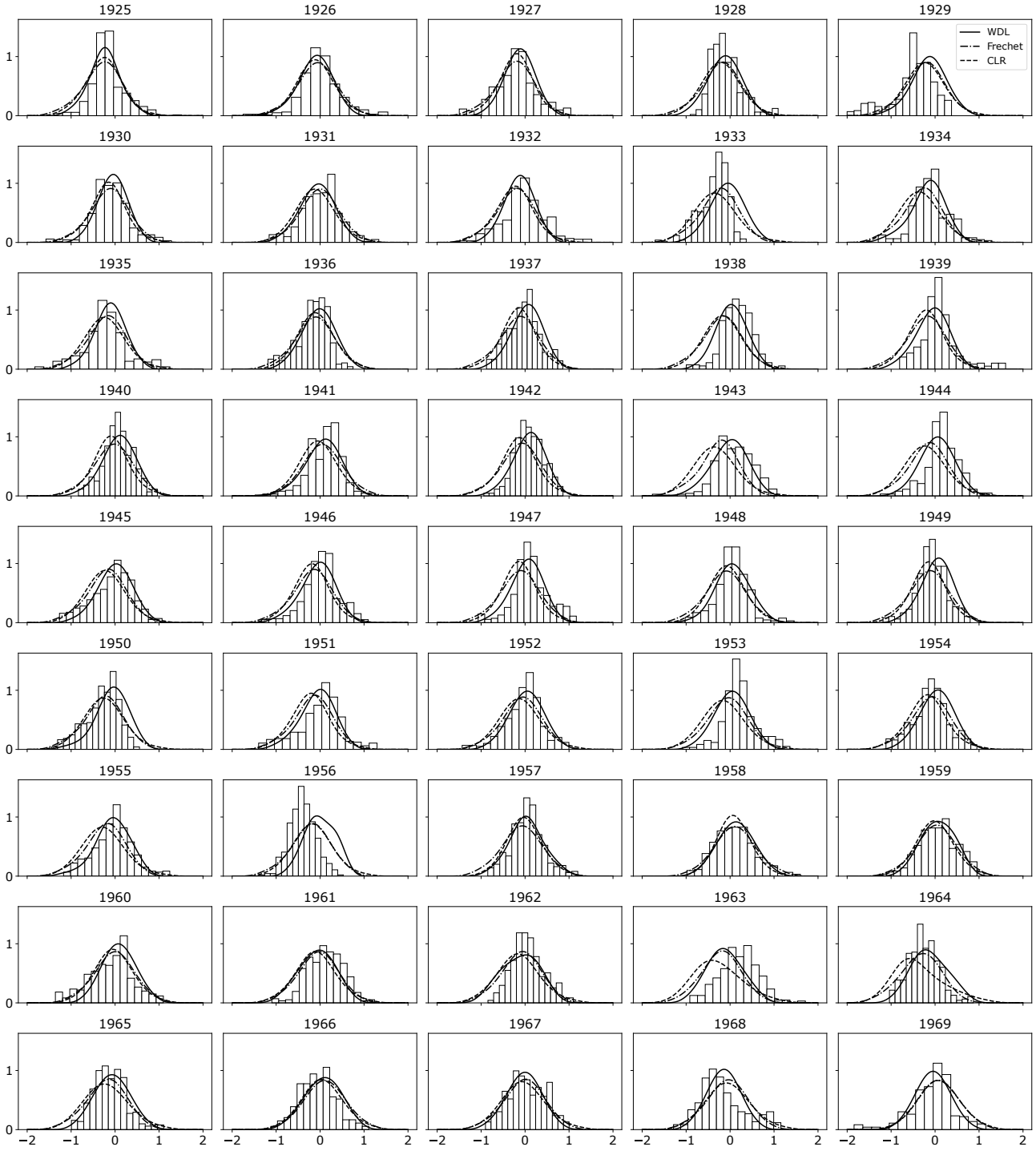

Figure A8: Predictions of annual temperature distributions (Part II).

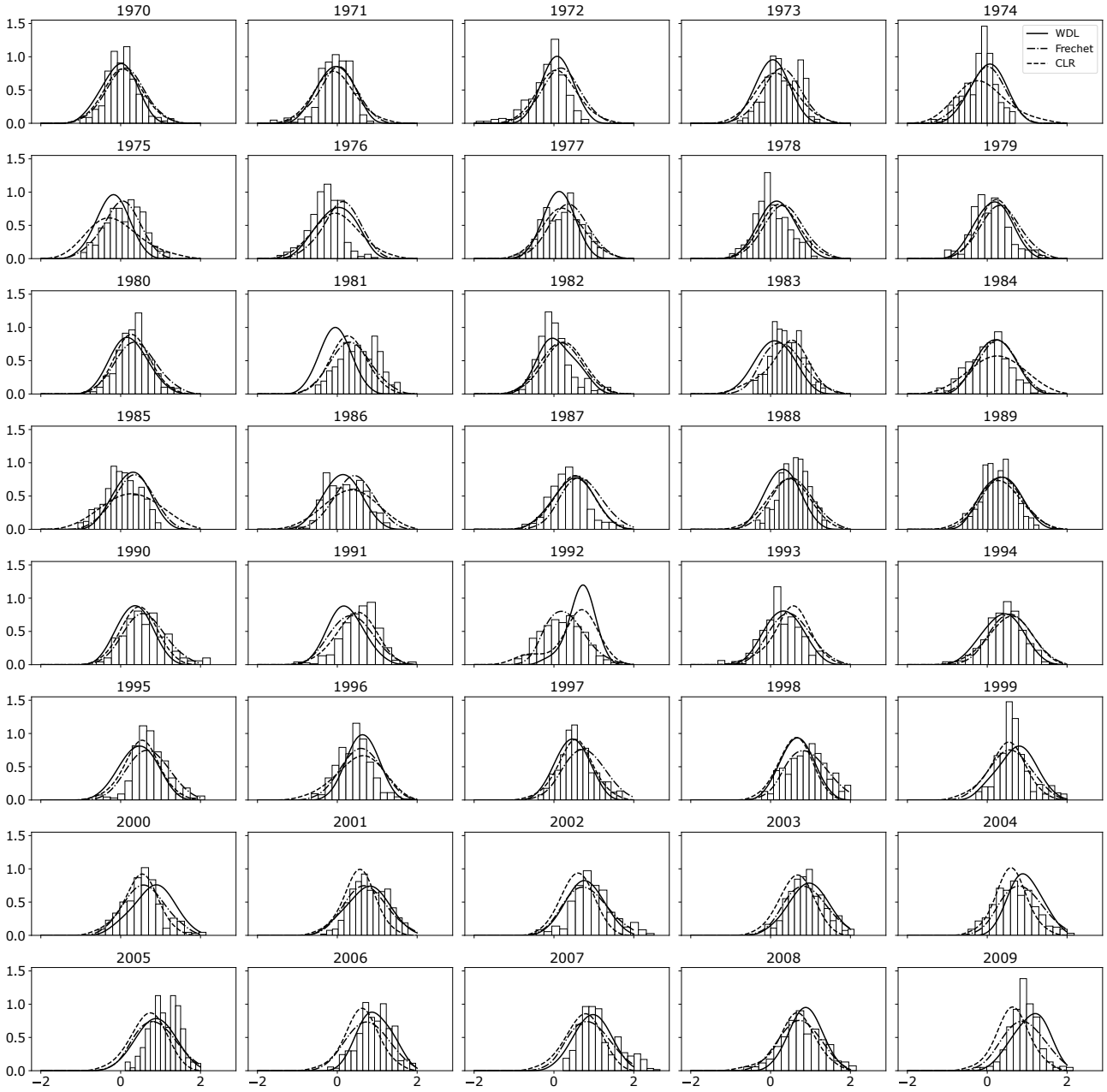

Figure A9: Predictions of annual temperature distributions (Part III).

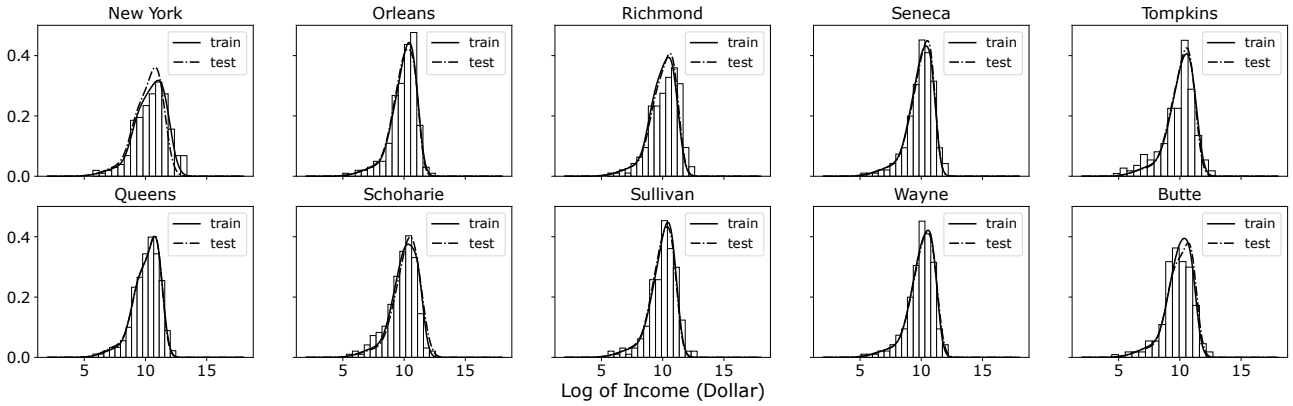

Figure A10: Selected predictions of regional income distributions.

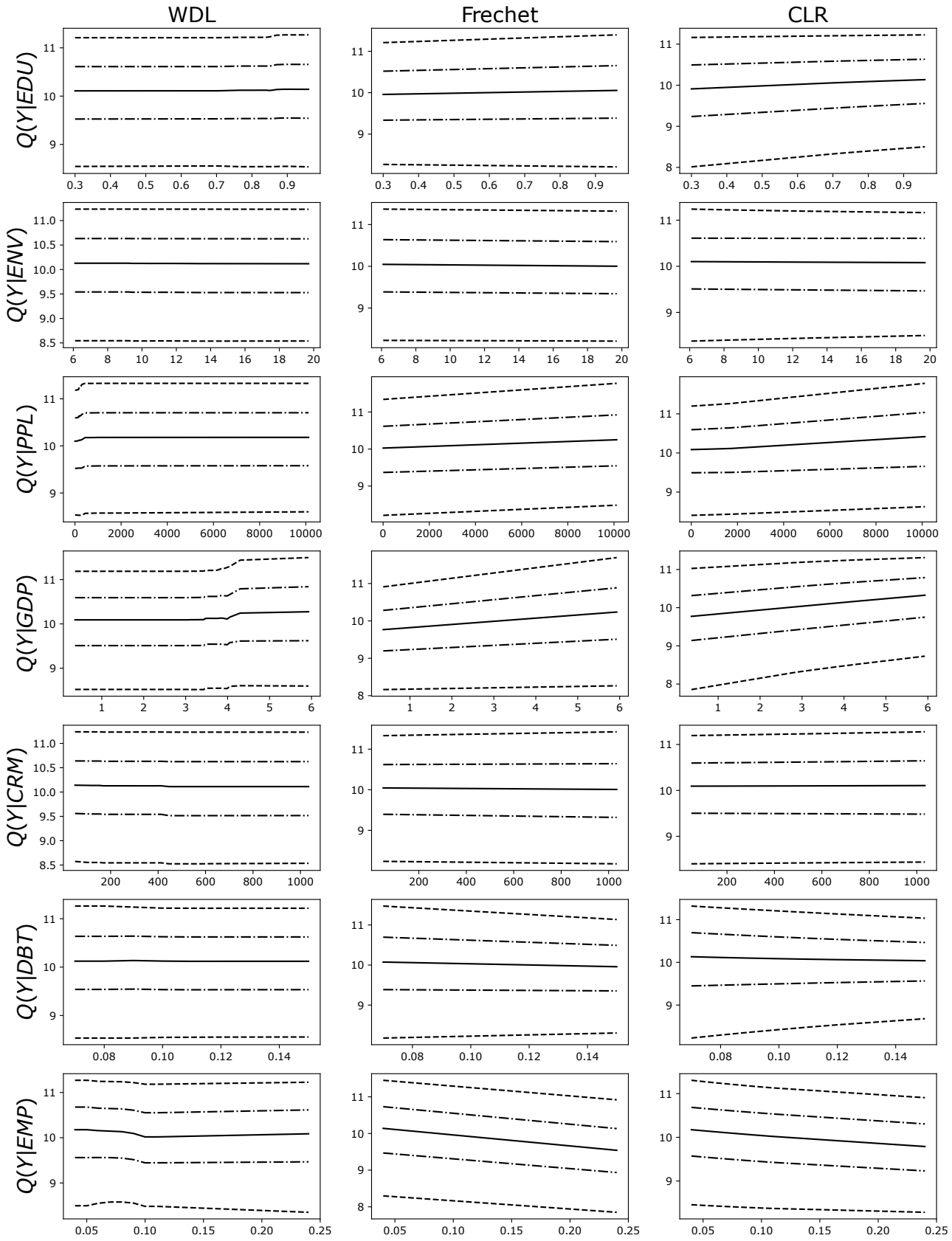

Figure A11: Functional partial dependence plot for predicted conditional quantiles versus the input scalar variables.