A Large-Scale Study of Probabilistic Calibration in Neural Network Regression

Victor Dheur [1] **Souhaib Ben Taieb** [1]

Abstract

Accurate probabilistic predictions are essential for optimal decision making. While neural network miscalibration has been studied primarily in classification, we investigate this in the less-explored domain of regression. We conduct the largest empirical study to date to assess the probabilistic calibration of neural networks. We also analyze the peformance of recalibration, conformal, and regularization methods to enhance probabilistic calibration. Additionally, we introduce novel differentiable recalibration and regularization methods, uncovering new insights into their effectiveness. Our findings reveal that regularization methods offer a favorable tradeoff between calibration and sharpness. Post-hoc methods exhibit superior probabilistic calibration, which we attribute to the finite-sample coverage guarantee of conformal prediction. Furthermore, we demonstrate that quantile recalibration can be considered as a specific case of conformal prediction. Our study is fully reproducible and implemented in a common code base for fair comparisons.

1. Introduction

Neural network predictions affect critical decisions in many applications, including medical diagnostics and autonomous driving (Gulshan et al., 2016; Guizilini et al., 2020). However, effective decision making often requires accurate probabilistic predictions (Gawlikowski et al., 2021; Abdar et al., 2021). For example, consider a probabilistic regression model that produces 90% prediction intervals. An important property would be that 90% of these prediction intervals contain the realizations.

For models that output a predictive distribution, *probabilistic calibration* is an important property that states that all quantiles must be calibrated, i.e., the frequency of realizations below these quantiles must match the corresponding quantile level. Additionally, predictive distributions should be sufficiently sharp (i.e., concentrated around the realizations) and leverage the information in the inputs.

In the classification setting, Guo et al. (2017) found that common neural architectures trained on image and text data were miscalibrated, sparking increased interest in neural network calibration. In a follow-up study, Minderer et al. (2021) showed that more recent neural architectures demonstrate improved calibration. However, there has been less research on calibration for neural probabilistic regression models compared to classification. Therefore, it remains uncertain whether the same results apply to the regression setting. This paper addresses this gap by conducting a comprehensive study on probabilistic calibration for regression using tabular data. We explore various calibration methods, including quantile recalibration (Kuleshov, Fenner, et al., 2018) and conformalized quantile regression (Romano, Patterson, et al., 2019). We also consider regularization methods, which have been shown to perform well in the classification setting (Karandikar et al., 2021; Popordanoska et al., 2022; Yoon et al., 2023).

We make the following main contributions:

1. We conduct the largest empirical study to date on probabilistic calibration of neural regression models using 57 tabular datasets (Sections 4 and 6). We consider multiple state-of-the-art calibration methods (Section 5), including post-hoc recalibration, conformal prediction, and regularization methods, with various scoring rules and predictive models.

2. Building on quantile recalibration, we propose a new differentiable calibration map using kernel density estimation, which provides improved negative log-likelihood compared to baselines. We also introduce two new regularization objectives based on the probabilistic calibration error (Section 5).

3. We show that quantile recalibration is a special case of conformal prediction, providing an explanation for

[1]Department of Computer Science, University of Mons, Mons, Belgium. Correspondence to: Victor Dheur <victor.dheur@umons.ac.be>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

their superior performance in terms of probabilistic calibration (Section 6).

2. Background

We consider a univariate regression problem where the target variable $Y \in \mathcal{Y}$ depends on an input variable $X \in \mathcal{X}$, with $\mathcal{Y} = \mathbb{R}$ representing the target space and $\mathcal{X}$ representing the input space. Our objective is to approximate the conditional distribution $P_{Y|X}$ using training data $\mathcal{D} = \{ (X_i, Y_i) \}_{i=1}^{N}$ where $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P \equiv P_X \times P_{Y|X}$.

A probabilistic predictor $F_\theta : \mathcal{X} \to \mathcal{F}$ is a function parametrized by $\theta \in \Theta$ that maps an input $x \in \mathcal{X}$ to a predictive cumulative distribution function (CDF) $F_\theta(\cdot \mid x)$ in the space $\mathcal{F}$ of distributions over $\mathbb{R}$. Additionally, given $x \in \mathcal{X}$, we denote the predictive quantile function (QF) by $Q_\theta(\cdot \mid x)$, and probability density function (PDF) by $f_\theta(\cdot \mid x)$. Similarly, the marginal CDF, QF, or PDF of a random variable R is denoted by F_R, Q_R, or f_R, respectively.

Probabilistic calibration. Given an input $x \in \mathcal{X}$, the model F_θ is ideal if it precisely matches the conditional distribution $P_{Y|X}$. However, learning the ideal model based on finite data is not possible without additional (strong) assumptions (Foygel Barber et al., 2021). To avoid additional assumptions, we can instead enforce certain desirable properties that are attainable in practice and that a good or ideal forecaster should exhibit. One such property is probabilistic calibration.

Let $Z = F_\theta(Y \mid X) \in [0, 1]$ denote the probability integral transform (PIT) of Y conditional on X. The model F_θ is *probabilistically calibrated* (also known as PIT-calibrated) if $\forall \alpha \in [0, 1]$,

$$F_Z(\alpha) \doteq \Pr(Z \leq \alpha) = \alpha. \tag{1}$$

Let $U \in [0, 1]$ be a uniform random variable independent of Z. The left and right hand sides of (1) can be interpreted as the CDF of Z and U, respectively, as a function of α. This shows that the uniformity of the PIT is equivalent to probabilistic calibration (Dawid, 1984).

Since the ideal forecaster is probabilistically calibrated, we can require this property from any competent forecaster. However, probabilistic calibration, though necessary, is not sufficient for making accurate probabilistic predictions. Additionally, as discussed by Gneiting and Resin (2021), probabilistic calibration primarily addresses unconditional aspects of predictive performance and is implied by more robust conditional notions of calibration, such as auto-calibration.

Probabilistic calibration error. The most common approach for evaluating probabilistic calibration is to consider distances of the form $\int_0^1 |F_Z(\alpha) - F_U(\alpha)|^p d\alpha$ where $p > 0$. The particular cases of $p = 1$ and $p = 2$ are known as the 1-Wasserstein distance and Cramér-von Mises distance, respectively. We denote the empirical CDF of the PIT as $\hat{F}_Z(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(Z_i \leq \alpha)$ where $Z_i = F_\theta(Y_i \mid X_i)$ are PIT realizations. A common approach to assess probabilistic calibration using Monte Carlo estimation is to evaluate it at equidistant values $\alpha_1 < \cdots < \alpha_M$ as follows:

$$\text{PCE}_p(F_\theta, \mathcal{D}) = \frac{1}{M} \sum_{j=1}^{M} \left| \alpha_j - \hat{F}_Z(\alpha_j) \right|^p. \tag{2}$$

This metric has been previously employed in literature such as Zhao et al. (2020) and Zhou et al. (2021) with $p = 1$, and Kuleshov, Fenner, et al. (2018) and Utpala and Rai (2020) with $p = 2$. It is important to note that, unlike the classical definition of the p-norm, we do not exponentiate $\frac{1}{p}$ in (2) to maintain consistency with prior literature. In the subsequent sections, we focus our analysis on PCE_1 and use the abbreviation PCE for brevity.

One limitation of scalar metrics like PCE is their inability to provide detailed information regarding calibration errors at individual quantile levels, $\alpha_1, \ldots, \alpha_M$. Instead, PIT reliability diagrams offer a visual assessment of probabilistic calibration across all quantile levels by plotting the empirical CDF of the PIT Z. These diagrams display the right side of (1) against its left side, with a perfectly calibrated model represented by a diagonal line (in the asymptotic case). Figure 2 provides examples of such reliability diagrams, which have been employed in studies by Pinson and Hagedorn (2012) and Kuleshov, Fenner, et al. (2018).

3. Related Work

Post-hoc calibration approaches involve adjusting the predictions of a trained model using a mapping learned from a separate calibration dataset. In the context of classification, temperature scaling (Guo et al., 2017) is a simple and effective method that adjusts predictive confidence while maintaining accuracy. For regression tasks, quantile recalibration (Kuleshov, Fenner, et al., 2018) aims to achieve probabilistic calibration. Conformal prediction (Vovk et al., 2020) is a general approach that provides prediction sets with a finite-sample coverage guarantee. Notable methods applied with deep learning include Conformal Quantile Regression (Romano, Patterson, et al., 2019) and Distributional Conformal Prediction (Izbicki et al., 2020; Chernozhukov et al., 2021). Furthermore, post-hoc approaches have also been proposed for conditional notions of calibration (Song et al., 2019; Kuleshov and Deshpande, 2022).

Regularization approaches aim to improve calibration during training by incorporating regularization techniques. Some methods, proposed by Zhao et al. (2020) and Feldman et al. (2021), utilize regularization to target different

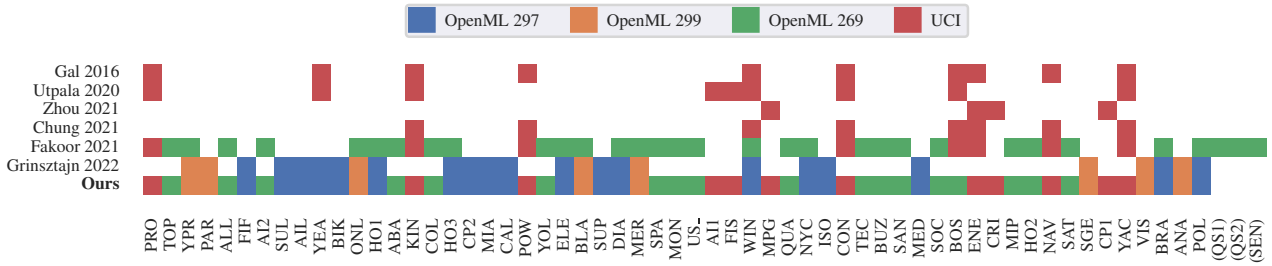


Figure 1: Multiple regression benchmark datasets with references. Datasets inside parentheses have not been considered in this study. Full dataset names are available in Table 1.

conditional notions of calibration based on the inputs. Zhou et al. (2021) introduced an alternative loss function involving the simultaneous training of two neural networks, while Pearce et al. (2018), Chung et al. (2021), and Thiagarajan et al. (2020) proposed objectives that allow control over the tradeoff between coverage and sharpness of prediction intervals. To our knowledge, the only regularization objective specifically targeting probabilistic calibration is quantile regularization (Utpala and Rai, 2020). Other types of uncertainty quantification methods include ensembling (Lakshminarayanan, Pritzel, et al., 2017) and Bayesian methods (Jospin et al., 2022)

4. Are Neural Regression Models Probabilistically Calibrated?

We conduct an extensive empirical study to evaluate the probabilistic calibration of neural regression models. To this end, we calculate the *probabilistic calibration error* defined in (2) for various state-of-the-art models across multiple benchmark datasets.

Benchmark datasets. We analyze a total of 57 datasets, including 27 from the OpenML curated benchmark (Grinsztajn et al., 2022), 18 from the AutoML Repository (Gijsbers et al., 2019), and 12 from the UCI Machine Learning Repository (Dua and Graff, 2017).

These datasets are widely used in the evaluation of deep probabilistic models and uncertainty quantification, as evidenced by previous studies such as Fakoor et al. (2021), Chung et al. (2021), Zhou et al. (2021), Utpala and Rai (2020), and Gal and Ghahramani (2016).

Figure 1 provides an overview of the utilization of these datasets in previous studies. To the best of our knowledge, our study represents the most comprehensive assessment of probabilistic calibration for neural regression models published to date.

Neural probabilistic regression models. We consider three state-of-the-art neural probabilistic regression models. The first model predicts a parametric distribution, where the parameters are obtained as outputs of a hypernetwork. Previous studies have often focused on the Gaussian distribution (Lakshminarayanan, Pritzel, et al., 2017; Utpala and Rai, 2020; Zhao et al., 2020). To introduce more flexibility, we consider a mixture of K Gaussian distributions. Given an input $x \in \mathcal{X}$, the hypernetwork parametrizes the means $\mu_k(x)$, standard deviations $\sigma_k(x)$, and weights $w_k(x)$ for each component $k = 1, ..., K$. To ensure positive standard deviations and that the mixture weights form a discrete probability distribution, we use the Softplus and Softmax activations, respectively. We have two variants of this model depending on the scoring rule used for training: the negative log-likelihood (NLL) or the continuous ranked probability score (CRPS). These models are denoted as MIX-NLL and MIX-CRPS, respectively. It is worth noting that the CRPS of a mixture of Gaussians has a closed-form expression (Grimit et al., 2006).

The second model predicts quantiles of the distribution (Tagasovska and Lopez-Paz, 2019; Chung et al., 2021; Feldman et al., 2021). Specifically, given an input $x \in \mathcal{X}$ and a quantile level $\alpha \in [0, 1]$, the model outputs a quantile $Q_\theta(\alpha \mid x)$. The full quantile function can be obtained by evaluating the model at multiple quantile levels. The model is trained by minimizing the quantile score at multiple levels, which is asymptotically equivalent to minimizing the CRPS (Bracher et al., 2021). We denote this model as SQR-CRPS, where SQR stands for simultaneous quantile regression (Tagasovska and Lopez-Paz, 2019).

Experimental setup. We adopt the large-sized regime introduced by Grinsztajn et al. (2022), which involves truncating the datasets to a maximum of 50,000 examples. Among the 57 datasets, the number of examples ranges from 135 to 50,000, and the number of features ranges from 3 to 3,611[1]. Each of the 57 datasets is divided into four sets: training (65%), validation (10%), calibration (15%), and test (10%). We normalize the input X and target Y using

[1]Please refer to Appendix D, specifically Table 1, for a detailed summary of each dataset.

the mean and standard deviation from the training split. The final predictions are then transformed back to the original scale. For our neural network models, we employ the same fully-connected architecture as previous studies conducted by Kuleshov, Fenner, et al. (2018), Chung et al. (2021), and Fakoor et al. (2021). Further details regarding the model hyperparameters can be found in Appendix C.

Results. In Figure 2, the first row displays the PCE (averaged over five random train-validation-test splits) for MIX-NLL in blue on each of the 57 datasets. For comparison, the PCE of a perfectly calibrated model, i.e. with uniformly distributed PITs, computed using 5×10^4 simulated values is shown in orange. The second row presents reliability diagrams for five datasets. Similar information is provided for MIX-CRPS and SQR-CRPS in Figures 12 and 13 in Appendix B.4, respectively. Additionally, reliability diagrams for all datasets can be found in Figure 15 in Appendix B.6.

The analysis reveals that the (average) PCE is generally high across many datasets, although there are significant variations between datasets. To test the statistical significance of these results, 10^4 samples were generated from the sampling distribution of the average PCE under the null hypothesis of probabilistic calibration. The resulting sampling distribution for all datasets is presented in Appendix B.5.

By computing the p-value associated with a one-sided test in the upper tail of the distribution (as illustrated in Appendix B.5), it was observed that most datasets have a p-value of zero. This indicates that the average PCE obtained for the considered model is higher than all the simulated average PCEs of the probabilistically calibrated model. Applying a threshold of 0.01 and a Holm correction for the 57 hypothesis tests, the null hypothesis is rejected for 11 datasets out of the 57.

Overall, the results indicate that the neural models considered in this study are generally probabilistically miscalibrated on a significant number of benchmark tabular datasets. In Section 6, we will further explore how calibration methods can substantially improve the PCE of neural models.

5. Calibration Methods

We begin by discussing the three main approaches to calibration: quantile recalibration, conformal prediction, and regularization-based calibration. Following that, we introduce two novel variants of regularization-based calibration.

Quantile recalibration and conformal prediction are post-hoc methods, meaning they are applied after model training. These approaches utilize a separate calibration dataset $\mathcal{D}' = \{ (X_i', Y_i') \}_{i=1}^{N'}$, where $(X_i', Y_i') \overset{\text{i.i.d.}}{\sim} P_{X,Y}$. On the other hand, regularization-based calibration operates directly during training and relies solely on the training data $\mathcal{D}$.

5.1. Quantile Recalibration

Quantile recalibration aims to transform a potentially miscalibrated CDF F_θ into a probabilistically calibrated CDF $F'_\theta = F_Z \circ F_\theta$, using the calibration map F_Z which represents the CDF of the PITs for F_θ. For a given quantile level $\alpha \in [0, 1]$, the recalibrated CDF F'_θ satisfies:

$$\begin{aligned} \Pr(F'_\theta(Y \mid X) \leq \alpha) &= \Pr(F_\theta(Y \mid X) \leq Q_Z(\alpha)) && (3) \\ &= F_Z(Q_Z(\alpha)) && (4) \\ &= \alpha. && (5) \end{aligned}$$

In practice, F_Z is not directly available and needs to be estimated from data. Kuleshov, Fenner, et al. (2018) proposed estimating it using isotonic regression, while Utpala and Rai (2020) showed that computing the empirical CDF is an equivalent and simpler method. Specifically, given a set of PIT values $Z'_i = F_\theta(Y'_i \mid X'_i), i = 1, \ldots, N'$, the calibration map ϕ^{EMP} is computed as:

$$\phi^{\text{EMP}}(\alpha; \{ Z'_i \}_{i=1}^{N'}) = \frac{1}{N'} \sum_{i=1}^{N'} \mathbb{1}(Z'_i \leq \alpha), \qquad (6)$$

where $\alpha \in [0, 1]$.

Similarly to Utpala and Rai (2020), we also consider a linear calibration map ϕ^{LIN}, which is continuous, and corresponds to a linear interpolation between the points $\Big\{ (0,0), (Z'_{(1)}, {}^{1}/_{N'+1}), \ldots, (Z'_{(N')}, {}^{N'}/_{N'+1}), (1,1) \Big\}$, where $Z'_{(k)}$ is the kth order statistic of $Z'_1, \ldots, Z'_{N'}$.

In addition, we propose a calibration map based on kernel density estimation (KDE), denoted as ϕ^{KDE}. This calibration map offers the advantage of being differentiable and can lead to improved NLL performance. The key idea is to use a relaxed approximation of the indicator function, which allows us to make the PIT CDF (6) differentiable. Specifically, we compute

$$\mathbb{1}_\tau(a \leq b) = \sigma(\tau(b - a)) \approx \mathbb{1}(a \leq b),$$

where $\tau > 0$ is an hyperparameter and $\sigma(x) = \frac{1}{1+e^{-x}}$ denotes the sigmoid function. The resulting smoothed empirical CDF is given by

$$\phi^{\text{KDE}}(\alpha; \{ Z'_i \}_{i=1}^{N'}) = \frac{1}{N'} \sum_{i=1}^{N'} \mathbb{1}_\tau(Z'_i \leq \alpha). \qquad (7)$$

This corresponds to estimating the CDF F_Z using KDE based on N' realizations of Z ($\{ Z'_i \}_{i=1}^{N'}$). Since σ is the CDF of the logistic distribution, we use the PDF of the logistic distribution as the kernel in the KDE. Algorithm 1 summarises this method.

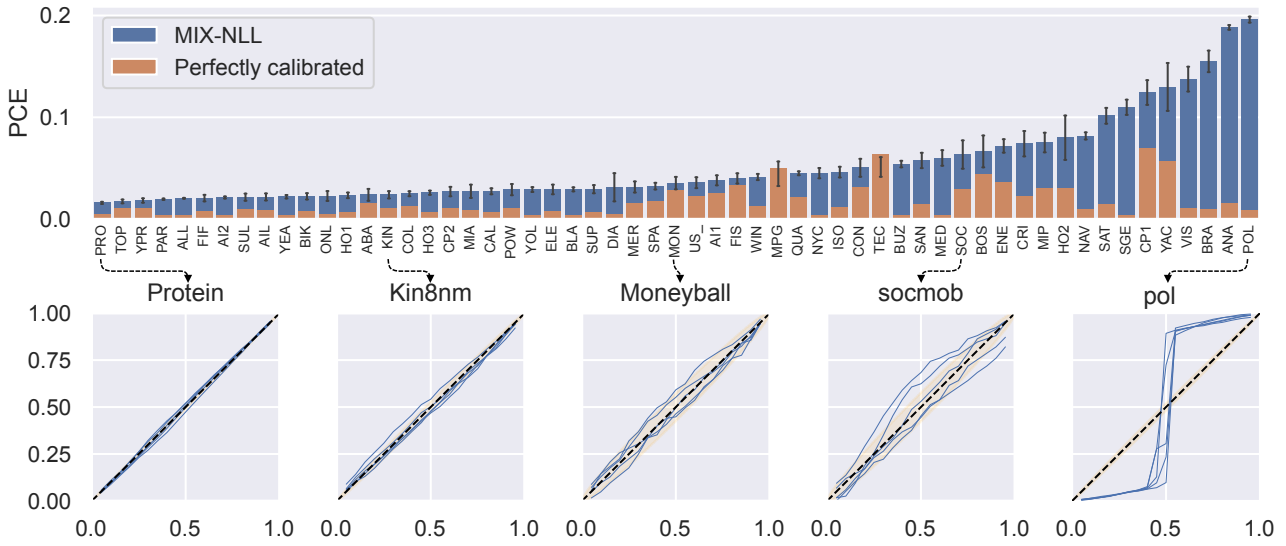


Figure 2: The top row shows the PCE for different datasets with one standard error (error bar). The bottom row gives examples of PIT reliability diagrams for five datasets.

Algorithm 1 Quantile recalibration

Input: Predictive CDF F_θ and $\mathcal{D}' = \{ (X_i', Y_i') \}_{i=1}^{N'}$.
Compute $Z_i' = F_\theta(Y_i' \mid X_i') \quad (i = 1, \ldots, N')$
Compute the calibration map ϕ, either ϕ^{EMP}, ϕ^{LIN}, or ϕ^{KDE}
Return: Recalibrated CDF $F_\theta' = \phi \circ F_\theta$.

5.2. Conformal Prediction

Let us assume the realizations of our calibration dataset $\mathcal{D}'$ are drawn exchangeably from $P_{X,Y}$[2]. Given a predictive model M_θ and a coverage level $\alpha \in [0, 1]$, (inductive) conformal prediction allows us to construct a prediction set $C_\alpha(X) \subseteq \mathcal{Y}$ for any input X, satisfying the property:

$$\Pr(Y \in C_\alpha(X)) = \frac{\lceil (N'+1)\alpha \rceil}{N'+1} \tag{8}$$

$$\approx \alpha. \tag{9}$$

Conformal prediction achieves this by utilizing a conformity score $s_\theta(Y \mid X)$, which intuitively quantifies the similarity between new samples and previously observed samples. When the conformity score increases with Y, an interval $C_\alpha(X) = (-\infty, s_\theta^{-1}(\alpha \mid X)]$ can be constructed, ensuring the conformal guarantee (8) at level α.

Let $Q_\theta'(\alpha \mid X) = s_\theta^{-1}(\alpha \mid X)$ represent the (revised) model obtained through conformal prediction from $Q_\theta(\alpha \mid X)$. Under the assumption that Q_θ' is continuous and strictly increasing, the conformal guarantee implies that $\Pr(Y \leq Q_\theta'(\alpha \mid X)) \approx \alpha$, which indicates approximate probabilistic calibration at level α.

Conformalized Quantile Regression (Romano, Patterson, et al., 2019) is an example of a conformal procedure, where the conformity score is defined as $s_\theta(Y \mid X) = Y - Q_\theta(\alpha \mid X)$, representing the quantile residual. Another example is Distributional Conformal Prediction (Izbicki et al., 2020; Chernozhukov et al., 2021), which employs the conformity score $s_\theta(Y \mid X) = F_\theta(Y \mid X)$, referring to the PIT. Algorithm 2 provides a summary of how to compute calibrated quantiles using inductive conformal prediction.

Algorithm 2 Calibrated quantiles with conformal prediction

Input: Trained model M_θ, $\mathcal{D}' = \{ (X_i', Y_i') \}_{i=1}^{N'}$, strictly increasing conformity score s, quantile level $\alpha \in [0, 1]$, input X.
Compute $S_i = s_\theta(Y_i' \mid X_i') \quad (i = 1, \ldots, N')$
Compute $\hat{q} = S_{(\lceil (N'+1)\alpha \rceil)}$ where $S_{(k)}$ denote the kth smallest value among $\{ S_1, \ldots, S_{N'}, +\infty \}$
Return: Calibrated quantile $Q_\theta'(\alpha \mid X) = s_\theta^{-1}(\hat{q} \mid X)$

5.3. Regularization-based Calibration

Regularization-based calibration methods aim to enhance model calibration by incorporating a regularization term into the training objective. Although widely used in classification, there are relatively fewer methods specifically designed for regression problems. In this section, we discuss two approaches: quantile regularization (Utpala and Rai, 2020) and the truncation method (Chung et al., 2021). The main

[2]This is implied by the common i.i.d. assumption.

steps of regularization-based calibration are summarized in Algorithm 3.

Algorithm 3 Regularization-based calibration

Input: Model M_θ, calibration regularizer $\mathcal{R}(\theta)$ and tuning parameter $\lambda \geq 0$.
Compute $\theta^* = \arg\min_{\theta\in\Theta} \mathcal{L}^{'}(\theta; \mathcal{D})$ where
$\mathcal{L}^{'}(\theta; \mathcal{D}) = {}^1\!/_N \sum_{i=1}^N \mathcal{L}(M_\theta(\cdot \mid X_i), Y_i) + \lambda\mathcal{R}(\theta; \mathcal{D})$
Return: Regularized model M_{θ^*}

5.3.1. Quantile Regularization

The regularizer proposed by Utpala and Rai (2020) aims to measure the deviation of the PIT variable Z from a uniform distribution, which is characteristic of a probabilistically calibrated model. This regularization penalty encourages the selection of calibrated models during training.

The authors observed that the KL divergence between Z and a uniform random variable is equivalent to the negative differential entropy of Z, denoted as $H(Z)$. To approximate $H(Z)$, they employed sample-spacing entropy estimation (Vasicek, 1976), resulting in the following regularizer:

$$\mathcal{R}_{\text{QR}}(\theta; \mathcal{D}) \tag{10}$$

$$= \frac{1}{N-k} \sum_{i=1}^{N-k} \log\left[\frac{N+1}{k}(Z_{(i+k)} - Z_{(i)})\right] \tag{11}$$

$$\approx H(Z), \tag{12}$$

where k is a hyperparameter satisfying $1 \leq k \leq N$, and $Z_{(i)}$ represents the ith order statistic of Z.

To ensure differentiability during optimization, the authors employed a differentiable relaxation technique called NeuralSort (Grover et al., 2019), as sorting is a non-differentiable operation.

5.3.2. Truncation-based Calibration

The regularization approach introduced by Chung et al. (2021), that we denote Trunc, involves truncating the predictive distribution based on the current level of calibration.

Given a quantile model Q_θ, let $\hat{F}_Z(\alpha) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}(Y_i \leq Q_\theta(\alpha \mid X_i))$ be the estimated PIT CDF evaluated at α and $\rho(x, y) = (y - x)\mathbb{1}(x < y)$. The regularization objective for level α is defined as follows:

$$\mathcal{R}_{\text{Trunc}}(\theta; \mathcal{D}, \alpha) \tag{13}$$

$$= \begin{cases} \frac{1}{N}\sum_{i=1}^N \rho(Q_\theta(\alpha \mid X_i), Y_i) & \text{if } \hat{F}_Z(\alpha) < \alpha \\ \frac{1}{N}\sum_{i=1}^N \rho(Y_i, Q_\theta(\alpha \mid X_i)) & \text{otherwise} \end{cases} \tag{14}$$

This regularization objective adjusts $\hat{F}_Z(\alpha)$ to match α by increasing it when $\hat{F}_Z(\alpha) < \alpha$, and vice versa. The final regularization objective is computed by averaging $\mathcal{R}_{\text{Trunc}}(\theta; \mathcal{D}, \alpha)$ over multiple quantile levels $\{\, \alpha_j \,\}_{j=1}^M$:

$$\mathcal{R}_{\text{Trunc}}(\theta; \mathcal{D}) = \frac{1}{M}\sum_{j=1}^M \mathcal{R}_{\text{trunc}}(\theta; \mathcal{D}, \alpha_j). \tag{15}$$

It is worth noting that Chung et al. (2021) combine the previous regularization objective with a sharpness objective that penalizes the width between the quantile predictions, given by $\frac{1}{M}\sum_{j=1}^M \frac{1}{N}\sum_{i=1}^N |Q_\theta(\alpha_j \mid X_i) - Q_\theta(1 - \alpha_j \mid X_i)|$. Instead, we combine it with a strictly proper scoring rule.

5.4. New Regularization-based Calibration Methods

Building upon the quantile calibration method discussed in Section 5.3.1, we propose two new regularization objectives which compute a differentiable PCE$_p$ using alternative statistical distances.

The first approach, named PCE-KDE, leverages the differentiable calibration map ϕ^{KDE} (7) based on KDE. Given a set of quantile levels $\{\, \alpha_j \,\}_{j=1}^M$, the regularization objective is given by

$$\mathcal{R}_{\text{PCE-KDE}}(\theta; \mathcal{D}) = \frac{1}{M}\sum_{j=1}^M \left|\alpha_j - \phi^{\text{KDE}}(\alpha_j; \{\, Z_i \,\}_{i=1}^N)\right|^p, \tag{16}$$

where $p > 0$. Note that $\mathcal{R}_{\text{PCE-KDE}}$ reduces to PCE$_p$ in (2) when τ in (7) goes to ∞.

The second approach considers distances of the form $\int_0^1 |Q_Z(\alpha) - Q_U(\alpha)|^p d\alpha$, where Q_Z and Q_U denote the quantile functions of the true and uniform distributions, respectively. When $p = 1$, this distance reduces to the 1-Wasserstein distance, equivalent to $\int_0^1 |F_Z(\alpha) - F_U(\alpha)| d\alpha$, which aligns with PCE (see Proposition 1 in Appendix A.1).

By exploiting the fact that $\mathbb{E}\left[F_Z(Z_{(i)})\right] = {}^i\!/_{N+1}$, we approximate $Q_Z({}^i\!/_{N+1})$ using the i-th order statistic $Z_{(i)}$. The regularization objective is given by

$$\mathcal{R}_{\text{PCE-Sort}}(\theta; \mathcal{D}) = \frac{1}{N}\sum_{i=1}^N \left|Z_{(i)} - \frac{i}{N+1}\right|^p, \tag{17}$$

where $p > 0$. Differentiable relaxations to sorting, such as those proposed by Blondel et al. (2020) and Cuturi et al. (2019), can be employed to obtain the order statistics.

6. A Comparative Study of Probabilistic Calibration Methods

In continuation of the empirical study described in Section 4, we now proceed to evaluate the performance of the probabilistic calibration methods outlined in the previous section.

Specifically, we apply eight distinct calibration methods to the three neural regression models introduced in Section 4. These methods are evenly divided into two categories: post-hoc methods and regularization-based methods.

To assess the effectiveness of these calibration methods, we employ four different evaluation metrics. The evaluation is conducted on a set of 57 datasets, utilizing the same experimental setup detailed in Section 4. To ensure a fair and consistent comparison, all the methods have been implemented within a unified codebase[3].

6.1. Experimental Setup

Base probabilistic models and calibration methods. We consider the three probabilistic models presented in Section 4, namely `MIX-NLL`, `MIX-CRPS`, and `SQR-CRPS`. For the `MIX` models, when applying quantile recalibration, we transform the CDF using the empirical CDF estimator (`Rec-EMP`), the linear estimator (`Rec-LIN`), or the KDE estimator (`Rec-KDE`). For `SQR-CRPS`, we transform multiple quantiles using conformalized quantile regression (`CQR`). For the three models, we consider the four regularization objectives presented in Sections 5.3 and 5.4 (with $p = 1$), namely $\mathcal{R}_{\text{PCE-KDE}}$ (`PCE-KDE`), $\mathcal{R}_{\text{PCE-Sort}}$ (`PCE-Sort`), $\mathcal{R}_{\text{QR}}$ (`QR`), and $\mathcal{R}_{\text{Trunc}}$ (`Trunc`). `PCE-Sort` is only shown in the Appendix because it performs similarly to `PCE-KDE`.

Metrics. We measure the accuracy of the probabilistic predictions using NLL and CRPS. For the SQR model, we estimate CRPS by averaging the quantile score at 64 equidistant quantile levels. Probabilistic calibration is measured using PCE, defined in (1). Finally, we measure sharpness using the mean standard deviation of the predictive distributions, denoted by STD.

Hyperparameters. In our experiments, `MIX-NLL` and `MIX-CRPS` output a mixture of 3 Gaussians, and `SQR-CRPS` outputs 64 quantiles. We justify the choice of these hyperparameters in Appendix C. The hyperparameter τ of `Rec-KDE` and `PCE-KDE` is fixed at 100, which was found to perform well empirically. For regularization methods, an important hyperparameter is the regularization factor λ. As previously observed in classification (Karandikar et al., 2021), we found that higher values of λ tend to improve calibration but worsen NLL, CRPS, and STD. Karandikar et al. (2021) proposed to limit the loss in accuracy by a maximum of 1%. We adopt a similar strategy by selecting λ which minimizes PCE with a maximum increase in CRPS of 10% in the validation set. For each dataset, we select λ in the set $\{0, 0.01, 0.05, 0.2, 1, 5\}$, which corresponds to various degrees of calibration regularization.

[3]The code can be accessed at the following GitHub repository: `https://github.com/Vekteur/probabilistic-calibration-study`

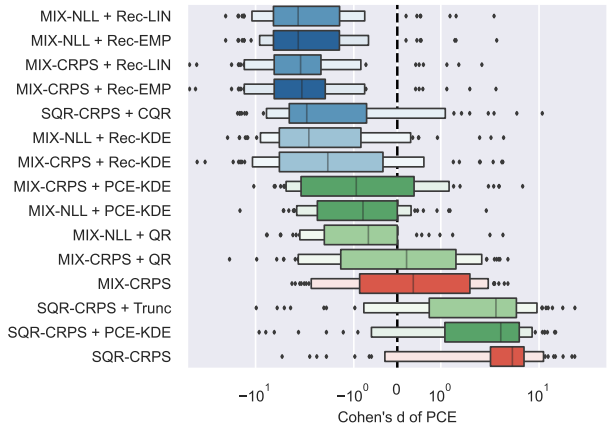
Comparison of multiple models over many datasets. As in Karandikar et al. (2021), and since NLL, CRPS and STD have different scales across datasets, we report Cohen's d, which is a standardized effect size comparing the mean of one method (over 5 runs, in our case) against a baseline. Values of -0.8 and -2 are considered large and huge, respectively. Due to the heterogeneity of the datasets that we consider, the performance of our models can vary widely across datasets. To visualize the results, we show the distribution of Cohen's d using letter-value plots (Hofmann et al., 2011), which indicate the quantiles at levels $1/8$, $1/4$, $1/2$, $3/4$ and $7/8$, as well as outliers. A median value below zero indicates that the model improved the metric on more than half the datasets.

In order to assess whether significant differences exist between different methods, we follow the recommendations of Ismail Fawaz et al. (2019), which are based on (Demšar, 2006). First, we test for a significant difference among model performances using the Friedman test (Friedman, 1940). Then, we use the pairwise post-hoc analysis recommended by Benavoli et al. (2016) using a Wilcoxon signed-rank test (Wilcoxon, 1945) with Holm's alpha correction (Holm, 1979). The results of this procedure are shown using a critical difference diagram (Demšar, 2006). The lower the rank (further to the right), the better performance of a model. A thick horizontal line shows a group of models whose performance is not significantly different, with a significance level of 0.05.

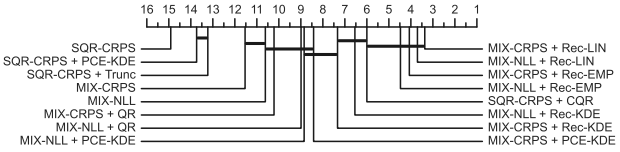
6.2. Results

Figure 3 shows the letter-values plots for the Cohen's d of PCE (top panel) as well as the associated critical diagram (bottom panel), for all methods and datasets. The reference model is `MIX-NLL`. The results with other models as reference are available in Appendix B.1. Blue, green, and red colors are used for the post-hoc methods, the regularization-based methods, and the base models, respectively. The same information is given in Figures 4 and 5 for the CRPS and the NLL, respectively.

Comparison of PCE. As expected, Figure 3 shows that the PCE of calibration methods is improved compared to the base models. Furthermore, independently of the base model, we can see that post-hoc methods achieve significantly better PCE than regularization methods. When comparing `PCE-KDE` with `QR`, we can see that there is a significantly larger decrease in PCE with the `MIX-CRPS` base model compared to `MIX-NLL`. Finally, both `PCE-KDE` and `Trunc` decrease PCE for `SQR-CRPS`, without a significant difference between them.

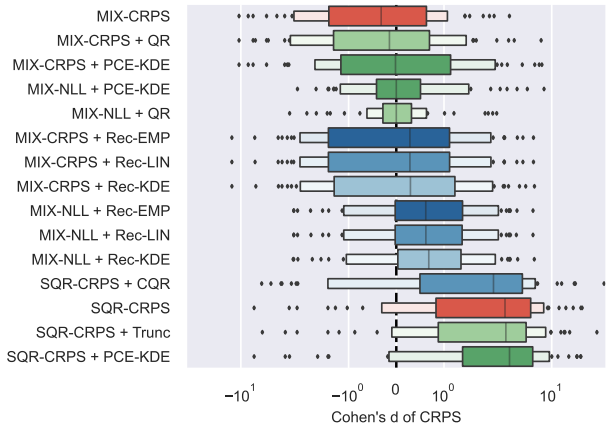


(a) Cohen's d of PCE with respect to the MIX−NLL model

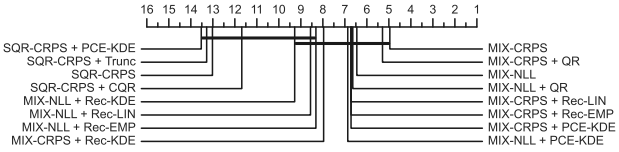


(b) Critical difference diagram

Figure 3: Comparison of PCE with multiple base losses and calibration methods.



(a) Cohen's d of CRPS with respect to the MIX−NLL model



(b) Critical difference diagram

Figure 4: Comparison of CRPS with multiple base losses and calibration methods.

Comparison of CRPS. While post-hoc methods outperform regularization methods in terms of PCE, Figure 4 shows they have a higher CRPS (except for the SQR base model). This can be explained by the fact that regularization methods prevent the CRPS from increasing exceedingly due to the selection criterion for λ.

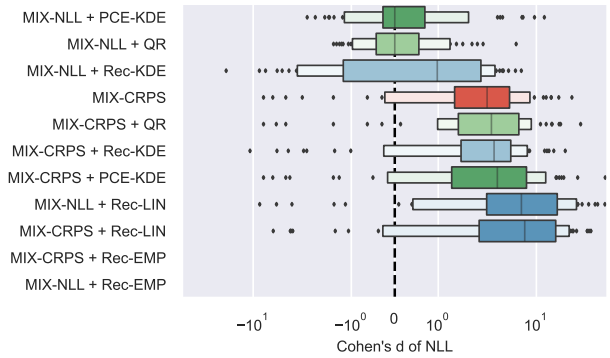
Comparison of NLL. Figure 5 shows the importance of the calibration map. In fact, quantile recalibration with a linear map significantly increases the NLL, while smooth interpolation decreases PCE without a large increase in NLL. Note that we only consider MIX models since we cannot compute the NLL for SQR.

On the choice of a calibration method. If probabilistic calibration is critical to the application, our experiments suggest that post-hoc methods such as quantile recalibration and conformal prediction should be preferred. However, when we also want to control the CRPS or the NLL, regularization methods can offer a better trade-off in terms of calibration and sharpness. In fact, as shown in Figure 6 in Appendix B.1, when the base model is MIX−NLL, all regularization methods provide a significant improvement in probabilistic calibration without deteriorating the CRPS, NLL or STD. For the MIX−CRPS model, Figure 7 shows that QR has limited impact on CRPS and NLL, while providing better calibration. For the SQR−CRPS base model, Figure 8 shows that the SQR−CRPS + CQR conformal method significantly outperforms the Trunc and PCE−KDE regularization methods both in terms of PCE and CRPS. Overall, Appendix B.1 suggests that MIX−NLL + PCE−KDE, MIX−CRPS + QR and SQR−CRPS + CQR are good choices for practitioners aiming to improve PCE without significantly impacting other aspects of the conditional distribution. Finally, since both regularization and post-hoc methods are able to improve calibration, we investigate whether a combination of these two methods can lead to better performance. Figure 9 in Appendix B.2 shows that such a combination does not significantly improve probabilistic calibration, with an increase in CRPS and NLL. This indicates that practitioners should exercise caution when applying regularization to a model that is already well-calibrated.

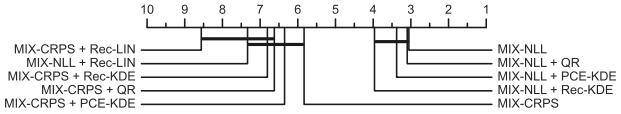
6.3. Link between Quantile Recalibration and Conformal Prediction

Conformal prediction methods are well-known for their finite-sample coverage guarantee. Interestingly, a specific implementation of quantile recalibration can be considered a special case of conformal prediction. This implies that quantile recalibration can also provide a finite-sample coverage guarantee. This observation could potentially explain why both methods, conformal prediction and quantile recalibration, are effective in improving probabilistic calibration.

Theorem 1. *Quantile recalibration is equivalent to Distributional Conformal Prediction (DCP) of left intervals at each coverage level* $\alpha \in [0, 1]$. *The equivalence is obtained*



(a) Cohen's d of NLL with respect to the `MIX-NLL` model



(b) Critical difference diagram

Figure 5: Comparison of NLL with multiple base losses and calibration methods.

when the estimator of the calibration map is defined by a slightly different estimator than the conventional one in (6)*, namely* $\phi_{DCP}(\alpha) = \frac{1}{N'+1} \sum_{i=1}^{N'} \mathbb{1}(Z'_i \leq \alpha)$.

Proof. Given a predictive distribution F_θ learned from a training dataset $\mathcal{D} = \{ (X_i, Y_i) \}_{i=1}^{N}$ where $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$, let $Z'_i = F_\theta(Y'_i \mid X'_i)$ represent the PIT values computed on a separate calibration dataset $\mathcal{D}' = \{ (X'_i, Y'_i) \}_{i=1}^{N'}$, where $(X'_i, Y'_i) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$.

In the DCP approach, as outlined in Algorithm 2, the conformal scores are given by the PIT values Z'_i. DCP first computes the α empirical quantile of the scores as $\hat{q} = Z'_{(\lceil (N'+1)\alpha \rceil)}$, where $Z'_{(k)}$ represents the kth smallest value among $\{ Z'_1, \ldots, Z'_{N'}, +\infty \}$. Then, the conformalized quantile is computed as $Q'_\theta(\alpha \mid X) = Q_\theta(\hat{q} \mid X)$, which corresponds to conformal prediction with coverage α for the left interval $(-\infty, Q'_\theta(\alpha \mid X)]$.

Let us consider quantile recalibration with the calibration map ϕ in Algorithm 1 given by $\phi_{\text{DCP}}(\alpha) = \frac{1}{N'+1} \sum_{i=1}^{N'} \mathbb{1}(Z'_i \leq \alpha)$. It computes a recalibrated CDF F'_θ by composing the original CDF F_θ with ϕ_{DCP}, yielding $F'_\theta(y \mid X) = \phi_{\text{DCP}}(F_\theta(y \mid X))$.

We observe that ϕ_{DCP} is the CDF of a discrete random variable, with $\phi^{-1}_{\text{DCP}}(\alpha) = Z'_{(\lceil (N'+1)\alpha \rceil)}$ representing its empirical quantile function. Furthermore, the composition $\phi_{\text{DCP}} \circ F_\theta(\cdot \mid X)$ acts as the inverse function of $Q_\theta(\cdot \mid X) \circ \phi^{-1}_{\text{DCP}}$. As a result, both the DCP approach and quantile recalibration yield QFs and CDFs that correspond to the same underlying distribution. □

Quantile recalibration with other recalibration maps (e.g., ϕ^{EMP}, ϕ^{LIN}, or ϕ^{KDE}) would correspond to DCP where the empirical quantile $\hat{q}$ is selected using other strategies which does not provide the exact conformal guarantee (8).

7. Conclusion

The observation that neural network classifiers tend to be miscalibrated (Guo et al., 2017) has prompted the development of various approaches for calibrating these models. In this paper, we present the largest empirical study conducted to date on the probabilistic calibration of neural regression models. Our study provides valuable insights into their performance and the selection of calibration methods. Notably, we introduce a novel differentiable calibration map based on kernel density estimation for quantile recalibration, as well as two novel regularization objectives derived from the PCE.

Our study reveals that regularization methods can provide a favorable tradeoff between calibration and sharpness. However, post-hoc methods demonstrate superior performance in terms of PCE. We attribute this finding to the finite-sample coverage guarantee offered by conformal prediction and demonstrate that quantile recalibration can be viewed as a specific case of conformal prediction.

Future investigations may extend the study of probabilistic calibration to other models, such as tree-based models, and explore alternative notions of calibration (Gneiting and Resin, 2021). Notably, distribution calibration represents a promising direction, as it has inspired the development of calibration methods (Song et al., 2019; Kuleshov and Deshpande, 2022).

References

[1] Moloud Abdar et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges". *An international journal on information fusion* 76 (Dec. 2021), pp. 243–297.

[2] Alessio Benavoli, Giorgio Corani, and Francesca Mangili. "Should We Really Use Post-Hoc Tests Based on Mean-Ranks?" *Journal of machine learning research: JMLR* 17.5 (2016), pp. 1–10.

[3] Mathieu Blondel et al. "Fast Differentiable Sorting and Ranking". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé Iii and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 950–959.

[4] Johannes Bracher et al. "Evaluating epidemic forecasts in an interval format". en. *PLoS computational biology* 17.2 (Feb. 2021), e1008618.

[5] Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. "Distributional conformal prediction". en. *Proceedings of the National Academy of Sciences of the United States of America* 118.48 (Nov. 2021).

[6] Youngseog Chung et al. "Beyond Pinball Loss: Quantile Methods for Calibrated Uncertainty Quantification". In: *Advances in Neural Information Processing Systems*. Ed. by M Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 10971–10984.

[7] Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. "Differentiable Ranks and Sorting using Optimal Transport" (28 5 2019). arXiv: 1905.11885 [cs.LG].

[8] A P Dawid. "Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach". *Journal of the Royal Statistical Society. Series A* 147.2 (1984), pp. 278–292.

[9] Janez Demšar. "Statistical Comparisons of Classifiers over Multiple Data Sets". *Journal of machine learning research: JMLR* 7 (Dec. 2006), pp. 1–30.

[10] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017.

[11] Rasool Fakoor et al. "Flexible Model Aggregation for Quantile Regression" (Feb. 2021). arXiv: 2103.00083 [stat.ML].

[12] Feldman, Bates, and Romano. "Improving conditional coverage via orthogonal quantile regression". *Advances in neural information processing systems* (2021).

[13] Rina Foygel Barber et al. "The limits of distribution-free conditional predictive inference". *Information and Inference: A Journal of the IMA* 10.2 (15 6 2021), pp. 455–482.

[14] Milton Friedman. "A Comparison of Alternative Tests of Significance for the Problem of m Rankings". en. *The Annals of Mathematical Statistics* 11.1 (Mar. 1940), pp. 86–92.

[15] Yarin Gal and Zoubin Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, USA: PMLR, 2016, pp. 1050–1059.

[16] Jakob Gawlikowski et al. "A Survey of Uncertainty in Deep Neural Networks" (July 2021). arXiv: 2107.03342 [cs.LG].

[17] Pieter Gijsbers et al. "An Open Source AutoML Benchmark" (July 2019). arXiv: 1907.00909 [cs.LG].

[18] Tilmann Gneiting and Johannes Resin. "Regression Diagnostics meets Forecast Evaluation: Conditional Calibration, Reliability Diagrams, and Coefficient of Determination" (Aug. 2021). arXiv: 2108.03210 [stat.ME].

[19] E P Grimit et al. "The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification". en. *Quarterly Journal of the Royal Meteorological Society* 132.621C (Oct. 2006), pp. 2925–2942.

[20] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on tabular data?" (July 2022). arXiv: 2207.08815 [cs.LG].

[21] Aditya Grover et al. "Stochastic Optimization of Sorting Networks via Continuous Relaxations". In: *International Conference on Learning Representations*. 2019.

[22] Vitor Guizilini et al. "3D Packing for Self-Supervised Monocular Depth Estimation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[23] Varun Gulshan et al. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs". en. *JAMA: the journal of the American Medical Association* 316.22 (13 12 2016), pp. 2402–2410.

[24] Chuan Guo et al. "On Calibration of Modern Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1321–1330.

[25] Heike Hofmann, Karen Kafadar, and Hadley Wickham. *Letter-value plots: Boxplots for large data*. Tech. rep. had.co.nz, 2011.

[26] Sture Holm. "A Simple Sequentially Rejective Multiple Test Procedure". *Scandinavian journal of statistics, theory and applications* 6.2 (1979), pp. 65–70.

[27] Hassan Ismail Fawaz et al. "Deep learning for time series classification: a review". *Data mining and knowledge discovery* 33.4 (Jan. 2019), pp. 917–963.

[28] Rafael Izbicki, Gilson Shimizu, and Rafael Stern. "Flexible distribution-free conditional predictive bands using density estimators". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 3068–3077.

[29] Laurent Valentin Jospin et al. "Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users". *IEEE Computational Intelligence Magazine* 17.2 (May 2022), pp. 29–48.

[30] Archit Karandikar et al. "Soft Calibration Objectives for Neural Networks" (July 2021). arXiv: 2108.00106 [cs.LG].

[31] Volodymyr Kuleshov and Shachi Deshpande. "Calibrated and Sharp Uncertainties in Deep Learning via Density Estimation". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 11683–11693.

[32] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. "Accurate Uncertainties for Deep Learning Using Calibrated Regression". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2796–2804.

[33] Lakshminarayanan, Pritzel, et al. "Simple and scalable predictive uncertainty estimation using deep ensembles". *Advances in neural information processing systems* (2017).

[34] Matthias Minderer et al. "Revisiting the Calibration of Modern Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by M Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 15682–15694.

[35] Tim Pearce et al. "High-Quality Prediction Intervals for Deep Learning: A Distribution-Free, Ensembled Approach". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 4075–4084.

[36] Pierre Pinson and Renate Hagedorn. "Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations". en. *Meteorological Applications* 19.4 (Dec. 2012), pp. 484–500.

[37] Teodora Popordanoska, Raphael Sayer, and Matthew B Blaschko. "A Consistent and Differentiable Lp Canonical Calibration Error Estimator". Oct. 2022.

[38] Romano, Patterson, et al. "Conformalized quantile regression". *Advances in neural information processing systems* (2019).

[39] Hao Song et al. "Distribution calibration for regression". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 5897–5906.

[40] Tagasovska and Lopez-Paz. "Single-model uncertainties for deep learning". *Advances in neural information processing systems* (2019).

[41] Jayaraman J Thiagarajan et al. "Designing accurate emulators for scientific processes using calibration-driven deep models". en. *Nature communications* 11.1 (June 2020), p. 5622.

[42] Saiteja Utpala and Piyush Rai. "Quantile Regularization: Towards Implicit Calibration of Regression Models" (Feb. 2020). arXiv: 2002.12860 [cs.LG].

[43] Oldrich Vasicek. "A Test for Normality Based on Sample Entropy". *Journal of the Royal Statistical Society. Series B, Statistical methodology* 38.1 (1976), pp. 54–59.

[44] Vladimir Vovk et al. "Conformal calibrators". In: *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*. Ed. by Alexander Gammerman et al. Vol. 128. Proceedings of Machine Learning Research. PMLR, 2020, pp. 84–99.

[45] Frank Wilcoxon. "Individual Comparisons by Ranking Methods". *Biometrics Bulletin* 1.6 (1945), pp. 80–83.

[46] Hee Suk Yoon et al. "ESD: Expected Squared Difference as a Tuning-Free Trainable Calibration Measure". In: *The Eleventh International Conference on Learning Representations*. 2023.

[47] Shengjia Zhao, Tengyu Ma, and Stefano Ermon. "Individual Calibration with Randomized Forecasting". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé Iii and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 11387–11397.

[48] Tianhui Zhou et al. "Estimating Uncertainty Intervals from Collaborating Networks". en. *Journal of machine learning research: JMLR* 22 (Jan. 2021).

A. Proofs

A.1. Integral of the Absolute Difference between CDFs or QFs

Proposition 1. Let $F_A, F_B : [0,1] \to [0,1]$ denote two strictly increasing CDFs of random variables defined on $[0,1]$ with corresponding QFs Q_A and Q_B. Then,

$$\int_0^1 |F_A(q) - F_B(q)|\, dq = \int_0^1 |Q_A(p) - Q_B(p)|\, dp. \tag{18}$$

Proof. We define two functions $r, s : [0,1] \times [0,1] \to \{\, 0, 1 \,\}$ where

$$r(q,p) = \begin{cases} 1 & \text{if } F_A(q) \le p \le F_B(q) \text{ or } F_B(q) \le p \le F_A(q) \\ 0 & \text{otherwise,} \end{cases} \tag{19}$$

$$\text{and } s(q,p) = \begin{cases} 1 & \text{if } Q_A(p) \le q \le Q_B(p) \text{ or } Q_B(p) \le q \le Q_A(p) \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

Let us show that r and s are equal. Considering $q \in [0,1]$ and $p \in [0,1]$, we can write

$$F_A(q) \le p \le F_B(q) \tag{21}$$
$$\iff (F_A(q) \le p) \wedge (p \le F_B(q)) \tag{22}$$
$$\iff (q \le Q_A(p)) \wedge (Q_B(p) \le q) \tag{23}$$
$$\iff Q_B(p) \le q \le Q_A(p), \tag{24}$$

where (23) holds since both F_A and F_B are strictly increasing.

Similarly, $F_B(q) \le p \le F_A(q) \iff Q_A(p) \le q \le Q_B(p)$. Hence $r(q,p) = 1 \iff s(q,p) = 1$ and r and s are equal.

By Fubini's theorem, we have

$$\int_0^1 \int_0^1 r(q,p)\, dp\, dq = \int_0^1 \int_0^1 s(q,p)\, dq\, dp. \tag{25}$$

Furthermore, upon evaluating the inner integrals, we obtain

$$\int_0^1 r(q,p)\, dp = \begin{cases} \displaystyle\int_{F_A(q)}^{F_B(q)} 1\, dp & \text{if } F_A(q) \le F_B(q) \\ \displaystyle\int_{F_B(q)}^{F_A(q)} 1\, dp & \text{otherwise} \end{cases} \tag{26}$$

$$= |F_A(q) - F_B(q)|. \tag{27}$$

Similarly, we have $\displaystyle\int_0^1 s(q,p)\, dq = |Q_A(p) - Q_B(p)|$. Finally, by substituting these results in (25), we prove (18).

□

B. Detailed Results

This section presents additional experimental results.

B.1. Comparison between Recalibration, Conformal Prediction and Regularization Approaches per Base Model

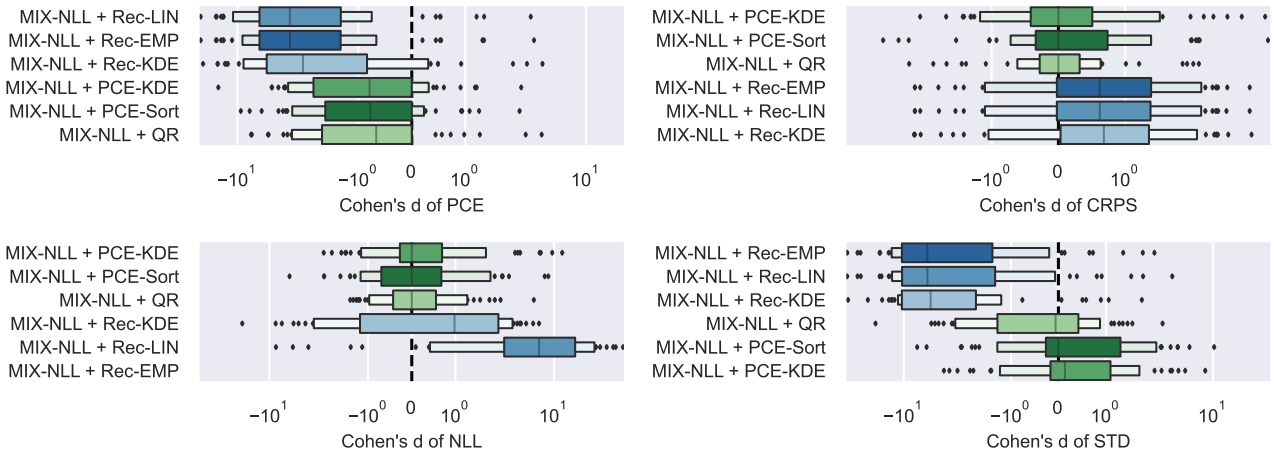
First, we present the results of our experiments comparing recalibration, conformal prediction, and regularization approaches. Our objective is to determine which metrics are improved by these methods compared to a vanilla model. We divide our comparisons based on the three base models considered: `MIX-NLL` (Figure 6), `MIX-CRPS` (Figure 7) and `SQR-CRPS` (Figure 8).

Since NLL, CRPS, and standard deviation cannot be directly compared across different datasets, we utilize Cohen's d as an effect size measure, with the baseline being a vanilla model of the same base model. For instance, the baseline for `MIX-CRPS + Rec-EMP` is `MIX-CRPS`. Additionally, we provide critical difference diagrams to assess the significance of differences.

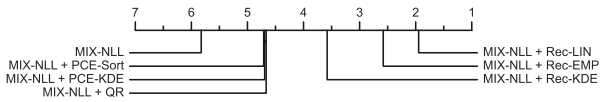
Overall, recalibration and conformal prediction demonstrate significantly improved PCE compared to the baseline, although there is a trade-off with other metrics. For both `MIX-NLL` and `MIX-CRPS`, `Rec-EMP` yields infinite NLL, `Rec-LIN` substantially increases NLL, while `Rec-KDE` has a lesser impact on NLL. However, `Rec-KDE` results in a significant degradation of CRPS compared to other recalibration methods when the base model is MIX-CRPS. In the case of quantile predictions, `CQR` significantly improves PCE.

While regularization methods generally lead to improved PCE, they are still outperformed by recalibration and conformal prediction in this regard. However, we observe that with the `MIX-NLL` base model, regularization methods (`PCE-KDE`, `PCE-Sort` and `QR`) have minimal impact on CRPS, NLL, and STD compared to recalibration methods. With the `MIX-CRPS` base model, the difference in CRPS between recalibration and regularization is less pronounced. Nevertheless, it is evident that regularization methods `PCE-KDE` and `PCE-Sort`, which rely on PCE, result in less sharp predictions compared to recalibration methods, which produce sharper predictions.

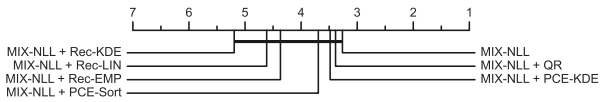
Regarding quantile predictions, the case is reversed: conformal prediction (`SQR-CRPS + CQR`) yields less sharp predictions, while regularization with `SQR-CRPS + Trunc` leads to sharper predictions.



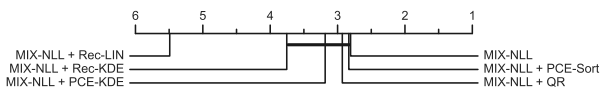
(a) Boxplots of Cohen's d of different metrics on all datasets, with respect to `MIX-NLL`.



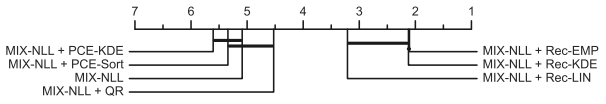
(b) PCE



(c) CRPS

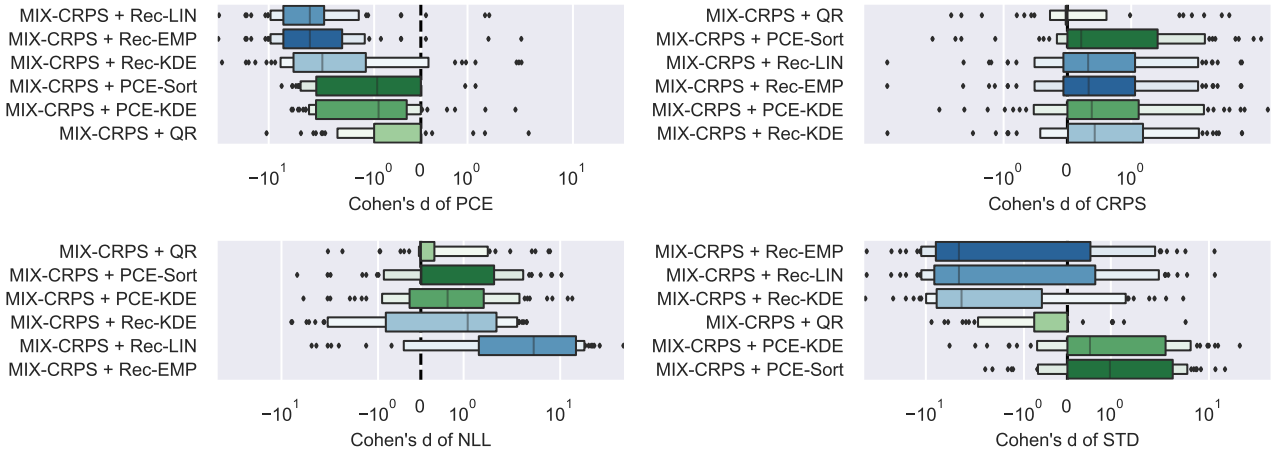


(d) NLL

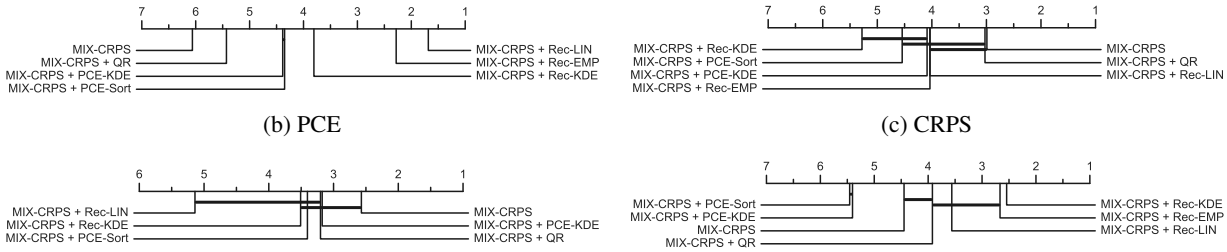


(e) STD

Figure 6: Comparison of different metrics where the base model is `MIX-NLL`.



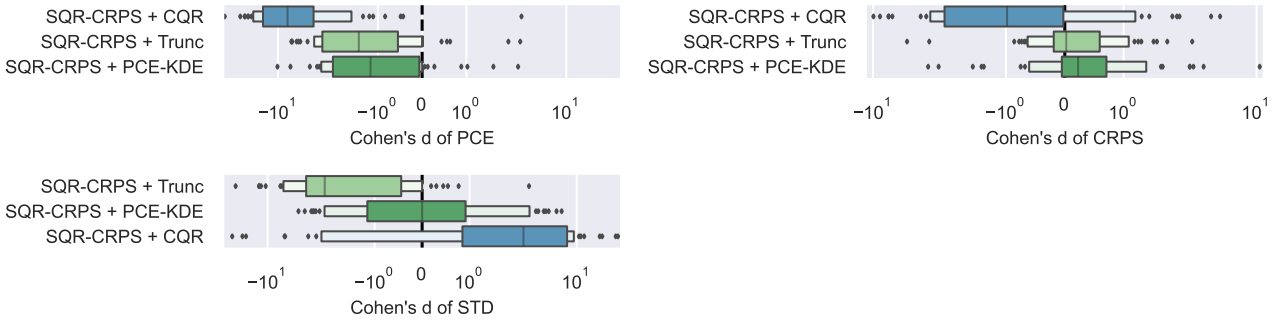
(a) Boxplots of Cohen's d of different metrics on all datasets, with respect to `MIX-CRPS`.



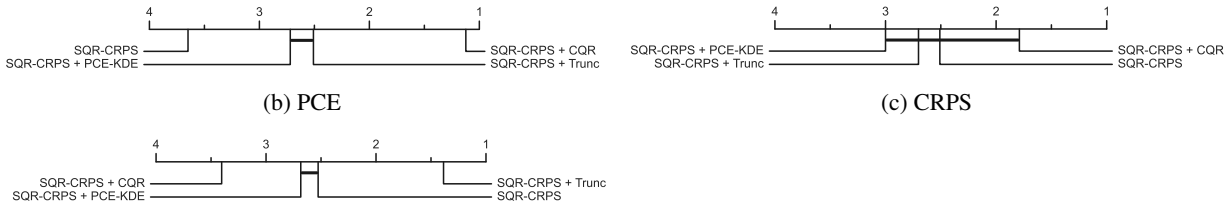
(b) PCE (c) CRPS

(d) NLL (e) STD

Figure 7: Comparison of different metrics where the base model is `MIX-CRPS`.



(a) Boxplots of Cohen's d of different metrics on all datasets, with respect to `SQR-CRPS`.



(b) PCE (c) CRPS

(d) STD

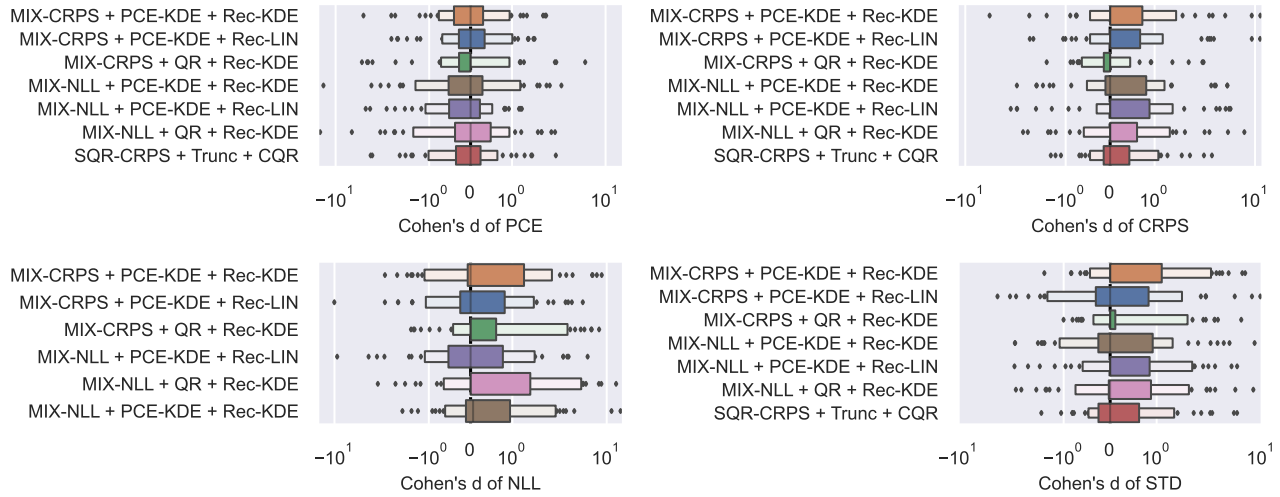
Figure 8: Comparison of different metrics where the base model is `SQR-CRPS`.

B.2. Combining Regularization and Post-hoc Methods

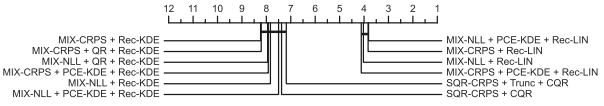
In this paper, we have established that post-hoc methods are generally more favorable than regularization methods when the primary objective is to enhance probabilistic calibration. Since regularization methods operate during training and do not alter the form of predictions (e.g., Gaussian mixture predictions), they can be easily combined with post-hoc methods. In this section, we address the question: "Which metrics do regularization methods improve when combined with a post-hoc method compared to the same model without regularization?"

To ensure clarity, we focus our presentation on a selection of paired regularization and post-hoc methods. Figure 9 illustrates the impact of regularization on various metrics for these pairs. In Figure 9(a), the baseline corresponds to the same post-hoc method without regularization, enabling a direct measurement of the effect of adding regularization to a post-hoc method. It is important to note that the boxplots in this figure cannot be directly compared due to the different baselines.

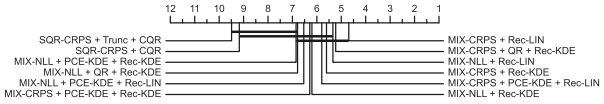
The critical difference diagrams provide a comparison of all methods, with and without regularization. Overall, when combined with post-hoc methods, regularization has a negative impact: no regularization method significantly improves probabilistic calibration, and they tend to negatively affect CRPS, NLL, and STD metrics.



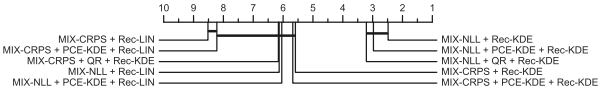
(a) Boxplots of Cohen's d of different metrics on all datasets, with respect to the same model except that regularization is not applied.



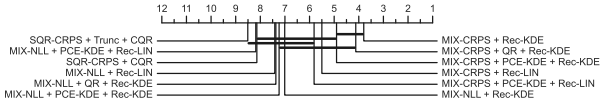
(b) PCE



(c) CRPS



(d) NLL



(e) STD

Figure 9: Comparison of different metrics showing the effect of regularization when combined with a post-hoc method, compared to the same model without regularization.

B.3. Post-hoc Calibration based on the Training Dataset

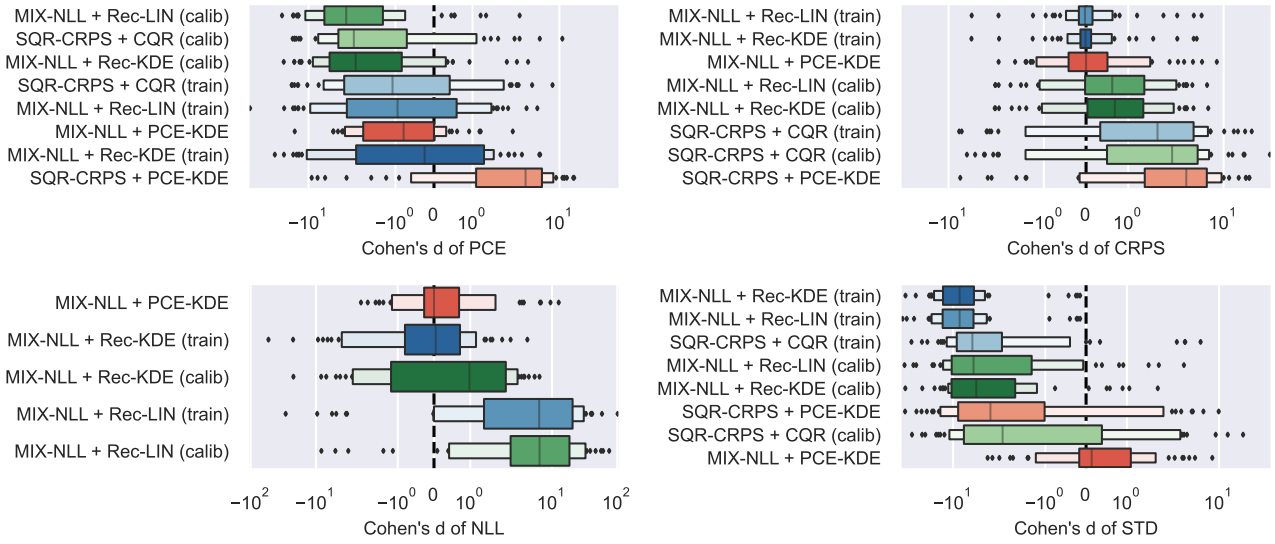
In this paper, the calibration map or conformity scores have been computed on a separate calibration dataset, following common practice in the literature. However, holding out data for post-hoc calibration reduces the quantity of training data. For the sake of clarity, we focus our analysis on the `MIX-NLL` and `SQR-CRPS` base losses

In this section, we compare post-hoc calibration based on the training dataset to post-hoc calibration based on the calibration dataset. We aim to answer the question: "Can it be beneficial to use post-hoc calibration based on the training dataset,

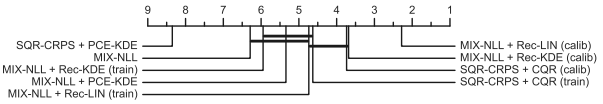
and should it be preferred over regularization methods when there is no calibration dataset available?" One advantage of regularization methods and post-hoc calibration methods based on the training dataset is that the base model can be trained on more data (80% in our experiments, compared to 65% when holding out the calibration dataset).

Figure 10 presents a comparison of different methods, with post-hoc methods trained on the calibration dataset indicated by `(calib)` and those trained on the training dataset indicated by `(train)`. We observe that post-hoc methods based on the calibration dataset tend to significantly outperform their counterparts based on the training dataset in terms of probabilistic calibration. Specifically, `MIX-NLL + Rec-LIN` and `MIX-NLL + Rec-KDE` achieve significantly better calibration when the calibration map is learned on the calibration dataset. Similarly, `SQR-CRPS + CQR` tends to improve calibration when conformal prediction is based on the calibration dataset. It is worth noting that even without a calibration dataset, post-hoc methods tend to be better calibrated than regularization methods.

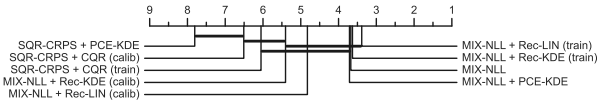
Finally, we observe that post-hoc methods based on the training dataset tend to achieve better CRPS and NLL scores, although not significantly. Additionally, they are also significantly sharper. This may be attributed to the larger training dataset available to the base model when there is no held-out dataset.



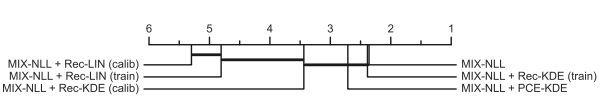
(a) Boxplots of Cohen's d of different metrics on all datasets, with respect to `MIX-NLL`.



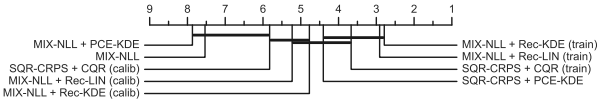
(b) PCE



(c) CRPS



(d) NLL



(e) STD

Figure 10: Comparison of different metrics.

B.4. Calibration of Vanilla Models

Figure 12 and Figure 13 provide additional results from our empirical study in Section 4, specifically focusing on the PCE obtained with `MIX-CRPS` and `SQR-CRPS`. The datasets are ordered in the same manner as shown in Figure 2 for comparison. We observe that `SQR-CRPS` is less calibrated compared to `MIX-NLL`.

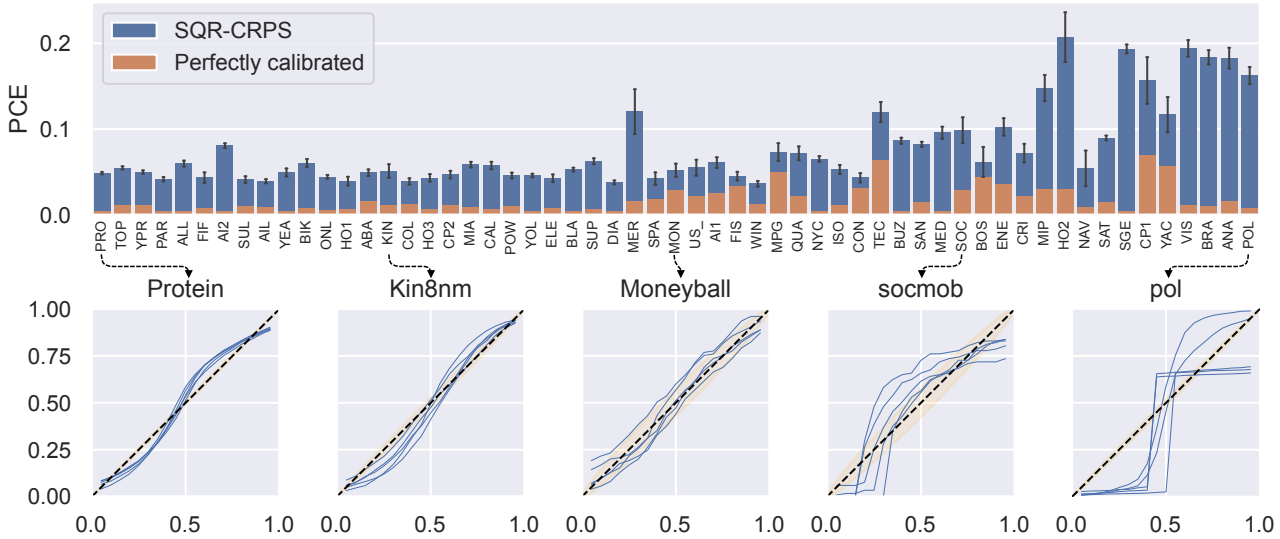


Figure 11: PCE obtained on different datasets, with examples of reliability diagrams. The height of each bar is the mean PCE of 5 runs with different dataset splits while the error bar represents the standard error of the mean. For 5 datasets, the PIT reliability diagrams of 5 runs are displayed in the bottom row.

Figure 12: PCE of `SQR-CRPS`, on all datasets.

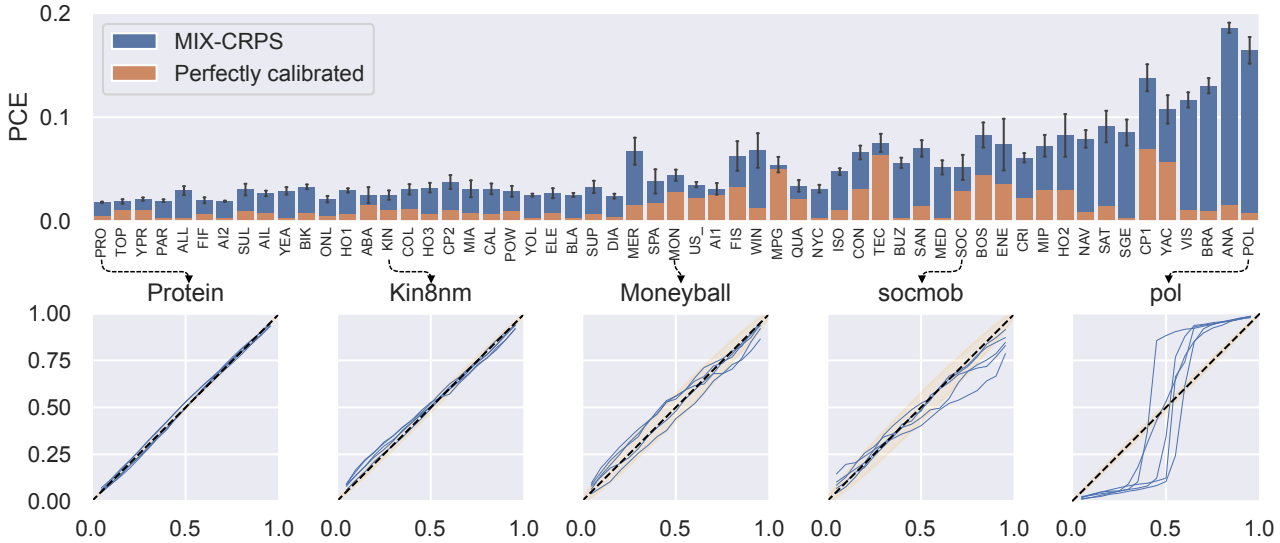


Figure 13: PCE of `MIX-CRPS`, on all datasets.

B.5. Distribution of the Test Statistic

Figure 14 shows the distribution of the test statistic, as described in Section 4. We observe that, in a lot of cases, the average PCE of the compared models is larger than all the 10^4 samples of the average PCE from a probabilistically calibrated model. Among the different calibration methods, post-hoc calibration with `MIX-NLL + Rec-EMP` achieves the highest level of calibration performance in the majority of cases.

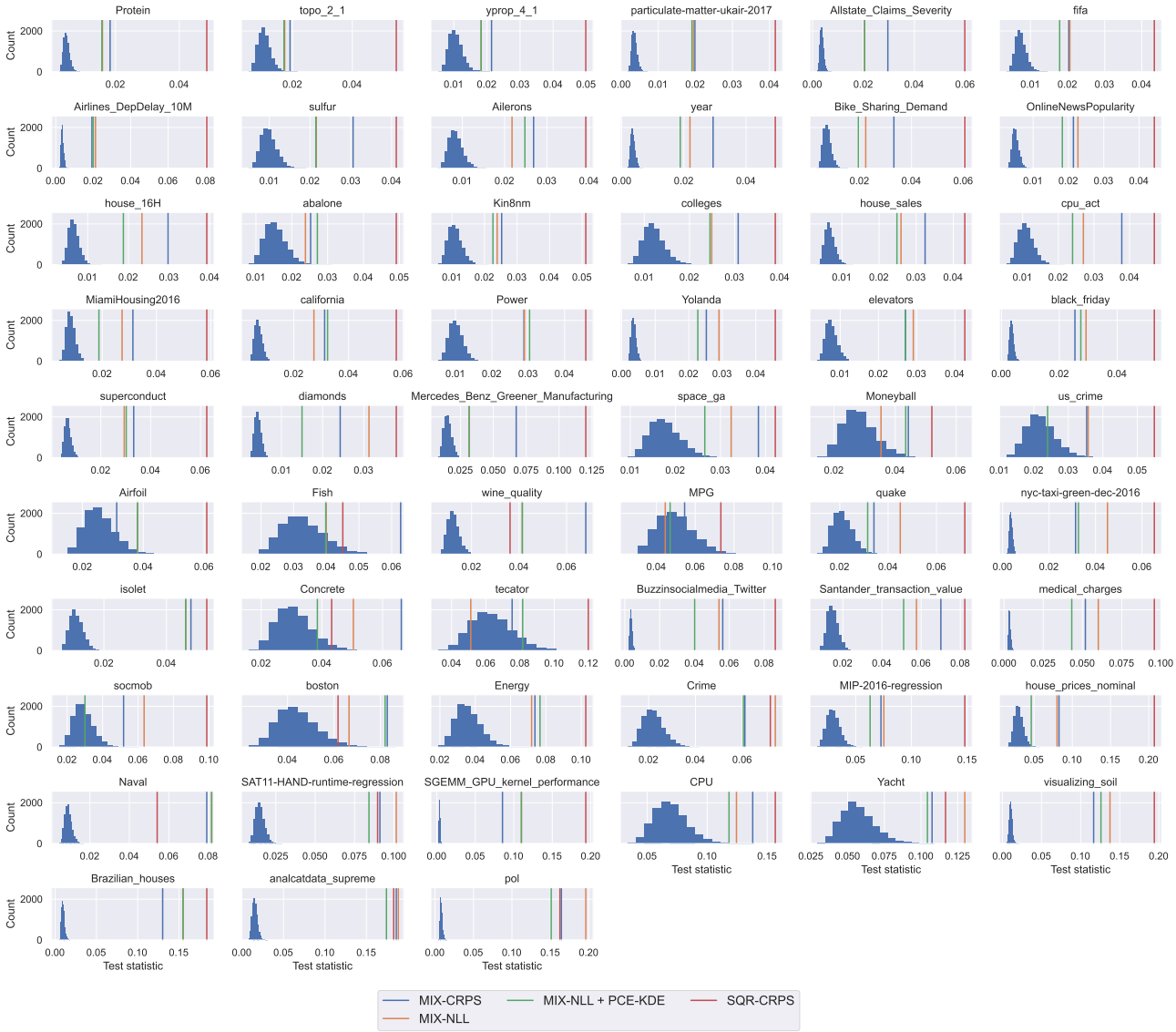


Figure 14: Distribution of the test statistic on all datasets for different models.

B.6. Reliability Diagrams

Figure 15 and Figure 16 compare reliability diagrams obtained on models with and without post-hoc calibration, respectively. With only a few exceptions, the post-hoc calibrated models exhibit a visual proximity to the diagonal line.

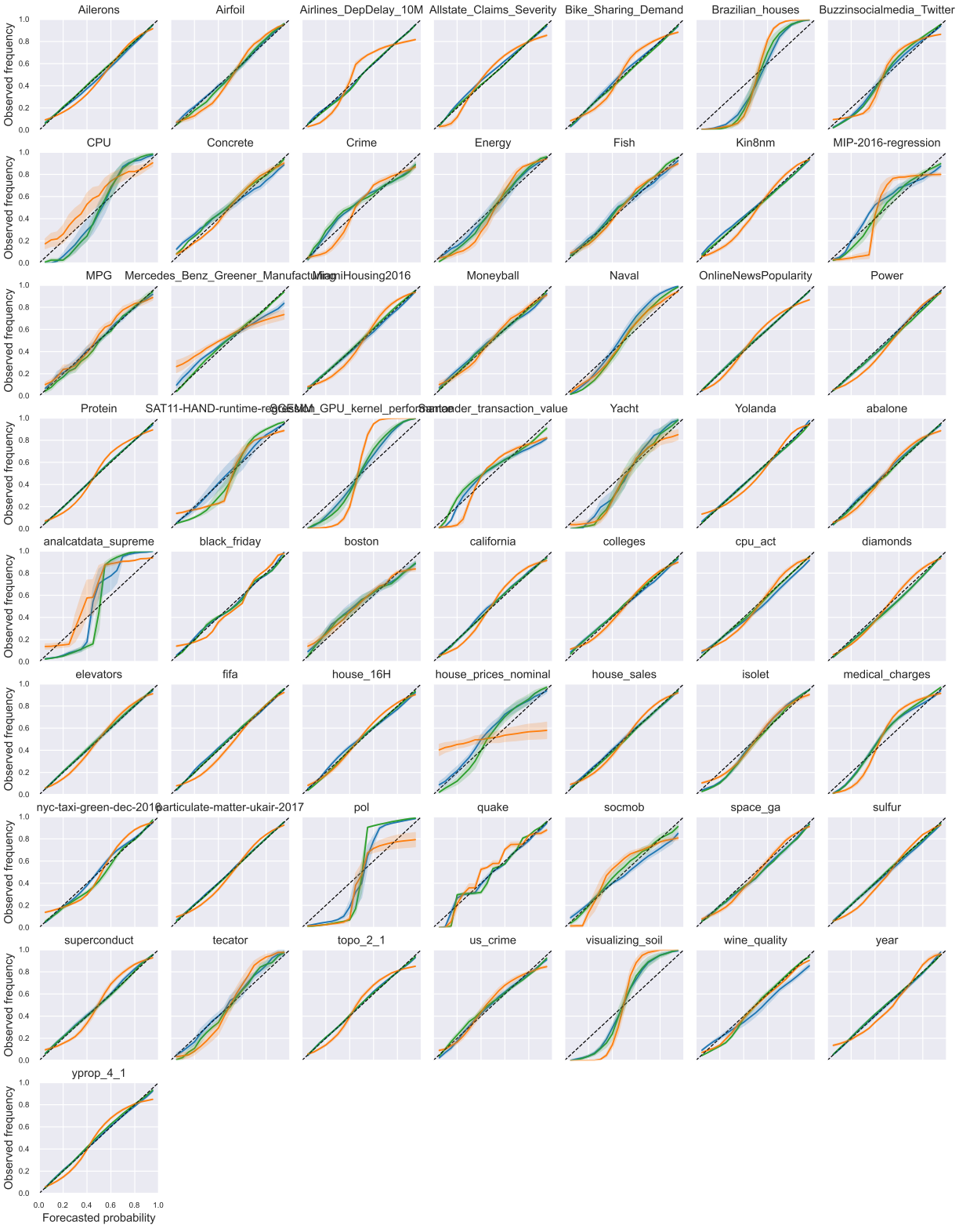


Figure 15: Reliability diagrams on all datasets for different models.

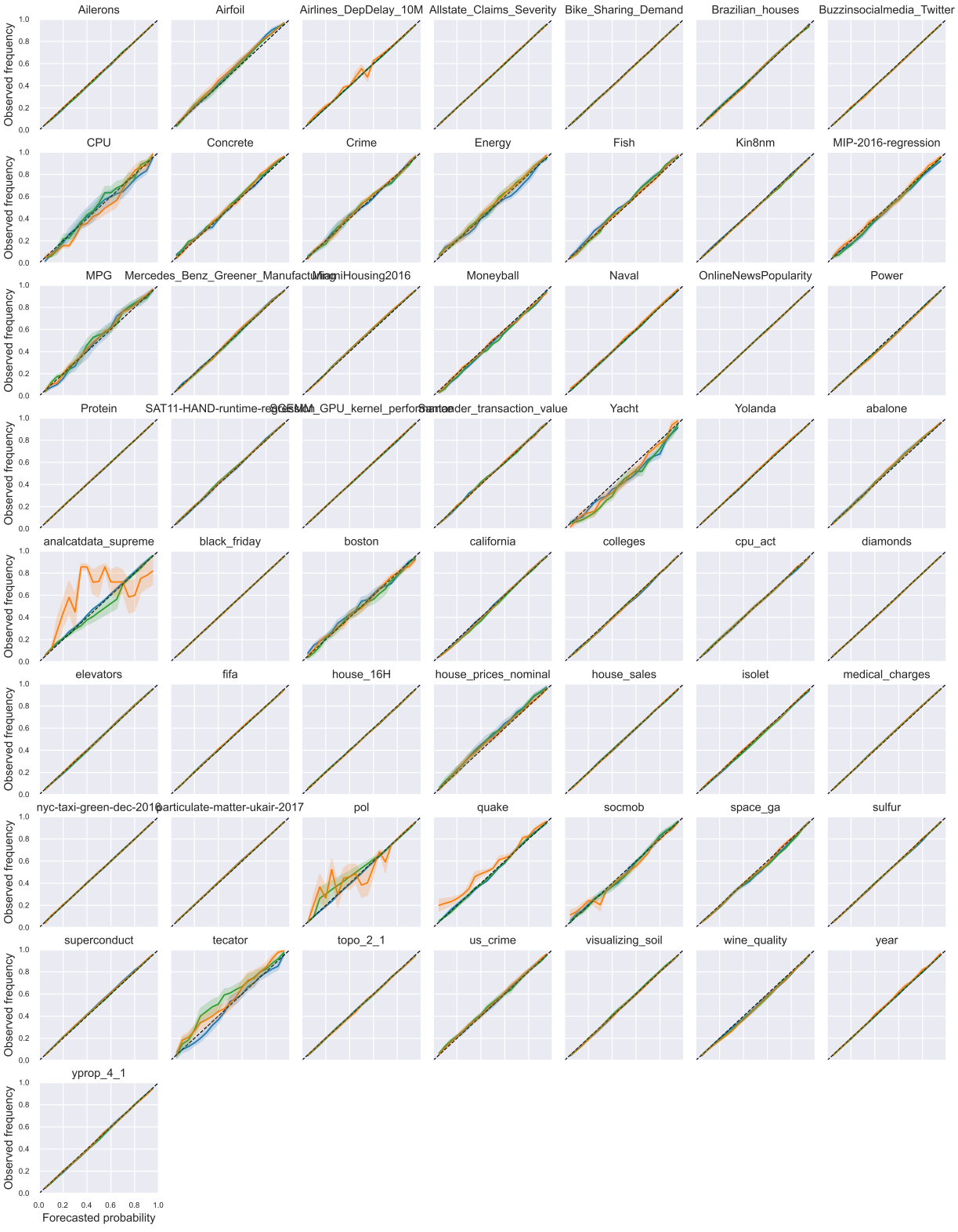


Figure 16: Reliability diagrams on all datasets for different models with post-hoc calibration.

C. Hyperparameters

In our experiments, we adopt a specific architecture consisting of 3 hidden layers with 100 units per layer, ReLU non-linearities, and a dropout rate of 0.2 on the last hidden layer. Early stopping with a patience of 30 is applied to select the epoch with the lowest base loss on the validation dataset.

In this section, we delve into the performance of different model parameters, including the number of components in Gaussian mixture predictions, the number of quantiles in quantile predictions, and the number of hidden layers in the underlying models.

Figure 17 compares models that predict mixtures with varying numbers of components compared to the reference of 3 components. Notably, when there is only 1 component (yielding a single Gaussian prediction), the model's performance significantly deteriorates in terms of CRPS, NLL, and sharpness. However, as the number of components increases beyond 3, the differences become less pronounced.

Figure 18 compares models with different numbers of quantiles compared to a reference of 64 quantiles. The results reveal a consistent pattern: predicting more quantiles consistently enhances performance in terms of probabilistic calibration, CRPS, and sharpness.

Figure 19 compares models with different numbers of layers relative to a 3-layer model. It highlights that models with 2, 3, or 5 layers tend to yield superior performance in terms of CRPS and NLL.

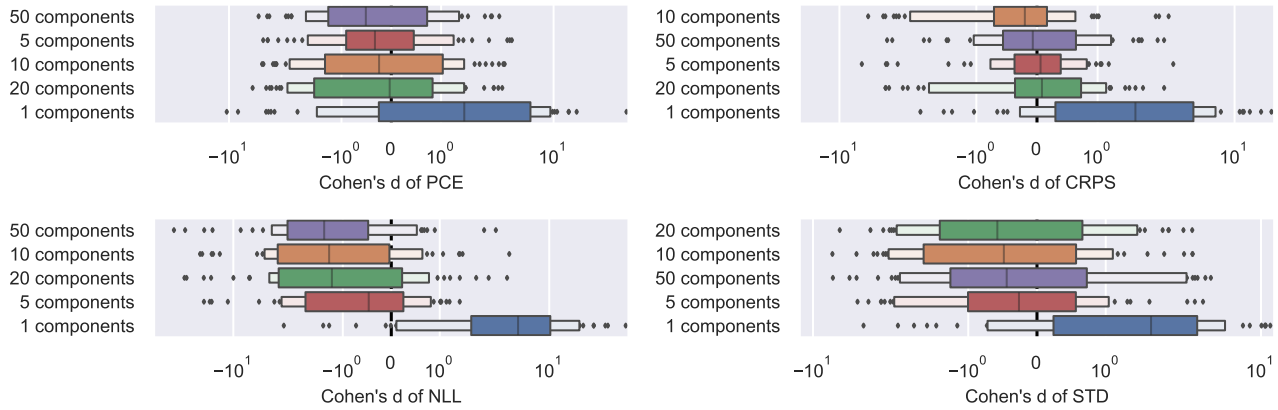


Figure 17: Comparison of models whose predictions are Gaussian mixtures with different numbers of components. All models are trained with NLL loss, without regularization or post-hoc method. The box plots show Cohen's d of different metrics on all datasets. Cohen's d is computed with respect to a model whose predictions are Gaussian mixtures with 3 components.

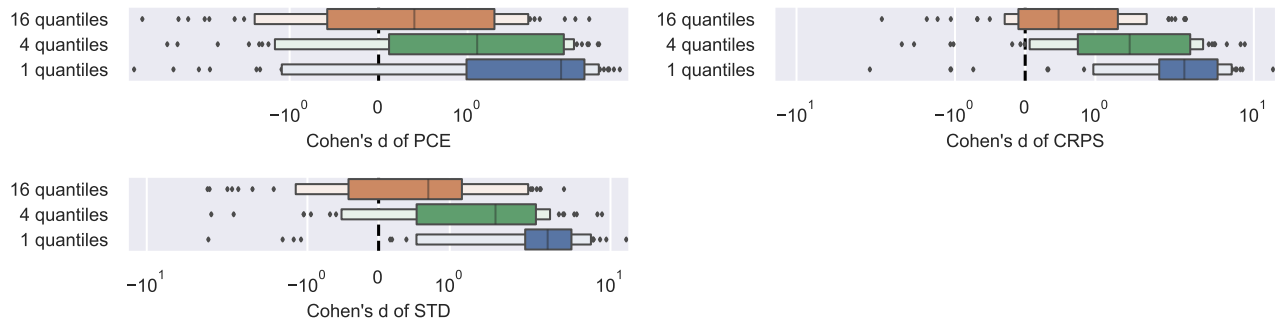


Figure 18: Comparison of models whose predictions are different numbers of quantiles. All models are trained with CRPS loss, without regularization or post-hoc method. The box plots show Cohen's d of different metrics on all datasets. Cohen's d is computed with respect to a model whose predictions are 64 quantiles.

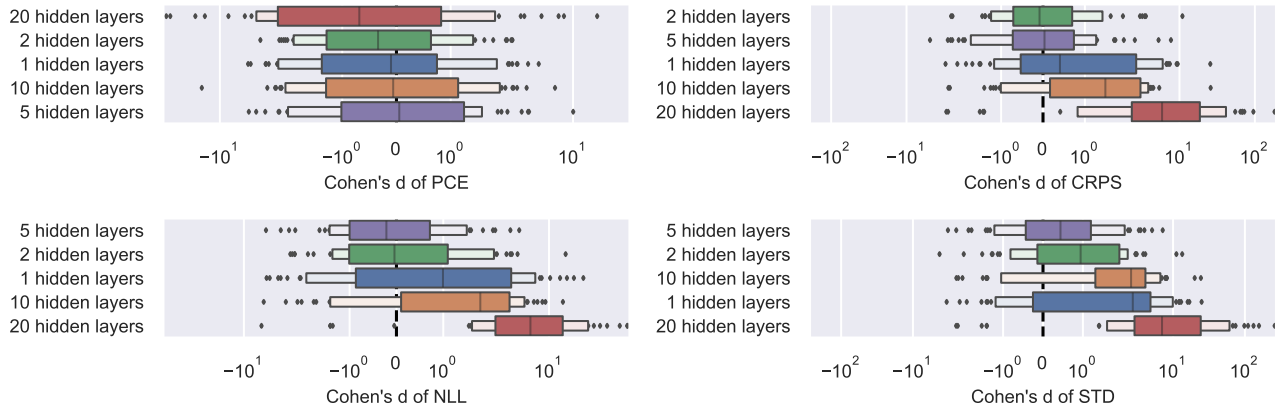


Figure 19: Comparison of models with different number of layers. All models predict Gaussian mixtures and are trained with NLL loss, without regularization or post-hoc method. The box plots show Cohen's d of different metrics on all datasets. Cohen's d is computed with respect to a model with 3 hidden layers.

D. Tabular Regression Datasets

Table 1 presents the datasets considered in our experiments. To ensure consistency, when datasets are available from multiple sources, we select one specific source per dataset, as indicated in Figure 1. Our selection prioritizes the suites 297, 299, and 269 of OpenML, followed by UCI datasets.

In the OpenML suite 297, we discovered that the datasets `houses` and `california` are identical, and thus, we only included the `california` dataset in our analysis. Moreover, the UCI archive for the dataset `wine_quality` contains two separate datasets for red and white wine. As there was no indication regarding the specific dataset(s) used in previous studies, we followed the approach of Grinsztajn et al. (2022) and solely considered the dataset related to white wine. In Figure 1, other studies may have employed the alternative dataset or a combination of both datasets.

Table 1: Datasets

Group	Dataset	Abbrev.	Nb of training instances	Nb of features
UCI	CPU	CP1	135	7
	Yacht	YAC	200	6
	MPG	MPG	254	7
	Energy	ENE	499	9
	Crime	CRI	531	104
	Fish	FIS	590	6
	Concrete	CON	669	8
	Airfoil	AI1	976	5
	Kin8nm	KIN	5324	8
	Power	POW	6219	4
	Naval	NAV	7757	17
	Protein	PRO	29724	9
OpenML 297	wine_quality	WIN	4223	11
	isolet	ISO	5068	613
	cpu_act	CP2	5324	21
	sulfur	SUL	6552	6
	Brazilian_houses	BRA	6949	8
	Ailerons	AIL	8937	33
	MiamiHousing2016	MIA	9055	13
	pol	POL	9750	26
	elevators	ELE	10789	16
	Bike_Sharing_Demand	BIK	11296	6
	fifa	FIF	11740	5
	california	CAL	13416	8
	superconduct	SUP	13820	79
	house_sales	HO3	14048	15
	house_16H	HO1	14809	16
	diamonds	DIA	35061	6
	medical_charges	MED	50000	3
	year	YEA	50000	90
	nyc-taxi-green-dec-2016	NYC	50000	9
OpenML 299	analcatdata_supreme	ANA	2633	12
	Mercedes_Benz _Greener_Manufacturing	MER	2735	735
	visualizing_soil	VIS	5616	5
	yprop_4_1	YPR	5775	82
	OnlineNewsPopularity	ONL	25768	73
	black_friday	BLA	50000	23
	SGEMM_GPU _kernel_performance	SGE	50000	15
	particulate-matter -ukair-2017	PAR	50000	26
OpenML 269	tecator	TEC	156	124
	boston	BOS	328	22
	MIP-2016-regression	MIP	708	111
	socmob	SOC	751	39
	Moneyball	MON	800	18
	house_prices_nominal	HO2	711	234
	us_crime	US_	1295	101
	quake	QUA	1415	3
	space_ga	SPA	2019	6
	abalone	ABA	2715	10
	SAT11-HAND- runtime-regression	SAT	2886	118
	Santander_transaction _value	SAN	2898	3611
	colleges	COL	4351	34
	topo_2_1	TOP	5775	252
	Allstate_Claims_Severity	ALL	50000	477
	Yolanda	YOL	50000	100
	Buzzinsocialmedia_Twitter	BUZ	50000	70
	Airlines_DepDelay_10M	AI2	50000	5