Better Uncertainty Calibration via Proper Scores for Classification and Beyond

Anonymous Author(s)
Affiliation
Address
email

Abstract

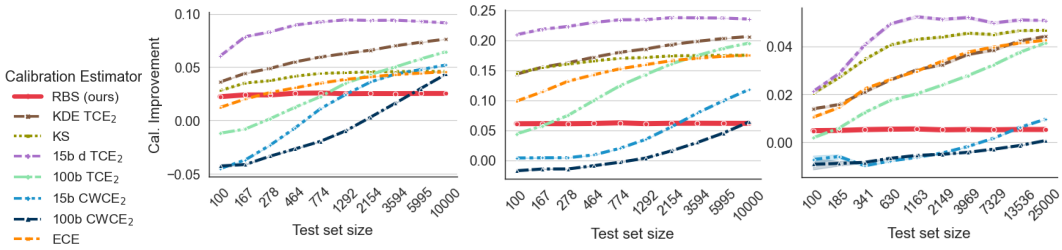
With model trustworthiness being crucial for sensitive real-world applications, practitioners are putting more and more focus on improving the uncertainty calibration of deep neural networks. Calibration errors are designed to quantify the reliability of probabilistic predictions but their estimators are usually biased and inconsistent. In this work, we introduce the framework of *proper calibration errors*, which relates every calibration error to a proper score and provides a respective upper bound with optimal estimation properties. This relationship can be used to reliably quantify the model calibration improvement. We theoretically and empirically demonstrate the shortcomings of commonly used estimators compared to our approach. Due to the wide applicability of proper scores, this gives a natural extension of recalibration beyond classification.

1 Introduction

Deep learning became a dominant cornerstone of machine learning research in the last decade and deep neural networks can surpass human-level predictive performance on a wide range of tasks [19, 46, 17]. However, Guo et al. [14] have shown that for modern neural networks, better classification accuracy can come at the cost of systematic overconfidence in their predictions. Practitioners in sensitive forecasting domains, such as cancer diagnostics [17], genotype-based disease prediction [25] or climate prediction [57], require for models to not only have high predictive power but also to reliably communicate uncertainty. This raises the need to quantify and improve the quality of predictive uncertainty, ideally via a dedicated metric. An uncertainty-aware model should give probabilistic predictions which represent the true likelihood of events depending on the very prediction. To quantify the extend to which this condition is violated, calibration errors have been introduced. In general, their estimators are usually biased [45] and inconsistent [52]. This, in turn, is highly problematic since we cannot quantify how reliable a model is if we do not know how reliable the metric is. Especially the medical field is a domain that requires high model trustworthiness, but with low expert availability and/or disease frequency we often encounter a small data regime. Resampling strategies can be viable options for optimization on small datasets but also reduce the evaluation set size even more. We will discover that little data exacerbates the estimation bias and propose reliable alternatives for improving uncertainty calibration.

Since deep neural networks often yield uncalibrated confidence scores [36], a variety of different post-hoc recalibration approaches have been proposed [7, 31]. These methods use the validation set to transform predictions returned by a trained neural network such that they become better calibrated. A key desired property of recalibration methods is to not reduce the accuracy after the transformation. Therefore, most modern approaches are restricted to accuracy preserving transformations of the model outputs [42, 14, 61, 44]. When recalibrating a model, it is crucial to have a reliable estimate of how much the chosen method improves the underlying model. However, when using current

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) TS of DenseNet 40 on CIFAR10

(b) ETS of Wide ResNet 32 on CIFAR100

(c) DIAG of DenseNet 161 on ImageNet

Figure 1: Estimated calibration improvement for various settings. The calibration error is estimated before and after a recalibration method (TS / ETS / DIAG) is applied and the difference (i.e. calibration improvement) is shown for increasing test set size. All common calibration errors are sensitive with respect to the test set size and can substantially over- or underestimate the effect of performing recalibration. Only RBS robustly estimates the improvement in calibration error for all test set sizes.

estimators for calibration errors, their biased nature results in estimates that are highly sensitive to the number of samples in the test set that are used to compute the calibration error before and after recalibration (Fig. 1; c.f. section 5).

Our **contributions** for better uncertainty calibration are summarized in the following. We...

- ... give an overview of current calibration error literature, place the errors into a taxonomy, and show which are insufficient for calibration quantification. This also includes several theoretical results, which highlight the shortcomings of current approaches.
- ... introduce the framework of **proper calibration errors**, which gives important guarantees and relates every element to a proper score. We can reliably estimate the improvement of an injective recalibration method w.r.t. a proper calibration error via its related proper score - even in non-classification settings.
- ... show that common calibration estimators are highly sensitive w.r.t. the test set size. We demonstrate that for commonly used estimators, the estimated improvement of recalibration methods is heavily biased and becomes monotonically worse with fewer test data.

2 Related Work

In this section, we give an extensive overview of published work regarding quantifying model calibration and model recalibration. Important definitions will be directly given, while others are placed in Appendix C. These will be the basis for our theoretical findings. Further, we will motivate the definition of proper calibration errors, which are directly related to proper scores. Consequently, we will also present important aspects of the framework around proper scores in later parts of this section.

2.1 Calibration errors

For clarity, we introduce shortly the notation used throughout this work. Assume we have random variables X and Y corresponding to feature and target variable, and feature and target space $\mathscr{X}$ and $\mathscr{Y}$. We have $\mathbb{P}_Y, \mathbb{P}_{Y|X} \in \mathscr{P}$, where $\mathbb{P}_Y$ refers to the distribution of Y, $\mathbb{P}_{Y|X}$ to the conditional distribution given X, and $\mathscr{P}$ a set of distributions on $\mathscr{Y}$. Even though some approaches explore calibration for regression tasks [47, 56, 6], it is most dominantly considered for classification. To distinguish between the general case and n-class classification, we refer to $\mathscr{P}_n$ as the n-dimensional simplex of corresponding categorical distributions.

A popular task is the calibration of the predicted top-label $C = \arg\max_k f_k(X)$ of a model $f\colon \mathscr{X} \to \mathscr{P}_n$ [14, 30, 34, 41, 52, 50, 44]. Here, the top-label confidence should represent the accuracy of this very prediction. Formally, we try to reach the condition $f_C(X) = \mathbb{P}(Y = C \mid f_C(X))$. However, the condition is weaker as one might expect, and referring to a model fulfilling this condition as

(perfectly) calibrated can give a false sense of security [52, 55]. This holds especially in forecasting domains, where low probability estimates can still be highly relevant. For example, assigning probability mass to an aggressive type of cancer can still trigger action even if it is not predicted as the most likely outcome. Further, we might also be interested in predictive uncertainty for non-classification tasks. Consequently, we use the stricter and more general condition that the complete prediction $f(X)$ should be equal to the complete conditional distribution of the target given this very prediction $\mathbb{P}_{Y|f(X)}$ as introduced by Widmann et al. [56]. More formally, we state:

Definition 2.1. A model $f\colon \mathscr{X} \to \mathscr{P}$ is **calibrated** if and only if $f(X) = \mathbb{P}_{Y|f(X)}$.

Note that $\mathscr{P}$ can be a set of arbitrary distributions, which incorporates $\mathscr{P}_n$ (classification) as a special case.

One of the first metrics for assessing model calibration that is still widely used in recent literature is the Brier score (BS) [38, 42, 16, 44]. For a model $f\colon \mathscr{X} \to \mathscr{P}_n$ the **Brier score** [3] is defined as

$$\mathrm{BS}(f) = \mathbb{E}\left[\left\|f(X) - Y'\right\|_2^2\right], \tag{1}$$

where Y' is one-hot-encoded Y. The estimator of the BS is equivalent to the mean squared error, illustrating that it does not purely capture model calibration. Rather, the BS can be interpreted as a comprehensive measure of model performance, simultaneously capturing model fit and calibration. This becomes more obvious via the canonical decomposition of the BS into a calibration and sharpness term [38]. Based on this decomposition, we can derive the following error. For $1 \leq p \in \mathbb{R}$, the ℓ_p **calibration error** (CE_p) of model $f\colon \mathscr{X} \to \mathscr{P}_n$ is defined as

$$\mathrm{CE}_p(f) = \mathbb{E}\left[\left\|f(X) - \mathbb{P}_{Y|f(X)}\right\|_p^p\right]. \tag{2}$$

The BS decomposition only supports the case $p = 2$, but a general ℓ_p formulation became more common in recent years [52, 55, 61]. The CE_p is intractable to reliably estimate in the general case. This is mostly due to the term $\mathbb{P}_{Y|f(X)}$ since we never have samples of every possible prediction for continuous models. This is in contrast to the original work of Murphy [38], where only models with a finite prediction space are considered so that computing the CE_2 does not require a special procedure. To assess the calibration of a continuous binary model Platt [42] used histogram estimation, transforming the infinite prediction space to a finite one. This is also referred to as equal width binning. Similarly, Nguyen & O'Connor [40] introduced an equal mass binning scheme for continuous binary models. Both, equal width and equal mass binning schemes, suffer from the requirement of setting a hyperparameter. This can significantly influence the estimated value [31] and there is no optimal default since every setting has a different bias-variance tradeoff [41]. The first calibration estimator for a continuous one-vs-all multi-class model was given by Naeini et al. [39] and is still the most commonly used measure to quantify calibration. It is referred to as the **expected calibration error** (ECE) of model $f\colon \mathscr{X} \to \mathscr{P}_n$ and defined as

$$\mathrm{ECE}(f) = \sum_{i=1}^{m} p_i \left|\mathrm{conf}_i - \mathrm{acc}_i\right| \tag{3}$$

with the bin frequency $p_i = \mathbb{P}(f_C(X) \in B_i)$, the bin-wise mean confidence $\mathrm{conf}_i = \mathbb{E}[f_C(X) \mid f_C(X) \in B_i]$, and the bin-wise accuracy $\mathrm{acc}_i = \mathbb{P}(Y = C \mid f_C(X) \in B_i)$, for $m \in \mathbb{N}$ bins $B_i \coloneqq \left(\frac{i-1}{m}, \frac{i}{m}\right]$. We can reliably estimate p_i, conf_i, and acc_i via the respective means. These are then 'plugged-in' to estimate the ECE.

Kull et al. [29] and Nixon et al. [41] independently introduced another calibration estimator, which also captures the extend to which the condition $\mathbb{P}(Y = k \mid f_k(X)) = f_k(X)$ is violated for each class $k \in \mathscr{Y}$. They respectively use equal width and equal mass binning. Similar to these estimators, Kumar et al. [31] introduced the **class-wise calibration error** (CWCE_p) and, similar to the ECE, the **top-label calibration error** (TCE_p). They only formulated the squared case $p = 2$ but suggested the extension of their definitions to general p-norms, which we provide in Appendix C.

Furthermore, Kumar et al. [31] and Vaicenavicius et al. [52] proved independently that using a fixed binning scheme for estimation leads to a lower bound of the expected error. Zhang et al. [61] circumvent binning schemes by using kernel density estimation to estimate the TCE_p.

Orthogonal ways to quantify model miscalibration have been proposed to not depend on binning or kernel density estimation schemes. Gupta et al. [16] introduced the **Kolmogorov-Smirnov**

calibration error (KS) (c.f. Appendix C), which is based on the Kolmogorov-Smirnov test between empirical cumulative distribution functions. Its estimator does not require setting a hyperparameter but the authors did not provide further theoretical aspects.

Estimators of the TCE_p and CWCE_p are in general not differentiable. Consequently, Kumar et al. [30] proposed the **Maximum mean calibration error** (MMCE) (c.f. Appendix C), which has a differentiable estimator. It is a kernel-based error, comparing the top-label confidence and the conditional accuracy, similar to the ECE.

Widmann et al. [55] argued that the MMCE is insufficient for quantifying calibration of a model, similar as the ECE. They further proposed the **Kernel calibration error** (KCE) (c.f. Appendix C). It is based on matrix-valued kernels and unlike the MMCE, which only uses the top-label prediction, includes the whole model prediction. The squared KCE has an unbiased estimator based on a U-statistic with quadratic runtime complexity with respect to the data size. However, even though the KCE is positive, the U-statistic estimator can give negative values. To this end, the authors use the estimated value as a test statistic w.r.t. the null hypothesis that the model is calibrated.

Furthermore, Widmann et al. [55] and Widmann et al. [56] proposed to unify different definitions of calibration errors in a theoretical framework, which also includes variance regression calibration. However, the framework allows calibration errors, which are zero even if the model is not calibrated at all.

2.2 Recalibration

A plethora of recalibration methods have been proposed to improve model calibration after training by transforming the model output probabilities [18, 42, 58, 59, 39, 40, 28, 14, 29, 31, 61, 16, 44]. These methods are optimized on a specific calibration set, which is usually the validation set. Key desiderata of these methods include for the algorithms to be accuracy-preserving and data-efficient [61], reflecting that typical use-cases include settings in sensitive domains where accuracy should remain unchanged and often little data is available to train and evaluate the models. Such accuracy-preserving methods only adjust the probability estimate in such a way that the predicted top-label remains the same. The most commonly used accuracy-preserving recalibration method is temperature scaling (TS) [14], where the model logits are divided by a single parameter $T \in \mathbb{R}_{>0}$ before computing the predictions via softmax. A more expressive extension of TS is ensemble temperature scaling (ETS) [61], where a weighted ensemble of TS output, model output, and label smoothing is computed. Rahimi et al. [44] proposed different classes of order-preserving transformations. A specifically interesting one is the class of diagonal intra order-preserving functions (DIAG). Here, the model logits are transformed elementwise with a scalar, monotonic, and continuous function, which is represented by neural networks of unconstrained monotonic functions [54].

2.3 Proper scores

Gneiting & Raftery [13] give an extensive overview of proper scores. Unfortunately, their presented definitions assume maximization as the model training objective. To stay in line with recent machine learning literature, we flip the sign when it is required in the following definitions, similar as in [4]. We specifically do not constraint ourselves to classification, which is a special case. Assume we give a prediction in $\mathscr{P}$ for an event in $\mathscr{Y}$ and we want to score how good the prediction was. A function $S : \mathscr{P} \times \mathscr{Y} \to \overline{\mathbb{R}}$ with $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ is called **scoring rule** or just **score**. Examples are the Brier score and the log score for classification, or the Dawid-Sebastiani score (DSS), which extends the MSE for variance regression [8, 12]. It is defined as $S(P, y) = \frac{(\mu_P - y)^2}{\sigma_P^2} + \log \sigma_P^2$. To compare distributions, we use the expected score $s_S \colon \mathscr{P} \times \mathscr{P} \to \mathbb{R}$ defined as $s_S(P, Q) = \mathbb{E}_{Y \sim Q}[S(P, Y)]$. A scoring rule S is defined to be **proper** if and only if $s_S(P, Q) \geq s_S(Q, Q)$ holds for all $P, Q \in \mathscr{P}$, and **strictly proper** if and only if $P \neq Q \implies s_S(P, Q) > s_S(Q, Q)$. In other words, a score is proper if predicting the target distribution gives the best expected value and strictly proper if no other prediction can achieve this value. Given a proper score S and $P, Q \in \mathscr{P}$, the associated **divergence** $d_S \colon \mathscr{P} \times \mathscr{P} \to \mathbb{R}_{\geq 0}$ is defined as $d_S(P, Q) = s_S(P, Q) - s_S(Q, Q)$ and the associated **generalized entropy** $g_S \colon \mathscr{P} \to \mathbb{R}$ as $g_S(Q) = s_S(Q, Q)$. For strictly proper S, d_S is only zero if $P = Q$; for (strictly) proper S, g_S is (strictly) concave. An example of a divergence-entropy combination is the Kullback-Leibler divergence and the Shannon entropy associated to the log score.

Associated entropies and divergences are used in the calibration-sharpness decomposition introduced by Bröcker [4] for proper scores of categorical distributions. In Lemma 4.1 we will prove that this result holds for proper scores of arbitrary distributions, as long as additional conditions are met. Further, associated divergences will be the backbone for our definition of *proper calibration errors* in Section 4.

3 Theoretical analysis of calibration errors

In this section, we present theoretical results regarding calibration errors used in current literature. First, we place these calibration errors into a taxonomy, which we introduce in Theorem 3.1. Next, we give an example of how much errors lower in the hierarchy can differ from CE_2 in Proposition 3.2. We illustrate the practical relevance by introducing a perfect calibration method according to most errors in Proposition D.2, which fails to be calibrated. Last, we analyse the ECE estimate with respect to the data size in Proposition 3.3. This proposition is the basis for explaining the empirical (miss-)behaviour of the ECE observed in Section 5. All proofs are presented in Appendix D.

To the best of the authors' knowledge, other publications provided relations between at most two distinct calibration errors or none at all while introducing a new one. The taxonomy in the following theorem is a novel contribution, improving overview of a whole body of work regarding quantifying calibration.

Theorem 3.1. *Given a model $f : \mathscr{X} \to \mathscr{P}_n$ and $1 \leq p \in \mathbb{R}$, we have*

$$BS\left(f\right) = 0 \Longrightarrow \begin{Bmatrix} CE_p\left(f\right) = 0 \\ KCE\left(f\right) = 0 \\ f \text{ calibrated} \end{Bmatrix} \Longrightarrow CWCE_p\left(f\right) = 0 \Longrightarrow \begin{Bmatrix} TCE_p\left(f\right) = 0 \\ MMCE\left(f\right) = 0 \\ KS\left(f\right) = 0 \end{Bmatrix} \Longrightarrow ECE\left(f\right) = 0$$

where statements inside curly brackets $\{\ldots\}$ are equivalent. Further, we have

$$2^{\frac{1}{p}} \geq \left[BS\left(f\right)\right]^{\frac{1}{p}} \overset{*}{\geq} \left[CE_p\left(f\right)\right]^{\frac{1}{p}} \geq CWCE_p\left(f\right) \geq TCE_p\left(f\right) \geq TCE_1\left(f\right) \geq \begin{cases} KS\left(f\right) \\ ECE\left(f\right) \\ c \cdot MMCE\left(f\right) \end{cases} \geq 0$$

*where * BS is only included for $p = 2$. The kernel dependent constant $c \in \mathbb{R}$ is given in Appendix D.2 according to Kumar et al. [30].*

From this theorem follows that it is ambiguous to refer to *perfect calibration* just because there exists a calibration error which is zero for a model. Now, we further demonstrate how large the gap between calibration errors can be.

Proposition 3.2. *Assume the model $f\colon \mathscr{X} \to \mathscr{P}_n$ is surjective. There exists a joint distribution $\mathbb{P}_{X,Y}$ such that for all $E \in \{MMCE, KS, ECE, TCE_p, CWCE_p \mid 1 \leq p \in \mathbb{R}\}$:*

$$E\left(f\right) = 0 \quad \textit{and} \quad CE_2\left(f\right) \geq 0.99 - \frac{1}{n}.$$

Recall that $\mathrm{CE}_2\left(f\right) = 0$ if and only if f is calibrated. In other words, most used calibration errors can be zero, but the model is still far from calibrated. An example of a model transformation with perfect ECE but uncalibrated outputs is given in Proposition D.2.

Among all calibration errors, the ECE is still the most commonly used [24, 26, 44, 50, 36, 49, 23, 35, 37, 15, 53, 10], even though its estimator is knowingly biased [31, 45]. Let $\hat{\mathrm{ECE}}_{(n)}$ denote the ECE estimator for n data instances as originally defined in Guo et al. [14]. The following gives an analysis how this estimate behaves approximately with respect to n and how this is further impacted by the ground truth ECE value. For simplicity, we omit the fixed model from the notation.

Proposition 3.3. *For $\mu_{(n)}$ defined in Appendix D.5, we have*

$$\mathbb{E}\left[\hat{ECE}_{(n)}\right] \approx \mu_{(n)} \geq ECE \quad , \quad \frac{\mathrm{d}\mu_{(n)}}{\mathrm{d}n} < 0 \quad , \quad \frac{\mathrm{d}^2\mu_{(n)}}{\left(\mathrm{d}n\right)^2} > 0 \quad \textit{and} \quad \frac{\mathrm{d}^2\mu_{(n)}}{\mathrm{d}n\ \mathrm{d}ECE} > 0.$$

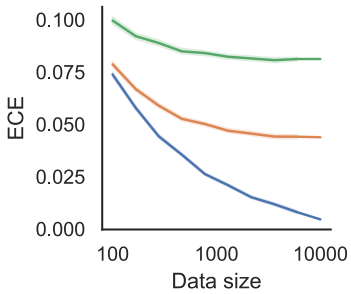


Figure 2: Estimated ECE of simulated models with perfect calibration (**blue**), mediocre calibration (**orange**), and bad calibration (**green**). Better calibration exacerbates ECE bias with respect to the data size, leading to unreliably calibration improvement quantification.

In words, the ECE estimator can be approximated by a function, which is strictly convex and monotonically decreasing in the data size. The smaller the data size, the larger the change of the function. Further, this change is also influenced by the true ECE value, such that, for low ground truth ECE, the function changes even more rapidly with the data size. Analogous results can be found for CWCE_1, CWCE_2, and TCE_2 with binning estimators as used in Kull et al. [29], Nixon et al. [41] and Kumar et al. [31].

To confirm the goodness of the approximation, we evaluated the ECE estimator on simulated models in Figure 2. The models are designed such that their true level of calibration is known. Including the results of Figure 1, the empirical curves are consistent with Proposition 3.3. Further details on the simulation are given in Appendix G.

The results in this section motivate a formal framework of proper calibration errors which are zero if and only if the model is calibrated and with robust estimators.

4 Proper calibration errors

In this section, we introduce the definition of *proper calibration errors*. We provide an easy-to-estimate upper bound and investigate some properties. As a preliminary step, we generalize a proper score decomposition. Again, all proofs are presented in Appendix D.

Bröcker [4] introduced a calibration-sharpness decomposition of proper scores w.r.t. categorical distributions. We extend this decomposition to proper scores of arbitrary distributions.

Lemma 4.1. *Let $\mathscr{P}$ be a set of arbitrary distributions for which exists a proper score S with some mild conditions. For random variables Q and Y with $Q, \mathbb{P}_Y, \mathbb{P}_{Y|Q} \in \mathscr{P}$, we have the decomposition*

$$\mathbb{E}\left[S\left(Q,Y\right)\right] = \underbrace{g_S\left(\mathbb{P}_Y\right)}_{\textit{generalized entropy}} - \underbrace{\mathbb{E}\left[d_S\left(\mathbb{P}_Y, \mathbb{P}_{Y|Q}\right)\right]}_{\textit{sharpness}} + \underbrace{\mathbb{E}\left[d_S\left(Q, \mathbb{P}_{Y|Q}\right)\right]}_{\textit{calibration}}.$$

Substituting Q with $f\left(X\right)$ and S with the Brier score, the calibration term equals the previously defined CE_2 of a model f. Lemma 4.1 motivates the following definition, which we introduce:

Definition 4.2. Given a model $f\colon \mathscr{X} \to \mathscr{P}$, we say

$$\text{CE}_S\left(f\right) := \mathbb{E}\left[d_S\left(f\left(X\right), \mathbb{P}_{Y|f(X)}\right)\right]$$

is a **(strictly) proper calibration error** if and only if d_S is a divergence associated with a (strictly) proper score S.

This gives $\text{CE}_{\text{BS}} \equiv \text{CE}_2$ as an example of a strictly proper calibration error for classification since the Brier score is a strictly proper score on $\mathscr{P}_n$. Strictly proper calibration errors have the highly desired property: $\text{CE}_S\left(f\right) = 0 \overset{a.s.}{\iff} f$ is calibrated. Since proper scores are not restricted to classification, the above definition gives a natural extension of calibration errors beyond classification.

Additionally, by generalizing the definition of proper scores, we can show that the squared KCE is a strictly proper calibration error (Appendix F). But, in general, there does not exist an unbiased estimator of a proper calibration error, since we cannot estimate $\mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]$ in an unbiased manner. Because we do not want lower bounds for errors used in sensitive applications, we introduce the following theorem about how to construct an upper bound.

Theorem 4.3. *For all proper calibration errors with $\inf_{P\in\mathscr{P}} g_S\left(P\right) \in \mathbb{R}$, there exists an associated **calibration upper bound***

$$\mathscr{U}_S\left(f\right) \geq \mathit{CE}_S\left(f\right)$$

defined as $\mathscr{U}_S\left(f\right) = \mathbb{E}\left[S\left(f\left(X\right), Y\right)\right] - \inf_{P\in\mathscr{P}} g_S\left(P\right)$. Under a classification setting and further mild conditions, we have $\lim_{ACC(f)\to 1} \mathscr{U}_S\left(f\right) - \mathit{CE}_S\left(f\right) = 0$.

In other words, we can always construct a non-trivial upper bound of a proper calibration error as long as the generalized entropy function has a finite infimum. The calibration upper bound approaches the true calibration error for models with high accuracy. For the Brier score we have $\inf_{P \in \mathscr{P}} g_{\text{BS}}(P) = 0$, which gives $\mathscr{U}_{\text{BS}}(f) = \mathbb{E}\left[\|f(X) - Y'\|^2\right]$. This makes the Brier score itself an upper bound of its induced calibration error.

Our proposed calibration upper bounds are provably reliable to use since they all have a minimum-variance unbiased estimator. Combining Theorem 3.1 and 4.3 gives $\sqrt{\mathscr{U}_{\text{BS}}(f)} = \sqrt{\text{BS}(f)} \geq \text{ECE}(f)$. Consequently, we might also be interested in transforming the estimated upper bound in a non-linear way for interpretability purposes. In general, any unbiased estimator $\hat{\theta}$ becomes biased after a non-linear transformation t, since $\mathbb{E}\left[t\left(\hat{\theta}\right)\right] \neq t\left(\mathbb{E}\left[\hat{\theta}\right]\right)$. But, if t is continuous, our estimator is still asymptotically unbiased and consistent [48].[1] We will further investigate the empirical robustness w.r.t. data size in Section 5 with t as the square root.

Furthermore, $\mathscr{U}_S$ has the following properties, which are helpful for the application of recalibration method optimization and selection.

Proposition 4.4. *Given injective functions $h, h' : \mathscr{P} \to \mathscr{P}$ we have*

$$\mathscr{U}_S(h \circ f) - \mathscr{U}_S(f) = CE_S(h \circ f) - CE_S(f) \quad ,$$

$$\mathscr{U}_S(h \circ f) > \mathscr{U}_S(h' \circ f) \iff CE_S(h \circ f) > CE_S(h' \circ f)$$

and (assuming S is differentiable)

$$\frac{\mathrm{d}\mathscr{U}_S(h \circ f)}{\mathrm{d}h} = \frac{\mathrm{d}CE_S(h \circ f)}{\mathrm{d}h}.$$

This is a generalization of Proposition 4.2 presented in [61]. It tells us that we can reliably estimate the improvement of any injective recalibration method via the upper bound. Furthermore, we get access to the calibration gradient and can compare different transformations. At first, injectivity seems like a significant restriction. But, we argue in the following that injectivity - rather than being accuracy-preserving - is a desired property of general recalibration methods. For example, we can construct a recalibration method, which is calibrated and accuracy-preserving, but only predicts a finite set of distinct values (see Appendix E). Specifically, we would only predict two distinct values for any input in binary classification. To exclude such naive solutions which substantially reduce model sharpness, we restrict ourselves to injective transformations of $\mathscr{P}_n \to \mathscr{P}_n$. These do provably not impact the model sharpness and preserve, at least partly, the continuity of the output space. Examples of injective transformations are TS, ETS, and DIAG. These state-of-the-art methods show very competitive performances even when compared to non-injective recalibration methods [61, 44]. Further, Proposition 4.4 also holds when replacing $\mathscr{U}_S$ with the expected score s_S without further conditions. This is useful when $\mathscr{U}_S$ does not exist, but we still want to perform recalibration as in Section 5 in the case of the DSS.

5 Experiments

In the following, we investigate the behavior of calibration error estimators in three settings. First, we use varying test set sizes for the estimators and compare their values. This will show how well the inequalities in Theorem 3.1 hold in practical settings and how robust the estimators are. Second, we explore what the estimated improvements of several recalibration methods are. This is done after the recalibration methods are already optimized on a given validation set; we only vary the size of the test set and compute calibration errors on these test sets before and after recalibration. In both settings, the straighter a line is, the more robust and, consequently, trustworthy is the estimator for practical applications. Third, we investigate how our framework can be used to improve calibration for tasks beyond classification by performing probabilistic regression with subsequent recalibration.

In all experiments we evaluate the following estimators: CWCE_2 with 15 equal width bins ('15b CWCE_2'), CWCE_2 with 100 equal width bins ('100b CWCE_2'), ECE with 15 equal width bins

[1] follows from Continuous Mapping Theorem and Theorem 3.2.6 of Takeshi [48]

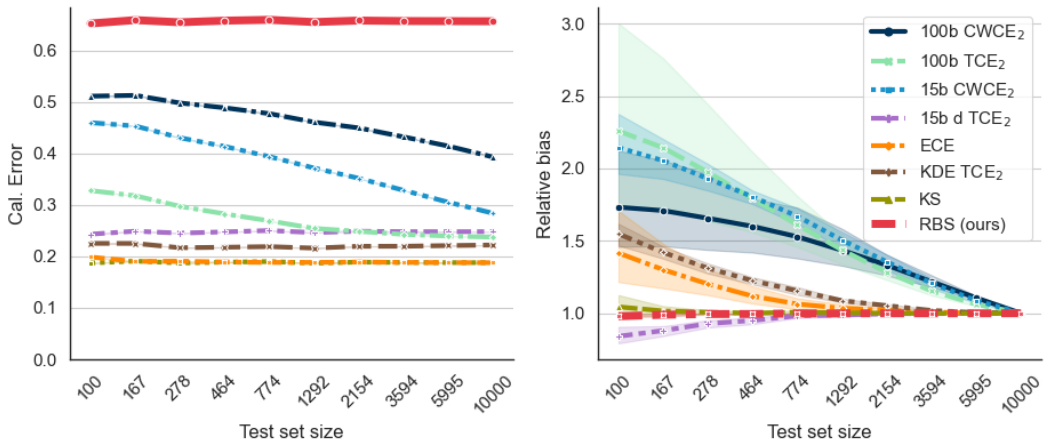


Figure 3: **Left:** Different calibration error estimates versus the test set size of ResNet Wide 32 and CIFAR100. The red line corresponds to the square root of the Brier score which is an upper bound of the true $\sqrt{\text{CE}_2}$. The other errors are lower bounds. **Right:** Relative change versus data size with respect to error at full size. Averaging across a multitude of models shows a systematic trend. An unbiased estimator would give a flat line.

('ECE'), TCE_2 with 100 equal width bins ('100b TCE_2'), TCE_2 with 15 equal mass bins and debias term ('15b d TCE_2'), TCE_2 with kernel density estimation ('KDE TCE_2'), KS ('KS') and the root calibration upper bound $\sqrt{\mathscr{U}_{\text{BS}}}$ ('RBS (ours)').The bin amounts are chosen based on past literature [31, 36]. The experiments are conducted across several model-dataset combinations, for which logit sets are openly accessible [2] [29, 44]. This includes the models Wide ResNet 32 [60], DenseNet 40, and DenseNet 161 [22] and the datasets CIFAR10, CIFAR100 [27], and ImageNet [9]. We did not conduct model training ourselves and refer to [29] and [44] for further details. We include TS, ETS, and DIAG as injective recalibration methods. Further details and results on additional models and datasets are reported in the Appendix G.

Robustness of calibration errors to test set size We illustrate the estimated values of our introduced upper bound and the other errors, which are lower bounds of the unknown CE_2 on the left of Figure 3. On the right, we aggregate across several models to show the systematic drop-off according to Proposition 3.3. The relative bias is computed by $Error(n)/Error(10000)$ and allows an aggregation of models with different calibration levels. Included models are DenseNet 40, Wide ResNet 32, ResNet 110 SD, ResNet 110, and LeNet 5, all trained on CIFAR10. All values represent the calibration of the given model without recalibration transformation. Only our proposed upper bound and KS are stable, and Appendix G shows this holds across a wide range of different settings. The theoretically highest lower bound (CWCE_2 with 100 bins) is also constantly the highest estimated lower bound, but it is sensitive to the test set size. Results for further settings presented in Appendix G show similar results.

Quantifying recalibration improvement Next, we assessed how well all estimators were able to quantify the improvement in calibration error after applying different injective recalibration methods (Fig. 1). Only our proposed upper bound estimator RBS is again robust throughout all settings. According to Proposition 4.4 and since RBS is asymptotically unbiased and consistent, it can be regarded as a reliable approximation of the real improvement of the presented recalibration methods. For all other estimators, there is a general trend to estimate recalibration improvement higher for large test set sizes. In other settings, especially for small test set sizes, calibration improvement is underestimated to the extent that negative improvements (poorer calibration than before) are suggested. Results on other settings presented in Appendix G show similar results. Taken together, these experiments demonstrate the unreliability of existing calibration estimators, in particular, when used to evaluate recalibration methods. In contrast, our proposed upper bound estimator is stable across different settings.

[2] https://github.com/markus93/NN_calibration/ and https://github.com/AmirooR/IntraOrderPreservingCalibration

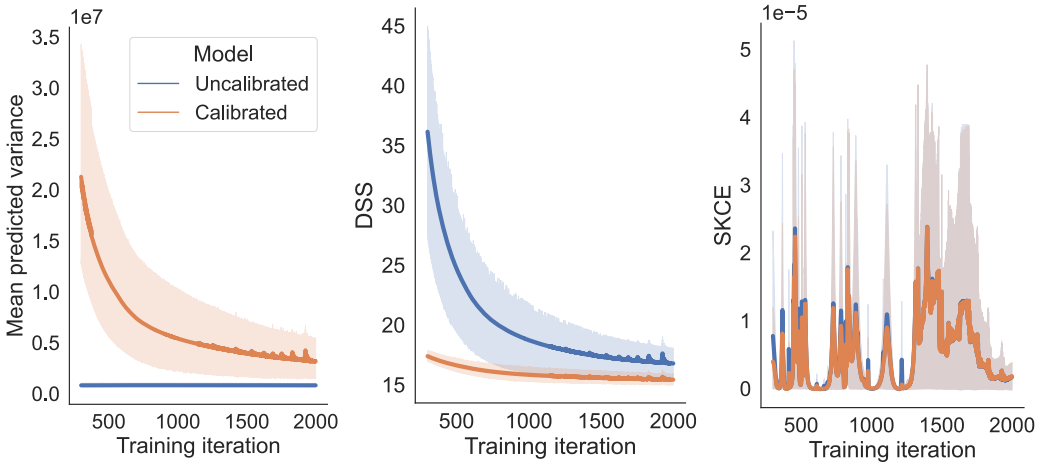


Figure 4: **Left:** Average predicted variance throughout model training before and after recalibration. Initially, due to a bad fit, recalibration adjusts the variance accordingly for better communicated uncertainty. Once the model fit improves, the predicted variance requires less adjustment due to less uncertainty in each prediction. **Middle:** DSS communicates reasonably changes in the variance due to recalibration. **Right:** SKCE fails to capture the variance trend and behaves erratically.

Variance regression calibration We consider variance regression to demonstrate the usefulness of proper calibration errors outside the classification setting. To this end, we predict sales prices with an uncertainty estimate in the UCI dataset *Residential Building*, which consists of a high feature (107) to data instances (372) ratio [43]. Our model of choice is a fully-connected mixture density network predicting mean and variance [2]. Similar to classification, we are interested in recalibration of the predicted variance to adjust possible under- or overconfidence. We use our proposed framework to derive a proper calibration error induced by the proper score DSS for recalibration. Further, we compare DSS [12] to squared KCE (SKCE) [56] and analyze the average predicted variance throughout model training. We use Platt scaling ($x \mapsto wx + b$ with parameters $w, b \in \mathbb{R}$ [42]) on the predicted variance in each training iteration to show how recalibration benefits uncertainty awareness. We expect high uncertainty awareness at the start of the model training with a drop-off at later iterations. As can be seen in Figure 4, recalibration is able to adjust the uncertainty estimate of a model as desired. Further, the DSS estimate, which captures predicted mean and variance correctness, directly communicates the improved variance fit. Contrary, the SKCE estimate appears more erratic between iteration steps and seemingly ignores the changes in variance and the recalibration improvement. We perform further variance regression experiments in Appendix G.

6 Conclusion

In this work, we address the problem of reliably quantifying the effect of recalibration on predictive uncertainty for classification and other probabilistic tasks. This is critical for adjusting under- or overconfidence via recalibration. To this end, we first provide a taxonomy of existing calibration errors. We discover that most errors are lower bounds of a proper calibration error and fail to assess if a model is calibrated. This motivates our definition of *proper calibration errors*, which provides a general class of errors for arbitrary probabilistic predictions. Since proper calibration errors cannot be estimated in the general case, we introduce upper bounds, which directly measure the calibration change for injective transformations. This allows us to reliably adjust model uncertainty via recalibration. We demonstrate theoretically and empirically that the estimated calibration improvement can be highly misleading for commonly used estimators, including the ECE. In stark contrast, our upper bound is robust to changes in data size and estimates robustly the improvement via injective recalibration. We further show in additional experiments that our approach can be applied successfully to variance regression.

References

[1] Aitchison, J. and Shen, S. M. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2): 261–272, 1980. ISSN 00063444. URL http://www.jstor.org/stable/2335470.

[2] Bishop, C. M. Mixture density networks. 1994.

[3] Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1): 1 – 3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

[4] Bröcker, J. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, Jul 2009. ISSN 1477-870X. doi: 10.1002/qj.456. URL http://dx.doi.org/10.1002/qj.456.

[5] Capiński, M. and Kopp, P. E. *Measure, integral and probability*, volume 14. Springer, 2004.

[6] Chung, Y., Neiswanger, W., Char, I., and Schneider, J. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification, 2021.

[7] Dawid, A. P. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379): 605–610, 1982. doi: 10.1080/01621459.1982.10477856. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1982.10477856.

[8] Dawid, A. P. and Sebastiani, P. Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, pp. 65–81, 1999.

[9] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

[10] Fan, H., Ferianc, M., Que, Z., Niu, X., Rodrigues, M. L., and Luk, W. Accelerating bayesian neural networks via algorithmic and hardware optimizations. *IEEE Transactions on Parallel and Distributed Systems*, 2022.

[11] Friedman, J. H. Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67, 1991.

[12] Gneiting, T. and Katzfuss, M. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1: 125–151, 2014.

[13] Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

[14] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

[15] Gupta, A., Kvernadze, G., and Srikumar, V. Bert & family eat word salad: Experiments with text understanding. *ArXiv*, abs/2101.03453, 2021.

[16] Gupta, K., Rahimi, A., Ajanthan, T., Mensink, T., Sminchisescu, C., and Hartley, R. Calibration of neural networks using splines. In *International Conference on Learning Representations*, 2020.

[17] Haggenmüller, S., Maron, R. C., Hekler, A., Utikal, J. S., Barata, C., Barnhill, R. L., Beltraminelli, H., Berking, C., Betz-Stablein, B., Blum, A., Braun, S. A., Carr, R., Combalia, M., Fernandez-Figueras, M.-T., Ferrara, G., Fraitag, S., French, L. E., Gellrich, F. F., Ghoreschi, K., Goebeler, M., Guitera, P., Haenssle, H. A., Haferkamp, S., Heinzerling, L., Heppt, M. V., Hilke, F. J., Hobelsberger, S., Krahl, D., Kutzner, H., Lallas, A., Liopyris, K., Llamas-Velasco, M., Malvehy, J., Meier, F., Müller, C. S., Navarini, A. A., Navarrete-Dechent, C., Perasole, A., Poch, G., Podlipnik, S., Requena, L., Rotemberg, V. M., Saggini, A., Sangueza, O. P., Santonja, C., Schadendorf, D., Schilling, B., Schlaak, M., Schlager, J. G., Sergon, M., Sondermann, W., Soyer, H. P., Starz, H., Stolz, W., Vale, E., Weyers, W., Zink, A., Krieghoff-Henning, E., Kather, J. N., von Kalle, C., Lipka, D. B., Fröhling, S., Hauschild, A., Kittler, H., and Brinker, T. J. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *European Journal of Cancer*, 156:202–216, 2021. ISSN 0959-8049. doi: https://doi.org/10.1016/j.ejca.2021.06.049. URL https://www.sciencedirect.com/science/article/pii/S0959804921004445.

[18] Hastie, T. and Tibshirani, R. Classification by pairwise coupling. *The annals of statistics*, 26(2):451–471, 1998.

[19] He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[20] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[21] Hoeffding, W. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293 – 325, 1948. doi: 10.1214/aoms/1177730196. URL https://doi.org/10.1214/aoms/1177730196.

[22] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[23] Islam, A., Chen, C.-F., Panda, R., Karlinsky, L., Radke, R. J., and Feris, R. S. A broad study on the transferability of visual representations with contrastive learning. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8825–8835, 2021.

[24] Joo, T., Chung, U., and Seo, M. Being bayesian about categorical probability. In *ICML*, 2020.

[25] Katsaouni, N., Tashkandi, A., Wiese, L., and Schulz, M. H. Machine learning based disease prediction from genotype data. *Biological Chemistry*, 402(8):871–885, 2021. doi: doi:10.1515/hsz-2021-0109. URL https://doi.org/10.1515/hsz-2021-0109.

[26] Kristiadi, A., Hein, M., and Hennig, P. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *ICML*, 2020.

[27] Krizhevsky, A. et al. Learning multiple layers of features from tiny images. 2009.

[28] Kull, M., Filho, T. S., and Flach, P. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 623–631. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/kull17a.html.

[29] Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in Neural Information Processing Systems*, 32:12316–12326, 2019.

[30] Kumar, A., Sarawagi, S., and Jain, U. Trainable calibration measures for neural networks from kernel mean embeddings. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2805–2814. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kumar18a.html.

[31] Kumar, A., Liang, P., and Ma, T. Verified uncertainty calibration. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 3792–3803, 2019.

[32] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[33] Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.-J., Fei-Fei, L., Yuille, A., Huang, J., and Murphy, K. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 19–34, 2018.

[34] Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32:13153–13164, 2019.

[35] Menon, A. K., Rawat, A. S., Reddi, S. J., Kim, S., and Kumar, S. A statistical perspective on distillation. In *ICML*, 2021.

[36] Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[37] Morales-Álvarez, P., Hernández-Lobato, D., Molina, R., and Hernández-Lobato, J. M. Activation-level uncertainty in deep neural networks. In *ICLR*, 2021.

[38] Murphy, A. H. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595 – 600, 1973. doi: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/apme/12/4/1520-0450_1973_012_0595_anvpot_2_0_co_2.xml.

[39] Naeini, M. P., Cooper, G. F., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 2901–2907. AAAI Press, 2015. ISBN 0262511290.

[40] Nguyen, K. and O'Connor, B. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1587–1598, 2015.

[41] Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2, 2019.

[42] Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large-Margin Classifiers*, pp. 61–74. MIT Press, 1999.

[43] Rafiei, M. H. and Adeli, H. A novel machine learning model for estimation of sale prices of real estate units. *Journal of Construction Engineering and Management*, 142(2):04015066, 2016.

[44] Rahimi, A., Shaban, A., Cheng, C.-A., Hartley, R., and Boots, B. Intra order-preserving functions for calibration of multi-class neural networks. *Advances in Neural Information Processing Systems*, 33: 13456–13467, 2020.

[45] Roelofs, R., Cain, N., Shlens, J., and Mozer, M. C. Mitigating bias in calibration error estimation, 2021.

[46] Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[47] Song, H., Diethe, T., Kull, M., and Flach, P. Distribution calibration for regression. In *International Conference on Machine Learning*, pp. 5897–5906. PMLR, 2019.

[48] Takeshi, A. *Advanced econometrics*. Harvard university press, 1985.

[49] Tian, J., Yung, D., Hsu, Y.-C., and Kira, Z. A geometric perspective towards neural calibration via sensitivity decomposition. In *NeurIPS*, 2021.

[50] Tomani, C., Gruber, S., Erdem, M. E., Cremers, D., and Buettner, F. Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10124–10132, June 2021.

[51] Tsagris, M., Beneki, C., and Hassani, H. On the folded normal distribution. *Mathematics*, 2(1):12–28, feb 2014. doi: 10.3390/math2010012. URL https://doi.org/10.3390%2Fmath2010012.

[52] Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., and Schön, T. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3459–3467. PMLR, 2019.

[53] Wang, X., Liu, H., Shi, C., and Yang, C. Be confident! towards trustworthy graph neural networks via confidence calibration. In *NeurIPS*, 2021.

[54] Wehenkel, A. and Louppe, G. Unconstrained monotonic neural networks. *Advances in Neural Information Processing Systems*, 32:1545–1555, 2019.

[55] Widmann, D., Lindsten, F., and Zachariah, D. Calibration tests in multi-class classification: A unifying framework. *Advances in Neural Information Processing Systems*, 32:12257–12267, 2019.

[56] Widmann, D., Lindsten, F., and Zachariah, D. Calibration tests beyond classification. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=-bxf89v3Nx.

[57] Yen, M.-H., Liu, D.-W., Hsin, Y.-C., Lin, C.-E., and Chen, C.-C. Application of the deep learning for the prediction of rainfall in southern taiwan. *Scientific reports*, 9(1):1–9, 2019.

[58] Zadrozny, B. and Elkan, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *ICML*, 1, 05 2001.

[59] Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pp. 694–699, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 158113567X. doi: 10.1145/775047.775151. URL https://doi.org/10.1145/775047.775151.

[60] Zagoruyko, S. and Komodakis, N. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016.

[61] Zhang, J., Kailkhura, B., and Han, T. Y.-J. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*, pp. 11117–11128. PMLR, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We mention required conditions for the proofs either directly or refer to Appendix D.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We see positive societal impacts from improved uncertainty awareness via recalibration.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Major conditions are directly stated; minor conditions in Appendix D as referred.
   (b) Did you include complete proofs of all theoretical results? [Yes] In Appendix D.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Full code is located in the supplementary material and further experimental details are in Appendix G.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Relevant training details are located in Appendix G. A lot of models are not trained by ourselves. Instead, we loaded their results from publicly accessible URLs. We refer to each URL and the original work, where the training was conducted.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We repeated experiments until the error bars are barely visible or not visible anymore. Only exception is Figure 4, where aggregations hinder visibility due to high variances in training different seeds.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix G.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We used assets from [29] and Rahimi et al. [44] located in https://github.com/markus93/NN_calibration/ and https://github.com/AmirooR/IntraOrderPreservingCalibration .
   (b) Did you mention the license of the assets? [No] Each lincense can be viewed in the respective github repository.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include the code for reproducing the experiments and figures in the supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] The data is publicly accessible and we cited each respective source.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The data is often abstract in nature (we used pretrained logits provided by a publicly accessible source) or downloaded the dataset from the UCI dataset database.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] No crowdsourcing or human subjects

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] No crowdsourcing or human subjects

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] No crowdsourcing or human subjects