Calibration in Context: A Case Study with Score Decompositions

Johannes Resin
Goethe University Frankfurt, Germany

Abstract

Decompositions of proper scores into measures of miscalibration (reliability), discrimination (resolution), and uncertainty have a long history in weather forecasting. In machine learning (ML), related calibration error metrics are now seeing a surge of interest. In this note, I review the close connection between these concepts and present a small case study on image classifiers from the literature. The study exemplifies that an exclusive focus on calibration error may lead to questionable conclusions when improvements in calibration come at the expense of a drastic decline in overall predictive performance. I critically examine histogram binning and show that isotonic regression produces better overall recalibration results. A simple linear interpolation of the isotonic fit is shown to further improve predictive performance without loss of calibration.

1 BACKGROUND & MOTIVATION

By now, the need for reliable uncertainty quantification is widely recognized in the ML literature (e.g., Hüllermeier and Waegeman, 2021). Probabilistic classifiers address this need by predicting class distributions. Calibration—the statistical compatibility between predictive distributions and true classes—is crucial for decision-making and trustworthy AI systems (e.g., Vaicenavicius et al., 2019). The predictive performance of such classifiers is typically assessed using proper scoring rules (Gneiting and Raftery, 2007), which can be decomposed into miscalibration (reliability), discrimination (resolution), and uncertainty components (e.g., Bröcker, 2009; Pohle, 2020), building on the seminal work of Sanders (1963) and Murphy (1973). In contrast, many authors in ML advocate the use of calibration error metrics (e.g., Naeini et al., 2015; Guo et al., 2017; Kumar et al., 2019; Nixon et al., 2019; Gupta and Ramdas, 2022; Błasiok et al., 2023; Rossellini et al., 2025), often treating them as primary evaluation criteria.

This note highlights the importance of considering calibration in the context of overall predictive performance, critically examining a finding by Gupta and Ramdas (2022). The study aligns with recent research that appreciates score decompositions and addresses the interplay and tensions between calibration and overall predictive ability (Gruber and Buettner, 2022; Silva Filho et al., 2023; Machado et al., 2024; Popordanoska et al., 2024; Berta et al., 2025; Chidambaram and Ge, 2025). A recent strand of statistical literature suggests the use of non-parametric isotonic regression as a robust alternative to traditional binning-based estimators of score components. Here, I adapt this approach to (probabilistic) multi-class classifiers, leveraging work on binary classifiers (Dimitriadis et al., 2021), and point and distributional forecasts for real-valued outcomes (Gneiting and Resin, 2023; Arnold et al., 2024). The decomposition is used to study class-wise recalibration approaches based on histogram binning and isotonic regression. The study shows that a simple smoothing of isotonic fits preserves discrimination effectively resulting in the best overall performance.

2 SETTING & METHODS

This section briefly introduces the problem setup and key tools for quantitative out-of-sample verification.

2.1 Basic Setting

Classification aims to predict the class label $Y \in \{1, \dots, k\}$ of an instance characterized by a feature vector $X \in \mathbb{R}^d$. A probabilistic classifier $c\colon \mathbb{R}^d \to \Delta_k = \{p = (p_j)_{j=1}^k \in [0,1]^k \mid \sum_j p_j = 1\}$ maps the features X to a probability distribution given by a vector of class probabilities from the probability simplex

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco.

Δ_k. The variables X and Y are assumed to follow a joint distribution $\mathbb{P}$ on a probability space Ω, where $\mathbb{P}_{Y|c(X)}$ denotes the conditional distribution of Y given the prediction $c(X)$, and $\mathbb{P}_Y$ denotes the marginal distribution of Y. Ideally, the classifier c should be calibrated (Vaicenavicius et al., 2019) in the sense that

$$\mathbb{P}_{Y|c(X)} = c(X), \tag{1}$$

while also effectively discriminating between the different classes. Together, calibration and discrimination are the key properties that drive the overall predictive performance of a classifier.

2.2 Recalibration Methods

To improve calibration, classifiers are often recalibrated using various methods (see Silva Filho et al., 2023). The common class-wise approach recalibrates each class probability individually using a binary calibrator and then normalizes the resulting probabilities (Zadrozny and Elkan, 2002). However, Gupta and Ramdas (2022) have recently suggested to omit the normalization step. The proposal raises significant conceptual and practical concerns: *How can binary class probabilities that do not sum to one be meaningfully interpreted or used for effective decision-making?* Rather than pursue these issues further, I reexamine the findings by Gupta and Ramdas (2022) empirically in this note.

The subsequent case study focuses on the class-wise approach, contrasting the use of histogram binning as used by Gupta and Ramdas (2022) and isotonic regression. Histogram binning partitions the unit interval into a fixed number of bins, recalibrating predictions in each bin to match the conditional event frequency of the bin. This method requires some tuning of the type of binning (equal-width or equal mass) and number of bins. In contrast, isotonic regression (Barlow et al., 1972) is a non-parametric, tuning-free method that finds (in-sample) optimal predictions under a natural monotonicity constraint via the pool-

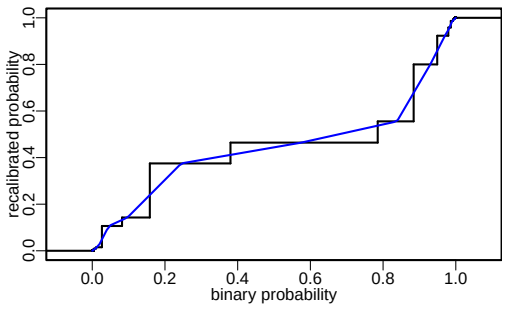

Figure 1: Isotonic fit (black) and smoothing via linear interpolation anchored at bin medians (blue).

adjacent-violators algorithm. The isotonic fit is characterized by its piecewise constant segments, which give rise to a data-driven binning. The case study also includes a simple smoothing of the isotonic fit that linearly interpolates between bin medians, as described in Appendix B and illustrated in Figure 1, to avoid discretizing recalibrated probabilities and preserve discrimination more effectively. This approach is closely related to the method by Jiang et al. (2011), who propose a piecewise cubic Hermite spline interpolation that produces a more refined smoothing.

2.3 Proper Scoring Rules

Scoring rules quantify overall predictive performance by assigning numeric scores to prediction-observation pairs $(p, y) \in \Delta_k \times \{1, \dots, k\}$. A scoring rule S is *strictly proper* if the expected score is always minimized by the true class distribution, i.e.,

$$\mathbb{E}_{Y\sim p}\, \mathrm{S}(p, Y) < \mathbb{E}_{Y\sim p}\, \mathrm{S}(q, Y) \text{ for all } p, q \in \Delta_k, p \neq q.$$

Here, I use the *Brier score*—a popular strictly proper scoring rule dating back to Brier (1950)—given by

$$\mathrm{BS}(p, y) = \sum_{j=1}^{k} (\mathbb{1}\{y = j\} - p_j)^2 \tag{2}$$

with indicator $\mathbb{1}\{y = j\}$ equal to 1 if $y = j$ and 0 otherwise. The Brier score is a sum of binary Brier scores $(p_j, \mathbb{1}\{y = j\}) \mapsto (\mathbb{1}\{y = j\} - p_j)^2$ for each class j.

2.4 Score Decompositions

The expected score $\overline{\mathrm{S}} = \mathbb{E}\, \mathrm{S}(c(X), Y)$ under a proper scoring rule can be decomposed as

$$\overline{\mathrm{S}} = \mathrm{MCB} - \mathrm{DSC} + \mathrm{UNC}, \tag{3}$$

where $\mathrm{MCB} = \overline{\mathrm{S}} - \mathbb{E}\, \mathrm{S}(\mathbb{P}_{Y|c(X)}, Y)$ quantifies *miscalibration*, $\mathrm{DSC} = \mathbb{E}\, \mathrm{S}(\mathbb{P}_Y, Y) - \mathbb{E}\, \mathrm{S}(\mathbb{P}_{Y|c(X)}, Y)$ quantifies *discrimination*, and $\mathrm{UNC} = \mathbb{E}\, \mathrm{S}(\mathbb{P}_Y, Y)$ (uncertainty) is a reference score attained by the marginal distribution (which is calibrated but does not discriminate at all).

2.5 Calibration Error

Calibration error is typically defined as the expected distance between predicted and true conditional distribution under a distance $d\colon \Delta_k^2 \to [0, \infty)$, i.e.,

$$\mathrm{CE} = \mathbb{E} d(c(X), \mathbb{P}_{Y|c(X)}).$$

When $d(q, p) = \mathbb{E}_{Y\sim p}[\mathrm{S}(q, Y) - \mathrm{S}(p, Y)]$ is the Bregman divergence associated with a proper scoring rule S (see

Gneiting and Raftery, 2007), a simple calculation confirms that CE = MCB. In what follows, I focus on the Brier score decomposition with its miscalibration component as calibration error.

2.6 Estimation of Score Components

In practice, the decomposition (3) is estimated on a test set of n prediction-observation pairs $(p_1, y_1), \ldots, (p_n, y_n)$, requiring estimates $\widehat{p}_1, \ldots, \widehat{p}_n$ of the conditional distribution $\mathbb{P}_{Y|c(X)=p_i}$ given prediction $i = 1, \ldots, n$. Using the empirical class frequencies $\widehat{p}_0$ to estimate the marginal distribution, an empirical decomposition of the average score $\widehat{\mathrm{S}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{S}(p_i, y_i)$ is obtained as

$$\widehat{\mathrm{S}} = \widehat{\mathrm{MCB}} - \widehat{\mathrm{DSC}} + \widehat{\mathrm{UNC}}, \tag{4}$$

$$\text{where } \widehat{\mathrm{MCB}} = \widehat{\mathrm{S}} - \tfrac{1}{n}\textstyle\sum_i \mathrm{S}(\widehat{p}_i, y_i),$$
$$\widehat{\mathrm{DSC}} = \tfrac{1}{n}\textstyle\sum_i \mathrm{S}(\widehat{p}_0, y_i) - \mathrm{S}(\widehat{p}_i, y_i), \text{ and}$$
$$\widehat{\mathrm{UNC}} = \tfrac{1}{n}\textstyle\sum_i \mathrm{S}(\widehat{p}_0, y_i).$$

Similar to the Brier score based decomposition of the CRPS in Arnold et al. (2024), the structure of the Brier score in (2) admits directly applying the decomposition by Dimitriadis et al. (2021) to the binary summands. This approach yields a *class-wise* decomposition and is equivalent to fitting an isotonic regression separately for each class, without normalizing the recalibrated class probabilities. Unlike standard binning estimators, the isotonic approach is independent of implementation choices while being based on a consistent estimator (Dimitriadis et al., 2021). However, the class-wise decomposition quantifies deviations from a weaker form of calibration, namely, *class-wise calibration* (Gupta and Ramdas, 2022) given by

$$\mathbb{P}(Y = j \mid p_j) = p_j \quad \text{for all} \quad j \in \{1, \ldots, k\}.$$

Appendix A contains additional results for an alternative *normalized* decomposition that uses isotonic regression with a normalization step to estimate the conditional class probabilities $\widehat{p}_i$ to be plugged into (4). The normalized decomposition quantifies deviations from the stronger form of calibration in (1).

3 CASE STUDY

The case study builds upon an experiment from Gupta and Ramdas (2022), who compared several recalibration methods for common image classifiers using a custom calibration error metric (CW-ECE) on CIFAR-10 and CIFAR-100 (Krizhevsky, 2009). In their study, they do not use other metrics such as proper scoring rules or classification accuracy. They propose omitting the standard normalization step in class-wise recalibration with binary histogram binning (CW-HB), contrasting it with the usual normalized approach (N-HB). Gupta and Ramdas (2022, p. 9, bold in original) report the following finding:

> "For CW-ECE, CW-HB is the best performing method across the two datasets and all four architectures. The N-HB method which has been used in many CW-ECE baseline experiments performs terribly. In other words, skipping the normalization step leads to a large improvement in CW-ECE. **This observation is one of our most striking findings.**"

Here, I reexamine this finding via the class-wise Brier score decomposition. The case study uses calibration and test sets of raw model predictions from the supplementary material for Gupta and Ramdas (2022). The datasets include predictions from seven base models (Zagoruyko and Komodakis, 2016; Huang et al., 2017; Mukhoti et al., 2020): ResNet-50, ResNet-110, WideResNet (each trained with either Brier score or focal loss), and DenseNet-121 (trained with focal loss). The present study features the following recalibration methods: histogram binning with normalization (N-HB) and without normalization (CW-HB) as used by Gupta and Ramdas (2022), isotonic regression with and without normalization (N-IR and CW-IR, respectively), and the smoothed version of isotonic regression (N-SIR and CW-SIR) introduced in Section 2.2.

Table 1 presents the score decompositions along with classification accuracy, i.e., the frequency with which the predicted (most likely) class matches the true class, averaged across all base models for each recalibration method and the two datasets. The MCB-DSC plots in Figure 2 provide a visualization of the individual performance for each model and recalibration method by plotting discrimination, DSC, against miscalibration, MCB. The isolines on these plots represent decompositions with equal overall predictive performance, illustrating trade-offs between miscalibration and discrimination. Similar plots have appeared in Gneiting et al. (2023); Arnold et al. (2024); Dimitriadis et al. (2024). Analogous tables and plots showing a normalized version of the Brier score decomposition are presented in Appendix A.

In line with the CW-ECE results reported by Gupta and Ramdas (2022), the class-wise decomposition yields a low calibration error, MCB, for CW-HB. However, this improvement comes at the cost of a drastic decline in discrimination, DSC, resulting in a doubled Brier score and twice as many misclassifications (1 - Acc). While the normalization step in N-HB partially recovers discrimination ability, predictive performance remains substantially reduced. In contrast, the

Table 1: Class-wise Brier score decomposition (×100) and classification accuracy (Acc) averaged across seven base models for both datasets (CIFAR-10 and CIFAR-100) and each recalibration method. The best average values per metric are highlighted in **bold**. Values in red indicate a drastic decline in predictive performance.

CIFAR-10					
Method	BS	MCB	DSC	UNC	Acc
base	0.781	0.063	8.282	9	**0.951**
CW-HB	1.832	**0.009**	7.177	9	0.890
N-HB	1.237	0.070	7.833	9	0.890
CW-IR	0.779	0.027	8.248	9	0.950
N-IR	0.770	0.047	8.277	9	0.950
CW-SIR	0.769	0.051	8.282	9	0.950
N-SIR	**0.764**	0.047	**8.283**	9	0.950

CIFAR-100					
Method	BS	MCB	DSC	UNC	Acc
base	0.327	0.039	0.702	0.99	**0.775**
CW-HB	0.865	**0.003**	0.128	0.99	0.483
N-HB	0.532	0.036	0.494	0.99	0.483
CW-IR	0.333	0.021	0.677	0.99	0.770
N-IR	0.324	0.034	0.699	0.99	0.770
CW-SIR	0.323	0.035	0.702	0.99	0.773
N-SIR	**0.320**	0.033	**0.703**	0.99	0.773

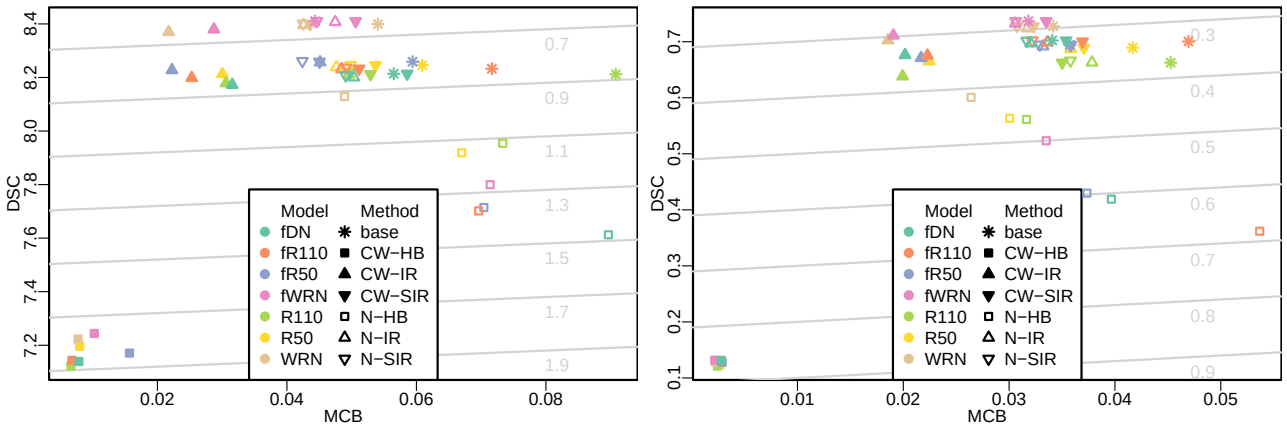

Figure 2: MCB-DSC plots for CIFAR-10 (left) and CIFAR-100 (right) showing class-wise Brier score decompositions (×100). The plots show results for seven base models that differ in architecture (**R**esNet-**50**, **R**esNet-**110**, **W**ide**R**es**N**et, and **D**ense**N**et-121) and training loss (**f**ocal loss or Brier score). The gray isolines correspond to MCB-DSC pairs that yield the same overall Brier score.

normalized score decomposition (presented in the appendix) attributes most of the score difference between CW-HB and N-HB to a large miscalibration component for CW-HB, because this decomposition quantifies deviations from the stronger notion of calibration in (1). Apart from this exception, the normalized decomposition shows fairly similar results with a slight increase in MCB for the class-wise (CW-X) approaches and a slight decrease in MCB for the normalized (N-X) versions. While CW-IR exhibits lower miscalibration than the other methods based on isotonic regression, this improvement comes at the cost of a moderate loss of discrimination, resulting in overall performance similar to the base classifier. Notably, the smoothed version CW-SIR does not show this trade-off, which may thus be due to CW-IR discretizing the recalibrated class probabilities. Ultimately, smoothed isotonic regression with normalization (N-SIR) achieves the best overall performance, reducing miscalibration while fully preserving discrimination ability.

4 CONCLUSIONS

In this note, I have illustrated the benefits of score decompositions for a balanced assessment of calibration and predictive performance of probabilistic multiclass classifiers. The results show a substantial decline in predictive performance with recalibration via histogram binning, while isotonic regression produced much better overall results. The smoothed version of isotonic regression with normalization (N-SIR) is found to be particularly effective at improving overall performance, preserving discrimination ability of the base classifiers perfectly while improving calibration in the case study.

In future research, I plan to study the statistical properties of the proposed isotonic estimators including possible biases (cf. Ferro and Fricker, 2012; Roelofs et al., 2022), and extend the study to additional classification tasks and recalibration methods.

Acknowledgements

I would like to thank Tilmann Gneiting and Alexander Jordan, as well as an anonymous reviewer for insightful comments and helpful discussion. The author gratefully acknowledges support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via project number 502572912.

Code and Data Availability

Code for replicating the case study is available at `https://github.com/resinj/calibration_AISTATS26`. Calibration and test data of raw model predictions are available from the supplementary material for Gupta and Ramdas (2022) at `https://openreview.net/forum?id=WqoBaaPHS-`.

References

Arnold, S., Walz, E.-M., Ziegel, J. F., and Gneiting, T. (2024). Decompositions of the mean continuous ranked probability score. *Electronic Journal of Statistics*, 18:4992–5044.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley, New York.

Berta, E., Holzmüller, D., Jordan, M. I., and Bach, F. (2025). Rethinking early stopping: Refine, then calibrate. Preprint, arXiv:2501.19195.

Błasiok, J., Gopalan, P., Hu, L., and Nakkiran, P. (2023). A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1727–1740.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135:1512–1519.

Chidambaram, M. and Ge, R. (2025). Reassessing how to compare and improve the calibration of machine learning models. In *The Thirteenth International Conference on Learning Representations*.

Dimitriadis, T., Gneiting, T., and Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences of the United States of America*, 118:e2016191118.

Dimitriadis, T., Gneiting, T., Jordan, A. I., and Vogel, P. (2024). Evaluating probabilistic classifiers: The triptych. *International Journal of Forecasting*, 40:1101–1122.

Ferro, C. A. and Fricker, T. E. (2012). A bias-corrected decomposition of the Brier score. *Quarterly Journal of the Royal Meteorological Society*, 138:1954–1960.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378.

Gneiting, T. and Resin, J. (2023). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, 17:3226–3286.

Gneiting, T., Wolffram, D., Resin, J., Kraus, K., Bracher, J., Dimitriadis, T., Hagenmeyer, V., Jordan, A. I., Lerch, S., Phipps, K., and Schienle, M. (2023). Model diagnostics and forecast evaluation for quantiles. *Annual Review of Statistics and Its Application*, 10:597–621.

Gruber, S. G. and Buettner, F. (2022). Better uncertainty calibration via proper scores for classification and beyond. In *Advances in Neural Information Processing Systems*, 35.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330.

Gupta, C. and Ramdas, A. (2022). Top-label calibration and multiclass-to-binary reductions. In *International Conference on Learning Representations*.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110:457–506.

Jiang, X., Osl, M., Kim, J., and Ohno-Machado, L. (2011). Smooth isotonic regression: a new method to calibrate predictive models. In *AMIA Joint Summits on Translational Science Proceedings*, pages 16–20.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Kumar, A., Liang, P. S., and Ma, T. (2019). Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, 32.

Machado, A. F., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024). Probabilistic scores of classifiers, calibration is not enough. Preprint, arXiv:2408.03421.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. H. S., and Dokania, P. K. (2020). Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, 33.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12:595–600.

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. (2019). Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Pohle, M.-O. (2020). The Murphy decomposition and the calibration-resolution principle: A new perspective on forecast evaluation. Preprint, arXiv:2005.01835.

Popordanoska, T., Gruber, S. G., Tiulpin, A., Buettner, F., and Blaschko, M. B. (2024). Consistent and asymptotically unbiased estimation of proper calibration errors. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 3466–3474.

Roelofs, R., Cain, N., Shlens, J., and Mozer, M. C. (2022). Mitigating bias in calibration error estimation. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 4036–4054.

Rossellini, R., Soloff, J. A., Barber, R. F., Ren, Z., and Willett, R. (2025). Can a calibration metric be both testable and actionable? In *Proceedings of Thirty Eighth Conference on Learning Theory*, pages 4937–4972.

Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology and Climatology*, 2:191–201.

Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., and Flach, P. (2023). Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112:3211–3260.

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., and Schön, T. B. (2019). Evaluating model calibration in classification. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. In *Proceedings of the British Machine Vision Conference*, pages 87.1–87.12.

Appendix

A RESULTS WITH NORMALIZED SCORE DECOMPOSITION

Table 2 and Figure 3 present results analogously to Section 3 for an alternative *normalized* Brier score decomposition obtained by separately recalibrating each class probability using isotonic regression and then normalizing the recalibrated class probabilities to obtain estimates of the conditional distributions $\widehat{p}_1, \ldots, \widehat{p}_n$ used in (4). As alluded to in the main text, conclusions are similar with the exception of the results for CW-HB, where much of the drop in overall score is attributed to miscalibration, because this decomposition quantifies deviations from the stronger notion of calibration in (1) instead of class-wise calibration.

Table 2: *Normalized* Brier score decomposition (×100) and classification accuracy (Acc) averaged across seven base models for both datasets (CIFAR-10 and CIFAR-100) and each recalibration method. The best average values per metric are highlighted in **bold**. Values in red indicate a drastic decline in predictive performance.

CIFAR-10					
Method	BS	MCB	DSC	UNC	Acc
base	0.781	0.059	8.279	9	**0.951**
CW-HB	1.832	0.604	7.772	9	0.890
N-HB	1.237	0.080	7.842	9	0.890
CW-IR	0.779	**0.030**	8.251	9	0.950
N-IR	0.770	0.045	8.275	9	0.950
CW-SIR	0.769	0.048	8.279	9	0.950
N-SIR	**0.764**	0.044	**8.280**	9	0.950

CIFAR-100					
Method	BS	MCB	DSC	UNC	Acc
base	0.327	0.032	0.696	0.99	**0.775**
CW-HB	0.865	0.331	0.457	0.99	0.483
N-HB	0.532	0.038	0.497	0.99	0.483
CW-IR	0.333	**0.024**	0.680	0.99	0.770
N-IR	0.324	0.029	0.694	0.99	0.770
CW-SIR	0.323	0.029	0.696	0.99	0.773
N-SIR	**0.320**	0.028	**0.699**	0.99	0.773

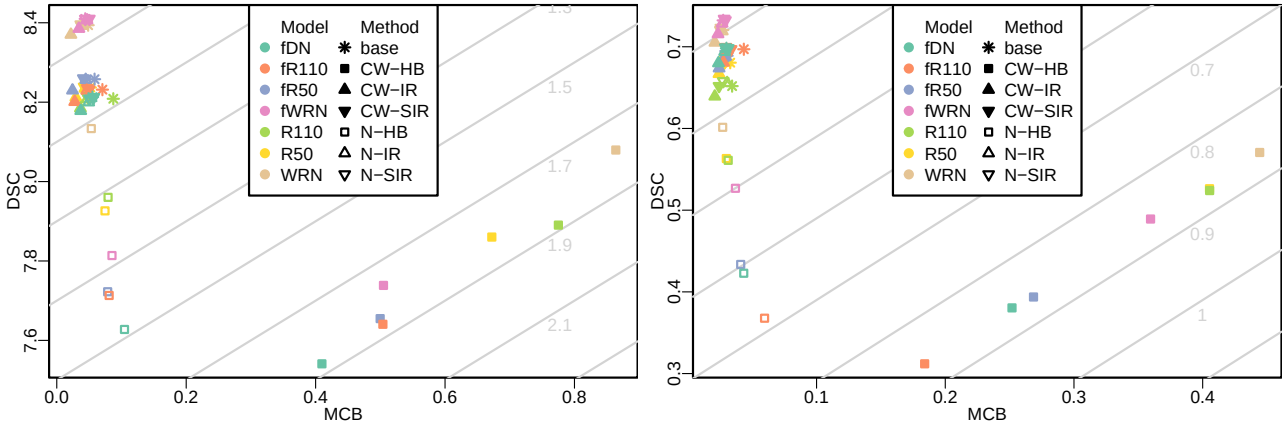

Figure 3: MCB-DSC plots for CIFAR-10 (left) and CIFAR-100 (right) showing *normalized* Brier score decompositions (×100). The plots show results for seven base models that differ in architecture (**R**esNet-**50**, **R**esNet-**110**, **W**ide**R**es**N**et, and **D**ense**N**et-121) and training loss (**f**ocal loss or Brier score). The gray isolines correspond to MCB-DSC pairs that yield the same overall Brier score.

B SMOOTHING OF ISOTONIC REGRESSION FITS

This section provides a brief description of the employed smoothing. Let $(p_1, y_1), \ldots, (p_n, y_n)$ be a sample of binary class probabilities and corresponding binary outcomes encoded as 0 or 1, where p_i is the predicted probability of the outcome $y_i = 1$, and assume w.l.o.g. that the sample is ordered by the the predictions, i.e., $p_1 \leq p_2 \leq \cdots \leq p_n$. Isotonic regression via the pool-adjacent violators algorithm returns a partition of the sample point indices into consecutive pools $B_1, \ldots, B_m \subset \{1, \ldots, n\}$ with separating indices $1 = i_1 < i_2 < \cdots < i_m < i_{m+1} = n + 1$ such that $B_j = \{i_j, i_j + 1, \ldots, i_{j+1} - 1\}$. For pool B_j, let $\bar{y}_j = \frac{1}{|B_j|} \sum_{i \in B_j} y_i$ be the respective fitted value and m_j be a median of the predictions in the pool, i.e., any value such that $|\{p_i \in B_j \mid p_i < m_j\}|/|B_j| \leq \frac{1}{2} \leq |\{p_i \in B_j \mid p_i \leq m_j\}|/|B_j|$. Additionally, set $\bar{y}_0 = m_0 = 0$ and $\bar{y}_{m+1} = m_{m+1} = 1$. Then the smoothed isotonic regression fit at $p \in [m_j, m_{j+1}]$ is given by

$$f(p) = \frac{\bar{y}_{j+1} - \bar{y}_j}{m_{j+1} - m_j}(p - m_j) + \bar{y}_j,$$

which is simply the piecewise linear interpolator of the points $(m_0, \bar{y}_0), \ldots, (m_{m+1}, \bar{y}_{m+1})$. In contrast, Jiang et al. (2011, Algorithm 1) "sample" the anchor points m_j from the intervals $[\min_{i \in B_j} p_i, \max_{i \in B_j} p_i]$ and apply piecewise cubic Hermite interpolation to obtain a smoothed fit. The resulting function is differentiable, while the simple piecewise linear interpolation only makes the regression function strictly increasing on $[p_1, p_n]$ to avoid mapping distinct predictions to a single recalibrated value.