Supplemental Materials for: *Mix-n-Match*: Ensemble and Compositional Methods for Accurate and Data-efficient Uncertainty Calibration

S1. Proofs of Proposition 3.1: Accuracy-Preserving Calibration Maps

For an arbitrary pair of classes ($\forall i, j \in [L]$, where $[L]$ denotes the set of positive integers up to L) of the prediction probability vector z, let us assume that $z_i < z_j$ without loss of generality. By definition, the strictly isotonic function g will output $g(z_i) < g(z_j)$ after the transformation. After dividing by the same normalization constant $G(z)$, the following relationship holds: $[T(z)]_i < [T(z)]_j$, where $[.]_i$ represents the i-th entry of the vector. Since the order of entries in the prediction vector is unchanged after the calibration, the classification accuracy is preserved.

S2. Proofs of Theorem 4.1: Statistical Properties of KDE-based ECE

Mirror image KDE for boundary correction. Considering that we work with the probability simplex Δ^L in the context of calibration, KDE can suffer from excessively large bias near the boundary of the simplex. To suitably correct for boundary bias without compromising the estimation quality, we adopt the *mirror image* KDE strategy (Singh & Póczos, 2014b). The convergence/consistency properties will be proved for such a choice.

Smoothness assumption on the underlying densities. Let β and N be positive numbers. Given a vector $s = (s_1, \dots, s_L)$ with non-negative integer entries, let us use $D^s := \frac{\partial^{\|s\|_1}}{\partial^{s_1} z_1 \dots \partial^{s_L} z_L}$ to denote the differential operator. The β-Hölder class of densities $\Sigma(\beta, N)$ contains those densities $p : [0,1]^L \to \mathbb{R}$ satisfying the following relationship:

$$|D^s p(z) - D^s p(z')| \leq N \|z - z'\|^{\beta - \|s\|_1},$$

for all $z, z' \in \mathcal{Z}$ and all s with $\|s\|_1 = \beta - 1$. We assume that the distribution for predictions $p(z)$ as well as the calibration probabilities π_l for each class $l \in [L]$ belongs to the β-Hölder class.

Assumption on the kernel function. We assume that the kernel function $K : \mathbb{R} \to \mathbb{R}_{\geq 0}$ has bounded support $[-1, 1]$ and satisfies:

$$\int_{-1}^{1} K(u)\, du = 1; \|K\|_1 = \int_{-1}^{1} |K(u)|\, du < \infty; \forall j \in [\beta - 1], \int_{-1}^{1} u^j K(u)\, du = 1.$$

Boundedness assumption. We denote $C_\pi := \sup_z \|z - \pi(z)\|_1$ and $C_z := \sup_z \tilde{p}(z)$ and assume they are both finite.

We can bound the KDE estimation error of $|\mathrm{ECE}(f) - \widetilde{\mathrm{ECE}}(f)|$ after applying the triangle inequality:

$$\begin{aligned} |\mathrm{ECE}(f) - \widetilde{\mathrm{ECE}}(f)| &= \left| \int \|z - \pi(z)\|_1 p(z)\, dz - \int \|z - \tilde{\pi}(z))\|_1 \tilde{p}(z)\, dz \right| \\ &\leq \int \left| p(z) \|z - \pi(z)\|_1 - \tilde{p}(z) \|z - \pi(z)\|_1 \right| dz + \int \left| \tilde{p}(z) \big(\|z - \pi(z)\|_1 - \|z - \tilde{\pi}(z)\|_1 \big) \right| dz \\ &\leq \sup_z \|z - \pi(z)\|_1 \int |p(z) - \tilde{p}(z)|\, dz + \sup_z \tilde{p}(z) \int \|\pi(z) - \tilde{\pi}(z)\|_1\, dz \\ &\leq C_\pi \int |p(z) - \tilde{p}(z)|\, dz + C_z \int \|\pi(z) - \tilde{\pi}(z)\|_1\, dz \end{aligned} \tag{S1}$$

which connects the absolute estimation error of ECE to the integrated estimation errors on the unknown densities $p(z)$ and $\pi(z)$. We then borrow the established convergence rate and consistency proofs for (conditional) density functional of mirror KDE (Singh & Póczos, 2014a) to derive the statistical properties for the proposed KDE-based ECE estimator.

Bias convergence rate. Taking the expectation and applying the Fubini's theorem on both sides in Eq. (S1), we can derive:

$$\begin{aligned}\mathbb{E}|\mathrm{ECE}(f)-\widetilde{\mathrm{ECE}}(f)| &\le C_\pi \int \mathbb{E}|p(z)-\tilde{p}(z)|\,dz + C_z \int \mathbb{E}\|\pi(z)-\tilde{\pi}(z))\|_1\,dz \\ &\le C_\pi C_{B1}(h^\beta + h^{2\beta} + \frac{1}{n_e h^L}) + C_z C_{B_2}(h^\beta + h^{2\beta} + \frac{1}{n_e h^L}) \le C(h^\beta + h^{2\beta} + \frac{1}{n_e h^L}),\end{aligned}$$

where C_{B1} and C_{B2} are constants given the sample size n_e and bandwidth h and $C = C_\pi C_{B1} + C_z C_{B2}$. The quantity $h^{2\beta}$ is introduced by the Bias Lemma (Singh & Póczos, 2014b) from the mirror image KDE. For $\int \mathbb{E}|p(z)-\tilde{p}(z)|\,dz$, we follow the standard KDE results (see Prop 1.1,1.2 and 1.2.3 (Tsybakov, 2008)) while the bound on the other term $\int \mathbb{E}\|\pi(z)-\tilde{\pi}(z))\|_1\,dz$ follows 6.2 in (Singh & Póczos, 2014a) or (Döring et al., 2016; Györfi et al., 2006).

The optimal bandwidth is $h \asymp n_e^{-1/(\beta+L)}$, leading to a convergence rate of $O\big(n_e^{-\beta/(\beta+L)}\big)$.

Consistency. Let $\tilde{p}'$ denote the KDE marginal density of z when an existing sample point is replaced by a new sample from the same distribution $p(z)$, and similarly $\tilde{\pi}'$ denote the KDE canonical calibration function after replacing a sample. Following (Singh & Póczos, 2014a), we can bound the density discrepancy before/after replacing a single sample by:

$$\int |\tilde{p}(z)-\tilde{p}'(z)|\,dz \le \frac{C_{V1}}{n_e}; \int \|\tilde{\pi}(z)-\tilde{\pi}'(z)\|_1\,dz \le \frac{C_{V2}}{n_e}, \tag{S2}$$

where C_{V1} and C_{V2} are constants in the class-dimension L and the kernel norm $\|K\|_1$ for exact values. Suppose that we use two sets of n_e independent samples to estimate p and π, respectively. Since $\widetilde{\mathrm{ECE}}(f)$ depends on $2n_e$ independent variables, combining Eq. (S2) with Eq. (S1), we can use McDiarmid's Inequality (McDiarmid, 1989) to derive that:

$$\mathbb{P}(|\widetilde{\mathrm{ECE}}(f)-\mathbb{E}\widetilde{\mathrm{ECE}}(f)|>\varepsilon) \le 2\exp\bigg(-\frac{2\varepsilon^2}{2n_e(2C_V/n_e)^2}\bigg) = 2\exp\bigg(-\frac{\varepsilon^2 n_e}{4C_V^2}\bigg).$$

for $C_V = \max(C_{V1}, C_{V2})$. As $\mathbb{P}(|\widetilde{\mathrm{ECE}}(f)-\mathbb{E}\widetilde{\mathrm{ECE}}(f)|>\varepsilon)$ approaches 0 when $n_e \to \infty$, the KDE-based ECE estimator is consistent.

S3. KDE Implementation Detail

Kernel function choice. Different types of kernel functions $K(u)$ can be used, such as the Gaussian and Epanechnikov functions. Our choice is the Triweight Kernel $K_h(u) = (1/h)\frac{35}{32}(1-(u/h)^2)^3$ on $[-1,1]$, since it has been recommended for problems with limited support interval (de Haan, 1999).

Bandwidth selection. We use the popular rule-of-thumb $h = 1.06\hat{\sigma}n_e^{-1/5}$ (Scott, 1992), where $\hat{\sigma}$ is the standard deviation of the samples.

S4. Proof for Proposition 4.2: Calibration Gain for Accuracy-Preserving Methods

According to the *calibration refinement decomposition* (Murphy, 1973), the expected calibration error ECE^2 is equal to:

$$\mathrm{ECE}^2(f) = \mathbb{E}\|z-\pi(z)\|_2^2 = \mathbb{E}\|z-y\|_2^2 - \mathbb{E}\|\pi(z)-y\|_2^2, \tag{S3}$$

where $\mathbb{E}\|z-y\|_2^2$ is the standard square loss and $\mathbb{E}\|\pi(z)-y\|_2^2$ is the *refinement error* (Murphy, 1973) that penalizes the existence of inputs sharing the same prediction but different class labels. Before proceeding further, we first introduce the definition of *injective* calibration maps:

Definition S4.1 (Injective Calibration Map)**.** The calibration map is *injective* if different prediction vectors remain different after calibration: $\forall z, z' \in \mathcal{Z}, T(z) \neq T(z)$ if $z \neq z'$.

Proposition S4.1. *The accuracy-preserving calibration map T in Def. 3.1is injective.*

Proof. Given $z \neq z'$, without loss of generality assume $G(z) \ge G(z')$ for their normalization constants. Since $z \neq z'$, there must exists at least one class l where $z_l < z'_l$. After the transformation by a strictly isotonic function g, we know that $g(z_l) < g(z'_l)$. Then we can derive that:

$$[T(z)]_l - [T(z')]_l = \frac{g(z_l)}{G(z)} - \frac{g(z'_l)}{G(z')} = \frac{g(z_l)G(z') - g(z'_l)G(z)}{G(z)G(z')} < 0.$$

Therefore, $T(z) \neq T(z')$ because their l-th entry is not equal. The calibration map is then injective.

Note that the canonical calibration function in Eq. (1) is essentially the conditional expectation of binary random variables Y, as $\pi_l(z) = \mathbb{P}[Y_l = 1|f(X) = z] = \mathbb{E}[Y_l|f(X) = z]$. By elementary properties of the conditional expectation, one can easily show that injective calibration maps will not change the canonical calibration probabilities, thus $\pi(z) = \pi(T(z))$ for injective T. Combining this with the decomposition relationship in Eq. (S3), we can show that:

$$\begin{aligned}\Delta\mathrm{ECE}^2(T) &= \mathrm{ECE}^2(f) - \mathrm{ECE}^2(T \circ f)\\ &= \mathbb{E}\|z-y\|_2^2 - \mathbb{E}\|\pi(z)-y\|_2^2 - \left(\mathbb{E}\|T(z)-y\|_2^2 - \mathbb{E}\|\pi(T(z))-y\|_2^2\right)\\ &= \mathbb{E}\|z-y\|_2^2 - \mathbb{E}\|T(z)-y\|_2^2.\end{aligned} \tag{S4}$$

Therefore, after applying an injective calibration map, which include the proposed accuracy-preserving ones, any changes in the squared loss will be due to the change in ECE^2. □

Remark. Most existing calibration methods are not injective. For example, in histogram binning (Zadrozny & Elkan, 2001) or the original isotonic regression method (Zadrozny & Elkan, 2002), all predictions inside certain intervals will be mapped to be identical, and violate the injective requirement. For parametric methods, such as vector, matrix (Guo et al., 2017), or Dirichlet scaling (Kull et al., 2019), different logits can be transformed to produce the same prediction probability vectors, and violate the injective requirement.

S5. Experimental Details and Additional Results in Section 5.1

S5.1. Experimental details

For the synthetic example, the labels (Y) and input features (X) are distributed as:

$$\mathbb{P}(Y_1 = 1) = \mathbb{P}(Y_2 = 1) = 1/2; \mathbb{P}(X = x|Y_1 = 1) = \mathcal{N}(x; -1, 1); \mathbb{P}(X = x|Y_2 = 1) = \mathcal{N}(x; 1, 1). \tag{S5}$$

The probability of observing the label $Y_1 = 1$ conditioned on the input x can be written as:

$$\mathbb{P}(Y_1 = 1|X = x) = 1/[1 + \exp(2x)].$$

We assume the prediction models to be in the following form, parameterized by β_0 and β_1:

$$z = f(x) = (z_1, z_2) = \left(\frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}, \frac{\exp(-\beta_0 - \beta_1 x)}{1 + \exp(-\beta_0 - \beta_1 x)}\right).$$

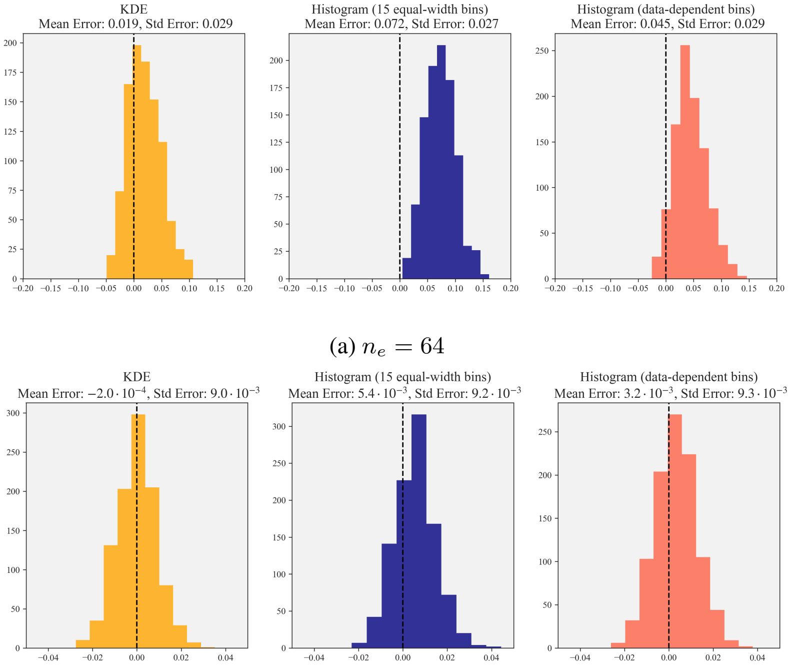
This leads to close-form expressions for the canonical calibration functions $\pi(z) = (\pi_1(z), \pi_2(z))$:

$$\pi_1(z) = [1 + \exp(-2\frac{\beta_0 + \log(1/z_1 - 1)}{\beta_1})]^{-1}, \pi_2(z)) = 1 - \pi_1(z). \tag{S6}$$

Finally, we estimate the ground-truth ECE^d based on Monte Carlo integration: (i) generate 10^6 random input-output sample pairs according to Eq. (S5), and (ii) record the sample average value of the quantity $|z_1 - \pi_1(z)|^d$ as the ground-truth.

S5.2. Additional results

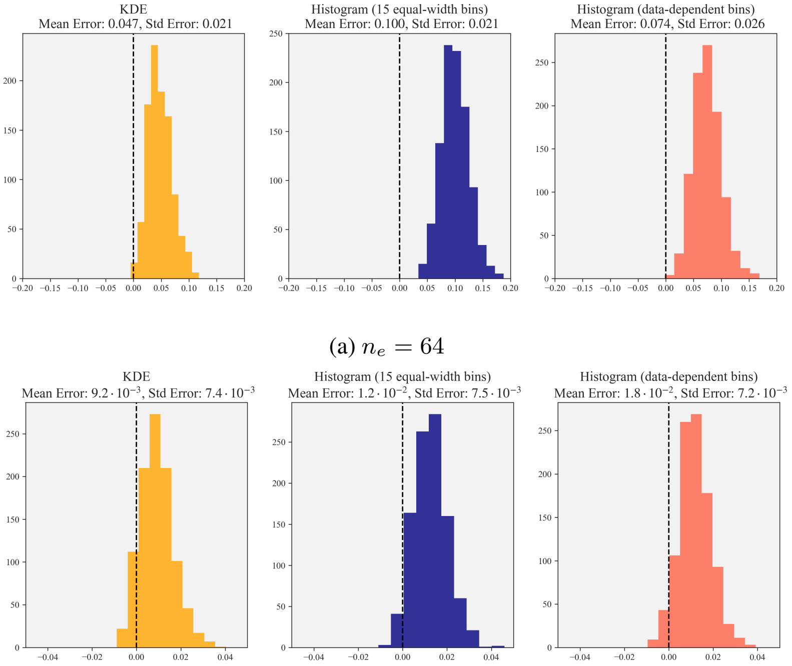
We plot the distribution of the errors between the ECE estimates and the ground-truth ECE in two representative scenarios: a data-limited scenario with $n_e = 64$ in Figure S1 and a data-rich scenario with $n_e = 1024$ in Figure S2. The KDE estimation errors are generally less biased (more concentrated around zero) as compared to histograms, corroborating the findings in Sec. 5.1. Judging from the variance of the estimation errors, the KDE estimation errors are generally less dispersed than the two histogram estimators, indicating that the KDE estimators are more reliable. In contrast, histogram ECE estimators tend to severely over-estimates ECE in data-limited regime, with the majority of their estimation errors being positive. Their sensitivity to the binning schemes can be also observed from the distribution discrepancies between using equal-width and data-dependent bins: the histogram estimator with data-dependent bins generally performs better than the one with equal-width bins, although it cannot reach the accuracy level of KDE estimators. However, it performs the worst in the data-rich scenario of Case 2 (Figure S2 bottom).



(a) $n_e = 64$

(b) $n_e = 1024$

Figure S1. Distribution of ECE estimation errors in Case 1: $\beta_0 = 0.5, \beta_1 = -1.5$ with (a) $n_e = 64$ and, (b) $n_e = 1024$.



(a) $n_e = 64$

(b) $n_e = 1024$

Figure S2. Distribution of ECE estimation errors in Case 2: $\beta_0 = 0.2, \beta_1 = -1.9$ with (a) $n_e = 64$, and (b) $n_e = 1024$.

Recently (Kumar et al., 2019) proposed a *debiased* histogram-based estimator for ECE$^{d=2}$, by leveraging error cancellations across different bins. We compare the proposed KDE ECE estimator with the debiased versions of histogram-based ECE estimators with both equal-width and data-dependent binning schemes. We vary the size of evaluation samples n_e from 64 to 1024 and plot the mean absolute error (averaged over 1000 independent experiments) between KDE/Debiased-Histograms estimates for ECE^2 and the ground truth in Figure S3. We observe that KDE-based estimator consistently performs better than the best-performing debiased histogram-based ECE estimators. This agrees with the findings in Sec. 5.1 and confirms the advantage of using KDE-based estimator over histograms-based estimators.

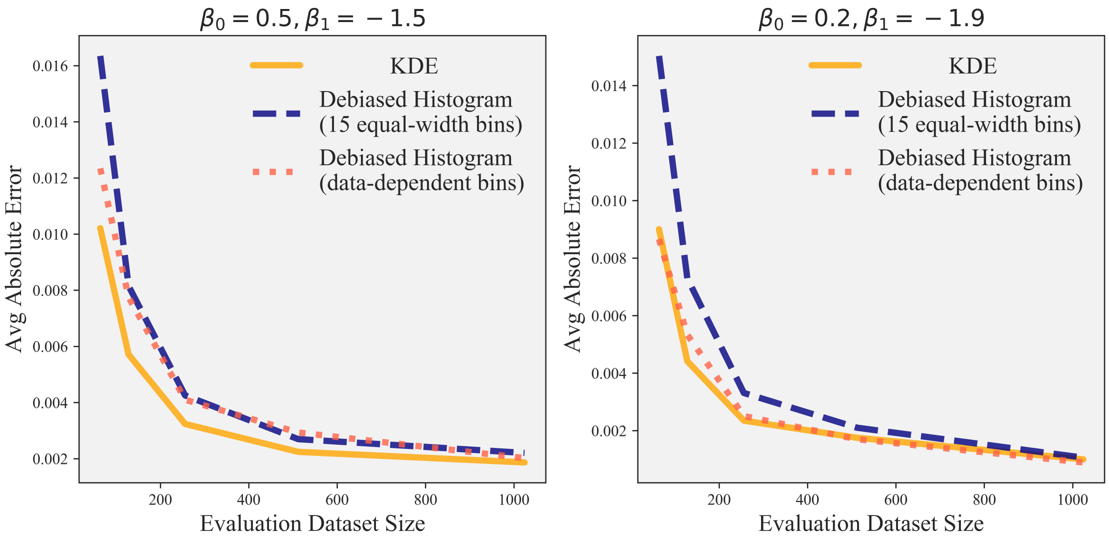


Figure S3. Average absolute error for KDE ECE and debiased histogram ECE estimators as a function of the number of samples used in the ECE estimation. KDE-based ECE estimator achieves lower estimation error, especially when the evaluation dataset is small.

S6. Experimental Detail and Additional Results in Section 5.2

S6.1. Training details

For training neural networks on CIFAR-10/100, we use SGD with Nesterov momentum and the cross-entropy loss. The weight decay is set to 0.0005, dampening to 0, momentum to 0.9, and minibatch size to 128. The initial learning rate is set to 0.1, and is dropped by a factor of 0.2 at 60, 120 and 160 epochs. For Wide ResNets, we use a dropout rate of 0.3. We train the models for a total of 500 epochs. We use standard mean/std normalization with flipping and data cropping augmentation as described in (Zagoruyko & Komodakis, 2016) on CIFAR-10/100 images.

S6.2. Expanded results

The quantitative measure of expressive power and data-efficiency is illustrated graphically in Figure S4 and discussed in Table S1 and Table S2. To summarize, ETS is comparably expressive to TS on CIFAR-10 and noticeably more expressive on CIFAR-100 and ImageNet. Both IRM and IROvA-TS are more efficient than IROvA. The relative efficiency gain of IRM increases as the problems become more complex. On the other hand, the relative efficiency gain of IROvA-TS appears to be quite stable on a wide range of problems.

We also provide expanded results on the learning curve analysis for additional neural network classifiers (see Figure S5 to Figure S7). From the visual comparison of the ECE learning curves of ETS and TS, or IRM/IROvA-TS and IROvA, we can confirm the importance to preserve the classification accuracy and the consistent benefit of employing the proposed *Mix-n-Match* strategies. Overall, in data-limited regime, parametric variants (TS, ETS) perform better than traditional non-parametric variants (IROvA, IROvA-TS) due to their high data-efficiency. ETS significantly outperforms TS and performs the best as the added expressive power from ensembles allow ETS to make further descent on ECE. The proposed accuracy-preserving non-parametric variant IRM also performs good: it is sometimes more effective than TS, although

Table S1. ECE[1] (%) with $n_c = 10000$ for CIFAR and $n_c = 45000$ for ImageNet; lower values imply more expressive power.

Dataset	Model	TS	ETS
CIFAR-10	DenseNet 40	1.32	1.32
CIFAR-10	LeNet 5	1.49	**1.48**
CIFAR-10	ResNet 110	2.12	2.12
CIFAR-10	WRN 28-10	1.67	1.67
CIFAR-100	DenseNet 40	1.73	**1.72**
CIFAR-100	LeNet 5	1.39	**1.31**
CIFAR-100	ResNet 110	2.81	**2.15**
CIFAR-100	WRN 28-10	3.23	**2.85**
ImageNet	DenseNet 161	1.95	**1.75**
ImageNet	ResNeXt 101	2.97	**2.22**
ImageNet	VGG 19	1.89	**1.83**
ImageNet	WRN 50-2	2.52	**2.10**

Table S2. Required calibration data amount n_c to reach IRM's performance with $n_c = 128$ samples; lower value means more data-efficient. All values are normalized (divided by 128) to show how many samples are equivalent to one sample in IRM for each method.

Dataset	Model	IRM	IROvA	IROvA-TS
CIFAR-10	DenseNet 40	**1.0**	2.45	2.10
CIFAR-10	LeNet 5	**1.0**	3.15	2.92
CIFAR-10	ResNet 110	**1.0**	1.84	1.70
CIFAR-10	WRN 28-10	**1.0**	1.98	1.90
CIFAR-100	DenseNet 40	**1.0**	37.6	17.9
CIFAR-100	LeNet 5	**1.0**	58.7	43.0
CIFAR-100	ResNet 110	**1.0**	28.1	10.7
CIFAR-100	WRN 28-10	**1.0**	24.2	15.6
ImageNet	DenseNet 161	**1.0**	282	174
ImageNet	ResNeXt 101	**1.0**	226	98.6
ImageNet	VGG 19	**1.0**	251	180
ImageNet	WRN 50-2	**1.0**	258	161

it cannot outperform ETS in most examined cases. The relatively good performance of IRM can be accredited to its high data-efficiency on complex calibration tasks. Going to the data-rich regime, the ECE reduction progress stalls for parametric methods. On complex problems, such as CIFAR-100 and ImageNet, this also applies to the accuracy-preserving non-parametric variants (IRM) due to its expressivity-efficiency trade-off. In contrast, the high expressive power of non-parametric variants (IROvA, IROvA-TS) allow them to keep minimizing the ECE and eventually outperform less expressive methods (ETS, TS or IRM) with sufficient amount of data – although, this cannot be verified in all the examined cases due to our limited data budget. In such regime, the composition method IROvA-TS significantly outperforms IROvA and performs the best due to its enhanced data-efficiency. Based on such observations, our further discussion on the guidelines of calibration methods will be restricted to ETS, IRM and IROvA-TS.

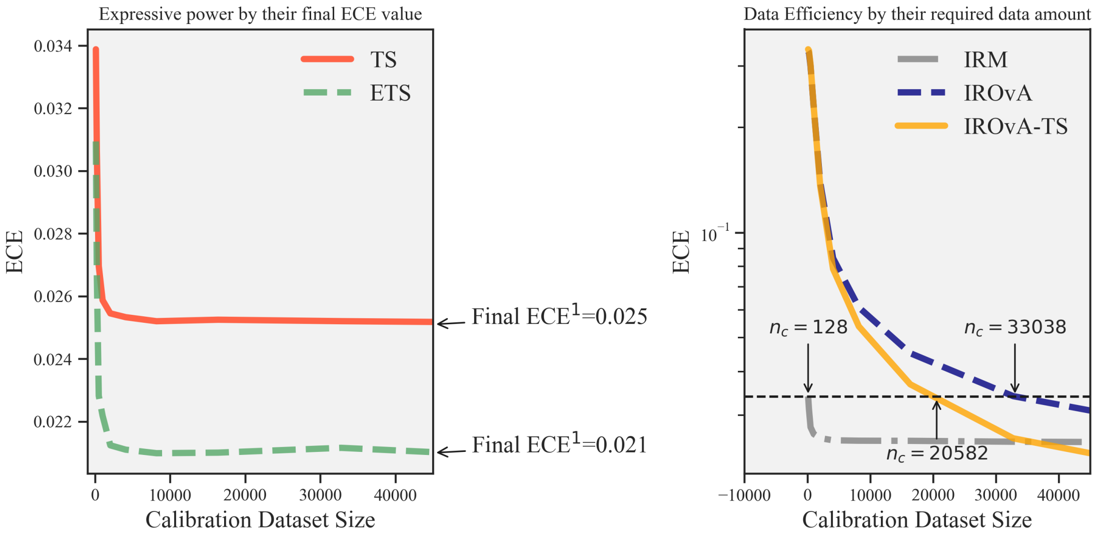
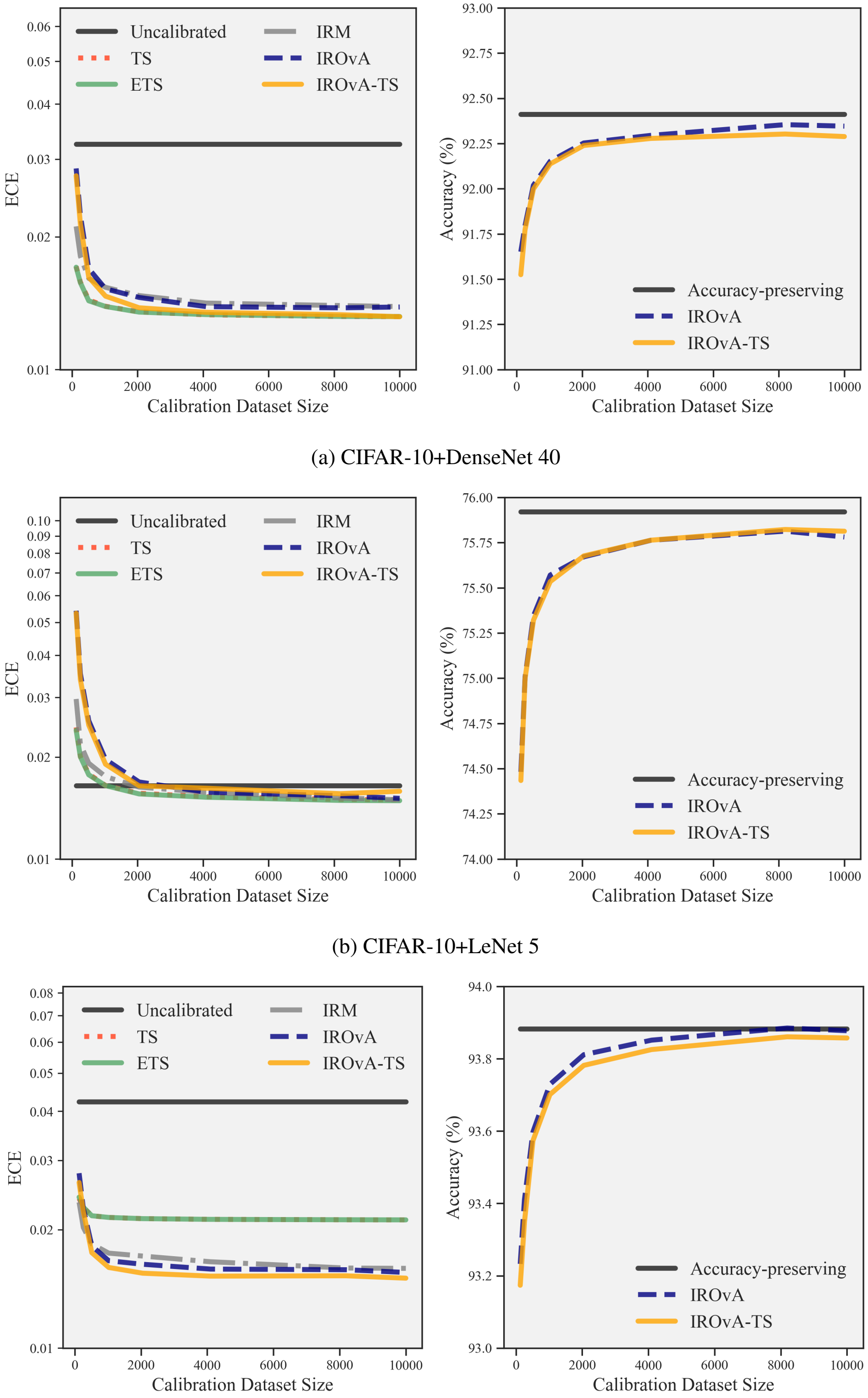


Figure S4. Graphical illustration for the proposed measure of expressive power and data-efficiency.

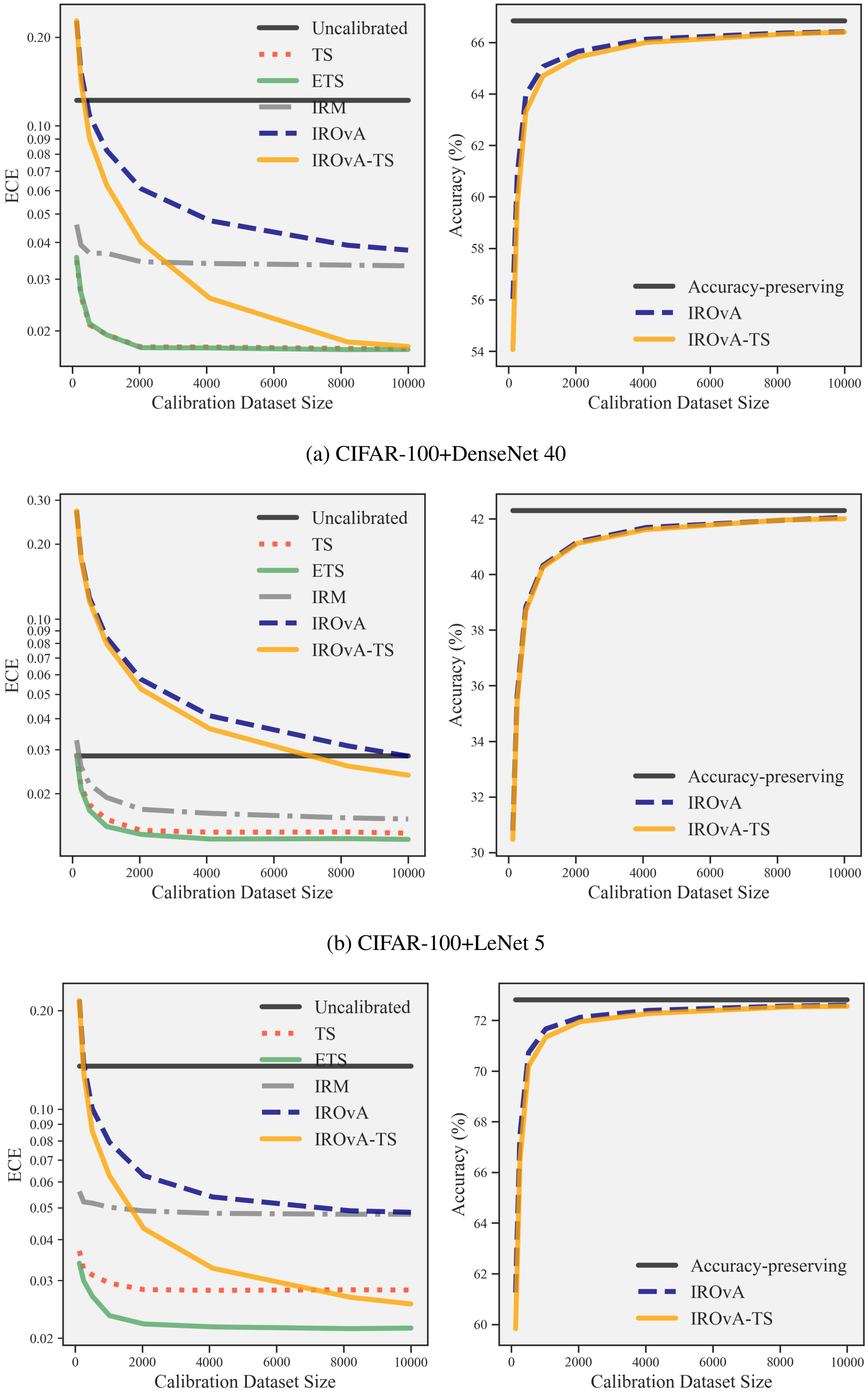


(a) CIFAR-10+DenseNet 40

(b) CIFAR-10+LeNet 5

(c) CIFAR-10+ResNet 110

Figure S5. Learning curve comparisons of various calibration methods on top-label ECE[1] (left) and the classification accuracy (right) on CIFAR-10 dataset with (a) DenseNet 40 model; (b) LeNet 5 model; and (c) ResNet 110 model.

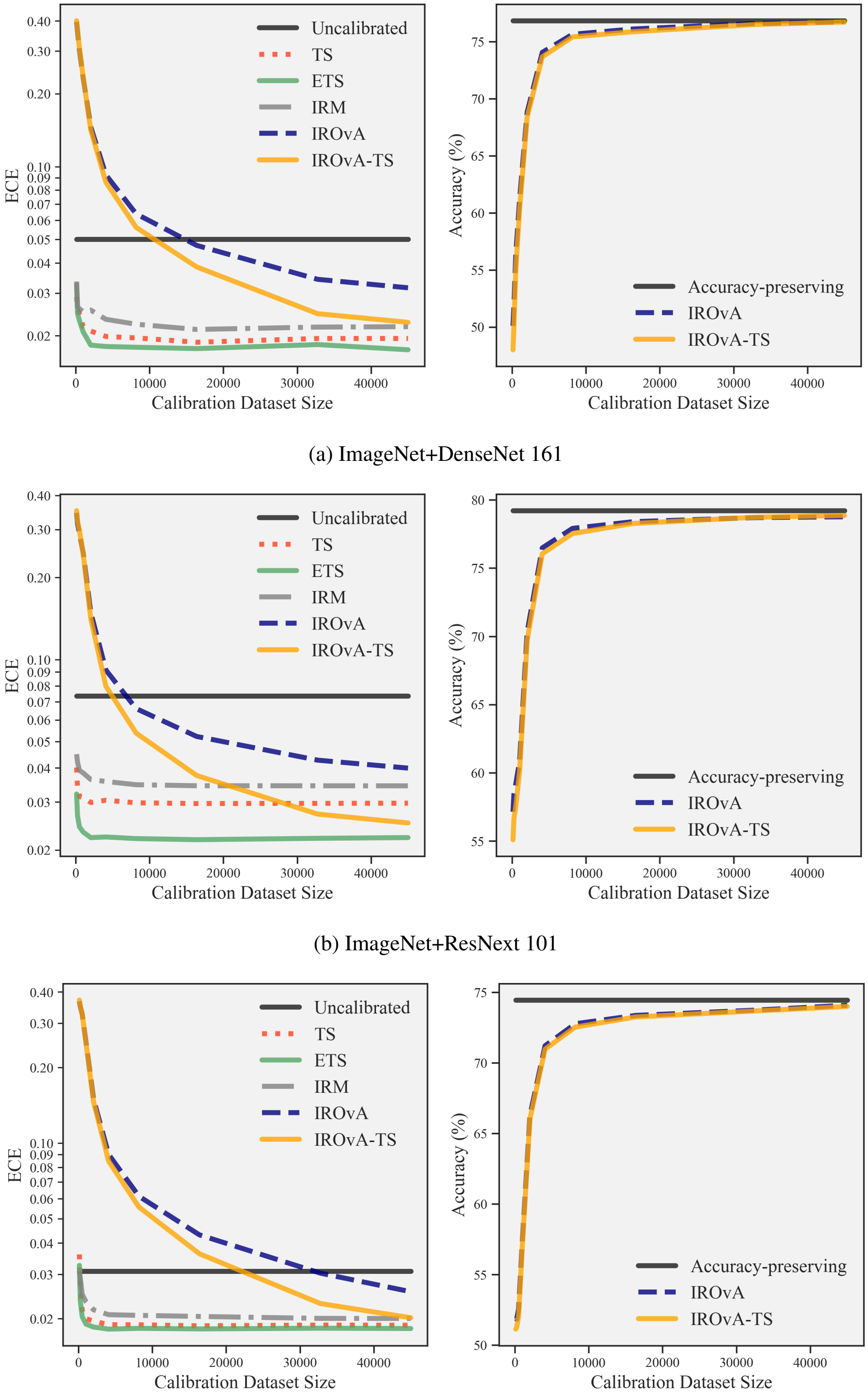


(a) CIFAR-100+DenseNet 40

(b) CIFAR-100+LeNet 5

(c) CIFAR-100+ResNet 110

Figure S6. Learning curve comparisons of various calibration methods on top-label ECE[1] (left) and the classification accuracy (right) on CIFAR-100 dataset with (a) DenseNet 40 model; (b) LeNet 5 model; and (c) ResNet 110 model.



(a) ImageNet+DenseNet 161

(b) ImageNet+ResNext 101

(c) ImageNet+VGG19

Figure S7. Learning curve comparisons of various calibration methods on top-label ECE[1] (left) and the classification accuracy (right) on ImageNet dataset with (a) DenseNet 161 model; (b) ResNext 101 model; and (c) VGG 19 model.

S7. Guidelines

First, we provide guidelines on choosing an appropriate calibration evaluation metric. If knowing the exact value of ECE is important, we recommend KDE-based top-label ECE estimator for its superior data-efficiency as compared to histograms. If the goal is to infer just the rankings not actual calibration errors, one should use the calibration gain metric. It provides a reliable and faithful comparison of different methods based on their actual calibration capabilities. We also note that calibration gain metric might be a lower bound for certain calibration methods, e.g., non accuracy-preserving methods.

We next provide general guidelines on selecting the best calibration method (ETS vs. IRM vs. IROvA-TS), based on: (a) the complexity of the calibration task which is a function of the model complexity (number of free parameters) and the data complexity (number of classes), and (b) resources at hand (the amount of the calibration data).

The complexity of the calibration task is directly related to the complexity of the canonical calibration function in Eq. (1). Although, we do not have the knowledge of the canonical calibration function, we expect a learning task with low model complexity and low data complexity to result in a low complexity calibration task (see Figure S5 (b)). Next, a learning task with low model complexity but high data complexity (Figure S6 (b)), or high model complexity but low data complexity (Figure S5 (a) and (c)) is expected to result in a moderately complex calibration task. Finally, we expect a learning task with high model & data complexity to result in a highly complex calibration task (Figure S6 (a) and (c), Figure S7).

For low complexity calibration tasks, we see that the performance of uncalibrated models are already satisfactory. This observation agrees with results in (Guo et al., 2017). Further, all the calibration methods perform similarly, however, proposed variants performing slightly better than the baseline approaches. The use case of the most practical interest is where the calibration task is expected to be complex. In such scenarios, an ideal calibration map should have enough expressive power to accurately approximate the canonical calibration function in Eq. (1). However, to fit an expressive calibration map, sufficiently large amount of calibration data is required which may or may not be available. In data limited regime, ETS is recommended as the first choice, while IRM is a potential alternative when the parametric assumptions of ETS are improper (see Figure 6 top left and Figure S5 (c)). In data rich regime, we recommend using IROvA-TS for its high expressive power. For moderate complexity calibration tasks, the patterns are similar to high complexity calibration tasks. The only difference is that the gain of the proposed *Mix-n-Match* strategies are not as drastic as of the case where calibration task is of high complexity. Of course, if the user has hard-constraints on accuracy-preservation, the choice would be limited to the accuracy-preserving calibrators regardless of the data size or the task complexity. In such scenarios, we recommend ETS and IRM. We also want to emphasize that both ETS and IRM are fairly efficient and perform well on all ranges of the calibration data size.

In summary, our take-home messages on the most appropriate calibration method are the following:

- For complex calibration task (poorly calibrated model, large number of classes), when a large calibration dataset is available and the user does not have hard constraints on preserving accuracy, the proposed compositional method IROvA-TS is recommended to achieve the best degree of calibration.
- For all other cases, the proposed ensemble method ETS is recommended when the parametric assumption is proper. IRM is a strong alternative to be considered in order to avoid the risk of parametric form mis-specification of ETS in certain cases. Both approaches preserve the classification accuracy, and one can conveniently compare them based on the proposed calibration gain metric.

References

de Haan, P. On the use of density kernels for concentration estimations within particle and puff dispersion models. *Atmospheric Environment*, 33(13):2007–2021, 1999.

Döring, M., Györfi, L., and Walk, H. Exact rate of convergence of kernel-based classification rule. In *Challenges in Computational Statistics and Data Mining*, pp. 71–91. Springer, 2016.

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, 2017.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Kull, M., Nieto, M. P., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, pp. 12295–12305, 2019.

Kumar, A., Liang, P. S., and Ma, T. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, pp. 3787–3798, 2019.

McDiarmid, C. On the method of bounded differences. *Surveys in Combinatorics*, 141(1):148–188, 1989.

Murphy, A. H. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

Scott, D. Multivariate density estimation. *Multivariate Density Estimation, Wiley, New York, 1992*, 1992.

Singh, S. and Póczos, B. Exponential concentration of a density functional estimator. In *Advances in Neural Information Processing Systems*, pp. 3032–3040, 2014a.

Singh, S. and Póczos, B. Generalized exponential concentration inequality for rényi divergence estimation. In *International Conference on Machine Learning*, pp. 333–341, 2014b.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

Zadrozny, B. and Elkan, C. Learning and making decisions when costs and probabilities are both unknown. In *International Conference on Knowledge Discovery and Data Mining*, pp. 204–213, 2001.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

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