Dataset	Cross-Entropy		LS-0.05		FLSD-53 (Ours)	
	Pre T	Post T	Pre T	Post T	Pre T	Post T
CIFAR-10/SVHN	61.71	59.66	68.68	68.68	**90.83**	**90.97**
CIFAR-10/CIFAR-10-C	77.54	75.16	72.17	72.17	**85.04**	**84.70**

Cross-Entropy		LS-0.05		FLSD-53 (Ours)	
Pre T	Post T	Pre T	Post T	Pre T	Post T
61.93	59.87	68.77	68.77	**90.29**	**90.37**
77.83	75.42	72.25	72.25	**85.25**	**85.02**

Table 1: AUROC (%) for out-of-distribution detection computed using ResNet-110 for CIFAR-10 and SVHN in the first row, and CIFAR-10 and CIFAR-10-Corrupted in the second row. In the left-hand table, we use softmax entropy to get the ROC curve. In the right-hand table, we use the confidence (maximum softmax probability) to get the ROC curve.

Dataset	Cross-Entropy		Brier Loss		MMCE		LS-0.05		FLSD-53 (Ours)	
	Pre T	Post T	Pre T	Post T	Pre T	Post T	Pre T	Post T	Pre T	Post T
CIFAR-10	5.62	3.79 (2.8)	4.96	3.31 (1.2)	6.37	5.15 (2.8)	6.81	6.81 (1.0)	**4.14**	**3.30 (1.1)**
CIFAR-100	20.97	7.26 (2.3)	11.75	5.09 (1.2)	20.87	6.49 (2.3)	14.65	10.00 (1.1)	**11.32**	**5.05 (1.2)**

Table 2: L2 calibration error (%) computed for ResNet-110 on CIFAR-10 and CIFAR-100 (both pre and post temperature scaling). The optimal temperature is indicated in brackets. Best results have been marked in bold.

We thank the reviewers for their insightful comments. They found our proposed approach simple to implement (R1) and practical (R3), our thorough analysis novel and valuable (R1,R3,R5) with good theoretical motivation (R1,R3), our experiments exhaustive and our results convincing (R1,R2,R3,R5). We address their concerns below.

OOD data detection (R2, R5): Thank you for suggesting the use of the AUROC metric. In Table 1, we present the results of comparing our approach to CE and LS-0.05 (R5), using SVHN and CIFAR-10-C (corrupted using severity 5 of the Gaussian noise, refer 1807.01697) (R5) separately as OOD data. We will report results from other corruption settings in the appendix of the final version. We use softmax entropy (Fig. 5 in the main paper) and confidence (maximum softmax probability) [32] respectively to compute the AUROC. **Note that FLSD-53 even without temperature scaling performs better than other methods with temperature scaling.** Having said that, we would like to clarify here that behavior on OOD data is not the main focus of the paper, and the primary purpose of this experiment is to show that (i) temperature scaling may not work on OOD samples, and (ii) interestingly, models that are trained on focal loss can be relatively more reliable. We will clarify this claim in the main paper.

L_2 calibration error and other regularizers (R2, R5): We show the evaluation of ResNet-110 using RMS calibration error [14] in Table 2 (R2). We will include numbers for other models and datasets in the main paper. We are aware of mixup regularization [32], but regularizers like mixup are orthogonal to our approach and can be added to many methods like ours (R5). Due to time constraints, we couldn't show its results in this rebuttal.

Weight norms (R1, R5): Thanks to R1 for pointing to the recent development in the evolution of weight norms as a result of gradient norms. We agree with R1 that weight norm analysis for the initial layers is complex due to batchnorm and weight decay. However, we do see the effect of weight magnification on miscalibration, as also shown in Section C and Fig C.1 of the Appendix, where we use a simple network without batchnorm or weight decay. We use weight decay (L2 penalty) in other experiments (R1). The decrease in weight norm of the last layer up to epoch 150 can be attributed to weight decay and batchnorm (R5), as also indicated by R1. Hence one may look at weight magnification *relative* to other loss functions (see Fig. 2(e) in the main paper). As R5 expects, the feature norm of the last layer output (Figure H.1 in Appendix) is increasing at least until epoch 150. We will clarify this in the paper and release the code.

NLL+entropy regularization (R1): We did run NLL + entropy regularization [26] experiments, but found those models to perform poorly for calibration, as the models were underconfident. However, please note that [26] shows that label smoothing [19] is equivalent to this entropy regularization if the order of the KL divergence between uniform distributions and model's outputs is reversed (Sec. 3.2 of [26]), and we show the results of label smoothing as a baseline.

Threshold on p_0 (R5): We use Propositions 1,2 and Fig 3(a) to set p_0. We select the threshold p_0 such that $g(p_0, \gamma) = 1$ for a γ that does not lead to dying gradients (high γ) or insufficient regularisation (low γ). $\gamma = 3$ proved to be a good candidate and $g(0.25, 3) \approx 1$. Hence, $p_0 = 0.25$. We will clarify this.

Table 1 (R2,R5): Thanks, we will move the vanilla focal loss results and accuracy results from the supplementary to the main paper, using the extra page in the final version. We will also add the confidence intervals to Table 1.

Citations (R1,R2,R3): We will cite 1508.05154 for AdaECE, although we use absolute differences instead of RMS error. We will cite ensembling methods and other papers (R1,R3) for context and make the citations more readable.

Imbalanced datasets (R1): Thanks for the suggestion. This would be an interesting experiment, but the calibration of imbalanced datasets is a major research topic on its own. Out of necessity, we therefore leave it for future work.

Figures, typo, clarity and real-life impact (R1, R2, R3, R5): Thanks, we will make the suggested changes to make the figures more intuitive to read. We will replace Figure 4 with the CIFAR-100 results from the supplementary (Figure K.1). We will also take into account all other suggestions when preparing the final version.