We thank the reviewers for the insightful comments. Due to space limitation, we only discuss major comments below.

R1,2,5 Intuitive Example. We will add a toy example suggested by R5 at the beginning of Sec. 3 to visualize different post-hoc calibration functions. This example is shown in Fig(a) below. Here we visualize OP and OI models trained using the same experiment setting as Fig. 2 of Kull *et al.* [14] on the 3-class Abalone UCI dataset. Each colored subset in the simplex denotes a region with the same input order (i.e., class prediction order); e.g., inside the red region, we have $\mathbf{x_3} > \mathbf{x_2} > \mathbf{x_1}$. Each arrow depicts how an input is mapped by a trained calibration function. Unlike UNCONSTRAINED model that can freely map the input probabilities and possibly change the order and accuracy, OP enforces the outputs to stay within the same colored region as the inputs, but the vector fields can be different across regions. OI further keeps the function permutation invariant, enforcing the vector fields to be the same among all the 6 colored regions (as reflected in the symmetry in the visualization). This invariance property of OI can significantly reduce the hypothesis space in learning, from the functions on whole simplex to functions on one colored region, for better generalization. DIAG figure is visually similar to OI since there are only 3 classes and is not shown due to the space limitation.

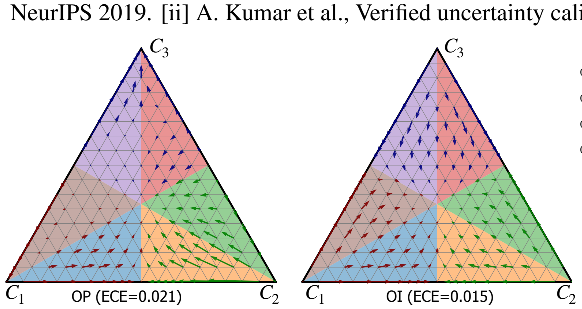
R1,3, 5 On Classwise-ECE Metric. We first note that the optimal score for classwise-ECE does not necessarily correspond to a perfect prediction as there are trivial solutions which yield optimal scores; therefore, as mentioned by R3, classwise-ECE must always be evaluated along with other proper scoring rules, such as NLL and Brier, for a meaningful comparison. This has been shown for ECE (e.g., Sec. 3 of [i], pointed out by R3), and we show the same applies to classwise-ECE in the Appendix Sec. D.1. Given this issue, DIR does not perform well overall: while DIR is superior in classwise-ECE, it is not good in NLL (Ranked 6th in Table 2) and Brier (Ranked 3rd in Table 6 in the Appendix), both of which are proper scoring rules [i]. To further understand this, in Sec. D.2 we evaluate the performance of all methods in terms of classwise Marginal Calibration Error (MCE) metric of [ii]. This metric is a debiased version of classwise-ECE metric and does not suffer from the fooling example in Sec. D.1 due to its adaptive binning scheme (see detailed discussion in D.1 and D.2). As Table 11 shows, DIAG has the best overall performance in this metric. Looking at NLL, Berier, and MCE metrics the hypothesis (brought up by R3) that our method is better on top-1 prediction while being weaker for other classes is not supported. Nonetheless, we acknowledge that future research is required to better understand the performance difference between classwise-ECE and other classwise metrics like Brier and MCE.

R1. • Showing accuracy or top-k in Table 1 is redundant since they are the same as the uncalibrated network; the proposed method is designed to enforce this constraint (which we verified in all experiments). • We are not limited to a single architecture, as cross-validation was employed for architecture search (see L298-300); Table 5 in the Appendix shows the selected architecture for each method. • In L215-216, $\mathbf{m}$ and σ are *one* factorization of $\mathbf{w}$, which we introduced to ease the implementation within deep neural networks; we still use $\mathbf{w}$ in Theorem 1 to keep the framework generic. • The order-invariant box in Fig. 2 denotes a decision box: if `No` (the order-preserving case), the network is fed with the original input; if `Yes` (the order-invariant case), the sorted input is used; We will update this figure and use the toy example above to build up the intuition before going to the math as the reviewer suggested. • *Main Message:* Our paper introduces the order-preserving family to address the accuracy drop issue in using multilayer networks for post-hoc calibration (see the performance of UNCONSTRAINED in Table 1). Searching inside this family allows one to use complex post-hoc calibration functions without losing accuracy. While DIAG works well in most experiments here, the best subfamily and architecture depend on many factors (e.g., the backbone model, metric, and calibration set size) and need to be determined by cross-validation.

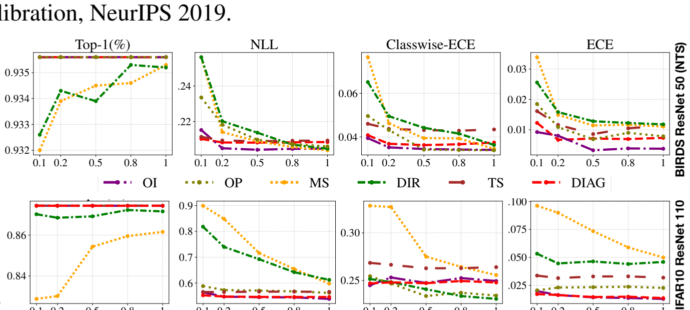
R3. We will update Sec D.1 as follows: Before giving the fooling example, we highlight that ECE is not a *proper scoring rule* based on the definition in [i] and refer to Sec. 3 of [i] for an example; at the end, we emphasize that these metrics should be used with other proper scoring rule metrics (e.g., NLL or Brier) in evaluation.

R5. • The UNCONSTRAINED results in Table 1 shows that a naive multi-layer perceptron can overfit. As requested, we analyse the methods' resilience to varying calibration set size in Fig(b). The plots show DIAG and OI methods are more stable when using a fraction of the calibration set (x-axis) compared to DIR and MS, highlighting the importance of the order-preserving family in low data regimes. We observe a similar trend in other datasets/models and will include these results in the Appendix. • We remark that the debiased ECE and classwise marginal calibration error metrics [ii] illustrated in Sec D.2 of Appendix (see Table 10 and 11 in Appendix for the evaluation) use ACE in addition to a debiasing technique to improve ECE and classwise-ECE, respectively. We will add a discussion about the importance of TACE and ACE. • We were not able to finish the OOD experiments on time and have to do it in future work.

References [i] Y. Ovadai et al., Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift, NeurIPS 2019. [ii] A. Kumar et al., Verified uncertainty calibration, NeurIPS 2019.



Fig(a) Learned calibration functions visualisation on simplex.



Fig(b) Performance vs. calibration set size