We thank the reviewers for their thoughtful reviews and feedback which will help us improve the paper. Overall, we found the reviews to be very positively worded, noting e.g. that the work is "comprehensive", "thorough", "impactful", "well-motivated", "clearly written", "addresses an important problem", etc. We thank the reviewers for their kind words and will address their comments below. We address common questions first and then address reviewers individually.

Code release: We would first like to share that our code and all model predictions from our experiments have been shared publicly online and are already being used by other research groups (for anonymity we will not share links here). We are also continuously refining this benchmark and have added a variational GP classifier as a last layer method.

"Why does SVI perform well on MNIST but poorly on every other dataset considered?" **(R2,R3)**: We agree that this should be discussed in the paper. Essentially, due to a variety of reasons (initialization, lack of priors for Bayesian neural networks, posterior collapse, optimization issues) we are not aware of a SVI variant that consistently achieves competitive accuracy as well as reliable uncertainty on large-scale problems (i.e. bigger than MNIST). Indeed, even after tremendous effort and incorporating a bag of tricks from the literature (careful initialization, empirical Bayes for the prior standard deviation), we were unable to get competitive accuracy on e.g. CIFAR. We will provide more details and analysis in the paper. Apart from optimization issues, another possibility is that SVI works well for datasets with well-specified models (e.g. models on MNIST achieve more than 99% accuracy), and underperforms on datasets where model is mis-specified. See also "Bootstrap prediction and Bayesian prediction under misspecified models" (Fushiki 2005) for theoretical arguments. To test this hypothesis, we will add an experiment on MNIST where we increase model mis-specification by reducing capacity (and accuracy) and compare the performance of ensemble and SVI.

R1: *"some of the baselines, seem to not be properly tuned... For example, for ImageNet and Skew Intensity 1, LL-SVI and Dropout underperform compared to Vanilla, ..."* In Sec A.7 we detail the tuning methodology that was used across models. However, R1 brings up a good point w.r.t. model capacity. We fixed the size of the models to have an "apples to apples" comparison, but dropout may require larger capacity models. Note, LL-SVI was trained with the rest of the model (under the ELBO) and not post-hoc which may work better. We will include these experiments in the final paper.

R2: *"In my opinion, the significance is somewhat reduced because only neural network models are used"* We empathize and agree that other models may allow for more principled uncertainty. However, given the widespread use of deep networks, we believe that understanding and benchmarking the uncertainty of this class of models is important. We struggled to find alternatives that were competitive across these tasks. Nevertheless, scaling up other models is a priority for us and we trained a variational GP on raw MNIST (incl a variety of kernels) to address this. However, we found that they were not competitive in accuracy. We will add this (and on Criteo) to the camera ready.

"Should one create one's own skew and OOD datasets? Are there any principles that are important to keep in mind while doing that? How hard will it be to use your eventually released code to do that?" We hope these experiments will generalize, but we believe it should be quite easy to add new datasets and experiments to our code (which has now been released). We will add details of how to do this, along with recommendations, to the paper.

"... Does more capacity mean better OOD calibration?" We discuss the tradeoffs in terms of memory and computational complexity in the appendix section A.8. However, the question of whether the additional capacity of ensembling gives an advantage is an interesting one. We will add an experiment to the paper where we train and evaluate higher capacity networks for the dropout and vanilla methods to explore this.

"details about ensemble": we used just a plain ensemble (no adversarial training, no heteroscedastic loss).

R3: *"Have the authors investigated the performance on different OOD datasets that are more / less similar to the source dataset...?"* We used the corruptions benchmark (Hendrycks and Dietterich, 2019) as it allows us to control the amount of similarity to source dataset, and see if consistent orderings are observed.

"analysis / actionable takeaways." This was limited due to space constraints, but we will add this to the supplement.

"I found the experimental section somewhat disorganized ... would have made for a significantly better read." Thanks for pointing this out. We will address the layout and presentation, particularly w.r.t. these figures, for the camera ready.

"... While SVI does seem to have lower Brier scores with shift, accuracies don't appear to be any better – how was this observation made?" Valid point. This wasn't well worded, but we meant specifically with regard to uncertainty as it's significantly better in Brier score but not accuracy - i.e. much better calibrated under significant shift. We will reword.

"... What are these additional properties that ECE and entropy capture?" Each proper scoring rule induces a calibration measure, see [Brocker, "Reliability, sufficiency, and the decomposition of proper scores", Quarterly Journal of the Royal Meteorological Society, 2009]. However, ECE is not the result of such decomposition and has no corresponding proper scoring rule; we instead chose to include ECE because it is popularly used. Each proper scoring rule is also associated with a corresponding entropy function and Shannon entropy is that for log probability [see Gneiting & Raftery, Journal of the American Statistical Association, 2007]. We will reword this and clarify our reasoning for the camera ready.