We thank the reviewers for their thoughtful comments and support. We want to emphasize that the paper provides many timely insights unified by a strong narrative around probabilistic model construction and generalization, showing the role of multimodal marginalization, neural networks priors, tempering, support, and inductive biases — with significant demonstrations, including an exhaustive empirical study of marginalization, a demonstration of the role of marginal likelihood in resolving questions around generalization, and a Bayesian resolution of double descent (tying together the opening narrative and the importance of multimodal marginalization). The material is particularly timely, given recent questions about Bayesian methods in deep learning, such as the treatment of deep ensembles as a competing approach to Bayesian neural networks, and the cold posterior experiments in Wenzel et. al (ICML 2020), which are resolved in our discussion of tempering. While papers with many types of contributions can be difficult to assess, we believe this paper makes an important and timely contribution to NeurIPS, and appreciate the strong support of reviewers.

Inspired by reviewer comments, we additionally evaluated the effect of deep ensembles. In the setting of Fig. 6 (c), deep ensembles achieve errors $60., 65., 68.6, 70., 70.4, 70.3, 71.2, 71.4$ for widths $5, 7, 10, 15, 20, 25, 30, 50$ respectively. In agreement with our Bayesian perspective of deep ensembles, they almost resolve double descent and provide similar but worse results compared to MultiSWAG. We will add the results to Figure 6.

- **R1.** Thank you for the thoughtful and supportive review. We describe the unifying narrative and contributions above. We appreciate the feedback and we will further clarify these connections in a final version. As you say, the main contribution of the MultiSWAG part of the paper is not the algorithm but the Bayesian perspective, the exhaustive empirical study, and the resolution of double descent. We believe these are major contributions, and that the paper overall has a significant degree of novelty (which we believe should not be constrained to algorithmic innovation).
- We will follow up with the AC on the paper you mention, but we can assure you this brief note is in no way in conflict with our submission and should not weigh in the decision.
- L138: in some applications such as continual learning we may want to approximate the posterior over the parameters rather than the posterior predictive distribution. The Bayesian posterior predictive is not necessarily the optimal model average under model misspecification. In particular, we argue for temperature scaling in appendix E, which leads to a different predictive distribution compared to the *true* posterior.
- Figure 3: Wasserstein distance can be easily computed from samples from the two distributions and provides a useful measure of their difference without the mode seeking or mode covering behavior of KL divergence.
- The weights of the components in MultiSWAG should reflect the relative mass of the different modes in the posterior. We expect these masses to be similar, but estimating them empirically is an interesting direction for future work.
- **R2.** Thank you for your thoughtful review. We think there are some misunderstandings, and hope you can consider our clarifications in your updated assessment. We would like to emphasize that the material in this paper should not be viewed as a background or tutorial on BDL, leading to the proposal of MultiSWAG. Our paper presents a novel perspective as well as truly exhaustive experiments that provide insights into BDL. The main purpose of MultiSWAG is to demonstrate properties of multi-basin marginalization, as part of the larger narrative of the paper.
- We will add a more detailed description of MultiSWA & MultiSWAG but note that these straightforwardly ensemble independently trained SWAG and SWA models (which are described in full papers) and the point here is about multibasin marginalization. MultiSWAG has the same computational complexity as Deep Ensembles at training time, and the memory overhead comes from storing 5-20 copies of the weights for each mode (note that these weights can be stored on the disk rather than in GPU memory and take significantly less memory compared to the activations). At test time, MultiSWAG has more overhead due to samples within each basin.
- In Fig. 3 the vague prior was used to prevent over-regularization of the network that would lead to a poor fit of the ground truth (obtained with HMC) to the data. The results do not hinge on the specific values of the hyper-parameters.
- We want to clarify that Fig. 3 is not intended to provide a like-for-like comparison. This experiment is intended to show the importance of multibasin representations for approximate BMA integration. In particular, we do not argue against VI in general, and a multibasin ensemble of VI approximations would also support our findings. We will clarify, and include a discussion of the reference you suggested.
- We do actually have accuracy results for Exp 2 in the Appendix, Fig. 18. We also report accuracy for double descent.
- We include deep ensembles to our experiment on double descent as you suggested (see above).
- **R3.** Thank you for your supportive review. We have now included a new experiment on deep ensembles performance in the context of double descent.
- **R4.** Thanks for your supportive review. (1) We view deep ensembles as a compelling mechanism for approximate BMA integration under constraints – different from both variational methods and MCMC, which are typically combined with simple MC. In this vein, we do not view deep ensemble weights as samples from an approximate posterior, and we would not recover the exact predictive distribution by taking an infinite number of ensembles. We also provide a related discussion in Appendix C and Appendix Figure 8. We will clarify in the final version. (2) Re: Fig 3. The point here is more about multimodal vs unimodal in approximating the BMA integral than about a deficiency of variational methods. Multiple independently trained VI models (for a multi-basin posterior) could indeed perform well and also make this point. Thanks for the question. We will clarify.