We thank all reviewers for their valuable comments. We'll further improve in the final version. Below, we address the detailed comments.

To Reviewer 1:

Q1: Beyond regression tasks: In this paper, we focus on regression tasks. As noted in the discussion part, we leave the classification task as future work and point out the potential challenge to be solved.

Q2: Model mis-specification: Thanks. It's indeed an interesting topic to systematically investigate on the robustness of our algorithm. We note that there is some recent work (e.g., [*1]) that studies the robustness of the MMD estimators, which can be applied to our MMD-based calibration method.

[*1] Briol et al., Statistical inference for generative models with maximum mean discrepancy. arXiv:1906.05944.

Q3. Results of larger networks: Guo et al. (ICML 2017) argued that the miscalibration was due to the sheer size of modern NNs and this conclusion was drawn from the image classification experiments. This conclusion is not suitable for the regression tasks because the larger network may overfit. The size of our networks is close to the previous works on regression problems ([11,16,18,20,30,35,38]). We'll make it clearer in the final version.

Q4: Lack of comparison with post-hoc calibration methods: Thanks. We have added the results of the related post-hoc methods (isotonic regression), and the ECPE of isotonic regression is $0.032 \pm 0.002, 0.042 \pm 0.002, 0.011 \pm 0.000, 0.023 \pm 0.002, 0.030 \pm 0.001, 0.061 \pm 0.003, 0.014 \pm 0.006, 0.213 \pm 0.212$ and 0.083 ± 0.006 respectively for Metro-traffic, Bike-sharing, Pickups, PM2.5, Air-quality, power-plant, protein-structure, naval-propulsion and wine. We can see that the calibration errors of our method are significantly smaller than those of isotonic regression. The updated results and error-bars will be included in the final version.

Q5: Ambiguity of the argument "essentially different from post-processing methods": Thanks for clarifying and we agree. Following the comment, We will change this argument to "our method also uses a post-processing procedure but learns the calibration in the model level, which means practitioners are not required to retrain the model but can enjoy the calibration performance".

Q6: Clarity: Thanks. We'll improve the clarity. In particular, we'll correct the minor errors in Eqn.7 and Eqn.10-11, revise the corresponding part of Eqn.7 and add the model parameters in Eqn.10-11. We address the other points below:
1) *Line 168*: It means that our method is different from the alternating fashion [35], our method optimizes two loss functions in turn.

2) *Line 173*: We mean that our distribution matching strategy produces the predictive distribution estimator (HNN) in the model level (see Q5). We will make it more precise in the final version.

3) *Fig. 2 about "sharpness"*: We prefer prediction intervals as tight as possible while accurately covering the ground truth in regression tasks. We measure the sharpness using the width of prediction intervals in this paper, which is detailed in Appendix C. Sharpness was previously used in [18, 35]. We will add an explanation in the main text.

Q7: newer BNN baselines: We have compared with DeepGP (Hugh et al, NIPS2017), which is a strong Bayesian baseline and can produce well-calibrated predictive uncertainties. We also add a BNN baseline method named fpovi [*2], whose ECPE is $0.065 \pm 0.001, 0.132 \pm 0.002, 0.173 \pm 0.001, 0.261 \pm 0.000$ and 0.276 ± 0.001 for Metro-traffic, Bike-sharing, Pickups, PM2.5 and Air-quality. The calibration error of our method is significantly smaller.

[*2] Ziyu et al. Function Space Particle Optimization for Bayesian Neural Networks. ICLR 2019.

To Reviewer 2: As for the amount of data is enough for a given problem, [12] shows that MMD has performance guarantees at finite sample sizes, based on uniform convergence bounds. For the results of our method, regardless of whether or not $p = q$, the empirical MMD converges in probability at rate $O((m+n)^{-\frac{1}{2}})$ to its population value, where m and n respectively represent the number of samples sampled from P and Q. And in practical applications, you can judge whether the data is sufficient according to MMD error. We will add the related discussion in the final version.

To Reviewer 3: Thanks for clarifying the concept of the distribution-level calibration and sorry for the confusion. The "distribution-level" in our submission means that we learn the calibrated predictive distribution Q by minimizing the kernel embedding measure. Yet, this claim does not mean to imply the pre-existing "distribution calibration" in [34]. According to [34], we agree that our method is for the global quantile-level calibration. To avoid ambiguity here, we will remove the words "distribution-level" in the final version and more precisely say that "we learn the calibrated predictive distribution Q by minimizing the kernel embedding measure".

To Reviewer 4: Thanks for the positive review.