



Figure 14: **Combined Accuracy of Uncertainty Estimation and Predictive Performance.** Comparing the NLL of all aleatoric uncertainty estimating models, aggregated over 10 experiments. For each dataset, the best model in terms of average NLL is marked with a star together with any other models not statistically worse based on a one-sided Mann-Whitney-Wilcoxon test. Apart from the Random Forest and Evidential models, all other models are trained with censored labels.