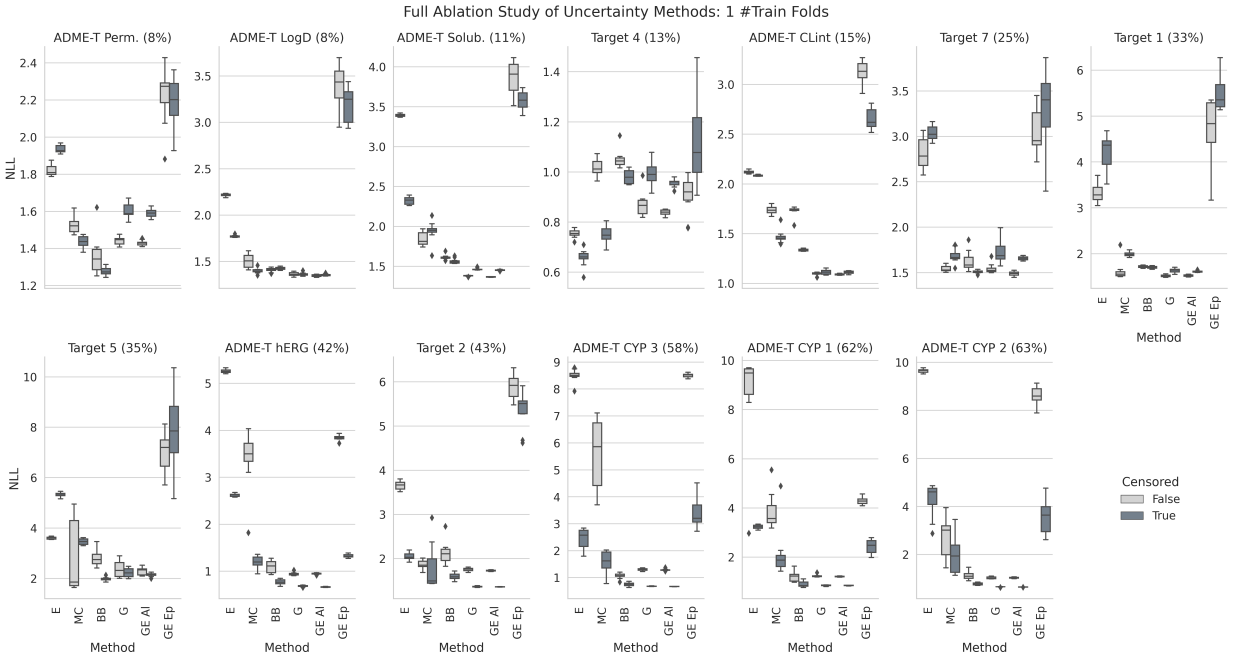

Figure 11: **Full Ablation Study: 1 Training Folds.** Comparison of NLL for models trained with and without additional censored labels on the first temporal setting containing one fold for training.

D Additional Model Comparison

In the following section, additional results from the model comparison are presented. First, the NLL scores for all aleatoric estimates are provided in Fig. 14. Next, the ENCE scores are provided in Fig. 15 to complement the NLL scores as a second metric for the global, intertwined evaluation of predictive performance and calibration of uncertainty estimates. Next, the confidence-based calibration curves from the first two temporal settings are shown in Fig. 16, as a complement to Fig. 5.