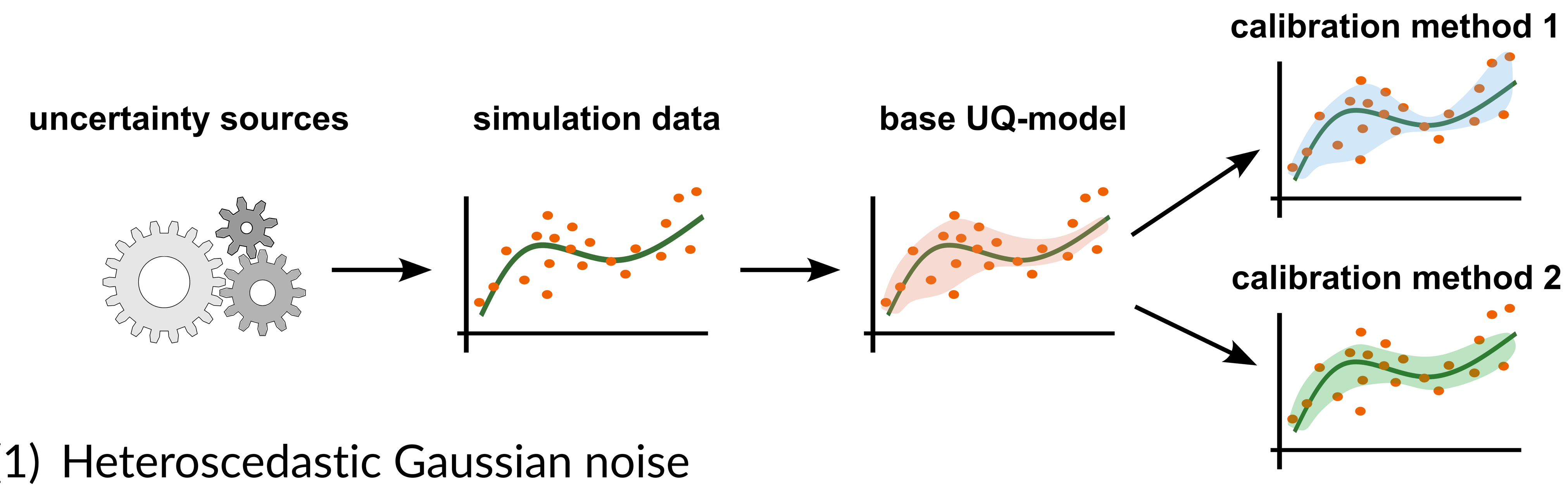
Comparing Conditional Calibration of Aleatoric Uncertainty Estimators

Kornelius Raeth

kornelius.raeth@uni-tuebingen.de

When do calibration methods fail?

Which sources of uncertainty can or can't be handled by existing UQ-models and calibration methods?



(1) Heteroscedastic Gaussian noise

(2) Simulation-based regression setup

(3) Base UQ-model with post-hoc calibration methods

$\rightarrow$ Evaluate conditional calibration of prediction intervals (PIs)

UQ-models	**Calibration methods**
▶ Deep Ensembles (DE)	▶ Quantile Calibration [Kuleshov, 2018]
▶ DE Bootstrap	▶ Calibration NN (CRPS loss) [Rasp, 2018]
▶ Mean-Var DE	▶ Distributional Conformal Prediction (DCP) [Chernozhukov, 2021]
▶ (MC-Dropout)	▶ (EMOS)

Evaluating conditional calibration

- ▶ We know $P_{X,Y}$ so we can compute PI coverage fraction $PICF(x) = P_{Y|x}(Y \in PI(x))$ in closed form.
- ▶ Calibration of a predictor cannot be assessed using a single training set [Sluijterman, 2021].

$D_m \sim P_{X,Y}$, $m = 1:M$

$\downarrow$ train

UQ-model

$\downarrow$ compute

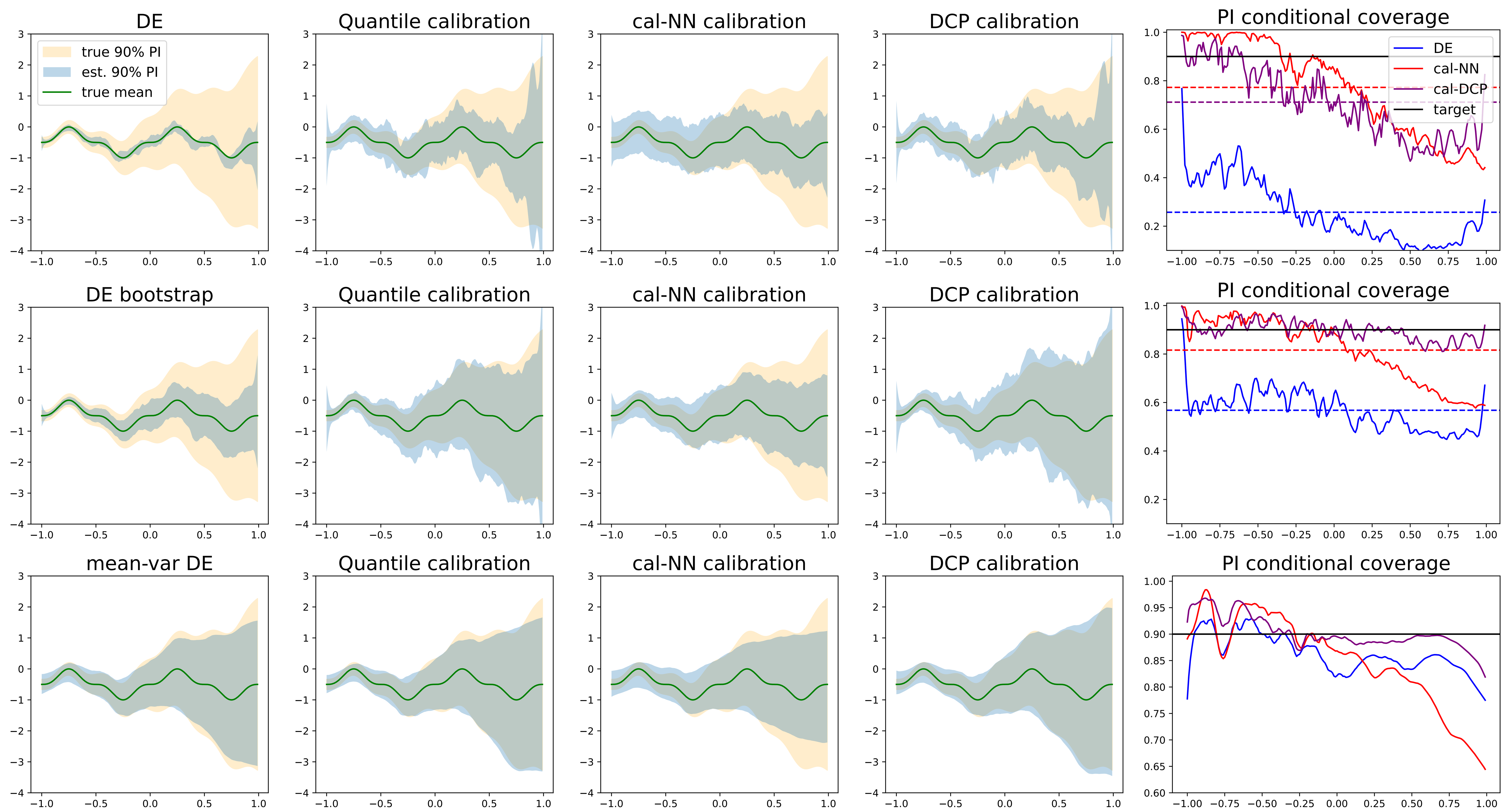
$PICF_m(x)$, $m = 1:M$

$$\rightarrow PICF(x) = \frac{1}{M}\sum_{m=1}^{M} PICF_m(x)$$

Perfect conditional calibration if

$$\forall x : PICF(x) = 1 - \alpha$$

Preliminary Results



Takeaways

- ▶ Mean-variance DE attains the best conditionally calibrated PIs among the base UQ-models.
- ▶ Calibration NN trained with CRPS loss does not attain good conditional calibration and can even lead to worse results.
- ▶ Global calibration methods can only achieve good conditional calibration if the base UQ-model already captures the heteroscedasticity.