Horoscope Framework: Evaluation Report

Keyvan Amiri Elyasi[1][0009−0007−3016−2392], Han van der Aa[2][0000−0002−4200−4937], and Heiner Stuckenschmidt[1][0000−0002−0209−3859]

[1] Data and Web Science Group, University of Mannheim, Germany
{keyvan,heiner}@informatik.uni-mannheim.de
[2] Faculty of Computer Science, University of Vienna, Austria
han.van.der.aa@univie.ac.at

1 Introduction

This report supplements our paper, Horoscope: A Probabilistic Remaining Time Prediction Framework for Business Processes, submitted to the 37th International Conference on Advanced Information Systems Engineering (CAiSE'25). It provides a detailed description of the experimental setup (see Section 2) and a comprehensive evaluation of our framework's performance (see Section 3).

2 Experimental Setup

This section describes the experimental setup, organized as follows: Section 2.1 introduces the event logs used in the experiments, while Section 2.2 provides a list of the UQ configurations that are supported by our framework and included in our experiments. Performance metrics for evaluating the quality of probabilistic predictions are detailed in Section 2.3, and Section 2.4 discusses the hyperparameter configurations for the different UQ configurations.

2.1 Event logs

We used 10 publicly available event logs in our experiments[1]:

- Five BPI Challenge 2020 logs: Domestic Declarations: BPIC20DD, International Declarations: BPIC20ID, Requests for Payment: BPIC20RFP, Prepaid Travel Cost: BPIC20PTC, Travel Permits: BPIC20TPD.
- BPI Challenge 2015 Municipality 1 (BPIC15-1)
- BPI Challenge 2013 Incidents (BPIC13I)
- BPI Challenge 2012 (BPIC12)
- BPI Challenge 2012 (Helpdesk)
- BPI Challenge 2012 (Sepsis)

Table 1 summarizes the characteristics of these event logs.

[1] https://data.4tu.nl/categories/13500?categories=13503

Table 1. Characteristics of the employed event logs (time-related attributes in days).

Event log	Cases	Events	Event Classes	Variants	Case length		Case duration	
					Avg.	Max	Avg.	Max
BPIC20DD	10 500	56 437	17	99	5.37	24	11.5	469.2
BPIC20ID	6449	72 151	34	753	11.19	27	86.5	742.0
BPIC20RFP	6886	36 796	19	89	5.34	20	12.0	406.0
BPIC20PTC	2099	18 246	29	202	8.69	21	36.8	324.5
BPIC20TPD	7065	86 581	51	1478	12.25	90	87.4	1190.3
BPIC15-1	1199	52 217	398	1170	43.55	101	95.9	1486.0
BPIC13I	7554	65 533	13	2278	8.68	123	12.1	771.4
BPIC12	13 087	262 200	36	4366	20.04	175	8.6	137.2
Helpdesk	4580	21 348	14	226	4.66	15	40.9	60.0
Sepsis	1050	15 214	16	846	14.49	185	28.5	422.3

2.2 UQ configurations

The uncertainty module in the Horoscope framework incorporates 12 UQ configurations to capture both epistemic and aleatoric uncertainty, summarized in Table 2. In the following, we refer to these configurations using the abbreviations introduced in this table.

Table 2. UQ configurations supported by uncertainty module.

Uncertainty Type	**UQ configurations**
Epistemic	Laplace approximation (LA), embedding-based random forests (E-RF), deep ensembles (DE), bootstrapping ensembles (BE), dropout approximation (DA), concrete dropout approximation (CDA)
Aleatoric	Heteroscedastic regression (H), classification and regression diffusion (CARD)
Epistemic+Aleatoric	DE+H, BE+H, DA+H, CDA+H

While our paper introduced the core concepts of these UQ configurations, it did not explore their computational efficiency in detail. This supplementary evaluation report addresses that gap. Experiments were conducted on an Nvidia RTX A6000 GPU, with training and inference times averaged across 10 event logs. Training time refers to the duration required to train a probabilistic model within the uncertainty module and is reported in hours. Once trained, the model can be deployed in real-time scenarios to predict the remaining time of an incomplete process instance (i.e., a prefix). Inference time, measured in milliseconds, represents the time taken to generate a prediction for one prefix. Table 3 summarizes the training and inference times for all 12 UQ configurations evaluated, with existing baselines marked by *. Additionally, it includes the training and inference times for the deterministic neural network serving as the backbone for these UQ configurations.

Table 3. Average and standard deviation for training and inference time across all 10 logs. Training time in hours, inference time in milliseconds.

	Training time (h)		Inference time (milli s)	
UQ configuration	Avg	Std	Avg	Std
LA	0.01	0.02	0.10	0.11
E-RF	0.01	0.03	0.29	0.36
DE	2.16	3.19	0.37	0.38
BE	1.07	2.02	0.51	0.59
DA^*	4.36	3.73	47.26	32.39
CDA^*	7.99	12.59	35.61	14.37
CARD	20.74	29.33	440.39	317.74
H^*	0.62	1.56	0.16	0.23
DE+H	1.29	2.37	0.27	0.29
BE+H	1.12	1.63	0.55	0.61
$DA + H^*$	5.17	8.29	36.12	16.07
$CDA + H^*$	8.60	15.89	37.65	15.68
Backbone	0.58	1.57	0.10	0.11

LA and *E-RF* are efficient techniques that can be applied post-hoc on pre-trained neural networks, requiring only 35 and 46 seconds for training, respectively. Another efficient technique is *H*, which requires training a modified neural network from scratch. However, its average training time is approximately 2.5 minutes longer than that of the deterministic backbone model.

For deep ensemble configurations (*DE*, *DE+H*), multiple models must be trained sequentially, leading to training times of 77 to 130 minutes. This issue can be mitigated by bootstrapping ensembles (*BE*, *BE+H*), which train multiple models on different data subsets and require roughly one hour for training.

MC dropout baselines (*DA*, *CDA*, *DA+H*, *CDA+H*) are less efficient than ensembles, requiring 4.5 to 8.5 hours due to the necessary sampling for inference on the validation set. This training time can be reduced by excluding validation from the training process, but this would eliminate the possibility of hyperparameter optimization and early stopping to prevent overfitting.

The most computationally expensive UQ configuration is the diffusion-based model (*CARD*), with an average training time of approximately 21 hours.

CARD and MC dropout, which require sampling for inference, have larger inference times compared to other UQ configurations. However, the inference times for all configurations are sufficiently small to make them suitable for real-time predictions. The longest inference time per prefix observed was for *CARD* on the BPIC12 event log, which took only 1.2 seconds.

2.3 Performance metrics

This section discusses key probabilistic prediction metrics focusing on accuracy, calibration, sharpness, and sparsification error.

Accuracy. To evaluate the prediction accuracy, we use Mean Absolute Error (MAE) which measures the average magnitude of errors between predicted and actual remaining time, providing an intuitive assessment of accuracy. Formally, MAE is defined as:

$$MAE = \frac{1}{|\mathcal{D}_{test}|} \sum_{y^* \in \mathcal{D}_{test}} |\hat{\mathbb{E}}[y^*] - y^*| \tag{1}$$

Calibration. Calibration measures the alignment between the predicted posterior and observed distributions of remaining time. To formally define it, we use the inverse cumulative distribution function (CDF), or quantile function, denoted as F^{-1}. This function returns the smallest value y such that the predicted remaining time is at most y with a given probability p:

$$F^{-1}(p) = inf\{y : F(y) \geq p\} \tag{2}$$

A probabilistic model is considered calibrated if the following condition is satisfied [5], where $\mathbb{I}$ is the indicator function:

$$\frac{\sum_{y^* \in \mathcal{D}_{test}} \mathbb{I}\{y^* \leq F^{-1}(p)\}}{|\mathcal{D}_{test}|} = p \quad \forall p \in [0, 1] \tag{3}$$

This ensures that the proportion of observed remaining times below the predicted quantile $F^{-1}(p)$ matches the corresponding probability p. From this definition, two key metrics can be derived: prediction interval coverage probability (PICP) and miscalibration area.

PICP quantifies the proportion of observed remaining times within the prediction interval $[L(p), U(p)]$ at a given confidence level p, as defined in Equation 4, where $\mathbb{I}$ denotes the indicator function [3]. For instance, the 95% credible interval must contain the actual remaining time for 95% of prefixes: $PICP(0.95) = 0.95$.

$$PICP(p) = \frac{1}{|\mathcal{D}_{test}|} \sum_{y^* \in \mathcal{D}_{test}} \mathbb{I}\{L(p) \leq y^* \leq U(p)\} \tag{4}$$

Computing PICP across different confidence levels yields average calibration metrics, often visualized in a calibration plot with $PICP(p)$ on the y-axis and $p \in [0, 1]$ on the x-axis. The area between the calibration plot and the parity line ($PICP(p) = p$), known as the miscalibration area, quantifies the model's calibration performance [1]. A perfectly calibrated model satisfies $PICP(p) = p$, resulting in a miscalibration area (MA) of zero. This metric highlights whether a model is systematically overconfident (underestimating uncertainty) or underconfident (overestimating uncertainty) [4].

Figure 1 illustrates the MA metric for three UQ configurations applied to remaining time prediction in the BPIC20ID event log. (a) shows the dropout approximation (*DA*) baseline [8], which performs poorly and is highly overconfident. (b) depicts the performance of bootstrap ensembles combined with heteroscedastic regression (*BE+H*), achieving well-calibrated predictions with MA

= 0.03. Finally, (c) presents the results for deep ensembles (*DE*), which result in an underconfident model.

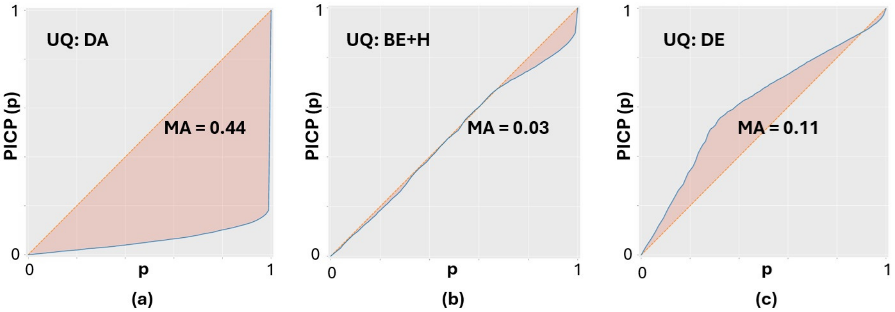

Fig. 1. Miscalibration Area (MA) for three different UQ configurations applied for remaining time prediction in the BPIC20ID event log: (a) an overconfident model, (b) a well-calibrated model, and (c) an underconfident model.

Sharpness. Sharpness measures the concentration of the posterior distribution of remaining time around its mean and is quantified by the average predicted standard deviation, $\overline{\sigma}$, across all prefixes in the test set:

$$\overline{\sigma} = \frac{\sum_{(\mathbf{x}^*,y^*)\in\mathcal{D}_{test}} \sigma_{\mathbf{x}^*}}{|\mathcal{D}_{test}|} \tag{5}$$

Unlike calibration, which assesses the alignment of predictions with observed values, sharpness focuses solely on the precision of probabilistic predictions, independent of model accuracy [1,5]. A lower $\overline{\sigma}$ indicates sharper predictions, suggesting greater confidence in the model's output.

Negative log likelihood (NLL). There is an inherent trade-off between calibration and sharpness: widening prediction intervals increases the likelihood that observed values will fall within them, but it also reduces precision. *Proper scoring rules*, such as the NLL calculated as in Equation 6 [4], balance both calibration and sharpness into a single metric [1]. Therefore, these metrics can be used as the optimization objective during the training of probabilistic neural networks [1,4]. We used NLL for hyper-parameter tuning in our experiments.

$$NLL = \frac{1}{2|\mathcal{D}_{test}|} \sum_{(\mathbf{x}^*,y^*)\in\mathcal{D}_{test}} \ln(2\pi{\sigma_{\mathbf{x}^*}}^2) + \frac{(\mu_{\mathbf{x}^*} - y^*)^2}{{\sigma_{\mathbf{x}^*}}^2} \tag{6}$$

Sparsification error. Sparsification error measures the alignment between a model's uncertainty and its prediction error. It is computed by sorting the test set predictions based on their uncertainty ($\sigma_{\mathbf{x}^*}$) and progressively removing subsets with the highest uncertainty (1% of the test examples at a time). The MAE on the remaining predictions is then plotted against the fraction of removed examples. A monotonically decreasing error curve indicates that the model assigns higher uncertainty to less accurate predictions. Based on the error curve, two important metrics for sparsification error can be derived:

- The ideal curve (oracle) would be derived by sorting predictions based on their actual MAE. The Area Under the Sparsification Error curve (AUSE) measures the area between the actual error curve and the oracle curve, offering a quantitative assessment of sparsification error [7]. A lower AUSE indicates greater model trustworthiness, as it shows that the model correctly signals higher uncertainty when its accuracy is in doubt.
- If the test set predictions are randomly sorted, the error curve will be approximately flat, representing the performance of a model with no correlation between prediction error and uncertainty. The area between the actual error curve and this random curve is called the Area Under the Random Gain curve (AURG). A higher AURG indicates that the model's predicted uncertainties are more informative than random guesses in relation to the actual error of the model [7].

Figure 2 illustrates the AUSE and AURG metrics for two UQ configurations applied to remaining time prediction in the BPIC15-1 event log. Panel (a) shows the performance of dropout approximation combined with heteroscedastic regression ($DA{+}H$), which exhibits high sparsification error, as indicated by a high AUSE and a low AURG. Panel (b) presents the performance of deep ensembles (DE), which demonstrates moderate sparsification error, with the error curve positioned in the middle of the oracle and random curves.

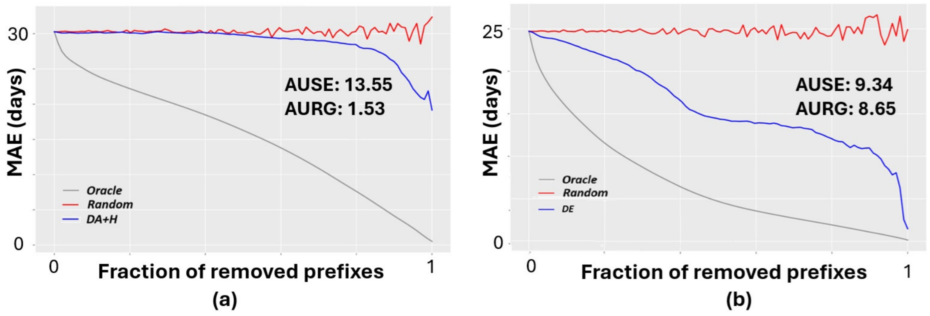

Fig. 2. Sparsification error for two different UQ configurations applied for remaining time prediction in the BPIC15-1 event log: (a) dropout approximation combined with heteroscedastic regression ($DA{+}H$), (b) deep ensembles (DE).

2.4 Hyper-parameter tuning

All experiments are based on DALSTM model [6] as the backbone deterministic neural network. For the backbone model, we adhered to the parameters suggested in [6], except for a dropout probability of 0.1, aligning with [8], which used MC dropout for probabilistic remaining time prediction.

Table 4 lists the UQ configurations and their hyper-parameters used in our experiments. Although the Horoscope framework supports various optimization metrics for hyper-parameter tuning, we used NLL on the validation set for this purpose.

We did not tune any hyperparameters for the post-hoc improvement module and conducted experiments using $\eta = 10$ equal-width intervals for range-based prediction adjustment (RPA).

Table 4. Hyper-parameters for Uncertainty Quantification Configurations

Configuration	Hyper-parameters
DE	The number of models in the ensemble, $m \in \{3, 4, \ldots, 10\}$, is treated as a hyper-parameter.
BE	The number of models in the ensemble, $m \in \{3, 4, \ldots, 10, 12, \ldots, 20, 24, 28, 32, 40\}$, is treated as a hyper-parameter. Each model is fitted using 25% of the training prefixes.
E-RF	The number of decision trees in the forest, $m \in \{10, 25, 50, 75, 100, 125, 150, 200\}$, and the maximum tree depth $RF_d \in \{3, 6, 9, 12, \infty\}$ are treated as hyper-parameters. For all other parameters, we used default values in RandomForestRegressor class in scikit-learn.
LA	Hyper-parameters $sigma\ noise \in \{0.5, 1.0, 2.0\}$, $prior\ precision \in \{1.0, 5.0, 10.0, 20.0, 30.0\}$, $temperature \in \{0.5, 1.0, 2.0, 4.0, 8.0\}$ are tuned post-hoc using empirical Bayes. We experimented with both full Hessian and KFAC factorization. See the original paper [2] for more information.
DA, CDA	We utilized the default values from [8]. To assess the impact of early stopping on performance, we trained two models: one with early stopping enabled and one without.
CARD	We used default hyper-parameters with a few adjustments. First, we jointly trained the backbone model $f_{\mathbf{w}}$ and the noise estimation network $q_\theta(y_{t-1}\|y_t)$ because post-hoc training of the noise estimation network led to poor performance. Second, we modified the diffusion schedule by reducing the variance of Gaussian noise for the first timestamp ($\beta_1 \in \{1e^{-5}, 1e^{-6}\}$), while keeping the ratio of first and last noise variances consistent with the original paper. Third, we also tested a width of 256 for noise estimation network in addition to the original dimension of 128. Finally, we experimented with window sizes $w_s \in \{3, 4, 5\}$ concatenating features from the last events of the prefix into the input of the noise estimation network. See the original paper [3] for more information.
H	No hyper-parameters.

3 Results

This section evaluates the performance of various UQ configurations across different metrics and is structured as follows: Section 3.1 examines the predictive accuracy of UQ configurations, along with the impact of post-hoc prediction adjustment on accuracy. Similar analyses are presented in Section 3.2, Section 3.3, and Section 3.4 for calibration, sharpness, and sparsification error, respectively.

3.1 Predictive accuracy

Table 5 summarizes the performance of different UQ configurations in terms of MAE. Configurations are separated by horizontal lines based on the uncertainty type. The star sign (*) indicates those already used for probabilistic remaining time prediction in existing baselines.

The predictive accuracy of LA matches that of its deterministic backbone model. Entries highlighted in blue represent UQ configurations that outperform the deterministic backbone model in predictive accuracy. For these models, uncertainty estimation improves accuracy, while entries in black achieve probabilistic predictions at the expense of accuracy.

Table 5. Predictive accuracy of UQ configurations, measured by mean absolute error (MAE) in days, before any post-hoc improvement. Entries in blue indicate results that are equal to or better than the accuracy of the deterministic backbone model.

UQ Configuration	BPIC20DD	BPIC20ID	BPIC20RFP	BPIC20PTC	BPIC20TPD	BPIC15-1	BPIC13I	BPIC12	Helpdesk	Sepsis
LA	4.12	21.72	5.19	7.73	27.79	25.28	5.53	7.84	9.45	15.73
E-RF	**3.97**	15.10	**4.97**	8.82	24.85	32.53	3.55	6.74	10.48	29.86
DE	5.92	17.26	7.35	11.08	27.12	**24.71**	3.36	6.05	11.18	15.44
BE	4.34	17.00	6.38	**6.94**	**24.57**	41.65	4.86	**5.77**	11.68	15.74
DA^*	5.27	14.68	6.24	8.28	25.30	53.53	11.56	6.37	10.32	16.35
CDA^*	7.50	15.07	7.31	8.27	25.00	37.78	7.09	6.82	17.70	15.86
CARD	4.50	22.26	5.71	11.86	30.37	100.43	4.69	7.03	12.87	24.50
H^*	4.13	15.40	5.59	8.36	26.01	26.93	3.15	5.86	**9.44**	**15.32**
DE+H	4.90	22.31	5.86	10.34	34.77	26.71	3.09	5.81	20.16	16.39
BE+H	4.70	**14.08**	6.62	10.45	27.80	27.23	**3.06**	5.95	18.65	15.69
$DA + H^*$	5.38	14.62	5.09	9.45	25.02	30.29	3.67	6.60	11.28	16.35
$CDA + H^*$	6.32	17.09	6.30	12.83	25.98	32.23	4.19	8.02	10.09	18.08

Overall, no UQ configuration consistently outperforms the others across all event logs in terms of accuracy, nor does any UQ configuration outperform the deterministic backbone model for every event log. Event logs can be categorized into two groups based on their level of difficulty for UQ configurations to achieve high predictive accuracy.

The first group includes event logs where the majority of UQ configurations improve predictive accuracy compared to the deterministic backbone model. This group comprises BPIC20ID, BPIC20TPD, BPIC13I, and BPIC12.

The second group consists of event logs where achieving probabilistic remaining time predictions often comes at the expense of reduced accuracy. This group includes BPIC15-1, BPIC20DD, BPIC20PTC, BPIC20RFP, Helpdesk, and Sepsis. Among these, BPIC15-1, BPIC20DD, and BPIC20PTC are the most challenging, with at least half of the UQ configurations degrading accuracy by 25% or more. BPIC15-1 and BPIC20DD are particularly difficult for MC dropout baselines, while BPIC20PTC challenges all UQ configurations that capture both epistemic and aleatoric uncertainty. The remaining logs in this group (BPIC20RFP, Helpdesk, and Sepsis) are moderately challenging but generally less problematic than the others.

This analysis highlights the variability in the accuracy of UQ configurations and the influence of event log characteristics on accuracy. Fig. 3, illustrates the distribution and variability of accuracy across different UQ configurations using box-and-whisker plots. Additionally, it shows the overall effect of applying range-based prediction adjustments (RPA) on the predictive accuracy of these configurations. Since the MAE metric depends on the average lead time of a process and varies across event logs, we normalized it by dividing by the average lead time for each event log prior to generating the box-and-whisker plots.

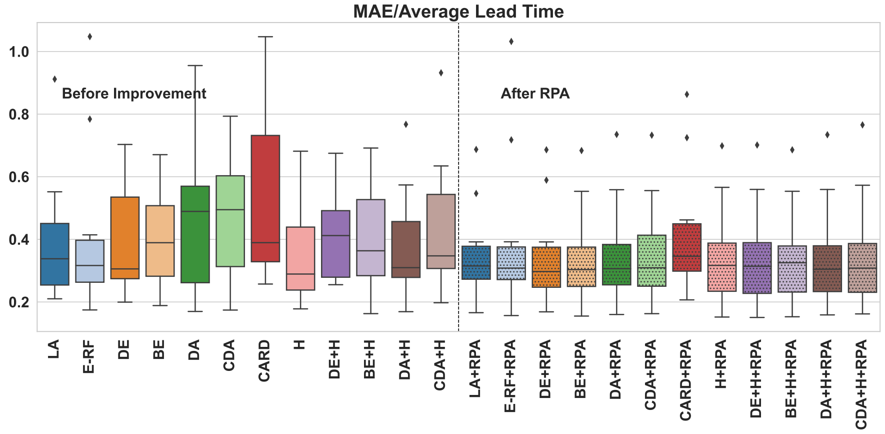

Fig. 3. Box-and-whisker plots showing the predictive accuracy of UQ configurations across 10 event logs. Results before post-hoc improvements are displayed on the left of the vertical line, while those after applying RPA are shown on the right.

Effectiveness of post-hoc improvement. As it is illustrated Fig. 3, RPA proves to be an effective post-hoc improvement strategy, because it generally reduces the MAE of the model. However, exceptions exist, as indicated by the individual points plotted beyond the whiskers.

Table 6 presents the MAE performance of various UQ configurations after applying RPA, while Table 7 reports the percentage accuracy improvements. Red entries in Table 7 denote cases where RPA reduces model accuracy.

Table 6. Predictive accuracy of UQ configurations, measured by mean absolute error (MAE) in days, after applying range-based prediction adjustment (RPA). Entries in blue indicate results that are equal to or better than the accuracy of the deterministic backbone model.

UQ Configuration	BPIC20DD	BPIC20ID	BPIC20RFP	BPIC20PTC	BPIC20TPD	BPIC15-1	BPIC13I	BPIC12	Helpdesk	Sepsis
LA	3.75	14.34	**4.70**	7.37	23.40	32.22	3.49	5.91	12.44	**15.58**
E-RF	3.75	13.55	4.71	7.58	23.69	30.10	3.30	6.17	12.38	29.42
DE	**3.73**	14.58	**4.70**	7.31	23.69	28.38	2.89	5.90	12.21	16.79
BE	3.74	13.41	**4.70**	**7.28**	23.41	29.66	2.95	**5.88**	**12.15**	15.77
DA^*	3.86	13.84	4.80	7.82	24.26	26.98	2.99	6.32	13.55	15.92
CDA^*	4.15	14.07	5.17	7.89	23.70	26.61	2.94	6.30	13.95	15.84
CARD	4.36	17.88	5.55	9.20	27.32	29.00	4.99	6.23	12.16	24.60
H^*	3.75	13.14	4.71	7.37	**23.31**	36.03	2.70	6.01	12.58	16.14
DE+H	3.74	**13.02**	4.71	7.72	23.45	36.42	**2.59**	6.03	12.41	15.95
BE+H	3.75	13.21	4.71	7.44	23.38	31.10	2.66	5.90	13.82	15.78
$DA + H^*$	3.88	13.73	4.72	7.83	24.27	24.93	2.71	6.31	13.56	15.94
$CDA + H^*$	3.88	14.01	4.83	8.08	24.33	**24.59**	2.69	6.58	13.88	16.33

Based on the results in Table 7, RPA's impact on the performance of UQ configurations varies across different event logs:

- For the BPI Challenge 2020 logs and the BPIC13I log, RPA improves model accuracy for almost all UQ configurations, achieving an average improvement of 10% to 30%.
- For the BPIC12 log, RPA enhances accuracy for 9 out of 12 UQ configurations. While accuracy decreases for three configurations, the degradation is minor (less than 4%). Overall, RPA either improves accuracy or has a negligible negative impact on this log.
- For the Sepsis and BPIC15-1 logs, RPA improves accuracy for approximately half of the UQ configurations and degrades it for the other half. However, the effect differs between the logs. For Sepsis, the impact is minimal, with only small changes in accuracy. In contrast, for BPIC15-1, RPA results in either significant improvements or substantial degradations in accuracy.

Table 7. Predictive accuracy gain after applying range-based prediction adjustment (RPA), measured in percentage. Red entries indicate cases where RPA reduces model accuracy.

UQ Configuration	BPIC20DD	BPIC20ID	BPIC20RFP	BPIC20PTC	BPIC20TPD	BPIC15-1	BPIC13I	BPIC12	Helpdesk	Sepsis
LA	8.99	33.96	9.40	4.74	15.82	−27.47	36.90	24.63	−31.71	0.95
E-RF	5.53	10.28	5.35	14.11	4.64	7.49	7.02	8.45	−18.12	1.48
DE	36.92	15.56	36.05	34.02	12.64	−14.82	13.83	2.41	−9.22	−8.76
BE	13.77	21.12	26.32	4.93	4.71	28.78	39.29	−1.97	−4.01	0.23
DA^*	26.85	5.73	23.10	5.51	4.13	49.60	74.12	0.72	−31.31	2.62
CDA^*	44.67	6.62	29.26	4.58	4.88	29.59	58.40	7.65	21.21	−0.17
CARD	2.98	19.66	2.82	22.41	10.07	71.13	−6.39	11.33	5.54	−0.42
H^*	9.20	14.67	15.82	11.85	10.41	−33.77	14.05	−2.50	−33.20	−5.35
DE+H	23.79	41.63	19.54	25.31	32.56	−36.38	16.16	−3.90	38.45	2.72
BE+H	20.24	6.16	28.80	28.85	15.90	−14.23	13.10	0.83	25.91	−0.54
$DA + H^*$	27.87	6.07	7.21	17.16	3.00	17.69	26.05	4.33	−20.23	2.53
$CDA + H^*$	38.54	18.03	23.38	36.97	6.34	23.71	35.74	17.85	−37.54	9.72
Average Gain(%)	19.64	19.10	17.76	14.38	10.77	4.14	27.04	6.55	−9.60	−0.10

– For the Helpdesk dataset, RPA reduces accuracy in 8 out of 12 UQ configurations. Exceptions include ensemble combined with heteroscedastic regression (*DE+H*, *BE+H*) and concrete dropout approximation (*CDA*), which benefit from RPA with accuracy gains of 20% to 40%. For most configurations, however, RPA is an ineffective post-hoc strategy for this dataset.

Since RPA applies different linear transformations to varying ranges of predictions, it not only impacts the accuracy of UQ configurations but also alters the distribution of predicted remaining times.

Figure 4 illustrates the probability density function of the remaining time for prefixes in the test set of the BPIC20ID log. An extreme case in this figure is *DE+H*, which consistently predicts expected values close to the mode of the remaining time distribution. For this UQ configurations, RPA significantly alters the prediction distribution, aligning it more closely with the observed distribution. In contrast, the prediction distributions of *LA* and *BE* are already well-aligned with the observed distribution. For these configurations, RPA primarily adjusts the modes by uniformly shifting the predicted values.

3.2 Calibration of UQ configurations

Table 8 summarizes the performance of different UQ configurations based on PICP (0.95), which is the proportion of observed remaining times within the

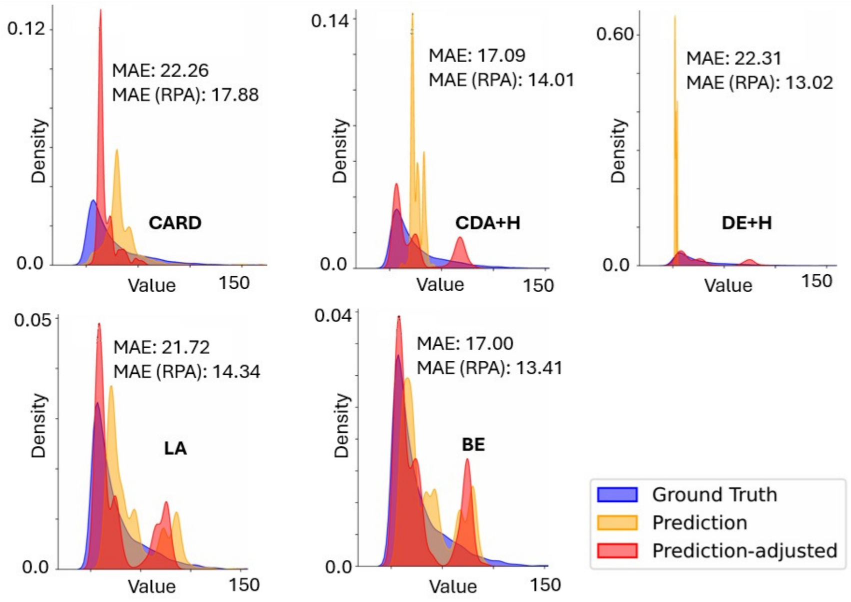

Fig. 4. probability density function of the remaining time for prefixes in the test set of the log BPIC20ID.

prediction interval at the confidence level 95%. For a perfect model the prediction interval must contain the actual remaining time for 95% of prefixes: $PICP(0.95) = 0.95$. The best result, marked in bold, has the minimum deviation from 0.95. The last row presents the average PICP across all configurations, highlighting the relative difficulty of achieving good calibration for each event log.

Table 9 summarizes the performance of different UQ configurations based on miscalibration area (MA). Configurations with an MA below 0.10, indicating well-calibrated uncertainty estimates, are highlighted in black. Overconfident results are marked in red, while underconfident results are shown in green.

Based on the results in Table 8 and Table 9, the worst calibration is observed for the MC dropout baselines (*DA*, *CDA*), which exhibit overconfidence across all 10 event logs. While combining these techniques with heteroscedastic regression (*DA+H*, *CDA+H*) slightly mitigates this issue, the combinations remain overconfident, likely because heteroscedastic regression itself suffers from overconfidence. Similarly, bootstrapping ensembles (*BE*) and embedding-based random forests (*E-RF*) face the same challenge.

In contrast, *CARD*, *LA*, *DE*, *DE+H*, and *BE+H* achieve acceptable calibration for most event logs. With the exception of *LA*, which shows underconfidence for some logs, all UQ configurations tend to be overconfident. Notably, the Helpdesk event log is the most challenging log. For this log, only *LA* produces

Table 8. Calibration of UQ configurations, measured by PICP. Results before any post-hoc improvement.

UQ Configuration	BPIC20DD	BPIC20ID	BPIC20RFP	BPIC20PTC	BPIC20TPD	BPIC15-1	BPIC13I	BPIC12	Helpdesk	Sepsis
LA	**0.94**	0.87	**0.98**	0.96	0.82	0.83	0.82	**0.98**	**0.97**	0.89
E-RF	0.12	0.25	0.13	0.29	0.20	0.99	**0.96**	0.25	0.87	**0.97**
DE	0.44	0.91	0.40	0.53	0.81	0.90	0.89	0.74	0.45	0.76
BE	0.57	0.92	0.19	0.75	0.62	**0.94**	0.72	0.29	0.28	0.43
DA^*	0.05	0.13	0.11	0.11	0.12	0.05	0.03	0.07	0.16	0.07
CDA^*	0.00	0.10	0.08	0.14	0.10	0.03	0.04	0.04	0.00	0.11
CARD	0.85	0.95	0.84	**0.95**	0.82	0.33	0.75	0.52	0.73	0.69
H^*	0.57	0.47	0.51	0.37	0.41	0.56	0.58	0.84	0.15	0.82
DE+H	0.72	0.62	0.71	0.56	0.64	0.76	0.81	0.88	0.58	0.82
BE+H	0.69	0.83	0.52	0.51	0.72	0.86	0.78	0.89	0.58	0.84
$DA + H^*$	0.34	0.83	0.55	0.51	**0.92**	0.99	0.82	0.55	0.46	0.65
$CDA + H^*$	0.26	**0.95**	0.50	0.11	0.88	0.40	0.75	0.63	0.37	0.30
Average	0.46	0.63	0.45	0.48	0.57	0.60	0.63	0.54	0.47	0.59

Table 9. Calibration of UQ configurations, measured by MA: results before any post-hoc improvement. Overconfident results are marked in red, while underconfident results are shown in green.

UQ Configuration	BPIC20DD	BPIC20ID	BPIC20RFP	BPIC20PTC	BPIC20TPD	BPIC15-1	BPIC13I	BPIC12	Helpdesk	Sepsis
LA	0.20	0.06	0.29	0.27	0.13	**0.06**	**0.04**	0.05	**0.10**	0.20
E-RF	0.44	0.38	0.44	0.37	0.41	0.22	0.09	0.39	0.18	0.13
DE	0.30	0.11	0.32	0.29	**0.05**	0.11	0.06	0.13	0.28	0.04
BE	0.25	0.07	0.41	0.11	0.24	0.08	0.16	0.37	0.37	0.32
DA^*	0.47	0.44	0.46	0.45	0.44	0.47	0.48	0.46	0.44	0.46
CDA^*	0.49	0.45	0.47	0.43	0.45	0.48	0.48	0.48	0.49	0.45
CARD	**0.05**	0.06	**0.03**	**0.05**	**0.05**	0.39	0.12	0.29	0.17	0.18
H^*	0.23	0.28	0.28	0.31	0.31	0.24	0.26	0.10	0.43	**0.07**
DE+H	0.10	0.18	0.07	0.21	0.13	0.10	0.13	**0.03**	0.31	0.14
BE+H	0.12	**0.03**	0.25	0.26	0.11	**0.06**	0.12	**0.03**	0.28	0.14
$DA + H^*$	0.32	0.11	0.22	0.23	0.16	0.23	0.07	0.23	0.21	0.10
$CDA + H^*$	0.39	0.06	0.25	0.46	0.10	0.26	0.16	0.13	0.36	0.34
Average	0.28	0.20	0.29	0.28	0.23	0.24	0.20	0.24	0.30	0.23

well-calibrated predictions, while the other UQ configurations exhibit overconfidence.

Table 10 summarizes the MA performance of UQ configurations after applying RPA and CR. Entries marked in red indicate combinations of UQ configurations and event logs where post-hoc improvement methods degrade calibration.

Table 10. Calibration of UQ configurations, measured by MA: results after RPA and CR.

UQ Configuration	BPIC20DD	BPIC20ID	BPIC20RFP	BPIC20PTC	BPIC20TPD	BPIC15-1	BPIC13I	BPIC12	Helpdesk	Sepsis
Results after Calibrated Regression										
LA	0.05	0.03	0.06	0.05	0.05	0.07	0.11	0.04	0.14	0.04
E-RF	0.04	0.03	0.06	0.04	0.09	0.11	0.09	0.07	0.09	0.13
DE	0.03	0.04	0.04	0.06	0.05	0.16	0.03	0.06	0.11	0.07
BE	0.05	0.05	0.06	0.06	0.03	0.07	0.02	0.06	0.14	0.05
DA^*	0.06	0.05	0.07	0.06	0.03	0.07	0.08	0.07	0.23	0.05
CDA^*	0.04	0.04	0.05	0.06	0.04	0.07	0.06	0.06	0.17	0.05
CARD	0.03	0.05	0.04	0.05	0.03	0.09	0.06	0.03	0.20	0.09
H^*	0.04	0.04	0.06	0.07	0.03	0.16	0.05	0.08	0.17	0.06
DE+H	0.05	0.04	0.06	0.06	0.07	0.13	0.03	0.08	0.17	0.08
BE+H	0.05	0.03	0.05	0.06	0.04	0.12	0.02	0.12	0.17	0.06
$DA + H^*$	0.05	0.05	0.05	0.08	0.03	0.10	0.04	0.06	0.22	0.08
$CDA + H^*$	0.03	0.07	0.06	0.06	0.04	0.19	0.04	0.04	0.10	0.08
Average	0.04	0.05	0.05	0.06	0.05	0.11	0.05	0.07	0.16	0.07
Results after Range-based Prediction Adjustment										
LA	0.05	0.05	0.06	0.07	0.10	0.10	0.11	0.06	0.18	0.04
E-RF	0.06	0.05	0.09	0.04	0.09	0.09	0.09	0.08	0.14	0.12
DE	0.05	0.03	0.06	0.03	0.07	0.05	0.08	0.07	0.17	0.04
BE	0.05	0.04	0.07	0.03	0.07	0.05	0.06	0.07	0.18	0.05
DA^*	0.06	0.07	0.07	0.06	0.08	0.16	0.13	0.06	0.20	0.04
CDA^*	0.04	0.06	0.06	0.07	0.09	0.15	0.12	0.06	0.24	0.05
CARD	0.05	0.04	0.05	0.03	0.05	0.11	0.20	0.05	0.19	0.10
H^*	0.05	0.04	0.06	0.06	0.07	0.09	0.11	0.07	0.18	0.03
DE+H	0.05	0.05	0.07	0.05	0.08	0.08	0.10	0.05	0.18	0.04
BE+H	0.05	0.03	0.06	0.05	0.07	0.08	0.08	0.05	0.21	0.04
$DA + H^*$	0.05	0.05	0.05	0.05	0.06	0.13	0.11	0.08	0.20	0.05
$CDA + H^*$	0.05	0.06	0.07	0.05	0.07	0.16	0.13	0.06	0.20	0.05
Average	0.05	0.05	0.06	0.05	0.08	0.10	0.11	0.06	0.19	0.06

The results show that RPA and CR significantly enhance the calibration of probabilistic models. Cases where these methods degrade calibration are rare (approximately 7-10% of all entries) and primarily involve configurations already achieving very good calibration. In these instances, the degradation is minor, and the models retain good calibration even after post-hoc adjustments.

Fig. 5, illustrates the distribution and variability of miscalibration area across different UQ configurations using box-and-whisker plots. Additionally, it shows the overall effect of applying calibrated regression (CR) and range-based prediction adjustments (RPA) on calibration of these configurations. While some of the configurations demonstrate poor calibration and high variability for miscalibration area before applying post-hoc adjustments, we can observe that nearly all UQ configurations achieve comparable calibration following these adjustments.

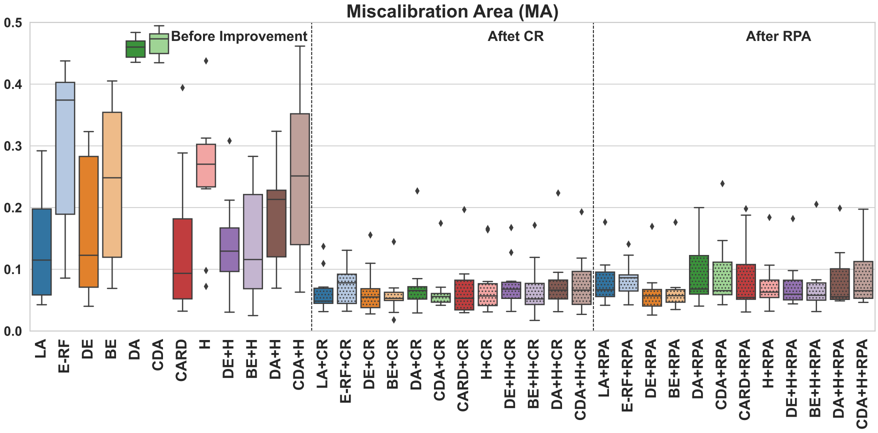

Fig. 5. Box-and-whisker plots showing calibration of UQ configurations across 10 event logs. Results before post-hoc improvement, and after applying CR and RPA are separated by dashed vertical lines.

3.3 Sharpness of UQ configurations

Table 11 summarizes the sharpness of different UQ configurations based on the average predicted standard deviation ($\overline{\sigma}$). To highlight which UQ configurations produce overly narrow or overly wide confidence intervals, the three configurations with the smallest $\overline{\sigma}$ for each event log are marked in red, while the three with the largest $\overline{\sigma}$ are marked in green.

There is a clear distinction in the sharpness of different UQ configurations. While *DA* and *CDA* tend to produce overly narrow confidence intervals, the widest intervals are usually generated by *CARD* and *LA*. Applying CR and

Table 11. Sharpness of UQ configurations, measured by average predicted standard deviation ($\overline{\sigma}$) in days: results before any post-hoc improvement.

UQ Configuration	BPIC20DD	BPIC20ID	BPIC20RFP	BPIC20PTC	BPIC20TPD	BPIC15-1	BPIC13I	BPIC12	Helpdesk	Sepsis
LA	7.88	22.06	17.14	21.72	22.20	22.45	5.01	11.51	15.57	22.38
E-RF	1.03	10.64	3.09	3.64	7.08	85.33	4.35	1.84	8.98	42.37
DE	1.87	24.26	2.06	4.72	27.90	36.85	5.59	4.08	6.18	11.00
BE	1.27	19.48	0.55	5.25	14.44	51.32	4.17	1.58	3.58	3.96
DA^*	0.39	1.78	0.22	0.90	2.24	4.99	1.58	0.46	0.67	0.90
CDA^*	0.21	1.23	0.28	1.00	1.66	2.43	0.61	0.32	1.32	0.52
CARD	4.23	35.62	6.50	16.36	33.03	46.43	7.42	4.00	10.81	20.68
H^*	1.40	4.02	1.35	1.38	9.08	9.57	2.06	5.53	1.95	13.00
DE+H	2.82	11.18	3.32	3.88	16.80	18.67	2.93	6.95	11.32	13.92
BE+H	2.38	12.61	2.04	3.26	16.14	25.93	2.80	7.51	10.74	18.12
$DA + H^*$	0.95	11.44	1.92	2.51	47.01	106.68	3.94	2.74	2.69	4.65
$CDA + H^*$	1.14	24.59	1.65	1.64	33.58	5.71	4.34	4.96	4.66	1.52
Average	2.04	14.94	3.10	5.18	18.48	31.42	3.73	4.32	6.60	11.35

RPA involves scaling the predicted standard deviation (i.e., predictive uncertainty) to enhance model calibration. Since most UQ configurations are overconfident, the scaling factor is often greater than one, which reduces sharpness. This highlights the inherent trade-off between calibration and sharpness, where post-hoc adjustments improve calibration at the expense of sharpness. However, for underconfident configurations (e.g., *LA*), post-hoc adjustments typically apply a scaling factor smaller than one, thereby improving both calibration and sharpness.

Figure 6 illustrates the distribution and variability of the average predicted standard deviation across UQ configurations using box-and-whisker plots. It also shows the impact of calibrated regression (CR) and range-based prediction adjustments (RPA) on sharpness. To account for variability in lead times across event logs, predicted standard deviations were normalized by dividing them by the average lead time for each event log before generating the plots.

As shown in the figure, both post-hoc improvement methods reduce the sharpness of UQ configurations. However, RPA outperforms CR by adjusting the predicted expected value of the remaining time before scaling the standard deviation. This approach is particularly beneficial when the model's predicted expected value significantly deviates from the observed value, making a substantial difference in such cases. Table 12 presents the scaling factors obtained through the RPA and CR methods. Entries smaller than one, indicating simultaneous improvement in calibration and sharpness, are marked in blue.

Table 12. Scaling factor obtained by post-hoc improvement methods (CR, RPA). Entries smaller than one, indicating simultaneous improvement in calibration and sharpness, are marked in blue.

UQ Configuration	BPIC20DD	BPIC20ID	BPIC20RFP	BPIC20PTC	BPIC20TPD	BPIC15-1	BPIC13I	BPIC12	Helpdesk	Sepsis
Results after Calibrated Regression										
LA	0.35	1.11	0.19	0.31	1.31	1.53	0.73	0.94	1.17	0.43
E-RF	17.46	6.25	11.32	6.18	5.92	0.66	1.00	6.55	1.6	0.98
DE	2.96	0.60	3.26	3.26	0.98	1.22	1.22	1.91	1.86	1.29
BE	2.31	1.19	8.54	1.72	2.08	1.26	1.58	5.70	3.02	3.62
DA^*	16.40	9.98	19.90	12.07	10.37	16.79	10.28	22.22	41.36	15.14
CDA^*	38.16	13.61	21.66	9.68	12.10	27.56	15.35	31.92	30.51	15.60
CARD	1.07	0.95	0.84	1.04	1.13	2.81	1.23	2.73	0.93	1.36
H^*	2.15	2.98	2.52	4.69	3.48	5.44	2.25	1.74	10.38	0.91
DE+H	1.36	2.10	1.07	2.80	2.03	2.26	1.39	1.20	3.62	0.53
BE+H	1.44	1.12	2.52	3.36	1.46	1.68	1.39	1.23	3.71	0.63
$DA + H^*$	4.82	1.64	2.04	3.74	0.51	0.62	1.09	3.34	11.10	1.51
$CDA + H^*$	5.40	1.05	2.60	9.66	0.72	11.21	1.45	2.70	3.54	7.36
Average	11.77	5.55	14.08	7.78	5.13	7.86	6.89	10.90	59.86	4.60
Results after Range-based Prediction Adjustment										
LA	0.33	0.52	0.18	0.28	0.69	1.86	0.62	0.72	0.62	0.43
E-RF	14.68	5.24	8.05	5.21	4.69	0.52	0.97	6.85	1.2	0.98
DE	1.42	0.61	1.53	1.68	0.81	0.80	0.62	2.01	1.56	1.59
BE	2.02	0.69	5.58	1.89	1.30	0.61	1.11	5.31	2.78	3.12
DA^*	7.12	7.24	14.64	9.06	8.57	10.57	1.08	21.95	15.27	11.92
CDA^*	15.30	10.62	12.89	7.65	10.83	18.05	3.06	32.72	7.15	20.99
CARD	0.84	0.59	0.77	0.59	0.77	1.00	0.59	2.32	0.90	0.82
H^*	1.89	3.00	2.32	4.77	1.93	3.77	1.87	1.54	5.00	1.00
DE+H	0.92	1.04	0.92	1.85	0.97	1.73	1.00	1.14	0.86	0.79
BE+H	1.11	1.06	1.53	2.39	1.14	1.71	1.22	1.02	0.95	0.64
$DA + H^*$	3.03	1.11	1.78	3.24	0.43	0.36	0.74	3.54	3.81	2.26
$CDA + H^*$	2.52	0.55	1.91	4.52	0.57	7.36	0.45	1.97	2.28	6.77
Average	7.98	3.54	11.77	6.38	3.50	5.73	4.10	10.77	51.74	4.82

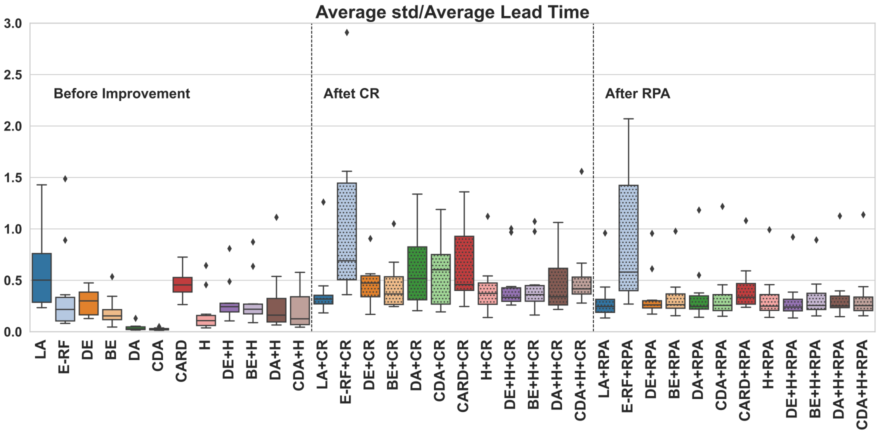

Fig. 6. Box-and-whisker plots showing sharpness of UQ configurations across 10 event logs. Results before post-hoc improvement, and after applying CR and RPA are separated by dashed vertical lines.

The results in Table 12 highlight the advantages of RPA over CR. Except for the Sepsis log, RPA consistently requires a smaller scaling factor than CR across the other nine event logs. RPA improves the sharpness of UQ configurations in one-third of all cases, compared to CR, which achieves this in only one-sixth of the cases. Notably, RPA simultaneously enhances both sharpness and calibration for *CARD* and *LA* in 9 out of 10 event logs.

3.4 Sparsification error of UQ configurations

Table 13 summarizes the performance of different UQ configurations based on AUSE, which measures how well the predicted uncertainty aligns with the model's actual error. Techniques such as *LA*, *CARD*, *DA*, and *CDA* generally exhibit higher sparsification error. In contrast, heteroscedastic regression, ensembles, and their combinations (*H*, *DE*, *BE*, *DE+H*, *BE+H*), as well as *E-RF*, tend to achieve lower AUSE. As a result, these methods make it easier for process analysts to differentiate between accurate and inaccurate predictions and better estimate model error based on predicted uncertainty.

Figure 7 presents the distribution and variability of AUSE across UQ configurations using box-and-whisker plots, as well as the impact of range-based prediction adjustments (RPA) on AUSE. To account for differences in lead times across event logs, AUSE values were normalized by dividing them by the average lead time for each event log before generating the plots. As shown in the figure, RPA reduces the sparsification error for MC dropout baselines and their combinations with heteroscedastic regression (*DA*, *CDA*, *DA+H*, *CDA+H*). However, this post-hoc adjustment has minimal impact on other UQ configurations.

Table 13. Area under sparsification error curve (AUSE) of UQ configurations: results before any post-hoc improvement.

UQ Configuration	BPIC20DD	BPIC20ID	BPIC20RFP	BPIC20PTC	BPIC20TPD	BPIC15-1	BPIC13I	BPIC12	Helpdesk	Sepsis
LA	2.84	8.30	4.38	3.72	18.59	13.28	2.44	3.77	5.08	13.16
E-RF	**1.39**	3.93	**1.43**	**1.83**	7.90	14.61	1.51	3.25	4.65	10.51
DE	1.86	4.72	2.22	2.08	6.78	9.34	1.30	2.31	4.82	7.87
BE	1.62	3.15	3.11	2.11	**6.33**	**8.53**	1.26	2.34	5.30	7.75
DA^*	2.56	7.10	3.86	2.61	15.35	18.74	3.70	3.52	7.53	10.80
CDA^*	3.10	8.62	3.90	3.76	14.71	14.56	2.19	4.01	8.45	12.91
CARD	2.46	8.15	4.51	3.27	16.54	34.81	1.61	3.29	6.21	7.96
H^*	1.52	3.12	1.68	1.86	7.32	13.52	1.28	1.92	**4.10**	**7.42**
DE+H	1.76	**2.86**	1.83	2.02	6.73	13.81	1.37	**1.83**	6.13	11.08
BE+H	1.69	3.34	2.17	1.86	6.49	11.06	**1.21**	2.03	6.80	8.71
$DA+H^*$	1.99	5.04	1.86	2.57	9.44	13.55	1.33	2.55	8.83	12.85
$CDA+H^*$	2.27	5.99	2.17	3.21	9.15	11.17	1.26	3.80	4.20	17.60

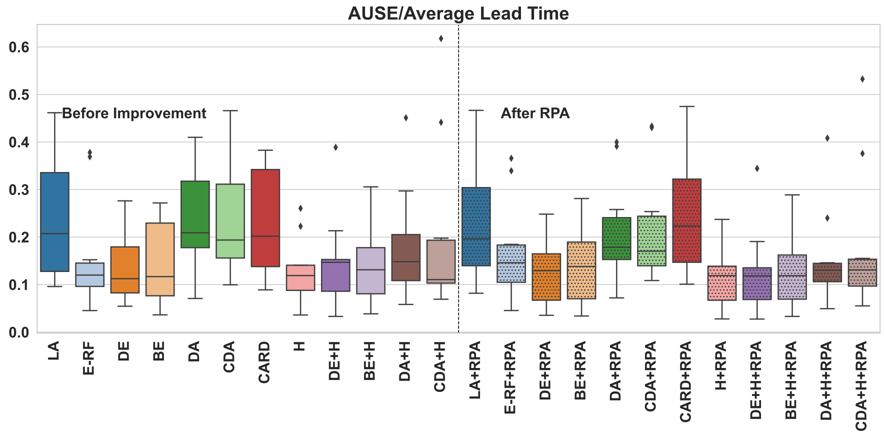

Fig. 7. Box-and-whisker plots showing sharpness of UQ configurations across 10 event logs. Results before post-hoc improvement, and after applying RPA are separated by dashed vertical line.

Table 14 summarizes the performance of different UQ configurations based on AURG, which measures how much better the predicted uncertainty aligns with the model's actual error when compared to uncertainty that is generated by random guessing (i.e., no correlation between predicted uncertainty and model error). Entries marked in red indicate cases where the UQ configuration performs worse than random guessing. Overall, the trends observed for AUSE also apply to AURG, with heteroscedastic regression, ensembles, and their combinations generally outperforming other UQ configurations. Notably, the Helpdesk log emerges as the most challenging dataset in our experiments. Except for *DE+H* and *BE+H*, the performance of other UQ configurations on this log is no better than random guessing.

Table 14. Area Under the Random Gain curve (AURG) of UQ configurations: results before any post-hoc improvement.

UQ Configuration	BPIC20DD	BPIC20ID	BPIC20RFP	BPIC20PTC	BPIC20TPD	BPIC15-1	BPIC13I	BPIC12	Helpdesk	Sepsis
LA	0.13	3.66	−0.59	1.65	−2.98	2.14	1.25	0.06	−0.60	−1.00
E-RF	1.31	6.75	1.92	4.33	9.02	1.14	−0.02	0.71	−0.20	0.31
DE	**2.05**	7.80	**2.61**	5.43	14.00	8.65	0.25	1.58	0.53	4.80
BE	1.50	6.90	1.40	3.01	9.50	**14.81**	**1.62**	0.99	0.30	4.25
DA^*	0.07	2.26	0.63	2.45	2.18	2.42	0.62	0.22	−0.57	0.87
CDA^*	1.01	1.20	1.05	1.50	3.25	2.18	1.02	0.17	−0.55	−0.08
CARD	0.35	4.36	−0.71	3.15	1.13	3.74	1.43	0.34	−0.08	**8.35**
H^*	1.45	8.36	2.36	4.52	10.45	5.89	0.27	1.63	0.70	4.87
DE+H	1.78	**12.79**	2.42	5.45	**19.10**	4.29	0.28	**1.87**	**1.65**	2.40
BE+H	1.71	7.01	2.45	**5.69**	14.44	5.43	0.54	1.79	1.06	3.53
$DA + H^*$	1.8	4.54	1.76	4.46	8.37	1.53	0.81	1.80	−1.15	0.60
$CDA + H^*$	1.74	3.30	2.36	4.67	9.35	12.66	1.09	1.59	0.14	−3.16

References

1. Chung, Y., Char, I., Guo, H., Schneider, J., Neiswanger, W.: Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification (2021)
2. Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., Hennig, P.: Laplace redux - effortless bayesian deep learning. In: Advances in Neural Information Processing Systems. vol. 34, pp. 20089–20103. Curran Associates, Inc. (2021)
3. Han, X., Zheng, H., Zhou, M.: CARD: Classification and regression diffusion models. Advances in Neural Information Processing Systems **35**, 18100–18115 (2022)

4. Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R., Coley, C.W.: Uncertainty quantification using neural networks for molecular property prediction. Journal of Chemical Information and Modeling **60**(8), 3770–3780 (2020), publisher: American Chemical Society
5. Kuleshov, V., Fenner, N., Ermon, S.: Accurate uncertainties for deep learning using calibrated regression. In: Proceedings of the 35th International Conference on Machine Learning. pp. 2796–2804. PMLR (2018), ISSN: 2640-3498
6. Navarin, N., Vincenzi, B., Polato, M., Sperduti, A.: LSTM networks for data-aware remaining time prediction of business process instances. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 1–7 (2017)
7. Nemani, V., Biggio, L., Huan, X., Hu, Z., Fink, O., Tran, A., Wang, Y., Zhang, X., Hu, C.: Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial. Mechanical Systems and Signal Processing **205**, 110796 (2023)
8. Weytjens, H., De Weerdt, J.: Learning uncertainty with artificial neural networks for predictive process monitoring. Applied Soft Computing **125**, 109134 (2022)