BOUNDED-ABSTENTION MULTI-HORIZON TIME-SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-horizon time-series forecasting involves making predictions for a consecutive sequence of subsequent time steps. This task arises in many application domains, such as healthcare and finance, where mispredictions can have a high cost and reduce trust. The learning with abstention framework tackles these problems by allowing a model to abstain from offering a prediction when it is at an elevated risk of making a misprediction. Unfortunately, existing abstention strategies are ill-suited for the multi-horizon setting because they ignore the structured nature of the predictions offered by a multi-horizon forecaster. We fill this gap by formalizing the problem of learning with abstention for the multi-horizon forecasting setting and show that its structured nature admits a richer set of abstention problems. Concretely, we propose three natural notions of how a model could abstain for multi-horizon forecasting. We theoretically analyze each problem to derive the optimal abstention strategy and propose an algorithm that implements it. Extensive evaluation on 24 benchmark datasets shows that our proposed algorithms significantly outperform existing baselines.

1 INTRODUCTION

Multi-horizon time-series forecasting involves simultaneously making predictions for a consecutive sequence of subsequent time steps. This problem arises in a wide range of applications, including predicting asset prices (Bao et al., 2017; Selvin et al., 2017), providing weather forecasts (Rasp et al., 2020), modeling a patient's prognosis (Alaa & van der Schaar, 2019; Lim, 2018), and predicting how an athlete's skill will change over time (van Arem et al., 2025).

Many multi-horizon problems occur in high-stakes settings (*e.g.*, medicine) where incorrect predictions can have a high cost (Rajkomar et al., 2019). Therefore, it is appealing to allow a model to abstain from making a prediction in certain situations (Ruggieri & Pugnana, 2025). For example, abstaining may be appropriate when a model is at heightened risk of making a misprediction, *e.g.*, because it is highly uncertain about its prediction. In these cases, a decision could be deferred to a human or more information can be collected (Hendrickx et al., 2024).

However, most existing abstention strategies target standard classification and regression problems where a single value (or class) is predicted for each instance. Unfortunately, these methods are ill-suited for the multi-horizon setting because they all treat each prediction separately. That is, for each time step in the horizon, they will independently decide whether to offer a prediction or abstain. This is disadvantageous because it disregards the clear structure between consecutive time steps, and ignoring this can lead to fragmented forecasts, as illustrated in Figure 1b. In many applications it is necessary to receive a continuous (sub)sequence of predictions for future time steps to identify trends or onset events (Pidan & El-Yaniv, 2011). For instance, an operator forecasting electricity demand needs an uninterrupted sequence of predictions to know when additional capacity must be brought online and for how long it must remain available.

Our work generalizes abstention strategies to account for the richer and more structured nature of the multi-horizon setting. Concretely, we investigate three natural multi-horizon abstention settings:

Full abstention is where the forecast is either given for all time steps or none as shown in Figure 1c. This naturally extends standard single-step abstention to the multi-horizon setting because for each test time series, either a forecast is offered or it is not. For example, this may be

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

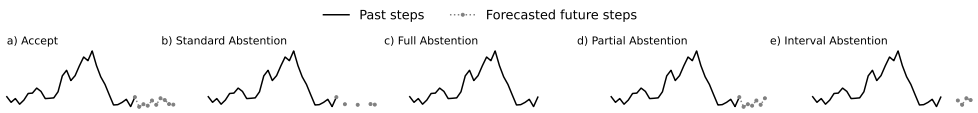

Figure 1: Possible forecasts with abstention in a multi-horizon setting. (a) The model offers a forecast for the entire horizon. (b) Standard abstention decides independently for each time step, hence the resulting forecast may be fragmented. (c) Full abstention only allows rejecting the forecast for the *entire* horizon. (d) Partial abstention lets the forecast start from the first time step and choose the end time step. (e) Interval abstention offers a forecast for a contiguous interval within the horizon.

appropriate when the test time series is overly noisy due to, *e.g.*, an EEG electrode with poor scalp contact, whose motion artifacts corrupt the entire recorded trial.

Partial abstention is where the forecast is offered for a consecutive sequence of time steps starting from the first time step in the prediction horizon as shown in Figure 1d. For example, in weather forecasting, predictive performance degrades sharply with time: a model may confidently forecast daily temperature for the coming few days, but the chaotic dynamics of the atmosphere make its longer-term estimates unreliable.

Interval abstention generalizes the partial abstention setting by offering a forecast for a consecutive sequence of time steps that may start from *any point* in the prediction horizon as shown in Figure 1e. For example, when forecasting daily COVID-19 case counts, the days right after a holiday are distorted by reporting lags, while the days after may be easier to predict.

We formalize each of these settings using the *bounded-abstention framework* (Franc et al., 2023), which aims to optimize forecasting performance subject to a target coverage constraint (*i.e.*, the proportion of examples for which a forecast is offered). For each abstention setting, we theoretically characterize the problem, provide the optimal abstention strategy, and propose an algorithm that approximates the optimal abstention strategy. We extensively validate our framework on 24 real-world datasets, demonstrating that our approaches significantly outperform several baselines.

2 Bounded-Abstention for Multi-Horizon Forecasting

We study the forecasting setting where the goal is to learn a forecasting model that generalizes to unseen time series (*e.g.*, learn from past patient data to provide forecasts for future patients). We assume time series of the form $y_{1:T} := (y_1, y_2, ..., y_T)$, where $y_i \in \mathbb{R}$ and i represents a time step.

We define a *selective multi-horizon forecaster* m as a pair (f, g), where

$f : \mathcal{Y}^{1:T} \to \mathcal{Y}^{T+1:T+H}$ is a *multi-horizon time-series forecaster* that, given an input time series $y_{1:T}$, predicts the values of its subsequent H time steps, where H is the forecast horizon;

$g : \mathcal{Y}^{1:T} \to \{1, \ldots, H\} \times \{0, \ldots, H\}$ is a *selection function* that determines the window within the horizon H for which a forecast is offered (accepted steps). For any time step outside this interval, the forecaster abstains (rejected steps). That is, $g\left(y_{1:T}\right) = (s, e)$ with $s \geq 1$ and $e \leq H$ denoting the start and the end of the forecast interval. For ease of notation, we omit the explicit dependence on the input time-series and use (s, e) instead of $(s(y_{1:T}), e(y_{1:T}))$. We adopt the convention that $(s = 1, e = 0)$ denotes that the entire horizon is rejected; apart from this exception we require $s \leq e$.

Formally, the output of m for a time point $T + 1 \leq t \leq T + H$ is defined as:

$$m_t\left(y_{1:T}\right) = \begin{cases} ® & \text{if } t < T + s \vee t > T + e \\ f_t\left(y_{1:T}\right) & \text{otherwise.} \end{cases} \tag{1}$$

That is, the forecasts for the time steps that satisfy the first condition are *rejected* whereas the forecasts for the remaining time steps are *accepted*.

We measure the performance of a *selective multi-horizon forecaster* using the *selective risk*, *i.e.*, the expected loss over the accepted steps. Letting $h(g) = e - s + 1$ be the number of accepted steps and

$\phi(g) = \mathbb{E}_{Y_{1:T}}[h(g)]$, we define the selective risk as:

$$\mathcal{R}(f, g) = \frac{\mathbb{E}_{Y_{1:T+H}}\left[\sum_{t=T+s}^{T+e} \ell\left(y_t, f_t\left(y_{1:T}\right)\right)\right]}{\phi(g)} \tag{2}$$

where $\ell\left(y_t, f_t\left(y_{1:T}\right)\right)$ is the loss at time-step t. Intuitively, the numerator represents the *total expected loss* over the accepted steps, while the denominator represents the *expected length* of the forecast interval. To avoid degenerate solutions where the forecaster always abstains, we enforce a *target coverage* c, requiring the model to offer forecasts for at least a fraction c of the steps in the horizon on average. Therefore, the objective is to find a selection function g that minimizes the *selective risk* subject to this coverage constraint (Pugnana et al., 2024). The general optimization problem can be formalized as:

$$g^* = \arg\min_g \mathcal{R}(f, g) \quad \text{s.t.} \quad \phi(g) \geq cH. \tag{3}$$

We now define three different ways to impose structural constraints on how g selects the indices s and e, each of which yields a distinct bounded-abstention problem. We theoretically analyze each problem and derive the optimal selection function. In our analysis, we assume *exchangeability across different time-series* (Stankeviciute et al., 2021): the joint distribution of the time series in the dataset is invariant to permutations of their order. We make *no assumption* about dependence *within* a series, so arbitrary temporal correlation across time steps is permitted. In practice, we focus on scenarios where we are given multiple time-series (*i.e.*, the training data), and we want to produce forecasts for new unseen time-series samples (*i.e.*, the test data).

2.1 Full abstention for multi-horizon forecasting

We first consider abstention at the *time-series level*:

Definition 1 (Full Abstention)**.** *In the **full abstention setting**, the selective forecaster must accept or reject the entire forecast horizon. Thus, the selection function is* $g_{FA} : \mathcal{Y}^{1:T} \to \{(1, H), (1, 0)\}$.

The goal is to minimize the risk over the accepted time-series while ensuring that, on average, forecasts are provided for at least a fraction c of the time-series.[1] Ideally, g_{FA} should abstain from those time-series with higher expected risk, *i.e.*, those for which the forecaster is most uncertain. The core challenge lies in evaluating the total expected risk across the entire horizon, while traditional abstention strategies evaluate it for a single prediction. We now show that the optimal selection function abstains whenever the sum of conditional risks over the entire horizon exceeds a threshold:

Theorem 1. *Given a forecaster f and a target coverage $c \in (0, 1]$, and assuming the cumulative conditional risk is not degenerate such that $\mathbb{P}\left(\sum_{t=T+1}^{T+H} \rho_t\left(y_{1:T}\right) < \tau_c^*\right) > 0$, the optimal selection function g_{FA}^* that is a solution to Equation 3 in full abstention is:*

$$g_{FA}^*\left(y_{1:T}\right) = \begin{cases} (1, H) & \text{if} \quad \sum_{t=T+1}^{T+H} \rho_t\left(y_{1:T}\right) < \tau_c^* \\ (1, \xi) & \text{if} \quad \sum_{t=T+1}^{T+H} \rho_t\left(y_{1:T}\right) = \tau_c^* \\ (1, 0) & \text{otherwise} \end{cases} \tag{4}$$

where ξ is a random variable such that $\xi = H$ with probability κ and $\xi = 0$ with probability $1 - \kappa$, and

$$\begin{aligned} \rho_t\left(y_{1:T}\right) &= \mathbb{E}_{Y_{T+1:T+H} | Y_{1:T}}\left[\ell\left(y_t, f_t(y_{1:T})\right) \mid Y_{1:T} = y_{1:T}\right] \\ \tau_c^* &= \inf\left\{v \mid \mathbb{E}_{Y_{1:T}}\left[\mathbb{1}\left\{\sum_{t=T+1}^{T+H} \rho_t\left(y_{1:T}\right) < v\right\}\right] \geq c\right\} \end{aligned} \tag{5}$$

$$\kappa = \begin{cases} 0 & \text{if} \quad \mathbb{P}\left(\sum_{t=T+1}^{T+H} \rho_t\left(y_{1:T}\right) = \tau_c^*\right) = 0, \\ \frac{c - \mathbb{P}\left(\sum_{t=T+1}^{T+H} \rho_t(y_{1:T}) < \tau_c^*\right)}{\mathbb{P}\left(\sum_{t=T+1}^{T+H} \rho_t(y_{1:T}) = \tau_c^*\right)} & \text{otherwise}. \end{cases}$$

[1] This is equivalent to forecasting cH time steps on average.

Proof. The proof can be found in Appendix B.1. □

Intuitively, the optimal solution is to abstain for those time-series for which the sum of conditional risks is greater than the c-th quantile of the distribution of the sum of conditional risks. Notice that ξ acts as a probabilistic tie-breaker, ensuring coverage constraint is exactly met when the distribution of the sum of the conditional risks is not continuous.

2.2 Partial abstention for multi horizon forecasting

Abstaining at the time-series level could be too strict, as the forecaster may provide accurate predictions for the initial time steps but becomes increasingly uncertain over time. Hence, it can be beneficial to abstain for all predictions after a given time step.

Definition 2 (Partial Abstention)**.** *The selective forecaster offers a prediction from $s = 1$ to e, where the chosen e differs for each considered time series. Therefore, the selection function in the **partial abstention setting** is defined as $g_{\texttt{PA}}(y_{1:T}) : \mathcal{Y}^{1:T} \to \{(1, 0), \ldots, (1, H)\}$.*

Here we aim to minimize the selective risk while ensuring that forecasts are provided for at least a fraction c of the horizon when averaged over all time series. Hence, for each time series, the selective forecaster must trade off between the increase in the conditional risk associated with a longer prediction with the need to meet the global coverage requirement. We formalize this intuition by showing that the optimal strategy arises from a Lagrangian relaxation, where the selection function balances the cumulative conditional risk against a linear reward for the forecast length.

Theorem 2. *Assume the joint vector $(\rho_{T+1}(y_{1:T}), \ldots, \rho_{T+H}(y_{1:T}))$ is absolutely continuous with respect to the Lebesgue measure.*[2] *Given a forecaster f and a target coverage $c \in (0, 1]$, the optimal selection function $g^*_{\texttt{PA}}$ in partial abstention returns a pair of indices (s^*, e^*), such that $s^* = 1$ and:*

$$e^* = \underset{e \in \{0,\ldots,H\}}{\arg\min} \left(\sum_{t=T+1}^{T+e} \rho_t(y_{1:T}) - \gamma^* e \right) \tag{6}$$

where $\gamma^* = \lambda^* + \eta^*$, $\lambda^* = \mathcal{R}(f, g^*_{\texttt{PA}})$ is the optimal risk, and η^* is the Lagrangian multiplier which is chosen such that the pair (e^*, η^*) satisfies the Karush–Kuhn–Tucker (KKT) conditions:

$$\eta^* \geq 0, \qquad \mathbb{E}_{Y_{1:T}}[e^*] \geq cH, \qquad \eta^* \left(\mathbb{E}_{Y_{1:T}}[e^*] - cH \right) = 0.$$

Proof. The proof can be found in Appendix B.2. □

Note that the parameter γ^* represents the reward of forecasting an additional step. For each test time series, the selective forecaster will provide a prediction for an additional time step only if the associated marginal increase in the cumulative conditional risk is lower than γ^*.

2.3 Interval abstention for multi horizon forecasting

Finally, we generalize the partial abstention setting by allowing the forecast to begin at any step.

Definition 3 (Interval Abstention)**.** *The selective forecaster selects an interval from s to e within the horizon and offers a prediction for all time steps in that interval. Therefore, the selection function in the **interval abstention setting** is defined as $g_{\texttt{IA}}(y_{1:T}) : \mathcal{Y}^{1:T} \to \{1, \ldots, H\} \times \{0, \ldots, H\}$. Note that the end time step e must be after the start time step s with the exception that (1,0) is possible (i.e., the entire forecast is rejected).*

Here the goal is to identify the interval that minimizes the selective risk, while ensuring that, on average, the model provides predictions for intervals of length at least cH. We now show that the optimal $g^*_{\texttt{IA}}$ is the minimizer of a Lagrangian trade-off between the cumulative conditional risk and a linear penalty on the interval length.

[2] See Appendix C for further discussion on the continuity assumption.

Theorem 3. *Assume the joint vector $(\rho_{T+1}(y_{1:T}), \ldots, \rho_{T+H}(y_{1:T}))$ is absolutely continuous with respect to the Lebesgue measure.*[2] *Given a forecaster f and a target coverage $c \in (0, 1]$, the optimal selection function g^*_{IA} in interval abstention selects a pair of indices (s^*, e^*) as:*

$$(s^*, e^*) = \underset{(s,e)|1\leq s\leq e\leq H\cup(1,0)}{\arg\min} \left(\sum_{t=T+s}^{T+e} \rho_t(y_{1:T}) - \gamma^* \cdot (e - s + 1) \right) \tag{7}$$

where $\gamma^* = \lambda^* + \eta^*$, and $\lambda^* = \mathcal{R}(f, g^*_{IA})$ is the optimal risk. The Lagrange multiplier η^* is chosen such that the pair $((s^*, e^*), \eta^*)$ satisfies the KKT conditions:

$$\eta^* \geq 0, \qquad \mathbb{E}_{Y_{1:T}}\left[e^* - s^* + 1\right] \geq cH, \qquad \eta^* \left(\mathbb{E}_{Y_{1:T}}\left[e^* - s^* + 1\right] - cH\right) = 0.$$

Proof. The proof can be found in Appendix B.3. □

Again, γ^* captures the expected reward for predicting another step. However, for each test time series, the optimal selective forecaster now will choose the start and end step of the contiguous interval whose cumulative conditional risk is smallest after subtracting the reward γ^* for each accepted step.

3 Learning a Selective Forecaster

We now describe how to learn a selective forecaster from data. This requires learning a forecaster, a conditional risk estimator, and calibrating a selection function. Note that the risk estimator and forecaster do not depend on the abstention settings, whereas the selection function does.

For the risk estimator, a natural choice in multi-horizon forecasting is considering the Mean Squared Error (MSE) over the accepted steps (Zhou et al., 2021; Challu et al., 2023). Thus, the conditional risk coincides with the conditional variance.[3] To estimate this quantity, we assume access to a dataset $\mathcal{D}_{\text{train}} = \left\{\left(y^i_{1:T}, y^i_{T+1:T+H}\right), i = 1, \ldots, n\right\}$, consisting of n exchangeable time series with observed past and future values. We construct an estimator $\hat{\rho}_t(y_{1:T})$ for each time step t by solving a regression task of the output variable

$$\hat{\rho}_t\left(y_{1:T}\right) = \hat{\sigma}^2_t(y_{1:T}) = \left(y_t - \hat{f}_t\left(y_{1:T}\right)\right)^2$$

where $\hat{\sigma}^2_t$ represents the conditional variance at time t. This quantity can be estimated by using a joint training approach, where a single neural network backbone is equipped with two output heads: one head outputs the H future steps $\hat{f}_t$ with $t \in \{T+1, \ldots, T+H\}$, and the other outputs the H future step conditional variances $\hat{\sigma}^2_t$ associated with $\hat{f}_t$. This model is effectively trained using the Gaussian β-NLL loss, which emphasizes well-calibrated variance estimates (Seitzer et al., 2022):

$$L^{Gauss}_{\beta-NLL}\Big(y, \hat{f}(y_{1:T}), \hat{\sigma}^2(y_{1:T})\Big) = \sum_{t=T+1}^{T+H} \mathrm{s}\Big(\hat{\sigma}^{2\beta}_t(y_{1:T})\Big) \left(\frac{\log\left(\hat{\sigma}^2_t(y_{1:T})\right)}{2} + \frac{\left(y_t - \hat{f}_t(y_{1:T})\right)^2}{2\hat{\sigma}^2_t(y_{1:T})} \right) \tag{8}$$

where $\mathrm{s}(\cdot)$ denotes the stop gradient operation and β controls the strength of the weighting.

All abstention strategies rely on the variance estimates $\hat{\sigma}^2_t$. Then, they enforce the coverage constraint using a separate calibration set $\mathcal{D}_{\text{calib}} = \{y^{n+i}_{1:T}\}_{i=1}^m$ of m exchangeable time series.[4] We show in Appendix F.5 that the ranking of these estimates correlates with that of the realized squared errors.

Full Abstention `FAbFor` (Full ABstention in multi-horizon FORecasting) approximates the optimal selection function derived in Theorem 1. We enforce the coverage constraint by computing $\sum_{t=T+1}^{T+H} \hat{\sigma}^2_t(y_{1:T})$ for all series in $\mathcal{D}_{calib}$ and setting $\hat{\tau}_c$ as the empirical c-quantile of these scores (see Equation 5). Thus the learned selection function accepts (in expected value) a fraction c of the forecasts, thereby satisfying the coverage constraint, and can be written as:

$$\hat{g}_{\texttt{FA}}\left(y_{1:T}\right) = \begin{cases} (1, H) & \text{if} \quad \sum_{t=T+1}^{T+H} \hat{\sigma}^2_t\left(y_{1:T}\right) < \hat{\tau}_c \\ (1, \hat{\xi}) & \text{if} \quad \sum_{t=T+1}^{T+H} \hat{\sigma}^2_t\left(y_{1:T}\right) = \hat{\tau}_c \\ (1, 0) & \text{otherwise} \end{cases}$$

[3] Other risk choices require estimating a different quantity (see Appendix F.8).

[4] Since the thresholds and parameters depend only on $y_{1:T}$, future values are not required for calibration.

where $\hat{\xi}$ is a random variable such that $\hat{\xi} = H$ with probability $\hat{\kappa}$ and $\hat{\xi} = 0$ otherwise, and

$$\hat{\kappa} = \begin{cases} 0 & \text{if} \quad \mathbb{P}\left(\sum_{t=T+1}^{T+H} \hat{\sigma}_t^2\left(y_{1:T}\right) = \hat{\tau}_c\right) = 0, \\ \frac{c - \mathbb{P}\left(\sum_{t=T+1}^{T+H} \hat{\sigma}_t^2(y_{1:T}) < \hat{\tau}_c\right)}{\mathbb{P}\left(\sum_{t=T+1}^{T+H} \hat{\sigma}_t^2(y_{1:T}) = \hat{\tau}_c\right)} & \text{otherwise.} \end{cases}$$

Partial abstention `PAbFor` (Partial ABstention in multi-horizon FORecasting) approximates the optimal selection function derived in Theorem 2. To efficiently estimate γ^* that satisfies the coverage constraint, we employ a binary search algorithm. This is possible because the coverage is monotonic with respect to γ (see Lemma 2 in Appendix B.2). However, since the empirical joint distribution of conditional risks may not be continuous, there may not exist a γ that satisfies the coverage constraint exactly. Therefore, we find $\hat{\gamma}_\ell$ and $\hat{\gamma}_r$ that provide the tightest lower and upper bounds on the target coverage on $\mathcal{D}_{calib}$:

$$\hat{\gamma}_\ell = \sup\left\{\gamma \geq 0 : \hat{\phi}\left(\hat{g}_\gamma\right) \leq cH\right\} \qquad \hat{\gamma}_r = \inf\left\{\gamma \geq 0 : \hat{\phi}\left(\hat{g}_\gamma\right) \geq cH\right\} \tag{9}$$

where $\hat{\phi}(\hat{g}_\gamma)$ denotes the empirical coverage on $\mathcal{D}_{calib}$ when using γ. Then, we enforce the coverage constraint exactly by defining $\hat{\gamma}$ as a random variable sampled independently for each test series taking the value $\hat{\gamma}_\ell$ with probability p and $\hat{\gamma}_r$ with probability $(1-p)$, where

$$p = \begin{cases} \frac{c_H - \hat{\phi}(\hat{g}_{\hat{\gamma}_r})}{\hat{\phi}(\hat{g}_{\hat{\gamma}_\ell}) - \hat{\phi}(\hat{g}_{\hat{\gamma}_r})}, & \text{if } \hat{\phi}(\hat{g}_{\hat{\gamma}_\ell}) \neq \hat{\phi}(\hat{g}_{\hat{\gamma}_r}), \\ 1, & \text{otherwise.} \end{cases} \tag{10}$$

The resulting selection function $\hat{g}_{\texttt{PA}}$ is obtained as:

$$\hat{g}_{\texttt{PA}}\left(y_{1:T}\right) = \left(1, \underset{e \in \{0,\ldots,H\}}{\arg\min} \left(\sum_{t=T+1}^{T+e} \hat{\sigma}_t^2\left(y_{1:T}\right) - \hat{\gamma} e\right)\right) \tag{11}$$

The full algorithm can be found in Appendix D.1.

Interval abstention Finally, `IntAbFor` (INTerval ABstention in multi-horizon FORecasting) approximates the optimal selection function derived in Theorem 3. Similar to `PAbFor`, this method relies on the variance estimates $\hat{\sigma}_t^2$, uses a binary search to identify the tightest bounds $\hat{\gamma}_\ell$ and $\hat{\gamma}_r$ around the target coverage, and obtains $\hat{\gamma}$ as a random variable induced by these bounds (see Equation 9 and Equation 10). The key difference lies in the local optimization step required for each time-series. For a fixed $\hat{\gamma}$, instead of simply truncating the horizon, `IntAbFor` estimates the optimal interval (s^*, e^*) by solving the minimization problem in Equation 7. We solve this optimization problem efficiently by decomposing it into two steps. First, for every possible forecast length $h \in \{0, \ldots, H\}$, we identify the start step $\hat{s}_h$ of the subsequence of length h with the lowest cumulative variance:

$$\hat{s}_h(y_{1:T}) = \underset{s \in \{1,\ldots,H-h+1\}}{\arg\min} \sum_{t=T+s}^{T+s+h-1} \hat{\sigma}_t^2(y_{1:T}),$$

Second, among these candidates, we select the interval length $\hat{h}$ that minimizes the Lagrangian cost:

$$\hat{h}\left(y_{1:T}\right) = \underset{h \in \{0,\ldots,H\}}{\arg\min} \left(\sum_{t=T+\hat{s}_h(y_{1:T})}^{T+\hat{s}_h(y_{1:T})+h-1} \hat{\sigma}_t^2\left(y_{1:T}\right) - \hat{\gamma} h\right).$$

Finally, the selection function $\hat{g}_{\texttt{IA}}$ is obtained as:

$$\hat{g}_{\texttt{IA}}\left(y_{1:T}\right) = \left(\hat{s}_{\hat{h}}\left(y_{1:T}\right), \hat{s}_{\hat{h}}\left(y_{1:T}\right) + \hat{h}(y_{1:T}) - 1\right).$$

The full algorithm can be found in Appendix D.2.

4 Experimental Evaluation

We address the following research questions:[5](**Q1**) In full abstention, does `FAbFor` achieve lower selective risk than its competitors? (**Q2**) Does giving the selection function more fine-grained control over the chosen forecasting interval yield a lower selective risk? (**Q3**) Is the coverage constraint satisfied at test time?

[5]The code is available in the supplementary material and will be publicly released upon acceptance

4.1 EXPERIMENTAL SETUP

Baselines. We compare FAbFor against three baselines. AdaptiveCF rejects the time series based on the sum of the conformal prediction intervals' widths over the H-step horizon. We use the adaptive approach, as the standard conformal method yields equal-width intervals for all time-series (Stankeviciute et al., 2021). MQ−RNN trains the selection function using a multi-quantile loss and rejects time series based on the sum of the predicted quantile intervals' width across the H-step horizon (Wen et al., 2018). DP−RNN employs Monte Carlo dropout during training and inference. The selection function is trained to reject time series with high predictive uncertainty, measured as the sum of the standard deviations across the H-step horizon (Gal & Ghahramani, 2016a).

Data. We evaluate the effectiveness of our strategies on the 24 datasets described in Table 2 (Appendix E.1) Dau et al. (2019). We include datasets with diverse characteristics, *e.g.*, the number of time-series ranges from 380 to 38400, and the forecast horizon ranges between 6 and 50. All datasets are publicly available. Appendix E.1 provides full details on the datasets and their preprocessing.

Evaluation metrics. As standard in the bounded-abstention framework, we consider two key quantities: the *empirical selective risk* and the *empirical coverage* (Geifman & El-Yaniv, 2017). Specifically, we compute the selective risk as the average MSE per accepted time step (see Equation 2). Moreover, we report in Appendix F.8 an experiment where Mean Absolute Error (MAE) is used as risk measure. We assess whether the empirical coverage is within a user-defined tolerance ϵ of the target coverage (Pugnana et al., 2024): $ConSat(\epsilon) = \mathbb{1}_{\hat{\phi}(\hat{g})/H \geq c-\epsilon}(\epsilon)$ where $\hat{\phi}(\hat{g})$ is the empirical coverage on the test set.

Setup. For each dataset, we employ the following procedure. (i) We randomly split the data into training, calibration, and test set $(60\%/20\%/20\%)$. (ii) We train the forecaster $\hat{f}$, the conditional variance estimator $\hat{\sigma}^2(y_{1:T})$ and the baselines on the training set. (iii) We calibrate the rejection parameters on the calibration set for six target coverage levels $c \in \{0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$. Specifically, we determine $\hat{\tau}_c$ (for FAbFor) or $\hat{\gamma}_\ell, \hat{\gamma}_r, p$ (for PAbFor and IntAbFor). For all baselines, we set the rejection threshold as the c-th quantile of their score distribution, ensuring that the $(1-c)$ fraction of time series with the largest uncertainty are rejected. (iv) For each coverage level we determine the selection function $\hat{g}$ and (v) compute the empirical selective risk and the empirical coverage on the test set. We report the average results across ten random seeds to split the data and initialize the network weights.

Model Architectures and Hyperparameters. The forecaster f relies on a common backbone across all methods, consisting of a single-layer Vanilla LSTM (Hochreiter & Schmidhuber, 1997) with 20 hidden units. We used LSTMs due to their ability to capture temporal correlations and long-range dynamics across time steps (Taieb & Atiya, 2016; Stankeviciute et al., 2021). We ablate the backbone using a Transformer in Appendix F.7 and consider forecasters that generate the H predictions autoregressively rather than simultaneously in Appendix F.9. For our proposed strategies (FAbFor, PAbFor, and IntAbFor), this backbone branches into two separate heads – one for prediction and one for variance estimation – where each head consists of an MLP with a single hidden layer of 40 neurons and ReLU activation. These are trained end-to-end using the loss in Equation 8 with the variance weight β set to 0.5 (Seitzer et al., 2022). An ablation study on this hyperparameter is provided in Appendix F.6. Similarly, for the baselines, the LSTM backbone feeds into a MLP with a single hidden layer of 40 neurons and ReLU activation. The baselines use this architecture but employ different loss functions (*e.g.*, quantile loss) to estimate the uncertainty (*e.g.*, width of the prediction intervals). Moreover, for AdaptiveCF, we set the sensitivity parameter β to 1; for MQ−RNN, we predict the 5th and 95th percentiles by setting the quantile loss parameters to $q = 0.05$; and for DP−RNN, we apply a dropout rate of 0.1 and compute the standard deviation over 100 Monte Carlo samples. All networks are trained for 500 epochs using the Adam optimizer with a learning rate of 0.001. See Appendix E.2 for full experimental details.

4.2 EXPERIMENTAL RESULTS

(Q1) FAbFor significantly outperforms the competitors. Figure 2 shows the average empirical selective risk $\hat{\mathcal{R}}$ across six target coverage levels for all considered methods in the full abstention setting on four representative datasets (see Appendix F.1 for full results). As expected, the selective risk decreases for all methods as more time-series are rejected. Overall, FAbFor consistently out-

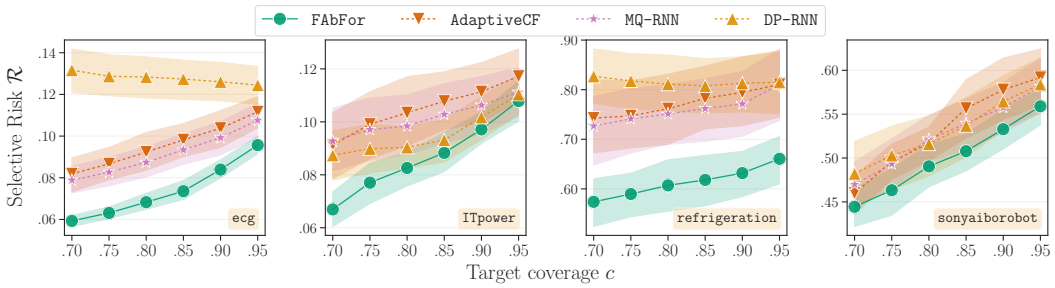

Figure 2: Average selective risk (with 95% c.i.) for all methods across six target coverages in the full abstention setting on four representative datasets.

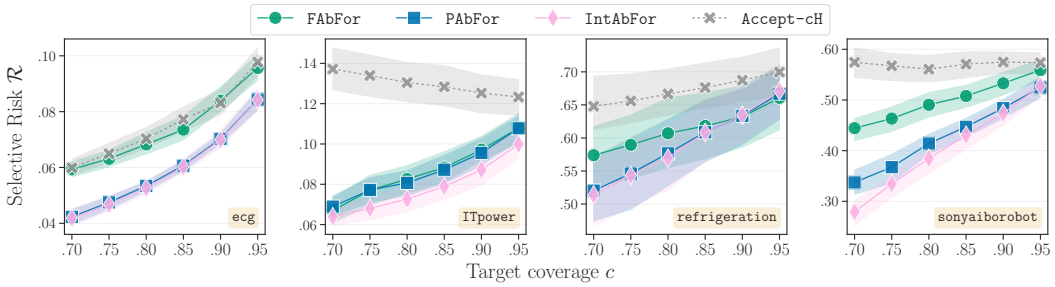

Figure 3: Average selective risk (with 95% c.i.) for the three proposed strategies and `Accept-cH` across six target coverages on four representative datasets.

performs all baselines on 21 out of 24 datasets, reducing the selective risk by an average of 14% vs `AdaptiveCF` and `MQ-RNN`, and 19% vs `DP-RNN`. Moreover, `FAbFor` achieves lower selective risk in around 80% of the experiments against the two runner-ups, `AdaptiveCF` and `MQ-RNN`.

To evaluate statistical significance, we apply the method proposed in Demšar (2006) separately for each target coverage and use the Bonferroni correction to account for multiple hypothesis tests. For each target coverage, the Friedman test rejects the null-hypothesis that all methods perform similarly. According to the Nemenyi post-hoc test (Demšar, 2006), `FAbFor` significantly outperforms the baselines for each considered target coverage; Figure 6 in Appendix F.1 shows all CD plots.

(Q2) Finer-grained control in the abstention strategy yields lower selective risk. To assess if having more fine-grained control over the selected interval yields a lower selective risk, we compare our three proposed methods to a naive baseline `Accept-cH` that predicts the first cH steps in the horizon.[6] Additionally, Appendix F.3 and Appendix F.4 show that `PAbFor` and `IntAbFor` also consistently outperform the competitors in the partial and interval abstention settings, respectively.

Figure 3 reports the average selective risk for the three proposed strategies and `Accept-cH` across six target coverage levels on four representative datasets (see Appendix F.2 for full results). All our approaches significantly outperform `Accept-cH`, confirming that an effective abstention strategy requires assessing the cumulative conditional risk. Next, we rank the methods from best (rank 1) to worst (rank 4) for each experiment and report the average ranks in Table 1. Across all coverages, we observe a performance hierarchy consistent with the level of control of the selection function: `IntAbFor` consistently achieves the lowest (best) rank, followed by `PAbFor` and `FAbFor`. Quantitatively, `PAbFor` reduces the average selective risk across all datasets by 7% compared to `FAbFor`, and `IntAbFor` achieves a further 2% reduction. The similar performance of `PAbFor` and `IntAbFor` arises because uncertainty typically increases over the time horizon. Hence, `IntAbFor`'s interval typically starts at $s = 1$, resulting in the same forecast interval as `PAbFor`. However, in some datasets the added flexibility is very important. For example, on `ITpower` and `sonyaiborobot`, `IntAbFor` achieves notably lower risk than `PAbFor`: on these datasets, `IntAbFor`'s selection function sets $s > 1$ for nearly 40% of the test time-series.

(Q3) Finer-grained control in the abstention strategy enables precise coverage satisfaction. Figure 4 reports the average coverage satisfaction $ConSat$ across five tolerances ϵ for all considered

[6] To exactly enforce the coverage constraint, it always accepts the first $\lfloor cH \rfloor$ steps and accepts the $(\lfloor cH \rfloor + 1)$-th step with probability $p = cH - \lfloor cH \rfloor$.

Table 1: Average rank ($\pm$ std.) for each method across all datasets for six target coverage levels.

c	**Ranks** (avg. $\pm$ std.) FAbFor	PAbFor	IntAbFor	Accept-cH
.70	2.77 ± 0.84	1.78 ± 0.63	$\mathbf{1.60 \pm 0.67}$	3.85 ± 0.40
.75	2.78 ± 0.84	1.73 ± 0.62	$\mathbf{1.64 \pm 0.70}$	3.85 ± 0.41
.80	2.78 ± 0.81	1.71 ± 0.64	$\mathbf{1.66 \pm 0.70}$	3.85 ± 0.40
.85	2.79 ± 0.76	1.74 ± 0.64	$\mathbf{1.62 \pm 0.72}$	3.85 ± 0.44
.90	2.69 ± 0.85	1.79 ± 0.64	$\mathbf{1.67 \pm 0.74}$	3.85 ± 0.45
.95	2.58 ± 0.86	1.82 ± 0.69	$\mathbf{1.74 \pm 0.77}$	3.86 ± 0.45

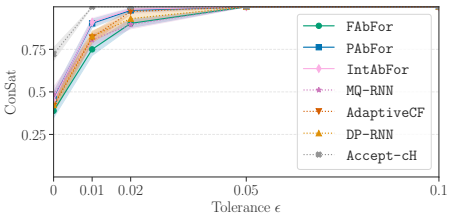

Figure 4: Constraint satisfaction ($ConSat$) as a function of the tolerance ϵ for our three strategies and the baselines.

methods. In the full abstention setting, all strategies exhibit comparable constraint satisfaction rates, consistently meeting the constraint when $\epsilon \geq 0.05$. Notably, `PAbFor` and `IntAbFor` outperform all other strategies except `Accept-cH`, which is designed to meet the constraint exactly. This suggests that the finer-grained control of these abstention strategies, *i.e.*, the ability to accept specific intervals within the horizon, enables more precise control over the coverage constraint.

5 Related Work

Selective Prediction. *Bounded-abstention* (or *Selective Prediction)* (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2019) methods aim to solve the trade-off between selective risk and coverage by abstaining from predicting when uncertainty is high (Hendrickx et al., 2024; Ruggieri & Pugnana, 2025). However, selective predictors have been predominantly studied in the context of classification (Geifman & El-Yaniv, 2017; Pugnana & Ruggieri, 2023; Franc et al., 2023) or regression (Zaoui et al., 2020), while limited attention has been devoted to other ML tasks (Perini & Davis, 2023; Stradiotti et al., 2025). Only a few works have studied abstention mechanisms for time-series inputs (Hatami & Chira, 2013): Pidan & El-Yaniv (2011) modeled the temporal trend as a Hidden Markov Model, while Brusokas et al. (2025) and Inácio et al. (2026) proposed selective forecasting methods that reject the entire forecast horizon based on heuristic confidence scores. In contrast, we formalize three abstention settings without assuming a specific generative process.

Uncertainty quantification in time-series forecasting. Several paradigms have been proposed for uncertainty quantification in time-series forecasting. Bayesian RNNs (Chien & Ku, 2016; Fortunato et al., 2019) capture epistemic uncertainty but require substantial architectural changes and depend heavily on the choice of the prior. Simplifications such as Monte Carlo dropout (Gal & Ghahramani, 2016a;b) reduce complexity, but often yield poorly calibrated intervals (Alaa & van der Schaar, 2020). Quantile RNNs (Wen et al., 2018) learn prediction intervals directly via the pinball loss, but are prone to quantile crossing and overfitting (Gasthaus et al., 2019). More recently, Conformal Forecasting RNNs (Stankeviciute et al., 2021) have emerged as a flexible, post-hoc approach for constructing prediction intervals with finite-sample guarantees. Other conformal approaches for time-series (Xu & Xie, 2023; Auer et al., 2023) calibrate each series' intervals on its own past errors, whereas we learn from multiple series to forecast unseen ones.

6 Conclusions

We formalized the general problem of learning with abstention in multi-horizon forecasting to handle the structured and correlated predictions offered by a multi-horizon forecaster. We proposed three natural ways to instantiate the problem (full, partial, and interval abstention) that give the model progressively more flexibility over the offered forecast. Using the bounded abstention framework, we derived the theoretical optimal selection function for each setting and provided practical algorithms to learn the corresponding models in practice. Empirically, our proposed approaches significantly outperformed several baselines. Moreover, abstention strategies with higher flexibility in the selection of the forecast interval yielded lower selective risks.

Limitations. Our framework assumes we have access to multiple time series and can treat them as exchangeable. Moreover, our approach trains the conditional risk estimator and the predictor jointly, which may be context-dependent in real-life scenarios. Finally, our approach focuses on ambiguity rejection; extending this to novelty rejection is left for future work.

AI USE STATEMENT

In this work, we used generative AI tools primarily for polishing the writing and partially for code development. All AI-assisted work was thoroughly reviewed; specifically, code was verified and tested for correctness by two authors. Generative AI tools were not used to generate research ideas. The authors take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All details necessary to reproduce our experiments are provided in Section 4.1 and Appendix E, including full descriptions of the models and datasets. Section 4.1 presents the overall experimental setup, the hyperparameters and training settings. Appendix E details the used datasets, which are publicly available online (Kelly et al., 2023; UK Health Security Agency, 2025; Mouatadid et al., 2023; Nguyen et al., 2023; Dau et al., 2019). The source code will be publicly released upon acceptance.

REFERENCES

Ahmed M. Alaa and Mihaela van der Schaar. Attentive state-space modeling of disease progression. In *Neural Information Processing Systems*, pp. 11334–11344, 2019.

Ahmed M. Alaa and Mihaela van der Schaar. Frequentist uncertainty in recurrent neural networks via blockwise influence functions. In *International Conference on Machine Learning*, pp. 175–190, 2020.

Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with modern hopfield networks. In *Neural Information Processing Systems*, 2023.

Wei Bao, Jun Yue, and Yulei Rao. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PloS one*, 12(7):e0180944, 2017.

Jonas Brusokas, Seshu Tirupathi, Dalin Zhang, and Torben Bach Pedersen. The time-energy model: Selective time-series forecasting using energy-based models. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=iHYCdTAOqF.

Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. Nhits: Neural hierarchical interpolation for time series forecasting. In *Association for the Advancement of Artificial Intelligence*, pp. 6989–6997, 2023.

Jen-Tzung Chien and Yuan-Chu Ku. Bayesian recurrent neural network for language modeling. *IEEE Transactions on Neural Networks and Learning Systems*, 27(2):361–374, 2016.

Jacob Cohen. A power primer. *Psychological bulletin*, 112(1):155, 1992.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(Jan):1–30, 2006.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, pp. 1605–1641, 2010.

Meire Fortunato, Charles Blundell, and Oriol Vinyals. Bayesian recurrent neural networks. *arXiv preprint arXiv:1704.02798*, 2019.

Vojtech Franc, Daniel Prruvsa, and Václav Vorácek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, pp. 11:1–11:49, 2023.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016a.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *NIPS*, pp. 1019–1027, 2016b.

Jan Gasthaus, Konstantinos Benidis, Yuyang Wang, Syama Sundar Rangapuram, David Salinas, Valentin Flunkert, and Tim Januschowski. Probabilistic forecasting with spline quantile function RNNs. In *Artificial Intelligence and Statistics*, pp. 1901–1910, 2019.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Neural Information Processing Systems*, pp. 4878–4887, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, pp. 2151–2159, 2019.

Nima Hatami and Camelia Chira. Classifiers with a reject option for early time-series classification. In *IEEE Symposium on Computational Intelligence and Ensemble Learning*, pp. 9–16, 2013.

Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine learning with a reject option: A survey. *Machine Learning*, pp. 1–38, 2024.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Ricardo Inácio, Vítor Cerqueira, Marília Barandas, and Carlos Soares. Selective time series forecasting via metalearning. *CoRR*, abs/2606.23448, 2026. doi: 10.48550/ARXIV.2606.23448. URL https://doi.org/10.48550/arXiv.2606.23448.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI machine learning repository, 2023.

Maurice G Kendall et al. A New Measure of Rank Correlation. *Biometrika*, 30(1-2):81–93, 1938.

Bryan Lim. Forecasting treatment responses over time using recurrent marginal structural networks. In *Neural Information Processing Systems*, pp. 7494–7504, 2018.

Soukayna Mouatadid, Paulo Orenstein, Genevieve Flaspohler, Miruna Oprescu, Judah Cohen, Franklyn Wang, Sean Knight, Maria Geogdzhayeva, Sam Levang, Ernest Fraenkel, and Lester Mackey. Subseasonalclimateusa: A dataset for subseasonal forecasting and benchmarking. In *Neural Information Processing Systems*, 2023.

Tung Nguyen, Jason Jewik, Hritik Bansal, Prakhar Sharma, and Aditya Grover. Climatelearn: Benchmarking machine learning for weather and climate modeling. In *Neural Information Processing Systems*, 2023.

Lorenzo Perini and Jesse Davis. Unsupervised anomaly detection with rejection. In *Neural Information Processing Systems*, 2023.

Dmitry Pidan and Ran El-Yaniv. Selective prediction of financial trends with hidden markov models. In *Neural Information Processing Systems*, pp. 855–863, 2011.

Andrea Pugnana and Salvatore Ruggieri. Auc-based selective classification. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, pp. 2494–2514. PMLR, 2023.

Andrea Pugnana, Lorenzo Perini, Jesse Davis, and Salvatore Ruggieri. Deep neural network benchmarks for selective classification. *Journal of Data-centric Machine Learning Research (DMLR)*, 1(17):1–58, 2024.

Stephan Rabanser and Nicolas Papernot. What does it take to build a performant selective classifier? In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/71460926102fade443ea7ec89ae8a73a-Abstract-Conference.html.

Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11), November 2020. ISSN 1942-2466. doi: 10.1029/2020ms002203. URL http://dx.doi.org/10.1029/2020MS002203.

Salvatore Ruggieri and Andrea Pugnana. Things machine learning models know that they don't know. In *Association for the Advancement of Artificial Intelligence*, pp. 28684–28693, 2025.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2022.

Sreelekshmy Selvin, R. Vinayakumar, E. A. Gopalakrishnan, Vijay Krishna Menon, and K. P. Soman. Stock price prediction using LSTM, RNN and CNN-sliding window model. In *International Conference on Advances in Computing, Communications and Informatics*, pp. 1643–1647. IEEE, 2017.

Kamile Stankeviciute, Ahmed M. Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. In *Neural Information Processing Systems*, pp. 6216–6228, 2021.

Luca Stradiotti, Dario Pesenti, Stefano Teso, and Jesse Davis. Learning to reject low-quality explanations via user feedback. *arXiv preprint arXiv:2507.12900*, 2025.

Souhaib Ben Taieb and Amir F. Atiya. A bias and variance analysis for multistep-ahead time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 27(1):62–76, 2016.

UK Health Security Agency. Ukhsa data dashboard. https://ukhsa-dashboard.data.gov.uk, 2025.

Koen W van Arem, Floris Goes-Smit, and Jakob Söhl. Forecasting the future development in quality and value of professional football players for applications in team management. *arXiv preprint arXiv:2502.07528*, 2025.

Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. A multi-horizon quantile recurrent forecaster. *arXiv preprint arXiv:1711.11053*, 2018.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, pp. 38707–38727, 2023.

Ahmed Zaoui, Christophe Denis, and Mohamed Hebiri. Regression with reject option and application to kNN. In *Neural Information Processing Systems*, 2020.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Association for the Advancement of Artificial Intelligence*, pp. 11106–11115, 2021.

A BROADER IMPACT

The deployment of multi-horizon forecasting models in high-stakes applications, such as healthcare and finance, is often hindered by the high cost of mistakes, which can have severe consequences. Enabling multi-horizon forecasters to abstain from making predictions in uncertain scenarios yields significant benefits across these domains.

First, by prioritizing the minimization of selective risk, our framework mitigates the potential consequences of erroneous forecasts, *e.g.*, preventing incorrect medical prognoses. By allowing the model to express uncertainty through full, partial, or interval abstention, we enhance the overall safety and reliability of automated decision-making systems.

Second, our approach facilitates more efficient human-AI collaboration. By identifying specific time steps or intervals where the model is uncertain, we allow human experts to focus their attention and limited resources solely on these challenging steps. This reduces the cognitive load on practitioners, ensuring that human intervention is prioritized for cases where the model is unreliable.

B PROOFS

For each theorem, we first repeat the statement in the main paper and then provide the proof.

B.1 PROOF OF THEOREM 1

Theorem. *Given a forecaster f and a target coverage $c \in (0, 1]$, and assuming the cumulative conditional risk is not degenerate such that $\mathbb{P}\left(\sum_{t=T+1}^{T+H} \rho_t\left(y_{1:T}\right) < \tau_c^*\right) > 0$, the optimal selection function g_{FA}^* that is a solution to Equation 3 in full abstention is:*

$$g_{FA}^*\left(y_{1:T}\right) = \begin{cases} (1, H) & \text{if} \quad \sum_{t=T+1}^{T+H} \rho_t\left(y_{1:T}\right) < \tau_c^* \\ (1, \xi) & \text{if} \quad \sum_{t=T+1}^{T+H} \rho_t\left(y_{1:T}\right) = \tau_c^* \\ (1, 0) & \text{otherwise} \end{cases} \tag{12}$$

where ξ is a random variable such that $\xi = H$ with probability κ and $\xi = 0$ with probability $1 - \kappa$, and

$$\begin{aligned} \rho_t\left(y_{1:T}\right) &= \mathbb{E}_{Y_{T+1:T+H}|Y_{1:T}}\left[\ell\left(y_t, f_t(y_{1:T})\right) \mid Y_{1:T} = y_{1:T}\right] \\ \tau_c^* &= \inf\left\{v \mid \mathbb{E}_{Y_{1:T}}\left[\mathbb{1}\left\{\sum_{t=T+1}^{T+H} \rho_t\left(y_{1:T}\right) < v\right\}\right] \geq c\right\} \end{aligned} \tag{13}$$

$$\kappa = \begin{cases} 0 & \text{if} \quad \mathbb{P}\left(\sum_{t=T+1}^{T+H} \rho_t\left(y_{1:T}\right) = \tau_c^*\right) = 0, \\ \frac{c - \mathbb{P}\left(\sum_{t=T+1}^{T+H} \rho_t(y_{1:T}) < \tau_c^*\right)}{\mathbb{P}\left(\sum_{t=T+1}^{T+H} \rho_t(y_{1:T}) = \tau_c^*\right)} & \text{otherwise}. \end{cases}$$

Proof. The selection function in full abstention returns a pair of indices where the start step is fixed at 1, while the end step can be 0 (*i.e.*, reject the entire horizon) or H (*i.e.*, accept the entire horizon):

$$g_{\mathrm{FA}}\left(y_{1:T}\right) = \left(1, e\left(y_{1:T}\right)\right)$$

Thus, we rewrite the selective risk in Equation 2 for the full abstention setting as:

$$\mathcal{R}(f, g_{\mathrm{FA}}) = \frac{\mathbb{E}_{Y_{1:T+H}}\left[\sum_{t=T+1}^{T+e(y_{1:T})} \ell\left(y_t, f_t\left(y_{1:T}\right)\right)\right]}{\mathbb{E}_{Y_{1:T}}\left[e\left(y_{1:T}\right)\right]} = \frac{\mathbb{E}_{Y_{1:T+H}}\left[\frac{e(y_{1:T})}{H} \sum_{t=T+1}^{T+H} \ell\left(y_t, f_t\left(y_{1:T}\right)\right)\right]}{\mathbb{E}_{Y_{1:T}}\left[e\left(y_{1:T}\right)\right]}$$

Thus, the optimization problem in Equation 3 can be solved by finding the optimal end step e^* as:

$$e^* = \arg\min_e \frac{\mathbb{E}_{Y_{1:T}}\left[\frac{e}{H} r_{1:H}\left(y_{1:T}\right)\right]}{\phi\left(e\right)} \quad \text{s.t.} \quad \phi\left(e\right) \geq cH \tag{14}$$

Let the total conditional risk over the entire horizon $r_{1:H}(y_{1:T})$ be defined as

$$r_{1:H}(y_{1:T}) = \sum_{t=T+1}^{T+H} \rho_t(y_{1:T}) = \mathbb{E}_{Y_{T+1:T+H}|Y_{1:T}}\left[\sum_{t=T+1}^{T+H} \ell(y_t, f_t(y_{1:T})) \mid Y_{1:T} = y_{1:T}\right].$$

We will prove the theorem by showing that any other selection function g_{FA} feasible to Equation 14 that selects a different e is not an optimal solution. We examine three cases.

Case 1. We consider functions where the end step is not always H when $r_{1:H}(y_{1:T}) < \tau_c^*$. Thus, we define $g_{\text{FA}}^{-}(y_{1:T}) = (1, \bar{e}(y_{1:T}))$ such that $\bar{e}(y_{1:T})$:

$$\mathbb{E}_{Y_{1:T}}[\bar{e}(y_{1:T})\mathbb{1}\{r_{1:H}(y_{1:T}) < \tau_c^*\}] < \mathbb{E}_{Y_{1:T}}[e^*(y_{1:T})\mathbb{1}\{r_{1:H}(y_{1:T}) < \tau_c^*\}] \tag{15}$$

Thus, there is a subset $Y'_{1:T} \subseteq \mathcal{Y}_{1:T}$ such that

$$\forall y_{1:T} \in Y'_{1:T} : r_{1:H}(y_{1:T}) \geq \tau_c^*$$

and

$$\mathbb{E}_{Y'_{1:T}}[\bar{e}(y_{1:T})] = \mathbb{E}_{Y_{1:T}}[H\mathbb{1}\{r_{1:H}(y_{1:T}) < \tau_c^*\}] - \mathbb{E}_{Y_{1:T}}[\bar{e}(y_{1:T})\mathbb{1}\{r_{1:H}(y_{1:T}) < \tau_c^*\}] > 0 \tag{16}$$

where the inequality holds for equation 15. Then, we denote the objective function in Equation 14 as

$$F(e) = \frac{\mathbb{E}_{Y_{1:T}}\left[\frac{e(y_{1:T})}{H} r_{1:H}(y_{1:T})\right]}{\mathbb{E}_{Y_{1:T}}[e(y_{1:T})]} = \frac{\mathbb{E}_{Y_{1:T}}\left[\frac{e(y_{1:T})}{H} r_{1:H}(y_{1:T})\right]}{\phi(e)}$$

and define $g'_{\text{FA}}(y_{1:T})$ that selects $(1, e'(y_{1:T}))$ as

$$g'_{\text{FA}}(y_{1:T}) = \begin{cases} (1, H) & \text{if} \quad r_{1:H}(y_{1:T}) < \tau_c^*, \\ (1, 0) & \text{if} \quad y_{1:T} \in Y'_{1:T}, \\ g_{\text{FA}}^{-}(y_{1:T}) & \text{otherwise} \end{cases}$$

where g'_{FA} is a solution to equation 14 since $\phi(\bar{e}) = \phi(e')$. Now, if we compute

$$\begin{aligned}
\phi(\bar{e})(F(\bar{e}) - F(e')) &= \mathbb{E}_{Y'_{1:T}}\left[r_{1:H}(y_{1:T})\frac{\bar{e}(y_{1:T})}{H}\right] - \mathbb{E}_{Y_{1:T}}[r_{1:H}(y_{1:T})\mathbb{1}\{r_{1:H}(y_{1:T}) < \tau_c^*\}] \\
&\quad + \mathbb{E}_{Y_{1:T}}\left[r_{1:H}(y_{1:T})\frac{\bar{e}(y_{1:T})}{H}\mathbb{1}\{r_{1:H}(y_{1:T}) < \tau_c^*\}\right] \\
&\geq \mathbb{E}_{Y_{1:T}}\left[\tau_c^*\left(1 - \frac{\bar{e}(y_{1:T})}{H}\right)\mathbb{1}\{r_{1:H}(y_{1:T}) < \tau_c^*\}\right] \\
&\quad - \mathbb{E}_{Y_{1:T}}\left[r_{1:H}(y_{1:T})\left(1 - \frac{\bar{e}(y_{1:T})}{H}\right)\mathbb{1}\{r_{1:H}(y_{1:T}) < \tau_c^*\}\right] \\
&= \mathbb{E}_{Y_{1:T}}\left[(\tau_c^* - r_{1:H}(y_{1:T}))\left(1 - \frac{\bar{e}(y_{1:T})}{H}\right)\mathbb{1}\{r_{1:H}(y_{1:T}) < \tau_c^*\}\right] > 0
\end{aligned}$$

Where the last inequality comes from linearity of the expectation given equation 16 and $(\tau_c^* - r_{1:H}(y_{1:T})) > 0$ over the set $\mathbb{1}\{r_{1:H}(y_{1:T}) < \tau_c^*\}$. Since, this value is greater than zero we have that g_{FA}^{-} is not an optimal solution.

Case 2. We want to show that for any other function $g_{\text{FA}}^{-} = (1, \bar{e})$ as

$$g_{\text{FA}}^{-}(y_{1:T}) = \begin{cases} g_{\text{FA}}^{*}(y_{1:T}) & \text{if} \quad r_{1:H}(y_{1:T}) < \tau_c^*, \\ (1, \bar{\xi}) & \text{if} \quad r_{1:H}(y_{1:T}) = \tau_c^* \\ (1, \tilde{\xi}) & \text{if} \quad r_{1:H}(y_{1:T}) > \tau_c^* \end{cases}$$

where $\bar{\xi}$ is a random variable such that $\bar{\xi} = H$ with probability $\bar{\kappa}$ and $\bar{\xi} = 0$ with probability $1 - \bar{\kappa}$ with $\kappa \neq \bar{\kappa}$, and $\tilde{\xi}$ is another random variable to accept the remaining required steps in the region $r_{1:H}(y_{1:T}) > \tau_c^*$ to satisfy the coverage constraint if needed, then $F(\bar{e}) > F(e^*)$.

We distinguish two sub-cases:

a. $\bar{\kappa} > \kappa$. Then,

$$\mathbb{E}_{Y_{1:T}}\left[\bar{e}\left(y_{1:T}\right)\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) < \tau_c^*\right\}\right] + \mathbb{E}_{Y_{1:T}}\left[\bar{e}\left(y_{1:T}\right)\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) = \tau_c^*\right\}\right]$$
$$> \mathbb{E}_{Y_{1:T}}\left[e^*\left(y_{1:T}\right)\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) < \tau_c^*\right\}\right] + \mathbb{E}_{Y_{1:T}}\left[\xi^*\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) = \tau_c^*\right\}\right] = cH$$

Since the coverage is already met, $\tilde{\xi}$ is not needed and set always to 0. Moreover, we have that $\phi\left(\bar{e}\right) > cH$ because $\bar{e}$ accepts more steps at the boundary condition than e^*. We can rewrite the objective function for $\bar{e}$ as:

$$F(\bar{e}) = \frac{\mathbb{E}_{Y_{1:T}}\left[\frac{e^*(y_{1:T})}{H} r_{1:H}\left(y_{1:T}\right)\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) \leq \tau_c^*\right\} + \left(\bar{\kappa} - \kappa\right)\tau_c^*\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) = \tau_c^*\right\}\right]}{\mathbb{E}_{Y_{1:T}}\left[e^*(y_{1:T}) + \left(\bar{\kappa} - \kappa\right)H\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) = \tau_c^*\right\}\right]}$$

Let $F(e^*) = \frac{N(e^*)}{\phi(e^*)}$, then:

$$F(\bar{e}) - F(e^*) = \frac{N(e^*) + \mathbb{E}_{Y_{1:T}}\left[\left(\bar{\kappa} - \kappa\right)\tau_c^*\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) = \tau_c^*\right\}\right]}{\phi(e^*) + \mathbb{E}_{Y_{1:T}}\left[\left(\bar{\kappa} - \kappa\right)H\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) = \tau_c^*\right\}\right]} - \frac{N(e^*)}{\phi(e^*)}$$
$$= \frac{\mathbb{E}_{Y_{1:T}}\left[\left(\bar{\kappa} - \kappa\right)\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) = \tau_c^*\right\}\right]\left(\frac{\tau_c^*}{H}\phi(e^*) - N(e^*)\right)}{\phi(e^*)\left(\phi(e^*) + \mathbb{E}_{Y_{1:T}}\left[\left(\bar{\kappa} - \kappa\right)H\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) = \tau_c^*\right\}\right]\right)} > 0.$$

The last inequality comes from the fact that (i) $\bar{\kappa} - \kappa > 0$ and (ii) the optimal policy e^* only accepts sequences whose cumulative risk is bounded by τ_c^*, thus its total expected risk $N(e^*)$ must be strictly less than the maximum possible risk over that same coverage, $\frac{\tau_c^*}{H}\phi(e^*)$. Hence, $g_{\bar{\text{FA}}}^-$ is not an optimal solution.

b. $\bar{\kappa} < \kappa$. In this case, we have that:

$$cH = \mathbb{E}_{Y_{1:T}}\left[e^*(y_{1:T})\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) < \tau_c^*\right\} + e^*(y_{1:T})\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) = \tau_c^*\right\}\right]$$
$$= \mathbb{E}_{Y_{1:T}}\left[\bar{e}(y_{1:T})\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) < \tau_c^*\right\}\right]$$
$$+ \mathbb{E}_{Y_{1:T}}\left[\bar{e}(y_{1:T})\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) = \tau_c^*\right\} + \bar{e}(y_{1:T})\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) > \tau_c^*\right\}\right]$$

Thus,

$$\mathbb{E}_{Y_{1:T}}\left[\xi^*\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) = \tau_c^*\right\}\right] = \mathbb{E}_{Y_{1:T}}\left[\bar{\xi}\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) = \tau_c^*\right\} + \tilde{\xi}\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) > \tau_c^*\right\}\right]$$

Now, if we compute

$$\phi\left(\bar{e}\right)\left(F\left(\bar{e}\right) - F\left(e^*\right)\right) = \mathbb{E}_{Y_{1:T}}\left[\frac{\bar{\xi}}{H}\tau_c^*\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) = \tau_c^*\right\} + \frac{\tilde{\xi}}{H}r_{1:H}(y_{1:T})\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) > \tau_c^*\right\}\right]$$
$$- \mathbb{E}_{Y_{1:T}}\left[\frac{\xi^*}{H}\tau_c^*\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) = \tau_c^*\right\}\right]$$
$$= \mathbb{E}_{Y_{1:T}}\left[\frac{\tilde{\xi}}{H}r_{1:H}(y_{1:T})\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) > \tau_c^*\right\} - \frac{\tilde{\xi}}{H}\tau_c^*\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) > \tau_c^*\right\}\right]$$
$$= \mathbb{E}_{Y_{1:T}}\left[\frac{\tilde{\xi}}{H}\left(r_{1:H}(y_{1:T}) - \tau_c^*\right)\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) > \tau_c^*\right\}\right] > 0$$

Where the last inequality comes from linearity of the expectation given $\left(r_{1:H}\left(y_{1:T}\right) - \tau_c^*\right) > 0$ over the set $\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) > \tau_c^*\right\}$. Since, this value is greater than zero we have that $g_{\bar{\text{FA}}}^-$ is not an optimal solution.

Case 3. In the final case, we consider functions $g_{\bar{\text{FA}}}^- = (1, \bar{e})$ that do not always return $(1, 0)$ when $r_{1:H}\left(y_{1:T}\right) > \tau_c^*$. In this case, we have that $\phi\left(\bar{e}\right) > cH$, and

$$F(\bar{e}) - F(e^*) = \frac{N(e^*) + \mathbb{E}_{Y_{1:T}}\left[\frac{\bar{e}(y_{1:T})}{H}r_{1:H}(y_{1:T})\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) > \tau_c^*\right\}\right]}{\phi(e^*) + \mathbb{E}_{Y_{1:T}}\left[\bar{e}(y_{1:T})\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) > \tau_c^*\right\}\right]} - \frac{N(e^*)}{\phi(e^*)}$$
$$= \frac{\mathbb{E}_{Y_{1:T}}\left[\bar{e}(y_{1:T})\left(\frac{r_{1:H}(y_{1:T})}{H}\phi(e^*) - N(e^*)\right)\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) > \tau_c^*\right\}\right]}{\phi(e^*)\left(\phi(e^*) + \mathbb{E}_{Y_{1:T}}\left[\bar{e}(y_{1:T})\mathbb{1}\left\{r_{1:H}\left(y_{1:T}\right) > \tau_c^*\right\}\right]\right)} > 0.$$

The last inequality holds because we strictly have $r_{1:H}(y_{1:T}) > \tau_c^*$ for the considered set. Furthermore, as established previously, the optimal policy e^* only accepts sequences whose cumulative risk is bounded by τ_c^*. Thus, its total expected risk $N(e^*)$ must be strictly less than the maximum possible risk over that same coverage, meaning $N(e^*) < \frac{\tau_c^*}{H}\phi(e^*) < \frac{r_{1:H}(y_{1:T})}{H}\phi(e^*)$ when $r_{1:H}(y_{1:T}) > \tau_c^*$. Hence, g_{FA}^- is not an optimal solution. □

B.2 Proof of Theorem 2

Theorem. *Assume the joint vector $(\rho_{T+1}(y_{1:T}), \ldots, \rho_{T+H}(y_{1:T}))$ is absolutely continuous. Given a forecaster f and a target coverage $c \in (0, 1]$, the optimal selection function g_{PA}^* in partial abstention returns a pair of indices (s^*, e^*), such that $s^* = 1$ and:*

$$e^* = \underset{e \in \{0,\ldots,H\}}{\arg\min} \left(\sum_{t=T+1}^{T+e} \rho_t\left(y_{1:T}\right) - \gamma^* e \right) \tag{17}$$

where $\gamma^* = \lambda^* + \eta^*$, $\lambda^* = \mathcal{R}\left(f, g_{\text{PA}}^*\right)$ is the optimal risk, and η^* is the Lagrangian multiplier which is chosen such that the pair (e^*, η^*) satisfies the Karush–Kuhn–Tucker (KKT) conditions:

$$\eta^* \geq 0, \qquad \mathbb{E}_{Y_{1:T}}\left[e^*\right] \geq cH, \qquad \eta^*\left(\mathbb{E}_{Y_{1:T}}\left[e^*\right] - cH\right) = 0.$$

First, we introduce two lemmas that will be used later in the proof.

Lemma 1. *Assume $D(e) > 0$ for any $e \in \mathcal{E}'$ and $\mathcal{R}(e) = \frac{N(e)}{D(e)}$. Then minimizing $\mathcal{R}(e)$ is equivalent to solving the following parametric linearized problem:*

$$e^* \in \underset{e \in \mathcal{E}'}{\arg\min} \frac{N(e)}{D(e)} \iff e^* \in \underset{e \in \mathcal{E}'}{\arg\min} \big(N(e) - \lambda^* D(e)\big)$$

where $\lambda^* = \mathcal{R}(e^*) = \frac{N(e^*)}{D(e^*)} \geq 0$.

Proof. We prove each direction separately.

⇒ Let $e^* \in \underset{e \in \mathcal{E}'}{\arg\min}\, \mathcal{R}(e)$, *i.e.*,

$$\frac{N(e^*)}{D(e^*)} \leq \frac{N(e)}{D(e)}, \qquad \forall e \in \mathcal{E}'. \tag{18}$$

Since $D(e) > 0$, we can rewrite this inequality as

$$N(e) - \frac{N(e^*)}{D(e^*)} D(e) \geq 0 \tag{19}$$

Let $\lambda^* = \frac{N(e^*)}{D(e^*)}$ be the risk of the optimal solution, then the inequality simplifies to:

$$N(e) - \lambda^* D(e) \geq 0 \tag{20}$$

Evaluating this parametric form at the optimal solution e^*, we obtain

$$N(e) - \lambda^* D(e) \geq 0 = N(e^*) - \frac{N(e^*)}{D(e^*)} D(e^*) = N(e^*) - \lambda^* D(e^*). \tag{21}$$

Hence, there exists λ^* such that e^* also minimizes the parametric form $N(e) - \lambda^* D(e)$.

⇐ Let $\lambda^* = \frac{N(e^*)}{D(e^*)}$, we have that

$$N(e^*) - \lambda^* D(e^*) \leq N(e) - \lambda^* D(e), \qquad \forall e \in \mathcal{E}'. \tag{22}$$

Dividing both sides by $D(e) > 0$, we obtain

$$\lambda^* = \frac{N(e^*)}{D(e^*)} \leq \frac{N(e)}{D(e)} \qquad \forall e \in \mathcal{E}', \tag{23}$$

so e^* also minimizes the fractional objective.

□

Lemma 2. *Given $\gamma > 0$, define for a fixed $y_{1:T}$*

$$e(\gamma) := \underset{e \in \{0,\dots,H\}}{\arg\min} \left\{ r_{1:e}(y_{1:T}) - \gamma e \right\}, \tag{24}$$

$$D(\gamma) := \mathbb{E}_{Y_{1:T}}\left[e(\gamma) \right]. \tag{25}$$

where we select the smallest minimizer in case of ties. Then for any fixed $y_{1:T}$, $e(\gamma)$ is non-decreasing in γ. Consequently, $D(\gamma)$ is also non-decreasing in γ.

Proof. We have to prove that for a given $y_{1:T}$:

$$\gamma_2 > \gamma_1 \Rightarrow e(\gamma_2) \geq e(\gamma_1) \tag{26}$$

Let $e_1 := e(\gamma_1)$ and $e_2 := e(\gamma_2)$. By optimality of e_1 at γ_1 and e_2 at γ_2, we have:

$$r_{1:e_1} - \gamma_1 e_1 \leq r_{1:e_2} - \gamma_1 e_2, \qquad r_{1:e_2} - \gamma_2 e_2 \leq r_{1:e_1} - \gamma_2 e_1. \tag{27}$$

Summing these inequalities gives

$$(\gamma_1 - \gamma_2)(e_2 - e_1) \leq 0. \tag{28}$$

Since $\gamma_2 > \gamma_1$, the first factor is negative, hence $e_2 - e_1 \geq 0$, *i.e.*, $e_2 \geq e_1$. Linearity and monotonicity of expectation imply that $D(\gamma)$ is also non-decreasing. □

We can now prove the theorem.

Proof. We define the cumulative conditional risk at step e as for a time series $y_{1:T}$:

$$r_{1:e}\left(y_{1:T}\right) = \sum_{t=T+1}^{T+e} \rho_t\left(y_{1:T}\right) = \mathbb{E}_{Y_{T+1:T+H}|Y_{1:T}}\left[\sum_{t=T+1}^{T+e} \ell\left(y_t, f_t\left(y_{1:T}\right)\right) \mid Y_{1:T} = y_{1:T} \right].$$

In the partial abstention setting, the start step is fixed at $s = 1$. Thus, the selection function $g_{\mathbb{PA}}$ needs only to determine the end step $e \in \{0, \dots, H\}$. The length of the forecast is $h(g_{\mathbb{PA}}) = e$. The optimization problem in Equation 3 can then be rewritten as:

$$\min_e \frac{\mathbb{E}_{Y_{1:T}}\left[r_{1:e(y_{1:T})}\left(y_{1:T}\right) \right]}{\mathbb{E}_{Y_{1:T}}\left[e\left(y_{1:T}\right) \right]} \quad \text{s.t.} \quad \mathbb{E}_{Y_{1:T}}\left[e\left(y_{1:T}\right) \right] \geq cH$$

Let $N(e) := \mathbb{E}_{Y_{1:T}}\left[r_{1:e(y_{1:T})}\left(y_{1:T}\right) \right]$ and $D(e) := \mathbb{E}_{Y_{1:T}}\left[e\left(y_{1:T}\right) \right]$. Given the set of admissible selection functions $\mathcal{E}' = \{e : D(e) \geq cH\}$, by Lemma 1 the objective to minimize becomes:

$$\min_{e \in \mathcal{E}'} \left\{ N(e) - \lambda^* D(e) \right\},$$

or, by making the coverage constraint explicit:

$$\min_e \left\{ N(e) - \lambda^* D(e) \right\} \quad \text{s.t.} \quad D(e) \geq cH.$$

Solving this is equivalent to the following parametric Lagrangian problem:

$$\mathcal{L}(e, \lambda^*, \eta) = N(e) - \lambda^* D(e) + \eta(cH - D(e)) = N(e) - (\lambda^* + \eta) D(e) + \eta cH,$$

where λ^* comes from the linearized transformation in Lemma 1 and $\eta \geq 0$ is the Lagrange multiplier associated with the coverage constraint.

Let $\gamma := \lambda^* + \eta$, the expectation is linear and e operates to each time-series $y_{1:T}$ independently. Consequently, minimizing the objective with respect to e is equivalent to selecting the end step that minimizes each local term independently:

$$\min_e \mathbb{E}_{Y_{1:T}}\left[r_{1:e(y_{1:T})}\left(y_{1:T}\right) - \gamma e\left(y_{1:T}\right) \right] = \mathbb{E}_{Y_{1:T}}\left[\min_{e \in \{0..H\}} \left(r_{1:e}\left(y_{1:T}\right) - \gamma e \right) \right]$$

Thus, for each sequence $y_{1:T}$, the optimal local decision is

$$e^*\left(y_{1:T}\right) = \underset{e\in\{0..H\}}{\arg\min}\left\{r_{1:e}\left(y_{1:T}\right) - \gamma e\right\}.$$

By Lemma 2, the expected coverage $D(\gamma) = \mathbb{E}_{Y_{1:T}}\left[e_\gamma(y_{1:T})\right]$ is non-decreasing in γ and

$$\lim_{\gamma\to 0^+} D(\gamma) = 0, \qquad \lim_{\gamma\to+\infty} D(\gamma) = H. \tag{29}$$

Since we assume the joint vector of the conditional risks is absolutely continuous, the expected coverage function $D(\gamma)$ is absolutely continuous with respect to γ. By the Intermediate Value Theorem, and given the boundary conditions, there exists a γ such that $D(\gamma) = cH$ exactly.

Let η^* be optimal dual variable associated with the constrained fractional problem, and let $\gamma^* = \lambda^* + \eta^*$. By construction, $e^* := e(\gamma^*)$ minimizes the Lagrangian $\mathcal{L}(e, \eta^*)$. The complementary slackness condition gives

$$\eta^*\left(cH - D(e^*)\right) = 0. \tag{30}$$

Hence, two cases arise:

- If $D(e^*) = cH$, the coverage constraint is active and $\eta^* \geq 0$ may be positive. In this case, $\lambda^* = \gamma^* - \eta^*$ and, by definition of the fractional optimum, $\lambda^* = \mathcal{R}(e^*)$.
- If $D(e^*) > cH$, the constraint is inactive and $\eta^* = 0$, so that $\gamma^* = \lambda^*$.

By solving the Lagrangian with complementary slackness, we obtain e^* that minimizes the linearized parametric problem under the constraint $\mathbb{E}_{Y_{1:T}}\left[e\left(y_{1:T}\right)\right] \geq cH$. By the equivalence established above, this e^* also minimizes the original fractional objective under the same constraint. Therefore, e^* is the solution to the constrained fractional problem.

Finally, the optimal selection function $g^*_{\texttt{PA}}$ is obtained as:

$$g^*_{\texttt{PA}}\left(y_{1:T}\right) = \left(1, e^*\left(y_{1:T}\right)\right) \tag{31}$$

□

B.3 Proof of Theorem 3

Theorem. *Assume the joint vector of $(\rho_{T+1}(y_{1:T}), \ldots, \rho_{T+H}(y_{1:T}))$ is absolutely continuous. Given a forecaster f and a target coverage $c \in (0, 1]$, the optimal selection function $g^*_{\texttt{IA}}$ in interval abstention returns a pair of indices (s^*, e^*) defined as:*

$$(s^*, e^*) = \underset{(s,e)|1\leq s\leq e\leq H\cup(1,0)}{\arg\min}\left(\sum_{t=T+s}^{T+e}\rho_t(y_{1:T}) - \gamma^*\cdot(e - s + 1)\right) \tag{32}$$

where $\gamma^* = \lambda^* + \eta^*$, and $\lambda^* = \mathcal{R}(f, g^*_{\texttt{IA}})$ is the optimal risk. The Lagrange multiplier η^* is chosen such that the pair $\left((e^*, s^*), \eta^*\right)$ satisfies the KKT conditions:

$$\eta^* \geq 0, \qquad \mathbb{E}_{Y_{1:T}}\left[e^* - s^* + 1\right] \geq cH, \qquad \eta^*\left(\mathbb{E}_{Y_{1:T}}\left[e^* - s^* + 1\right] - cH\right) = 0.$$

Proof. We first introduce an alternative formulation for the selection function $g^-_{\texttt{IA}}(y_{1:T}) = (s(y_{1:T}), h(y_{1:T}))$, where $s(y_{1:T})$ is the start step and $h(y_{1:T})$ is the length of the forecast interval. Solving the problem in Equation 3 with respect to $g_{\texttt{IA}}$ is equivalent to solving the same problem with respect to $g^-_{\texttt{IA}}$, since the first can be obtained from the second just by computing the end step as $e(y_{1:T}) = s(y_{1:T}) + h(y_{1:T}) - 1$. Thus, in the following, we will obtain the optimal $g^-_{\texttt{IA}}{}^*$, from which the optimal $g^*_{\texttt{IA}}$ can be directly derived.

We denote the cumulative conditional risk for a forecast interval starting at step s with interval length h as:

$$r_{s:(s+h-1)}(y_{1:T}) = \sum_{t=T+s}^{T+s+h-1}\rho_t(y_{1:T}) = \mathbb{E}_{Y_{T+1:T+H}|Y_{1:T}}\left[\sum_{t=T+s}^{T+s+h-1}\ell\left(y_t, f_t\left(y_{1:T}\right)\right) \mid Y_{1:T} = y_{1:T}\right].$$

Given $g^{-}_{\texttt{IA}}(y_{1:T}) = (s(y_{1:T}), h(y_{1:T}))$, the optimization problem in this setting can be written as:

$$\min_{s,h} \frac{\mathbb{E}_{Y_{1:T}}\left[r_{s:s+h-1}\left(y_{1:T}\right)\right]}{\mathbb{E}_{Y_{1:T}}\left[h\right]} \quad \text{s.t.} \quad \mathbb{E}_{Y_{1:T}}\left[h\right] \geq cH$$

Following the result of Lemma 1, this constrained fractional problem is equivalent to minimizing the linearized Lagrangian objective:

$$\mathcal{L}(s, h, \gamma) = \mathbb{E}_{Y_{1:T}}\left[r_{s:(s+h-1)}(y_{1:T}) - \gamma \cdot h\right] + \eta cH,$$

where $\gamma = \lambda^* + \eta$, λ^* comes from the linearized transformation in Lemma 1 and $\eta \geq 0$ is the Lagrange multiplier associated with the coverage constraint.

Since the expectation is linear and the optimization is made independently for each time series, we can minimize the inner term point-wise. Thus, for a fixed γ, the optimal selection strategy $g^{-}_{\texttt{IA}}{}^*(y_{1:T}) = \left(s^*\left(y_{1:T}\right), h^*\left(y_{1:T}\right)\right)$ is the minimizer of the local cost:

$$\left(s^*\left(y_{1:T}\right), h^*\left(y_{1:T}\right)\right) = \underset{\substack{s\in\{1,\ldots,H\}\\ h\in\{0,\ldots,H-s+1\}}}{\arg\min} \left(\sum_{t=T+s}^{T+s+h-1} \rho_t(y_{1:T}) - \gamma h \right).$$

where $\left(s^*\left(y_{1:T}\right), h^*\left(y_{1:T}\right)\right) = (1, 0)$ means the entire horizon is rejected and the objective value is 0. This optimization can be decomposed into two steps:

1. **Best start for fixed length:** For each time series $y_{1:T}$, we compute a candidate start time $s_h\left(y_{1:T}\right)$ for every possible interval length $h \in \{0, \ldots, H\}$ that minimizes the conditional risk for that specific duration:

$$s^*_h(y_{1:T}) = \underset{s\in\{1,\ldots,H-h+1\}}{\arg\min} \sum_{t=T+s}^{T+s+h-1} \rho_t(y_{1:T})$$

2. **Optimal length:** Select the optimal duration h^* that balances the minimal conditional risk against the length reward γ:

$$h^*\left(y_{1:T}\right) = \underset{h\in\{0,\ldots,H\}}{\arg\min} \left(\sum_{t=T+s_h(y_{1:T})}^{T+s_h(y_{1:T})+h-1} \rho_t(y_{1:T}) - \gamma h \right)$$

As in Theorem 2, under the assumption that the joint vector of conditional risks is absolutely continuous, the expected length $\mathbb{E}[h^*_\gamma(y_{1:T})]$ is a continuous, non-decreasing function of γ. Thus, there exists a γ such that the constraint $\mathbb{E}[h^*(y_{1:T})] = cH$ is satisfied exactly.

Finally, the optimal selection function $g^*_{\texttt{IA}}$ is obtained as:

$$g^*_{\texttt{IA}}\left(y_{1:T}\right) = \left(s^*_{h^*}\left(y_{1:T}\right), s^*_{h^*}\left(y_{1:T}\right) + h^*\left(y_{1:T}\right) - 1\right) \tag{33}$$

□

C Further discussion on continuity assumption

In Theorem 2 and Theorem 3, we assume that the joint distribution of the conditional risk vector is absolutely continuous with respect to the Lebesgue measure. This assumption is required to guarantee the existence of a deterministic γ that satisfies the target coverage exactly. While we acknowledge that this continuity may not hold in practice, it does not limit the practical applicability of our framework. As detailed in Section 3, the optimal strategy can be implemented via a randomized policy to handle discrete jumps in coverage. This guarantees exact constraint satisfaction and demonstrates robust performance in our empirical evaluation (see Section 4.2).

D Algorithmic implementations

D.1 `PAbFor`

Algorithm 1 provides the full details for `PAbFor` outlined in Section 3 (Partial abstention). The algorithm first trains the forecaster and the conditional variance estimator on the training set, then it estimates $\hat{\gamma}_\ell$, $\hat{\gamma}_r$, and p using the calibration set.

Algorithm 1 PAbFor ($\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{calib}}$, c, H, ϵ)

Inputs: $\mathcal{D}_{\text{train}}$ - training data, $\mathcal{D}_{\text{calib}}$ - calibration data, c - target coverage, H - forecast horizon, ϵ - tolerance for γ

Output: $\hat{\gamma}_\ell$, $\hat{\gamma}_r$ - the lower and upper bounds for γ, p - the probability to select the policy

forecaster = Forecaster(); forecaster.fit($\mathcal{D}_{\text{train}}$)
$\hat{f}_{\text{calib}}$, $\hat{\sigma}^2_{\text{calib}}$ = forecaster.predict($\mathcal{D}_{\text{calib}}$)
$\hat{r}$ = $cumSum(\hat{\sigma}^2_{\text{calib}}$, axis=1)
$\hat{r}$ = $hstack([0,...0], \hat{r})$ // to account for the rejection of the full horizon
$\hat{\gamma}_\ell = 0$, $\hat{\gamma}_r = max(\hat{r})$
while $\hat{\gamma}_r - \hat{\gamma}_\ell > \epsilon$ **do**
    $\hat{\gamma} = 0.5\,(\hat{\gamma}_r + \hat{\gamma}_\ell)$
    $\hat{e} = [\text{argmin}_e\,(\hat{r}(y)[e] - \hat{\gamma}e)$ **for** $y \in \mathcal{D}_{\text{calib}}]$
    $\hat{\phi}_{\hat{\gamma}} = mean(\hat{e})$
    if $\hat{\phi}_{\hat{\gamma}} = cH$ **then**
        return $\hat{\gamma}$, $\hat{\gamma}$, 1 // if coverage meets the target level, no randomized policy
    else if $\hat{\phi}_{\hat{\gamma}} < cH$ **then**
        $\hat{\gamma}_\ell = \hat{\gamma}$ // if coverage is below the target level, increase $\hat{\gamma}$
    else if $\hat{\phi}_{\hat{\gamma}} > cH$ **then**
        $\hat{\gamma}_r = \hat{\gamma}$ // if coverage is above the target level, decrease $\hat{\gamma}$
$\hat{e}_{\hat{\gamma}_\ell} = [\text{argmin}_e\,(\hat{r}(y)[e] - \hat{\gamma}_\ell e)$ **for** $y \in \mathcal{D}_{\text{calib}}]$
$\hat{e}_{\hat{\gamma}_r} = [\text{argmin}_e\,(\hat{r}(y)[e] - \hat{\gamma}_r e)$ **for** $y \in \mathcal{D}_{\text{calib}}]$
$\hat{\phi}_{\hat{\gamma}_\ell} = mean(\hat{e}_{\hat{\gamma}_\ell})$, $\hat{\phi}_{\hat{\gamma}_r} = mean(\hat{e}_{\hat{\gamma}_r})$
$p = \frac{cH - \hat{\phi}_{\hat{\gamma}_r}}{\hat{\phi}_{\hat{\gamma}_\ell} - \hat{\phi}_{\hat{\gamma}_r}}$
return $\hat{\gamma}_\ell$, $\hat{\gamma}_r$, p

D.2 IntAbFor

Algorithm 2 provides the full details for IntAbFor outlined in Section 3 (Interval abstention). The algorithm first trains the forecaster and the conditional variance estimator on the training set. Next, for each time-series in the calibration set, it identifies the subsequence with the lowest cumulative variance for every possible interval length. Finally, it estimates $\hat{\gamma}_\ell$, $\hat{\gamma}_r$, and p considering only these optimal subsequences.

E Experimental setup : additional details

E.1 Data

Table 2 summarizes the characteristics of the 24 publicly available datasets used in our evaluation. Prior to training, we apply standard normalization. To strictly prevent data leakage, the scaling parameters are computed solely on the training set. These parameters are then used to normalize the corresponding test time-series. Detailed descriptions of each dataset and specific preprocessing steps are provided in the following paragraphs.

eeg (Kelly et al., 2023) The data is available at https://archive.ics.uci.edu/ml/datasets/EEG+Database. We used the medium version of the dataset, aiming to forecast the responses of ten control subjects to visual stimuli of three types. We excluded alcoholic subjects from our experiments, as prior work and summary statistics indicate that EEG responses from control subjects are more challenging to predict (Stankeviciute et al., 2021). Each subject had repeated trials for each type of stimulus, and each trial produced a time series for 64 EEG channels, corresponding to 64 sensors. We treated each individual trial and each channel as a separate time-series resulting in a total of 38400 time-series. Following Stankeviciute et al. (2021), we downsampled the sequences (originally of length 255) to a length of 50, where we used the first 40 time steps as input and the final 10 as the prediction horizon. We split the data such that trials belonging to the same patient are always in the same set.

Algorithm 2 `IntAbFor` ($\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{calib}}$, c, H, ϵ)

Inputs: $\mathcal{D}_{\text{train}}$ - training data, $\mathcal{D}_{\text{calib}}$ - calibration data, c - target coverage, H - forecast horizon, ϵ - tolerance for γ

Output: $\hat{\gamma}_\ell$, $\hat{\gamma}_r$ - the lower and upper bounds for γ, p - the probability to select the policy

forecaster = Forecaster(); forecaster.fit($\mathcal{D}_{\text{train}}$)
$\hat{f}_{\text{calib}}$, $\hat{\sigma}^2_{\text{calib}}$ = forecaster.predict($\mathcal{D}_{\text{calib}}$)
for h in $\{0..H\}$ **do**
    $\hat{s}[h] = \left[\text{argmin}_{s\in\{1,\dots,H-h+1\}} \left(sum\left(\hat{\sigma}^2[y, s:s+h]\right)\right) \textbf{ for } y \in \mathcal{D}_{\text{calib}}\right]$
// compute the start of the best subsequence for each length (h=0 means full abstention)
$\hat{\gamma}_\ell = 0$, $\hat{\gamma}_r = max(cumSum(\hat{\sigma}^2_{\text{calib}}, axis = 1))$
while $\hat{\gamma}_r - \hat{\gamma}_\ell > \epsilon$ **do**
    $\hat{\gamma} = 0.5\,(\hat{\gamma}_r + \hat{\gamma}_\ell)$
    $\hat{h} = \left[\text{argmin}_h \left(sum\left(\hat{\sigma}^2\left[y, \hat{s}[y,h]:\hat{s}[y,h]+h\right]\right) - \hat{\gamma}h\right) \textbf{ for } y \in \mathcal{D}_{\text{calib}}\right]$
    $\hat{\phi}_{\hat{\gamma}} = mean(\hat{h})$
    if $\hat{\phi}_{\hat{\gamma}} = cH$ **then**
        return $\hat{\gamma}$, $\hat{\gamma}$, 1 // if coverage meets the target level, no randomized policy
    else if $\hat{\phi}_{\hat{\gamma}} < cH$ **then**
        $\hat{\gamma}_\ell = \hat{\gamma}$ // if coverage is below the target level, increase $\hat{\gamma}$
    else if $\hat{\phi}_{\hat{\gamma}} > cH$ **then**
        $\hat{\gamma}_r = \hat{\gamma}$ // if coverage is above the target level, decrease $\hat{\gamma}$
$\hat{h}_{\hat{\gamma}_\ell} = \left[\text{argmin}_h \left(sum\left(\hat{\sigma}^2\left[y, \hat{s}[y,h]:\hat{s}[y,h]+h\right]\right) - \hat{\gamma}_\ell h\right) \textbf{ for } y \in \mathcal{D}_{\text{calib}}\right]$
$\hat{h}_{\hat{\gamma}_r} = \left[\text{argmin}_h \left(sum\left(\hat{\sigma}^2\left[y, \hat{s}[y,h]:\hat{s}[y,h]+h\right]\right) - \hat{\gamma}_r h\right) \textbf{ for } y \in \mathcal{D}_{\text{calib}}\right]$
$\hat{\phi}_{\hat{\gamma}_\ell} = mean(\hat{h}_{\hat{\gamma}_\ell})$, $\hat{\phi}_{\hat{\gamma}_r} = mean(\hat{h}_{\hat{\gamma}_r})$
$p = \frac{cH - \hat{\phi}_{\hat{\gamma}_r}}{\hat{\phi}_{\hat{\gamma}_\ell} - \hat{\phi}_{\hat{\gamma}_r}}$
return $\hat{\gamma}_\ell$, $\hat{\gamma}_r$, p

`covid` (UK Health Security Agency, 2025) The data is available at `https://coronavirus.data.gov.uk/`. We used regional-level data from the United Kingdom, divided into 380 lower-tier local authorities, resulting in 380 time series. We considered daily new COVID-19 cases over a 110-day period, starting in mid-September 2020 and ending in January 2021. Each time series was split into an input sequence of 100 days and a prediction horizon of 10 days.

`temperature` (Mouatadid et al., 2023) The data was retrieved from `https://github.com/microsoft/subseasonal_data/`. This dataset contains weather data for 832 different locations in the United States. Our objective is to forecast the daily temperature computed as the average between the max and the min temperature for the years $2022 - 2023$. Each time series (*i.e.*, location) is split into an input sequence of 698 days and a forecast horizon of 30 days (*i.e.*, the final month).

`ERA5` (Nguyen et al., 2023) The data is available at `https://github.com/aditya-grover/climate-learn`. Since the full ERA5 dataset is too large for most deep learning models, ClimateLearn provides access to a reduced version derived from WeatherBench (Rasp et al., 2020). This version regrids the raw data to a lower spatial resolution of $5.625°$, resulting in a total of 2048 distinct geographic locations (*i.e.*, time-series). Our objective is to forecast the daily maximum temperature for the year 2018. Each time series is split into an input sequence of 358 days and a prediction horizon of 7 days (*i.e.*, the final week).

UCR repository (Dau et al., 2019) The remaining datasets were retrieved from the UCR repository. We removed the original target variable and treat each dataset as a forecasting task, with different forecasting horizons ranging from 6 to 50 steps.

Table 2: Characteristics of the 24 datasets used for the experiments.

dataset	# (training)	# (calibration)	# (test)	T	H
chlorine	2584	861	861	150	16
covid	228	76	76	100	10
ERA5	1228	410	410	358	7
ecg	3000	1000	1000	130	10
eeg	23040	7680	7680	40	10
faceall	1350	450	450	111	20
fetalecg	2259	753	753	700	50
fiftywords	543	181	181	250	20
ford	2952	984	984	450	50
freezer	1800	600	600	271	30
handoutlines	822	274	274	270	30
inlineskate	390	130	130	200	20
ITpower	657	219	219	18	6
melbourne	1365	455	455	17	6
middlephalanx	534	178	178	70	10
mixedshapes	1755	585	585	1000	24
refrigeration	450	150	150	700	20
sonyaiborobot	372	124	124	45	20
swedishleaf	675	225	225	100	28
symbols	612	204	204	350	48
temperature	500	166	166	698	30
wafer	4298	1432	1432	140	12
wordrecognition	345	115	115	134	10
yoga	1980	660	660	400	26

E.2 HARDWARE

We used Python language and an AMD EPYC 9334 32-Core Processor machine with 256 GB RAM and two NVIDIA L40S GPUs to run all our experiments. Our whole evaluation pipeline requires 12 days to complete.

F EXPERIMENTS: ADDITIONAL RESULTS

F.1 Q1: DETAILED RESULTS

For completeness, Figure 5 reports each dataset's average selective risk $\hat{\mathcal{R}}$ (with 95% c.i.) for FAbFor and the considered baselines across six target coverage levels. We complement this graphical representation with Table 3, which provides the detailed numerical values, reporting each dataset's average empirical selective risk ($\pm$ std) for the considered methods. Additionally, Figure 6 shows the Critical Difference (CD) diagrams for the statistical tests conducted in Q1. These plots confirm that FAbFor significantly outperforms the competitors, while the other baselines perform similarly across all considered target coverage levels.

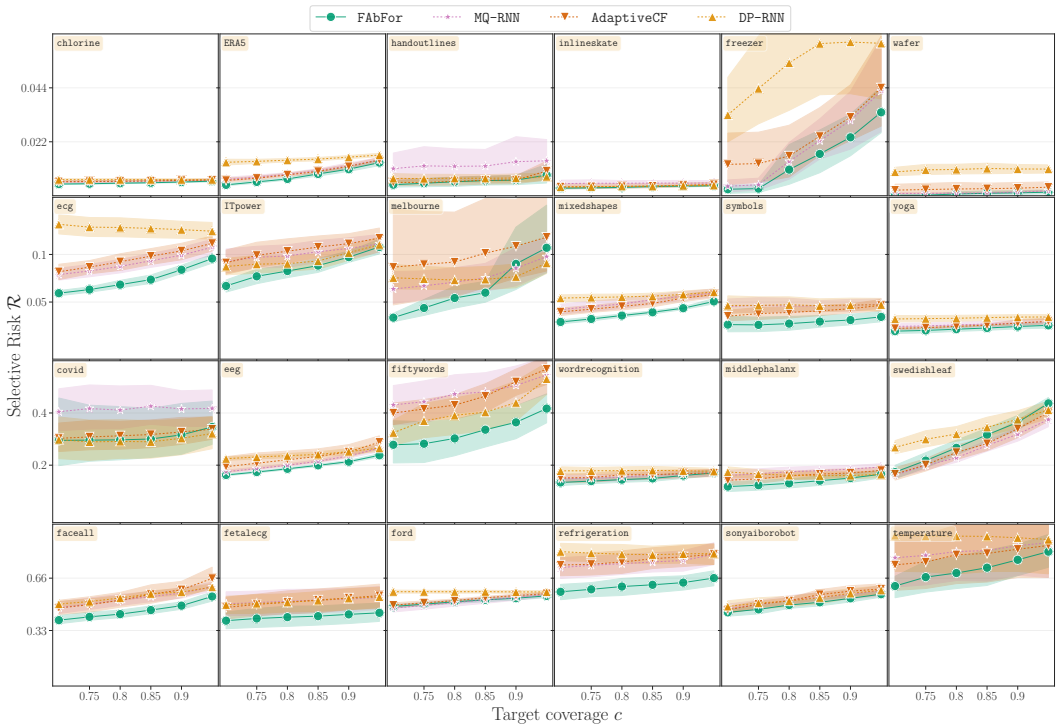

Figure 5: Each dataset's average selective risk (with 95% c.i.) for FAbFor and the competitors.

Table 3: Each dataset's average empirical selective risk (± std) for all methods across six target coverages in the full abstention setting. FAbFor achieves a lower selective risk than the baselines in most of the dataset-coverage pairs.

dataset	c	FAbFor	AdaptiveCF	DP-RNN	MQ-RNN
chlorine	.70	**.0048 ± .0003**	.0057 ± .0006	.0066 ± .0006	.0056 ± .0006
	.75	**.0049 ± .0003**	.0058 ± .0006	.0065 ± .0006	.0057 ± .0006
	.80	**.0051 ± .0003**	.0059 ± .0006	.0066 ± .0006	.0059 ± .0006
	.85	**.0053 ± .0003**	.0060 ± .0006	.0066 ± .0006	.0060 ± .0005
	.90	**.0056 ± .0004**	.0062 ± .0006	.0065 ± .0006	.0062 ± .0005
	.95	**.0059 ± .0004**	.0064 ± .0006	.0065 ± .0006	.0064 ± .0005
ERA5	.70	**.0044 ± .0006**	.0062 ± .0014	.0136 ± .0019	.0067 ± .0014
	.75	**.0056 ± .0009**	.0071 ± .0012	.0140 ± .0014	.0076 ± .0012
	.80	**.0068 ± .0007**	.0085 ± .0011	.0145 ± .0011	.0087 ± .0011
	.85	**.0089 ± .0014**	.0099 ± .0013	.0149 ± .0012	.0103 ± .0012
	.90	**.0108 ± .0015**	.0120 ± .0013	.0157 ± .0010	.0122 ± .0019
	.95	**.0135 ± .0020**	.0145 ± .0018	.0166 ± .0011	.0146 ± .0013
handoutlines	.70	**.0045 ± .0021**	.0055 ± .0023	.0071 ± .0030	.0110 ± .0089
	.75	**.0051 ± .0027**	.0054 ± .0021	.0070 ± .0028	.0122 ± .0116
	.80	**.0057 ± .0030**	.0059 ± .0028	.0071 ± .0026	.0120 ± .0108
	.85	**.0062 ± .0029**	.0063 ± .0027	.0071 ± .0025	.0121 ± .0100
	.90	**.0064 ± .0031**	**.0064 ± .0027**	.0072 ± .0026	.0139 ± .0149
	.95	.0084 ± .0062	.0101 ± .0062	**.0078 ± .0023**	.0142 ± .0136
inlineskate	.70	**.0031 ± .0005**	.0036 ± .0018	.0036 ± .0004	.0049 ± .0019
	.75	**.0032 ± .0006**	.0037 ± .0017	.0036 ± .0004	.0051 ± .0018
	.80	**.0034 ± .0005**	.0038 ± .0016	.0039 ± .0002	.0051 ± .0016
	.85	**.0038 ± .0005**	.0039 ± .0015	**.0038 ± .0002**	.0052 ± .0016
	.90	**.0039 ± .0005**	.0042 ± .0014	.0041 ± .0003	.0051 ± .0015
	.95	**.0041 ± .0004**	.0046 ± .0013	.0042 ± .0004	.0052 ± .0014

dataset	c	FAbFor	AdaptiveCF	DP-RNN	MQ-RNN
freezer	.70	**.0026 ± .0026**	.0128 ± .0183	.0329 ± .0230	.0038 ± .0039
	.75	**.0029 ± .0026**	.0131 ± .0180	.0437 ± .0285	.0045 ± .0040
	.80	**.0106 ± .0139**	.0162 ± .0176	.0542 ± .0367	.0140 ± .0124
	.85	**.0170 ± .0141**	.0242 ± .0155	.0620 ± .0400	.0224 ± .0136
	.90	**.0238 ± .0139**	.0320 ± .0174	.0627 ± .0401	.0307 ± .0188
	.95	**.0340 ± .0154**	.0440 ± .0294	.0622 ± .0410	.0427 ± .0305
wafer	.70	**.0002 ± .0001**	.0023 ± .0035	.0097 ± .0030	.0009 ± .0005
	.75	**.0002 ± .0001**	.0025 ± .0036	.0106 ± .0035	.0010 ± .0004
	.80	**.0007 ± .0014**	.0027 ± .0035	.0107 ± .0032	.0010 ± .0004
	.85	**.0010 ± .0014**	.0029 ± .0034	.0111 ± .0034	.0011 ± .0004
	.90	**.0012 ± .0013**	.0031 ± .0033	.0109 ± .0031	.0015 ± .0006
	.95	**.0014 ± .0013**	.0035 ± .0033	.0109 ± .0027	.0020 ± .0005
ecg	.70	**.0593 ± .0043**	.0819 ± .0130	.1317 ± .0166	.0789 ± .0103
	.75	**.0631 ± .0045**	.0867 ± .0121	.1287 ± .0163	.0827 ± .0107
	.80	**.0683 ± .0065**	.0923 ± .0140	.1283 ± .0166	.0874 ± .0117
	.85	**.0736 ± .0072**	.0982 ± .0131	.1274 ± .0157	.0934 ± .0122
	.90	**.0839 ± .0072**	.1038 ± .0139	.1261 ± .0149	.0992 ± .0131
	.95	**.0957 ± .0072**	.1116 ± .0126	.1245 ± .0146	.1076 ± .0133
ITpower	.70	**.0669 ± .0104**	.0913 ± .0209	.0874 ± .0147	.0927 ± .0192
	.75	**.0770 ± .0126**	.0991 ± .0209	.0898 ± .0152	.0971 ± .0191
	.80	**.0826 ± .0109**	.1034 ± .0202	.0904 ± .0129	.0984 ± .0178
	.85	**.0883 ± .0125**	.1078 ± .0185	.0931 ± .0132	.1028 ± .0190
	.90	**.0972 ± .0097**	.1113 ± .0179	.1017 ± .0136	.1063 ± .0178
	.95	**.1079 ± .0126**	.1172 ± .0158	.1104 ± .0119	.1112 ± .0149
melbourne	.70	**.0336 ± .0074**	.0866 ± .0861	.0755 ± .0214	.0637 ± .0264
	.75	**.0437 ± .0135**	.0895 ± .0821	.0743 ± .0199	.0670 ± .0246
	.80	**.0542 ± .0188**	.0920 ± .0782	.0731 ± .0189	.0709 ± .0236
	.85	**.0598 ± .0173**	.1014 ± .0801	.0739 ± .0199	.0754 ± .0238
	.90	.0901 ± .0599	.1087 ± .0744	**.0766 ± .0203**	.0855 ± .0328
	.95	.1070 ± .0607	.1182 ± .0713	**.0909 ± .0193**	.0977 ± .0300
mixedshapes	.70	**.0291 ± .0033**	.0397 ± .0066	.0542 ± .0058	.0421 ± .0084
	.75	**.0322 ± .0034**	.0425 ± .0058	.0548 ± .0057	.0453 ± .0085
	.80	**.0358 ± .0028**	.0454 ± .0066	.0554 ± .0051	.0478 ± .0084
	.85	**.0393 ± .0027**	.0486 ± .0059	.0562 ± .0056	.0526 ± .0075
	.90	**.0435 ± .0025**	.0541 ± .0063	.0581 ± .0062	.0561 ± .0071
	.95	**.0506 ± .0042**	.0579 ± .0057	.0605 ± .0053	.0595 ± .0068
symbols	.70	**.0263 ± .0133**	.0352 ± .0092	.0464 ± .0157	.0453 ± .0115
	.75	**.0260 ± .0124**	.0378 ± .0104	.0465 ± .0137	.0465 ± .0124
	.80	**.0273 ± .0113**	.0391 ± .0094	.0472 ± .0122	.0457 ± .0115
	.85	**.0295 ± .0098**	.0407 ± .0086	.0461 ± .0109	.0460 ± .0098
	.90	**.0311 ± .0092**	.0430 ± .0081	.0469 ± .0108	.0463 ± .0095
	.95	**.0344 ± .0093**	.0466 ± .0097	.0471 ± .0100	.0478 ± .0082
yoga	.70	**.0194 ± .0039**	.0223 ± .0020	.0322 ± .0055	.0244 ± .0049
	.75	**.0202 ± .0041**	.0231 ± .0022	.0326 ± .0054	.0249 ± .0046
	.80	**.0215 ± .0042**	.0244 ± .0018	.0329 ± .0050	.0260 ± .0044
	.85	**.0227 ± .0042**	.0257 ± .0019	.0337 ± .0045	.0269 ± .0046
	.90	**.0243 ± .0045**	.0280 ± .0048	.0342 ± .0042	.0279 ± .0046
	.95	**.0255 ± .0048**	.0297 ± .0045	.0343 ± .0042	.0293 ± .0042
covid	.70	**.2950 ± .2312**	.3029 ± .0946	.2979 ± .1309	.4046 ± .1386
	.75	.2959 ± .1931	.3081 ± .0888	**.2893 ± .1275**	.4170 ± .1338
	.80	.2970 ± .1755	.3124 ± .0890	**.2910 ± .1225**	.4110 ± .1297
	.85	.3002 ± .1653	.3175 ± .0847	**.2896 ± .1146**	.4260 ± .1218
	.90	.3162 ± .1599	.3271 ± .0817	**.3032 ± .1111**	.4155 ± .1130
	.95	.3464 ± .1417	.3381 ± .0721	**.3212 ± .1031**	.4175 ± .1112

dataset	c	FAbFor	AdaptiveCF	DP-RNN	MQ-RNN
eeg	.70	**.1445 ± .0044**	.1623 ± .0147	.2235 ± .0316	.1669 ± .0143
	.75	**.1554 ± .0049**	.1753 ± .0179	.2303 ± .0344	.1806 ± .0155
	.80	**.1677 ± .0054**	.1929 ± .0231	.2379 ± .0350	.1955 ± .0171
	.85	**.1810 ± .0053**	.2109 ± .0256	.2436 ± .0357	.2126 ± .0217
	.90	**.1974 ± .0061**	.2352 ± .0301	.2538 ± .0369	.2322 ± .0279
	.95	**.2218 ± .0065**	.2694 ± .0382	.2719 ± .0379	.2588 ± .0298
fiftywords	.70	**.2781 ± .1311**	.3999 ± .0639	.3244 ± .0803	.4310 ± .1196
	.75	**.2822 ± .1250**	.4152 ± .0633	.3698 ± .1070	.4430 ± .1339
	.80	**.3025 ± .1158**	.4299 ± .0756	.3908 ± .1085	.4727 ± .1185
	.85	**.3358 ± .1096**	.4643 ± .0898	.4033 ± .1133	.4805 ± .1199
	.90	**.3641 ± .1081**	.5195 ± .0811	.4395 ± .1241	.5056 ± .1231
	.95	**.4168 ± .0917**	.5681 ± .0870	.5307 ± .0972	.5457 ± .1122
wordrecognition	.70	**.1366 ± .0266**	.1572 ± .0245	.1951 ± .0167	.1645 ± .0297
	.75	**.1464 ± .0232**	.1599 ± .0263	.1981 ± .0180	.1709 ± .0252
	.80	**.1524 ± .0235**	.1638 ± .0234	.2023 ± .0225	.1733 ± .0264
	.85	**.1581 ± .0236**	.1690 ± .0222	.1997 ± .0228	.1785 ± .0243
	.90	**.1641 ± .0247**	.1723 ± .0187	.1979 ± .0192	.1850 ± .0198
	.95	**.1700 ± .0228**	.1859 ± .0175	.1960 ± .0206	.1886 ± .0183
middlephalanx	.70	**.1180 ± .0320**	.1425 ± .0330	.1742 ± .0294	.1604 ± .0323
	.75	**.1233 ± .0307**	.1473 ± .0291	.1664 ± .0274	.1672 ± .0305
	.80	**.1304 ± .0286**	.1600 ± .0268	.1610 ± .0252	.1749 ± .0281
	.85	**.1405 ± .0274**	.1658 ± .0286	.1595 ± .0236	.1790 ± .0284
	.90	**.1505 ± .0276**	.1701 ± .0255	.1606 ± .0220	.1847 ± .0259
	.95	.1654 ± .0240	.1799 ± .0207	**.1644 ± .0197**	.1931 ± .0245
swedishleaf	.70	.1735 ± .0432	.1658 ± .0380	.2684 ± .0393	**.1619 ± .0154**
	.75	.2176 ± .0461	.2012 ± .0381	.2992 ± .0484	**.1988 ± .0311**
	.80	.2664 ± .0490	.2488 ± .0441	.3192 ± .0479	**.2281 ± .0285**
	.85	.3164 ± .0440	.2826 ± .0431	.3455 ± .0584	**.2754 ± .0341**
	.90	.3641 ± .0462	.3397 ± .0540	.3752 ± .0554	**.3185 ± .0423**
	.95	.4369 ± .0369	.4014 ± .0560	.4115 ± .0511	**.3750 ± .0431**
faceall	.70	**.3953 ± .0438**	.4719 ± .0732	.4972 ± .0259	.4738 ± .0562
	.75	**.4160 ± .0408**	.4924 ± .0725	.5124 ± .0285	.4974 ± .0562
	.80	**.4333 ± .0395**	.5205 ± .0718	.5349 ± .0250	.5223 ± .0530
	.85	**.4588 ± .0393**	.5580 ± .0905	.5618 ± .0298	.5434 ± .0516
	.90	**.4865 ± .0371**	.5870 ± .0830	.5721 ± .0356	.5658 ± .0542
	.95	**.5446 ± .0466**	.6583 ± .1043	.6039 ± .0299	.6051 ± .0473
fetalecg	.70	**.3919 ± .0863**	.4755 ± .1081	.4922 ± .0938	.4978 ± .1245
	.75	**.4058 ± .0956**	.4945 ± .1115	.5003 ± .0953	.5058 ± .1239
	.80	**.4141 ± .1007**	.5081 ± .1156	.5080 ± .0972	.5150 ± .125
	.85	**.4204 ± .0996**	.5223 ± .1176	.5195 ± .1006	.5218 ± .1235
	.90	**.4317 ± .0983**	.5351 ± .1149	.5329 ± .1052	.5285 ± .1225
	.95	**.4414 ± .0963**	.5498 ± .1193	.5420 ± .1071	.5388 ± .1194
ford	.70	.4816 ± .0162	.4898 ± .0247	.5745 ± .0155	**.4708 ± .0368**
	.75	.4958 ± .0159	.5041 ± .0215	.5746 ± .0140	**.4882 ± .0336**
	.80	.5109 ± .0156	.5182 ± .0179	.5744 ± .0142	**.5058 ± .0320**
	.85	.5229 ± .0175	.5364 ± .0151	.5749 ± .0140	**.5225 ± .0297**
	.90	**.5338 ± .0168**	.5504 ± .0143	.5731 ± .0147	.5388 ± .0257
	.95	**.5482 ± .0186**	.5624 ± .0154	.5733 ± .0136	.5563 ± .0205
refrigeration	.70	**.5739 ± .0763**	.7427 ± .1127	.8270 ± .0889	.7268 ± .1109
	.75	**.5896 ± .0732**	.7472 ± .1151	.8178 ± .0896	.7411 ± .1060
	.80	**.6071 ± .0840**	.7615 ± .1128	.8113 ± .0907	.7512 ± .0989
	.85	**.6181 ± .0829**	.7826 ± .1052	.8076 ± .0898	.7620 ± .0972
	.90	**.6319 ± .0746**	.7946 ± .1104	.8131 ± .0943	.7715 ± .1026
	.95	**.6607 ± .0778**	.8104 ± .1071	.8151 ± .1031	.8099 ± .1133

dataset	c	FAbFor	AdaptiveCF	DP-RNN	MQ-RNN
sonyaiborobot	.70	**.4444 ± .0358**	.4599 ± .0365	.4822 ± .0609	.4699 ± .0433
	.75	**.4633 ± .0405**	.4958 ± .0305	.5030 ± .0586	.4933 ± .0462
	.80	**.4905 ± .0381**	.5168 ± .0317	.5160 ± .0526	.5208 ± .0453
	.85	**.5078 ± .0345**	.5567 ± .0496	.5368 ± .0551	.5395 ± .0463
	.90	**.5329 ± .0368**	.5779 ± .0551	.5645 ± .0503	.5575 ± .0457
	.95	**.5590 ± .0331**	.5918 ± .0505	.5839 ± .0504	.5815 ± .0524
temperature	.70	**.6102 ± .1080**	.7435 ± .2824	.9294 ± .4112	.7884 ± .3538
	.75	**.6666 ± .1562**	.7617 ± .2836	.9249 ± .4060	.8035 ± .3532
	.80	**.6923 ± .1644**	.8074 ± .2969	.9249 ± .4017	.8273 ± .3476
	.85	**.7259 ± .1787**	.8160 ± .2989	.9233 ± .3949	.8340 ± .3524
	.90	**.7757 ± .1708**	.8423 ± .3285	.9145 ± .3885	.8644 ± .3659
	.95	**.8270 ± .1648**	.8658 ± .3315	.9037 ± .3784	.8837 ± .3703

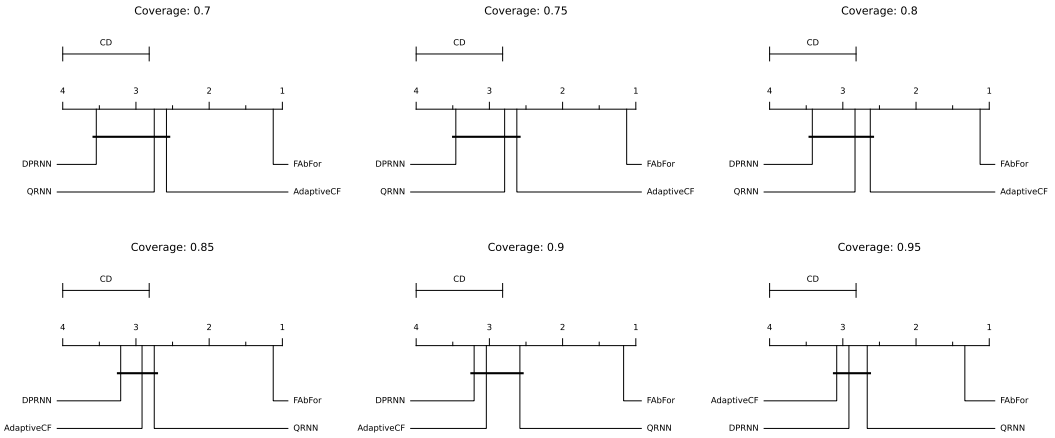

Figure 6: CD plots of empirical selective risk for different target coverages. FAbFor significantly outperforms the baselines for all considered target coverages.

F.2 Q2 : DETAILED RESULTS

For completeness, Figure 7 reports each dataset's average selective risk $\hat{\mathcal{R}}$ (with 95% c.i.) for our three proposed strategies and Accept-cH across six target coverage levels. We complement this graphical representation with Table 4, which provides the detailed numerical values, reporting each dataset's average empirical selective risk ($\pm$ std) for the considered methods.

Table 4: Each dataset's average empirical selective risk ($\pm$ std) for the three proposed strategies and Accept-cH across six target coverages. IntAbFor achieves often the lowest risk, followed by, in order, PAbFor and FAbFor. All strategies outperform Accept-cH.

dataset	c	FAbFor	PAbFor	IntAbFor	Accept-cH
chlorine	.70	.0048 ± .0003	.0038 ± .0002	**.0034 ± .0003**	.0105 ± .0034
	.75	.0049 ± .0003	.0040 ± .0002	**.0036 ± .0003**	.0103 ± .0032
	.80	.0051 ± .0003	.0042 ± .0003	**.0039 ± .0003**	.0101 ± .0030
	.85	.0053 ± .0003	.0045 ± .0002	**.0042 ± .0003**	.0100 ± .0028
	.90	.0056 ± .0004	.0049 ± .0003	**.0046 ± .0004**	.0100 ± .0026
	.95	.0059 ± .0004	.0056 ± .0003	**.0052 ± .0004**	.0101 ± .0025

dataset	c	FAbFor	PAbFor	IntAbFor	Accept-cH
ERA5	.70	.0044 ± .0006	**.0043 ± .0005**	**.0043 ± .0006**	.0160 ± .0015
	.75	.0056 ± .0009	**.0055 ± .0007**	**.0055 ± .0007**	.0163 ± .0016
	.80	.0068 ± .0007	**.0067 ± .0008**	.0069 ± .0008	.0164 ± .0015
	.85	.0089 ± .0014	**.0085 ± .0011**	.0086 ± .0011	.0166 ± .0015
	.90	.0108 ± .0015	**.0106 ± .0014**	.0107 ± .0014	.0168 ± .0015
	.95	.0135 ± .0020	.0134 ± .0019	**.0133 ± .0019**	.0169 ± .0016
handoutlines	.70	.0045 ± .0021	**.0041 ± .0020**	.0044 ± .0023	.0140 ± .0075
	.75	.0051 ± .0027	**.0046 ± .0023**	.0050 ± .0026	.0154 ± .0083
	.80	.0057 ± .0030	**.0053 ± .0025**	**.0053 ± .0027**	.0172 ± .0095
	.85	.0062 ± .0029	**.0057 ± .0025**	.0058 ± .0031	.0187 ± .0104
	.90	.0064 ± .0031	**.0063 ± .0030**	.0064 ± .0035	.0207 ± .0117
	.95	.0084 ± .0062	.0086 ± .0057	**.0080 ± .0056**	.0227 ± .0129
inlineskate	.70	.0031 ± .0005	**.0028 ± .0004**	**.0028 ± .0004**	.0037 ± .0005
	.75	.0032 ± .0006	**.0029 ± .0004**	**.0029 ± .0004**	.0039 ± .0006
	.80	.0034 ± .0005	**.0031 ± .0004**	**.0031 ± .0003**	.0041 ± .0006
	.85	.0038 ± .0005	**.0033 ± .0003**	**.0033 ± .0003**	.0044 ± .0007
	.90	.0039 ± .0005	**.0036 ± .0003**	**.0036 ± .0003**	.0047 ± .0007
	.95	**.0041 ± .0004**	.0043 ± .0008	.0043 ± .0008	.0049 ± .0007
freezer	.70	.0026 ± .0026	**.0023 ± .0024**	.0057 ± .0053	.0592 ± .0225
	.75	.0029 ± .0026	**.0022 ± .0022**	.0055 ± .0045	.0599 ± .0224
	.80	.0106 ± .0139	**.0023 ± .0021**	.0062 ± .0051	.0610 ± .0211
	.85	.0170 ± .0141	.0105 ± .0133	**.0065 ± .0049**	.0618 ± .0203
	.90	.0238 ± .0139	.0150 ± .0131	**.0113 ± .0131**	.0624 ± .0202
	.95	.0340 ± .0154	**.0271 ± .0160**	.0282 ± .0183	.0630 ± .0198
wafer	.70	**.0002 ± .0001**	.0003 ± .0003	.0003 ± .0005	.0133 ± .0045
	.75	**.0002 ± .0001**	.0004 ± .0007	.0005 ± .0011	.0133 ± .0045
	.80	.0007 ± .0014	**.0004 ± .0007**	.0005 ± .0010	.0133 ± .0047
	.85	.0010 ± .0014	**.0005 ± .0009**	.0006 ± .0012	.0132 ± .0045
	.90	.0012 ± .0013	.0010 ± .0012	**.0009 ± .0013**	.0132 ± .0047
	.95	.0014 ± .0013	**.0013 ± .0013**	.0015 ± .0019	.0133 ± .0048
ecg	.70	.0593 ± .0043	.0423 ± .0043	**.0419 ± .0044**	.0599 ± .0051
	.75	.0631 ± .0045	.0475 ± .0036	**.0467 ± .0044**	.0650 ± .0055
	.80	.0683 ± .0065	.0534 ± .0028	**.0527 ± .0038**	.0703 ± .0061
	.85	.0736 ± .0072	.0607 ± .0030	**.0603 ± .0039**	.0772 ± .0069
	.90	.0839 ± .0072	.0703 ± .0029	**.0699 ± .0050**	.0831 ± .0072
	.95	.0957 ± .0072	.0843 ± .0058	**.0842 ± .0061**	.0978 ± .0080
ITpower	.70	.0669 ± .0104	.0689 ± .0081	**.0639 ± .0065**	.1372 ± .0172
	.75	.0770 ± .0126	.0772 ± .0098	**.0680 ± .0087**	.1340 ± .0166
	.80	.0826 ± .0109	.0807 ± .0095	**.0727 ± .0105**	.1305 ± .0163
	.85	.0883 ± .0125	.0872 ± .0110	**.0789 ± .0104**	.1284 ± .0155
	.90	.0972 ± .0097	.0956 ± .0108	**.0872 ± .0129**	.1253 ± .0155
	.95	.1079 ± .0126	.1078 ± .0104	**.0999 ± .0120**	.1233 ± .0146
melbourne	.70	.0336 ± .0074	.0307 ± .0061	**.0305 ± .0058**	.1171 ± .0414
	.75	.0437 ± .0135	**.0341 ± .0061**	.0363 ± .0091	.1313 ± .0552
	.80	.0542 ± .0188	**.0439 ± .0090**	.0440 ± .0092	.1359 ± .0505
	.85	.0598 ± .0173	**.0564 ± .0160**	**.0564 ± .0160**	.1436 ± .0542
	.90	.0901 ± .0599	**.0675 ± .0154**	.0677 ± .0152	.1490 ± .0541
	.95	.1070 ± .0607	**.1010 ± .0576**	**.1010 ± .0576**	.1601 ± .0591
mixedshapes	.70	.0291 ± .0033	**.0198 ± .0032**	**.0198 ± .0026**	.0423 ± .0043
	.75	.0322 ± .0034	**.0233 ± .0031**	**.0233 ± .0024**	.0464 ± .0047
	.80	.0358 ± .0028	**.0275 ± .0028**	.0276 ± .0023	.0505 ± .0051
	.85	.0393 ± .0027	**.0326 ± .0026**	.0328 ± .0023	.0543 ± .0056
	.90	.0435 ± .0025	**.0388 ± .0023**	.0389 ± .0028	.0586 ± .0063
	.95	.0506 ± .0042	**.0467 ± .0031**	.0474 ± .0041	.0622 ± .0063

dataset	c	FAbFor	PAbFor	IntAbFor	Accept-cH
symbols	.70	.0263 ± .0133	**.0233 ± .0105**	**.0233 ± .0105**	.0356 ± .0085
	.75	.0260 ± .0124	.0245 ± .011	**.0244 ± .0110**	.0371 ± .0088
	.80	.0273 ± .0113	**.0263 ± .0108**	**.0263 ± .0108**	.0384 ± .0091
	.85	.0295 ± .0098	**.0284 ± .0101**	**.0284 ± .0100**	.0394 ± .0092
	.90	.0311 ± .0092	**.0308 ± .0096**	.0310 ± .0095	.0406 ± .0094
	.95	**.0344 ± .0093**	.0359 ± .0095	.0359 ± .0095	.0416 ± .0096
yoga	.70	.0194 ± .0039	**.0144 ± .0034**	.0145 ± .0034	.0205 ± .0039
	.75	.0202 ± .0041	**.0155 ± .0035**	.0156 ± .0035	.0221 ± .0040
	.80	.0215 ± .0042	**.0169 ± .0035**	.0170 ± .0035	.0239 ± .0043
	.85	.0227 ± .0042	**.0187 ± .0036**	.0188 ± .0036	.0257 ± .0046
	.90	.0243 ± .0045	**.0210 ± .0040**	.0211 ± .0041	.0276 ± .0050
	.95	.0255 ± .0048	**.0240 ± .0046**	.0242 ± .0047	.0300 ± .0005
covid	.70	.2950 ± .2312	.2927 ± .2304	**.1974 ± .1894**	.3663 ± .1642
	.75	.2959 ± .1931	.2900 ± .1879	**.2075 ± .1765**	.3761 ± .1600
	.80	.2970 ± .1755	.2929 ± .1823	**.2199 ± .1670**	.3946 ± .1493
	.85	.3002 ± .1653	.3058 ± .1651	**.2405 ± .1612**	.4131 ± .1391
	.90	.3162 ± .1599	.3176 ± .1586	**.2896 ± .1397**	.4594 ± .1500
	.95	.3464 ± .1417	**.3281 ± .1502**	.3462 ± .1207	.4536 ± .1658
eeg	.70	.1445 ± .0044	**.1444 ± .0035**	**.1444 ± .0039**	.2903 ± .0082
	.75	.1554 ± .0049	.1556 ± .0040	**.1551 ± .0044**	.2922 ± .0082
	.80	.1677 ± .0054	.1673 ± .0044	**.1672 ± .0051**	.2940 ± .0081
	.85	.1810 ± .0053	**.1807 ± .0046**	.1808 ± .0055	.2937 ± .0083
	.90	.1974 ± .0061	.1976 ± .0048	**.1964 ± .0067**	.2929 ± .0084
	.95	.2218 ± .0065	.2207 ± .0054	**.2189 ± .0077**	.2918 ± .0089
fiftywords	.70	.2781 ± .1311	**.1865 ± .0623**	.2039 ± .0720	.3784 ± .0799
	.75	.2822 ± .1250	**.2153 ± .0724**	.2339 ± .0757	.4018 ± .0859
	.80	.3025 ± .1158	**.2572 ± .0913**	.2808 ± .0987	.4245 ± .0940
	.85	.3358 ± .1096	**.2980 ± .1048**	.3174 ± .1100	.4460 ± .1011
	.90	.3641 ± .1081	**.3388 ± .1011**	.3606 ± .1227	.4666 ± .1041
	.95	.4168 ± .0917	**.4098 ± .0968**	.4140 ± .1227	.4852 ± .1051
wordrecognition	.70	.1366 ± .0266	**.1354 ± .0248**	.1448 ± .0233	.1616 ± .0159
	.75	.1464 ± .0232	**.1402 ± .0248**	.1505 ± .0230	.1662 ± .0175
	.80	.1524 ± .0235	**.1517 ± .0229**	.1564 ± .0233	.1698 ± .0171
	.85	.1581 ± .0236	**.1572 ± .0220**	.1609 ± .0244	.1734 ± .0189
	.90	**.1641 ± .0247**	.1654 ± .0237	.1658 ± .0241	.1782 ± .0193
	.95	**.1700 ± .0228**	.1711 ± .0227	.1715 ± .0232	.1819 ± .0204
middlephalanx	.70	.1180 ± .0320	.1112 ± .0232	**.1107 ± .0211**	.1950 ± .0256
	.75	.1233 ± .0307	**.1205 ± .0205**	.1208 ± .0155	.2013 ± .0266
	.80	.1304 ± .0286	**.1273 ± .0195**	.1292 ± .0170	.2024 ± .0261
	.85	.1405 ± .0274	**.1354 ± .0206**	.1421 ± .0254	.1999 ± .0256
	.90	.1505 ± .0276	**.1467 ± .0258**	.1481 ± .0260	.1969 ± .0255
	.95	.1654 ± .0240	**.1580 ± .0252**	.1601 ± .0261	.1918 ± .0257
swedishleaf	.70	.1735 ± .0432	.1694 ± .0393	**.1465 ± .0232**	.4847 ± .0597
	.75	.2176 ± .0461	.2045 ± .0407	**.1815 ± .0289**	.4946 ± .0588
	.80	.2664 ± .0490	.2425 ± .0507	**.2228 ± .0249**	.5040 ± .0588
	.85	.3164 ± .0440	.2972 ± .0433	**.2602 ± .0267**	.5113 ± .0566
	.90	.3641 ± .0462	.3571 ± .0461	**.3277 ± .0315**	.5126 ± .0543
	.95	.4369 ± .0369	.4346 ± .0417	**.4097 ± .0453**	.5152 ± .0528
faceall	.70	.3953 ± .0438	.3868 ± .0438	**.3855 ± .0416**	.7552 ± .1072
	.75	.4160 ± .0408	**.4109 ± .0389**	.4137 ± .0429	.7567 ± .1033
	.80	.4333 ± .0395	**.4307 ± .0477**	.4372 ± .0490	.7562 ± .0998
	.85	.4588 ± .0393	**.4529 ± .0544**	.4633 ± .0541	.7551 ± .0996
	.90	**.4865 ± .0371**	.4885 ± .0497	.4972 ± .0495	.7518 ± .0973
	.95	**.5446 ± .0466**	.5552 ± .0601	.5689 ± .0650	.7504 ± .0927

dataset	c	FAbFor	PAbFor	IntAbFor	Accept-cH
fetalecg	.70	.3919 ± .0863	**.3261 ± .0576**	.3262 ± .0575	.4072 ± .0737
	.75	.4058 ± .0956	**.3446 ± .0624**	.3447 ± .0619	.4222 ± .0799
	.80	.4141 ± .1007	.3665 ± .0681	**.3643 ± .0669**	.4359 ± .0861
	.85	.4204 ± .0996	.3879 ± .0776	**.3870 ± .0773**	.4468 ± .0918
	.90	.4317 ± .0983	.4116 ± .0822	**.4103 ± .0831**	.4571 ± .0956
	.95	.4414 ± .0963	.4415 ± .0973	**.4357 ± .0926**	.4687 ± .0981
ford	.70	.4816 ± .0162	.4199 ± .0162	**.4180 ± .0114**	.4639 ± .0093
	.75	.4958 ± .0159	.4490 ± .0138	**.4463 ± .0116**	.4814 ± .0098
	.80	.5109 ± .0156	.4731 ± .0126	**.4718 ± .0114**	.4989 ± .0096
	.85	.5229 ± .0175	.4993 ± .0143	**.4976 ± .0120**	.5164 ± .0091
	.90	.5338 ± .0168	.5216 ± .0161	**.5190 ± .0122**	.5338 ± .0104
	.95	.5482 ± .0186	.5423 ± .0169	**.5400 ± .0130**	.5496 ± .0125
refrigeration	.70	.5739 ± .0763	.5203 ± .0779	**.5138 ± .0785**	.6479 ± .0603
	.75	.5896 ± .0732	.5462 ± .0845	**.5431 ± .0759**	.6560 ± .0590
	.80	.6071 ± .0840	.5767 ± .0778	**.5694 ± .0766**	.6665 ± .0579
	.85	.6181 ± .0829	.6087 ± .0766	**.6083 ± .0713**	.6762 ± .0556
	.90	**.6319 ± .0746**	.6335 ± .0681	.6345 ± .0642	.6874 ± .0544
	.95	**.6607 ± .0778**	.6669 ± .0655	.6711 ± .0644	.6997 ± .0570
sonyaiborobot	.70	.4444 ± .0358	.3373 ± .0404	**.2794 ± .0381**	.5744 ± .0463
	.75	.4633 ± .0405	.3675 ± .0432	**.3344 ± .0371**	.5674 ± .0445
	.80	.4905 ± .0381	.4141 ± .0324	**.3847 ± .0457**	.5607 ± .0420
	.85	.5078 ± .0345	.4470 ± .0316	**.4305 ± .0435**	.5707 ± .0390
	.90	.5329 ± .0368	.4839 ± .0317	**.4736 ± .0343**	.5748 ± .0366
	.95	.5590 ± .0331	**.5248 ± .0334**	.5277 ± .0325	.5737 ± .0344
temperature	.70	.6102 ± .1080	.4934 ± .0958	**.4919 ± .0950**	.6720 ± .1296
	.75	.6666 ± .1562	.5364 ± .0920	**.5363 ± .0922**	.7062 ± .1364
	.80	.6923 ± .1644	**.5725 ± .1011**	.5728 ± .1011	.7421 ± .1454
	.85	.7259 ± .1787	**.6142 ± .1010**	.6146 ± .1001	.7776 ± .1535
	.90	.7757 ± .1708	.6775 ± .1253	**.6771 ± .1239**	.8048 ± .1572
	.95	.8270 ± .1648	.7523 ± .1433	**.7509 ± .1452**	.8394 ± .1650

F.3 PAbFor AGAINST THE COMPETITORS IN PARTIAL ABSTENTION

In this section, we evaluate the performance of PAbFor against the Q1 baselines (AdaptiveCF, MQ-RNN, and DP-RNN) under the *partial abstention* setting. All baselines employ the exact selection function derived in Theorem 2; thus, the models differ solely in how they estimate the conditional risk $\hat{r}(x)$.

Figure 8 shows the average empirical selective risk ($\hat{\mathcal{R}}$) across six target coverage levels for PAbFor and the baselines. Similar to the full abstention setting, PAbFor outperforms the competitors in 18 datasets out of 24 and reduces the selective risk on average by **27%** across all evaluated target coverage levels.

F.4 IntAbFor AGAINST THE COMPETITORS IN INTERVAL ABSTENTION

In this section, we evaluate the performance of IntAbFor against the Q1 baselines (AdaptiveCF, MQ-RNN, and DP-RNN) under the *interval abstention* setting. All baselines employ the exact selection function derived in Theorem 3; thus, the models differ solely in how they estimate the conditional risk $\hat{r}(x)$.

Figure 9 shows the average empirical selective risk ($\hat{\mathcal{R}}$) across six target coverage levels for IntAbFor and the baselines. Similar to the full and partial abstention setting, IntAbFor outperforms the competitors in 17 datasets out of 24 and reduces the selective risk on average by **25%** across all evaluated target coverage levels.

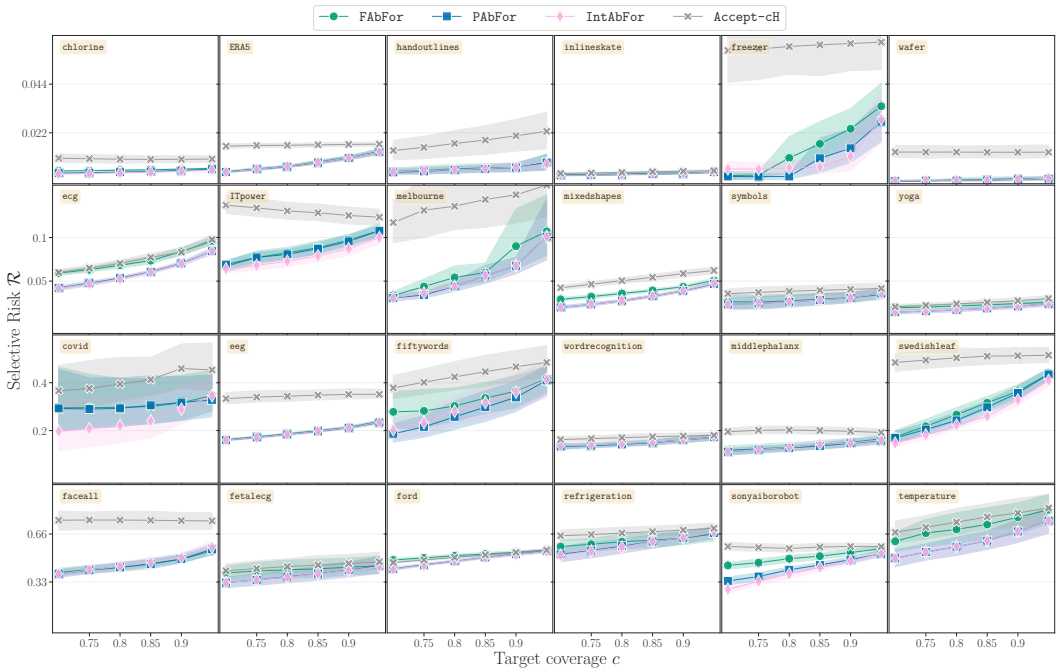

Figure 7: Each dataset's average selective risk (with 95% c.i.) per step for `FAbFor`, `PAbFor`, `IntAbFor`, and `Accept-cH` across six target coverages.

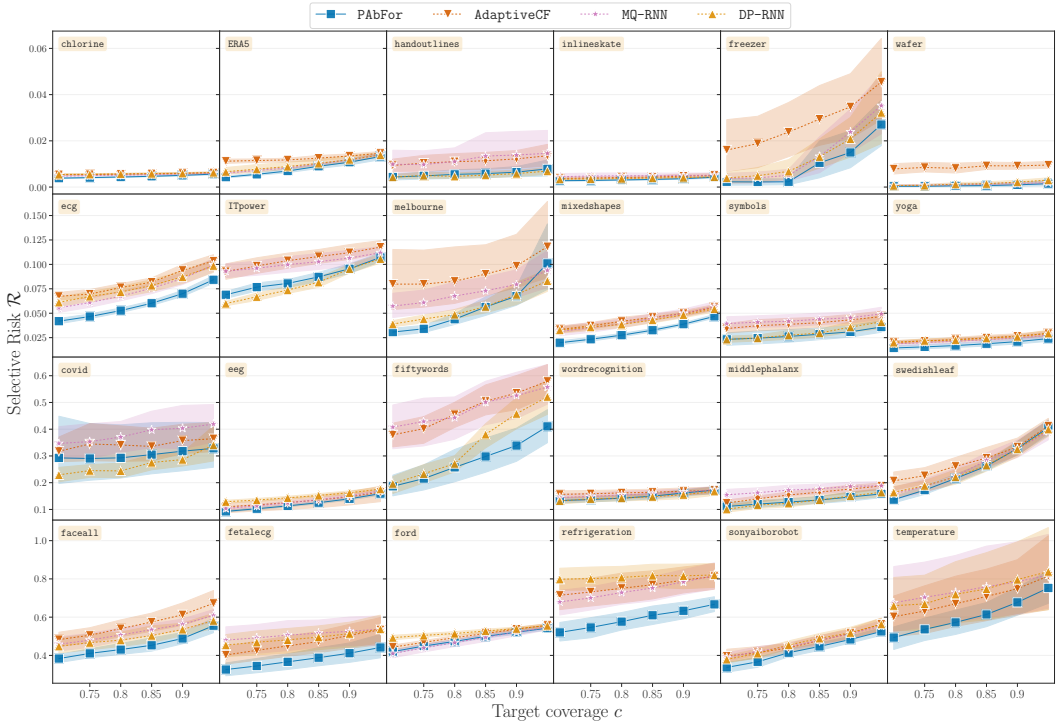

Figure 8: Each dataset's average selective risk (with 95% c.i.) for `PAbFor` and the competitors.

F.5 Ranking quality of the conditional variance estimates

Bounded-abstention strategies depend on the ranking induced by the uncertainty estimates rather than on their absolute calibration (Rabanser & Papernot, 2025). We therefore evaluate our esti-

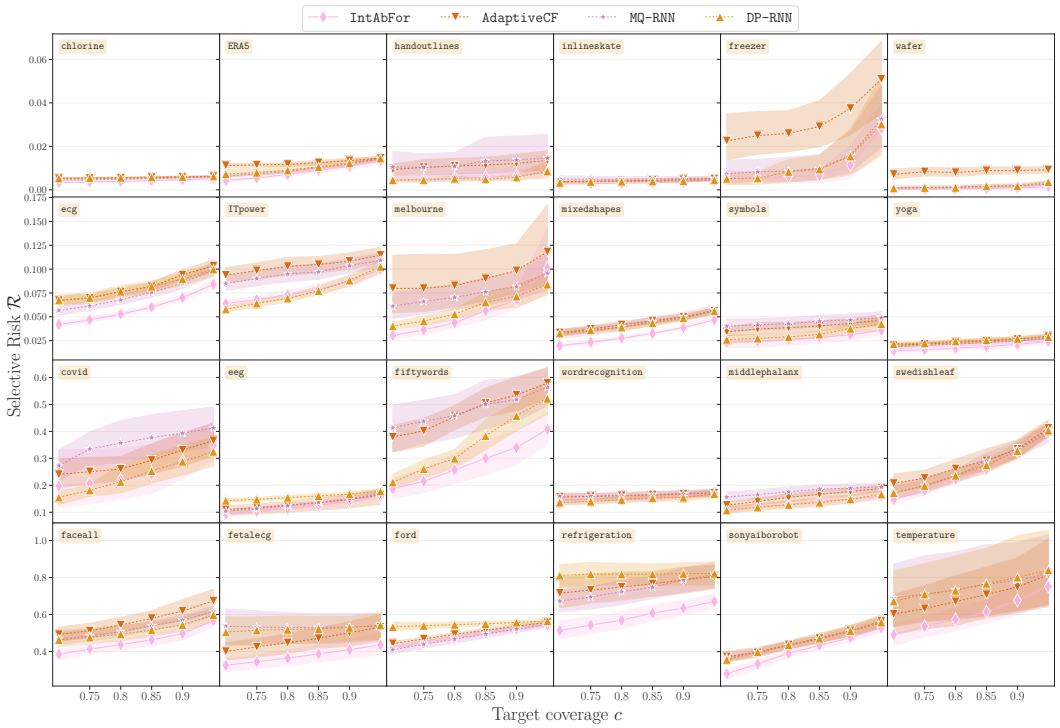

Figure 9: Each dataset's average selective risk (with 95% c.i.) for `IntAbFor` and the competitors.

Table 5: Each dataset's Kendall's τ rank correlation ($\pm$ std) between the estimated conditional variances and the realized squared errors.

dataset	Kendall's τ	dataset	Kendall's τ
ERA5	0.583 ± 0.038	melbourne	0.522 ± 0.030
chlorine	0.323 ± 0.030	middlephalanx	0.486 ± 0.049
covid	0.453 ± 0.066	mixedshapes	0.394 ± 0.031
ecg	0.319 ± 0.022	fetalecg	0.270 ± 0.081
eeg	0.409 ± 0.009	refrigeration	0.377 ± 0.074
faceall	0.450 ± 0.030	sonyaiborobot	0.479 ± 0.040
fiftywords	0.318 ± 0.063	swedishleaf	0.611 ± 0.018
ford	0.312 ± 0.021	symbols	0.290 ± 0.050
freezer	0.559 ± 0.104	temperature	0.304 ± 0.072
handoutlines	0.333 ± 0.042	wafer	0.347 ± 0.065
inlineskate	0.302 ± 0.070	wordrecognition	0.258 ± 0.098
ITpower	0.432 ± 0.032	yoga	0.315 ± 0.036

mates using the Kendall τ rank correlation coefficient (Kendall et al., 1938) between the estimated conditional variances $\hat{\sigma}_t^2$ and the realized squared errors on the test set.

Table 5 reports the average Kendall's τ ($\pm$ std) per dataset. Our estimates attain an average τ of 0.38, remaining consistently above 0.3 for almost all datasets. This confirms that the variance head learns a reliable ranking signal of predictive uncertainty (Cohen, 1992).

F.6 Ablation study: β hyperparameter

To assess the impact of β, we train our three proposed strategies (`FAbFor`, `PAbFor`, and `IntAbFor`) using varying values for this hyperparameter. Specifically, we test five different values of β: 0 (*i.e.*, the stop-gradient operation is not used), 0.1, 0.25, 0.5 (default value), and 1. To reduce

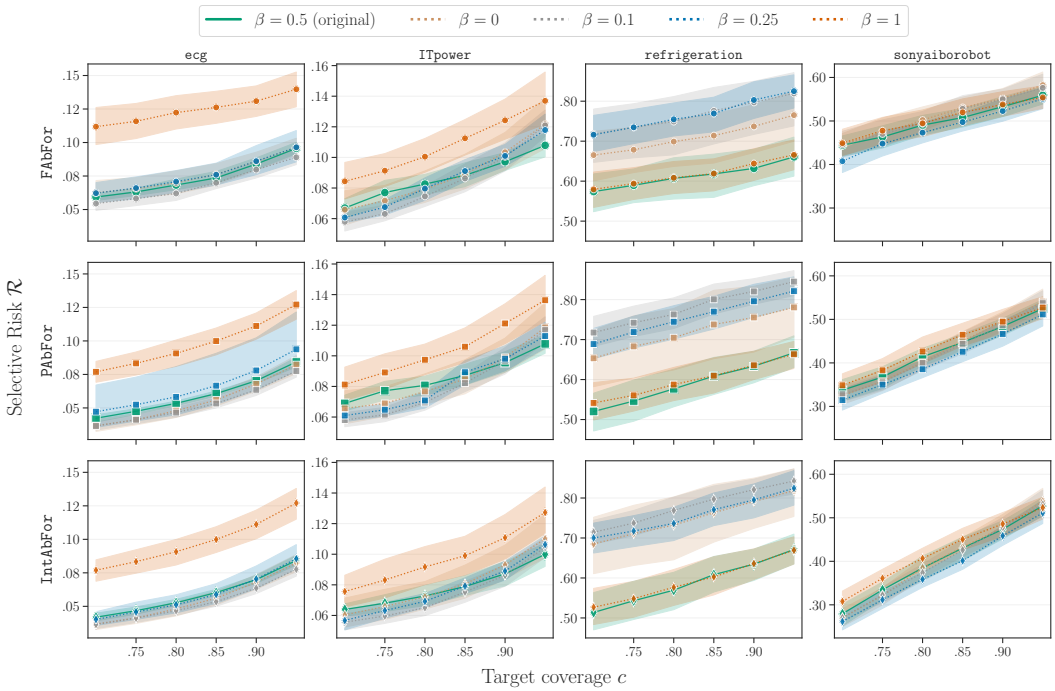

Figure 10: Average selective risk (with 95% c.i.s) for the three proposed strategies using five different values of β across six target coverages on four representative datasets. The model with $\beta = 1$ tends to underperform, while the models using intermediate values of β perform similarly.

computational complexity, we run this ablation only on the four representative datasets included in the main paper.

Fig. 10 reports the average empirical selective risk across the six target coverages for these configurations. The model with $\beta = 1$ underperforms on `ecg` and `ITpower`, while the models with $\beta = 0$, $\beta = 0.1$, and $\beta = 0.25$ underperform on `refrigeration`. The chosen default value of $\beta = 0.5$ offers consistently good performance across all datasets, though the overall sensitivity to β appears limited.

F.7 Architecture ablation: Selective Forecasters with a Transformer-based backbone

In this section, we repeat the experiments in Q1 and Q2 but employ a Transformer-based sequence encoder as the backbone of the selective forecasters. Specifically, the raw time-series are first projected into a higher-dimensional embedding space via a linear layer. Then, we add a standard sinusoidal positional encoding to inject temporal order information. This sequence of embeddings is then processed by a Transformer encoder consisting of three identical layers, each equipped with four attention heads. Finally, we extract the encoder's hidden state corresponding to the final input time step and fed this latent representation into two output heads that jointly predict the forecast values and their corresponding conditional variances.

`FAbFor` **significantly outperforms its competitors when a transformer-based backbone is used.** Figure 11 shows each dataset's average empirical selective risk $\hat{\mathcal{R}}$ across six target coverage levels for all considered methods in the full abstention setting when a transformer-based backbone is used. Overall, `FAbFor` reduces the selective risk by an average of 17% vs `MQ-RNN`, 20% vs `AdaptiveCF`, and 28% vs `DP-RNN`. Moreover, `FAbFor` achieves lower selective risk in around 87% of the experiments against the best runner-up `MQ-RNN`.

To evaluate statistical significance, we apply the method proposed in Demšar (2006) separately for each target coverage and use the Bonferroni correction to account for multiple hypothesis tests. For each target coverage, the Friedman test rejects the null-hypothesis that all methods perform similarly.

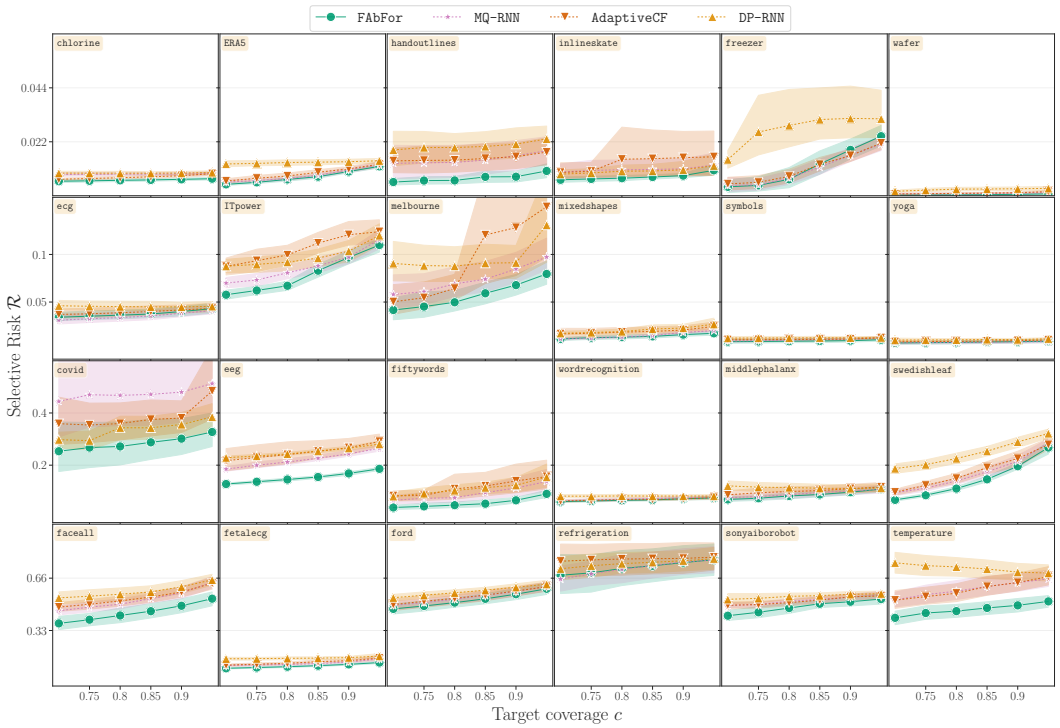

Figure 11: Each dataset's average selective risk (with 95% c.i.) for all methods across six target coverages in the full abstention setting using a transformer-based backbone. For all the coverage level considered, FAbFor achieves a lower selective risk than the baselines in most of the datasets.

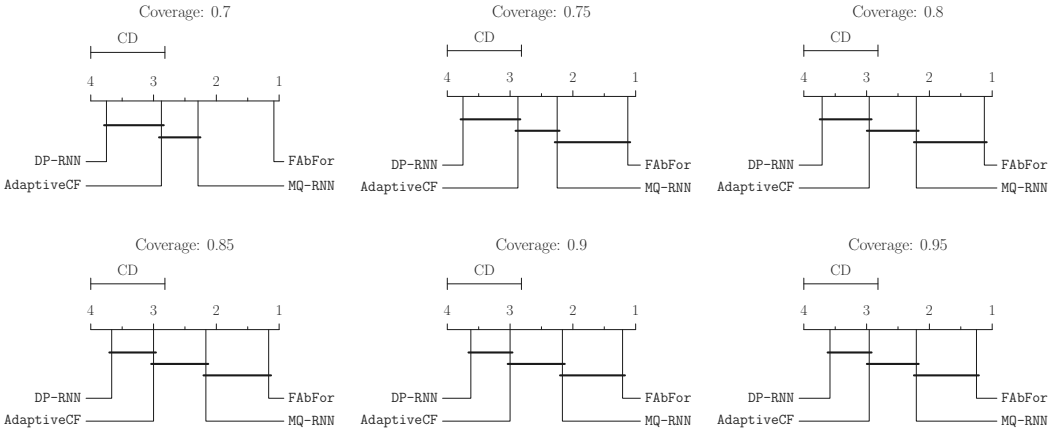

Figure 12: CD plots of empirical selective risk for different target coverages when a transformer-based backbone is used. FAbFor significantly outperforms the baselines for all considered target coverages.

According to the Nemenyi post-hoc test (Demšar, 2006), FAbFor significantly outperforms the baselines for each considered target coverage; Figure 12 shows all critical difference plots.

Finer-grained control in the abstention strategy yields lower selective risk when a transformer-based backbone is used. We evaluate our three proposed strategies against the naive baseline Accept-cH using the Transformer backbone. Figure 13 shows the average selective risk across six target coverage levels for all datasets. Consistent with the results in Q2, our approaches outperform Accept-cH and maintain the exact same performance hierarchy. Specifically, ranking the methods from best (rank 1) to worst (rank 4) for each experiment, IntAbFor achieves the best average rank

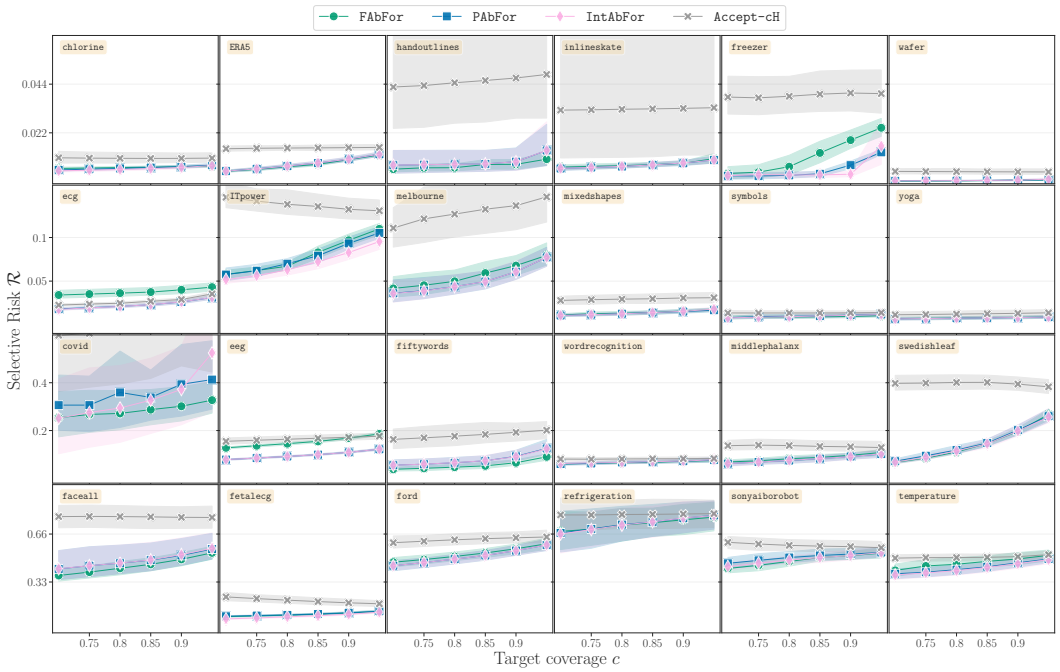

Figure 13: Each dataset's average selective risk (with 95% c.i.) for `FAbFor`, `PAbFor`, `IntAbFor`, and `Accept-cH` across six target coverages using a Transformer backbone. `IntAbFor` achieves often the lowest risk, followed by, in order, `PAbFor` and `FAbFor`. All strategies outperform `Accept-cH`.

(1.73), followed closely by `PAbFor` (1.98) and `FAbFor` (2.37), while `Accept-cH` consistently ranks last (3.88). Again, `PAbFor` and `IntAbFor` perform similarly on many datasets because uncertainty naturally increases over the forecast horizon.

F.8 Additional experiment: Mean Absolute Error (MAE) as risk measure

In this section, we repeat the experiments in Q1 and Q2 but measure the performance using the Mean Absolute Error (MAE) over the accepted steps as the risk measure. Thus, the conditional risk coincides in this setting with the conditional absolute deviation.

Estimating this quantity requires two important modifications to the algorithms introduced in Section 3. First, we replace the output head estimating the conditional variance with one estimating the conditional absolute deviation. This estimator $\hat{b}_t$ can be constructed by solving a regression task of the output variable:

$$\hat{b}_t(y_{1:T}) = \left| y_t - \hat{f}_t(y_{1:T}) \right|$$

Second, we train the model using the β-NLL loss under the Laplace distribution:

$$L^{Laplace}_{\beta-NLL}(y, \hat{f}\,(y_{1:T}), \hat{b}\,(y_{1:T})) = \sum_{t=T+1}^{T+H} \mathrm{s}\left(\hat{b}_t^{\beta}(y_{1:T})\right) \left(\frac{\left| y_t - \hat{f}_t(y_{1:T}) \right|}{\hat{b}_t(y_{1:T})} + \log\left(2\hat{b}_t(y_{1:T})\right) \right)$$

Then, we simply replace the conditional variance estimator with this absolute difference estimator $\hat{b}_t$ in `FAbFor`, `PAbFor`, and `IntAbFor`. Moreover, we train `DP-RNN` using the MAE loss in this experiment.

`FAbFor` **significantly outperforms its competitors when MAE is used as the risk measure.** Figure 14 shows each dataset's average empirical selective risk across six target coverage levels for all considered methods in the full abstention setting when the selective risk is measured as the MAE over the accepted steps. Overall, `FAbFor` reduces the selective risk by an average of 12%

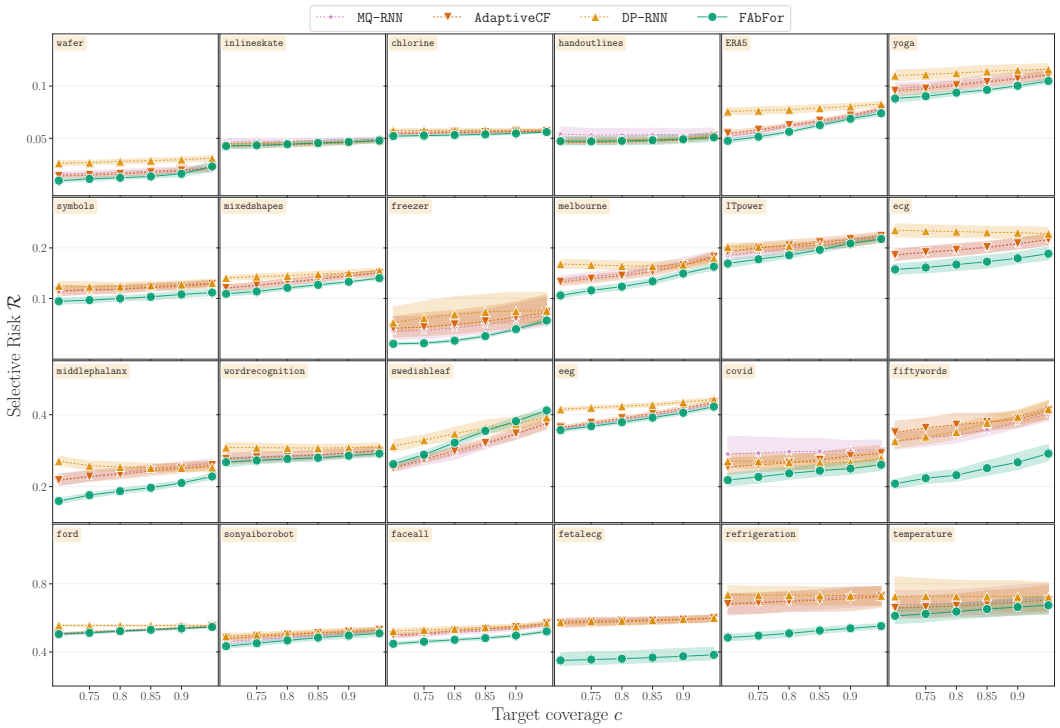

Figure 14: Each dataset's average selective risk (measured as MAE, with 95% c.i.) for all methods across six target coverages in the full abstention setting. For all the coverage levels considered, FAbFor achieves a lower selective risk than the baselines in most of the datasets.

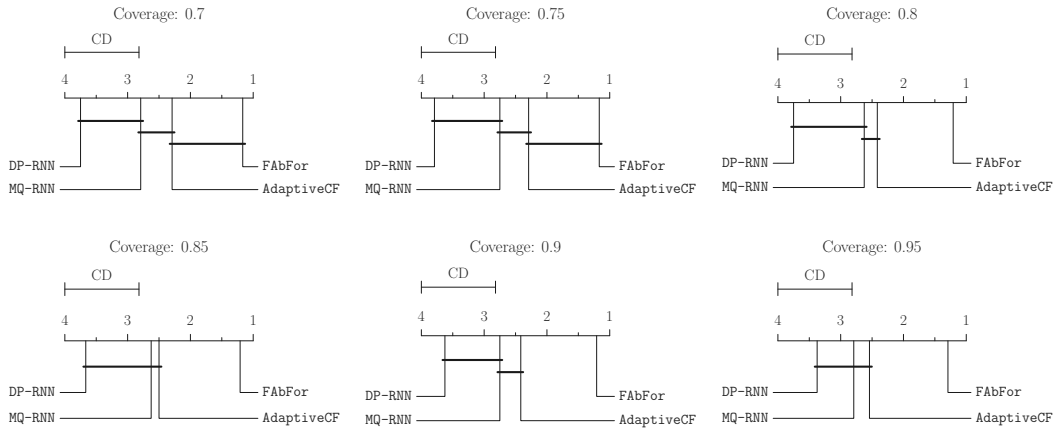

Figure 15: CD plots of empirical selective risk (measured as MAE) for different target coverages. FAbFor significantly outperforms the baselines for all considered target coverages.

vs AdaptiveCF, 13% vs MQ−RNN, and 19% vs DP−RNN. Moreover, FAbFor achieves lower selective risk in around 88% of the experiments against the best runner-up MQ−RNN.

To evaluate statistical significance, we apply the method proposed in Demšar (2006) separately for each target coverage and use the Bonferroni correction to account for multiple hypothesis tests. For each target coverage, the Friedman test rejects the null-hypothesis that all methods perform similarly. According to the Nemenyi post-hoc test (Demšar, 2006), FAbFor significantly outperforms DP−RNN and MQ−RNN for all considered target coverages, and outperforms AdaptiveCF when $c \geq 0.8$; Figure 15 shows all critical difference plots.

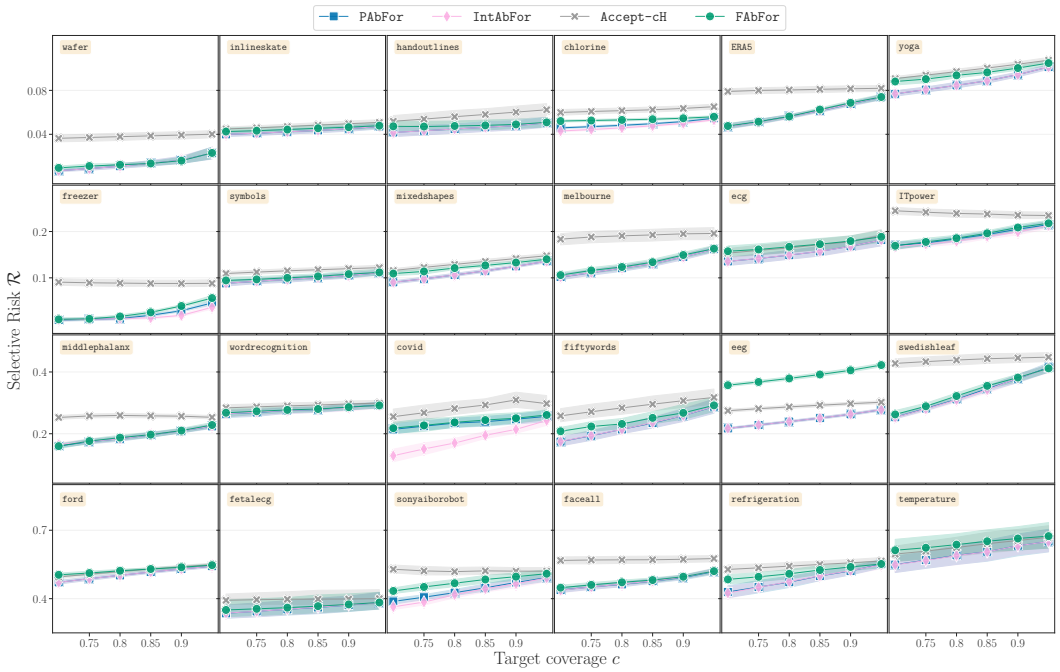

Figure 16: Each dataset's average selective risk (with 95% c.i.) for `FAbFor`, `PAbFor`, `IntAbFor`, and `Accept-cH` across six target coverages using MAE as risk measure. `IntAbFor` achieves often the lowest risk, followed by, in order, `PAbFor` and `FAbFor`. All strategies almost alwaysoutperform `Accept-cH`.

Finer-grained control in the abstention strategy yields lower selective risk when MAE is used as risk measure. We evaluate our three proposed strategies against the naive baseline `Accept-cH` using MAE as risk measure. Figure 16 shows the average selective risk across six target coverage levels for all datasets. Consistent with the results in Q2, our approaches outperform `Accept-cH` and maintain the exact same performance hierarchy. Specifically, ranking the methods from best (rank 1) to worst (rank 4) for each experiment, `IntAbFor` achieves the best average rank (1.56), followed closely by `PAbFor` (1.68) and `FAbFor` (2.90), while `Accept-cH` consistently ranks last (3.85). Again, `PAbFor` and `IntAbFor` perform similarly on many datasets because uncertainty naturally increases over the forecast horizon.

F.9 Additional experiment: autoregressive forecasters

In this section, we repeat the experiments in Q1 and Q2 but generate the H-step-ahead forecasts autoregressively. Concretely, the encoder still compresses the time-series $y_{1:T}$ into a hidden state, but the forecast of each step is fed back as the input of the next one, so that $\hat{f}_{T+k}$ is conditioned on $\hat{f}_{T+1:T+k-1}$. Every step-wise head additionally receives the normalized position k/H in the horizon, since it does not have a dedicated output slot per step. During training we use scheduled sampling: the fed-back value is the ground truth with probability ρ and the model's own output otherwise, where ρ decays linearly from 1 to 0 over the first half of the epochs. We apply the same modification to all the forecasters and keep every other component of our strategies and the baselines unchanged. In particular, we roll out only the step forecasts and leave the variance estimator conditioned on the context alone: in `FAbFor`, `PAbFor`, and `IntAbFor` the variance head keeps reading $y_{1:T}$ and emitting H outputs to estimate the *marginal* conditional risk.

`FAbFor` **outperforms its competitors when the forecasts are generated autoregressively.** Figure 17 shows each dataset's average empirical selective risk $\hat{\mathcal{R}}$ across six target coverage levels for all considered methods in the full abstention setting using autoregressive generation. Overall, `FAbFor` reduces the selective risk by an average of 13% vs `AdaptiveCF`, 15% vs `DP-RNN`, and 23% vs `MQ-RNN`, and achieves a lower selective risk in around 81% of the experiments against the best runner-up `AdaptiveCF`.

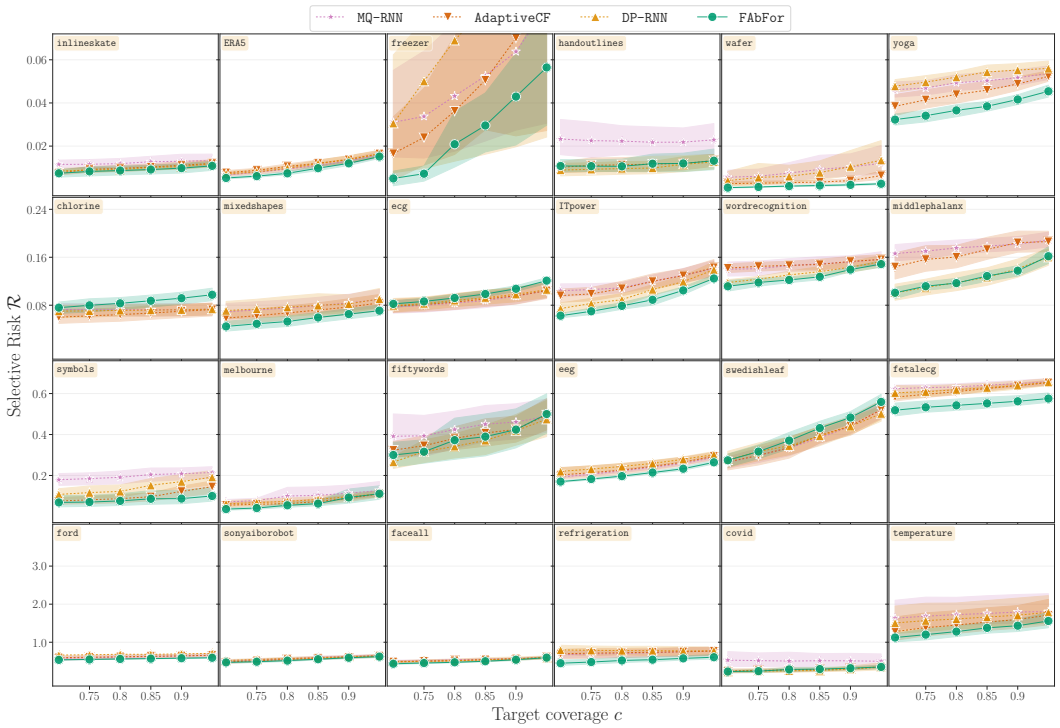

Figure 17: Each dataset's average selective risk (with 95% c.i.) for all methods across six target coverages in the full abstention setting when the forecasts are produced autoregressively. For all the coverage levels considered, `FAbFor` achieves a lower selective risk than the baselines in most of the datasets.

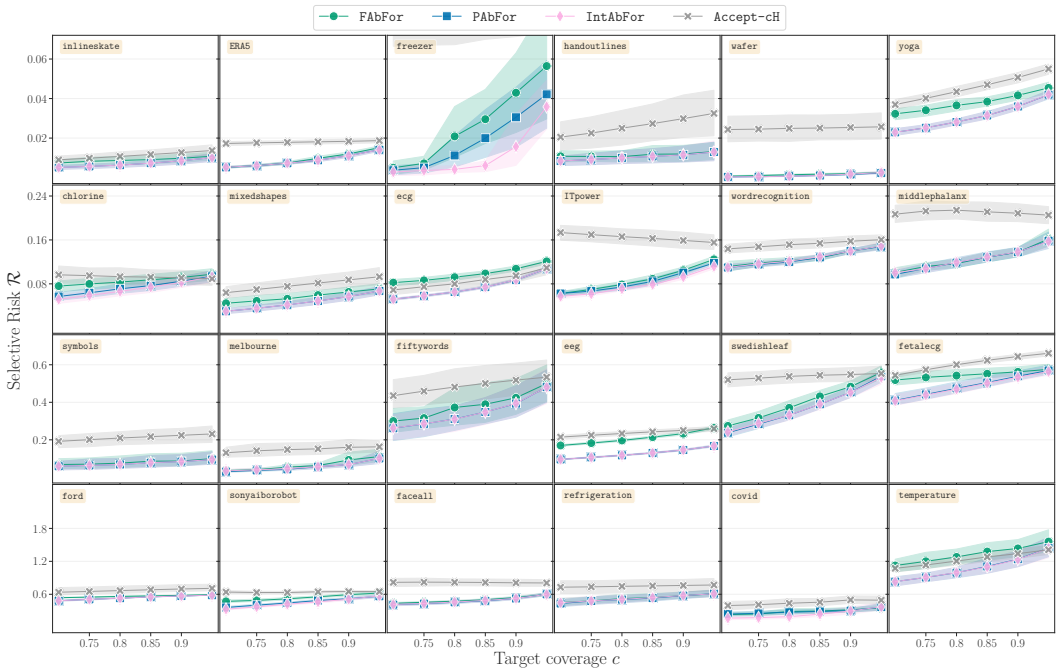

Figure 18: Each dataset's average selective risk (with 95% c.i.) for `FAbFor`, `PAbFor`, `IntAbFor`, and `Accept-cH` across six target coverages when the forecasts are produced autoregressively.

Finer-grained control in the abstention strategy yields lower selective risk when the forecasts are generated autoregressively. We evaluate our three proposed strategies against the naive base-

line `Accept-cH` under autoregressive generation. Figure 18 shows the average selective risk across six target coverage levels for all datasets. Consistent with the results in Q2, our approaches outperform `Accept-cH` and maintain the exact same performance hierarchy: across all dataset and coverage level combinations, `IntAbFor` achieves the best performance in 81 (dataset coverage) pairs, followed by `PAbFor` in 56 pairs, and `FAbFor` in 7 ones. Hence, the benefits of our abstention strategies do not depend on how the multi-horizon forecast is produced, but only on the availability of a per-step marginal uncertainty estimate.