Conformalized Adaptive Forecasting of Heterogeneous Trajectories

Yanfei Zhou [1] **Lars Lindemann** [2] **Matteo Sesia** [1]

Abstract

This paper presents a new conformal method for generating *simultaneous* forecasting bands guaranteed to cover the *entire path* of a new random trajectory with sufficiently high probability. Prompted by the need for dependable uncertainty estimates in motion planning applications where the behavior of diverse objects may be more or less unpredictable, we blend different techniques from online conformal prediction of single and multiple time series, as well as ideas for addressing heteroscedasticity in regression. This solution is both principled, providing precise finite-sample guarantees, and effective, often leading to more informative predictions than prior methods.

1. Introduction

Time series forecasting is a crucial problem with numerous applications in science and engineering. Many machine learning algorithms, including deep neural networks, have been developed to address this task, but they are typically designed to produce point predictions and struggle to quantify uncertainty. This limitation is especially problematic in domains involving intrinsic unpredictability, such as human behavior, and in high-stakes situations like autonomous driving (Lindemann et al., 2023; Lekeufack et al., 2023) or wildfire forecasting (Xu et al., 2022; 2023a).

A popular framework for endowing any model with reliable uncertainty estimates is that of *conformal prediction* (Vovk et al., 2005; Lei et al., 2018a). The idea is to observe and quantify the model's predictive performance on a *calibration* data set, independent of the training sample. If those data are sampled from the test population, the calibration performance is representative of the performance at test time. Thus, it becomes possible, with suitable algorithms, to convert any model's point predictions into *intervals* (or sets) with guaranteed coverage properties for future observations.

Conformal prediction typically hinges on *exchangeability*—an assumption less stringent than the requirement for calibration and test data to be independent and identically distributed. Under data exchangeability, conformal prediction can provide reliable statistical safeguards for any predictive model. Its flexibility enables applications across many tasks, including regression (Lei & Wasserman, 2014; Romano et al., 2019; Sesia & Romano, 2021), classification (Lei et al., 2013; Sadinle et al., 2019; Podkopaev & Ramdas, 2021), outlier detection (Bates et al., 2023; Marandon et al., 2024; Liang et al., 2024), and time series forecasting (Xu & Xie, 2021; Stankeviciute et al., 2021; Xu & Xie, 2023b; Ajroldi et al., 2023). This paper focuses on the last topic.

Conformal methods for time series tend to fall into one of two categories: *multi-series* and *single-series*. Methods in the former category aim to predict a new trajectory by leveraging other *jointly exchangeable* trajectories from the same population (Stankeviciute et al., 2021; Lindemann et al., 2023; Lekeufack et al., 2023). In the single-series setting, the aim shifts to forecasting future values based on historical observations from a fixed series, typically avoiding strict exchangeability assumptions (Gibbs & Candès, 2021; 2022; Angelopoulos et al., 2024). This paper draws inspiration from both areas and addresses a remaining limitation of current methods for multi-series forecasting.

The challenge addressed in this paper is that of *data heterogeneity*—distinct time series with different levels of unpredictability. For instance, in motion planning, forecasting the paths of pedestrians may be complicated by the relatively erratic behavior of some individuals, such as small children or intoxicated adults. This variability aligns with the classical issue of *heteroscedasticity*. The latter has recently gained some recognition within the conformal prediction literature, particularly for regression (Romano et al., 2019) and classification (Romano et al., 2020b; Einbinder et al., 2022). In this paper, we address heteroscedasticity within the more complex setting of trajectory forecasting.

Related Work

The challenge of conformal inference for non-exchangeable data is receiving significant attention, both from more gen-

[1]Department of Data Sciences and Operations, University of Southern California, Los Angeles, CA, USA [2]Department of Computer Science, University of Southern California, Los Angeles, CA, USA. Correspondence to: Matteo Sesia <sesia@marshall.usc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

eral perspectives (Tibshirani et al., 2019; Barber et al., 2023; Qiu et al., 2023) and in the context of time-series forecasting. An important line of research has focused on forecasting a single series, including recent works inspired by Gibbs & Candès (2021) such as Gibbs & Candès (2022); Bastani et al. (2022); Zaffran et al. (2022); Feldman et al. (2023); Dixit et al. (2023); Angelopoulos et al. (2024); Bhatnagar et al. (2023). Further, other approaches that combine conformal prediction with single-series forecasting include those of Chernozhukov et al. (2018); Xu & Xie (2021; 2023a;b); Sousa et al. (2022); Auer et al. (2023); Xu et al. (2023b). The present paper builds on this extensive body of work, drawing particular inspiration from Gibbs & Candès (2021). However, our approach is distinct in its pursuit of stronger simultaneous coverage guarantees, a goal justified by motion planning applications, for example, but not achievable within the constraints of single-series forecasting.

Conformal prediction in multi-series forecasting has so far received relatively less attention. Lin et al. (2022) explored a somewhat related yet distinct problem. Their work focused on ensuring different types of "longitudinal" and "cross-sectional" coverage, which is a different goal compared to our objective of simultaneously forecasting an *entire* new trajectory. We conduct direct comparisons between our method and those of Stankeviciute et al. (2021) and Yu et al. (2023); Cleaveland et al. (2024). These address problems akin to ours but adopt different approaches and do not focus on heteroscedasticity. Specifically, Stankeviciute et al. (2021) implemented a Bonferroni correction, which is often very conservative, while Yu et al. (2023), Cleaveland et al. (2024), and Sun & Yu (2023) used a technique more aligned with ours but lacking in adaptability to heteroscedastic conditions.

2. Background and Motivation

2.1. Problem Statement and Notation

We consider a data set comprising n observations of arrays of length $(T+1)$, namely $\mathcal{D} := \{\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(n)}\}$. For $i \in [n] := \{1, \ldots, n\}$, the array $\boldsymbol{Y}^{(i)} = (Y_0^{(i)}, Y_1^{(i)}, \ldots, Y_T^{(i)})$ represents $T+1$ observations of some d-dimensional vector $Y_t^{(i)} = (Y_{t,1}^{(i)}, \ldots, Y_{t,d}^{(i)}) \in \mathbb{R}^d$, measured at distinct time steps $t \in \{0, \ldots, T\}$. We will assume throughout the paper that the n trajectories are sampled exchangeably from some arbitrary and unknown distribution P. However, it is worth emphasizing that we make no assumptions about the potentially complex time dependence with each series $(Y_0^{(i)}, Y_1^{(i)}, \ldots, Y_T^{(i)})$. Intuitively, our goal is to leverage the data in $\mathcal{D}$ to construct an informative *prediction band* for the trajectory of a new series $\boldsymbol{Y}^{(n+1)}$, which is assumed to be also sampled exchangeably from the same distribution.

For simplicity, we focus on *one-step-ahead* forecasting, which means that we want to construct a prediction band for $\boldsymbol{Y}^{(n+1)}$ one step at a time. That is, we imagine that the initial position $Y_0^{(n+1)}$ is given and then wait to observe $Y_{t-1}^{(n+1)}$ before predicting $Y_t^{(n+1)}$, for each $t \in [T]$. This perspective is often useful, for example in motion planning applications, but it is of course not the only possible one. Fortunately, though, our solution for the one-step-ahead problem can easily be extended to *multiple-step-ahead* forecasting, as explained in Appendix A6, or even *one-shot* forecasting of an entire trajectory.

Let $\hat{C}(\boldsymbol{Y}^{(n+1)}) := (\hat{C}_1(\boldsymbol{Y}^{(n+1)}), \ldots, \hat{C}_T(\boldsymbol{Y}^{(n+1)}))$ represent the output prediction band, where each $\hat{C}_t(\boldsymbol{Y}^{(n+1)}) \subseteq \mathbb{R}^d$ is a prediction region for the vector $Y_t^{(n+1)}$ that may depend on past observations $Y_s^{(n+1)}$ for $s < t$, as well as on the data in $\mathcal{D}$. As we develop a method to construct $\hat{C}(\boldsymbol{Y}^{(n+1)})$, one goal is to ensure the following notion of *simultaneous marginal coverage*:

$$\mathbb{P}\left[Y_t^{(n+1)} \in \hat{C}_t(\boldsymbol{Y}^{(n+1)}),\ \forall t \in [T]\right] \geq 1 - \alpha. \quad (1)$$

Simply put, the entire trajectory should lie within the band with probability at least $1-\alpha$, for some chosen level $\alpha \in (0, 1)$. This property is called *marginal* because it treats both $\boldsymbol{Y}^{(n+1)}$ and the data in $\mathcal{D}$ as random samples from P.

2.2. Benefits and Limitations of Marginal Coverage

Marginal coverage is not only convenient, since it is achievable under quite realistic assumptions, but also useful. For example, in motion planning, prediction bands with simultaneous marginal coverage can help autonomous vehicles decide on a path that is unlikely to collide with another vehicle or pedestrian at any point in time. However, the marginal nature of Equation (1) is not always fully satisfactory, particularly because it may obscure the adverse impacts of heterogeneity across trajectories, as explained next.

Imagine forecasting the movement of pedestrians crossing a street at night. Suppose that 90% of them are sober, walking in highly predictable patterns, while the remaining 10% are intoxicated. See Figure 1 for a visualization of this scenario. It is clear that uncertainty estimation is of paramount concern while forecasting the harder-to-predict drunk trajectories. Addressing this issue is crucial, for example, to ensure that autonomous vehicles navigate such environments with the necessary level of caution. However, not all prediction bands with marginal coverage are equally useful in this context. For example, 90% marginal coverage could be easily attained even by a trivial algorithm that provides valid prediction bands only for trajectories of the "easy" type. This thought experiment shows that despite their general theoretical guarantees, conformal prediction methods still require careful design to provide informative uncertainty estimates, particularly in the case of heterogeneous data.

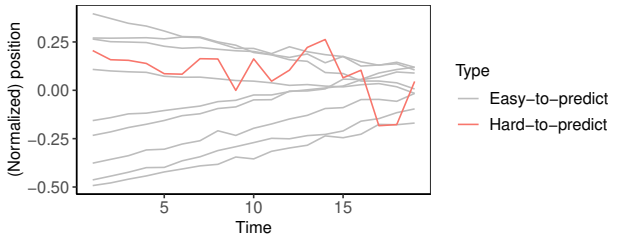

Figure 1. One-dimensional representations of 10 pedestrian trajectories, one of which is intrinsically less predictable.

The aforementioned limitations of marginal coverage have been acknowledged before. While achieving stronger theoretical guarantees in finite samples is generally unfeasible (Vovk, 2012; Barber et al., 2021a), some approaches practically tend to work better in this regard than others. In particular, methods have been developed for regression (Romano et al., 2019; Izbicki et al., 2020), classification (Romano et al., 2020b; Cauchois et al., 2021; Einbinder et al., 2022), and sketching (Sesia et al., 2023) to seek *approximate conditional coverage* guarantees stronger than (1).

2.3. Towards Approximate Conditional Coverage

The goal in this paper is to construct prediction bands that are valid not only for a large fraction of all trajectories but also with high probability for distinct "types" of trajectory. In our street crossing example, this means we would like to have valid coverage not only marginally but also *conditional* on some relevant features of the pedestrian. For example, one may want $\hat{C}^{(n+1)}$ to approximately satisfy

$$\mathbb{P}\left[Y_t^{(n+1)} \in \hat{C}_t(\boldsymbol{X}^{(n+1)}),\ \forall t \mid \phi(\boldsymbol{X}^{(n+1)})\right] \geq 1-\alpha, \quad (2)$$

where ϕ could represent the indicator of whether $\boldsymbol{X}^{(n+1)}$ corresponds to an intoxicated pedestrian.

While there exist algorithms providing coverage conditional on a limited set of discrete features (Romano et al., 2020a), our challenge exceeds the capabilities of available approaches. One issue is that the relevant features might not be directly observable. For example, an autonomous vehicle might only detect a pedestrian's movements in real time, lacking broader contextual information about that person, such as knowing whether they are intoxicated or sober. Therefore, our problem requires an innovative approach.

2.4. Preview of Main Contributions

We introduce a novel approach for constructing prediction bands for (multi-dimensional) trajectories, called Conformalized Adaptive Forecaster for Heterogeneous Trajectories (CAFHT). This method guarantees simultaneous marginal coverage as defined in (1) and is shown to achieve superior conditional coverage in practice compared to existing methods, as indicated by (2). A key feature of CAFHT is that it does not require pre-specified labels of intrinsic difficulty but rather it automatically adjusts the width of its prediction bands to each new trajectory in an online manner. This adaptability is derived from the capabilities of Adaptive Conformal Inference (ACI) (Gibbs & Candès, 2021), which dynamically adjusts the prediction intervals to reflect the ease or challenge of predicting subsequent steps in a given trajectory. Additionally, our method inherits from ACI the ability to produce prediction bands that are generally valid even for worst-case trajectories, provided these trajectories are of sufficient length (Gibbs & Candès, 2021).

Figure 2 offers a glimpse into the effectiveness of CAFHT applied to the pedestrian trajectories from Figure 1. Our method's advantage over state-of-the-art techniques (Stankeviciute et al., 2021; Yu et al., 2023) lies in its ability to automatically generate narrower bands for easier trajectories and wider ones for harder paths. As shown through extensive experiments, this leads to more useful uncertainty estimates with higher conditional coverage. In contrast, existing methods struggle to accommodate heterogeneity, often resulting in uniform prediction bands for all trajectories.

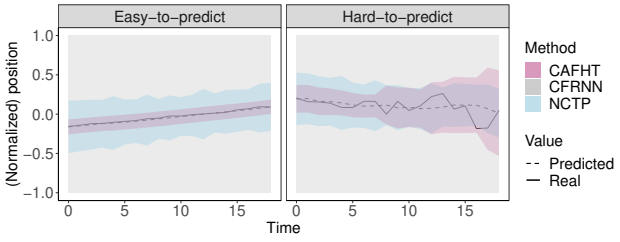

Figure 2. Conformal forecasting bands constructed using different methods, for the heterogeneous pedestrian trajectories from Figure 1. All methods guarantee simultaneous marginal coverage at the 90% level. Our method (CAFHT) can automatically adapt to the unpredictability of each trajectory. Here, the CFRNN bands so wide as to be uninformative, spanning from -1 to +1.

In the next section, we explain how our approach integrates traditional split-conformal inference with online conformal prediction (Gibbs & Candès, 2021; 2022; Angelopoulos et al., 2024). Originally designed for single-series forecasting, these methods are adapted in our setting to construct flexible prediction bands that automatically adjust to the varying unpredictability of each trajectory. For clarity, we begin by describing an implementation of our method based on ACI (Gibbs & Candès, 2021), though other methods could also be accommodated, including the conformal PID approach from Angelopoulos et al. (2024) (discussed further in Section 3.7). It is crucial to note that all implementations of CAFHT are designed to provide the same guarantee of simultaneous marginal coverage and the same capability to accommodate heteroscedasticity.

3. Methodology

3.1. Training a Black-Box Forecasting Model

The preliminary step in our CAFHT method consists of randomly partitioning the data set $\mathcal{D}$ into two distinct subsets of trajectories, $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{cal}}$. The subset $\mathcal{D}_{\text{train}}$ is used to train a forecasting model $\hat{g}$. This model could be almost anything, including a long short-term memory network (LSTM) (Hochreiter & Schmidhuber, 1997; Alahi et al., 2016), a transformer network (Nayakanti et al., 2022; Zhou et al., 2023), or a traditional autoregressive moving average model (Wei et al., 2023). Our only assumption regarding $\hat{g}$ is that it is able to generate point predictions for future steps based on partial observations from a new time series.

In this paper, we choose an LSTM model for demonstration and focus on one-step-ahead predictions. While the ability of CAFHT to guarantee simultaneous marginal coverage does not depend on the forecasting accuracy of $\hat{g}$, more accurate models generally tend to yield more informative conformal predictions (Lei et al., 2018b).

3.2. Initializing the Adaptive Prediction Bands

After training the forecaster $\hat{g}$ on the data in $\mathcal{D}_{\text{train}}$, our method will convert its one-step-ahead point predictions for any new trajectory $\boldsymbol{Y}$ into suitable *prediction bands*. This is achieved by applying the ACI algorithm of Gibbs & Candès (2021). For simplicity, we begin by focusing on the special case of one-dimensional trajectories ($d = 1$). An extension of our solution to higher-dimensional trajectories is deferred to Section 3.6.

ACI was designed to generate one-step-ahead forecasts for a single one-dimensional time series, without requiring a pre-trained forecaster $\hat{g}$. In the single-series framework, Gibbs & Candès (2021) suggested training $\hat{g}$ in an online manner. In our setting, where we have access to multiple trajectories from the same population, it is logical to pre-train it. In any case, pre-training does not exclude the potential for further online updates of $\hat{g}$ with each subsequent one-step-ahead prediction. However, to simplify the notation, our discussion now focuses on a static model.

A review of ACI (Gibbs & Candès, 2021) can be found in Appendix A1. Here, we briefly highlight two critical aspects of that algorithm. Note that the main ideas of our method can also be straightforwardly applied in combinations with other variations of the ACI method, as shown in Section 3.7.

Firstly, the ACI algorithm involves a "learning rate" parameter $\gamma > 0$, controlling the adaptability of the prediction bands to the evolving time series. The adjustment mechanism operates as follows: at each time t, ACI modifies the width of the upcoming prediction interval for $Y^{(t+1)}$. If the previous interval failed to encompass $Y^{(t)}$, the next interval is expanded; conversely, if it was sufficient, the next interval is narrowed. Thus, larger values of γ result in more substantial adjustments at each time step. In contrast, lower values of γ generally lead to "smoother" prediction bands.

Secondly, the width of the ACI prediction band is also influenced by a parameter $\alpha \in (0, 1)$, which represents the nominal level of the method. The design of the ACI algorithm aims to ensure that, over an extended period, the generated prediction bands will accurately contain the true value of Y_t approximately a $1 - \alpha$ fraction of the time. Consequently, a smaller α leads to broader bands.

Within our context, ACI is useful to transform the point predictions of $\hat{g}$ into *uncertainty-aware* prediction bands, but it is not satisfactory on its own. Firstly, it is not always clear how to choose a good learning rate. Secondly, the ACI prediction bands lack finite-sample guarantees. Specifically, they do not guarantee simultaneous marginal coverage (1). Our method overcomes these limitations as follows.

3.3. Calibrating the Adaptive Prediction Bands

We now discuss how to calibrate the ACI prediction bands discussed in the previous section to achieve simultaneous marginal coverage (1). For simplicity, we begin by taking the learning rate parameter γ as fixed. We will then discuss later how to optimize the choice of γ in a data-driven way.

Let $\hat{C}^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \gamma) = [\hat{\ell}^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \gamma), \hat{u}^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \gamma)]$ denote the prediction band constructed by ACI, with learning rate γ and level $\alpha_{\text{ACI}} \in (0, 1)$, for each *calibration* trajectory $i \in \mathcal{D}_{\text{cal}}$. Note that this band is constructed one step at a time, based on the point predictions of $\hat{g}$ at each step $t \in [T]$ and past observations of $Y_s^{(i)}$ for all $s < t$; see Appendix A1 for further details on ACI. We will refer to the cross-sectional prediction interval identified by this band at time $t \in [T]$ as $\hat{C}_t^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \gamma) = [\hat{\ell}_t^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \gamma), \hat{u}_t^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \gamma)]$.

Our method will transform these ACI bands, which can only achieve a weaker notion of *asymptotic average coverage* because they do not leverage any exchangeability, into simultaneous prediction bands satisfying (2). For each $i \in \mathcal{D}_{\text{cal}}$, CAFHT evaluates a *conformity score* $\hat{\epsilon}_i(\gamma)$:

$$\hat{\epsilon}_i(\gamma) := \max_{t \in [T]} \left\{ \max \left\{ \left[\hat{\ell}_t^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \gamma) - Y_t^{(i)} \right]_+, \left[Y_t^{(i)} - \hat{u}_t^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \gamma) \right]_+ \right\} \right\}, \tag{3}$$

where $[x]_+ := \max(0, x)$ for any $x \in \mathbb{R}$. Intuitively, $\hat{\epsilon}_i(\gamma)$ measures the largest margin by which $\hat{C}^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \gamma)$ should be expanded in both directions to simultaneously cover the entire trajectory $\boldsymbol{Y}^{(i)}$ from $t = 1$ to $t = T$. This is in-

spired by the method of Romano et al. (2019) for quantile regression, although one difference is that their scores may be negative. Other choices of conformity scores are also possible in our context, however, as discussed in Section 3.5.

Let $\hat{Q}(1-\alpha, \gamma)$ denote the $\lceil (1-\alpha)(1+|\mathcal{D}_{\text{cal}}|) \rceil$-th smallest value of $\hat{\epsilon}_i(\gamma)$ among $i \in \mathcal{D}_{\text{cal}}$. CAFHT constructs a prediction band $\hat{C}(\boldsymbol{Y}^{(n+1)}, \gamma)$ for $\boldsymbol{Y}^{(n+1)}$, one step at a time, as follows. Let $\hat{C}_t^{\text{ACI}}(\boldsymbol{Y}^{(n+1)}, \gamma)$ denote the ACI prediction interval for $Y_t^{(n+1)}$ at time $t \in [T]$. (Recall this depends on $\hat{g}$ and $\boldsymbol{Y}_s^{(n+1)}$ for all $s < t$.) Then, define the interval

$$\hat{C}_t(\boldsymbol{Y}^{(n+1)}, \gamma) = \Big[\hat{\ell}_t^{\text{ACI}}(\boldsymbol{Y}^{(n+1)}, \gamma) - \hat{Q}(1-\alpha, \gamma), \\ \hat{u}_t^{\text{ACI}}(\boldsymbol{Y}^{(n+1)}, \gamma) + \hat{Q}(1-\alpha, \gamma)\Big]. \tag{4}$$

Our prediction band $\hat{C}(\boldsymbol{Y}^{(n+1)}, \gamma)$ for one-step-ahead forecasting is then obtained by concatenating the intervals in (4) for all $t \in [T]$. More compactly, we can write $\hat{C}(\boldsymbol{Y}^{(n+1)}, \gamma) = \hat{C}^{\text{ACI}}(\boldsymbol{Y}^{(n+1)}, \gamma) \pm \hat{Q}(1-\alpha, \gamma)$.

The next result establishes finite-sample simultaneous coverage guarantees for this method.

Theorem 1. *Assume that the calibration trajectories in $\mathcal{D}_{cal}$ are exchangeable with $\boldsymbol{Y}^{(n+1)}$. Then, for any $\alpha \in (0,1)$, the prediction band output by CAFHT, applied with fixed parameters α, α_{ACI}, and γ, satisfies (1).*

The proof of Theorem 1 is relatively simple and can be found in Appendix A2. We remark that this guarantee holds at the desired level α regardless of the value of the ACI parameter α_{ACI}. However, it is typically intuitive to set $\alpha_{\text{ACI}} = \alpha$. A notable advantage of this choice is that it leaves us with the challenge of tuning only one ACI parameter, γ.

Further, it is important to note that CAFHT can only expand the ACI prediction bands, since its conformity scores are non-negative. Thus, our method retains the ACI guarantee of asymptotic average coverage at level $1-\alpha$ (Gibbs & Candès, 2021), almost surely for *any* trajectory $\boldsymbol{Y}^{(n+1)}$:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}[Y_t \notin \hat{C}_t(\boldsymbol{Y}^{(n+1)}, \gamma)] \overset{\text{a.s.}}{=} \alpha. \tag{5}$$

See Appendix A1 for details about how ACI achieves (5).

3.4. Data-Driven Parameter Selection

The ability of the ACI algorithm to produce informative prediction bands can sometimes be sensitive to the choice of the learning rate γ (Gibbs & Candès, 2021; Angelopoulos et al., 2024). This leads to a question: how can we select γ in a data-driven manner? In our scenario, which involves multiple relevant trajectories from the same population, addressing this tuning challenge is somewhat simpler than in the original single-series context for which the ACI algorithm was designed. Nonetheless, careful consideration is still required in the tuning process of γ, as we discuss next.

As a naive approach, one may feel tempted to apply the CAFHT method described above using different learning rates, with the idea of then cherry-picking the value of γ leading to the most appealing prediction bands. Unsurprisingly, however, such an unprincipled approach would invalidate the coverage guarantee because it breaks the exchangeability between the test trajectory and the calibration data. This issue is closely related to problems of conformal prediction after model selection previously studied by Yang & Kuchibhotla (2021) and Liang et al. (2023). Therefore, we propose two alternative solutions inspired by their works.

The simplest approach to explain involves an additional data split. Let us randomly partition $\mathcal{D}_{\text{cal}}$ into two subsets of trajectories, $\mathcal{D}_{\text{cal}}^1$ and $\mathcal{D}_{\text{cal}}^2$. The trajectories in $\mathcal{D}_{\text{cal}}^1$ can be utilized to select a good choice of γ in a data-driven way. In particular, we seek the value of γ leading to the most informative prediction bands—a goal that can be quantified by minimizing the average width of our prediction bands produced for the trajectories in $\mathcal{D}_{\text{cal}}^1$. Then, the calibration procedure described in Section 3.3 will be applied using only the data in $\mathcal{D}_{\text{cal}}^2$ instead of the full $\mathcal{D}_{\text{cal}}$. The fact that the selection of γ does not depend on the calibration trajectories in $\mathcal{D}_{\text{cal}}^2$ means that γ can be essentially regarded as fixed, and therefore our output bands enjoy the marginal simultaneous coverage guarantee of Theorem 1. This version of our CAFHT method is outlined in Algorithm 1. The parameter tuning module of this procedure is summarized by Algorithm A1 in Appendix A3.

Alternatively, it is also possible to carry out the selection of $\hat{\gamma}$ in a rigorous way without splitting $\mathcal{D}_{\text{cal}}$. However, this would require replacing the empirical quantile $\hat{Q}(1-\alpha, \hat{\gamma})$ in the CAFHT method with a more conservative quantity $\hat{Q}(1-\alpha', \hat{\gamma})$, where the value of $\alpha' < \alpha$ depends on the number L of candidate parameter values considered. We refer to Appendix A4 for further details.

Our method employs a grid search to optimize the ACI hyper-parameters, a standard practice for hyper-parameter tuning. It is important to note that the more computationally demanding components of CAFHT, such as training the models and selecting γ via grid search, are conducted offline and require completion only once. After these preliminary steps, the real-time component of CAFHT, which constructs prediction bands for new test trajectories, is fast and efficient.

Algorithm 1 CAFHT

Input: A pre-trained forecaster $\hat{g}$ producing one-step-ahead predictions; calibration trajectories $\mathcal{D}_{\text{cal}}$; the initial position $Y_0^{(n+1)}$ of a test trajectory $\boldsymbol{Y}^{(n+1)}$; the desired nominal level $\alpha \in (0,1)$; a grid of candidate learning rates $\{\gamma_1, \ldots, \gamma_L\}$.
Randomly split $\mathcal{D}_{\text{cal}}$ into $\mathcal{D}^1_{\text{cal}}$ and $\mathcal{D}^2_{\text{cal}}$.
Select a learning rate $\hat{\gamma} \in \{\gamma_1, \ldots, \gamma_L\}$, applying Algorithm A1 using the trajectory data in $\mathcal{D}^1_{\text{cal}}$.
Construct $\hat{C}^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \hat{\gamma})$ using ACI, for $i \in \mathcal{D}^2_{\text{cal}}$.
Evaluate $\hat{\epsilon}_i(\hat{\gamma})$ using (3), for $i \in \mathcal{D}^2_{\text{cal}}$.
Compute the empirical quantile $\hat{Q}(1-\alpha, \hat{\gamma})$.
for $t \in [T]$ **do**
  Compute $\hat{C}^{\text{ACI}}_t(\boldsymbol{Y}^{(n+1)}, \hat{\gamma})$ with ACI, using the past of the trajectory $(Y_0^{(n+1)}, Y_1^{(n+1)}, \ldots, Y_{t-1}^{(n+1)})$.
  Compute a prediction interval $\hat{C}_t(\boldsymbol{Y}^{(n+1)}, \hat{\gamma})$ for the next step, using (4).
  Observe the next step of the trajectory, $Y_t^{(n+1)}$.
end for
Output: An online prediction band $\hat{C}(\boldsymbol{Y}^{(n+1)})$.

3.5. CAFHT with Multiplicative Scores

A potential shortcoming of Algorithm 1 is that it can only add a constant margin of error to the prediction band constructed by the ACI algorithm. While straightforward, this approach may not be always optimal. In many cases, it would seem more natural to utilize a multiplicative error. The rationale behind this is intuitive: trajectories that are inherently more unpredictable, resulting in wider ACI prediction bands, may necessitate larger margins of error to ensure valid simultaneous coverage. This concept can be seamlessly integrated into the CAFHT method by replacing the conformity scores initially outlined in (3) with these:

$$\tilde{\epsilon}_i(\gamma) = \max_{t \in T} \left\{ \max \left\{ \frac{\left[\hat{\ell}^{\text{ACI}}_t(\boldsymbol{Y}^{(i)}, \gamma) - Y_t^{(i)}\right]_+}{|\hat{C}^{\text{ACI}}_t(\boldsymbol{Y}^{(i)}, \gamma)|}, \frac{\left[Y_t^{(i)} - \hat{u}^{\text{ACI}}_t(\boldsymbol{Y}^{(i)}, \gamma)\right]_+}{|\hat{C}^{\text{ACI}}_t(\boldsymbol{Y}^{(i)}, \gamma)|} \right\} \right\}. \tag{6}$$

Then, the counterpart of Equation (4) becomes

$$\hat{C}(\boldsymbol{Y}^{(n+1)}, \gamma) = \hat{C}^{\text{ACI}}(\boldsymbol{Y}^{(n+1)}, \gamma) \pm \tilde{Q}(1-\alpha, \gamma) \cdot |\hat{C}^{\text{ACI}}(\boldsymbol{Y}^{(n+1)}, \gamma)|,$$

where $\tilde{Q}(1-\alpha, \gamma)$ is the $\lceil (1-\alpha)(1+|\mathcal{D}_{\text{cal}}|)\rceil$-th smallest value in $\{\tilde{\epsilon}_i(\gamma), i \in \mathcal{D}_{\text{cal}}\}$.

At this point, it is easy to prove that the prediction bands obtained produced by CAFHT with these multiplicative conformity scores still enjoy the same marginal simultaneous coverage guarantee established by Theorem 1.

We refer to Figures A27–A28 in Appendix A5.4 for empirical illustrations and comparisons of prediction bands generated with multiplicative and additive scores; see also Table A29 for a summary of their corresponding empirical quantiles $\hat{Q}(1-\alpha, \hat{\gamma})$.

3.6. Extension to Multi-Dimensional Trajectories

The problem of forecasting trajectories with $d > 1$ (e.g., a two-dimensional walk), can be addressed with an intuitive extension of CAFHT. In fact, ACI extends naturally to the multidimensional case and the first component of our method that requires some special care is the computation of the empirical quantile $\hat{Q}(1-\alpha, \hat{\gamma})$. Yet, even this obstacle can be overcome quite easily. Consider evaluating a vector-valued version of the additive scores from (3):

$$\hat{\epsilon}_{ij}(\gamma) := \max_{t \in [T]} \left\{ \max \left\{ \left[\hat{\ell}^{\text{ACI}}_{t,j}(\boldsymbol{Y}^{(i)}, \gamma) - Y_{t,j}^{(i)}\right]_+, \left[Y_{t,j}^{(i)} - \hat{u}^{\text{ACI}}_{t,j}(\boldsymbol{Y}^{(i)}, \gamma)\right]_+ \right\} \right\}, \tag{7}$$

for each dimension $j \in [d]$. Then, we can recover a one-dimensional problem prior to computing $\hat{Q}(1-\alpha, \hat{\gamma})$ by taking (for example) the maximum value of $\hat{\epsilon}_{ij}(\gamma)$; i.e., $\hat{\epsilon}^{\infty}_i(\gamma) = \max_{j \in [d]} \hat{\epsilon}_{ij}(\gamma)$. Ultimately, each $\hat{C}_t(\boldsymbol{Y}^{(n+1)}, \gamma)$ is obtained by applying (4) with $\hat{Q}(1-\alpha, \hat{\gamma})$ defined as the $\lceil (1-\alpha)(1+|\mathcal{D}_{\text{cal}}|)\rceil$-th smallest value of $\hat{\epsilon}^{\infty}_i(\gamma)$.

We conclude this section by remarking that this general idea could also be implemented using the multiplicative conformity scores described in Section 3.5, as well as by using different dimension reduction functions in (3). For example, one may consider replacing the infinity-norm in (3) with an ℓ^2 norm, leading to a "spherical" margin of error around the ACI prediction bands instead of a "square" one.

3.7. Leveraging Conformal PID Prediction Bands

CAFHT is not heavily reliant on the specific mechanics of ACI. The crucial aspect of ACI is its capability to transform black-box point forecasts into prediction bands that approximately mirror the unpredictability of each trajectory. Thus, our method can integrate with any variation of ACI.

Some of our demonstrations in Appendix A5 include an alternative implementation of CAFHT that employs the conformal PID algorithm of Angelopoulos et al. (2024) instead of ACI. To minimize computational demands, our demonstrations will primarily utilize the quantile tracking feature of the original conformal PID method. This simpli-

fied version of conformal PID is influenced only by a single hyper-parameter—a learning rate γ, similar to ACI.

3.8. Direct Comparison to ACI

CAFHT utilizes ACI as an internal component and is designed to leverage several exchangeable trajectories to construct prediction bands for a new trajectory sampled from the same population, while guaranteeing simultaneous marginal coverage as defined in (1). In contrast, ACI handles a single (arbitrary) trajectory and focuses on a different notion of asymptotic average coverage, which allows for temporary deviations of the true trajectory from the output prediction band. This crucial distinction between CAFHT and ACI is highlighted by the numerical experiments detailed in Figures A25–A26 in Appendix A5.3.

4. Numerical Experiments

4.1. Setup and Benchmarks

This section demonstrates the empirical performance of our method. We focus on applying CAFHT with multiplicative scores, based on the ACI algorithm, and tuning the learning rate through data splitting. Additional results pertaining to different implementations of CAFHT are in Appendix A5. In all experiments, the candidate values for the ACI learning rate parameter γ range from 0.001 to 0.1 at increments of 0.01, and from 0.2 to 0.9 at increments of 0.1.

The CAFHT method is compared with two benchmark approaches that also provide simultaneous marginal coverage (1). The first one is the Conformal Forecasting Recurrent Neural Network (CFRNN) approach of Stankeviciute et al. (2021), which relies on a Bonferroni correction for multiple testing. In particular, the CFRNN method produces a prediction band satisfying (1) for a trajectory of length T by separately computing T conformal prediction intervals at level α/T, one for each time step, each obtained using regression techniques typical to the regression setting. An advantage of this approach is that it is conceptually intuitive, but it can become quite conservative if T is large.

The second benchmark is the Normalized Conformal Trajectory Predictor (NCTP) of Yu et al. (2023). This method is closer to ours but utilizes different scores and does not leverage ACI to adapt to heterogeneity. In short, NCTP directly takes as input a forecaster $\hat{g}$ providing one-step-ahead point predictions $\hat{Y}_t^{(i)}$ and evaluates the scores $\hat{\epsilon}_i = \max_{t\in[T]}\{(|\hat{Y}_t^{(i)} - Y_t^{(i)}|)/\sigma_t\}$ for each $i \in \mathcal{D}_{\text{cal}}$, where σ_t are suitable data-driven normalization constants. This approach is similar to that of Cleaveland et al. (2024), which deviates only in the computation of the σ_t constants, and it tends to work quite well if the trajectories are homogeneous.

While there exist other methods, such as CopulaCPTS (Sun & Yu, 2023), which can achieve simultaneous marginal coverage as defined in (1), they, like NCTP, lack adaptability to heteroscedastic conditions, and are thus expected to perform similarly under such conditions. For clarity and conciseness, we focus on CFRNN and NCTP as the benchmarks in our primary experiments. Additional experiments involving CopulaCPTS, detailed in Appendix A5.6, demonstrate performance comparable to NCTP, as anticipated.

For all methods, the underlying forecasting model is a recurrent neural network with 4 stacked LSTM layers followed by a linear layer. The learning rate is set equal to 0.001, for an AdamW optimizer with weight decay 1e-6. The models are trained for a total of 50 epochs, so that the mean squared error loss loss approximately converges.

Prior to the beginning of our analyses, all trajectories will be pre-processed with a batch normalization step based on $\mathcal{D}_{\text{train}}$, so that all values lie within the interval $[-1, 1]$. This is useful to ensure a numerically stable learning process and more easily interpretable performance measures.

In all experiments, we evaluate the performance of the prediction bands in terms of their simultaneous marginal coverage (i.e., the proportion of test trajectories entirely contained within the prediction bands), the average width (over all times $t \in [T]$ and all test trajectories, which have a maximum value of 2 after standardizing our data to fall within the range $[-1, 1]$), and the simultaneous coverage conditional on a trajectory being "hard-to-predict", as made more precise in the next subsection.

It is crucial to note that while we, as experiment designers, are aware of the "difficulty label" for each trajectory, the methods used in this study do not have access to this information. Therefore, achieving high simultaneous conditional coverage is inherently challenging. Although not theoretically guaranteed to exceed any specific threshold, higher values of this measure are preferable for practical purposes.

4.2. Synthetic Trajectories

We begin by considering univariate ($d = 1$) synthetic trajectories generated from an autoregressive (AR) model, $X_t = 0.9X_{t-1} + 0.1X_{t-2} - 0.2X_{t-3} + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma_t^2)$, for all $t \in [T]$ with $T = 100$. Similar to Stankeviciute et al. (2021), we consider two noise profiles: a dynamic profile in which σ_t^2 is increasing with time, and a static profile in which σ_t^2 is constant. The results based on the dynamic profile are presented here, while the others are discussed in Appendix A5. To make the problem more interesting, we ensure that some trajectories are intrinsically more unpredictable than the others. Specifically, in the dynamic noise setting, we set $\sigma_t^2 = t \cdot k$, with $k = 10$, for a fraction $\delta = 0.1$ of the trajectories, while $\sigma_t^2 = t$ for the remaining ones.

Figure 3 summarizes the performance of the three methods

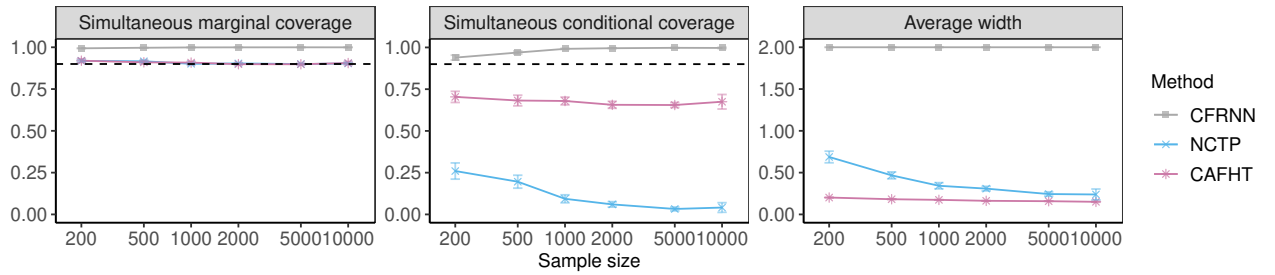

Figure 3. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories (of which 25% are utilized for calibration). All methods achieve 90% simultaneous marginal coverage. Our method (CAFHT) leads to more informative bands with lower average width and higher conditional coverage. The error bars indicate 2 standard errors. Note that the CFRNN bands here are so wide as to be uninformative.

as a function of the number of trajectories in $\mathcal{D}$, which is varied between 200 and 10,000. The results are averaged over 500 test trajectories and 100 independent experiments. See Table A1 in Appendix A5 for standard errors. In each case, 75% of the trajectories are used for training and the remaining 25% for calibration. Our method utilizes 50% of the calibration trajectories to select the ACI learning rate γ. All experiments target 90% simultaneous marginal coverage, with additional results for higher coverage levels presented in Appendix A5.5.

All methods attain 90% simultaneous marginal coverage, aligning with theoretical predictions. Notably, CAFHT yields the most informative bands, characterized by the narrowest average width and higher conditional coverage compared to NCTP. This can be explained by the fact that NCTP is not designed to account for the varying noise levels inherent in different trajectories. Consequently, NCTP generates less adaptive bands, too wide for the easier trajectories and too narrow for the harder ones. CAFHT also surpasses CFRNN; while CFRNN seems to attain the highest conditional coverage, it generates very wide bands that are practically uninformative for all trajectories. This is due to its rigid approach to handling time dependencies via a Bonferroni correction.

Figure 4 summarizes the results of similar experiments investigating the performances of different methods as a function of the prediction horizon T, which is varied between 5 and 100; see Table A2 in Appendix A5 for the corresponding standard errors. Here, the number of trajectories in $\mathcal{D}$ is fixed equal to 2000. The results highlight how CFRNN becomes more conservative as T increases. By contrast, NCTP produces relatively narrower bands but also achieves the lowest conditional coverage. Meanwhile, our CAFHT method again yields the most informative prediction bands, with low average width and high conditional coverage.

Appendix A5 describes additional experimental results that are qualitatively consistent with the main findings. These experiments investigate the effects of the data dimensions (Figure A1 and Table A3), of the proportion of hard trajectories (Figure A2 and Table A4), and evaluate the robustness of different methods against distribution shifts (Figure A3 and Table A5). Additionally, these experiments are replicated using synthetic data from an AR model with a static noise profile; see Figures A4–A8 and Tables A6–A10.

Furthermore, we conducted several experiments to investigate the performance of various implementations of our method. Figures A9–A13 and Tables A11–A15 focus on comparing alternative model selection approaches while applying the multiplicative conformity scores defined in (6). Figures A14–A18 and Tables A16–A20 summarize similar experiments based on the additive scores defined in (3).

4.3. Pedestrian Trajectories

We now apply the three methods to forecast pedestrian trajectories generated from the ORCA simulator (Van den Berg et al., 2008), which follow nonlinear dynamics and are intrinsically harder to predict than the synthetic trajectories discussed before. The data include 2-dimensional position measurements for 1,291 pedestrians, tracked over $T = 20$ time steps. To make the problem more challenging, we introduce dynamic noise to the trajectories of 10% of randomly selected pedestrians, making their paths more unpredictable. Figure 1 plots ten representative trajectories.

All trajectories are normalized as in the previous section, and we train the same LSTM for 50 epochs. In each experiment, the training and calibration sets use 1000 randomly chosen trajectories, and the test set consists of the remaining 291 trajectories. All results are averaged over 100 repetitions.

Figure 5 investigates the effect of varying the noise level, setting $\sigma_t^2 \propto t \cdot$noise level (varied from 1.5 to 5) for the hard trajectories and $\sigma_t^2 \propto t$ for the easy ones. Again, all methods attain 90% simultaneous marginal coverage, but CAFHT produces the most informative bands, with relatively narrow width and higher conditional coverage compared to NCTP. Meanwhile, CFRNN leads to very conservative bands, as

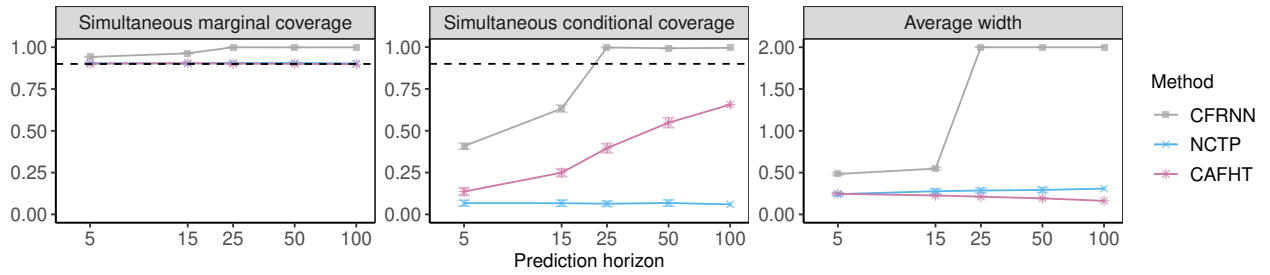

Figure 4. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the prediction horizon. Other details are as in Figure 3. For large prediction horizon, the CFRNN bands so wide as to be uninformative.

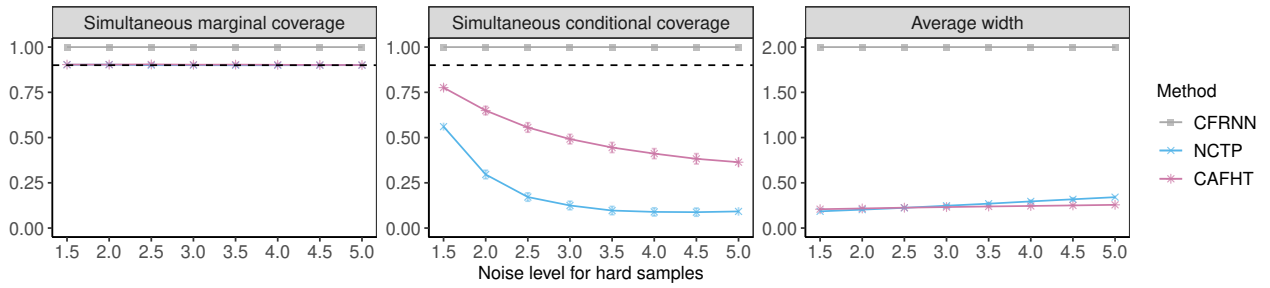

Figure 5. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the noise level controlling the intrinsic unpredictability of the more difficult trajectories. Note that the CFRNN bands so wide as to be uninformative.

in the previous section. See Table A21 in Appendix A5 for further details.

Additional numerical experiments are summarized in Appendix A5. Figure A19 and Table A22 investigate the effect of having a larger fraction of hard trajectories. Figure A20 and Table A23 compare the performances of different methods as a function of the sample size used for training and calibration. Figures A21–A24 and Tables A24–A27 perform a comparative analysis of different implementations of our methods under varying noise levels, using both multiplicative and additive conformity scores.

5. Discussion

This work opens several directions for future research. On the theoretical side, one may want to understand the conditions under which our method can asymptotically achieve *optimal* prediction bands in the limit of large sample sizes, potentially drawing inspiration from Lei et al. (2018b) and Sesia & Candès (2020). Moreover, there are several potential ways to further enhance our method and address some of its remaining limitations. For example, it could be adapted to provide even stronger types of coverage guarantees beyond those considered in this paper by conditioning on the calibration data or on some other observable features. Another possible direction is to study how to best reduce the algorithmic randomness caused by data splitting (Vovk, 2015), possibly using cross-conformal methods (Barber et al., 2021b) or E-value approaches (Bashari et al., 2024). Additionally, our method could be further improved by incorporating time dependency into the ACI learning rate or by relaxing the exchangeability assumption by leveraging weighted conformal inference ideas (Tibshirani et al., 2019). Lastly, it would be especially interesting to apply this method in real-world motion planning scenarios.

Software implementing the algorithms and data experiments are available online at `https://github.com/FionaZ3696/CAFHT.git`.

Acknowledgements

The authors thank anonymous referees for helpful comments, and the Center for Advanced Research Computing at the University of Southern California for providing computing resources. M. S. and Y. Z. were partly supported by NSF grant DMS 2210637. M. S. was also partly supported by an Amazon Research Award.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ajroldi, N., Diquigiovanni, J., Fontana, M., and Vantini, S. Conformal prediction bands for two-dimensional functional time series. *Computational Statistics & Data Analysis*, 187:107821, 2023.

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., and Savarese, S. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 961–971, 2016.

Angelopoulos, A., Candès, E., and Tibshirani, R. J. Conformal PID control for time series prediction. *Adv. Neural Inf. Process. Syst.*, 36, 2024.

Auer, A., Gauch, M., Klotz, D., and Hochreiter, S. Conformal prediction for time series with modern hopfield networks. In *Adv. Neural Inf. Process. Syst.*, 2023.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference*, 10(2):455–482, 2021a.

Barber, R. F., Candès, E. J., Ramdas, A., Tibshirani, R. J., et al. Predictive inference with the jackknife+. *Ann. Stat.*, 49(1):486–507, 2021b.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *Ann. Stat.*, 51(2):816–845, 2023.

Bashari, M., Epstein, A., Romano, Y., and Sesia, M. Derandomized novelty detection with FDR control via conformal e-values. *Adv. Neural Inf. Process. Syst.*, 36, 2024.

Bastani, O., Gupta, V., Jung, C., Noarov, G., Ramalingam, R., and Roth, A. Practical adversarial multivalid conformal prediction. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Adv. Neural Inf. Process. Syst.*, 2022.

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. Testing for outliers with conformal p-values. *Ann. Stat.*, 51(1):149 – 178, 2023.

Bhatnagar, A., Wang, H., Xiong, C., and Bai, Y. Improved online conformal prediction via strongly adaptive online learning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Cauchois, M., Gupta, S., and Duchi, J. C. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *J. Mach. Learn. Res.,*, 22 (1):3681–3722, 2021.

Chernozhukov, V., Wüthrich, K., and Yinchu, Z. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On learning theory*, pp. 732–749. PMLR, 2018.

Cleaveland, M., Lee, I., Pappas, G. J., and Lindemann, L. Conformal prediction regions for time series using linear complementarity programming. 38(19):20984–20992, 2024.

Dixit, A., Lindemann, L., Wei, S. X., Cleaveland, M., Pappas, G. J., and Burdick, J. W. Adaptive conformal prediction for motion planning among dynamic agents. In *Learning for Dynamics and Control Conference*, pp. 300–314. PMLR, 2023.

Einbinder, B.-S., Romano, Y., Sesia, M., and Zhou, Y. Training uncertainty-aware classifiers with conformalized deep learning. In *Adv. Neural Inf. Process. Syst.*, volume 35, 2022.

Feldman, S., Ringel, L., Bates, S., and Romano, Y. Achieving risk control in online learning settings. *arXiv preprint arXiv:2307.16895*, 2023.

Gibbs, I. and Candès, E. Adaptive conformal inference under distribution shift. *Adv. Neural Inf. Process. Syst.*, 34:1660–1672, 2021.

Gibbs, I. and Candès, E. Conformal inference for online prediction with arbitrary distribution shifts. *arXiv preprint arXiv:2208.08401*, 2022.

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

Izbicki, R., Shimizu, G., and Stern, R. Flexible distribution-free conditional predictive bands using density estimators. In *International Conference on Artificial Intelligence and Statistics*, pp. 3068–3077. PMLR, 2020.

Lei, J. and Wasserman, L. Distribution-free prediction bands for non-parametric regression. *J. R. Stat. Soc. (B)*, 76(1): 71–96, 2014.

Lei, J., Robins, J., and Wasserman, L. Distribution-free prediction sets. *J. Am. Stat. Assoc.*, 108(501):278–287, 2013.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.*, 113(523):1094–1111, 2018a.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.*, 113(523):1094–1111, 2018b.

Lekeufack, J., Angelopoulos, A. A., Bajcsy, A., Jordan, M. I., and Malik, J. Conformal decision theory: Safe autonomous decisions from imperfect predictions. *arXiv preprint arXiv:2310.05921*, 2023.

Liang, Z., Zhou, Y., and Sesia, M. Conformal inference is (almost) free for neural networks trained with early stopping. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Liang, Z., Sesia, M., and Sun, W. Integrative conformal p-values for out-of-distribution testing with labelled outliers. *J. R. Stat. Soc. (B)*, pp. qkad138, 2024.

Lin, Z., Trivedi, S., and Sun, J. Conformal prediction with temporal quantile adjustments. In *Adv. Neural Inf. Process. Syst.*, 2022.

Lindemann, L., Cleaveland, M., Shim, G., and Pappas, G. J. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics and Automation Letters*, 2023.

Marandon, A., Lei, L., Mary, D., and Roquain, E. Adaptive novelty detection with false discovery rate guarantee. *Ann. Stat.*, 52(1):157–183, 2024.

Nayakanti, N., Al-Rfou, R., Zhou, A., Goel, K., Refaat, K. S., and Sapp, B. Wayformer: Motion forecasting via simple & efficient attention networks. *arXiv preprint arXiv:2207.05844*, 2022.

Podkopaev, A. and Ramdas, A. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pp. 844–853. PMLR, 2021.

Qiu, H., Dobriban, E., and Tchetgen, E. Prediction sets adaptive to unknown covariate shift. *J. R. Stat. Soc. (B)*, 07 2023.

Romano, Y., Patterson, E., and Candès, E. J. Conformalized quantile regression. In *Adv. Neural Inf. Process. Syst.*, pp. 3538–3548, 2019.

Romano, Y., Barber, R. F., Sabatti, C., and Candès, E. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 4 2020a.

Romano, Y., Sesia, M., and Candès, E. J. Classification with valid and adaptive coverage. *Adv. Neural Inf. Process. Syst.*, 33, 2020b.

Sadinle, M., Lei, J., and Wasserman, L. Least ambiguous set-valued classifiers with bounded error levels. *J. Am. Stat. Assoc.*, 114(525):223–234, 2019.

Sesia, M. and Candès, E. J. A comparison of some conformal quantile regression methods. *Stat*, 9(1), 2020.

Sesia, M. and Romano, Y. Conformal prediction using conditional histograms. *Adv. Neural Inf. Process. Syst.*, 34, 2021.

Sesia, M., Favaro, S., and Dobriban, E. Conformal frequency estimation using discrete sketched data with coverage for distinct queries. *J. Mach. Learn. Res.*, 24(348): 1–80, 2023.

Sousa, M., Tomé, A. M., and Moreira, J. A general framework for multi-step ahead adaptive conformal heteroscedastic time series forecasting. *arXiv preprint arXiv:2207.14219*, 2022.

Stankeviciute, K., M Alaa, A., and van der Schaar, M. Conformal time-series forecasting. *Adv. Neural Inf. Process. Syst.*, 34:6216–6228, 2021.

Sun, S. H. and Yu, R. Copula conformal prediction for multi-step time series prediction. In *The Twelfth International Conference on Learning Representations*, 2023.

Tibshirani, R. J., Foygel Barber, R., Cand'es, E., and Ramdas, A. Conformal prediction under covariate shift. *Adv. Neural Inf. Process. Syst.*, 32, 2019.

Van den Berg, J., Lin, M., and Manocha, D. Reciprocal velocity obstacles for real-time multi-agent navigation. In *2008 IEEE international conference on robotics and automation*, pp. 1928–1935. IEEE, 2008.

Vovk, V. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25, pp. 475–490, 2012.

Vovk, V. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):9–28, 2015.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer, 2005.

Wei, S. X., Dixit, A., Tomar, S., and Burdick, J. W. Moving obstacle avoidance: A data-driven risk-aware approach. *IEEE Control Systems Letters*, 7:289–294, 2023.

Xu, C. and Xie, Y. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pp. 11559–11569. PMLR, 2021.

Xu, C. and Xie, Y. Conformal prediction for time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023a.

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

Xu, C., Vazquez, D. A. Z., Yao, R., Qiu, F., and Xie, Y. Wildfire modeling with point process and conformal prediction. *arXiv preprint arXiv:2207.13250*, 2022.

Xu, C., Xie, Y., Vazquez, D. A. Z., Yao, R., and Qiu, F. Spatio-temporal wildfire prediction using multi-modal data. *IEEE Journal on Selected Areas in Information Theory*, 2023a.

Xu, H., Mei, S., Bates, S., Taylor, J., and Tibshirani, R. Uncertainty intervals for prediction errors in time series forecasting. *arXiv preprint arXiv:2309.07435*, 2023b.

Yang, Y. and Kuchibhotla, A. K. Finite-sample efficient conformal prediction. *arXiv preprint arXiv:2104.13871*, 2021.

Yu, X., Zhao, Y., Yin, X., and Lindemann, L. Signal temporal logic control synthesis among uncontrollable dynamic agents with conformal prediction. *arXiv preprint arXiv:2312.04242*, 2023.

Zaffran, M., Dieuleveut, A., F'eron, O., Goude, Y., and Josse, J. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, 2022.

Zhou, Z., Wang, J., Li, Y.-H., and Huang, Y.-K. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17863–17873, 2023.

A1. Further Details on the ACI Algorithm

A1.1. Background on ACI

In this section, we briefly review some relevant components of the *adaptive conformal inference* (ACI) method introduced by Gibbs & Candès (2021) in the context of forecasting a single time series. The goal of ACI is to construct prediction bands in an online setting, while accounting for possible changes in the data distribution across different times. Specifically, ACI is designed to create prediction bands with a long-term average coverage guarantee. Intuitively, this guarantee means that, for an indefinitely long time series, a sufficiently large proportion of the series should be contained within the output band. This objective is notably distinct from the one investigated in our paper. However, since our method builds upon ACI, it can be useful to recall some relevant technical details of the latter method.

In the online learning setting considered by Gibbs & Candès (2021), one observes covariate-response pairs $\{(X_t, Y_t)\}_{t\in\mathbb{N}} \subset \mathbb{R}^d \times \mathbb{R}$ in a sequential fashion. At each time step $t \in \mathbb{N}$, the goal is to form a prediction set $\hat{C}_t$ for Y_t using the previously observed data $\{(X_r, Y_r)\}_{1\leq r\leq t-1}$ as well as the new covariates X_t. Given a target coverage level $\alpha \in (0, 1)$, the constructed prediction set should guarantee that, over long time, at least $100(1-\alpha)\%$ of the time Y_t lies within the set.

Recall that standard split-conformal prediction methods require a calibration dataset $\mathcal{D}_{\text{cal}} \subseteq \{(X_r, Y_r)\}_{1\leq r\leq t-1}$ that is independent of the data used to fit the regression model. The standard approach involves constructing a prediction set as $\hat{C}_t(\alpha) = \{y : S(X_t, y) \leq \hat{Q}(1-\alpha)\}$, where $S(X_t, y)$ is a score that measures how well y conforms with the prediction of the fitted model. For example, if we denote the fitted model as $\hat{g}$, a classical example of scoring function would be $S(X_t, y) = |\hat{g}(X_t) - y|$. Then, in general, the score $S(X_t, y)$ is compared to a suitable empirical quantile, $\hat{Q}(1-\alpha)$, of the analogous scores evaluated on the calibration data: $\hat{Q}(1-\alpha) = \inf\{s : (|\mathcal{D}_{\text{cal}}|^{-1} \sum_{(X_r,Y_r)\in\mathcal{D}_{\text{cal}}} \mathbb{1}_{\{S(X_r,Y_r)\leq s\}}) \geq 1-\alpha\}$. If the observations taken at different times are not exchangeable with one another, however, standard conformal prediction algorithms cannot achieve valid coverage. This is where ACI comes into play.

The core concept of ACI involves dynamically updating the functions $\hat{g}$, $S(\cdot)$, and $\hat{Q}(\cdot)$ at each time step, utilizing newly acquired data. Concurrently, ACI modifies the nominal miscoverage target level α_t of its conformal predictor for each time increment. The purpose of adjusting the α level at each time step is to calibrate future predictions to be more or less conservative depending on their empirical performance in covering past values of the time series. For instance, if a prediction band is found to be excessively broad, it will be narrowed in subsequent steps, and the opposite applies if it's too narrow. This strategy enables ACI to continuously adapt to potential dependencies and distribution changes within the time series, maintaining relevance and accuracy in an online context. Specifically, ACI employs the following α-update rule:

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \text{err}_t),$$

where

$$\text{err}_t = \begin{cases} 1, & \text{if } Y_t \notin \hat{C}_t^{\text{ACI}}(\alpha_t), \\ 0, & \text{otherwise.} \end{cases},$$

and $\hat{C}_t^{\text{ACI}}(\alpha_t) = \{y : S_t(X_t, y) \leq \hat{Q}_t(1-\alpha_t)\}$. Equivalently,

$$\hat{C}_t^{\text{ACI}}(\alpha_t) = [\hat{\ell}_t^{\text{ACI}}, \hat{u}_t^{\text{ACI}}] = [\hat{g}(X_t) - \hat{Q}_t(1-\alpha_t), \hat{g}(X_t) + \hat{Q}_t(1-\alpha_t)].$$

The hyperparameter $\gamma > 0$ controls the magnitude of each update step. Intuitively, a larger γ means that ACI can rapidly adjust to observed changes in the data distribution. However, this may come at the expense of increased instability in the prediction bands. Consequently, the ideal value of γ tends to be specific to the application at hand, requiring careful consideration to balance responsiveness and stability. This is why our CAFHT method involves a data-driven parameter tuning component.

The main theoretical finding established by Gibbs & Candès (2021) is that ACI always attains valid long-term average coverage. Notably, this result is achieved without the necessity for any assumptions regarding the distribution of the unique time series in question. More precisely, with probability one,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \text{err}_t - \alpha \right| \leq \frac{\max(\alpha_1, 1-\alpha_1) + \gamma}{T\gamma},$$

which implies

$$\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} \text{err}_t \overset{\text{a.s.}}{=} \alpha.$$

This result is not essential for proving our simultaneous marginal coverage guarantee, but it offers an intuitive rationale for our methodology. Indeed, the capacity of ACI to adaptively encompass the inherent variability in each time series is key to our method's enhanced conditional coverage compared to other conformal prediction approaches for multi-series forecasting. Further, our method inherits the same long-term average coverage property of ACI because it can only expand the prediction bands of the latter.

In this paper, we implement ACI without re-training the forecasting model $\hat{g}$ at each step. This approach is viable due to our access to additional "training" time series data from the same population, and it aids in diminishing the computational cost of our numerical experiments. Nonetheless, our methodology is flexible enough to incorporate ACI with periodic re-training, aligning with the practices suggested by Gibbs & Candès (2021) and the very recent related conformal PID method of Angelopoulos et al. (2024).

A1.2. Warm Starts

As originally designed, ACI primarily aimed at achieving asymptotic coverage in the limit of a very long trajectory, sometimes tolerating very narrow prediction intervals in the initial time steps. However, we have observed that this behavior can negatively impact the performance of our method in finite-horizon scenarios. To address this issue, we introduce in this paper a simple warm-start approach for ACI. This involves incorporating artificial conformity scores at the start of each trajectory. These scores are generated as uniform random noise, with values falling within the range of observed residuals in the training dataset. Consequently, ACI typically begins with a wider interval for its first forecast. Importantly, this modification does not affect the long-term asymptotic properties of ACI when applied to a single trajectory, nor does it impact our guarantee of finite-sample simultaneous marginal coverage. However, it often results in more informative (narrower) prediction bands.

The solution described above is applied in our experiments using 5 warm-start scores, denoted as $\hat{\epsilon}_{-4}, \ldots, \hat{\epsilon}_0$, and setting the initial value of α_{-4} equal to 0.1. A similar warm-start approach is also utilized when we apply the conformal PID algorithm of Angelopoulos et al. (2024) instead of ACI. However, for the algorithm the warm start simply consists of setting the initial quantile q_0 equal to the $(1-\alpha)$-th quantile evaluated on the empirical distribution of scores computed using the training set.

A2. Proof of Theorem 1

Proof of Theorem 1. The proof follows directly from the exchangeability of the conformity scores, as it is often the case for split-conformal prediction methods. Denote $\hat{\epsilon}_{n+1}(\gamma)$ the conformity score of the test trajectory $\boldsymbol{Y}^{(t+1)}$ evaluated using the ACI prediction band constructed with step size γ. For any fixed α and $\gamma > 0$, we have that $Y_t^{(n+1)} \in \hat{C}_t^{(n+1)} \forall t \in [T]$ if and only if $\hat{\epsilon}_{n+1}(\gamma) \leq \hat{Q}(1-\alpha, \gamma)$, where $\hat{Q}(1-\alpha, \gamma)$ is the $\lceil (1-\alpha)(1+|\mathcal{D}_{\text{cal}}|)\rceil$-th smallest value of $\hat{\epsilon}_i(\gamma)$ for all $i \in \mathcal{D}_{\text{cal}}$. Since the test trajectory is exchangeable with $\mathcal{D}_{\text{cal}}$, its score $\hat{\epsilon}_{n+1}(\gamma)$ is also exchangeable with $\{\hat{\epsilon}_i(\gamma), i \in \mathcal{D}_{\text{cal}}\}$. Then by Lemma 1 in Romano et al. (2019), it follows that $\mathbb{P}(Y_t^{(n+1)} \in \hat{C}_t^{(n+1)} \forall t \in [T]) = \mathbb{P}(\hat{\epsilon}_{n+1}(\gamma) \leq \hat{Q}(1-\alpha, \gamma)) \geq 1-\alpha$. □

A3. Algorithms

Algorithm A1 Model selection component of CAFHT

Input: A pre-trained forecaster $\hat{g}$ producing one-step-ahead predictions; calibration trajectories $\mathcal{D}^1_{\text{cal}}$; a grid of candidate learning rates $\{\gamma_1, \ldots, \gamma_L\}$.
for $\ell \in [L]$ **do**
  Construct $\hat{C}^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \gamma_\ell)$ using ACI, for $i \in \mathcal{D}^1_{\text{cal}}$.
  Evaluate $\hat{\epsilon}_i(\gamma_\ell)$ using (3), for $i \in \mathcal{D}^1_{\text{cal}}$.
  Compute $\hat{Q}(1-\alpha, \gamma_\ell)$, the $(1-\alpha)(1+1/|\mathcal{D}^1_{\text{cal}}|)$-th quantile of $\{\hat{\epsilon}_i(\gamma_\ell), i \in \mathcal{D}^1_{\text{cal}}\}$.
  Construct $\hat{C}(\boldsymbol{Y}^{(i)}, \gamma_\ell) = (\hat{C}_1(\boldsymbol{Y}^{(i)}, \gamma_\ell), \ldots, \hat{C}_T(\boldsymbol{Y}^{(i)}, \gamma_\ell))$ using (4) for $i \in \mathcal{D}^1_{\text{cal}}$.
end for
Pick $\hat{\gamma}$ such that,

$$\hat{\gamma} := \underset{\ell \in [L]}{\arg\min}\, \text{AvgWidth}(C(\boldsymbol{Y}^{(i)}, \gamma_\ell)). \tag{A8}$$

Output: Selected learning rate parameter $\hat{\gamma}$.

Algorithm A2 CAFHT - multiplicative scores

Input: A pre-trained forecaster $\hat{g}$ producing one-step-ahead predictions; calibration trajectories $\mathcal{D}_{\text{cal}}$; the initial position $Y_0^{(n+1)}$ of a test trajectory $\boldsymbol{Y}^{(n+1)}$; the desired nominal level $\alpha \in (0,1)$; a grid of candidate learning rates $\{\gamma_1, \ldots, \gamma_L\}$.
Randomly split $\mathcal{D}_{\text{cal}}$ into $\mathcal{D}^1_{\text{cal}}$ and $\mathcal{D}^2_{\text{cal}}$.
Select a learning rate $\hat{\gamma} \in \{\gamma_1, \ldots, \gamma_L\}$, applying Algorithm A3 using the trajectory data in $\mathcal{D}^1_{\text{cal}}$.
Construct $\hat{C}^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \hat{\gamma})$ using ACI, for $i \in \mathcal{D}^2_{\text{cal}}$.
Evaluate $\hat{\epsilon}_i(\hat{\gamma})$ using (6), for $i \in \mathcal{D}^2_{\text{cal}}$.
Compute the empirical quantile $\hat{Q}(1-\alpha, \hat{\gamma})$.
for $t \in [T]$ **do**
  Compute $\hat{C}^{\text{ACI}}_t(\boldsymbol{Y}^{(n+1)}, \hat{\gamma})$ with ACI, using the past of the test trajectory $(Y_0^{(n+1)}, Y_1^{(n+1)}, \ldots, Y_{t-1}^{(n+1)})$.
  Compute a prediction interval $\hat{C}_t(\boldsymbol{Y}^{(n+1)}, \hat{\gamma})$ for the next step, using the multiplicative version of (4).
  Observe the next step of the trajectory, $Y_t^{(n+1)}$.
end for
Output: An online prediction band $\hat{C}(\boldsymbol{Y}^{(n+1)})$.

Algorithm A3 Model selection component of CAFHT - multiplicative scores

Input: A pre-trained forecaster $\hat{g}$ producing one-step-ahead predictions; calibration trajectories $\mathcal{D}^1_{\text{cal}}$; a grid of candidate learning rates $\{\gamma_1, \ldots, \gamma_L\}$.
for $\ell \in [L]$ **do**
  Construct $\hat{C}^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \gamma_\ell)$ using ACI, for $i \in \mathcal{D}^1_{\text{cal}}$.
  Evaluate $\hat{\epsilon}_i(\gamma_\ell)$ using (6), for $i \in \mathcal{D}^1_{\text{cal}}$.
  Compute $\hat{Q}(1-\alpha, \gamma_\ell)$, the $(1-\alpha)(1+1/|\mathcal{D}^1_{\text{cal}}|)$-th quantile of $\{\hat{\epsilon}_i(\gamma_\ell), i \in \mathcal{D}^1_{\text{cal}}\}$.
  Construct $\hat{C}(\boldsymbol{Y}^{(i)}, \gamma_\ell) = (\hat{C}_1(\boldsymbol{Y}^{(i)}, \gamma_\ell), \ldots, \hat{C}_T(\boldsymbol{Y}^{(i)}, \gamma_\ell))$ for $i \in \mathcal{D}^1_{\text{cal}}$, using the multiplicative version of (4).
end for
Pick $\hat{\gamma}$ such that,

$$\hat{\gamma} := \underset{\ell \in [L]}{\arg\min}\, \text{AvgWidth}(C(\boldsymbol{Y}^{(i)}, \gamma_\ell)). \tag{A9}$$

Output: Selected learning rate parameter $\hat{\gamma}$.

A4. Parameter Tuning for CAFHT Without Data Splitting

Here, we outline an alternate implementation of CAFHT which, in contrast to the primary method described in Section 3.4, obviates the need for additional subdivision of the calibration data in $\mathcal{D}_{\text{cal}}$ for selecting an optimal value of the ACI learning rate parameter γ. In essence, this version of CAFHT employs the same calibration dataset $\mathcal{D}_{\text{cal}}$ for both choosing $\hat{\gamma}$ and calibrating the conformal margin of error via $\hat{Q}(1-\alpha', \hat{\gamma})$. It does so by using a judiciously selected $\alpha' < \alpha$ to compensate for the selection step. Enabled by the theoretical results of Yang & Kuchibhotla (2021) and Liang et al. (2023), this method is outlined below by Algorithms A4–A5 using additive conformity scores, and by Algorithms A6–A7 using multiplicative conformity scores.

In the following, we will assume that the goal is for CAFHT to select a good $\hat{\gamma}$ from a list of L candidate parameter values, $\gamma_1, \ldots, \gamma_L$, for some fixed integer $L \geq 1$.

Using the DKW inequality, Yang & Kuchibhotla (2021) proves that, when calibrating at the nominal level α, a conformal prediction set $\hat{C}^{(n+1)}$ constructed after using the same calibration set $\mathcal{D}_{\text{cal}}$ to select the best model among L candidates may have an inflated coverage rate in the following form:

$$\mathbb{P}(Y^{(n+1)} \in \hat{C}^{(n+1)}) \geq \left(1 + \frac{1}{|\mathcal{D}_{\text{cal}}|}\right)(1-\alpha) - \frac{\sqrt{\log(2L)/2} + c(L)}{\sqrt{|\mathcal{D}_{\text{cal}}|}}, \tag{A10}$$

where $c(L)$ is a constant that is generally smaller than $1/3$ and can be computed explicitly,

$$c(L) = \frac{\sqrt{2}Le^{-\log(2L)}}{\sqrt{\log(2L)} + \sqrt{\log(2L) + 4/\pi}}.$$

This justifies applying CAFHT, without data splitting, using $\hat{Q}(1-\alpha'_{\text{DKW}}, \hat{\gamma})$ instead of $\hat{Q}(1-\alpha, \hat{\gamma})$, where

$$\alpha'_{\text{DKW}} = 1 - \frac{1-\alpha+\text{err}}{1+1/|\mathcal{D}_{\text{cal}}|}, \qquad \text{err} = \frac{\sqrt{\log(2L)/2} + c(L)}{\sqrt{|\mathcal{D}_{\text{cal}}|}}.$$

A further refinement of this approach was proposed by Liang et al. (2023), which suggested instead using

$$\alpha' = \max\{\alpha'_{\text{Markov}}, \alpha'_{\text{DKW}}\}, \tag{A11}$$

where α'_{Markov} is computed as follows. By combining the results of Vovk (2012) with Markov's inequality, Liang et al. (2023) proved the following inequality in the same context of (A10):

$$\mathbb{P}(Y^{(n+1)} \in \hat{C}^{(n+1)}) \geq I^{-1}\left(\frac{1}{bL}; |\mathcal{D}_{\text{cal}}| + 1 - l, l\right) \cdot (1 - 1/b), \tag{A12}$$

where $I^{-1}(x; |\mathcal{D}_{\text{cal}}| + 1 - l, l)$ is the inverse Beta cumulative distribution function with $l = \lfloor \alpha(|\mathcal{D}_{\text{cal}}| + 1) \rfloor$, and $b > 1$ is any fixed constant. Therefore, the desired value of α'_{Markov} can be calculated by inverting (A12) numerically, with the choice of $b = 100$ recommended by Liang et al. (2023). In particular, we generate a grid of $\hat{\alpha}$ candidates, evaluate the Markov lower bounds associated with each $\hat{\alpha}$, and then return the largest possible $\hat{\alpha}$ such that its Markov bound is greater than $1-\alpha$.

A potential advantage of the bound in (A12) relative to (A10) is that the $[\sqrt{\log(2L)/2} + c(L)]/\sqrt{|\mathcal{D}_{\text{cal}}|}$ term in the latter does not depend on α. That makes (A10) sometimes too conservative when α is small; see Appendix A1.2 in Liang et al. (2023). However, neither bound always dominates the other, hence why we adaptively follow the tighter one using (A11).

The performance of CAFHT applied without data splitting, relying instead on the theoretical correction for parameter tuning described above, is investigated empirically in Appendix A5.

Algorithm A4 CAFHT (theory)

Input: A pre-trained forecaster $\hat{g}$ producing one-step-ahead predictions; calibration trajectories $\mathcal{D}_{\text{cal}}$; the initial position $Y_0^{(n+1)}$ of a test trajectory $\boldsymbol{Y}^{(n+1)}$; the desired nominal level $\alpha \in (0,1)$; a grid of candidate learning rates $\{\gamma_1, \ldots, \gamma_L\}$.

Select a learning rate $\hat{\gamma} \in \{\gamma_1, \ldots, \gamma_L\}$, applying Algorithm A5 using the trajectory data in $\mathcal{D}_{\text{cal}}$.
Construct $\hat{C}^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \hat{\gamma})$ using ACI, for $i \in \mathcal{D}_{\text{cal}}$.
Evaluate $\hat{\epsilon}_i(\hat{\gamma})$ using (6), for $i \in \mathcal{D}_{\text{cal}}$.
Compute the empirical quantile $\hat{Q}(1-\alpha', \hat{\gamma})$, where α' is defined in (A11).
for $t \in [T]$ **do**
    Compute $\hat{C}_t^{\text{ACI}}(\boldsymbol{Y}^{(n+1)}, \hat{\gamma})$ with ACI, using the past of the test trajectory $(Y_0^{(n+1)}, Y_1^{(n+1)}, \ldots, Y_{t-1}^{(n+1)})$.
    Compute a prediction interval $\hat{C}_t(\boldsymbol{Y}^{(n+1)}, \hat{\gamma})$ for the next step, using (4) with $\hat{Q}(1-\alpha', \hat{\gamma})$.
    Observe the next step of the trajectory, $Y_t^{(n+1)}$.
end for
Output: An online prediction band $\hat{C}(\boldsymbol{Y}^{(n+1)})$.

Algorithm A5 Model selection component of CAFHT (theory)

Input: A pre-trained forecaster $\hat{g}$ producing one-step-ahead predictions; calibration trajectories $\mathcal{D}_{\text{cal}}$; a grid of candidate learning rates $\{\gamma_1, \ldots, \gamma_L\}$.
for $\ell \in [L]$ **do**
    Construct $\hat{C}^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \gamma_\ell)$ using ACI, for $i \in \mathcal{D}_{\text{cal}}$.
    Evaluate $\hat{\epsilon}_i(\gamma_\ell)$ using (3), for $i \in \mathcal{D}_{\text{cal}}$.
    Compute $\hat{Q}(1-\alpha', \gamma_\ell)$, the $(1-\alpha')(1+1/|\mathcal{D}_{\text{cal}}|)$-th smallest value of $\{\hat{\epsilon}_i(\gamma_\ell), i \in \mathcal{D}_{\text{cal}}\}$, where α' is defined in (A11).
    Construct $\hat{C}(\boldsymbol{Y}^{(i)}, \gamma_\ell) = (\hat{C}_1(\boldsymbol{Y}^{(i)}, \gamma_\ell), \ldots, \hat{C}_T(\boldsymbol{Y}^{(i)}, \gamma_\ell))$ using (4) for $i \in \mathcal{D}_{\text{cal}}$.
end for
Pick $\hat{\gamma}$ such that,

$$\hat{\gamma} := \underset{\ell \in [L]}{\arg\min}\, \text{AvgWidth}(C(\boldsymbol{Y}^{(i)}, \gamma_\ell)). \tag{A13}$$

Output: Selected learning rate parameter $\hat{\gamma}$.

Algorithm A6 CAFHT (theory) - multiplicative scores

Input: A pre-trained forecaster $\hat{g}$ producing one-step-ahead predictions; calibration trajectories $\mathcal{D}_{\text{cal}}$; the initial position $Y_0^{(n+1)}$ of a test trajectory $\boldsymbol{Y}^{(n+1)}$; the desired nominal level $\alpha \in (0,1)$; a grid of candidate learning rates $\{\gamma_1, \ldots, \gamma_L\}$.

Select a learning rate $\hat{\gamma} \in \{\gamma_1, \ldots, \gamma_L\}$, applying Algorithm A7 using the trajectory data in $\mathcal{D}_{\text{cal}}$.
Construct $\hat{C}^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \hat{\gamma})$ using ACI, for $i \in \mathcal{D}_{\text{cal}}$.
Evaluate $\hat{\epsilon}_i(\hat{\gamma})$ using the multiplicative version of (4), for $i \in \mathcal{D}_{\text{cal}}$.
Compute the empirical quantile $\hat{Q}(1-\alpha', \hat{\gamma})$, where α' is defined in (A11).
for $t \in [T]$ **do**
    Compute $\hat{C}_t^{\text{ACI}}(\boldsymbol{Y}^{(n+1)}, \hat{\gamma})$ with ACI, using the past of the test trajectory $(Y_0^{(n+1)}, Y_1^{(n+1)}, \ldots, Y_{t-1}^{(n+1)})$.
    Compute $\hat{C}_t(\boldsymbol{Y}^{(n+1)}, \hat{\gamma})$ for the next step, using the multiplicative version of (4) with $\hat{Q}(1-\alpha', \hat{\gamma})$.
    Observe the next step of the trajectory, $Y_t^{(n+1)}$.
end for
Output: An online prediction band $\hat{C}(\boldsymbol{Y}^{(n+1)})$.

Algorithm A7 Model selection component of CAFHT (theory) - multiplicative scores

Input: A pre-trained forecaster $\hat{g}$ producing one-step-ahead predictions; calibration trajectories $\mathcal{D}_{\text{cal}}$; a grid of candidate learning rates $\{\gamma_1, \ldots, \gamma_L\}$.
for $\ell \in [L]$ **do**
  Construct $\hat{C}^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \gamma_\ell)$ using ACI, for $i \in \mathcal{D}_{\text{cal}}$.
  Evaluate $\hat{\epsilon}_i(\gamma_\ell)$ using the multiplicative version of (4), for $i \in \mathcal{D}_{\text{cal}}$.
  Compute $\hat{Q}(1-\alpha', \gamma_\ell)$, the $(1-\alpha')(1+1/|\mathcal{D}_{\text{cal}}|)$-th quantile of $\{\hat{\epsilon}_i(\gamma_\ell), i \in \mathcal{D}_{\text{cal}}\}$, where α' is defined in (A11).
  Construct $\hat{C}(\boldsymbol{Y}^{(i)}, \gamma_\ell) = (\hat{C}_1(\boldsymbol{Y}^{(i)}, \gamma_\ell), \ldots, \hat{C}_T(\boldsymbol{Y}^{(i)}, \gamma_\ell))$ for $i \in \mathcal{D}_{\text{cal}}$.
end for
Pick $\hat{\gamma}$ such that,

$$\hat{\gamma} := \underset{\ell \in [L]}{\arg\min}\, \text{AvgWidth}(C(\boldsymbol{Y}^{(i)}, \gamma_\ell)). \tag{A14}$$

Output: Selected learning rate parameter $\hat{\gamma}$.

A5. Additional Experimental Results

A5.1. Synthetic Data

A5.1.1. Main Results — Comparing CAFHT to CFRNN and NCTP

AR data with dynamic noise profile. Firstly, we investigate the performance of the three methods considered in this paper, namely CAFHT, CFRNN, and NCTP, using synthetic data from an AR model with dynamic noise profile. The default settings of the experiments are as described in Section 4, but this appendix contains more detailed results.

Figure 3 and Table A1 report on the average performance on simulated heterogeneous trajectories of prediction bands constructed by different methods as a function of the total number of training and calibration trajectories. The number of trajectories is varied between 200 and 10,000. All methods achieve 90% simultaneous marginal coverage. As discussed earlier in Section 4, these results show that our method (CAFHT) leads to more informative bands with lower average width and higher conditional coverage.

Figure 4 and Table A2 show the performance of prediction bands constructed by different methods, as a function of the prediction horizon, which is varied between 5 and 100. As the prediction horizon increases, the CFRNN method becomes more and more conservative, while the CAFHT method can consistently produce small predicting bands while maintaining relatively high conditional coverage.

Figure A1 and Table A3 report on the performance of all methods as a function of the dimensionality of the trajectories, which is varied between 1 and 10. Again, the results show that the CAFHT method leads to more informative bands with lower average width and higher conditional coverage.

Figure A2 and Table A4 report on the performances of these methods as a function of the proportion $\delta \in [0, 1]$ of hard trajectories in the population. We assess these results at δ values of 0.1, 0.2, and 0.5. It is observed that when the dataset contains a small number of hard-to-predict trajectories, the CAFHT method achieves superior conditional coverage and yields a narrower prediction band compared to the NCTP method. As the fraction of difficult-to-predict trajectories increases, the performance of NCTP improves (there would be no heterogeneity issue if all trajectories were "hard to predict"). Nonetheless, the CAFHT method consistently produces the narrowest, and thereby the most informative, prediction bands across the range of δ values considered.

Finally, Figure A3 and Table A5 investigate the robustness of all methods to distribution shifts. To this end, we kept the proportion of difficult-to-predict trajectories at 0.1 in both the training and calibration datasets, but varied this proportion in the test dataset, altering δ from 0.2 to 0.9 in the test set. Under these circumstances, as the calibration set and test set are not exchangeable, no method can ensure marginal coverage at the intended 90% level. However, as shown in Figure A3 and Table A5, CAFHT, in practice, tends to achieve higher marginal coverage compared to NCTP. This is consistent with the fact that CAFHT typically leads to higher conditional coverage in the absence of distribution shifts (Einbinder et al., 2022). Additionally, the increasing width of the CAFHT prediction bands as the strength of the distribution shift grows demonstrates its enhanced ability to accurately measure predictive uncertainty.

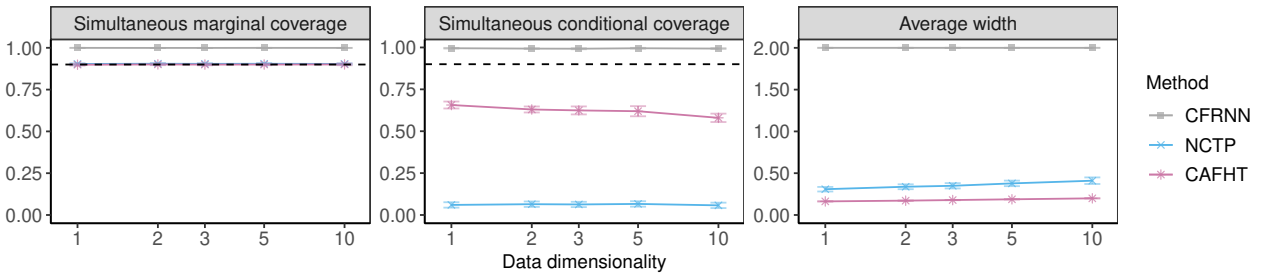

Figure A1. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the data dimensionality. See Table A3 for more detailed results and standard errors.

Table A1. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure 3.

			Simultaneous coverage		
Sample size	Method	Average width	Conditional-hard	Conditional-easy	Marginal
200					
200	CFRNN	2.000 (0.000)	0.939 (0.007)	1.000 (0.000)	0.994 (0.001)
200	NCTP	0.687 (0.035)	0.260 (0.024)	0.992 (0.002)	0.919 (0.004)
200	CAFHT	0.202 (0.003)	0.704 (0.017)	0.944 (0.005)	0.920 (0.005)
500					
500	CFRNN	2.000 (0.000)	0.969 (0.005)	1.000 (0.000)	0.997 (0.000)
500	NCTP	0.467 (0.021)	0.196 (0.019)	0.994 (0.002)	0.916 (0.003)
500	CAFHT	0.182 (0.002)	0.682 (0.016)	0.934 (0.003)	0.910 (0.004)
1000					
1000	CFRNN	2.000 (0.000)	0.992 (0.002)	1.000 (0.000)	0.999 (0.000)
1000	NCTP	0.343 (0.018)	0.093 (0.012)	0.992 (0.001)	0.901 (0.003)
1000	CAFHT	0.174 (0.001)	0.679 (0.012)	0.934 (0.003)	0.908 (0.003)
2000					
2000	CFRNN	2.000 (0.000)	0.995 (0.001)	1.000 (0.000)	1.000 (0.000)
2000	NCTP	0.308 (0.014)	0.060 (0.008)	0.996 (0.001)	0.903 (0.002)
2000	CAFHT	0.163 (0.001)	0.656 (0.010)	0.926 (0.002)	0.899 (0.003)
5000					
5000	CFRNN	2.000 (0.000)	0.998 (0.001)	1.000 (0.000)	1.000 (0.000)
5000	NCTP	0.244 (0.013)	0.033 (0.006)	0.997 (0.001)	0.900 (0.002)
5000	CAFHT	0.158 (0.001)	0.655 (0.007)	0.925 (0.002)	0.899 (0.002)
10000					
10000	CFRNN	2.000 (0.000)	0.997 (0.002)	1.000 (0.000)	1.000 (0.000)
10000	NCTP	0.239 (0.033)	0.041 (0.015)	0.997 (0.001)	0.903 (0.003)
10000	CAFHT	0.151 (0.002)	0.675 (0.022)	0.931 (0.004)	0.906 (0.005)

Table A2. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the prediction horizon. The red numbers indicate smaller prediction bands or higher conditional coverage. See corresponding plot in Figure 4.

			Simultaneous coverage		
Prediction horizon	Method	Average width	Conditional-hard	Conditional-easy	Marginal
5					
5	CFRNN	0.484 (0.007)	0.408 (0.009)	1.000 (0.000)	0.942 (0.001)
5	NCTP	0.242 (0.011)	0.067 (0.009)	0.996 (0.001)	0.904 (0.002)
5	CAFHT	0.247 (0.005)	0.136 (0.011)	0.985 (0.001)	0.902 (0.003)
15					
15	CFRNN	0.548 (0.007)	0.632 (0.011)	1.000 (0.000)	0.964 (0.001)
15	NCTP	0.277 (0.014)	0.067 (0.009)	0.995 (0.001)	0.903 (0.002)
15	CAFHT	0.227 (0.003)	0.249 (0.012)	0.976 (0.001)	0.904 (0.002)
25					
25	CFRNN	2.000 (0.000)	0.998 (0.001)	1.000 (0.000)	1.000 (0.000)
25	NCTP	0.286 (0.014)	0.064 (0.008)	0.997 (0.001)	0.906 (0.002)
25	CAFHT	0.212 (0.002)	0.396 (0.014)	0.957 (0.002)	0.902 (0.003)
50					
50	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
50	NCTP	0.293 (0.015)	0.068 (0.010)	0.997 (0.001)	0.906 (0.002)
50	CAFHT	0.192 (0.002)	0.548 (0.015)	0.939 (0.002)	0.901 (0.002)
100					
100	CFRNN	2.000 (0.000)	0.995 (0.001)	1.000 (0.000)	1.000 (0.000)
100	NCTP	0.308 (0.014)	0.060 (0.008)	0.996 (0.001)	0.903 (0.002)
100	CAFHT	0.163 (0.001)	0.656 (0.010)	0.926 (0.002)	0.899 (0.003)

Table A3. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the data dimensionality. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A1.

				Simultaneous coverage		
Data dimensionality		Method	Average width	Conditional-hard	Conditional-easy	Marginal
1						
	1	CFRNN	2.000 (0.000)	0.995 (0.001)	1.000 (0.000)	1.000 (0.000)
	1	NCTP	0.308 (0.014)	0.060 (0.008)	0.996 (0.001)	0.903 (0.002)
	1	CAFHT	0.163 (0.001)	0.656 (0.010)	0.926 (0.002)	0.899 (0.003)
2						
	2	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
	2	NCTP	0.338 (0.015)	0.064 (0.008)	0.996 (0.001)	0.905 (0.002)
	2	CAFHT	0.172 (0.001)	0.630 (0.009)	0.930 (0.002)	0.901 (0.002)
3						
	3	CFRNN	2.000 (0.000)	0.993 (0.002)	1.000 (0.000)	0.999 (0.000)
	3	NCTP	0.349 (0.016)	0.063 (0.008)	0.996 (0.001)	0.904 (0.002)
	3	CAFHT	0.179 (0.001)	0.624 (0.012)	0.930 (0.002)	0.900 (0.002)
5						
	5	CFRNN	2.000 (0.000)	0.995 (0.001)	1.000 (0.000)	1.000 (0.000)
	5	NCTP	0.378 (0.017)	0.066 (0.008)	0.996 (0.001)	0.904 (0.002)
	5	CAFHT	0.188 (0.001)	0.619 (0.015)	0.931 (0.002)	0.900 (0.002)
10						
	10	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
	10	NCTP	0.410 (0.020)	0.057 (0.008)	0.996 (0.001)	0.903 (0.002)
	10	CAFHT	0.199 (0.001)	0.580 (0.013)	0.936 (0.002)	0.901 (0.003)

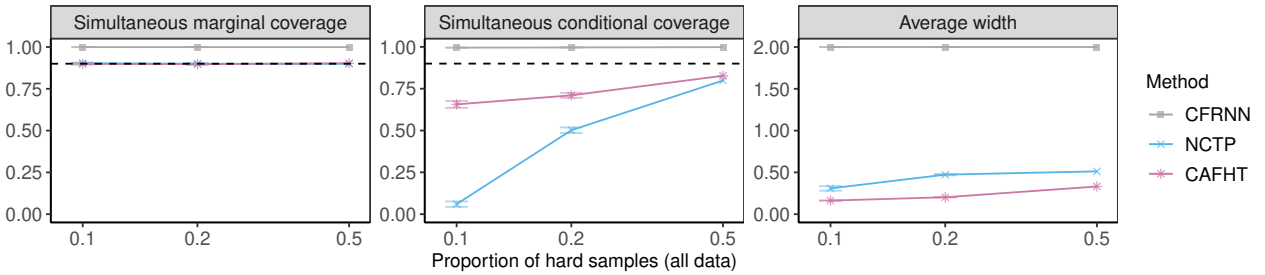

Figure A2. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the proportion of hard-to-predict trajectories. See Table A4 for more detailed results and standard errors.

Table A4. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the proportion of hard-to-predict trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A2.

				Simultaneous coverage		
Proportion of hard samples (all data)		Method	Average width	Conditional-hard	Conditional-easy	Marginal
0.1						
	0.1	CFRNN	2.000 (0.000)	0.995 (0.001)	1.000 (0.000)	1.000 (0.000)
	0.1	NCTP	0.308 (0.014)	0.060 (0.008)	0.996 (0.001)	0.903 (0.002)
	0.1	CAFHT	0.163 (0.001)	0.656 (0.010)	0.926 (0.002)	0.899 (0.003)
0.2						
	0.2	CFRNN	2.000 (0.000)	0.997 (0.001)	1.000 (0.000)	0.999 (0.000)
	0.2	NCTP	0.473 (0.004)	0.502 (0.008)	1.000 (0.000)	0.900 (0.002)
	0.2	CAFHT	0.203 (0.001)	0.710 (0.007)	0.944 (0.002)	0.897 (0.003)
0.5						
	0.5	CFRNN	2.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.999 (0.000)
	0.5	NCTP	0.512 (0.004)	0.799 (0.004)	1.000 (0.000)	0.899 (0.002)
	0.5	CAFHT	0.331 (0.002)	0.827 (0.005)	0.978 (0.001)	0.903 (0.003)

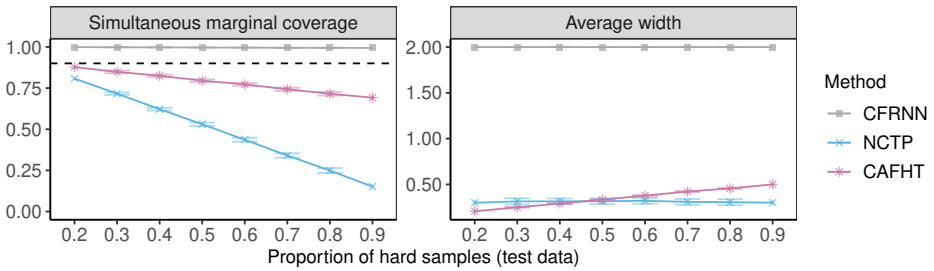

Figure A3. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods under distributional shift. The results are shown as a function of the proportion of hard-to-predict trajectories in the test data. See Table A5 for more detailed results and standard errors.

Table A5. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods under distributional shift. The results are shown as a function of the proportion of hard-to-predict trajectories in the test data. The red numbers indicate higher marginal coverage. See the corresponding plot in Figure A3.

Proportion of hard samples (test data)	Method	Length	Marginal coverage
0.2			
0.2	CFRNN	2.000 (0.000)	0.999 (0.000)
0.2	NCTP	0.300 (0.016)	0.808 (0.003)
0.2	CAFHT	0.207 (0.002)	0.877 (0.003)
0.3			
0.3	CFRNN	2.000 (0.000)	0.998 (0.000)
0.3	NCTP	0.315 (0.017)	0.716 (0.004)
0.3	CAFHT	0.250 (0.002)	0.849 (0.004)
0.4			
0.4	CFRNN	2.000 (0.000)	0.997 (0.000)
0.4	NCTP	0.315 (0.016)	0.622 (0.004)
0.4	CAFHT	0.291 (0.003)	0.824 (0.004)
0.5			
0.5	CFRNN	2.000 (0.000)	0.996 (0.000)
0.5	NCTP	0.316 (0.016)	0.530 (0.005)
0.5	CAFHT	0.334 (0.003)	0.795 (0.004)
0.6			
0.6	CFRNN	2.000 (0.000)	0.996 (0.001)
0.6	NCTP	0.321 (0.016)	0.436 (0.006)
0.6	CAFHT	0.378 (0.003)	0.772 (0.004)
0.7			
0.7	CFRNN	2.000 (0.000)	0.994 (0.001)
0.7	NCTP	0.310 (0.016)	0.341 (0.006)
0.7	CAFHT	0.422 (0.003)	0.743 (0.005)
0.8			
0.8	CFRNN	2.000 (0.000)	0.994 (0.001)
0.8	NCTP	0.306 (0.017)	0.249 (0.008)
0.8	CAFHT	0.457 (0.004)	0.715 (0.005)
0.9			
0.9	CFRNN	2.000 (0.000)	0.994 (0.001)
0.9	NCTP	0.301 (0.016)	0.152 (0.007)
0.9	CAFHT	0.500 (0.004)	0.692 (0.006)

AR data with static noise profile. Next, we present the results based on data generated from the AR model with the static noise profile.

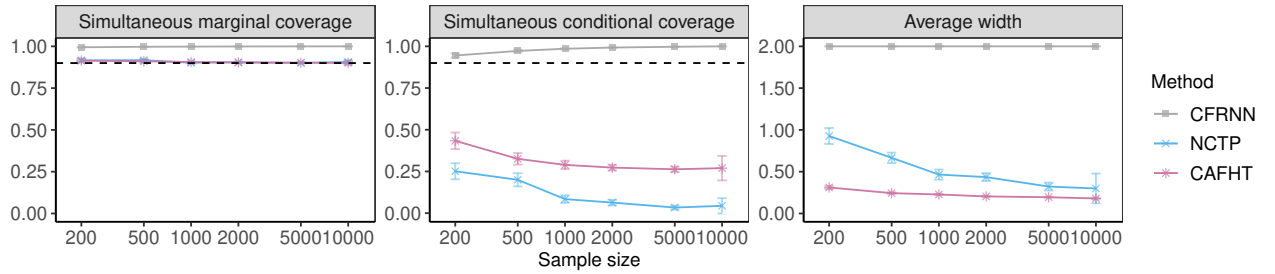

Figure A4. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. See Table A6 for more detailed results and standard errors.

Table A6. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A4.

			Simultaneous coverage		
Sample size	Method	Average width	Conditional-hard	Conditional-easy	Marginal
200					
200	CFRNN	2.000 (0.000)	0.945 (0.006)	1.000 (0.000)	0.994 (0.001)
200	NCTP	0.927 (0.048)	0.251 (0.024)	0.990 (0.002)	0.916 (0.004)
200	CAFHT	0.311 (0.010)	0.434 (0.025)	0.968 (0.005)	0.914 (0.006)
500					
500	CFRNN	2.000 (0.000)	0.972 (0.004)	1.000 (0.000)	0.997 (0.000)
500	NCTP	0.666 (0.031)	0.200 (0.020)	0.995 (0.002)	0.917 (0.003)
500	CAFHT	0.243 (0.005)	0.326 (0.017)	0.974 (0.003)	0.910 (0.004)
1000					
1000	CFRNN	2.000 (0.000)	0.986 (0.002)	1.000 (0.000)	0.999 (0.000)
1000	NCTP	0.466 (0.030)	0.084 (0.011)	0.994 (0.001)	0.902 (0.002)
1000	CAFHT	0.227 (0.003)	0.289 (0.012)	0.976 (0.002)	0.906 (0.003)
2000					
2000	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
2000	NCTP	0.435 (0.024)	0.063 (0.008)	0.996 (0.001)	0.904 (0.002)
2000	CAFHT	0.204 (0.002)	0.273 (0.008)	0.973 (0.002)	0.904 (0.002)
5000					
5000	CFRNN	2.000 (0.000)	0.997 (0.001)	1.000 (0.000)	1.000 (0.000)
5000	NCTP	0.322 (0.024)	0.034 (0.006)	0.996 (0.001)	0.901 (0.002)
5000	CAFHT	0.194 (0.002)	0.263 (0.008)	0.973 (0.001)	0.902 (0.002)
10000					
10000	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
10000	NCTP	0.299 (0.089)	0.044 (0.023)	0.997 (0.002)	0.907 (0.005)
10000	CAFHT	0.180 (0.005)	0.270 (0.037)	0.966 (0.004)	0.901 (0.006)

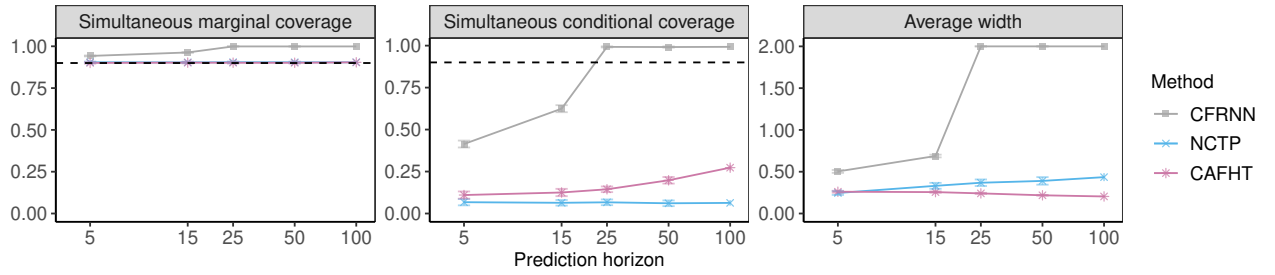

Figure A5. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the prediction horizon. See Table A7 for more detailed results and standard errors.

Table A7. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the prediction horizon. The red numbers indicate smaller prediction bands or higher conditional coverage. See corresponding plot in Figure A5.

				Simultaneous coverage		
Prediction horizon		Method	Average width	Conditional-hard	Conditional-easy	Marginal
5						
	5	CFRNN	0.504 (0.009)	0.414 (0.010)	1.000 (0.000)	0.942 (0.001)
	5	NCTP	0.246 (0.012)	0.067 (0.009)	0.996 (0.001)	0.904 (0.002)
	5	CAFHT	0.260 (0.006)	0.110 (0.011)	0.987 (0.001)	0.901 (0.002)
15						
	15	CFRNN	0.688 (0.008)	0.624 (0.010)	1.000 (0.000)	0.963 (0.001)
	15	NCTP	0.331 (0.018)	0.064 (0.008)	0.996 (0.001)	0.904 (0.002)
	15	CAFHT	0.256 (0.004)	0.125 (0.011)	0.987 (0.001)	0.902 (0.002)
25						
	25	CFRNN	2.000 (0.000)	0.992 (0.002)	1.000 (0.000)	0.999 (0.000)
	25	NCTP	0.368 (0.020)	0.067 (0.009)	0.996 (0.001)	0.905 (0.002)
	25	CAFHT	0.241 (0.003)	0.144 (0.008)	0.984 (0.001)	0.902 (0.002)
50						
	50	CFRNN	2.000 (0.000)	0.992 (0.002)	1.000 (0.000)	0.999 (0.000)
	50	NCTP	0.390 (0.023)	0.061 (0.008)	0.996 (0.001)	0.904 (0.002)
	50	CAFHT	0.218 (0.002)	0.198 (0.010)	0.978 (0.002)	0.901 (0.002)
100						
	100	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
	100	NCTP	0.435 (0.024)	0.063 (0.008)	0.996 (0.001)	0.904 (0.002)
	100	CAFHT	0.204 (0.002)	0.273 (0.008)	0.973 (0.002)	0.904 (0.002)

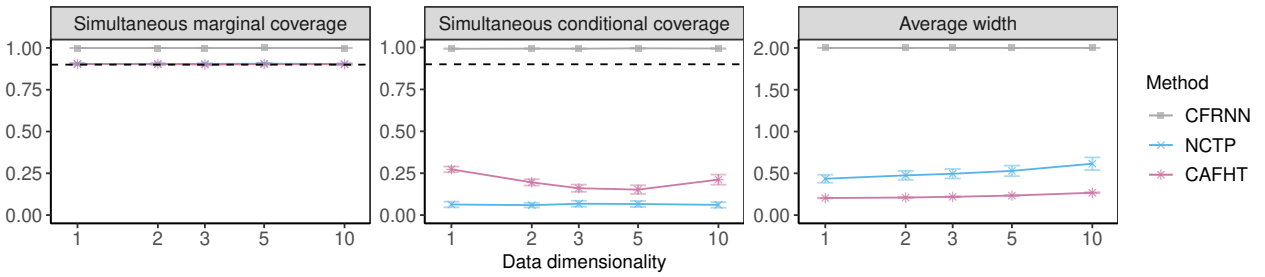

Figure A6. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the data dimensionality. See Table A8 for more detailed results and standard errors.

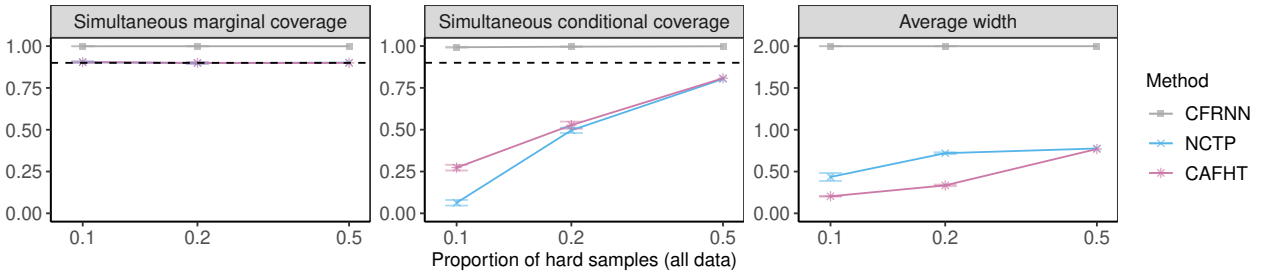

Figure A7. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the proportion of hard-to-predict trajectories. See Table A9 for more detailed results and standard errors.

Table A8. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the data dimensionality. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A6.

			Simultaneous coverage		
Data dimensionality	Method	Average width	Conditional-hard	Conditional-easy	Marginal
1					
1	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
1	NCTP	0.435 (0.024)	0.063 (0.008)	0.996 (0.001)	0.904 (0.002)
1	CAFHT	0.204 (0.002)	0.273 (0.008)	0.973 (0.002)	0.904 (0.002)
2					
2	CFRNN	2.000 (0.000)	0.994 (0.001)	1.000 (0.000)	0.999 (0.000)
2	NCTP	0.475 (0.027)	0.059 (0.007)	0.996 (0.001)	0.903 (0.002)
2	CAFHT	0.210 (0.002)	0.195 (0.009)	0.981 (0.001)	0.904 (0.002)
3					
3	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
3	NCTP	0.495 (0.028)	0.068 (0.009)	0.996 (0.001)	0.904 (0.002)
3	CAFHT	0.219 (0.003)	0.160 (0.011)	0.982 (0.001)	0.900 (0.002)
5					
5	CFRNN	2.000 (0.000)	0.995 (0.001)	1.000 (0.000)	1.000 (0.000)
5	NCTP	0.529 (0.032)	0.066 (0.009)	0.997 (0.001)	0.906 (0.002)
5	CAFHT	0.234 (0.003)	0.152 (0.013)	0.985 (0.002)	0.903 (0.002)
10					
10	CFRNN	2.000 (0.000)	0.994 (0.001)	1.000 (0.000)	0.999 (0.000)
10	NCTP	0.615 (0.038)	0.061 (0.008)	0.997 (0.001)	0.904 (0.002)
10	CAFHT	0.267 (0.003)	0.212 (0.015)	0.978 (0.002)	0.902 (0.003)

Table A9. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the proportion of hard-to-predict trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A7.

			Simultaneous coverage		
Proportion of hard samples (all data)	Method	Average width	Conditional-hard	Conditional-easy	Marginal
0.1					
0.1	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
0.1	NCTP	0.435 (0.024)	0.063 (0.008)	0.996 (0.001)	0.904 (0.002)
0.1	CAFHT	0.204 (0.002)	0.273 (0.008)	0.973 (0.002)	0.904 (0.002)
0.2					
0.2	CFRNN	2.000 (0.000)	0.996 (0.001)	1.000 (0.000)	0.999 (0.000)
0.2	NCTP	0.720 (0.006)	0.496 (0.008)	1.000 (0.000)	0.898 (0.002)
0.2	CAFHT	0.335 (0.005)	0.527 (0.011)	0.993 (0.001)	0.900 (0.003)
0.5					
0.5	CFRNN	2.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.999 (0.000)
0.5	NCTP	0.776 (0.005)	0.803 (0.003)	1.000 (0.000)	0.901 (0.002)
0.5	CAFHT	0.770 (0.007)	0.808 (0.004)	0.992 (0.001)	0.900 (0.002)

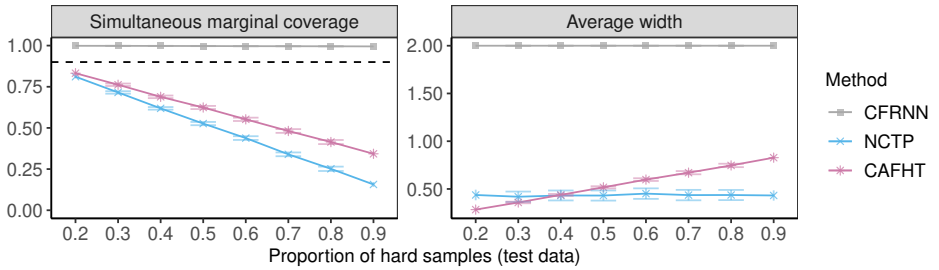

Figure A8. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods under distributional shift. The results are shown as a function of the proportion of hard-to-predict trajectories in the test data. See Table A10 for detailed results and standard errors.

Table A10. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods under distributional shift. The results are shown as a function of the proportion of hard-to-predict trajectories in the test data. The red numbers indicate higher marginal coverage. See the corresponding plot in Figure A8.

Proportion of hard samples (test data)	Method	Length	Marginal coverage
0.2			
0.2	CFRNN	2.000 (0.000)	0.999 (0.000)
0.2	NCTP	0.437 (0.028)	0.811 (0.003)
0.2	CAFHT	0.284 (0.003)	0.832 (0.003)
0.3			
0.3	CFRNN	2.000 (0.000)	0.998 (0.000)
0.3	NCTP	0.419 (0.027)	0.715 (0.004)
0.3	CAFHT	0.359 (0.004)	0.762 (0.004)
0.4			
0.4	CFRNN	2.000 (0.000)	0.998 (0.000)
0.4	NCTP	0.432 (0.026)	0.619 (0.004)
0.4	CAFHT	0.438 (0.005)	0.689 (0.004)
0.5			
0.5	CFRNN	2.000 (0.000)	0.997 (0.000)
0.5	NCTP	0.431 (0.027)	0.526 (0.005)
0.5	CAFHT	0.517 (0.006)	0.624 (0.005)
0.6			
0.6	CFRNN	2.000 (0.000)	0.996 (0.001)
0.6	NCTP	0.451 (0.028)	0.438 (0.006)
0.6	CAFHT	0.598 (0.007)	0.552 (0.005)
0.7			
0.7	CFRNN	2.000 (0.000)	0.996 (0.001)
0.7	NCTP	0.436 (0.027)	0.340 (0.006)
0.7	CAFHT	0.671 (0.008)	0.481 (0.006)
0.8			
0.8	CFRNN	2.000 (0.000)	0.996 (0.001)
0.8	NCTP	0.437 (0.027)	0.252 (0.007)
0.8	CAFHT	0.746 (0.009)	0.414 (0.006)
0.9			
0.9	CFRNN	2.000 (0.000)	0.995 (0.001)
0.9	NCTP	0.432 (0.027)	0.157 (0.007)
0.9	CAFHT	0.828 (0.010)	0.343 (0.007)

A5.1.2. Supplementary Results — Comparing Different Versions of CAFHT

In this subsection, we add different versions of CAFHT into comparison. We will separately analyze the CAFHT prediction bands constructed using multiplicative conformity scores (6) and those constructed using additive conformity scores (3). The conclusions from the results evaluated using synthetic data with the dynamic profile and with the static profile are very similar. To save space, we only demonstrate the results using data with the static profile.

We consider the following implementations of CAFHT:

- CAFHT: the main method. It is the CAFHT method based on the ACI prediction band using the data splitting strategy; see Algorithm A2.
- CAFHT - PID: the CAFHT method based on the conformal PID prediction band using the data splitting strategy. It can be implemented simply by substituting $\hat{C}^{\text{ACI}}$ to $\hat{C}^{\text{PID}}$ in Algorithm A2 wherever possible.
- CAFHT (theory): the CAFHT method based on the ACI prediction band after correcting the theoretical coverage; see Appendix A4 and Algorithm A6.
- CAFHT (theory) - PID: the CAFHT method based on the conformal PID prediction band after correcting the theoretical coverage. It can be implemented simply by substituting $\hat{C}^{\text{ACI}}$ to $\hat{C}^{\text{PID}}$ in Algorithm A6 wherever possible.

CAFHT — Multiplicative Scores

The results of CAFHT with multiplicative conformity scores (6) are first presented.

Similar to what we have observed from the results in subsection A5.1.1, CAFHT outperforms the benchmark methods (CFRNN and NCTP) across all configurations we considered. Generally, CAFHT produces narrower, more informative bands with higher conditional coverage. Among the different versions of CAFHT, the prediction bands generated using the theoretical correction approach (outlined in A4) tend to be more conservative compared to those from the data-splitting approach. Additionally, in our experiments, the performance of prediction bands constructed by CAFHT with ACI is empirically similar to those created using PID.

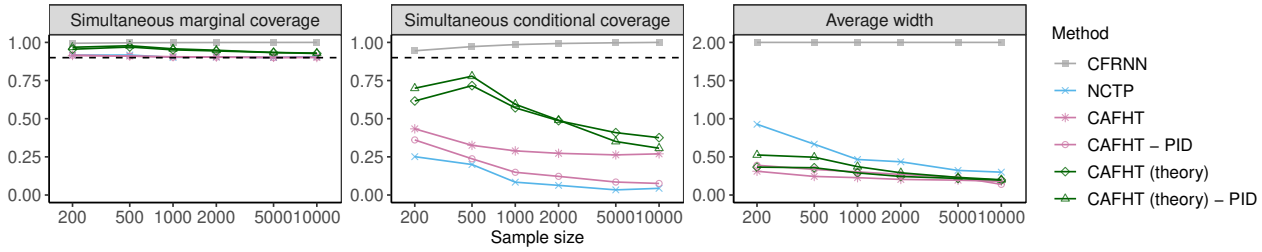

Figure A9. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. See Table A11 for detailed results and standard errors.

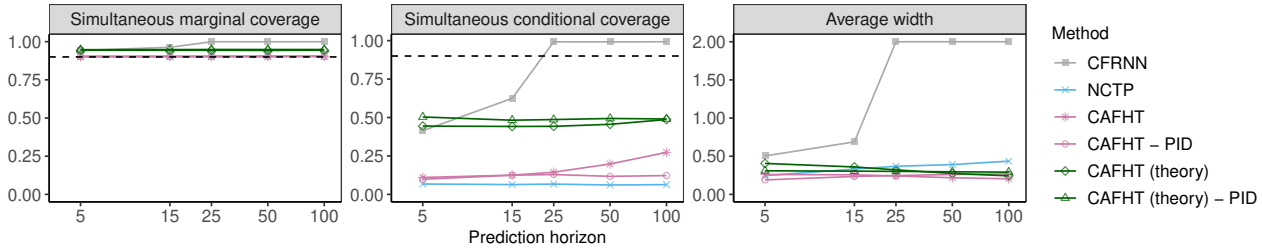

Figure A10. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the prediction horizon. See Table A12 for detailed results and standard errors.

Table A11. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. See corresponding plot in Figure A9.

			Simultaneous coverage		
Sample size	Method	Average width	Conditional-hard	Conditional-easy	Marginal
200					
200	CFRNN	2.000 (0.000)	0.945 (0.006)	1.000 (0.000)	0.994 (0.001)
200	NCTP	0.927 (0.048)	0.251 (0.024)	0.990 (0.002)	0.916 (0.004)
200	CAFHT	0.311 (0.010)	0.434 (0.025)	0.968 (0.005)	0.914 (0.006)
200	CAFHT - PID	0.390 (0.022)	0.361 (0.027)	0.978 (0.004)	0.916 (0.006)
200	CAFHT (theory)	0.365 (0.009)	0.616 (0.016)	0.993 (0.001)	0.955 (0.002)
200	CAFHT (theory) - PID	0.525 (0.014)	0.700 (0.018)	0.998 (0.001)	0.968 (0.002)
500					
500	CFRNN	2.000 (0.000)	0.972 (0.004)	1.000 (0.000)	0.997 (0.000)
500	NCTP	0.666 (0.031)	0.200 (0.020)	0.995 (0.002)	0.917 (0.003)
500	CAFHT	0.243 (0.005)	0.326 (0.017)	0.974 (0.003)	0.910 (0.004)
500	CAFHT - PID	0.335 (0.021)	0.237 (0.022)	0.983 (0.003)	0.910 (0.004)
500	CAFHT (theory)	0.358 (0.006)	0.717 (0.012)	0.997 (0.001)	0.970 (0.001)
500	CAFHT (theory) - PID	0.495 (0.009)	0.779 (0.013)	1.000 (0.000)	0.978 (0.001)
1000					
1000	CFRNN	2.000 (0.000)	0.986 (0.002)	1.000 (0.000)	0.999 (0.000)
1000	NCTP	0.466 (0.030)	0.084 (0.011)	0.994 (0.001)	0.902 (0.002)
1000	CAFHT	0.227 (0.003)	0.289 (0.012)	0.976 (0.002)	0.906 (0.003)
1000	CAFHT - PID	0.303 (0.018)	0.149 (0.014)	0.988 (0.002)	0.903 (0.003)
1000	CAFHT (theory)	0.290 (0.004)	0.571 (0.013)	0.994 (0.001)	0.951 (0.002)
1000	CAFHT (theory) - PID	0.371 (0.007)	0.595 (0.014)	0.999 (0.000)	0.958 (0.002)
2000					
2000	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
2000	NCTP	0.435 (0.024)	0.063 (0.008)	0.996 (0.001)	0.904 (0.002)
2000	CAFHT	0.204 (0.002)	0.273 (0.008)	0.973 (0.002)	0.904 (0.002)
2000	CAFHT - PID	0.270 (0.016)	0.122 (0.012)	0.989 (0.001)	0.904 (0.002)
2000	CAFHT (theory)	0.243 (0.003)	0.487 (0.010)	0.994 (0.001)	0.944 (0.001)
2000	CAFHT (theory) - PID	0.291 (0.004)	0.490 (0.011)	0.999 (0.000)	0.949 (0.002)
5000					
5000	CFRNN	2.000 (0.000)	0.997 (0.001)	1.000 (0.000)	1.000 (0.000)
5000	NCTP	0.322 (0.024)	0.034 (0.006)	0.996 (0.001)	0.901 (0.002)
5000	CAFHT	0.194 (0.002)	0.263 (0.008)	0.973 (0.001)	0.902 (0.002)
5000	CAFHT - PID	0.233 (0.015)	0.085 (0.009)	0.990 (0.001)	0.900 (0.002)
5000	CAFHT (theory)	0.217 (0.002)	0.409 (0.008)	0.991 (0.001)	0.933 (0.001)
5000	CAFHT (theory) - PID	0.231 (0.003)	0.351 (0.010)	0.998 (0.000)	0.934 (0.001)
10000					
10000	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
10000	NCTP	0.299 (0.089)	0.044 (0.023)	0.997 (0.002)	0.907 (0.005)
10000	CAFHT	0.180 (0.005)	0.270 (0.037)	0.966 (0.004)	0.901 (0.006)
10000	CAFHT - PID	0.140 (0.016)	0.076 (0.039)	0.989 (0.004)	0.903 (0.005)
10000	CAFHT (theory)	0.196 (0.005)	0.376 (0.033)	0.985 (0.002)	0.927 (0.006)
10000	CAFHT (theory) - PID	0.200 (0.008)	0.307 (0.045)	0.994 (0.001)	0.930 (0.004)

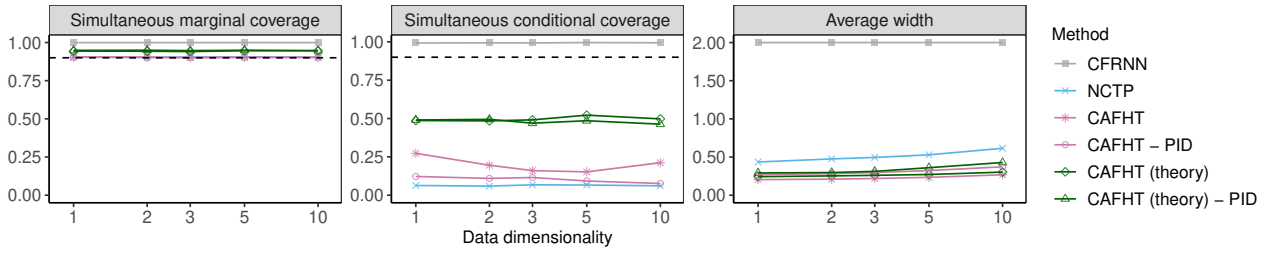

Figure A11. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the data dimensionality. See Table A13 for detailed results and standard errors.

Table A12. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the prediction horizon. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A10.

				Simultaneous coverage	
Prediction horizon	Method	Average width	Conditional-hard	Conditional-easy	Marginal
5					
5	CFRNN	0.504 (0.009)	0.414 (0.010)	1.000 (0.000)	0.942 (0.001)
5	NCTP	0.246 (0.012)	0.067 (0.009)	0.996 (0.001)	0.904 (0.002)
5	CAFHT	0.260 (0.006)	0.110 (0.011)	0.987 (0.001)	0.901 (0.002)
5	CAFHT - PID	0.190 (0.008)	0.098 (0.011)	0.991 (0.001)	0.903 (0.002)
5	CAFHT (theory)	0.406 (0.007)	0.444 (0.012)	0.998 (0.000)	0.944 (0.002)
5	CAFHT (theory) - PID	0.309 (0.005)	0.503 (0.014)	0.996 (0.001)	0.947 (0.002)
15					
15	CFRNN	0.688 (0.008)	0.624 (0.010)	1.000 (0.000)	0.963 (0.001)
15	NCTP	0.331 (0.018)	0.064 (0.008)	0.996 (0.001)	0.904 (0.002)
15	CAFHT	0.256 (0.004)	0.125 (0.011)	0.987 (0.001)	0.902 (0.002)
15	CAFHT - PID	0.235 (0.014)	0.125 (0.012)	0.987 (0.002)	0.902 (0.003)
15	CAFHT (theory)	0.359 (0.005)	0.442 (0.012)	0.998 (0.000)	0.944 (0.001)
15	CAFHT (theory) - PID	0.304 (0.006)	0.483 (0.012)	0.999 (0.000)	0.948 (0.002)
25					
25	CFRNN	2.000 (0.000)	0.992 (0.002)	1.000 (0.000)	0.999 (0.000)
25	NCTP	0.368 (0.020)	0.067 (0.009)	0.996 (0.001)	0.905 (0.002)
25	CAFHT	0.241 (0.003)	0.144 (0.008)	0.984 (0.001)	0.902 (0.002)
25	CAFHT - PID	0.247 (0.015)	0.129 (0.013)	0.987 (0.002)	0.903 (0.003)
25	CAFHT (theory)	0.321 (0.004)	0.442 (0.012)	0.998 (0.000)	0.944 (0.002)
25	CAFHT (theory) - PID	0.302 (0.005)	0.487 (0.011)	0.999 (0.000)	0.949 (0.001)
50					
50	CFRNN	2.000 (0.000)	0.992 (0.002)	1.000 (0.000)	0.999 (0.000)
50	NCTP	0.390 (0.023)	0.061 (0.008)	0.996 (0.001)	0.904 (0.002)
50	CAFHT	0.218 (0.002)	0.198 (0.010)	0.978 (0.002)	0.901 (0.002)
50	CAFHT - PID	0.260 (0.017)	0.117 (0.014)	0.991 (0.001)	0.905 (0.003)
50	CAFHT (theory)	0.274 (0.003)	0.456 (0.011)	0.996 (0.000)	0.943 (0.002)
50	CAFHT (theory) - PID	0.296 (0.005)	0.494 (0.011)	0.998 (0.000)	0.949 (0.002)
100					
100	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
100	NCTP	0.435 (0.024)	0.063 (0.008)	0.996 (0.001)	0.904 (0.002)
100	CAFHT	0.204 (0.002)	0.273 (0.008)	0.973 (0.002)	0.904 (0.002)
100	CAFHT - PID	0.270 (0.016)	0.122 (0.012)	0.989 (0.001)	0.904 (0.002)
100	CAFHT (theory)	0.243 (0.003)	0.487 (0.010)	0.994 (0.001)	0.944 (0.001)
100	CAFHT (theory) - PID	0.291 (0.004)	0.490 (0.011)	0.999 (0.000)	0.949 (0.002)

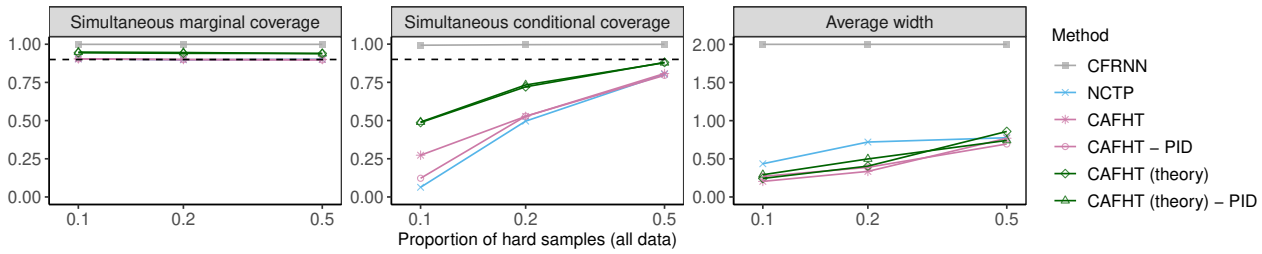

Figure A12. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the proportion of hard-to-predict trajectories. See Table A14 for detailed results and standard errors.

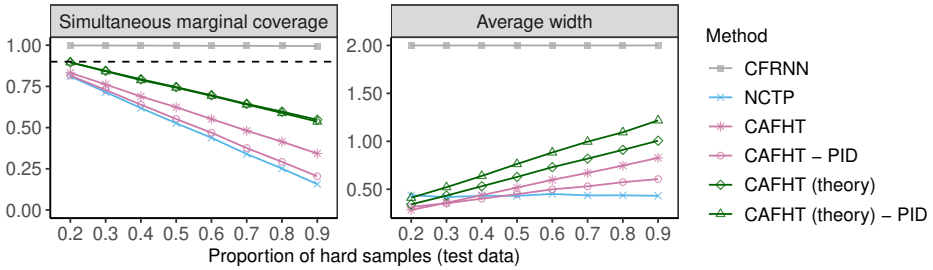

Figure A13. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods under distributional shift. The results are shown as a function of the proportion of hard-to-predict trajectories in the test data. See Table A15 for detailed results and standard errors.

Table A13. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the data dimensionality. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A11.

			Simultaneous coverage		
Data dimensionality	Method	Average width	Conditional-hard	Conditional-easy	Marginal
1					
1	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
1	NCTP	0.435 (0.024)	0.063 (0.008)	0.996 (0.001)	0.904 (0.002)
1	CAFHT	0.204 (0.002)	0.273 (0.008)	0.973 (0.002)	0.904 (0.002)
1	CAFHT - PID	0.270 (0.016)	0.122 (0.012)	0.989 (0.001)	0.904 (0.002)
1	CAFHT (theory)	0.243 (0.003)	0.487 (0.010)	0.994 (0.001)	0.944 (0.001)
1	CAFHT (theory) - PID	0.291 (0.004)	0.490 (0.011)	0.999 (0.000)	0.949 (0.002)
2					
2	CFRNN	2.000 (0.000)	0.994 (0.001)	1.000 (0.000)	0.999 (0.000)
2	NCTP	0.475 (0.027)	0.059 (0.007)	0.996 (0.001)	0.903 (0.002)
2	CAFHT	0.210 (0.002)	0.195 (0.009)	0.981 (0.001)	0.904 (0.002)
2	CAFHT - PID	0.283 (0.017)	0.110 (0.010)	0.988 (0.001)	0.902 (0.002)
2	CAFHT (theory)	0.253 (0.002)	0.485 (0.012)	0.993 (0.001)	0.943 (0.001)
2	CAFHT (theory) - PID	0.297 (0.004)	0.495 (0.012)	0.999 (0.000)	0.949 (0.001)
3					
3	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
3	NCTP	0.495 (0.028)	0.068 (0.009)	0.996 (0.001)	0.904 (0.002)
3	CAFHT	0.219 (0.003)	0.160 (0.011)	0.982 (0.001)	0.900 (0.002)
3	CAFHT - PID	0.297 (0.018)	0.115 (0.012)	0.988 (0.001)	0.902 (0.002)
3	CAFHT (theory)	0.260 (0.003)	0.491 (0.012)	0.990 (0.001)	0.941 (0.001)
3	CAFHT (theory) - PID	0.312 (0.004)	0.470 (0.011)	1.000 (0.000)	0.947 (0.001)
5					
5	CFRNN	2.000 (0.000)	0.995 (0.001)	1.000 (0.000)	1.000 (0.000)
5	NCTP	0.529 (0.032)	0.066 (0.009)	0.997 (0.001)	0.906 (0.002)
5	CAFHT	0.234 (0.003)	0.152 (0.013)	0.985 (0.002)	0.903 (0.002)
5	CAFHT - PID	0.323 (0.021)	0.093 (0.010)	0.991 (0.001)	0.903 (0.002)
5	CAFHT (theory)	0.271 (0.003)	0.522 (0.011)	0.992 (0.001)	0.946 (0.002)
5	CAFHT (theory) - PID	0.360 (0.004)	0.486 (0.012)	1.000 (0.000)	0.949 (0.002)
10					
10	CFRNN	2.000 (0.000)	0.994 (0.001)	1.000 (0.000)	0.999 (0.000)
10	NCTP	0.615 (0.038)	0.061 (0.008)	0.997 (0.001)	0.904 (0.002)
10	CAFHT	0.267 (0.003)	0.212 (0.015)	0.978 (0.002)	0.902 (0.003)
10	CAFHT - PID	0.372 (0.023)	0.076 (0.011)	0.993 (0.001)	0.902 (0.002)
10	CAFHT (theory)	0.303 (0.003)	0.498 (0.009)	0.995 (0.001)	0.946 (0.001)
10	CAFHT (theory) - PID	0.428 (0.003)	0.463 (0.012)	1.000 (0.000)	0.947 (0.001)

Table A14. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the proportion of hard-to-predict trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A12.

			Simultaneous coverage		
Proportion of hard samples (all data)	Method	Average width	Conditional-hard	Conditional-easy	Marginal
0.1					
0.1	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
0.1	NCTP	0.435 (0.024)	0.063 (0.008)	0.996 (0.001)	0.904 (0.002)
0.1	CAFHT	0.204 (0.002)	0.273 (0.008)	0.973 (0.002)	0.904 (0.002)
0.1	CAFHT - PID	0.270 (0.016)	0.122 (0.012)	0.989 (0.001)	0.904 (0.002)
0.1	CAFHT (theory)	0.243 (0.003)	0.487 (0.010)	0.994 (0.001)	0.944 (0.001)
0.1	CAFHT (theory) - PID	0.291 (0.004)	0.490 (0.011)	0.999 (0.000)	0.949 (0.002)
0.2					
0.2	CFRNN	2.000 (0.000)	0.996 (0.001)	1.000 (0.000)	0.999 (0.000)
0.2	NCTP	0.720 (0.006)	0.496 (0.008)	1.000 (0.000)	0.898 (0.002)
0.2	CAFHT	0.335 (0.005)	0.527 (0.011)	0.993 (0.001)	0.900 (0.003)
0.2	CAFHT - PID	0.388 (0.005)	0.528 (0.010)	0.993 (0.001)	0.899 (0.002)
0.2	CAFHT (theory)	0.404 (0.005)	0.721 (0.007)	0.996 (0.001)	0.940 (0.001)
0.2	CAFHT (theory) - PID	0.498 (0.005)	0.733 (0.007)	0.999 (0.000)	0.946 (0.001)
0.5					
0.5	CFRNN	2.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.999 (0.000)
0.5	NCTP	0.776 (0.005)	0.803 (0.003)	1.000 (0.000)	0.901 (0.002)
0.5	CAFHT	0.770 (0.007)	0.808 (0.004)	0.992 (0.001)	0.900 (0.002)
0.5	CAFHT - PID	0.694 (0.005)	0.796 (0.004)	1.000 (0.000)	0.898 (0.002)
0.5	CAFHT (theory)	0.858 (0.007)	0.880 (0.003)	0.998 (0.000)	0.939 (0.001)
0.5	CAFHT (theory) - PID	0.741 (0.005)	0.878 (0.003)	1.000 (0.000)	0.939 (0.002)

Table A15. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods under distributional shift. The results are shown as a function of the proportion of hard-to-predict trajectories in the test data. The red numbers indicate higher marginal coverage. See the corresponding plot in Figure A13.

Proportion of hard samples (test data)	Method	Length	Marginal coverage
0.2			
0.2	CFRNN	2.000 (0.000)	0.999 (0.000)
0.2	NCTP	0.437 (0.028)	0.811 (0.003)
0.2	CAFHT	0.284 (0.003)	0.832 (0.003)
0.2	CAFHT - PID	0.316 (0.018)	0.816 (0.003)
0.2	CAFHT (theory)	0.342 (0.005)	0.896 (0.002)
0.2	CAFHT (theory) - PID	0.411 (0.008)	0.896 (0.003)
0.3			
0.3	CFRNN	2.000 (0.000)	0.998 (0.000)
0.3	NCTP	0.419 (0.027)	0.715 (0.004)
0.3	CAFHT	0.359 (0.004)	0.762 (0.004)
0.3	CAFHT - PID	0.353 (0.017)	0.728 (0.004)
0.3	CAFHT (theory)	0.434 (0.006)	0.844 (0.003)
0.3	CAFHT (theory) - PID	0.521 (0.010)	0.842 (0.003)
0.4			
0.4	CFRNN	2.000 (0.000)	0.998 (0.000)
0.4	NCTP	0.432 (0.026)	0.619 (0.004)
0.4	CAFHT	0.438 (0.005)	0.689 (0.004)
0.4	CAFHT - PID	0.402 (0.017)	0.640 (0.005)
0.4	CAFHT (theory)	0.532 (0.007)	0.794 (0.003)
0.4	CAFHT (theory) - PID	0.639 (0.011)	0.788 (0.004)
0.5			
0.5	CFRNN	2.000 (0.000)	0.997 (0.000)
0.5	NCTP	0.431 (0.027)	0.526 (0.005)
0.5	CAFHT	0.517 (0.006)	0.624 (0.005)
0.5	CAFHT - PID	0.448 (0.017)	0.552 (0.006)
0.5	CAFHT (theory)	0.628 (0.009)	0.746 (0.004)
0.5	CAFHT (theory) - PID	0.764 (0.013)	0.743 (0.005)
0.6			
0.6	CFRNN	2.000 (0.000)	0.996 (0.001)
0.6	NCTP	0.451 (0.028)	0.438 (0.006)
0.6	CAFHT	0.598 (0.007)	0.552 (0.005)
0.6	CAFHT - PID	0.499 (0.020)	0.469 (0.007)
0.6	CAFHT (theory)	0.730 (0.011)	0.696 (0.005)
0.6	CAFHT (theory) - PID	0.883 (0.016)	0.693 (0.006)
0.7			
0.7	CFRNN	2.000 (0.000)	0.996 (0.001)
0.7	NCTP	0.436 (0.027)	0.340 (0.006)
0.7	CAFHT	0.671 (0.008)	0.481 (0.006)
0.7	CAFHT - PID	0.531 (0.022)	0.376 (0.008)
0.7	CAFHT (theory)	0.819 (0.012)	0.644 (0.005)
0.7	CAFHT (theory) - PID	0.996 (0.019)	0.640 (0.007)
0.8			
0.8	CFRNN	2.000 (0.000)	0.996 (0.001)
0.8	NCTP	0.437 (0.027)	0.252 (0.007)
0.8	CAFHT	0.746 (0.009)	0.414 (0.006)
0.8	CAFHT - PID	0.574 (0.023)	0.292 (0.008)
0.8	CAFHT (theory)	0.911 (0.013)	0.597 (0.005)
0.8	CAFHT (theory) - PID	1.094 (0.020)	0.589 (0.008)
0.9			
0.9	CFRNN	2.000 (0.000)	0.995 (0.001)
0.9	NCTP	0.432 (0.027)	0.157 (0.007)
0.9	CAFHT	0.828 (0.010)	0.343 (0.007)
0.9	CAFHT - PID	0.606 (0.027)	0.205 (0.010)
0.9	CAFHT (theory)	1.006 (0.014)	0.548 (0.006)
0.9	CAFHT (theory) - PID	1.218 (0.023)	0.537 (0.009)

CAFHT — Additive Scores

Finally, the results of CAFHT with additive conformity scores (3) are presented.

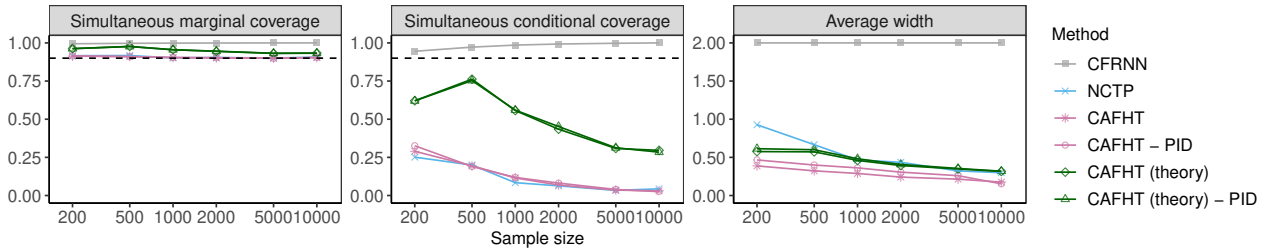

Figure A14. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. See Table A16 for detailed results and standard errors.

Table A16. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A14.

			Simultaneous coverage		
Sample size	Method	Average width	Conditional-hard	Conditional-easy	Marginal
200					
200	CFRNN	2.000 (0.000)	0.945 (0.006)	1.000 (0.000)	0.994 (0.001)
200	NCTP	0.927 (0.048)	0.251 (0.024)	0.990 (0.002)	0.916 (0.004)
200	CAFHT	0.388 (0.018)	0.290 (0.030)	0.982 (0.004)	0.913 (0.006)
200	CAFHT - PID	0.466 (0.026)	0.326 (0.031)	0.980 (0.004)	0.915 (0.006)
200	CAFHT (theory)	0.576 (0.015)	0.621 (0.023)	1.000 (0.000)	0.963 (0.002)
200	CAFHT (theory) - PID	0.614 (0.016)	0.620 (0.024)	1.000 (0.000)	0.962 (0.002)
500					
500	CFRNN	2.000 (0.000)	0.972 (0.004)	1.000 (0.000)	0.997 (0.000)
500	NCTP	0.666 (0.031)	0.200 (0.020)	0.995 (0.002)	0.917 (0.003)
500	CAFHT	0.323 (0.013)	0.195 (0.023)	0.988 (0.003)	0.911 (0.004)
500	CAFHT - PID	0.399 (0.024)	0.190 (0.022)	0.988 (0.003)	0.910 (0.004)
500	CAFHT (theory)	0.572 (0.008)	0.762 (0.016)	1.000 (0.000)	0.977 (0.002)
500	CAFHT (theory) - PID	0.600 (0.009)	0.753 (0.015)	1.000 (0.000)	0.976 (0.001)
1000					
1000	CFRNN	2.000 (0.000)	0.986 (0.002)	1.000 (0.000)	0.999 (0.000)
1000	NCTP	0.466 (0.030)	0.084 (0.011)	0.994 (0.001)	0.902 (0.002)
1000	CAFHT	0.289 (0.010)	0.115 (0.015)	0.993 (0.001)	0.904 (0.003)
1000	CAFHT - PID	0.362 (0.021)	0.119 (0.015)	0.993 (0.001)	0.904 (0.003)
1000	CAFHT (theory)	0.460 (0.006)	0.557 (0.015)	1.000 (0.000)	0.955 (0.002)
1000	CAFHT (theory) - PID	0.481 (0.006)	0.559 (0.013)	1.000 (0.000)	0.955 (0.002)
2000					
2000	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
2000	NCTP	0.435 (0.024)	0.063 (0.008)	0.996 (0.001)	0.904 (0.002)
2000	CAFHT	0.241 (0.008)	0.069 (0.010)	0.993 (0.001)	0.901 (0.002)
2000	CAFHT - PID	0.306 (0.019)	0.081 (0.012)	0.994 (0.001)	0.904 (0.002)
2000	CAFHT (theory)	0.393 (0.004)	0.434 (0.011)	1.000 (0.000)	0.944 (0.001)
2000	CAFHT (theory) - PID	0.407 (0.004)	0.452 (0.011)	1.000 (0.000)	0.946 (0.001)
5000					
5000	CFRNN	2.000 (0.000)	0.997 (0.001)	1.000 (0.000)	1.000 (0.000)
5000	NCTP	0.322 (0.024)	0.034 (0.006)	0.996 (0.001)	0.901 (0.002)
5000	CAFHT	0.214 (0.007)	0.035 (0.006)	0.994 (0.001)	0.899 (0.002)
5000	CAFHT - PID	0.260 (0.017)	0.040 (0.006)	0.994 (0.001)	0.899 (0.002)
5000	CAFHT (theory)	0.348 (0.003)	0.308 (0.008)	1.000 (0.000)	0.931 (0.001)
5000	CAFHT (theory) - PID	0.354 (0.003)	0.313 (0.009)	1.000 (0.000)	0.932 (0.001)
10000					
10000	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
10000	NCTP	0.299 (0.089)	0.044 (0.023)	0.997 (0.002)	0.907 (0.005)
10000	CAFHT	0.176 (0.023)	0.030 (0.025)	0.994 (0.002)	0.903 (0.004)
10000	CAFHT - PID	0.154 (0.026)	0.025 (0.023)	0.997 (0.001)	0.905 (0.005)
10000	CAFHT (theory)	0.315 (0.005)	0.296 (0.016)	1.000 (0.000)	0.933 (0.004)
10000	CAFHT (theory) - PID	0.319 (0.007)	0.284 (0.042)	1.000 (0.000)	0.932 (0.005)

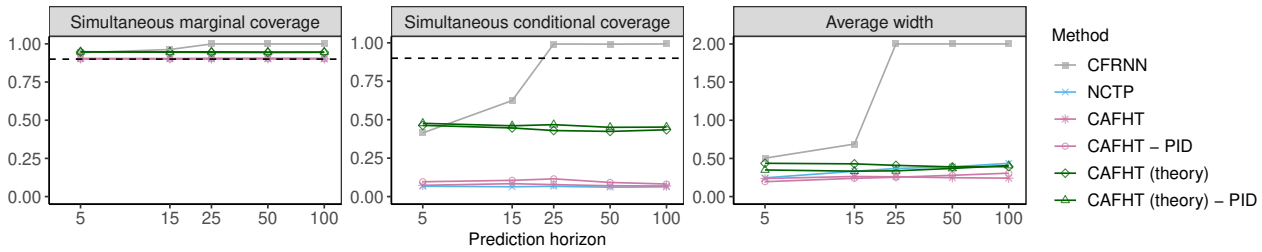

Figure A15. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the prediction horizon. See Table A17 for detailed results and standard errors.

Table A17. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the prediction horizon. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A15.

				Simultaneous coverage	
Prediction horizon	Method	Average width	Conditional-hard	Conditional-easy	Marginal
5					
5	CFRNN	0.504 (0.009)	0.414 (0.010)	1.000 (0.000)	0.942 (0.001)
5	NCTP	0.246 (0.012)	0.067 (0.009)	0.996 (0.001)	0.904 (0.002)
5	CAFHT	0.237 (0.008)	0.073 (0.009)	0.993 (0.001)	0.902 (0.002)
5	CAFHT - PID	0.195 (0.008)	0.096 (0.010)	0.990 (0.001)	0.902 (0.002)
5	CAFHT (theory)	0.436 (0.008)	0.463 (0.012)	1.000 (0.000)	0.947 (0.002)
5	CAFHT (theory) - PID	0.348 (0.008)	0.477 (0.012)	1.000 (0.000)	0.948 (0.002)
15					
15	CFRNN	0.688 (0.008)	0.624 (0.010)	1.000 (0.000)	0.963 (0.001)
15	NCTP	0.331 (0.018)	0.064 (0.008)	0.996 (0.001)	0.904 (0.002)
15	CAFHT	0.263 (0.008)	0.084 (0.010)	0.992 (0.001)	0.902 (0.002)
15	CAFHT - PID	0.239 (0.014)	0.106 (0.010)	0.988 (0.002)	0.901 (0.002)
15	CAFHT (theory)	0.429 (0.006)	0.446 (0.012)	1.000 (0.000)	0.945 (0.001)
15	CAFHT (theory) - PID	0.333 (0.006)	0.460 (0.012)	1.000 (0.000)	0.947 (0.002)
25					
25	CFRNN	2.000 (0.000)	0.992 (0.002)	1.000 (0.000)	0.999 (0.000)
25	NCTP	0.368 (0.020)	0.067 (0.009)	0.996 (0.001)	0.905 (0.002)
25	CAFHT	0.259 (0.008)	0.078 (0.010)	0.991 (0.001)	0.902 (0.002)
25	CAFHT - PID	0.253 (0.016)	0.115 (0.011)	0.989 (0.002)	0.904 (0.003)
25	CAFHT (theory)	0.409 (0.005)	0.430 (0.013)	1.000 (0.000)	0.944 (0.002)
25	CAFHT (theory) - PID	0.337 (0.006)	0.468 (0.013)	1.000 (0.000)	0.948 (0.002)
50					
50	CFRNN	2.000 (0.000)	0.992 (0.002)	1.000 (0.000)	0.999 (0.000)
50	NCTP	0.390 (0.023)	0.061 (0.008)	0.996 (0.001)	0.904 (0.002)
50	CAFHT	0.247 (0.008)	0.070 (0.010)	0.993 (0.001)	0.902 (0.002)
50	CAFHT - PID	0.279 (0.019)	0.092 (0.014)	0.993 (0.001)	0.904 (0.003)
50	CAFHT (theory)	0.389 (0.004)	0.424 (0.013)	1.000 (0.000)	0.943 (0.002)
50	CAFHT (theory) - PID	0.368 (0.005)	0.450 (0.012)	1.000 (0.000)	0.946 (0.002)
100					
100	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
100	NCTP	0.435 (0.024)	0.063 (0.008)	0.996 (0.001)	0.904 (0.002)
100	CAFHT	0.241 (0.008)	0.069 (0.010)	0.993 (0.001)	0.901 (0.002)
100	CAFHT - PID	0.306 (0.019)	0.081 (0.012)	0.994 (0.001)	0.904 (0.002)
100	CAFHT (theory)	0.393 (0.004)	0.434 (0.011)	1.000 (0.000)	0.944 (0.001)
100	CAFHT (theory) - PID	0.407 (0.004)	0.452 (0.011)	1.000 (0.000)	0.946 (0.001)

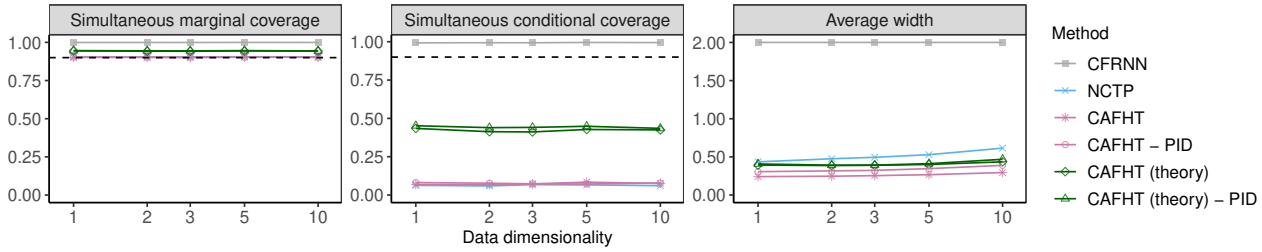

Figure A16. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the data dimensionality. See Table A18 for detailed results and standard errors.

Table A18. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the data dimensionality. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A16.

			Simultaneous coverage		
Data dimensionality	Method	Average width	Conditional-hard	Conditional-easy	Marginal
1					
1	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
1	NCTP	0.435 (0.024)	0.063 (0.008)	0.996 (0.001)	0.904 (0.002)
1	CAFHT	0.241 (0.008)	0.069 (0.010)	0.993 (0.001)	0.901 (0.002)
1	CAFHT - PID	0.306 (0.019)	0.081 (0.012)	0.994 (0.001)	0.904 (0.002)
1	CAFHT (theory)	0.393 (0.004)	0.434 (0.011)	1.000 (0.000)	0.944 (0.001)
1	CAFHT (theory) - PID	0.407 (0.004)	0.452 (0.011)	1.000 (0.000)	0.946 (0.001)
2					
2	CFRNN	2.000 (0.000)	0.994 (0.001)	1.000 (0.000)	0.999 (0.000)
2	NCTP	0.475 (0.027)	0.059 (0.007)	0.996 (0.001)	0.903 (0.002)
2	CAFHT	0.247 (0.008)	0.067 (0.009)	0.993 (0.001)	0.902 (0.002)
2	CAFHT - PID	0.317 (0.020)	0.077 (0.011)	0.993 (0.001)	0.902 (0.002)
2	CAFHT (theory)	0.386 (0.003)	0.413 (0.011)	1.000 (0.000)	0.942 (0.001)
2	CAFHT (theory) - PID	0.392 (0.003)	0.440 (0.010)	1.000 (0.000)	0.945 (0.001)
3					
3	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
3	NCTP	0.495 (0.028)	0.068 (0.009)	0.996 (0.001)	0.904 (0.002)
3	CAFHT	0.254 (0.008)	0.073 (0.010)	0.992 (0.001)	0.901 (0.002)
3	CAFHT - PID	0.323 (0.021)	0.073 (0.010)	0.993 (0.001)	0.902 (0.002)
3	CAFHT (theory)	0.390 (0.004)	0.412 (0.013)	1.000 (0.000)	0.942 (0.002)
3	CAFHT (theory) - PID	0.392 (0.004)	0.441 (0.012)	1.000 (0.000)	0.944 (0.002)
5					
5	CFRNN	2.000 (0.000)	0.995 (0.001)	1.000 (0.000)	1.000 (0.000)
5	NCTP	0.529 (0.032)	0.066 (0.009)	0.997 (0.001)	0.906 (0.002)
5	CAFHT	0.267 (0.009)	0.084 (0.012)	0.993 (0.001)	0.904 (0.002)
5	CAFHT - PID	0.346 (0.024)	0.072 (0.010)	0.993 (0.001)	0.903 (0.002)
5	CAFHT (theory)	0.401 (0.004)	0.428 (0.012)	1.000 (0.000)	0.944 (0.001)
5	CAFHT (theory) - PID	0.411 (0.004)	0.449 (0.012)	1.000 (0.000)	0.946 (0.002)
10					
10	CFRNN	2.000 (0.000)	0.994 (0.001)	1.000 (0.000)	0.999 (0.000)
10	NCTP	0.615 (0.038)	0.061 (0.008)	0.997 (0.001)	0.904 (0.002)
10	CAFHT	0.295 (0.009)	0.076 (0.011)	0.993 (0.001)	0.902 (0.002)
10	CAFHT - PID	0.389 (0.027)	0.078 (0.011)	0.994 (0.001)	0.903 (0.002)
10	CAFHT (theory)	0.435 (0.004)	0.425 (0.011)	1.000 (0.000)	0.943 (0.001)
10	CAFHT (theory) - PID	0.469 (0.004)	0.435 (0.011)	1.000 (0.000)	0.944 (0.001)

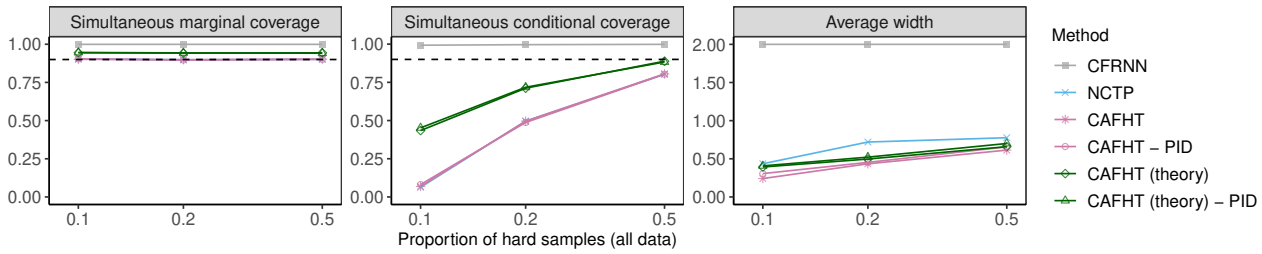

Figure A17. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the proportion of hard-to-predict trajectories. See Table A19 for detailed results and standard errors.

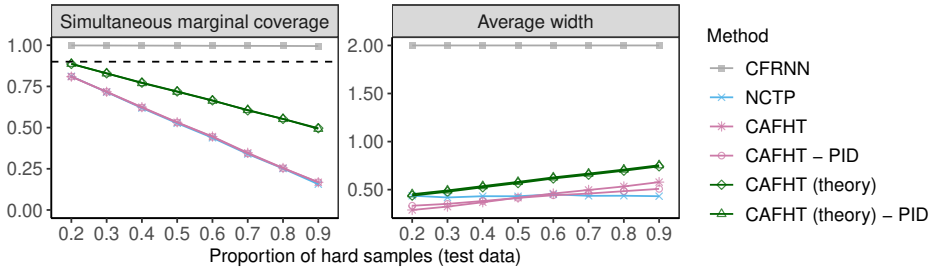

Figure A18. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods under distributional shift. The results are shown as a function of the proportion of hard-to-predict trajectories in the test data. See Table A20 for detailed results and standard errors.

Table A19. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the proportion of hard-to-predict trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A17.

			Simultaneous coverage		
Proportion of hard samples (all data)	Method	Average width	Conditional-hard	Conditional-easy	Marginal
0.1					
0.1	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
0.1	NCTP	0.435 (0.024)	0.063 (0.008)	0.996 (0.001)	0.904 (0.002)
0.1	CAFHT	0.241 (0.008)	0.069 (0.010)	0.993 (0.001)	0.901 (0.002)
0.1	CAFHT - PID	0.306 (0.019)	0.081 (0.012)	0.994 (0.001)	0.904 (0.002)
0.1	CAFHT (theory)	0.393 (0.004)	0.434 (0.011)	1.000 (0.000)	0.944 (0.001)
0.1	CAFHT (theory) - PID	0.407 (0.004)	0.452 (0.011)	1.000 (0.000)	0.946 (0.001)
0.2					
0.2	CFRNN	2.000 (0.000)	0.996 (0.001)	1.000 (0.000)	0.999 (0.000)
0.2	NCTP	0.720 (0.006)	0.496 (0.008)	1.000 (0.000)	0.898 (0.002)
0.2	CAFHT	0.435 (0.004)	0.496 (0.010)	1.000 (0.000)	0.899 (0.002)
0.2	CAFHT - PID	0.453 (0.004)	0.488 (0.010)	1.000 (0.000)	0.897 (0.002)
0.2	CAFHT (theory)	0.497 (0.004)	0.712 (0.007)	1.000 (0.000)	0.942 (0.001)
0.2	CAFHT (theory) - PID	0.522 (0.004)	0.718 (0.006)	1.000 (0.000)	0.943 (0.001)
0.5					
0.5	CFRNN	2.000 (0.000)	0.999 (0.000)	1.000 (0.000)	0.999 (0.000)
0.5	NCTP	0.776 (0.005)	0.803 (0.003)	1.000 (0.000)	0.901 (0.002)
0.5	CAFHT	0.614 (0.004)	0.805 (0.005)	1.000 (0.000)	0.902 (0.002)
0.5	CAFHT - PID	0.657 (0.005)	0.803 (0.004)	1.000 (0.000)	0.901 (0.002)
0.5	CAFHT (theory)	0.659 (0.005)	0.888 (0.003)	1.000 (0.000)	0.944 (0.002)
0.5	CAFHT (theory) - PID	0.701 (0.005)	0.883 (0.003)	1.000 (0.000)	0.941 (0.002)

Table A20. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods under distributional shift. The results are shown as a function of the proportion of hard-to-predict trajectories in the test data. The red numbers indicate higher marginal coverage. See the corresponding plot in Figure A18.

Proportion of hard samples (test data)		Method	Length	Marginal coverage
0.2				
	0.2	CFRNN	2.000 (0.000)	0.999 (0.000)
	0.2	NCTP	0.437 (0.028)	0.811 (0.003)
	0.2	CAFHT	0.289 (0.010)	0.809 (0.003)
	0.2	CAFHT - PID	0.333 (0.021)	0.810 (0.004)
	0.2	CAFHT (theory)	0.438 (0.005)	0.886 (0.002)
	0.2	CAFHT (theory) - PID	0.453 (0.005)	0.887 (0.003)
0.3				
	0.3	CFRNN	2.000 (0.000)	0.998 (0.000)
	0.3	NCTP	0.419 (0.027)	0.715 (0.004)
	0.3	CAFHT	0.324 (0.010)	0.717 (0.004)
	0.3	CAFHT - PID	0.352 (0.020)	0.717 (0.005)
	0.3	CAFHT (theory)	0.478 (0.005)	0.829 (0.003)
	0.3	CAFHT (theory) - PID	0.492 (0.005)	0.829 (0.004)
0.4				
	0.4	CFRNN	2.000 (0.000)	0.998 (0.000)
	0.4	NCTP	0.432 (0.026)	0.619 (0.004)
	0.4	CAFHT	0.369 (0.010)	0.624 (0.005)
	0.4	CAFHT - PID	0.382 (0.019)	0.622 (0.005)
	0.4	CAFHT (theory)	0.524 (0.005)	0.773 (0.004)
	0.4	CAFHT (theory) - PID	0.537 (0.005)	0.770 (0.004)
0.5				
	0.5	CFRNN	2.000 (0.000)	0.997 (0.000)
	0.5	NCTP	0.431 (0.027)	0.526 (0.005)
	0.5	CAFHT	0.417 (0.011)	0.532 (0.006)
	0.5	CAFHT - PID	0.414 (0.018)	0.531 (0.006)
	0.5	CAFHT (theory)	0.568 (0.005)	0.719 (0.005)
	0.5	CAFHT (theory) - PID	0.581 (0.006)	0.718 (0.005)
0.6				
	0.6	CFRNN	2.000 (0.000)	0.996 (0.001)
	0.6	NCTP	0.451 (0.028)	0.438 (0.006)
	0.6	CAFHT	0.460 (0.012)	0.446 (0.008)
	0.6	CAFHT - PID	0.441 (0.020)	0.444 (0.008)
	0.6	CAFHT (theory)	0.616 (0.006)	0.665 (0.006)
	0.6	CAFHT (theory) - PID	0.627 (0.006)	0.663 (0.007)
0.7				
	0.7	CFRNN	2.000 (0.000)	0.996 (0.001)
	0.7	NCTP	0.436 (0.027)	0.340 (0.006)
	0.7	CAFHT	0.497 (0.012)	0.347 (0.008)
	0.7	CAFHT - PID	0.459 (0.020)	0.343 (0.008)
	0.7	CAFHT (theory)	0.655 (0.007)	0.605 (0.007)
	0.7	CAFHT (theory) - PID	0.667 (0.007)	0.607 (0.007)
0.8				
	0.8	CFRNN	2.000 (0.000)	0.996 (0.001)
	0.8	NCTP	0.437 (0.027)	0.252 (0.007)
	0.8	CAFHT	0.533 (0.012)	0.255 (0.009)
	0.8	CAFHT - PID	0.484 (0.020)	0.253 (0.009)
	0.8	CAFHT (theory)	0.697 (0.007)	0.553 (0.008)
	0.8	CAFHT (theory) - PID	0.709 (0.007)	0.553 (0.009)
0.9				
	0.9	CFRNN	2.000 (0.000)	0.995 (0.001)
	0.9	NCTP	0.432 (0.027)	0.157 (0.007)
	0.9	CAFHT	0.578 (0.013)	0.168 (0.010)
	0.9	CAFHT - PID	0.508 (0.022)	0.168 (0.011)
	0.9	CAFHT (theory)	0.742 (0.007)	0.496 (0.009)
	0.9	CAFHT (theory) - PID	0.754 (0.007)	0.493 (0.010)

A5.2. Pedestrian Data

In this subsection, we present the experimental results of the pedestrian data described in 4. Recall that we preprocess the dataset by adding heteroskedasticity such that 10% of the data are designed to be hard-to-predict by adding a random noise follows $N(0, \sigma_t^2)$, where $\sigma_t^2 \propto t \cdot$ noise level. The easy-to-predict data are added a random noise with $\sigma_t^2 \propto t$. By default, 10% of the trajectories are set to be hard-to-predict.

Similar to the previous section, we first demonstrate the main result by using the CAFHT method with ACI and multiplicative scores as the main method to be compared with the benchmark methods CFRNN and NCTP. The results after adding the dynamic noise profile to the data are presented here for demonstrative purposes.

A5.2.1. Main Results — Comparing CFRNN, NCTP, and the Main Implementation of CAFHT

Figure 5 and Table A21 show the average performance on pedestrian heterogeneous trajectories of prediction bands constructed by different methods, as a function of the noise level. The noise level is varied from 1.5 to 5. All methods achieve 90% simultaneous marginal coverage. Our method (CAFHT) leads to more informative bands with lower average width and higher conditional coverage.

The results of another experiment in which 20% of the trajectories are hard-to-predict are presented in Figure A19 and Table A22. Again, we observe that even though a larger percentage of hard trajectories on the pedestrian data can increase the empirical conditional coverage of all methods, CAFHT maintains clear advantages relative to the baselines.

Additionally, Figure A20 and Table A23 present results for varying numbers of trajectories in the training and calibration sets, from 200 to 1000, with the noise level set at 3 and the percentage of hard trajectories set to 10%. Again, CAFHT outperforms the other benchmarks.

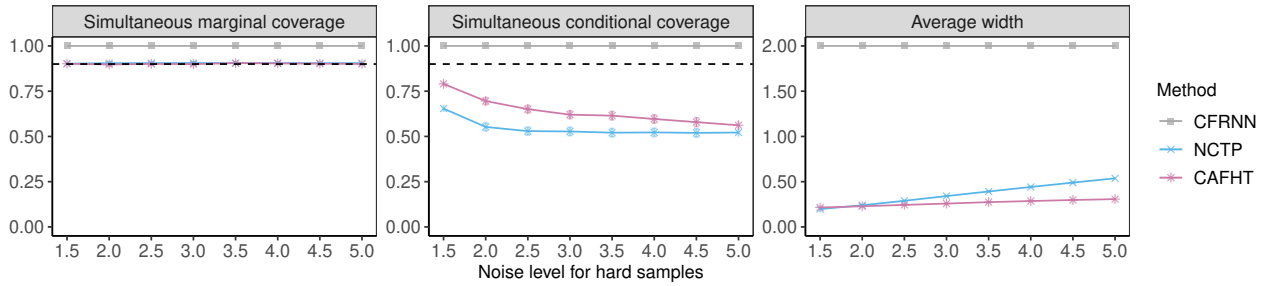

Figure A19. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the noise level. 20% of the trajectories are set to be hard-to-predict.

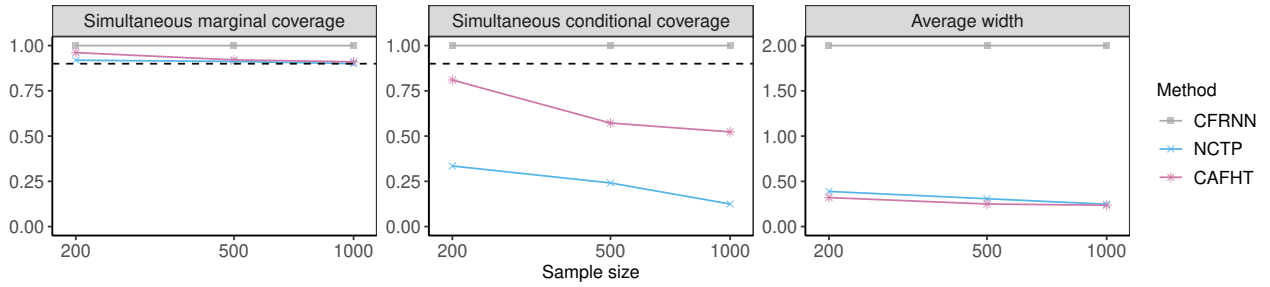

Figure A20. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories.

Table A21. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the noise level. The red numbers indicate smaller prediction bands or higher conditional coverage. 10% of the trajectories are set to be hard-to-predict. See the corresponding plot in Figure 5.

			Coverage		
Noise level for hard samples	Method	Length	Marginal	Conditional-easy	Conditional-hard
1					
1.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
1.0	NCTP	0.172 (0.001)	0.898 (0.003)	0.898 (0.003)	0.901 (0.006)
1.0	CAFHT	0.201 (0.001)	0.902 (0.003)	0.901 (0.003)	0.912 (0.005)
1.5					
1.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
1.5	NCTP	0.185 (0.002)	0.902 (0.003)	0.941 (0.002)	0.561 (0.011)
1.5	CAFHT	0.208 (0.001)	0.903 (0.003)	0.918 (0.003)	0.776 (0.009)
2					
2.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
2.0	NCTP	0.204 (0.002)	0.903 (0.002)	0.973 (0.002)	0.296 (0.012)
2.0	CAFHT	0.216 (0.001)	0.903 (0.003)	0.932 (0.002)	0.649 (0.011)
2.5					
2.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
2.5	NCTP	0.224 (0.003)	0.900 (0.003)	0.984 (0.001)	0.171 (0.011)
2.5	CAFHT	0.225 (0.001)	0.904 (0.003)	0.944 (0.002)	0.556 (0.013)
3					
3.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
3.0	NCTP	0.247 (0.004)	0.900 (0.003)	0.989 (0.001)	0.125 (0.012)
3.0	CAFHT	0.232 (0.001)	0.903 (0.003)	0.949 (0.002)	0.492 (0.013)
3.5					
3.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
3.5	NCTP	0.270 (0.005)	0.900 (0.002)	0.992 (0.001)	0.097 (0.011)
3.5	CAFHT	0.239 (0.002)	0.903 (0.003)	0.955 (0.002)	0.445 (0.014)
4					
4.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
4.0	NCTP	0.295 (0.007)	0.900 (0.002)	0.993 (0.001)	0.089 (0.011)
4.0	CAFHT	0.244 (0.002)	0.902 (0.003)	0.958 (0.002)	0.411 (0.014)
4.5					
4.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
4.5	NCTP	0.318 (0.008)	0.900 (0.002)	0.993 (0.001)	0.088 (0.011)
4.5	CAFHT	0.250 (0.002)	0.901 (0.003)	0.961 (0.002)	0.383 (0.014)
5					
5.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
5.0	NCTP	0.341 (0.010)	0.900 (0.002)	0.993 (0.001)	0.092 (0.012)
5.0	CAFHT	0.257 (0.002)	0.901 (0.003)	0.963 (0.002)	0.364 (0.014)

Table A22. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the noise level. The red numbers indicate smaller prediction bands or higher conditional coverage. 20% of the trajectories are set to be hard-to-predict. See the corresponding plot in Figure A19.

				Coverage		
Noise level for hard samples		Method	Length	Marginal	Conditional-easy	Conditional-hard
1.5						
	1.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	1.5	NCTP	0.197 (0.002)	0.902 (0.003)	0.963 (0.002)	0.654 (0.010)
	1.5	CAFHT	0.214 (0.001)	0.902 (0.003)	0.930 (0.003)	0.790 (0.007)
2						
	2.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	2.0	NCTP	0.239 (0.002)	0.905 (0.003)	0.992 (0.001)	0.552 (0.010)
	2.0	CAFHT	0.228 (0.001)	0.897 (0.003)	0.948 (0.002)	0.695 (0.009)
2.5						
	2.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	2.5	NCTP	0.289 (0.002)	0.905 (0.002)	0.998 (0.000)	0.529 (0.011)
	2.5	CAFHT	0.243 (0.001)	0.901 (0.003)	0.963 (0.002)	0.651 (0.010)
3						
	3.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	3.0	NCTP	0.341 (0.003)	0.906 (0.002)	0.999 (0.000)	0.527 (0.011)
	3.0	CAFHT	0.258 (0.002)	0.899 (0.003)	0.969 (0.002)	0.620 (0.010)
3.5						
	3.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	3.5	NCTP	0.392 (0.003)	0.905 (0.002)	0.999 (0.000)	0.521 (0.011)
	3.5	CAFHT	0.274 (0.002)	0.905 (0.003)	0.977 (0.001)	0.615 (0.011)
4						
	4.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	4.0	NCTP	0.442 (0.004)	0.905 (0.002)	0.999 (0.000)	0.522 (0.011)
	4.0	CAFHT	0.286 (0.002)	0.903 (0.003)	0.979 (0.002)	0.596 (0.011)
4.5						
	4.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	4.5	NCTP	0.490 (0.004)	0.905 (0.002)	1.000 (0.000)	0.520 (0.011)
	4.5	CAFHT	0.298 (0.002)	0.902 (0.003)	0.982 (0.001)	0.579 (0.012)
5						
	5.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
	5.0	NCTP	0.537 (0.004)	0.905 (0.002)	1.000 (0.000)	0.522 (0.011)
	5.0	CAFHT	0.307 (0.002)	0.901 (0.003)	0.984 (0.001)	0.562 (0.012)

Table A23. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A20.

			Coverage		
Sample size	Method	Length	Marginal	Conditional-easy	Conditional-hard
200					
200	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
200	NCTP	0.388 (0.009)	0.919 (0.004)	0.986 (0.002)	0.335 (0.023)
200	CAFHT	0.321 (0.006)	0.961 (0.004)	0.979 (0.003)	0.810 (0.018)
500					
500	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
500	NCTP	0.308 (0.006)	0.912 (0.003)	0.989 (0.001)	0.241 (0.017)
500	CAFHT	0.249 (0.002)	0.921 (0.004)	0.961 (0.003)	0.572 (0.016)
1000					
1000	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
1000	NCTP	0.247 (0.004)	0.900 (0.003)	0.989 (0.001)	0.125 (0.012)
1000	CAFHT	0.236 (0.002)	0.911 (0.003)	0.955 (0.002)	0.523 (0.014)

A5.2.2. Supplementary Results — Comparing Different CAFHT Implementations

CAFHT - multiplicative scores

The results of CAFHT with multiplicative conformity scores (6) are first presented in Figures A21–A22 and Tables A24–A25.

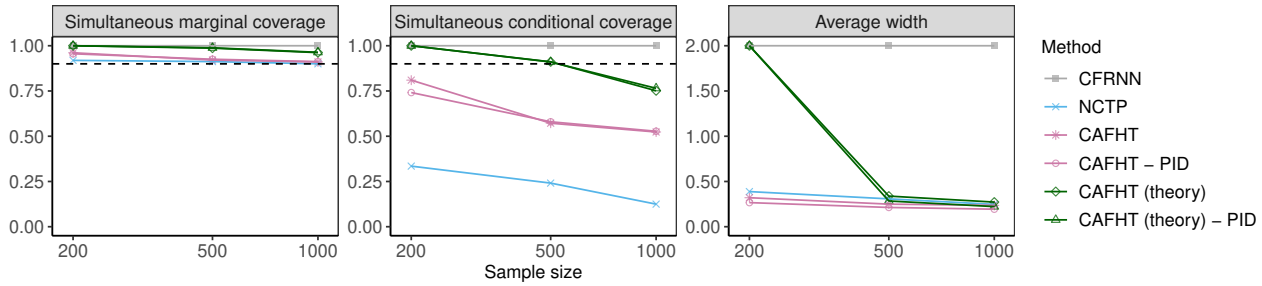

Figure A21. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories.

Table A24. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A21.

			Coverage		
Sample size	Method	Length	Marginal	Conditional-easy	Conditional-hard
200					
200	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
200	NCTP	0.388 (0.009)	0.919 (0.004)	0.986 (0.002)	0.335 (0.023)
200	CAFHT	0.321 (0.006)	0.961 (0.004)	0.979 (0.003)	0.810 (0.018)
200	CAFHT - PID	0.266 (0.005)	0.956 (0.005)	0.980 (0.003)	0.741 (0.022)
200	CAFHT (theory)	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
200	CAFHT (theory) - PID	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
500					
500	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
500	NCTP	0.308 (0.006)	0.912 (0.003)	0.989 (0.001)	0.241 (0.017)
500	CAFHT	0.249 (0.002)	0.921 (0.004)	0.961 (0.003)	0.572 (0.016)
500	CAFHT - PID	0.213 (0.003)	0.925 (0.004)	0.965 (0.003)	0.580 (0.018)
500	CAFHT (theory)	0.338 (0.004)	0.987 (0.001)	0.995 (0.001)	0.911 (0.008)
500	CAFHT (theory) - PID	0.284 (0.004)	0.989 (0.001)	0.998 (0.000)	0.910 (0.008)
1000					
1000	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
1000	NCTP	0.247 (0.004)	0.900 (0.003)	0.989 (0.001)	0.125 (0.012)
1000	CAFHT	0.236 (0.002)	0.911 (0.003)	0.955 (0.002)	0.523 (0.014)
1000	CAFHT - PID	0.194 (0.001)	0.912 (0.003)	0.956 (0.002)	0.527 (0.017)
1000	CAFHT (theory)	0.272 (0.001)	0.962 (0.002)	0.986 (0.001)	0.751 (0.010)
1000	CAFHT (theory) - PID	0.222 (0.001)	0.964 (0.001)	0.987 (0.001)	0.765 (0.009)

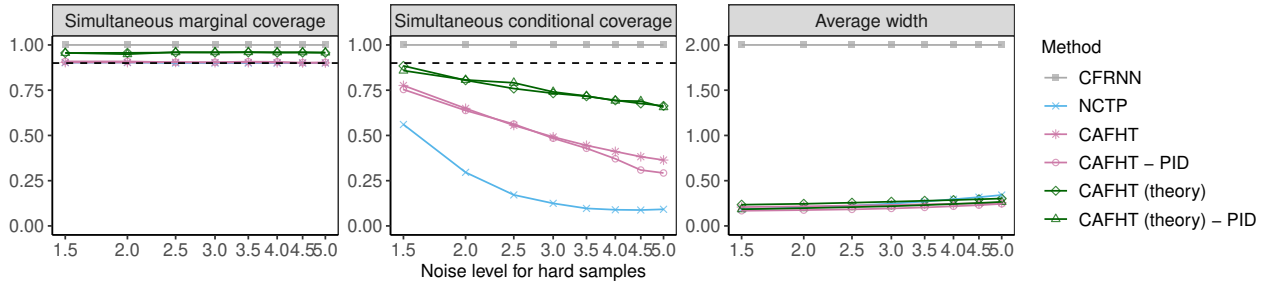

Figure A22. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the noise level.

Table A25. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the noise level. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A22.

				Coverage		
Noise level for hard samples	Method	Length	Marginal	Conditional-easy	Conditional-hard
1					
1.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
1.0	NCTP	0.172 (0.001)	0.898 (0.003)	0.898 (0.003)	0.901 (0.006)
1.0	CAFHT	0.201 (0.001)	0.902 (0.003)	0.901 (0.003)	0.912 (0.005)
1.0	CAFHT - PID	0.159 (0.001)	0.906 (0.003)	0.906 (0.003)	0.909 (0.005)
1.0	CAFHT (theory)	0.226 (0.001)	0.955 (0.002)	0.955 (0.002)	0.959 (0.004)
1.0	CAFHT (theory) - PID	0.174 (0.001)	0.957 (0.002)	0.956 (0.002)	0.960 (0.004)
1.5					
1.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
1.5	NCTP	0.185 (0.002)	0.902 (0.003)	0.941 (0.002)	0.561 (0.011)
1.5	CAFHT	0.208 (0.001)	0.903 (0.003)	0.918 (0.003)	0.776 (0.009)
1.5	CAFHT - PID	0.168 (0.001)	0.909 (0.003)	0.927 (0.003)	0.754 (0.011)
1.5	CAFHT (theory)	0.235 (0.001)	0.956 (0.002)	0.964 (0.001)	0.884 (0.006)
1.5	CAFHT (theory) - PID	0.184 (0.001)	0.956 (0.002)	0.967 (0.001)	0.858 (0.008)
2					
2.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
2.0	NCTP	0.204 (0.002)	0.903 (0.002)	0.973 (0.002)	0.296 (0.012)
2.0	CAFHT	0.216 (0.001)	0.903 (0.003)	0.932 (0.002)	0.649 (0.011)
2.0	CAFHT - PID	0.175 (0.001)	0.909 (0.003)	0.940 (0.002)	0.638 (0.015)
2.0	CAFHT (theory)	0.245 (0.001)	0.957 (0.002)	0.974 (0.001)	0.805 (0.008)
2.0	CAFHT (theory) - PID	0.197 (0.001)	0.949 (0.010)	0.965 (0.010)	0.807 (0.014)
2.5					
2.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
2.5	NCTP	0.224 (0.003)	0.900 (0.003)	0.984 (0.001)	0.171 (0.011)
2.5	CAFHT	0.225 (0.001)	0.904 (0.003)	0.944 (0.002)	0.556 (0.013)
2.5	CAFHT - PID	0.183 (0.001)	0.905 (0.003)	0.944 (0.003)	0.563 (0.018)
2.5	CAFHT (theory)	0.257 (0.001)	0.958 (0.002)	0.981 (0.001)	0.759 (0.009)
2.5	CAFHT (theory) - PID	0.208 (0.001)	0.960 (0.001)	0.980 (0.001)	0.791 (0.008)
3					
3.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
3.0	NCTP	0.247 (0.004)	0.900 (0.003)	0.989 (0.001)	0.125 (0.012)
3.0	CAFHT	0.232 (0.001)	0.903 (0.003)	0.949 (0.002)	0.492 (0.013)
3.0	CAFHT - PID	0.193 (0.002)	0.905 (0.003)	0.953 (0.003)	0.485 (0.018)
3.0	CAFHT (theory)	0.268 (0.001)	0.958 (0.002)	0.984 (0.001)	0.733 (0.010)
3.0	CAFHT (theory) - PID	0.218 (0.001)	0.960 (0.001)	0.985 (0.001)	0.741 (0.009)
3.5					
3.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
3.5	NCTP	0.270 (0.005)	0.900 (0.002)	0.992 (0.001)	0.097 (0.011)
3.5	CAFHT	0.239 (0.002)	0.903 (0.003)	0.955 (0.002)	0.445 (0.014)
3.5	CAFHT - PID	0.204 (0.003)	0.908 (0.003)	0.963 (0.002)	0.429 (0.019)
3.5	CAFHT (theory)	0.278 (0.001)	0.959 (0.002)	0.987 (0.001)	0.717 (0.011)
3.5	CAFHT (theory) - PID	0.230 (0.001)	0.960 (0.001)	0.988 (0.001)	0.718 (0.010)
4					
4.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
4.0	NCTP	0.295 (0.007)	0.900 (0.002)	0.993 (0.001)	0.089 (0.011)
4.0	CAFHT	0.244 (0.002)	0.902 (0.003)	0.958 (0.002)	0.411 (0.014)
4.0	CAFHT - PID	0.218 (0.004)	0.907 (0.003)	0.968 (0.002)	0.371 (0.019)
4.0	CAFHT (theory)	0.286 (0.002)	0.958 (0.002)	0.988 (0.001)	0.694 (0.012)
4.0	CAFHT (theory) - PID	0.242 (0.001)	0.960 (0.002)	0.991 (0.001)	0.692 (0.011)
4.5					
4.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
4.5	NCTP	0.318 (0.008)	0.900 (0.002)	0.993 (0.001)	0.088 (0.011)
4.5	CAFHT	0.250 (0.002)	0.901 (0.003)	0.961 (0.002)	0.383 (0.014)
4.5	CAFHT - PID	0.229 (0.005)	0.902 (0.003)	0.970 (0.003)	0.309 (0.016)
4.5	CAFHT (theory)	0.295 (0.002)	0.958 (0.002)	0.990 (0.001)	0.677 (0.012)
4.5	CAFHT (theory) - PID	0.254 (0.002)	0.960 (0.002)	0.991 (0.001)	0.689 (0.013)
5					
5.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
5.0	NCTP	0.341 (0.010)	0.900 (0.002)	0.993 (0.001)	0.092 (0.012)
5.0	CAFHT	0.257 (0.002)	0.901 (0.003)	0.963 (0.002)	0.364 (0.014)
5.0	CAFHT - PID	0.244 (0.007)	0.904 (0.003)	0.974 (0.002)	0.293 (0.017)
5.0	CAFHT (theory)	0.303 (0.002)	0.957 (0.002)	0.991 (0.001)	0.663 (0.011)
5.0	CAFHT (theory) - PID	0.266 (0.002)	0.958 (0.002)	0.993 (0.001)	0.656 (0.013)

CAFHT — Additive Scores

The results of CAFHT with additive conformity scores (3) are presented in Figures A23–A24 and Tables A26–A27.

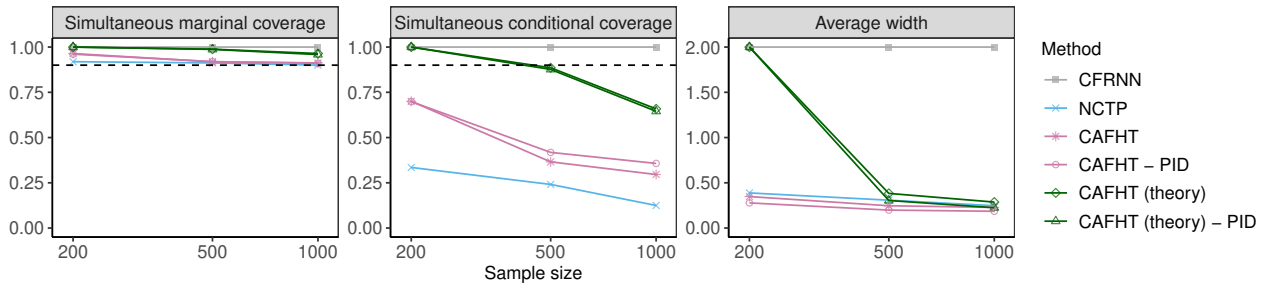

Figure A23. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories.

Table A26. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A23.

			Coverage		
Sample size	Method	Length	Marginal	Conditional-easy	Conditional-hard
200					
200	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
200	NCTP	0.388 (0.009)	0.919 (0.004)	0.986 (0.002)	0.335 (0.023)
200	CAFHT	0.348 (0.009)	0.964 (0.003)	0.994 (0.001)	0.700 (0.024)
200	CAFHT - PID	0.278 (0.008)	0.961 (0.004)	0.991 (0.002)	0.699 (0.025)
200	CAFHT (theory)	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
200	CAFHT (theory) - PID	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
500					
500	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
500	NCTP	0.308 (0.006)	0.912 (0.003)	0.989 (0.001)	0.241 (0.017)
500	CAFHT	0.246 (0.003)	0.919 (0.004)	0.982 (0.002)	0.366 (0.019)
500	CAFHT - PID	0.198 (0.002)	0.920 (0.004)	0.977 (0.002)	0.418 (0.018)
500	CAFHT (theory)	0.383 (0.005)	0.988 (0.001)	1.000 (0.000)	0.886 (0.010)
500	CAFHT (theory) - PID	0.304 (0.005)	0.987 (0.001)	1.000 (0.000)	0.876 (0.011)
1000					
1000	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
1000	NCTP	0.247 (0.004)	0.900 (0.003)	0.989 (0.001)	0.125 (0.012)
1000	CAFHT	0.228 (0.002)	0.912 (0.003)	0.983 (0.002)	0.296 (0.015)
1000	CAFHT - PID	0.185 (0.001)	0.911 (0.003)	0.974 (0.002)	0.357 (0.015)
1000	CAFHT (theory)	0.287 (0.002)	0.964 (0.002)	0.999 (0.000)	0.658 (0.013)
1000	CAFHT (theory) - PID	0.225 (0.002)	0.959 (0.002)	0.994 (0.001)	0.645 (0.013)

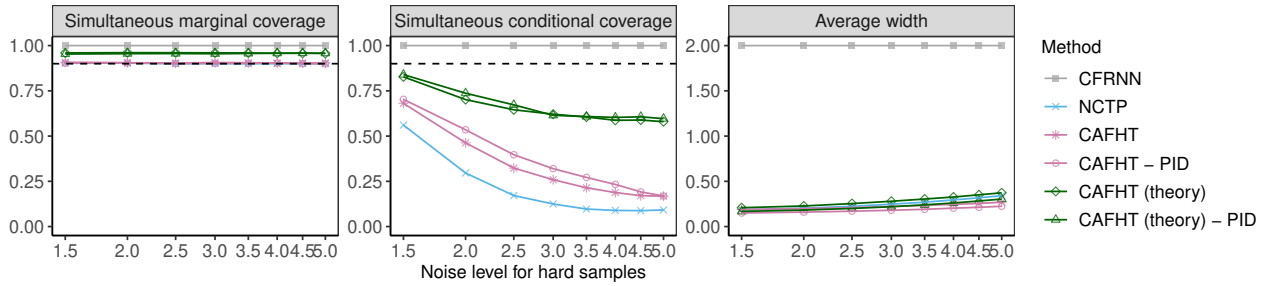

Figure A24. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the noise level.

Table A27. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the noise level. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A24.

				Coverage		
Noise level for hard samples	Method	Length	Marginal	Conditional-easy	Conditional-hard
1					
1.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
1.0	NCTP	0.172 (0.001)	0.898 (0.003)	0.898 (0.003)	0.901 (0.006)
1.0	CAFHT	0.180 (0.001)	0.907 (0.003)	0.907 (0.004)	0.907 (0.005)
1.0	CAFHT - PID	0.141 (0.001)	0.901 (0.003)	0.901 (0.003)	0.901 (0.006)
1.0	CAFHT (theory)	0.197 (0.001)	0.958 (0.002)	0.958 (0.002)	0.961 (0.003)
1.0	CAFHT (theory) - PID	0.158 (0.001)	0.955 (0.002)	0.954 (0.002)	0.957 (0.004)
1.5					
1.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
1.5	NCTP	0.185 (0.002)	0.902 (0.003)	0.941 (0.002)	0.561 (0.011)
1.5	CAFHT	0.188 (0.001)	0.908 (0.003)	0.934 (0.003)	0.681 (0.010)
1.5	CAFHT - PID	0.150 (0.001)	0.903 (0.003)	0.926 (0.003)	0.703 (0.010)
1.5	CAFHT (theory)	0.208 (0.001)	0.960 (0.002)	0.975 (0.001)	0.827 (0.007)
1.5	CAFHT (theory) - PID	0.167 (0.001)	0.953 (0.002)	0.967 (0.001)	0.840 (0.007)
2					
2.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
2.0	NCTP	0.204 (0.002)	0.903 (0.002)	0.973 (0.002)	0.296 (0.012)
2.0	CAFHT	0.199 (0.001)	0.906 (0.003)	0.957 (0.002)	0.463 (0.011)
2.0	CAFHT - PID	0.160 (0.001)	0.904 (0.003)	0.946 (0.002)	0.535 (0.012)
2.0	CAFHT (theory)	0.228 (0.001)	0.961 (0.002)	0.990 (0.001)	0.702 (0.011)
2.0	CAFHT (theory) - PID	0.182 (0.001)	0.955 (0.002)	0.980 (0.001)	0.737 (0.010)
2.5					
2.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
2.5	NCTP	0.224 (0.003)	0.900 (0.003)	0.984 (0.001)	0.171 (0.011)
2.5	CAFHT	0.209 (0.001)	0.904 (0.003)	0.971 (0.002)	0.323 (0.013)
2.5	CAFHT - PID	0.170 (0.001)	0.901 (0.003)	0.959 (0.002)	0.397 (0.012)
2.5	CAFHT (theory)	0.254 (0.002)	0.961 (0.002)	0.996 (0.000)	0.645 (0.012)
2.5	CAFHT (theory) - PID	0.200 (0.001)	0.956 (0.002)	0.988 (0.001)	0.673 (0.010)
3					
3.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
3.0	NCTP	0.247 (0.004)	0.900 (0.003)	0.989 (0.001)	0.125 (0.012)
3.0	CAFHT	0.222 (0.002)	0.907 (0.003)	0.981 (0.002)	0.259 (0.014)
3.0	CAFHT - PID	0.180 (0.001)	0.902 (0.003)	0.969 (0.002)	0.320 (0.014)
3.0	CAFHT (theory)	0.280 (0.002)	0.960 (0.002)	0.999 (0.000)	0.621 (0.013)
3.0	CAFHT (theory) - PID	0.219 (0.001)	0.955 (0.002)	0.993 (0.001)	0.615 (0.013)
3.5					
3.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
3.5	NCTP	0.270 (0.005)	0.900 (0.002)	0.992 (0.001)	0.097 (0.011)
3.5	CAFHT	0.233 (0.002)	0.905 (0.003)	0.985 (0.002)	0.214 (0.014)
3.5	CAFHT - PID	0.191 (0.002)	0.902 (0.003)	0.975 (0.002)	0.271 (0.015)
3.5	CAFHT (theory)	0.305 (0.003)	0.959 (0.002)	0.999 (0.000)	0.605 (0.014)
3.5	CAFHT (theory) - PID	0.240 (0.002)	0.956 (0.002)	0.996 (0.001)	0.608 (0.013)
4					
4.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
4.0	NCTP	0.295 (0.007)	0.900 (0.002)	0.993 (0.001)	0.089 (0.011)
4.0	CAFHT	0.246 (0.003)	0.904 (0.003)	0.986 (0.002)	0.188 (0.015)
4.0	CAFHT - PID	0.203 (0.003)	0.904 (0.003)	0.981 (0.002)	0.233 (0.015)
4.0	CAFHT (theory)	0.326 (0.003)	0.958 (0.002)	1.000 (0.000)	0.587 (0.015)
4.0	CAFHT (theory) - PID	0.261 (0.002)	0.958 (0.002)	0.998 (0.000)	0.603 (0.013)
4.5					
4.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
4.5	NCTP	0.318 (0.008)	0.900 (0.002)	0.993 (0.001)	0.088 (0.011)
4.5	CAFHT	0.255 (0.004)	0.904 (0.003)	0.988 (0.002)	0.171 (0.016)
4.5	CAFHT - PID	0.213 (0.004)	0.901 (0.003)	0.982 (0.002)	0.192 (0.015)
4.5	CAFHT (theory)	0.351 (0.003)	0.958 (0.002)	1.000 (0.000)	0.588 (0.014)
4.5	CAFHT (theory) - PID	0.283 (0.002)	0.959 (0.002)	0.999 (0.000)	0.606 (0.014)
5					
5.0	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
5.0	NCTP	0.341 (0.010)	0.900 (0.002)	0.993 (0.001)	0.092 (0.012)
5.0	CAFHT	0.269 (0.004)	0.904 (0.003)	0.989 (0.002)	0.168 (0.017)
5.0	CAFHT - PID	0.224 (0.004)	0.899 (0.003)	0.983 (0.002)	0.169 (0.016)
5.0	CAFHT (theory)	0.373 (0.003)	0.957 (0.002)	1.000 (0.000)	0.580 (0.014)
5.0	CAFHT (theory) - PID	0.304 (0.003)	0.958 (0.002)	0.999 (0.000)	0.596 (0.015)

A5.3. Comparing ACI and CAFHT

As previously explained, the objectives of CAFHT and ACI are very different. CAFHT leverages information from multiple exchangeable trajectories to construct prediction bands for a trajectories from the same population, ensuring simultaneous coverage as per Equation (1). By contrast, ACI constructs an online prediction band for a single trajectory, aiming to achieve long-term average coverage.

Consider a motion planning scenario: CAFHT's objective is to maintain most vehicles within their predicted zones throughout a specified period, ensuring a high probability of reaching their destinations without incident. On the other hand, ACI aims for asymptotic average coverage, which tolerates frequent, albeit temporary, deviations from the predicted path for each vehicle. In practical terms, this means each vehicle might exit and re-enter the ACI-predicted region numerous times, spending about 90% of the time within the prediction band on average. If exiting these regions could lead to severe accidents, CAFHT's approach would ensure that 90% (or any pre-specified percentage) of vehicles safely arrive at their destinations, whereas ACI's approach could potentially result in none of the vehicles reaching their destinations safely.

This concept is demonstrated in Figure A25, which contrasts the prediction bands created using ACI and CAFHT for two pedestrian trajectories. The figure clearly shows that ACI does not fully encompass the trajectories, thus failing to meet our objective of achieving simultaneous coverage.

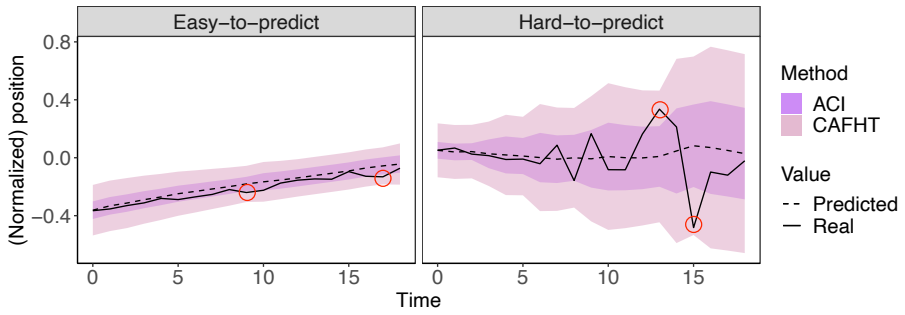

Figure A25. Forecasting bands constructed using ACI and CAFHT, for the heterogeneous pedestrian trajectories. Red circles indicate scenarios where the real values exceed ACI prediction bands.

Figure A26 and Table A28 provide additional insight, reporting on experiments that replicate the analysis from Figure 3 but include results from ACI. Unlike CAFHT and the two other benchmark methods, ACI is unable to meet the simultaneous marginal coverage guarantee.

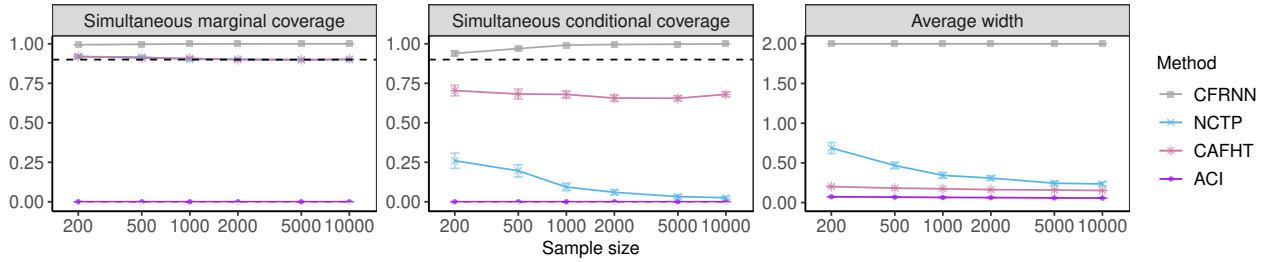

Figure A26. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories.

Table A28. Performance on heterogeneous pedestrian trajectories of conformal prediction bands constructed by different methods, as a function of the noise level. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A26.

			Simultaneous coverage		
Sample size	Method	Average width	Conditional-hard	Conditional-easy	Marginal
200					
200	CFRNN	2.000 (0.000)	0.939 (0.007)	1.000 (0.000)	0.994 (0.001)
200	NCTP	0.687 (0.035)	0.260 (0.024)	0.992 (0.002)	0.919 (0.004)
200	CAFHT	0.202 (0.003)	0.704 (0.017)	0.944 (0.005)	0.920 (0.005)
200	ACI	0.073 (0.001)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
500					
500	CFRNN	2.000 (0.000)	0.969 (0.005)	1.000 (0.000)	0.997 (0.000)
500	NCTP	0.467 (0.021)	0.196 (0.019)	0.994 (0.002)	0.916 (0.003)
500	CAFHT	0.182 (0.002)	0.682 (0.016)	0.934 (0.003)	0.910 (0.004)
500	ACI	0.069 (0.001)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
1000					
1000	CFRNN	2.000 (0.000)	0.992 (0.002)	1.000 (0.000)	0.999 (0.000)
1000	NCTP	0.343 (0.018)	0.093 (0.012)	0.992 (0.001)	0.901 (0.003)
1000	CAFHT	0.174 (0.001)	0.679 (0.012)	0.934 (0.003)	0.908 (0.003)
1000	ACI	0.065 (0.001)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
2000					
2000	CFRNN	2.000 (0.000)	0.995 (0.001)	1.000 (0.000)	1.000 (0.000)
2000	NCTP	0.308 (0.014)	0.060 (0.008)	0.996 (0.001)	0.903 (0.002)
2000	CAFHT	0.163 (0.001)	0.656 (0.010)	0.926 (0.002)	0.899 (0.003)
2000	ACI	0.063 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
5000					
5000	CFRNN	2.000 (0.000)	0.998 (0.001)	1.000 (0.000)	1.000 (0.000)
5000	NCTP	0.244 (0.013)	0.033 (0.006)	0.997 (0.001)	0.900 (0.002)
5000	CAFHT	0.158 (0.001)	0.655 (0.007)	0.925 (0.002)	0.899 (0.002)
5000	ACI	0.059 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
10000					
10000	CFRNN	2.000 (0.000)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)
10000	NCTP	0.235 (0.011)	0.026 (0.004)	0.998 (0.000)	0.900 (0.001)
10000	CAFHT	0.152 (0.001)	0.680 (0.007)	0.928 (0.001)	0.903 (0.002)
10000	ACI	0.057 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

A5.4. Comparing the Multiplicative Scores and the Additive Scores

The CAFHT prediction bands are constructed in two stages: initially, the underlying ACI bands are established, followed by adding a conformalized correction term. When corrections employ the additive nonconformity scores specified in Equation (A20), the heterogeneity of the trajectories is managed exclusively via the ACI bands. In contrast, using the multiplicative scores from Equation (A23) allows both components to adapt to heteroscedasticity, though the primary adjustment is through ACI.

More precisely, multiplicative scores impose proportionally wider margins of error on broader ACI intervals than on narrower ones. Hence, while adjusting for heteroscedasticity is chiefly the responsibility of ACI, the use of multiplicative scores arises from the recognition that ACI residuals might still display heteroscedastic traits. In such instances, multiplicative scores are better suited to capturing this variability than their additive counterparts.

As shown in Figure A27, additive scores impose a constant correction term (the empirical quantile $\hat{Q}$) on ACI intervals. In comparison, multiplicative scores adjust the ACI bands by a non-constant amount (the empirical quantile $\hat{Q}$ multiplied by the size of the ACI bands).

In line with established conformal inference methodologies, we prefer to delegate the more complex "adaptability" functions to the underlying machine learning model (in this case, the forecaster integrated with ACI). The next phase of conformalization simply involves a clear, straightforward adjustment to secure the simultaneous marginal coverage guarantee. Nonetheless, future developments might introduce more intricate scoring designs, potentially enhancing empirical performance but at the expense of simplicity in the methodology.

Figure A28 presents a side-by-side comparison of CAFHT using multiplicative scores, CAFHT using additive scores, NCTP, and CFRNN for two example heterogeneous pedestrian trajectories. The plot demonstrates CAFHT, with both scoring approaches, effectively manages heterogeneity, though the multiplicative scores offer superior adaptability. In contrast, NCTP and CFRNN do not adjust to heterogeneity. The empirical quantile $\hat{Q}$ for this experiment is recorded in Table A29.

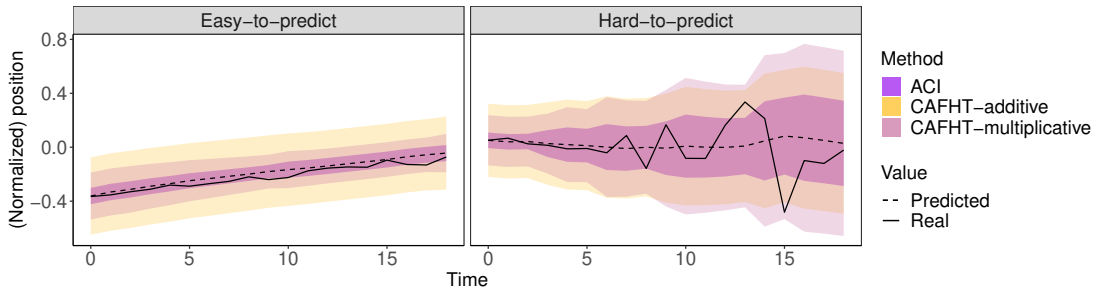

Figure A27. Forecasting bands constructed using ACI and CAFHT, for the heterogeneous pedestrian trajectories. Red circles indicate scenarios where the real values exceed ACI prediction bands.

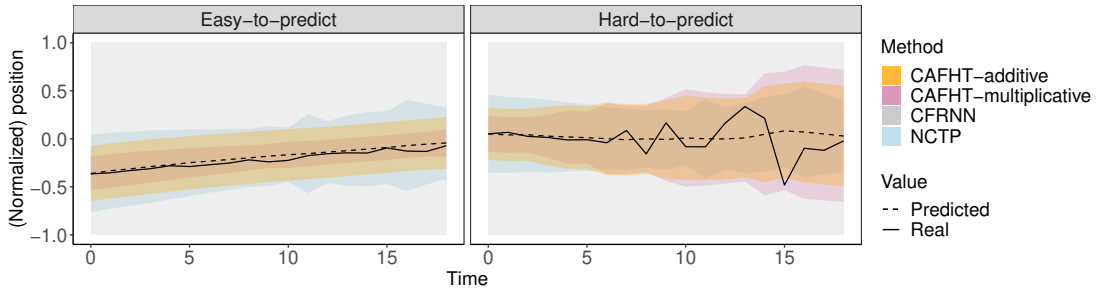

Figure A28. Forecasting bands constructed using different methods for the heterogeneous pedestrian trajectories.

Table A29. Empirical quantiles obtained from each method in Figure A28.

Method	Empirical quantile $\hat{Q}$	Remark
CFRNN	∞	$\hat{Q} = \infty$ for every time step.
NCTP	0.5268	$\hat{Q}$ is multiplied with the standard error at each time step before adding or subtracting from the point prediction.
CAFHT - multiplicative	0.5883	$\hat{Q}$ is multiplied with the width of the ACI bands at each time step before adding or subtracting from the ACI bands.
CAFHT - additive	0.2058	$\hat{Q}$ is directly added or subtracted from the ACI bands.

A5.5. Prediction Bands at Higher Coverage Levels

This section presents the results of additional experiments conducted using $\alpha = 0.05$ and $\alpha = 0.01$, seeking simultaneous coverage at the 95% level and the 99% level respectively. We continue to use the main implementation of the CAFHT method, which utilizes multiplicative scores based on the ACI algorithm and optimizes the learning rate through data splitting.

When higher coverage levels are employed, it is necessary to increase the number of samples in the calibration data to ensure that the adjusted empirical quantile level $(1-\alpha)(1-1/|\mathcal{D}_{\text{cal}}|)$ remains below 1. In our experiments, we cap the adjusted level at 1 whenever it exceeds this value.

EXPERIMENTS WITH 95% COVERAGE LEVEL

Figure A29 shows that all methods achieve 95% simultaneous marginal coverage. Our method (CAFHT) leads to more informative bands with lower average width and higher conditional coverage.

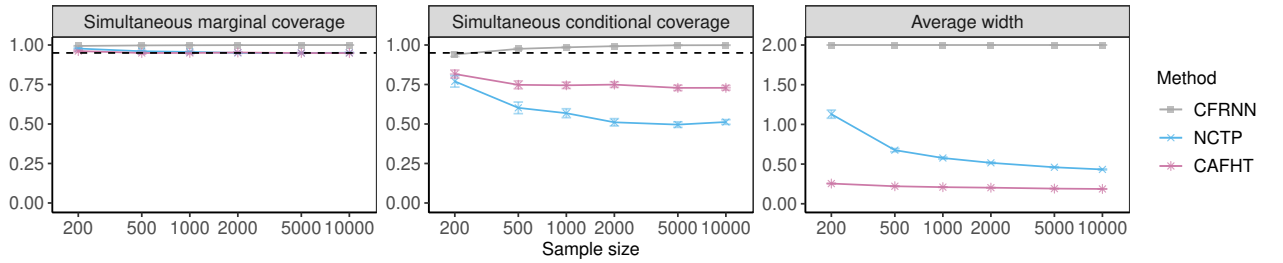

Figure A29. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories (25% are randomly assigned to calibration set). The target simultaneous marginal coverage level is 95%. See Table A30 for detailed results and standard errors.

Table A30. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. Target simultaneous marginal coverage level is 95%. See corresponding plot in Figure A29.

			Simultaneous coverage		
Sample size	Method	Average width	Conditional-hard	Conditional-easy	Marginal
200					
200	CFRNN	2.000 (0.000)	0.940 (0.006)	1.000 (0.000)	0.994 (0.001)
200	NCTP	1.129 (0.026)	0.770 (0.018)	1.000 (0.000)	0.978 (0.002)
200	CAFHT	0.255 (0.003)	0.816 (0.013)	0.978 (0.003)	0.962 (0.003)
500					
500	CFRNN	2.000 (0.000)	0.976 (0.003)	1.000 (0.000)	0.998 (0.000)
500	NCTP	0.676 (0.010)	0.602 (0.018)	1.000 (0.000)	0.961 (0.002)
500	CAFHT	0.221 (0.002)	0.748 (0.013)	0.972 (0.002)	0.949 (0.003)
1000					
1000	CFRNN	2.000 (0.000)	0.985 (0.002)	1.000 (0.000)	0.999 (0.000)
1000	NCTP	0.577 (0.007)	0.568 (0.014)	1.000 (0.000)	0.956 (0.002)
1000	CAFHT	0.209 (0.002)	0.745 (0.010)	0.973 (0.002)	0.950 (0.002)
2000					
2000	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
2000	NCTP	0.516 (0.005)	0.511 (0.012)	1.000 (0.000)	0.950 (0.001)
2000	CAFHT	0.203 (0.001)	0.749 (0.009)	0.976 (0.001)	0.953 (0.002)
5000					
5000	CFRNN	2.000 (0.000)	0.998 (0.001)	1.000 (0.000)	1.000 (0.000)
5000	NCTP	0.460 (0.003)	0.496 (0.009)	1.000 (0.000)	0.949 (0.001)
5000	CAFHT	0.191 (0.001)	0.729 (0.008)	0.974 (0.001)	0.949 (0.001)
10000					
10000	CFRNN	2.000 (0.000)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)
10000	NCTP	0.433 (0.004)	0.513 (0.007)	1.000 (0.000)	0.951 (0.001)
10000	CAFHT	0.187 (0.001)	0.729 (0.007)	0.974 (0.001)	0.949 (0.001)

EXPERIMENTS WITH 99% COVERAGE LEVEL

When seeking a 99% coverage level, using a relatively small sample size will result in the adjusted level being very close to, or equal to, 1, mapping the empirical quantile $\hat{Q}$ to infinity. Consequently, as depicted in Figure A30, NCTP and CAFHT generate regions that span the entire space $[-1, 1]$ when the sample size is small. CAFHT requires slightly more calibration samples than NCTP to produce practically useful prediction regions when employing a data-splitting strategy. When the prediction bands are practically useful, CAFHT tends to produce narrower and thus more informative results compared to NCTP while maintaining similarly high conditional coverage.

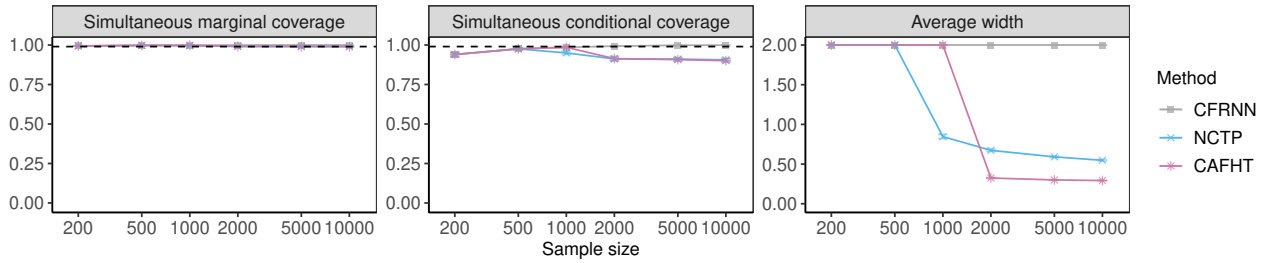

Figure A30. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. Target simultaneous marginal coverage level is 99%. See Table A31 for detailed results and standard errors.

Table A31. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. Target simultaneous marginal coverage level is 99%. See corresponding plot in Figure A30.

			Simultaneous coverage		
Sample size	Method	Average width	Conditional-hard	Conditional-easy	Marginal
200					
200	CFRNN	2.000 (0.000)	0.940 (0.006)	1.000 (0.000)	0.994 (0.001)
200	NCTP	2.000 (0.000)	0.940 (0.006)	1.000 (0.000)	0.994 (0.001)
200	CAFHT	2.000 (0.000)	0.940 (0.006)	1.000 (0.000)	0.994 (0.001)
500					
500	CFRNN	2.000 (0.000)	0.976 (0.003)	1.000 (0.000)	0.998 (0.000)
500	NCTP	2.000 (0.000)	0.976 (0.003)	1.000 (0.000)	0.998 (0.000)
500	CAFHT	2.000 (0.000)	0.976 (0.003)	1.000 (0.000)	0.998 (0.000)
1000					
1000	CFRNN	2.000 (0.000)	0.985 (0.002)	1.000 (0.000)	0.999 (0.000)
1000	NCTP	0.846 (0.013)	0.950 (0.005)	1.000 (0.000)	0.995 (0.000)
1000	CAFHT	2.000 (0.000)	0.985 (0.002)	1.000 (0.000)	0.999 (0.000)
2000					
2000	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
2000	NCTP	0.673 (0.008)	0.912 (0.005)	1.000 (0.000)	0.991 (0.001)
2000	CAFHT	0.325 (0.003)	0.914 (0.007)	1.000 (0.000)	0.991 (0.001)
5000					
5000	CFRNN	2.000 (0.000)	0.998 (0.001)	1.000 (0.000)	1.000 (0.000)
5000	NCTP	0.591 (0.005)	0.911 (0.004)	1.000 (0.000)	0.991 (0.000)
5000	CAFHT	0.300 (0.002)	0.907 (0.005)	1.000 (0.000)	0.990 (0.001)
10000					
10000	CFRNN	2.000 (0.000)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)
10000	NCTP	0.548 (0.005)	0.906 (0.005)	1.000 (0.000)	0.991 (0.000)
10000	CAFHT	0.293 (0.001)	0.901 (0.005)	1.000 (0.000)	0.990 (0.001)

A5.6. Comparisons with CopulaCPTS

For completeness, this subsection presents empirical results that compare our CAFHT method with CopulaCPTS (Sun & Yu, 2023), which uses the copula of prediction residuals across the entire horizon. Similar to NCTP, CopulaCPTS struggles with adaptability under heteroscedastic conditions and is thus expected to achieve conditional coverage akin to that of NCTP. We conducted these comparisons using synthetic AR data with dynamic profiles. The CopulaCPTS method is considered suitable only for situations with ample calibration data, as noted by Sun & Yu (2023). Accordingly, we performed experiments with large datasets of 5,000 and 10,000 trajectories, designating 25% randomly for calibration and the remainder for training. The findings were validated against an additional 100 independently generated test trajectories.

The results, displayed in Figures A31–A32 and Tables A32–A33, confirm the anticipated outcomes. CopulaCPTS delivers results comparable to NCTP, while CAFHT surpasses the CopulaCPTS baseline by producing narrower prediction bands and achieving higher conditional coverage.

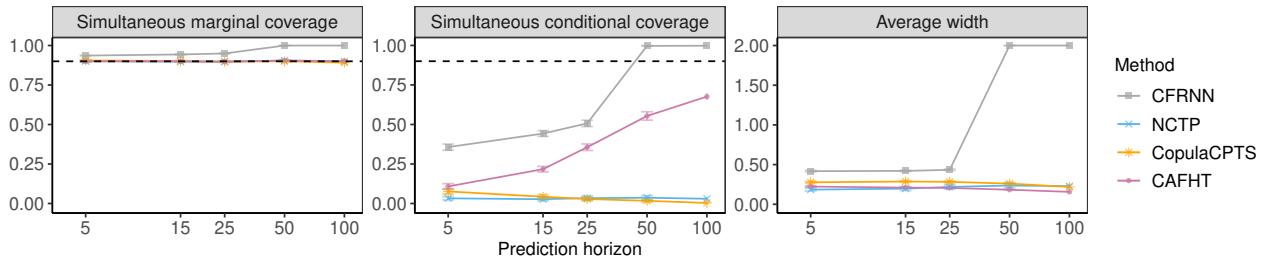

Figure A31. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the prediction horizon. See Table A32 for detailed results and standard errors.

Table A32. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of prediction horizon. The red numbers indicate smaller prediction bands or higher conditional coverage. See corresponding plot in Figure A31.

			Simultaneous coverage		
Prediction horizon	Method	Average width	Conditional-hard	Conditional-easy	Marginal
5					
5	CFRNN	0.417 (0.006)	0.357 (0.010)	1.000 (0.000)	0.936 (0.001)
5	NCTP	0.185 (0.009)	0.033 (0.005)	0.996 (0.001)	0.901 (0.002)
5	CopulaCPTS	0.277 (0.006)	0.076 (0.009)	0.997 (0.001)	0.906 (0.002)
5	CAFHT	0.223 (0.004)	0.108 (0.009)	0.987 (0.001)	0.900 (0.002)
15					
15	CFRNN	0.420 (0.005)	0.442 (0.009)	1.000 (0.000)	0.943 (0.001)
15	NCTP	0.197 (0.012)	0.027 (0.005)	0.996 (0.001)	0.898 (0.002)
15	CopulaCPTS	0.287 (0.005)	0.043 (0.006)	0.998 (0.001)	0.901 (0.002)
15	CAFHT	0.210 (0.002)	0.217 (0.009)	0.977 (0.001)	0.900 (0.002)
25					
25	CFRNN	0.434 (0.005)	0.507 (0.010)	1.000 (0.000)	0.950 (0.001)
25	NCTP	0.218 (0.013)	0.034 (0.006)	0.996 (0.001)	0.898 (0.002)
25	CopulaCPTS	0.282 (0.006)	0.028 (0.006)	0.997 (0.001)	0.898 (0.002)
25	CAFHT	0.206 (0.001)	0.356 (0.010)	0.959 (0.002)	0.897 (0.002)
50					
50	CFRNN	2.000 (0.000)	0.997 (0.001)	1.000 (0.000)	1.000 (0.000)
50	NCTP	0.236 (0.013)	0.036 (0.007)	0.997 (0.001)	0.905 (0.001)
50	CopulaCPTS	0.261 (0.007)	0.017 (0.004)	0.993 (0.001)	0.900 (0.002)
50	CAFHT	0.183 (0.001)	0.554 (0.013)	0.941 (0.002)	0.904 (0.002)
100					
100	CFRNN	2.000 (0.000)	0.998 (0.001)	1.000 (0.000)	1.000 (0.000)
100	NCTP	0.229 (0.013)	0.030 (0.006)	0.996 (0.001)	0.899 (0.002)
100	CopulaCPTS	0.218 (0.008)	0.003 (0.002)	0.991 (0.001)	0.892 (0.002)
100	CAFHT	0.156 (0.001)	0.676 (0.010)	0.927 (0.002)	0.902 (0.002)

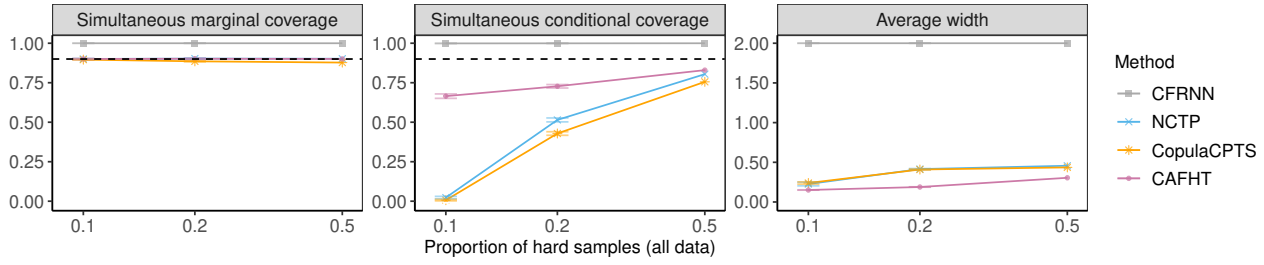

Figure A32. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the proportion of hard-to-predict trajectories. See Table A33 for detailed results and standard errors.

Table A33. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the proportion of hard-to-predict trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. See corresponding plot in Figure A32.

			Simultaneous coverage		
Proportion of hard samples (all data)	Method	Average width	Conditional-hard	Conditional-easy	Marginal
0.1					
0.1	CFRNN	2.000 (0.000)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)
0.1	NCTP	0.226 (0.011)	0.023 (0.004)	0.999 (0.000)	0.901 (0.001)
0.1	CopulaCPTS	0.239 (0.007)	0.007 (0.002)	0.995 (0.001)	0.896 (0.002)
0.1	CAFHT	0.151 (0.001)	0.665 (0.007)	0.927 (0.001)	0.900 (0.002)
0.2					
0.2	CFRNN	2.000 (0.000)	0.999 (0.000)	1.000 (0.000)	1.000 (0.000)
0.2	NCTP	0.415 (0.003)	0.515 (0.006)	1.000 (0.000)	0.903 (0.002)
0.2	CopulaCPTS	0.409 (0.003)	0.429 (0.006)	1.000 (0.000)	0.885 (0.002)
0.2	CAFHT	0.189 (0.001)	0.727 (0.005)	0.945 (0.001)	0.902 (0.002)
0.5					
0.5	CFRNN	2.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
0.5	NCTP	0.457 (0.003)	0.804 (0.003)	1.000 (0.000)	0.902 (0.002)
0.5	CopulaCPTS	0.436 (0.003)	0.756 (0.003)	1.000 (0.000)	0.878 (0.002)
0.5	CAFHT	0.305 (0.002)	0.830 (0.003)	0.972 (0.001)	0.901 (0.002)

A6. Extension to Multi-Step Forecasting

This section extends CAFHT to the multiple-step-ahead forecasting setting. Similar to section 2.1, consider a data set containing n observations of trajectories of length $T+1$, namely $\mathcal{D} := \{\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(n)}\}$. For $i \in [n] := \{1, \ldots, n\}$, the array $\boldsymbol{Y}^{(i)} = (Y_0^{(i)}, \ldots, Y_T^{(i)})$ represents $T+1$ observations of some d-dimensional vector $Y_t^{(i)} = (Y_{t,1}^{(i)}, \ldots, Y_{t,d}^{(i)}) \in \mathbb{R}^d$, measured at distinct time steps $t \in \{0, \ldots, T+1\}$. Let g denote a trainable trajectory predictor that can make H-steps-ahead forecasts.

Consider a new trajectory $\boldsymbol{Y}^{(n+1)}$ sampled exchangeably with $\mathcal{D}$. Given the initial position $Y_0^{(n+1)}$, at every time t for $t \in \{1, \ldots, T\}$, the real value $Y_t^{(n+1)}$ is revealed, and we aim to construct prediction regions $(\hat{C}_t^1(\boldsymbol{Y}^{(n+1)}), \ldots, \hat{C}_t^H(\boldsymbol{Y}^{(n+1)}))$ for $(Y_{t+1}^{(n+1)}, \ldots, Y_{t+H}^{(n+1)})$ using the predictions $(\hat{Y}_{t+1}^{(n+1)}, \ldots, \hat{Y}_{t+H}^{(n+1)})$ made by $\hat{g}$.

Let $\hat{C}_t^\tau(\boldsymbol{Y}^{(n+1)})$ represent the τ-th-step-ahead prediction band for $Y_{t+\tau}^{(n+1)}$ output at time t from the CAFHT method. We aim to achieve the marginal simultaneous coverage, similar to Equation (1):

$$\mathbb{P}\left[Y_{t+\tau}^{(n+1)} \in \hat{C}_t^\tau(\boldsymbol{Y}^{(n+1)}),\ \forall t \in [T],\ \forall \tau \in [H]\right] \geq 1-\alpha. \tag{A15}$$

Similar to the one-step-ahead setting, we first initialize the adaptive prediction bands by extending the original ACI method to leverage the information of multi-step-ahead forecasting and to construct a multi-step-ahead prediction band. After that, we will calibrate the initialized adaptive bands and perform data-driven parameter selection.

A6.1. Multi-Step-Ahead ACI

In this section, we explain how to extend the original one-step-ahead ACI to produce multi-steps-ahead prediction regions. Although this approach is intuitive, it may be possible to improve it in the future.

Consider a similar online setting as in Gibbs & Candès (2021), where one observes covariate-response pairs $\{(X_t, Y_t)\}_{t\in\mathbb{N}} \subset \mathbb{R}^d \times \mathbb{R}$ in the sequential order. Denote the fitted model that can make H steps ahead predictions as $\hat{g}$. At each time step t, assume that we observe pairs up until $\{(X_t, Y_t)\}$ and make H steps ahead forecasts $(\hat{Y}_{t+1}, \ldots, \hat{Y}_{t+H})$ for the future values $(Y_{t+1}, \ldots, Y_{t+H})$ using $\hat{g}$. To construct the prediction regions for $(Y_{t+1}, \ldots, Y_{t+H})$, consider running H many ACI in parallel using the lagged nonconformity scores proposed by Dixit et al. (2023).

First, to construct the prediction region for a single time step $Y_{t+\tau}$ in the future for any $\tau \in [H]$, we compute the lagged nonconformity score, defined as:

$$S_t^\tau(X_t, y) = \|y - \hat{g}(X_t)\| = \|y - \hat{Y}_t^\tau\|. \tag{A16}$$

Intuitively, this measures the distance between y and the prediction for $Y_{t+\tau}$ made at the current time. Then, the standard split conformal prediction approach to construct the prediction region for $Y_{t+\tau}$ at miscoverage level α would become $\hat{C}_t^\tau(\alpha) = \{y : S_t^\tau(X_t, y) \leq \hat{Q}(1-\alpha)\}$, where $\hat{Q}(1-\alpha) = \inf\{s : (|\mathcal{D}_{\text{cal}}|^{-1} \sum_{(X_r, Y_r)\in\mathcal{D}_{\text{cal}}} \mathbb{1}_{\{S_{r-\tau}^\tau(X_{r-\tau}, Y_r) \leq s\}}) \geq 1-\alpha\}$. To incorporate the core idea of ACI to continuously adapt the potential distribution changes within the time series, we run the following modified α-update rule:

$$\alpha_{t+1}^\tau = \alpha_t^\tau + \gamma^\tau(\alpha - \text{err}_t^\tau), \tag{A17}$$

where

$$\text{err}_t^\tau = \begin{cases} 1, & \text{if } Y_t \notin \hat{C}_{t-\tau}^{\text{ACI},\tau}(\alpha_{t-1}^\tau), \\ 0, & \text{otherwise.} \end{cases}, \tag{A18}$$

Above, γ^τ denotes the step size, which can be different for each τ, and $\hat{C}_{t-\tau}^{\text{ACI},\tau}(\alpha_{t-1}^\tau)$ is the prediction region constructed for Y_t at τ steps ago as $\hat{C}_{t-\tau}^{\text{ACI},\tau}(\alpha_{t-1}^\tau) = \{y : S_{t-\tau}^\tau(X_{t-\tau}, y) \leq \hat{Q}_{t-\tau}(1-\alpha_{t-1})\}$. Equivalently,

$$\hat{C}_{t-\tau}^{\text{ACI},\tau}(\alpha_{t-1}) = [\hat{\ell}_{t-\tau}^{\text{ACI},\tau}, \hat{u}_{t-\tau}^{\text{ACI},\tau}] = [\hat{Y}_{t-\tau}^\tau - \hat{Q}_{t-\tau}(1-\alpha_{t-1}), \hat{Y}_{t-\tau}^\tau + \hat{Q}_{t-\tau}(1-\alpha_{t-1})].$$

The prediction region of $Y_{t+\tau}$ is then formed by:

$$\hat{C}_t^{\text{ACI},\tau}(\alpha_{t+1}^\tau) = [\hat{Y}_t^\tau - \hat{Q}_t(1-\alpha_{t+1}), \hat{Y}_t^\tau + \hat{Q}_t(1-\alpha_{t+1})]. \tag{A19}$$

To construct multiple steps ahead prediction regions of $(Y_{t+1}, \ldots, Y_{t+H})$ at time t, we run the above procedure for every $\tau \in [H]$, and form the prediction regions $(\hat{C}_t^{\text{ACI},1}(\alpha_{t+1}^1), \ldots, \hat{C}_t^{\text{ACI},H}(\alpha_{t+1}^H))$; see Algorithm A8.

Algorithm A8 Multi-step-ahead ACI

1: **Input**: A pre-trained forecaster $\hat{g}$ producing H-step-ahead predictions; current time t; time trajectory with observed past values $(Y_1, \ldots, Y_{t-1})$.
2: Observe the true value at current time Y_t.
3: Make H-step-ahead predictions $(\hat{Y}_{t+1}, \ldots, \hat{Y}_{t+H})$ for $(Y_{t+1}, \ldots, Y_{t+H})$.
4: **for** $\tau \in [H]$ **do**
5: Evaluate err_t^τ using Equation (A18).
6: Update α_{t+1}^τ using Equation (A17).
7: Construct prediction region $\hat{C}_t^{\text{ACI},\tau}(\alpha_{t+1}^\tau)$ for $Y_{t+\tau}$ using Equation (A19).
8: **end for**
9: **Output**: Online multi-steps-ahead prediction regions $(\hat{C}_t^{\text{ACI},1}(\alpha_{t+1}^1), \ldots, \hat{C}_t^{\text{ACI},H}(\alpha_{t+1}^H))$.

A6.2. Calibrating the Adaptive Prediction Bands

In the previous section, we discussed how to form multiple steps ahead prediction regions using ACI at every time t. We now proceed to calibrate these regions to achieve simultaneous coverage guarantee (A15). For simplicity, we start by taking the learning rate γ^τ as fixed and constant for all $\tau \in [H]$.

Different from the one-step-ahead setting, with multi-step-ahead ACI, at every time t we can construct H prediction regions for the following H values. As we move on to observe the next trajectory value, we can update the future prediction regions using the more recent information. In fact, at every t, we will have $H-1$ different prediction regions, separately constructed from $H-1, H-2, \ldots, 1$ steps ago, denoted as $\hat{C}_{t-H}^{\text{ACI},H}, \ldots, \hat{C}_{t-1}^{\text{ACI},1}$. To perform calibration, we need to summarize the information obtained from those into a single region, which we will explain next.

For any $\tau \in [H]$, let $\hat{C}_{t-\tau}^{\text{ACI},\tau}(\boldsymbol{Y}^{(i)}, \gamma) = [\hat{\ell}_{t-\tau}^{\text{ACI},\tau}(\boldsymbol{Y}^{(i)}, \gamma), \hat{u}_{t-\tau}^{\text{ACI},\tau}(\boldsymbol{Y}^{(i)}, \gamma)]$ denote the prediction band for Y_t constructed at τ steps ago with learning rate γ. For each calibration trajectory $i \in \mathcal{D}_{\text{cal}}$, CAFHT evaluates the nonconformity score $\hat{\epsilon}_i(\gamma)$ using the following equation:

$$\hat{\epsilon}_i(\gamma) := \max_{t \in \{1,\ldots,T\}} \left\{ \max \left\{ \left[\max_{\tau \in [H]} \left\{ \hat{\ell}_{t-\tau}^{\text{ACI},\tau}(\boldsymbol{Y}^{(i)}, \gamma) \right\} - Y_t^{(i)} \right]_+, \left[Y_t^{(i)} - \min_{\tau \in [H]} \left\{ \hat{u}_{t-\tau}^{\text{ACI},\tau}(\boldsymbol{Y}^{(i)}, \gamma) \right\} \right]_+ \right\} \right\}, \tag{A20}$$

Intuitively, $\hat{\epsilon}_i(\gamma)$ measures the maximum absolute distance of Y_t from the prediction regions constructed at different historical time steps $\hat{C}_{t-H}^{\text{ACI},H}, \ldots, \hat{C}_{t-1}^{\text{ACI},1}$.

The remaining components of our method then follow the same logic as the one-step-ahead CAFHT. Let $\hat{Q}(1-\alpha, \gamma)$ denote the $\lceil (1-\alpha)(1+|\mathcal{D}_{\text{cal}}|) \rceil$-th smallest value of $\hat{\epsilon}_i(\gamma)$ among $i \in \mathcal{D}_{\text{cal}}$. At every time step $t \in [T]$, CAFHT constructs H-steps ahead prediction bands $\hat{C}_t^\tau(\boldsymbol{Y}^{(n+1)}, \gamma), \forall \tau \in [H]$ using the following equation:

$$\hat{C}_t^\tau(\boldsymbol{Y}^{(n+1)}, \gamma) = \left[\hat{\ell}_t^{\text{ACI},\tau}(\boldsymbol{Y}^{(n+1)}, \gamma) - \hat{Q}(1-\alpha, \gamma), \hat{u}_t^{\text{ACI},\tau}(\boldsymbol{Y}^{(n+1)}, \gamma) + \hat{Q}(1-\alpha, \gamma) \right]. \tag{A21}$$

The next result establishes finite-sample simultaneous coverage guarantees for this method.

Theorem A1. *Assume that the calibration trajectories in $\mathcal{D}_{cal}$ are exchangeable with $\boldsymbol{Y}^{(n+1)}$. Then, for any $\alpha \in (0,1)$, the prediction band output by the multi-step-ahead CAFHT, applied with fixed parameters α, α_{ACI}, and γ, satisfies* (A15).

Proof. The proof is very similar to the proof of Theorem 1, and it follows directly from the exchangeability of the conformity scores. Denote $\hat{\epsilon}_{n+1}(\gamma)$ the conformity score of the test trajectory $\boldsymbol{Y}^{(t+1)}$ evaluated using Equation (A20). For any fixed α and $\gamma > 0$, we have that $Y_{t+\tau}^{(n+1)} \in \hat{C}_t^\tau(\boldsymbol{Y}^{(n+1)}, \gamma)\ \forall \tau \in [H]\ \forall t \in [T]$ if and only if $\hat{\epsilon}_{n+1}(\gamma) \leq \hat{Q}(1-\alpha, \gamma)$, where $\hat{Q}(1-\alpha, \gamma)$ is the $\lceil (1-\alpha)(1+|\mathcal{D}_{\text{cal}}|) \rceil$-th smallest value of $\hat{\epsilon}_i(\gamma)$ for all $i \in \mathcal{D}_{\text{cal}}$. Since the test trajectory is exchangeable with $\mathcal{D}_{\text{cal}}$, its score $\hat{\epsilon}_{n+1}(\gamma)$ is also exchangeable with $\{\hat{\epsilon}_i(\gamma), i \in \mathcal{D}_{\text{cal}}\}$. Then by Lemma 1 in Romano et al. (2019), it follows that $\mathbb{P}(Y_{t+\tau}^{(n+1)} \in \hat{C}_t^\tau(\boldsymbol{Y}^{(n+1)}, \gamma)\ \forall \tau \in [H]\ \forall t \in [T]) = \mathbb{P}(\hat{\epsilon}_{n+1}(\gamma) \leq \hat{Q}(1-\alpha, \gamma)) \geq 1-\alpha$. □

A6.3. Data-Driven Parameter Selection

Similar to the one-step-ahead CAFHT, we can choose the step size parameter γ in a data-driven way. For simplicity, we start by selecting among a grid of candidate $\{\gamma_1, \ldots, \gamma_L\}$, but assuming that the step size stays the same for every time step $\tau \in [H]$. Later in the experiments, we discuss using alternative options, such as setting γ decaying as τ increases, which is more intuitive in practice as predictions made longer steps ahead are usually less reliable than the predictions made more recently.

Algorithm A9 Model selection component of multi-steps-ahead CAFHT

Input: A pre-trained forecaster $\hat{g}$ producing H-step-ahead predictions; calibration trajectories $\mathcal{D}^1_{\text{cal}}$; a grid of candidate learning rates $\{\gamma_1, \ldots, \gamma_L\}$.
for $\ell \in [L]$ **do**
  Construct $\hat{C}_t^{\text{ACI},\tau}(\boldsymbol{Y}^{(i)}, \gamma_\ell)\ \forall t \in [T], \forall \tau \in [H]$ using Algorithm A8, for $i \in \mathcal{D}^1_{\text{cal}}$.
  Evaluate $\hat{\epsilon}_i(\gamma_\ell)$ using (A20), for $i \in \mathcal{D}^1_{\text{cal}}$.
  Compute $\hat{Q}(1-\alpha, \gamma_\ell)$, the $(1-\alpha)(1+1/|\mathcal{D}^1_{\text{cal}}|)$-th quantile of $\{\hat{\epsilon}_i(\gamma_\ell), i \in \mathcal{D}^1_{\text{cal}}\}$.
  Construct $\hat{C}_t^{\tau}(\boldsymbol{Y}^{(i)}, \gamma_\ell)\ \forall \tau \in [H] \forall t \in [H]$ using (A21) for $i \in \mathcal{D}^1_{\text{cal}}$.
end for
Pick $\hat{\gamma}$ such that,

$$\hat{\gamma} := \underset{\ell \in [L]}{\arg\min}\ \text{AvgWidth}(\{C_t^{\tau}(\boldsymbol{Y}^{(i)}, \gamma_\ell)\}_{t \in [T], \tau \in [H]}). \tag{A22}$$

Output: Selected learning rate parameter $\hat{\gamma}$.

Algorithm A10 Multi-step-ahead CAFHT

Input: A pre-trained forecaster $\hat{g}$ producing multi-step-ahead predictions; calibration trajectories $\mathcal{D}_{\text{cal}}$; the initial position $Y_0^{(n+1)}$ of a test trajectory $\boldsymbol{Y}^{(n+1)}$; the desired nominal level $\alpha \in (0,1)$; a grid of candidate learning rates $\{\gamma_1, \ldots, \gamma_L\}$.
Randomly split $\mathcal{D}_{\text{cal}}$ into $\mathcal{D}^1_{\text{cal}}$ and $\mathcal{D}^2_{\text{cal}}$.
Select a learning rate $\hat{\gamma} \in \{\gamma_1, \ldots, \gamma_L\}$, applying Algorithm A1 using the trajectory data in $\mathcal{D}^1_{\text{cal}}$.
Construct $\hat{C}^{\text{ACI}}(\boldsymbol{Y}^{(i)}, \hat{\gamma})$ using ACI, for $i \in \mathcal{D}^2_{\text{cal}}$.
Evaluate $\hat{\epsilon}_i(\hat{\gamma})$ using (A20), for $i \in \mathcal{D}^2_{\text{cal}}$.
Compute the empirical quantile $\hat{Q}(1-\alpha, \hat{\gamma})$.
for $t \in [T]$ **do**
  Observe the current step $Y_t^{(n+1)}$.
  Compute $\hat{C}_t^{\text{ACI},\tau}(\boldsymbol{Y}^{(n+1)}, \hat{\gamma})\ \forall \tau \in [H]$ with the multi-step-ahead ACI stated in Algorithm A8, using the past of the test trajectory $(Y_1^{(n+1)}, \ldots, Y_t^{(n+1)})$.
  Compute prediction bands $\hat{C}_t^{\tau}(\boldsymbol{Y}^{(n+1)}, \hat{\gamma}), \forall \tau \in [H]$ for the next H steps, using (A21).
end for
Output: Online prediction bands $\hat{C}(\boldsymbol{Y}^{(n+1)})$.

A6.4. Multi-step-ahead CAFHT using Multiplicative Scores

Similar to the one-step-ahead cases, we can utilize a multiplicative score for the multi-step-ahead settings. This can be simply accomplished by replacing the nonconformity scores defined in (A20) with these:

$$\tilde{\epsilon}_i(\gamma) := \max_{t \in \{1,\ldots,T\}} \left\{ \max \left\{ \max_{\tau \in [H]} \left\{ \frac{\left[\hat{\ell}_{t-\tau}^{\text{ACI},\tau}(\boldsymbol{Y}^{(i)}, \gamma) - Y_t^{(i)}\right]_+}{|\hat{C}_{t-\tau}^{\text{ACI},\tau}(\boldsymbol{Y}^{(i)}, \gamma)|} \right\}, \max_{\tau \in [H]} \left\{ \frac{\left[Y_t^{(i)} - \hat{u}_{t-\tau}^{\text{ACI},\tau}(\boldsymbol{Y}^{(i)}, \gamma)\right]_+}{|\hat{C}_{t-\tau}^{\text{ACI},\tau}(\boldsymbol{Y}^{(i)}, \gamma)|} \right\} \right\} \right\}, \tag{A23}$$

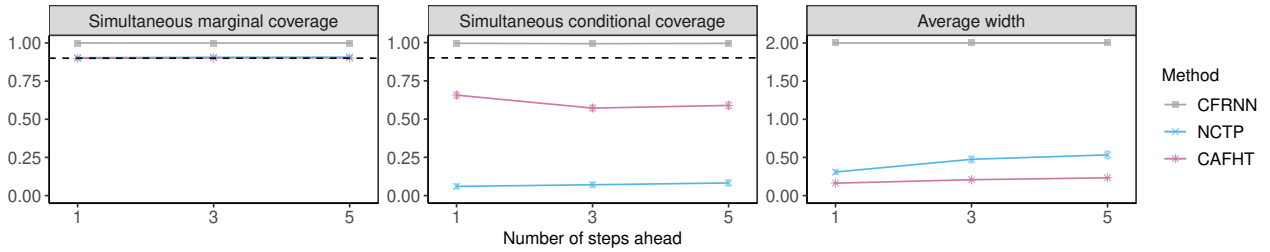

Figure A33. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the steps-ahead parameter H utilized by the forecaster. Other details are as in Table A34.

and the counterpart of Equation (A21) becomes

$$\tilde{C}_t^{\tau}(\boldsymbol{Y}^{(n+1)}, \gamma) = \Big[\hat{\ell}_t^{\mathrm{ACI},\tau}(\boldsymbol{Y}^{(n+1)}, \gamma) - \hat{Q}(1-\alpha, \gamma) \cdot |\hat{C}_t^{\mathrm{ACI},\tau}(\boldsymbol{Y}^{(i)}, \gamma)|, \\ \hat{u}_t^{\mathrm{ACI},\tau}(\boldsymbol{Y}^{(n+1)}, \gamma) + \hat{Q}(1-\alpha, \gamma) \cdot |\hat{C}_t^{\mathrm{ACI},\tau}(\boldsymbol{Y}^{(i)}, \gamma)|\Big]. \tag{A24}$$

A6.5. Numerical Experiments

We utilize the same synthetic settings as in Section 4, but modify the LSTM models so that they can make multiple steps ahead of predictions. Again, we choose the ACI-based multiplicative scores as the main CAFHT method.

Figure A33 summarizes the performance of the three methods as a function of the number of steps ahead predictions made by the forecaster, which is varied from 1 to 5. When number of steps is equal to 1, we recover the one-step-ahead CAFHT results. In each case, 75% of the trajectories are used for training and the remaining 25% for calibration. Our method utilizes 50% of the calibration trajectories to select the ACI learning rate γ. The results are averaged over 500 test trajectories and 100 independent experiments.

As we can see, all methods attain 90% simultaneous coverage as defined in (A15). However, our method yields the most efficient results in terms of obtaining the smallest size of the prediction band and higher conditional coverage than the NCTP benchmark. See Table A34 for standard errors.

Table A34. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A33.

				Simultaneous coverage	
Number of steps ahead	Method	Average width	Conditional-hard	Conditional-easy	Marginal
1					
1	CFRNN	2.000 (0.000)	0.995 (0.001)	1.000 (0.000)	1.000 (0.000)
1	NCTP	0.308 (0.014)	0.060 (0.008)	0.996 (0.001)	0.903 (0.002)
1	CAFHT	0.163 (0.001)	0.656 (0.010)	0.926 (0.002)	0.899 (0.003)
3					
3	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
3	NCTP	0.475 (0.022)	0.071 (0.008)	0.996 (0.001)	0.906 (0.002)
3	CAFHT	0.208 (0.002)	0.572 (0.010)	0.938 (0.002)	0.902 (0.002)
5					
5	CFRNN	2.000 (0.000)	0.995 (0.001)	1.000 (0.000)	0.999 (0.000)
5	NCTP	0.534 (0.025)	0.083 (0.010)	0.997 (0.001)	0.907 (0.002)
5	CAFHT	0.233 (0.002)	0.589 (0.012)	0.936 (0.002)	0.902 (0.002)

In another experiment, the steps-ahead parameter is fixed as $H = 3$, and the total number of trajectories in the training and calibration sets are varied from 200 to 2000. Again, our method yields the most informative bands.

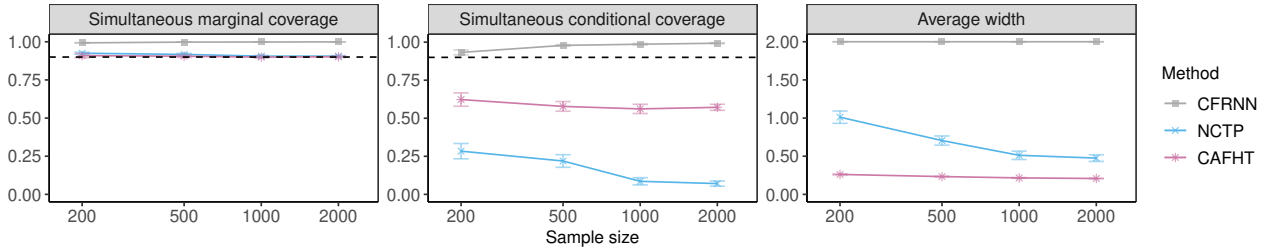

Figure A34. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the number of trajectories in the training and calibration sets, made by the 3-steps-ahead forecaster. Other details are as in Table A35.

Table A35. Performance on simulated heterogeneous trajectories of prediction bands constructed by different methods, as a function of the total number of training and calibration trajectories. The red numbers indicate smaller prediction bands or higher conditional coverage. See the corresponding plot in Figure A34.

			Simultaneous coverage		
Sample size	Method	Average width	Conditional-hard	Conditional-easy	Marginal
200					
200	CFRNN	2.000 (0.000)	0.932 (0.008)	1.000 (0.000)	0.993 (0.001)
200	NCTP	1.012 (0.040)	0.284 (0.025)	0.997 (0.001)	0.925 (0.003)
200	CAFHT	0.261 (0.004)	0.623 (0.022)	0.942 (0.006)	0.909 (0.007)
500					
500	CFRNN	2.000 (0.000)	0.979 (0.003)	1.000 (0.000)	0.998 (0.000)
500	NCTP	0.705 (0.030)	0.219 (0.021)	0.995 (0.002)	0.917 (0.003)
500	CAFHT	0.233 (0.002)	0.578 (0.016)	0.942 (0.003)	0.906 (0.004)
1000					
1000	CFRNN	2.000 (0.000)	0.986 (0.002)	1.000 (0.000)	0.999 (0.000)
1000	NCTP	0.512 (0.027)	0.086 (0.012)	0.995 (0.001)	0.906 (0.002)
1000	CAFHT	0.216 (0.002)	0.561 (0.015)	0.937 (0.003)	0.900 (0.003)
2000					
2000	CFRNN	2.000 (0.000)	0.993 (0.001)	1.000 (0.000)	0.999 (0.000)
2000	NCTP	0.475 (0.022)	0.071 (0.008)	0.996 (0.001)	0.906 (0.002)
2000	CAFHT	0.208 (0.002)	0.572 (0.010)	0.938 (0.002)	0.902 (0.002)