Conformal Prediction for Time-Dependent PDEs

Joshua Stiller[*]
LMU Munich,
MCML

Annika Schneider
TU Munich, Helmholtz AI,
relAI, MCML

Eyke Hüllermeier
LMU Munich,
MCML, DFKI

Abstract

Uncertainty quantification is crucial in scientific machine learning, where models inform safety-critical tasks such as flood forecasting, financial risk management, and thermal control in machines. Conformal prediction provides distribution-free coverage guarantees, but in time-dependent settings common to physics and engineering, these guarantees can break down, leading to systematic undercoverage. We study this problem in the context of surrogate models for time-dependent physical systems described by partial differential equations. We prove that in a function space setting, distributions at arbitrarily close times can be mutually singular, making exact coverage guarantees impossible. We further show that in discretized settings, the total variation distance between distributions often saturates exponentially fast with the grid resolution, mandating restrictions on the spatial discretization to ensure coverage. Finally, for certain settings, we show how to use principled weighted conformal prediction to obtain finite-sample coverage guarantees over growing time horizons.

1 Introduction

Many problems in physics and engineering, such as weather prediction, aerodynamics, and financial modeling, are governed by partial differential equations (PDEs). Classical numerical solvers are accurate but computationally expensive, often scaling poorly with the dimensionality of the data. AI-based surrogate models have emerged as a promising alternative, providing fast approximations of PDE solutions. Prominent examples include physics-informed neural networks (PINNs) [29], DeepONets [22], and Neural Operators [20, 21].

Despite these advances, surrogate models still lack principled mechanisms for uncertainty quantification, which are vital, as model predictions often inform safety-critical decisions in science and engineering. Conformal prediction (CP) [33] provides a principled framework, producing distribution-free uncertainty sets with guaranteed marginal coverage. These guarantees, however, rely on exchangeability between calibration and test samples, a condition that is frequently violated in time-dependent PDEs.

Non-stationarity in time-dependent PDEs Let u_t denote the solution of a time-dependent PDE at time t. In practice, we are interested in predicting $u_{t+\delta}$ for several time steps $\delta > 0$, beyond the available training and calibration data. Unless u_t is a stationary process, test samples follow a different distribution than observed calibration samples, breaking the exchangeability assumption required by conformal prediction (see Figure E.1). This type of non-stationarity is ubiquitous: sudden shocks (e.g., stock market crashes), long-term structural changes (e.g., climate trends), and limited development windows (e.g., laboratory testing) all produce systematic shifts in the data distribution (see Figure 1).

[*]Corresponding author (joshua.stiller@lmu.de)

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

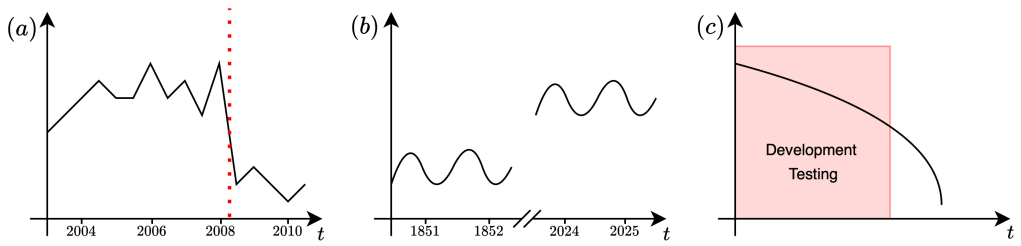


Figure 1: Examples of different types of non-stationarity that can arise in time-dependent systems. *(a)* Sudden distribution shift due to external shocks, illustrated by a stock market crash. *(b)* Long-term structural changes, as in climate time series. *(c)* Limited observation window during system development, where calibration and testing occur on different parts of the trajectory.

Why existing methods fall short Existing CP approaches for time-dependent settings either rely on assumptions like mixing, which generally fail for PDE dynamics, or relax to asymptotic guarantees that average coverage over time and cannot guarantee per-step coverage [9, 10]. In theory, weighted conformal prediction (WCP) [32, 3] provides finite-sample coverage under distribution shift, but it requires likelihood ratios $p_{\text{test}}/p_{\text{cal}}$ that are generally unavailable for PDE surrogates. As a result, no current approach provides finite-sample coverage at every future time step. This raises a structural question: when can CP guarantees hold for time-dependent PDEs, and how should CP be applied to obtain them? We answer with theoretical limits and practical guidelines.

Our contributions

1. We show that in function-space, already for the heat equation with Gaussian initial conditions, the total variation (TV) distance between solutions at any two distinct times is maximal. This rules out non-exchangeable CP guarantees on the function-space view common in the neural operator literature.
2. We show that for discretized PDEs the TV distance often approaches 1 exponentially fast with the grid granularity. This gives a practical guideline: rather than calibrating on the full prediction grid, we should restrict to a coarser grid or the spatial subregion of interest, as we do in our real-world experiments.
3. We identify assumptions under which closed-form likelihood ratios and finite-sample WCP guarantees at every future time step exist in the discretized setting and further lift this discrete guarantee to the true continuous solution.

We empirically validate these results on synthetic PDEs and three real-world datasets.

2 Related work

CP for PDEs is closely related to conformal prediction for time series. Existing approaches can be roughly categorized as follows:

Access to exchangeable data Some work treats entire solution trajectories as exchangeable units (e.g., Ma et al. [23] for Neural Operators, Moya et al. [25] for DeepONets, or Gray et al. [13] and Gopakumar et al. [12, 11] for general surrogate models). These approaches fail when extrapolated to future time steps beyond the observed trajectory, as the distribution of the solution changes with time.

Relaxing to asymptotic guarantees Instead of requiring marginal coverage at every time step, asymptotic coverage provides guarantees on average over time. The seminal work on adaptive conformal prediction [9] tightens or loosens conformal bands based on past miscoverage. Several refinements have been proposed [e.g., 1, 37, 10]. For non-critical tasks and slowly drifting time series, this can be a viable approach. In many PDE problems, however, undercoverage at a single time step can be fatal and it is a priori unknown how fast changes appear.

Exploiting additional properties To extend CP beyond exchangeability, some work assumes additional properties on the data. For example, stationarity, (strong) mixing of the process or residuals, or requirements on the accuracy of the predictive model [e.g., 27, 37, 2, 38, 36]. Depending on the assumptions, both asymptotic and per-time-step guarantees can be obtained.

Distributional assumptions Generally, CP is distribution-free, but when certain properties of the distribution are known, they can be exploited to obtain coverage guarantees without exchangeability. In full generality, Barber et al. [3] show that coverage can be lower bounded in terms of the total variation (TV) distance between calibration and test distributions. Farinhas et al. [8] extend this to conformal risk control. Under covariate or label shift, more specialized approaches achieve marginal coverage when the distribution shift can be quantified [32, 28]. In practice, these methods are rarely used in a principled way, as the required density ratios are typically unknown. A related approach comes from Harris and Liu [15], which replaces global exchangeability with a local assumption across nearby time points.

3 Background

3.1 Conformal prediction

Conformal prediction (CP) is a framework for constructing prediction sets with marginal finite-sample coverage guarantees [33]. In the standard split setting, a model is trained on a dataset $\mathcal{D}_{\text{train}}$ and calibrated on a dataset $\mathcal{D}_{\text{cal}}$, yielding a set-valued predictor $\mathcal{C}$ such that, for a test sample $(x_{\text{test}}, y_{\text{test}})$,

$$\mathbb{P}\big(y_{\text{test}} \in \mathcal{C}(x_{\text{test}})\big) \;\geq\; 1-\alpha,$$

where α denotes the miscoverage level. This guarantee relies on **exchangeability** of calibration and test samples, meaning that their joint distribution is invariant under permutations. When exchangeability is violated, coverage may fail.

Conformal prediction beyond exchangeability When calibration and test distributions differ but are related through a likelihood ratio, weighted conformal prediction (WCP) provides a natural extension of standard CP [32]. A canonical example is *covariate shift*, where calibration samples (x_i, y_i) for $i \in \{1, \ldots, n\}$ are drawn from $\mathcal{P}_X \times \mathcal{P}_{Y|X}$, while test samples $(x_{\text{test}}, y_{\text{test}})$ follow $\mathcal{Q}_X \times \mathcal{P}_{Y|X}$, i.e., the conditional distribution of Y given X remains unchanged. In this setting, valid coverage guarantees can be recovered by reweighting the calibration quantiles according to the density ratio

$$w_i \;\propto\; \frac{p_{\text{test}}(x_i)}{p_{\text{cal}}(x_i)}, \quad \forall x_i \in \mathcal{D}_{\text{cal}} \cup \{x_{\text{test}}\}, \quad \sum_i w_i = 1, \tag{1}$$

where p_{cal} and p_{test} denote the densities of $\mathcal{P}_X$ and $\mathcal{Q}_X$ (see Appendix F for more details).

In the most general case, Barber et al. [3] and, similarly, Farinhas et al. [8] provide corrections for the conformal guarantees based on the TV distance between calibration and test distribution:[1]

$$\mathbb{P}\big(y_{\text{test}} \in \mathcal{C}(x_{\text{test}})\big) \;\geq\; 1-\alpha-\sum_{i=1}^{n} \tilde{w}_i \, d_{\text{TV}}(\boldsymbol{z}, \boldsymbol{z}^i), \tag{2}$$

where $\tilde{w}_i = w_i/(w_1 + \cdots + w_n + 1)$, with $w_i \in [0,1]$, $\boldsymbol{z} = ((x_1, y_1), \ldots, (x_n, y_n), (x_{\text{test}}, y_{\text{test}}))$ and $\boldsymbol{z}^i$ arises from permuting the test pair in $\boldsymbol{z}$ with the ith calibration pair. These approaches are rarely used in a principled way because they require knowing the corresponding distributions. We show when and how they can be used on PDE problems to regain coverage over time.

3.2 PDEs as operator mappings

Many dynamical systems in physics and engineering can be described by evolution equations of the form

$$\frac{\partial u}{\partial t}(\boldsymbol{x}, t) = \mathcal{L}_{\boldsymbol{x}} u(\boldsymbol{x}, t),$$

[1] The TV distance is originally defined on probability measures, and whenever we write $d_{\text{TV}}(\boldsymbol{x}, \boldsymbol{y})$ for random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, or $d_{\text{TV}}(\mathcal{P}, \mathcal{Q})$ for probability distributions $\mathcal{P}$ and $\mathcal{Q}$, we refer to the TV distance between their corresponding probability measures.

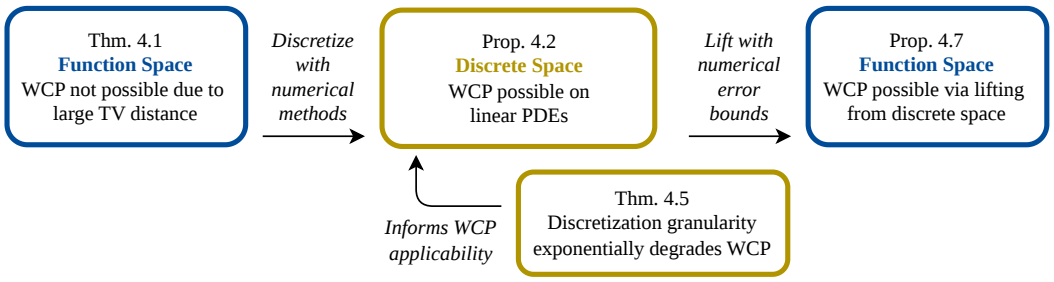


Figure 2: Overview of the theoretical results. The colors differentiate between function space (blue) and discretized space (yellow).

where $u : \Omega \times [0, \infty) \to \mathbb{R}$ is the state variable, $\boldsymbol{x}$ denotes spatial coordinates on a bounded domain $\Omega \subset \mathbb{R}^d$, $t \geq 0$ is the time, and $\mathcal{L}_{\boldsymbol{x}}$ is a (possibly nonlinear) differential operator acting on the spatial variable $\boldsymbol{x}$. In this paper, we are interested in the Cauchy-type problem, where we consider boundary conditions on $\bar{\Omega}$ and initial conditions $u(\boldsymbol{x}, 0)$ from some Banach space of functions $(\mathcal{A}, \|\cdot\|_{\mathcal{A}})$ and are interested in a solution $u(\boldsymbol{x}, t)$ in some Banach space of functions $(\mathcal{U}_t, \|\cdot\|_{\mathcal{U}_t})$. We will further only consider well-posed problems with a time-invariant differential operator, where we can define solution operators $\mathcal{G}_t : \mathcal{A} \to \mathcal{U}_t$, $\mathcal{G}_t(a) = u(\cdot, t)$ that uniquely map an initial condition to a solution function $u(\boldsymbol{x}, t)$ and the map $t \mapsto \mathcal{G}_t$ is continuous in t. In the rest of the paper, we will assume that all functions come from the same space, so $\mathcal{A} = \mathcal{U}_t$ for all $t \geq 0$, to simplify the notation, but the results apply more generally.

4 WCP for time-dependent PDEs

We begin by specifying the conformal setup, followed by the main theoretical contributions which we summarize in Figure 2.

4.1 Problem setting

From initial conditions to solutions To apply CP in the PDE setting, we first specify the underlying distributions. Let $u(\boldsymbol{x}, t)$ be the solution of a PDE and $S_\delta : \mathcal{U} \to \mathcal{U}$ be the solution operator mapping $u(\boldsymbol{x}, t)$ to $u(\boldsymbol{x}, t + \delta)$ for any $t \geq 0$. We may assume the initial condition (or in practice: our initial measurement) is distributed according to $u(\boldsymbol{x}, 0) \sim \mathcal{P}_0$. Then, S_δ induces a pushforward measure[2] $\mathcal{P}_\delta := (S_\delta)_{\#}\mathcal{P}_0$. If we now consider consecutive time points, our distribution changes over time:

$$\mathcal{P}_0 \xrightarrow{S_\delta} \mathcal{P}_\delta \xrightarrow{S_\delta} \mathcal{P}_{2\delta} \xrightarrow{S_\delta} \cdots$$

Thus, we obtain a sequence of probability distributions $\{\mathcal{P}_{k\delta}\}_{k\in\mathbb{N}}$ on the same function space, evolving under the PDE dynamics. Calibration and test data drawn from different $\mathcal{P}_t$ are therefore **not exchangeable**: although they live in the same function space, their distributions shift with time.

Datasets Datasets of size n typically consist of pairs $(u_i(\boldsymbol{x}, t), u_i(\boldsymbol{x}, t + \delta))$, for $i \in \{1, \ldots, n\}$, generated from initial conditions $u_i(\boldsymbol{x}, 0) \sim \mathcal{P}_0$. In practice, the u_i can either arise from measurements or numerical simulations. In both simulations and real-world settings, the states $u_i(\boldsymbol{x}, t)$ are discretized on a grid with p points, yielding vectors in $\mathbb{R}^p$. In this case, we omit the spatial dependence and denote the discretized solution at time t by $\boldsymbol{u}_i^t := (u_i(x_0, t), \ldots, u_i(x_{p-1}, t))^T \in \mathbb{R}^p$ (bold font). Nevertheless, since neural operators can generalize across grids, the function-space viewpoint remains relevant and is commonly adopted in the literature [15, 13, 24]. Accordingly, we consider both the functional representation (Section 4.2) and the discretized setting (Section 4.3). Throughout the paper we assume that our calibration dataset consists of pairs $(u_{\text{cal},i}(\boldsymbol{x}, t_{\text{cal}}), u_{\text{cal},i}(\boldsymbol{x}, t_{\text{cal}} + \delta))$ from some time t_{cal} and our test dataset consists of pairs $(u_{\text{test},i}(\boldsymbol{x}, t), u_{\text{test},i}(\boldsymbol{x}, t + \delta))$ for some $t \geq t_{\text{cal}}$. We omit the subscripts *cal* and *test* for notational simplicity.

[2]We slightly abuse notation here by writing the pushforward in terms of the distribution instead of the measure corresponding to the distribution.

4.2 Distribution shifts in function spaces

We will now consider CP on PDEs, starting with the function-space perspective, where the data points $u_i(\boldsymbol{x}, t)$ are elements of a suitable function space. As the data is not exchangeable, we must resort to a more general CP method. WCP is not suitable because the function-space densities[3] cannot be evaluated, but if the TV distance between the distributions at times t and $t + \delta$ were moderate, CP coverage guarantees for the prediction at $t + \delta$ could still be recovered using the approach from Equation (2). The following theorem shows that even for simple problems, like the heat equation with Gaussian initial distribution, the TV distance between the solution distributions $\mathcal{P}_t, \mathcal{P}_{t+\delta}$ at different time points $t, t+\delta$ is always maximal $d_{\mathrm{TV}}(\mathcal{P}_t, \mathcal{P}_{t+\delta}) = 1$, for all $t \geq 0, \delta > 0$. This is representative of a broader phenomenon that "[...] measures in infinite-dimensional spaces have a strong tendency of being mutually singular." [14]. As a direct consequence, regular CP and any method requiring distributional similarity between calibration and test fails. In particular, exact WCP via Equation (1) is inapplicable.

Theorem 4.1. *Consider the one-dimensional heat equation $\partial_t u = \partial_{xx} u$ on the domain $\Omega = (0, 1)$ with Dirichlet boundary conditions, $t \geq 0$. Suppose the initial condition $u_0(x)$ is sampled from a Gaussian random field $\mathcal{P}_0 = \mathcal{N}(\mathbf{0}, (\boldsymbol{I} - \boldsymbol{\Lambda})^{-1})$, where $\boldsymbol{\Lambda}$ is the Laplace operator on Ω. Then, the TV distance between the solution distributions at any two time points is maximal:*

$$d_{\mathrm{TV}}(\mathcal{P}_t, \mathcal{P}_{t+\delta}) = 1, \quad \forall\, t \geq 0,\ \delta > 0.$$

Proof sketch: The proof uses the Feldman–Hájek theorem [6], which shows that two Gaussian measures on a Hilbert space are either equivalent or mutually singular. We then calculate the Cameron–Martin spaces of both measures and show they differ, which tells us that the measures cannot be equivalent. The full proof is provided in Appendix G.1. □

We will now discuss how, despite the issue above, coverage guarantees can be achieved in practice.

4.3 Coverage guarantees in practice

Theorem 4.1 shows that a pure function-space perspective renders TV-based corrections vacuous, even for arbitrarily close time points. This is not a contradiction to practical CP on PDE surrogate models: in practice, we always observe or simulate on a finite number of degrees of freedom. Therefore, we shift the perspective to the *discretized state* $\boldsymbol{u}_i^t \in \mathbb{R}^p$. As we still consider the non-exchangeable setting, we either need additional inductive biases or rely on TV-based corrections. We will consider both approaches, starting with the former.

While CP is inherently distribution-free, related methods such as WCP [32, 8] give up this property in exchange for relaxing the exchangeability assumption. We will now identify a setting relevant to time-dependent PDEs where closed-form likelihood ratios are available so that WCP can be applied.

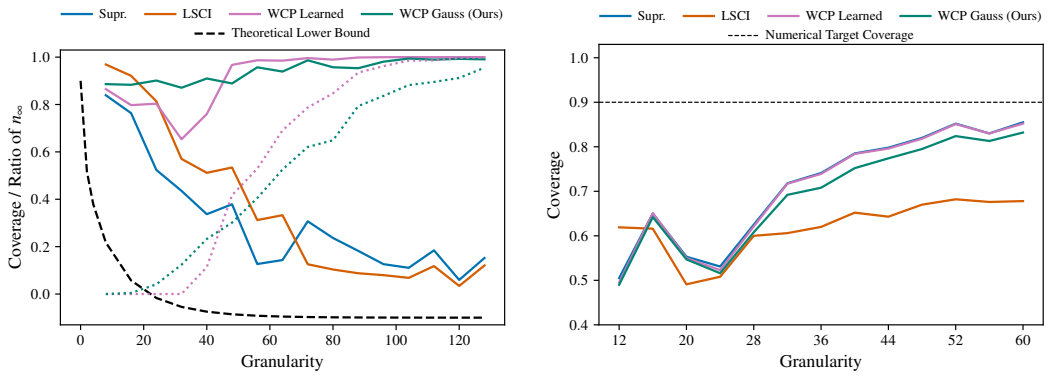
First, note that in physics, Gaussianity is often assumed for the initial distribution, because the aggregate effect of many small independent perturbations forming the initial condition is approximately Gaussian by the central limit theorem. This is well supported by the literature [21, 30, 12, 39, 11].

Second, observe that when considering *linear* PDEs solved with standard numerical schemes, the time evolution is affine $\boldsymbol{u}_i^{t+\delta} = \boldsymbol{A}_\delta \boldsymbol{u}_i^t + \boldsymbol{b}_\delta$, with $\boldsymbol{A}_\delta \in \mathbb{R}^{p \times p}$ and $\boldsymbol{b}_\delta \in \mathbb{R}^p$. If the initial condition is Gaussian, $\boldsymbol{u}_i^0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, then $\boldsymbol{u}_i^t$ remains Gaussian for all $t \geq 0$, with mean $\boldsymbol{\mu}_t$ and covariance $\boldsymbol{\Sigma}_t$ evolving explicitly. Consequently, for linear PDEs with Gaussian initial distribution the calibration and test distributions admit closed-form likelihood ratios, enabling WCP.

Proposition 4.2. *Consider a linear evolution equation on a bounded domain with linear boundary conditions and a deterministic source term. If the calibration dataset at time t_{cal} and a test sample u_{test}^t at time $t \geq t_{\mathrm{cal}}$ are generated from the same affine numerical scheme on independent initial conditions sampled from a Gaussian distribution, then we can construct a likelihood-weighted conformal predictor such that*

$$\mathbb{P}\big(\boldsymbol{u}_{\mathrm{test}}^{t+\delta} \in \mathcal{C}_{t+\delta}^{\mathrm{WCP}}(\boldsymbol{u}_{\mathrm{test}}^t)\big) \geq 1 - \alpha.$$

[3] See Appendix C for a concrete example of such a function-space distribution.



(a) Drift PDE, Proposition 4.4

(b) Backward heat equation, Proposition 4.7

Figure 3: Visualization for Proposition 4.4 and Proposition 4.7. *(a)* Coverage decays exponentially with grid granularity for the baselines, while WCP grows infinite bands (dotted lines). The theoretical lower bound follows Barber et al. [3]. *(b)* Coverage of the analytical solution of the backward heat equation with numerical conformal bands. With finer granularity, coverage over the analytic solution converges to the target as the reconstruction error vanishes. Also here, the coverage decline effect of (a) would start dominating at higher granularity. More in Appendix H.4.

Proof sketch: As the initial distribution is Gaussian and the PDE linear, we can show that the solution at each time step is Gaussian with known mean and covariance. With this, we calculate the exact weights for WCP. The full proof is provided in Appendix G.2. □

Remark 4.3 (Other initial distributions). This result extends to any other initial distribution that is closed under affine transformations (e.g., Cauchy, Laplace, and Logistic). We include additional experiments sampling from different initial distributions, including heavy-tailed Cauchy initial conditions, in Appendix K.

We only know the input-time distribution, not the prediction-time one (Gaussian with unknown moments), so **the distribution itself cannot be used to infer uncertainty directly**.

Even though WCP is theoretically attainable in the setting we just identified, the resulting conformal bands can become too large in practice to provide useful information. In the next section, we present a way to alleviate this: We show that the *grid granularity* is a powerful lever for controlling the TV distance and conformal band size, and with this, the applicability of WCP.

4.4 Effect of grid granularity

So far we worked with $\boldsymbol{u}_i^t$ at arbitrary granularity. Refining the grid recovers the function-space limit, where Theorem 4.1 gives $d_{\mathrm{TV}} = 1$, raising the question of how granularity affects CP. We analyze this question for the *drift PDE* which represents slight non-stationary behavior. The following proposition shows that for this PDE, the TV distance approaches 1 exponentially fast in the grid size, so standard CP approaches become inapplicable even under slight drifts.

Proposition 4.4. *Consider the drift PDE $\partial_t u(x,t) = c$, with $x \in (0,1), t \geq 0, c \in \mathbb{R}$, and compatible boundary conditions. Define $\boldsymbol{z}$ and $\boldsymbol{z}^i$ as in Equation* (2) *with n i.i.d. calibration pairs at times $(t_1, t_1+\delta)$ and a test pair at time $(t_2, t_2+\delta)$ from an affine scheme on a grid with p points and Gaussian initial conditions. Then, the total variation distance between the laws of $\boldsymbol{z}$ and $\boldsymbol{z}^i$ is*

$$d_{\mathrm{TV}}(\boldsymbol{z}, \boldsymbol{z}^i) \;=\; 2\Phi\Big(\frac{\sqrt{2pc^2}\,|t_1 - t_2|}{2}\Big) - 1,$$

where Φ is the cumulative distribution function of the standard normal distribution. Hence, $d_{\mathrm{TV}}(\boldsymbol{z}, \boldsymbol{z}^i) \to 1$ exponentially fast in the grid size p and with it, the coverage guarantee from Barber et al. [3] vanishes exponentially.

Moreover, for likelihood-ratio weighting as in Equation (7) *evaluated at the distribution centers* $\boldsymbol{x}_{\text{test}} := \mathbb{E}[\boldsymbol{u}_{n+1}^{t_2}]$ *and* $\boldsymbol{x}_{\text{cal}} := \mathbb{E}[\boldsymbol{u}_i^{t_1}]$*, the test weight satisfies*

$$w_{\text{test}} = 1 - O\big(ne^{-pc^2(t_1-t_2)^2}\big), \quad w_{\text{cal}} = O\big(e^{-pc^2(t_1-t_2)^2}\big).$$

Hence, the weighted conformal bands become infinite at an exponential rate.

Proof sketch: The TV distance follows from Theorem 4.5 with $\boldsymbol{w} = \mathbf{1}$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{I}_p$, and for the weights, we calculate the explicit density ratio. The full proof is provided in Appendix G.6. □

We visualize Proposition 4.4 in Figure 3a. The drift example provides a good heuristic that a slight non-stationarity can lead to exponential decay of the coverage guarantees for general (non)linear PDEs. We extend this result to general linear PDEs in Proposition G.1.

The exponential growth in p relies on white-noise initial conditions, where every grid point carries independent information about the drift. In real data, neighboring grid points are usually correlated, which requires a more general statement. The following theorem replaces the grid size p by an *effective dimension* p_{eff}, which is small when the drift points along directions of high initial variance. Restricting CP to a subregion of the domain reduces the effective dimension and thereby improves the CP guarantees, though other approaches can be more suitable depending on the covariance structure of the data.

Theorem 4.5. *Consider the drift PDE on a grid with* p *points,* $\partial_t \boldsymbol{u}^t = c\,\boldsymbol{w}$ *with* $t \geq 0$*, drift strength* $c \in \mathbb{R}$*, and drift profile* $\boldsymbol{w} \in \mathbb{R}^p$ *(e.g.,* $\boldsymbol{w} = \mathbf{1}$ *for a constant drift), and let the initial conditions be non-degenerate Gaussian,* $\boldsymbol{u}^0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$*. For* $\boldsymbol{z}$ *and* $\boldsymbol{z}^i$ *as in Proposition 4.4,*

$$d_{\text{TV}}\big(\boldsymbol{z}, \boldsymbol{z}^i\big) \;=\; 2\Phi\Big(\frac{\sqrt{2\,p_{\text{eff}}\,c^2}\,|t_1 - t_2|}{2}\Big) - 1, \quad \textit{with } p_{\text{eff}} := \boldsymbol{w}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{w}.$$

Proof sketch: Whitening the covariance reduces the statement to the equal-covariance identity used for Proposition 4.4. The full proof is provided in Appendix G.5. □

For $\boldsymbol{w} = \mathbf{1}$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{I}_p$, we have $p_{\text{eff}} = p$, which recovers Proposition 4.4. Thus, the TV distance is dependent on the smoothness of the initial state and how the PDE drifts relative to it, even approaching 1 exponentially fast in the grid size.

Knowing that many PDEs of interest are non-stationary, and that fine grids are often required for accurate predictions, it may seem futile to apply non-exchangeable CP on PDEs at all. However, while the full grid is usually necessary for prediction, we often do not need uncertainty quantification over the whole space. Referencing our real-world examples in Section 5, we may find that the temperature of Lake Maggiore is most relevant at depths where certain fish are living, or that bleaching of the FRAP target is most interesting in the middle of the plate where the laser is focused. In such cases, Theorem 4.5 tells us to restrict CP to coherent regions of interest, such as the center of the bleach spot or a few neighboring depths in the lake, and leverage the more favorable properties of the marginal distribution. We demonstrate this in our experiments.

4.5 Applications and extensions of the framework

The previous theorem showed that non-exchangeable CP should only be applied to a subset of the grid to avoid exponential decay of the coverage guarantees and uninformative conformal bands. While WCP methods "only" suffer from vacuous prediction intervals instead of undercoverage, they face additional practical challenges: For numerical data, we can evolve the exact mean and covariance from the initial distribution. For real-world measurements, however, the true mean and covariance must be *estimated* from data, introducing additional uncertainty. While it is common in CP literature that some assumptions cannot be checked on real-world data (e.g., [32, 3]), we nonetheless want to give a *theoretical outlook* on how to account for this estimation uncertainty and recover coverage guarantees when additional assumptions can be made.

Proposition 4.6 (Robust WCP under Gaussian distributions)**.** *Under the assumptions of Proposition 4.2, suppose we additionally observe independent estimation samples* $\{\boldsymbol{u}_j^{t_{\text{cal}}}\}_{j=1}^{m_0}$ *and* $\{\boldsymbol{u}_j^{t}\}_{j=1}^{m_1}$

with $m_k > p$ for $k \in \{0, 1\}$, and fix an arbitrary estimation uncertainty $\eta > 0$. Then, we can construct a conformal prediction set $C^{\text{rob}}_{t+\delta}$ with coverage guarantee

$$\mathbb{P}\big(\boldsymbol{u}^{t+\delta}_{\text{test}} \in C^{\text{rob}}_{t+\delta}(\boldsymbol{u}^{t}_{\text{test}})\big) \;\geq\; 1 - \eta - \alpha.$$

Smaller η will usually correspond to wider bands.

Proof sketch: We use the sample distributions of the estimated means and covariances to construct a $1 - \eta$ confidence region over the true means and covariances. Then, we take the worst case quantiles over these regions. The full proof can be found in Appendix G.3. □

Without additional structure on the covariance, this procedure is intractable in practice, and only of theoretical value. Exploring tractable problem classes (e.g., with covariances diagonal in Fourier space) is left for future work. Empirically, such corrections are often unnecessary, as mean and covariance estimates converge rapidly with the number of samples. In particular, we find that constructing prediction bands from estimated mean and covariance yields consistent coverage, whereas estimating the full distribution shift, as the original WCP does, often leads to undercoverage.

4.6 Recovering guarantees for the true solution

While we saw in Section 4.2 that function-space coverage is generally intractable, numerical schemes typically come with deterministic error bounds of order $O(h^k)$ in the grid size h. The following proposition lifts a discrete CP set C_h to a conformal set on the true solution by inflating C_h with this error: the convergence *rate* carries over (even though unknown constants typically prevent computing C explicitly), and depends only on the prediction granularity rather than that used for C_h. See Figure 3b for an empirical demonstration.

Proposition 4.7. *Let $u(\cdot, t)$ be a PDE solution in some function space $(\mathcal{U}, \|\cdot\|_{\mathcal{U}})$ and $\boldsymbol{u}^t$ be its approximation via some numerical scheme on grid size h. Let C_h be a conformal prediction set for the discretized state with coverage $1 - \alpha$ and assume that an error bound $\big\|u(\cdot, t) - \boldsymbol{I}_h \boldsymbol{u}^t\big\|_{\mathcal{U}} \;\leq\; \varepsilon_{h,t}(u_0)$ is available via some reconstruction operator $\boldsymbol{I}_h : \mathbb{R}^p \to \mathcal{U}$ (e.g., piecewise interpolation). Then we can define a conformal set for the function-space setting, $C \subset \mathcal{U}$, with $\mathbb{P}\big(u(\cdot, t) \in C\big) \;\geq\; 1 - \alpha$ by*

$$C := \big\{v \in \mathcal{U} : \exists\, \boldsymbol{z} \in C_h \text{ with } \|v - \boldsymbol{I}_h \boldsymbol{z}\|_{\mathcal{U}} \leq \varepsilon_{h,t}(u_0)\big\}. \tag{3}$$

Proof sketch: We use the error bound to extend the conformal set for the numerical solution using the triangle inequality. The full proof is given in Appendix G.4 and a visualization in Figure 3b. □

5 Experiments

We demonstrate our theoretical results on both synthetic and real-world data, in particular the coverage of WCP with Gaussian likelihood ratio.

Experimental setup For the **synthetic case**, we choose the general second-order PDE framework, which covers many well-known PDEs $u_t = au_{xx} + bu_x + cu, x \in (0, 2\pi),\ t \geq 0$ where $a, b, c \in \mathbb{R}$ are tunable parameters and u_0 is sampled from a multivariate Gaussian. We target setups where the PDE becomes more unstable over time ($a < 0$). Otherwise, the PDE smooths out over time, leading to trivial coverage at future time steps. Hence, we consider $a = \{-0.005, -0.0075, -0.01\}$ and for the remaining parameters we choose $b, c \in \{-0.5, 0, 0.5\}$. See Figure H.1 for exemplary data. For **real-world scenarios**, we use pulsed-thermography measurements from Wei et al. [34], diffusion dynamics from FRAP-microscopy (Fluorescence recovery after photobleaching) [35] and heat-diffusion from sensors in Lake Maggiore under seasonal drift [31] (details in Appendices I.1, I.2, and I.3). We consider the following tasks, targeting 90% coverage:[4] **(1) Recalibration case:** Starting from $t_0 = 0$, we generate states at time steps $t_i = t_0 + i\delta$. We use pairs (t_i, t_{i+1}) for calibration and test on (t_{i+1}, t_{i+2}) for $i \in \{0, \ldots, 18\}$ (we also include the calibration time coverage as the first step). The inverse scenario is obtained by reversing time. **(2) General case:** As above, but no recalibration is allowed after t_0, corresponding to deployment at an unknown future time.

[4] A sample is considered covered only if all spatial points are covered. The score is defined as the maximum absolute deviation over the spatial dimension.

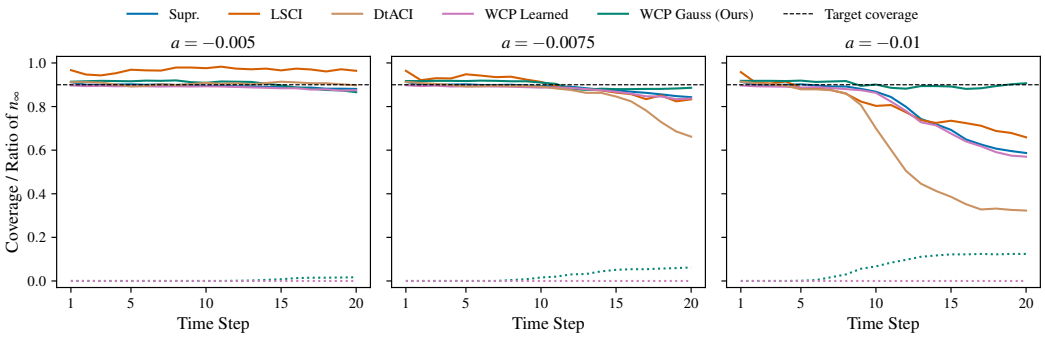


Figure 4: Coverages for varying a and fixed $b = 0, c = 0.5$ on the recalibration task. The 90% target is marked by a black dashed line. For WCP Gauss and WCP Learned, the fraction of samples with infinite bandwidth n_∞ is indicated by green and pink dotted lines, respectively. Coverage holds for some parameters, but small variations induce undercoverage at later time steps. Only WCP Gauss meets the target throughout.

Baselines We define four baselines for our experiments. The first is a naïve implementation (**Supr.**) which depends on exchangeability [7]. Secondly, we use WCP with learned weights, as described in Tibshirani et al. [32] (**WCP Learned**), which serves as our main baseline. Thirdly, we use the **LSCI** method [15], a novel approach for CP on PDEs (note Appendix D). Finally, we use Dynamically-Tuned Adaptive Conformal Inference (**DtACI**) by Gibbs and Candès [10], an updated version of the original Adaptive CP [9] with automatic parameter selection. However, as adaptive CP requires an evaluation of the error at every step, this method is only applicable to the recalibration task. We compare these to our WCP method (**WCP Gauss**), which only estimates mean and covariance for each model input to construct the weights. In cases where the distributional dissimilarity of u_t and $u_{t+\delta}$ is too large, the WCP methods predict infinite bands. We report the fraction of these samples as n_∞. A band is wider the larger the likelihood ratio of its test sample, and infinite once the normalized test weight exceeds α (Appendix J.3). Note that in safety-critical applications, reporting trivial bands is often a more valuable result than delivering bands with undercoverage.

Additional experiments To demonstrate the effect of our derived results, Figure 3 provides visualizations of Proposition 4.4 and Proposition 4.7. More details in Appendix H.4. We additionally include an ablation study in Appendix H.3 that stress-tests WCP Gauss under reduced sample sizes and increased dimensionality. As expected, when the number of samples is insufficient relative to the problem dimensionality, instability in covariance estimation can lead to undercoverage. Coverage is stable across random calibration/test splits and, for the pulsed-thermography data, across plate assignments with retrained surrogates (Appendix J.1).

Results Exemplary figures for synthetic and real-world coverage results can be seen in Figure 4 and Figure I.2 respectively. Results for all remaining parameter combinations (incl. tables) are provided in Appendix K.[5] Across both synthetic and real-world tasks, **WCP Gauss** is the only method that consistently attains the target coverage. In general, WCP Gauss can estimate the weights more easily than WCP Learned, and DtACI would require longer time windows to adapt (which may require costly measurements). This does not mean that coverage with other methods is impossible. While other methods can achieve coverage in specific settings or with careful tuning or more expressive models (e.g., for *WCP Learned*), this behavior does not generalize reliably. In practice, small changes such as grid resolution, PDE parameters, or evaluation time steps can already lead to undercoverage. This issue is often hidden in smoothing forward problems, where residuals shrink over time, but becomes apparent in inverse problems, unstable dynamics, or blow-up regimes. Overall, **WCP is the only method providing formal guarantees**, and we can see empirically that this is a clear advantage as soon as our system exhibits significant dynamics.

[5] Bandwidth results are reported in Appendix K. Our WCP method does not optimize for bandwidth, but could certainly be extended with more advanced score functions or CP algorithms.

6 Limitations

The convergence results for increasing grid granularity (Proposition 4.4) should be read as a heuristic for (non)linear PDEs: with increasing granularity, even small drifts can cause the TV distance to deteriorate rapidly. Of course, one can construct PDEs with no drift at all, which may exhibit more favorable TV behavior. Similar considerations apply to the vacuous TV bound in function spaces. Identifying the conditions under which these effects occur more generally is left for future work. Finally, while the lifting proposition (Proposition 4.7) is independent of the choice of PDE, the proposed WCP method (Proposition 4.2 together with Proposition 4.6) relies strictly on linearity and loses its coverage guarantees for nonlinear PDEs, where coverage depends on the PDE parameters (Appendix J.2).

7 Discussion

Conformal prediction for time-dependent physical phenomena is often constrained by non-exchangeable data. In this work, we investigated whether coverage guarantees can be maintained beyond the exchangeability assumption and show that this depends strongly on the problem representation. On function spaces, measures are typically mutually singular, making coverage guarantees unattainable. On discretized domains, however, we derived how weighted CP can be applied to linear PDEs to obtain coverage guarantees. We further show that these guarantees can be preserved for real-world measurements, and that CP should be restricted to a coarse, low-dimensional grid to provide meaningful guarantees. A natural next step is to extend these results to specific classes of (non)linear PDEs.

Acknowledgments and Disclosure of Funding

The authors gratefully acknowledge the computational and data resources provided by the Leibniz Supercomputing Centre (LRZ). JS acknowledges funding from the Munich Center for Machine Learning (MCML), as well as the support he received from the METEOR Research Training Group, which is sponsored by the German Research Foundation. AS is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Research, Technology and Space. The authors declare no competing interests.

References

[1] Anastasios Angelopoulos, Emmanuel Candès, and Ryan J Tibshirani. Conformal PID control for time series prediction. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 23047–23074. Curran Associates, Inc., 2023. doi: 10.52202/075280-1000. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/47f2fad8c1111d07f83c91be7870f8db-Paper-Conference.pdf.

[2] Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with modern Hopfield networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 56027–56074. Curran Associates, Inc., 2023. doi: 10.52202/075280-2444. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/aef75887979ae1287b5deb54a1e3cbda-Paper-Conference.pdf.

[3] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, April 2023. ISSN 0090-5364, 2168-8966. doi: 10.1214/23-AOS2276. URL https://projecteuclid.org/journals/annals-of-statistics/volume-51/issue-2/Conformal-prediction-beyond-exchangeability/10.1214/23-AOS2276.full.

[4] S.S. Barsov and V.V. Ul'yanov. Estimates of the proximity of Gaussian measures. *Soviet Mathematics–Doklady*, 34:462–466, 1987.

[5] Lukas Biewald. Experiment Tracking with Weights and Biases, 2020. URL https://www.wandb.com/.

[6] Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*. Cambridge University Press, 1 edition, December 1992. ISBN 978-0-521-38529-9 978-0-521-05980-0 978-0-511-66622-3. URL https://www.cambridge.org/core/product/identifier/9780511666223/type/book.

[7] Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Conformal prediction bands for multivariate functional data. *Journal of Multivariate Analysis*, 189:104879, May 2022. ISSN 0047259X. doi: 10.1016/j.jmva.2021.104879. URL https://linkinghub.elsevier.com/retrieve/pii/S0047259X21001573.

[8] António Farinhas, Chrysoula Zerva, Dennis Ulmer, and Andre Martins. Non-exchangeable conformal risk control. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, pages 50952–50966, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/de04896f011beff76c91e094f72727f4-Paper-Conference.pdf.

[9] Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0d441de75945e5acbc865406fc9a2559-Paper.pdf.

[10] Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024. URL https://jmlr.org/papers/v25/22-1218.html.

[11] Vignesh Gopakumar, Ander Gray, Lorenzo Zanisi, Timothy Nunn, Daniel Giles, Matt Kusner, Stanislas Pamela, and Marc Peter Deisenroth. Calibrated physics-informed uncertainty quantification. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 20103–20141. PMLR, July 2025. URL https://proceedings.mlr.press/v267/gopakumar25a.html.

[12] Vignesh Gopakumar, Ander Gray, Joel Oskarsson, Lorenzo Zanisi, Daniel Giles, Matt J Kusner, Stanislas Pamela, and Marc Peter Deisenroth. Uncertainty quantification of surrogate models using conformal prediction. *Machine Learning: Science and Technology*, 7(1):015025, 2026. URL http://dx.doi.org/10.1088/2632-2153/ae2e7b.

[13] Ander Gray, Vignesh Gopakumar, Sylvain Rousseau, and Sebastien Destercke. Guaranteed Prediction Sets for Functional Surrogate Models. In Silvia Chiappa and Sara Magliacane, editors, *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 1569–1585. PMLR, July 2025. URL https://proceedings.mlr.press/v286/gray25a.html.

[14] Martin Hairer. An Introduction to Stochastic PDEs, July 2023. URL http://arxiv.org/abs/0907.4178. arXiv:0907.4178 [math.PR].

[15] Trevor Harris and Yan Liu. Locally adaptive conformal inference for operator models, July 2025. URL http://arxiv.org/abs/2507.20975.

[16] Haiyun He and Ziv Goldfeld. Information-theoretic generalization bounds for deep neural networks. *IEEE Transactions on Information Theory*, 71(8):6227–6247, 2025. doi: 10.1109/TIT.2025.3568697. URL https://doi.org/10.1109/TIT.2025.3568697.

[17] Wolfgang Karl Härdle and Léopold Simar. *Applied Multivariate Statistical Analysis*. Springer International Publishing, Cham, 2019. ISBN 978-3-030-26005-7 978-3-030-26006-4. doi: 10.1007/978-3-030-26006-4. URL http://link.springer.com/10.1007/978-3-030-26006-4.

[18] Lazaros Karageorgiou and Kyprianos-Iason Prodromidis. Limit profile for the high-temperature Curie-Weiss model, October 2025. URL http://arxiv.org/abs/2510.18793.

[19] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, February 2004. ISSN 0047259X. doi: 10.1016/S0047-259X(03)00096-4. URL https://linkinghub.elsevier.com/retrieve/pii/S0047259X03000964.

[20] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural Operator: Graph Kernel Network for Partial Differential Equations, March 2020. URL http://arxiv.org/abs/2003.03485.

[21] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

[22] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, March 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00302-5. URL https://www.nature.com/articles/s42256-021-00302-5.

[23] Ziqi Ma, David Pitt, Kamyar Azizzadenesheli, and Anima Anandkumar. Calibrated uncertainty quantification for operator learning via conformal prediction. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. doi: 10.48550/arXiv.2402.01960. URL https://openreview.net/forum?id=cGpegxy12T.

[24] Amirhossein Mollaali, Gabriel Zufferey, Gonzalo Constante-Flores, Christian Moya, Can Li, Meng Yue, and Guang Lin. Conformalized prediction of post-fault voltage trajectories using pre-trained and finetuned attention-driven neural operators. *Neural Networks*, 192:107809, 2025. doi: 10.1016/J.NEUNET.2025.107809. URL https://doi.org/10.1016/j.neunet.2025.107809.

[25] Christian Moya, Amirhossein Mollaali, Zecheng Zhang, Lu Lu, and Guang Lin. Conformalized-DeepONet: a distribution-free framework for uncertainty quantification in deep operator networks. *Physica D: Nonlinear Phenomena*, 471:134418, January 2025. ISSN 01672789. doi: 10.1016/j.physd.2024.134418. URL https://linkinghub.elsevier.com/retrieve/pii/S0167278924003683.

[26] Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Statistics. Wiley, 1 edition, March 1982. ISBN 978-0-471-09442-5 978-0-470-31655-9. doi: 10.1002/9780470316559. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316559.

[27] Roberto I. Oliveira, Paulo Orenstein, Thiago Ramos, and João Vitor Romano. Split Conformal Prediction and Non-Exchangeable Data. *Journal of Machine Learning Research*, 25(225):1–38, 2024. URL http://jmlr.org/papers/v25/23-1553.html.

[28] Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 844–853. PMLR, July 2021. URL https://proceedings.mlr.press/v161/podkopaev21a.html.

[29] M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, February 2019. ISSN 00219991. doi: 10.1016/j.jcp.2018.10.045. URL https://linkinghub.elsevier.com/retrieve/pii/S0021999118307125.

[30] Fabricio Dos Santos, Tara Akhound-Sadegh, and Siamak Ravanbakhsh. Physics-informed transformer networks. In *The Symbiosis of Deep Learning and Differential Equations III*, 2023. URL https://openreview.net/forum?id=zu80h9YryU.

[31] Kara Scheu, Oliver Fringer, Derek Fong, and Stephen Monismith. Observation data and numerical model files for river plumes in Pallanza Bay of Lake Maggiore, Italy, April 2018. URL https://zenodo.org/record/1206187.

[32] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf.

[33] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer International Publishing, Cham, 2022. ISBN 978-3-031-06648-1 978-3-031-06649-8. URL https://link.springer.com/10.1007/978-3-031-06649-8.

[34] Ziang Wei, Ahmad Osman, Bernd Valeske, and Xavier Maldague. Pulsed thermography dataset for training deep learning models. *Applied Sciences*, 13(5), 2023. ISSN 2076-3417. doi: 10.3390/app13052901. URL https://www.mdpi.com/2076-3417/13/5/2901.

[35] Victor Wåhlstrand Skärström, Annika Krona, Niklas Lorén, and Magnus Röding. DeepFRAP: fast fluorescence recovery after photobleaching data analysis using deep neural networks. *Journal of Microscopy*, 282(2):146–161, May 2021. ISSN 0022-2720, 1365-2818. doi: 10.1111/jmi.12989. URL https://onlinelibrary.wiley.com/doi/10.1111/jmi.12989.

[36] Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11559–11569. PMLR, July 2021. URL https://proceedings.mlr.press/v139/xu21h.html.

[37] Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38707–38727. PMLR, July 2023. URL https://proceedings.mlr.press/v202/xu23r.html.

[38] Chen Xu, Hanyang Jiang, and Yao Xie. Conformal prediction for multi-dimensional time series by ellipsoidal sets. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 55076–55099. PMLR, July 2024. URL https://proceedings.mlr.press/v235/xu24m.html.

[39] Anthony Zhou and Amir Barati Farimani. Predicting change, not states: An alternate framework for neural PDE surrogates. *Computer Methods in Applied Mechanics and Engineering*, 441:117990, June 2025. ISSN 00457825. doi: 10.1016/j.cma.2025.117990. URL https://linkinghub.elsevier.com/retrieve/pii/S0045782525002622.

A LLM disclaimer

Large language models (LLMs) were used to support the development and execution of this project. They contributed to improving the clarity and readability of the manuscript, suggesting alternative phrasings, helping with the code, offering feedback on mathematical arguments, and providing ideas throughout the research and coding process. All research contributions, results, and final formulations were manually verified.

B Licenses of external assets

All external assets used in this work are publicly available. The pulsed-thermography dataset of Wei et al. [34] is released under CC-BY 4.0 and hosted on Kaggle.[6] The Lake Maggiore thermistor data of Scheu et al. [31] is archived on Zenodo under CC-BY 4.0.[7] The FRAP dataset and reference code of Wåhlstrand Skärström et al. [35] are archived on Zenodo under CC-BY 4.0.[8]

C Example: a Gaussian random field on a function space

A probability distribution on a function space is very similar to a distribution on $\mathbb{R}^p$. Where a draw from a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ on $\mathbb{R}^p$ produces a vector of p numbers, a draw from a distribution on $L^2(\Omega)$ produces an entire function $u : \Omega \to \mathbb{R}$. Each sample is a function over Ω, and different draws give different functions.

The simplest and most common such distributions are *Gaussian random fields* (GRFs), the infinite-dimensional analogue of multivariate Gaussians. Like a Gaussian on $\mathbb{R}^p$, a GRF is determined by two ingredients: a mean function $\boldsymbol{m} \in L^2(\Omega)$ playing the role of $\boldsymbol{\mu}$, and a covariance operator $\boldsymbol{C}$ playing the role of $\boldsymbol{\Sigma}$. The covariance operator $\boldsymbol{C} : L^2(\Omega) \to L^2(\Omega)$ must be positive, symmetric, and trace class [14]. The trace-class condition is what makes the infinite sum of variances finite, ensuring samples actually lie in $L^2(\Omega)$. We write $u \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{C})$.

The most useful way to picture a sample is via its *Karhunen–Loève expansion*. If $(\mu_n, e_n)_{n\geq 1}$ are the eigenvalues and eigenfunctions of $\boldsymbol{C}$ (so $\boldsymbol{C}e_n = \mu_n e_n$), then a draw $u \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C})$ can be written as

$$u(x) \;=\; \sum_{n=1}^{\infty} \sqrt{\mu_n}\, \xi_n\, e_n(x), \qquad \xi_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \tag{4}$$

i.e., a random superposition of the basis functions e_n with independent Gaussian coefficients whose variances μ_n are set by $\boldsymbol{C}$. The decay rate of μ_n controls how rough or smooth the samples look: fast decay suppresses high-frequency e_n and gives smooth samples, slow decay gives rough ones.

For Theorem 4.1, we take $\Omega = (0,1)$, $\boldsymbol{m} = \mathbf{0}$, and $\boldsymbol{C} = (\boldsymbol{I} - \boldsymbol{\Lambda})^{-1}$ with $\boldsymbol{\Lambda}$ the Laplace operator under homogeneous Dirichlet boundary conditions. The eigenfunctions $e_n(x) = \sqrt{2}\sin(n\pi x)$ are sine modes of increasing frequency, and the eigenvalues $\mu_n = 1/(1 + (n\pi)^2)$ decay like $1/n^2$. The trace condition $\sum_n \mu_n < \infty$ holds, so $\mathcal{N}(\mathbf{0}, \boldsymbol{C})$ is a well-defined GRF on $L^2(\Omega)$. Each sample is a continuous function on $(0,1)$ that vanishes at the endpoints (because all e_n do), with most of its energy in the low-frequency modes.

D Validating local exchangeability

To have formal guarantees with the LSCI approach from [15], it is required that the model's *residuals* r_t are locally exchangeable. Considering this for the simplest setup of two time points t and $t + \delta$ for some $\delta > 0$, this means that it must hold that

$$d_{\mathrm{TV}}\Big(\begin{pmatrix} r_t \\ r_{t+\delta} \end{pmatrix}, \begin{pmatrix} r_{t+\delta} \\ r_t \end{pmatrix}\Big) \leq d(t, t+\delta)$$

for a pre-metric d on the time domain. Note, however, that since we do not have access to the laws of the random vectors above, we cannot reason about their TV distance. Even though in Proposition 4.2

[6] `https://www.kaggle.com/datasets/ziangwei/irtpvc`
[7] `https://doi.org/10.5281/zenodo.1206188`
[8] `https://doi.org/10.5281/zenodo.3874218`

we derive the laws of the solutions u_t, it is not clear how to reason about the above left-hand side without further assumptions (like independence of the residuals over time, which is not plausible as the solution has a clear time-dependence in all non-stationary PDEs).

E Additional figures

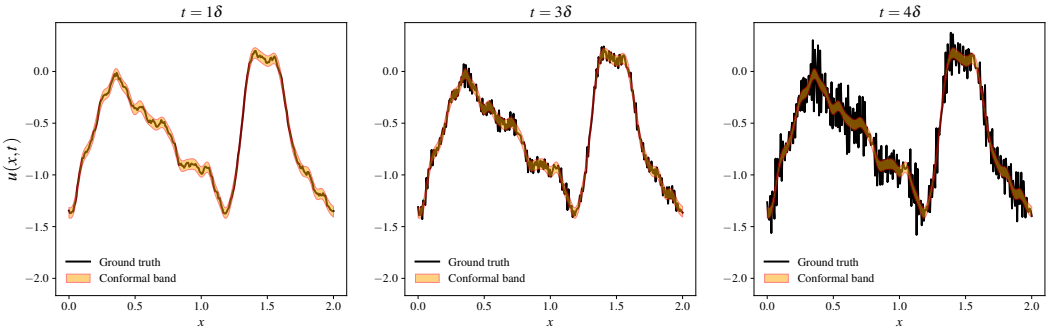


Figure E.1: CP bands of the backward heat equation calibrated at $t = 0$. Each panel shows one time point of the solution $u(x, t)$ and CP bands over the spatial domain x. While at time $t = \delta$, the solution is properly covered, coverage is lost at later time steps.

F Weighted conformal prediction formula

For completeness, we recall the explicit form of the weighted split conformal predictor of Tibshirani et al. [32] that underlies our *WCP Gauss* method (Proposition 4.2).

Given a calibration set $\mathcal{D}_{\text{cal}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, a fixed score function S, and unnormalized weights w_i on $\mathcal{D}_{\text{cal}} \cup \{\boldsymbol{x}_{\text{test}}\}$ (e.g., Equation (1) or, in our case, Equation (7)), define the normalized weights

$$\tilde{w}_i := \frac{w_i}{\sum_{j=1}^n w_j + w_{\text{test}}},\ i = 1, \dots, n, \qquad \tilde{w}_{n+1} := \frac{w_{\text{test}}}{\sum_{j=1}^n w_j + w_{\text{test}}}, \tag{5}$$

where w_{test} is the weight evaluated at $\boldsymbol{x}_{\text{test}}$. Writing $S_i := S(\boldsymbol{x}_i, y_i)$ for the calibration scores, the weighted split conformal prediction set at level $1 - \alpha$ is

$$\mathcal{C}^{\text{WCP}}(\boldsymbol{x}_{\text{test}}) := \left\{ y \in \mathbb{R} \ :\ S(\boldsymbol{x}_{\text{test}}, y) \le Q_{1-\alpha}\Big(\sum_{i=1}^n \tilde{w}_i\, \delta_{S_i} + \tilde{w}_{n+1}\, \delta_{+\infty} \Big) \right\}, \tag{6}$$

where δ_a denotes a point mass at a and $Q_{1-\alpha}(\mu)$ the $(1-\alpha)$-quantile of the discrete measure μ. The atom at $+\infty$ forces $\mathcal{C}^{\text{WCP}} = \mathbb{R}$ whenever $\tilde{w}_{n+1} > \alpha$, i.e., whenever the test point is so unlike the calibration set that no finite quantile carries enough mass. This corresponds to the infinite bands reported as n_∞ in our experiments. With weights set to the true density ratio between test and calibration distributions, the predictor satisfies $\mathbb{P}\big(y_{\text{test}} \in \mathcal{C}^{\text{WCP}}(\boldsymbol{x}_{\text{test}})\big) \ge 1 - \alpha$ [32]. Standard split CP (Section 3.1) is recovered by setting $w_i \equiv 1$, which collapses all normalized weights to $1/(n+1)$.

Gaussian WCP algorithm Algorithm 1 below instantiates the WCP set in Equation (6) using Gaussian density estimates at calibration and prediction times, with covariance estimated by the Ledoit–Wolf shrinkage estimator [19] for stability in the high-dimensional regime. We list the algorithm with different sets for calibration and test. In practice, it should be common to use the same trajectories at different time points for calibration and testing. We show both in our real-world tasks, where the pulsed-thermography and FRAP datasets use the same trajectories, but Lake Maggiore splits the data (at least it splits them into early-spring and late-spring).

Algorithm 1: WCP for time-dependent PDEs.

Input: calibration pairs $\{(\boldsymbol{u}_i^{t_{\mathrm{cal}}}, \boldsymbol{u}_i^{t_{\mathrm{cal}}+\delta})\}_{i=1}^n$, test-time inputs $\{\tilde{\boldsymbol{u}}_j^t\}_{j=1}^{m+1}$ at time $t \geq t_{\mathrm{cal}}$ (with $\tilde{\boldsymbol{u}}_{m+1}^t$ the query), nonconformity score S, miscoverage level α.
Output: prediction set $\mathcal{C}_{t+\delta}^{\mathrm{WCP}}(\tilde{\boldsymbol{u}}_{m+1}^t)$.

$\mathcal{D}_{t_{\mathrm{cal}}} \leftarrow \{\boldsymbol{u}_i^{t_{\mathrm{cal}}}\}_{i=1}^n,\ \mathcal{D}_t \leftarrow \{\tilde{\boldsymbol{u}}_j^t\}_{j=1}^m$
for $\tau \in \{t_{\mathrm{cal}},\ t\}$ **do**
  $\hat{\boldsymbol{\mu}}_\tau \leftarrow \frac{1}{|\mathcal{D}_\tau|} \sum_{\boldsymbol{u} \in \mathcal{D}_\tau} \boldsymbol{u}$ `// sample mean`
  $\hat{\boldsymbol{\Sigma}}_\tau \leftarrow \mathrm{LedoitWolf}(\mathcal{D}_\tau)$ `// Ledoit-Wolf shrinkage`
for $i = 1, \ldots, n$ **do**
  $w_i \leftarrow \varphi(\boldsymbol{u}_i^{t_{\mathrm{cal}}}; \hat{\boldsymbol{\mu}}_t, \hat{\boldsymbol{\Sigma}}_t) / \varphi(\boldsymbol{u}_i^{t_{\mathrm{cal}}}; \hat{\boldsymbol{\mu}}_{t_{\mathrm{cal}}}, \hat{\boldsymbol{\Sigma}}_{t_{\mathrm{cal}}})$ `// importance weight`
  $S_i \leftarrow S(\boldsymbol{u}_i^{t_{\mathrm{cal}}}, \boldsymbol{u}_i^{t_{\mathrm{cal}}+\delta})$ `// calibration score`
$w_{m+1} \leftarrow \varphi(\tilde{\boldsymbol{u}}_{m+1}^t; \hat{\boldsymbol{\mu}}_t, \hat{\boldsymbol{\Sigma}}_t) / \varphi(\tilde{\boldsymbol{u}}_{m+1}^t; \hat{\boldsymbol{\mu}}_{t_{\mathrm{cal}}}, \hat{\boldsymbol{\Sigma}}_{t_{\mathrm{cal}}})$ `// test weight`
$\tilde{w}_i \leftarrow w_i / (\sum_j w_j + w_{m+1}),\ \tilde{w}_{n+1} \leftarrow w_{m+1} / (\sum_j w_j + w_{m+1})$ `// normalize, cf. Equation (5)`
$\hat{q}_{1-\alpha} \leftarrow Q_{1-\alpha}\big(\sum_{i=1}^n \tilde{w}_i\, \delta_{S_i} + \tilde{w}_{n+1}\, \delta_{+\infty}\big)$
return $\mathcal{C}_{t+\delta}^{\mathrm{WCP}}(\tilde{\boldsymbol{u}}_{m+1}^t) = \{\, \boldsymbol{y} : S(\tilde{\boldsymbol{u}}_{m+1}^t, \boldsymbol{y}) \leq \hat{q}_{1-\alpha} \,\}$

G Proofs

G.1 Proof of Theorem 4.1

Theorem 4.1. *Consider the one-dimensional heat equation $\partial_t u = \partial_{xx} u$ on the domain $\Omega = (0, 1)$ with Dirichlet boundary conditions, $t \geq 0$. Suppose the initial condition $u_0(x)$ is sampled from a Gaussian random field $\mathcal{P}_0 = \mathcal{N}(\mathbf{0}, (\boldsymbol{I} - \boldsymbol{\Lambda})^{-1})$, where $\boldsymbol{\Lambda}$ is the Laplace operator on Ω. Then, the TV distance between the solution distributions at any two time points is maximal:*

$$d_{\mathrm{TV}}(\mathcal{P}_t, \mathcal{P}_{t+\delta}) = 1, \quad \forall\, t \geq 0,\ \delta > 0.$$

Proof. Notice that our whole setup is on a Hilbert space. We begin by showing that $(\boldsymbol{I} - \boldsymbol{\Lambda})^{-1}$ is a well-defined covariance operator on $L^2(\Omega)$. For this, according to [14, Proposition 4.17], the operator must be positive, symmetric and trace class. As both $\boldsymbol{I}$ and $-\boldsymbol{\Lambda}$ are positive and symmetric, so is their sum. As $-\boldsymbol{\Lambda}$ is positive, it follows that $(\boldsymbol{I} - \boldsymbol{\Lambda})$ is invertible, due to strictly positive eigenvalues. Further, the eigenvalues of $-\boldsymbol{\Lambda}$ with Dirichlet boundary conditions are given by $\lambda_n = (n\pi)^2$, with corresponding eigenfunctions $e_n(x) = \sqrt{2}\sin(n\pi x)$, which form an orthonormal basis of $L^2(\Omega)$. Therefore, the eigenvalues of $(\boldsymbol{I} - \boldsymbol{\Lambda})^{-1}$ are given by $\mu_n = 1/(1 + (n\pi)^2)$ and as

$$\sum_{n=1}^{\infty} \mu_n < \infty,$$

we conclude that $(\boldsymbol{I} - \boldsymbol{\Lambda})^{-1}$ is a trace class operator and thus defines a Gaussian measure on $L^2(\Omega)$.

Now our proof will be based on the Feldman–Hájek theorem [6, Theorem 2.23], which gives a characterization of when two Gaussian measures on a Hilbert space are either equivalent or mutually singular. We will briefly state the whole chain of reasoning, and then provide the necessary details.

We will show that our measure at all times is Gaussian. By the Feldman–Hájek theorem, two Gaussian measures $\mathcal{N}(\boldsymbol{m}_1, \boldsymbol{C}_1)$ and $\mathcal{N}(\boldsymbol{m}_2, \boldsymbol{C}_2)$ on a Hilbert space are either equivalent or mutually singular. A necessary condition for equivalence is that the Cameron–Martin spaces, as given by $\boldsymbol{C}^{1/2}$, of the two measures are equal as sets [6, Theorem 2.23]. Thus, if the ranges of the covariance operators $\boldsymbol{C}_1^{1/2}$ and $\boldsymbol{C}_2^{1/2}$ are not equal, then the measures are mutually singular and their TV distance is 1.

Calculating the covariance operators Starting with a measure μ_0 of the initial distribution, the heat equation induces a semigroup $S(t) = \exp(t\boldsymbol{\Lambda})$, which maps the initial condition u_0 to the solution at time t, i.e., $u(\cdot, t) = S(t)u_0$. Therefore, the measure μ_t of $u(\cdot, t)$ is induced by the

pushforward of μ_0 under $S(t)$, i.e., $\mu_t = S(t)_{\#}\mu_0$. As $S(t)$ is linear, μ_t is also a Gaussian measure with mean $\mathbf{0}$ and covariance operator

$$\boldsymbol{C}_t = S(t)(\boldsymbol{I} - \boldsymbol{\Lambda})^{-1}S(t)^* = \exp(t\boldsymbol{\Lambda})(\boldsymbol{I} - \boldsymbol{\Lambda})^{-1}\exp(t\boldsymbol{\Lambda}),$$

where $S(t)^*$ denotes the adjoint of $S(t)$ [14, Chap. 4.3]. As we have seen above, the eigenvalues of $(\boldsymbol{I} - \boldsymbol{\Lambda})^{-1}$ are given by $\mu_n = 1/(1 + (n\pi)^2)$, with corresponding eigenfunctions $e_n(x)$. Further, the eigenfunctions of $\boldsymbol{\Lambda}$ are also given by $e_n(x)$, with corresponding eigenvalues $\lambda_n = -(n\pi)^2$. Lastly, by functional calculus, the eigenfunctions of $\exp(t\boldsymbol{\Lambda})$ are also given by $e_n(x)$ with corresponding eigenvalues $\nu_n = \exp(-t(n\pi)^2)$. With this, we can compute

$$\begin{aligned}\boldsymbol{C}_t\boldsymbol{e}_n &= \exp(t\boldsymbol{\Lambda})(\boldsymbol{I} - \boldsymbol{\Lambda})^{-1}\exp(t\boldsymbol{\Lambda})\boldsymbol{e}_n \\ &= (\exp(-t(n\pi)^2))\left(\frac{1}{1 + (n\pi)^2}\right)(\exp(-t(n\pi)^2))\boldsymbol{e}_n = \frac{\exp(-2t(n\pi)^2)}{1 + (n\pi)^2}\boldsymbol{e}_n.\end{aligned}$$

Thus, the eigenvalues of $\boldsymbol{C}_t$ are given by $\lambda_n(t) = \nu_n^2\mu_n = \exp(-2t(n\pi)^2)/(1 + (n\pi)^2)$, with corresponding eigenfunctions $e_n(x)$.

Calculating the Cameron–Martin spaces The functions $e_n(x)$ form an orthonormal basis of $L^2(\Omega)$, so we can express every element $f \in L^2(\Omega)$ as

$$f = \sum_{n=1}^{\infty} c_n e_n, \quad \sum_{n=1}^{\infty} c_n^2 < \infty.$$

The Cameron–Martin space H_t of $\mathcal{P}_t$ is given by the range of $\boldsymbol{C}_t^{1/2}$, which is given by

$$\mathrm{Ran}(\boldsymbol{C}_t^{1/2}) = \left\{\boldsymbol{C}_t^{1/2}f \mid f \in L^2(\Omega)\right\} = \left\{\sum_{n=1}^{\infty}\sqrt{\lambda_n(t)}c_n e_n \mid f \in L^2(\Omega)\right\}.$$

Therefore, $g \in H_t$ if and only if g can be expressed as

$$g = \sum_{n=1}^{\infty} d_n e_n, \quad \sum_{n=1}^{\infty} \frac{d_n^2}{\lambda_n(t)} < \infty.$$

Inserting the expression for $\lambda_n(t)$, we see that $g \in H_t$ if and only if

$$\sum_{n=1}^{\infty} d_n^2 \frac{1 + (n\pi)^2}{\exp(-2t(n\pi)^2)} < \infty.$$

Showing that the Cameron–Martin spaces are not equal Now it is easy to see that for any $t \geq 0$, $\delta > 0$, the Cameron–Martin spaces H_t and $H_{t+\delta}$ are not equal. For example, the function

$$h(x) = \sum_{n=1}^{\infty} \exp(-(t + \delta)(n\pi)^2)e_n(x)$$

is an element of H_t, as

$$\sum_{n=1}^{\infty}(\exp(-(t + \delta)(n\pi)^2))^2 \frac{1 + (n\pi)^2}{\exp(-2t(n\pi)^2)} = \sum_{n=1}^{\infty}(1 + (n\pi)^2)\exp(-2\delta(n\pi)^2) < \infty,$$

but it is not an element of $H_{t+\delta}$, as

$$\sum_{n=1}^{\infty}(\exp(-(t + \delta)(n\pi)^2))^2 \frac{1 + (n\pi)^2}{\exp(-2(t + \delta)(n\pi)^2)} = \sum_{n=1}^{\infty}(1 + (n\pi)^2) = \infty.$$

Therefore, by the Feldman–Hájek theorem, the measures $\mathcal{P}_t$ and $\mathcal{P}_{t+\delta}$ are mutually singular, and their TV distance is 1. □

G.2 Proof of Proposition 4.2

Proposition 4.2. *Consider a linear evolution equation on a bounded domain with linear boundary conditions and a deterministic source term. If the calibration dataset at time t_{cal} and a test sample u^t_{test} at time $t \geq t_{\mathrm{cal}}$ are generated from the same affine numerical scheme on independent initial conditions sampled from a Gaussian distribution, then we can construct a likelihood-weighted conformal predictor such that*

$$\mathbb{P}\big(\boldsymbol{u}^{t+\delta}_{\mathrm{test}} \in \mathcal{C}^{\mathrm{WCP}}_{t+\delta}(\boldsymbol{u}^t_{\mathrm{test}})\big) \geq 1-\alpha.$$

Proof. By assumption, the numerical time–δ map is affine: there exist a matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ and vector $\boldsymbol{b} \in \mathbb{R}^p$ such that

$$\boldsymbol{u}^{t+\delta} = \boldsymbol{A}\boldsymbol{u}^t + \boldsymbol{b}.$$

Iterating yields for $k \in \mathbb{N}$,

$$\boldsymbol{u}^{t_{\mathrm{cal}}+k\delta} = \boldsymbol{A}^k \boldsymbol{u}^{t_{\mathrm{cal}}} + \sum_{j=0}^{k-1} \boldsymbol{A}^j \boldsymbol{b},$$

and in particular each $\boldsymbol{u}^t$ is an affine function of $\boldsymbol{u}^0$.

Since $\boldsymbol{u}^0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and the multivariate normal family is closed under affine transformations, we have for every $t \geq 0$

$$\boldsymbol{u}^t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$$

for some mean $\boldsymbol{\mu}_t$ and covariance $\boldsymbol{\Sigma}_t$ (e.g., $\boldsymbol{\mu}_t = \boldsymbol{M}_t\boldsymbol{\mu}_0 + \boldsymbol{c}_t$ and $\boldsymbol{\Sigma}_t = \boldsymbol{M}_t\boldsymbol{\Sigma}_0\boldsymbol{M}_t^\top$ for the corresponding affine representation $\boldsymbol{u}^t = \boldsymbol{M}_t\boldsymbol{u}^0 + \boldsymbol{c}_t$). Hence the calibration and test laws admit Lebesgue densities $p_{t_{\mathrm{cal}}}$ and p_t on $\mathbb{R}^p$ and the likelihood ratio

$$r(\boldsymbol{x}) \;:=\; \frac{p_t(\boldsymbol{x})}{p_{t_{\mathrm{cal}}}(\boldsymbol{x})} \tag{7}$$

is available in closed form. In particular, weighting by Equation (7) corresponds to importance weights proportional to $p_t/p_{t_{\mathrm{cal}}}$.

Now apply likelihood-weighted split conformal prediction (weighted CP) with these weights to any fixed score function. The standard validity theorem for weighted split CP (under covariate shift with known likelihood ratio) implies that the resulting predictor $\mathcal{C}^{\mathrm{WCP}}_{t+\delta}$ satisfies

$$\mathbb{P}\big(\boldsymbol{u}^{t+\delta} \in \mathcal{C}^{\mathrm{WCP}}_{t+\delta}\big) \;\geq\; 1-\alpha$$

for each target time index $t \geq t_{\mathrm{cal}}$. This proves the claim. □

G.3 Proof of Proposition 4.6

Proposition 4.6 (Robust WCP under Gaussian distributions)**.** *Under the assumptions of Proposition 4.2, suppose we additionally observe independent estimation samples $\{\boldsymbol{u}_j^{t_{\mathrm{cal}}}\}_{j=1}^{m_0}$ and $\{\boldsymbol{u}_j^{t}\}_{j=1}^{m_1}$ with $m_k > p$ for $k \in \{0,1\}$, and fix an arbitrary estimation uncertainty $\eta > 0$. Then, we can construct a conformal prediction set $C^{\mathrm{rob}}_{t+\delta}$ with coverage guarantee*

$$\mathbb{P}\big(\boldsymbol{u}^{t+\delta}_{\mathrm{test}} \in C^{\mathrm{rob}}_{t+\delta}(\boldsymbol{u}^t_{\mathrm{test}})\big) \;\geq\; 1-\eta-\alpha.$$

Smaller η will usually correspond to wider bands.

Proof. As this proof is purely statistical, we will use random variables X_i instead of PDE-solutions u_i^t to ease the notation with more familiar symbols. In particular, we will use $Z_i^{(0)}$ and $Z_i^{(1)}$ for the estimation samples of the calibration and test distribution, respectively.
The proof extends WCP [32] to use worst case weights over a confidence region. Specifically, we know that we can construct a valid conformal set with the true parameters $\theta^\star = (\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ using WCP. Now we construct a $(1-\eta)$ confidence region Θ for the unknown true parameters $\theta^\star$ using the covariate-only samples. The robust construction then takes a worst-case weighted conformal

threshold (meaning the threshold that yields the largest conformal set over Θ) computed using the WCP weights

$$p_{i,\theta}(x) = \frac{w_\theta(X_i)}{\sum_{j=1}^n w_\theta(X_j) + w_\theta(x)},\ i = 1, \ldots, n, \qquad p_{n+1,\theta}(x) = \frac{w_\theta(x)}{\sum_{j=1}^n w_\theta(X_j) + w_\theta(x)},$$

where $w_\theta = dP_1^\theta / dP_0^\theta$ is the covariate likelihood ratio of normal distributions with parameters θ, X_i are calibration samples and x is the test sample. On the event $\{\theta^\star \in \Theta\}$, the robust set contains the oracle WCP set. The stated coverage then follows by a union bound.

Throughout, $\varphi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the p-variate Gaussian density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma} \succ 0$.

Getting the $(1-\eta)$ confidence region for $\theta^\star$ From the covariate-only samples $\{Z_j^{(k)}\}_{j=1}^{m_k}$ for $k \in \{0, 1\}$, define the sample mean and sample covariance

$$\widehat{\boldsymbol{\mu}}_k := \frac{1}{m_k} \sum_{j=1}^{m_k} Z_j^{(k)}, \qquad \widehat{\boldsymbol{\Sigma}}_k := \frac{1}{m_k - 1} \sum_{j=1}^{m_k} (Z_j^{(k)} - \widehat{\boldsymbol{\mu}}_k)(Z_j^{(k)} - \widehat{\boldsymbol{\mu}}_k)^\top.$$

Since $m_k > p$ and the data are Gaussian, $\widehat{\boldsymbol{\Sigma}}_k$ is positive definite almost surely [26, Thm. 3.1.4].

Fix numbers $\eta_1, \eta_2, \eta_3, \eta_4 > 0$ with $\eta_1 + \eta_2 + \eta_3 + \eta_4 \leq \eta$.

Mean confidence sets. By Härdle and Simar [17, Cor. 5.3], for each $k \in \{0, 1\}$, the Hotelling T^2 statistic

$$T_k^2 := m_k (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)$$

satisfies

$$\frac{m_k - p}{p(m_k - 1)} T_k^2 \sim F_{p, m_k - p}.$$

Let $c_k^{(\boldsymbol{\mu})}$ be the $(1 - \eta_{3+k})$ quantile of $\frac{p(m_k-1)}{m_k-p} F_{p,m_k-p}$ and define

$$\mathcal{M}_k := \left\{ \boldsymbol{\mu} \in \mathbb{R}^p : m_k (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu})^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}) \leq c_k^{(\boldsymbol{\mu})} \right\}.$$

Then $\mathcal{M}_k$ is compact and $\mathbb{P}(\boldsymbol{\mu}_k \in \mathcal{M}_k) \geq 1 - \eta_{3+k}$. Notice that $\mathcal{M}_k$ is an ellipsoid around the sample mean.

Covariance confidence sets. For the sample covariance matrix, we know $W_k = (m_k - 1)\widehat{\boldsymbol{\Sigma}}_k \sim \mathrm{Wishart}_p(\boldsymbol{\Sigma}_k, m_k - 1)$ [17, Thm. 5.7 (c)]. The eigenvalues of $\boldsymbol{\Sigma}_k^{-1/2} \widehat{\boldsymbol{\Sigma}}_k \boldsymbol{\Sigma}_k^{-1/2}$ have a distribution independent of $\boldsymbol{\Sigma}_k$ [17, Thm. 5.5]. Choose constants $0 < a_k < b_k < \infty$ such that

$$\mathbb{P}\Big(a_k I \preceq \boldsymbol{\Sigma}_k^{-1/2} \widehat{\boldsymbol{\Sigma}}_k \boldsymbol{\Sigma}_k^{-1/2} \preceq b_k I \Big) \geq 1 - \eta_{k+1}.$$

This is well-defined, as the eigenvalues of $\boldsymbol{\Sigma}_k^{-1/2} \widehat{\boldsymbol{\Sigma}}_k \boldsymbol{\Sigma}_k^{-1/2}$ are almost surely strictly positive and finite. Hence $L_k := \lambda_{\min}(\boldsymbol{\Sigma}_k^{-1/2} \widehat{\boldsymbol{\Sigma}}_k \boldsymbol{\Sigma}_k^{-1/2})$ and $U_k := \lambda_{\max}(\boldsymbol{\Sigma}_k^{-1/2} \widehat{\boldsymbol{\Sigma}}_k \boldsymbol{\Sigma}_k^{-1/2})$ are well-defined random variables with support on a subset of $(0, \infty)$.
Define

$$\mathcal{S}_k := \left\{ \boldsymbol{\Sigma} \in \mathbb{S}_{++}^p : \frac{1}{b_k} \widehat{\boldsymbol{\Sigma}}_k \preceq \boldsymbol{\Sigma} \preceq \frac{1}{a_k} \widehat{\boldsymbol{\Sigma}}_k \right\}.$$

Each set $\mathcal{S}_k$ is compact. Indeed, $\mathcal{S}_k$ is bounded since, for any $\boldsymbol{\Sigma} \in \mathcal{S}_k$, its spectral norm satisfies

$$\|\boldsymbol{\Sigma}\|_2 = \lambda_{\max}(\boldsymbol{\Sigma}),$$

which is bounded above by the largest eigenvalue allowed by the defining inequality of $\mathcal{S}_k$.

Closedness follows as well. Let $(\boldsymbol{\Sigma}_n)_{n \in \mathbb{N}} \subset \mathcal{S}_k$ be a convergent sequence with limit $\boldsymbol{\Sigma}_\infty$. By definition of $\mathcal{S}_k$, for every n and every vector $z \in \mathbb{R}^p$,

$$z^\top \left(\frac{1}{a_k} \widehat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_n \right) z \geq 0.$$

By continuity of the quadratic form, passing to the limit yields

$$z^\top \left(\frac{1}{a_k} \widehat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_\infty \right) z \geq 0 \quad \text{for all } z \in \mathbb{R}^p.$$

An analogous argument applies to the lower bound in the definition of $\mathcal{S}_k$. Hence $\mathbf{\Sigma}_\infty \in \mathcal{S}_k$, and $\mathcal{S}_k$ is closed.

Since $\mathcal{S}_k$ is both closed and bounded in the space of symmetric matrices, it is compact. Moreover, under the stated assumptions, the matrices in $\mathcal{S}_k$ are almost surely bounded away from singularity.

Define the joint confidence region

$$\Theta := \mathcal{M}_0 \times \mathcal{S}_0 \times \mathcal{M}_1 \times \mathcal{S}_1.$$

By a union bound,

$$\mathbb{P}(\theta^\star \in \Theta) \geq 1 - \eta. \tag{8}$$

Let $E := \{\theta^\star \in \Theta\}$.

Defining the weights and quantiles For $\theta = (\boldsymbol{\mu}_0, \mathbf{\Sigma}_0, \boldsymbol{\mu}_1, \mathbf{\Sigma}_1) \in \Theta$, define the Gaussian likelihood ratio

$$w_\theta(x) := \frac{\varphi(x; \boldsymbol{\mu}_1, \mathbf{\Sigma}_1)}{\varphi(x; \boldsymbol{\mu}_0, \mathbf{\Sigma}_0)}.$$

For a test covariate value $x \in \mathbb{R}^p$, define the associated WCP weights

$$p_{i,\theta}(x) := \frac{w_\theta(X_i)}{\sum_{j=1}^n w_\theta(X_j) + w_\theta(x)}, \quad i = 1, \ldots, n, \qquad p_{n+1,\theta}(x) := \frac{w_\theta(x)}{\sum_{j=1}^n w_\theta(X_j) + w_\theta(x)}.$$

Then $p_{i,\theta}(x) \geq 0$ and $\sum_{i=1}^{n+1} p_{i,\theta}(x) = 1$.

Let $\theta^\star$ denote the true parameters and $w^\star := w_{\theta^\star}$. The *oracle* WCP set is the WCP set obtained by using $w^\star$ in the weights above, and we denote it by $\widehat{C}_{\theta^\star}$ below.

Fix $x \in \mathbb{R}^p$ and a candidate label $y \in \mathbb{R}$. Let S_i, $i = 1, \ldots, n$, be the nonconformity scores for the sample (X_i, Y_i).

For each $\theta \in \Theta$ and test point (x, y) with score S_{n+1}, define the discrete probability measure on $[0, \infty]$,

$$\widehat{\nu}_\theta^{(x,y)} := \sum_{i=1}^n p_{i,\theta}(x)\, \delta_{S_i} + p_{n+1,\theta}(x)\, \delta_\infty, \tag{9}$$

and define the corresponding CDF on $\mathbb{R}$ by

$$F_\theta(t; x, y) := \widehat{\nu}_\theta^{(x,y)}\big((-\infty, t]\big) = \sum_{i=1}^n p_{i,\theta}(x)\, \mathbf{1}\{S_i \leq t\}, \qquad t \in \mathbb{R},$$

since the atom at ∞ contributes no mass to $(-\infty, t]$ for finite t.

Define the $(1 - \alpha)$ weighted quantile as

$$q_\theta(x, y) := \inf\Big\{t \in \mathbb{R} :\ F_\theta(t; x, y) \geq 1 - \alpha\Big\}. \tag{10}$$

Equivalently, $q_\theta(x, y)$ is the quantile of level $1 - \alpha$ under the discrete law $\widehat{\nu}_\theta^{(x,y)}$ in Equation (9).

Robust threshold and robust prediction set Define the robust threshold as the worst-case quantile over Θ:

$$q_{\mathrm{rob}}(x, y) := \sup_{\theta \in \Theta} q_\theta(x, y). \tag{11}$$

Define the robust WCP prediction set by

$$\widehat{C}_{\mathrm{rob}}(x) := \Big\{y \in \mathbb{R} :\ S_{n+1} \leq q_{\mathrm{rob}}(x, y)\Big\}. \tag{12}$$

Also, define for any fixed $\theta \in \Theta$, the WCP set

$$\widehat{C}_\theta(x) := \Big\{y \in \mathbb{R} :\ S_{n+1} \leq q_\theta(x, y)\Big\}.$$

In particular, $\widehat{C}_{\theta^\star}$ is the oracle WCP set.

On the event $E = \{\theta^\star \in \Theta\}$, we have for all (x, y),

$$q_{\mathrm{rob}}(x, y) = \sup_{\theta \in \Theta} q_\theta(x, y) \ \geq\ q_{\theta^\star}(x, y),$$

and therefore, for all x,

$$\widehat{C}_{\mathrm{rob}}(x) \supseteq \widehat{C}_{\theta^\star}(x) \qquad \text{on } E. \tag{13}$$

Robust coverage Let $F := \{Y_{n+1} \notin \widehat{C}_{\mathrm{rob}}(X_{n+1})\}$. By Equation (13),

$$F \cap E \subseteq \{Y_{n+1} \notin \widehat{C}_{\theta^\star}(X_{n+1})\} \cap E.$$

Hence,

$$\mathbb{P}(F) \leq \mathbb{P}\Big(Y_{n+1} \notin \widehat{C}_{\theta^\star}(X_{n+1})\Big) + \mathbb{P}(E^c).$$

By the oracle WCP guarantee (i.e., the covariate-shift WCP coverage result applied with the true likelihood ratio $w^\star = w_{\theta^\star}$),

$$\mathbb{P}\Big(Y_{n+1} \in \widehat{C}_{\theta^\star}(X_{n+1})\Big) \geq 1 - \alpha,$$

so $\mathbb{P}(Y_{n+1} \notin \widehat{C}_{\theta^\star}(X_{n+1})) \leq \alpha$. Also, by Equation (8), $\mathbb{P}(E^c) \leq \eta$. Therefore,

$$\mathbb{P}\Big(Y_{n+1} \in \widehat{C}_{\mathrm{rob}}(X_{n+1})\Big) \geq 1 - \eta - \alpha.$$

□

G.4 Proof of Proposition 4.7

Proposition 4.7. *Let $u(\cdot, t)$ be a PDE solution in some function space $(\mathcal{U}, \|\cdot\|_{\mathcal{U}})$ and $\boldsymbol{u}^t$ be its approximation via some numerical scheme on grid size h. Let C_h be a conformal prediction set for the discretized state with coverage $1-\alpha$ and assume that an error bound $\big\|u(\cdot,t) - \boldsymbol{I}_h \boldsymbol{u}^t\big\|_{\mathcal{U}} \leq \varepsilon_{h,t}(u_0)$ is available via some reconstruction operator $\boldsymbol{I}_h : \mathbb{R}^p \to \mathcal{U}$ (e.g., piecewise interpolation). Then we can define a conformal set for the function-space setting, $C \subset \mathcal{U}$, with $\mathbb{P}\big(u(\cdot,t) \in C\big) \geq 1-\alpha$ by*

$$C := \big\{v \in \mathcal{U} : \exists\, \boldsymbol{z} \in C_h \text{ with } \|v - \boldsymbol{I}_h \boldsymbol{z}\|_{\mathcal{U}} \leq \varepsilon_{h,t}(u_0)\big\}. \tag{3}$$

Proof. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be an underlying probability space supporting the random initial condition $u_0 : \Omega \to \mathcal{U}$ and hence the (random) PDE solution at time t,

$$u(\cdot, t) : \Omega \to \mathcal{U},$$

as well as the random numerical state

$$\boldsymbol{u}^t : \Omega \to \mathbb{R}^p.$$

Assume $\boldsymbol{I}_h : \mathbb{R}^p \to \mathcal{U}$ is measurable and that the deterministic error bound assumed in Proposition 4.7 holds for all $\omega \in \Omega$,

$$\big\|u(\cdot,t)(\omega) - \boldsymbol{I}_h \boldsymbol{u}^t(\omega)\big\|_{\mathcal{U}} \leq \varepsilon_{h,t}(u_0(\omega)). \tag{14}$$

Let $C_h \subset \mathbb{R}^p$ be a (measurable) prediction set such that

$$\mathbb{P}\big(\boldsymbol{u}^t \in C_h\big) \geq 1 - \alpha.$$

Define the lifted set $C \subset \mathcal{U}$ by Equation (3), i.e.

$$C := \Big\{u(\cdot,t) \in \mathcal{U} :\ \exists\, \boldsymbol{z} \in C_h \text{ s.t. } \|u(\cdot,t) - \boldsymbol{I}_h \boldsymbol{z}\|_{\mathcal{U}} \leq \varepsilon_{h,t}(u_0)\Big\},$$

where the right-hand side is interpreted pointwise in ω through the random radius $\varepsilon_{h,t}(u_0(\omega))$.

Consider the event $E := \{\omega :\ \boldsymbol{u}^t(\omega) \in C_h\} \in \mathcal{F}$. Fix $\omega \in E$. Then $\boldsymbol{z} := \boldsymbol{u}^t(\omega) \in C_h$. By the pathwise bound Equation (14),

$$\big\|u(\cdot,t)(\omega) - \boldsymbol{I}_h \boldsymbol{z}\big\|_{\mathcal{U}} = \big\|u(\cdot,t)(\omega) - \boldsymbol{I}_h \boldsymbol{u}^t(\omega)\big\|_{\mathcal{U}} \leq \varepsilon_{h,t}(u_0(\omega)).$$

Hence, by definition of C, we have $u(\cdot,t)(\omega) \in C$. Since $\omega \in E$ was arbitrary, this proves the event inclusion

$$\{\omega : \boldsymbol{u}^t(\omega) \in C_h\} \subseteq \{\omega : u(\cdot,t)(\omega) \in C\}.$$

Taking probabilities yields

$$\mathbb{P}(u(\cdot,t) \in C) \ \geq\ \mathbb{P}(\boldsymbol{u}^t \in C_h) \ \geq\ 1 - \alpha,$$

which is the desired marginal coverage statement. □

G.5 Proof of Theorem 4.5

Theorem 4.5. *Consider the drift PDE on a grid with p points, $\partial_t \boldsymbol{u}^t = c\,\boldsymbol{w}$ with $t \geq 0$, drift strength $c \in \mathbb{R}$, and drift profile $\boldsymbol{w} \in \mathbb{R}^p$ (e.g., $\boldsymbol{w} = \mathbf{1}$ for a constant drift), and let the initial conditions be non-degenerate Gaussian, $\boldsymbol{u}^0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. For $\boldsymbol{z}$ and $\boldsymbol{z}^i$ as in Proposition 4.4,*

$$d_{\mathrm{TV}}\big(\boldsymbol{z}, \boldsymbol{z}^i\big) \;=\; 2\Phi\Big(\frac{\sqrt{2\,p_{\mathrm{eff}}\,c^2}\,|t_1 - t_2|}{2}\Big) - 1, \quad \textit{with } p_{\mathrm{eff}} := \boldsymbol{w}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{w}.$$

Proof. We write $\boldsymbol{v} := c\,\boldsymbol{w}$ for the drift vector. In addition to the TV distance, we derive the weights for likelihood-ratio weighting, which we use in the proof of Proposition 4.4. The drift PDE $\partial_t \boldsymbol{u}^t = \boldsymbol{v}$ has the solution

$$\boldsymbol{u}^t = \boldsymbol{u}^0 + t\boldsymbol{v}.$$

Since $\boldsymbol{u}^0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, we have for every $t \geq 0$,

$$\boldsymbol{u}^t \sim \mathcal{N}(\boldsymbol{m}_t, \boldsymbol{\Sigma}_0), \qquad \boldsymbol{m}_t := \boldsymbol{\mu}_0 + t\boldsymbol{v},$$

i.e., the drift shifts the mean and leaves the covariance unchanged.

Reduction to TV distance on inputs For $m \in \{1, 2\}$, each pair $(\boldsymbol{u}_i^{t_m}, \boldsymbol{u}_i^{t_m+\delta})$ can be written as

$$(\boldsymbol{u}_i^{t_m}, \boldsymbol{u}_i^{t_m+\delta}) = \big(\boldsymbol{u}_i^{t_m}, \mathcal{G}_\delta(\boldsymbol{u}_i^{t_m})\big)$$

for some solution operator $\mathcal{G}_\delta$. Consequently, the joint distribution of $(\boldsymbol{u}_i^{t_m}, \boldsymbol{u}_i^{t_m+\delta})$ is supported on a p-dimensional submanifold of $\mathbb{R}^{2p}$. This is easiest to see in the case $p = 1$, where the joint distribution is concentrated on the graph of $\mathcal{G}_\delta(x)$ in the x–y plane.

The only difference between the joint distributions of $(\boldsymbol{u}_i^{t_1}, \boldsymbol{u}_i^{t_1+\delta})$ and $(\boldsymbol{u}_i^{t_2}, \boldsymbol{u}_i^{t_2+\delta})$ lies in how probability mass is distributed along this graph. In particular, any measurable set either has probability zero under both joint distributions (if it does not intersect the graph), or its probability is entirely determined by the corresponding subset of x-values.

Define

$$\boldsymbol{\xi}_m := \mathcal{L}(\boldsymbol{u}_i^{t_m}) \quad \text{and} \quad \boldsymbol{\nu}_m := \mathcal{L}\big(\boldsymbol{u}_i^{t_m}, \boldsymbol{u}_i^{t_m+\delta}\big), \qquad m \in \{1, 2\},$$

where $\mathcal{L}$ denotes the law of a random variable, and define the mapping

$$T(x) := \big(x, \mathcal{G}_\delta(x)\big).$$

Then $\boldsymbol{\nu}_m = T_{\#}\boldsymbol{\xi}_m$. Since T is injective, with inverse given by projection onto the first component restricted to $\mathrm{graph}(\mathcal{G}_\delta)$, it follows that

$$\|\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2\|_{\mathrm{TV}} = \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|_{\mathrm{TV}}.$$

From here on, we will use $\hat{\boldsymbol{z}} := (\boldsymbol{u}_1^{t_1}, \ldots, \boldsymbol{u}_n^{t_1}, \boldsymbol{u}_{n+1}^{t_2})$ and $\hat{\boldsymbol{z}}^i$ for $\hat{\boldsymbol{z}}$ when the ith and last entry are swapped.

Exact TV distance for equal-covariance Gaussians As each $\boldsymbol{u}_i^t$ is multivariate normal and independent of the others, the joint distributions of $\hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{z}}^i$ are again multivariate normal with the same block-diagonal covariance $\boldsymbol{I}_{n+1} \otimes \boldsymbol{\Sigma}_0 \in \mathbb{R}^{p(n+1)\times p(n+1)}$. The mean of $\hat{\boldsymbol{z}}$ is $\bar{\boldsymbol{\mu}} := (\boldsymbol{m}_{t_1}, \ldots, \boldsymbol{m}_{t_1}, \boldsymbol{m}_{t_2}) \in \mathbb{R}^{p(n+1)}$, consisting of n blocks $\boldsymbol{m}_{t_1}$ followed by one block $\boldsymbol{m}_{t_2}$. The mean $\bar{\boldsymbol{\mu}}^i$ of $\hat{\boldsymbol{z}}^i$ equals $\bar{\boldsymbol{\mu}}$ with the ith and last block swapped.

TV distance is invariant under bijective measurable maps, so we may apply the whitening map $\boldsymbol{x} \mapsto (\boldsymbol{I}_{n+1} \otimes \boldsymbol{\Sigma}_0^{-1/2})\boldsymbol{x}$ to both distributions, which turns the covariance into $\boldsymbol{I}_{p(n+1)}$. For identity covariances we use the following identity, which is used in various papers (e.g., He and Goldfeld [16], Karageorgiou and Prodromidis [18]) and attributed to Barsov and Ul'yanov [4] (as we cannot find access to the original source we use the citation directly from other papers that used the identity).

$$d_{\mathrm{TV}}(\boldsymbol{z}, \boldsymbol{z}^i) = d_{\mathrm{TV}}(\hat{\boldsymbol{z}}, \hat{\boldsymbol{z}}^i) = d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{I}_{p(n+1)}), \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{I}_{p(n+1)})) = 2\Phi(\frac{\big\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\big\|_2}{2}) - 1$$

Together with the whitening, it gives

$$d_{\mathrm{TV}}(\boldsymbol{z}, \boldsymbol{z}^i) = d_{\mathrm{TV}}(\hat{\boldsymbol{z}}, \hat{\boldsymbol{z}}^i) = 2\Phi\Big(\frac{\big\|(\boldsymbol{I}_{n+1} \otimes \boldsymbol{\Sigma}_0^{-1/2})(\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}^i)\big\|_2}{2}\Big) - 1.$$

The vector $\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}^i$ is zero except for the ith block, $\boldsymbol{m}_{t_1} - \boldsymbol{m}_{t_2} = (t_1 - t_2)\boldsymbol{v}$, and the last block, $(t_2 - t_1)\boldsymbol{v}$. Hence

$$\big\|(\boldsymbol{I}_{n+1} \otimes \boldsymbol{\Sigma}_0^{-1/2})(\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}^i)\big\|_2^2 = 2(t_1 - t_2)^2\, \boldsymbol{v}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{v}$$

and

$$d_{\mathrm{TV}}\big(\boldsymbol{z}, \boldsymbol{z}^i\big) = 2\Phi\Big(\frac{\sqrt{2\, \boldsymbol{v}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{v}}\, |t_1 - t_2|}{2}\Big) - 1.$$

Moreover, since $1 - \Phi(z) \le e^{-z^2/2}$, we get

$$1 - d_{\mathrm{TV}}(\boldsymbol{z}, \boldsymbol{z}^i) = 2\Big(1 - \Phi\big(\tfrac{\sqrt{2\, \boldsymbol{v}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{v}}\, |t_1 - t_2|}{2}\big)\Big) \le 2\exp\Big(-\frac{(t_1 - t_2)^2\, \boldsymbol{v}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{v}}{4}\Big),$$

so $d_{\mathrm{TV}}(\boldsymbol{z}, \boldsymbol{z}^i) \to 1$ exponentially fast in $(t_1 - t_2)^2\, \boldsymbol{v}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{v}$. With $\boldsymbol{v} = c\, \boldsymbol{w}$, we have $\boldsymbol{v}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{v} = c^2 p_{\mathrm{eff}}$, which gives Theorem 4.5.

Likelihood ratio at the centers Let p_t denote the density of $\mathcal{N}(\boldsymbol{m}_t, \boldsymbol{\Sigma}_0)$ and $\Delta := \boldsymbol{m}_{t_2} - \boldsymbol{m}_{t_1} = (t_2 - t_1)\boldsymbol{v}$. The Gaussian likelihood ratio is

$$r(\boldsymbol{x}) := \frac{p_{t_2}(\boldsymbol{x})}{p_{t_1}(\boldsymbol{x})} = \exp\Big(\Delta^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{x} - \frac{1}{2}\big(\boldsymbol{m}_{t_2}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{m}_{t_2} - \boldsymbol{m}_{t_1}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{m}_{t_1}\big)\Big).$$

Evaluating at $\boldsymbol{x}_{\mathrm{test}} := \boldsymbol{m}_{t_2}$ and $\boldsymbol{x}_{\mathrm{cal}} := \boldsymbol{m}_{t_1}$ gives

$$\log r(\boldsymbol{m}_{t_2}) = \tfrac{1}{2}\Delta^\top \boldsymbol{\Sigma}_0^{-1} \Delta, \qquad \log r(\boldsymbol{m}_{t_1}) = -\tfrac{1}{2}\Delta^\top \boldsymbol{\Sigma}_0^{-1} \Delta.$$

Since $\Delta^\top \boldsymbol{\Sigma}_0^{-1} \Delta = (t_1 - t_2)^2\, \boldsymbol{v}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{v} =: \kappa$, we obtain

$$r(\boldsymbol{x}_{\mathrm{test}}) = e^{\kappa/2}, \qquad r(\boldsymbol{x}_{\mathrm{cal}}) = e^{-\kappa/2}.$$

Normalized weights If the calibration set contains n points with ratio $r(\boldsymbol{x}_{\mathrm{cal}})$ and we normalize over these n points plus the test point, then

$$w_{\mathrm{test}} = \frac{r(\boldsymbol{x}_{\mathrm{test}})}{n r(\boldsymbol{x}_{\mathrm{cal}}) + r(\boldsymbol{x}_{\mathrm{test}})} = \frac{e^{\kappa}}{n + e^{\kappa}} = 1 - O\big(n e^{-\kappa}\big),$$

and for any such calibration point,

$$w_{\mathrm{cal}} = \frac{r(\boldsymbol{x}_{\mathrm{cal}})}{n r(\boldsymbol{x}_{\mathrm{cal}}) + r(\boldsymbol{x}_{\mathrm{test}})} = \frac{1}{n + e^{\kappa}} = O\big(e^{-\kappa}\big).$$

□

G.6 Proof of Proposition 4.4

Proposition 4.4. *Consider the drift PDE $\partial_t u(x,t) = c$, with $x \in (0,1), t \ge 0, c \in \mathbb{R}$, and compatible boundary conditions. Define $\boldsymbol{z}$ and $\boldsymbol{z}^i$ as in Equation (2) with n i.i.d. calibration pairs at times $(t_1, t_1 + \delta)$ and a test pair at time $(t_2, t_2 + \delta)$ from an affine scheme on a grid with p points and Gaussian initial conditions. Then, the total variation distance between the laws of $\boldsymbol{z}$ and $\boldsymbol{z}^i$ is*

$$d_{\mathrm{TV}}\big(\boldsymbol{z}, \boldsymbol{z}^i\big) \;=\; 2\Phi\Big(\frac{\sqrt{2pc^2}\, |t_1 - t_2|}{2}\Big) - 1,$$

where Φ is the cumulative distribution function of the standard normal distribution. Hence, $d_{\mathrm{TV}}(\boldsymbol{z}, \boldsymbol{z}^i) \to 1$ exponentially fast in the grid size p and with it, the coverage guarantee from Barber et al. [3] vanishes exponentially.

Moreover, for likelihood-ratio weighting as in Equation (7) evaluated at the distribution centers $\boldsymbol{x}_{\mathrm{test}} := \mathbb{E}[\boldsymbol{u}_{n+1}^{t_2}]$ and $\boldsymbol{x}_{\mathrm{cal}} := \mathbb{E}[\boldsymbol{u}_i^{t_1}]$, the test weight satisfies

$$w_{\mathrm{test}} = 1 - O\big(n e^{-pc^2(t_1 - t_2)^2}\big), \quad w_{\mathrm{cal}} = O\big(e^{-pc^2(t_1 - t_2)^2}\big).$$

Hence, the weighted conformal bands become infinite at an exponential rate.

Proof. Proposition 4.4 is the special case $\boldsymbol{w} = \mathbf{1}$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{I}_p$ of the proof in Appendix G.5. Then $p_{\mathrm{eff}} = p$ and $\kappa = pc^2(t_1 - t_2)^2$, which gives the TV distance, its exponential convergence in p, and the weights. □

G.7 General TV distance bound

Proposition G.1 (Extension of Proposition 4.4)**.** *Consider a linear partial differential equation discretized on a spatial grid with p points. Let $\mathbf{u}^{t_1}, \mathbf{u}^{t_2} \in \mathbb{R}^p$ be the Gaussian numerical solutions $\mathcal{N}(\boldsymbol{\mu}_{t_1}, \boldsymbol{\Sigma}_{t_1})$ and $\mathcal{N}(\boldsymbol{\mu}_{t_2}, \boldsymbol{\Sigma}_{t_2})$ at two distinct time steps.*

Suppose there exist uniform spectral bounds $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$ such that for $t \in \{t_1, t_2\}$, the eigenvalues of the covariance matrices satisfy $\lambda(\boldsymbol{\Sigma}_t) \in [\underline{\lambda}, \overline{\lambda}]$. Then the total variation distance between the laws of the numerical solutions satisfies

$$d_{\mathrm{TV}}(\mathbf{u}^{t_1}, \mathbf{u}^{t_2}) \leq \frac{1}{2}\sqrt{p\left(\frac{\overline{\lambda}}{\underline{\lambda}} - 1 + \log\frac{\overline{\lambda}}{\underline{\lambda}}\right) + \frac{\|\boldsymbol{\mu}_{t_1} - \boldsymbol{\mu}_{t_2}\|^2}{\underline{\lambda}}}.$$

For a consistent discretization of a smooth Gaussian random field, $\underline{\lambda} \to 0$ as p grows, so the uniform spectral bounds only hold for a fixed range of grid sizes.

Proof. The result follows by bounding the Kullback–Leibler (KL) divergence and applying Pinsker's inequality. Let $P = \mathcal{N}(\boldsymbol{\mu}_{t_1}, \boldsymbol{\Sigma}_{t_1})$ and $Q = \mathcal{N}(\boldsymbol{\mu}_{t_2}, \boldsymbol{\Sigma}_{t_2})$. By Pinsker's inequality,

$$d_{\mathrm{TV}}(P, Q) \leq \sqrt{\frac{1}{2}\,\mathrm{KL}(P\|Q)}.$$

The KL divergence for p-dimensional Gaussian measures is well known as

$$\mathrm{KL}(P\|Q) = \frac{1}{2}\left[\mathrm{tr}(\boldsymbol{\Sigma}_{t_2}^{-1}\boldsymbol{\Sigma}_{t_1}) - p + \log\frac{\det\boldsymbol{\Sigma}_{t_2}}{\det\boldsymbol{\Sigma}_{t_1}} + (\boldsymbol{\mu}_{t_1} - \boldsymbol{\mu}_{t_2})^\top\boldsymbol{\Sigma}_{t_2}^{-1}(\boldsymbol{\mu}_{t_1} - \boldsymbol{\mu}_{t_2})\right].$$

First, we bound the terms related to the covariance structure. Under the assumption that the eigenvalues of $\boldsymbol{\Sigma}_{t_1}$ and $\boldsymbol{\Sigma}_{t_2}$ are bounded by $\underline{\lambda}$ and $\overline{\lambda}$, the trace of the product satisfies

$$\mathrm{tr}(\boldsymbol{\Sigma}_{t_2}^{-1}\boldsymbol{\Sigma}_{t_1}) \leq p \cdot \lambda_{\max}(\boldsymbol{\Sigma}_{t_2}^{-1})\lambda_{\max}(\boldsymbol{\Sigma}_{t_1}) \leq p\frac{\overline{\lambda}}{\underline{\lambda}},$$

where the first inequality follows from von Neumann's trace inequality as the matrices are symmetric and positive definite.
Similarly, the log-determinant ratio is bounded by the sum of the logs of the eigenvalue ratios:

$$\log\frac{\det\boldsymbol{\Sigma}_{t_2}}{\det\boldsymbol{\Sigma}_{t_1}} = \sum_{i=1}^{p}\log\frac{\lambda_i(\boldsymbol{\Sigma}_{t_2})}{\lambda_i(\boldsymbol{\Sigma}_{t_1})} \leq p\log\frac{\overline{\lambda}}{\underline{\lambda}}.$$

Next, the quadratic form involving the mean shift is bounded using the spectral norm of the inverse covariance matrix:

$$(\boldsymbol{\mu}_{t_1} - \boldsymbol{\mu}_{t_2})^\top\boldsymbol{\Sigma}_{t_2}^{-1}(\boldsymbol{\mu}_{t_1} - \boldsymbol{\mu}_{t_2}) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{t_2}^{-1})\|\boldsymbol{\mu}_{t_1} - \boldsymbol{\mu}_{t_2}\|^2 \leq \frac{\|\boldsymbol{\mu}_{t_1} - \boldsymbol{\mu}_{t_2}\|^2}{\underline{\lambda}}.$$

Substituting these bounds into the KL divergence expression, we obtain

$$\mathrm{KL}(P\|Q) \leq \frac{1}{2}\left[p\left(\frac{\overline{\lambda}}{\underline{\lambda}} - 1 + \log\frac{\overline{\lambda}}{\underline{\lambda}}\right) + \frac{\|\boldsymbol{\mu}_{t_1} - \boldsymbol{\mu}_{t_2}\|^2}{\underline{\lambda}}\right].$$

Applying the square root from Pinsker's inequality yields the stated bound. □

H Experimental details

We tracked our results with Weights & Biases [5].

H.1 Methods

Supr. The Supr. method follows the suggestion by Diquigiovanni et al. [7]. We calculate the scores at calibration time as the maximum over the spatial dimension x. Then, at test time the respective quantile of these scores is added to each spatial location of the test prediction to generate the conformal interval. No further fine-tuning is necessary.

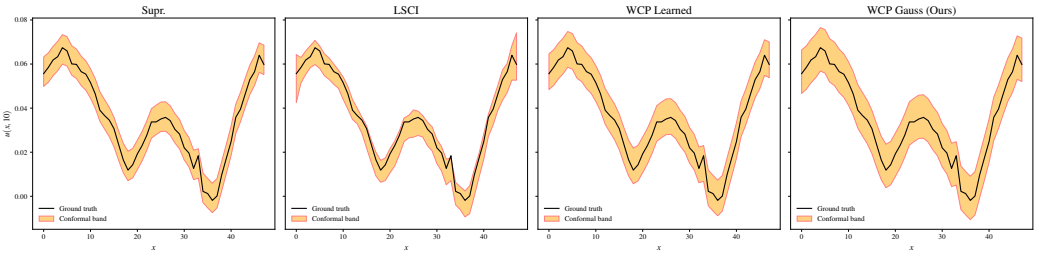


Figure H.1: Exemplary coverage plots for the methods (from left to right) Supr., LSCI, WCP Learned, and WCP Gauss (Ours) for a sample at time step 10 with parameters $a = -0.0075, b = 0, c = 0.5$.

LSCI The LSCI method is difficult to fine-tune. In particular, we have four different parameters: λ, the embedding dimension, the metric, and the number of samples to use for the conformal bands. As the computational burden is high for this method, we opted for a mid-range setup, and chose (following the original work) $\lambda = 3$, embedding dimension 20, the supremum norm, and varied the number of samples between 500 and 1000, depending on what gave the closest to the target coverage for the first time step. This fine-tuning would not be feasible in practice, but as it favors the baseline we compare against, it does not harm the validity of our method.

DtACI To showcase adaptive methods, we use Dynamically-Tuned Adaptive Conformal Inference (DtACI) by Gibbs and Candès [10]. They dynamically adjust their α value according to previous miscoverage to satisfy the original $1 - \alpha$ guarantee. In comparison to previous methods [e.g., 9], this method does not need manual tuning of the correction parameter, and the paper suggests calculation for some default parameters which we followed. For the rest of the free parameters, we followed the examples from their paper and chose 8 agents with γ one of $[0.001, 0.002, 0.004, 0.008, 0.016, 0.032, 0.064, 0.128]$ and a window size of 5 (due to the short time frame we observe).

WCP Learned The WCP Learned method is a WCP method according to the original proposal of Tibshirani et al. [32]. There, they suggest estimating the (unknown) weights by the probabilistic output of a classifier. We experimented with multiple versions of LogisticRegression, Random Forests, and Support Vector Machines. We found gradient boosting to yield the best results in our experiments (n_estimators=50, max_depth=3, learning_rate=0.1). There might be better models, e.g., certain deep learning models could model the underlying distributions better. However, the quality of a model depends on the data, i.e., a working model on one dataset might undercover on another. Also, as the models need to be trained at every prediction step, large deep learning models are practically infeasible.

WCP Gauss Our proposed method is also based on Tibshirani et al. [32]. However, instead of estimating the weights, we use the knowledge of the underlying distribution to calculate them precisely. The mean and covariance are always estimated from the data using the average for the mean and Ledoit–Wolf [19] for the covariance.

H.2 Data generation

We generated multiple synthetic datasets from a general second-order formulation for periodic PDEs, each discretized with finite-difference schemes with exact exponential matrix evolution. Initial conditions were sampled from Gaussian processes with covariance $625(-\partial_x^2 + 25I)^{-2}$ to provide smooth but nontrivial trajectories. Each dataset is stored as a compressed `.npz` file containing $(n_{\text{samples}}, n_t + 1, n_x)$ trajectories together with grid and metadata. The PDE looks as follows

$$u_t(x,t) = au(x,t)_{xx} + bu(x,t)_x + cu(x,t), \quad x \in (0, 2\pi),\ t \geq 0,$$
$$u(x,0) = u_0(x), \quad x \in (0, 2\pi),$$

with periodic boundary conditions, and where $a, b, c \in \mathbb{R}$ are tunable parameters. We sampled data from all combinations of $a \in \{-0.005, -0.0075, -0.01\}$, $b \in \{-0.5, 0, 0.5\}$, and $c \in \{-0.5, 0, 0.5\}$. Note that we focused on negative values for a, as positive values tend to

smooth the trajectory over time which leads to trivial coverage even for naïve CP. We sampled 21 time steps (incl. initial condition) between 0 and 1 (for the recalibration task, we extended the time to 2, as otherwise all methods cover too easily) and sampled 5000 trajectories for training, 500 for validation, 5000 for calibration, and 1000 for testing, and use a spatial resolution of $n_x = 48$. As surrogate, we train a 1D Fourier Neural Operator [21] (3 layers, 12 Fourier modes, 32 hidden channels) to predict one time step ahead, using AdamW (learning rate 10^{-3}, cosine schedule) on the MSE loss for 5 epochs.

H.3 Ablation study

The critical part of the WCP Gauss method is the accurate estimation of the covariance. If the dimensionality increases or the number of samples for estimation drops, we expect instability in the covariance estimation and, therefore, instability in the coverage results. In Figure H.2 we see an example of the synthetic experiment when we increase grid granularity (i.e., the dimensionality of the problem) or reduce the number of calibration samples. The same would hold for reducing the number of test samples, but we omit this from further plots so as not to overload the appendix. We can see that a higher granularity results in an earlier onset of infinite bands (as suggested in Proposition 4.4) and that the method becomes unstable when the number of estimation samples is too low for the dimensionality of the problem.

H.4 Visualizations of theoretical results

Proposition 4.7 We use the backward heat equation to visualize this proposition. As we do not know the exact approximation error, we cannot construct conformal sets as described in the proposition. It is, however, reasonable to assume that the miscoverage should decrease as the approximation error drops (because what we would theoretically add to the numerical conformal intervals will get smaller with finer granularity). Therefore, we sampled numerical trajectories for the backward heat equation (with parameter $\nu = -0.001$ and the granularities seen in the figure) as described in Appendix H.2. For each granularity, we used the sampled initial conditions to generate the analytic solution from its Fourier representation using 16 modes for sine and cosine each. We then proceeded as in the synthetic case (general task), calibrated with the numerical data, predicted from numerical input, and checked coverage with the analytical output. The depicted coverages come from the first time step after the calibration (at later time steps, the negative effect from Proposition 4.4 reduced the observed effect, and the methods started undercovering or getting infinite bandwidth).

Proposition 4.4 To show that a higher granularity implies a higher TV distance and faster infinite bandwidth, we sampled from the PDE in the proposition (with drift parameter 1 and the granularities seen in the figure) as described in Appendix H.2. For each granularity, we calculated the coverage as in the synthetic case (general task). The depicted time step is 10, as this was the time step where the effect for our chosen PDE was best visible (earlier time steps would need higher granularity to drop in coverage, and later ones dropped too quickly). The theoretical lower bound was calculated as described in Barber et al. [3] as $1 - \alpha - \sum_{i=1}^{n} \tilde{w}_i d_{\mathrm{TV}}(\boldsymbol{z}, \boldsymbol{z}^i)$ with equal weights w_i. Here, $\boldsymbol{z}$ and $\boldsymbol{z}^i$ are described in Proposition 4.4.

Correlated initial conditions Figure H.3 evaluates the TV distance of Theorem 4.5 for the constant drift $\boldsymbol{v} = \mathbf{1}$ under correlated initial conditions. If the initial field and the drift are smooth, neighboring grid points are strongly correlated and $\boldsymbol{v}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{v}$ stays bounded as the grid is refined, so the TV distance levels off below 1. A drift that is rougher than the initial field, such as a step, still drives the TV distance to 1.

I Real-world examples

I.1 Pulsed-thermography data

The dataset of Wei et al. [34] contains 19 samples of 2D surfaces measured over 1810 time steps. We allocate 3 samples for training a Fourier Neural Operator [21], 1 for validation, 13 for calibration and testing. 2 samples were excluded due to large outlier noise. We crop the borders of the measurement data, since they correspond to non-heated background regions (see Figure I.1). We then determine

the time frame where cooling begins, by finding the frame with the largest temperature and use the 20 following time frames as the dataset. As we consider an inverse problem, we flip the time axis. For the recalibration task, we additionally use a time stride (i.e., we remove every second frame from the data), because otherwise all methods always covered (as they provide temporally very fine-grained data). Because the dataset is small, we divide each surface into square patches of size 10×10 and treat them as identically distributed samples (to mitigate dependency between the patches, we leave an empty space of size 5 to all neighboring patches). While this induces some dependency between patches, and the variations in the metal plates (e.g., differently drilled backside holes) introduce slight distribution shifts, both effects are minor for a real-world measurement dataset and have a negligible impact on CP performance. We further do not split the trajectories into different calibration and test sets. We calibrate at an earlier time than we test, but for the same trajectories. This also introduces dependency structures, but again mitigates the low dataset size and mirrors the real-world setting, where the same processes are observed over time. We can see in Figure I.2 that WCP Gauss is the only method with consistent coverage over time.

I.2 Lake Maggiore thermistor data

The Lake Maggiore thermistor dataset of Scheu et al. [31][9] contains 1-minute-resolution thermistor chain measurements at multiple depths in Pallanza Bay. We use 50 nights in early spring for calibration and 14 nights in late spring for testing. A 1D Fourier Neural Operator [21] (3 layers, 3 Fourier modes, 32 hidden channels) is trained on 15 further early-spring nights (and validated on 5) to predict one time step ahead, for up to 200 epochs with early stopping. We subsample to 30-minute resolution across 7 depth sensors and extract nighttime windows as independent trajectories of vertical heat diffusion without solar forcing, which fits the linear-PDE setting of Proposition 4.2. We z-scored with the global mean and standard deviation of the training nights. As for the pulsed-thermography data, we frame the prediction as an inverse problem and reverse the time axis. Following the discussion after Theorem 4.5, we restrict CP to the middle depth sensors (27, 30, and 35 m above bed) rather than the full water column, exploiting the more favorable marginal distribution. Figure I.3 shows the resulting coverages for both tasks: WCP Gauss again maintains the target coverage almost everywhere (note that with a test size of 14, single-step deviations from 90% are statistically expected).

I.3 FRAP data

The FRAP dataset of Wåhlstrand Skärström et al. [35][10] contains fluorescence-recovery-after-photobleaching experiments on 32 wt% sucrose solutions, whose post-bleach intensity field obeys the standard diffusion equation and again fits Proposition 4.2. We train a 2D Fourier Neural Operator [21] on 400 frame pairs from 20 simulated recovery fields of size 256×256 pixels, and use 20 real experiments for calibration and test (we use different time steps for calibration and test, not different trajectories). The training fields are produced by a spectral diffusion simulator with randomly sampled diffusion coefficient, bleach depth, additive-noise level, and bleach radius according to the original paper. Real recoveries are computed as $\Delta F/F_0$ relative to the per-pixel pre-bleach mean, percentile-and-MAD clipped to suppress imaging outliers, Gaussian-smoothed with σ=0.6 px, and per-trajectory z-scored before training. The 5×5 box filter is applied only at CP score extraction and is not seen by the FNO. From every 256×256 frame we extract a 3×3 grid of pixels around the bleach center, spaced 5 pixels apart and each smoothed with a 5×5 box filter, yielding a 9-dimensional time series. The time axis is then truncated, time-strided, and reversed to frame the problem as an inverse one. Following the discussion after Theorem 4.5, restricting CP to the center of the bleach region (where the laser is focused and the recovery dynamics are concentrated) rather than to the full 256×256 field exploits the more favorable marginal distribution. Figure I.4 shows the resulting coverages together with sample recovery snapshots. Only WCP Gauss attains non-trivial target coverage at every time step on both tasks (with the same test-size caveat as above).

[9] https://doi.org/10.5281/zenodo.1206188
[10] https://doi.org/10.5281/zenodo.3874218

J Robustness experiments

J.1 Multi-seed stability

Figure J.1 repeats the general task of the synthetic experiment (Figure K.5, $a = -0.0075$) over 10 reshuffled calibration/test splits and the pulsed-thermography experiment over 10 plate-level reshuffles of the train/validation/calibration/test assignment, each with a retrained surrogate. Since each seed uses a different plate assignment and surrogate, the pulsed-thermography curves differ from those in Figure I.2. The standard deviation across seeds is small and the ranking of the methods is the same as in the single-seed figures. LSCI is omitted due to its computational cost.

J.2 Nonlinear PDEs

We add a cubic reaction term to the synthetic PDE, $u_t = au_{xx} + cu - \gamma u^3$ with $a = -0.0075$ and $b = 0$, where $\gamma = 0$ recovers the linear PDE. For $\gamma > 0$, the solutions are no longer Gaussian and Proposition 4.2 does not apply. Figure J.2 shows that coverage then depends on the PDE parameters: for $c = 0.5$, WCP Gauss attains the target for all γ, while for $c = 3$ the solutions saturate into a bimodal distribution and at $\gamma = 10$ the coverage of WCP Gauss drops to about 0.5. Coverage on nonlinear PDEs is therefore not guaranteed and can be lost by changing the parameters.

J.3 Bandwidths

The bandwidth of WCP Gauss is determined by the likelihood ratio of the test sample. A test sample that is likely under the test distribution but unlikely under the calibration distribution has a large weight $\tilde{w}_{n+1}$, which moves the weighted quantile in Equation (6) upwards. Once $\tilde{w}_{n+1} > \alpha$, the band is infinite. Figure J.3 shows this for the synthetic PDE: the width increases monotonically with the likelihood ratio, and infinite bands occur exactly for the samples with $\tilde{w}_{n+1} > \alpha$.

J.4 Other initial distributions

We further show results when starting sampling from other initial distributions. In Figure K.7, we can see coverages when initially sampling from a Laplace distribution and in Figure K.8 when sampling from a Logistic distribution. As the results are almost equivalent to those with Gaussian initial conditions, we only show the graphs for $b = 0$ and refrain from further plots or tables. In Figure K.9, we show coverages of WCP Gauss (with Cauchy instead of Gaussian likelihood-ratio weights) when sampling from heavy-tailed Cauchy initial conditions over multiple seeds. The method stays stable over the seed distribution.

K Additional results on synthetic data

We present the additional results (plots and tables) for the synthetic experiments. For the recalibration task, we find them in Figure K.1 and Table K.1 for $b = -0.5$, in Figure K.2 and Table K.2 for $b = 0$, and in Figure K.3 and Table K.3 for $b = 0.5$. For the general task, we find them in Figure K.4 and Table K.4 for $b = -0.5$, in Figure K.5 and Table K.5 for $b = 0$, and in Figure K.6 and Table K.6 for $b = 0.5$.

All results show similar structures. If the parameter set yields a simple dynamic, all methods cover even far beyond the calibration points. But when the dynamics get rougher, all but the WCP Gauss method undercover. WCP Gauss refrains from false predictions and instead yields infinite bands. DtACI yields worse results than is usually seen in the literature. The reason is that we observe only a short prediction horizon. When using adaptive CP on stock market data for example, thousands of time steps are available and changes happen slowly.

L Computing resources

All the experiments were conducted on a mix of CPUs and GPUs (A40, A100, H100), based on the availability of our academic cluster. Most runs are comfortably executed on CPUs, but for the image data, GPUs speed up the process. No run takes longer than a few minutes. For the LSCI method,

Table K.1: Coverage and mean bandwidth for the recalibration task for varying a, c and fixed $b = -0.5$. We also report the fraction of samples where infinite bands were reported n_∞. The mean bandwidth is taken over finite bands only. The gray font is chosen for better readability.

			$c = -0.5$					$c = 0.0$					$c = 0.5$				
		T:	1	5	10	15	20	1	5	10	15	20	1	5	10	15	20
$a = -0.005$	Supr.	Cov.	.91	.92	.91	.90	.90	.91	.91	.90	.91	.89	.91	.89	.88	.89	.88
		Bw.	.0125	.0109	.0089	.0073	.0065	.0133	.0134	.0135	.0145	.0161	.0142	.0171	.0243	.0419	.0793
	LSCI	Cov.	.96	.96	.92	.91	.90	.96	.95	.91	.94	.86	.97	.96	.97	.97	.96
		Bw.	.0121	.0069	.0052	.0052	.0056	.0126	.0108	.0101	.0148	.0152	.0133	.0167	.0238	.0508	.0988
	DtACI	Cov.	.91	.97	.96	.96	.94	.91	.93	.93	.93	.86	.91	.88	.89	.91	.91
		Bw.	.0125	.0117	.0094	.0080	.0068	.0133	.0129	.0131	.0138	.0150	.0142	.0145	.0212	.0344	.0650
	WCP Learned	Cov.	.90	.89	.88	.89	.89	.90	.89	.89	.89	.87	.90	.88	.88	.89	.87
		Bw.	.0117	.0100	.0083	.0070	.0063	.0124	.0125	.0128	.0138	.0158	.0133	.0165	.0238	.0410	.0768
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.91	.91	.91	.93	.98	.91	.91	.90	.93	.97	.91	.92	.91	.89	.86
		Bw.	.0127	.0106	.0087	.0080	.0086	.0134	.0138	.0135	.0158	.0218	.0143	.0192	.0283	.0430	.0708
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	2%
$a = -0.0075$	Supr.	Cov.	.91	.92	.91	.87	.81	.91	.91	.89	.84	.75	.91	.89	.88	.87	.83
		Bw.	.0126	.0110	.0093	.0089	.0122	.0133	.0136	.0142	.0179	.0316	.0142	.0172	.0251	.0461	.1042
	LSCI	Cov.	.95	.95	.89	.83	.84	.96	.93	.86	.81	.77	.98	.94	.91	.89	.85
		Bw.	.0122	.0070	.0068	.0086	.0129	.0127	.0110	.0132	.0182	.0333	.0134	.0142	.0256	.0584	.1263
	DtACI	Cov.	.91	.97	.96	.88	.64	.91	.93	.91	.73	.47	.91	.88	.88	.78	.69
		Bw.	.0126	.0118	.0097	.0090	.0111	.0133	.0130	.0137	.0165	.0279	.0142	.0146	.0220	.0394	.0914
	WCP Learned	Cov.	.90	.89	.88	.85	.81	.90	.89	.88	.81	.74	.90	.88	.87	.85	.82
		Bw.	.0118	.0102	.0088	.0087	.0122	.0124	.0128	.0138	.0175	.0314	.0133	.0168	.0246	.0442	.1017
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.92	.91	.90	.98	.96	.92	.91	.89	.96	.97	.91	.91	.90	.88	.89
		Bw.	.0127	.0107	.0091	.0112	.0146	.0136	.0140	.0141	.0223	.0395	.0144	.0192	.0276	.0453	.1134
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	1%	2%	0%	0%	2%	5%	6%
$a = -0.01$	Supr.	Cov.	.91	.92	.87	.72	.69	.91	.91	.84	.65	.60	.91	.89	.84	.68	.59
		Bw.	.0126	.0112	.0102	.0154	.0320	.0133	.0139	.0160	.0312	.0829	.0143	.0174	.0274	.0696	.2281
	LSCI	Cov.	.95	.92	.78	.75	.78	.96	.89	.74	.68	.68	.97	.89	.80	.74	.66
		Bw.	.0123	.0075	.0090	.0161	.0349	.0127	.0094	.0151	.0328	.0897	.0134	.0140	.0304	.0810	.2472
	DtACI	Cov.	.91	.96	.90	.47	.43	.91	.93	.75	.34	.33	.91	.87	.70	.38	.32
		Bw.	.0126	.0118	.0102	.0137	.0282	.0133	.0130	.0149	.0275	.0731	.0143	.0147	.0242	.0617	.2040
	WCP Learned	Cov.	.90	.89	.84	.72	.70	.90	.89	.81	.64	.61	.90	.88	.83	.66	.59
		Bw.	.0118	.0103	.0099	.0154	.0320	.0124	.0129	.0155	.0310	.0829	.0134	.0171	.0269	.0685	.2275
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.92	.91	.97	.93	.94	.92	.91	.93	.95	.93	.92	.92	.91	.93	.93
		Bw.	.0128	.0111	.0125	.0179	.0394	.0137	.0144	.0190	.0390	.1034	.0146	.0196	.0318	.0876	.2969
		n_∞	0%	0%	0%	2%	2%	0%	0%	2%	5%	5%	0%	0%	7%	12%	12%

it is highly advised to use the computation cache that JAX provides. Otherwise, these calculations can take far longer, even on GPUs. Memory and storage do not exceed 10 GB per dataset. The full research project required more compute than the experiments reported in the paper, due to failed runs, parameter tuning, or experiments that were not relevant enough for the final cut.

Table K.2: Coverage and mean bandwidth for the recalibration task for varying a, c and fixed $b = 0.0$. We also report the fraction of samples where infinite bands were reported n_∞. The mean bandwidth is taken over finite bands only. The gray font is chosen for better readability.

			$c = -0.5$					$c = 0.0$					$c = 0.5$				
		T:	1	5	10	15	20	1	5	10	15	20	1	5	10	15	20
$a = -0.005$	Supr.	Cov.	.91	.92	.92	.92	.91	.91	.91	.91	.91	.89	.91	.90	.90	.89	.88
		Bw.	.0125	.0108	.0089	.0073	.0063	.0132	.0133	.0136	.0144	.0157	.0141	.0172	.0249	.0417	.0798
	LSCI	Cov.	.96	.97	.93	.90	.88	.96	.95	.93	.89	.87	.97	.97	.98	.97	.96
		Bw.	.0120	.0073	.0049	.0048	.0051	.0124	.0126	.0098	.0110	.0145	.0131	.0187	.0235	.0495	.0972
	DtACI	Cov.	.91	.97	.96	.97	.95	.91	.94	.94	.95	.91	.91	.89	.91	.91	.90
		Bw.	.0125	.0116	.0091	.0076	.0065	.0132	.0127	.0128	.0135	.0149	.0141	.0144	.0204	.0336	.0634
	WCP Learned	Cov.	.90	.90	.90	.90	.89	.90	.90	.90	.89	.88	.90	.89	.89	.88	.88
		Bw.	.0116	.0099	.0083	.0069	.0061	.0122	.0124	.0129	.0137	.0152	.0132	.0164	.0241	.0394	.0759
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.91	.91	.91	.93	.97	.91	.92	.91	.93	.95	.91	.92	.91	.90	.87
		Bw.	.0125	.0103	.0086	.0078	.0078	.0133	.0135	.0135	.0153	.0192	.0143	.0188	.0276	.0437	.0681
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	2%
$a = -0.0075$	Supr.	Cov.	.91	.92	.92	.88	.78	.91	.91	.91	.83	.72	.91	.90	.89	.88	.84
		Bw.	.0125	.0109	.0092	.0087	.0116	.0132	.0136	.0144	.0172	.0300	.0142	.0173	.0257	.0460	.1006
	LSCI	Cov.	.96	.95	.89	.82	.81	.96	.93	.85	.76	.76	.96	.95	.91	.86	.83
		Bw.	.0121	.0068	.0062	.0079	.0121	.0125	.0097	.0111	.0164	.0316	.0132	.0156	.0250	.0567	.1235
	DtACI	Cov.	.91	.97	.96	.89	.64	.91	.94	.92	.73	.48	.91	.89	.89	.85	.66
		Bw.	.0125	.0117	.0093	.0086	.0107	.0132	.0128	.0133	.0159	.0269	.0142	.0144	.0210	.0382	.0876
	WCP Learned	Cov.	.90	.90	.90	.86	.77	.90	.90	.89	.81	.71	.90	.89	.89	.87	.84
		Bw.	.0116	.0101	.0086	.0085	.0116	.0123	.0127	.0137	.0168	.0298	.0132	.0164	.0248	.0449	.0991
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.91	.91	.92	.97	.96	.92	.91	.92	.95	.95	.92	.92	.91	.88	.89
		Bw.	.0126	.0105	.0094	.0111	.0140	.0135	.0136	.0153	.0220	.0368	.0144	.0189	.0279	.0447	.1084
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	1%	2%	0%	0%	2%	5%	6%
$a = -0.01$	Supr.	Cov.	.91	.92	.89	.70	.66	.91	.91	.86	.63	.59	.91	.90	.87	.69	.59
		Bw.	.0125	.0111	.0100	.0147	.0307	.0133	.0137	.0156	.0300	.0796	.0143	.0175	.0275	.0669	.2207
	LSCI	Cov.	.95	.92	.76	.75	.77	.96	.87	.76	.67	.69	.96	.89	.80	.74	.66
		Bw.	.0121	.0074	.0085	.0154	.0342	.0126	.0097	.0145	.0314	.0882	.0133	.0148	.0294	.0792	.2418
	DtACI	Cov.	.91	.97	.91	.48	.43	.91	.94	.80	.35	.33	.91	.88	.70	.39	.32
		Bw.	.0125	.0118	.0099	.0133	.0272	.0133	.0129	.0144	.0266	.0707	.0143	.0145	.0229	.0591	.1976
	WCP Learned	Cov.	.90	.90	.86	.70	.67	.90	.89	.84	.63	.59	.90	.89	.86	.68	.57
		Bw.	.0117	.0103	.0095	.0147	.0308	.0124	.0129	.0152	.0299	.0797	.0132	.0169	.0270	.0663	.2184
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.92	.91	.97	.97	.93	.92	.91	.94	.95	.95	.92	.92	.90	.89	.91
		Bw.	.0128	.0108	.0132	.0183	.0378	.0135	.0137	.0192	.0376	.1030	.0145	.0190	.0301	.0770	.2767
		n_∞	0%	0%	0%	2%	2%	0%	0%	2%	5%	5%	0%	0%	7%	12%	12%

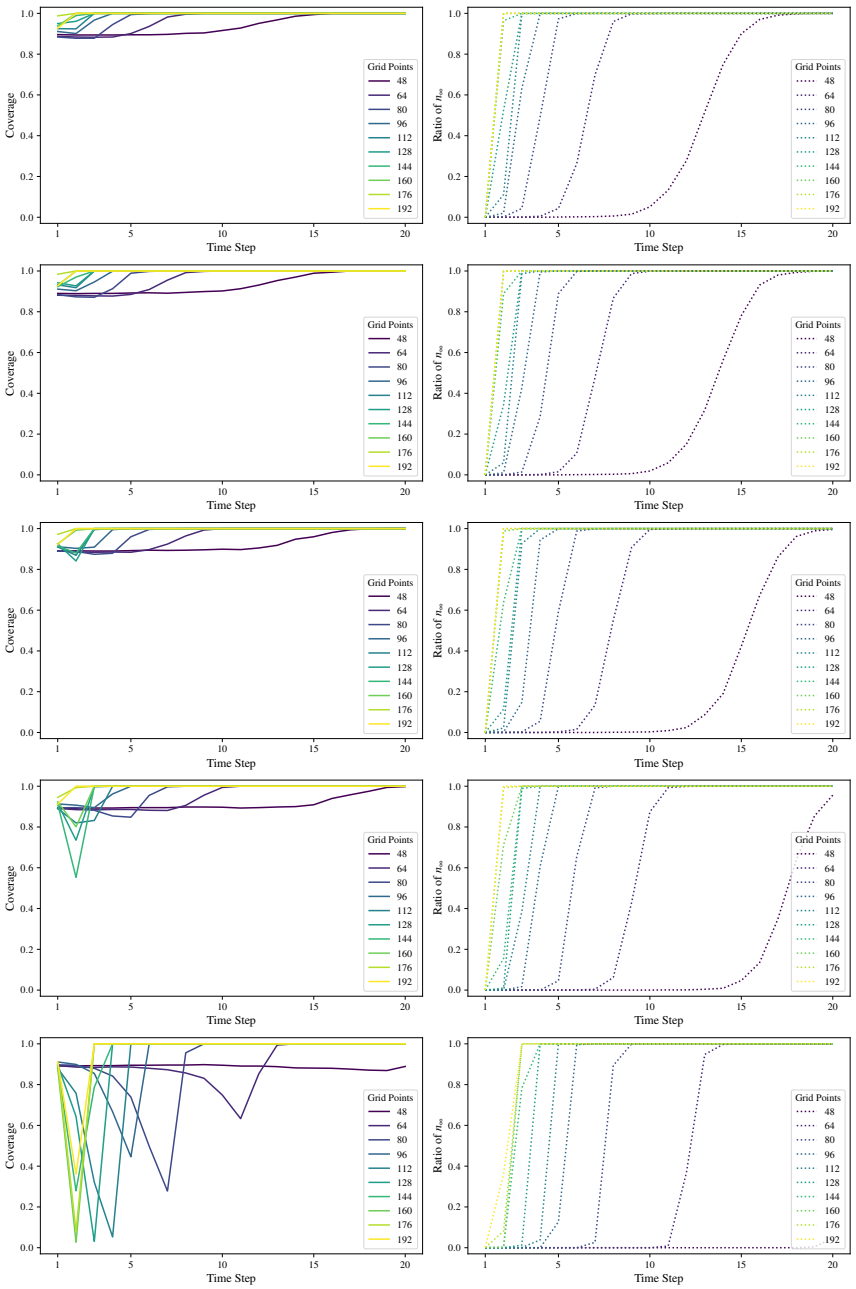


Figure H.2: Ablation study for method stability. We use data from the synthetic task (no recalibration) with parameters $a = -0.0075, b = 0, c = 0$. From top to bottom, the rows correspond to 5000, 4000, 3000, 2000, and 1000 calibration samples. Each plot reports coverage (left) and the fraction of infinite conformal bands (right) over time, where color corresponds to grid granularity. We see that higher granularity leads to earlier onset of infinite bands, which aligns with Proposition 4.4. When the number of samples is insufficient relative to the dimensionality, covariance estimation degrades and undercoverage can occur.

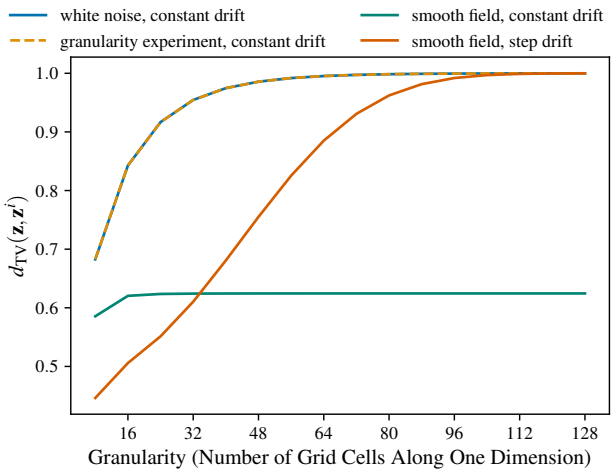


Figure H.3: TV distance of Theorem 4.5 at $|t_1 - t_2| = 0.5$ over the grid granularity. For the constant drift $\boldsymbol{v} = \mathbf{1}$ under white noise and the covariance of the granularity experiment in Figure 3a, $\mathbf{1}^\top \boldsymbol{\Sigma}_0^{-1} \mathbf{1} = p$ and both curves coincide. For the smooth Gaussian random field $\mathcal{N}(0, 625(-\partial_x^2 + 25\boldsymbol{I})^{-2})$ on $(0, 2\pi)$, evaluated on the grid, the TV distance levels off below 1 for the constant drift, but tends to 1 for a step drift ($v_j = 1$ on the left half of the domain, 0 otherwise).

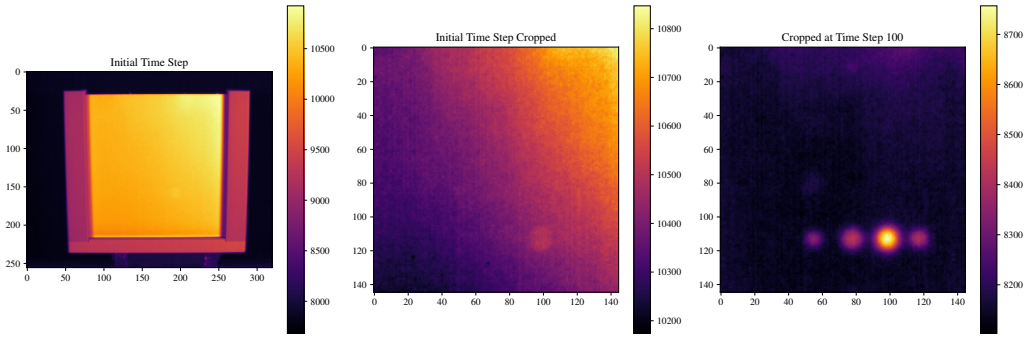


Figure I.1: Exemplary frames from the real-world pulsed-thermography dataset. The left image shows a raw data frame from the beginning of the cooling period. The middle image shows how we cropped the frames to remove the background. The right image shows how the data looks after 100 time steps.

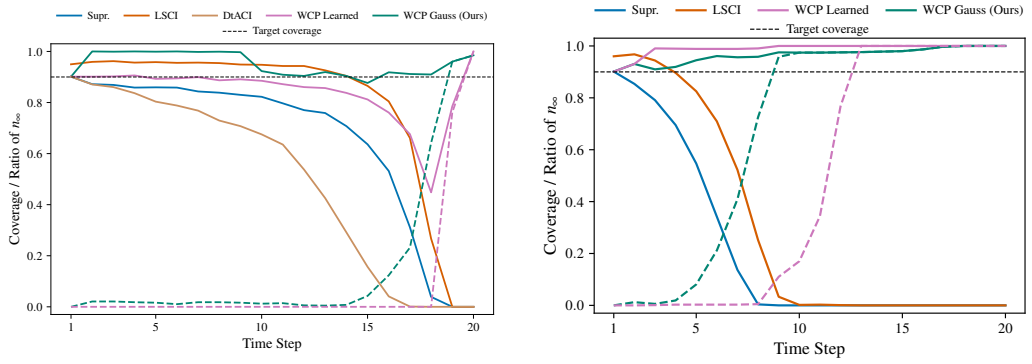


Figure I.2: Coverages for the pulsed-thermography dataset for the recalibration task (left) and the general task (right). The 90% coverage target and, for WCP Gauss (resp. Learned), the fraction of samples with infinite bandwidth n_∞ are marked as a black dashed line and a green (resp. pink) dashed line, respectively. We see that all methods achieve target coverage at first, but only WCP Gauss can preserve coverage over time.

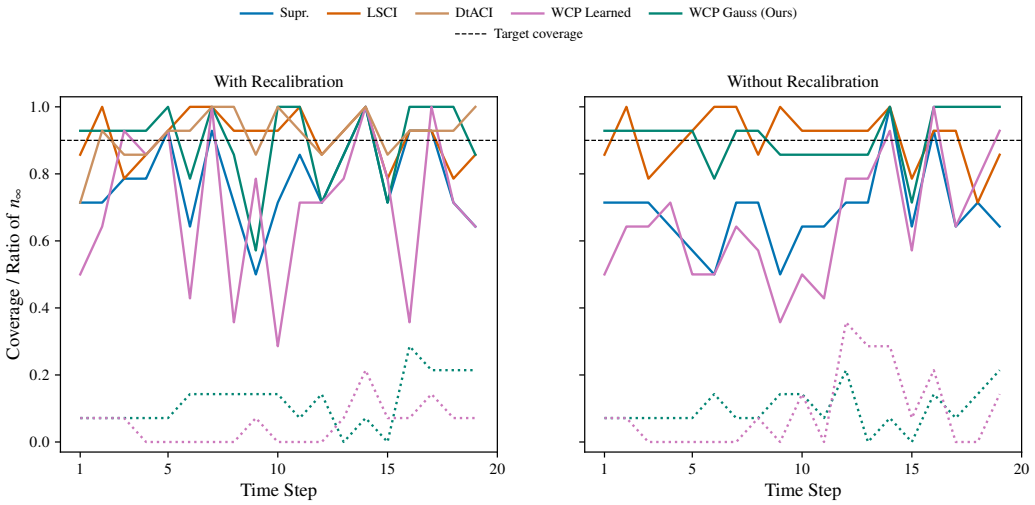


Figure I.3: Coverages for the Lake Maggiore nighttime cooling dataset for the recalibration task (left) and the general task (right). The 90% target is marked by a black dashed line. For WCP Gauss (resp. Learned), the fraction of samples with infinite bandwidth n_∞ is shown as a green (resp. pink) dotted line. WCP Gauss and LSCI maintain the target coverage on both tasks. Deviations at single time steps are statistically expected given the test size of 14.

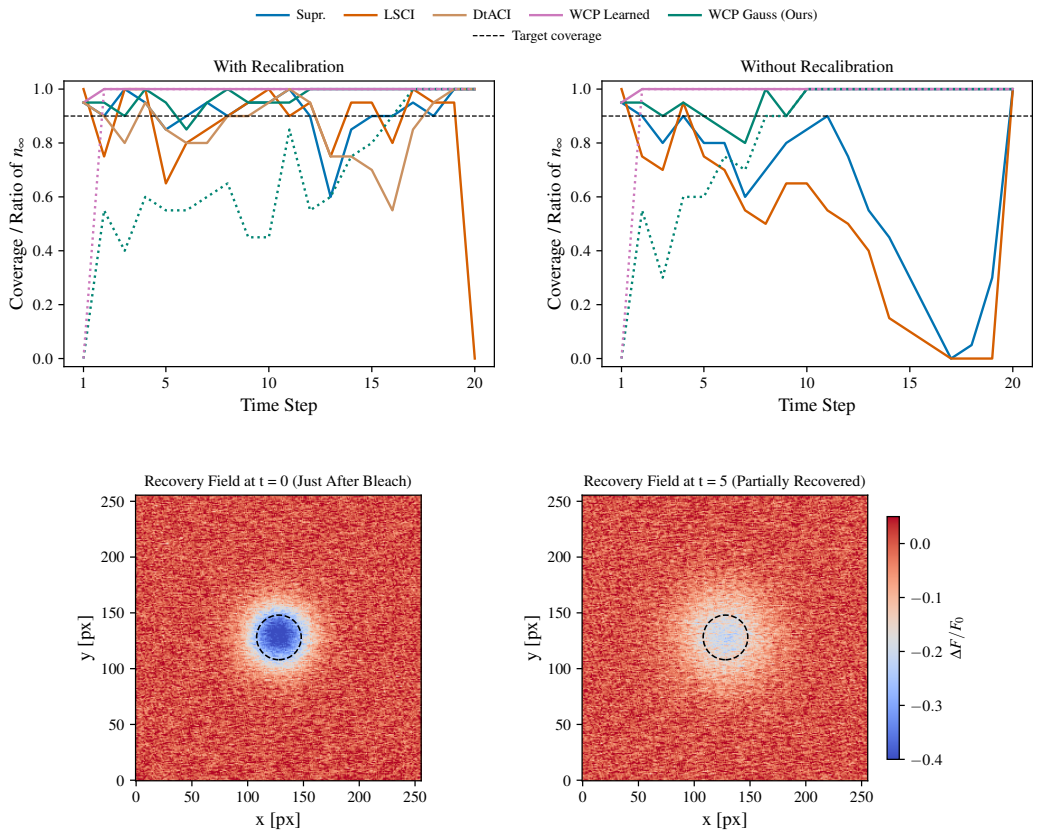


Figure I.4: (*Top*) Coverages for the FRAP dataset for the recalibration task (left) and the general task (right). The 90% target is marked by a black dashed line. For WCP Gauss (resp. Learned), the fraction of samples with infinite bandwidth n_∞ is shown as a green (resp. pink) dotted line. (*Bottom*) Recovery field snapshots from a real FRAP experiment on 32 wt% sucrose solution, showing the normalized fluorescence change $\Delta F/F_0$ just after photobleaching ($t = 0$) and after partial recovery ($t = 5$). The dashed circle marks the bleach center. WCP Gauss is the only method that provides non-trivial target coverage for every time step on both tasks, where WCP Learned immediately provides trivial bands.

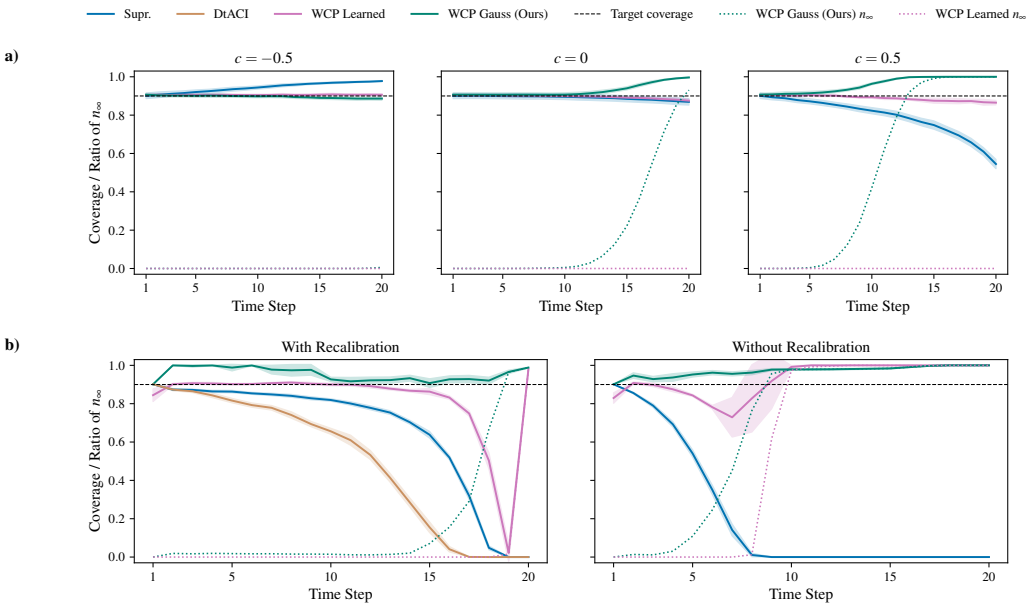


Figure J.1: Multi-seed stability. *(a)* Coverages for the general task of the synthetic PDE ($a = -0.0075$, $b = 0$) for $c \in \{-0.5, 0, 0.5\}$, mean $\pm$ one standard deviation over 10 reshuffled calibration/test splits. *(b)* Coverages for the pulsed-thermography dataset for the recalibration task (left) and the general task (right), mean $\pm$ one standard deviation over 10 plate-level reshuffles with a retrained surrogate per seed. The 90% target is marked by a black dashed line. For WCP Gauss (resp. Learned), the fraction of samples with infinite bandwidth n_∞ is shown as a green (resp. pink) dotted line.

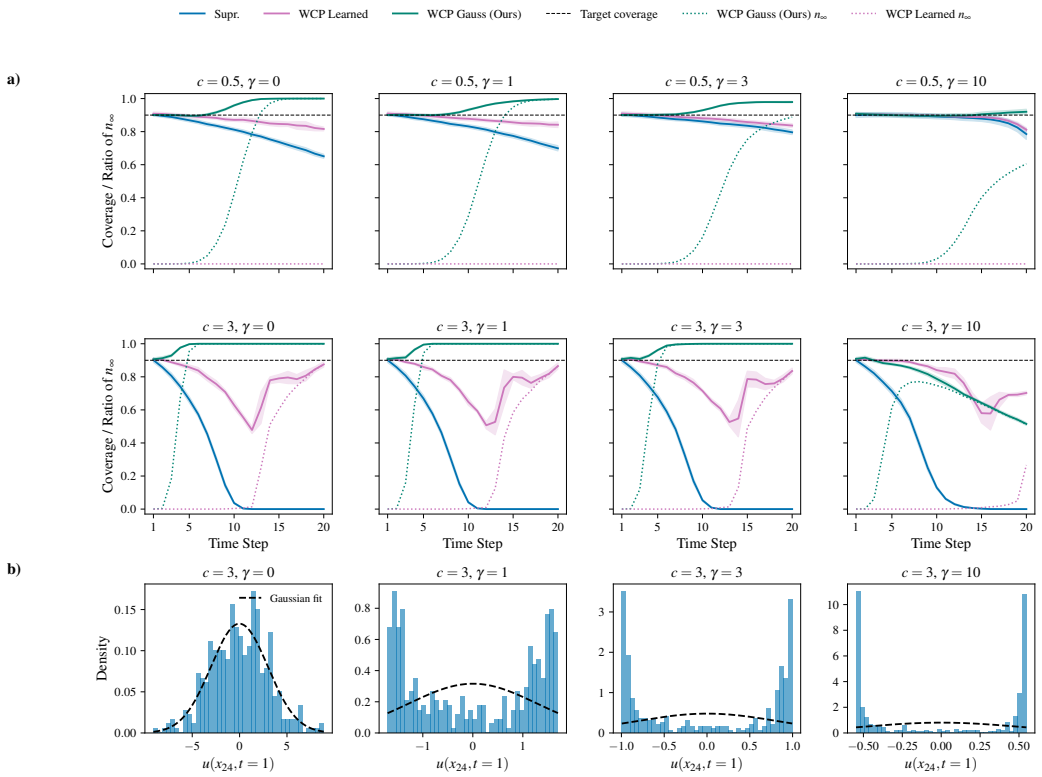


Figure J.2: Nonlinear ablation for $u_t = au_{xx} + cu - \gamma u^3$ ($a = -0.0075$, $b = 0$). *(a)* Coverages, mean $\pm$ one standard deviation over 10 reshuffled calibration/test splits, for $c = 0.5$ (top row) and $c = 3$ (bottom row). The 90% target is marked by a black dashed line. For WCP Gauss (resp. Learned), the fraction of samples with infinite bandwidth n_∞ is shown as a green (resp. pink) dotted line. *(b)* Marginal distribution of the middle grid point at $t = 1$ for $c = 3$ with a Gaussian fit.

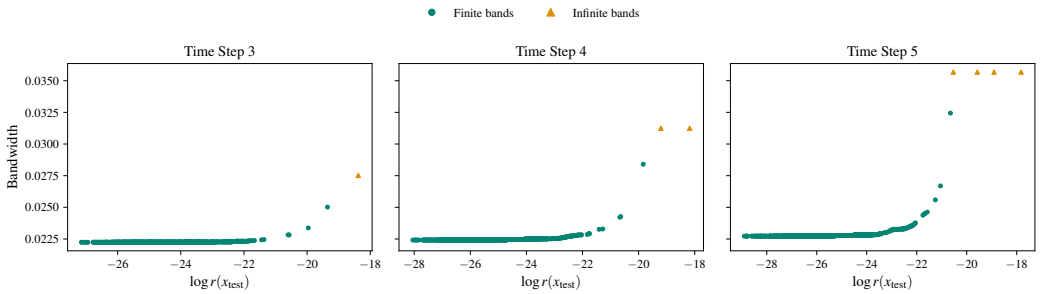


Figure J.3: Bandwidth of WCP Gauss per test sample against the test log likelihood-ratio $\log r(x_{\text{test}})$ for the synthetic PDE ($a = -0.0075$, $b = 0$, $c = 0.5$, general task) at the first time steps with infinite bands. Infinite bands are marked by triangles at the top of each panel.

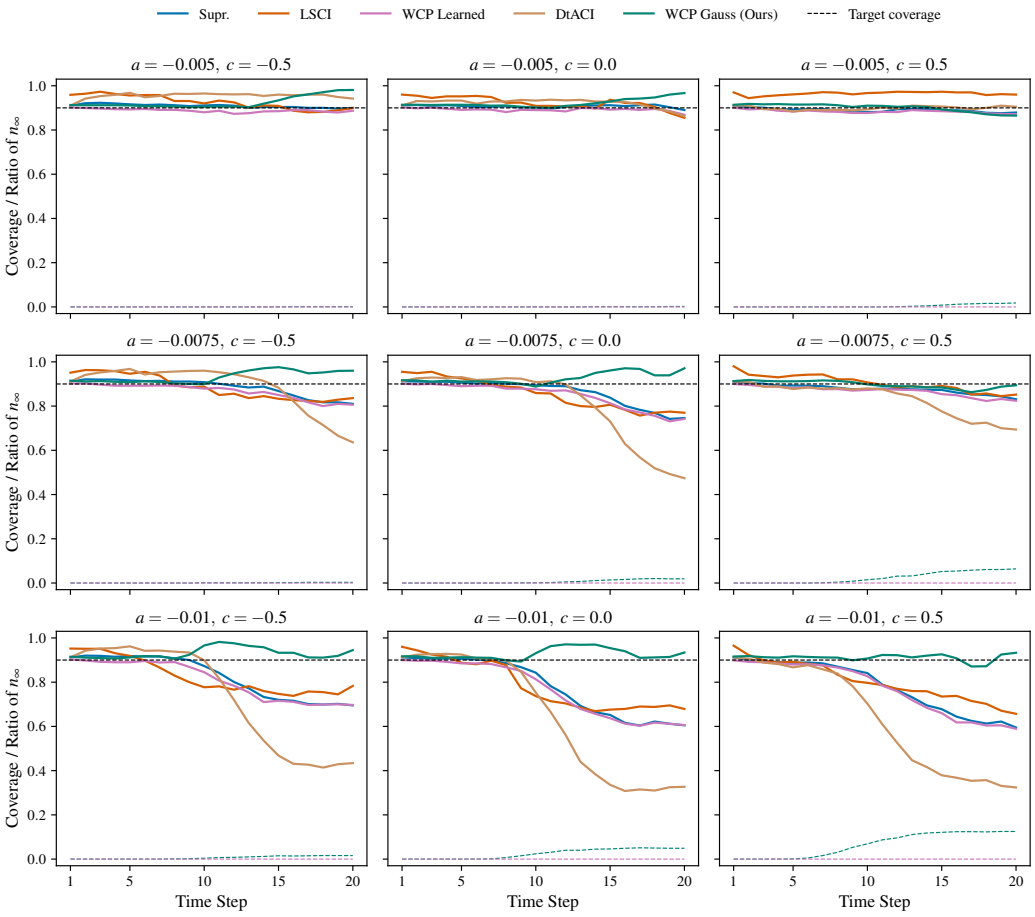


Figure K.1: Coverages for the recalibration task for varying a, c and fixed $b = -0.5$ across increasing prediction horizon. The 90% coverage target is marked as black dashed line and the fraction of samples with infinite bandwidth n_∞ for WCP Learned and WCP Gauss as dotted line in their respective color.

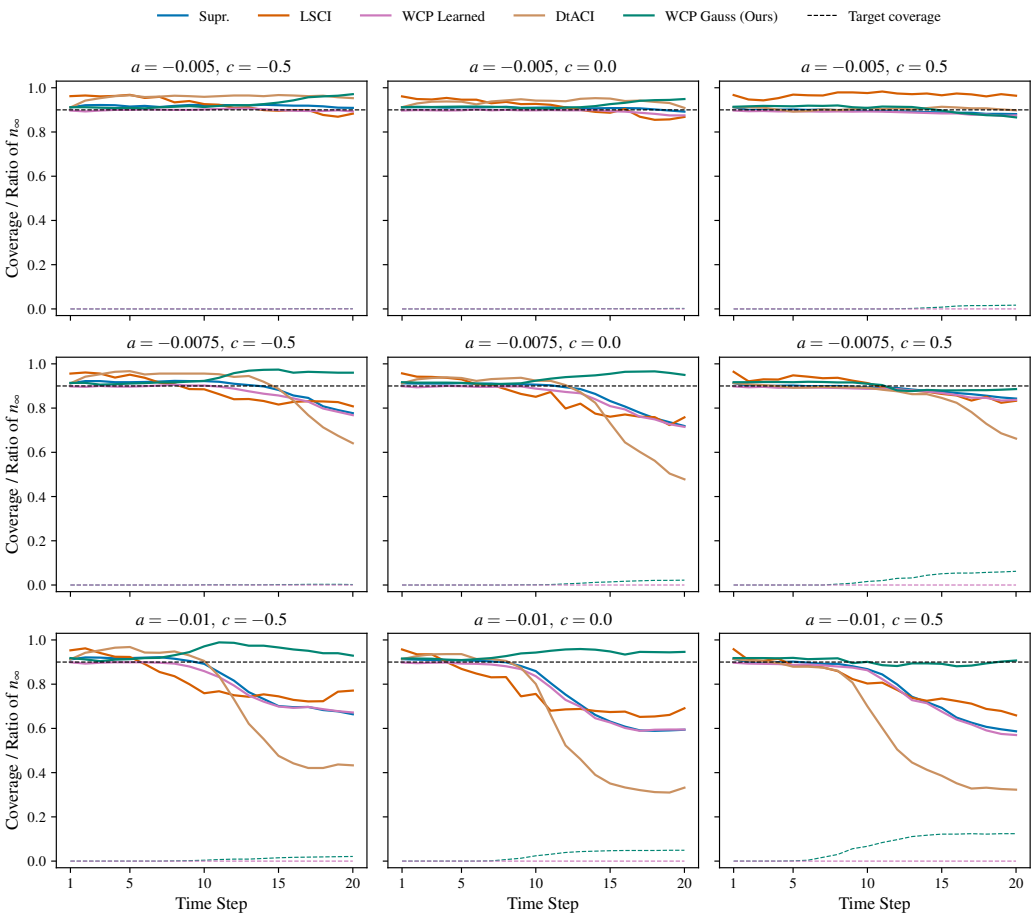


Figure K.2: Coverages for the recalibration task for varying a, c and fixed $b = 0$ across increasing prediction horizon. The 90% coverage target is marked as black dashed line and the fraction of samples with infinite bandwidth n_∞ for WCP Learned and WCP Gauss as dotted line in their respective color.

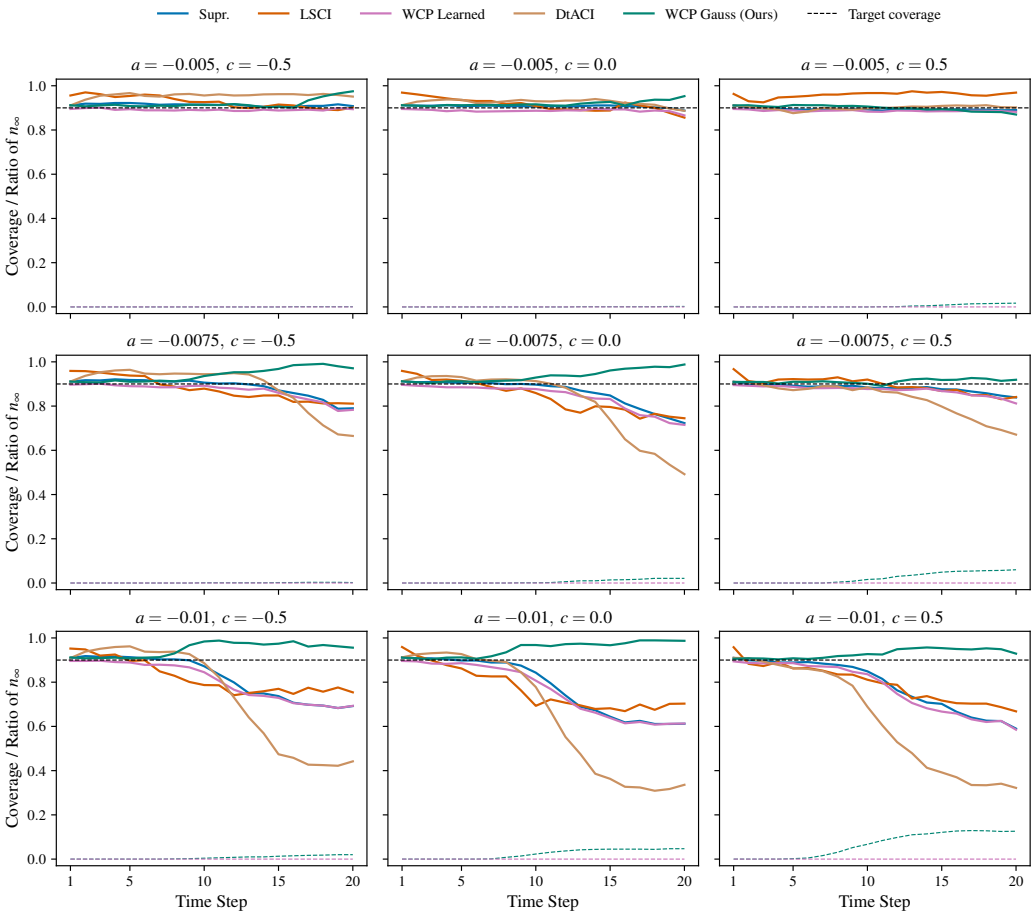


Figure K.3: Coverages for the recalibration task for varying a, c and fixed $b = 0.5$ across increasing prediction horizon. The 90% coverage target is marked as black dashed line and the fraction of samples with infinite bandwidth n_∞ for WCP Learned and WCP Gauss as dotted line in their respective color.

Table K.3: Coverage and mean bandwidth for the recalibration task for varying a, c and fixed $b = 0.5$. We also report the fraction of samples where infinite bands were reported n_∞. The mean bandwidth is taken over finite bands only. The gray font is chosen for better readability.

			$c = -0.5$					$c = 0.0$					$c = 0.5$				
		T:	1	5	10	15	20	1	5	10	15	20	1	5	10	15	20
$a = -0.005$	Supr.	Cov.	.91	.92	.92	.91	.91	.91	.91	.91	.91	.89	.91	.89	.89	.89	.89
		Bw.	.0127	.0109	.0090	.0075	.0067	.0134	.0135	.0139	.0147	.0164	.0144	.0172	.0248	.0426	.0809
	LSCI	Cov.	.96	.95	.93	.92	.90	.97	.94	.91	.89	.86	.96	.95	.97	.97	.97
		Bw.	.0100	.0067	.0053	.0053	.0056	.0102	.0096	.0095	.0127	.0150	.0108	.0137	.0238	.0505	.0986
	DtACI	Cov.	.91	.97	.96	.96	.95	.91	.94	.93	.93	.89	.91	.88	.90	.91	.90
		Bw.	.0127	.0117	.0096	.0080	.0068	.0134	.0130	.0133	.0139	.0154	.0144	.0146	.0213	.0346	.0638
	WCP Learned	Cov.	.90	.89	.89	.89	.89	.90	.89	.89	.89	.87	.90	.89	.88	.89	.88
		Bw.	.0117	.0101	.0084	.0072	.0064	.0124	.0127	.0129	.0140	.0157	.0133	.0168	.0238	.0403	.0766
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.91	.91	.91	.91	.97	.91	.91	.92	.93	.95	.91	.91	.91	.89	.87
		Bw.	.0129	.0106	.0089	.0074	.0080	.0134	.0137	.0143	.0155	.0199	.0144	.0187	.0276	.0431	.0683
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	2%
$a = -0.0075$	Supr.	Cov.	.91	.92	.91	.87	.79	.91	.91	.90	.85	.72	.91	.89	.88	.88	.84
		Bw.	.0127	.0110	.0093	.0092	.0123	.0134	.0137	.0145	.0182	.0316	.0144	.0174	.0256	.0472	.1050
	LSCI	Cov.	.96	.94	.88	.85	.81	.96	.91	.86	.80	.75	.97	.92	.92	.87	.84
		Bw.	.0092	.0070	.0068	.0086	.0129	.0098	.0091	.0135	.0177	.0333	.0102	.0129	.0255	.0580	.1262
	DtACI	Cov.	.91	.96	.94	.87	.67	.91	.93	.91	.74	.49	.91	.87	.89	.80	.67
		Bw.	.0127	.0118	.0098	.0090	.0112	.0134	.0131	.0138	.0168	.0282	.0144	.0147	.0219	.0396	.0908
	WCP Learned	Cov.	.90	.89	.89	.86	.78	.90	.89	.88	.83	.71	.90	.89	.88	.87	.81
		Bw.	.0118	.0102	.0088	.0090	.0122	.0125	.0128	.0136	.0179	.0314	.0133	.0170	.0244	.0456	.1014
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.91	.91	.94	.97	.97	.91	.91	.93	.96	.99	.91	.91	.90	.92	.92
		Bw.	.0128	.0108	.0102	.0113	.0153	.0134	.0139	.0163	.0231	.0436	.0143	.0188	.0274	.0559	.1276
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	1%	2%	0%	0%	2%	5%	6%
$a = -0.01$	Supr.	Cov.	.91	.91	.87	.74	.69	.91	.90	.84	.64	.61	.91	.89	.85	.70	.59
		Bw.	.0127	.0112	.0103	.0155	.0319	.0135	.0138	.0161	.0313	.0825	.0144	.0175	.0277	.0699	.2270
	LSCI	Cov.	.95	.90	.79	.77	.75	.96	.86	.69	.68	.70	.96	.86	.81	.72	.67
		Bw.	.0088	.0071	.0091	.0163	.0349	.0090	.0094	.0144	.0331	.0907	.0093	.0124	.0305	.0802	.2482
	DtACI	Cov.	.91	.96	.89	.47	.44	.91	.93	.78	.36	.34	.91	.86	.69	.39	.32
		Bw.	.0127	.0118	.0104	.0138	.0282	.0135	.0131	.0150	.0277	.0731	.0144	.0149	.0241	.0616	.2033
	WCP Learned	Cov.	.90	.89	.84	.73	.69	.90	.89	.81	.64	.61	.89	.89	.84	.67	.58
		Bw.	.0118	.0104	.0100	.0155	.0319	.0125	.0132	.0155	.0312	.0825	.0134	.0172	.0271	.0685	.2266
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.91	.91	.98	.97	.96	.91	.91	.97	.97	.99	.91	.91	.93	.95	.93
		Bw.	.0127	.0110	.0149	.0202	.0403	.0134	.0142	.0235	.0405	.1204	.0141	.0189	.0358	.0997	.2914
		n_∞	0%	0%	0%	1%	2%	0%	0%	2%	4%	5%	0%	0%	7%	12%	13%

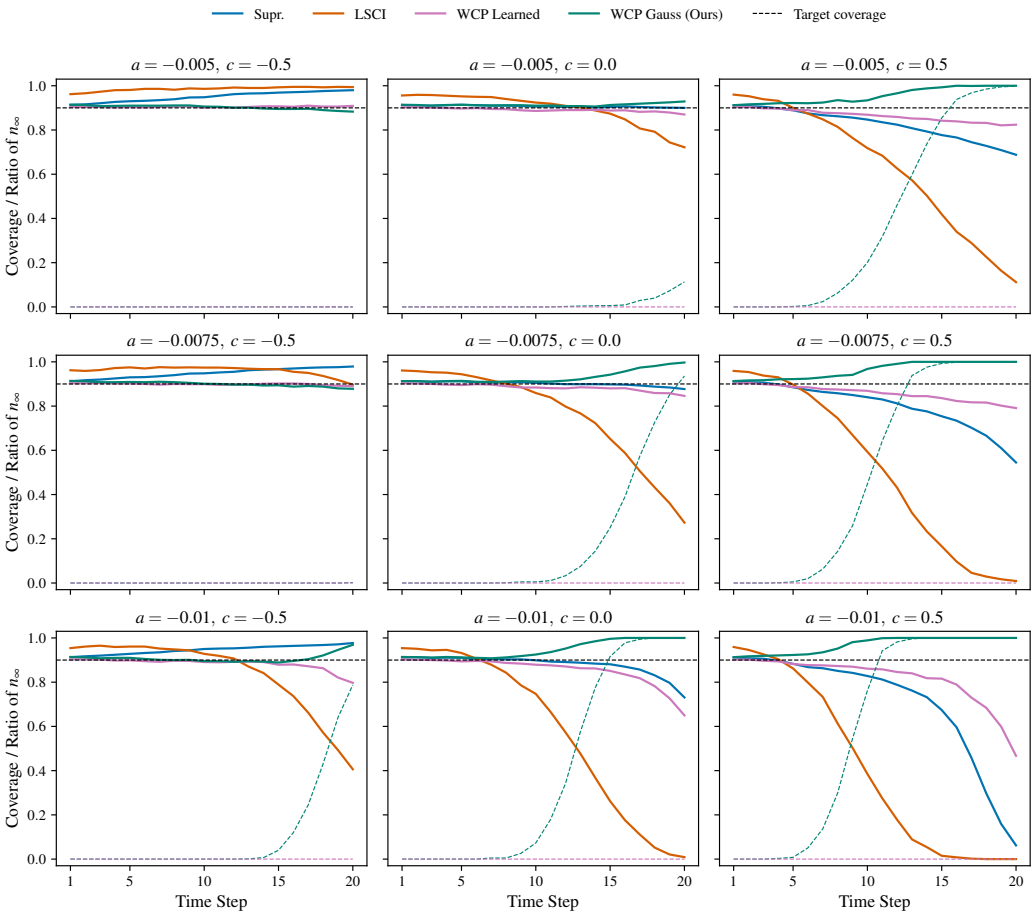


Figure K.4: Mean coverages for the general task (no recalibration) for varying a, c and fixed $b = -0.5$ across an increasing prediction horizon. The 90% coverage target is marked as black dashed line and the fraction of samples with infinite bandwidth n_∞ for WCP Learned and WCP Gauss as dotted line in their respective color.

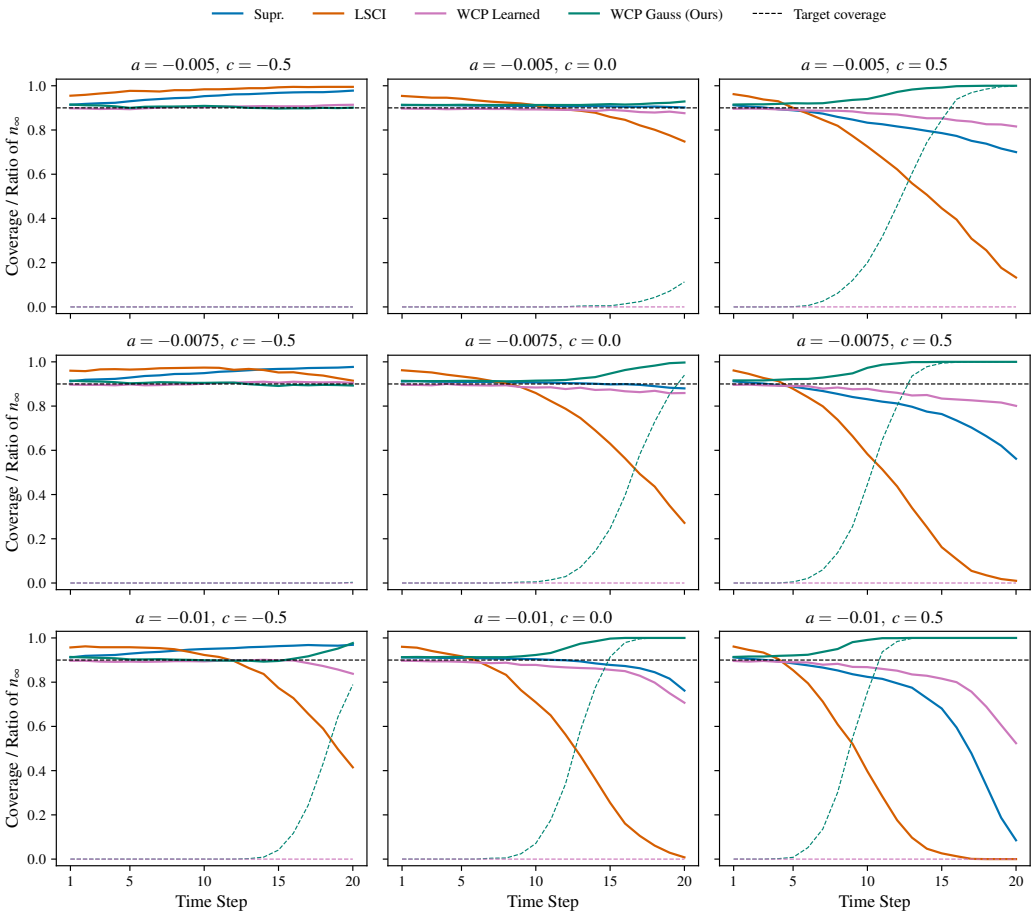


Figure K.5: Mean coverages for the general task (no recalibration) for varying a, c and fixed $b = 0$ across increasing prediction horizon. The 90% coverage target is marked as black dashed line and the fraction of samples with infinite bandwidth n_∞ for WCP Learned and WCP Gauss as dotted line in their respective color.

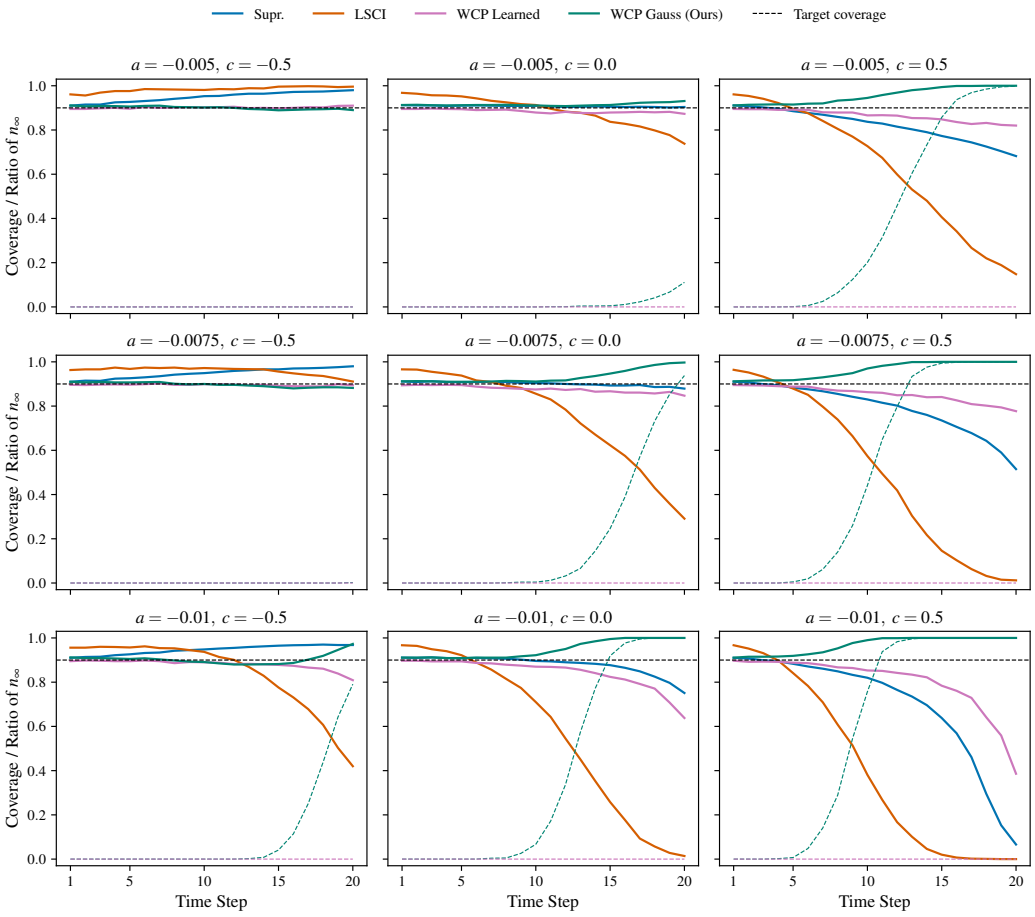


Figure K.6: Mean coverages for the general task (no recalibration) for varying a, c and fixed $b = 0.5$ across increasing prediction horizon. The 90% coverage target is marked as black dashed line and the fraction of samples with infinite bandwidth n_∞ for WCP Learned and WCP Gauss as dotted line in their respective color.

Table K.4: Coverage and mean bandwidth for the general task (no recalibration) for varying a, c and fixed $b = -0.5$. We also report the fraction of samples where infinite bands were reported n_∞. The mean bandwidth is taken over finite bands only. A dash (–) marks cells where every band was infinite. The gray font is chosen for better readability.

			$c = -0.5$					$c = 0.0$					$c = 0.5$				
		T:	1	5	10	15	20	1	5	10	15	20	1	5	10	15	20
$a = -0.005$	Supr.	Cov.	.91	.93	.95	.97	.98	.91	.91	.91	.90	.90	.91	.89	.85	.78	.69
		Bw.	.0129	.0129	.0129	.0129	.0129	.0132	.0132	.0132	.0132	.0132	.0136	.0136	.0136	.0136	.0136
	LSCI	Cov.	.96	.98	.99	.99	.99	.96	.95	.92	.87	.72	.96	.90	.72	.42	.11
		Bw.	.0122	.0122	.0123	.0123	.0123	.0124	.0124	.0124	.0124	.0124	.0127	.0127	.0126	.0125	.0124
	WCP Learned	Cov.	.90	.90	.90	.91	.91	.90	.90	.89	.89	.87	.90	.89	.87	.84	.82
		Bw.	.0119	.0111	.0103	.0095	.0087	.0122	.0123	.0122	.0124	.0122	.0127	.0136	.0153	.0165	.0184
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.91	.91	.91	.90	.88	.91	.91	.91	.91	.93	.91	.92	.93	.99	1.00
		Bw.	.0129	.0115	.0104	.0093	.0083	.0133	.0133	.0133	.0136	.0148	.0137	.0164	.0191	.0231	.0264
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	1%	11%	0%	0%	20%	85%	>99%
$a = -0.0075$	Supr.	Cov.	.91	.93	.95	.97	.98	.91	.91	.91	.90	.88	.91	.89	.84	.75	.55
		Bw.	.0128	.0128	.0128	.0128	.0128	.0132	.0132	.0132	.0132	.0132	.0137	.0137	.0137	.0137	.0137
	LSCI	Cov.	.96	.98	.97	.97	.90	.96	.94	.86	.65	.27	.96	.90	.59	.17	.01
		Bw.	.0122	.0123	.0123	.0123	.0123	.0125	.0125	.0125	.0125	.0125	.0128	.0127	.0126	.0125	.0124
	WCP Learned	Cov.	.90	.90	.90	.90	.89	.90	.90	.88	.88	.85	.90	.89	.87	.84	.79
		Bw.	.0119	.0111	.0104	.0095	.0087	.0123	.0124	.0123	.0125	.0122	.0127	.0137	.0154	.0165	.0180
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.91	.91	.90	.89	.88	.91	.91	.91	.94	1.00	.91	.92	.97	1.00	1.00
		Bw.	.0129	.0115	.0104	.0093	.0086	.0133	.0134	.0136	.0157	.0201	.0138	.0165	.0204	.0252	–
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	25%	94%	0%	0%	45%	99%	100%
$a = -0.01$	Supr.	Cov.	.91	.93	.95	.96	.98	.91	.91	.90	.88	.73	.91	.88	.83	.67	.06
		Bw.	.0129	.0129	.0129	.0129	.0129	.0132	.0132	.0132	.0132	.0132	.0137	.0137	.0137	.0137	.0137
	LSCI	Cov.	.95	.96	.93	.79	.41	.95	.93	.75	.26	.01	.96	.86	.39	.02	.00
		Bw.	.0122	.0123	.0123	.0123	.0123	.0125	.0125	.0125	.0125	.0125	.0128	.0127	.0126	.0126	.0124
	WCP Learned	Cov.	.90	.90	.89	.88	.80	.90	.89	.88	.85	.65	.90	.88	.86	.82	.47
		Bw.	.0119	.0111	.0103	.0095	.0087	.0123	.0123	.0124	.0124	.0125	.0127	.0136	.0153	.0167	.0183
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.91	.91	.90	.89	.97	.91	.91	.93	1.00	1.00	.91	.92	.99	1.00	1.00
		Bw.	.0129	.0116	.0105	.0099	.0100	.0133	.0134	.0146	.0182	–	.0138	.0167	.0217	–	–
		n_∞	0%	0%	0%	4%	79%	0%	0%	7%	92%	100%	0%	1%	76%	100%	100%

Table K.5: Coverage and mean bandwidth for the general task (no recalibration) for varying a, c and fixed $b = 0.0$. We also report the fraction of samples where infinite bands were reported n_∞. The mean bandwidth is taken over finite bands only. A dash (–) marks cells where every band was infinite. The gray font is chosen for better readability.

			$c = -0.5$					$c = 0.0$					$c = 0.5$				
		T:	1	5	10	15	20	1	5	10	15	20	1	5	10	15	20
$a = -0.005$	Supr.	Cov.	.91	.93	.95	.97	.98	.91	.91	.91	.91	.90	.91	.89	.83	.79	.70
		Bw.	.0128	.0128	.0128	.0128	.0128	.0132	.0132	.0132	.0132	.0132	.0136	.0136	.0136	.0136	.0136
	LSCI	Cov.	.96	.98	.98	.99	1.00	.95	.94	.91	.86	.75	.96	.90	.73	.45	.13
		Bw.	.0121	.0121	.0122	.0122	.0122	.0123	.0123	.0123	.0123	.0123	.0127	.0126	.0125	.0124	.0123
	WCP Learned	Cov.	.90	.89	.90	.91	.91	.90	.90	.89	.89	.88	.90	.89	.88	.85	.82
		Bw.	.0118	.0110	.0102	.0095	.0087	.0122	.0122	.0123	.0123	.0122	.0127	.0138	.0151	.0167	.0178
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.92	.90	.91	.90	.90	.91	.91	.91	.92	.93	.92	.92	.94	.99	1.00
		Bw.	.0129	.0112	.0103	.0091	.0083	.0133	.0133	.0134	.0138	.0151	.0138	.0164	.0196	.0250	.0275
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	11%	0%	0%	20%	85%	>99%
$a = -0.0075$	Supr.	Cov.	.91	.93	.95	.97	.98	.91	.91	.91	.90	.88	.91	.89	.83	.76	.56
		Bw.	.0128	.0128	.0128	.0128	.0128	.0132	.0132	.0132	.0132	.0132	.0136	.0136	.0136	.0136	.0136
	LSCI	Cov.	.96	.97	.97	.95	.92	.96	.93	.86	.63	.27	.96	.88	.58	.16	.01
		Bw.	.0122	.0122	.0122	.0123	.0123	.0124	.0124	.0124	.0124	.0124	.0127	.0126	.0125	.0125	.0123
	WCP Learned	Cov.	.90	.90	.90	.91	.91	.90	.90	.88	.88	.86	.90	.89	.88	.83	.80
		Bw.	.0118	.0111	.0102	.0095	.0087	.0122	.0122	.0122	.0123	.0123	.0127	.0139	.0153	.0164	.0184
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.92	.90	.91	.89	.89	.91	.91	.92	.95	1.00	.92	.92	.97	1.00	1.00
		Bw.	.0129	.0113	.0103	.0092	.0085	.0133	.0133	.0137	.0160	.0188	.0138	.0165	.0207	.0240	–
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	25%	94%	0%	0%	45%	99%	100%
$a = -0.01$	Supr.	Cov.	.91	.93	.95	.96	.97	.91	.91	.90	.88	.76	.91	.89	.82	.68	.09
		Bw.	.0128	.0128	.0128	.0128	.0128	.0132	.0132	.0132	.0132	.0132	.0136	.0136	.0136	.0136	.0136
	LSCI	Cov.	.96	.96	.92	.77	.42	.96	.92	.71	.26	.01	.96	.85	.40	.03	.00
		Bw.	.0122	.0122	.0123	.0123	.0123	.0124	.0124	.0124	.0124	.0124	.0127	.0127	.0126	.0125	.0124
	WCP Learned	Cov.	.90	.89	.89	.90	.84	.90	.89	.88	.86	.71	.90	.89	.87	.82	.52
		Bw.	.0119	.0111	.0102	.0095	.0088	.0122	.0122	.0123	.0123	.0125	.0127	.0139	.0154	.0164	.0185
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.91	.90	.90	.90	.98	.91	.91	.92	1.00	1.00	.91	.92	.99	1.00	1.00
		Bw.	.0129	.0114	.0104	.0097	.0099	.0133	.0134	.0147	.0194	–	.0138	.0166	.0228	–	–
		n_∞	0%	0%	0%	4%	79%	0%	0%	7%	91%	100%	0%	1%	76%	100%	100%

Table K.6: Coverage and mean bandwidth for the general task (no recalibration) for varying a, c and fixed $b = 0.5$. We also report the fraction of samples where infinite bands were reported n_∞. The mean bandwidth is taken over finite bands only. A dash (–) marks cells where every band was infinite. The gray font is chosen for better readability.

			$c = -0.5$					$c = 0.0$					$c = 0.5$				
		T:	1	5	10	15	20	1	5	10	15	20	1	5	10	15	20
$a = -0.005$	Supr.	Cov.	.91	.93	.95	.97	.98	.91	.91	.91	.91	.90	.91	.89	.84	.77	.68
		Bw.	.0129	.0129	.0129	.0129	.0129	.0133	.0133	.0133	.0133	.0133	.0137	.0137	.0137	.0137	.0137
	LSCI	Cov.	.96	.98	.98	1.00	1.00	.97	.95	.91	.84	.74	.96	.90	.73	.41	.15
		Bw.	.0122	.0122	.0122	.0123	.0123	.0124	.0124	.0124	.0124	.0124	.0127	.0126	.0125	.0124	.0123
	WCP Learned	Cov.	.90	.90	.90	.90	.91	.90	.89	.88	.88	.87	.90	.89	.87	.85	.82
		Bw.	.0119	.0109	.0103	.0095	.0087	.0123	.0122	.0122	.0123	.0122	.0127	.0140	.0149	.0167	.0182
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.91	.91	.90	.89	.89	.91	.91	.91	.91	.93	.91	.92	.94	.99	1.00
		Bw.	.0130	.0113	.0102	.0092	.0084	.0134	.0134	.0135	.0138	.0149	.0138	.0158	.0196	.0232	.0241
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	11%	0%	0%	20%	86%	>99%
$a = -0.0075$	Supr.	Cov.	.91	.93	.95	.97	.98	.91	.91	.90	.89	.88	.91	.88	.83	.73	.52
		Bw.	.0129	.0129	.0129	.0129	.0129	.0133	.0133	.0133	.0133	.0133	.0137	.0137	.0137	.0137	.0137
	LSCI	Cov.	.96	.97	.97	.96	.91	.97	.94	.86	.62	.29	.96	.88	.57	.15	.01
		Bw.	.0122	.0123	.0123	.0123	.0123	.0124	.0125	.0125	.0125	.0125	.0127	.0127	.0126	.0125	.0124
	WCP Learned	Cov.	.90	.90	.90	.89	.89	.89	.90	.88	.87	.85	.90	.89	.86	.84	.78
		Bw.	.0119	.0110	.0102	.0095	.0087	.0123	.0123	.0122	.0121	.0123	.0127	.0140	.0151	.0170	.0179
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.91	.91	.90	.89	.88	.91	.91	.91	.95	1.00	.91	.92	.97	1.00	1.00
		Bw.	.0129	.0113	.0103	.0092	.0086	.0134	.0134	.0137	.0156	.0203	.0139	.0160	.0208	.0244	–
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	25%	94%	0%	0%	44%	99%	100%
$a = -0.01$	Supr.	Cov.	.91	.93	.95	.96	.97	.91	.91	.90	.88	.75	.91	.88	.82	.64	.07
		Bw.	.0129	.0129	.0129	.0129	.0129	.0133	.0133	.0133	.0133	.0133	.0138	.0138	.0138	.0138	.0138
	LSCI	Cov.	.96	.96	.94	.78	.42	.97	.92	.71	.26	.01	.97	.84	.38	.02	.00
		Bw.	.0123	.0123	.0123	.0123	.0124	.0125	.0125	.0125	.0125	.0125	.0128	.0127	.0126	.0125	.0124
	WCP Learned	Cov.	.90	.90	.89	.88	.81	.90	.89	.87	.82	.64	.90	.89	.85	.78	.39
		Bw.	.0119	.0111	.0103	.0096	.0088	.0123	.0124	.0124	.0122	.0123	.0128	.0140	.0150	.0165	.0175
		n_∞	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	WCP Gauss (Ours)	Cov.	.91	.91	.89	.88	.97	.91	.91	.92	.99	1.00	.91	.92	.99	1.00	1.00
		Bw.	.0130	.0115	.0103	.0098	.0101	.0133	.0134	.0146	.0186	–	.0139	.0162	.0222	–	–
		n_∞	0%	0%	0%	4%	79%	0%	0%	7%	92%	100%	0%	1%	76%	100%	100%

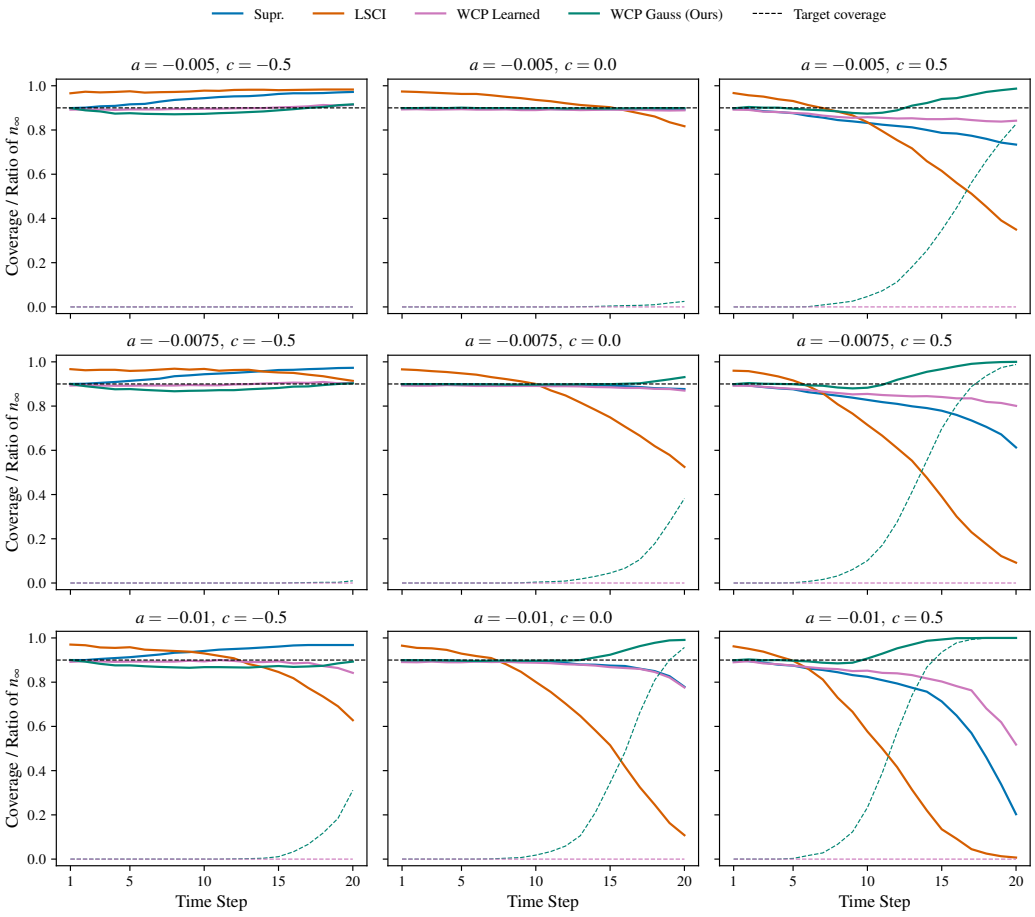


Figure K.7: Coverages for the general task for varying a, c and fixed $b = 0$ with Laplace initial distribution across increasing prediction horizon. The 90% coverage target is marked as black dashed line and the fraction of samples with infinite bandwidth n_∞ for WCP Learned and WCP Gauss as dotted line in their respective color.

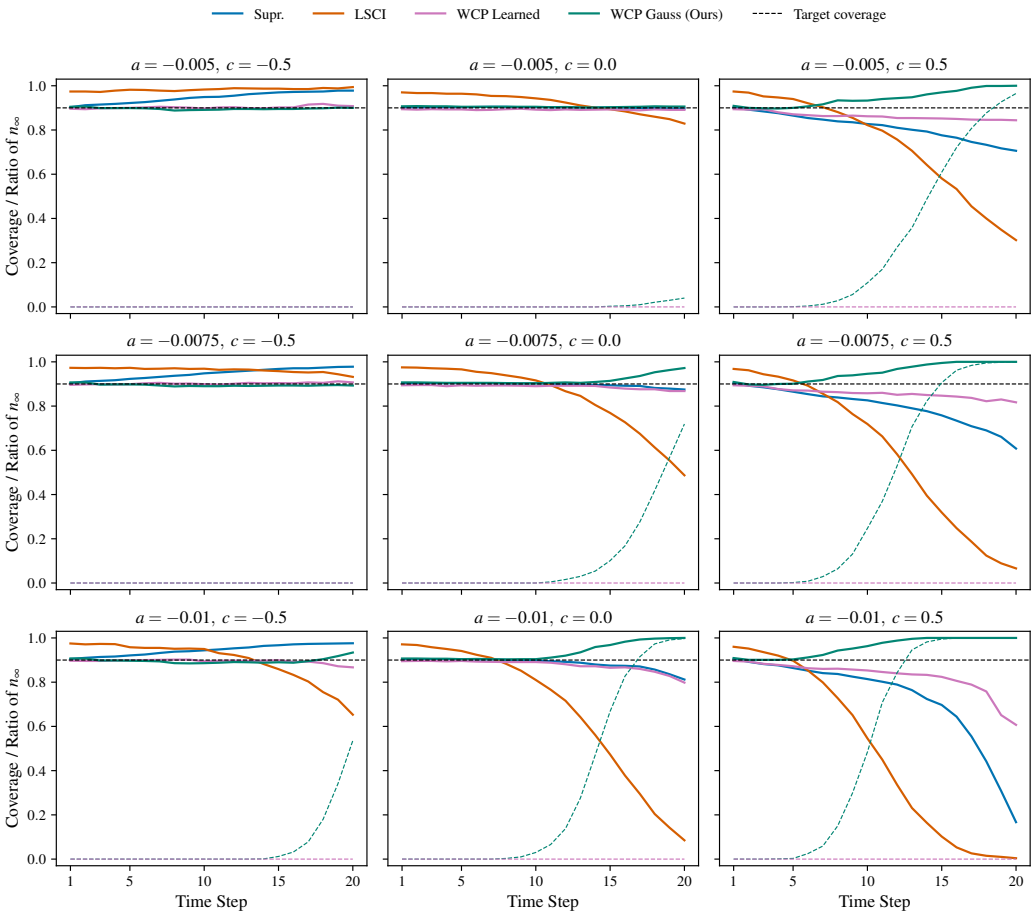


Figure K.8: Coverages for the general task for varying a, c and fixed $b = 0$ with Logistic initial distribution across increasing prediction horizon. The 90% coverage target is marked as black dashed line and the fraction of samples with infinite bandwidth n_∞ for WCP Learned and WCP Gauss as dotted line in their respective color.

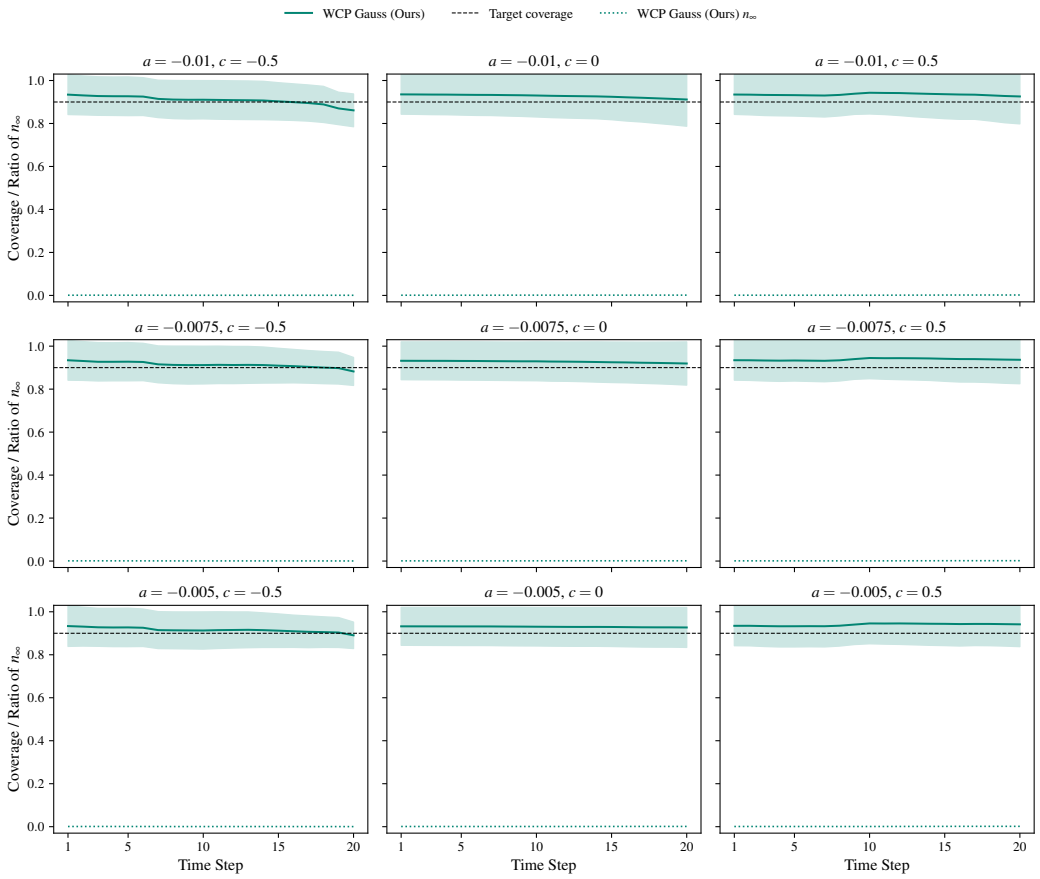


Figure K.9: Coverages of WCP with Cauchy likelihood-ratio weights for the general task for varying a, c and fixed $b = 0$ with heavy-tailed elliptical Cauchy initial distribution, mean $\pm$ one standard deviation over 10 reshuffled calibration/test splits. The 90% coverage target is marked as black dashed line and the fraction of samples with infinite bandwidth n_∞ as dotted line. The Cauchy likelihood ratios stay bounded, so no infinite bands occur. The heavy tails increase the variance of the coverage across splits, and for $c = -0.5$ the mean coverage falls slightly below the target at late time steps.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We investigate time-dependent PDEs. Our first theorem shows the claim about function spaces and the rest is dedicated to PDEs on discretized spaces. We especially show that weighted conformal prediction on linear PDEs can use the density ratio arising from Gaussianity.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a dedicated section on limitations in the main paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Each of the theoretical results has a small proof sketch in the main paper and a complete proof in the appendix. The assumptions are clearly stated in the main paper and referenced in the proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We detail our experimental setting in the appendix and provide code to reproduce our results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide code and a readme with instructions to reproduce our results at `https://github.com/StillJosh/Conformal_Prediction_for_Time-Dependent_PDEs`. The data is either generated by our code or publicly available via the references we provide.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All the training and test details are specified in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not claim competitive advantage over other methods. Our main contribution is a theoretical one, and the experiments are meant to illustrate the theory. We just want to show that CP methods that have no guarantees in the non-exchangeable setting will often lose coverage, while our method will instead predict infinite intervals. We further highlight that the fact that other models lose coverage over time is highly dependent on the choice of hyperparameters. Providing error bars would provide a false sense of significance to these results, which we do not want to claim. We explicitly want to highlight that coverage guarantees can be lost over time, but that this does not necessarily happen. Following the rebuttal, we added one representative robustness experiment to the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have a section in the appendix with details about the computing resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: No potentially harmful consequences are expected to arise from this research, neither during the research process nor from the results. We have reviewed the NeurIPS Code of Ethics and confirm that our research conforms to it in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: There are no negative societal impacts. However, we briefly discuss the benefits of better uncertainty quantification and highlight their importance for safety-critical physical applications.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: No such risks are posed by our paper.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All external datasets and the FRAP reference code are cited in the paper and listed with their licenses and URLs in Appendix B. All are used in accordance with their licenses (CC-BY 4.0).

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our only asset is the code released, which is well documented.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We added a section in the appendix describing how we used LLMs in our research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.