Trajectory-Level Conformal Prediction for Epidemiological Forecasting under Scenario Shifts

Garv Patel
garvp@umich.edu
University of Michigan
USA

ABSTRACT

Scenario modeling in epidemiology often relies on surrogate models that generate ensembles of possible futures at low computational cost. However, under distribution shifts such as policy interventions, holiday effects, or new variants, these ensemble spreads are frequently miscalibrated, undermining their reliability for public health decision-making. We address this reliability gap by constructing a fast and accurate surrogate model and applying Adaptive Conformal Prediction over Trajectory Ensembles (CP-Traj) to epidemiological forecasting. CP-Traj calibrates entire trajectories by constructing stepwise uncertainty bands that achieve target coverage while remaining as sharp as possible. Using simulated data, we evaluate CP-Traj on simulation tasks. We demonstrate that our surrogate model reliably reproduces SIR epidemic dynamics across different epidemiological scenarios, achieving low validation error and strong generalization on varying trajectories. However, applying CP-Traj to this setting reveals substantial challenges in maintaining stable and meaningful uncertainty throughout the epidemic's trajectory. Running CP-Traj with calibration maintains controlled miscoverage parameters but achieves only limited effective coverage (averaging roughly 19%) mostly on the epidemic's tail. These findings suggest that while CP-Traj provides interpretable calibration behavior, additional methodological advances or more accurate calibration is needed for reliable uncertainty quantification under evolving epidemiological scenarios.

Project Website: https://garv-p.github.io/598_project_website/

KEYWORDS

Conformal Prediction, Time Series Forecasting, Epidemiology, COVID-19, Uncertainty Quantification, Ensemble Forecasting, Surrogate models, CP-Traj

1 INTRODUCTION

Many critical decisions must be made under scenario shifts—such as climate change in weather forecasting or sudden policy changes in epidemiology, yet running full-fidelity simulators for every possible scenario is computationally prohibitive. Surrogate models offer a far more efficient alternative for simulation, producing trajectory ensembles at a fraction of the cost. However, when the target scenario's distribution departs from training conditions such as greenhouse gas–driven climate shifts in weather, or intervention-driven dynamics in epidemiology, this stresses the underlying assumptions that the surrogate models rely on.[5]

As a result, their ensemble spreads can become miscalibrated. This hampers the ability for policy makers and others informed by these simulations to make these decisions with confidence.[5]

This project targets the reliability gap from this distribution shift. We adopt established scenario-modeling settings from the epidemiology communities such as policy interventions, holidays, and vaccine adoption[5] to evaluate forecasts where calibrated uncertainty matters most for planning, resource allocation, and risk communication.

Conformal prediction (CP) offers a statistically rigorous way of creating uncertainty intervals for the predictions of models without making distributional assumptions. We specifically build on *Adaptive Conformal Prediction over Trajectory Ensembles (CP-Traj)* [6], which calibrates entire paths: an inter-step optimization allocates stepwise radii so the resulting trajectory-level set achieves target coverage and can form discontinuous bands around distinct modes.

Within this broader context, our goals are to create a model that can model different scenarios dictated by certain parameters and evaluate how CP-Traj can be used to help improve the confidence in trajectory ensembles in the scenario modeling. We evaluate in the domain of epidemiology.

2 PROBLEM DEFINITION

Lay problem definition:

Surrogate scenario modelers are used to model many possible futures but under scenario shifts the ensemble spread of a given model can be miscalibrated. We train an Epidemiology Informed Neural Network to take in some scenario parameters to model this possible scenario. We use Cp-Traj that takes the sample trajectories from the model to create an uncertainty set that attains our target confidence interval, stays as narrow as possible, and can capture multiple possible futures as trajectories diverge. Surrogate forecasters can sample many possible futures but under scenario shifts the ensemble spread given is often miscalibrated.

Formal setting. Let $x_t = (S_t, I_t, R_t)$ denote the percentage of the population that is currently susceptible(S), Infected(I), and Recovered(R) at time t, with parameters $\Omega = \beta, \gamma$ where β is the infectivity rate and γ is the mean infectious period, be the outputs and parameters of our teacher model(a mechanistic epidemiological model like SIR).

We generate training data from the simulator model by running many simulations with $\Omega_s = \beta_s, \gamma_s$ used to model different scenarios S. A scenario $\mathcal{S}$ specifies future covariate paths in areas like vaccine adoption, holiday effects, mobility assumptions.

We train the surrogate to model the conditional distribution from the teacher data. This surrogate is trained such that given the initial epidemic state $x_i = (S_i, I_i, R_i)$, and an epidemic scenario as defined by $\Omega_e = \beta_e, \gamma_e)$, it is able to output future predictions of cases and admissions in that scenario.

For a horizon H and K scenarios $S = S^1, ..., S^K$, a surrogate simulator produces an ensemble of K sampled trajectories at each time step. For lead h, define $y_t^h := Y_{t+h}$ for $h \in [1, H]$. the samples are generated as:

$$\widehat{y}^{(k)}_{t,1:H} \sim \hat{p}\left(\cdot \,\middle|\, X_{1:t},\, X^{\mathcal{S}}_{t+1:t+H}\right), \quad k = 1, \ldots, K.$$

Each sampled trajectory provides predictions at every lead point in $[1, H]$, giving a scenario-conditioned distribution over the future path. We keep a short rolling buffer B_t of recent forecast-truth pairs. This buffer is used for calibration and to help us determine the confidence intervals for each of our trajectories and to update them as new information comes in from the real scenario S_T.

Goal. We aim to have trajectory level confidence intervals that attain a user specified coverage rate $1 - \alpha$ over the time steps t and across the forecast horizons $h \in [1, H]$, that is as small as possible by the user defined objective J.

3 RELATED WORK

3.1 Conformal Prediction and Distribution Shift

Conformal prediction has been widely studied as a way to provide distribution-free uncertainty quantification. A central challenge is handling non-exchangeable or shifted data, which arises naturally in time series and epidemiological contexts. Recent work on Adaptive Conformal Inference under Distribution Shift [3] develops methods to adjust conformal sets when the test distribution may differ from training, while Probabilistic Conformal Prediction [1] proposes probabilistic refinements of coverage guarantees. These approaches highlight the tension between adaptivity and validity when forecasting under regime changes, such as new variants or interventions during the COVID-19 pandemic.

3.2 Trajectory-Level Calibration

The most relevant advance for our purposes is CP-Traj [6], which introduces a unified framework for conformalizing ensembles of sampled trajectories. Unlike marginal conformal prediction, CP-Traj enforces trajectory-level coverage across horizons by solving an optimization problem over radius parameters, yielding sharper and more adaptive uncertainty sets. This aligns directly with our methodological goals.

3.3 Epidemiological Ensemble Forecasting

During COVID-19, collaborative platforms such as the U.S. COVID-19 Scenario Modeling Hub were developed to combine forecasts from multiple teams. The evaluation study by Johansson et al. [4] showed that ensemble forecasts provided useful guidance but often suffered from poorly calibrated intervals, particularly under sharp epidemic transitions. These findings motivate approaches like CP-Traj that explicitly address coverage.

3.4 Neural Forecasting Models

Alongside statistical baselines, recent work has developed neural architectures that incorporate epidemiological structure. The Epidemiologically Informed Neural Network (EINN) [7] embeds compartmental dynamics into deep learning models to improve generalization and interpretability. These models provide for a robust hybrid model that we can use for our scenario prediction, pairing with conformal prediction.

4 METHOD

4.1 Overview

We use the surrogate model to produce K sampled trajectories over horizon H at each time step t. We instantiate the trajectory-level calibration using *Adaptive Conformal Prediction over Trajectory Ensembles (CP-Traj)* [6] to obtain trajectory-level uncertainty bands that attain the user-specified target coverage and are as small as possible.

4.2 Neural Network Surrogate Model

We plan to use a neural network to capture the dynamics of the mechanistic model and to use as our surrogate model. In order to incorporate the temporal dynamics from the simulation data we integrate the transformer architecture with self attention. This design enables the model to learn and understand long-range interactions across timesteps that traditional feedforward networks would not. This surrogate would be trained on simulation data generated across many (β, γ) pairs and time steps.

4.3 Mechanstic SIR Simulator Model

For our teacher model, we use a mechanistic SIR model, a compartmental model that splits the population into four parts: Susceptible(people who can catch the disease), Infectious(can spread the disease), and Recovered(those who have either recovered or died from the disease). These dynamics are controlled by the rates of infection β which determine the rate at which susceptible become exposed and the recovery rate γ which determine the rate at which those infected become recovered.[2] Using these dynamics and parameters, we can determine the change in each of these compartments and gain information about the spread of disease on the population. Importantly we can model scenarios by changing the parameters of this model, for example a stricter mask mandate causing β to decrease.

4.4 CP-Traj

To quantify uncertainty across the simulated trajectories we use the methods described in Adaptive Conformal Prediction over Trajectory ensembles. Given an ensemble of K trajectories from the surrogate for a chosen scenario and horizon H, CP-Traj converts those samples into a horizon wide, trajectory level uncertainty bands that attain the user's target confidence level $1 - \alpha$ and are as small as possible under the user defined objective function. It does this by maintaining a Buffer B of recent forecast-truth pairs which is used to compute nonconformity scores for each trajectory, measuring how far each trajectory is from the observed truth. From these nonformity scores S, CP-Traj creates prediction intervals for each forescast step.

As new observations come in, the buffer is updated and the coverage interval are recalibrated.

5 EXPERIMENTS

5.1 Data

We generate the dataset by running our SIR simulator model across a range of epidemiological parameters. Specifically we sample $n = 100$ random pairs of (β, γ) where $\beta \in [0, 1], \gamma \in [0, 1]$. For each sampled pair we simulate the dynamics over $T = 200$ steps using fixed initial conditions to obtain trajectories of each of the compartments across time. To generate trajectories for our calibration and test sets we choose a base beta and gamma to center our ensemble around. Using these base parameters we generate 10 random beta-gamma pairs to create our ensemble and we take the mean of these parameters to get our ground truth beta and gamma for that ensemble

5.2 Tasks

Our approach is successful if the surrogate model can reliably reproduce the dynamics of the mechanistic SIR simulations and if CP-Traj produces calibrated and concise uncertainty bounds across the scenarios. Specifically our success criteria entails:

- The surrogate accurately captures the relationship between the epidemiological parameters (β, γ) and the resulting trajectories as well as generalization to unseen parameter combinations.
- The conformal prediction procedure maintains a coverage close to the desired confidence level $1 - \alpha$ while producing narrow intervals.

5.3 Experimental Setup

To evaluate the effectiveness of CP-Traj in quanitfying uncertainty in our simulation, we conduct experiments using a synthetic SIR dataset designed to mimic outbreak scenarios. To train our surrogate model we use a dataset of 2000 trajectories split evenly in choice of beta-gamma pairs to model epidemics that outbreak and decay.

To create the ensemble for our testing phase, we choose 10 beta-gamma pair scenarios to form the ensemble used for CP-Traj. Each pair represents a distinct infection and recovery dynamic that represent individual scenarios. For each selected pair we run the surrogate model with parameters (β_i, γ_i, t_i) for each of the ten scenarios. We select another pair of parameters (β_G, γ_G) that is the closest to the 10 selected scenario pairs to be our ground truth and run the mechanistic SIR model to produce the observed trajectory against which the predictions are compared. This ensures that the reality of the situation lies within our predicted conditions in the ensemble.

The CP-Traj procedure is applied iteratively at each timestep, updating the calibration buffer and uncertainty intervals as new ground truth information becomes available. To handle the shifts of the epidemic we employ an adaptation rate of 0.35 and a memory window of 10 days. This setup allows the system to rapidly forget errors made at the very beginning of the model trajectory while also being able to adapt to the peak and recovery of the epidemic, allowing us to evaluate the adaptive calibration and coverage of CP-Traj across the different phases of the epidemic. To initialize the system, we calibrate using 100 trajectory ensembles, ensuring that the system established a baseline uncertainty with $\alpha = .1$ before onilne adaptation began.

5.4 Results

5.4.1 *Surrogate Model performance.*

We evaluate the performance of the Surrogate model against the ground truth generated by the SIR model. Initial experiments using Mean Squared Error yielded suboptimal performance particularly in the tail end of the pandemic where infections come close to 0. We shifted to using the Mean Square Logarithmic Error to enforce sensitivity across the smaller orders of magnitude around the tail. We also found that the model was not able to predict beta-gamma pairs that lead to a decay in infection rates effectively. After inspecting the dataset it was found that only a small minority of the combinations in the training set led to a decay. To combat this imbalance in training data we implemented a 50% split in the dataset builder between pairs that cause an outbreak and those that cause a decay. These changes ultimately lead to a Validation loss of 0.013153. We then evaluated on a test set of 50 held out trajectories spanning a variety of different outcomes. Quantitatively, the model achieves a mean MSLE of 0.0169 and a mean absolute error of 122.8 individuals. These results indicate that the model generalizes reliably across epidemiological regimes.

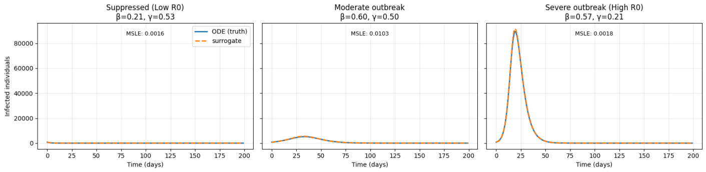


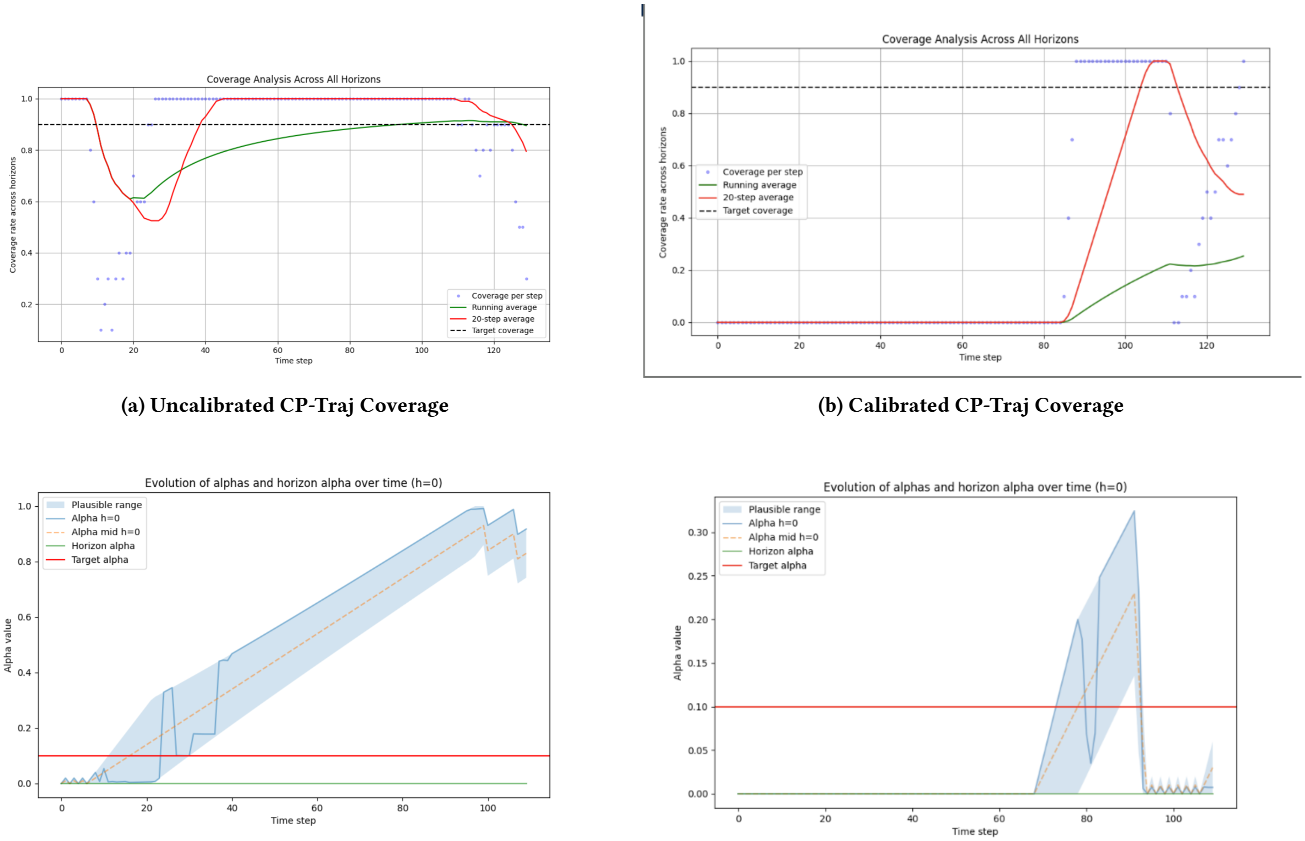
Figure 1: Surrogate model predictions compared against the simulator model solution across three different outbreak regimes.

5.4.2 *Adaptive coverage and CP-Traj.*

While the surrogate model provides accurate point predictions across epidemiological regimes, real world application requires uncertainty quantification across multiple different scenarios. For this we use Cp-Traj, at each time step and forecast horizon, CP-Traj updates the miscoverage parameters using feedback from incoming ground truth, enabling the prediction intervals to expand or contract in response.

When CP-Traj is ran without precomputing scores on the calibration set, the resulting prediction intervals produce high coverage at early timesteps with a dramatic drop as the epidemic climbs in infections. In the latter half of the infection though, the coverage is high. This success in coverage comes at the cost of a consistently increasing miscoverage parameter and growing alpha, meaning the system is becoming more uncertain over time rather than remaining properly calibrated.

To help alleviate this we run CP traj with a calibration set. The inclusion of calibration data anchors the score distribution to past regimes and the models performance and prevents the parameters from growing uncontrollably. When ran with the calibration set the coverage stays consistently near zero until the tail end of the epidemic where infections come close to zero. While this delays the availability of useful uncertainty early in the outbreak, it eliminates the deceptive early overconfidence observed in the uncalibrated setting and ensures that any future coverage has statistical basis.



(a) Uncalibrated CP-Traj Coverage

(b) Calibrated CP-Traj Coverage

(c) Uncalibrated Alpha evolution

(d) Calibrated Alpha evolution

Figure 2: Comparison of CP-Traj behavior with and without calibration. The uncalibrated method exhibits false early-time confidence and peak failure, while the calibrated variant shows a phase of zero coverage followed by delayed coverage. The corresponding alpha plots show that without calibration the alpha grows rapidly while with calibration remains consistently low until reliable behavior is observed.

5.5 Evaluation

Our approach was able to successfully create a surrogate model that is reliably able to capture the relationship between epidemiological parameters from the SIR model.The low validation loss and test performance across a suite of different beta gamma demonstrate that the model is able to accurately capture the epidemic trajectories and is able to generalize to different unseen parameter combinations.

In regards to creating a conformal prediction procedure that maintains a coverage close to the desired confidence level, we were unable to reliably achieve stable coverage across the full trajectory under different epidemiological regimes. While the uncalibrated setting exhibited high early coverage the behavior was unfounded and quickly collapsed during the rise of the epidemic. While the configuration was able to achieve coverage for most of the trajectory the accompanying skyrocketing alpha parameter indicates that this success is driven by increasingly uncertain prediction intervals and become more uninformative over time. For the calibrated CP-Traj, the miscoverage parameter remained tightly under control and avoided any illusions of coverage by the prediction intervals. This stability comes at the cost of near-zero coverage over the whole trajectory, except when the epidemic stabilize near its end. Specifically the average coverage over the trajectory we were able to achieve was around 19% This indicated that the calibration successfully prevents early confidence in the ensemble but also does not adapt to the changing epidemic dynamics for most of the trajectory.

6 CONCLUSION AND DISCUSSION

In this project we developed a neural surrogate model for SIR epidemic dynamics and examined applying conformal prediction in the form of CP-Traj to a set of plausible ensemble of simulation parameters to create uncertainty intervals as new information of the epidemic comes in. The surrogate model achieves strong performance and generalization across epidemic parameters, demonstrating that neural surrogates can serve as efficient and accurate alternatives to mechanistic methods. However our application of CP-Traj was less successful in maintaining reliable and informative uncertainty throughout the full course of an epidemic's trajectory. These results highlight the challenges in applying conformal prediction in a properly calibrated way to dynamical systems.

Future work could include experimenting with alternative ways of composing the calibration set to better capture the different sections of the epidemic. This could include increasing the size of the calibration set or having calibration sets that more exactly match the testing conditions.

REFERENCES

[1] Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Tijana Zrnic. 2022. Probabilistic Conformal Prediction. *Journal of the Royal Statistical Society: Series B* (2022).

[2] William Faris. 2021. The SIR model of an epidemic. (04 2021). https://doi.org/10.48550/arXiv.2104.12029

[3] Emmanuel Gibbs, Isaac. Candes. 2021. Adaptive Conformal Inference Under Distribution Shift. *Advances in Neural Information Processing Systems* (2021).

[4] Michael A. Johansson et al. 2021. Evaluation of the US COVID-19 Scenario Modeling Hub for Informing Pandemic Response Under Uncertainty. *Nature Communications* (2021).

[5] Reich Lab. 2021. COVID-19 Forecast Hub. https://github.com/reichlab/covid19-forecast-hub. Accessed: 2025-10-03.

[6] Ruipu Li, Daniel Menacho, and Alexander Rodríguez. 2023. Adaptive Conformal Prediction Intervals Over Trajectory Ensembles. *arXiv preprint arXiv:2306.00000* (2023).

[7] Alexander Rodriguez et al. 2021. Epidemiologically Informed Neural Networks for Pandemic Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* (2021).