SCENARIO FORECASTING FROM TIME SERIES VIA MECHANISM INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series forecasting supports timely decisions in public health, macroeconomics, and online markets. However, the same observed history can be consistent with different evolution rules, which may imply different future outcomes and preparation needs. A distribution over future trajectories alone does not make these alternatives explicit or show how predictions depend on the underlying dynamics. We introduce **mechanism-informed scenario forecasting**, which groups plausible evolution mechanisms into scenarios and predicts their probabilities and conditional future distributions. Both the scenario definitions and their number depend on the observed history. Since real histories typically lack mechanism labels, we train our proposed method, **TimeStone**, on synthetic histories paired with their generating mechanisms. Following the prior-data fitted network approach, the model learns to infer a distribution over mechanisms from a new history. It simulates futures under sampled mechanisms and groups these mechanisms into scenarios. Experiments on three real-world datasets show strong scenario quality and competitive performance in trajectory forecasting and event prediction.

1 INTRODUCTION

Time-series data capture how real-world systems evolve, from disease outbreaks and economic activity to user behavior in online markets. Time-series forecasting and event prediction use these observations to anticipate future values and events. These predictions inform planning and resource allocation, where considering plausible alternatives and their probabilities can help users prepare for changing conditions.

Existing research models uncertainty over future outcomes (Ansari et al., 2024; Halawi et al., 2024) and generates possible future trajectories (Zarifis et al., 2025). SCATTER generates diverse natural-language hypotheses about future events (Chang et al., 2026). However, diversity in predicted outcomes alone does not distinguish the alternative evolution rules that could generate them. This distinction matters because similar histories can be consistent with different rules.

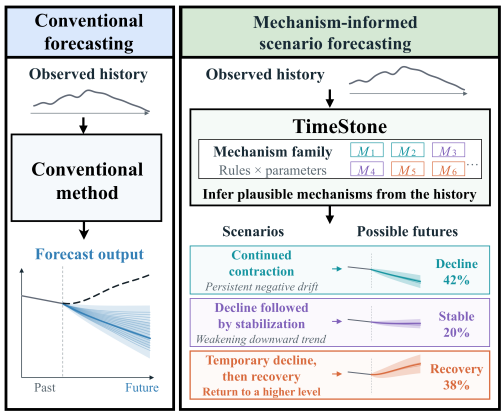


Figure 1: **One history, multiple mechanism-informed scenarios.** TimeStone links inferred mechanisms to alternative futures and scenario probabilities. Black dashes show the observed future; shading indicates predictive uncertainty.

In Figure 1, for example, declining industrial production may reflect persistent downward dynamics or a temporary deviation from a higher long-run level, implying continued decline or recovery. We call a set of evolution rules and parameter values within a domain a *mechanism* (cf. Table 1).

To describe these alternative futures together with their underlying mechanisms, we formulate **mechanism-informed scenario forecasting**. Given an observed history, the task is to construct scenarios from plausible mechanisms within a specified family and predict each scenario's probability and conditional future distribution. Each scenario groups mechanisms with similar rules and parameters.

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

We introduce **TimeStone** to address this task. Following the prior-data fitted network approach (Müller et al., 2022), we train an inference model on synthetic histories paired with their generating mechanisms. Given a new history, TimeStone infers a distribution over mechanisms, simulates their futures, and groups similar mechanisms into scenarios. Each group's probability mass determines its scenario probability. Figure 1 illustrates how mechanisms compatible with the same observed decline can support continued decline, a plateau, or a renewed increase.

We evaluate TimeStone on COVID-19 case counts, Amazon event sequences, and FRED-MD macroeconomic time series. Its scenarios more closely match observed continuations of similar histories than the evaluated baselines across all three datasets. We also compare with the COVID-19 Scenario Modeling Hub (SMH), which generates projections under expert-defined assumptions about future epidemic conditions (Loo et al., 2024; Howerton et al., 2023). Over a four-week horizon, our scenario sets achieve lower nearest-median error and higher reference coverage than SMH. Forecasting and event-probability evaluations further show competitive predictive performance.

We summarize our contributions as follows:

1. **New Task:** We formulate mechanism-informed scenario forecasting, which constructs history-dependent mechanism groups and predicts their probabilities and conditional future distributions.
2. **New Method:** We develop **TimeStone**, which learns mechanism inference from synthetic histories without requiring mechanism labels for real observations.
3. **Evaluation:** Across three datasets, TimeStone achieves lower scenario distances than five baselines and improves four-week scenario-set quality relative to SMH while maintaining competitive predictive performance.

2 Related Work

2.1 Forecasting Trajectories and Events

Probabilistic time-series forecasting models the conditional distribution $p(f \mid D)$ through likelihoods, quantiles, or sampled trajectories. DeepAR learns across related time series, while Chronos and Moirai use large-scale pretraining for zero-shot forecasting (Salinas et al., 2020; Ansari et al., 2024; Woo et al., 2024). LPTM learns transferable time-series representations through adaptive segmentation and cross-domain pretraining (Kamarthi & Prakash, 2024).

Event forecasting models event occurrence, timing, and type. Temporal point processes typically use predefined event types (Du et al., 2016; Shchur et al., 2020; Zuo et al., 2020). Open-ended methods expand the output space: ForestCast organizes news events into story trees for forecasting, while SCATTER generates diverse natural-language hypotheses about future events (Yu et al., 2025; Chang et al., 2026).

2.2 Prior Fitting and Posterior Inference

Prior-data fitted networks (PFNs) learn amortized posterior prediction from synthetic tasks sampled from a prior (Müller et al., 2022). TabPFN applies this principle to tabular classification (Hollmann et al., 2023). ForecastPFN and TimePFN extend prior fitting to univariate and multivariate forecasting, respectively (Dooley et al., 2023; Taga et al., 2025), while TabPFN-TS applies a pretrained tabular model to forecasting through temporal features (Hoo et al., 2025).

Neural posterior estimation learns parameter posterior distributions from simulated parameter–observation pairs (Papamakarios & Murray, 2016). Related temporal PFNs use structural causal priors for counterfactual prediction or causal-effect estimation (Zare et al., 2026; Thumm & Chen, 2026).

2.3 Scenario Construction

Scenario-reduction methods approximate a predictive distribution with a smaller weighted set of representative futures (Dupačová et al., 2003). Diffusion-based scenario trees organize sampled

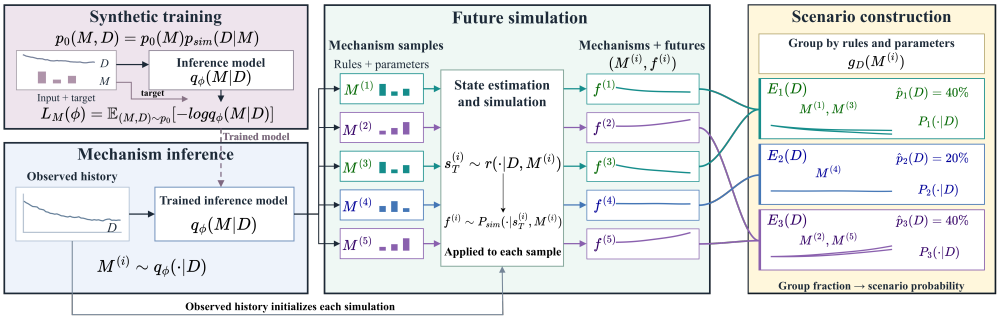


Figure 2: **Overview of TimeStone.** Colors link mechanisms to their simulated futures and scenarios. The five samples and scenario probabilities are illustrative.

trajectories into branching alternatives through clustering (Zarifis et al., 2025). Expert-guided approaches such as the COVID-19 Scenario Modeling Hub (SMH) generate projections under prespecified epidemic conditions (Loo et al., 2024; Howerton et al., 2023). Our task constructs scenarios by grouping inferred mechanisms within a specified family, associating each group with a probability and a conditional future distribution.

3 Proposed Method

Problem formulation. Let D denote the observed history up to time T and f a future trajectory over a specified horizon. A mechanism M specifies statistical evolution rules and their parameters. Given D, the task is to construct scenarios from plausible mechanisms within a specified family and estimate each scenario's probability and conditional future distribution. The output is

$$\mathcal{F}(D) = \{(E_k(D), p_k(D), P_k(\cdot \mid D))\}_{k=1}^{K(D)}, \tag{1}$$

where $E_k(D)$ is a group of mechanisms with similar rules and parameters, $p_k(D)$ is the probability that a mechanism belongs to this group given D, and $P_k(\cdot \mid D)$ is the conditional distribution of futures generated by mechanisms in the group. Both the composition and the number of scenarios $K(D)$ depend on the observed history.

3.1 Overall Workflow

Figure 2 summarizes our method in four steps:

1. **Synthetic training.** We generate histories from sampled mechanisms and train the inference model using each history as input and its generating mechanism as the target.
2. **Mechanism inference.** Given an observed history, the trained model infers a distribution over mechanisms. We sample mechanisms from this distribution.
3. **Future simulation.** For each sampled mechanism, we estimate the current state from the observed history and simulate a future.
4. **Scenario construction.** We group mechanisms with similar rules and parameters. Each group's sample fraction estimates its scenario probability, while its simulated futures approximate the conditional future distribution.

3.2 Mechanism Inference

To estimate scenario probabilities, we first need a distribution over mechanisms given the observed history. We learn this distribution, $q_\phi(M \mid D)$, from synthetic histories whose generating mechanisms are known.

Mechanism family. We define a mechanism M as a statistical generative model with specified evolution rules and parameter values. For each domain, we specify a family of such models, summarized in Table 1. We use these models to generate synthetic histories for training, as described below. Full specifications are provided in Appendix B.1.

Synthetic training pairs. To train $q_\phi(M \mid D)$, we need histories paired with their generating mechanisms. We sample a mechanism M from the prior $p_0(M)$ over the specified family, then generate a history D using the simulator under that mechanism. Empirical quantities in the prior are calibrated using only real training observations. Repeating this procedure produces synthetic training pairs (D, M), with the history as input and its generating mechanism as the target.

Table 1: Mechanism families and selected components.

Domain	Mechanism (or model) M and its construction
COVID-19	Level–trend model $y_t \sim \mathrm{NegBin}(\mu_t, r)$ M: level/trend, noise scales, dispersion
Amazon	Event transitions and waiting times $\Pr(e_{t+1} = b \mid e_t = a, M) = A_{ab}$ M: transition matrix A, waiting-time tables
FRED-MD	Dynamic factor model $\mathbf{y}_t = \Lambda \mathbf{z}_t + \mathbf{u}_t$ M: factor choices, loadings, AR coefficients, volatility

We denote the distribution of these training pairs by

$$p_0(D, M) = p_0(M)\, p_{\mathrm{sim}}(D \mid M), \tag{2}$$

where $p_{\mathrm{sim}}(D \mid M)$ is the simulator's distribution over histories. The training objective below averages over this distribution. Its conditional distribution $p_0(M \mid D)$ is the mechanism posterior that $q_\phi(M \mid D)$ is trained to approximate.

Training objective. We learn the parameters ϕ of $q_\phi(M \mid D)$ by minimizing the negative log-likelihood (NLL) of the generating mechanism in each synthetic training pair (D, M):

$$\begin{aligned} \mathcal{L}_M(\phi) &= \mathbb{E}_{(D,M)\sim p_0}\left[-\log q_\phi(M \mid D)\right] \\ &= -\mathbb{E}_{(D,M)\sim p_0}\left[\log p_0(M \mid D)\right] + \mathbb{E}_{D\sim p_0(D)}\left[D_{\mathrm{KL}}\left(p_0(\cdot \mid D) \,\|\, q_\phi(\cdot \mid D)\right)\right]. \end{aligned} \tag{3}$$

Here, $p_0(M \mid D)$ is the mechanism posterior under the synthetic training distribution. Only the KL term depends on ϕ, so minimizing the expected NLL minimizes the expected KL divergence from this posterior to $q_\phi(M \mid D)$ (Proposition 1). Training evaluates the NLL on synthetic pairs and does not require computing $p_0(M \mid D)$.

PFNs use the same conditional-likelihood principle for posterior prediction (Müller et al., 2022). Here, the prediction target is the generating mechanism.

Inference. Given a new history D, the trained model outputs a mechanism distribution $q_\phi(M \mid D)$. To simulate futures under this distribution, we sample N mechanisms:

$$M^{(i)} \sim q_\phi(\cdot \mid D), \qquad i = 1, \ldots, N. \tag{4}$$

Each sample specifies the evolution rules and parameter values used for future simulation.

3.3 Future Simulation

For each sampled mechanism $M^{(i)}$, we generate a future trajectory $f^{(i)}$ over the forecast horizon by simulating its evolution rules.

State estimation and simulation. Simulation requires a starting state s_T at the end of the observed history. We estimate this state from D under $M^{(i)}$ and denote the resulting distribution by $r(s_T \mid D, M^{(i)})$, allowing a fixed state as a special case. We then sample a starting state and simulate forward:

$$\begin{aligned} s_T^{(i)} &\sim r(\cdot \mid D, M^{(i)}), \\ f^{(i)} &\sim P_{\mathrm{sim}}(\cdot \mid s_T^{(i)}, M^{(i)}). \end{aligned} \tag{5}$$

Here, P_{sim} is the simulator's conditional distribution over future trajectories. The mechanism remains fixed throughout simulation. We retain each pair $(M^{(i)}, f^{(i)})$ for scenario construction.

3.4 SCENARIO CONSTRUCTION

Grouping mechanisms. We group mechanisms sampled from $q_\phi(M \mid D)$ according to similarity in their rules and parameters. Representatives are selected from these samples, and each mechanism is assigned to its nearest representative using weighted Euclidean distance. We denote the resulting mechanism groups by $E_k(D)$. We select the number of groups $K(D)$ using a separate set of mechanism samples, balancing their distance to the nearest representatives against a penalty for additional groups. Feature scaling and selection details are provided in Appendix B.4.

Scenario probabilities and futures. We estimate each group's probability $p_k(D)$ as the fraction of fresh mechanism samples that belong to $E_k(D)$. These samples are drawn from $q_\phi(M \mid D)$ independently of those used to construct the groups:

$$\widehat{p}_k(D) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{M^{(i)} \in E_k(D)\}, \qquad M^{(i)} \sim q_\phi(\cdot \mid D). \tag{6}$$

We simulate futures for the fresh mechanisms using Equation 5. These futures estimate each group's conditional distribution $P_k(\cdot \mid D)$, completing the scenario output $\mathcal{F}(D)$.

3.5 PREDICTIVE ADAPTATION

The procedure above defines the zero-shot variant of TimeStone. The adapted variant adds an optional training stage using real history–future pairs to adjust mechanism probabilities.

We keep the inference model and simulator fixed and train only a weighting network. Its parameters ω determine positive weights $w_\omega(D, M)$, giving the adapted mechanism distribution

$$q_{\phi,\omega}(M \mid D) = \frac{q_\phi(M \mid D)\, w_\omega(D, M)}{\mathbb{E}_{M' \sim q_\phi(\cdot \mid D)}[w_\omega(D, M')]}. \tag{7}$$

Using $q_{\phi,\omega}$ in the future simulation procedure gives the adapted predictive distribution $\widehat{P}_{\phi,\omega}(\cdot \mid D)$. We learn ω by comparing this distribution with the observed future f in each real training pair (D, f):

$$\begin{aligned}\mathcal{L}_{\text{adapt}}(\omega) = \mathbb{E}_{(D,f)\sim\text{train}}\big[\,&\mathrm{ES}\big(\widehat{P}_{\phi,\omega}(\cdot \mid D), f\big) \\ &+ \lambda_a D_{\mathrm{KL}}(q_{\phi,\omega}(\cdot \mid D) \,\|\, q_\phi(\cdot \mid D))\big].\end{aligned} \tag{8}$$

Here, the energy score (ES) evaluates the predicted future distribution against the observed future (Gneiting & Raftery, 2007). The KL term penalizes deviations from the base mechanism distribution, with strength controlled by λ_a.

Adaptation requires observed futures but no mechanism labels. The adapted distribution is used with the same future simulation and scenario construction procedures. The energy score definition, finite-sample estimation, and checkpoint selection are provided in Appendix E.1 and B.5.

4 EXPERIMENTS

Given an observed history, mechanism-informed scenario forecasting constructs scenarios from plausible mechanisms and predicts their probabilities and conditional future distributions. Real datasets provide no mechanism labels and only one observed future per history. We therefore evaluate two aspects:

1. **Scenario quality.** Do generated futures agree with observed examples, and do scenario representatives cover different observed alternatives? We use observed continuations of similar histories as reference futures.
2. **Predictive performance.** Does the full predictive distribution accurately predict the observed future and its event category? We evaluate distributional forecasts using ES and event predictions using Event-NLL and accuracy. Event categories are shared across methods and are distinct from mechanism groups.

Table 2: **Scenario quality.** Each method is evaluated on up to three scenarios with the highest positive probabilities. Best values are shown in bold pink and second-best values in green. ‡ marks methods adapted to generate numerical future trajectories.

	COVID-19		Amazon		FRED-MD	
Method	**Scenario dist.↓**	**Coverage error↓**	**Scenario dist.↓**	**Coverage error↓**	**Scenario dist.↓**	**Coverage error↓**
Chronos-T5	1.5120	1.7165	1.0710	1.1653	1.1650	1.5616
Moirai	6.2559	2.2038	1.5440	1.3025	4.5292	2.0032
LPTM (adapted, 4 comp.)‡	2.7077	1.9156	1.3293	1.1909	1.7679	1.9101
SCATTER (adapted)‡	1.5259	2.0588	1.0632	1.1232	1.5521	1.7539
ForestCast (adapted)‡	1.7835	2.1834	1.0717	1.1235	1.4616	1.7259
TimeStone (zero-shot)	1.0493	1.7088	**1.0253**	**1.1108**	1.0736	**1.2452**
TimeStone (adapted)	**1.0365**	**1.6749**	1.0260	1.1579	**1.0693**	1.2629

We evaluate our method on COVID-19 case counts, Amazon event sequences, and FRED-MD macroeconomic time series. We also compare our scenario sets with expert-designed SMH scenarios and use ablations to examine the contributions of mechanism inference and scenario construction. Experimental settings are provided in Appendix C.

4.1 Scenario Quality

We evaluate scenario quality against observed continuations of similar histories, which serve as reference futures. For each test history, we retrieve reference cases using history features alone and use their observed futures as a shared reference set for all methods. These references provide observed examples of alternative developments following similar histories. They are used only for evaluation, not for model training or scenario construction. See Appendix D.1 for details.

For all baselines, we use the same procedure to group similar predicted trajectories and select one trajectory from each group as its representative. For SCATTER, ForestCast, and LPTM (‡), we train probabilistic decoders on real training history–future pairs to generate future trajectories from pretrained history representations. For LPTM, we freeze the pretrained backbone and evaluate four-component heads that also use numerical history features and point forecasts.

We adapt average nearest-neighbor set distances (Schutze et al., 2012) to assess two complementary properties. *Scenario distance* measures how closely generated futures agree with the reference futures. *Coverage error* measures how well scenario representatives cover the alternatives present in those references. Together, these metrics assess agreement with observed examples and coverage of their variation. Both use Euclidean distances in a shared future-feature space:

$$\begin{aligned} \text{Scenario dist.} &= \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\frac{1}{|X_k|}\sum_{f\in X_k}\min_{r\in R}\|f-r\|_2, \\ \text{Coverage error} &= \frac{1}{|R|}\sum_{r\in R}\min_{k\in\mathcal{K}}\|r-c_k\|_2. \end{aligned} \tag{9}$$

Here, R is the reference set, X_k contains the generated futures assigned to scenario k, and c_k is its representative trajectory. The set $\mathcal{K}$ indexes the evaluated scenarios. For scenarios without evaluation samples, we use the representative alone, setting $X_k = \{c_k\}$. Both metrics are computed per test history and averaged over the test set.

Results. Both TimeStone variants achieve lower scenario distances than all evaluated baselines across the three datasets (Table 2). The zero-shot model also achieves lower coverage errors than every baseline on all three datasets. Together, these results indicate that its scenario sets better represent the alternative continuations observed after similar histories: generated futures remain closer to observed examples, while scenario representatives better cover the variation among those examples.

Table 3: **Predictive performance.** Best and second-best values within each training setting are shown in bold pink and green, respectively. ‡ denotes methods extended to generate numerical future trajectories.

	COVID-19			Amazon			FRED-MD		
Method	**ES↓**	**Event-NLL↓**	**Accuracy↑**	**ES↓**	**Event-NLL↓**	**Accuracy↑**	**ES↓**	**Event-NLL↓**	**Accuracy↑**
Without target-domain predictive training									
Chronos-T5	0.601	1.354	0.391	0.648	0.442	0.871	2.314	1.102	0.807
Moirai	0.784	1.706	0.335	0.679	0.417	0.873	2.394	0.839	0.842
ForecastPFN	1.867	5.091	0.263	0.682	0.875	0.872	2.561	2.363	0.658
TimePFN	2.310	4.322	0.374	0.677	0.877	0.872	2.529	1.272	0.816
TabPFN-TS	0.586	2.675	0.399	0.645	0.692	0.867	2.421	0.772	0.833
SCATTER‡	1.166	1.940	0.330	0.754	0.831	0.863	2.365	1.248	0.696
ForestCast‡	0.962	1.724	0.328	0.758	0.793	0.861	2.295	1.323	0.627
LPTM	1.133	4.977	0.279	0.752	0.868	0.873	2.615	1.454	0.789
TimeStone (zero-shot)	**0.578**	**1.197**	**0.448**	**0.596**	**0.300**	**0.878**	**2.282**	**0.724**	**0.853**
With target-domain predictive training									
RMTPP	0.593	1.206	0.424	0.608	0.352	0.874	2.562	0.691	0.810
IFTPP	0.603	1.316	0.352	0.602	0.307	0.874	2.742	0.837	0.826
THP	0.621	1.230	0.413	0.613	0.373	0.874	2.619	0.625	0.846
DeTPP	0.585	1.202	0.441	0.617	0.388	0.874	2.566	0.707	0.817
DeepCOVID	3.869	2.230	0.333	0.616	0.334	0.877	2.858	0.749	0.851
SCATTER (adapted)‡	0.803	1.513	0.410	0.602	0.319	0.877	2.353	1.412	0.468
ForestCast (adapted)‡	0.888	1.667	0.362	0.606	0.330	0.874	2.346	1.482	0.557
TimeStone (adapted)	**0.577**	**1.199**	**0.443**	**0.596**	**0.300**	**0.878**	**2.278**	**0.614**	**0.848**

4.2 Predictive Performance

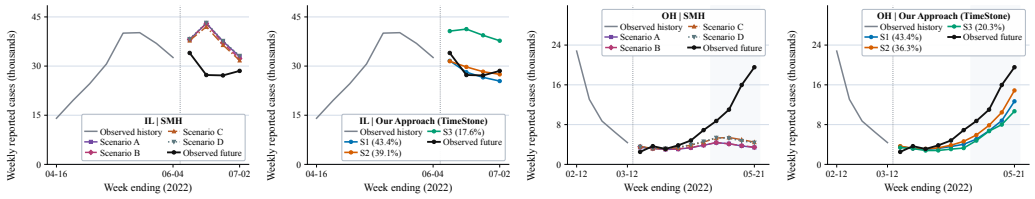
We also evaluate whether TimeStone supports accurate forecasting and event prediction in addition to scenario construction. The energy score (ES) assesses the full predictive distribution against the observed future (Gneiting & Raftery, 2007). For event prediction, we estimate category probabilities from predictive sample fractions using shared categories of future observations, distinct from each method's scenarios. Event negative log-likelihood (Event-NLL) evaluates the probability assigned to the observed category, while accuracy measures whether the highest-probability category is correct. Evaluation details and prediction-set results are provided in Appendix E.

Results. TimeStone achieves the lowest ES and Event-NLL within each training setting across all three datasets (Table 3). The zero-shot model also achieves the highest accuracy across all three datasets, while the adapted model does so on COVID-19 and Amazon. These results show that the predictive distributions underlying its scenarios support accurate forecasting and event prediction. On FRED-MD, adaptation further reduces Event-NLL from 0.724 to 0.614, while accuracy decreases from 0.853 to 0.848, indicating improved event probability estimates with a small trade-off in classification accuracy.

4.3 Comparison with the Scenario Modeling Hub

The COVID-19 Scenario Modeling Hub (SMH) combines epidemiological projections under expert-defined vaccination and variant assumptions to support longer-term planning (Howerton et al., 2023). We compare its published scenario sets with TimeStone's to assess agreement with realized futures and coverage of observed alternatives. TimeStone constructs its scenarios from observed histories without prescribing future conditions for each scenario.

Evaluation. We evaluate four-week futures across 47 US jurisdictions at two forecast origins (94 cases), comparing SMH's four scenarios with up to four TimeStone scenarios ranked by inferred probability. The defining booster and variant changes begin beyond this window, so the comparison assesses short-term scenario sets (COVID-19 Scenario Modeling Hub, 2022a;b). Since SMH does not assign occurrence probabilities to its scenarios, we evaluate their medians as a set. Nearest-median RMSE measures the minimum RMSE between the observed future and any scenario median, selected retrospectively. Reference coverage measures the fraction of 32 observed continuations



(a) Illinois (SMH) (b) Illinois (TimeStone) (c) Ohio (SMH) (d) Ohio (TimeStone)

Figure 3: **Scenario examples from SMH and TimeStone (zero-shot).** Figure (a),(b) show Illinois over four weeks, and Figure (c),(d) show Ohio over ten weeks. Colored curves show scenario medians, gray curves show observed histories, and black curves show observed futures. Parentheses indicate TimeStone's scenario probabilities.

from similar training histories matched by at least one scenario median. Evaluation details are provided in Appendix F.

Results. TimeStone (zero-shot) reduces nearest-median RMSE by 31.9% and increases reference coverage from 14.89% to 26.51% (Table 4). Its scenario sets therefore include candidates closer to realized futures while covering more observed alternatives over the evaluated four-week horizon. Compared with SMH's projections under prespecified future conditions, TimeStone's inferred scenario sets include candidates closer to realized futures and cover more observed alternatives over the evaluated four-week horizon.

Table 4: **Four-week scenario-set comparison with SMH.** Results are averaged over 94 cases across 47 US jurisdictions and two forecast origins.

Method	Nearest-median RMSE↓	Reference coverage (%)↑
SMH	1.2033	14.89
TimeStone (zero-shot)	0.8191	26.51
TimeStone (adapted)	0.9596	25.12

Visualization. Figure 3 shows four-week Illinois and ten-week Ohio examples. In Illinois, TimeStone's two most probable scenarios follow the observed decline and subsequent lower case counts, while a third retains a higher-case alternative. In Ohio, TimeStone's scenarios represent the later rebound, although they underestimate its magnitude, while the SMH medians remain at lower levels. The Ohio example illustrates behavior over a longer horizon and is separate from the four-week quantitative comparison. Additional visualizations are provided in Appendix F.4.

5 Analysis and Ablation Studies

We examine whether mechanism inference is informative, whether mechanism groups distinguish future distributions, and how the results depend on the prior.

5.1 Mechanism Inference

We first test whether inference provides more precise and accurate information about the generating mechanisms using 256 synthetic histories per domain with known mechanisms (Table 5(a)). On real data, we replace the inferred mechanism distribution with the prior or a distribution inferred from another history. All variants use the query history for state estimation and generate 256 futures. History-conditioned inference achieves lower ES than both controls across all three datasets, with all 95% confidence intervals for the ES reductions

Table 5: **Mechanism inference.** (a) Mean marginal coverage and median parameter-wise posterior/prior ratios. (b) Control-minus-zero-shot ES. All ES reductions have 95% CIs above zero. Definitions and intervals are provided in Appendix G.1.

	COVID-19	**Amazon**	**FRED-MD**
(a) Synthetic mechanism inference			
90% coverage (%)	91.17	88.83	89.71
Width / prior	0.670	0.288	0.672
RMSE / prior	0.746	0.499	0.652
(b) Real-data ES reduction			
Prior	0.1728	0.0253	0.3255
Other history	0.2174	0.0579	0.3654

Table 6: **Effect of mechanism grouping.** Entries are mechanism-group scores minus control scores, with 95% confidence intervals. Negative values favor mechanism grouping. The random and trajectory controls use separate evaluation subsets.

	Random grouping	Trajectory grouping	
Dataset	ΔConditional ES	ΔScenario distance	ΔCoverage error
COVID-19	-0.024065 [-0.029212, -0.019647]	-0.2285 [-0.2954, -0.1765]	$+0.0038$ [0.00302, 0.0046]
Amazon	-0.002986 [-0.003468, -0.002574]	-0.0070 [-0.0099, -0.0041]	$+0.0045$ [0.0033, 0.0059]
FRED-MD	-0.000873 [-0.001554, -0.000279]	-0.0662 [-0.1081, -0.0320]	$+0.0083$ [-0.0003, 0.0177]

Table 7: **Sensitivity to prior and family choices.** (a) Score differences are variant minus baseline, with 95% confidence intervals. Negative values favor the variant. (b) TV compares probability vectors over fixed mechanism regions. Ranges span the tested prior variants.

(a) Predictive performance

Dataset	**Setting**	ΔES	ΔEvent-NLL
COVID-19	Prior variant 1	-0.02097 [-0.03109, -0.01143]	-0.01416 [-0.03035, 0.00170]
	Prior variant 2	-0.00843 [-0.01778, 0.00077]	-0.00609 [-0.01831, 0.00603]
Amazon	Prior variant 1	-0.00118 [-0.00251, 0.00020]	$+0.00029$ [-0.00451, 0.00473]
	Prior variant 2	-0.00172 [-0.00337, -0.00012]	-0.00570 [-0.01217, -0.00010]

(b) Changes in mechanism-region probabilities

Dataset	**Prior change (TV)**	**Resampling (TV)**
COVID-19	0.124–0.130	0.017
Amazon	0.191–0.192	0.020

above zero (Table 5(b)). These results support the predictive value of inferring mechanisms from the query history, beyond using that history to initialize the simulator.

5.2 Scenario Construction

On independent simulated futures with known generating mechanisms, mechanism groups achieve lower conditional ES than random groups matched in number and size, supporting predictive information in group assignments under the specified simulator (Table 6). Using the same future samples and matched displayed scenario counts, mechanism grouping lowers scenario distance relative to trajectory grouping across all three datasets. These results support mechanism-based grouping as an informative way to organize predictive uncertainty. The resulting scenarios connect alternative future distributions to explicit evolution rules and parameters, making inferred dynamics part of the forecast. Protocols are provided in Appendix G.2.

5.3 Prior Sensitivity

We retrain prior variants on COVID-19 and Amazon under matched budgets, using training data for prior calibration and development data for configuration selection. Predictive effects vary by setting, while probability shifts over fixed mechanism regions, measured by total variation (TV), exceed those from resampling the same posterior (Table 7). These shifts indicate prior sensitivity beyond Monte Carlo variation. Protocols are provided in Appendix G.3.

6 Conclusion

We introduced **mechanism-informed scenario forecasting**, which connects alternative futures to the evolution rules and parameters that may generate them. TimeStone learns mechanism inference from synthetic histories and uses the inferred distribution to define scenarios, their probabilities, and their conditional future distributions. Experiments across three domains show strong predictive performance and coverage of observed alternatives, while controlled analyses support the predictive value of history-conditioned mechanism inference. This framework makes inferred dynamics part of the forecast, describing both how a system may evolve and the modeling assumptions behind those possible futures.

Ethics statement. TimeStone is a research framework for exploring plausible future scenarios. Its outputs depend on the specified mechanism family and available data and may be inaccurate under changing conditions. Public-health applications, including COVID-19 forecasting, require further validation, expert review, and clear communication of uncertainty.

Reproducibility Statement. To ensure the reproducibility and completeness of this paper, we make our code available at `https://anonymous.4open.science/r/TimeStone-DFE4`. We give details on our experimental protocol in the Appendix C.

The Use of Large Language Models (LLMs). We used ChatGPT as a writing assistant for polishing language and checking notation. No research ideas or content generation were conducted by LLMs.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

He Chang, Zhulin Tao, Lifang Yang, Xianglin Huang, and Yunshan Ma. Scattered hypothesis generation for open-ended event forecasting. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17288–17304, San Diego, California, United States, July 2026. Association for Computational Linguistics.

COVID-19 Scenario Modeling Hub. Round 14 Scenarios, 2022a. URL https://github.com/midas-network/covid19-scenario-modeling-hub_archive/blob/master/previous-rounds/README_Round14.md. Scenario specifications. Last updated May 16, 2022.

COVID-19 Scenario Modeling Hub. Round 15 Scenarios, 2022b. URL https://github.com/midas-network/covid19-scenario-modeling-hub_archive/blob/master/previous-rounds/README_Round15.md. Scenario specifications. Last updated July 22, 2022.

Samuel Dooley, Gurnoor Singh Khurana, Chirag Mohapatra, Siddartha V. Naidu, and Colin White. ForecastPFN: Synthetically-trained zero-shot forecasting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 2403–2426, 2023.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1555–1564. Association for Computing Machinery, 2016.

Jitka Dupačová, Nicole Gröwe-Kuska, and Werner Römisch. Scenario reduction in stochastic programming: An approach using probability metrics. *Mathematical Programming*, 95(3):493–511, 2003.

Andrew Foong, Wessel Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, and Richard Turner. Meta-learning stationary stochastic process prediction with convolutional neural processes. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8284–8295, 2020. URL https://papers.nips.cc/paper/2020/hash/5df0385cba256a135be596dbe28fa7aa-Abstract.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 2007. doi: 10.1198/016214506000001437. URL https://sites.stat.washington.edu/raftery/Research/PDF/Gneiting2007jasa.pdf.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. *Advances in Neural Information Processing Systems*, 37:50426–50468, 2024.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From tables to time: Extending TabPFN-v2 to time series forecasting. *arXiv preprint arXiv:2501.02945*, 2025. URL https://arxiv.org/abs/2501.02945.

Emily Howerton, Lucie Contamin, Luke C Mullany, Michelle Qin, Nicholas G Reich, Samantha Bents, Rebecca K Borchering, Sung-mok Jung, Sara L Loo, Claire P Smith, et al. Evaluation of the us covid-19 scenario modeling hub for informing pandemic response under uncertainty. *Nature communications*, 14(1):7260, 2023.

Harshavardhan Kamarthi and B. Aditya Prakash. Large pre-trained time series models for cross-domain time series analysis tasks. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2311.11413.

Sara L Loo, Emily Howerton, Lucie Contamin, Claire P Smith, Rebecca K Borchering, Luke C Mullany, Samantha Bents, Erica Carcelen, Sung-mok Jung, Tiffany Bogich, et al. The us covid-19 and influenza scenario modeling hubs: Delivering long-term projections to guide policy. *Epidemics*, 46:100738, 2024.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022.

George Papamakarios and Iain Murray. Fast ϵ-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, volume 29, pp. 1028–1036, 2016. URL https://arxiv.org/abs/1605.06376.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3):1181–1191, 2020.

Oliver Schutze, Xavier Esquivel, Adriana Lara, and Carlos A Coello Coello. Using the averaged hausdorff distance as a performance measure in evolutionary multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, 16(4):504–522, 2012.

Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. Intensity-free learning of temporal point processes. In *International Conference on Learning Representations*, 2020.

Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. TimePFN: Effective multivariate time series forecasting with synthetic data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20761–20769, 2025.

Dennis Thumm and Ying Chen. Interventional time series priors for causal foundation models, 2026.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR, 2024.

Zi Yu, Shaoxiang Wang, Guozheng Li, Yu Zhang, and Chi Harold Liu. ForestCast: Open-ended event forecasting with semantic news forest. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12667–12681, Suzhou, China, November 2025. Association for Computational Linguistics.

Amirhossein Zare, Amirhessam Zare, Herlock Rahimi, Reza Salarikia, and Mohammad Kashkooli. Causal longitudinal prior-fitted networks for counterfactual outcome prediction, 2026.

Stelios Zarifis, Ioannis Kordonis, and Petros Maragos. Diffusion-based scenario tree generation for multivariate time series prediction and multistage stochastic optimization, 2025. URL https://arxiv.org/abs/2509.14832.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer Hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11692–11702. PMLR, 2020.

We introduce **TimeStone**[1], a framework for mechanism-informed scenario forecasting.

A Inference Model and Synthetic Training

TimeStone implements the mechanism inference procedure in Section 3.2 using a history representation and a mechanism posterior decoder. The domain-specific implementations differ in their observation representations and posterior families, while sharing the mechanism negative log-likelihood (NLL) objective.

A.1 Bayesian Interpretation of Mechanism Inference

We apply the standard conditional likelihood argument (Foong et al., 2020; Müller et al., 2022) to mechanism inference. The following proposition shows that our NLL objective targets the mechanism posterior induced by the prior and simulator.

Proposition 1 (Mechanism posterior approximation)**.** *Fix the prior and simulator, and let training pairs follow $p_0(M, D) = p_0(M)p_{\mathrm{sim}}(D \mid M)$, with marginal $p_0(D)$ and posterior $p_0(M \mid D)$. Let $q_\phi(M \mid D)$ be a normalized conditional density. Assume that $p_0(\cdot \mid D)$ and $q_\phi(\cdot \mid D)$ are defined with respect to a common reference measure, and that $\log p_0(M \mid D)$ and $\log q_\phi(M \mid D)$ are integrable under $p_0(M, D)$. Then the loss in Equation 3 satisfies*

$$\mathcal{L}_M(\phi) = C + \mathbb{E}_{D\sim p_0(D)}\left[D_{\mathrm{KL}}\left(p_0(\cdot \mid D) \,\|\, q_\phi(\cdot \mid D)\right)\right], \tag{10}$$

where $C = -\mathbb{E}_{(M,D)\sim p_0}[\log p_0(M \mid D)]$ is independent of ϕ. If the model class contains this posterior, every global minimizer $\phi^\star$ of the expected NLL satisfies $q_{\phi^\star}(\cdot \mid D) = p_0(\cdot \mid D)$ for $p_0(D)$-almost every history.

Proof. Adding and subtracting $\log p_0(M \mid D)$ gives

$$\begin{aligned}\mathcal{L}_M(\phi) = &-\mathbb{E}_{(M,D)\sim p_0}\left[\log p_0(M \mid D)\right] \\ &+ \mathbb{E}_{(M,D)\sim p_0}\left[\log \frac{p_0(M \mid D)}{q_\phi(M \mid D)}\right].\end{aligned}$$

Conditioning the second expectation on D yields Equation 10. Since KL divergence is nonnegative, $\mathcal{L}_M(\phi) \geq C$. If the model class contains the posterior, this lower bound is attained. Every global minimizer must therefore have zero expected KL divergence. Nonnegativity then implies that the conditional KL divergence is zero for $p_0(D)$-almost every history, giving $q_{\phi^\star}(\cdot \mid D) = p_0(\cdot \mid D)$. □

Training uses sampled (M, D) pairs and does not require evaluating $p_0(M \mid D)$. The proposition characterizes approximation on average over histories generated by the specified prior and simulator. Exact posterior recovery requires that the model can represent the posterior and reaches a global minimum of the expected NLL. Performance with finite training data and real observations is assessed empirically.

B Method Details

This appendix summarizes the implementation of mechanism inference, future simulation, scenario construction, and predictive adaptation in Section 3.

B.1 Mechanism Families and Priors

We instantiate Section 3.2 with a domain-specific prior $p_0(M)$ and observation simulator $p_{\mathrm{sim}}(D \mid M)$. Each synthetic episode uses one sampled mechanism; initial states and process and observation noise vary across realizations. Empirical prior quantities are calibrated from the real TRAIN split and then fixed during posterior training. The resulting families permit continuous parameter variation and combinations of components.

[1]The name is inspired by Doctor Strange's Time Stone and the idea of exploring multiple possible futures.

B.1.1 COVID-19

The COVID-19 prior models weekly reported counts. Each mechanism is $M = (c, s, r, \rho, \kappa, \varphi)$, representing the latent log-level center, stationary level scale, count-dispersion shape, velocity persistence, restoring coefficient, and initial phase. The two-dimensional latent state $x_t = (u_t, v_t)^\top$ evolves as

$$x_{t+1} = A(M)x_t + \sigma_M \epsilon_t \mathbf{1}, \qquad A(M) = \begin{pmatrix} 1-\kappa & \rho \\ -\kappa & \rho \end{pmatrix}, \qquad \epsilon_t \sim \mathcal{N}(0,1). \tag{11}$$

Let C solve $C = ACA^\top + \mathbf{1}\mathbf{1}^\top$ and set $\sigma_M = s/\sqrt{C_{11}}$. The initial state is Gaussian with covariance $\sigma_M^2 C$, with mean $(-\varphi s/2,\ \varphi\sigma_M\sqrt{C_{22}}/2)^\top$. Writing $z_t = c + u_t$, the observation mean is

$$\mu_t = \begin{cases} \exp(z_t), & z_t \le \log B, \\ B\left[1 + (z_t - \log B)/p\right]^p, & z_t > \log B, \end{cases} \qquad B = 1{,}146{,}200, \quad p = 2. \tag{12}$$

Counts are sampled by $\lambda_t \sim \mathrm{Gamma}(r, r/\mu_t)$, with the second argument denoting rate, followed by $y_t \sim \mathrm{Poisson}(\lambda_t)$. Thus $\mathrm{Var}(y_t \mid x_t, M) = \mu_t + \mu_t^2/r$. The mean link is positive and unbounded; B is a transition scale, not a clipping threshold.

The prior uses 68 equally weighted templates from two nonoverlapping 24-week windows in each of 34 calibration regions. The actual template observations span March 7, 2020–January 30, 2021. Each template provides anchors (c_j, s_j, r_j). Conditional on a uniformly sampled template j, the coefficient draws are independent with

$$c = c_j + U[-0.15, 0.15], \qquad s = s_j \exp(U[-0.15, 0.15]), \tag{13}$$

$$r = 2r_j \exp(U[-0.25, 0.25]), \qquad \rho \sim U[0.875, 0.95], \tag{14}$$

$$\kappa = 0.04 \exp(U[-0.15, 0.15]), \qquad \Pr(\varphi = -1, 0, 1) = (0.3, 0.4, 0.3). \tag{15}$$

Here each occurrence of U is an independent draw. Template anchors are calibrated from smoothed log counts and moment summaries. The link scale is four times the largest eligible calibration count. The observed input consists of reported counts, and the mechanism parameters describe their aggregate statistical dynamics.

B.1.2 Amazon

The Amazon prior generates event times and marks from 16 categories. A mechanism contains a 16×16 mark-transition table A, a 6×2 short/long waiting-band gate G, and a 2×10 within-band gap-bin table Q. Conditional on a prior component j, table rows are independent Dirichlet draws. Row concentrations are 80 for A, 5 for G, and 80 for Q. The prior is a uniform mixture of 5,143 such components, whose row centers are calibrated from prefixes of official TRAIN sequences. These components specify continuous distributions over probability tables, rather than a finite catalogue of complete mechanisms.

Starting from mark ℓ_t, the next mark is drawn from $A_{\ell_t,:}$. The gate row is indexed by whether the mark repeats and by the previous waiting band: $a_t = \mathbf{1}[\ell_{t+1} = \ell_t] + 2b_{t-1}$. We sample $b_t \sim \mathrm{Cat}(G_{a_t,:})$, a gap bin from $Q_{b_t,:}$, and then a waiting time within that bin. The waiting bands are $b_t \in \{0, 1\}$. Synthetic sequences use a uniform initial mark and the startup-band sentinel $b_{-1} = 2$. The initial mark is conditioned on when computing mechanism evidence. All gap bins except the final long-gap bin use uniform densities; the final long-gap bin uses a shifted exponential tail. The short/long threshold is 0.0965 in native event-time units, and the tail begins at 0.79999 with exponential scale 0.099991. Bin boundaries are fixed from TRAIN calibration.

Of the 6,454 official TRAIN sequences, a deterministic sequence-index hash split assigns 5,143 to calibration and 1,311 to TRAIN-internal development. Each calibration sequence contributes its first $\max(2, \lfloor 0.6n \rfloor)$ events. Row centers use pooled smoothing and a calibrated repeat-propensity adjustment. Synthetic histories have a uniformly sampled observed length from 12 through 60 events. Real forecasting uses at most the last 60 events and predicts the next six recorded events; this evaluation is conditional on those six future events being recorded.

B.1.3 FRED-MD

The FRED-MD prior combines dynamic factors, autoregressive residuals, and stochastic volatility. It uses 18 transformed monthly indicators and six calibration templates, each fitted on a 60-month

block within January 1960–December 1989. Templates are estimated using a global principal component, sector-residual components, autoregressive moments, residual variances, and housing trend scales. We apply $\Delta \log$ to RPI, W875RX1, INDPRO, PAYEMS, and CLAIMSx; Δ to UNRATE, FEDFUNDS, TB3MS, GS10, and AAA; $\Delta^2 \log$ to CPIAUCSL, PCEPI, PPICMM, M2SL, BUSLOANS, and NONREVSL; and $\log$ to HOUST and PERMIT. The data are a revised June-2025 snapshot; the experiment is retrospective rather than a real-time-vintage evaluation.

A mechanism combines a global factor with optional real-activity, price, housing, interest-rate, and money/credit factors. The global factor is always present and each other factor has prior activation probability 0.85. A base template b is uniform over the six templates. Each active factor selects template j with probability $0.8\mathbf{1}[j = b] + 0.2/6$. Its two partial-autocorrelation coordinates are sampled as $k = \tanh(\operatorname{atanh} \widehat{k} + 0.2Z)$ and its amplitude as $\exp(0.2Z)$, with independent standard-normal draws. The resulting factor transition is AR(2), with coefficients $k_1(1-k_2)$ and k_2. Residual AR(1) coefficients use Fisher-coordinate perturbation scale 0.15, and the common residual amplitude has log scale 0.2. The baseline residual variance v is split into latent residual variance $0.9v$ and measurement variance $0.1v$. Before stochastic-volatility scaling, the residual AR(1) innovation variance is $0.9v(1-\rho^2)$. The prior also includes a shared housing random-walk component with scale $\sigma_{\mathrm{RW}} = \widehat{\sigma}_{\mathrm{RW},b} \exp(0.2Z)$ and a housing-factor lag of 0, 1, or 2 months with probabilities $(0.5, 0.25, 0.25)$ when that factor is active.

The common log-volatility process has persistence 0.9 and stationary standard deviation 0.7. Four additional contrast coordinates induce relative volatility across five sector blocks through a fixed orthonormal Helmert basis. Their persistence is $0.5 + 0.495U$, $U \sim \mathrm{Beta}(8, 2)$, and their stationary standard deviation is $0.6 \exp(0.35Z)$. Conditional on the volatility paths, observations follow a linear-Gaussian state-space model.

B.2 Mechanism Inference and Training

Encoder and posterior. COVID-19 and FRED-MD use separately trained instances of the same encoder: a width-192 input projection, a one-layer bidirectional GRU with 192 units per direction, and two Transformer layers with four attention heads and feed-forward width 384. Attention combines one full-history token and four temporal-segment tokens. A parallel MLP encodes means, standard deviations, endpoints, covariances, and lagged covariances; its output is merged with the sequence representation. Inputs contain normalized observations, masks, and temporal information; COVID-19 counts are log-transformed.

The COVID-19 head is a four-component mixture with categorical template and phase variables and five transformed Beta coordinates. Compatible templates are marginalized in density evaluation. The FRED-MD head combines categorical and Bernoulli structural choices with conditional Gaussian and Beta parameter distributions. Both support normalized density evaluation with transformation Jacobians.

Amazon uses 312 count statistics and a shared additive spline with 5,664 trainable coefficients. Learned component weights are combined with analytic Dirichlet updates:

$$q_\phi(M \mid D) = \sum_{j=1}^{5143} w_{\phi,j}(D) \prod_r \mathrm{Dir}(M_r \mid \alpha_{j,r} + n_r(D)) \,. \tag{16}$$

Here $n_r(D)$ contains observed table-row counts. The component weights are learned; the conditional concentration updates are analytic.

Training and selection. All base models minimize synthetic mechanism NLL, $\mathbb{E}[-\log q_\phi(M \mid D)]$, without a future-prediction loss. Amazon omits parameter-independent constants from this objective. COVID-19 and Amazon reuse fixed synthetic banks with replacement; FRED-MD generates fresh episodes online. All use AdamW, weight decay 10^{-4}, gradient clipping at 5, and seed 17.

COVID-19 and Amazon use constant learning rates and synthetic DEV NLL for checkpoint selection, checked every 300 updates. Amazon architecture selection averages DEV NLL over three seeds. FRED-MD uses 1,024 warmup updates and cosine decay to 10% of its initial rate; training cycles complete and masked histories. Synthetic DEV NLL selects posterior checkpoints, while

Table 8: Base-model training settings. Updates count minibatch steps. History units are weeks, events, and months, respectively.

Setting	COVID-19	Amazon	FRED-MD
Synthetic TRAIN pairs	8,192	65,536	8,388,608 online
Synthetic DEV pairs	2,048	1,024	1,024
History length	24	12–60	36/48/60
Batch size	64	64	256
Total / selected update	3,000 / 900	6,000 / 5,400	32,768 / 32,768
Learning rate	5×10^{-4}	3×10^{-2}	2×10^{-4}

Table 9: Grouping settings. Each fitting and validation set contains n mechanisms. Weights follow the descriptor-block order given above; FRED-MD uses equal weights.

Domain	Setting	$K_{\max}$	λ	$m_{\min}$	n	Weights
COVID-19	Zero-shot	3	0.04	0.025	128	$(1, 2, 1)$
COVID-19	Adapted	6	0.08	0.050	128	$(1, 1, 1)$
Amazon	Zero-shot	3	0.04	0.025	256	$(2, 1, 1)$
Amazon	Adapted	8	0.04	0.025	512	$(2, 1, 1)$
FRED-MD	Both	6	0.08	0.050	128	Equal

real DEV ES, with Event-NLL as a tie-breaker, selects among FRED-MD checkpoint and prior configurations.

B.3 State Estimation and Future Simulation

Each sampled mechanism conditions a simulation initialized from the query history:

$$M^{(i)} \sim q_\phi(\cdot \mid D), \qquad s_T^{(i)} \sim r(\cdot \mid D, M^{(i)}), \qquad f^{(i)} \sim P_{\text{sim}}(\cdot \mid s_T^{(i)}, M^{(i)}). \tag{17}$$

COVID-19 uses a guided particle filter with 512 particles per mechanism. Amazon starts from the last observed mark and waiting band. FRED-MD uses a Rao–Blackwellized particle filter with 128 volatility particles and conditional Kalman updates. Each mechanism remains fixed throughout its rollout. Particle filtering approximates the conditional state distribution. State-filter likelihoods are not reapplied as mechanism weights. The simulator is the same one used for synthetic generation; no separate trajectory network is trained.

Forecast evaluation uses 256 mechanisms and one rollout each, with three Monte Carlo repetitions. Horizons are four weeks, six recorded events, and six months for COVID-19, Amazon, and FRED-MD. Every path retains its generating mechanism. A representative rollout is an actual simulated path; a plotted pointwise group median is a distributional summary.

B.4 Scenario Construction

We group posterior mechanisms using rule and parameter descriptors, without using future trajectories or realized query futures. COVID-19 descriptors cover location, variation, and dynamics; Amazon uses square-root transition, gate, and emission probabilities; FRED-MD uses structural choices, dynamic parameters, covariance contributions, and volatility. Descriptor blocks are scaled using 1,024 independent prior draws and compared by weighted squared Euclidean distance.

For each candidate K, medoids are fitted to one posterior sample set and evaluated on an independent set. Let V_K be mean validation distortion. Candidates minimize $J_K = V_K / \max(V_1, 10^{-12}) + \lambda(K-1)$ subject to minimum group mass $m_{\min}$. The final choice is the smallest eligible K within one paired standard error of the minimum; $K = 1$ is always allowed. Medoids are actual sampled mechanisms, and the resulting assignment g_D is then frozen.

COVID-19/Amazon configurations are selected using DEV histories and TRAIN-reference support and coverage. This domain-level selection is separate from selecting K for each query. Resampling stability is diagnostic and does not discard partitions or force multiple groups.

Probabilities and conditional futures. Scenario probability is posterior group mass, $p_k(D) = \Pr_{q_\phi}(g_D(M) = k)$. Each future inherits its generating mechanism's group, defining the conditional future distribution $P_k(\cdot \mid D)$. Consequently, $\sum_k p_k P_k$ equals the model's predictive distribution. Different groups may have overlapping futures.

Reported probabilities use 256 ranking draws, and conditional-future evaluation uses a separate bank of 256 linked rollouts; both are separate from primary partition fitting. Their finite-sample estimates therefore need not satisfy the mixture identity exactly. FRED-MD reuses its three forecast banks for fitting, ranking, and future evaluation. Auxiliary stability refits reuse ranking draws, and FRED-MD stability anchors reuse evaluation draws. These mechanism groups are distinct from the common frozen event partition used for cross-method realized-event scoring.

B.5 Predictive Adaptation and Data Use

Adaptation freezes the base posterior and simulator and learns a bounded reweighting:

$$q_{\phi,\omega}(M \mid D) \propto q_\phi(M \mid D) w_\omega(D, M), \qquad w_\omega(D, M) = 0.1 + 0.9\,\mathrm{sigmoid}(g_\omega(D, M)). \quad (18)$$

The network concatenates standardized history and mechanism features and applies a width-64 hidden layer with SiLU and a scalar output. Zero output-layer initialization preserves the base posterior. Training caches 64 mechanisms with four rollouts each per query, and minimizes predictive ES plus 0.01 times discrete KL from uniform weights on the cached mechanism draws. ES includes distances between distinct rollouts of the same mechanism. Prediction uses rejection sampling from the base posterior with acceptance probability $w_\omega(D, M)$.

Optimization. We use AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 32, gradient clipping at 1, seed 17, and 1,200 updates. TRAIN/DEV query counts are 2,048/128, 2,048/128, and 534/54. Checkpoints are evaluated every 100 updates, including initialization, using mean DEV ES plus the same KL penalty. Selected updates are 1,200, 0, and 100 for COVID-19, Amazon, and FRED-MD. Thus Amazon's adapted posterior equals its base posterior, although grouping settings and Monte Carlo draws differ.

C Experimental Setup

We evaluate COVID-19 counts, Amazon event sequences, and FRED-MD time series. Methods share query histories, horizons, observable representations, and scoring rules within each comparison. Preprocessing uses fixed training/calibration statistics or statistics from the observed query history.

C.1 Datasets and Preprocessing

COVID-19. We use Google COVID-19 Open Data v3, aggregating daily reported cases for 50 US states and Washington, DC into Saturday-ending weeks. The downloaded series span March 2020–September 2022. TRAIN forecast origins run from April 2020 to November 2021, DEV from December 2021 to March 12, 2022, and TEST from March 19 to August 13, 2022, with weekly stride. Training windows are purged at the DEV boundary. Completely missing weeks are zero-filled. Predictive ES uses $\log(1 + \max(y, 0))$ counts.

Amazon. We use the EasyTPP Amazon release, with 16 nominal event categories and native relative time units. Official TRAIN/DEV/TEST splits contain 6,454/922/1,851 sequences. Forecast origins start after 16 observations and advance by four events, retaining cases with six recorded subsequent events. Histories contain at most 60 events. Evaluation represents marks categorically and scales gaps by the observed history's median gap; the task predicts the next six recorded events, not events within a fixed calendar horizon.

FRED-MD. We use 18 transformed monthly indicators from the June-2025 FRED-MD snapshot. Monthly TRAIN/DEV/TEST forecast origins span 1965–June 2009, 2010–June 2014, and 2015–June 2024, respectively. Each query contains 60 historical months and six future months; the matched TimeStone histories and targets must be finite. Forecast scoring standardizes each indicator using the query history. This is a retrospective fixed-vintage evaluation.

Table 10: Matched evaluation cohorts. Scenario-distance evaluation uses fixed common subsets; TRAIN-reference support additionally requires reference eligibility.

Dataset	History	Horizon	Forecast/event TEST	Scenario-distance TEST
COVID-19	24 weeks	4 weeks	1,034	237
Amazon	16–60 events	6 events	12,035	256
FRED-MD	60 months	6 months	114	114

Table 11: Pretrained forecasting baselines and their prediction interfaces. Learned probabilistic extensions are reported separately from point configurations.

Method	Weights / implementation	Predictive output
Chronos-T5	Chronos-T5-base	Native autoregressive samples; independent FRED channels.
Moirai	Moirai-1.1-R-base	Native distribution samples with multivariate history conditioning.
ForecastPFN	Released TensorFlow SavedModel	Channelwise point forecasts.
TimePFN	Released TimePFN checkpoint	Multivariate point forecasts.
TabPFN-TS	Time-series package 1.2.0; TabPFN 8.5.0, v3 time-series weights (2026-05-06)	Quantile interpolation with shared temporal ranks and independent channel ranks.
LPTM	LPTM-large2, frozen	Point forecasts and separately trained probabilistic heads.

C.2 Baselines and Training Access

We separate methods by their use of real target-domain history–future pairs for predictive training. A frozen backbone with a target-trained head is adapted. TimeStone's zero-shot setting uses synthetic posterior training with real TRAIN prior calibration; its adaptation is specified in Appendix B.5.

Amazon foundation/PFN baselines forecast log gaps and 16 category-indicator channels; projected indicator probabilities generate nominal marks. This is an explicit event adapter, not scalar regression on category IDs. Outside the Amazon event adapter, point forecasts define degenerate predictive distributions unless a probabilistic extension is identified.

RMTPP, IFTPP, THP, and DeTPP use task-specific event/count adapters. Amazon uses recurrent event sampling; DeTPP adds a TRAIN-calibrated log-gap residual model. RMTPP and THP use documented recurrent/intensity implementations, while IFTPP uses the authors' lognormal-mixture model. COVID transfers use discretized weekly growth marks and TRAIN-derived within-mark changes. DeepCOVID is a count-only, direct-four-week bootstrap adaptation with 80 members. These adapters are distinguished from reproductions of the original complete forecasting pipelines.

SCATTER and ForestCast. We report numerical extensions inspired by these methods. SCATTER combines frozen Qwen2.5-3B-Instruct history embeddings with numerical summaries and a joint predictive decoder trained by NLL and accuracy/diversity rewards. ForestCast uses frozen Qwen2.5-7B-Instruct embeddings of a binary history-summary tree, with NLL and a prediction-error-based DPO-style objective. Adapted decoders use real TRAIN pairs; separately evaluated versions without target-domain predictive training use 4,096 synthetic pairs.

LPTM probabilistic extension. A frozen LPTM backbone supplies pooled history embeddings and point forecasts to a two-layer width-128 head. The head predicts one- or four-component Gaussian residual mixtures with diagonal-plus-rank-two covariance; Amazon adds component-conditioned categorical mark probabilities. We train on real TRAIN pairs using joint NLL, AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 128, gradient clipping at 5, and 2,000 updates. TRAIN/DEV query counts are 2,048/128, 2,048/128, and 534/54 for COVID-19, Amazon, and FRED-MD. Checkpoints are selected by DEV joint NLL within each component configuration and training seed. The four-component configuration is reported, averaging seeds 17, 29, and 43; component count is not scenario count.

C.3 Common Scenario Evaluation

Forecasting baselines group trajectories in a common domain-specific observable space and use sampled trajectories as representatives. Candidate groups use 128 fitting and 128 validation samples. Independent 256-sample banks estimate ranking probabilities and evaluate conditional futures. TimeStone groups mechanisms as described in Appendix B.4. We evaluate up to three positive-mass scenarios in probability order, allowing fewer when selected by the method. This matches the displayed-scenario budget, not grouping compute.

Scenario distance and coverage error use eight history-matched DEV or eligible TEST reference futures, shared across methods. COVID-19 and FRED-MD references must finish by the query cutoff; COVID-19 also excludes the query state. Amazon excludes the query's split-local sequence, without claiming calendar or cross-split customer isolation. References enter scoring only after forecasts, groups, and representatives are fixed.

C.4 Statistical Reporting

Repeated predictions are averaged within each query before aggregation. For scenario and mechanism comparisons, 95% percentile intervals use paired cluster bootstrap, with clusters defined by COVID-19 state, Amazon sequence identifier, or FRED-MD year. We use 2,000 resamples, except for the FRED-MD mechanism-inference controls, which use 5,000. These intervals condition on the fitted models and reference banks; training-seed standard deviations are reported separately. No multiple-comparison correction is applied.

Amazon's selected TimeStone adaptation is the identity initialization; differences between its reporting arms do not demonstrate learned reweighting gains.

D Scenario Quality Evaluation

D.1 Reference Futures and Distance Metrics

We evaluate each method's scenario groups with predictions and group definitions held fixed. TimeStone groups mechanisms by their rules and parameters and assigns each generated future to its mechanism's group.

Reference selection. For Table 2, we select the eight nearest reference histories from eligible validation and test records using history features alone, and use their observed continuations as reference futures. All methods use the same reference set for each query. We exclude records with the same ID or an identical history as the query. For COVID-19, we also exclude references from the same US state as the query. Amazon references come from a different data split. COVID-19 and FRED-MD reference futures must be fully observed by the query's forecast origin. The selected FRED-MD reference futures do not overlap in time, although their history windows may overlap.

History features. COVID-19 uses log-count changes over the latest eight weeks, the latest per-capita reported level, and volatility (10 dimensions). Amazon uses event-type frequencies over the latest 16 events, one-hot encodings of the last four event types, the latest eight inter-event times, and the median inter-event time over the latest 16 events (89 dimensions). FRED-MD uses the latest 12 months of changes normalized by standard deviations from the 60-month history, together with the latest values and volatility (252 dimensions). We center and scale each feature coordinate using medians and interquartile ranges (IQRs) from real training histories, then retrieve neighbors using Euclidean distance with feature blocks scaled to account for their dimensions.

Future features and distances. Let $\Phi(f; D)$ denote the future features of trajectory f computed using history D. COVID-19 uses the four-week log-count trajectory relative to the last observed log-count, divided by the standard deviation over the eight-week history and by $\sqrt{4}$. Amazon uses six future inter-event times transformed as $\log(1+\mathrm{gap}/m)/\sqrt{6}$, where m is the median inter-event time over the latest 16 observed events, together with event-type one-hot encodings divided by $\sqrt{12}$ (102 dimensions). FRED-MD uses all six months and 18 variables relative to their last observed values,

divided by their respective 60-month history standard deviations and by $\sqrt{108}$. Standard deviations and IQRs used for normalization are floored at 0.1. No additional training-IQR transformation is applied to future features.

Each reference is normalized using its own history. Thus, the distances in Equation 9 are computed in the shared feature space as

$$\|\Phi(f; D) - \Phi(r; D_r)\|_2 ,$$

where D_r is the history associated with reference future r. Generated futures and scenario representatives use the query history D. The metrics therefore compare future developments expressed relative to their corresponding histories.

Scenario representatives. TimeStone uses the sampled future generated by the medoid mechanism in each group. Each baseline uses a trajectory medoid, chosen to minimize the sum of squared distances to the other trajectories in its group in the grouping feature space. Both representatives are actual generated samples and are selected without using reference futures or the query's observed future.

Reported metrics. We assess up to three scenarios with the highest positive probabilities and average both metrics over all evaluated test histories. Scenario distance weights the selected scenarios equally and uses $X_k = \{c_k\}$ when a scenario has no evaluation samples. These metrics are used only for evaluation.

E Additional Predictive Evaluation

E.1 Energy Score and Shared Event Categories

We evaluate predictive trajectories using the shared observation-space maps below. Event-NLL, accuracy, and prediction-set metrics use fixed categories of these observations: five rule-based categories for COVID-19 and eight TRAIN-fitted clusters for Amazon and FRED-MD. These evaluation categories are separate from the mechanism groups used to construct scenarios.

Future representations. For history D and future f, let $\Phi_D(f)$ denote the evaluation vector. The same map is applied to predictions and the observed future.

- **COVID-19:** the four weekly counts are mapped to $(\log(1 + \max(f_h, 0)))_{h=1}^4$.
- **Amazon:** for six future gaps Δ_h and marks m_h, we use

$$\Phi_D(f) = \left[\frac{(\log(1 + \Delta_h / s_D))_{h=1}^6}{\sqrt{6}}, \frac{(e_{m_h})_{h=1}^6}{\sqrt{12}} \right], \tag{19}$$

  yielding 102 features, where e_m is a 16-dimensional one-hot vector and s_D is the history-gap lower median, floored at 10^{-8}. For an even number of gaps, the lower of the two middle values is used. Gaps are measured from the forecast cutoff and then between consecutive future events.

- **FRED-MD:** the six-month, 18-indicator transformed future is standardized channelwise by the preceding 60 months and vectorized, with an additional factor $1/\sqrt{108}$. History standard deviations use denominator 60; a value at most 10^{-8} is replaced by 1. This yields $\Phi_D(f) \in \mathbb{R}^{108}$.

Finite-sample energy score. Each Monte Carlo repetition supplies $S = 256$ equally weighted trajectories. Model mixture or ensemble probabilities enter through trajectory generation; the scorer uses the resulting uniform sample bank without further reweighting. Repeated trajectories, including point predictions, are retained. For $x_s = \Phi_D(f^{(s)})$ and $y = \Phi_D(f^{\text{obs}})$, the finite-sample energy

score is

$$\widehat{\mathrm{ES}} = \frac{1}{S}\sum_{s=1}^{S}\|x_s - y\|_2 - \frac{1}{2}\widehat{D},$$
$$\widehat{D}_{\mathrm{C}} = \frac{1}{128}\sum_{s=1}^{128}\|x_s - x_{s+128}\|_2, \tag{20}$$
$$\widehat{D}_{\mathrm{A/F}} = \frac{2}{S(S-1)}\sum_{1\le s<t\le S}\|x_s - x_t\|_2.$$

Here $\widehat{D}_{\mathrm{C}}$ is used for COVID-19, with the fixed first-half/second-half pairing of stored samples; $\widehat{D}_{\mathrm{A/F}}$ is used for Amazon and FRED-MD, with all $S(S-1)/2$ distinct index pairs. Neither estimator includes self-pairs. These conventions are shared across methods within each domain.

We score each query separately for Monte Carlo seeds 0, 1, and 2, then average the scores. We do not pool the 768 trajectories before calculating nonlinear metrics such as Event-NLL or prediction sets. When a method has multiple training seeds, its scores are additionally averaged equally across those seeds. The matched TEST cohorts contain 1,034 COVID-19, 12,035 Amazon, and 114 FRED-MD queries.

COVID-19 categories. Let $h_{-3}, \ldots, h_0$ be the last four recorded counts, $f_1, \ldots, f_4$ the nonnegative future counts, and P the region's population. The event-rule history retains recorded reporting revisions. Set

$$b = 5P/10^5,$$
$$R_f = \frac{\frac{1}{4}\sum_{t=1}^{4} f_t}{\max(\frac{1}{4}\sum_{t=-3}^{0} h_t, b)}, \tag{21}$$
$$\epsilon = 0.02\max\left(\frac{h_0 + \sum_{t=1}^{4} f_t}{5}, b\right).$$

Define $f_0 = h_0$. Remove differences $f_t - f_{t-1}$ with absolute value at most ϵ, then count sign changes among the remaining differences. Apply the following rules in order, assigning the first matching category:

1. **Multiple turns (ID 0):** at least two sign changes.
2. **Stable or declining (ID 1):** $R_f \le 1.1$.
3. **Rise then peak (ID 2):** some $t \in \{1, 2, 3\}$ satisfies $f_t > f_{t-1}$ and $f_t > f_{t+1}$.
4. **Rise without an interior peak (ID 3):** $f_4 \ge f_1$.
5. **Other (ID 4):** all remaining futures.

The ordered rules are mutually exclusive and exhaustive. Their names summarize observable patterns and are not inferred mechanism labels.

Amazon and FRED-MD categories. We fit one $K = 8$ Euclidean k-means partition per domain using 8,192 Amazon and 534 FRED-MD TRAIN future vectors, each transformed relative to its own history. Amazon windows are selected at evenly spaced indices from eligible TRAIN windows. Initialization uses seed 20260913: the first center is a uniform random training vector, and subsequent centers are sampled proportionally to squared distance from the nearest selected center, with weights floored at 10^{-15}. We perform 100 assignment/mean-update iterations, retaining the previous center for an empty cluster. No additional whitening or centroid normalization is applied.

The centers are frozen and shared across methods. A predicted or observed future is assigned to the nearest center in Φ_D space, with exact ties resolved by the smaller category ID. Fitting uses neither TEST futures nor model predictions. This joint-future partition differs from next-mark-only or per-indicator event definitions.

Table 12: **Prediction-set coverage and size.** Coverage is a fraction and is interpreted jointly with set size; the nominal probability mass is 0.9. The LPTM row uses the frozen point-forecast adapter. ‡ denotes methods extended to generate numerical future trajectories.

	COVID-19		Amazon		FRED-MD	
Method	**Set cov.**	**Set size**	**Set cov.**	**Set size**	**Set cov.**	**Set size**
Without target-domain predictive training						
Chronos-T5	0.920	3.046	0.937	1.507	0.833	1.073
Moirai	0.778	2.493	0.981	2.696	0.868	1.368
ForecastPFN	0.263	1.000	0.873	1.009	0.658	1.000
TimePFN	0.374	1.000	0.873	1.008	0.816	1.000
TabPFN-TS	0.651	2.350	0.895	1.188	0.877	1.345
SCATTER‡	0.675	2.213	0.872	1.487	0.858	1.827
ForestCast‡	0.761	2.503	0.879	1.555	0.865	1.962
LPTM	0.279	1.000	0.874	1.004	0.789	1.000
TimeStone (zero-shot)	0.943	3.021	0.987	1.848	0.933	2.629
With target-domain predictive training						
RMTPP	0.964	3.230	0.961	1.577	0.919	1.642
IFTPP	0.974	3.717	0.990	1.855	0.947	2.638
THP	0.944	3.161	0.951	1.509	0.923	1.513
DeTPP	0.949	3.083	0.974	1.759	0.920	1.686
DeepCOVID	0.643	2.143	0.966	1.619	0.877	1.135
SCATTER (adapted)‡	0.812	2.479	0.984	1.839	0.905	2.850
ForestCast (adapted)‡	0.794	2.561	0.978	1.820	0.845	2.673
TimeStone (adapted)	0.944	3.067	0.987	1.849	0.924	2.626

E.2 Event Probability Metrics

For query i, Monte Carlo repetition r, and category j, let $\widehat{\pi}_{irj}$ be the fraction of its 256 predictive samples assigned to j, and let $y_i^\star$ be the observed future's category. With N queries and $R = 3$ repetitions, we report

$$\begin{aligned} \text{Event-NLL} &= -\frac{1}{NR}\sum_{i=1}^{N}\sum_{r=1}^{R}\log\max(\widehat{\pi}_{ir,y_i^\star}, 10^{-3}), \\ \text{Accuracy} &= \frac{1}{NR}\sum_{i=1}^{N}\sum_{r=1}^{R}\mathbf{1}\{\arg\max_j \widehat{\pi}_{irj} = y_i^\star\}. \end{aligned} \tag{22}$$

The floor is applied only inside the NLL; it does not change or renormalize probabilities used for accuracy or prediction sets. Accuracy ties are resolved by the smaller category ID.

E.3 Prediction Sets

Sort categories by decreasing $\widehat{\pi}_{irj}$, resolving ties by increasing category ID. The prediction set $C_{ir}^{0.9}$ is the shortest prefix whose cumulative probability is at least 0.9. Equal-probability categories are not automatically included together. We report

$$\begin{aligned} \text{Set coverage} &= \frac{1}{NR}\sum_{i=1}^{N}\sum_{r=1}^{R}\mathbf{1}\{y_i^\star \in C_{ir}^{0.9}\}, \\ \text{Set size} &= \frac{1}{NR}\sum_{i=1}^{N}\sum_{r=1}^{R}|C_{ir}^{0.9}|. \end{aligned} \tag{23}$$

Coverage and size are interpreted jointly. The 0.9 threshold concerns predicted probability mass and does not guarantee 90% empirical coverage. ES, Event-NLL, and prediction-set metrics are computed from the same saved predictions and matched queries, with shared event partitions for the latter two metrics.

E.4 Prediction-Set Results

The numerical adapters for FRED-MD point-process baselines and DeepCOVID outside COVID-19 follow the domain-specific baseline setup. The LPTM row reports its frozen point-forecast adapter, including categorical projection for Amazon; it does not use an additionally trained mixture head.

On COVID-19, TimeStone zero-shot attains coverage 0.943 with mean set size 3.021, compared with 0.920 and 3.046 for Chronos-T5. On Amazon, coverage 0.987 accompanies a larger set than Chronos-T5 (1.848 versus 1.507). On FRED-MD, higher zero-shot coverage also accompanies larger sets than the baselines without target-domain predictive training. These are coverage–size comparisons, not evidence that larger coverage alone implies a better forecast.

F Comparison with the Scenario Modeling Hub

F.1 Forecast Origins and Scenario Sets

The quantitative comparison covers 47 US jurisdictions at forecast origins June 4 and July 30, 2022, yielding 94 cases with four-week observed futures. We use SMH rounds 14 and 15, respectively (COVID-19 Scenario Modeling Hub, 2022a;b), and compare their four published scenarios with up to four TimeStone scenarios ranked by inferred probability. The defining booster and variant changes begin beyond these windows, so the comparison concerns short-term scenario sets. Since SMH does not assign occurrence probabilities to its scenarios, we evaluate the scenario medians as sets.

F.2 Scenario-Set Metrics

Nearest-median RMSE. For case i, let $\widetilde{f}_{i,k}$ denote the median trajectory of scenario k and $f_i^\star$ the observed future, expressed in the same evaluation coordinates. The case-level score is

$$\min_k \sqrt{\frac{1}{4}\sum_{h=1}^{4}\left(\widetilde{f}_{i,k,h} - f^\star_{i,h}\right)^2}. \tag{24}$$

The closest median is selected after observing the future, so this score measures whether the scenario set contains a close candidate.

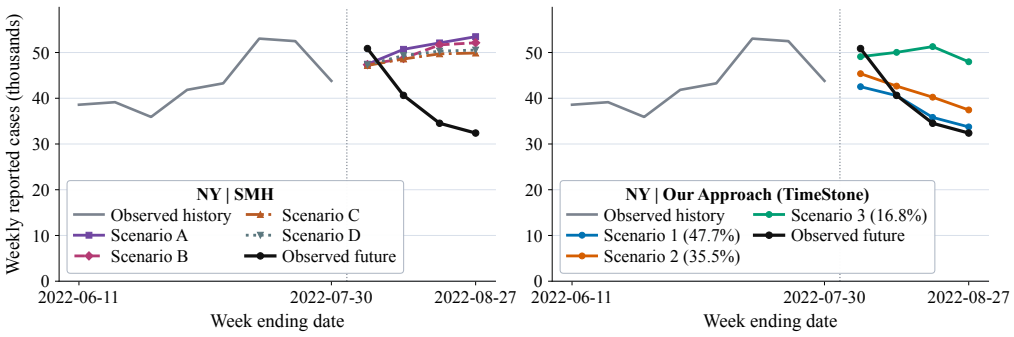
Reference coverage. For each case, we use 32 observed continuations from similar training histories as reference futures. Reference coverage is the fraction of these futures matched by at least one scenario median. It measures the fraction of matched references, whereas the coverage error in Section 4.1 measures their average distance to the nearest representative. Aggregate results are reported in Table 4.

F.3 Selected Examples

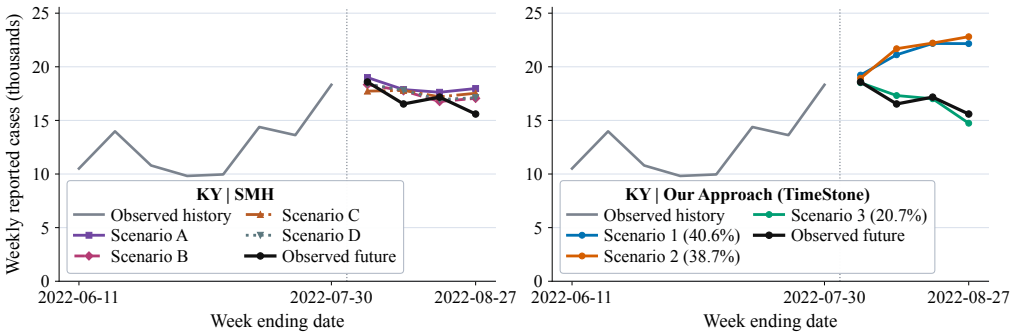
Figure 3 compares SMH with TimeStone (zero-shot) for Illinois over four weeks following June 4, 2022, and Ohio over ten weeks following March 12, 2022. Illinois is included in the quantitative comparison. Ohio provides an additional qualitative example of a rebound following a decline and is not included in the 94-case evaluation. Additional visualizations are provided in Appendix F.4.

F.4 Additional Scenario Visualizations

We show additional four-week examples from the forecast origins used in the SMH comparison. These selected cases illustrate different observed developments and how they are represented by the scenario sets. Colored curves show scenario medians, gray curves show observed histories, and black curves show observed futures. Probabilities in parentheses refer to TimeStone's mechanism groups. Aggregate performance is reported in Table 4.

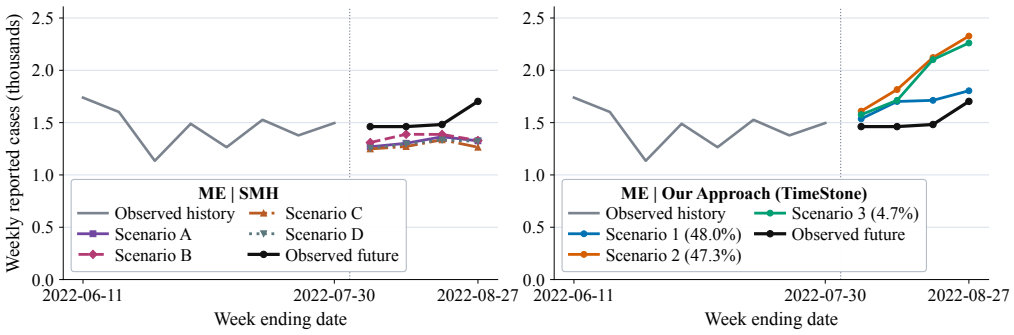


(a) New York (adapted, July 30, 2022)

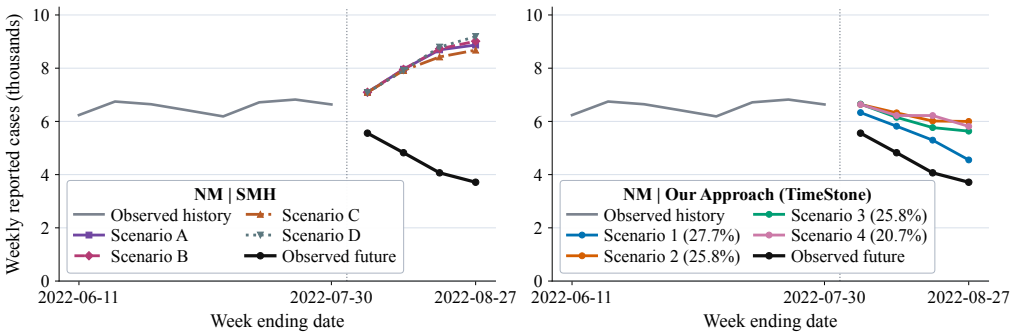


(b) Kentucky (zero-shot, July 30, 2022)

Figure 4: **Observed declines represented by different scenario ranks.** In New York, TimeStone's highest-probability scenario (47.7%) follows the observed decline, while the SMH medians increase. In Kentucky, a lower-probability scenario (20.7%) follows the observed trajectory more closely than TimeStone's two higher-probability scenarios, which increase. The closest continuation need not be the highest-probability one.

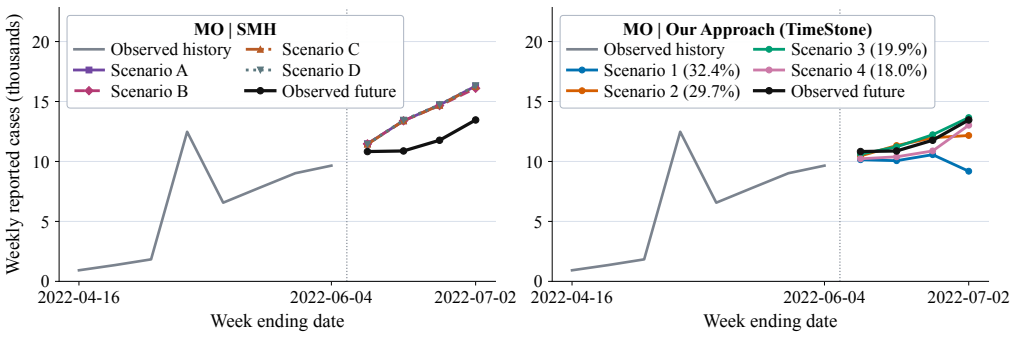


(a) Maine (zero-shot, July 30, 2022)

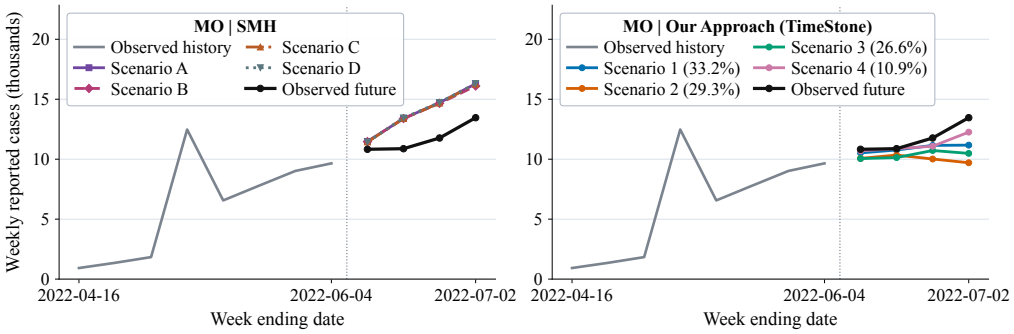


(b) New Mexico (adapted, July 30, 2022)

Figure 5: **Direction and magnitude of alternative continuations.** In Maine, TimeStone includes increasing trajectories consistent with the final observed rise, although they increase earlier than the observations. In New Mexico, all displayed TimeStone medians decrease, with different magnitudes, while the SMH medians increase. TimeStone captures the direction of the observed decline but overestimates case counts.



(a) Missouri (adapted, scenario set 1)



(b) Missouri (adapted, scenario set 2)

Figure 6: **Two scenario sets for Missouri at the June 4, 2022 forecast origin.** In (a), a lower-probability scenario (19.9%) follows the observed increase. In (b), an increasing alternative remains present, but all displayed medians underestimate the final observed count. The two panels show different TimeStone outputs for the same case and are not independent evaluation cases. Scenario indices and probabilities refer to each displayed output separately.

G Additional Analyses and Ablation Studies

We provide additional analyses of mechanism inference, scenario construction, and prior sensitivity, following Section 5. Unless stated otherwise, analyses use the zero-shot model. Statistical procedures and evaluation scope are described in Appendix C.4.

G.1 Mechanism Inference

We evaluate inference against known synthetic mechanisms and test its predictive value on real histories.

G.1.1 Synthetic Mechanism Inference

We generate 256 independent synthetic histories per domain, containing 24 weekly observations for COVID-19, 60 events for Amazon, and 60 monthly observations for FRED-MD. For each history, we draw 256 mechanisms from the inferred distribution and 256 from the prior. We evaluate prespecified continuous parameter summaries using nominal 90% marginal interval coverage, interval width, and posterior-mean estimation error. Table 5(a) reports coverage averaged over summaries and the median of summary-wise posterior-to-prior width ratios.

Prior RMSE baseline. For each case, the prior and posterior point estimates are the means of their respective 256 draws. We compute RMSE against the known generating mechanisms over the same 256 cases and report the median, across parameter summaries, of posterior RMSE divided by prior RMSE. Both point estimates therefore use the same Monte Carlo sample budget.

A separate six-week COVID-19 diagnostic gives 79.3% marginal coverage, below the nominal 90%. Its protocol and cases differ from the matched-forecast-origin experiment in Appendix G.1.4.

Expected posterior KL on Amazon. Amazon admits an analytical mechanism posterior under the specified prior and simulator. We estimate the expected forward KL from this posterior to the model distribution using 2,048 held-out synthetic pairs sampled from the prior joint distribution:

$$\widehat{\mathrm{KL}} = \frac{1}{2048} \sum_{i=1}^{2048} \left[\log p_0(M^{(i)} \mid D^{(i)}) - \log q_\phi(M^{(i)} \mid D^{(i)}) \right], \quad (M^{(i)}, D^{(i)}) \sim p_0(M, D). \tag{25}$$

Because each generating mechanism is conditionally distributed according to $p_0(M \mid D)$, this average estimates

$$\mathbb{E}_{D \sim p_0(D)} \left[D_{\mathrm{KL}}(p_0(\cdot \mid D) \,\|\, q_\phi(\cdot \mid D)) \right]. \tag{26}$$

The estimate is 0.0559, with a 95% interval of $[0.0414, 0.0710]$. It estimates the forward KL for the full mechanism distribution, averaged over synthetic histories, matching the mechanism NLL objective. The evaluated Amazon model combines learned mixture weights with analytical updates within components.

G.1.2 History-Conditioned Mechanism Inference

We replace the query-conditioned mechanism distribution with either the prior or a distribution inferred from another history. Each variant estimates the simulator state from the query history under its sampled mechanisms and generates 256 futures. State estimates can therefore vary with the mechanisms. The evaluation includes 1,034 COVID-19 cases, 12,035 Amazon cases, and 114 FRED-MD cases.

Table 13 reports control scores minus query-conditioned scores, so positive differences favor history-conditioned inference. The query-conditioned reference forecasts in this ablation have mean ES values of 0.5779, 0.5962, and 2.282 on COVID-19, Amazon, and FRED-MD, respectively.

Both controls have higher ES and Event-NLL across all three datasets, with paired intervals excluding zero. Conditioning the mechanism distribution on the query history therefore provides predictive information beyond using that history for state estimation.

Table 13: **Predictive value of history-conditioned mechanism inference.** Entries show control scores minus zero-shot scores with paired 95% confidence intervals. Positive values favor the distribution inferred from the query history.

Dataset	Control	ΔES	ΔEvent-NLL
COVID-19	Prior	0.1728 [0.1135, 0.2426]	0.0809 [0.0283, 0.1423]
	Other history	0.2174 [0.1699, 0.2733]	0.2318 [0.1748, 0.2943]
Amazon	Prior	0.0253 [0.0226, 0.0281]	0.0181 [0.0120, 0.0240]
	Other history	0.0579 [0.0538, 0.0620]	0.0664 [0.0535, 0.0787]
FRED-MD	Prior	0.3255 [0.0443, 0.6678]	0.2594 [0.0056, 0.4660]
	Other history	0.3654 [0.0954, 0.6658]	0.1548 [0.0069, 0.2973]

Table 14: **Predictive value of retaining multiple mechanisms.** Positive score differences favor the full mechanism distribution. Brackets show paired 95% confidence intervals.

Dataset	ΔES	ΔEvent-NLL
COVID-19	0.02707 [0.01702, 0.03875]	0.06371 [0.03774, 0.09422]
Amazon	0.00533 [0.00501, 0.00567]	0.00302 [0.00152, 0.00452]
FRED-MD	0.00654 [−0.00687, 0.02359]	0.01572 [−0.01763, 0.05121]

G.1.3 Retaining Mechanism Uncertainty

The single-mechanism control selects the sampled mechanism with the smallest total mechanism distance to the other posterior samples. It generates 256 stochastic futures from this representative, matching the full-distribution forecast's future sample budget. Table 14 reports single-mechanism score minus full-distribution score.

Retaining multiple mechanisms improves both scores over this representative-mechanism control on COVID-19 and Amazon. The FRED-MD intervals include zero.

Differences among mechanism-specific forecasts. We measure disagreement among mechanism-specific future distributions without using the observed future and fix the disagreement ranges using development data. The ES benefit of retaining multiple mechanisms increases from 0.0127 to 0.2703 between the low and high ranges on COVID-19, and from 0.00293 to 0.00806 on Amazon. The high COVID-19 range contains 36 cases. On FRED-MD, the high-minus-low difference is positive, although the high range's benefit interval includes zero. Table 15 reports these associations, evaluated in the same distance representation used by ES.

G.1.4 Effect of History Length

We vary the amount of observed history while keeping the forecast origin fixed. For each domain, we use 128 synthetic cases and three Monte Carlo evaluations, removing earlier observations to obtain shorter histories. Table 16 reports mechanism spread relative to the prior, the mean number of scenario groups, and average coverage of nominal 90% marginal mechanism intervals.

Longer histories reduce mechanism spread in all three domains, with paired intervals excluding zero, while the number of groups does not decrease consistently. Amazon's lower spread is accompanied by lower marginal coverage. These comparisons include any change in initial-state distribution caused by truncation. The six-week COVID-19 coverage differs from the separate 79.3% diagnostic in Appendix G.1.1 because the protocols and cases differ.

G.1.5 Unseen Mechanism Combinations

We examine whether inference generalizes to combinations of mechanism components excluded from training. The components themselves remain present in other combinations. The excluded combinations involve high persistence together with strong restoring dynamics on COVID-19, high repetition propensity together with a large gate difference on Amazon. For FRED-MD, we exclude the joint occurrence of signed-AR dynamics and the base component appearing most frequently in 4,096 prior draws. Both elements remain present separately in other training and development combinations. These combinations are excluded from both training and development data.

Table 15: **Mechanism uncertainty and differences between forecasts.** Entries compare the ES benefit of retaining multiple mechanisms in the high-disagreement range with that in the low-disagreement range. Positive differences indicate a larger benefit when mechanism-specific forecasts differ more. Brackets show 95% confidence intervals.

Dataset	Difference in ES benefit (high minus low)
COVID-19	$0.25758\ [0.11743,\ 0.36990]$
Amazon	$0.00514\ [0.00430,\ 0.00595]$
FRED-MD	$0.02197\ [0.00239,\ 0.04931]$

Table 16: **Effect of history length on inference and scenario construction.** Entries show results for shorter and longer histories at the same forecast origin. History lengths are measured in weeks for COVID-19, events for Amazon, and months for FRED-MD.

Dataset	History length	Spread / prior	Mean $K(D)$	Coverage (%)
COVID-19	$6 \to 24$	$0.503 \to 0.470$	$2.93 \to 2.95$	$82.4 \to 91.5$
Amazon	$15 \to 60$	$0.406 \to 0.138$	$2.78 \to 1.96$	$90.3 \to 87.8$
FRED-MD	$36 \to 60$	$0.307 \to 0.288$	$2.52 \to 2.64$	$90.4 \to 90.3$

Seen and unseen cases. For each domain, we generate 256 new cases from combinations represented during training and 256 from the excluded combinations. Table 17 reports scores on the excluded combinations minus scores on the represented combinations. We evaluate both future ES and mechanism NLL, the negative log-density assigned to the generating mechanism. Mechanism NLL depends on the parameterization and dimension, so its magnitude should not be compared across domains.

Unseen combinations have higher mechanism NLL on COVID-19 and higher losses on both measures on FRED-MD. The COVID-19 future ES and both Amazon differences are inconclusive. These comparisons use different seen and unseen cases; the following control evaluates the effect of training exclusion on matched cases.

Training comparison on the same unseen cases. Under a separate training protocol, we train COVID-19 and Amazon models for 12,000 steps with or without the designated combinations. Both models are evaluated on the same unseen cases: 128 per domain for future ES, and 517 COVID-19 and 542 Amazon cases for mechanism NLL. Table 18 reports excluded-combination minus all-combination scores. These results are separate from Table 17.

Exclusion increases COVID-19 mechanism NLL, while neither domain shows a conclusive future ES difference. Aggregate forecast scores can therefore be less sensitive to changes in mechanism inference; the intervals do not establish equivalent performance.

G.1.6 Dynamics outside the Mechanism Family

We compare forecasts under the specified dynamics with forecasts under two departures: asymmetric COVID-19 latent shocks with matched mean and variance, and Amazon events that depend on earlier marks beyond the modeled transitions.

The reported COVID-19 and Amazon ES and Event-NLL changes have intervals including zero. Under the stronger COVID-19 perturbation, coverage of the nominal 90% event prediction set decreases by 4.30 percentage points, with a 95% interval of $[-8.20, -0.39]$. This measures coverage of observable event categories.

High-probability COVID-19 predictions are infrequent and Amazon event categories are concentrated, limiting the assessment of overconfidence under these perturbations.

G.2 Scenario Construction

We evaluate the predictive information, organization, and stability of mechanism groups, the numerical accuracy of their probabilities, and the value of retaining within-scenario future distributions. Each analysis specifies its evaluation subset and training setting.

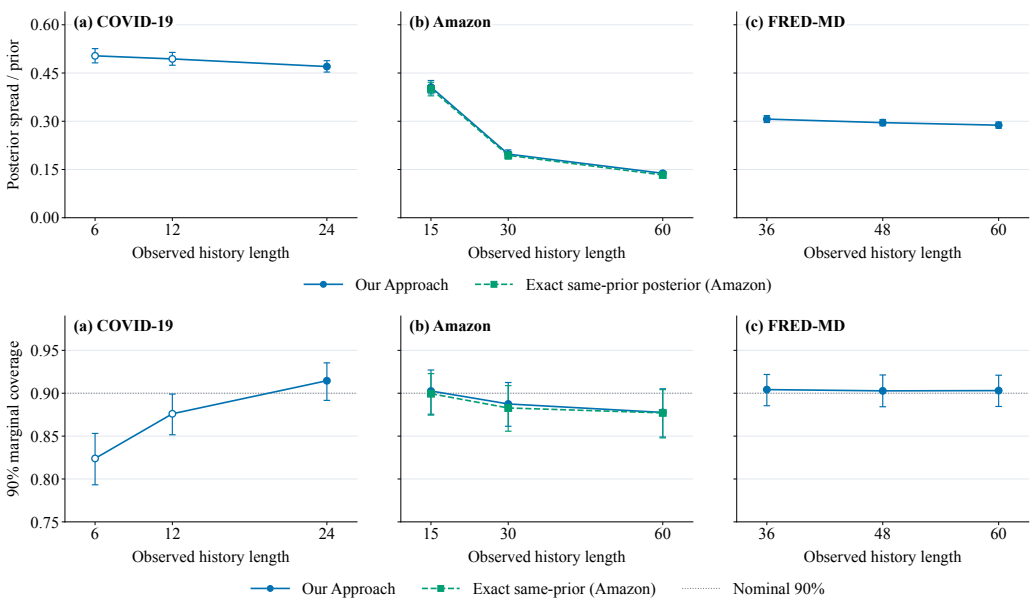


Figure 7: **Mechanism inference with different history lengths.** Top: mechanism spread relative to the prior. Bottom: coverage of nominal 90% marginal intervals. History lengths are measured in weeks for COVID-19, events for Amazon, and months for FRED-MD. The Amazon panels also show inference under the exact posterior with the same prior. The two panels distinguish posterior concentration from interval coverage.

Table 17: **Generalization to excluded mechanism combinations.** Score differences compare unseen with seen combinations for a model trained without the excluded combinations. Positive differences indicate higher loss on unseen combinations. Brackets show 95% confidence intervals.

Dataset	ΔFuture ES	ΔMechanism NLL
COVID-19	-0.04796 $[-0.12575, 0.03084]$	2.097 $[1.810, 2.384]$
Amazon	-0.00703 $[-0.06291, 0.04961]$	-75.554 $[-526.641, 401.722]$
FRED-MD	0.05782 $[0.02928, 0.08692]$	11.060 $[9.523, 12.623]$

G.2.1 Predictive Information in Mechanism Groups

We compare mechanism grouping with random assignments that preserve group counts and sizes within each sample set. For each independent simulated target, we assign its known generating mechanism to a group and evaluate ES against independent futures from that group. The random control replaces these assignments. This tests group-conditional prediction under the simulator.

Table 19 uses 237 COVID-19, 256 Amazon, and 114 FRED-MD histories. Differences are mechanism-group minus random-group ES, so negative values favor mechanism grouping. Relative reductions divide the ES improvement by random-group ES, using unrounded scores.

All three differences favor mechanism grouping, with intervals excluding zero. The reductions are small on Amazon and FRED-MD, supporting predictive distinctions between groups under the specified simulator.

G.2.2 Grouping Stability and Selection Rules

Stability under resampling. We repeat grouping using independent samples and measure agreement between the resulting assignments using the adjusted Rand index (ARI). Table 20 compares mechanism and trajectory grouping at the same number of groups. The table includes only cases for which ARI can be evaluated, and reports the eligible fraction for each domain.

Table 18: **Effect of excluding combinations during training.** Both models are evaluated on the same unseen cases. Positive differences indicate higher loss when the combinations are excluded. Brackets show 95% confidence intervals.

Dataset	ΔFuture ES	ΔMechanism NLL
COVID-19	-0.00267 $[-0.00859, 0.00297]$	0.52048 $[0.45143, 0.59143]$
Amazon	0.00011 $[-0.00035, 0.00056]$	0.05923 $[-0.00030, 0.12238]$

Table 19: **Predictive information in mechanism groups.** Conditional ES is evaluated on independent simulated targets. The random control preserves group sizes. Brackets show 95% confidence intervals. Negative ΔES favors mechanism grouping.

Dataset	Histories	Mechanism groups	Random groups	ΔES	Reduction
COVID-19	237	0.886405	0.910471	-0.024065 $[-0.029212, -0.019647]$	2.643%
Amazon	256	0.634278	0.637264	-0.002986 $[-0.003468, -0.002574]$	0.469%
FRED-MD	114	0.994492	0.995365	-0.000873 $[-0.001554, -0.000279]$	0.088%

Mechanism grouping has higher ARI than trajectory grouping. A fixed-hash control also achieves high stability, so predictive information is assessed separately through conditional ES in Appendix G.2.1.

G.2.3 Accuracy of Scenario Probability Estimation

We test the numerical accuracy of estimating scenario probabilities by mechanism sample fractions. For each of 64 histories per domain, we fix the fitted groups and draw fresh mechanism samples at budgets of 100, 256, 1,000, and 4,096, with 20 repetitions per budget. An independent set of 65,536 draws provides a reference probability vector. We repeat the evaluation for the zero-shot and adapted mechanism distributions.

For sample budget N, let $\widehat{p}_k^{(N)}(D)$ denote the estimated probability of fixed group k, and let $\widehat{p}_k^{(\mathrm{ref})}(D)$ denote its reference estimate. We compute the total variation error

$$\mathrm{TV}_N(D) = \frac{1}{2} \sum_{k=1}^{K(D)} \left| \widehat{p}_k^{(N)}(D) - \widehat{p}_k^{(\mathrm{ref})}(D) \right|. \tag{27}$$

Table 21 reports results at budgets of 256 and 4,096. Values are multiplied by 100 for readability and measure discrepancies between probability vectors.

Larger sample budgets reduce TV error in every domain and setting. The COVID-19 and Amazon subsets contain no single-group cases. These results assess Monte Carlo accuracy relative to a finite reference estimate, with the model and groups fixed; they do not assess calibration against real outcomes.

G.2.4 Value of Within-Scenario Future Distributions

We compare decisions based on all 256 predictive samples with decisions based on one representative per scenario and the corresponding probabilities. Both summarize the same forecast. Using shortfall-to-excess cost ratios of $1{:}1$, $4{:}1$, and $9{:}1$, we evaluate the value of retaining within-scenario uncertainty.

Table 22 reports representative-based cost minus full-sample cost at a $4{:}1$ ratio. Positive differences indicate higher costs after compression. Cost units differ across domains, so magnitudes should only be compared within each domain.

Representative-based costs are higher in all 18 combinations of domain, training setting, and cost ratio, with 15 intervals excluding zero. Results also depend on representative choice. These findings support retaining each scenario's conditional future distribution for decision calculations: combining all group distributions with their probabilities recovers the overall forecast, while a single representative discards within-group variation.

Table 20: **Grouping stability under resampling.** Higher ARI indicates more consistent assignments. Both groupings use the same number of groups. Eligible fractions identify the cases included in the reported comparison.

Dataset	Mechanism ARI	Trajectory ARI	Eligible (%)
COVID-19	0.6478	0.4830	100.00
Amazon	0.6367	0.1749	96.09
FRED-MD	0.5192	0.1466	98.25

Table 21: **Monte Carlo accuracy of scenario probabilities.** Entries report $100 \times$ TV relative to an independent 65,536-draw reference. The inferred mechanism distribution and group definitions remain fixed as the sample budget changes.

Dataset	Setting	$N = 256$	$N = 4096$
COVID-19	Zero-shot	3.277	0.868
	Adapted	3.560	0.917
Amazon	Zero-shot	3.099	0.813
	Adapted	3.701	0.957
FRED-MD	Zero-shot	3.005	0.782
	Adapted	2.876	0.744

Amazon's adapted model retains identity weights, so differences between its reporting arms do not measure learned adaptation effects.

G.3 Prior Sensitivity

G.3.1 Configurations and Training Budgets

For COVID-19 and Amazon, we compare a baseline prior with two variants, labeled in Table 23. Each configuration is trained for 12,000 steps with three seeds and evaluated on 237 COVID-19 cases or 256 Amazon cases. For FRED-MD, we compare a common-SV baseline family with a sector-dynamic family, each trained for 32,768 steps with three seeds and evaluated on 114 cases. This family comparison is separate from the main benchmark, where zero-shot and adapted TimeStone use the same mechanism family.

G.3.2 Predictive Performance

Table 7(a) reports variant-minus-baseline differences with 95% confidence intervals under matched training budgets.

Table 23 reports means and standard deviations across three training seeds for COVID-19 and Amazon. For FRED-MD, mean ES decreases from 2.3149 with the common-SV family to 2.2780 with the sector-dynamic family. The paired ES difference excludes zero, while the Event-NLL difference remains inconclusive.

G.3.3 Probabilities over Fixed Mechanism Regions

For 64 real histories per domain on COVID-19 and Amazon, we compare inferred probabilities over fixed mechanism regions using $\mathrm{TV}(u, v) = \frac{1}{2}\sum_j |u_j - v_j|$. Table 7(b) compares prior-induced changes with resampling variation under the same posterior; ranges span the tested variants. Fixed regions provide corresponding quantities across models without equating separately fitted scenario groups.

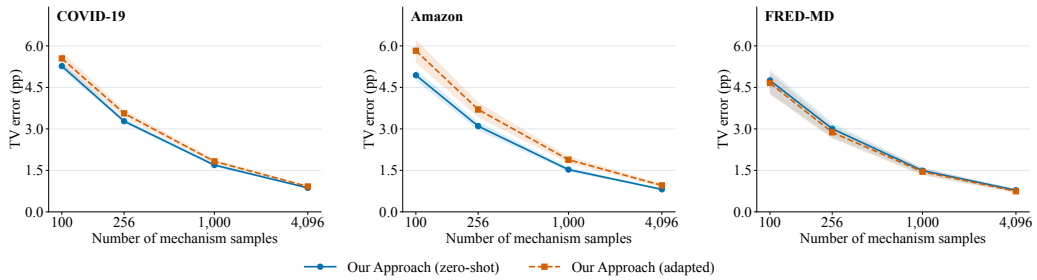


Figure 8: **Convergence of scenario probability estimates.** The inferred mechanism distribution and group assignments are fixed. Curves report $100 \times$ TV between sample-fraction estimates and an independent 65,536-draw reference as the mechanism sample budget increases. The evaluation uses 64 histories per domain and 20 repetitions at each budget. This measures Monte Carlo approximation of fixed scenario masses.

Table 22: **Decision cost of replacing conditional distributions with representative trajectories.** Entries report representative-based cost minus full-sample cost at a 4:1 shortfall-to-excess cost ratio. Brackets show 95% confidence intervals. Positive values indicate higher costs after compression.

Dataset	Zero-shot	Adapted
COVID-19	1.127 [0.480, 1.845]	3.419 [0.547, 8.814]
Amazon	0.214 [0.175, 0.250]	0.137 [0.098, 0.176]
FRED-MD	0.135 [0.022, 0.275]	0.268 [0.186, 0.368]

Table 23: **Predictive scores across prior settings.** Entries are means and standard deviations across three training seeds. Parentheses indicate the corresponding labels in the prior-sensitivity figures.

Dataset	Prior setting	ES	Event-NLL
COVID-19	Baseline (H4/P2)	0.6274 ± 0.0033	1.2065 ± 0.0040
	Prior variant 1 (H16/P2)	0.6064 ± 0.0156	1.1923 ± 0.0155
	Prior variant 2 (H4/P4)	0.6189 ± 0.0048	1.2004 ± 0.0124
Amazon	Baseline (M80/G5)	0.5902 ± 0.0002	0.2820 ± 0.0011
	Prior variant 1 (M20/G5)	0.5891 ± 0.0002	0.2823 ± 0.0003
	Prior variant 2 (M20/G20)	0.5885 ± 0.0004	0.2763 ± 0.0017

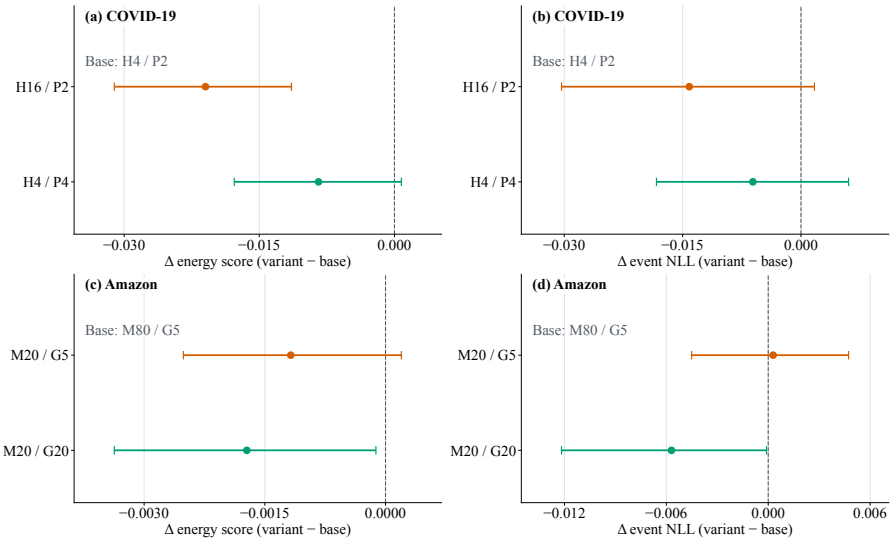


Figure 9: **Predictive sensitivity to prior settings.** Points show changes in ES and Event-NLL relative to each domain's baseline prior under matched training budgets. Negative values favor the alternative prior. Horizontal bars show 95% confidence intervals. Prior labels correspond to Table 23.

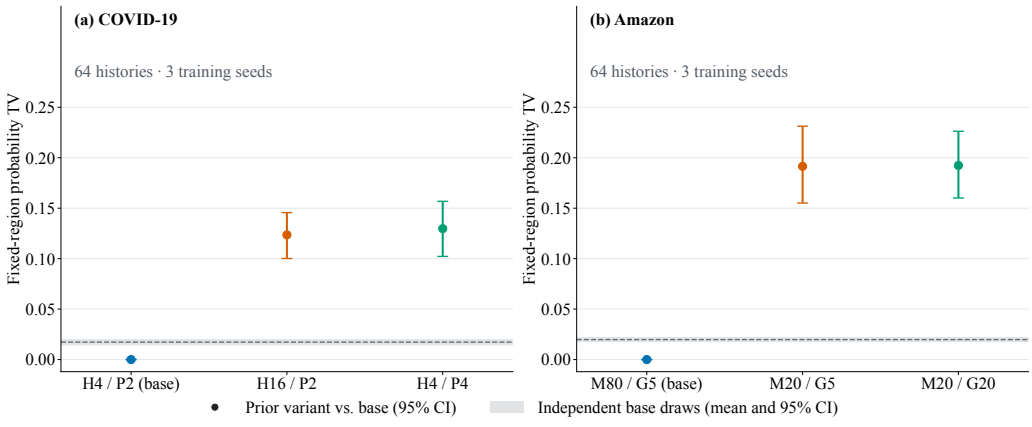


Figure 10: **Prior sensitivity of mechanism-region probabilities.** Total variation compares each prior variant with the base model over the same fixed regions. The gray reference summarizes differences between independent draws from the base posterior. The evaluation uses 64 histories and three training seeds per domain. Fixed regions permit corresponding probabilities to be compared without equating separately fitted scenario groups.