Causal Inference with Time Series Foundation Models

Cyrus Illick
Columbia University

Saeyoung Rho
Columbia University

Vishal Misra
Columbia University

Abstract

We study counterfactual estimation with time-series panel data where multiple units are observed and only one unit undergoes an intervention. Synthetic control (SC) addresses this problem by treating post-intervention outcomes of the treated unit as missing and imputing them using the rest of the panel. This can be viewed as a special forecasting problem in which future observations of untreated units are available. Motivated by this perspective, we explore the use of time-series foundation models (TSFMs) in the SC setting and compare with classical linear SC methods. The results show that linear models remain strong baselines, while TSFMs offer advantages in settings with stronger temporal trends, greater nonlinearity, or when the low-rank assumption fails. Finally, we discuss future research directions highlighting the need for better adaptation of TSFMs for SC.

1 INTRODUCTION

In observational causal inference, counterfactual estimation often involves inference with time series data. One of the most widely adopted practices is synthetic control (SC), a method for predicting counterfactual time series of a treated unit by leveraging observations from untreated units. For example, if California (i.e., *target* unit) decides to adopt a new policy, SC estimates what would have happened to California had that intervention not occurred by creating the counterfactual time series using data from the states that did not go through a similar intervention (i.e., *donor* units). The original SC modeled the target time series as a simplex combination of donor time series (Abadie and Gardeazabal, 2003). More recently, methods have been extended that view it as linear regression (Doudchenko and Imbens, 2016; Amjad et al., 2018), matrix completion (Athey et al., 2021), and representation learning (Xu, 2017; Rho et al., 2026), adopting more advanced machine learning techniques.

In the meantime, the machine learning community's focus has evolved from developing a task-specific model to learning a foundational model that can be adapted to a wide range of downstream tasks. The most prominent example is large language models (LLMs), where the model is trained on vast, diverse data. This new paradigm has been adopted in time-series forecasting as well, and multiple technology companies have announced time series foundation models (TSFMs) in recent years, e.g., TimesFM (by Google, Das et al. (2024)), Chronos (by Amazon, Ansari et al. (2024, 2025)), TabPFN (by PriorLabs, Grinsztajn et al. (2025)) and MOIRAI (by SalesForce, Woo et al. (2024)). The main purpose of TSFMs is to predict the future time series given historical measurements. The SC prediction task can be viewed as time-series forecasting if we can utilize post-intervention observations from the donor units in the prediction process.

In this paper, we explore how TSFMs can be adopted to facilitate causal inference, especially focusing on SC applications. We introduce technical details of SC and TSFMs (Section 2), discuss how to use TSFMs for SC (Section 3), present experimental results (Section 4), and discuss future work (Section 5).

2 BACKGROUND

2.1 Synthetic Control

The time-series panel dataset for synthetic control consists of the following components: Let $Y_{i,t} \in \mathbb{R}$ be the observation from i-th unit at time t. We have a (treated) target unit with index 1 and n (untreated) donor units indexed by $i \in \{2, ..., n+1\}$. The untreated observation matrix is of size $(n+1) \times T$, where the target unit's values after intervention $t > T_0$ are missing because the intervention occurs at T_0.

The classical algorithm, `SC` (Abadie and Gardeazabal, 2003), learns SC weights $f \in \mathbb{R}^n$ of donors to estimate the target, i.e. $\hat{Y}_{1,t} = (Y_{2:n+1,t})^T f$, under the simplex constraint ($\sum f_i = 1, f_i \geq 0 \ \forall i$). Later, Ro-

bust Synthetic Control (`RSC`, Amjad et al. (2018)) was suggested by focusing on the approximate low-rank behavior of the observational matrix. It performs principal component regression by selecting top few singular values of the matrix as a de-noising step prior to regression. Recently, Time-Aware Synthetic Control (`TASC`, Rho et al. (2026)) was proposed to use temporal trend of the data by adopting a state space model. It uses Kalman filtering and RTS smoothing to estimate the hidden states that evolve over time, while also preserving the approximately low-rank structure.

We use the three algorithms (`SC, RSC,` and `TASC`) as traditional baseline SC approaches. They are all learning from only a given dataset (as opposed to multiple time-series examples that the model can train on) and the model operates on a linear class of functions[1].

2.2 Time Series Foundation Models (TSFMs)

In this section, we discuss TSFMs that allow the use of post-intervention donor data.

`TabPFN`(v2.5, Grinsztajn et al. (2025)) is a foundation model designed for tabular data, that can be utilized for time-series panel data by treating time steps as examples and covariates (e.g. donor-series) as features. The model forms embeddings from small groups of features, and applies alternating attention mechanisms to model dependencies between examples (time steps) and features (units). `TabPFN-TS`(Hoo et al., 2026) explicitly reformulates time-series forecasting as a tabular regression problem and solves it using `TabPFN`(v2). Given a target series $y_{1:T}$ and optional covariates, a tabular dataset is constructed with each time index corresponding to a row of temporal features (e.g. time encodings, seasonal indicators), paired with observed values. Forecasting reduces to predicting future rows, whose temporal features and covariates are known.

`Chronos`(v2, Ansari et al. (2025)) is specifically designed for time-series forecasting. Given a target time series and covariates (e.g. donor-series), the model partitions each series into fixed-length patches, and generates patch embeddings. The model approximates the predictive distribution of the target's post-intervention time series using two attention mechanisms: *time attention* between patch embeddings in the same series and *group attention* between series at the same patch time index. Time attention uses positional embedding to model temporal evolution.

`TimesFM` (v2.5, Das et al. (2024); Google Research (2026)) is a TSFM that operates on fixed-length temporal patch embeddings but models dependence solely along the time dimension, without cross-series attention. To incorporate covariates, it uses an external regression component in which the target series is first modeled as a function of covariates (e.g. donor-series), and the residual component is then forecast using a decoder-only attention model. The final prediction combines the regression-based and residual forecast.

As opposed to the traditional algorithms to learn SC models, these foundational models are trained on a vast amount of time-series datasets coming from a wide range of distributions, utilizing the Transformer structure and moving beyond the classical linear regime.

3 USING TSFMS FOR SC

TSFMs were originally developed for time-series forecasting, where long historical contexts are available and no known future covariates are assumed. As a result, their architectures are not optimized for SC settings[2]. For example, patch sizes ($P = 16$ for `Chronos` (v2) and $P = 32$ for `TimesFM`) are relatively large for datasets typically used in SC applications. Moreover, known future covariates play only an auxiliary role, but are central to SC. Hence, the key to using TSFMs for SC is to explore which architecture is most suited for the SC setting and how to incorporate the future donor information as an input.

To do so, we consider three methods. First, we can adopt the panel data as input `(*-panel)`. Recent TSFMs allow future known covariates as an additional input, which corresponds to the donor time-series in SC. Second, we can flatten the panel structure into a univariate time series `(*-flatten)` by reading the matrix column by column, from bottom to top. For example, the panel data $Y \in \mathbb{R}^{(n+1)\times T}$ can be flattened to a sequence $(Y_{n+1,1}, Y_{n,1}, \ldots, Y_{1,1}), \ldots, (Y_{n+1,T}, Y_{n,T}, \ldots, Y_{1,T})$. Now, the target time series $Y_{i,\cdot}$ can be predicted by using all previous observations in an autoregressive way. Third, we can subtract the donor mean from the target observation (`*-mean`) and predict the residual.

4 EXPERIMENTAL RESULTS

In this section, we evaluate TSFM and classic linear approaches for SC with synthetic data. Results with real-world data are deferred to Appendix B.

[1] This comes from linear factor model assumed by SC (See Abadie (2021)) and also approximately low-rank assumption (Amjad et al., 2018)

[2] See Yu et al. (2025a) for the impact of design choices on TSFM's performance

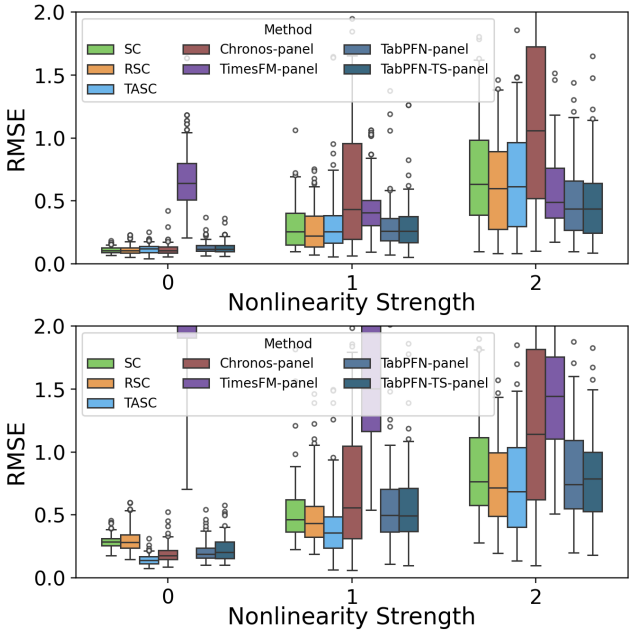


Figure 1: Prediction error with varying nonlinearity strength, with small(top) and large(bottom) R.

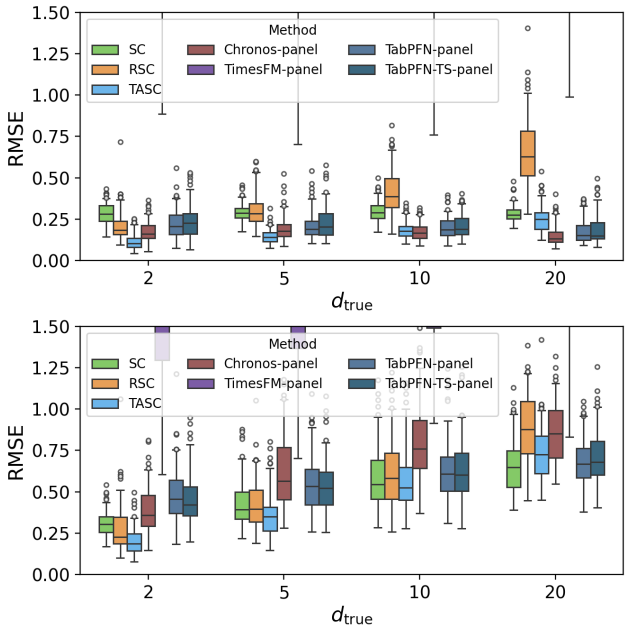


Figure 2: Prediction error with varying latent dimension d_{true}, with small(top) and large(bottom) Q.

4.1 Synthetic Data Generation

Let $y_t \in \mathbb{R}^{n+1}$ be the observation and $x_t \in \mathbb{R}^d$ be the hidden state at time t. We generate synthetic data following a modified state-space model:

$$x_t = Ax_{t-1} + \alpha \tanh(Bx_{t-1}) + q_{t-1} \quad (1)$$

$$y_t = Hx_t + \beta \tanh(Cx_t) + r_t, \quad (2)$$

where $q_{t-1} \sim \mathcal{N}(0, Q)$ is the latent perturbation and $r_t \sim \mathcal{N}(0, R)$ is the observation noise. The parameters $\theta = \{A, H, B, C, Q, R\}$ are randomly initialized while ensuring the stability of the generated time series. The latent signal x_t evolves following a linear trend A and a nonlinear trend $\tanh(Bx_{t-1})$ with strength α. Then, the observation y_t is the sum of a linear projection Hx_t and a nonlinear transformation $\tanh(Cx_t)$ with strength β. We consider $\alpha, \beta \in \{0, 1, 2\}$. Full explanation is deferred to Appendix A.

4.2 Ablation Study with Simulated Data

In this section, we compare linear approaches and TSFM approaches on simulated datasets. All TSFMs are implemented using the `*-panel` method, and the datasets are generated with $n = 49$, $T_0 = 50$, and $T = 100$. The prediction error is computed as Root Mean Squared Error(RMSE) of post-intervention target estimate with respect to its signal $y^{sig}_{t,1} = y_{t,1} - r_{t,1}$, assuming zero intervention effect.[3]

First, we vary the nonlinearity strength ($\alpha = \beta$) in the simulated data. Figure 1 presents the post-intervention RMSE under low (top panel, small R) and high (bottom panel, large R) observation noise. All methods except for `TimesFM` tend to yield lower prediction error when the data is linear ($\alpha = \beta = 0$). While linear models show stronger performance in most cases, `TimesFM`, `TabPFN` and `TabPFN-TS` achieve lower prediction error compared to linear models when nonlinearity strength is large ($\alpha = \beta = 2$) and observation noise is low.

Next, we vary the hidden state dimension $d_{true} \in \{2, 5, 10, 20\}$ while fixing $\alpha = \beta = 0$ and large R. Figure 2 shows the post-intervention RMSE under strong (top panel, small Q) and weak (bottom panel, large Q) temporal trends. For `RSC` and `TASC`, we set the approximate rank hyperparameter[4] to $d = d_{true}$ to evaluate their best-case performance. TSFMs are less sensitive to the increasing d_{true}. When temporal trend is strong, TSFMs (except `TimesFM`) outperform linear approaches when d_{true} increases, with `Chronos` achieving the lowest median RMSE when $d_{true} = 10, 20$. On the other hand, when the temporal trend is weak, TSFM approaches offer no advantage to linear SC approaches across all choices of d_{true}.

[3] We evaluate accuracy on 100 generated datasets in each setting.

[4] This corresponds to the number of singular values to keep in `RSC` and hidden state dimension in `TASC`.

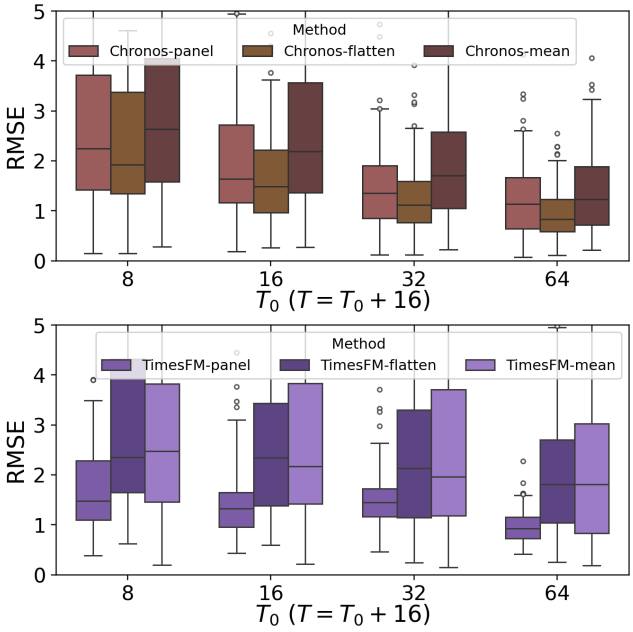


Figure 3: Post-intervention RMSE with varied T_0 and different data input methods for `Chronos` (top, $n = 15$) and `TimesFM` (bottom, $n = 31$).

4.3 Input Representation for TSFMs

Now, we evaluate three input representation styles (`*-panel`, `*-flatten`, `*-mean`) from Section 3, focusing on `Chronos`(v2, $P = 16$) and `TimesFM`($P = 32$).

Figure 3 shows the prediction error as T_0 increases ($T_0 \in 8, 16, 32, 64$), with $T = T_0 + 16$, $d_{true} = 5$, and $\alpha = \beta = 2$, focusing on settings where TSFMs are expected to perform well. Note that the number of donors n is chosen based on the model's patch size to allow repeating patterns between each patch[5]. In the top panel, `Chronos-flatten` outperforms the other two methods, suggesting potential for flattening method. In the bottom panel, `TimesFM-panel` still remains the best method for `TimesFM`. This shows that TimesFM benefits more from the regression mechanism utilized in the `*-panel` input style than from temporal attention in the univariate inputs `*-flatten` and `*-mean`. In evaluations on real-world datasets in Appendix B, we show cases where `*-mean` input achieves lower median RMSE than `*-panel` input.

5 DISCUSSION

We explored how TSFMs can be adapted for SC. The results show that TSFMs have potential to be advantageous in the following regimes: (1) when the data exhibit stronger nonlinearity and low observation noise (Figure 1, top), and (2) when the data has bigger rank and temporal trend is strong (Figure 2, top).

Nonetheless, linear models (`SC`, `RSC`, and `TASC`) remain strong baselines in most settings, particularly (1) when observation noise is high for any levels of nonlinearity strength tested (Figure 1, bottom) and (2) when temporal trends in the latent state are less pronounced (Figure 2, bottom). Importantly, linear models are strong baselines when the panel data has approximate low-rank structure, which is commonly observed in real-world settings (Udell and Townsend, 2019). Yu et al. (2025b) show that time-series data can induce low-rank input embeddings in Transformer models, similar to `TASC`. This suggests that low-rank structure in the latent space may be a key characteristic in SC that needs to be investigated. A promising direction for future work is to better understand where the boundary lies between low- and high-rank regimes, and how this distinction can guide the choice between linear approaches and TSFMs.

Furthermore, the different architectures and training data of TSFMs result in performance distinctions that depend on the structure of the panel data and should be investigated further in future work. For example, in Appendix C we observe that when observation noise is correlated: `TimesFM`, `TabPFN` and `TabPFN-TS` trend toward predicting the noisy observation, rather than the true signal, when pre-intervention length increases whereas `Chronos` (as well as linear approaches) predict the true signal.

Future work could explore fine-tuning TSFMs for SC tasks and guiding prediction with alternate input representations. For example, the improved performance of `Chronos-flatten` shows how altering the input representation has potential for SC tasks with short pre-intervention periods.

In practice, computational efficiency is an important trade-off. Linear models are far more memory-efficient than TSFMs (e.g., n parameters in `SC` versus millions in TSFMs). They are also quicker to train. However, if the same TSFM is reused for many SC instances, the initial training cost may eventually be amortized. At inference time, linear models remain substantially faster than TSFMs. Another limitation of TSFMs for SC is that, unlike classical SC, they do not provide a transparent donor-weight decomposition, making it harder to assess which donors drive the counterfactual estimation.

Lastly, future work should evaluate these methods on a broader range of datasets, including more complex data-generating processes, varying data sizes, and additional real-world datasets.

[5]This choice is made for investigation purposes and is not intended as a general recommendation.

References

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review*, 93(1):113–132.

Amjad, M., Shah, D., and Shen, D. (2018). Robust synthetic control. *Journal of Machine Learning Research*, 19(22):1–51.

Ansari, A. F., Shchur, O., Küken, J., Auer, A., Han, B., Mercado, P., Rangapuram, S. S., Shen, H., Stella, L., Zhang, X., Goswami, M., Kapoor, S., Maddix, D. C., Guerron, P., Hu, T., Yin, J., Erickson, N., Desai, P. M., Wang, H., Rangwala, H., Karypis, G., Wang, Y., and Bohlke-Schneider, M. (2025). Chronos-2: From univariate to universal forecasting.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., and Wang, Y. (2024). Chronos: Learning the language of time series.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116:1716–1730.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. (2015). Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1):247–274.

Das, A., Kong, W., Sen, R., and Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting.

Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. NBER Working Paper 22791.

Google Research (2026). TimesFM: Time Series Foundation Model. `https://github.com/google-research/timesfm`.

Grinsztajn, L., Flöge, K., Key, O., Birkel, F., Roof, B., Jund, P., Jäger, B., Hayler, A., Safaric, D., Simone Alessi, F. J., Manium, M., Yu, R., Garg, A., Robertson, J., Hoo, S. B. L., Moroshan, V., Bühler, M., Purucker, L., Cornu, C., Wehrhahn, L. C., Bonetto, A., Gambhir, S., Hollmann, N., and Hutter, F. (2025). Tabpfn-2.5: Advancing the state of the art in tabular foundation models.

Hoo, S. B., Müller, S., Salinas, D., and Hutter, F. (2026). From tables to time: Extending tabpfn-v2 to time series forecasting.

Rho, S., Illick, C., Narasipura, S., Abadie, A., Hsu, D., and Misra, V. (2026). Time-aware synthetic control.

Udell, M. and Townsend, A. (2019). Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. (2024). Unified training of universal time series forecasting transformers.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76.

Yu, A., Maddix, D. C., Han, B., Zhang, X., Ansari, A. F., Shchur, O., Faloutsos, C., Wilson, A. G., Mahoney, M. W., and Wang, Y. (2025a). Understanding the implicit biases of design choices for time series foundation models.

Yu, A., Maddix, D. C., Han, B., Zhang, X., Ansari, A. F., Shchur, O., Faloutsos, C., Wilson, A. G., Mahoney, M. W., and Wang, Y. (2025b). Understanding transformers for time series: Rank structure, flow-of-ranks, and compressibility. *arXiv preprint arXiv:2510.03358*.

Causal Inference with Time Series Foundation Models: Supplementary Materials

A HOW WE GENERATED DATA

In this section we provide more in-depth detail on how we generated synthetic datasets for evaluation. Each dataset is generated from the nonlinear state-space model,

$$x_t = Ax_{t-1} + \alpha \tanh(Bx_{t-1}) + q_{t-1}, \tag{3}$$
$$y_t = Hx_t + \beta \tanh(Cx_t) + r_t, \tag{4}$$

where $x_t \in \mathbb{R}^d$ denotes the latent state, $y_t \in \mathbb{R}^N$ denotes the observed panel with $N = n + 1$ (one target and n donors), $q_t \sim \mathcal{N}(0, Q)$ is the latent state temporal perturbation, $r_t \sim \mathcal{N}(0, R)$ is the observation noise, and the initial state is drawn as $x_0 \sim \mathcal{N}(m_0, P_0)$. The parameter set is $\theta = \{A, H, B, C, Q, R, m_0, P_0\}$.

The transition matrix $A \in \mathbb{R}^{d\times d}$ is constructed by drawing a matrix with i.i.d. standard normal entries and applying a QR decomposition to obtain an orthonormal matrix. To ensure stability of the latent dynamics, we control the spectral radius $\rho(A)$; if $\rho(A) > 0.95$, we rescale $A \leftarrow A \cdot (0.95/\rho(A))$ so that $\rho(A) \le 0.95$, preventing explosive trajectories. $H \in \mathbb{R}^{N\times d}$ is generated row-wise. For each row $i \in [N]$, a concentration vector $a_i \in (0, 1)^d$ is drawn with i.i.d. Uniform$(0, 1)$ entries, and $H_{i,:}$ is sampled from a Dirichlet distribution with parameter a_i.

The perturbation covariance Q is generated by first drawing a random matrix Z with entries sampled independently from Uniform$([a, b))$ scaled by $1/\sqrt{d}$, form $\tilde{Q} = ZZ^\top + 10^{-6}I$. Then, we randomly flip the sign of off-diagonal entries while keeping it symmetric. Unless otherwise specified, R is generated as a diagonal matrix where each diagonal entry is drawn independently from Uniform$([a, b))$. We also evaluate performance when observation noise is correlated with R generated following the same procedure as Q.

The initial latent mean is sampled as $m_0 \sim$ Uniform$(0, 1)^d$. The initial covariance P_0 is generated using the same procedure as for Q described above, with noise range $a = 0.01, b = 0.1$. The initial state is then drawn as $x_0 \sim \mathcal{N}(m_0, P_0)$.

We consider four covariance regimes: (small Q, small R), (large Q, small R), (small Q, large R), (large Q, large R). Where a "small" matrix is generated with $a = 0.01, b = 0.1$, and a "large" matrix is generated with $a = 0.1, b = 1.0$.

The nonlinear components are controlled by α (latent nonlinearity) and β (observation nonlinearity), with $\alpha, \beta \in \{0, 1, 2\}$. Latent-only nonlinearity corresponds to $(\alpha, \beta) = (s, 0)$, observation-only to $(0, s)$, and joint nonlinearity to (s, s) for $s \in \{0, 1, 2\}$.

With these parameters, we generate the true signal $y_t^{sig} = Hx_t + \beta \tanh(Cx_t)$, as well as the noisy observation $y_t = y_t^{sig} + r_t$. In our evaluations, the model only sees y_t but the prediction accuracy is measured by comparing against the target component $y_{t,1}^{sig}$ of the true signal y_t^{sig}. In settings where we evaluate performance on data generated with correlated observation noise (non-diagonal R), we evaluate performance accuracy measuring against the target unit's true signal $y_{t,1}^{sig}$ as well as against the noisy observation $y_{t,1} = y_{t,1}^{sig} + r_{t,1}$. We first evaluate prediction accuracy in settings with diagonal observation noise covariance matrix R, as this provides a clean benchmark in which the optimal predictor of the noisy observed outcome coincides with the optimal predictor of the true signal. In this case, comparisons across methods directly reflect differences in signal recovery. In contrast, when R is non-diagonal and observation noise is correlated, donor observations contain information about contemporaneous target noise, and the optimal predictor of the noisy observed outcome differs from that of the true signal. In such settings, we evaluate both true signal and noisy observation prediction.

B EMPIRICAL EVALUATION ON REAL-WORLD DATA

In this section, we evaluate TSFM approaches (`Chronos` (v2), `TimesFM`, `TabPFN` (v2.5), and `TabPFN-TS`) for SC on real-world datasets and compare performance to linear approaches `TASC, SC, RSC`. We maintain a consistent methodology in terms of choice of coefficients and experimental setup as described in (Rho et al., 2026). In (Rho et al., 2026), evaluations compare performance between `TASC`, `RSC`, `SC`, and Causal Impact Model (`CIM`) (Brodersen et al., 2015) - where `TASC` showed strong performance against these other methods. In our evaluations, we focus on the performance of TSFM approaches and their comparative performance to linear approaches (`TASC`, `SC`, `RSC`).

For TSFM approaches, we consider two data-preparation inputs fed into the TSFM (discussed in Section 3): (`*-panel`) and (`*-mean`).

B.1 Proposition 99

Our first real world evaluation is on the classic synthetic control application from (Abadie et al., 2010): evaluating the effect of Proposition 99. Proposition 99 was a policy enacted in California in 1988 that significantly increased the state's cigarette tax. To assess the causal impact of this policy intervention, synthetic control methods can be applied to estimate the counterfactual outcome for California - i.e. estimate what cigarette sales would have looked like had the policy not been implemented. Some SC methods can incorporate auxiliary variables such as average retail price of cigarettes, per capita state personal income, etc. However, in our analysis, we evaluate the methods using the single predictor of per-capita cigarette sales for a consistent comparison. In order to evaluate the capabilities of each method in predicting California's post-intervention counterfactual time series from the point of intervention in 1988 until 2000, we conduct *placebo tests*. Since the true counterfactual is unobservable, we simulate it by treating each donor unit (states who did not undergo a policy intervention) as if it were the target. In each placebo test, we predict a donor unit's time series using the remaining donors and assess whether the synthetic control method can accurately reconstruct the observed outcomes.

Figure 4 shows post-intervention root mean squared error (RMSE) from the placebo test for TSFM and linear SC approaches. Following (Rho et al., 2026) in this setting, we set the ridge coefficient of `RSC` as 0.1 and approximate rank $d = 2$, and `TASC` uses the same $d = 2$ for hidden state dimension. The ridge coefficient used in `TimesFM-panel` is set at 0.1 to match that of `RSC` in this setting. We include two methods of data input (`*-panel` and `*-mean`), described in Section 3, for `Chronos` and `TimesFM`.

Across linear and TSFM approaches, `TASC` achieves the lowest median and smallest variance of RMSE, suggesting that `TASC` may provide the most reliable post-intervention prediction of California. Out of the four TSFM methods with `*-panel` input, `TimesFM` achieves the lowest median RMSE. However, over all data preparation methods `Chronos-mean` achieves the lowest median RMSE out of TSFM methods and has the second overall lowest median RMSE (after `TASC`).

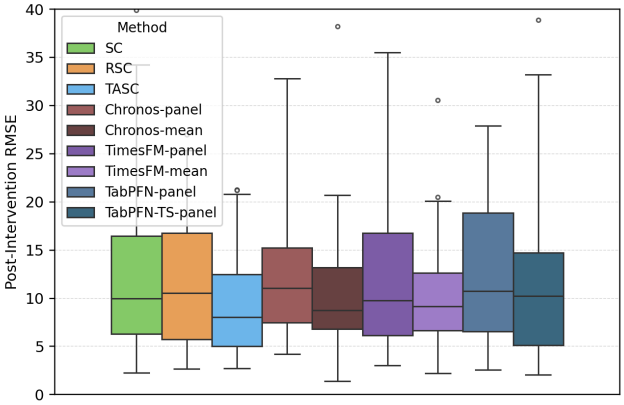


Figure 4: Post-intervention RMSE from placebo test, Proposition 99

B.2 Cricket

Although synthetic control is primarily designed for counterfactual inference, it can also be applied to prediction in *episodic time series*, such as game score trajectories. By aligning matches at their start time, we can construct panel datasets for SC application.

In this section, we extend our evaluation to sports data with an evaluation over cricket score trajectory data from Indian Premier League (IPL). The evaluation is consistent with what is followed in (Rho et al., 2026). The dataset utilized in this study is created from ball-by-ball records of IPL matches from April 18, 2008 to March 25th, 2025. In the T20 format, a standard match consists of two innings of 20 overs each, with 6 balls bowled per over, and one team batting per inning. The analysis is restricted to the first 120 legal deliveries in each inning and does not include the additional deliveries that can sometimes exceed 120 due to penalty balls. There are 1524 score trajectories in the dataset that have at least 120 balls delivered and are used in generating placebo tests to evaluate on. The scores are aggregated cumulatively over the course of each inning per ball.

Each inning in the dataset is represented as a time series of length $T = 120$. The intervention point is fixed at $T_0 = 72$, corresponding to the transition from power-play to later overs. The remaining 48 deliveries ($t = T_0 + 1, \ldots, T$) are used for evaluation under a placebo assumption (no true intervention effect).

To simulate a realistic forecasting setting, a target match is randomly selected from all eligible matches, and the donor pool consists of the n most recent matches prior to the target match date, with $n \in \{18, 36, 72, 144\}$. For each method and donor size, the experiment is repeated 100 times, each time with a newly selected random target match. Median post-intervention RMSE is reported across repetitions.

The data is mean-centered prior to fitting by subtracting the mean score trajectory computed from the selected donors; for `TASC` and `RSC`, the hidden state dimension and number of singular values to keep were both set to be $d = 5$ and `RSC` is implemented with the ridge coefficient of 10^3 - following (Rho et al., 2026). `TimesFM` ridge coefficient is set at 10^3 to match that of `RSC` in this setting. For TSFM methods in this setting, we evaluate only (`*-panel`) and (`*-mean`) approaches.

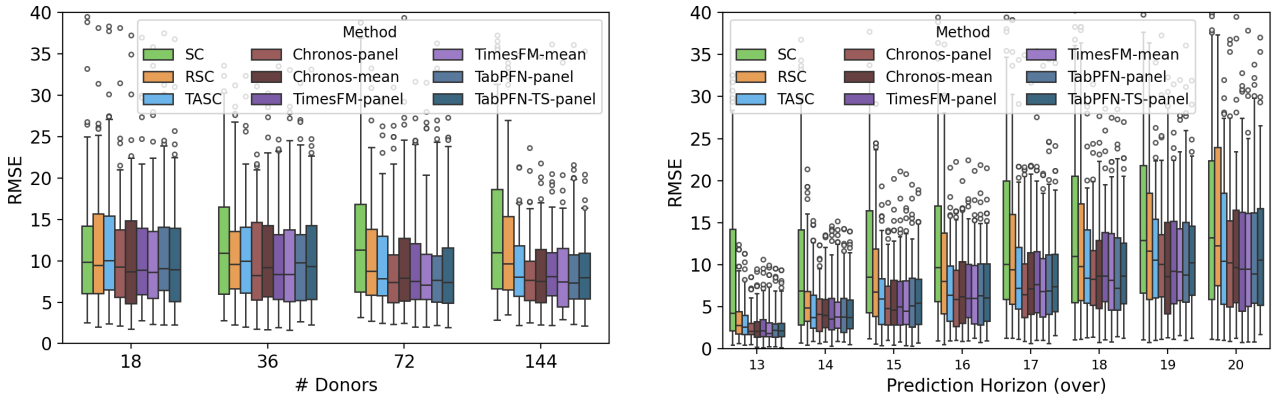
Figure 5a presents overall post-intervention RMSE as the donor pool size varies. In this setting, we observe that `Chronos-panel` achieves the lowest median RMSE compared to `TASC`, `RSC`, `SC`, and other foundation model approaches when $n = 36, 72$ and `TabPFN` achieves the lowest median RMSE when $n = 144$. When $n = 18$ `Chronos-mean` achieves the lowest median RMSE out-performing `Chronos-panel`. Across all donor counts, `TimesFM-mean` achieves a lower median RMSE than `TimesFM-panel`. We observe that `SC` prediction accuracy degrades as n increases with `SC` achieving the highest median RMSE when the donor count is largest ($n = 144$). `TASC` and TSFM approaches all achieve the lowest median RMSE in the setting with the largest donor count ($n = 144$) suggesting that these methods are able to make more accurate predictions with the additional context of more donor time-series. Despite having the lowest median RMSE when $n = 144$, `TabPFN` exhibits a larger variance than `Chronos-panel` at this donor count. Figure 5b presents RMSE in horizon blocks per over (6-balls intervals), when $n = 144$ showing the impact of short-term and long-term prediction horizons.

B.3 Basketball

We next consider NBA game score trajectories, following the same experimental design as in (Rho et al., 2026). The dataset consists of cumulative score trajectories sampled at 15-second intervals, yielding time series of length $T = 192$ for each game. The intervention point is fixed at $T_0 = 96$, corresponding to halftime.

A target game is randomly selected from the full dataset of eligible games. The donor pool consists of the n most recent games prior to the target date, with $n \in \{24, 48, 96, 192\}$. For each donor size and method, the experiment is repeated 100 times using the same set of randomly selected targets. The data is mean-centered using the selected donors prior to fitting. The hidden state dimension and number of singular values to keep were both set to $d = 5$ for `TASC` and `RSC` respectively. `RSC` and `TimesFM` ridge coefficient is set at 10^3 - matching the ridge coefficient in the cricket-data setting.

Figure 6a reports median post-intervention RMSE across donor sizes. In this domain, `Chronos-mean` achieves the lowest median RMSE when $n = 24$ and $n = 48$, `TabPFN-TS` achieves the lowest median RMSE when $n = 96$ and `TASC` achieves the lowest median RMSE when $n = 192$, which is also the lowest median RMSE achieved over all possible donor counts considered. Consistent with performance in the cricket-data setting, we observe that

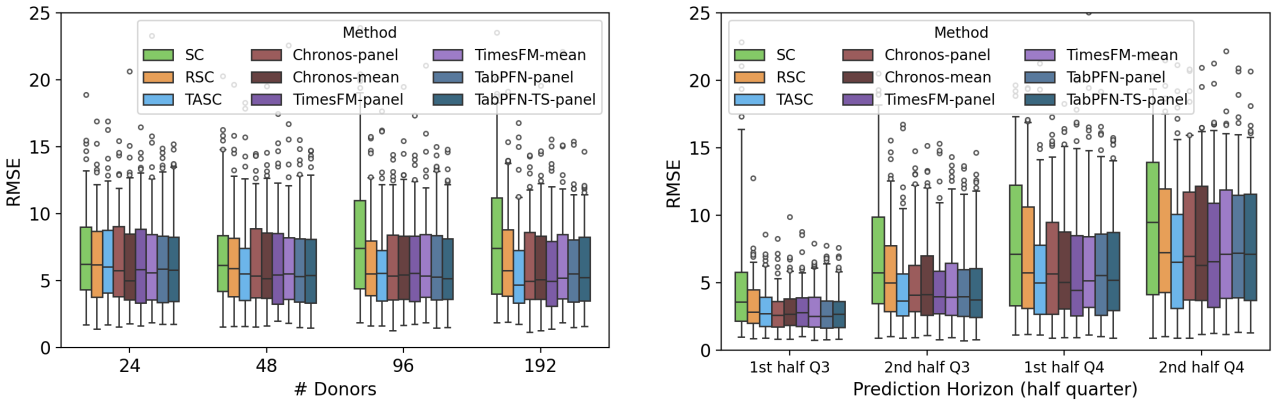


(a) Post-intervention RMSE, IPL

(b) Prediction RMSE per over, IPL ($n = 144$)

Figure 5: Evaluations on cricket (IPL) score trajectory panel data

`SC` exhibits lower prediction accuracy when the number of donors increases. Figure 6b presents RMSE in horizon blocks per half quarter, when $n = 192$, showing the impact of short-term and long-term prediction horizons. `TabPFN` achieves the lowest median RMSE among all methods in short term prediction (predicting the 1st half of the third quarter), but achieves a higher median RMSE than other TSFM approaches and `TASC` in long term prediction (predicting the 2nd half of the fourth quarter).



(a) Post-intervention RMSE, NBA

(b) Prediction RMSE per half-quarter, NBA ($n = 192$)

Figure 6: Evaluations on basketball (NBA) score trajectory panel data

C Ablation Study on Synthetic Data

In this section, we provide additional results on simulated datasets. We evaluate the impact of nonlinearity, correlated and uncorrelated observation noise, temporal trend strength, and pre-intervention period length. First, in C.1 we examine performance when observation noise is correlated (non-diagonal R) in the same settings discussed in Section 4.2 where observation noise is uncorrelated. Then, in C.2, we provide results in settings where observation noise (both correlated and uncorrelated) is low. In C.3 we compare prediction accuracy measured against the target unit's true signal $y_{t,1}^{sig}$ and noisy observation $y_{t,1} = y_{t,1}^{sig} + r_{t,1}$ and evaluate how TSFM approaches perform when pre-intervention period $t = 1, ..., T_0$ increases. Finally in C.4 we analyze observation and latent nonlinearity strengths (α, β) separately. In the experiments reported (unless otherwise stated), we use $T = 100$, $T_0 = 50$, $N = 50$ ($n = 49$), $d_{true} = 5$, and all TSFM approaches are implemented with the `*-panel` input method. We evaluate prediction accuracy on 100 generate datasets for each setting.

The following four observations summarize the main findings of this section. (1) When observation noise is high,

both linear and TSFM approaches (except `TimesFM`) achieve more accurate true signal predictions when observation noise is uncorrelated than when it is correlated. (2) TSFM approaches `TabPFN, TabPFN-TS, TimesFM` offer advantages over linear approaches when nonlinearity strength is large and observation noise is low. (3) Among TSFM approaches, `Chronos` achieves the best true signal prediction accuracy when temporal trends are strong and linear. (4) When observation noise is correlated: `TimesFM` predicts the noisy observation rather than the true signal; `TabPFN` and `TabPFN-TS` trend toward predicting the noisy observation when pre-intervention length increases; and `Chronos` (as well as linear approaches) predict the true signal.

C.1 ABLATION STUDY ON CORRELATED OBSERVATION NOISE

In this section, we evaluate prediction accuracy on synthetic datasets where observation noise covariance matrix R is non-diagonal and thus observation noise between donors and target series is correlated. We first evaluate prediction accuracy measuring against the target unit's true signal $y_{t,1}^{sig}$ and evaluate prediction accuracy to the noisy observation $y_{t,1} = y_{t,1}^{sig} + r_{t,1}$ in C.3.

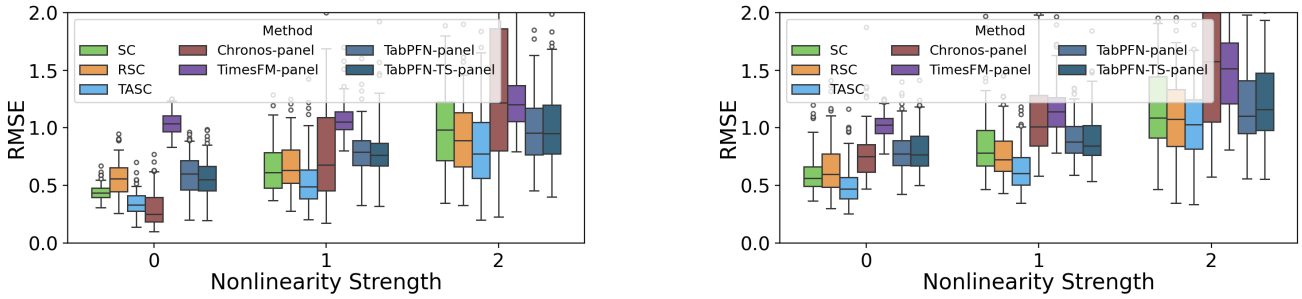


Figure 7: Post-intervention RMSE with varied nonlinearity strength when temporal trend is strong (left) and weak (right). In both cases, observation noise is large and correlated.

First, we vary the nonlinearity strength ($\alpha = \beta$) in the simulated data. Figure 7 presents the post-intervention RMSE under strong (left panel, small Q) and weak (right panel, large Q) temporal trends when correlated observation noise is high (large non-diagonal R). All methods tend to yield lower prediction error when the temporal trend is strong than when it is weak.

In the linear setting ($\alpha = \beta = 0$) with strong temporal trend, `Chronos` achieves the lowest median RMSE among TSFM and linear approaches. In this setting, all methods except `TimesFM`, achieve a higher median RMSE with correlated observation noise than is achieved in the comparable setting where observation noise is uncorrelated (diagonal R). While still achieving the highest median RMSE among all methods in the correlated observation noise setting, `TimesFM` achieves a significantly lower median RMSE than when observation noise is uncorrelated. The performance of `Chronos` degrades the most as nonlinearity strength increases and `Chronos` achieves the highest median RMSE when nonlinearity strength is high ($\alpha = \beta = 2$).

Consistent with performance when observation noise is uncorrelated, `TASC` achieves the lowest median RMSE across all levels of nonlinearity strength when the temporal trend is weak (large Q) and observation noise is high (large R).

Next, we vary the hidden state dimension $d_{true} \in \{2, 5, 10, 20\}$ while fixing nonlinearity strength $\alpha = \beta = 0$ and evaluate post-intervention RMSE when observation noise is correlated.

Figure 8 shows the post-intervention RMSE under strong (left, small Q) and weak (right, large Q) temporal trends when correlated observation noise is high (large non-diagonal R).

When temporal trend is strong, `Chronos` achieves the lowest median RMSE among TSFM and linear approaches when $d_{true} \in \{5, 10, 20\}$ and `TASC` achieves the lowest median RMSE when $d_{true} = 2$. While the performance of `TASC` and `RSC` degrades as the latent dimension increases, TSFM approaches and `SC` are less sensitive to variations in the latent dimension. The performance of `Chronos` improves when the latent dimension increases with `Chronos` achieving it's lowest median RMSE when $d_{true} = 20$.

When temporal trend is weak, linear approaches `TASC` and `SC` achieve lower median RMSE than TSFM approaches across all $d_{true} \in \{2, 5, 10, 20\}$. However, the performance of linear approaches degrades as d_{true} increases -

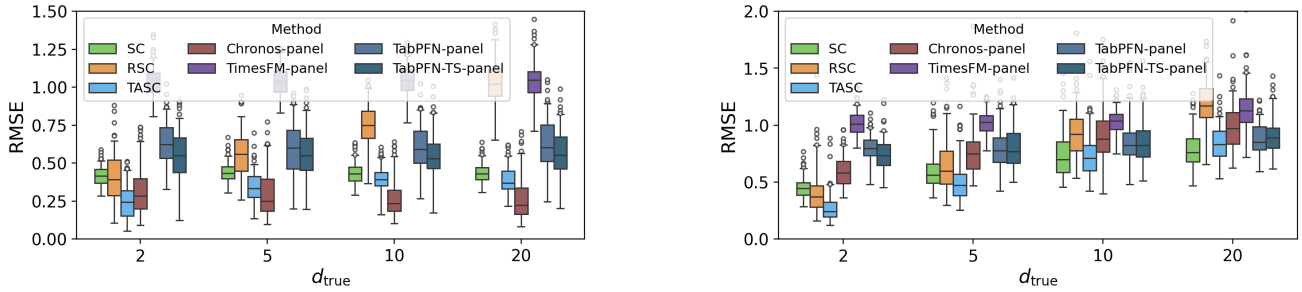


Figure 8: Post-intervention RMSE with varied hidden state dimension (d_{true}) when temporal trend is strong (left) and weak (right). In both cases, observation noise is large and correlated.

showing a strong performance improvement over TSFM approaches when $d_{true} = 2$ but similar median RMSE when $d_{true} = 20$. TSFM approaches are less sensitive to variations in the latent dimension d_{true} compared to linear approaches.

C.2 Settings With Low Observation Noise

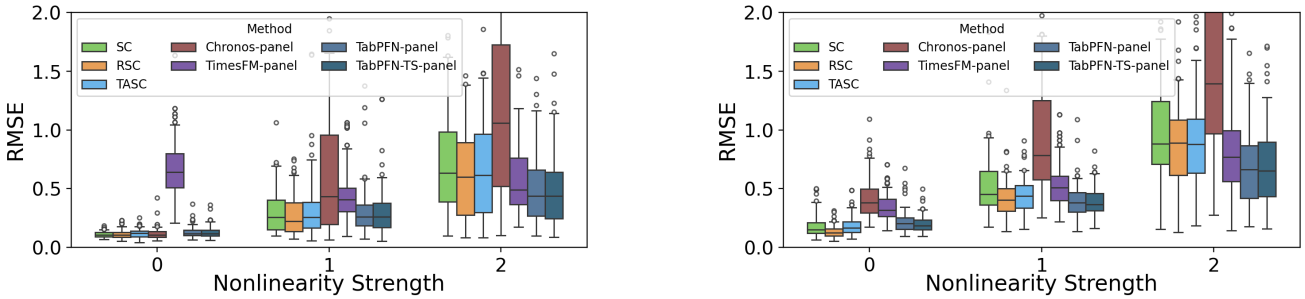


Figure 9: Post-intervention RMSE with varied nonlinearity strength ($\alpha = \beta$) when temporal trend is strong (left) and weak (right). In both cases, observation noise is small and uncorrelated.

In this section, we evaluate the impact of nonlinearity strength and latent dimension when observation noise (both uncorrelated and correlated) is low.

Figure 9 shows the results of each approach when observation noise is low and uncorrelated (small diagonal R), in different settings of temporal perturbation and nonlinearity strength.

TSFM approaches (`TabPFN`, `TabPFN-TS`, and `TimesFM`) achieve lower median RMSE than linear SC approaches when nonlinearity is large ($\alpha = \beta = 2$) with `TabPFN` and `TabPFN-TS` achieving the lowest median RMSE among TSFM approaches across all levels of nonlinearity strength.

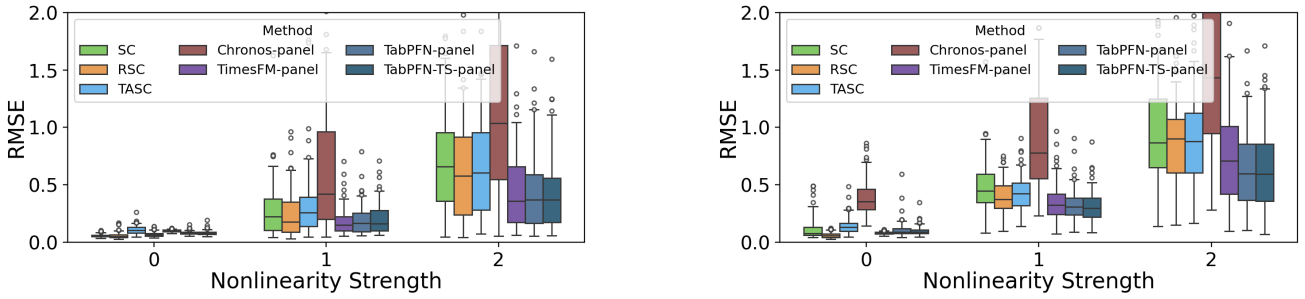


Figure 10: Post-intervention RMSE with varied nonlinearity strength when temporal trned is strong (left) and weak (right). In both cases, observation noise is small and correlated.

Figure 10 provides results when observation noise is correlated (non-diagonal small R). Compared to setting

with uncorrelated observation noise, `TimesFM` post-intervention RMSE is significantly lower with correlated observation noise. In both cases, where the temporal trend is strong (left, small Q) and weak (right, large Q), TSFM approaches `TabPFN`, `TabPFN-TS`, and `TimesFM` outperform linear approaches when nonlinearity strength increases and these improvements are more significant when observation noise is correlated.

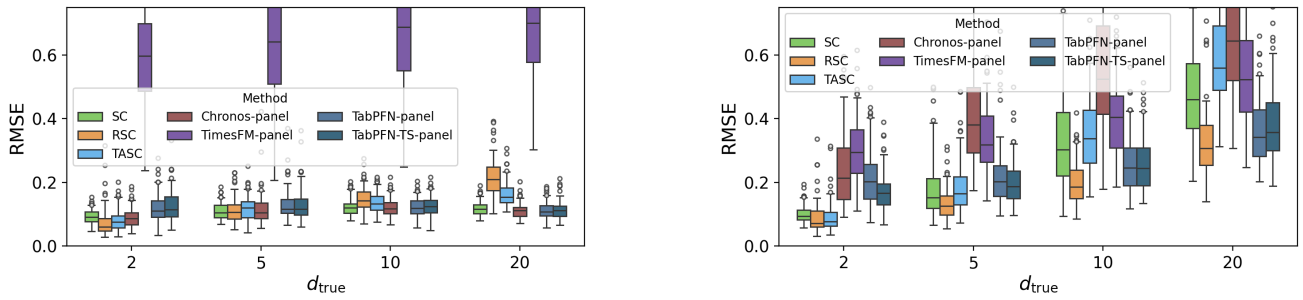


Figure 11: Post-intervention RMSE with varied hidden state dimension (d_{true}) when temporal trend is strong (left) and weak (right). In both cases, observation noise is small and uncorrelated.

We next consider the impact of the hidden latent dimension d_{true}. Figure 11 shows the results of each method when we vary the hidden dimension d_{true} of the data generation process in the linear setting ($\alpha = \beta = 0$) when observation noise is uncorrelated and low (small diagonal R). We measure post-intervention RMSE with $d_{true} \in \{2, 5, 10, 20\}$ and we set the hidden dimension of `TASC` and the number of singular values to keep for `RSC` as $d = d_{true}$.

When there is a strong temporal trend (small Q) and the rank increases, TSFM approaches outperform linear approaches: `Chronos` achieves the smallest median RMSE across all methods when $d_{true} \in \{10, 20\}$. On the other hand, when there is a weaker temporal trend (large Q) TSFM approaches offer no advantage over linear approaches with `RSC` achieving the lowest median RMSE at all settings of d_{true}.

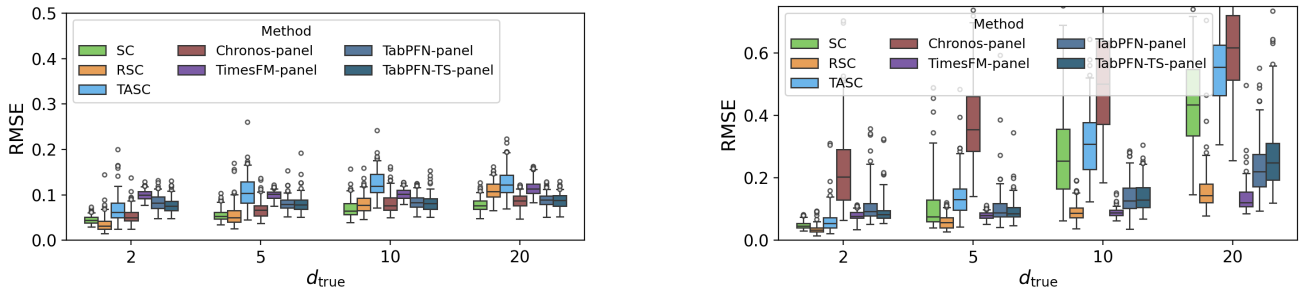


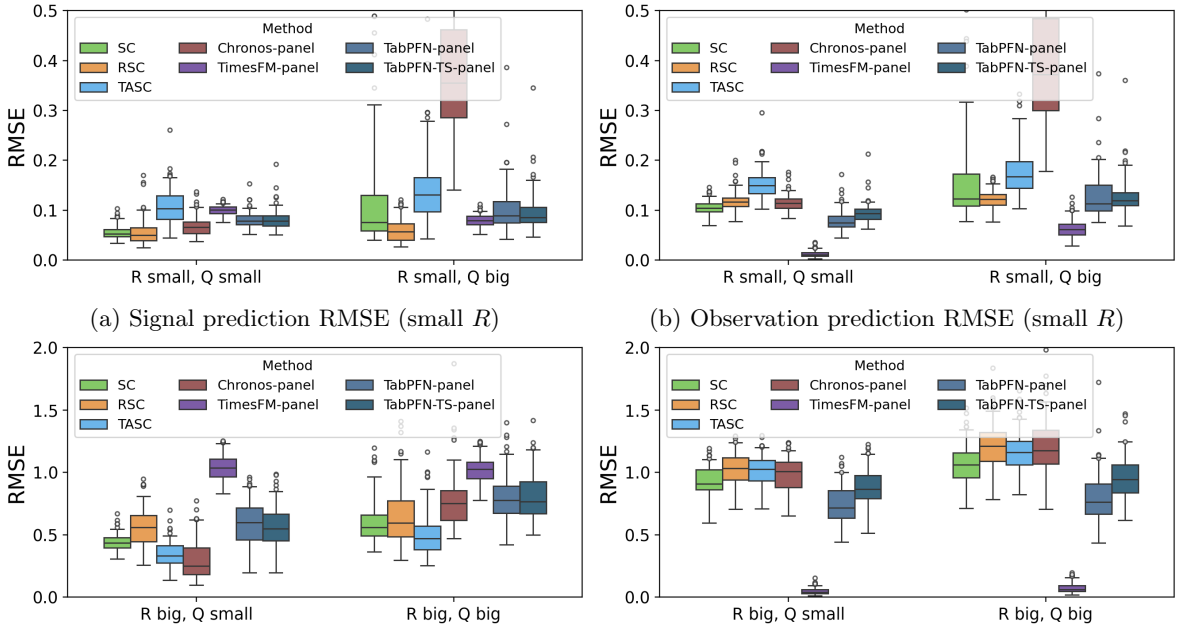
Figure 12: Post-intervention RMSE with varied hidden state dimension (d_{true}) when temporal trend is strong (left) and weak (right). In both cases, observation noise is small and correlated.

Figure 12 shows the post-intervention RMSE under strong (left, small Q) and weak (right, large Q) temporal trends when observation noise is low and correlated (small non-diagonal R). `TimesFM` achieves a significantly lower median RMSE when observation noise is correlated than when it is uncorrelated with `TimesFM` achieving lower median RMSE than other TSFM approaches when there is a weak temporal trend across all $d_{true} \in \{2, 5, 10, 20\}$.

C.3 True signal vs noisy observation prediction

In this section, we report post-intervention prediction accuracy when evaluated against the true signal and against the noisy observation. When observation noise is correlated across units, these become distinctly different prediction tasks.

In the linear setting ($\alpha = \beta = 0$), this distinction can be formalized by comparing the Bayes-optimal predictors of $y_{t,1}^{\text{sig}}$ and $y_{t,1}$ given the observed data. To account for the missing post-intervention target observations, partition the observations and parameters as $y_t = \begin{bmatrix} y_{t,1} \\ y_{t,2} \end{bmatrix}$, $r_t = \begin{bmatrix} r_{t,1} \\ r_{t,2} \end{bmatrix}$, $H = \begin{bmatrix} h_1^\top \\ H_2 \end{bmatrix}$, and $R = \begin{bmatrix} R_1 & R_{12} \\ R_{12}^\top & R_2 \end{bmatrix}$, where



(a) Signal prediction RMSE (small R)

(b) Observation prediction RMSE (small R)

(c) Signal prediction RMSE (large R)

(d) Observation prediction RMSE (large R)

Figure 13: Post-intervention RMSE measured against the true signal and noisy observation in settings with small and large correlated observation noise.

$y_{t,1}, r_{t,1} \in \mathbb{R}$, $y_{t,2}, r_{t,2} \in \mathbb{R}^n$, $h_1 \in \mathbb{R}^d$, $H_2 \in \mathbb{R}^{n \times d}$, $R_1 \in \mathbb{R}$, $R_{12} \in \mathbb{R}^{1 \times n}$, and $R_2 \in \mathbb{R}^{n \times n}$. Then the target and donor observation equations can be written as $y_{t,1} = h_1^\top x_t + r_{t,1}$ and $y_{t,2} = H_2 x_t + r_{t,2}$.

Let $\mathcal{D} = \{y_{1:T_0,1}, y_{1:T,2}\}$ denote the observed data, and let $\hat{x}_t = \mathbb{E}[x_t \mid \mathcal{D}]$. The Bayes-optimal predictor of the target signal $y_{t,1}^{\text{sig}} = h_1^\top x_t$ is $\mathbb{E}[y_{t,1}^{\text{sig}} \mid \mathcal{D}] = \mathbb{E}[h_1^\top x_t \mid \mathcal{D}] = h_1^\top \hat{x}_t$. For the noisy observation $y_{t,1}$, correlated observation noise implies that, conditional on x_t, the donor residual $y_{t,2} - H_2 x_t$ carries information about the target noise $r_{t,1}$. Because $r_{t,1}$ and $r_{t,2}$ are jointly Gaussian with covariance blocks R_1, R_{12}, and R_2, the conditional mean of $r_{t,1}$ given $r_{t,2}$ is $R_{12} R_2^{-1} r_{t,2}$. It follows that $\mathbb{E}[y_{t,1} \mid \mathcal{D}] = h_1^\top \hat{x}_t + R_{12} R_2^{-1}\big(y_{t,2} - H_2 \hat{x}_t\big)$. Therefore, $\mathbb{E}[y_{t,1} \mid \mathcal{D}] - \mathbb{E}[y_{t,1}^{\text{sig}} \mid \mathcal{D}] = R_{12} R_2^{-1}\big(y_{t,2} - H_2 \hat{x}_t\big)$. Thus, when $R_{12} = 0$, the Bayes-optimal predictors coincide, whereas when $R_{12} \neq 0$, predicting the noisy target includes an additional donor-residual correction induced by correlated observation noise.

Figure 13 shows post-intervention RMSE for signal prediction (left) and noisy-observation prediction (right) in the linear setting ($\alpha = \beta = 0$) under small correlated observation noise (top) and large correlated observation noise (bottom). The linear methods `SC`, `RSC`, and `TASC`, as well as `Chronos`, achieve lower median RMSE for signal prediction than for noisy-observation prediction, with the gap becoming more pronounced when observation noise is high (R is large). In contrast, `TimesFM` achieves substantially lower RMSE for noisy-observation prediction than for signal prediction. In the high observation noise setting, `TimesFM` has the highest median signal prediction RMSE among all methods but the lowest median noisy observation prediction RMSE, suggesting a stronger exploitation of correlated observation noise rather than recovery of the true latent signal. `TabPFN` and `TabPFN-TS` show more similar RMSEs across signal and observation prediction. In Section C.3.1, we show that as the pre-intervention period increases, `TabPFN` and `TabPFN-TS` also trend toward more accurate noisy observation prediction relative to true signal prediction.

C.3.1 Ablation study on pre-intervention period length

In this section, we continue our evaluation of true signal and noisy observation prediction and analyze the impact of correlated and uncorrelated observation noise as pre-intervention period ($t = 1, ..., T_0$) increases.

Figure 14 shows prediction RMSE for true signal (left) and noisy observation (right) when observation noise is high and correlated (large non-diagonal R). When $T_0 = 50$ all methods except `TimesFM` achieve a lower signal

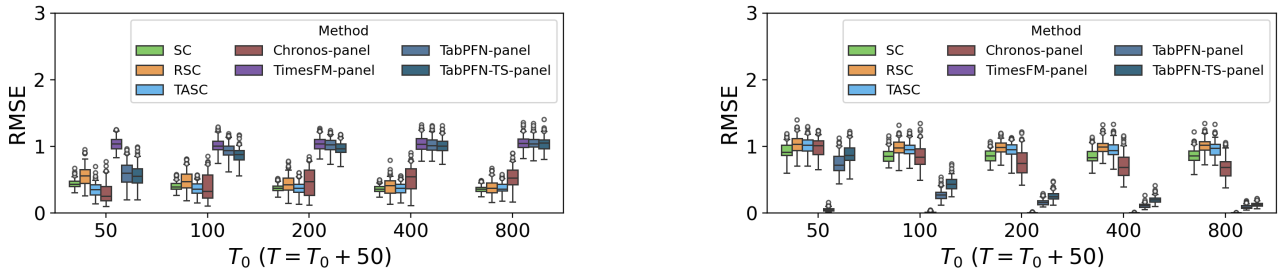


Figure 14: Post-intervention RMSE with varied pre-intervention period length T_0 when data generation is linear and observation noise is correlated. Prediction RMSE is measured against the true signal (left) and noisy observation (right).

prediction RMSE than noisy observation prediction RMSE. However, when T_0 increases, `TabPFN` and `TabPFN-TS` trend toward predicting the noisy observation, achieving comparable true signal prediction RMSE to `TimesFM`. Linear approaches and `Chronos` achieve lower true signal prediction RMSE than observation prediction RMSE across all pre-intervention lengths T_0. `Chronos` achieves the lowest median true signal prediction RMSE when $T_0 = 50$ but accuracy degrades as T_0 increases and `Chronos` achieves a higher true signal prediction RMSE than linear approaches when $T_0 = 800$. In General, TSFM approaches trend toward predicting the noisy observation rather than the true signal when pre-intervention period increase with `TimesFM, TabPFN, TabPFN-TS` showing significant preference to noisy observation prediction.

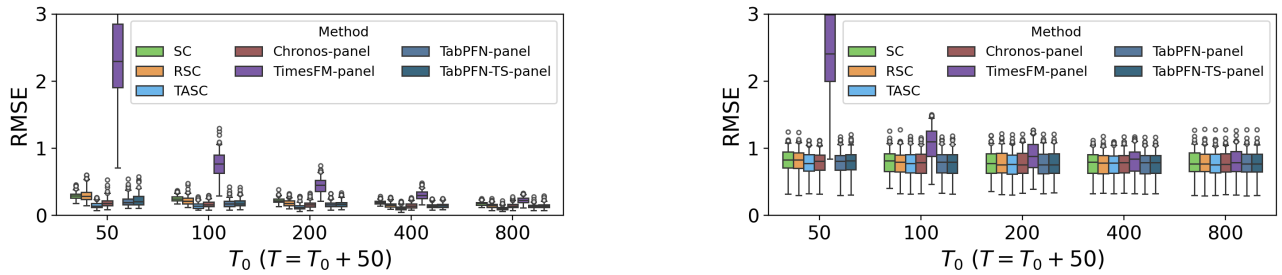


Figure 15: Post-intervention RMSE with varied pre-intervention period length T_0 when data generation is linear and observation noise is uncorrelated. Prediction RMSE is measured against the true signal (left) and noisy observation (right).

Figure 15 shows prediction RMSE in the same setting as above but when observation noise is uncorrelated (large diagonal R). Unsurprisingly, all methods achieve a lower signal prediction RMSE than noisy observation prediction RMSE when observation noise is uncorrelated. `TASC` achieves the lowest median true signal prediction RMSE when observation noise is uncorrelated - which is likely the result of `TASC` assumption that observation noise is uncorrelated. `Chronos, TabPFN`, and `TabPFN-TS`, achieve lower true signal prediction RMSE than `SC` and `RSC` showing that in the absence of correlated observation noise, these approaches can recover the true signal comparably to linear SC approaches.

Figure 16 (correlated observation noise) and 17 (uncorrelated observation noise) show prediction RMSE in the same two settings as discussed above but with non-linearity strength $\alpha = \beta = 2$. In the nonlinear setting, predicting the noisy observation and the true signal is a more difficult task, but the general trend persists: `TimesFM, TabPFN`, and `TabPFN-TS` trend toward predicting the noisy observation when observation noise is correlated and all methods achieve a lower true signal prediction RMSE than noisy observation prediction RMSE when R is uncorrelated. When observation noise is correlated and $T_0 = 50$, `Chronos` achieves the highest median RMSE but as T_0 increase achieves a lower signal prediction RMSE than other TSFM methods. When observation noise is uncorrelated (Figure 17) `TabPFN`, and `TabPFN-TS` achieve lower true signal prediction RMSE than linear approaches when T_0 increases showing, again, that in the absence of correlated noise, TSFM approaches `TabPFN` and `TabPFN-TS` can recover the true signal.

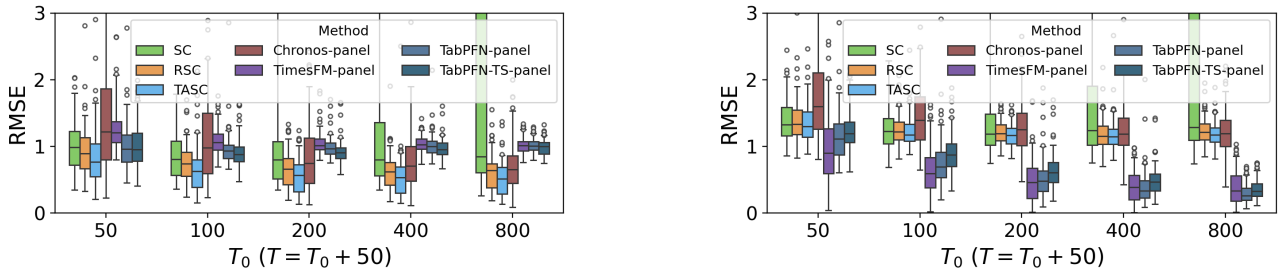


Figure 16: Post-intervention RMSE with varied pre-intervention period length T_0 when data generation is nonlinear and observation noise is correlated. Prediction RMSE is measured against the true signal (left) and noisy observation (right).

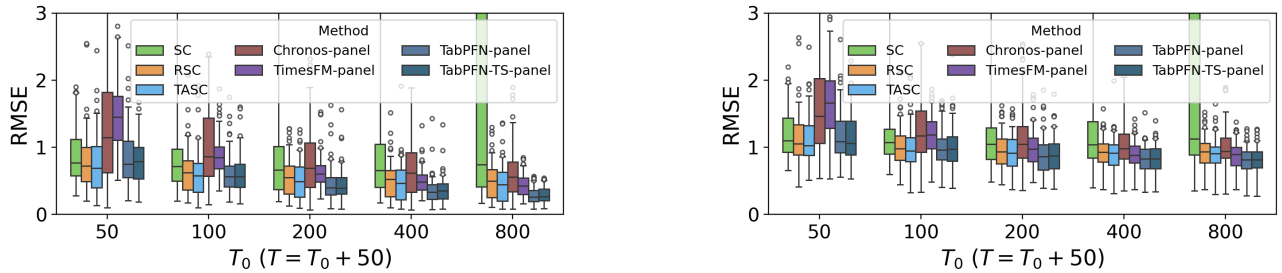


Figure 17: Post-intervention RMSE with varied pre-intervention period length T_0 when data generation is nonlinear and observation noise is uncorrelated. Prediction RMSE is measured against the true signal (left) and noisy observation (right).

C.4 Comparing Observation and Latent nonlinearity

In this section, we study the impact of latent and observation nonlinearity separately. To evaluate latent nonlinearity we vary $\alpha \in \{0, 1, 2\}$ when fixing observation nonlinearity strength $\beta = 0$. Similarly, we fix latent nonlinearity strength $\alpha = 0$ and vary $\beta \in \{0, 1, 2\}$ to evaluate observation nonlinearity.

Figure 18 shows the results of observation and latent nonlinearity when observation noise R is small in settings where temporal perturbation Q is both small and large. Linear approaches (`SC`, `RSC`, and `TASC`) achieve lower median RMSE than TSFM approaches in settings with only latent nonlinearity. `TabPFN` and `TabPFN-TS` achieve lower median RMSE than linear approaches in settings with high observation nonlinearity and high temporal perturbation (large Q).

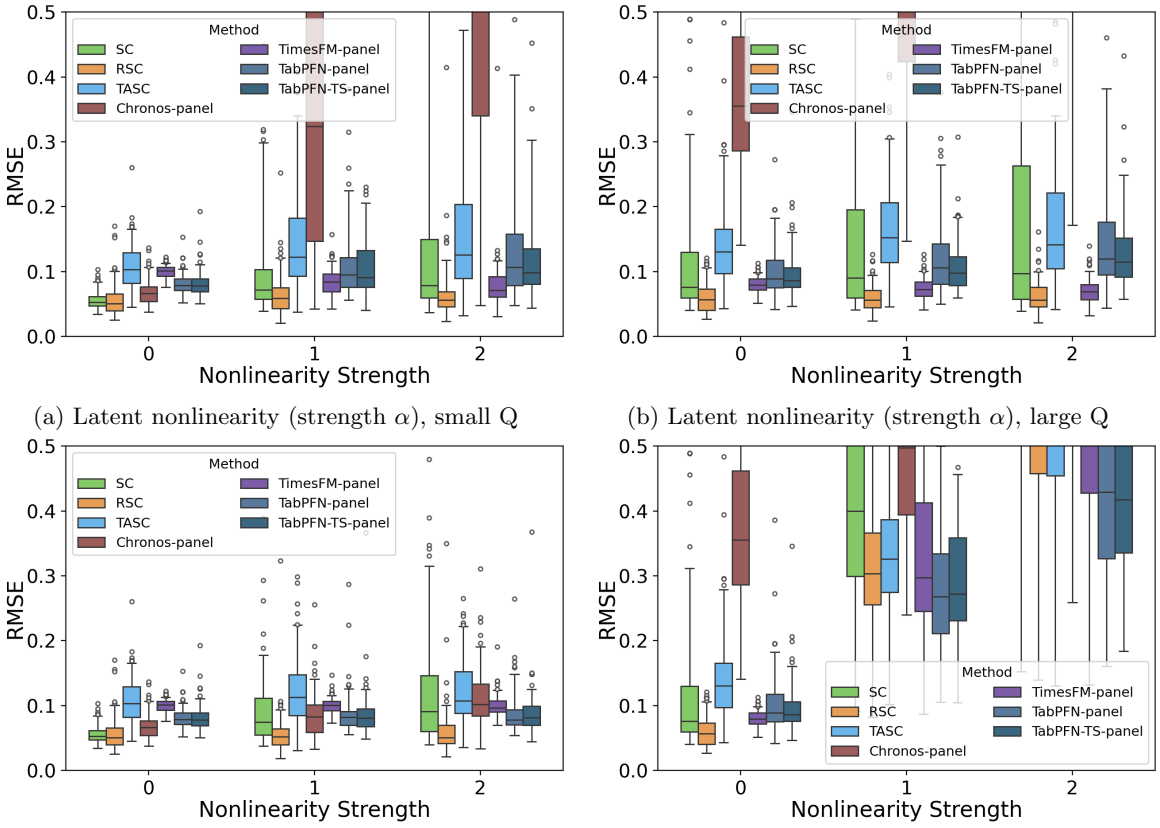
Figure 19 shows the results of observation and latent nonlinearity when observation noise R is large in settings where temporal perturbation Q is both small and large. We observe that when observation noise is large, TSFM approaches offer no benefits (and often show higher median RMSE) when compared to linear approaches (`SC`, `RSC`, and `TASC`).

Across these different settings of nonlinearity, we find that the performance of `Chronos` is very sensitive to the strength of both observation and latent nonlinearity with its median RMSE increasing significantly across many settings when nonlinearity strength increases from 1 to 2.

D Additional Ablation Tests on (`*-flatten`) Input Method

In this section, we provide further explanation and evaluation of the (`*-flatten`) input method described in Section 3.

Let the target unit be indexed by $i = 1$ and donors by $i \in \{2, \ldots, n + 1\}$. Instead of treating each donor time series $Y_{i,1:T}$ as a separate covariate, we construct a single flattened sequence that interleaves donor and target values across time.



(a) Latent nonlinearity (strength α), small Q

(b) Latent nonlinearity (strength α), large Q

(c) Observation nonlinearity (strength β), small Q

(d) Observation nonlinearity (strength β), large Q

Figure 18: Post-intervention RMSE on simulation data with low correlated observation noise (small non-diagonal R)

We flatten the pre-intervention outcome matrix $Y_{1:n+1,1:T_0}$ in time-major order and append donor outcomes at $T_0 + 1$, yielding

$$\big[Y_{n+1,1}, \ldots, Y_{2,1}, Y_{1,1},\ Y_{n+1,2}, \ldots, Y_{2,2}, Y_{1,2},$$
$$\ldots,\ Y_{n+1,T_0}, \ldots, Y_{2,T_0}, Y_{1,T_0},\ Y_{n+1,T_0+1}, \ldots, Y_{2,T_0+1}\big].$$

This flattened sequence is treated as the historical context for a single "target" series, and the TSFM is tasked with predicting the next value $\hat{Y}_{1,T_0+1}$.

`Chronos`(v2) and `TimesFM` process inputs using non-overlapping patches of fixed length P. When the total number of units equals the patch length ($n + 1 = P$), each time slice aligns exactly with one patch.

Under this alignment, each patch corresponds to a full cross-sectional snapshot:

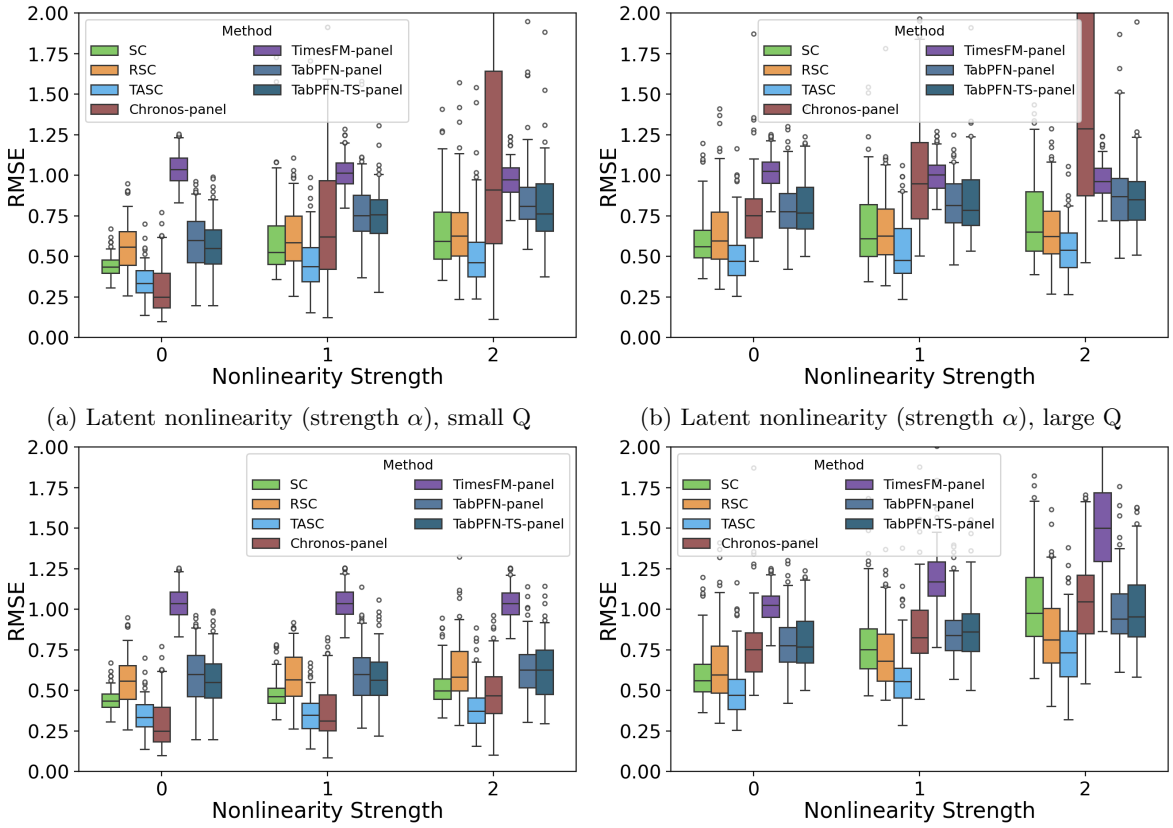
$$\big[Y_{n+1,t}, Y_{n,t}, \ldots, Y_{2,t}, Y_{1,t}\big].$$

After shifting by one step for autoregressive prediction, the effective predictive patch takes the form

$$\big[Y_{1,t-1}, Y_{n+1,t}, Y_{n,t}, \ldots, Y_{2,t}\big],$$

which combines the previous target value with the current donor outcomes.

Because the default patch size in `Chronos`(v2) is $P = 16$ and for `TimesFM` it is $P = 32$, we run experiments with $n = 15$ and $n = 31$ so that $n + 1 = 16$ and $n + 1 = 32$ for `Chronos`(v2) and `TimesFM` respectively, achieving exact patch alignment. In this regime, the patch embedding mechanism encodes the entire donor cross-section jointly within a single token embedding. This allows the model to process cross-sectional donor structure in a structured manner before applying temporal attention across patches.



(a) Latent nonlinearity (strength α), small Q

(b) Latent nonlinearity (strength α), large Q

(c) Observation nonlinearity (strength β), small Q

(d) Observation nonlinearity (strength β), large Q

Figure 19: Post-intervention RMSE on simulation data with high correlated observation noise (large non-diagonal R)

Figure 20 shows results for three different data-input methods `*-panel`, `*-flatten`, and `*-mean` for `Chronos` and `TimesFM`. For both methods, we run experiments where $T_0 = 8, 16, 32, 64$, $T = T_0 + 16$, $d_{true} = 5$, observation noise is high (large R) and temporal perturbation is low (small Q). To ensure that the flattened time-series will be aligned with the patch size of each method, we set $n = 15$ for `Chronos` and $n = 31$ for `TimesFM`. From these results we observe that input `Chronos-flatten` outperforms `Chronos-panel`. However, `TimesFM-panel` outperforms `TimesFM-flatten` and `TimesFM-mean` across all $T_0 \in \{8, 16, 32, 64\}$. The success of `Chronos-flatten` in this setting without any finetuning to this particular data preparation scheme suggests that this input method could be useful for SC when pre-intervention periods are short and future work is required to finetune the model to this particular data preparation mode.

To further study if the benefits of `Chronos-flatten` are the result of patch alignment we next provide results on varying the number of donors n for different methods of input for `Chronos`. Figure 21 reports post-intervention RMSE for `Chronos-panel`,`*-flatten`,`*-mean` across $n \in \{10, 15, 20\}$.

We observe that `*-flatten` achieves a lower median RMSE than `Chronos-panel` when $n = 15$, the case in which patch alignment holds ($n + 1 = P$). However, `*-flatten` performs worse than `Chronos-panel` when $n = 10$ or $n = 20$, where this alignment is broken.

These results indicate that the performance gains of (`*-flatten`) depend on the structural alignment with the TSFM's patch embedding mechanism. When $n + 1 \neq P$, donor time slices are split across patches, disrupting the cross-sectional encoding and degrading predictive performance.

This experiment highlights that `Chronos`' inductive bias is sensitive not only to the presence of donor information but also to how that information aligns with the model's fixed patching architecture.

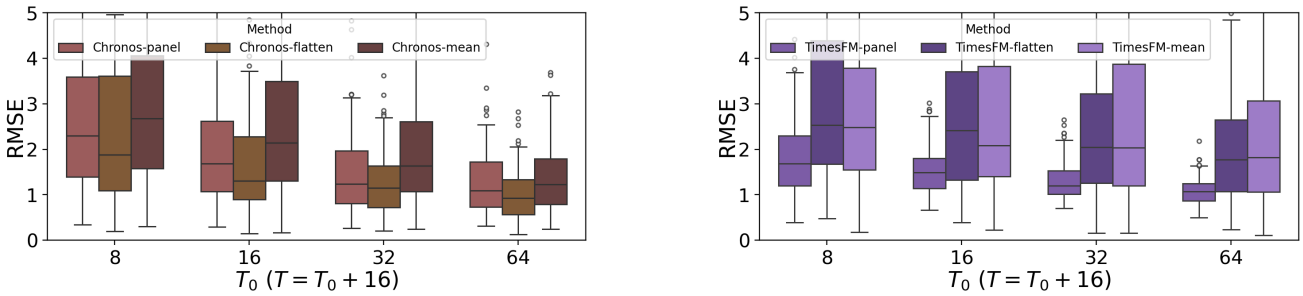


Figure 20: Post-intervention RMSE with varied T_0 and $T = T_0 + 16$, nonlinearity strength ($\alpha = \beta = 2$), large R, small Q and different data prep methods (left, `Chronos` with $n = 15$) and (right, `TimesFM` with $n = 31$).

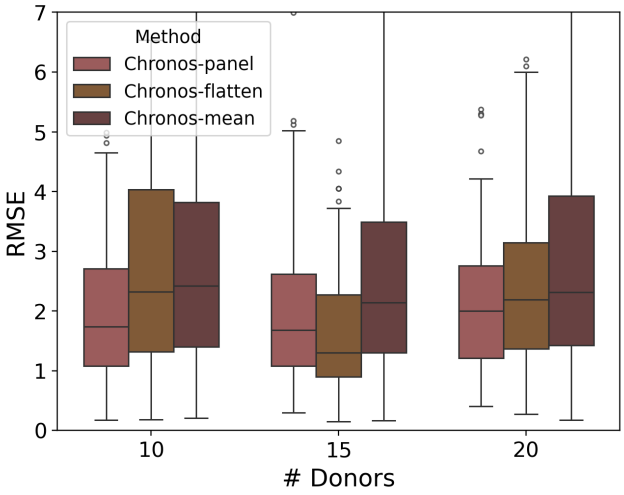


Figure 21: Post-intervention RMSE for different variants of `Chronos` input at different number of donors $n \in \{10, 15, 20\}$. Patch alignment for `Chronos-flatten` when $n = 15$