When Is a Known Future Covariate an Intervention? Decision Feedback, an Accuracy–Validity Identity and a Placebo Audit for Foundation-Model What-If Forecasts

Anonymous submission

Abstract

Time-series foundation models (TSFMs) accept known future covariates, and practitioners use them as what-if simulators: supply a different future price, setpoint or dose and read off the forecast. We ask when such a scenario forecast answers the interventional question. Conditioning on a supplied decision path equals intervening on it only if no supplied decision responded to outcome shocks after the forecast origin; for plans fixed in advance, the model must also know when the plan was issued. When decisions react inside the window, the most accurate forecaster gives biased and often wrong-signed what-if answers, and its factual-accuracy advantage over the true causal model equals their mean squared disagreement, so factual benchmarks rank it first. We propose a placebo audit that needs neither ground truth nor declared effect signs: shift only decisions that act after the forecast target. In a pilot of 110k forecasts (pre-registered for Chronos-2 and TimesFM-2.5; Timestral-209M added later and confirmed on fresh seeds), under reactive control, feedback reversed the sign of all three models' what-if answers (to $-4.4\times$ the true effect at the longest context), the bias did not fade with longer context, and no model showed awareness of when plans were issued. The audit flagged 22/22 displaced cells of the model that reads the whole future window (20/21 on fresh seeds), with no false flags on feedback-free replays; it was borderline for Chronos-2 and, as predicted, blind to TimesFM's same-time covariate pathway. A building-thermal stress test with switching controllers reproduces the displacement for Chronos-2 and Timestral, and exposes a second limit of the audit: future decisions can stand in for omitted drivers.

1 Introduction

Covariate-aware TSFMs are marketed and used for planning. Chronos-2 lists known future covariates as a native capability (Ansari et al. 2025), TimesFM-2.5 (Das et al. 2023) supports covariates through in-context regression (Won 2026), Timestral advertises known-future covariates read through covariate-aware attention (Mistral AI 2026), and vendor tutorials show what-if price scenarios on retail data (Nixtla 2026). The same tutorial warns that because prices "were typically set in response to demand rather than at random," the measured price–demand relationship is "an association, not a causal effect." Covariate benchmarks nevertheless score factual accuracy (Shchur et al. 2025), and recent work reports what-if failures of TSFMs on closed-loop and confounded data (Lu 2026; Amria et al. 2026; Won 2026; Wang 2026), finds factual skill an unstable guide to interventional validity (Lu 2026), and traces sign reversal in trained world models to reactive controllers (Zhang et al. 2026). As far as we can tell (for Lu (2026) from its abstract only), missing are a condition for when a zero-shot scenario forecast is interventional, an exact account of why factual benchmarks reward the biased model, and a check that needs no declared effect signs. For example, a thermostat heats when the room is cold, so on its logs more heating predicts a colder room.

This paper makes four contributions.

1. **A timing rule.** A scenario forecast is interventional when the supplied decision window is exogenous: no supplied decision depends on outcome shocks after the origin. Reactive control, receding-horizon re-planning and plans re-issued inside the window violate it. A plan fixed at the origin satisfies it only if the model can tell when the plan was issued; a model pooled over plan phases is biased even at issue times.
2. **An accuracy–validity identity.** The full-path conditioner's factual-MSE advantage over the true causal model equals their mean squared disagreement; any forecaster's gain requires leakage, which is zero on exogenous windows.
3. **A placebo audit** that needs no ground truth and no declared effect signs, with its blind spots stated.
4. **An evaluation against replay twins**, pre-registered for Chronos-2 and TimesFM-2.5 and extended under the same frozen rules to Timestral-209M (confirmed on fresh seeds), with an independent recomputation and a building-thermal stress test. Feedback reverses all three models' what-if answers under reactive control, longer context does not help, and factual skill ranks the most biased models first.

2 When Conditioning Equals Intervening

Setting. An outcome Y is driven by decisions A that act with a one-step delay. A forecaster at origin t sees the history $\mathcal{H}_t = (Y_{\le t}, A_{<t})$ and a supplied window $A_{t:t+H-1}$,

This is an anonymized submission for review purposes only. Distribution, citation, or public sharing of this manuscript is strictly prohibited. Copyright and publication details will appear in the final version if accepted.

where A_s acts on Y_{s+1}. Let $I_t = (\mathcal{H}_t, A_{t:t+H-1})$. The *full-path conditioner* $F_h = \mathbb{E}[Y_{t+h} \mid I_t]$ is what factual training targets. The *interventional forecast* $C_h(a) = \mathbb{E}[Y_{t+h} \mid \mathcal{H}_t, \mathrm{do}(A_{t:t+H-1} = a)]$ is what a planner needs; write C_h for C_h evaluated at the realised window. A what-if response is the change in a forecast per unit sustained shift of the supplied decisions.

Timing rule. Suppose $Y_{s+1} = g(Y_{\le s}, A_{\le s}) + \varepsilon_{s+1}$ with innovations independent of the past (no hidden confounder). If the window is exogenous, $A_{t:t+H-1} \perp \varepsilon_{t+1:t+H} \mid \mathcal{H}_t$, then conditioning on the window leaves the distribution of future shocks unchanged and $F_h = C_h$. If a supplied decision depends on a post-origin shock, conditioning on it reveals that shock and $F_h \neq C_h$ generically. This is the strict-exogeneity condition behind Granger–Sims tests (Sims 1972; Chamberlain 1982), applied to the forecast window. Dropping the decisions that act after the target (*causally ordered* conditioning) does not repair it beyond one step, because intermediate decisions respond to intermediate outcomes; only rolling a one-step causal model forward (g-computation) does (Melnychuk, Frauen, and Feuerriegel 2022; Xiong et al. 2024; Zare et al. 2026).

Exact illustration. Take $Y_{t+1} = aY_t + bA_t + \sigma e_{t+1}$ with $b = 0.4$, $\sigma = 1$, a stabilising policy gain $k = 0.5$ and closed-loop coefficient $a - bk = 0.5$, so the true one-step effect is 0.40. Decisions follow one of four regimes: open loop $A_t = s\,d_t$; reactive control $A_t = -kY_t + s\,d_t$; rolling re-planning $A_t = -k\mathbb{E}[Y_t \mid \text{past}] + s\,d_t$; and block plans issued every 8 steps. Here s is excitation and d_t is i.i.d. noise. We compute population responses exactly from Gaussian loadings. At $s = 0.3$ the full-path conditioner's one-step response is -1.58 under reactive control and -2.31 under rolling re-planning; it is exact in open loop and for block plans forecast from their issue time. Moving the origin 0–7 steps past issue gives $0.40, 0.40, 0.11, -0.09, -0.37, -0.78, -1.45, -2.95$ (Fig. 1a). A conditioner pooled over plan phases, the population limit of an in-context learner that is not told the issue time, gives -1.13 at every origin; with 64-step plans it still gives 0.11.

Accuracy–validity identity. C_h is a function of I_t and F_h is the conditional mean given I_t, so $Y - F_h$ is orthogonal to $F_h - C_h$ and

$$\mathrm{MSE}(C_h) - \mathrm{MSE}(F_h) = \mathbb{E}\big[(F_h - C_h)^2\big]. \qquad (1)$$

For any forecaster G built from I_t, $\mathrm{MSE}(C_h) - \mathrm{MSE}(G) = 2L - D$ with $D = \mathbb{E}[(G - C_h)^2]$ and leakage $L = \mathbb{E}[(Y - C_h)(G - C_h)]$. On an exogenous window $Y - C_h$ consists of future shocks independent of I_t, so $L = 0$ and no forecaster beats the causal model. Under feedback, every factual gain over the causal model is bought with leakage, and the most accurate forecaster is the one whose what-if answers disagree most with the truth. In the linear model the identity holds to 10^{-9} in all 264 cells computed, and F beats C on factual MSE by 2.3–87.2% in all 72 reactive and rolling cells computed. The identity is elementary. Lu (2026) and Zhang et al. (2026) observe that skill and validity diverge; the identity gives the size of the factual advantage and shows that it vanishes on exogenous windows.

Placebo audit. With a one-step delay and no anticipation, C_h does not depend on decisions that act after $t + h$. We shift only those decisions and measure the placebo response P; any $P \neq 0$ shows non-causal use. This is a negative-control exposure (Lipsitch, Tchetgen Tchetgen, and Cohen 2010) applied to the model's responses rather than the data. With R_{eff} the response to the decisions that can affect the target, we flag a cell when $\pi = |P|/(|R_{\text{eff}}| + |P|) \ge 0.2$ and $|P|$ exceeds three across-seed standard errors. In the linear model the rule flags 56 of 57 displaced feedback cell-horizons and none of 27 exogenous ones. Frozen and applied to policy gains it was not tuned on ($k = 0.25$, $k = 1$), it flags 43/44 and 70/72 with no false flags in 2×126 exogenous cases. A zero placebo response is necessary, not sufficient. A forecaster that ignores covariates passes trivially. So does a forecaster whose bias runs only through decisions that can affect the target: a same-time covariate regression, which is TimesFM-2.5's default covariate pathway, has $P = 0$ exactly yet a reactive response of -0.71 against a truth of 0.40.

3 Pilot Evaluation

Protocol. For Chronos-2 and TimesFM-2.5, hypotheses, thresholds, the extension rule and the decision rules were fixed in a configuration file before any forecast; Timestral was run later under the same rules; expectations were recorded before every run and failed expectations are reported. Each instance is an independent simulated series. The grid covers open loop, reactive control, rolling re-planning, block plans forecast from issue time, and block plans forecast four steps after issue, at $s \in \{0.3, 1\}$ and context $T \in \{256, 1024\}$, with 50 instances per cell plus one pre-registered extension to 100. Each response is an antisymmetric paired query $(Q(+\delta) - Q(-\delta))/2\delta$, with δ one standard deviation of the context decisions and $h \in \{1, 4\}$; the readout is the median forecast.

Replay twins. An open-loop baseline cannot separate feedback from a model's own attenuation or from its reaction to autocorrelated decisions. Each closed-loop instance therefore gets a *replay twin*: the same decision sequence fed into the same plant with independent noise, which is exogenous by construction. The feedback displacement is $\Delta = \rho - \rho_{\text{twin}}$, with $\rho = R/\tau_h$ the response relative to the true effect τ_h ($\tau_1 = 0.40$, $\tau_4 = 1.01$). A second, σ-matched twin rescales the noise so that the outcome variance matches; A *cell* is one model, regime, excitation s and context length T; a cell-horizon adds h. A cell-horizon is *displaced* if $\Delta \le -0.25$ and its 95% percentile bootstrap interval over instances lies below 0, and *undisplaced* if $|\Delta| \le 0.10$ with the interval inside $(-0.25, 0.25)$. Reactive and rolling cells count as displaced only against both twins. A cell is *assessable* if the lower 95% bound of ρ_{twin}, computed on 20 separate twins, exceeds 0.2.

Hypotheses.

- *H1*: at least 12 of the 16 reactive and rolling cell-horizons are displaced downward, and the bias does not fade with context (median $|\Delta_{1024}|/|\Delta_{256}| \ge 0.75$).

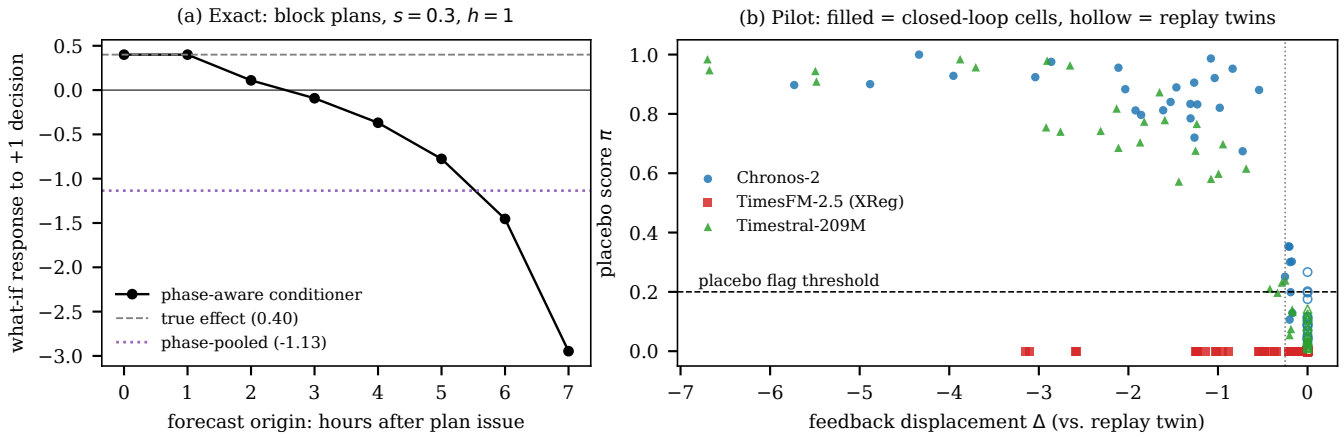


Figure 1: (a) Exact one-step what-if response of the full-path conditioner under block plans, by forecast origin; the phase-pooled conditioner (dotted line) is biased at every origin. (b) Pilot: placebo score π against feedback displacement Δ for every closed-loop cell-horizon (filled) and its replay twin (hollow, plotted at $\Delta = 0$). The dashed line is the flag threshold $\pi = 0.2$. Timestral separates cleanly; TimesFM's placebo response is exactly zero; Chronos-2 has two flagged twin cells, one just above the π threshold ($\pi = 0.20$) and one just above the three-SE bar ($\pi = 0.27$).

- *H2*: does the model know when plans were issued (paired difference between after-issue and issue-time displacement)?
- *H3a*: flag sensitivity of at least 75% on displaced cells, with at most 10% of replay twins flagged.
- *H4*: predictive intervals ignore non-identification. This is tested on two worlds with identical histories but causal gains 0.4 and 1.2; an interval narrower than 3.07 at a query three decision-standard-deviations off-policy must under-cover in one world.

Models and harness. We run Chronos-2 (chronos-forecasting 2.3.2), TimesFM-2.5 with XReg in its default `xreg + timesfm` mode, and Timestral-209M, all zero-shot on CPU. Timestral was released during the study and added after the other two models' results were known. Its outcome was therefore confirmed on a fresh seed base before we report it. The harness gate passed with no repair:

- reference conditioners and an ARX oracle matched their exact population values in 159/160 comparisons;
- a planted leaky forecaster was flagged in 36/36 cell-horizons;
- TimesFM's decision-time response on reactive logs was -1.580 ± 0.005, against -1.58 predicted.

An independent re-implementation of the analysis, run on the raw Chronos-2 and TimesFM-2.5 forecasts, agreed on every count and outcome, and the same code reproduced Timestral's outcomes; all outcomes are also unchanged under 50 alternative bootstrap seeds.

4 Results

Feedback reverses all three models. Under reactive control at low excitation, the models' what-if responses are -4.02, -1.76 and -4.36 times the true effect. On the replay twins of the same instances, which have the same decisions without feedback, they are $+0.32$, $+1.39$ and $+1.13$. H1 is supported for every model (Table 1). The bias does not fade with four times more context: the context ratios are 1.30, 1.01 and 1.17, so it grows for Chronos-2 and Timestral. TimesFM matched every population prediction we recorded before running it, including its two cells displaced by less than the 0.25 bar ($\Delta \approx -0.21$) and its reactive response ($\rho = -1.76$ predicted; -1.76 and -1.77 observed at $T = 1024$ and 256).

No model knows when plans were issued. Forecast from a plan's issue time, where a phase-aware conditioner is exact, all three models are as displaced as four steps later (Chronos-2 and TimesFM: $|D| \leq 0.10$ in 7/8 pairs; Timestral inconclusive on the pilot seeds and pooled, 6/8, on the fresh seeds). This matches the pooled-conditioner prediction: absent an issue-time covariate, advance plans are not a safe harbour.

The audit separates the three interfaces. Timestral reads the whole future window and responds to decisions after the target. The audit flags all 22 of its displaced cells, flags none of its 16 replay twins, and repeats this on fresh seeds (20/21, 0/16). Chronos-2 is flagged in all 12 displaced cells, but two of its 16 twins are also flagged ($\pi = 0.20$ and 0.27). It thus reacts slightly to future decisions even when they are exogenous but autocorrelated, a specificity cost we had named in advance. The 12.5% twin rate falls between the support and refutation bars. TimesFM's in-context regression reads only the decision that acts on each target step, so its placebo response is exactly zero and the audit cannot see a bias that H1 shows is there. This blind spot was predicted from the interface's source code before any forecast.

Factual skill rewards the wrong model. The identity predicts that a full-path model can beat the true causal model on factual MSE exactly where its what-if answers are worst. Timestral beats the causal model in 8/8 feedback cell-horizons, with an MSE ratio as low as 0.19, while its

	Chronos-2	TimesFM-2.5	Timestral
H1 displaced/assessable	12/12	14/16	16/16
context ratio	1.30	1.01	1.17
outcome	supported	supported	supported
H2 issue time	pooled	pooled	inconcl.
H3a sensitivity	12/12	0/18	22/22
twin flags	2/16	0/16	0/16
outcome	inconcl.	refuted	supported
H4 coverage at c=0	0.30	0.775	0.115
outcome	n/a	supported	n/a
ρ reactive, s=0.3, h=1	-4.02	-1.76	-4.36
replay twin ρ_{twin}	0.32	1.39	1.13
Beats causal (factual MSE)	7/8 (6)	3/8 (0)	8/8 (6)

Table 1: Pilot outcomes after the single pre-registered extension (pre-registered for Chronos-2 and TimesFM-2.5; Timestral is a later arm run under the same rules). "n/a": the H4 precondition (calibration at the logged policy) failed. Timestral on fresh seeds: H1 16/16 (ratio 1.13), H3a 20/21 with 0/16 twin flags, H2 pooled. "Beats causal": feedback cell-horizons at T=1024 where factual MSE is below the true causal model's; in brackets, those with the 95% interval of the MSE ratio below 1. H3a sensitivity also counts displaced block-plan cells forecast four steps after issue. c: query offset in decision standard deviations; intervals are 10–90% quantiles (nominal 80%). Reference: the ideal full-path conditioner gives $\rho = -3.91$ (twin 1.01) in the same cell; the causal model gives 1.

what-if answer is $-4.4\times$ the truth; Chronos-2 does so in 7/8. TimesFM, which leaks less, has a point ratio below 1 in 3/8, none with an interval below 1 (against 6/8 for both other models). In open loop no model reliably beats it. A factual covariate benchmark built on feedback logs would therefore rank Timestral and Chronos-2 above the true causal model.

Intervals and non-identification. TimesFM's interval width is identical for every query (2.52), because, in its source code, the covariate term is added to every residual quantile without widening them. Its coverage falls from 0.775 on-policy to 0.005–0.015 in the world with the larger causal gain at queries three standard deviations off-policy (H4 supported). Chronos-2 and Timestral were too narrow at the logged policy (coverage 0.30 and 0.115) for the test to apply. Timestral's intervals do widen off-policy (from 0.32 to 8.06).

5 Building-Thermal Stress Test

To test whether the findings depend on the linear toy, we simulate an hourly two-node thermal room heated by a heat pump. Its efficiency depends on outdoor temperature T_o, which follows a diurnal cycle plus persistent weather noise and solar gain, and power is clipped to $[0, 3]$ kW. The plant is linear in applied power given T_o, so true effects follow exactly by shared-noise simulation; the controllers are nonlinear. We log four regimes: a hysteresis thermostat; a one-step receding-horizon controller (MPC); a day-ahead schedule forecast from midnight, its issue time; and the same schedule forecast from 20:00, a window that straddles the next plan.

	Δ at h=1		displaced	twin flags
	thermostat	MPC	(of 4)	(of 8)
Linear (u only)	-0.40	-0.40	4	8
Linear (u, T_o,hour)	-0.85	-1.13	4	0
Chronos-2	-0.50	-0.86	3	4
TimesFM-2.5	-0.05	-0.15	0	0
Timestral	-0.55	-0.61	4	1

Table 2: Thermal stress test (100 instances per cell, $T = 1024$ hours). Displacement against replay twins sharing the decision and weather paths. No forecaster is displaced ($\Delta > -0.25$ throughout) for the day-ahead schedule at either origin.

Models receive power and T_o (a perfect weather forecast) as known covariates; queries shift power only. The gate passed: an oracle with the true dynamics was exact, a planted leak was flagged 8/8, and feedback was present (correlation of heat-pump power u with indoor temperature T_i: -0.58 and -0.59 under thermostat and MPC, -0.01 in open loop).

Findings. Feedback displacement survives switching control. Chronos-2 is displaced in 3/4 thermostat and MPC cell-horizons, and its MPC response is wrong-signed ($\rho = -0.35$ against $+0.51$ on the twin). Timestral is displaced in 4/4, attenuated rather than reversed ($\rho = +0.31$ against $+0.86$ under the thermostat), and the audit flags all 4 displaced cells with 1/8 twin flags. TimesFM is barely displaced here ($|\Delta| \leq 0.16$). Its regression includes T_o, which may absorb much of the confounding, but we have not established why. The thermal setting exposes a second limit of the audit. A linear conditioner given only the decisions flags every replay twin: the logged decisions were computed from weather forecasts and the solar cycle, so future decisions stand in for future drivers the conditioner was not given. Given T_o and the hour, the same conditioner flags no twin. Chronos-2, which also receives T_o, still flags 4/8 twins. A placebo flag therefore indicates that a model responds to decisions after the target, which is non-causal use, but not necessarily feedback; replay twins that share the drivers are what separate the two.

6 Related Work

TSFM what-if evaluation. Lu (2026) evaluate four covariate TSFMs on three simulators with observed, off-support and hidden confounders. They find that factual skill is an unstable guide to interventional validity (rank correlation $+0.80$ and -0.80 across simulators) and report sign inversion under hidden confounding. We could compare only against its public abstract. We study feedback with no hidden confounder, and add the timing rule, the identity, which predicts when skill and validity align, and the placebo audit. Amria et al. (2026) show that TSFMs used for control need independent excitation in context. Won (2026) report memoryless covariate interfaces in TimesFM-2.5 and TabPFN-TS and attenuated responses in Chronos-2 on open-loop plants. Wang (2026) document wrong-signed effects of naive fits on closed-loop building data. WhatIf-Bench (Ashok et al. 2026) measures extrapolation to novel covariate settings, without

feedback. Berthelier et al. (2026) probe simple covariate relations. Zhang et al. (2026) show that accuracy and mechanism consistency diverge in trained time-series world models and attribute sign reversal to reactive controllers. Their check needs declared effect signs, ours does not, and we add the timing rule and the identity.

Causal forecasting. The gap between conditional and interventional forecasts is classical (Vankadara et al. 2021; Antolín-Díaz, Petrella, and Rubio-Ramírez 2021), as are strict versus sequential exogeneity (Chamberlain 1982; Bonhomme 2025), time-varying confounding and its g-computation repairs (Melnychuk, Frauen, and Feuerriegel 2022; Xiong et al. 2024; Zare et al. 2026), and causal evaluation of deep forecasters (Crasson, Nabet, and Lécuyer 2024). Synthetic interventional benchmarks exist (Thumm, Anthony, and Chen 2026; Jang et al. 2026). Our contribution is to turn these principles into a timing rule, an identity and an audit for zero-shot covariate interfaces, and to test them.

7 Limitations and Recommendations

Our theory is exact only for the linear-Gaussian model, and all evaluations use simulators: real logs carry no interventional ground truth. The audit is a one-sided test. It is blind to same-time covariate interfaces, it can fire on exogenous autocorrelated decisions (Chronos-2) or on decisions that proxy omitted drivers (thermal), and it assumes no anticipation, which announced decisions such as promotions violate. The models are specific released versions. For practice:

1. read scenario forecasts as interventional only on exogenous windows, and supply the plan-issue time;
2. run the placebo audit: a flag shows that the model uses decisions that cannot affect the target (through feedback or proxied drivers), and a pass is uninformative for same-time covariate interfaces; a Sims regression of outcomes on future decisions tests the log itself, which we did not evaluate here;
3. do not select models by factual covariate benchmarks on feedback logs;
4. when the one-step effect is identified, roll a causal one-step model forward instead.

Code, pre-recorded configurations and all run records will be released.

References

Amria, M.; Hoffmann, J.; Bordne, P.; Rothenhäusler, A.; Frison, L.; Walnum, H. T.; Gros, S.; and Bödecker, J. 2026. Time-series Foundation Models for Predictive Control: The Role of Excitation. arXiv:2610.06447.

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. arXiv:2510.15821.

Antolín-Díaz, J.; Petrella, I.; and Rubio-Ramírez, J. F. 2021. Structural Scenario Analysis with SVARs. *Journal of Monetary Economics*, 117: 798–815.

Ashok, A.; Parmar, M.; Goyal, P.; Mishra, B. D.; Li, C.-L.; Rish, I.; Drouin, A.; Yoon, J.; and Pfister, T. 2026. WhatIf-Bench: Benchmarking Intervention-Aware Time Series Forecasting. arXiv:2610.26582.

Berthelier, G.; Baranova, M.; Pantea, A.-T.; Naour, E. L.; Petralia, A.; Nabil, T.; and Palpanas, T. 2026. Investigating simple target-covariate relationships for Chronos-2 and TabPFN-TS. arXiv:2605.12200.

Bonhomme, S. 2025. Back to Feedback: Dynamics and Heterogeneity in Panel Data. arXiv:2512.17576.

Chamberlain, G. 1982. The General Equivalence of Granger and Sims Causality. *Econometrica*, 50(3).

Crasson, T.; Nabet, Y.; and Lécuyer, M. 2024. Training and Evaluating Causal Forecasting Models for Time-Series. arXiv:2411.00126.

Das, A.; Kong, W.; Sen, R.; and Zhou, Y. 2023. A decoder-only foundation model for time-series forecasting. arXiv:2310.10688.

Jang, J.; Jin, H.; Park, H.; Chae, K.; and Kim, T. 2026. What If TSF: Reframing Time Series Forecasting as Scenario-Guided Multimodal Forecasting. arXiv:2601.08509.

Lipsitch, M.; Tchetgen Tchetgen, E.; and Cohen, T. 2010. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology*.

Lu, S. 2026. Forecasting Skill Is Not Interventional Validity: Evaluating Covariate-Aware Time-Series Foundation Models as Zero-Shot Counterfactual Predictors. NeurIPS 2026 Workshop on Foundation Models for Temporal Systems. Compared on the basis of the public abstract; full text not accessible to us.

Melnychuk, V.; Frauen, D.; and Feuerriegel, S. 2022. Causal Transformer for Estimating Counterfactual Outcomes. arXiv:2204.07258.

Mistral AI. 2026. Introducing Timestral. Blog post and model release, 15 October 2026.

Nixtla. 2026. What-If Forecasting: Price Effects in Retail. TimeGPT documentation. Accessed 19 October 2026.

Shchur, O.; Ansari, A. F.; Turkmen, C.; Stella, L.; Erickson, N.; Guerron, P.; Bohlke-Schneider, M.; and Wang, Y. 2025. fev-bench: A Realistic Benchmark for Time Series Forecasting. arXiv:2509.26468.

Sims, C. A. 1972. Money, Income, and Causality. *American Economic Review*, 62(4): 540–552.

Thumm, D.; Anthony, B. T.; and Chen, Y. 2026. DoTime: A Synthetic Benchmark Generator for Interventional and Counterfactual Time Series. arXiv:2607.27263.

Vankadara, L. C.; Faller, P. M.; Hardt, M.; Minorics, L.; Ghoshdastidar, D.; and Janzing, D. 2021. Causal Forecasting:Generalization Bounds for Autoregressive Models. arXiv:2111.09831.

Wang, Y. 2026. ThermoForce: A Physics-Structured Interventional World Model for Building HVAC Control. arXiv:2607.03942.

Won, H.-I. 2026. Attenuated in-context identification in time-series foundation models: diagnosis under counterfactual inputs and repair by synthetic forced-system fine-tuning. arXiv:2610.08118.

Xiong, H.; Wu, F.; Deng, L.; Su, M.; and wei H Lehman, L. 2024. G-Transformer: Counterfactual Outcome Prediction under Dynamic and Time-varying Treatment Regimes. arXiv:2406.05504.

Zare, A.; Zare, A.; Rahimi, H.; Salarikia, R.; and Kashkooli, M. 2026. Causal Longitudinal Prior-Fitted Networks for Counterfactual Outcome Prediction. arXiv:2606.05797.

Zhang, H.; Guo, J.; Xu, Z.; Plotkin, Z.; Konz, N.; Tan, Z.; and Chen, T. 2026. On the Divergence of Accuracy and Mechanism Consistency in Time Series World Models. arXiv:2610.01842.