CONFOUNDING-AWARE COUNTERFACTUAL PREDICTION WITH PARTIAL IDENTIFICATION AND CONFORMAL INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual prediction under an externally specified trajectory arises when a continuous intervention affects an entire system, leaving no contemporaneous control group (e.g. financial stress testing). In this setting, latent variables that jointly drive the intervention path and outcomes can make the causal target differ from the corresponding predictive quantity, even when predictive accuracy is high. We develop a model-agnostic framework for dynamic counterfactual prediction under unobserved confounding. An additive latent-confounding model and a partial-R^2 sensitivity specification yield a closed-form identified set for the causal mean. We estimate its predictive center with flexible learners and combine the resulting confounding radius with importance-weighted conformal calibration. The final interval separates estimation uncertainty from identification uncertainty and provides finite-sample coverage up to explicit temporal-dependence and extrapolation remainder terms. Semi-synthetic experiments show that standard predictive intervals can fail while predictive fit remains nearly unchanged, whereas the proposed intervals substantially improve causal coverage across linear and nonlinear data-generating processes. Credit stress testing provides an important motivating application: on Freddie Mac mortgages under the realized 2020 unemployment path, the framework widens predictive bands to reflect plausible confounding and flags horizons that lie beyond historical support.

1 INTRODUCTION

Stress testing[1], a central modeling task at large financial institutions, asks: how would their credit losses change if the macroeconomy—say, unemployment and house prices—were forced into a severe downturn? Stress-testing approaches typically use predictive machine learning models trained on historical data to project losses under specified macroeconomic scenarios (Moffo, 2024). When these projections are interpreted as the effects of externally imposed macroeconomic paths, an additional causal identification question arises.

Fundamentally, no machine learning model—however accurate its historical fit—can answer this causal question without an identification strategy [2]. In past recessions, unemployment rose alongside deteriorating financial conditions and tightening credit supply; models trained on such data may capture both the association with unemployment and the effects of co-moving unobserved factors, so their projections need not coincide with the effect of an intervention on unemployment alone (Bernanke et al., 1999; Adrian et al., 2019; Ramey, 2016; Bica et al., 2020). Surprisingly, this tension has received little attention, with Gao et al. (2017) and Mauri et al. (2023) the notable exceptions, both of which propose learning a causal network from data and using it to simulate stress scenarios. While a useful first step, such approaches do not provide a formalized framework for causal stress testing.

[1] Stress testing requires financial institutions to project losses on their portfolios under hypothetical macroeconomic scenarios prescribed by regulators. Prominent frameworks include the U.S. Federal Reserve's CCAR (Board of Governors of the Federal Reserve System, 2024), the EBA EU-wide stress tests, and the Basel Committee's Stress Testing Principles (Basel Committee on Banking Supervision, 2018).

[2] An identification strategy is a set of assumptions, typically about the data-generating process, that allows one to isolate a causal effect from observational data.

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.

Two features of stress testing make it particularly resistant to standard causal inference. First, the macro scenario affects all units simultaneously, leaving no control group—ruling out methods that rely on untreated comparisons (Abadie et al., 2010; Athey et al., 2021). Second, the treatment varies continuously rather than taking a binary form, precluding recent approaches to counterfactual forecasting without controls that require a binary intervention (Cerqua et al., 2023). Together, these two features put causal stress testing beyond the reach of existing causal inference methods.

In this paper, we propose a partial identification and estimation framework for causaly implementing stress testing. Standard causal inference methods aim at point identification—recovering a single value for the causal effect. Partial identification instead asks how large the causal effect could be for a given level of confounding, producing an identified set rather than a point estimate (Manski, 1990; 2003). This is closely related to sensitivity analysis in the causal machine learning literature (Cinelli & Hazlett, 2020; Kallus et al., 2019; Frauen et al., 2024).

Our framework operates on individual-level panel data[3]—where the goal is to estimate how credit losses change under a hypothetical macro scenario. As discussed above, unobserved factors that jointly drive the macro path and credit losses make the causal effect unidentifiable from data alone. To overcome these challenges, we introduce a key identifying assumption: the unobserved confounder enters both the outcome and macro equations additively. This is not a restrictive modeling choice—it is the structure implied by the workhorse models of latent financial conditions in macroeconomics and credit risk (Bernanke et al., 1999; Adrian et al., 2019).

Under this assumption, we derive a closed-form identified set for the causal effect, expressed in terms of two interpretable sensitivity parameters (Carnegie et al., 2016; Cinelli & Hazlett, 2020) that govern the strength of confounding in the macro and outcome equations. The width of the identified set directly reflects the assumed strength of confounding: stronger confounding yields wider sets, and when confounding is set to zero the set collapses to a point estimate.

For inference, we combine the identified set with importance-weighted conformal prediction (Tibshirani et al., 2019; Barber et al., 2023). Conformal prediction provides finite-sample coverage guarantees without requiring a correctly specified predictive model. We reweight calibration residuals to account for distributional shift between historical macro paths and the stress scenario, so that calibration places greater emphasis on historical periods that are more relevant to the target path. The resulting guarantee allows for explicit remainder terms capturing temporal dependence and limited overlap; when the stress path is weakly supported by historical data, the method flags extrapolation rather than treating nominal coverage as assured. Combining the identified set with the conformal band yields a confidence interval that decomposes into two interpretable components—estimation uncertainty and confounding bias—each with an explicit finite-sample guarantee.

To evaluate the framework[4], we design semi-synthetic experiments that combine real U.S. unemployment paths with synthetic credit outcomes from a known data-generating process (in short, DGP). We test under both linear and nonlinear DGPs, and the framework is agnostic to the choice of base predictor—any machine learning model can be plugged into our framework. Our experiments show that standard predictive models fit historical data well but systematically miss the causal effect under stress, and their confidence intervals are far too narrow to detect this failure. We then show that only our full framework—conformal bands combined with the confounding envelope—achieves nominal coverage across all horizons on the linear DGP and yields substantial coverage gains on the nonlinear DGP, outperforming standard uncertainty quantifications including linear regression with standard errors and ensemble conformal prediction (Xu & Xie, 2021).

We further apply the framework to 1.35 million Freddie Mac mortgages, using the realized 2020 COVID unemployment path as a stress scenario. While semi-synthetic experiments provide ground-truth causal effects, the real-data application provides a complementary assessment of the framework's behavior under a genuine macroeconomic shock. Because realized delinquency reflects all conditions accompanying the unemployment path, it benchmarks the predictive rather than the unobserved causal mean. The full interval contains the realized outcomes at all twelve horizons once

[3]For example, a credit card portfolio where each account's monthly balance, payment, and delinquency status are tracked over time, while the macroeconomy (unemployment, interest rates) moves in the background.

[4]Real data do not provide ground truth for counterfactual outcomes—we cannot observe what credit losses would have been under a macro scenario that did not materialize. Semi-synthetic experiments, where the data-generating process is known, are the standard approach for validating causal inference methods.

the confounding envelope is included, whereas the conformal-only interval contains only three of twelve. We therefore interpret this exercise as illustrating the practical behavior of the interval and its extrapolation diagnostics, rather than as validation of causal coverage.

Our contributions are twofold. To the financial risk management literature, we demonstrate that the gap between predictive and causal inference in stress testing is structural and invisible to standard validation, and provide a framework that delivers valid causal coverage with a concrete calibration protocol using observable financial conditions data. To the causal machine learning literature, we propose a partial identification method for panel data with a common continuous treatment and no control group—a setting beyond the reach of existing causal inference methods. The method yields a closed-form identified set with finite-sample inference guarantees that decompose uncertainty into estimation error and confounding bias.

2 Setup and Estimands

Consider panel data with units $i = 1, \ldots, N$ observed over periods $t = 1, \ldots, T$. Let $Y_{i,t}$ denote the outcome (e.g., credit loss) and A_t a macro variable common across units (e.g., unemployment rate). A stress scenario specifies a hypothetical path $a^S = (a_1^S, \ldots, a_H^S)$ over a forecast horizon $h = 1, \ldots, H$, optionally contrasted against a baseline path a^B. We reindex the forecast origin as time zero. Let $\mathbb{E}_{\text{hist}}$ denote expectation under the distribution of pre-stress rolling origins and $\mathbb{E}_{\text{future}}$ expectation at the forecast origin.

Definition 1 (Three mean targets)**.** *For a macro path $a = (a_1, \ldots, a_H)$, define the historical predictive, future predictive, and causal means as*

$$\begin{aligned} \mu_h^{\text{hist}}(a) &:= \mathbb{E}_{\text{hist}}[Y_{i,t+h} \mid A_{t+1:t+h} = a_{1:h}], \\ \mu_h(a) \equiv \mu_h^{\text{future}}(a) &:= \mathbb{E}_{\text{future}}[Y_{i,h} \mid A_{1:h} = a_{1:h}], \\ \mu_h^*(a) &:= \mathbb{E}[Y_{i,h}(a_{1:h})]. \end{aligned}$$

The first two means answer the same predictive question under different time regimes. Their difference, $\mu_h^{\text{future}}(a) - \mu_h^{\text{hist}}(a)$, is a *temporal generalization gap*: the relationship learned from historical data may not persist at the forecast origin. The future predictive and causal means instead differ in how the macro path is generated. Under intervention, A is set externally; under observation, conditioning on $A_{1:h} = a_{1:h}$ also selects on unobserved factors that drove A to that path. Their difference is therefore a *confounding gap*. For example, losses observed when unemployment reaches 10% may partly reflect deteriorating financial conditions that accompanied the increase, rather than the effect of unemployment alone. The stress test asks for $\mu_h^*(a^S)$, whereas standard prediction targets $\mu_h^{\text{future}}(a^S)$.

For any estimator $\widehat{\mu}_h(a)$ trained on historical data, total error decomposes as

$$\mu_h^*(a) - \widehat{\mu}_h(a) = \underbrace{\mu_h^*(a) - \mu_h^{\text{future}}(a)}_{\text{confounding gap}} + \underbrace{\mu_h^{\text{future}}(a) - \mu_h^{\text{hist}}(a)}_{\text{temporal generalization gap}} + \underbrace{\mu_h^{\text{hist}}(a) - \widehat{\mu}_h(a)}_{\text{estimation error}}.$$

This decomposition organizes the remainder of the paper. Assumption 1 sets the temporal generalization gap to zero for supported paths, and Proposition 1 then identifies the future predictive mean from the historical distribution. The conformal radius later controls estimation error, while the confounding envelope in Section 3 bounds the gap between the future predictive and causal means. A structural break violates the temporal stability condition and falls outside the guarantee: importance weighting can change the relative frequency of supported macro paths, but cannot repair a change in the path-conditional mean itself.

Assumption 1 (Path-conditional mean stability and overlap)**.** *For every $h \leq H$ and target path $a_{1:h}$,*

$$\mathbb{E}_{\text{future}}[Y_{i,h} \mid A_{1:h} = a_{1:h}] = \mathbb{E}_{\text{hist}}[Y_{i,t+h} \mid A_{t+1:t+h} = a_{1:h}].$$

The historical conditional mean is invariant across pre-stress rolling origins t, and $a_{1:h}$ lies in the support of the pre-stress distribution of $A_{t+1:t+h}$.

Proposition 1 (Identification of the future predictive mean)**.** *Under Assumption 1,*

$$\mu_h(a) = \mu_h^{\text{future}}(a) = \mu_h^{\text{hist}}(a),$$

and this common predictive mean is identified from the pre-stress rolling-origin distribution for every $h \leq H$ and every target path in its support.

Proof in Appendix B.1.

The two parts of Assumption 1 play distinct roles: mean stability transports the predictive relationship from historical rolling origins to the forecast origin, while overlap makes the relevant historical conditional mean observable in principle. Proposition 1 is therefore a purely predictive identification result. It neither depends on the estimator used nor claims that $\mu_h(a) = \mu_h^*(a)$; the latter gap is the causal problem addressed next.

For the structural and estimation results below, let $\mathcal{H}_{i,t} := \{Y_{i,s}, A_s\}_{s \leq t}$ denote unit i's history up to time t, and let $I_{i,t} = g(\mathcal{H}_{i,t})$ be a finite-dimensional state used for prediction.

3 Causal Set Identification under Structural Confounding

The core economic concern is that unobserved factors—such as financial conditions, credit market tightness, or investor sentiment—simultaneously drive both the macro variable and the outcome. When financial conditions deteriorate, unemployment rises and credit losses increase, but the losses are partly caused by the financial conditions themselves, not by unemployment alone. We model this through a single latent process U_t that enters both the outcome and the macro variable additively. Formally,

Assumption 2 (Additive latent confounding)**.** *The data-generating process satisfies:*

- *(a) Outcome equation: $Y_{i,t+1} = m(I_{i,t}, A_{t+1}) + \gamma_Y U_t + \xi_{i,t+1}$, where $\xi_{i,t+1}$ is an idiosyncratic outcome shock with $\mathbb{E}[\xi_{i,t+1} \mid I_{i,t}, A_{t+1}, U_t] = 0$ and conditional variance σ_ξ^2.*
- *(b) Macro equation: $A_{t+1} = f(A_t) + \gamma_A U_t + \eta_{t+1}$, where η_{t+1} is an idiosyncratic macro shock with $\mathbb{E}[\eta_{t+1} \mid A_t, U_t] = 0$ and conditional variance σ_η^2.*
- *(c) Latent process: $\{U_t\}$ is stationary with $\mathbb{E}[U_t] = 0$, $\mathrm{Var}(U_t) = \sigma_U^2$, and finite autocovariances $\mathrm{Cov}(U_t, U_{t+k})$.*
- *(d) Innovation independence: the outcome shocks $\{\xi_{i,t}\}_{i,t}$, macro shocks $\{\eta_t\}_t$, and latent process $\{U_t\}_t$ are mutually independent.*

Assumption 3 (Partial-R^2 sensitivity specification)**.** *For fixed $\sigma_U, \sigma_\eta, \sigma_\xi > 0$, define the sensitivity parameters*

$$\rho_A := \frac{\gamma_A^2 \sigma_U^2}{\gamma_A^2 \sigma_U^2 + \sigma_\eta^2}, \qquad \rho_Y := \frac{\gamma_Y^2 \sigma_U^2}{\gamma_Y^2 \sigma_U^2 + \sigma_\xi^2}.$$

We restrict the sensitivity class to

$$0 \leq \rho_A \leq \bar{\rho}_A < 1, \qquad 0 \leq \rho_Y \leq \bar{\rho}_Y < 1,$$

where $\bar{\rho}_A$ and $\bar{\rho}_Y$ may be chosen by the analyst or calibrated from auxiliary data.

Remark 1 (Interpretation of the sensitivity parameters). The parameters ρ_A and ρ_Y measure the fractions of structural residual variance attributable to the latent confounder U_t in the macro and outcome equations, respectively. Both parameters lie in $[0, 1)$ and have a directly interpretable scale. For instance, $\rho_A = 0.3$ means that U_t accounts for 30% of the structural macro residual variance, with the analogous interpretation for ρ_Y. $\rho_A = 0$ or $\rho_Y = 0$ removes the corresponding confounding channel.

This parameterization is in one-to-one correspondence with the magnitudes of the structural coefficients:

$$|\gamma_A| = \frac{\sigma_\eta}{\sigma_U} \sqrt{\frac{\rho_A}{1 - \rho_A}}, \qquad |\gamma_Y| = \frac{\sigma_\xi}{\sigma_U} \sqrt{\frac{\rho_Y}{1 - \rho_Y}}.$$

Thus, Assumption 3 expresses the allowed confounding strength on a common variance-share scale.

The additive confounding structure follows the sensitivity analysis literature (Carnegie et al., 2016; Zheng et al., 2024). The outcome model m may be arbitrarily nonlinear in observables; the additive structure restricts only how the unobserved U_t enters. Exogeneity ($\gamma_A = 0$ or $\gamma_Y = 0$) would close the gap entirely but is implausible. Under these assumptions, we can bound the gap between $\mu_h(a)$ and $\mu_h^*(a)$:

Theorem 1 (Partial identification)**.** *Under Assumptions 1–3, for any macro path a, $\mu_h^*(a) \in [\mu_h(a) - c_h,\ \mu_h(a) + c_h]$, where*

$$c_h = \Big|\gamma_Y \mathbb{E}[U_{h-1} \mid A_{1:h} = a_{1:h}] - \big(\mathbb{E}[m(I^*_{i,h-1}, a_h)] - \mathbb{E}[m(I_{i,h-1}, a_h) \mid A_{1:h} = a_{1:h}]\big)\Big|,$$

with $I^*_{i,t}$ the state under the causal distribution and $I_{i,t}$ under the predictive distribution.

The bound c_h has two components. The first term is the direct channel: at horizon h, the confounder U_{h-1} enters the outcome with coefficient γ_Y, and conditioning on $A_{1:h} = a_{1:h}$ can shift our expectation of U_{h-1} away from zero. The magnitude of this shift depends on how strongly U drives the macro variable (γ_A) and on the stress path. The second term is the state propagation channel: earlier confounders affect past outcomes $Y_{i,1}, \ldots, Y_{i,h-1}$, which feed into the state $I_{i,h-1}$ and affect the endpoint outcome through m.

Theorem 1 allows general nonlinear outcome dynamics and an arbitrary stationary confounder process, but leaves c_h in an abstract form. To obtain a computable expression, we now impose three additional restrictions: U_t is Gaussian AR(1), the macro shocks are Gaussian white noise, and the outcome dynamics are affine in the lagged outcome. Gaussianity makes the conditional confounder mean explicit, AR(1) determines its covariance across horizons, and the affine outcome equation gives simple persistence weights. These restrictions are used only for the closed form below.

Stress tests start from the portfolio and macro state observed at the forecast date. We condition on this common starting point throughout; because it is the same under prediction and intervention, it cancels from their difference.

Corollary 1 (Closed-form sensitivity set under Gaussian-affine dynamics)**.** *Under Assumptions 2 and 3, suppose that U_t is a stationary Gaussian AR(1) with coefficient $|\phi_U| < 1$, that $\eta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\eta^2)$, and that $m(I_{i,t}, A_{t+1}) = \beta Y_{i,t} + g(A_{t+1})$.*

Given $(f, \beta, \phi_U, \sigma_\eta, \sigma_\xi)$, define

$$r_j := a_j - f(a_{j-1}), \qquad \mathbf{r} = (r_1, \ldots, r_h)^\top, \qquad \mathbf{q}_h = (\beta^{h-1}, \ldots, \beta, 1)^\top.$$

Let $\widetilde{\Sigma}_h$ be the AR(1) correlation matrix with $[\widetilde{\Sigma}_h]_{jk} = \phi_U^{|j-k|}$. The absolute predictive–causal gap is

$$d_h(a; \rho_A, \rho_Y) = \frac{\sigma_\xi}{\sigma_\eta}\sqrt{\frac{\rho_A \rho_Y}{(1-\rho_A)(1-\rho_Y)}}\left|\mathbf{q}_h^\top \widetilde{\Sigma}_h \left(\frac{\rho_A}{1-\rho_A}\widetilde{\Sigma}_h + I\right)^{-1} \mathbf{r}\right|.$$

Therefore, the worst-case sensitivity radius is

$$c_h(a; \bar\rho_A, \bar\rho_Y) = \sup_{\substack{0 \le \rho_A \le \bar\rho_A \\ 0 \le \rho_Y \le \bar\rho_Y}} d_h(a; \rho_A, \rho_Y),$$

and $\mu_h^*(a) \in [\mu_h(a) - c_h(a; \bar\rho_A, \bar\rho_Y),\ \mu_h(a) + c_h(a; \bar\rho_A, \bar\rho_Y)]$.

Here $\mathbf{r}$ measures how surprising the stress path is relative to the macro baseline f, while $\mathbf{q}_h$ records how outcome persistence carries earlier confounding forward. The gap is zero if either confounding channel is absent. Because the gap need not increase monotonically with ρ_A, we maximize over the sensitivity region rather than simply plugging in both upper bounds. Proof in Appendix B.3.

4 Inference: Confidence Intervals for the Causal Estimand

The identified set is centered at the predictive mean $\mu_h(a)$, which we estimate using the recursive rollout standard in stress-testing practice: fit a one-step predictor $\widehat{m}$ on pre-stress data and iteratively feed each prediction into the next-period state. We quantify its error using rolling-origin conformal scores and importance weighting to account for the shift from historical macro paths to the stress path, following existing weighted and dependent conformal methods (Tibshirani et al., 2019; Barber et al., 2023). Their exact construction, regularity conditions, and finite-sample guarantee are given in Appendix A. Let $q^w_{1-\alpha}$ denote the resulting weighted conformal error quantile. Combining its predictive-mean coverage with the confounding radius c_h from Theorem 1 yields

$$\mathrm{CI}_h(a) = \big[\widehat{\mu}_h(a) \pm (q^w_{1-\alpha} + c_h)\big], \qquad \mathbb{P}(\mu_h^*(a) \in \mathrm{CI}_h(a)) \ge 1 - \alpha - R_{\mathrm{mix}} - R_{\mathrm{weight}}.$$

Here $q^w_{1-\alpha}$ captures estimation and distribution-shift uncertainty, c_h captures identification uncertainty under $(\bar\rho_A, \bar\rho_Y)$, and $R_{\mathrm{mix}}, R_{\mathrm{weight}}$ quantify residual temporal dependence and extrapolation. Appendix Theorem 2 states the formal result.

5 SEMI-SYNTHETIC EXPERIMENTS

The theory establishes an irreducible gap c_h between the predictive mean $\mu_h(a)$ and the causal mean $\mu_h^*(a)$. The experiments ask three questions. First, does this gap matter in practice—can a model that passes standard backtesting validation still be systematically wrong about the causal quantity? Second, does adding the confounding envelope c_h to the weighted conformal band improve coverage of μ_h^*? Third, how important is the confounding envelope relative to estimation uncertainty?

To answer these questions, we design semi-synthetic experiments that combine realistic macroeconomic dynamics with fully known causal ground truth; the complete data-generating process and parameter values are reported in Appendix Table 1. The "semi-synthetic" label reflects the data construction: the macro variable is built from historical U.S. unemployment rates[5], inheriting real macro dynamics including the 2001 and 2008 recessions. The confounder U_t and outcomes $Y_{i,t}$ are synthetic with known parameters. The learner observes the confounded macro variable A^{obs} and outcomes, but not the confounder itself. We consider two specifications for m:

$$\begin{aligned}
\text{Linear:} \quad & m(Y_{i,t}, X_i, A^{\text{obs}}_{t+1}) = \alpha + \beta_1 Y_{i,t} + \beta_2 A^{\text{obs}}_{t+1} + \beta_3 X_i, \\
\text{Nonlinear:} \quad & m(Y_{i,t}, X_i, A^{\text{obs}}_{t+1}) = \alpha + \beta_1 Y_{i,t} + \beta_2 A^{\text{obs}}_{t+1} + \beta_4 (A^{\text{obs}}_{t+1})^2 \\
& \qquad\qquad + \beta_5 \max(Y_{i,t} - \bar{Y},\, 0) + \beta_3 X_i.
\end{aligned}$$

The linear model is the standard specification in industry stress testing. The nonlinear model adds a quadratic macro response and a threshold amplifier. We use ordinary least squares (OLS) for the linear DGP and a two-hidden-layer multilayer perceptron (MLP)[6] for the nonlinear DGP. Models are trained on 1990–2014, conformal calibration scores are computed on 2015–2019 without updating model parameters, and the stress rollout is evaluated over a 12-month horizon from January 2020.[7] The law of motion for the macro variable A^{obs}_t is fitted on 1990–2019 data. The stress scenario is a flat unemployment path at approximately 9%—two standard deviations above the historical mean, comparable in severity to the 2008 financial crisis.[8] The analyst computes c_h at the true confounding strength.[9] The real-data application estimates γ_A and γ_Y separately.

5.1 DOES CONFOUNDING CAUSE SILENT FAILURE?

We run three experiments to show that a model passing standard validation can be systematically wrong about the causal quantity. First, the predictive and causal means diverge even when both are computed exactly in closed form. Second, a projection using the true outcome function—with zero estimation error—still misses the causal truth by far more than its standard-error band. Third, one-step-forward backtesting R^2 is nearly unchanged as the causal gap grows, so standard validation cannot detect the failure.

Figure 1 summarizes all three results. In the top panel, holding $\gamma_A = 0.5$ fixed, the causal 12-month loss remains unchanged as γ_Y increases, whereas the predictive loss rises linearly; at $\gamma_Y = 1.0$, the predictive loss overstates the causal loss by 7.1%—a material gap that arises before any model is fitted. The middle panel shows that the causal and predictive projections diverge immediately. Across horizons, the causal outcome lies 58–212 standard errors from the predictive mean, far outside its uncertainty band.[10] The gap therefore persists after model-estimation error has been removed. In

[5] Monthly civilian unemployment rate from FRED (series UNRATE), `https://fred.stlouisfed.org/series/UNRATE`, 1990–2024, $T = 420$ months. We acknowledge that the observed unemployment rate is itself influenced by latent factors, but it provides a realistic approximation of macro dynamics for our simulation.

[6] Two hidden layers with 64 and 32 units respectively, tanh activation, trained to predict $Y_{i,t+1}$ from the same $(Y_{i,t}, X_i, A^{\text{obs}}_{t+1})$ features as OLS.

[7] The experiments use the first 372 months of the series; the remaining 48 months (2021–2024) are unused.

[8] Over the stress horizon, the model input is set directly to a^S rather than the confounded A^{obs}, so the confounder no longer enters the macro input. This mirrors the definition of the causal estimand: the stress path is externally specified, not generated by the historical macro dynamics.

[9] In practice, the true confounding strengths are unknown. The analyst specifies sensitivity bounds $(\bar{\rho}_A, \bar{\rho}_Y)$ as in Assumption 3 and computes $c_h(a; \bar{\rho}_A, \bar{\rho}_Y)$ as the worst-case gap over the resulting sensitivity region. The resulting bound is valid provided the true confounding strengths lie within that region.

[10] The band spans two standard errors on either side of the predictive mean and has an approximate half-width of 0.01, so it is visually indistinguishable from the predictive trajectory at the scale of the figure.

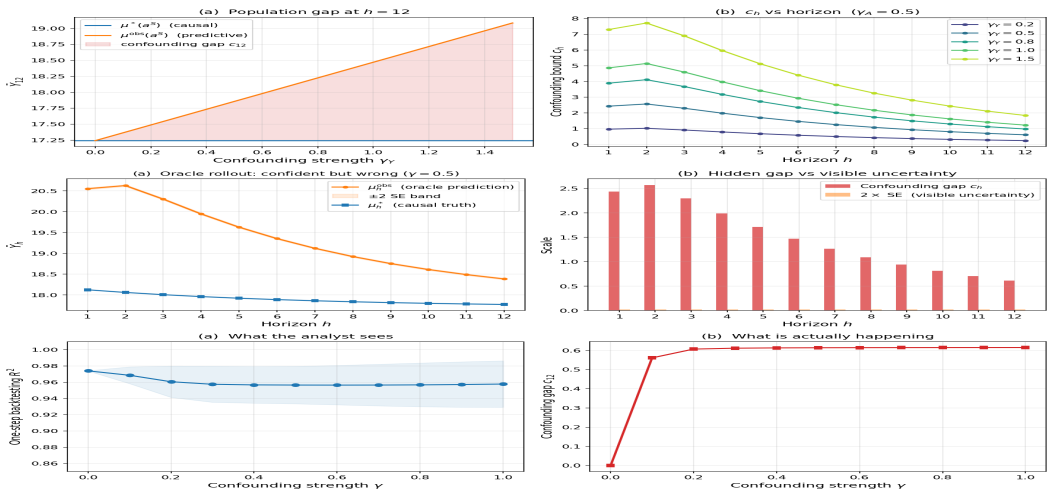


Figure 1: Silent failure of predictive stress testing under confounding.

the bottom panel, one-step backtesting performance changes little as confounding increases: R^2 falls only from 0.974 to 0.957, while the 12-month causal gap rises from zero to 0.61. The model therefore continues to appear accurate in validation even as its stress projection becomes materially biased. Together, the three experiments show that the gap is structural, large relative to estimation uncertainty, and invisible to standard validation—the conditions under which a stress test fails silently.

5.2 Does the Confounding Envelope Improve Causal Coverage?

The previous section showed that a predictive stress projection can miss the true causal mean even when the model appears accurate. We now ask whether the full interval from our framework can cover the true causal mean. Coverage measures how often an interval contains the truth across repeated samples. We therefore repeat the experiment on 50 independently generated datasets[11] and compare four interval-construction methods, described below. For each dataset, each method constructs an interval at every horizon under the same stress scenario, and we check whether it contains the known causal mean. Coverage is measured by the percentage of the 50 intervals that contain the truth.

We compare our approach with two benchmark methods. First, OLS, which constructs intervals from compounded regression standard errors across the rollout. Second, an unweighted conformal band (UCB), which takes the quantile of the per-horizon historical errors directly, without reweighting toward the stress path. We also compare two versions of our method: one reweights those errors by importance weights to account for the distributional shift to the stress scenario, and the other adds the confounding envelope c_h on top of the reweighted interval.

To calibrate the confounding strength, we use the Chicago Fed National Financial Conditions Index (NFCI), a weekly index of U.S. financial conditions in which positive values indicate stress. We aggregate the NFCI to monthly frequency over 1990–2019 and use it as a proxy for U_t. An AR(1) fit yields $\hat{\phi}_U = 0.973$ and $\hat{\sigma}_\nu = 0.124$, while regressing UNRATE AR(1) residuals on lagged standardized NFCI gives $\hat{\gamma}_A = 0.091$. We use the symmetric calibration $\gamma_Y = \gamma_A$ in this coverage experiment and the interval-width decomposition below.

Figure 2 shows coverage under this empirically calibrated confounding. On the linear DGP (top), OLS intervals achieve 2–8% coverage, UCB reaches 84–94%, ours (cal) reaches 82–92%, and ours (full) achieves 100% across all 12 horizons. With the importance-weighted conformal calibration held fixed, adding c_h raises coverage by 8–18 percentage points. On the nonlinear DGP (bottom), UCB and ours (cal) achieve 22–70% coverage, whereas ours (full) reaches 74–100% and

[11] Each replication draws a fresh (U, ξ) realization, and the causal mean is computed from the closed-form structural equations.

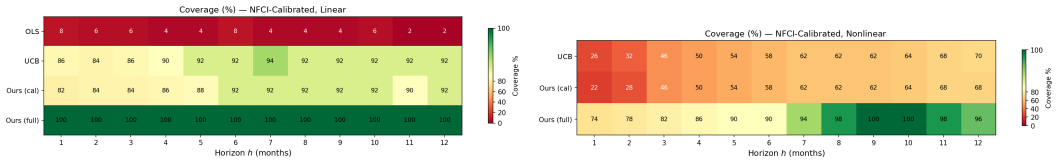


Figure 2: Coverage under NFCI-calibrated confounding. Left: linear DGP; right: nonlinear DGP.

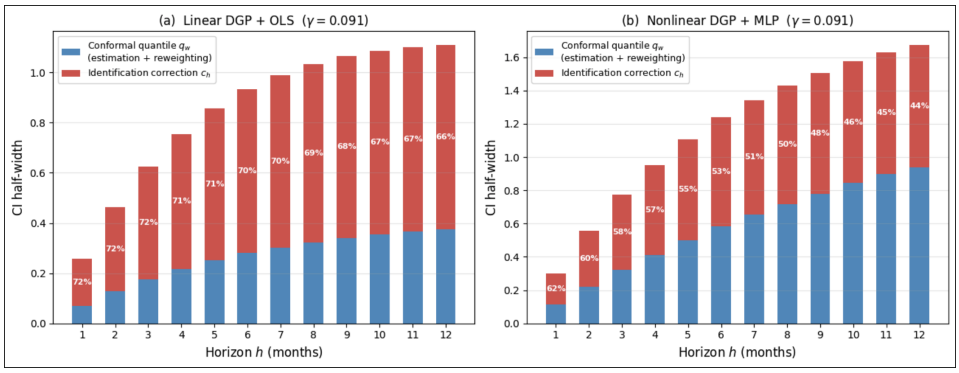


Figure 3: Decomposition of interval half-widths under NFCI-calibrated confounding ($\gamma_A = \gamma_Y = 0.091$). Left: linear DGP with OLS; right: nonlinear DGP with MLP.

attains at least 90% coverage from $h = 5$ onward. At $h = 1$, adding c_h raises coverage from 22% to 74%, a 52-percentage-point gain. Thus, even modest empirically calibrated confounding has a material effect on causal coverage.

5.3 Decomposing Interval Width

Having established the coverage gains under empirically calibrated confounding, we next examine what drives the width of the resulting interval. Figure 3 decomposes each interval half-width into the importance-weighted conformal quantile q_h^w, which captures estimation error and reweighting uncertainty, and the identification correction c_h. Under the linear DGP, c_h accounts for 72% of the interval half-width at $h = 1$ and remains the dominant component at 66% at $h = 12$. Under the nonlinear DGP, its share declines from 62% to 44% as the conformal component grows more rapidly with the horizon, becoming the larger component from $h = 9$ onward. Thus, even at the empirically calibrated confounding strength, identification uncertainty is quantitatively important, while nonlinear long-horizon forecasts also face substantial estimation uncertainty.

6 A Realized Stress Test on Freddie Mac Mortgages

We apply the framework to Freddie Mac mortgages under the realized 2020 unemployment path. Realized delinquency reflects all conditions accompanying unemployment and therefore benchmarks the predictive mean μ_h, not the unobserved intervention mean μ_h^*; the exercise demonstrates practical interval behavior rather than causal coverage. We aggregate active loans into 62 origination-vintage $\times$ FICO cohorts and forecast their 90+DPD rates over January–December 2020.

The model trains on 1999–2014, calibrates on 2015–2019, and recursively rolls out through 2020; weighted validation selects OLS. Rolling backtests produce nominal 90% importance-weighted conformal bands. NFCI-based regressions yield $\hat{\gamma}_A = 0.0715$ and two outcome-side calibrations, $\hat{\gamma}_Y = 0.000316$ and 0.000202. Appendix G.5 and Table 2 report the data construction, learner comparison, weighting procedure, and calibration equations.

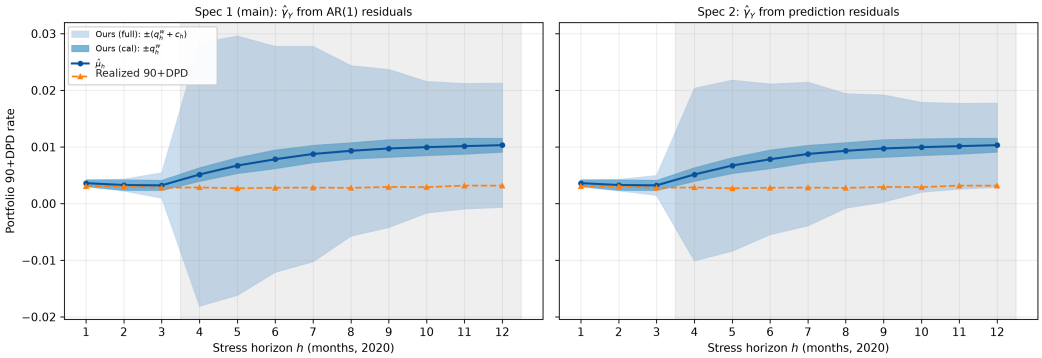


Figure 4: 2020 stress test under two calibrations of γ_Y.

Figure 4 shows that the predictive forecast rises from 0.36% to 1.03%, whereas realized 90+DPD peaks at 0.41%. The conformal-only interval contains 3 of 12 outcomes; adding c_h contains all 12 under both γ_Y calibrations, so the improvement comes entirely from accounting for confounding. Importance weights become highly concentrated after unemployment reaches 14.8% in April, motivating the gray abstention region for horizons 4–12. The alternative calibration and a $g = 3$ dependence check retain 12-of-12 containment, illustrating both the value of the confounding envelope and the need to flag unsupported extrapolation.

7 Conclusion

Regulatory stress testing asks a causal question but is answered by predictive methods, and the gap is quantitatively large: under NFCI-calibrated confounding, standard uncertainty methods cover the causal stress-period mean 0–8% of the time versus a nominal 90%. We have presented a framework that closes this gap through three modular pieces: a closed-form partial identification bound c_h, non-asymptotic error control for the industry-standard recursive rollout, and importance-weighted conformal calibration with abstention diagnostics. The combined interval $\widehat{\mu}_h(a) \pm (q^w_{1-\alpha} + c_h)$ separates estimation from identification uncertainty, each with its own guarantee, giving practitioners and regulators a transparent basis for decision-making. The framework does not eliminate uncertainty—it makes uncertainty visible and interpretable.

AI Use Statement

Generative AI tools were used to assist with grammar and language editing, manuscript revision, LaTeX formatting, and checking the clarity and internal consistency of mathematical derivations and proofs. The authors reviewed every AI-assisted change and independently verified all mathematical claims, calculations, and empirical results. The authors take full responsibility for the final content and presentation of this work.

Reproducibility Statement

The paper states the identifying assumptions and core interval construction in the main text, gives the conformal calibration details, regularity conditions, and proofs in the appendix, and reports the complete semi-synthetic data-generating process, learner specifications, calibration windows, and evaluation protocol in Appendix G. The macroeconomic series used in the experiments is publicly available from FRED.

References

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. 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, 2010.

Tobias Adrian, Nina Boyarchenko, and Domenico Giannone. Vulnerable growth. *American Economic Review*, 109(4):1263–1289, 2019.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Basel Committee on Banking Supervision. Stress testing principles. Technical report, Bank for International Settlements, October 2018. URL https://www.bis.org/bcbs/publ/d450.htm.

Ben S Bernanke, Mark Gertler, and Simon Gilchrist. The financial accelerator in a quantitative business cycle framework. In *Handbook of Macroeconomics*, volume 1, pp. 1341–1393. Elsevier, 1999.

Ioana Bica, Ahmed M Alaa, and Mihaela Van Der Schaar. Time series deconfounder: estimating treatment effects over time in the presence of hidden confounders. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 884–895, 2020.

Board of Governors of the Federal Reserve System. Stress testing under the prior capital frameworks. https://www.federalreserve.gov/supervisionreg/stress-tests-capital-planning.htm, 2024.

Nicole Bohme Carnegie, Masataka Harada, and Jennifer L Hill. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, 9(3):395–420, 2016.

Augusto Cerqua, Marco Letta, and Fiammetta Menchetti. Causal inference and policy evaluation without a control group. *arXiv preprint arXiv:2312.05858*, 2023.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B*, 82(1):39–67, 2020.

Dennis Frauen, Fergus Imrie, Alicia Curth, Valentyn Melnychuk, Stefan Feuerriegel, and Mihaela van der Schaar. A neural framework for generalized causal sensitivity analysis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ikX6D1oM1c.

Gelin Gao, Bud Mishra, and Daniele Ramazzotti. Efficient simulation of financial stress testing scenarios with Suppes-Bayes causal networks. *Procedia Computer Science*, 108:272–284, 2017.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd international conference on artificial intelligence and statistics*, pp. 2281–2290. PMLR, 2019.

Charles F Manski. Nonparametric bounds on treatment effects. *American Economic Review*, 80(2): 319–323, 1990.

Charles F Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Andrea Mauri, Danilo Rubicondo, Nicola Polino, and Tommaso Ferretti. Advancements in bank stress tests: from Bayesian averaging to causal AI. Research Paper Series 53, Iason Consulting, 2023.

Ahmadou Mustapha Fonton Moffo. A machine learning approach in stress testing us bank holding companies. *International Review of Financial Analysis*, 95:103476, 2024.

Valerie A Ramey. Macroeconomic shocks and their propagation. In *Handbook of Macroeconomics*, volume 2, pp. 71–162. Elsevier, 2016.

Paul R Rosenbaum. *Observational Studies*. Springer, 2nd edition, 2002.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel J Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, volume 32, pp. 2530–2540, 2019.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 11559–11569, 2021.

Jiajing Zheng, Jiaxi Wu, Alexander D'Amour, and Alexander Franks. Sensitivity to unobserved confounding in studies with factor-structured outcomes. *Journal of the American Statistical Association*, 119(547):2026–2037, 2024.

A Inference Details

A.1 Weighted Conformal Calibration

Fix a target horizon h. Split the pre-stress data into a training segment (used to fit $\hat{m}$) and a held-out calibration segment. Within the calibration segment, select B rolling origins spaced g periods apart;[12] each origin b must have at least h subsequent periods of observed data. At each origin b, run the rollout h steps forward along the realized macro path and record the prediction error

$$S_b := \left| N^{-1} \sum_i \left(\hat{Y}_{i,h}(b) - Y_{i,h}(b) \right) \right|,$$

where $\hat{Y}_{i,h}(b)$ is the h-step rollout prediction from origin b and $Y_{i,h}(b)$ is the corresponding realized outcome. Each calibration score S_b is computed under the historical macro path from origin b, but the quantity of interest is the prediction error under the stress path $a^S = (a_1^S, \ldots, a_H^S)$. To bridge this distribution shift, assign each score an importance weight

$$w_b(a^S) := \prod_{j=1}^{h} \frac{p(a_j^S \mid J_{b,j-1})}{p(a_{b,j} \mid J_{b,j-1})},$$

where $a_{b,j}$ is the realized macro value at the j-th step from origin b and $J_{b,j}$ is the aggregate macro state at that point. The conditional density $p(\cdot \mid J_{b,j-1})$ is estimated from pre-stress data using any consistent density model for A_{t+1} given the macro state.

The weighted scores $\{(S_b, w_b)\}_{b=1}^{B}$ define an empirical distribution of prediction errors reweighted to approximate stress conditions. The estimation uncertainty band is the weighted conformal quantile

$$q_{1-\alpha}^{w} := \inf \left\{ q : \frac{\sum_b w_b \mathbf{1}\{S_b \le q\} + W_{\max}}{\sum_b w_b + W_{\max}} \ge 1 - \alpha \right\},$$

where $W_{\max} := \max_b w_b$ acts as a conservative correction for the possibility that the stress-period prediction error exceeds all observed historical errors.

A.2 Finite-Sample Guarantee

Standard conformal prediction guarantees exact $1 - \alpha$ coverage under exchangeability, which our setting violates in two ways: calibration scores are temporally dependent, and the stress macro path differs from the historical paths used for calibration. Importance weighting (Tibshirani et al., 2019) and spacing out rolling origins (Barber et al., 2023) address these violations, with two explicit remainder terms: R_{mix} for residual temporal dependence and R_{weight} for extrapolation cost.

Assumption 4 (Decaying temporal dependence)**.** *The joint process $\{(Y_{i,t}, A_t)\}_t$ is β-mixing with mixing coefficients $\beta(k) \le C_\beta e^{-\lambda k}$.*

Assumption 5 (Bounded likelihood ratio)**.** *$w_b(a^S) \le W_{\max} < \infty$ for all b.*

Assumption 6 (Score path-homogeneity)**.** *The distribution of the rollout error score, conditional on $J_{b,j}$, depends on the macro path only through the likelihood ratio.*

Assumption 4 requires that temporal dependence decays exponentially. Assumption 5 ensures the stress path is not so extreme that no historical period receives meaningful weight. Assumption 6 requires prediction-error behavior to remain stable across macro environments after conditioning on the aggregate state.

Proposition 2 (Weighted calibration coverage)**.** *Under Assumptions 4–6:*

$$P\big(\mu_h(a) \in [\hat{\mu}_h(a) \pm q_{1-\alpha}^{w}]\big) \ge 1 - \alpha - R_{\text{mix}} - R_{\text{weight}},$$

where $R_{\text{mix}} = 2(B+1)\beta(g)$ and $R_{\text{weight}} = W_{\max}/(\sum_b w_b + W_{\max})$.

Combining Proposition 2 with Theorem 1 by triangle inequality gives the causal interval.

[12] The gap g balances two considerations: larger g reduces temporal dependence between consecutive scores, but yields fewer calibration points $B \approx \lfloor (T_{\text{cal}} - h)/g \rfloor$.

Theorem 2 (CI for $\mu_h^*(a)$)**.** *Under Assumptions 1–6 and* $|\mu_h^*(a) - \mu_h(a)| \le c_h$*,*

$$\mathrm{CI}_h(a) := \big[\widehat{\mu}_h(a) \pm (q^w_{1-\alpha} + c_h)\big], \qquad \mathbb{P}\big(\mu_h^*(a) \in \mathrm{CI}_h(a)\big) \ge 1 - \alpha - R_{\mathrm{mix}} - R_{\mathrm{weight}}.$$

The interval separates estimation uncertainty, $q^w_{1-\alpha}$, from identification uncertainty, c_h. Coverage is conditional on the partial-R^2 bounds $(\bar{\rho}_A, \bar{\rho}_Y)$ containing the true confounding strengths, as in standard sensitivity analysis (Rosenbaum, 2002; Cinelli & Hazlett, 2020).

B Proofs for Identification

B.1 Identification of the Future Predictive Mean

Proof of Proposition 1. By overlap, $\mu_h^{\mathrm{hist}}(a)$ is a well-defined conditional-mean functional of the observed pre-stress rolling-origin distribution and is therefore identified from that distribution. Path-conditional mean stability then gives

$$\mu_h^{\mathrm{future}}(a) = \mu_h^{\mathrm{hist}}(a).$$

Definition 1 sets $\mu_h(a) \equiv \mu_h^{\mathrm{future}}(a)$, so $\mu_h(a)$ is identified from the pre-stress distribution. This argument neither identifies $\mu_h^*(a)$ nor invokes a particular estimator. □

B.2 Proof of Theorem 1

Proof. Define $b_j := \mathbb{E}[U_{j-1} \mid A_{1:j} = a_{1:j}]$, and let $I^*_{i,t}, I_{i,t}$ denote the state under the causal and predictive distributions.

$h = 1$**.** By Assumption 2(a), $Y_{i,1} = m(I_{i,0}, a_1) + \gamma_Y U_0 + \xi_{i,1}$. Since $I_{i,0}$ is observed and the same under both distributions,

$$\mu_1^*(a) = \mathbb{E}[m(I_{i,0}, a_1)] + \gamma_Y \mathbb{E}[U_0] = \mathbb{E}[m(I_{i,0}, a_1)], \quad \mu_1(a) = \mathbb{E}[m(I_{i,0}, a_1)] + \gamma_Y b_1.$$

The causal mean uses $\mathbb{E}[U_0] = 0$, since the intervention leaves the law of U unchanged; the predictive mean uses $\mathbb{E}[U_0 \mid A_1 = a_1] = b_1$. Thus $\Delta_1(a) := \mu_1^*(a) - \mu_1(a) = -\gamma_Y b_1$, which is finite under the stated conditions.

The state then updates: earlier outcomes enter $I_{i,1}$, so the effect of U_0 on $Y_{i,1}$ can propagate through the state to the next period.

$h = 2$**.** The structural equation is $Y_{i,2} = m(I_{i,1}, a_2) + \gamma_Y U_1 + \xi_{i,2}$, so

$$\mu_2^*(a) = \mathbb{E}[m(I^*_{i,1}, a_2)], \quad \mu_2(a) = \mathbb{E}[m(I_{i,1}, a_2) \mid A_{1:2} = a_{1:2}] + \gamma_Y b_2.$$

The gap has two terms: $\Delta_2(a) = D_2 - \gamma_Y b_2$, where $D_2 := \mathbb{E}[m(I^*_{i,1}, a_2)] - \mathbb{E}[m(I_{i,1}, a_2) \mid A_{1:2} = a_{1:2}]$. Here D_2 captures state propagation. All terms are finite under the stated conditions.

General h**.** Define $D_j := \mathbb{E}[m(I^*_{i,j-1}, a_j)] - \mathbb{E}[m(I_{i,j-1}, a_j) \mid A_{1:j} = a_{1:j}]$, $D_1 = 0$. Applying the same argument to the period-h structural equation gives

$$\mu_h^*(a) = \mathbb{E}[m(I^*_{i,h-1}, a_h)], \quad \mu_h(a) = \mathbb{E}[m(I_{i,h-1}, a_h) \mid A_{1:h} = a_{1:h}] + \gamma_Y b_h.$$

Subtracting yields $\Delta_h(a) = D_h - \gamma_Y b_h$. The effect of earlier confounders through past outcomes is already incorporated in D_h. Since D_h and b_h are finite, $|\Delta_h(a)| = |D_h - \gamma_Y b_h| =: c_h < \infty$. Hence $\mu_h^*(a) \in [\mu_h(a) - c_h, \mu_h(a) + c_h]$. □

B.3 Proof of Corollary 1

Proof. All expectations below are conditional on the fixed forecast-origin state. Iterating the affine outcome equation gives

$$Y_{i,h} = \beta^h Y_{i,0} + \sum_{j=0}^{h-1} \beta^{h-1-j} \left[g(a_{j+1}) + \gamma_Y U_j + \xi_{i,j+1} \right].$$

Under intervention, the macro path is imposed externally and the latent process retains its zero-mean law. Under prediction, conditioning on $A_{1:h} = a_{1:h}$ shifts the conditional means of $U_0, \ldots, U_{h-1}$.

The initial portfolio and the $g(a_{j+1})$ terms are identical under both regimes, while Assumption 2(a) gives zero conditional mean for the outcome shocks. Therefore

$$\mu_h^*(a) - \mu_h(a) = -\gamma_Y \sum_{j=0}^{h-1} \beta^{h-1-j} \mathbb{E}[U_j \mid A_{1:h} = a_{1:h}].$$

Let $\mathbf{U} = (U_0, \ldots, U_{h-1})^\top$. From the structural macro equation and the definition $r_j = a_j - f(a_{j-1})$,

$$\mathbf{r} = \gamma_A \mathbf{U} + \boldsymbol{\eta}.$$

The Gaussian assumptions and Assumption 2(d) imply

$$\mathrm{Cov}(\mathbf{U}, \mathbf{r}) = \gamma_A \Sigma_{UU}, \qquad \mathrm{Var}(\mathbf{r}) = \gamma_A^2 \Sigma_{UU} + \sigma_\eta^2 I.$$

Because $A_0 = a_0$ and f are fixed, conditioning on the macro path is equivalent to conditioning on $\mathbf{r}$. Gaussian conditioning therefore gives

$$\mathbb{E}[\mathbf{U} \mid \mathbf{r}] = \gamma_A \Sigma_{UU} (\gamma_A^2 \Sigma_{UU} + \sigma_\eta^2 I)^{-1} \mathbf{r}.$$

Substituting $\Sigma_{UU} = \sigma_U^2 \widetilde{\Sigma}$ and $\gamma_A^2 \sigma_U^2 = \sigma_\eta^2 \rho_A / (1 - \rho_A)$ yields

$$\mathbb{E}[\mathbf{U} \mid \mathbf{r}] = \frac{\gamma_A \sigma_U^2}{\sigma_\eta^2} \widetilde{\Sigma} \left(\frac{\rho_A}{1 - \rho_A} \widetilde{\Sigma} + I \right)^{-1} \mathbf{r}.$$

Substituting this expression into the persistence-weighted gap and using the one-to-one correspondence between $|\gamma_A|, |\gamma_Y|$ and ρ_A, ρ_Y, the scalar prefactor becomes

$$\frac{|\gamma_Y \gamma_A| \sigma_U^2}{\sigma_\eta^2} = \frac{\sigma_\xi}{\sigma_\eta} \sqrt{\frac{\rho_A \rho_Y}{(1 - \rho_A)(1 - \rho_Y)}}.$$

This proves the expression for $d_h(a; \rho_A, \rho_Y)$. Taking the supremum over $0 \le \rho_A \le \bar{\rho}_A$ and $0 \le \rho_Y \le \bar{\rho}_Y$ gives $c_h(a; \bar{\rho}_A, \bar{\rho}_Y)$; Theorem 1 then gives the stated sensitivity set. □

C Proofs for Estimation

To formalize the tradeoff between two estimators, we need two regularity conditions: the outcome model and state transition are smooth, and the learner achieves uniform one-step error.

Assumption 7 (Smooth state dynamics)**.** *Let ι, ι' denote two states, a the macro value, y, y' realized outcomes, and ι^+, ι'^+ the corresponding next-period states. There exist constants $L_m, L_u \ge 0$ such that:*

- *(i) Outcome model: $|m(\iota, a) - m(\iota', a)| \le L_m \|\iota - \iota'\|$.*
- *(ii) State transition: Let ι^+ and ι'^+ be the next-period states obtained by updating ι with outcome y and ι' with outcome y', respectively, under the same macro value a. Then $\|\iota^+ - \iota'^+\| \le L_u(\|\iota - \iota'\| + |y - y'|)$.*

Assumption 8 (One-step excess risk)**.** *With probability $\ge 1 - \delta_n$: $\sup_{\iota, a} |\hat{m}(\iota, a) - m(\iota, a)| \le \epsilon_n$.*

When $I_{i,t} = (Y_{i,t}, Y_{i,t-1}, \ldots)$ is a finite lag vector, the update shifts entries by one, so Assumption 7(ii) holds with $L_u = 1$. The *amplification rate* $\rho := L_u(1 + L_m)$ governs how one-step errors compound: if $\rho < 1$, errors shrink with each step; if $\rho > 1$, they grow exponentially. The following theorem makes this precise.

Theorem 3 (Recursive estimation error)**.** *Under Assumptions 7–8, with $\rho := L_u(1 + L_m)$ and $\Gamma_h := \sum_{j=0}^{h-1} \rho^j$,*

$$|\widehat{\mu}_h(a) - \mu_h(a)| \le \underbrace{\epsilon_n \Gamma_h}_{\textit{learner}} + \underbrace{|b_h(a)|}_{\textit{mean-state bias}} + \underbrace{O_p(N^{-1/2})}_{\textit{sampling}},$$

where $b_h(a)$ is zero for affine m.

Proof in Appendix C.1. The first term compounds one-step errors through the rollout; $\Gamma_h \to 1/(1-\rho)$ when $\rho < 1$ and grows exponentially when $\rho > 1$. The second arises because the rollout feeds means back into m rather than averaging over the stochastic outcome distribution. The third is cross-sectional noise. On the other side, the direct estimator avoids this compounding entirely.

Corollary 2 (Direct estimator error)**.** *If* $\sup_{\iota,a_{1:h}} |\widehat{m}_h(\iota, a_{1:h}) - m_h(\iota, a_{1:h})| \le \epsilon^{\text{dir}}_{n,h}$ *with prob.* $\ge 1-\delta_n$*, then* $|\widehat{\mu}_h(a) - \mu_h(a)| \le \epsilon^{\text{dir}}_{n,h} + O_p(N^{-1/2})$.

The amplification rate ρ provides a practical rule. When $\rho < 1$, $\Gamma_h \to 1/(1-\rho)$ stabilizes and recursive is reliable. When $\rho > 1$, switch to direct beyond $h^* := \min\{h : \epsilon_n \Gamma_h > \epsilon^{\text{dir}}_{n,h}\}$. When $\rho \approx 1$, compare empirically. In practice ρ need not be computed analytically—it suffices to observe whether recursive rollout errors grow or stabilize with horizon.

C.1 Proof of Theorem 3

Proof. Define the *true-model rollout* $\bar{Y}_{i,j}$: starting from $I_{i,0}$, iterate $\bar{Y}_{i,j} = m(\bar{I}_{i,j-1}, a_j)$ and $\bar{I}_{i,j}$ via the same update but with $\bar{Y}_{i,j}$ in place of the true outcome. By definition $\widehat{\mu}_h(a) = N^{-1}\sum_i \widehat{Y}_{i,h}$; adding/subtracting $\bar{Y}_{i,h}$ and $\mathbb{E}[\bar{Y}_{i,h}]$:

$$\widehat{\mu}_h(a) - \mu_h(a) = \underbrace{N^{-1}\textstyle\sum_i(\widehat{Y}_{i,h} - \bar{Y}_{i,h})}_{\text{(A)}} + \underbrace{N^{-1}\textstyle\sum_i \bar{Y}_{i,h} - \mathbb{E}[\bar{Y}_{i,h}]}_{\text{(B)}} + \underbrace{\mathbb{E}[\bar{Y}_{i,h}] - \mu_h(a)}_{\text{(C)}}.$$

(A) Learner error. At initialization, $\widehat{I}_{i,0} = \bar{I}_{i,0} = I_{i,0}$. At step j:

$$\begin{aligned}
|\widehat{Y}_{i,j} - \bar{Y}_{i,j}| &= |\widehat{m}(\widehat{I}_{i,j-1}, a_j) - m(\bar{I}_{i,j-1}, a_j)| \\
&\le \underbrace{|\widehat{m}(\widehat{I}_{i,j-1}, a_j) - m(\widehat{I}_{i,j-1}, a_j)|}_{\le \epsilon_n} + \underbrace{|m(\widehat{I}_{i,j-1}, a_j) - m(\bar{I}_{i,j-1}, a_j)|}_{\le L_m \|\widehat{I}_{i,j-1} - \bar{I}_{i,j-1}\|} \\
&\le \epsilon_n + L_m \|\widehat{I}_{i,j-1} - \bar{I}_{i,j-1}\|,
\end{aligned}$$

by Assumptions 8 and 7(i). The state error updates by Assumption 7(ii):

$$\|\widehat{I}_{i,j} - \bar{I}_{i,j}\| \le L_u(\|\widehat{I}_{i,j-1} - \bar{I}_{i,j-1}\| + |\widehat{Y}_{i,j} - \bar{Y}_{i,j}|) \le \rho\|\widehat{I}_{i,j-1} - \bar{I}_{i,j-1}\| + L_u\epsilon_n.$$

Unrolling from $\|\widehat{I}_{i,0} - \bar{I}_{i,0}\| = 0$: $\|\widehat{I}_{i,j} - \bar{I}_{i,j}\| \le L_u\epsilon_n\Gamma_j$. Substituting Γ_{h-1} back:

$$|\widehat{Y}_{i,h} - \bar{Y}_{i,h}| \le \epsilon_n + L_m L_u \epsilon_n \Gamma_{h-1} \le \epsilon_n + \rho\epsilon_n\Gamma_{h-1} = \epsilon_n(1 + \rho\Gamma_{h-1}) = \epsilon_n\Gamma_h,$$

using $L_m L_u \le \rho$ and $\Gamma_h = 1 + \rho\Gamma_{h-1}$. All steps use the single event $\{\sup|\widehat{m} - m| \le \epsilon_n\}$; no horizon union bound needed.

(B) Sampling noise. $\bar{Y}_{i,h}$ is a deterministic function of $I_{i,0}$, which varies across units. By the LLN, (B) $= O_p(N^{-1/2})$.

(C) Mean-state bias. The true-model rollout feeds $m(\bar{I}_{i,j-1}, a_j)$ back into the state rather than averaging over the stochastic outcome distribution. (C) $= b_h(a)$ vanishes when m is affine in the state (Jensen's gap is zero).

Combining: $|\widehat{\mu}_h(a) - \mu_h(a)| \le \epsilon_n\Gamma_h + |b_h(a)| + O_p(N^{-1/2})$. □

D Proofs for Inference

D.1 Proof of Proposition 2

Proof. The event $\mu_h(a) \in [\widehat{\mu}_h(a) \pm q^w_{1-\alpha}]$ is equivalent to $S_{B+1} \le q^w_{1-\alpha}$ with $S_{B+1} := |\widehat{\mu}_h(a) - \mu_h(a)|$. We bound $\mathbb{P}(S_{B+1} > q^w_{1-\alpha})$.

Step 1 (exchangeable scores). Assume $S_1, \ldots, S_B, S_{B+1}$ are exchangeable. Assumption 6 ensures the calibration-to-stress shift is fully captured by w_b; by Theorem 2 of Tibshirani et al. (2019), $\mathbb{P}(S_{B+1} > q^w_{1-\alpha}) \le \alpha + R_{\text{weight}}$ with $R_{\text{weight}} = W_{\max}/(\sum_b w_b + W_{\max})$.

Step 2 (non-exchangeable scores). The scores come from rolling origins g periods apart, so they are temporally dependent. By Berbee's coupling lemma, for a β-mixing process there exists an exchangeable sequence $S'_1, \ldots, S'_{B+1}$ with $\mathbb{P}(S_b \neq S'_b) \leq \beta(g)$ per b. By Step 1 applied to S': $\mathbb{P}(S'_{B+1} > q'^{w}_{1-\alpha}) \leq \alpha + R_{\text{weight}}$. By union bound, $\mathbb{P}(\exists b : S_b \neq S'_b) \leq (B+1)\beta(g)$. On the agreement event, $q^w_{1-\alpha} = q'^{w}_{1-\alpha}$ and Step 1 applies directly. Therefore:

$$\mathbb{P}(S_{B+1} > q^w_{1-\alpha}) \leq (\alpha + R_{\text{weight}}) + (B+1)\beta(g) \leq \alpha + R_{\text{weight}} + R_{\text{mix}},$$

with $R_{\text{mix}} = 2(B+1)\beta(g)$ (the factor 2 absorbs the Berbee-based constant). Taking complements gives the claim. □

D.2 Proof of Theorem 2 and Contrast Extension

Proof of Theorem 2. By Proposition 2, on an event of probability $\geq 1 - \alpha - R_{\text{mix}} - R_{\text{weight}}$, $|\widehat{\mu}_h(a) - \mu_h(a)| \leq q^w_{1-\alpha}$. By the maintained condition, $|\mu^*_h(a) - \mu_h(a)| \leq c_h$. Triangle inequality: $|\mu^*_h(a) - \widehat{\mu}_h(a)| \leq |\mu^*_h(a) - \mu_h(a)| + |\mu_h(a) - \widehat{\mu}_h(a)| \leq c_h + q^w_{1-\alpha}$. □

Corollary 3 (CI for the path contrast)**.** *Let* $\widehat{\tau}_h := \widehat{\mu}_h(a^S) - \widehat{\mu}_h(a^B)$ *and* $c_h := \max\{c_h(a^S), c_h(a^B)\}$. *Then*

$$\mathbb{P}\big(\tau^*_h \in [\widehat{\tau}_h \pm (2q^w_{1-\alpha} + 2c_h)]\big) \geq 1 - 2\alpha - 2R_{\text{mix}} - 2R_{\text{weight}}.$$

Proof. Apply Theorem 2 to a^S and a^B separately; by union bound, both CIs hold with prob. $\geq 1 - 2\alpha - 2R_{\text{mix}} - 2R_{\text{weight}}$. On this event, $|\tau^*_h - \widehat{\tau}_h| \leq |\mu^*_h(a^S) - \widehat{\mu}_h(a^S)| + |\mu^*_h(a^B) - \widehat{\mu}_h(a^B)| \leq 2(q^w_{1-\alpha} + c_h)$. □

E Algorithms

Algorithm 1 Recursive Rollout Estimator

Require: Learned predictor $\widehat{m}$, observed initial state $I_{i,0}$, macro path $a = (a_1, \ldots, a_H)$

1: Initialize $\widehat{I}_{i,0} \leftarrow I_{i,0}$
2: **for** $j = 1, \ldots, H$ **do**
3: $\quad \widehat{Y}_{i,j} \leftarrow \widehat{m}(\widehat{I}_{i,j-1}, a_j)$ {one-step prediction}
4: $\quad \widehat{I}_{i,j} \leftarrow$ update state by replacing true outcome with $\widehat{Y}_{i,j}$ {feed back}
5: **end for**
6: **return** $\widehat{\mu}_h(a) := N^{-1} \sum_{i=1}^N \widehat{Y}_{i,h}$

F Practical Diagnostics

(i) **Rolling backtests**: one-step and multi-step prediction errors via panel cross-validation.
(ii) **Placebo tests**: estimated contrasts at fake t'_0 should be near zero (tests stationarity).
(iii) **Extrapolation diagnostics**: report B_{eff} and R_{weight} from Proposition 2.
(iv) **Local amplification**: perturb $\widehat{I}_{i,t}$ along rollout trajectories, measure $|\Delta\widehat{Y}_{i,t+h}|/\|\delta\|$ (local ρ). If > 1, flag expanding-system risk.
(v) **Mean-state bias check**: compare the rollout against Monte Carlo g-formula (sample innovations from fitted residuals). Large discrepancy ⇒ prefer direct.
(vi) **Covariate screening**: exclude variables affected by the macro path unless jointly modeled.

G Experiment Details

G.1 DGP Parameters

Table 1 reports the complete laws of motion and parameter values used in the semi-synthetic experiments.

Table 1: Simulation DGP and parameter values.

Laws of motion

Outcome: $Y_{i,t+1} = m(Y_{i,t}, X_i, A^{\text{obs}}_{t+1}) + \gamma_Y U_t + \xi_{i,t+1}$, $\xi_{i,t+1} \sim \mathcal{N}(0, \sigma^2_\xi)$

Confounder: $U_{t+1} = \phi_U U_t + \nu_{t+1}$, $\nu_{t+1} \sim \mathcal{N}(0, \sigma^2_\nu)$

Observed macro: $A^{\text{obs}}_{t+1} = \text{UNRATE}_{t+1} + \gamma_A U_t$

UNRATE: $\text{UNRATE}_{t+1} = \hat{\mu}_A + \hat{\phi}_A(\text{UNRATE}_t - \hat{\mu}_A) + \eta_{t+1}$

$\eta_{t+1} \sim \mathcal{N}(0, \hat{\sigma}^2_\eta)$; fitted on 1990–2019 data

Symbol	Value	Description	Symbol	Value	Description
α	1.0	Intercept	ϕ_U	0.85	Confounder persistence
β_1	0.85	Outcome persistence	σ_ν	0.5	Confounder innov. std
β_2	0.15	Macro sensitivity	(γ_A, γ_Y)	varies	By experiment
β_3	0.3	Cross-sect. covariate			
β_4	0.02	Quadratic (nonlin.)	$\hat{\phi}_A$	0.995	UNRATE AR coeff.
β_5	0.12	Threshold (nonlin.)	$\hat{\mu}_A$	5.84	Mean unemp. rate (%)
$\bar{Y}$	18.0	Distress thresh. (nonlin.)	$\hat{\sigma}_\eta$	0.154	UNRATE innov. std
σ_ξ	0.15	Idiosyncratic noise	$\hat{\sigma}_A$	1.59	Marginal std (%)
N	3,000	Cross-sect. units	a^S	9.03	Stress: $\hat{\mu}_A + 2\hat{\sigma}_A$ (%)
H	12	Horizon (months)			

G.2 Temporal Structure and Learners

The 420-month sample is divided into three non-overlapping windows: **Training** (1990–2014, $t = 1, \ldots, 300$): learner fits $\widehat{m}$; includes 2001 and 2008 recessions. **Calibration** (2015–2019, $t = 301, \ldots, 360$): rolling-origin conformal scores; no parameter updates. **Stress horizon** (2020 onward, $t = 361, \ldots, 372$): rollout under $a^S_{1:H}$; coverage evaluated vs. $\mu^*_h(a^S)$.

Linear DGP estimated by OLS; nonlinear by two-layer MLP (64–32 units, tanh activation). Both learners observe $(Y_{i,t}, X_i, A^{\text{obs}}_t)$ but not U_t.

G.3 NFCI Calibration Details

The Chicago Fed NFCI is a weekly index of U.S. financial conditions (stress is positive). We aggregate to monthly (end-of-month) over 1990–2019. An AR(1) fit yields $\hat{\phi}_U = 0.973$, $\hat{\sigma}_\nu = 0.124$. Regressing $\hat{\eta}^A_t$ (UNRATE AR(1) residuals) on lagged NFCI (standardized) gives $\hat{\gamma}_A = 0.091$ with robust SE. Setting $\gamma_Y = \hat{\gamma}_A$ yields the symmetric conservative calibration used in §5.2.

G.4 Coverage Evaluation Protocol

For each $(\gamma, \text{method}, h)$ cell in the coverage heatmaps: 50 replications, each generating a fresh (U, ξ) realization, fitting the learner on 1990–2014, computing q^w_h on 2015–2019, rolling out on the flat stress path $a^S_{1:12} = 9.03\%$, and checking whether $\mu^*_h(a^S)$—computed from the closed-form structural equations—lies within the reported CI. Coverage is the fraction of replications in which it does. Nominal coverage is 90%.

G.5 Freddie Mac Application Details

We use Freddie Mac's Single-Family Loan-Level Dataset, covering approximately 56 million mortgages originated since 1999.[13] We retain loans active in December 2019 and originated during 1999–2019, aggregate them by origination vintage and FICO bucket, and retain cohorts with at least 100 active loans. The outcome is the cohort share of active loans that are 90 or more days past due. Portfolio weights are fixed at December 2019 unpaid principal balance; the realized benchmark excludes assistance- or disaster-flagged loan-months from the numerator but not the denominator.

The one-month predictor uses lagged cohort delinquency, forecast-month unemployment, cohort age, and FICO indicators. Active-loan-weighted rolling validation compares OLS, LASSO, random forest, LightGBM, and a multilayer perceptron, selecting OLS. For each horizon, rolling origins

[13] See `https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset`.

Table 2: Freddie Mac panel, prediction task, and empirical calibration.

Empirical specification

Panel/outcome: origination-vintage × FICO cohort-months; $Y_{i,t+1}$ is the cohort 90+DPD rate.
Predictor: $I_{i,t} = (Y_{i,t}, \mathrm{Age}_{i,t+1}, \mathrm{FICO}_i)$, $A_{t+1} = \mathrm{UNRATE}_{t+1}$; learner: OLS selected by weighted rolling-validation RMSE.
Windows/weights: training 1999–2014; calibration 2015–2019; stress test 2020; December 2019 fixed UPB weights.

Calibration equations

NFCI: $Z_{t+1} = \phi_U Z_t + \nu_{t+1}$; UNRATE: $\mathrm{UNRATE}_{t+1} = c_A + \phi_A \mathrm{UNRATE}_t + \eta^A_{t+1}$.
Loadings: $\widehat{\eta}^A_{t+1} = \gamma_A Z_t + e^A_{t+1}$ and $\widehat{\eta}^Y_{t+1} = \gamma_Y Z_t + e^Y_{t+1}$, where $\widehat{\eta}^Y_{t+1}$ is the portfolio-delinquency AR(1) residual.

Symbol	Value	Description	Symbol	Value	Description
N	62	Cohorts	$\hat{\phi}_U$	0.972	NFCI persistence
Minimum size	100	Active loans per cohort	$\hat{\sigma}_\nu$	0.233	NFCI innovation std.
H	12	Stress horizon (months)	$\hat{\phi}_A$	0.9980	UNRATE persistence for c_h
α	0.10	Conformal miscoverage	$\hat{\sigma}_\eta$	0.1538	UNRATE innovation std.
g	1	Rolling-origin gap	$\hat{\gamma}_A$	0.0715	Macro loading (HC1 SE 0.0064)
B	48–59	Origins across horizons	$\hat{\gamma}_Y$	0.000316	Outcome loading (HC1 SE 0.000032)
$\hat{\beta}$	0.9996	Lagged-outcome coefficient	$\hat{\gamma}_Y^{\mathrm{alt}}$	0.000202	Prediction-residual calibration

Notes: $g = 3$ is a sensitivity check. Importance weights use a 1999–2014 UNRATE AR(1) ($\hat{\phi}_A = 0.9932$, $\hat{\sigma}_\eta = 0.1670$); the reported UNRATE parameters use 1990–2019 for calibrating c_h.

over 2015–2019 produce 48–59 absolute forecast errors. The main specification uses $g = 1$ and their importance-weighted 90th percentile; $g = 3$ is a temporal-dependence check. Standardized month-end NFCI over 1990–2019 calibrates γ_A, while portfolio-delinquency and prediction residuals provide the two calibrations of γ_Y.

Table 2 reports the panel construction, prediction task, and calibration parameters for the realized mortgage stress test in Section 6.